J Mol Model (2012) 18:1679–1689
DOI 10.1007/s00894-011-1199-9

ORIGINAL PAPER

Conformational flexibility, binding energy, role of salt bridge and alanine-mutagenesis for c-Abl kinase complex

Kshatresh Dutta Dubey · Rajendra Prasad Ojha

Received: 26 November 2010 / Accepted: 20 July 2011 / Published online: 3 August 2011

Abstract Abl kinase plays a decisive role in the mechanism of the most fatal human pathogen chronic mylogenous leukemia (CML). Here, we have carried out a comprehensive study about the conformational flexibility, role of salt bridge and the protein- ligand interaction for this kinase with its well-known inhibitor, Imatinib. We have performed molecular dynamics simulations for conformational behavior, investigated the salt bridges and calculated the binding free energy of Imatinib with MM-PB/SA method for Abl kinase complex. We also explored the role of salt-bridge in the kinase complex and its effect on binding activity of inhibitors. Furthermore, to investigate the importance of those residues which form salt bridges, we mutated them by Alanine with the help of Alanine scanning program. We noticed significant variations in total free energy of Imatinib in all possible mutations. The binding free energy of ligand for kinase receptor was analyzed by molecular mechanics Poission Boltzmann surface area (MM-PB/SA) method. These results suggest that conserved glutamic acid and lysine are necessary for stability of complex.

Keywords Alanine-mutagenesis · MM-PBSA method · Molecular Dynamics Simulations · Protein-hydration · Salt-bridge

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1199-9) contains supplementary material, which is available to authorized users.

K. D. Dubey (✉) · R. P. Ojha (✉)
Biophysics Unit, Department of Physics,
DDU Gorakhpur University,
Gorakhpur 273 009, India
e-mail: kshatresh@gmail.com
e-mail: rp_ojha@yahoo.com

Introduction

The calculation of binding affinity of a drug ligand with a protein receptor is a major preoccupation of early stage drug discovery. Even though this is not an easy task, yet, there are some theories, particularly Linear Interaction Energy (LIE) and molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) which are commonly used to examine this binding affinity. These are free energy pathway methods based on the recognition that the free energy of binding is the change in free energy when one protein and one ligand react to form a complex. Here we report the energetics and conformation of an enzymatic system c-Abl protein kinase through molecular dynamics and the MM-PBSA calculation. Many groups are engaged in cramming the molecular structure of this kinase during the last few decades [1–4]. Beside the crystallographic study there are many workers who have explained the mechanism of these tyrosine kinases [5–10]. Akin to other protein kinases, Abl-kinase is activated by toggling on and off states of their complex structures. Leukemia occurs due to over activity of this kinase, so a drug that obstructs the activity of this enzyme can be treated as a potential drug target for cancer. Fortunately, crystal structures of some protein kinases have been determined and numerous small molecule inhibitors of such kinases have been also developed [1, 11–13]. Among such drugs, Imatinib is ATP competitor inhibitors which binds between the cleft of N and C terminals of the kinase and blocks the ATP binding. It interacts with a number of hydrophobic residues, including several aromatic amino acids, which are often important in interaction with polar surface of inhibitors. It has been approved for the use in the patient affected by the CML. The crystallographic structure [14] of Imatinib with c-Abl protein kinase suggests that due to binding of

Imatinib, activation loop is folded inward which prevents aspartate from ligating with magnesium.

The molecular biology reveals that E286, T315, M318, I360, D381, F382, G383 are conserved in many tyrosine kinases and are associated in binding with the ligand or forming the channels for its binding. Furthermore, a good number of highly conserved charged residues, Lys, Arg, Glu, and Asp, capable of forming salt-bridge, are also observed. A similar salt-bridge has been observed in the structural analysis of a kinase [15]. There are certain amino acids for example Glu/Asp:Lys/Arg that are being conserved in diverse sequences of all known kinases whose structural as well as functional roles in salt bridge has yet to be explicated. The present study describes the characterization of number of salt bridges for the first time using computational and molecular modeling technique. These finding sheds light on molecular basis of the conservation of Glu/Asp:Lys/Arg salt bridges in kinases. In addition to this, since protein hydration is very important for three dimensional structures and activity [16–19]; all experiments are being carried out in water in order to understand its behavior on the stability of the kinase-ligand complex.

Methods and materials

We have used Amber 10 for molecular dynamics and MM-PBSA calculations to study the behavior of the human c-Abl protein kinase complexed with Imatinib. We have performed molecular dynamics simulation of kinase with and without Imatinib to investigate the interaction of ligand with target receptor.

Molecular dynamics

The starting structure of the inactivated human kinase bounded with Imatinib was taken from protein data bank with PDB id 2HYY [14]. The missing heavy atoms and residue Glu 275 in the crystallographic structure were corrected with the Leap module of the amber package [20]. Partial atomic charges of drug Imatinib was calculated using restrained electrostatic potential (RESP) procedure [21, 22]. Ab initio calculations for ligand were carried out using Gaussian 03 program at the HF/6-31 G* level of theory. The kinase structure was initially a tetramer from which a monomer was taken for studies. We adopted a different numbering of the residue with respect to the crystallographic structure [14], we took the residue 235 to 498 of human kinase. Hence in our present study the residue one corresponds to TRP235 and residue 264 for GLN498 of the crystallographic structure. The initial parameters for the Imatinib were prepared by antechamber module of Amber 10 package and then by the leap module during the parameter preparation for the complete system. Prior to the MD simulation the structure was subjected to minimization in order to remove the steric clashes. The complete system was neutralized with Na+ion and immersed into the truncated octahedral shell of TIP3P [23] water of dimension extending up to 75 Å. The system was then gently annealed from 10 to 300 K over a period of 200 ps and then maintained in the isothermal–isobaric ensemble (NPT) and thereafter at a target temperature of 300 K and target pressure of 1 bar using Langevin thermostat [24] and Barendsen barostat [25] with collision frequency 2 ps and pressure relaxation time as 1 ps. The hydrogen bonds were constrained using SHAKE [26]. After this the dynamics was continued up to 1.5 ns to equilibration. For the analysis of the system, molecular mechanical production phase was initiated and again continued for another 14 ns maintaining the same parameters. The structures in the trajectories were collected at every 10 ps intervals. All analysis of trajectories was done with the Ptraj module of Amber10. VMD 1.6.7 [27], Chimera-1.3 [28] and Maestro [29] graphical programs were used for the visualization purpose.

Free energy simulation

The free energy analysis of the production trajectories employ a single trajectory MM-PBSA[30, 31] method combined with a determination of the change in the configurational entropy using the harmonic approximation of normal mode analysis. The principles of these methods are well established and have been discussed by many workers [32–35]. At this juncture we describe the specific parameters employed in our approach. The free energy difference of binding is composed of the following terms:

$$\Delta G_{bind} = \Delta G_{ele} + \Delta G_{vdw} + \Delta G_{pol} + \Delta G_{nonpol} - T\,\Delta S. \quad (1)$$

Here, the first two components in the right hand side represent the van der Waals and electrostatic components of the gas phase molecular mechanics free energy difference, the third term is the electrostatic polar components of the solvation free energy, and the fourth term is the nonpolar component of the solvation free energy. All terms are calculated using the standard MM-PBSA method implemented in Amber 10. The last term is the contribution from the change in the configurational entropy (T∆S). A detailed explanation of these components can be seen in our recent publication [34]. The change in configurational entropy upon ligand association are estimated by all atom normal mode analysis performed with the Amber NMODE module. Prior to the MMPBSA analysis all water molecules and the sodium ions were stripped from the trajectory. The dielectric constant used for the solute and surrounding solvent was 1 and 80 respectively. During the MMPBSA calculation snapshot were generated between the trajectories of the 4 ns-12 ns with time interval of 100 ps. The calculations for solute entropy were performed with NMODE module of the Amber 10. The structures were minimized in the gas phase using conjugate gradient method for 5000 steps, using a distance dependent dielectric. The frequencies of vibration mode were computed by the normal mode for the minimized structures at 300 K.

Alanine-scanning mutagenesis

The alanine-scanning mutagenesis of protein-protein interfacial residues has generated a large amount of information that allowed the discovery of energetically important determinants of specificity at intermolecular protein interfaces [36–38]. We have used alanine-scanning program implemented in MM-PB/SA method of Amber 10 for the mutating Alanine. We made 14 alanine mutations in residues Asp7Ala, Lys11Ala, Lys13Ala, Glu21Ala, Glu24ala, Lus37Ala, Glu52Ala, Glu82Ala, Glu118Ala, Lys122Ala, Lys144Ala, Asp157Ala and Lys170Ala for the equilibrated Abl complex. We performed standard MM-PB/SA method to calculate the binding free energy of all mutated system separately.

Sietraj calculation

In this method a binding energy function that consists of force field term is supplemented by solvation term. This function is further used to calibrate the solvation model along with the solvation interaction terms in a self-consistent manner. The incentive for this approach was that the solute dielectric constant dependence of calculated hydration gas to water transfer free energy is markedly different from that of the binding free energies and thus model solvation is directly calibrated in the context of the binding free energy calculations. In this calculation high internal dielectric constant and strong van der Waals scaling was used for analysis and presents the parameterization of continuum solvation model, based directly on the experimental data. In this method binding free energy was calculated by -

$$\Delta G_{\mathrm{bind}} = \alpha(E_{\mathrm{vdw}} + E_{\mathrm{coul}} + E_{\mathrm{RF}} + E_{\mathrm{cav}}) + \mathrm{constant(C)}. \tag{2}$$

Here, E_{vdw} and E_{coul} are the intermolecular van der Waals and columbic interaction energy which is calculated using the Amber molecular mechanics force field. E_{RF} is the change in the reaction field energy between the bound and the unbound state that is calculated using the Sietraj (BRIUMM) program [39, 40]. Here E_{cav} is the cavity energy taken to the change in the molecular surface area ΔSA and it is calculated by the formula $E_{cav}=\gamma'\Delta SA$. Where γ' is the surface area coefficient for optimized parameters and it has the value of 0.012894 kcal $^{mol-}$12.α and C are fitting parameters whose values are 0.104758 and - 2.89 kcal mol^{-1}, respectively.

Results and discussion

Conformational analysis

The initial structure prior to simulation of inactive complex formed by kinase with Imatinib, shown in Fig. 1, illustrate that Imatinib binds in the cleft between the N- and C-terminal lobes and the DFG motif (D147-F148-G149) flips out to make a channel beyond the gatekeeper residue Thr81 for benzamide and N-methyl piperazine groups of Imatinib. The interaction of Imatinib with receptor is shown in Fig. 2. The N methyl piperazine group of Imatinib was found to have a strong interaction with the protein via hydrogen bonds with the main-chain carbonyl group of Ile126 and His127. Other hydrogen-bond interactions were found between pyridine N of ligand and the backbone NH of Met84 in the hinge region, the anilino NH of ligand and the side chain of the gatekeeper residue Thr81, the amide NH and the side chain of Glu52 from C-helix, and the amide carbonyl and the backbone NH of Ala146 (which just precedes the highly conserved DFG motif). The four snapshots of the kinase complex from the production trajectory, at 1 ns, 5 ns and finally at the 13 ns are shown in Fig. 1S of supplementary material for the structural illustration. The overall flexibility of the protein can be explained by B-factor calculation from the MD trajectories. The Debye -Waller factor or isotropic temperature factor (B-Factor), plotted as a function of residue number with respect to center of mass as Imatinib, and was obtained from the following equation

$$\mathrm{B_i} = 8/3\left(\pi^2 \Delta \mathrm{r}^2\right) \tag{3}$$

Here, Δr^2 is the mean square fluctuation for the C_α atom of the residue. In a typical B- factor pattern low B factor values show the well structured regions and a high values show the loosely structured loop regions or domains termini. The graph representing the B-factor with respect to residue of the protein is shown in Fig. 3. Here, we notice that residues 15–21, 49–55, 78–84, 147–149 have a smaller B-factor. It is noteworthy that in these regions the regulatory elements (activation loop, DFG motif and alpha helix) are located. The lesser B-factor suggests that these

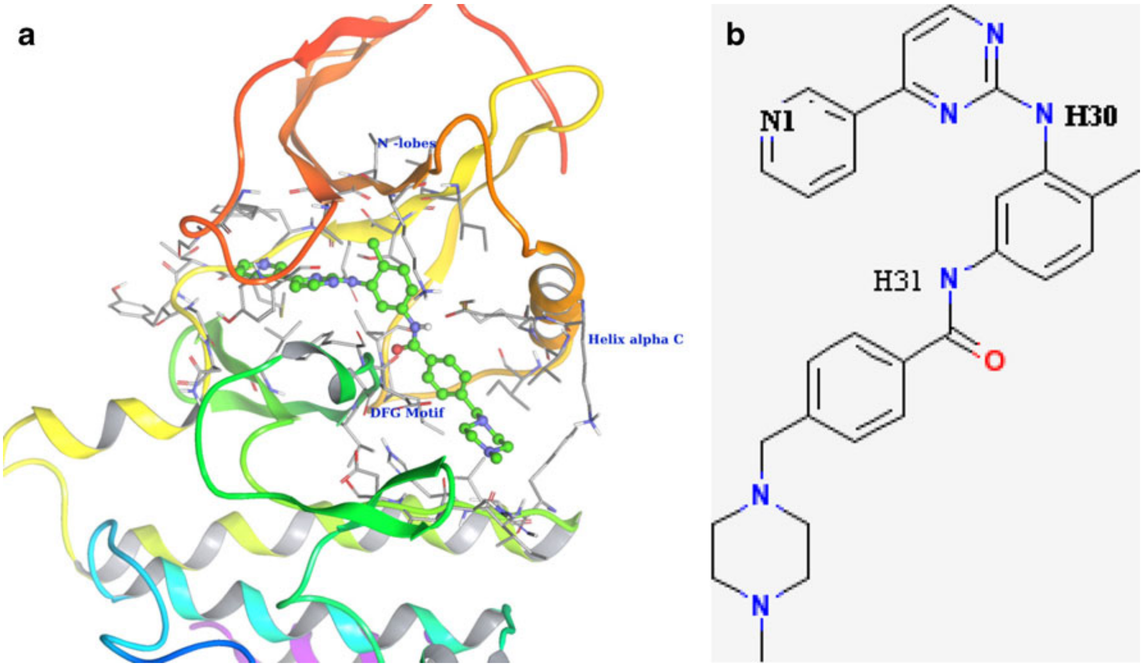
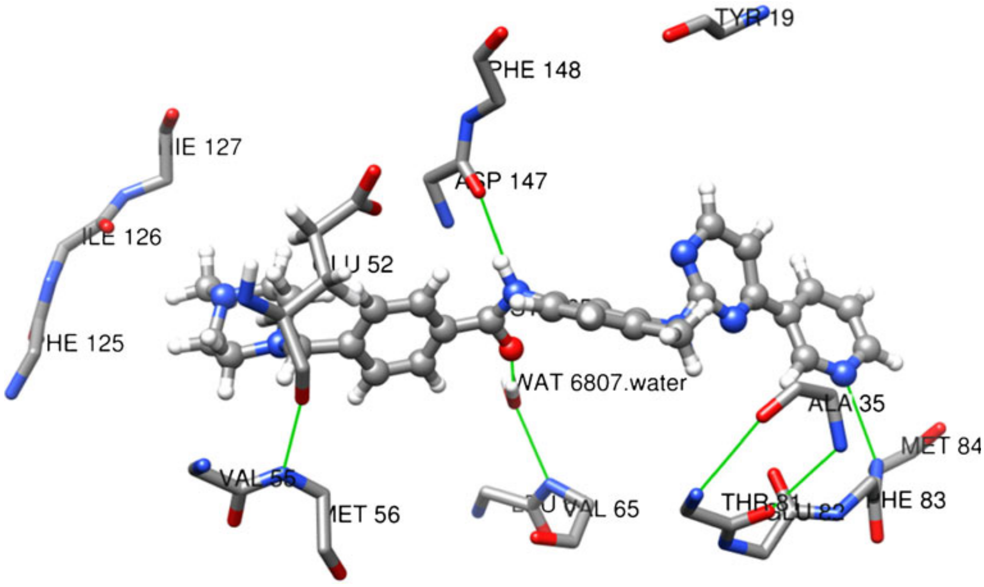


Fig. 1 (**a**) Initial structure of c-Abl protein kinase bounded with Imatinib. (**b**) Two dimensional structure of Imatinib

residues are stabilized due to interaction with some other groups. Apart from the very flexible termini, the most significant fluctuation corresponds to the loop regions. It can be seen that C-terminal chain, which consists of alpha helices, is the most flexible region (for both unbounded protein and bounded protein). Among the N terminal chains, the two alpha helices and the associated loops near the hinge region, show higher B-factor. Due to high flexibility of this region (followed by the activation loop and preceded by DFG motif), it has the capability to move up to 10 Å to interact with ATP binding site. A comparison of the structure of c-abl at various steps during the simulation and with final structure suggests that the RMS fluctuation for the activation loop is up to 14 Å. It slowly moves at each step from initial structure to final structure. This conformational flexibility of activation loop suggests that it has the capability to release and facilitate the nucleotide binding at this site when needed. After monitoring the atomic positions and their positional fluctuations in Fig. 3, it is noteworthy that the atoms and residues involved

Fig. 2 The interaction of Imatinib with backbone atom of Abl kinase. Hydrogen bonds are represented by green lines



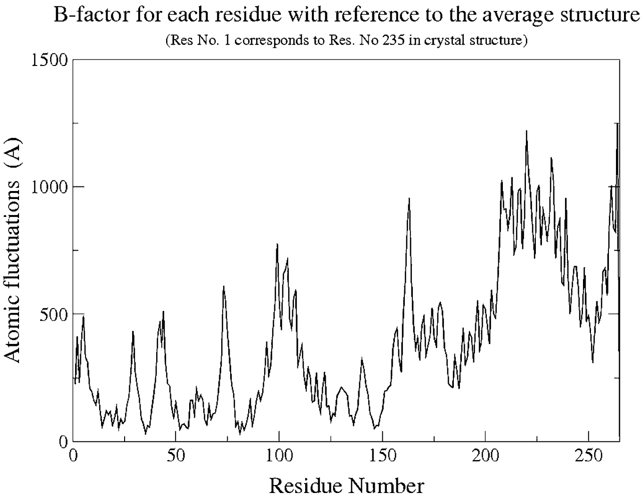


Fig. 3 Atomic fluctuations (B-Factor) during the whole production simulation with respect to average structure

in hydrogen bonding have lower B- factor. The loops are more flexible than other regions in the complex (Fig. 3) and thus, are more difficult to characterize reliably by crystallography. That is why, there are a number of conformational differences between simulated and crystallographic structures for the exposed side chains from these flexible loops, for example residue Glu275, Asp276 and Thr277 are missing in some of the chains, while some of the side chains of Arg239, His246, Lys274, Met278, Lys285, Lys294, Arg307, Glu308, Lys356, Lys357, Arg386, Leu387, Asp391, Glu466, Lys467, Glu470, Gln491, and Glu494 residues are missing from the crystallographic structure of the complex, which are parts of the loops discussed above. The RMSD (root mean square deviation) of the backbone (in bounded and unbounded form) for the 14 ns simulation period is shown in Fig. 4. The RMSD of 1.5 Å was observed for the receptor in the bound state while it was almost 2.25 Å for the unbounded form. The ligand RMSD is 0.7 Å, which is relatively smaller than receptor in unbound and bound form. We see that there is relatively high deviation in the receptor in the absence of Imatinib, which shows the effect of ligand binding with the receptor. It is apparent that more fluctuation was observed for the receptor as well as for the ligand due to local conformational changes during the simulation time 9.1 ns to 10 ns. The change in the receptor conformation suggest the change in the ligand conformation also which is obvious from the graph. To explain above conformational change, hydrogen bonding between the ligand and the regulatory elements have been calculated and shown in Fig. 5. It should be noted that the carbonyl oxygen of Asp147 forms a strong hydrogen bond with average distance of 2.8 Å while the side chain of Glu52 forms slightly weak hydrogen bond with Imatinib. It may occur due to involvement of the O$^{\varepsilon}$ in the salt bridge formation, discussed below. The proton acceptors of these two residues are oriented at an angle of approximately 75^{o} with respect to each other toward the N5 atom of Imatinib. By monitoring the trajectory motion, a variation of 60–90 degree has been observed during the simulation period. This variation in the orientation is governed by the salt bridge formed between the Glu52 and Lys37. Since the Asp147 and Glu52 is interacting with Imatinib through the same amide group which causes a change in the dihedral angles between the aromatic rings. The exchange of hydrogen-bonded structures causes a change in the orientation of the aromatic ring of inhibitor and as a result there is a local conformational variation and fluctuation in the RMSD values. The graph shows the flexibility of these regulatory elements which is in agreement with the earlier results by other techniques [7, 10, 36]. During the simulation, side chain of Thr81 makes a strong hydrogen

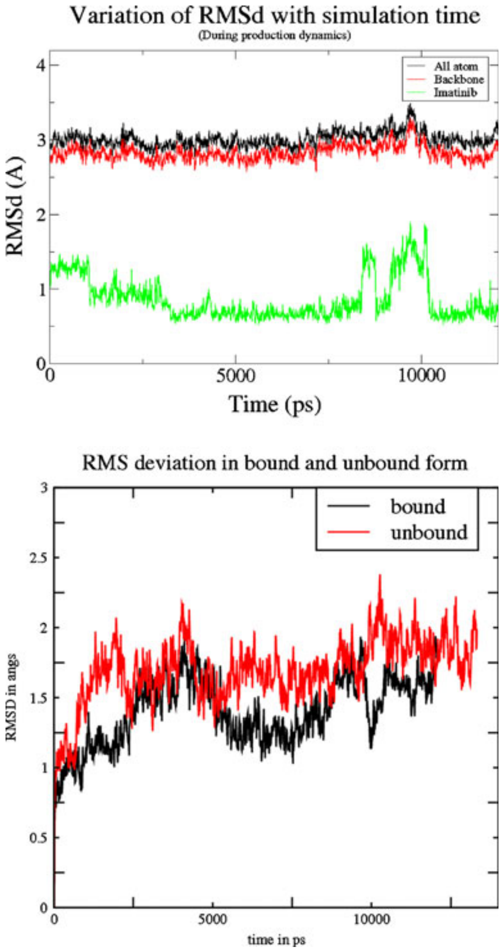


Fig. 4 RMSD of tyrosine kinase complex, backbone and ligand with reference to the average structure during whole production dynamics (Top). Comparison of RMS variation of backbone of c-Abl in bounded and in unbounded form (Down). The RMS variation in later one is plotted with reference to unbounded structure

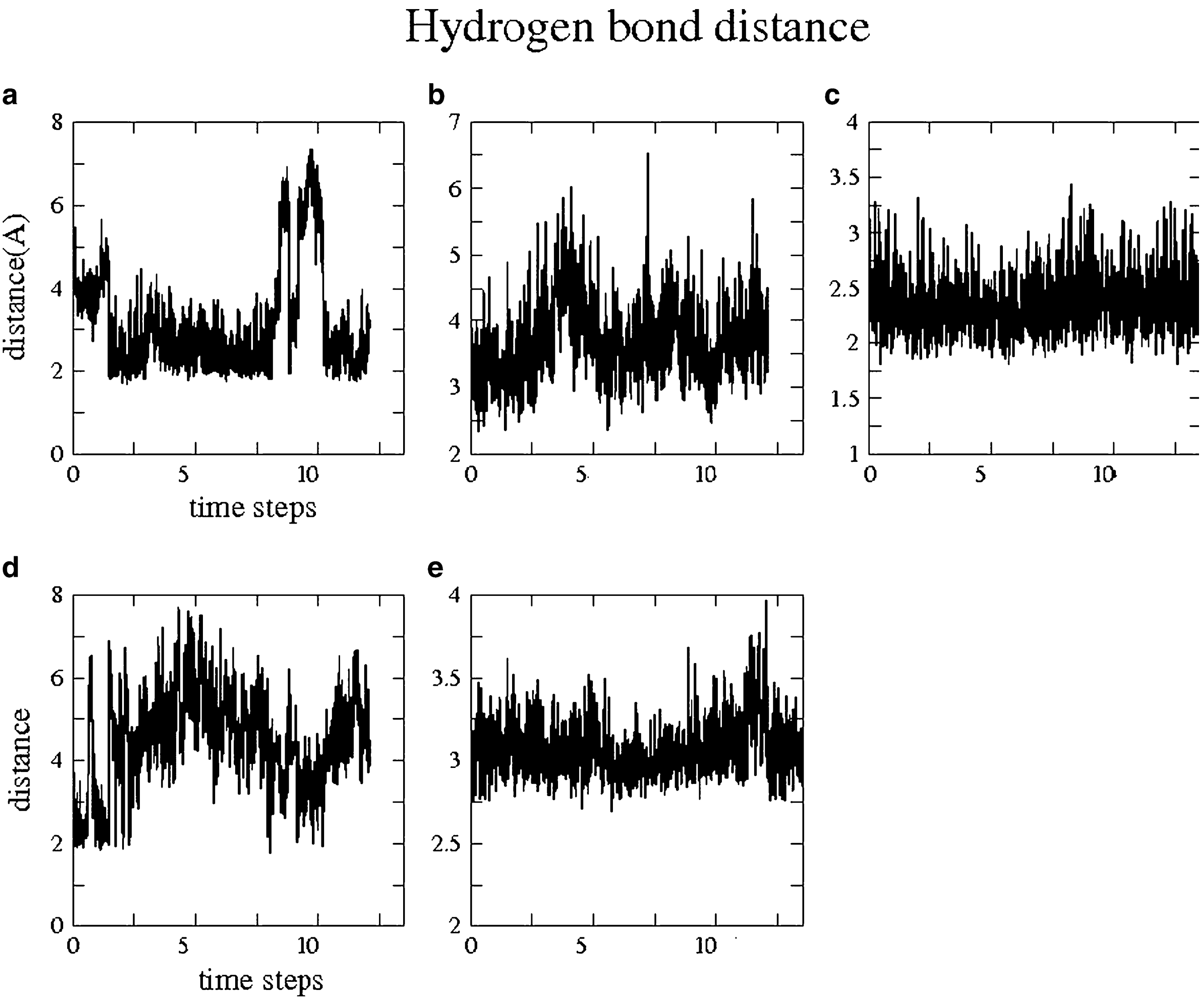


Fig. 5 Hydrogen bond distance of some regulatory elements with Imatinib during whole production dynamics (2–14 ns). Here, time is represented in nano-second (**a**) distance between Asp147@O with STI265@H31 (**b**) Distance between Tyr 19@H with STI265@H (**c**) distance between Thr 81@OG1 with STI265@H30 (**d**) distance between Glu 52@OE2 with STI265@H31 (**e**) distance between Met84@N with STI265@N1

bonding with amide nitrogen of Imatinib (Fig. 5c). Similarly the amide group of Met84 is also forming strong hydrogen bonds with the N1 atom of the Imatinib (Fig. 5e). The side chain of Tyr19 of P-loop is orienting itself to form the close interactions with kinase inhibitor, including an edge-to-face aromatic interaction between Tyr19 and the pyridine and pyrimidine rings of Imatinib. The other contacts are mainly van der Waals interaction, between the inhibitor and the protein, which contain the hydrophobic cleft created by various amino acids. The interacting amino acids are stabilized and show small fluctuations. From Fig. 3, it is clear that interacting residues show relatively low b-factor due to the interactions with the ligand or other residues.

Binding free energy

We used the single complex trajectory protocol, the MM-PBSA method, to predict the binding affinity and calculation of entropic contributions. The results obtained by the MM-PBSA are shown in the Table 1. It is clear from the table that main contribution in the total energy comes from the electrostatic energy for all system (receptor, complex, and ligand), while the difference of the van der Waals energy is more negative with respect to the electrostatic energy that indicates about the good pose of ligand to acquire the cavity of receptor. Strong van der Waals interaction implies a reasonable geometry for binding with kinase receptor. High contributions of polar group indicate

Table 1 Binding free energy components of the c-Abl kinase complex for snapshots extracted from production dynamics[a] . Here, Std stands for standard deviation

Contributions[c]	Complex		Receptor		Ligand		Delta[b]	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
E_{ele}	−6168.34	77.52	−5573.08	77.69	−571.88	3.10	−23.38	3.67
E_{vdw}	−1085.76	25.68	−1046.25	25.42	30.98	3.20	−70.49	3.48
E_{int}	6031.67	42.54	−5912.37	43.69	119.30	7.27	00.00	0.00
E_{gas}	−1222.43	93.84	−706.97	94.97	−421.60	5.83	−93.86	4.32
G_{np}	102.63	1.29	105.33	1.33	6.15	0.04	−8.85	0.16
G_{pol}	−3603.70	67.16	−3632.21	67.38	−29.13	0.85	57.64	4.10
G_{sol}	−3501.07	66.86	−3526.87	67.11	−22.98	0.85	48.79	4.07
G_{bind}	−4723.50	52.69	−4233.84	52.46	−444.59	5.75	−45.07	4.80
G_{cav}							61.64	
TΔS							−39.03	1.83
ΔG_{free}							−06.04	
ΔG_{exp}							−10.37	

[a] All values are given in kcal mol^{-1}

[b] Contributions(complex) –contributions(receptor+ligand)

[c] E_{ele}: Columbic energy; E_{vdw}: van der Waals energy; E_{gas}=E_{ele}+E_{vdw}+E_{int}; G_{np}: non polar salvation free energy; G_{pol}: polar salvation free energy; G_{sol}=G_{np}+G_{pol} ; G_{bind}=E_{gas}+G_{sol}; G_{cav}: cavity energy; TdS : total entropic contributions as determined by normal mode analysis; G_{free}= G_{bind} –TdS.

that inhibition of kinase is mostly supported by the hydrophilic interactions. The absolute free energy comes out to be −06.04 kcal mol^{-1}, which is the indication of the strong binding affinity. Furthermore, the negative binding energy also supports the tight binding of ligand with the receptor. Here, cavity energy is 61.64 kcal mol^{-1} which indicates that the work has to be done in the reorganization of solvent molecules. Since the cavity energy is directly proportional to the accessible surface area, hence its high value indicates larger solvent accessible surface area. The experimental binding energy of Imatinib with the Bcr-Abl kinase [41] is about −10.37 kcal mol^{-1}. The SIE method was calibrated on binding affinity between protein and small molecule ligands in order to test its ability to predict the binding affinity for this particular system, for which the MM-PBSA computation data is available. The results of the Sietraj calculation are shown in Table 2. It is obvious from the table that binding affinity of Imatinib comes about −11.08 kcal mol^{-1} which is close to the experimental value. During the Sietraj calculation we have used same MD trajectories as that was used for MM-PBSA calculations.

Role of salt bridge in kinase activity

Since salt bridges play an important role for the folding of proteins, hence we took account of the salt bridges formed in the simulated structures. The occupancy of the salt-bridge formation has been plotted and shown in Fig. 6. We perceive several salt bridges between positive and negative charged amino acids. Some of these, e.g., Lys-Glu [Lys271-Glu286 or Lys37-Glu52 (present nomenclature as in this text)] salt-bridges have been observed and reported in crystal structures [14]. This salt-bridge is important for maintaining an active kinase conformation and orienting the lysine side chain for interaction with ATP phosphates in the Abl bounded Dasatinib structure (PDB id 2GQQ). A salt-bridge arranged in similar way has also been observed with LCK kinase (PDB id 3LCK) [15]. As shown in Fig. 6, the occupancy of salt-bridge between 52E -37 K and 82E-144 K is more than 80%, while salt- bridge between the residue 21E- 13 K, 118E-122 K and 157D-170 K have occupancy more than 40%. For other salt-bridges the occupancy is smaller. From the atomistic view, these salt-bridges are formed by N^{ε} and N^{n2} of the Arg or N^{ε} of Lys and $O^{\varepsilon1}$ or $O^{\varepsilon2}$ of Glu or $O^{\delta1}$ or $O^{\delta2}$ of Asp. The smaller histograms shown in above figure represent the bridges observed during simulations with lower occupancy. The side chains of these residues are also involved in hydrogen

Table 2 Binding energy components calculated from Sietraj method[a]

E_{ele}: Coulombic energy; E_{vdw}: van der Waals energy; E_{RF}: reaction field energy; ΔG: binding free energy(formula given in material & methods)

[a] All values are given in kcal mol^{-1}

Contribution	Mean	Std
E_{ele}	−70.61	3.53
E_{vdw}	−10.52	1.62
E_{RF}	−15.77	1.90
E_{cav}	−12.83	0.35
ΔG	−11.08	0.44
ΔG_{exp}	−10.37	

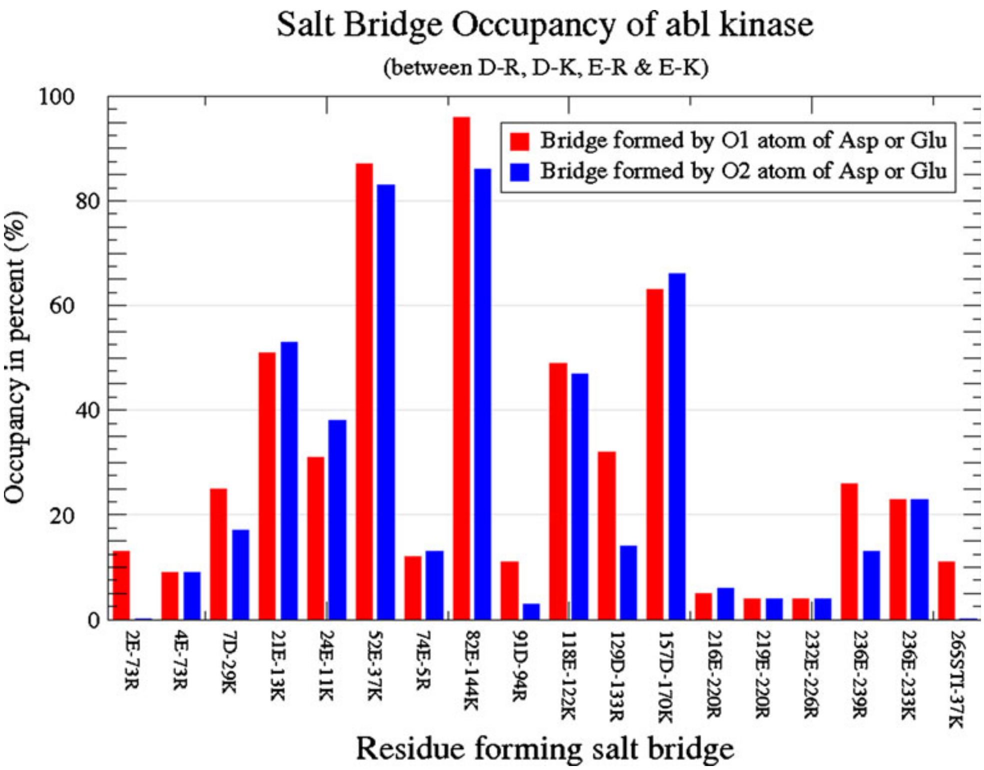


Fig. 6 Percentage occupancy of the hydrogen bonds formed by salt bridges. The bonded amino acids are mentioned below the corresponding histogram

bonding with other residues in the protein molecule. Figure 6 infers one another strong salt bridge between Glu82 and Lys144. Here, we speculate that salt-bridge Glu82:Lys144 brings gatekeeper residue and activation loop together at the time of folding, which may be responsible for the activity of the kinases. The salt-bridges between Asp157:Lys170, which is observed during the dynamics, are part of the activation loop. Here, Asp157 is very close to the Tyr159 which may be a hot spot for the ATP binding (it is phosphorylated in the presence of divalent ions) and Lys170 is at the end of the activation loop and is a small part of alpha helix known as the peptide binding site. Again we speculate that this salt bridge will be helpful in bringing the active Tyr159 at the active binding site during phosphorylation process. It may be responsible to hold the molecule close to reactant for the specific period of reaction needed for this activity. The other salt-bridge having more than 40 % occupancy is Glu21:Lys13. It is to be noted that these two residues are part of the P-loop. As discussed above among Glu52 and Asp147 only one binds tightly with the inhibitor Imatinib. During the simulation it is observed that when Glu52 is bonded with Imatinib, simultaneously it binds with Lys37 to form a salt-bridge. On the other hand, when Asp147 binds with Imatinib then Glu52 does not show any bonding during this period. Therefore, probably, salt-bridge between Glu21:Lys13 is needed to fix the position of P-loop such that Glu52 is in proximity of inhibitor. The absence of this salt-bridge may force the loop to orient in other side, farther to the inhibitor, which may hinder the Glu52 to approach the active site or bonding site with the inhibitor. The P-loop mutation causes the loss of two hydrogen bonds that stabilize the inactive conformation and would therefore tend to shift the equilibrium distribution of the kinase conformational states toward the active conformation, to which Imatinib does not bind. This effect has been already observed in experiments (IC50=6.7 μM for Glu21Lys) [15].

Alanine-mutagenesis

The mutation studies of these amino acids at specific sites may be helpful to understand the activity played by these residues. We used alanine-mutagenesis to investigate the effect of mutation on salt-bridges. The results of the alanine-mutagenesis are shown in Table 3. From the table it is clear that mutations do not change the binding energy of Imatinib significantly, because the variations in intermolecular energy ΔE_{gas} and solvation energy ΔE_{solv} cancel the net change in binding energy except for Mut2. However, the total free energy ΔG_{free} has significantly different value. This variation in free energy arises due to difference in entropic contributions. Consequently, one can speculate that the salt-bridge does not play a vital role in the activity of inhibitors. However, mutation causes a large change in entropic contribution of the complex. Here, we recall that the entropic contribution implies the vibrational and rotational motion of the system, hence a major change in entropy results to a large motion in the complex which

Table 3 The computational alanine-mutagenesis results for Abl kinase complex

Cont	Mut1	Mut2	Mut3	Mut4	Mut5	Mut6	Mut7
ΔE_{ele}	−20.6	−19.7	−30.3	−19.7	−28.3	−30.7	−29.6
ΔE_{vdw}	−73.3	−71.5	−71.6	−73.8	−70.2	−70.8	−71.2
ΔE_{gas}	−94	−91.2	−98.7	−93.6	−96.2	−98.3	−98.3
ΔG_{np}	−7.6	−7.6	−7.5	−7.6	−7.5	−7.5	−7.5
ΔG_{pol}	56.1	59	58.6	55.3	58.6	59.8	61.4
ΔG_{solv}	48.5	51.4	51	47.7	51	52.2	53.8
ΔG_{bind}	−45.5	−39.8	−47.7	−45.8	−45.2	−46	−44.5
TΔS	−27.4	−31.1	−29.5	−40	−24	−27.7	−11.6
ΔG_{free}	−18	−8.6	−18.1	−5.8	−21.1	−18.3	−32.8

Here

Mut1=Mutation of residue Lys37 and Glu52 with Ala

Mut 2=Mutation of residue Glu82 and Lys144 with Ala

Mut 3=Mutation of residue Asp157 and Lys170 with Ala

Mut 4=Mutation of residue Asp7 and Lys29 with Ala

Mut 5=Mutation of residue Lys13 and Glu21 with Ala

Mut 6=Mutation of residue Lys11 and Glu24 with Ala

Mut 7=Mutation of residue Glu118 and Lys122 with Ala

might been occurred due to breaking of salt-bridges in receptor due to above mutations. Hence, we conclude that the salt-bridges do not affect the inhibitor activity but it should be conserved for the stability of kinase complex. This assumption is supported by data available in the literature [41] which suggests that kinase is less stable at pH <4.5, and effects the folding pattern of the protein. It may happen because of the unavailability of salt-bridge forming residue. At lower pH a proton is attached at the side chain of Glu and Asp and these residues does not possess the negative charge,

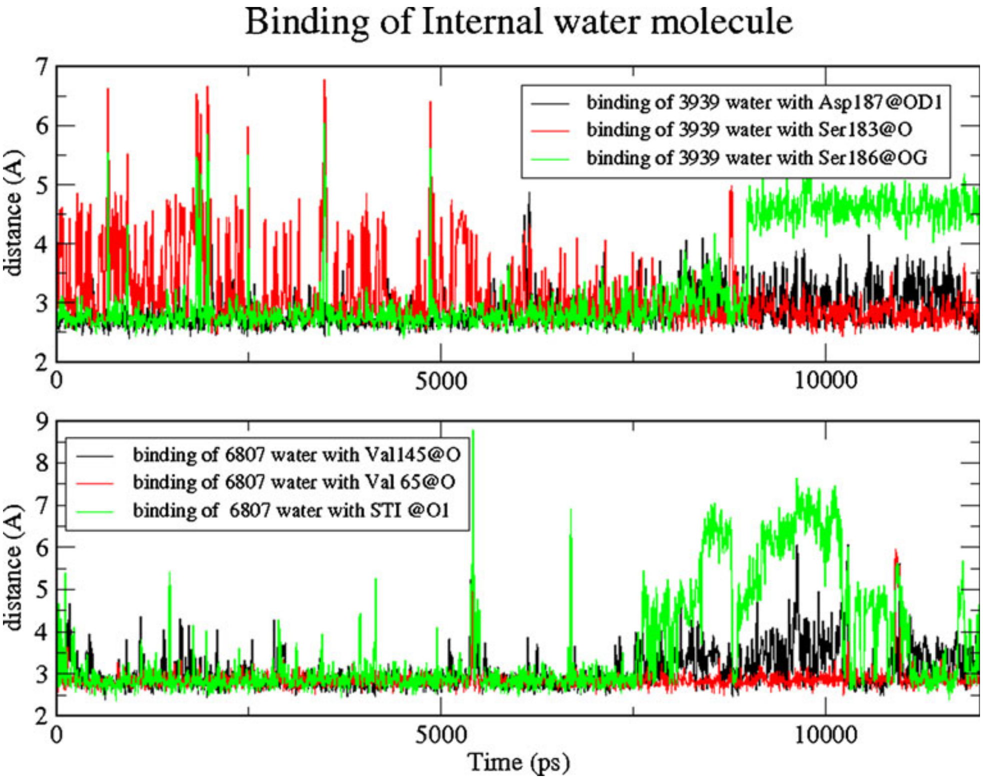


Fig. 7 Life time binding of water molecules with Imatinib and receptor during whole production dynamics with 3939 (upper) and 6807(lower) water molecule. The distances between heavy atoms are plotted with respect to time

and proton attracting capability is lost. Unfortunately, the folding pattern at low pH and its three dimensional structure is not known for c-Abl kinase. Furthermore, the mutation of Glu82Ala and Lys144Ala has a considerable effect on the activity of Imatinib. By the structural insight of kinase complex, we see that the salt-bridge between Glu 82 and Lys 144 forms a cavity for Imatinib. Due to alanine mutation, this salt-bridge breaks and increases the distance of inhibitor from activation loop which causes a weak bonding and hence there is a decrement in binding free energy.

Effect of hydration

All the experiments are carried out in water in order to understand its role on the stability of the kinase ligand complex. The water molecule can bridge between the carbonyl oxygen atoms and amide protons of different amino acids to catalyze the formation of a particular structure and its reversal for kinase-inhibitor and protein-protein interface. A long lived water molecule (WAT 6807) has been observed during the simulations. As shown in Fig. 7, this particular molecule lives in the same range during the whole simulation period which may be termed as the life partner and forms linkage between the ligand STI and amino acid Val65. Another water molecule WAT 3939 has longer time with the kinase along with Imatinib. It is bound close to Asp 187, Ser 183 and Ser186 and remains there in the same region during the whole simulation period. A spine of water network has been observed in first hydration shell which is shown in supplementary materiel Fig. 2S. The hydrogen bonds holding these water molecules to the kinase are stronger with longer life time than bulk water, and this water is available for colligative effects. The specific positions occupied by water molecules observed in the first hydration shell are same as observed in the crystal structure of various kinases of crystal structures of kinases Src, Abl and others [41].

Conclusions

The results obtained from MD simulation and MM-PB/SA method to evaluate binding free energy led us to conclude that Imatinib binds to the cleft of this inactivated kinase to form a stable complex. The absolute free energy of −06.04 kcal mol^{-1} supports the tight binding of Imatinib with c-Abl tyrosine kinase. A cavity is created by replacing the solvent molecule from the large solvent accessible surface area which is confirmed by the cavity energy of 61.64 kcal mol^{-1}. The binding energy of −11.87 kcal mol^{-1} obtained from SIE method is in accordance of the experimental binding energy of 10.37 kcal mol^{-1} [41]. It is concluded that the hydration play an important role for the stability of the kinase-Imatinib complex.

We propose that Glu21:Lys13, Glu52:Lys37, Glu82: Lys144, Glu157:Lys170 salt-bridges are conserved for the in-vivo stability of kinases and are necessary for optimum activity of kinase. The loop regions are preferred sites for ATP because of conformational adaptability required for active-site binding. Therefore, structural stabilization in the loop region is important for kinase functions. The mutational experiments for some of the salt-bridges shows that deformation in most salt bridges do not effect the activity of kinase considerably, but it increases the entropy, i.e., disorderness of the complex. It is also obvious that, mutations of amino acid near the binding site are hot spot and they result in reduction in the activity because of the loss of conformational rigidity and effectively converted the property of the inhibitor. Our finding highlights the importance of considering kinase conformation in the rational design of inhibitors for cancer targets.

Acknowledgments The partial computational work is carried out at Supercomputing facility for bioinformatics and computational biology, Indian Institute of Technology, New Delhi, India, their support is gratefully acknowledged. Special thanks to Prof. B. Jayaram, coordinator SCFbio, Indian Institute of Technology, New Delhi (IITD) to facilitate the access of supercomputer at IITD. Authors are grateful to Dr. Azara Praveen, Zakir Hussain College of Engineering and Technology, Aligarh Muslim University, Aligarh, for proofing the manuscript. KDD thanks to Council for Scientific and Industrial Research (CSIR) for the award of senior research fellowship.

References

1. Rix U, Hantschel O, Durnberger G, Rix LLR, Planyavsky M, Fernbach NV, Kaupe I, Bennett KL, Valent P, Colinge J, Kocher T, Furga GS (2007) Chemical proteomic profiles of the BCR-ABL inhibitors Imatinib, Nilotinib and Dasatinib reveals novel kinases and non kinases targets. Blood 110:4055–4063
2. Seeliger MA, Nagar B, Frank F, Cao X, Henderson MN, Kuriyan J (2007) c-Src binds to the cancer drug imatinib with an inactive abl/c-kit conformation and a distributed thermodynamic penalty. Structure 15:299–311
3. Deininger MWN, GoldmanJM MJV (2000) The molecular Biology of chronic myeloid leukemia. Blood 96:3343–3356
4. Goodsell DS (2005) The molecular Biology of chronic myeloid leukemia. Oncologist 10:758–759
5. Nagar B, Hantschel O, Young MA, Scheffzek K, Veach D, Bornmann W, Clarkson B, Superti-Furga G, Kuriyan J (2003) Structural basis for the auto inhibition of c-Abl tyrosine kinase. Cell 112:859–871
6. Nagar B (2007) c – Abl Tyrosine Kinase and inhibition by the cancer Drug Imatinib. J Nutr 137:1518S–1523S
7. Li W, Miller WT (2006) Role of activation loop tyrosine in regulation of the insulin-like growth factor I receptor tyrosine kinase. J Bio Chem 281:23785–23791
8. Emrick MA, Lee T, Starkey PJ, Mumby MC, Resing KA, Ahn NG (2006) The gatekeeper residue controls, auto activation of

ERK2 via a pathway of intramolecular connectivity. PNAS 103:18101–18106

9. Shan Y, Seelinger MA, Eastwood MP, Frank F, Xu H, Jensen MO, Dror RO, Kuriyan J, Shaw DE (2009) A conserved protonation dependent switch controls drug binding in the Abl kinase. PNAS 106:139–144
10. Buettner R, Mesa T, Vulture A, Lee F, Jove R (2008) Inhibition of Src family kinases with dasatinib blocks migration and invasion of human melanoma cells. Mol Cancer Res 6:1766–1774
11. Carpinelli P, Ceruti R, Giorgini ML, Cappella P, Gianellini L, Croci V, Degrassi A, Texido G, Rocchetti M, Vianello P, Rusconi L, Storici P, Zugnoni P, Arrigoni C, Soncini C, Alli C, Patton V, Marsiglio A, Ballinari D, Pesenti E, Fancelli D (2007) PHA-739358, a potent inhibitor of Aurora kinases with a selective target inhibition profile relevant to cancer. J Mol Cancer Res 6:3158–3168
12. Mol CD, Dougan DR, Schneider TR, Skene RJ, Kraus ML, Scheibe DN, Snell GP, Zou H, Sang BC, Wilson KP (2004) Structural basis for the auto inhibition and STI-571 inhibition of c-Kit tyrosine kinase. J Biol Chem 279:31655–31663
13. Nagar B, Bornmann W, Pellicena P, Schindler T, Veach D, Miller WT, Clarkson B, Kuriyan J (2002) Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and Imatinib (STI-571). Cancer Res 62:4236–4243
14. Cowan-Jacob SW, Fendrich J, Floersheimer A, Furel P, Liebentanz J, Rummel G, Rheinberger P, Centeleghe M, Fabbro D, Manely PW (2007) Structural bilogy contribution to the discovery of drug to treat chronic myelogenous leukaemia. Acta Cryst D 63:80–93
15. Tokarski JS, Newitt JA, Chang CHYJ, Cheng JD, Wittekind M, Kiefer SE, Kish K, Lee FYF, Brozillerri R, Lombardo LJ, Xie D, Zhang Y, Klei HE (2006) The structure of dasatinib (BMS-354825) bound to activated Abl kinase domain elucidates its inhibitory activity against Imatinib resistance Abl-Mutant. Cancer research 66:5790–5797
16. Denisov VP, Halle B (1996) Protein hydration dynamics in aqueous solution. Faraday Disc 103:227–244
17. Bryant RG (1996) The dynamics of water-protein interactions. Annu Rev Biophys Biomol Struct 25:29–53
18. Rupley JA, Careri G (1991) Protein hydration and function. Adv Protein Chem 41:37–172
19. Meyer E (1992) Internal water molecules and h-bonding in biological macromolecules: a review of structural features with functional implications. Protein Sci 1:1543–1562
20. Case DA, Cheatham TE, Daren T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The amber biomolecular simulation programs. J Comput Chem 26:1668–1688
21. Bayly CI, Cieplak P, Cornell W, Kollman PA (1993) A well behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. J Phys Chem 97:10269–10280
22. Cornell WD, Cieplak P, Bayly CI, Kollman PA (1993) Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. J Am Chem Soc 115:9620–9631
23. Jorgenson WL, Chandrashekhar J, Madura JD, Imprey RW, Klein M (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
24. Izaguirre JA, Catarello DP, Wozanaik JM, Skeel RD (2001) Langevin stabilization of molecular dynamics. J Chem Phys 114:2090–2098
25. Berendsen HJC, Postama JPM, van-Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:3684–3690
26. Ryckaert JP, Cicotti G, Barendsen HJC (1977) Numerical integration of the Cartesian equation of motion of a system with constraints: Molecular dynamics of n-alkanes. J Comput Phys 23:327–341
27. Humphrey W, Dalke A, Schulten K (1996) VMD – virtual molecular dynamics. J Mol Graph Model 14:33–38
28. Pettersen EF, Goddard TD, Haung CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF chimera – A visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612
29. Maestro, Schrodinger Inc, USA (2008)
30. Gohlke H, Case DA (2003) Converging free energy estimates: MM-PB (GB) SA studies on the Protein-Protein Complex Ras-Raf. J Comput Chem 25:238–250
31. Fogolari F, Brigo A, Molinari H (2003) Protocols for MM/PBSA molecular dynamics simulations of proteins. Biophys J 85:159–166
32. Grochowaski P, Trylska J (2007) Continuum molecular electrostatics, salt effects, and counterion binding- a review of the poisson-boltzmann theory and its modifications. Biopolymers 89:93–113
33. Tsui V, Case DA (2001) Theory and application of generalized born solvation model in macromolecular simulations. Biopolymers 56:271–291
34. Dubey KD, Ojha RP (2011) Binding free energy calculation with QM/MM hybrid methods for Abl-Kinase inhibitor. J Biol Phys 37:69–78
35. Leach AR (2003) Molecular Modeling: Principle and Application, 2nd edn. Prentice Hall
36. Moreira IS, Fernandes PA, Ramos MJ (2006) Unraveling the importance of protein-protein interaction: application of computational alanine-mutagenesis to the study of the IgG1 streptococcal protein G (C2 fragment) complex. J Phys Chem B 110:10962–10969
37. Moreira IS, Fernandes PA, Ramos MJ (2006) Computational alanine scanning mutagenesis, an improved methodological approach. J Comput Chem 28:644–654
38. Masso M, Lu Z, Vaismann II (2006) Computational mutagenesis of protein structure function correlation. Proteins 64:234–245
39. Naim M, Bhat S, Rankin KN, Dennis S, Chowdhury SF, Siddiqi I, Drabik P, Sulea T, Bayly C, Jakalian A, Purisima EO (2007) Solvated interaction energy (SIE) for scoring protein ligand binding affinities. 1. Exploring the parameter space. J Chem Inf Model 47:122–133
40. Cui Q, Sulea T, Schrag JD, Munger C, Hung MN, Naim M, Cygler M, Purisima EO (2008) Molecular dynamics and solvated interaction energy studies of protein –protein interaction: The MP1-p14 scaffolding complex. Structural biology contribution to tyrosine kinase drug discovery. J Mol Biol 379:787–802
41. Pricl S, Fermeglia M, Ferrone M, Tamborini E (2005) T315I-mutated Bcr-Abl in chronic myeloid leukemia and Imatinib: insights from a computational study. Mol Cancer Ther 4:1167–1174

J Mol Model (2012) 18:1691–1700
DOI 10.1007/s00894-011-1150-0

ORIGINAL PAPER

Computational modeling on the recognition of the HRE motif by HIF-1: molecular docking and molecular dynamics studies

Pandian Sokkar · Vani Sathis ·
Murugesan Ramachandran

Received: 15 April 2011 / Accepted: 9 June 2011 / Published online: 5 August 2011

Abstract Hypoxia inducible factor-1 (HIF-1) is a bHLH-family transcription factor that controls genes involved in glycolysis, angiogenesis, migration, as well as invasion factors that are important for tumor progression and metastasis. HIF-1, a heterodimer of HIF-1α and HIF-1β, binds to the hypoxia responsive element (HRE) present in the promoter regions of hypoxia responsive genes, such as vascular endothelial growth factor (VEGF). Neither the structure of free HIF-1 nor that of its complex with HRE is available. Computational modeling of the transcription factor–DNA complex has always been challenging due to their inherent flexibility and large conformational space. The present study aims to model the interaction between the DNA-binding domain of HIF-1 and HRE. Experiments showed that rigid macromolecular docking programs (HEX and GRAMM-X) failed to predict the optimal dimerization of individually modeled HIF-1 subunits. Hence, the HIF-1 heterodimer was modeled based on the phosphate system positive regulatory protein (PHO4) homodimer. The duplex VEGF-DNA segment containing HRE with flanking nucleotides was modeled in the B form and equilibrated via molecular dynamics (MD) simulation. A rigid docking approach was used to predict the crude binding mode of HIF-1 dimer with HRE, in which the putative contacts were found to be present. An MD simulation (5 ns) of the HIF-1–HRE complex in explicit water was performed to account for its flexibility and to optimize its interactions. All of the conserved amino acid residues were found to play roles in the recognition of HRE. The present work, which sheds light on the recognition of HRE by HIF-1, could be beneficial in the design of peptide or small molecule therapeutics that can mimic HIF-1 and bind with the HRE sequence.

Keywords Hypoxia-inducible factor-1 · Hypoxia responsive element · DNA recognition · Homology modeling · Molecular dynamics

Introduction

Hypoxia plays an important role in tumor progression through angiogenesis and resistance to programmed cell death [1–3]. A hypoxic tumor occurs due to the increased metabolic rate and oxygen consumption of rapidly proliferating tumor cells [4]. A hypoxic tumor can also arise as the distance from the local capillary increases (due to the expanding cell mass) and during photodynamic cancer therapy [4, 5]. The hypoxia-responsive pathway allows tumor cells to overcome harsh conditions. The most important mediator identified in this pathway is hypoxia inducible factor-1 (HIF-1), a transcription factor for various angiogenic factors such as vascular endothelial growth factor (VEGF), and for genes encoding proteins involved in energy metabolism, cell survival, red blood cell production, and vasomotor regulation [6, 7].

HIF-1 is a heterodimer consisting of HIF-1α and HIF-1β subunits. HIF-β is a nuclear protein, whereas HIF-1α shuttles between the cytoplasm and nucleus [8]. The α and β subunits both belong to the basic helix-loop-helix (bHLH) PER-ARNT-SIM (PAS) domain family of transcription factors [8]. In HIF-1α, the N-terminal (bHLH-PAS) domain is required for dimerization and DNA binding, whereas the C-terminal domains are required for hypoxia-induced nuclear localization, protein stabilization

P. Sokkar · M. Ramachandran
School of Chemistry, Madurai Kamaraj University,
Madurai 625 021, Tamil Nadu, India

V. Sathis · M. Ramachandran (✉)
School of Biological Sciences, Madurai Kamaraj University,
Madurai, 625 021, Tamil Nadu, India
e-mail: rammurugesan@yahoo.com

and transactivation [9–11]. HIF-1α is stable only under hypoxia, and the accumulation of HIF-1α is followed by its entry into the nucleus, where HIF-1α binds with HIF-1β. The two subunits then bind with a specific five-nucleotide DNA sequence (5′-RCGTG-3′), known as the hypoxia responsive element (HRE), located in the promoter regions of hypoxia-responsive genes [10].

VEGF is a key mediator for angiogenesis in several forms of cancer, and it is induced most importantly by hypoxia [12, 13]. In addition, VEGF is an appealing target for anticancer therapeutics [14]. The HIF-1 dimer binds to the HRE sequence (5′-TACGTG-3′) in the VEGF promoter and induces the expression of VEGF. Echinomycin, a quinoxaline class of cyclic peptide antibiotic, is known to bind to the VEGF-HRE sequence and inhibit VEGF expression [15]. Interestingly, echinomycin has also been reported to induce apoptosis in several types of cancer cell [16, 17]. Therefore, targeting the HRE sequence with small molecules could be a potential therapeutic option to treat cancer. A short peptide model (with nanomolar affinity) of yeast GCN4 transcription factor has been developed, based on the binding mode of GCN4 with its cognate DNA, by Talanian et al. [18]. Hence, the mechanism of HRE sequence recognition by HIF-1 dimer would greatly aid in the design of peptides or small molecules that can specifically bind to HRE. Unfortunately, neither the structure of free HIF-1 nor that of its complex with HRE is available. However, the DNA-binding (bHLH) domains of HIF-1α and HIF-1β can be modeled from the crystal structures of several HLH transcription factors. In general, the computational modeling of macromolecular complexes (protein–protein and protein–DNA complexes) is challenging due to the large conformational space, the inherent structural flexibility of such complexes, and the conformational changes induced upon complex formation [19, 20]. Hence, modeling the HIF-1–DNA complex from the subunits HIF-1α and HIF-1β is a rather challenging task.

Despite the presence of crystal structures of several transcription factors bound to DNA, the molecular modeling approach has not been widely adopted for unknown structures of this kind. In the present study, the computational modeling of the bHLH domains of the HIF-1 dimer and the interaction of the HIF-1 dimer with HRE are discussed.

Methodology

Sequence alignment

To find suitable templates to model the DNA-binding domain of HIF-1, bHLH domain (both HIF-1α and HIF-1β) sequences were aligned with structures in the Protein Data Bank [21] (PDB: http://www.pdb.org/) using the NCBI-BLASTp tool [22], which is available on the NCBI website (http://www.ncbi.nlm.nih.gov/) using a default threshold *E* value of 10 and an inclusion threshold value of 0.005. Multiple sequence alignments were created using the ClustalX tool [23].

HIF-1 dimer modeling

The crystal structure of the PHO4 homodimer bound to DNA (1A0A) [24] was selected as a template to model the HIF-1 dimer. The sequences of the DNA-binding regions of ten bHLH-transcription factors, including PHO4, were aligned with HIF-1α and HIF-1β using ClustalX with a Gonnet weight matrix [25] (gap opening penalty 10 and gap extension penalty 0.2). The alignment between PHO4 and HIF-1α/HIF-1β was used for model building in Modeller 9v7 [26]. To model the HIF-1 dimer, the HIF-1 subunits were modeled from the two subunits of the PHO4 homodimer. The resulting HIF-1 dimer was refined by the “slow_large” optimization protocol of Modeller [26].

DNA modeling

The HRE sequence (–TACGTG–), along with fourteen flanking nucleotides, was selected from the promoter region of human VEGF-1. The canonical B-form structure of this twenty-nucleotide sequence was modeled using the model.it server [27] available at http://hydra.icgeb.trieste.it/dna/.

Molecular dynamics of DNA

Simulation system

The Visual Molecular Dynamics (VMD) tool was used to prepare the simulation system [28]. The CHARMM27 force field [29] was used for parameterization and the program NAMD [30] was used for all energy minimization and molecular dynamics (MD) simulation work, unless otherwise specified. All of the MD simulations were carried out in explicit water, employing periodic boundary conditions. The TIP3P water model [31] was used to solvate the DNA molecule in a rectangular water box, with a distance of at least 12 Å imposed between the solute atoms and the edge of the box. No counterions were added to neutralize the charge. The system was first energy minimized for 1000 steps with the atoms of DNA fixed, and then unrestrained energy minimization was performed for 1000 steps.

Simulation settings

The system was equilibrated at 250 K for 10 ps with the DNA atoms fixed, followed by 10 ps MD without restraints. The system was subsequently simulated for 100 ps at 300 K with the following settings. The classical equations of motion were integrated by a leapfrog integrator using a time step of 1 fs. The impulse-based Verlet-I/r-RESPA method was used to perform multiple time stepping: 4 fs for long-range electrostatics, 2 fs for short-range nonbonded forces, and 1 fs for bonded forces [32]. The Swift function was used to cut off the Lennard–Jones potential, with the first cutoff at 10 Å and the second cutoff at 12 Å. Particle mesh Ewald (PME) was employed to calculate long-range electrostatic interactions with a grid size of ~1 unit in all directions [33]. Short-range interactions were calculated at intervals of 2 fs and long-range interactions were calculated at intervals of 4 fs. All bonds involving hydrogen atoms were constrained to their equilibrium bond parameters using the SHAKE algorithm [34]. Langevin dynamics were employed to maintain the temperature at 300 K using a damping coefficient of 1/ps. The Langevin piston was employed to maintain the pressure at 1 atm, with a Langevin piston period of 100 fs and an oscillation decay time of 50 fs. The Langevin piston was coupled to the heat bath. Trajectories were recorded every 200 fs.

DNA–HIF-1 docking

MD simulation was performed with the modeled DNA structure for 100 ps at 300 K, and the equilibrated structure was used for docking studies. The interaction between HIF-1 (bHLH dimer) and DNA was studied using the program HEX 6.1 [35]. HEX is a rigid macromolecular docking program employing spherical polar Fourier (SPF) correlations in terms of shape and electrostatics. In many respects, this approach is similar to conventional fast Fourier transform (FFT) docking methods which use Cartesian grid representations of the molecular shape and electrostatic properties and translational FFTs to perform docking correlations. The default docking control parameters of the HEX program were used to arrive at 100 docked conformations.

MD simulation of the HIF-1–DNA complex

The HIF-1–DNA complex was placed in a TIP3P water box, with a distance of at least 12 Å imposed between the solute atoms and the edge of the box. The positions of the solute atoms were fixed during the first energy minimization for 1000 steps, and no atoms were fixed during the second energy minimization for 1000 steps. MD simulation of the complex was carried out with the aforementioned protocol. The system was equilibrated at 250 K for 10 ps with the solute atoms fixed, and this was followed by 40 ps of unrestrained MD. The equilibrated system was subjected to 5 ns MD at 300 K.

Analysis of model quality, MD simulation and docking

The quality of the modeled HIF-1 structure was assessed based on the Ramachandran plot occupancy of residues using the PROCHECK server [36]. VMD was used to analyze the molecular dynamics trajectory. The protein and DNA atoms in all of the frames were superimposed on the first frame of the trajectory to remove global rotational and translational movements. The root mean square deviation (RMSD) was calculated with reference to the starting structure. The secondary structure of the protein residues during the simulation was analyzed with the STRIDE program [37], as implemented in the VMD-TIMELINE plug-in. The PyMol molecular viewer (http://www.pymol.org/) was employed to analyze the docked structures.

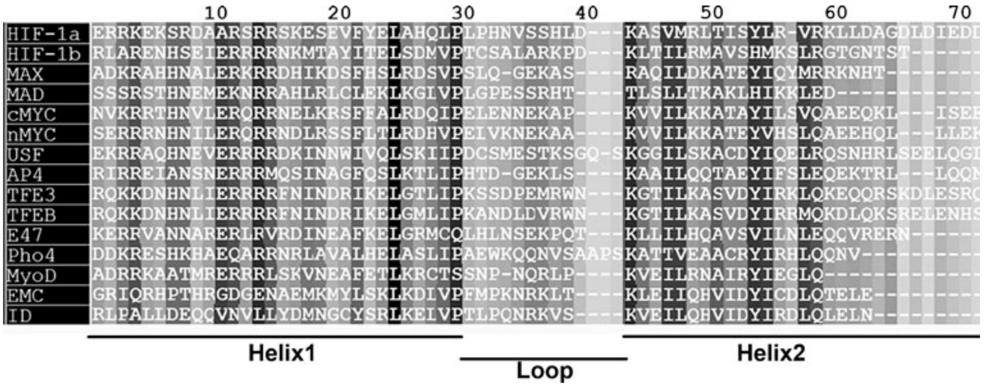

10
20
30
40
50
60
70
HIF-1a ERRKEKSRDAARSRRSKESEVFYELAHQLPLPHNVSSHLD---KASVMRLTISYLR-VRKLLDAGDLDIEDD
HIF-1b RLARENHSEIERRRRNKMTAYITELSDMVPTCSALARKPD---KLTILRMAVSHMKSLRGTGNTST------
MAX ADKRAHHNALERKRRDHIKDSFHSLRDSVPSLQ-GEKAS----RAQILDKATEYIQYMRRKNHT--------
MAD SSSRSTHNEMEKNRRAHLRLCLEKLKGIVPLGPESSRHT----TLSLLTKAKLHIKKLED-----------
cMYC NVKRRTHNVLERQRRNELKRSFFALRDQIPELENNEKAP----KVVILKKATAYILSVQAEEQKL---ISEE
nMYC SERRRNHNILERQRRNDLRSSFLTLRDHVPEIVKNEKAA----KVVILKKATEYVHSLQAEEHQL---LLEK
USF EKRRAQHNEVERRRRDKINNWIVQLSKIIPDCSMESTKSGQ-SKGGILSKACDYIQELRQSNHRLSEELQGL
AP4 RIRREIANSNERRRMQSINAGFQSLKTLIPHTD-GEKLS----KAAILQQTAEYIFSLEQEKTRL---LQQN
TFE3 RQKKDNHNLIERRRRFNINDRIKELGTLIPKSSDPEMRWN---KGTILKASVDYIRKLQKEQQRSKDLESRQ
TFEB RQKKDNHNLIERRRRFNINDRIKELGMLIPKANDLDVRWN---KGTILKASVDYIRRMQKDLQKSRELENHS
E47 KERRVANNARERLRVRDINEAFKELGRMCQLHLNSEKPQT---KLLILHQAVSVILNLEQQVRERN------
Pho4 DDKRESHKHAEQARRNRLAVALHELASLIPAEWKQQNVSAAPSKATTVEAACRYIRHLQQNV---------
MyoD ADRRKAATMRERRRLSKVNEAFETLKRCTSSNP-NQRLP----KVEILRNAIRYIEGLQ-----------
EMC GRIQRHPTHRGDGENAEMKMYLSKLKDIVPFMPKNRKLT----KLEIIQHVIDYICDLQTELE-------
ID RLPALLDEQQVNVLLYDMNGCYSRLKEIVPTLPQNRKVS----KVEILQHVIDYIRDLQLELN-------
Helix1
Loop
Helix2

Fig. 1 Partial sequence alignment of the DNA binding domains of several bHLH transcription factors with the HIF-1 subunits. Amino acids are numbered locally. *Color gradations of white to black* indicate the extent of amino acid conservation. Highly conserved residues are indicated by *dark shading*, and *light shading* indicates variable residues. The alignment shows three regions: the basic residue-rich N-terminal helix (helix1), the highly variable loop region, and the hydrophobic residue-rich C-terminal helix (helix2)

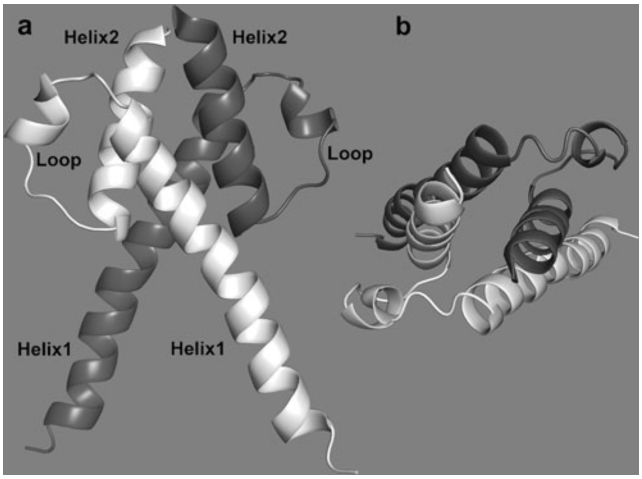


Fig. 2 **a** Cartoon representation of the HIF-1 complex modeled using the crystal structure of the PHO4 homodimer. The model is a four-helix bundle formed by HIF-1α and HIF-1β. HIF-1α is colored *white* and HIF-1β is shown in *dark gray*. N-terminal helices (helix1), loops and C-terminal helices (helix2) are labeled. **b** Orthogonal view of **a** that provides a better view of the packing of the four-helix bundle

Results and discussion

Template selection

The main aim of the work described in this paper was to investigate the DNA-binding mechanism of HIF-1. Hence, the bHLH domain sequence was used for all of the studies reported here. Although the PAS domain is involved in the stabilization of the HIF-1 complex, it does not play a major role in the specific recognition of DNA base pairs [9]. Sequence similarity searches for both HIF-1α and HIF-1β against PDB, using the NCBI-BLASTP program, revealed that the bHLH domains of both HIF-1α and HIF1β did not present high sequence similarities with any known protein structures. The only hit from the bHLH transcription factor family was found to be the crystal structure of PHO4 (1A0A), which showed poor sequence similarity to HIF-1α (sequence identities: 27%, *E* value: 9, positives: 53%, gaps: 5%, and query coverage: 80%). In contrast to HIF-1α, HIF-1β showed fair sequence similarity to several bHLH

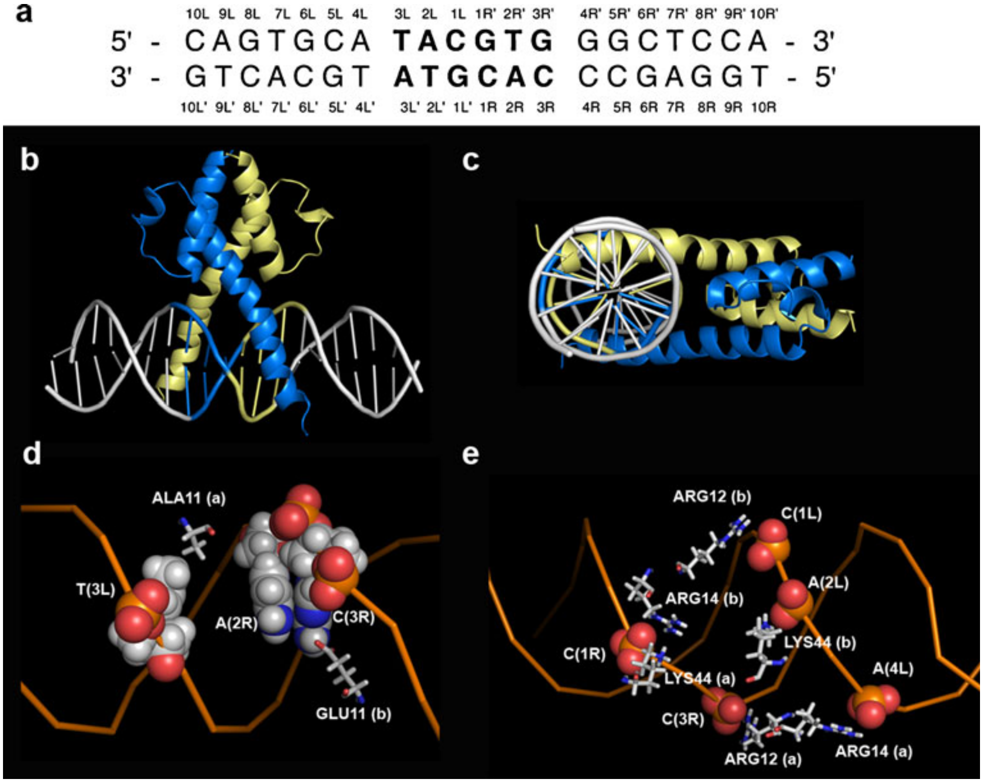

a
10L 9L 8L 7L 6L 5L 4L 3L 2L 1L 1R' 2R' 3R' 4R' 5R' 6R' 7R' 8R' 9R' 10R'
5' - CAGTGCA TACGTG GGCTCCA - 3'
3' - GTCACGT ATGCAC CCGAGGT - 5'
10L' 9L' 8L' 7L' 6L' 5L' 4L' 3L' 2L' 1L' 1R 2R 3R 4R 5R 6R 7R 8R 9R 10R
b
c
d
ALA11 (a)
T(3L)
A(2R)
C(3R)
GLU11 (b)
e
ARG12 (b)
C(1L)
ARG14 (b)
A(2L)
LYS44 (b)
C(1R)
LYS44 (a)
A(4L)
C(3R)
ARG12 (a)
ARG14 (a)

Fig. 3 **a**–**e** The binding mode of the HIF-1 dimer with HRE, as predicted by the rigid macromolecular docking program, HEX 6.1. **a** DNA numbering scheme according to Ferré-D'Amaré et al. [39]. The core HRE sequence is shown in *bold letters*. **b** Overview of the binding of HIF-11 to the HRE (TACGTG) motif. The half-site recognition of each HIF-1 subunit is labeled in the figure. HIF-1α (*blue helices*) is bound to the TAC half-site (*yellow coloration in the DNA structure*). HIF-1β (*yellow helices*) is bound to the GTG half-site (*blue lines in the DNA*). **c** Visualization of **b** alongside the helical axis of the DNA. **d** Vital interactions of amino acid residue (*stick representation, colored by atom type*) E11 of HIF-1β with amino groups of A(2R) and C(3R), and the A11 residue of HIF-1 with T(3L). DNA bases are represented as *spheres* and the DNA backbone is shown as an *orange ribbon*. **e** Several conserved basic residues (*stick representation*) make contact with backbone phosphate groups (*spheres*) of DNA. Color code for atom types: *gray* carbon, *blue* nitrogen, *red* oxygen, *orange* phosphorus, *white* hydrogen transcription factors, including PHO4. The best hit was found to be the structure of the USF transcription factor–DNA complex (1AN4, sequence identities: 38%, *E* value: 1 × 10^{-4}, positives: 60%, gaps: 3%, and query coverage: 87%). The next best hit was found to be PHO4 (sequence identities: 28%, *E* value: 0.013, positives: 53%, gaps: 4%, and query coverage: 95%). Hence, PHO4 and USF were selected as templates for modeling HIF-1α and HIF-1β, respectively.

Homology modeling of HIF-1

We initially opted to model HIF-1α and HIF-1β individually using these two templates. However, rigid protein–protein docking programs such as the GRAMM-X server [38] and HEX 6.1 were unable to predict the optimal HIF-1 dimer from individual subunits for DNA binding (data not shown). This may be due to either the absence of the PAS domain in our models or the absence of flexibility treatment in the docking methods. Hence, it was decided to model the dimer from the PHO4 dimer template due to the significant sequence similarity of PHO4 with both subunits of HIF-1.

The accuracy of homology modeling depends largely on the quality of the alignment between the target and template sequences. The low sequence similarity between the HIF-1 subunits and PHO4 could introduce errors into the alignment. To improve the alignment quality, a multiple sequence alignment was generated with various bHLH transcription factors, including the HIF-1 subunits and PHO4 (Fig. 1). The sequence number was assigned locally for ease of discussion. The alignment can be divided into three regions: basic helix1 (1–30), loop (31–43) and helix2 (44–65). The alignment showed that certain residues were highly conserved in their bHLH transcription factors. Almost all of the residues at positions 3, 4, 7, 8, 12, 14 and 15 are positively charged, and R14 is absolutely conserved. It is interesting to note that glutamate is always present at position 11 in these transcription factors, except in the case of HIF-1α. Hydrophobic residues are conserved in the helix1 region, with the invariable presence of L25 and P30. These hydrophobic residues are required for the packing of helices, and P30 is required to terminate helix1. The loop region displays much variation, and the helix2 region shows conserved hydrophobic residues that are required for dimerization.

The alignment between the HIF-1 subunits and PHO4 present in the multiple sequence alignment was used for model building via the Modeller program. Both HIF-1α and HIF-1β were modeled together using different chains of the PHO4 homodimer. The modeled HIF-1 dimer is shown in Fig. 2.

Structure of HIF-1

Each subunit has a relatively long alpha helix, rich in basic residues for DNA binding, and a shorter helix. These two helices are connected by a long loop containing a short turn of a helix that makes the loop compact. The loop determines the directionality of the two helices. The dimer is a four-helix bundle with a packed hydrophobic interior. Unlike leucine zipper proteins, the second helix is very short, which might affect the tight HIF-1 complex formation. However, the PAS domains from the respective subunits of HIF-1 dimerize to give additional support to the complex. The structure quality of this bHLH dimer of HIF-1 was assessed using PROCHECK server. The structure was found to have 94.8% of its residues in the

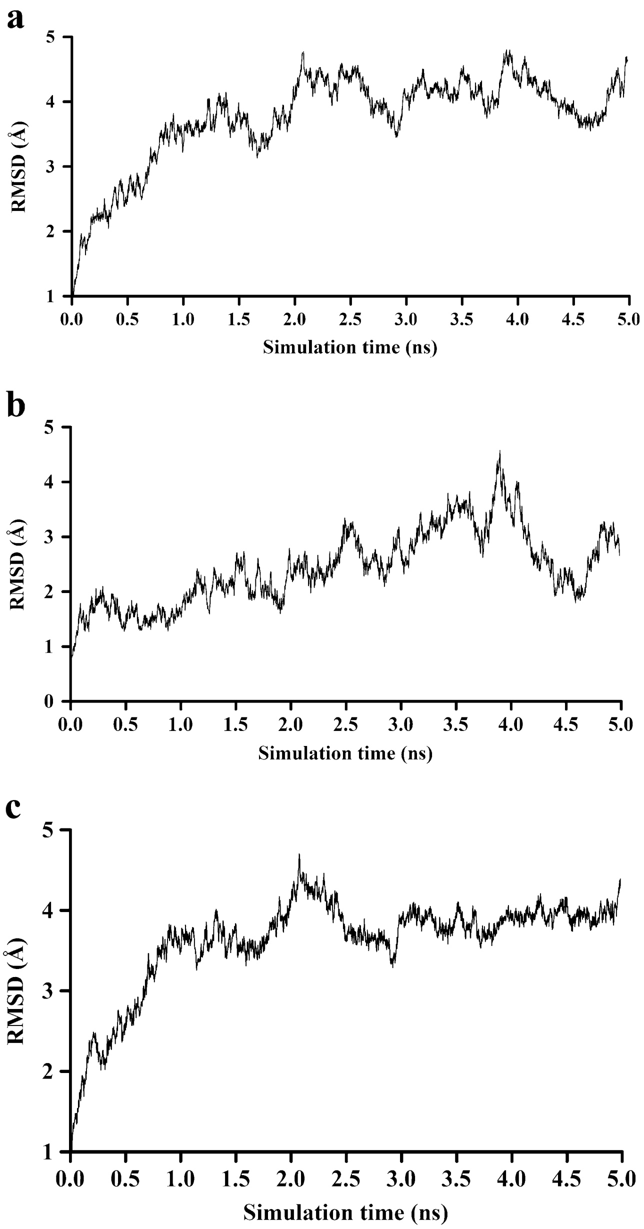


Fig. 4 **a**–**c** Root mean square deviation (RMSD) of non-hydrogen atoms in coordinates as a function of simulation time calculated for **a** the HIF-1 and DNA complex, **b** DNA alone, and **c** HIF-1 alone during the simulation of the HIF-1–DNA complex

most favored regions and the remaining 5.2% of its residues in additionally allowed regions of the Ramachandran plot, thus suggesting that the model is of good quality. This model was used for docking studies with DNA.

DNA structure modeling

The HRE sequence (5′-TACGTG-3′) from the VEGF-1 promoter region, along with 14 flanking bases, was selected to model the three-dimensional structure. The B form of the DNA conformation was selected to mimic the in vivo conformation. It has been observed that the bHLH family of transcription factors generally does not introduce distortions (such as bending) in the DNA structure [24, 39–41]. The unbent DNA structure was therefore used in this study. MD simulation (100 ps at 300 K) of the DNA model was carried out to gently relax the structure. The equilibrated structure from the simulation was selected for further studies. The numbering scheme for DNA follows that of Ferré-D'Amaré et al. (Fig. 3a) [40].

HIF-1–DNA docking

The interaction of HIF-1 dimer with HRE was studied using the HEX tool, which offers rigid macromolecular docking based on shape and electrostatic correlations. Although the interaction between the protein and DNA is inherently flexible, the incorporation of flexibility to calculate docking interactions can be cumbersome. Hence, rigid docking was used to arrive at an approximate model for the interaction, and subsequent refinement of this model was preferred to account for the flexibility. The HEX tool generated 100 different docked conformations of the protein and DNA. In the case of macromolecular docking, the scoring functions are generally inefficient at selecting an optimal binding conformation. Hence, in the present study, conformations in which the binding orientation of HIF-1 with HRE was similar to that of reported bHLH protein–DNA complex structures [24, 39–41] were chosen. From this subset, the docking pose with the highest number of interactions (van der Waals and electrostatic) between the basic region of HIF-1 and HRE was selected for further studies. The residues were considered to be interacting if they were within 3 Å of each other. The various interactions present in this conformation were analyzed.

The initial crude docking pose of HIF-1 with HRE is given in Fig. 3. The two basic helices are postioned compactly in the major groove of DNA, and head in opposite directions (Fig. 3b and c). HIF-1α makes contact with first half-site of the HRE (TAC) duplex, and HIF-1β with the second half-site of the HRE (GTG) duplex (Fig. 3b). The hypoxia specificity of HIF-1α correlates with the recognition of the HRE-specific half-site TAC. On the other hand, HIF-1β is capable of forming heterodimers with other bHLH proteins,

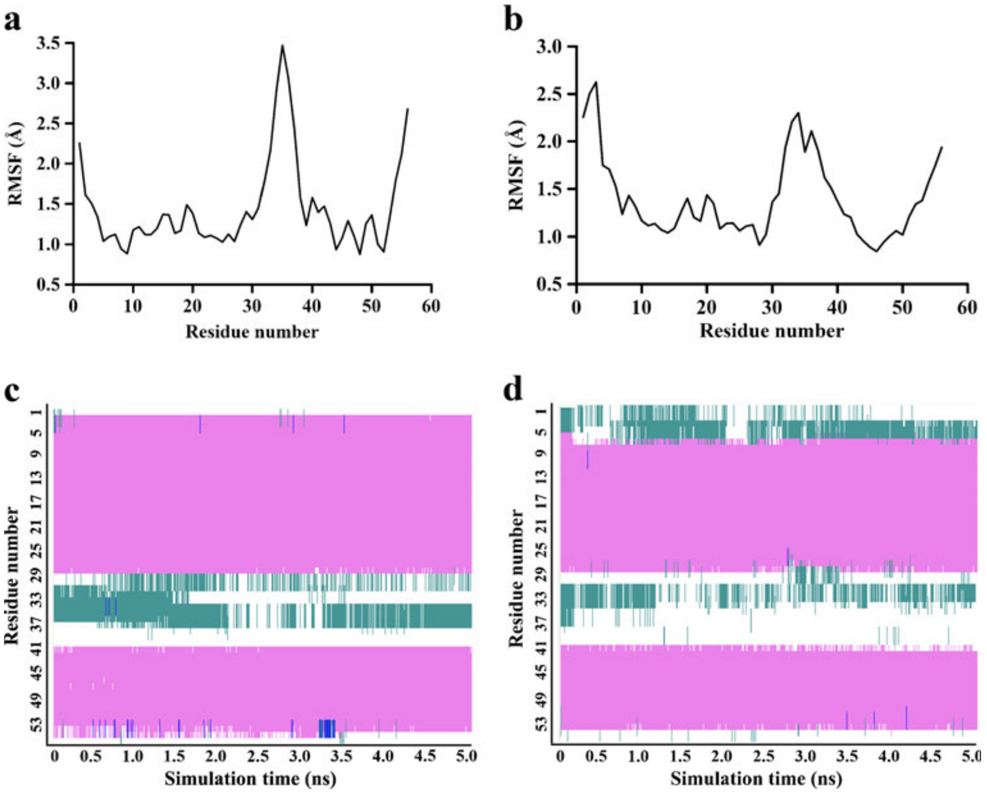


Fig. 5 a–d Root mean square fluctuations (RMSFs) of C-alpha atoms in coordinates for each residue averaged over the simulation time. The RMSFs of HIF-1α (**a**) and HIF-β (**b**) are shown in the picture. **c** and **d** depict the evolution of the secondary structure during the simulation time for HIF-1α and HIF-1β, respectively. Each color represents a secondary structure element: *pink* α-helix, *green* turns, *blue* 3_{10} helix, *white* loops

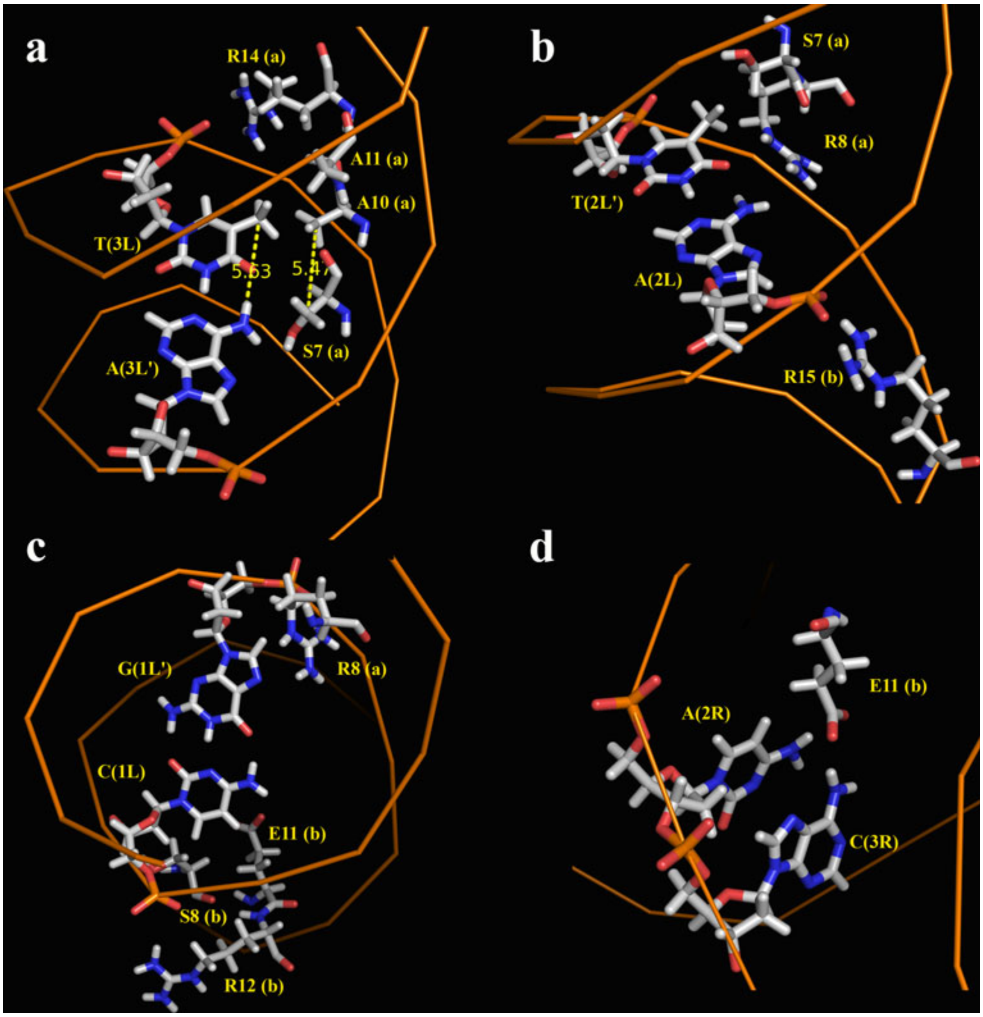


Fig. 6 a–**d** The vital interactions present in the complex of HIF-1 and HRE. **a** The T(3L)–A(3L′) base pair is perfectly complemented by A11 and S7 of HIF-1α. The distance between A11 and S7 is very similar to that of their interaction partners. The distances shown in the picture (as *yellow dotted lines*) are in Å. The recognition of the A(2L)–T (2L′) pair (**b**) and the recognition of the C(1L)–G(1L′) pair (**c**) are given. The phosphate group recognition is also shown in the figure. **d** Simultaneous recognition of A(2R) and C(3R) by the conserved E11 residue of HIF-1β through a hydrogen-bonding bridge is depicted in the figure

and also exhibits a preference for the GTG half-site [42]. The highly conserved E11 residue in HIF-1β interacts with two amino groups, one from C(3R) and the other from A(2R) (Fig. 3d). It is interesting to note that the crystal structures of PHO4 and several other transcription factors display similar interactions through the conserved glutamate residue. In addition, it has been reported that the conserved glutamate in PHO4 is essential for DNA binding activity [43]. In contrast, HIF-1α has an alanine residue instead of glutamate at position 11. However, the hydrophobic residue was found to favorably interact with the methyl group of T(3L) (Fig. 3d). In addition to these interactions, most of the conserved basic residues interact with negatively charged phosphate groups (Fig. 3e). Nevertheless, these interactions may be suboptimal due to the absence of a flexibility treatment.

MD simulation of the HIF-1–DNA complex

MD simulation of this HIF-1–DNA complex was carried out to assess its stability and to optimize the interactions between the protein and DNA. The RMSD of non-hydrogen atoms (non-H RMSD) compared to the initial structure was calculated to investigate the stability of the modeled complex. The RMSD plot showed a general trend for a stable MD trajectory (Fig. 4a). An initial rise in RMSD (up to ~3.5Å) was observed for the first 1 ns, indicating equilibration of the system. After this equilibration period, the RMSD of the protein–DNA complex hovers around 4.0Å for the rest of the simulation period, indicating fluctuations in the structure of the protein and/or DNA. We also calculated the non-H RMSD of the DNA structure alone during the MD simulation of the protein–DNA complex (Fig. 4b). The plot shows that the RMSD of DNA increases gradually (up to 4 Å) for an initial 4 ns and drops to 2–3Å afterwards. This indicates greater fluctuation in the DNA structure during the simulation. However, the fluctuations were observed mostly in the terminal nucleotides, and the protein-binding site was found to be stable. It has generally been observed that an unspecific protein–DNA interaction greatly distorts the DNA structure. The absence of DNA structural distortion during the MD simulation suggests that the protein–DNA interaction is close to optimal in our model. The non-H RMSD

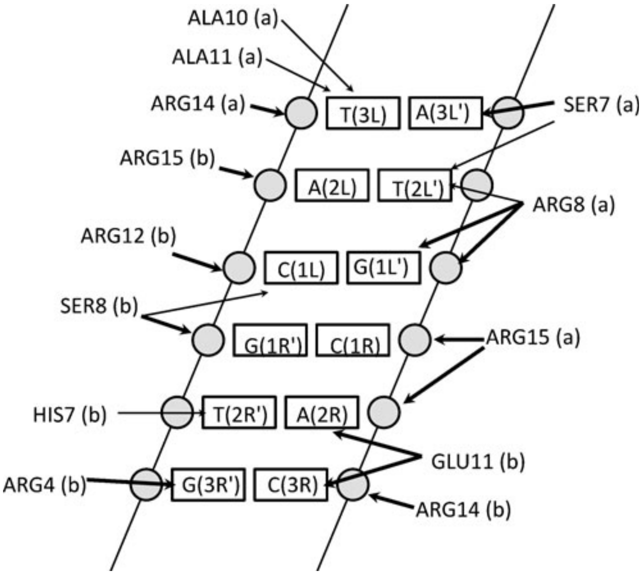


Fig. 7 Schematic diagram summarizing the contacts involved in the recognition of HRE by HIF-1. DNA bases are represented as *cylinders* and phosphate groups are shown as *circles*. The *arrows* represent the contact between amino acids and bases/phosphate. *Thick arrows* indicate hydrogen-bonding interactions and *thin arrows* indicate hydrophobic interactions. The amino acids of HIF-1α and HIF-1β are indicated by *(a)* and *(b)*, respectively

profile of the HIF-1 complex is shown in Fig. 4c. After an initial rise in RMSD (up to 4 Å) for the first 1 ns, the HIF-1 complex relaxes at 4 Å for the remaining simulation time, suggesting that the HIF-1 structure is fairly stable.

Analyzing the root mean square fluctuations (RMSFs) of Cα atoms of HIF-1 can provide information on the flexibility of specific areas of the protein. RMSF plots of both HIF-1α (Fig. 5a) and HIF-1β (Fig. 5b) showed very similar patterns of flexibility. HIF-1α exhibited larger fluctuations (up to ~3.5 Å) in the loop region (residues 30–40) compared to HIF-1β (~2.5 Å). Aside from that, the models showed fluctuations at both N- and C-termini. C-terminal helices are involved in the dimerization of the HIF-1β complex, and fluctuations in these regions could be due to the absence of PAS domains. Higher fluctuations were observed in the N-terminal region of HIF-1β, but not in HIF-1α.

To find out whether these fluctuations can cause changes in the secondary structure, we analyzed the evolution of secondary structure elements during the simulation period (Fig. 5c and d). Both HIF-1 subunits showed two stable helices (pink bands) and one highly fluctuating loop (residues 30–40). The N-terminal helix of HIF-1α was found to be more stable. The conformation of residues 1–7 of HIF-1β was found to fluctuate greatly between turns and loops, which correlates with the observed higher RMSF in the corresponding region. In both HIF-1 subunits, the loop region was found to fluctuate more. A structure was selected at the end of simulation trajectory in order to study the interaction between HIF-1 and DNA.

HIF-1–DNA interaction

The optimized structure was found to be similar to the initial structure. T(3L) interacts with the A11 residue of HIF-1α through its methyl group, as seen in the initial structure. The interaction of A11 with T(3L) could be key to the recognition of HRE, since the T(3L) is specific for HRE and HIF-1α has an alanine residue at position 11 instead of a conserved glutamate. The interaction between the methyl group of T(3L) and A11 did not vary during the MD simulation, suggesting that this interaction is stable. The T(3L) methyl group was found to be juxtaposed with hydrophobic side chains of A10 and A11 from HIF-1α. The complementary base of T(3L), A(3L′), forms a hydrogen bond through its –NH_2 group with the oxygen of S7 (HIF-1α). Since four residues separate S7 and A11, S7 is positioned exactly underneath the A11 in the helix. The N-terminal helical axis of HIF-1α lies almost parallel to the plane of the T(3L)–A(3L′) base pair. In this conformation, T (3L) is located above the A(3L′) and perfectly complements A11 and S7 of HIF-1α. Interestingly, the distance between T(3L)–CH_3 and A(3L′)–NH_2 (5.63 Å) was found to be comparable to the distance between A11–CH_3 and S7–OH (5.47 Å) (Fig. 6a). The negatively charged phosphate group of T(3L) was found to interact with the positive charged side chain of R14 (HIF-1α). The phosphate group of A(2L) interacts with R15 of HIF-1β. The complementary base of A(2L), T(2L′), forms a hydrophobic interaction through its methyl group with the alkyl chain of S7 and R8 of HIF-1α (Fig. 6b). The nonpolar part of the C(1L) base is located in a hydrophobic cavity formed by the alkyl groups of S8 and E11 from HIF-1β. C(1L) phosphate interacts with R8 of HIF-1β (Fig. 6c). The hydroxyl group of S8 (HIF-1β) forms a strong hydrogen bond with the G(1R′) phosphate group. The phosphate group of C(1R) forms two hydrogen bonds with R15 guanidine (HIF-1α). The T(2R′)–CH_3 group forms a σ–π stacking interaction with H7 of the (HIF-1β) aromatic ring. G(3R′) forms two hydrogen bonds, through the nitrogen of its five-membered ring and the CO group of the six-membered ring, with the guanidine group of R4 (HIF-1β). The highly conserved E11 (HIF-1β) recognizes A(2R) and C(3R) by interacting with both A (2R)–NH_2 and C(3R)–NH_2 through its carboxylate group (Fig. 6d). Interestingly, several crystal structures of bHLH transcription factors demonstrate very similar interactions through the conserved glutamate residue [24, 39–41]. Figure 7 summarizes the interactions involved in the recognition of HRE by HIF-1. Thus, most of the conserved residues make contact with either the base or the phosphodiester group.

Previously, Michel et al. generated four mutants (S7A, A10S, A11E and R15A) of HIF-1α through site-directed mutagenesis and investigated the effect of mutations using a reporter gene assay and a DNA-binding assay [44]. The S7A mutant was found to induce the reporter gene level by ~4-fold, while the other mutants failed to induce the reporter gene. In our model, the methyl groups of A10 and A11 are involved in hydrophobic interactions with the methyl group of T(3L). A10S and A11E mutants with polar side chains would destabilize the binding. R15 was found to interact with the phosphate groups of C(1R) and A(2R), and hence the mutation of R15 is not expected to influence DNA binding significantly. However, Blackwell et al. reported that the residues that do not interact directly with the bases could influence the specific DNA recognition [45]. Hence, R15 might play an indirect role in the recognition of HRE. The side chain of S7 is located ~3 Å away from the methyl group of T(2L′), which becomes nonpolar upon mutation (S7A), and this in turn could facilitate stronger binding.

Conclusions

HIF-1 is a bHLH transcription factor that plays a crucial role in hypoxia-triggered angiogenesis and tumor growth. Neither the structure of HIF-1 nor its complex with HRE is available. In the present work, the DNA-binding domain of HIF-1 and its interaction with HRE were modeled for the first time using homology modeling, macromolecular docking and molecular dynamics. The various interactions demonstrated by our model of HIF-1–DNA binding comply with the previously reported crystal structures of the bHLH transcription factor–DNA complex and site-directed mutagenesis data. Elucidating the fundamental interactions that govern the recognition of HRE by HIF-1, as achieved in the present study, could be beneficial for the design and development of small-molecule therapeutics that can bind to HRE.

Acknowledgments We thank the University Grants Commission (UGC), New Delhi, for awarding a Senior Research Fellowship to S. Pandian. Thanks are due to the UGC Networking Resource Centre in Biological Sciences (NRCBS), Madurai Kamaraj University, for lab facilities.

References

1. Graeber TG, Osmanian C, Jacks T, Housman DE, Koch CJ, Lowe SW, Giaccia AJ (1996) Hypoxia-mediated selection of cells with diminished apoptotic potential in solid tumours. Nature 379:88–91
2. Kim CY, Tsai MH, Osmanian C, Graeber TG, Lee JE, Giffard RG, DiPaolo JA, Peehl DM, Giaccia AJ (1997) Selection of human cervical epithelial cells that possess reduced apoptotic potential to low-oxygen conditions. Cancer Res 57:4200–4204
3. Liao D, Johnson RS (2007) Hypoxia: a key regulator of angiogenesis in cancer. Cancer Metast Rev 26:281–290
4. Powis G, Kirkpatrick L (2004) Hypoxia inducible factor-1α as a cancer drug target. Mol Cancer Ther 3:647–654
5. Sitnik TM, Hampton JA, Henderson BW (1998) Reduction of tumour oxygenation during and after photodynamic therapy in vivo: effects of fluence rate. Brit J Cancer 77:1386–1394
6. Harris AL (2002) Hypoxia—a key regulatory factor in tumour growth. Nat Rev Cancer 2:38–47
7. Forsythe JA, Jiang BH, Iyer NV, Agani F, Leung SW, Koos RD, Semenza GL (1996) Activation of vascular endothelial growth factor gene transcription by hypoxia-inducible factor-1. Mol Cell Biol 16:4604–4613
8. Wang GL, Jiang BH, Rue EA, Semenza GL (1995) Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O_2 tension. Proc Natl Acad Sci USA 92:5510–5514
9. Crews ST (1998) Control of cell lineage-specific development and transcription by bHLH-PAS proteins. Genes Dev 12:607–620
10. Lando D, Peet DJ, Whelan DA, Gorman JJ, Whitelaw ML (2002) Asparagine hydroxylation of the HIF transactivation domain a hypoxic switch. Science 295:858–861
11. Pugh CW, O'Rourke JF, Nagao M, Gleadle JM, Ratcliffe PJ (1997) Activation of hypoxia-inducible factor-1; definition of regulatory domains within the α subunit. J Biol Chem 272:11205–11214
12. Plate KH, Breier G, Weich HA, Risau W (1992) Vascular endothelial growth factor is a potential tumour angiogenesis factor in human gliomas in vivo. Nature 359:845–848
13. Toi M, Hoshina S, Takayanagi T, Tominaga T (1994) Association of vascular endothelial growth factor expression with tumor angiogenesis and with early relapse in primary breast cancer. Cancer Sci 85:1045–1049
14. Schneider BP, Sledge GW (2007) Drug insight: VEGF as a therapeutic target for breast cancer. Nat Rev Clin Oncol 4:181–189
15. Kong D, Park EJ, Stephen AG, Calvani M, Cardellina JH, Monks A, Fisher RJ, Shoemaker RH, Melillo G (2005) Echinomycin, a small-molecule inhibitor of hypoxia-inducible factor-1 DNA-binding activity. Cancer Res 65:9047–9055
16. Hu Y, Kirito K, Yoshida K, Mitsumori T, Nakajima K, Nozaki Y, Hamanaka S, Nagashima T, Kunitama M, Sakoe K, Komatsu N (2009) Inhibition of hypoxia-inducible factor-1 function enhances the sensitivity of multiple myeloma cells to melphalan. Mol Cancer Ther 8:2329–2338
17. Park JY, Park SJ, Shim KY, Lee KJ, Kim YB, Kim YH, Kim SK (2004) Echinomycin and a novel analogue induce apoptosis of HT-29 cells via the activation of MAP kinase pathway. Pharmacol Res 50:201–207
18. Talanian RV, McKnight CJ, Kim PS (1990) Sequence-specific DNA binding by a short peptide dimer. Science 249:769–771
19. Andrusier N, Mashiach E, Nussinov R, Wolfson HJ (2008) Principles of flexible protein–protein docking. Proteins 73:271–289
20. Pabo CO, Nekludova L (2000) Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? J Mol Biol 301:597–624
21. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242
22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
23. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882
24. Shimizu T, Toumoto A, Ihara K, Shimizu M, Kyogoku Y, Ogawa N, Oshima Y, Hakoshima T (1997) Crystal structure of PHO4 bHLH domain–DNA complex: flanking base recognition. EMBO J 16:4689–4697
25. Gonnet GH, Cohen MA, Benner SA (1992) Exhaustive matching of the entire protein sequence database. Science 256:1443–1445

26. Fiser A, Sali A (2003) Modeller: generation and refinement of homology-based protein sequence models. Methods Enzymol 374:461–491
27. Vlahovicek K, Pongor S (2000) Model.it: building three dimensional DNA models from sequence data. Bioinformatics 16:1044–1045
28. Humphrey W, Dalke A, Schulten K (1996) VMD—visual molecular dynamics. J Mol Graph 14:33–38
29. MacKerell AD Jr et al (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616
30. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26:1781–1802
31. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
32. Benz RW, Castro-Román F, Tobias DJ, White SH (2005) Experimental validation of molecular dynamics simulations of lipid bilayers: a new approach. Biophys J 88:805–817
33. Darden TA, York DM, Pedersen LG (1993) Particle mesh Ewald. An N.log(N) method for Ewald sums in large systems. J Chem Phys 98:10089
34. Van Gunsteren WF, Berendsen HJC (1977) Algorithms for macromolecular dynamics and constraint dynamics. Mol Phys 34:1311–1327
35. Ritchie DW, Kemp GJ (2000) Protein docking using spherical polar Fourier correlations. Proteins 39:178–194
36. Laskowski RA, MacArthur MW, Moss DS, Thorton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26:283–291
37. Andersen CAF, Palmer AG, Brunak S, Rost B (2002) Continuum secondary structure captures protein flexibility. Structure 10:175–185
38. Tovchigrechko A, Vakser IA (2006) GRAMM-X public web server for protein–protein docking. Nucleic Acids Res 34:W310–W314
39. Ellenberger T, Fass D, Arnaud M, Harrison SC (1994) Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. Genes Dev 8:970–980
40. Ferré-D'Amaré AR, Pognonec P, Roeder RG, Burley SK (1994) Structure and function of the b/HLH/Z domain of USF. EMBO J 13:180–189
41. Ferré-D'Amaré AR, Prendergast GC, Ziff EB, Burley SK (1993) Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. Nature 363:38–45
42. Swanson HI, Chan WK, Bradfield CA (1995) DNA binding specificities and pairing rules of the Ah receptor, ARNT, and SIM proteins. J Biol Chem 270:26292–26302
43. Fisher F, Goding CR (1992) Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANNTG motif. EMBO J 11:4103–4109
44. Michel G, Minet E, Mottet D, Remacle J, Michiels C (2002) Site-directed mutagenesis studies of the hypoxia-inducible factor-1α DNA-binding domain. Biochim Biophys Acta 1578:73–83
45. Kophengnavong T, Michnowicz JE, Blackwell TK (2000) Establishment of distinct MyoD, E2A, and twist DNA binding specificities by different basic region-DNA conformations. Mol Cell Biol 20:261–272

J Mol Model (2012) 18:1701–1711
DOI 10.1007/s00894-011-1187-0

ORIGINAL PAPER

3D-QSAR based pharmacophore modeling and virtual screening for identification of novel pteridine reductase inhibitors

Divya Dube · Vinita Periwal · Mukesh Kumar · Sujata Sharma · Tej P. Singh · Punit Kaur

Received: 22 March 2011 / Accepted: 12 July 2011 / Published online: 9 August 2011

Abstract Pteridine reductase is a promising target for development of novel therapeutic agents against Trypanosomatid parasites. A 3D-QSAR pharmacophore hypothesis has been generated for a series of *L. major* pteridine reductase inhibitors using Catalyst/HypoGen algorithm for identification of the chemical features that are responsible for the inhibitory activity. Four pharmacophore features, namely: two H-bond donors (D), one Hydrophobic aromatic (H) and one Ring aromatic (R) have been identified as key features involved in inhibitor-PTR1 interaction. These features are able to predict the activity of external test set of pteridine reductase inhibitors with a correlation coefficient (r) of 0.80. Based on the analysis of the best hypotheses, some potent Pteridine reductase inhibitors were screened out and predicted with anti-PTR1 activity. It turned out that the newly identified inhibitory molecules are at least 300 fold more potent than the current crop of existing inhibitors. Overall the current SAR study is an effort for elucidating quantitative structure-activity relationship for the PTR1 inhibitors. The results from the combined 3D-QSAR modeling and molecular docking approach have led to the prediction of new potent inhibitory scaffolds.

Keywords Docking · Methotrexate · Neglected diseases · Pteridine reductase · Virtual screening

D. Dube · V. Periwal · M. Kumar · S. Sharma · T. P. Singh ·
P. Kaur (✉)
Department of Biophysics,
All India Institute of Medical Sciences,
Ansari Nagar,
New Delhi, 110029, India
e-mail: punitkaur1@hotmail.com

P. Kaur
e-mail: punit@aiims.ac.in

Introduction

The protozoan parasites of *Trypanosoma and Leishmania* species are causal organism for serious tropical diseases like African sleeping sickness, Chagas' disease and Leishmaniasis [1–4]. These diseases are grouped under neglected tropical diseases as the most affected population are the poorest living in remote, rural areas and urban slums or in the conflict zones. According to WHO latest estimates over 1 billion people or about one sixth of the world's population is suffering from at least one or more neglected tropical diseases. The drugs presently available for the treatment of these fatal diseases are expensive and often toxic. Moreover, there exists a lack of effective and adequate treatment due to the development of resistant strains [5–8]. This further necessitates the development of safe, efficient and cost-effective drugs against these neglected diseases.

Pteridine reductase (PTR1) has emerged as a promising target for the development of novel therapeutics against the protozoan parasites *Trypanosoma* and *Leishmania* [1, 9, 10]. PTR1 is an essential broad spectrum enzyme responsible for pteridine salvage in these organisms. These pterins and folates are essential for the growth of the parasites, however, unlike their mammalian hosts, they do not possess the genes to encode for their de novo synthesis. Thus, these parasites depend on exogenous sources for their uptake which is facilitated by the enzymes bifunctional dihydrofolate reductase (DHFR) – thymidylate synthase (TS) together with pteridine reductase [3]. The DHFR-TS enzyme catalyzes the reduction of folate but shows no activity toward pterins [9]. PTR1 has the ability to reduce conjugated (folates) and unconjugated (biopterins) pterins into their dihydro or oxidized states [9]. This ability of PTR1 to reduce folates in addition to pteirns provides the means to bypass the DHFR-TS pathway. Consequently, it contributes to antifolate drug resistance and is responsible for the failure of conventional therapies against the trypanosomatids. This suggests that it can be targeted for inhibitor design. Therefore, extensive efforts are in progress for the development of newer, specific and safe therapeutic agents targeting it.

Table 1 Training and test set molecules. The test set molecules are highlighted in bold fonts and underlined. R/R^1 and R^2 are the side chains of the scaffolds. The molecules selected as test set with their respective K_i values are shown in bold [11–16]

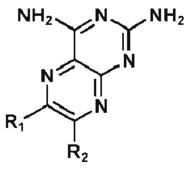
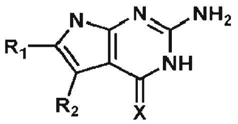
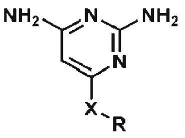
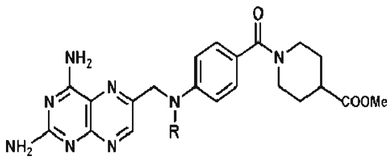
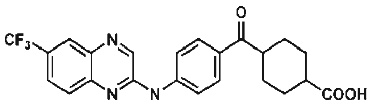
Scaffold	PDB-Id/ Nomenclature	R/R1	R2	R3/X	K_i (μM) *L.major*
	3JQ6	$CH(CH_3)_2$	$CH(CH_3)_2$	-	0.24
	3JQ7	C_6H_5	NH_2	-	**3.4**
	3JQ8	CH_3	$(CH_3)_2$	-	12
	1E7W	$CH_2N(CH_3)C_6H_5CO$ NH $CH(COOH)CH_2CH_2$ COOH	H	-	0.039
	3JQA	H	H	S	>27
	3JQB	H	CH_2 $CH_2C_6H_5$	O	>27
	3JQC	Br	CN	O	>27
	3JQD	C_6H_5	CN	O	>27
	3JQE	$C_6H_4OCH_3$	CN	O	3.4
	3JQ9	$C_7H_5O_2$	CN	O	2.6
	J10	H	H	O	**>27**
	J11	H	CN	O	>27
	J12	C_6H_4 CH_2CH_2	CN	O	16.4
	J13	C_6H_4CHO(meta)	CN	O	**4.2**
	3JQF		-	NH_2	>27
	3BMN	c-C_3H_5	-	NH	>27
	3BMO	$C_6H_4CH_3$	-	S	~27
	3BMQ	$CH_2C_6H_5$	-	S	0.60
	3JQG	$CH_2C_6H_4OCH_3$(para)	-	S	2.7
	3H4V	H	-	-	0.1
	2QHX	CH_3	-	-	**0.037**
	L2	-	-	-	7

Table 1 (continued)

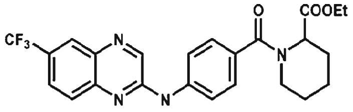
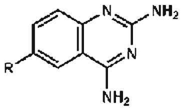
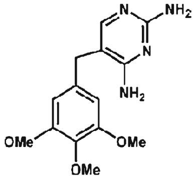
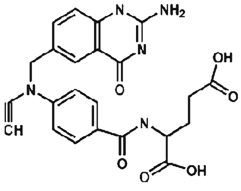
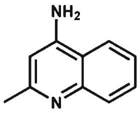
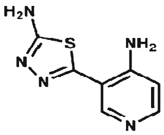
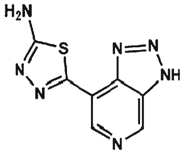
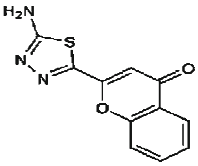
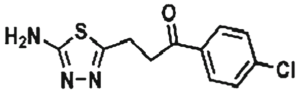
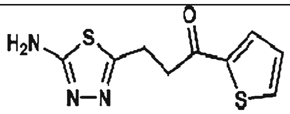
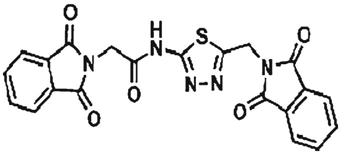
Scaffold	PDB-Id/ Nomenclature	R/R1	R2	R3/X	K_i (μM) *L.major*
	L3	-	-	-	6
	1W0C	NH_2	-	-	2.0
	2VZ0	$C_6H_5CH_3$	-	-	0.10
	**2BFM**	-	-	-	**12**
	2BFA	-	-	-	>10
	L4	-	-	-	390
	**L5**	-	-	-	**31**
	L6	-	-	-	309
	L7	-	-	-	22

Table 1 (continued)

Scaffold	PDB-Id/ Nomenclature	R/R1	R2	R3/X	K_i (μM) *L.major*
	L8	-	-	-	89
	L9	-	-	-	29
	L10	-	-	-	116

Recently, various rational structural techniques like structure-based drug design have been used to identify inhibitors against this enzyme [11–18]. This further necessitates exploration of the binding preferences of the known inhibitors in the context of structure activity relationship and the identification of potential novel lead molecules against PTR1. Pharmacophore modeling is one of the best 3D-QSAR methods which have been successfully applied to drug discovery process [19, 20]. In the present study, based on the knowledge of the *L. major* inhibitors of PTR1 in the literature and the co-crystal structures of the complexes of the same; pharmacophore modeling, docking, and 3D QSAR studies have been performed.

Pharmacophore models of PTR1 inhibitors have been established by using the HypoGen algorithm implemented in the DS (2.0) package (M/S Accelrys Inc., San Diego, USA). The best quantitative pharmacophore model (Hypo1) as well as all the top pharmacophore models consist of pharmacophore features: two hydrogen bond donors, one hydrophobic feature, and one Ring aromatic feature. Thus, Hypo1 was used as a 3D query to perform virtual screening by molecular fingerprint matching using ligand databases including ZINC [21] drug diversity test. The screened compounds were further cross-checked by docking with LigandFit [22]. The most potent inhibitor identified is predicted to be 300 fold better than the reported ones and could serve as the possible lead scaffold for the design of novel and potent PTR1 inhibitors.

Experimental section

Dataset collection

The diverse inhibitor molecules of PTR1 from *L. major* reported in literature with experimentally determined K_i values were selected for 3D-QSAR studies [11–16]. The inhibitor molecules were divided into training and test set so as to cover varied range of the binding affinities and structural diversity. The *L. major* training set consists of 28 molecules with their K_i values ranging from 0.037 to 390 μM over four orders of magnitude. The *L. major* test set consists of six molecules with K_i values ranging from 0.037 to 31 μM. All the inhibitory scaffolds are shown in the Table 1 with the molecules selected as test set being bold and underlined. The program Discovery Studio v2.0 and its various protocols were used for all the molecular modeling and docking studies.

Pharmacophore model generation

Pharmacophore modeling was carried out using *3D QSAR pharmacophore generation* protocol. Using the training sets, a set of predictive Pharmacophore QSAR models have been derived for *L. major* inhibitors (mostly diaminopteridines and quinazolines). All the compounds in the training as well as test set were built using the DS 2D/3D visualizer. For each compound, the geometries were corrected, atoms typed and based on the modified CHARMm force field energy minimization was performed [23]. Diverse conformational models for each compound were generated using an energy constraint of 20 kcal mol^{-1} and 250 as the maximum number of conformers.

The *3D QSAR pharmacophore generation* protocol uses the catalyst HypoGen algorithm [24] to derive SAR hypothesis models (pharmacophores) from a set of ligands with known activity values on a given biological target. The input ligands should contain two molecular properties namely, known activity (activ) and uncertainty (uncert). The uncertainty factor for each compound represents the range of uncertainty in the activity value based on the expected statistical distribution of biological data collection. Here, this factor was defined as the default value of three. Four pharmacophoric features HB_donor, HB_ acceptor, Hydrophobic_aromatic and Ring_aromatic were selected to generate a maximum of ten pharmacophores. Remaining parameters were kept as default. The activity of each training set compound is estimated using regression parameters. The parameters were computed by regression analysis using the relationship of geometric fit value and negative logarithm of activity. The greater the geometric fit, the greater the activity prediction of the compound.

Pharmacophore model cross-validation

The main purpose of validating a quantitative model is to cross-check whether the generated hypotheses are correct or not. When generating hypotheses, the cross validation is carried out by calculating cost analysis, RMSD and best fit values simultaneously. The cross validation algorithm attempts to minimize a cost function consisting of three terms: weight cost, error cost, and configuration cost. The overall cost (total cost) of a hypothesis is calculated by summing over three cost factors. Weight cost is a value that increases as the feature weight in a model deviates from an ideal value of two. The deviation between the estimated activities of the training set and their experimentally determined values, adds to the error cost and is reflected in the correlation coefficient *r*. The third term, i.e., the configuration cost, penalizes the complexity of the hypothesis. This is a fixed cost, which is equal to the entropy of the hypothesis space. The more the number of features (a maximum of five) in a generated hypothesis, the higher is the entropy with a subsequent increase in this cost. Further, cross validation was performed by using the Fischer's randomization test where the biological activity data are randomized within a fixed chemical data set and the HypoGen process is initiated to explore possibilities of other hypotheses of good predictive values. For a statistically relevant pharmacophoric model, the hypothesis generated prior to scrambling should be better than rest. In addition to cost analysis of the generated hypothesis, the predictive ability of the pharmacophore models obtained was evaluated using a test set of compounds with *ligand pharmacophore mapping* protocol. This protocol uses catalyst to identify ligands that map to a pharmacophore,

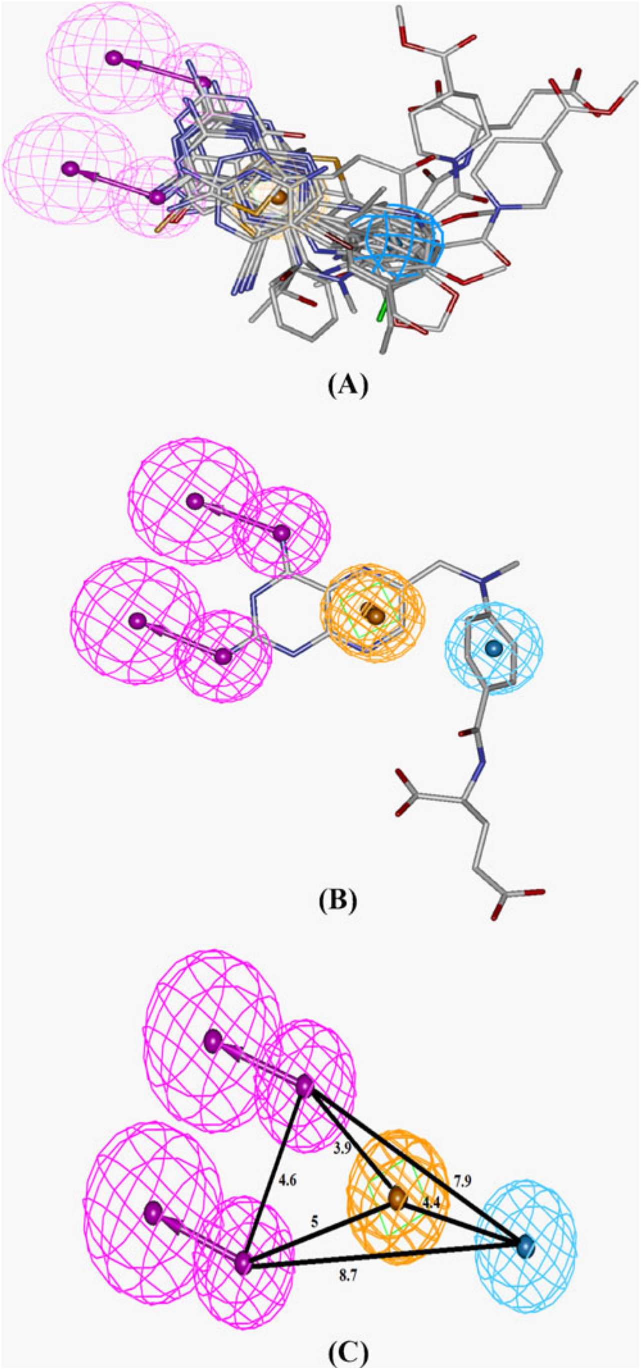


Fig. 1 Hypo1 consists of a hydrophobic aromatic (H, light blue), ring aromatic (R, orange) and H-bond donors (D, magenta) features. (A) Hypo1 is shown aligned to *L. major* inhibitors form the training set. (B) Hypo1 with best fit inhibitory scaffold 1E7W. (C) The distances (in Å) between pharmacophore features of Hypo1 are marked in black lines

and aligns the ligands to the query. Only the best mapping for each ligand was allowed.

Virtual screening for database searching

The pharmacophore model which scored least RMSD value, high 'best fit value', cost difference and lastly able to predict the test set with a high correlation was used as query for the three-dimensional similarity searches. The ZINC chemical database with drug-like compounds consisting of over 13 million structurally diverse small molecules was screened with this query using the *screen library* protocol. A molecule which fits well with the pharmacophoric features of the pharmacophore model was retrieved as a hit. DS (2.0) was also used to further filter out drug-like compounds by applying various filters.

Docking and validation

The crystal structure of receptor PTR1 (PDB ID: 1E7W for *L. major*) were used as the docking template for this study. The receptor molecules were prepared by removing crystallographic water molecules along with other heteroatoms. The chemical structure of NADP was corrected for valency, bond order and alternate conformations. Methotrexate (MTX) was extracted from their PDB complex. The receptor was typed by applying CHARMm Forcefield (version c28) and hydrogen atom positions were optimized using conjugate gradient algorithm. Since NADP was observed in all the DHFR/PTR1 complexes with inhibitors, therefore it was taken as part of receptor for molecular docking studies.

Defining MTX as the ligand, the force field typed and energy minimized receptors were first used to find active sites using the *Find sites as volume of selected ligand* tool. The molecular docking of pteridine reductase inhibitors so obtained was carried out with docking engine LigandFit interfaced with DS (2.0) to further strengthen our predictions. The LigandFit software makes use of a cavity detection algorithm for detecting potential active site regions. A shape comparison filter is combined with a Monte Carlo conformational search for generating ligand poses consistent with the active site shape. Candidate poses are minimized in the context of the active site using a grid-based method for evaluating protein-ligand interaction energies. The docked conformations of the ligands so generated were scored using eleven different scoring functions available with Cscore in Ligandfit, namely, LigScore1, LigScore2, PLP1, PLP2, Jain, PMF, PMF04, DockScore, Ludi I, Ludi II and Ludi III. Since LUDI III scoring function provided the best correlation, this was chosen as the scoring function. LUDI III, an empirical scoring function, is a fast and accurate scoring function originally developed to reproduce the binding affinities of protein-ligand complexes. Additionally, LUDI III has also been found to be the most consistent scoring function that we have used to validate and predict the binding constant for different targets. The relation between the experimental and the predicted binding affinity determined by the Ludi III scores can be expressed by Eq. 1.

$$\textit{Predicted Binding Affinity} = -\log of(\textit{Empirical score}/100) \qquad (1)$$

Validation of docking algorithm

The validation of the docking algorithm is carried out in order to ensure that the docking and the scoring are able to reproduce experimentally determined conformation of the ligand accurately within allowed limits of deviation (2Å). The docking algorithm is verified by comparing the docked conformation with that of experimental one, in a procedure commonly called as control docking. This is done by superimposing both the conformations and calculating the *rmsd*. The lower the *rmsd* values the better is the pose prediction by the docking algorithm.

Table 2 Statistical values of the top ten pharmacophore hypotheses generated by the *3D QSAR pharmacophore generation* protocol for *L. major* inhibitors

Training set						
Hypo no.	Features	Best fit value	Total cost	Δcost	RMSD(Å)	Correlation (r)
1	DDHR	7.6	111.9	45.9	1.0	0.87
2	DDHR	4.81	110.3	47.5	1.0	0.86
3	DDHR	5.87	110.1	49.8	1.0	0.86
4	DDHR	5.77	108.0	52.4	1.0	0.86
5	DDHR	5.06	105.4	52.5	1.0	0.86
6	DDHR	7.09	105.3	58.9	1.05	0.85
7	ADR	5.86	98.9	59.6	1.1	0.82
8	DDHR	5.81	98.2	59	1.1	0.84
9	DDHR	6.95	98.8	61.8	1.1	0.83
10	DDHR	6.95	96.0	61.8	1.1	0.84

Δcost=null cost-total cost; null cost=157.8; fixed cost=115.3; configuration cost=15.6; D-Hydrogen bond donor, H-Hydrophobic aromatic, R-Aromatic

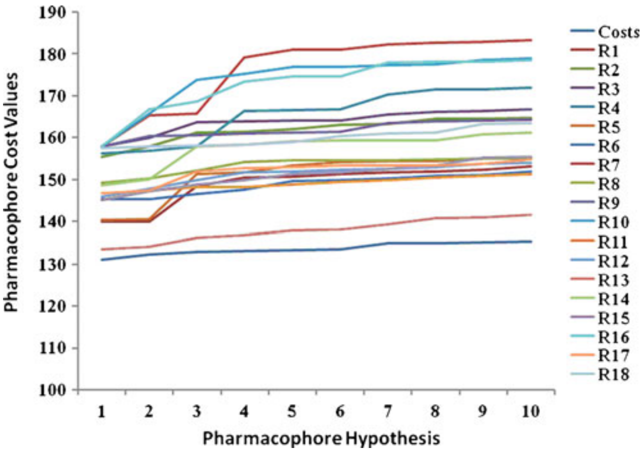


Fig. 2 Fischer's validation performed at 95% confidence level. The difference in costs between Hypogen runs and the scrambled runs is depicted for *L. major* inhibitors

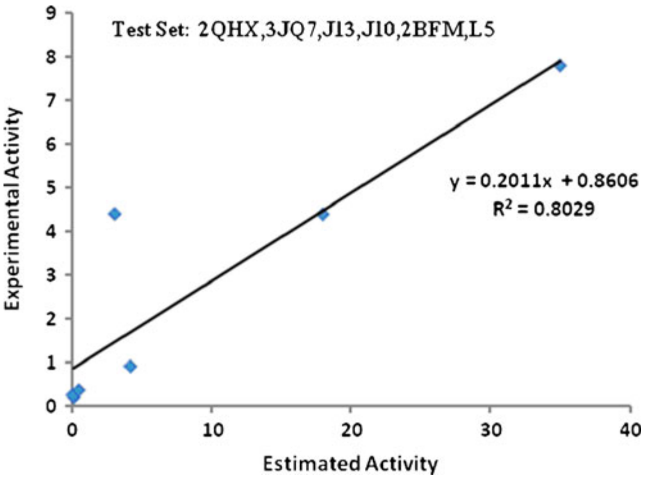


Fig. 4 Correlation plot of between experimental and predicted activities of Hypo1 for *L. major* test set

Results

3D QSAR pharmacophore modeling and validation

Ten 3D-QSAR pharmacophore hypotheses were predicted using the training set comprising diverse scaffolds. Pharmacophore models were generated using the HypoGen algorithm implemented in the *3D QSAR Pharmacophore Generation* protocol. All the ligand molecules were divided into the training as well as the test set (Table 1). These models were validated by applying cost analysis, 'RMSD analysis' (root mean square deviation) and 'best fit' values. First of all the HypoGen module aligns all the molecules to a reference scaffold (here MTX) and selects the best hypotheses from many possibilities by applying a cost analysis. The molecular alignment is standardized and optimized using combination of different pharmacophore features. Finally four distinct chemical features were picked which are able to describe the pharmacophore required for PTR1 inhibition in *L. major*. These features include one or two H-bond donors (D), one or two hydrophobic aromatic (H) and one ring aromatic (R) (Fig. 1a).

A total of ten *3D QSAR pharmacophore* hypotheses were predicted by 28*L. major* inhibitors called as training set (Table 1). Out of ten pharmacophore models, nine possessed the following four chemical features: two hydrogen bond donors (D), one ring aromatic (R) and one Hydrophobic aromatic (H). The best hypothesis with all the molecules aligned on the training set is shown in Fig. 1a. The ligand recognition and the substrate catalysis in the PTR1 is dependent upon the stacking interaction formed by the substrate with the nicotinamide moiety of NADP with Tyr194 and Phe113 residues as well as the hydrogen bonds with Ser111 and Tyr194 present in the active site. The stacking interaction is reflected by the ring aromatic feature of the Hypo1 (Hypothesis I) while the hydrogen bonds with Ser-111 and Tyr-194 are represented by the two H-bond donor features.

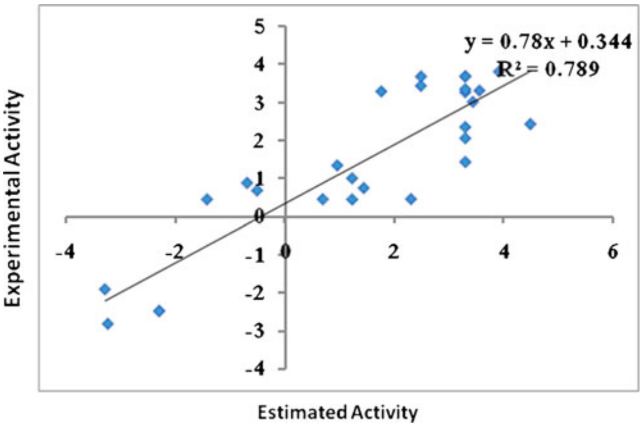


Fig. 3 Correlation plot of between experimental and predicted activities of Hypo1 for training set of *L. major*

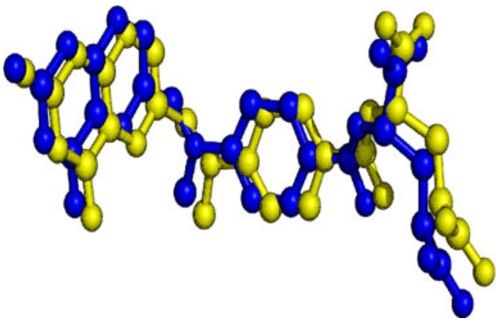


Fig. 5 Structurally superimposed co-crystal (blue) and docked MTX (yellow) displayed in ball & stick representation. The overall *rmsd* is 0.6Å between both the structures

Table 3 Docking scores for the best pose of docked MTX calculated by LigandFit

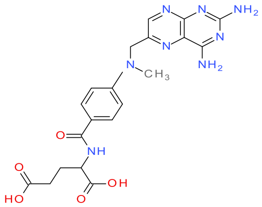
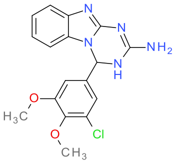
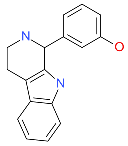
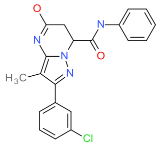
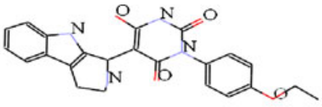
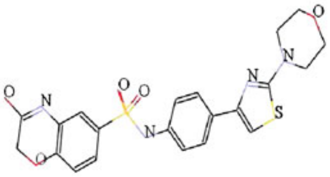
LS1	LS2	PLP1	PLP2	Jain	PMF	PMF04	DockSc-ore	Ludi1	Ludi2	Ludi3	RMSD	Consensus
L. major with MTX												
3.73	5.54	105.78	88.01	1.43	155.8	101.25	34.913	425	381	718	0.69	11

This means that the four chemical features could effectively map the chemical features of the training set compounds. All ten predictive *3D QSAR pharmacophore* hypotheses derived for *L. major* inhibitors and the relevant statistics are shown in Table 2.

Cost analysis

The significance of the selected hypothesis was estimated by comparing the cost values which should lie between the null and the fixed costs. The hypothesis closest to the fixed

Table 4 Virtual screening results of PTR1 inhibitors identified from zincdatabase (zinc-ID)

S. No.	**Zinc-ID**	**Structure**	**Estimate d activity (by Eq. 1)**
***L. major* Inhibitors**			
1.	**MTX**		0.06μM
2.	**ZINC07127727**		0.12nM
3.	**ZINC19688794**		4.7nM
4.	**ZINC32101443**		2.3nM
5.	**ZINC17015141**		0.06nM
6.	**ZINC24262782**		0.15nM

cost and farthest from the null cost is considered to be statistically more significant. The fixed cost is a representation of the cost of the ideal theoretical hypothesis which may predict the activity of compounds used in the training set whereas the null costs consider the cost of each hypothesis with no features such that each activity is taken to be the average activity. In this study the null cost of top ten hypotheses is 157.8 and fixed cost is 115.3 with a difference of 45.9 (Table 2) and this may lead to a meaningful pharmacophore model. The configuration cost value for hypothesis I is 15.6. This value is well within the range observed for a standard HypoGen model where it should not be greater than 17.0. In simple terms, there should be a large difference between fixed cost and null cost with a value of 40–60 bits, which would imply a 75-90% probability of correlating the experimental and estimated activity data. In this study, all ten hypotheses have total cost close to the fixed cost and thus the hypothesis is significant.

RMSD and best fit values

The root mean square deviation (RMSD) is an indication of the quality of 'prediction' and depicts deviation in the alignment of the inhibitory scaffolds. The lower rmsd value is indication of better alignment of molecules. The RMS deviation of all ten hypotheses ranged from 1.0 Å-1.1 Å, which is an acceptable range (Table 2). 'Best fit' value indicates the overall fitness of all the training set compounds on a particular pharmacophore model during pharmacophore generation. These values shown in Table 2 for each pharmacophore hypothesis lie within the acceptable range. The best fit molecule (1E7W) superposed on the pharmacophore features of Hypothesis 1 of Table 2 is shown in Fig. 1b with the inter feature distances shown in Fig. 1c.

Fischer's randomization test

Additionally, cross validation was carried out by applying Fischer's randomization test at 95% confidence level. Out of a total of 19 random hypothesis generated, one scramble run did not produce any valid hypotheses. Thus these were automatically excluded from the data tables. Out of remaining 18 runs only single runs had cost close to our hypotheses, and were thus left out as outliers. The total costs of random pharmacophore models obtained from the Fischer's randomization test as well as original cost are shown in Fig. 2. The original pharmacophore hypothesis is superior to those produced randomly as observed in the figure where the original hypothesis (lowermost) statistics is unique among all the random hypothesis (R1-R18). These results further provide confidence on pharmacophore Hypothesis I (Hypo 1) as the best hypothesis.

Besides the statistical relevance, the predictive ability of the Hypo1 was checked by its ability to predict the activity of the training set compounds as well as an external set of compounds called as test set. The test set compounds exhibit wide range of activity and structural diversity (see Material and Methods Section). In conclusion, Hypo1 is able to correctly predict activity of most of the training set compounds with a correlation value of 0.78 and test set compounds with a correlation value of 0.80. The correlation between the experimental and the predicted activities by hypothesis I for the training and the test sets are shown in Fig. 3 and Fig. 4 respectively.

Four out of ten hypotheses were able to produce a maximum of four features; out of them hypotheses 1 (Hypo1) had the best fit value of 7.6, least RMSD of 1.0 Å and highest correlation of 0.87. Hypo1 mapped on all four features: two H-bond donors (D), one hydrophobic aromatic (H) and one ring aromatic (R) and predicted test set compounds with a high correlation and best fit value. It was able to map all the compounds in the test set. The Hypo1 pharmacophore hypothesis satisfies all the validation criteria and can be considered as statistically significant. The Hypo1 mapped onto all the *L. major* inhibitory scaffolds present in the training set as shown in Fig. 1a. Hence, Hypo1 was chosen as the best hypotheses and later used as the template for database searching in all the virtual screening experiments.

ZINC database screening with hypo1, filtering and docking

Hypo1 from *L. major* predictive 3D-QSAR pharmacophores were used to screen ZINC database drug-like

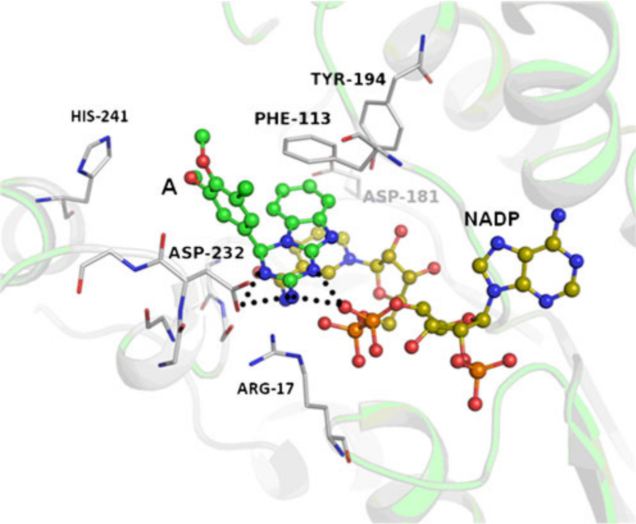


Fig. 6 Molecular interactions of the inhibitor with in the PTR1 binding pocket of *L. major* bound with ligand ZINC07127727 labeled as A. Hydrogen bonds are marked by dashed lines. The inhibitor is sandwiched between NADP from one end and Phe113 and Tyr194 from other end in *L. major* PTR1

compounds consisting of over 1 million compounds with the *screen library* protocol. In this study only 80,000 of the zinc compounds were used for the screening purpose.

L. major Hypo1 screened out a total of about 400 compounds in which all four pharmacophoric features of Hypo1 were present. These screened molecules were then subjected to *prepare ligand* protocol to generate their 3D structure and to filter out non-drug like compounds. This resulted in generation of total 1330 of these ligand poses. In order to reduce the chance of false positives, the ligand poses so obtained were further prioritized and filtered by docking with *LigandFit* docking engine. For the comparative purpose, the control docking experiments were performed with MTX as inhibitors from *L. major* PTR1-MTX co-complex (PDB code: 1E7W). The docking site was defined by selecting the volume of MTX as ligand. Since a large part of pteridine binding site is occupied by nicotinamide which is essential for ligand recognition and catalysis, the docking was carried out by retaining NADP in the binding site. The Ludi-III was used to predict binding affinities of the hits identified by virtual screening with Zinc databases as well for control docking experiments. The second pose with the least RMSD value of 0.6Å was selected as the best pose. The superimposition of this pose onto the crystal structure of the same is indicated in Fig. 5. The second pose has the highest consensus for the docking scores and is shown in Table 3. The predicted binding affinity calculated from Eq. 1, for MTX is 0.06 μM which is in good agreement with the experimental binding affinity of 0.039 μM (Table 1).

In the next step, the high throughput molecular docking of screened inhibitors were performed with *L. major* PTR1 using the same parameters as that of control docking run. Based on Ludi-III scores, 20 inhibitor molecules were finally screened out as potential PTR1 inhibitors. Out of these, only five compounds having estimated activity in nano-molar range are reported here and shown in Table 4. The binding affinities are calculated from Ludi III scores by Eq. 1. All these compounds have the benzimidazole and indole like moieties. These basic parent scaffold constituents are similar to the substrate (pterins and folates) like pyrrolo[2,3-d] pyrimidine framework of the PTR1 enzyme. On the basis of the molecular interactions and the presence of all the pharmacophore features, the inhibitor ZINC07127727 can be proposed as the best inhibitor of *L .major* PTR1. The docking pose of the best inhibitors in the *L. major* PTR1 binding site is shown in Fig. 6.

Conclusions

3D QSAR pharmacophore modeling approach used in combination with pharmacophore-based virtual screening has led to the identification of four different scaffolds as novel Pteridine reductase inhibitors from *L. major*. The pharmacophore model were generated from the recently reported series of PTR1 inhibitors. The 3D QSAR hypothesis suggests that the hydrogen bond donor, aromatic and hydrophobic_aromatic are the essential features required for the inhibitory activity against PTR1. The presence of two hydrogen bond donors in the predictive pharmacophore model (Hypo1) indicates the importance of these interactions in the PTR1 ligand recognition and catalysis.

In addition to the identification of pharmacophores, the best pharmacophore hypothesis was used to identify similar drug-like compounds from ZINC database for *L. major*. The hits so obtained after similarity searches were further prioritized by molecular docking approach using LigandFit docking tool. The docking protocol and scoring functions were cross validated by control docking experiments. Amongst all the inhibitors identified, the compound ZINC07127727 from Zinc database showed the highest predicted K_d value of 0.06nM for *L .major* PTR1. The predicted K_d value is more than 300 fold higher than that of the presently known inhibitor methotrexate. Thus, these identified inhibitors could be promising lead compounds that can be further taken up for the development of novel therapeutic entities against the trypanosomatid parasite *Leishmania.* Overall, the current study is an attempt for elucidating structure-activity relationship for PTR1 inhibitors. This study has led to the understanding of the protein–ligand interactions involved between the enzyme and its inhibitors and could probably aid in the development of highly selective and potent inhibitors against PTR1 as anti-parasitic agents.

Acknowledgments Authors acknowledge Indian Council of Medical Research, India for grant of Bio-Medical Informatics Centre. DD thanks Department of Science & Technology, India for financial support.

References

1. Bello AR, Nare B, Freedman D, Hardy L, Beverley SM (1994) PTR1: a reductase mediating salvage of oxidized pteridines and methotrexate resistance in the protozoan parasite Leishmania major. Proc Natl Acad Sci USA 91:11442–11446
2. Nare B, Hardy LW, Beverley SM (1997) The roles of pteridine reductase 1 and dihydrofolate reductase-thymidylate synthase in pteridine metabolism in the protozoan parasite Leishmania major. J Bio Chem 272:13883–13891
3. Nare B, Luba J, Hardy LW, Beverley S (1997) New approaches to Leishmania chemotherapy: pteridine reductase 1 (PTR1) as a target and modulator of antifolate sensitivity. Parasitology 114 (Suppl):S101–110
4. Soto J, Arana BA, Toledo J, Rizzo N, Vega JC, Diaz A, Luz M, Gutierrez P, Arboleda M, Berman JD, Junge K, Engel J, Sindermann H (2004) Miltefosine for new world cutaneous leishmaniasis. Clin Infect Dis 38:1266–1272
5. Arevalo J, Ramirez L, Adaui V, Zimic M, Tulliano G, Miranda-Verastegui C, Lazo M, Loayza-Muro R, De Doncker S, Maurer A,

Chappuis F, Dujardin JC, Llanos-Cuentas A (2007) Influence of Leishmania (Viannia) species on the response to antimonial treatment in patients with American tegumentary leishmaniasis. J Infect Disease 195:1846–1851

6. Badaro R, Lobo I, Munos A, Netto EM, Modabber F, Campos-Neto A, Coler RN, Reed SG (2006) Immunotherapy for drug-refractory mucosal leishmaniasis. J Infect Disease 194:1151–1159
7. Sundar S, Chakravarty J, Agarwal D, Rai M, Murray HW (2010) Single-dose liposomal amphotericin B for visceral leishmaniasis in India. N Engl J Med 362:504–512
8. Sundar S, Jha TK, Thakur CP, Sinha PK, Bhattacharya SK (2007) Injectable paromomycin for Visceral leishmaniasis in India. N Engl J Med 356:2571–2581
9. Gourley DG, Schuttelkopf AW, Leonard GA, Luba J, Hardy LW, Beverley SM, Hunter WN (2001) Pteridine reductase mechanism correlates pterin metabolism with drug resistance in trypanosomatid parasites. Nat Struct Biol 8:521–525
10. Sienkiewicz N, Ong HB, Fairlamb AH (2010) Trypanosoma brucei pteridine reductase 1 is essential for survival in vitro and for virulence in mice. Mol Microbiol 77:658–671
11. Tulloch LB, Martini VP, Iulek J, Huggan JK, Lee JH, Gibson CL, Smith TK, Suckling CJ, Hunter WN (2010) Structure-based design of pteridine reductase inhibitors targeting African sleeping sickness and the leishmaniases. J Med Chem 53:221–229
12. Cavazzuti A, Paglietti G, Hunter WN, Gamarro F, Piras S, Loriga M, Allecca S, Corona P, McLuskey K, Tulloch L, Gibellini F, Ferrari S, Costi MP (2008) Discovery of potent pteridine reductase inhibitors to guide antiparasite drug development. Proc Natl Acad Sci USA 105:1448–1453
13. Schüttelkopf AW, Hardy LW, Beverley SM, Hunter WN (2005) Structures of Leishmania major pteridine reductase complexes reveal the active site features important for ligand binding and to guide inhibitor design. J Mol Biol 352:105–116
14. McLuskey K, Gibellini F, Carvalho P, Avery MA, Hunter WN (2004) Inhibition of Leishmania major pteridine reductase by 2,4,6-triaminoquinazoline: structure of the NADPH ternary complex. Acta Crystallogr D Biol Crystallogr 60:1780–1785
15. Shanks EJ, Ong HB, Robinson DA, Thompson S, Sienkiewicz N, Fairlamb AH, Frearson JA (2010) Development and validation of a cytochrome c-coupled assay for pteridine reductase 1 and dihydrofolate reductase. Anal Biochem 396:194–203
16. Ferrari S, Morandi F, Motiejunas D, Nerini E, Henrich S, Luciani R, Venturelli A, Lazzari S, Calò S, Gupta S, Hannaert V, Michels PA, Wade RC, Costi MP (2010) ScreeningiIdentification of nonfolate compounds, including a CNS drug, as antiparasitic agents inhibiting pteridine reductase. J Med Chem 54:211–221
17. Mpamhanga CP, Spinks D, Tulloch LB, Shanks EJ, Robinson DA, Collie IT, Fairlamb AH, Wyatt PG, Frearson JA, Hunter WN, Gilbert IH, Brenk R (2009) One scaffold, three binding modes: novel and selective pteridine reductase 1 inhibitors derived from fragment hits discovered by virtual screening. J Med Chem 52:4454–4465
18. Spinks D, Ong HB, Mpamhanga CP, Shanks EJ, Robinson DA, Collie IT, Read KD, Frearson JA, Wyatt PG, Brenk R, Fairlamb AH, Gilbert IH (2011) Design, synthesis and biological evaluation of novel inhibitors of Trypanosoma brucei pteridine reductase 1. Chem Med Chem 6:302–308
19. Chen JJ, Liu TL, Yang LJ, Li LL, Wei YQ, Yang SY (2009) Pharmacophore modeling and virtual screening studies of check-point kinase 1 inhibitors. Chem Pharm Bull 57:704–709
20. Thangapandian S, Krishnamoorthy NK, John S, Sakkiah S, Lazar P, Lee Y, Lee KW (2010) Pharmacophore modeling, virtual screening and molecular docking studies for identification of new inverse agonists of human histamine H 1 receptor. Bull Korean Chem Soc 31:52–58
21. Irwin JJ, Shoichet BK (2005) ZINC–a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182
22. Venkatachalam CM, Jiang X, Oldfield T, Waldman M (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model 21:289–307
23. Kumar M, Verma S, Sharma S, Srinivasan A, Singh TP, Kaur P (2010) Structure-based in silico design of a high-affinity dipeptide inhibitor for novel protein drug target Shikimate kinase of Mycobacterium tuberculosis. Chem Biol Drug Des 76:277–284
24. Güner OF (2000) Pharmacophore perception, development, and use in drug design. International University Line, La Jolla, CA

J Mol Model (2012) 18:1713–1722
DOI 10.1007/s00894-011-1194-1

ORIGINAL PAPER

3D modeling and molecular dynamics simulation of an immune-regulatory cytokine, interleukin-10, from the Indian major carp, *Catla catla*

Bikash R. Sahoo · Banikalyan Swain · Madhubanti Basu · Padmaja Panda · Nikhil K. Maiti · Mrinal Samanta

Received: 12 March 2011 / Accepted: 19 July 2011 / Published online: 10 August 2011

Abstract Interleukin-10 (IL-10) is a pleiotropic immune-regulatory cytokine that is expressed in various species of fish and higher vertebrates, and is activated during infection. In spite of its important role, IL-10 has not been well characterized either functionally or structurally in fish. To analyze its properties and function, we constructed a 3D model of IL-10 in the Indian major carp, the catla (*Catla catla*), which is a highly preferred fish species and the most commercially important one in the Indian subcontinent. The catla IL-10 model was constructed by comparative modeling using human IL-10 (2ILK) as the template, and a 5 ns molecular dynamics (MD) simulation was carried out to characterize its structural and dynamical features, which was validated by the SAVES, WHAT IF and MolProbity servers. Analysis using the VAST server revealed a comparatively low level of homology between catla and human IL-10 amino acids at the N-terminal (22.7%) compared to the C-terminal (38.29%). Six conserved domains (A–F) were predicted in catla that threaded well with human IL-10, but their putative interaction sites varied significantly. The amino acid residues in helices A and F differed in length between catla and human IL-10, which may lead to the differences in the IL-10/IL-10R complexes of these two species. The existence of two highly conserved amino acid residues (Cys5 and Cys10) in fish IL-10 but not in higher vertebrate (including human) IL-10 was analyzed in this 3D model. CastP, cons-PPISP and InterProSurf server identified several binding pockets with various probe radii, but Cys5 and Cys10 did not form any significant bonds relating to structural stabilization or protein–protein interactions.

Keywords Indian major carp · *Catla catla* · IL-10 · Comparative modeling · Molecular dynamics

Introduction

Cytokines are low molecular weight proteins or glycoproteins that play a vital role in immunity by initiating and regulating the inflammatory processes. Interleukin-10 (IL-10) is a pleiotropic cytokine that influences the activities of many cell types in the immune system and has important immune regulatory functions [1]. It is produced by activated T cells, B cells, monocytes/macrophages, mast cells, and keratinocytes [2]. IL-10 was initially described as a cytokine synthesis inhibitory factor (CSIF) that shifts the body's immune reaction away from the inflammatory response induced by a pathogen or by the immune system [3]. It aids in the regulation, differentiation and proliferation of several immune cells, such as T cells, B cells and natural killer cells [4], and also mediates immune-stimulatory properties in order to eliminate infectious and noninfectious particles with limited inflammation [5].

India ranks third among the world's freshwater fish producers, and among the freshwater-cultured fish species, the catla (*Catla catla*) is the most commercially important and highly favored fish, so Indian aquaculture is greatly dependent on catla. Recently we cloned and sequenced the catla IL-10 gene, analyzed its expressional upregulation

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1194-1) contains supplementary material, which is available to authorized users.

B. R. Sahoo · B. Swain · M. Basu · P. Panda · N. K. Maiti · M. Samanta (✉)
Fish Health Management Division,
Central Institute of Freshwater Aquaculture (CIFA),
Kausalyaganga,
Bhubaneswar, Orissa, India 751002
e-mail: msamanta1969@yahoo.com

during diseases, and identified the presence of an IL-10-dependent signaling pathway that partially downregulates the expression of pro-inflammatory cytokines and is closely related to higher vertebrates.

IL-10 binds to its high-affinity receptor IL-10R1 [6], and exerts its biological activity as a homodimer consisting of two noncovalently linked monomers [7]. Each subunit is stabilized by two intra-chain disulfide bonds. Catla IL-10 contains six cysteine residues. Among these, four are common to fish, birds and mammals, and the other two cysteine residues are highly conserved across the fish species and are not present in higher vertebrates.

Considering these facts together, we investigated the sequence-based prediction of the 3D structure of mature catla IL-10 in order to analyze the properties and functions of the protein, as well as to determine the significance of two highly conserved cysteine residues in fish that may play a significant role in forming disulfide bridges during protein folding or may possess high solvent accessibility potential for the interaction with the IL-10 receptor.

Materials and methods

Computational resources

Computational studies were carried on a Pentium 4, 2.40 GHz PC equipped with the CentOS environment. Sequence alignment was carried out using Clustal W [8], and was displayed with ESPript 2.2 [9]. Comparative modeling was performed with the MODELLER 9v8 program (academic version) [10]. Loop refinement was carried out using the same MODELLER 9v8 program manually, without any constraint refinement in the alignment. Molecular dynamics (MD) simulations, energy minimization, and trajectory analysis were carried out using GROMACS 4.0.3 (http://www.gromacs.org) and the GROMOS9643a1 force field. The modeled protein was validated by PROCHECK [11], Verify3D [12], ERRAT [13] and PROVE [14] by SAVES (the Structure Analysis and Verification Server) (http://nihserver.mbi.ucla.edu). The feasibility of the structure was also investigated using the WHAT IF [15] and MolProbity [16] servers. Protein visualization and superimposition was carried out using Discovery Studio Visualizer 2.0 (http://accelrys.com/) and PyMOL (educational license version) (http://www.pymol.org/). All graphs were created using Xmgr 4.1.2 (http://plasma-gate.weizmann.ac.il/Xmgr/).

Comparative modeling of catla IL-10

The full length IL-10 gene of catla was cloned by RACE (5′ and 3′) from the EST sequence (GenBank acc. no.: GU256643); after aligning the 5′ and 3′ sequences, an 1137-bp cDNA sequence was obtained (GenBank acc. no. HQ221996). The open reading frame (ORF) consisted of 540 bp nucleotides that encoded 179 amino acids. The signal peptide (22 aa) was predicted using the SignalP 3.0 server [17], and was excluded from model building. Mature catla IL-10 consisted of 157 amino acid residues with a predicted molecular mass of 18.50 kDa, and was used for template searching. The template search, based on the functional domains concept, was carried out in GeneSilico Metaserver [18], 3D-Jury [19] and Pcons.net [20]. All of these searches suggested that the human IL-10 crystal structure from the Protein Data Bank (PDB code: 2ILK) at 1.6Å resolution was the best template for catla IL-10, with ~29% identity and 57% positives. We also searched for a template using the MODELLER 9v8 program, and it also indicated that 2ILK was the best template, with 28.86% identity, an expected value of 0, and a query coverage of ~99%. Although viral interleukin-10 (PDB code: 1VLK) showed a slightly higher level of identity (30.50%) than 2ILK in MODELLER 9v8 program, its query coverage (92%) and alignment score were lower.

The protein threading approach used by I-TASSER [21] was also employed to determine the best template in terms of fold recognition. It predicted a catla IL-10 model that utilized 2ILK as the best template (rank 1). To ensure the sensitivity and accuracy of our selected template, the FUGUE (Find Homologs of Uncharacterized Gene Products Using Environment-specific substitution tables) program [22] was used to perform a sequence–structure comparison between the target and template. FUGUE utilizes a curated database of structure-based alignments for homologous protein families, HOMSTRAD (Homologous Structure Alignment Database) [23]. From the FUGUE search, the HOMSTRAD family with the top Z-score against the cut-off score (Z-score>6.0) was considered to be the best template for modeling. The result was presented using JOY annotation [24]. Using this program, the highest Z-score of 24.30 was obtained with the same 2ILK template, followed by 1M4R (interleukin 22, *Homo sapiens*), with a Z-score of 17.85.

After validating that 2ILK is the most appropriate template, 10 models were generated by MODELLER, and loop refinement was carried out locally. The model with the lowest value of the normalized discrete optimized protein energy (DOPE) was chosen as the best model, and subjected to molecular dynamics (MD) simulation.

Swiss-Pdb Viewer [25] was implemented for energy minimization, using a harmonic constraint of 100 kJ mol^{-1} $\mathrm{Å}^{-2}$ and the steepest descent and conjugate gradient techniques along with the GROMOS 43B1 force field [26]. The refined model was validated with PROCHECK, Verify3D, ERRAT and PROVE in order to confirm that all bond lengths, dihedral angles and torsion angles attain a stable configuration. The structure was also checked using the WHAT IF and MolProbity servers.

Molecular dynamics (MD) simulations

MD simulations were conducted for the modeled systems in explicit solvent using the GROMACS (Groningen Machine for Chemical Simulations) 4.0.3 package [27]. The model was solvated by 15,735 water molecules in an octahedral box with edges that were 0.9 nm from the molecular boundary. To obtain a neutral system, three CL^- ions were added (charge −3.00) to the catla IL-10 model (which has a net positive charge of 3.00). The solvated system was then subjected to further energy minimization (maximum number of steps: 2000) to remove steric conflicts between the protein and water molecules, using the steepest descent integrator. Convergence was achieved in the energy minimization when the maximum force was smaller than 1000 kJ mol^{-1} nm^{-1}. The energy-minimized model was subjected to position-restrained MD under NPT conditions, keeping the number of particles (N), the system pressure (P) and the temperature (T) constant. This was carried out for 50,000 steps for a total of 100 ps. The reference temperature for coupling (via Berendsen temperature coupling) was 300 K, and a pressure of 1 atm was maintained by the Parrinello–Rahman algorithm. Snapshots of the trajectory were taken every 0.5 ps. The final MD of 2,500,000 steps was carried out for 5,000 ps (5 ns) using the particle mesh Ewald (PME) electrostatics method under NPT conditions. The final model was set for validation by the SAVES server.

The human IL-10 dimer was built by performing a symmetry transformation on the atoms of the monomer using the tool Build Crystallographic Symmetry implemented in the Swiss-Pdb Viewer software. The Protein–Protein Interaction Server [28] and NACCESS v.2.1.1 (Atomic Solvent Accessible Area Calculations) program [29] were used to identify the amino acids at the protein–protein interface of the complex.

Pocket information for catla IL-10 was predicted by CASTp [30] (using the default settings), cons-PPISP [31] and the InterProSurf server [32], in order to locate the binding-site amino acids.

Results and discussion

Model building

To create the model of catla IL-10, we followed a comparative modeling protocol and evaluated several templates using sequence comparison and functionally conserved domain search methods. We found that 2ILK (human IL-10) was the best template, in agreement with the previous observations of Pinto et al. [33]. The fold recognition approach employed in I-TASSER also identified that 2ILK was the best template using the TM-align program [34]; it had a TM score of 0.9397 (the TM score lies within (0,1) and is >0.5, meaning that the structures share the same fold), an RMSD of 0.92, a sequence identity of 28%, and a query coverage of 97%. The error associated with creating models using only sequence comparison was minimized with this approach. The percentage (%) identity and the alignment scores of 2ILK (the template) obtained by different programs are presented in Table 1. The alignments of the catla IL-10 protein with the IL-10 proteins of another eleven species are shown in Fig. 1, and additionally conserved cysteine residues (Cys5 and Cys10) in fish are indicated on this figure. The sequence–structure alignment of catla IL-10 and 2ILK was also generated in FUGUE and presented by the JOY annotation program (Fig. 2), which indicated the possible regions of helices, turns and β-sheets of catla IL-10 with reference to the template, and was validated by a secondary structure comparison.

The secondary structure of the catla IL-10 amino acid sequence was predicted by PSIPRED [35] (Fig. 3a), and the secondary structure of the catla IL-10 model was assigned by the STRIDE program [36], which used hydrogen-bond energies and main chain dihedral angles to identify helices, coils and strands; this secondary structure is shown in Fig. 3b. The helical positions of catla IL-10 were aligned with human IL-10, and this accurately highlighted the conserved domains, with approximately the same percentage of helical residues obtained for both the sequence-based and the 3D model based secondary structure approaches. This signified that the fits to both of the secondary structures were good.

Catla IL-10 protein model validation

To validate the catla IL-10 model, a Ramachandran plot was drawn and the structure was analyzed by PROCHECK. It

Table 1 Sequence identity of the 2ILK template, as obtained by different servers

Tools	% Identity	Score
pdbblast	29 (2nd)	1.06892×10^{-46}
csblast	29 (top)	7.00×10^{-41}
prc	29 (2nd)	1.4×10^{-38}
blastp	29 (top)	3.87513×10^{-15}
mGenthreader	28 (top)	1.00×10^{-7}
sp3	28 (top)	−351.694

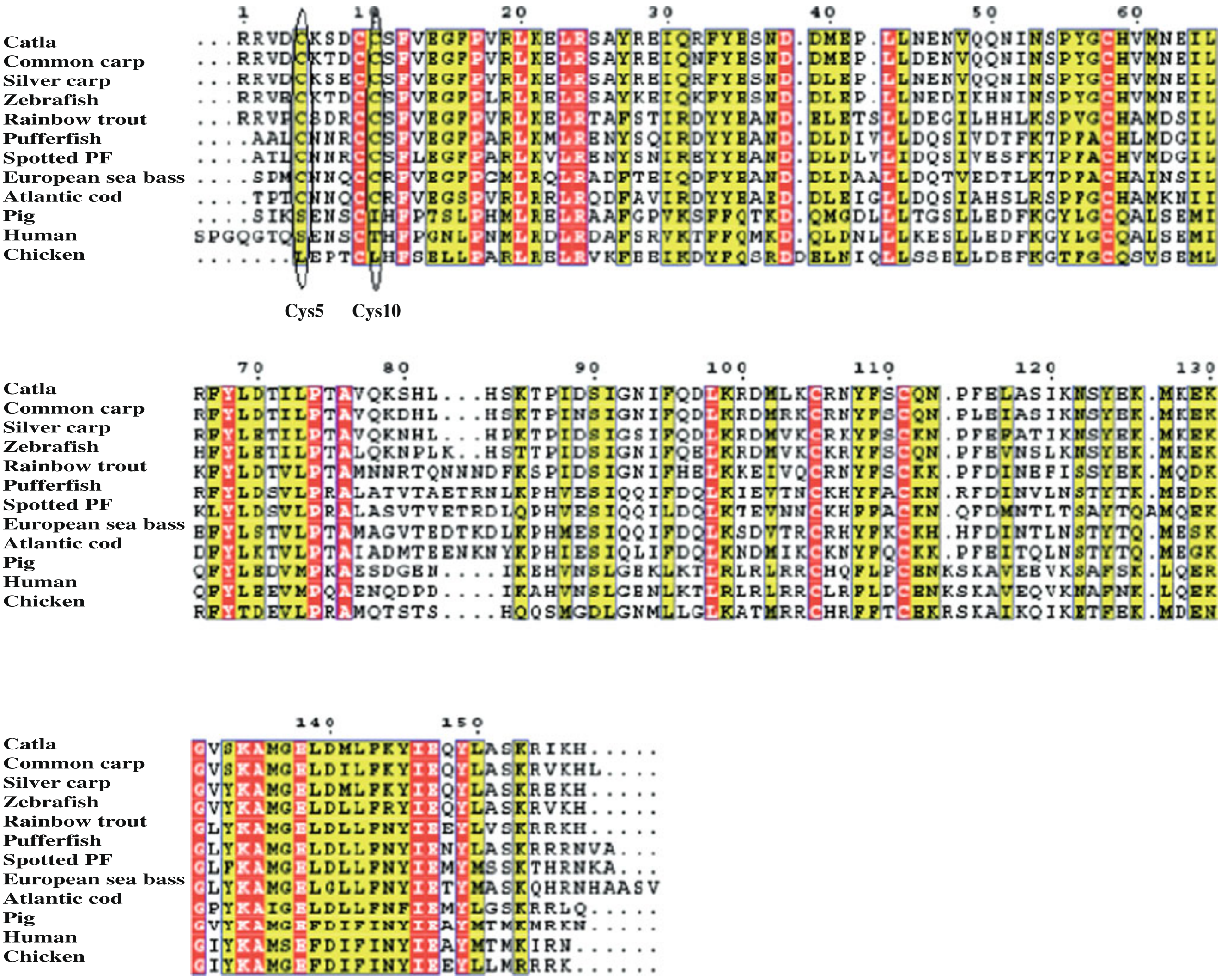


Fig. 1 Multiple sequence alignment of catla IL-10 with the IL-10 proteins of other species. Amino acid sequences were retrieved from the GenBank data base, aligned with Clustal W, and the block was prepared using the ESPript 2.2 program. Conserved amino acids are shown in *red boxes* and >50% consensus residues are shown in *yellow boxes*. Two additional cysteine residues (Cys5 and Cys10) that are conserved in fish are indicated by *circles* and labeled. GenBank accession numbers: *Catla catla* (catla) HQ221996, *Cyprinus carpio* (common carp) BAC76885, *Hypophthalmichthys molitrix* (silver carp) AAY99196, *Danio rerio* (zebrafish) AAW78362, *Oncorhynchus mykiss* (rainbow trout) BAD20648, *Takifugu rubripes* (*Fugu rubripes*) CAD62446, *Tetraodon nigroviridis* (spotted green pufferfish) CAD67786, *Dicentrarchus labrax* (European sea bass) ABH09454, *Gadus morhua* (Atlantic cod) ABV64720, *Sus scrofa* (pig) ABP68816, *Homo sapiens* (human) CAG46825, and *Gallus gallus* (chicken) Q6A2H4

was observed that the phi–psi angles of 86.8% of the residues were in the most favored regions, 11.8% were in the additional allowed regions, 1.4% were in the generously allowed regions, and no residues fell in the disallowed regions (Fig .4). The overall G-factor scores for the catla IL-10 model after MD simulation was −0.17. This indicated that the model was accepted, as it was greater than the recommended value (−0.50). The Verify 3D results for the catla IL-10 model showed that 45.57% of the amino acids had an average 3D-1D score of >0.2, and 89.18% of the residues showed positive scores (cut-off score was >0), indicating the reliability of the proposed model. The catla IL-10 model was analyzed by the PROVE program to measure the average magnitude of the volume irregularities in terms of the Z-score root mean square deviation (Z-score RMS). The Z-score RMS values for the catla and human IL-10 models were 1.398 and 1.139, respectively (a Z-score RMS value of ~1.0 indicates good resolution of structures). The overall quality of the catla IL-10 model determined by the ERRAT program was 97.917 (a value of ~95% shows high resolution). All of these programs together indicated that the catla IL-10 protein model was valid. The average coarse packing quality, anomalous bond length, planarity, packing quality and the collision with symmetry axis obtained by the WHAT IF server showed good acceptance of the model. The MolProbity server predicted that 0% of

Catla RRVDCKSDCCSFVEGFPVRLKELRSAYREIQRFYESNDDMEPLLNENVQQ

2ILK TQSENSÇTHFPGNLPNMLRDLRDAFSRVKTFFQMKDQLDNLLLKESLLED

Catla NINSPYGCHVMNEILRFYLDTILPTAVQKSHLHSKTPIDSIGNIFQDLKR

2ILK FKGYLGÇQALSEMIQFYLEEVMPQAENQDPDIKAHVNSLGENLKTLRLRL

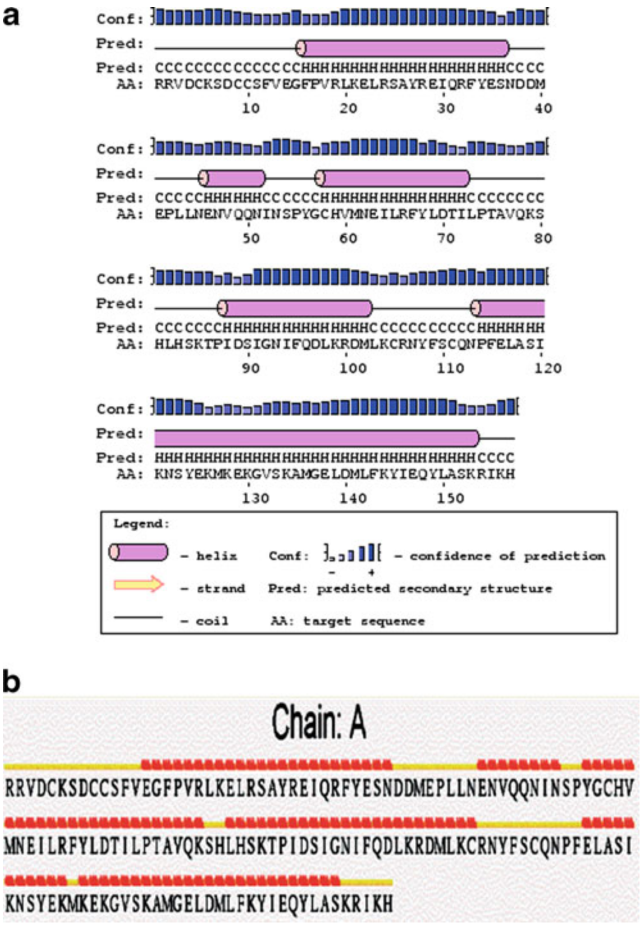
Catla DMLKCRNYFSCQNPFELASIKNSYEKMKEKGVSKAMGELDMLFKYIEQYL

2ILK RRÇHRFLPÇENKSKAVEQVKNAFNKLQEKGIYKAMSEFDIFINYIEAYMT

Catla ASKRIKH

2ILK MKIRN

Fig. 2 Sequence–structure alignment of the target and template. The alignment of the catla IL-10 sequence (target) with the structure of 2ILK (template) was performed by FUGUE, and the results were presented with the JOY annotation program. Alpha-helices are shown in *red*, *underlined residues* represent H bonds to the main chain, *ç* indicates a disulfide bridge, and *italicized letters* represent positive phi torsion angles



```
Conf:
Pred:
Pred: CCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHCCCC
  AA: RRVDCKSDCCSFVEGFPVRLKELRSAYREIQRFYESNDDM
              10        20        30        40

Conf:
Pred:
Pred: CCCCCHHHHHHCCCCCCHHHHHHHHHHHHHHHHCCCCCCC
  AA: EPLLNENVQQNINSPYGCHVMNEILRFYLDTILPTAVQKS
              50        60        70        80

Conf:
Pred:
Pred: CCCCCCCHHHHHHHHHHHHHHHHCCCCCCCCCCHHHHHHH
  AA: HLHSKTPIDSIGNIFQDLKRDMLKCRNYFSCQNPFELASI
              90       100       110       120

Conf:
Pred:
Pred: HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCC
  AA: KNSYEKMKEKGVSKAMGELDMLFKYIEQYLASKRIKH
              130       140       150

Legend:
 - helix     Conf: - confidence of prediction
                     -   +
 - strand    Pred: predicted secondary structure
 - coil      AA: target sequence
```

RRVDCKSDCCSFVEGFPVRLKELRSAYREIQRFYESNDDMEPLLNENVQQNINSPYGCHV

MNEILRFYLDTILPTAVQKSHLHSKTPIDSIGNIFQDLKRDMLKCRNYFSCQNPFELASI

KNSYEKMKEKGVSKAMGELDMLFKYIEQYLASKRIKH

Fig. 3a–b Secondary structure of catla IL-10. **a** Predicted secondary structure of the catla IL-10 amino acid sequence obtained by PSIPRED. *H* denotes a helical region and *C* denotes a coil region. **b** Secondary structure assignment of the catla IL-10 model by STRIDE. *Red* indicates a helical domain and *yellow* indicates a loop region

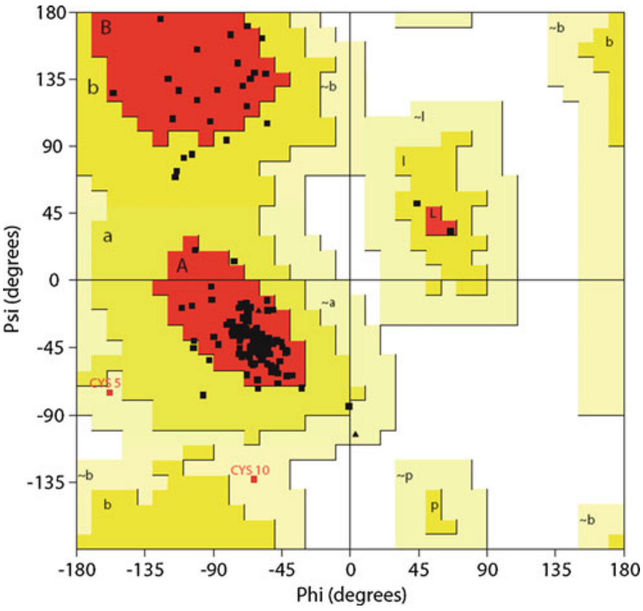


Fig. 4 Ramachandran plot of the catla IL-10 model. The plot was calculated with the PROCHECK program

the residues had bad bonds (goal 0%), 0% of the residues had bad angles (goal<0.1%), and that 0% of the Cβ deviations were >0.25 Å (goal 0%). This further strengthened the reliability of the catla IL-10 model. The 3D models of catla and human IL-10 (2ILK) were depicted graphically by the PyMOL visualization tool along with the sequences and disulfide bridges (Fig. 5).

To further validate the accuracy of the structure of the homolog and the methods used to generate the 3D model of the target sequence, a cross-check validation approach was used. In this strategy, the catla IL-10 model was chosen as the template and the 2ILK sequence was considered the target. The MODELLER program was used and 3D coordinates were generated for 2ILK. The validation report for the proposed model was compared to the PDB structure, and this comparison is presented in Table 2. The RMSD values between the PDB coordinates of 2ILK and the model of 2ILK generated by MODELLER was calculated to be 1.389 Å for backbone atoms and 2.4 Å for all atoms by the PyMOL superimposition program. These data validated the reliability of our proposed model of catla IL-10.

To functionally characterize catla IL-10, we queried the model structure in the VAST [37] and Dali [38] servers. The VAST server predicted the domains 1–110 (N-terminal) and 111–157 (C-terminal) in catla IL-10. Human IL-10 contained similar domains at the N-terminus (1–113) and C-terminus (114–160). The Clustal W result revealed 38.29% identity at the C-terminal end and 22.7% identity at the N-terminal end of catla IL-10 with the human IL-10 (Fig. 6). Hence, a comparatively low level of homology between catla and human IL-10 was predicted at the

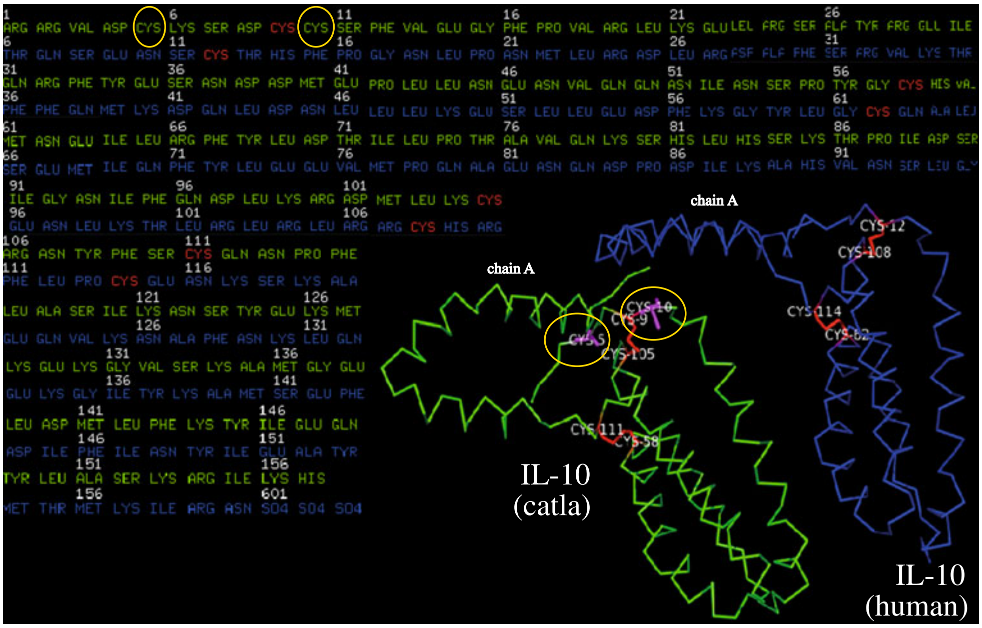


Fig. 5 3D structures of catla and human IL- 10. The 3D structures of catla and human IL-10 are shown in *green* and *blue*, respectively. The cysteine residues that form disulfide bridges are shown in *red*. Additional cysteine residues (Cys5 and Cys10) in catla IL-10 are *encircled by yellow rings* and represented by *pink sticks* in the 3D structure

N-terminus. Pair-wise structural alignment was performed in the Dali server to locate each functional domain, and the highest Z-score value (12.4) suggested significant target–template domain alignments (Fig. 7).

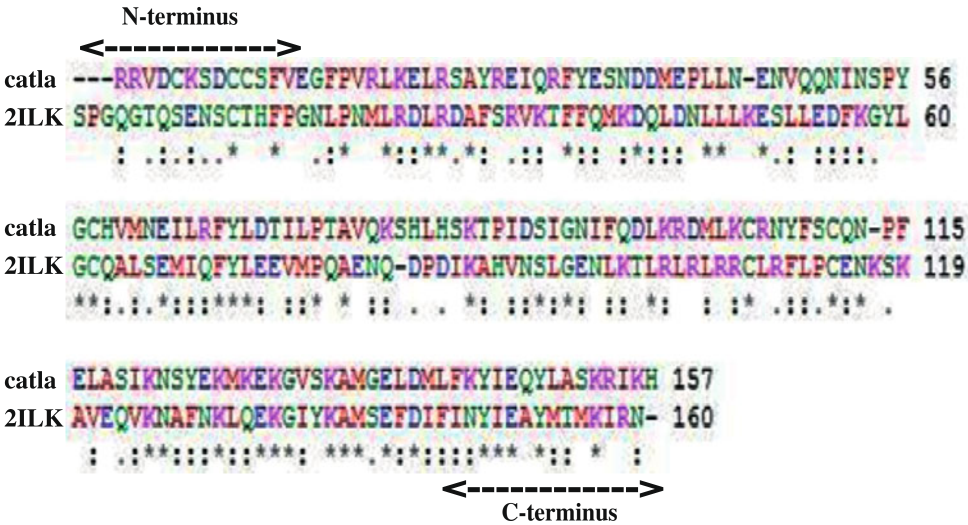
Molecular dynamics simulation

To gain insight into the stability and MD properties of the structure of the homolog, explicit solvent MD simulation was performed. The steepest descent method of energy minimization for the solvated protein model showed that the maximum force dropped below the defined value after 574 steps. Position-restrained molecular dynamics lasting for 100 ps fixed all bond lengths in the system, and an improved validation score was observed. The RMSD value of the catla IL-10 model was compared after performing simulation with the template (2ILK) for 5 ns (Fig. 8). The results showed that the RMSD became stable at the end. This suggested that an accepted structure was obtained by the end of the simulation.

To locate the flexible regions, the root mean square fluctuations (RMSFs) for the Cα atoms of each of the residues were examined by GROMACS in order to measure the average displacement, and these data are presented graphically in Fig. 9. In addition, the time profile of the secondary structure changes during the trajectory was constructed in VMD [39], and this is shown in Fig. S1 of the Electronic supplementary material (ESM). The mean RMSFs of catla IL-10 and human IL-10 were 0.323 nm and 0.327 nm, and their standard deviations were 0.193 nm and 0.122 nm, respectively. All these data indicate minute fluctuations, highlighting the reliability of the model structure.

Table 2 Cross-validation of 2ILK, where catla IL-10 is used as the template by MODELLER

Ramachandran plot analysis	2ILK plot statistics		2ILK model plot statistics	
	Residues	Percentage	Residues	Percentage
Residues in most favored regions	130	90.9	127	88.8
Residues in additional allowed regions	9	6.3	14	9.8
Residues in generously allowed regions	3	2.1	1	0.7
Residues in disallowed regions	1	0.7	1	0.7
Number of non-glycine and non-proline residues	143	100	143	100
Number of end residues (excluding glycine and proline)	2		2	
Number of glycine residues	5		5	
Number of proline residues	5		5	
G-factor	−0.14		0.06	

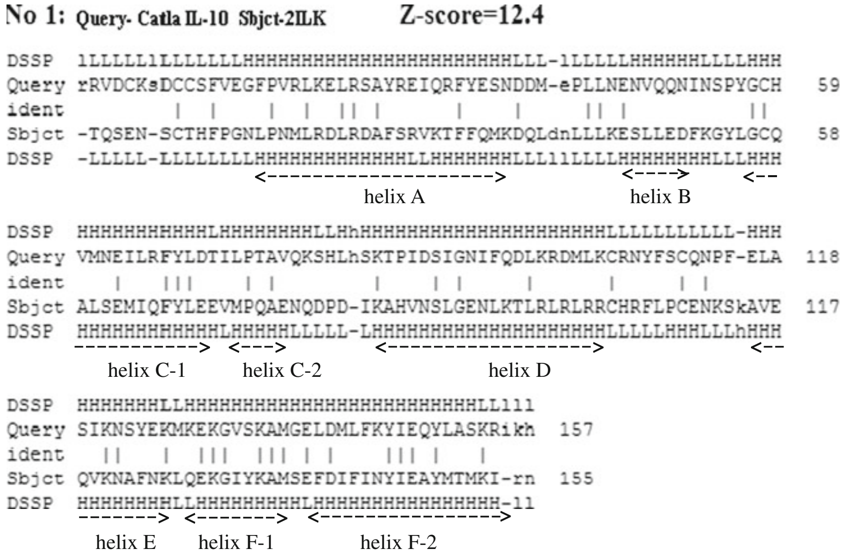


```
catla ---RRVDCKSDCCSFVEGFPVRLKELRSAYREIQRFYESNDDMEPLLN-ENVQQNINSPY 56
2ILK  SPGQGTQSENSCTHFPGNLPNMLRDLRDAFSRVKTFFQMKDQLDNLLLKESLLEDFKGYL 60
         : .:.:..*  *  .:*  *::**.*: .:: *:: :*::: **  *.: ::::.

catla GCHVMNEILRFYLDTILPTAVQKSHLHSKTPIDSIGNIFQDLKRDMLKCRNYFSCQN-PF 115
2ILK  GCQALSEMIQFYLEEVMPQAENQ-DPDIKAHVNSLGENLKTLRLRLRRCLRFLPCENKSK 119
      **:.:.*:::***: ::* * :: . . *: ::*:*: :: *:  : :* .::.*:* .

catla ELASIKNSYEKMKEKGVSKAMGELDMLFKYIEQYLASKRIKH 157
2ILK  AVEQVKNAFNKLQEKGIYKAMSEFDIFINYIEAYMTMKIRN- 160
       : .:**:::*::***: ***.*:*::::*** *:: *  :
```

Fig. 6 Sequence alignment of catla and human IL-10. The IL-10 amino acid sequence alignment was generated by Clustal W, and identical (*) or similar (. or :) amino acids are marked. Amino acids at the N-terminus and C-terminus are shown by *arrows*

The monomer of 2ILK (MW 18.2 kDa) showed good threading with the monomer of catla IL-10 (MW 18.5 kDa). The Cα-based superimposition RMSD value for each domain between catla IL-10 and 2ILK was calculated by PyMOL program, and these values are shown in Table 3 and Fig. 10. The backbone and all-atom superimposition RMSDs were calculated to be 1.148 Å and 1.731 Å, respectively. The structural alignment of the backbone for the functional domains revealed very low deviations, indicating the acceptance of the catla IL-10 model.

We then investigated the possible interaction of catla IL-10 with its receptor on the basis of the complex formed between the human IL-10 dimer and its receptor molecule (IL-10R1; PDB code: 1J7V) [6]. The crystal structure of the cytokine–receptor complex consisted of one IL-10 homodimer with two IL-10R1 molecules. Each receptor binds the identical twofold-related surfaces of IL-10. Residues at the interface derive from two backbone segments. The first segment includes the AB loop and is centered on the bend of helix F2, and the second segment is located near the N-terminus of helix A and the C-terminus of helix F2. The amino acid residues present in helices A and F in human and catla IL-10 differed in length, which may signify a difference in the IL-10/IL-10R complexes of these two species. The relative accessibility percentage (solvent exposure) was calculated by the NACCESS program for each residue involved in the interaction for both human and catla IL-10 (Table 4). As is evident from Table 4, there are



```
No 1: Query- Catla IL-10  Sbjct-2ILK          Z-score=12.4

DSSP  lLLLLLlLLLLLLLHHHHHHHHHHHHHHHHHHHHHHHHHLLL-lLLLLLHHHHHHLLLLHHH
Query rRVDCKsDCCSFVEGFPVRLKELRSAYREIQRFYESNDDM-ePLLNENVQQNINSPYGCH   59
ident         |  |    |  |  || |      |    |     || |          ||
Sbjct -TQSEN-SCTHFPGNLPNMLRDLRDAFSRVKTFFQMKDQLdnLLLKESLLEDFKGYLGCQ   58
DSSP  -LLLLL-LLLLLLLLHHHHHHHHHHHHHHLLHHHHHHHHLLLllLLLLHHHHHHHHLLLHHH
                     <-------------------->         <---->    <--
                             helix A                helix B

DSSP  HHHHHHHHHHHLHHHHHHHHHLLHhHHHHHHHHHHHHHHHHHHHHHHHLLLLLLLLLL-HHH
Query VMNEILRFYLDTILPTAVQKSHLhSKTPIDSIGNIFQDLKRDMLKCRNYFSCQNPF-ELA  118
ident    |   |||    | |        |    | |     |      |     | |
Sbjct ALSEMIQFYLEEVMPQAENQDPD-IKAHVNSLGENLKTLRLRLRRCHRFLPCENKSkAVE  117
DSSP  HHHHHHHHHHHHLHHHHHLLLLL-LHHHHHHHHHHHHHHHHHHHHHHLLLLLHHHLLLhHHH
      -----------> <--->       <------------------>            <--
         helix C-1    helix C-2          helix D

DSSP  HHHHHHHLLHHHHHHHHHHHHHHHHHHHHHHHHHHHLLlll
Query SIKNSYEKMKEKGVSKAMGELDMLFKYIEQYLASKRikh  157
ident   ||   |  |||  ||| |    ||| |   |
Sbjct QVKNAFNKLQEKGIYKAMSEFDIFINYIEAYMTMKI-rn  155
DSSP  HHHHHHHHLLHHHHHHHHHHLHHHHHHHHHHHHHHHHH-ll
      -------> <--------> <---------------->
       helix E   helix F-1       helix F-2
```

Fig. 7 Pair-wise structural alignment of catla and human IL-10 by the Dali server. The six helices found in the IL-10 structures are indicated by *arrows*. *H/h* represent helices, *L/l* represent loops. Structurally equivalent residues are shown in *uppercase letters*, structurally nonequivalent residues (loops) are in *lowercase letters*. Amino acid identities are indicated by *vertical bars*

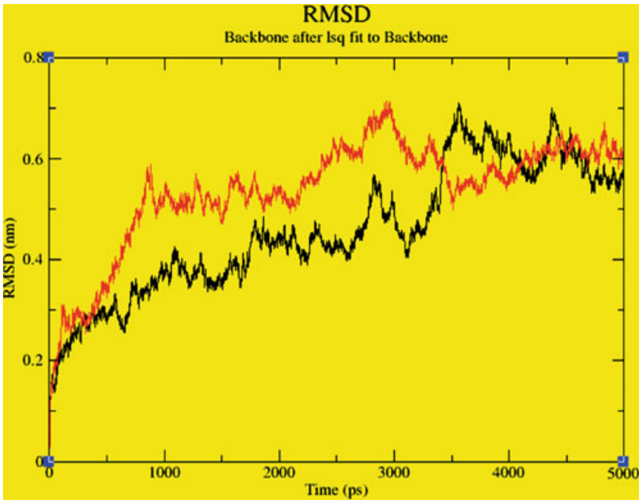


Fig. 8 Comparison of the root mean square deviations (RMSDs) of catla and human IL-10 versus simulation time. The RMSDs of the Cα backbone atoms in catla and human IL-10 are shown in *black* and *red*, respectively

marked differences (shown in bold and italics) in the solvent exposure properties of catla and human IL-10 amino acids. This strengthens the possibility that the complex in catla is stabilized in a different way to the complex in human.

Table 3 Cα-based RMSD values for catla IL-10 obtained by the PyMOL program

Catla IL-10 vs 2ILK	Cα RMSD (Å)
Helix A	0.461
Helix B	0.315
Helix C1	0.293
Helix C2	0.182
Helix D	0.332
Helix E	0.714
Helix F1	0.384
Helix F2	0.440

In the catla IL-10 model, 17 pockets were predicted by the CASTp server, and this highlighted some interesting results. Amino acids that were conserved only in fish species (Cys5 and Cys10) did not show any interactions for any pocket. Different probe radii were used by the CASTp server to check the significances of these conserved cysteine residues in the protein–receptor interaction, but no significant pocket information was obtained for these residues. Further analysis in cons-PPISP and the InterProSurf server showed that these cysteine residues did not occur in any clusters that may take part in the protein interaction. These data strongly support the interacting residues of the catla IL-10 model having different positional distributions with higher confidence clusters as compared to the human IL-10.

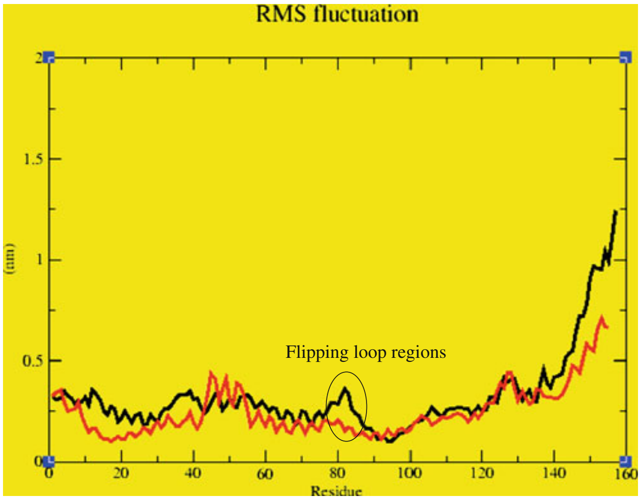


Fig. 9 Comparison of the root mean square fluctuations (RMSFs) of catla and human IL-10. The RMSFs of the Cα backbone atoms in catla and human IL-10 are shown in *black* and *red*, respectively

Conclusions

Cytokines play a pivotal role in modulating fish immunity. Among the cytokines, the role of IL-10 seems to be a

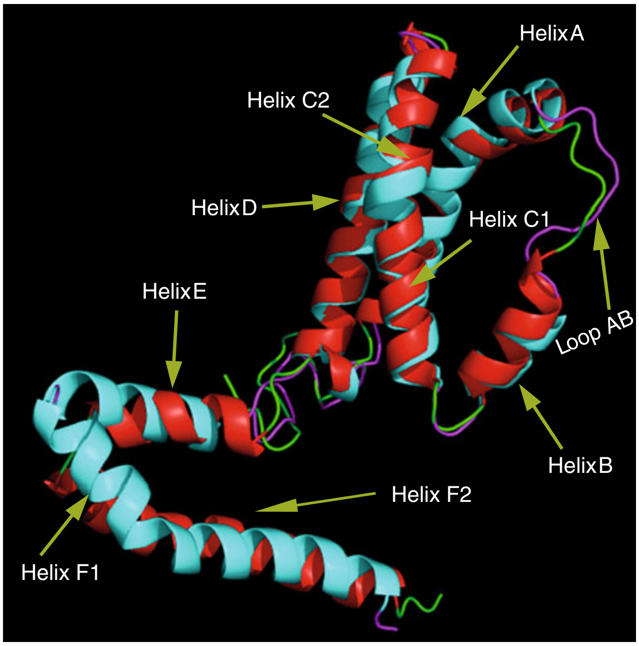


Fig. 10 Superimposition of the catla IL-10 model onto the template (human IL-10, PDB ID: 2ILK). In catla IL-10, helices are shown in *red* and loops are shown in *green*. In the template, helices are shown in *cyan* and loops are in *magenta*

Table 4 Solvent exposure of amino acids in IL-10 interacting with IL-10R in catla and human

	Amino acids	% Solvent exposure in human IL-10	Amino acids	% Solvent exposure in catla IL-10
helix A	***LEU19***	***14.9***	***PHE16***	***33.1***
	PRO20	56.5	PRO17	63.2
	ASN21	44.2	VAL18	49.1
	LEU23	***46.8***	***LEU20***	***35.5***
	ARG24	56.6	LYS21	59.3
	ARG27	74.4	ARG24	68.1
	ASP28	***55.5***	***SER25***	***69.4***
	PHE30	48.0	TYR27	53.1
	SER31	***58.0***	***ARG28***	***75.0***
	LYS34	***41.5***	***GLN31***	***25.4***
	THR35	***49.9***	***ARG32***	***65.1***
	GLN38	59.1	GLU35	67.5
	MET39	77.4	SER36	78.4
loop AB	ASP41	63.2	ASP38	65.6
	GLN42	87.5	ASP39	95.6
	LEU43	***62.9***	***MET40***	***82.8***
	ASP44	***5.0***	***GLU41***	***98.4***
	ASN45	71.1	PRO42	68.5
	LEU46	61.0	LEU43	67.7
	LEU47	80.3	LEU44	75.7
helix F	LYS138	52.7	LYS134	46.3
	SER141	66.5	GLY137	66.2
	GLU142	42.2	GLU138	45.1
	ASP144	61.2	ASP140	65.3
	ILE145	***41.2***	***MET141***	***30.8***
	ASN148	60.7	LYS144	58.4
	TYR149	***72.4***	***TYR145***	***31.7***
	ILE150	40.6	ILE146	45.5
	GLU151	60.9	GLU147	57.4
	MET154	52.0	LEU150	58.1
	THR155	***40.9***	***ALA151***	***52.4***
	ILE158	***69.3***	***ARG154***	***90.0***
	ARG159	71.0	ILE155	63.3

The amino acids in bold and italic shows significant percentage difference

crucial one, due to its inductive expression during diseases and its pleotropic role in immunity. To understand the functional biology of IL-10 in fish, a 3D modeled structure of IL-10 may aid in locating the putative sites of IL-10, and in understanding the interaction with its receptor. In this study, the 3D model of catla IL-10 showed considerable structural differences from human IL-10, which may lead to functional differences. Small peptides spanning the putative regions identified by 3D modeling could be synthesized for *in vivo* experimental investigations of IL-10-mediated cell signaling and functional characterization. The 3D modeling data showed that two additional cysteine residues (Cys5 and Cys10) in fish did not form any significant bonds involved in structural stabilization or the protein–receptor interaction, so it is speculated that, during the course of evolution, they have mutated in higher vertebrates. Further investigations are required to validate these predictions.

Acknowledgments The work was financially supported by a grant from the National Agricultural Innovation Project, Indian Council of Agricultural Research (NAIP-ICAR; project code: C4-C30018). The authors gratefully acknowledge the aid of the Bioinformatics Resources and Applications Facility (BRAF) at the Center for Development of Advanced Computing (C-DAC), Pune, India, for the MD simulation studies. We thank the Director of the Central Institute of Freshwater Aquaculture (CIFA) for providing the necessary facilities to carry out this study.

References

1. Moore KW, O'Garra A, de Waal MR et al (1993) Interleukin-10. Annu Rev Immunol 11:165–190
2. Mosmann TR (1994) Properties and function of interleukin-10. Adv Immunol 56:1–26
3. Fiorentino DF, Bond MW, Mosmann TR (1989) Two types of mouse helper T cell. IV. Th2 clones secrete a factor that inhibits cytokine production by Th1 clones. J Exp Med 170:2081–2095
4. Zdanov A (2004) Structural features of the interleukin-10 family of cytokines. Curr Pharm Des 10:3873–3884
5. Rousset F, Garcia E, Defrance T, Peronne C, Vezzio N, Hsu DH, Kastelein R, Moore KW, Banchereau J (1992) Interleukin 10 is a potent growth and differentiation factor for activated human B lymphocytes. Proc Natl Acad Sci USA 89:1890–1893
6. Josephson K, Logsdon JN, Walter RM (2001) Crystal structure of the IL-10/IL-10R1 complex reveals a shared receptor binding site. Immunity 14:35–46
7. Tan JC, Indelicato SR, Narula SK, Zavodny PJ, Chou CC (1993) Characterization of interleukin-10 receptors on human and mouse cells. J Biol Chem 268:21053–21059
8. Thompson JD, Higgins DG, Gibson TJ (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 22:4673–4680
9. Gouet P, Courcelle E, Stuart DI, Metoz F (1999) ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics 15:305–308
10. Sali A, Blundell TL (1993) Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 234:779–815
11. Laskoswki RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereo chemical quality of protein structures. J Appl Crystallogr 26:283–291
12. Eisenberg D, Luthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol 277:396–404
13. Colovos C, Yeates TO (1993) Verification of protein structures: patterns of non bonded atomic interactions. Protein Sci 2:1511–1519
14. Pontius J, Richelle J, Wodak SJ (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol 264:121–136
15. Vriend G (1990) WHAT IF: a molecular modeling and drug design program. J Mol Graph 8:52–6,29
16. Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D 66:12–21
17. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795
18. Kurowski MA, Bujnicki JM (2003) GeneSilico protein structure prediction meta-server. Nucleic Acids Res 31:3305–3307
19. Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19:1015–1018
20. Wallner B, Elofsson A (2005) Pcons5: combining consensus, structural evaluation and fold recognition scores. Bioinformatics 21:4248–4254
21. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40
22. Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310:243–257
23. Mizuguchi K, Deane CM, Blundell TL, Overington JP (1998) HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 7:2469–2471
24. Mizuguchi K, Deane CM, Blundell TL, Overington JP (1998) JOY: protein sequence–structure representation and analysis. Bioinformatics 14:617–623
25. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723
26. Walter RP, Scott PH et al (1999) The GROMOS biomolecular simulation program package. J Phys Chem 103:3596–3607
27. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) GROMACS: fast, flexible, and free. J Comput Chem 26:1701–1718
28. Jones S, Thornton JM (1996) Principles of protein–protein interactions. Proc Natl Acad Sci USA 93:13–20
29. Hubbard SJ, Thornton JM (1993) NACCESS. University College, London
30. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 34:116–118
31. Chen H, Zhou HX (2005) Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. Proteins 61:21–35
32. Negi SS, Schein CH, Oezquen N, Power TD, Braun W (2007) InterProSurf: a web server for predicting interacting sites on protein surfaces. Bioinformatics 23:3397–3399
33. Pinto RD, Nascimento DS, Reis MIR, do vale A, dos Santos NM (2007) Molecular characterization, 3D modelling and expression analysis of sea bass (*Dicentrarchus labrax* L.) interleukin-10. Mol Immunol 44:2066–2075
34. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM score. Nucleic Acids Res 33:2302–2309
35. McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16:404–405
36. Heinig M, Frishman D (2004) STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res 32:500–502
37. Gibrat JF, Madej T, Bryant SH (1996) Surprising similarities in structure comparison. Curr Opin Struct Biol 6:377–385
38. Holm L, Sander C (1995) Dali: a network tool for protein structure comparison. Trends Biochem Sci 20:478–480
39. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14(33–8):27–28

J Mol Model (2012) 18:1723–1728
DOI 10.1007/s00894-011-1203-4

ORIGINAL PAPER

Introducing NOB-NOBs: nitrogen-oxygen-boron cycles with potential high-energy properties

Aloysus K. Lawong · David W. Ball

Received: 13 June 2011 / Accepted: 26 July 2011 / Published online: 11 August 2011

Abstract As a follow-up on a study of a family of boron-oxygen-nitrogen compounds composed of two datively bonded B–O–N backbones, we investigate a similar series of compounds that have similar fragments but are covalently bonded. B3LYP/6-31G(d,p) quantum mechanical calculations have been performed to determine the minimum-energy geometries, vibrational frequencies, and thermochemical properties of the parent compound and a series of nitro-substituted derivatives. Our results indicate that some of the derivatives have at least appropriate thermodynamics for possible high-energy materials, in some cases being favorable over similar dimeric compounds with coordinate covalent B–N bonds.

Keywords NOB-NOB compounds · B3LYP calculations · High-energy materials

Introduction

In 1963, Kuhn and Inatome [1] published a report on an air-stable boron–oxygen–nitrogen molecule that they determined was composed of two B–O–N molecules in the form of a six-membered ring. They later presented evidence, in the form of measured dipole moments, that the ring existed as a chair conformer [2]. Because experimental evidence suggested that the nitrogen atom in the B–O–N monomer made a coordinate covalent bond with the boron atom of the other monomer, Kuhn et al. referred to these molecules as “BON-BON” species. Their derivatives had several *n*-butyl groups bonded to either the boron atom or the nitrogen atom (or both) in the ring. Recently, we (Lawong AK, Ball DW, 2011, manuscript in preparation) performed computational chemical analyses of the parent BON-BON molecule (the moniker of which we are choosing to give in all capitals, unlike Kuhn et al. [1, 2], to emphasize the atomic constitution of the six-membered ring) and a variety of nitro-substituted BON-BON molecules in order to study their potential as new high-energy (HE) materials. The parent BON-BON molecule, cyclo-$BH_2ONH_2BH_2ONH_2$, is shown in Fig. 1 (Lawong AK, Ball DW, 2011, manuscript in preparation). Because both the B and N atoms are tetracoordinated, the molecule adopts a cyclohexane-like central ring structure.

In the course of our study, we realized that there is another way to link two B–O–N moieties: using actual covalent bonds between the B and N atoms between the two monomers, rather than coordinate covalent bonds. That is, a nitrogen atom on one B–O–N fragment would covalently bond with the boron atom on a second B–O–N fragment, with a similar covalent bond occurring between the other ends of the fragments. To differentiate this bonding arrangement from that found in BON-BON molecules, we propose the name “NOB-NOB” in reference to the covalently bonded six-membered ring. In this work, as a follow-up to our BON-BON study, we present a computational chemical study of the structures and properties (including the thermochemical properties) of the parent NOB-NOB molecule and nitro derivatives of NOB-NOB.

Computational details

All calculations were performed using the Gaussian 09 computational chemistry program [3] on an IBM cluster 1350 supercomputer at the Ohio Supercomputer Center in Columbus, Ohio. We used the density functional theoretical method, as defined by combining Becke’s three-parameter exchange functional with the correlation functional of Lee,

A. K. Lawong · D. W. Ball (✉)
Department of Chemistry, Cleveland State University,
2121 Euclid Avenue,
Cleveland, OH 44115, USA
e-mail: d.ball@csuohio.edu

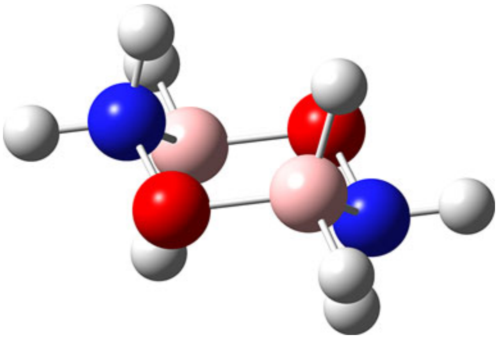


Fig. 1 The general structure of BON-BON-type six-membered rings. Because of the tetracoordinated B and N atoms, the ring adopts a cyclohexane-type structure, in this case the chair conformer. (From Lawong AK, Ball DW, 2011, manuscript in preparation)

Yang, and Parr (abbreviated to B3LYP in the Gaussian program) [4, 5], along with the standard Gaussian basis set labeled 6-31G(d,p) [6]. Minimum-energy geometries of the NOB-NOB molecules were determined using default settings, and vibrational frequency calculations were performed to verify that a minimum-energy geometry was found. Once the proper structure of the NOB-NOB molecule was established, the enthalpy of formation was determined by calculating the enthalpy change for the molecule formed from its gaseous elements, and then corrected for the enthalpy of formation of gas-phase boron. For example, the reaction for the parent molecule was

$$2H_2(g) + O_2(g) + N_2(g) + 2B(g)$$
$$\rightarrow \text{cyclo} - \text{NHOBHNHOBH}(g).$$

The energy change associated with this reaction was determined from the calculations and then corrected for the formation of two moles of B(g):

$$2[B(s) \rightarrow B(g)] \qquad \Delta H = 2\left[565.0\ \text{kJ mol}^{-1}\right].$$

The enthalpy of formation for B(g) was taken from the NIST Chemistry Webbook website [7]. Once corrected for the formation of B(g), the energy represents the enthalpy of formation of the NOB-NOB molecule. After this, enthalpies of decomposition and/or combustion can be determined using standard balanced reactions, assuming that the products are $B_2O_3(s)$, $H_2O(\ell)$, and $N_2(g)$. When necessary, $O_2(g)$ is added as a reactant for the complete oxidation of B and H in the molecules.

Results and discussion

The non-nitrated NOB-NOB compound has the formula (cyclo-)NHOBHNHOBH. There are nine nitro-NOB-NOB compounds: two nitro-NOB-NOB isomers, four dinitro-NOB-NOB isomers, two trinitro-NOB-NOB isomers, and one tetranitro-NOB-NOB molecule. Thus, here we are reporting on a total of ten NOB-NOB compounds. For much of the presentation that follows, we will focus on the non-nitrated NOB-NOB molecule (referred to as the "parent NOB-NOB") and the tetranitro-NOB-NOB molecule, (cyclo)-$N(NO_2)OB(NO_2)N(NO_2)OB(NO_2)$. The less nitrated molecules have properties intermediate between the two extremes, and (except for their thermodynamics) this will be assumed unless there is something noteworthy about a particular nitro-NOB-NOB. Readers interested in learning more about the partially nitrated NOB-NOB compounds can contact the corresponding author.

Figure 2 shows the optimized geometries of the parent NOB-NOB compound and tetranitro-NOB-NOB, while Table 1 lists some representative bonding parameters of these two molecules. The structure of the parent NOB-NOB molecule should be compared to that of the parent BON-BON molecule, shown in Fig. 1 (Lawong AK, Ball DW, 2011, manuscript in preparation): the parent NOB-NOB optimizes as a flat molecule, suggesting that the nitrogen atoms have strong sp^2 character, as opposed to the cyclohexane-like ring adopted by the BON-BON derivatives. As if to belie this, however, the bond angle that the N–H bond makes with the oxygen atom in the ring is close to the expected near-tetrahedral angle: 108.3° rather than the ideal 120°. This is likely due to an intramolecular interaction between the electropositive H atom and the electronegative O atom. On the other hand, the B–H bond is oriented almost exactly 120° (actually slightly less: 117°) from the other ring atoms. The six-membered ring is not a perfect hexagon. The O–N, N–B, and B–O bonds are slightly different lengths (1.427, 1.398, and 1.378 Å, respectively), while the bond angles vary, sometimes significantly, from 120°. The relative orientations of the NO_2 groups in tetranitro-NOB-NOB show an interesting pattern: the NO_2 group bonded to the nitrogen atom in the ring lies in the plane of the ring, while the NO_2 group

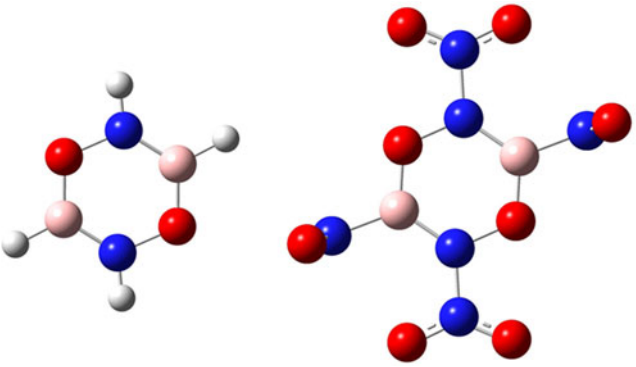


Fig. 2 The parent NOB-NOB molecule (cf. Fig. 1) and the tetranitro-NOB-NOB molecule

Table 1 Representative bonding parameters of the parent NOB-NOB molecule and the tetranitro-NOB-NOB derivative. Distances in Å, angles in degrees

	NOB-NOB	Tetranitro-NOB-NOB
r(N–O)	1.427	1.399
r(O–B)	1.378	1.363
r(B–N)	1.398	1.419
r(N–H)	1.006	–
r(B–H)	1.189	–
r(N–N)	–	1.443
r(B–NO_2)	–	1.516
α(B–O–N)	113.9	117.3
α(O–N–B)	124.1	121.7
α(N–B–O)	122	121.1
α(O–B–H)	117.8	–
α(O–N–H)	108.3	–
α(O–N–O)	–	127.0, 130.6

bonded to the boron atom in the ring lies perpendicular to the ring. This pattern is repeated in all NOB-NOB derivatives that have NO_2 groups on adjacent atoms (which is seen for three derivatives in the group of ten molecules studied here): the NO_2 group bonded to the nitrogen atom is always in the plane of the ring, while the NO_2 group bonded to the boron atom is always perpendicular to the plane of the ring. This is unusual for an ortho-substituted di- (or greater) nitro compound with a planar central ring. For example, 1,2-dinitrobenzene has its adjacent nitro groups rotated ~41° out of the plane of the planar C_6 ring [8], while the nitro groups in hexanitrobenzene are all 43–45° out of the plane of the ring [9].

The general structure of the BON-BON molecules studied previously (Lawong AK, Ball DW, 2011, manuscript in preparation) included a central six-membered ring that resembled the chair conformation of cyclohexane. The first major difference in the structures of BON-BON and NOB-NOB (in both cases referring to the parent molecule) is a nonplanar chair conformation for BON-BON and a flat, distorted hexagon for NOB-NOB. As for bond distances within the six-membered rings, only the N–O bond distance remains close to being the same for the two types of rings (1.427 Å here versus 1.429 Å for the parent BON-BON molecule). The B–O bond is slightly shorter in NOB-NOB (1.378 Å here versus 1.476 Å in BON-BON). The B–N bond is significantly shorter in NOB-NOB (1.427 Å vs. 1.624 Å), as might be expected for a B atom bonding to a trivalent N atom. Bonding to and bonding within the NO_2 groups were similar in the two types of molecules, except for the N–N(nitro) bond; again, as befitting a bond to a trivalent N atom, the N–N bond distance in tetranitro-NOB-NOB was found to be 1.443 Å, down significantly from the 1.696–1.899 Å bond distance found in octanitro-BON-BON (Lawong AK, Ball DW, 2011, manuscript in preparation).

A look at some of the molecular orbitals explains why the six-membered rings of NOB-NOB are close to being planar. Figure 3 shows HOMO-7 and HOMO-21 for the parent NOB-NOB and the tetranitro derivative, respectively. These molecular orbitals show the delocalization of electrons over the entire ring and even, in the case of the tetranitro derivative, into the NO_2 groups. This is very reminiscent of the π orbitals of aromatic systems, and it would not be surprising if this molecule were found to have aromatic character.

Figure 4 shows the calculated vibrational spectra of the parent NOB-NOB molecule and the tetranitro derivative, which should help in identifying these substances should synthesis be attempted. The vibrational spectrum of NOB-NOB is unremarkable, with the N–H, B–H, and ring vibrations appearing in predictable ranges. The vibrational spectrum of tetranitro-NOB-NOB is more interesting. Some features that stand out are what appear to be doublets of absorptions throughout the spectrum, especially in the mid- to high-energy range. Visualization of the normal vibrational modes using the GaussView program [10] demonstrates the reason for these near-doublets. Each pair represents two similar motions that would otherwise be degenerate if the symmetry of the molecule were higher. For example, the strong absorption at 1765 cm^{-1} is the asymmetric O–N–O stretch of the two NO_2 groups bonded to the nitrogen atoms in the ring. The strong absorption near it, at 1629 cm^{-1}, is the asymmetric O–N–O stretch of the two NO_2 groups bonded to the boron atoms in the ring. Within 2 cm^{-1} of each of these strong absorptions is a zero-intensity absorption identifiable as the corresponding symmetric stretches of the same atoms. Similar correspondences can be assigned to other absorptions of similar intensities and close frequencies. Also, because of the symmetry of the molecule, fully 25 of the 48 normal modes of vibration have exactly zero intensity (compared to only 10 out of 24 for the parent NOB-NOB molecule).

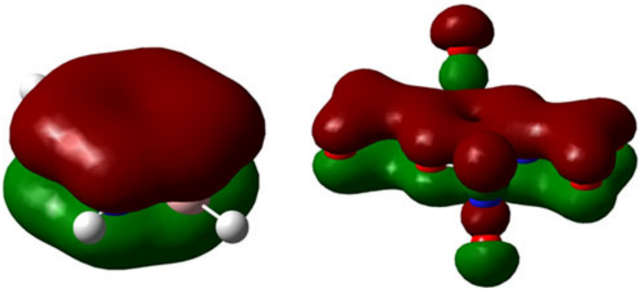


Fig. 3 Molecular orbitals showing the delocalization of electrons in the ring (*left*, parent NOB-NOB) and extending into the NO_2 groups planar to the ring (*right*, tetranitro-NOB-NOB)

Because our main focus is the thermodynamics of the NOB-NOB molecules, here we will include information about all of the isomers. After determining the enthalpy of formation of the molecules as described above, the combustion or decomposition enthalpy was also determined. Whether the relevant reaction is considered a combustion or decomposition depends on the oxygen balance (OB%) of the compound. The oxygen balance is given by the expression [11]

$$\mathrm{OB\%} = -\frac{3200\left(\frac{3}{4}b + \frac{1}{4}h + 0n - \frac{1}{2}o\right)}{\mathrm{MW}},$$

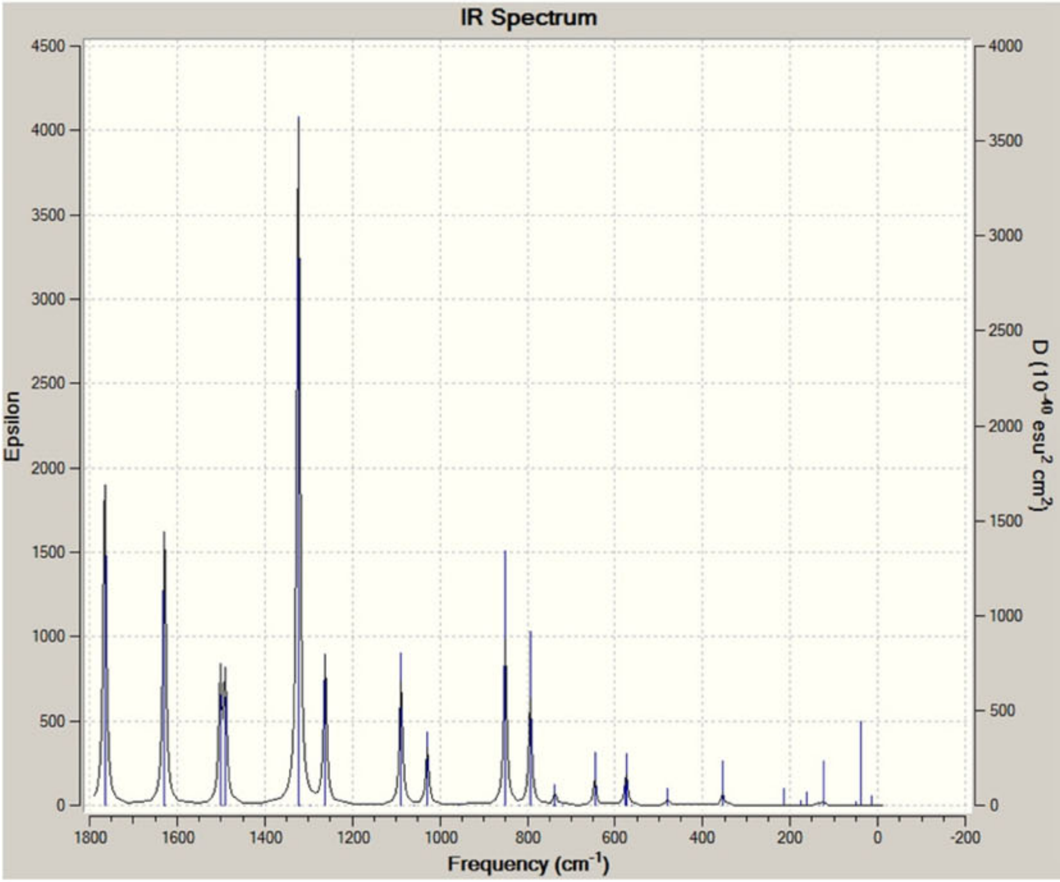
where b, h, n, and o are the numbers of boron, hydrogen, nitrogen, and oxygen atoms in the molecular formulae, respectively, and MW is the molar mass of the molecule. An OB% that is less than zero indicates that a molecular formula does not have sufficient oxidizer (here, oxygen) present to oxidize all other atoms present, while an OB% of greater than zero indicates that a molecular formula does contain sufficient oxidizer to oxidize all other atoms fully. Thus, substances with negative OB% values need additional oxidizer (assumed here to be molecular oxygen), and the enthalpy changes of reaction with said oxidizer are appropriately labeled enthalpies of combustion (ΔH_{comb}). Substances with positive OB% values have sufficient oxidizer atoms to oxidize themselves, so the enthalpy changes of reaction are more appropriately labeled enthalpies of decomposition (ΔH_{dec}). Table 2 lists oxygen balances, calculated enthalpies of formation, and resulting enthalpies of decomposition or combustion for the ten NOB-NOB derivatives. There are considerably fewer nitro-NOB-NOB molecules than nitro-BON-BON molecules because of the fewer hydrogen atoms that can be substituted for NO_2 groups, which in turn leads to fewer substitutional isomers. In the labeling of the nitro-NOB-NOB isomers, the point of substitution is given, with the prime (′) implying that the additional NO_2 substitution (where appropriate) is in the other NOB monomer as well. According to the OB% values, the parent NOB-NOB and nitro-NOB-NOB require extra oxidizer, so the enthalpies of reaction are enthalpies of combustion. For greater NO_2 substitution, the positive OB% values indicate sufficient oxygen to oxidize completely, so enthalpies are better described as decomposition enthalpies.

Table 2 shows that all of the NOB-NOB-based compounds have strongly negative enthalpies of formation, likely due in part to the strong B–N bonds in the six-membered rings. Upon nitration, the thermodynamics of the isomers shows a similar trend to the respective BON-BON compounds, but not as extreme. Once again, in the nitro-substituted compound, the site of nitration significantly affects the energy values, with the B-substituted nitro-NOB-NOB predicted to be more stable than the N-substituted nitro-NOB-NOB. However, the difference in ΔH_{f} values is only about 120 kJ mol^{-1}, rather than the 220 kJ mol^{-1} seen between the two nitro-BON-BON derivatives. The trend is clear in the NOB-NOB derivatives, as it was in the BON-BON molecules: all other things being the same, an NO_2 group bonded to an N atom of the six-membered ring leads to a less-stable isomer than a similar molecule with the NO_2 group bonded to a B atom of the ring.

Fig. 4 Calculated vibrational spectra of the parent NOB-NOB molecule ▶ (*bottom*) and the tetranitro derivative (*top*). Note that the horizontal scales are different for the two spectra

The calculated enthalpies of formation generally increase (that is, get less negative) as the level of nitration increases; however, the increase is not monotonic. The least stable isomer, relative to the constituent elements, is N,N′-dinitro-NOB-NOB. In this molecule, both of the relatively stable N–H bonds from the parent compound are substituted for NO_2 groups, so they are replaced with less-stable N–NO_2 bonds. As mentioned in the BON-BON paper (Lawong AK, Ball DW, 2011, manuscript in preparation), these N–N bonds are the most likely to initiate decomposition in this case too.

With enthalpies of formation determined, enthalpies of combustion or decomposition can be determined using standard combustion or decomposition reactions. The enthalpy of combustion of the parent NOB-NOB molecule is −957 kJ mol^{-1}. Per unit mass, this molar enthalpy of combustion is recalculated to a value of −11.2 kJ g^{-1}. This is about twice as much energy per unit mass as current HE materials like RDX and HMS, whose specific enthalpies of decomposition are both about 5 kJ g^{-1} [12]. However, this is significantly lower than the specific enthalpy of combustion for the parent BON-BON compound, which is −16.6 kJ g^{-1}. With four less hydrogen atoms and stronger B–N bonds, the parent NOB-NOB compound not only has a more negative enthalpy of formation than the parent BON-BON, but it gives off two less H_2O molecules as combustion products. The enthalpies of combustion of the two nitro-NOB-NOB isomers are slightly less negative than that of the parent molecule, and because of the rather dramatic increase in mass brought on by a single NO_2 group (89.6 u for the parent molecule, but 134.6 u for nitro-NOB-NOB: a 50.4% increase), the energy given off per gram decreases by about half, to −5.9 to −6.9 kJ g^{-1}. Upon increasing the nitro content, the calculated enthalpies of decomposition vary between −570 and −855 kJ mol^{-1}, varying more because of the position of the NO_2 group rather than the number of NO_2 groups. This is in part because higher levels of nitration lead to the formation of more N_2 and O_2 as products, which have enthalpies of formation of zero and thus contribute nothing to the


IR Spectrum
Epsilon
0
500
1000
1500
2000
2500
3000
3500
4000
4500
D (10^{-40} esu^2 cm^2)
0
500
1000
1500
2000
2500
3000
3500
4000
Frequency (cm^{-1})
1800
1600
1400
1200
1000
800
600
400
200
0
-200

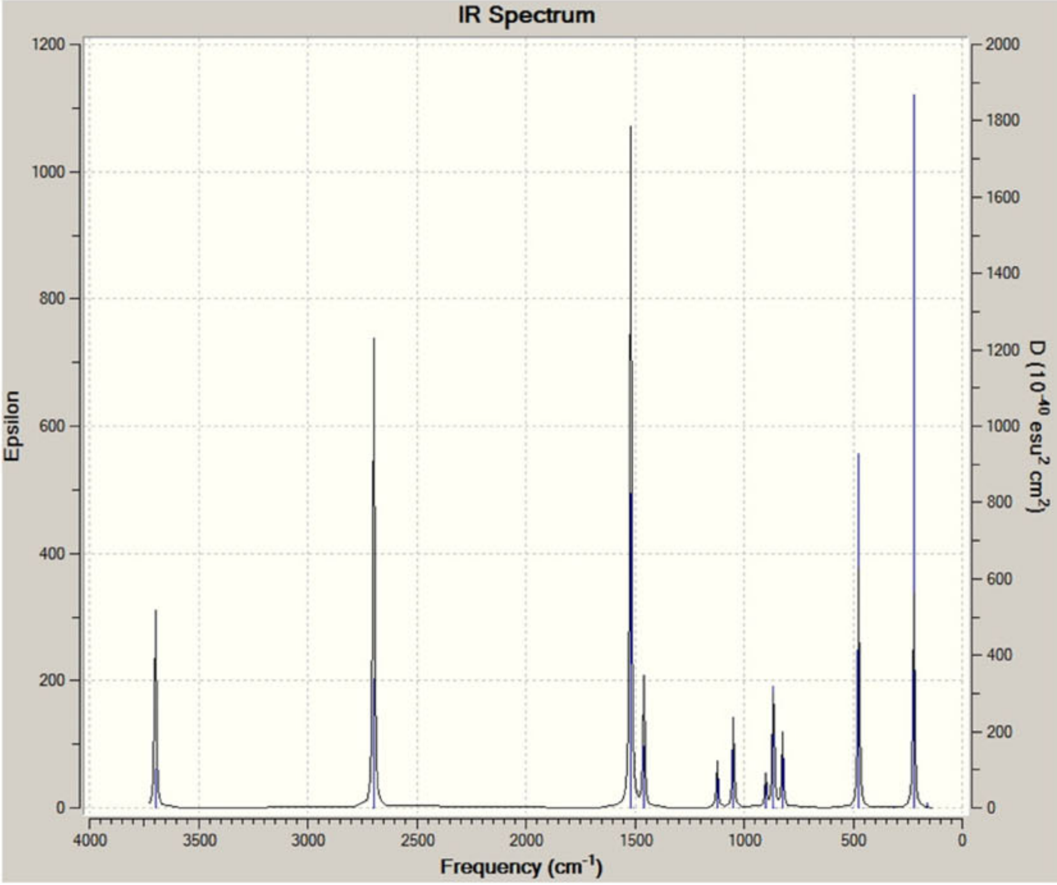

IR Spectrum
Epsilon
0
200
400
600
800
1000
1200
D (10^{-40} esu^2 cm^2)
0
200
400
600
800
1000
1200
1400
1600
1800
2000
Frequency (cm^{-1})
4000
3500
3000
2500
2000
1500
1000
500
0

Table 2 Oxygen balances, enthalpies of formation and combustion/decomposition, and specific enthalpies of reaction for NOB-NOB and its nitrated derivatives

Molecule	OB%	ΔH_f (kJ mol^{-1})	$\Delta H_{comb/dec}$ (kJ mol^{-1})	$\Delta H_{comb/dec}$ (kJ g^{-1})
NHOBH–NHOBH	−56.1	−362.7	−957	−11.2
Nitro-NOB-NOB	−6.1			
B-		−426.7	−772	−5.91
N-		−300.9	−897.8	−6.87
Dinitro-NOB-NOB	18.2			
B,B′-		−478.3	−599.5	−3.41
B,N-		−349.1	−728.7	−4.15
B,N′-		−334.9	−742.9	−4.23
N,N′-		−222.6	−855.2	−4.87
Trinitro-NOB-NOB	32.6			
B,N,B′-		−375.2	−581.7	−2.64
B,N,N′-		−246.4	−710.5	−3.22
Tetranitro-NOB-NOB	43.6	−264.24	−571.7	−2.23

generation of stable products. The fact that one-half of an H_2O molecule less is formed as a product with the addition of each NO_2 group apparently has only minimal impact on the resulting enthalpy of decomposition.

However, although the enthalpies of decomposition are fluctuating about a mean (which is about −680 kJ mol^{-1}), the mass of the molecule is increasing by a net 45.0 u per nitro group, so the enthalpies of decomposition per unit gram are decreasing noticeably. The specific enthalpies of formation for all dinitro-NOB-NOB isomers are less than those for nitro-NOB-NOB, and trinitro-NOB-NOB isomers even lower. The tetranitro-NOB-NOB derivative has the lowest specific enthalpy of decomposition, −2.23 kJ mol^{-1}. Even this value is not entirely out of range for potential HE materials; Akhavan lists [12] the specific enthalpy of reaction for nitroguanidine at −2.47 kJ g^{-1}, just slightly more energy per gram than that of tetranitro-NOB-NOB.

We point out that even this lowest value for tetranitro-NOB-NOB is more energy than six types of nitrated BON-BON molecules (Lawong AK, Ball DW, 2011, manuscript in preparation), which can accommodate more NO_2 groups and hence achieve higher molar masses, reducing their energy density despite their more negative enthalpies of decomposition. Thus, nitrated NOB-NOB derivatives may be potential HE candidates that are worthy of additional exploration. Other factors need to be considered before nitrated NOB-NOB molecules would be deemed "good" HE materials, like velocity of detonation and impact sensitivity. However, for at least some NOB-NOB compounds, their thermodynamics of combustion and decomposition are promising.

Acknowledgments We would like to thank the Ohio Supercomputer Center in Columbus, OH, for granting us time to perform this work. AKL thanks the McNair Scholars Program at Cleveland State University for support.

References

1. Kuhn LP, Inatome N (1963) J Am Chem Soc 85:1206–1207
2. Thomson HB, Kuhn LP, Inatome N (1964) J Phys Chem 68:421422
3. Frisch MJ, Trucks GW, Schlegel HB et al (2009) Gaussian 09, revision A.01. Gaussian, Inc., Wallingford
4. Becke AD (1993) J Chem Phys 98:5648–5652
5. Lee C, Yang W, Parr RG (1988) Phys Rev 37:785–789
6. Hariharan PC, Pople JA (1973) Theor Chem Acc 28:213–222
7. National Institute of Standards and Technology (2011) NIST chemistry webbook. http://webbook.nist.gov. Accessed April 28, 2011
8. Herbstein FH, Kapon M (1990) Acta Cryst B46:567–572
9. Akopyan ZA, Struchkov YT, Dashevskii VG (2006) J Struct Chem 7:385–392
10. Dennington R, Keith T, Millam J (2007) GaussView, version 4.1. Semichem Inc., Shawnee Mission
11. Persson PA, Holmberg R, Lee J (1993) Rock blasting and explosives engineering. CRC Press, Boca Raton
12. Akhavan J (2004) The chemistry of explosives, 2nd edn. Royal Society of Chemistry, London

J Mol Model (2012) 18:1729–1734
DOI 10.1007/s00894-011-1202-5

ORIGINAL PAPER

Benchmarking of ONIOM method for the study of NH_3 dissociation at open ends of BNNTs

Ali Ahmadi · Javad Beheshtian · Mohammad Kamfiroozi

Received: 8 May 2011 / Accepted: 26 July 2011 / Published online: 12 August 2011

Abstract The reliability of ONIOM approach have been examined in calculations of adsorption energies, transition structures, change of HOMO-LUMO energy gaps and equilibrium geometries of the interaction between NH_3 and N-enriched (**A**) or B-enriched (**B**) open ended boron nitride nanotubes. To these ends, four models of the **A** or **B**, with different inner and outer layers have been studied. In addition, various low-levels including, AM1, PM3, MNDO and UFF have been examined, applying B3LYP/6-31 G* in all high-levels. It was shown, that in the case of **A**, (choosing two atom layers of the tube open-end as inner layer) the results of ONIOM approach are in best agreement with those of the pure density functional theory (DFT) calculations, while their results significantly differ from those of DFT in the case of **B** in same conditions. All above and population analysis demonstrate that the ONIOM may be a reliable scheme in the study of weak interactions while it is a controversial approach and should be applied cautiously in the case of strong interactions. We also probed the effect of tube length and diameter on the consistency between ONIOM and DFT results, showing that this consistency is independent of the mentioned parameters.

Keywords Adsorption · B3LYP · Boron nitride nanotubes · DFT · NH_3 · ONIOM

A. Ahmadi
Young Researchers Club, Islamic Azad University,
Islamshahr branch,
Islamshahr, Iran

J. Beheshtian (✉)
Department of Chemistry,
Shahid Rajaee Teacher Training University,
P.O. Box: 16875–163, Tehran, Iran
e-mail: J.Beheshtian@srttu.edu

M. Kamfiroozi
Department of Chemistry, Islamic Azad University,
Shiraz branch,
Shiraz, Iran

Introduction

The applications of computational chemistry span predicting the structure, spectra, transition states and reactivity of complicated molecules. To serve as a predictive tool, however, the methods should be applicable to a large enough portion of the system, reflecting the features of the real system. Among the all quantum mechanical methods, a few of them can be easily applied to the study of thermodynamics and reaction mechanisms in large systems such as proteins, nanotubes, etc. In fact, the calculation time in accurate ab initio methods grows much faster than the number of atoms. This growth is roughly relative to the third power of the number of atomic basis functions used to solve the Schrödinger equation, at least within the density functional theory (DFT) context.

Developed by Morokuma et al., ONIOM is a method to study the large molecules by dividing them into two or three layers, where a high-level calculation is performed on the smallest layer (inner layer) and the rest layer (outer layer) effects are included at a low-level of theory [1, 2].

In spite of several theoretical studies on single-walled carbon nanotubes, applying ONIOM method, [3, 4] only few works have been published on the case of boron nitride nanotubes (BNNTs) [5]. BNNTs as inorganic nanomaterials, have received considerable research interests [6, 7] because of their remarkable electronic, mechanical, and thermal properties.

The ONIOM method is still a controversial method, it has been recently reported that the often-used ONIOM (B3LYP:AM1) approach is not appropriate for some nanotube systems [8, 9]. In the present work, we are interested in ONIOM study of the N-H bond cleavage of NH_3 at the open ends of BNNTs, evaluating its reliability by comparing its results with our previous reported full DFT ones [10]. To this aim, several combinations of different levels of theory as well as different partitions of the inner and outer layers were considered.

The N-H bond cleavage is a challenging problem not only toward the transformation of NH_3 into a useful amino compound but also toward the starting of many catalytic reactions [11–13]. Since only methodological aspects of the ONIOM method have been studied in the present work, no discussion was addressed with respect to either experimental data or absolute accuracy of the chosen levels of calculations.

Computational details

All calculations were carried out with the Gaussian 98 suite of programs [14]. A zigzag (4, 0) BNNT, $B_{20}N_{20}H_4$ was chosen with open ends, in which only one end was saturated with four H atoms. Existing two different terminals for zigzag BNNTs, two forms of open-ended types were used in order to model the NH_3 dissociation at the tube ends, including N-enriched (**A**) and B-enriched (**B**) types (Fig. 1).

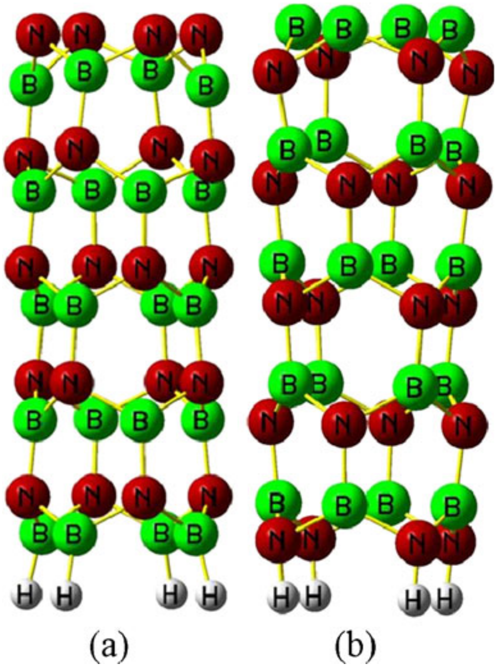


Fig. 1 (**a**) The model of **A** (N-enriched open-ended BNNTs), (**b**) The model of **B** (B-enriched open ended BNNTs)

Firstly, four models of the **A** tube were selected in which the inner layers consist of N4B4 (**A1**), N8B4 (**A2**), N8B8 (**A3**) and N5B5 (**A4**), Fig. 2. In models of **A1**, **A2** and **A3**, two, three and four rows of atoms were placed respectively in the inner layer, and in the **A4** two hexagonal rings were selected as inner layer. All of models with and without NH_3 were optimized using ONIOM (B3LYP/6-31 G*:AM1) and adsorption energy (E_{ad}) were computed (Table 1). The E_{ad} is defined here as follows:

$$E_{ad} = E_{tot}(NH_3 + \text{open} - \text{ended BNNT}) - E_{tot}(\text{open} - \text{ended BNNT}) - E_{tot}(NH_3),$$

where E_{tot} is the total energy of a given system.

In addition, the other low-levels were applied in the ONIOM study of **A3** model including the semiempirical MNDO and PM3 methods and also the molecular mechanics universal force field (UFF). All calculations performed on **A** model were repeated for **B** model, as well. Finally, to explore the effect of tube length and diameter on the consistency of ONIOM and DFT, the A3 model of three other BNNTs were studied, including: (5,0), (6,0) and (7,0) zigzag types.

Results and discussion

At first, we probed reliability of the ONIOM(B3LYP/6-31 G*:AM1) level of theory in calculation of geometrical parameters and E_{ad} of NH_3 dissociation at open ends of **A1**, **A2**, **A3**, and **A4** models. B3LYP, has the most generality and predictive capacity providing a sufficiently accurate description of finite-size nanotubes [15–19]. As we have recently showed, during the NH_3 adsorption process, a single NH_3 molecule dissociates into an -H atom and an -NH_2 group at the open ends of BNNTs [10].

In the present work, we compare the results of ONIOM method with those of full DFT in ref. [10]. The calculated E_{ad} amounts for **A1**, **A2**, **A3** and **A4** models are −80.8, -70.5, -64.8 and −74.7 kcal mol^{-1}, respectively (Table 1). The results indicate that only the E_{ad} of NH_3 on **A3** model partly agrees with B3LYP/6-31 G* result (−63.1 kcal mol^{-1}) and the energy difference is about 1.7 kcal mol^{-1} (with 3% error). As shown in Table 1, the other models show significant errors. The calculated values for three representative bond lengths including, **N1**-NH_2, **N2**-H and **N1**-**N2** (Table 1, Fig. 3), indicate that geometrical parameters have no dependency upon the type of tube models.

However, adopting the **A3** as the appropriate model, we subsequently compared the reliability of four different low-levels (AM1, PM3, UFF and MNDO) in E_{ad}, activation energy (E_{act}), HOMO-LUMO energy gap (E_g), charge transfer (Q_T), dipole moment (μ) and structure geometry

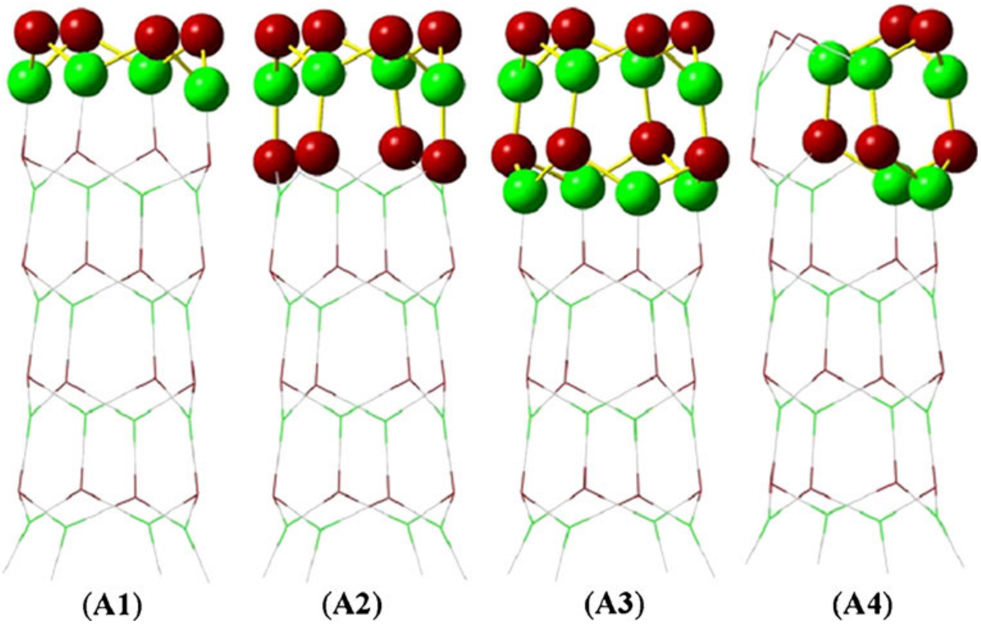


Fig. 2 Four optimized models of the **A** tube at ONIOM (B3LYP/6-31 G*:AM1)

calculations. The E_g, μ and Q_T were computed using full B3LYP/6-31 G* level of theory (performing a single point (SP) calculation on the optimized structures of ONIOM), due to inability of the ONIOM method in their calculations. Frequency calculations verified the obtained transition structures (with one imaginary frequency). Ugliengo et al. showed that the results of frequency calculations using ONIOM method are comparable with full DFT ones [20–22].

All calculated parameters are shown in Table 2, indicating that the results of E_{ad} are near to that of ref. [10], except those of UFF, showing an error of 7.4%. However, among all low-levels, E_{ad} of MNDO is the nearest to that of B3LYP. We once more performed a SP energy calculations on the **A3** and H-**A3**-NH_2 complexes, using B3LYP/6-31 G*. It is observed that the difference of E_{ad} between ONIOM and full DFT is relatively reduced in the all cases (Table 2). It seems that performing a high-level SP calculation on the ONIOM-based optimized structures improve the initial results of E_{ad}.

The UFF has the most deviation in E_g calculations. Either PM3 or MNDO overestimates E_g about 0.11 eV, while AM1 underestimates it about 0.13 eV. Generally, all methods give eligible results, except the UFF. In E_{act} calculations, we observe a good consistency between ONIOM and DFT as the result of PM3 is in the best agreement with that of full DFT.

In the cases of structure geometry and Q_T calculations, the results of the different ONIOM methods show no significant difference in comparison to those of DFT, indicating that these properties are not dependent upon low-levels. Finally, the largest and lowest deviations belong to the UFF and PM3, respectively, in the case of μ calculation.

Table 1 E_{ad} (kcal mol^{-1}) and representative bond lengths (Å) of different models of **A** tube, which are compared with the results of full DFT of ref. [10]

Model	E_{ad}	Error	N1-NH_2	N2-H	[a]N1, N2
A1	−80.8	28%	1.449	1.023	2.562
A2	−70.5	12%	1.434	1.025	2.541
A3	−64.8	3%	1.454	1.023	2.497
A4	−74.7	18%	1.459	1.025	2.553
A	−63.1	-	1.455	1.022	2.495

[a] N1, N2 is the distance between N1 and N2

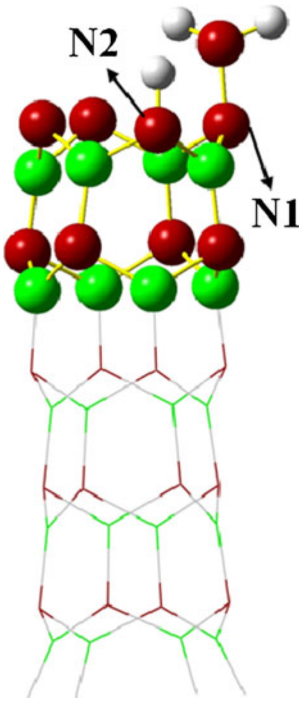


Fig. 3 The structural geometry of **A3**/NH_3 complex with different ONIOM methods

Table 2 The E_{ad} and E_{act} (in kcal mol^{-1}) and Q_T for NH_3 adsorption at open ends of **A3** model and the μ (debye), E_g and representative bond lengths (Å) of **A3** model, calculated at various low levels. These date are compared with those of full DFT of ref. [10]

Method	E_{ad}	E_{ad}(SP)	E_{act}	Q_T (e)	E_g (eV)	μ	N1-NH_2	N2-H	[b]N1, N2
AM1	−64.8	−63.5	11.3	0.39	2.64	2.60	1.454	1.022	2.496
PM3	−62.2	−63.4	13.9	0.39	2.88	1.96	1.455	1.022	2.495
MNDO	−63.0	−63.4	15.1	0.39	2.88	2.60	1.455	1.022	2.498
UFF	−67.8	−65.9	[a]-	0.40	3.33	3.12	1.452	1.022	2.495
Ref. [10]	−63.1	−63.1	13.7	0.39	2.77	1.74	1.455	1.022	2.494

[a] Using this low level of theory, the TS structure was not found. [b] N1, N2 is the distance between N1 and N2

However, PM3 is the most reliable among all low-level methods and the results of UFF are the most misleading, especially in the calculations of E_{ad}, E_{act} and μ. The results of MNDO are somewhat similar to those of PM3 by experience, the PM3 usage is difficult in comparison to that of MNDO, due to many convergence failures in the Gaussian program.

Subsequently, we assessed the reliability of the ONIOM method in E_{ad} calculation of NH_3 dissociation at the open end of **B** model, using the same three semiempirical low-levels. The data of Table 3 show that the results of ONIOM method for all **B** models are not in agreement with DFT. The best agreement belongs to the **B3** model with 25% error. It is noteworthy to mention that the designation of **B** models is similar to those of the **A** models. In the **B3** case, the calculated E_{ad} values for AM1, PM3, and MNDO are −156.5, -202.5 and −182.6 kcal mol^{-1} while that of full DFT is −131.1 kcal mol^{-1} [10].

In contrast to the case of **A**, the results of ONIOM method significantly differ with that of DFT in all models of **B**. For example in **B3** model, the calculated errors are 26%, 63% and 47% for AM1, PM3, and MNDO, respectively. This induces very cautiously usage of ONIOM method in computational studies.

Table 3 E_{ad} values (kcal mol^{-1}) and representative bond lengths (Å) of different models of **B** tube, which compared with the results of full DFT of ref. [10]

Model	E_{ad}	Error(%)	B1-NH_2	B2-H	[a]B1,B2
B1	−491.6	275	1.385	1.192	2.419
B2	−866.2	561	1.383	1.191	2.346
B3	−163.6	25	1.386	1.190	2.481
B4	−863.8	559	1.382	1.196	2.427
B	−131.1	-	1.386	1.191	2.424

[a] B1, B2 is the distance between B1 and B2

Here, we interpret in detail why the ONIOM method is appropriate for **A3** model, while it is not for **B3** (with the same atoms in low-level). To this end, we performed Mulliken population analysis on the **A3** and **B3** tubes and their NH_3 adsorbed complexes. It is noteworthy to say that our main objective is a comparative study and, the exact values of Mulliken charges (MCs) is not the purpose of the present manuscript. However, during the NH_3 adsorption process, the MCs of atoms change within the tubes, whereas the amounts of changes reduce going away from their ends.

The percentage of MC changes were computed during the adsorption process for five layers at the open end of both **A3** and **B3**. We have designated the name of “layer” for a row of N atoms with a row of B atoms. The results are depicted in Fig. 4 and collected in Table 4. In the case of **A3**, the changes are 34.7% and 10.1% for layers 1 and 2 (region **1**) while those are only 2.6%, 0.3% and 0.1% for layers 3, 4 and 5 (region **2**), respectively. It demonstrates that the MCs of atoms of region **1** significantly change during the adsorption process, while their changes in region **2** are not significant.

In other words, region **1** is a chemically active site and it is necessary to locate in high-level of ONIOM scheme

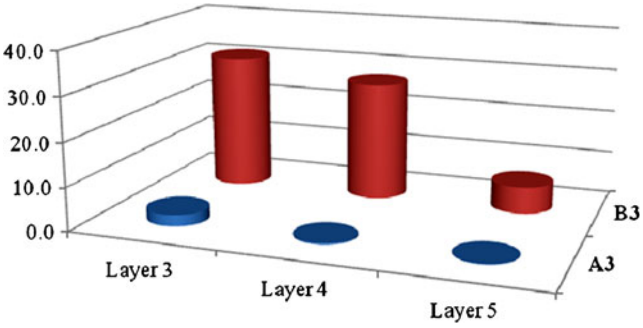


Fig. 4 The percentage change of Mulliken charges in layers 3, 4 and 5 of both **A3** and **B3** models

Table 4 The percentage Mulliken charges changes in layers 3, 4 and 5 of both **A3** and **B3** models

Layer	A3	B3
3	2.6%	31.3%
4	0.3%	27.3%
5	0.1%	6.2%

while region **2** is not a sufficiently active area and can be placed in low-level. As discussed above, this strategy has been applied in our ONIOM calculations for the case of **A3**, justifying the reliable results. In the models of **A1**, **A2** and **A4** some atoms of region **1** are located in low-level of theory, justifying the large errors in the E_{ad} results.

In the case **B3**, we observed different results so as the changes of MCs in region **2** are not negligible. These changes are 31.3%, 27.3% and 6.2% in layers 3, 4 and 5, respectively. This suggests that region **2** may be a part of chemically active site that should be studied in high-level of theory. As shown before, this region has been placed in low-level of theory and it may be the origin of the large errors in E_{ad} of **B3** model. As a result, we conclude that the ONIOM may be a reliable scheme in the study of weak interactions, while in the case of strong interactions it is a controversial approach and should be applied cautiously. In other words, in the case of ONIOM-based strong interaction studies, more atoms should be located in high-level in comparison to that of weak interaction studies.

Subsequently, we probed the effect of tube length and diameter on reliability of ONIOM results. Here we only considered case **A**, because the ONIOM results of **B** models are very different from those of DFT. To this end, we calculated E_{ad} values of NH_3 dissociation at open ends of **A3** models of (5,0), (6,0) and (7,0) tubes with ONIOM (B3LYP/6-31 G*:AM1) and full B3LYP/6-31 G*. The data of Table 5 indicate that the results of ONIOM and full DFT are in best agreement for different diameter tubes. In addition, to investigate the effect of length, we considered the (5,0)-**A3** model with various length including: 9, 8, 7, 6 and 5 layers. The results (Table 6) show that E_{ad} values for all tubes with various lengths do not differ significantly for both ONIOM and DFT and there is good consistency between these approaches. Generally, we conclude that the agreement between the results of ONIOM and DFT approaches are independent of tube length and diameter.

Table 5 The E_{ad} values of NH_3 dissociation at the open ends of **A3** model achieved from ONIOM and DFT. The energy unit is kcal mol^{-1}

BNNT	E_{ad}(DFT)	E_{ad}(ONIOM)
(4,0)	−63.1	−64.8
(5,0)	−73.7	−73.4
(6,0)	80.1	−82.4
(7,0)	−86.5	−87.2

Table 6 The calculated E_{ad} for NH_3 dissociation at open ends of (5,0)-**A3** models with different models. The unit of energy is kcal mol^{-1}. The first column shows the number of tube layers

(5,0)-A3	E_{ad}(DFT)	E_{ad}(ONIOM)
5	−73.7	−73.3
6	−73.6	−73.3
7	−74.2	−73.3
8	−74.3	−73.3
9	−75.6	−73.2

We mention that the absolute values of E_{ad} are increased as the tube diameter is elongated. This phenomenon is justified as the weakening of tube edge bonds resulting from the bond length elongation because of diameter enlargement.

Conclusions

We used the ONIOM method to calculate the adsorption energies (E_{ad}), transition structures, the change of HOMO-LUMO energy gaps (E_g) and structure geometries of the NH_3 adsorption on the **A** (N-enriched) and **B** (B-enriched) models of open ended BNNTs. Different low-levels including, AM1, PM3, MNDO and UFF have been investigated, applying B3LYP/6-31 G* in all high-levels of ONIOM calculations. PM3 method is the most reliable among all low-levels used here especially in calculation of E_{ad}, E_{act} and dipole moment and UFF is the most misleading. Either PM3 or MNDO overestimates the E_g, while AM1 underestimates it. We showed that in the case of **A**, by selecting two atom layers of the open end of the tube as inner layer, the results of ONIOM approach is in best agreement with those of the pure DFT calculations while in the case of **B** with the same condition, the results of ONIOM significantly differ with those of DFT. Finally, the results demonstrated that the ONIOM might be a reliable method in the study of weak interactions while in the case of strong interactions it is a controversial approach and should be applied cautiously. In addition, we showed that the agreement between the results of ONIOM and DFT approaches are independent of tube length and diameter.

References

1. Froese R, Humbel S, Svensson M, Morokuma K (1997) IMOMO (G2MS): a new high-level G2-like method for large molecules and its applications to Diels−Alder reactions. J Phys Chem A 101:227–233
2. Svensson M, Humbel S, Froese R, Matsubara T, Sieber S, Morokuma K (1996) ONIOM: a multilayered integrated MO + MM method for geometry optimizations and single point energy predictions a test for Diels−Alder reactions and $Pt(P(t\text{-}Bu)_3)_2 + H_2$ oxidative addition. J Phys Chem 100:19357–19363
3. Wang L, Yi C, Zou H, Gan H, Xu J, Xu W (2011) Initial reactions of methyl-nitramine confined inside armchair (5,5) single-walled carbon nanotube. J Mol Model. doi:10.1007/s00894-011-0967-x
4. Lu X, Yuan Q, Zhang Q (2003) Sidewall epoxidation of single-walled carbon nanotubes: a theoretical prediction. Org Lett 5:3527–3530
5. Li F et al. (2006) Theoretical study of hydrogen atom adsorbed on carbon-doped BNnanotubes. Phys Lett A 357:369–373
6. Peralta-Inga Z et al. (2002) Characterization of surface electrostatic potentials of some (5,5) and (n,1) carbon and boron/nitrogen model nanotubes. Nano Lett 3:21–28
7. Wu HS, Cui XY, Qin XF, Strout D, Jiao H (2006) Boron nitride cages from B12N12 to B36N36: square–hexagon alternants vs boron nitride tubes. J Mol Model 12:537–542
8. Chu Y-Y, Su M-D (2004) Theoretical study of addition reactions of carbene, silylene, and germylene to carbon nanotubes. Chem Phys Lett 394:231–237
9. Bettinger HF (2004) Effects of finite carbon nanotube length on sidewall addition of fluorine atom and methylene. Org Lett 6:731–734
10. Ahmadi A, Beheshtian J, Hadipour N (2011) Chemisorption of NH_3 at the open ends of boron nitride nanotubes: a DFT study. Struct Chem 22:183–188
11. Blum O, Milstein D (2002) Oxidative addition of water and aliphatic alcohols by $IrCl(trialkylphosphine)_3$. J Am Chem Soc 124:11456–11467
12. Frey GD, Lavallo V, Donnadieu B, Schoeller WW, Bertrand G (2007) Facile splitting of hydrogen and ammonia by nucleophilic activation at a single carbon center. Science 316:439–441
13. Zhao J, Goldman AS, Hartwig JF (2005) Oxidative addition of ammonia to form a stable monomeric amido hydride complex. Science 307:1080–1082
14. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski J, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, AlLaham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PM, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (1998) Gaussian Inc. Pittsburgh, PA
15. Kilina S, Badaeva E, Piryatinski A, Tretiak S, Saxena A, Bishop AR (2009) Bright and dark excitons in semiconductor carbon nanotubes: insights from electronic structure calculations. Phys Chem Chem Phys 11:4113–4123
16. Tabtimsai C, Keawwangchai S, Wanno B, Ruangpornvisuti V (2011) Gas adsorption on the Zn–, Pd– and Os–doped armchair (5,5) single–walled carbon nanotubes. J Mol Model. doi:10.1007/s00894-011-1047-y
17. Ruangpornvisuti V (2010) Molecular modeling of dissociative and non-dissociative chemisorption of nitrosamine on close-ended and open-ended pristine and Stone-Wales defective (5,5) armchair single-walled carbon nanotubes. J Mol Model 16:1127–1138
18. Ahmadi A, Beheshtian J, Hadipour NL (2011) Interaction of NH_3 with aluminum nitride nanotube: electrostatic vs. covalent. Physica E 43:1717–1719
19. Ahmadi A, Kamfiroozi M, Beheshtian J, Hadipour N (2011) The effect of surface curvature of aluminum nitride nanotubes on the adsorption of NH_3. Struct Chem. doi:10.1007/s11224-011-9820-1
20. Rimola A, Tosoni S, Sodupe M, Ugliengo P (2006) Does silica surface catalyse peptide bond formation? New insights from first-principles calculations. ChemPhysChem 7:157–163
21. Rimola A, Zicovich-Wilson CM, Dovesi R, Ugliengo P (2010) Search and characterization of transition state structures in crystalline systems using valence coordinates. J Chem Theor Comput 6:1341–1350
22. Roggero I, Civalleri B, Ugliengo P (2001) Modeling physisorption with the ONIOM method: the case of NH_3 at the isolated hydroxyl group of the silica surface. Chem Phys Lett 341:625–632

J Mol Model (2012) 18:1735–1753
DOI 10.1007/s00894-011-1179-0

ORIGINAL PAPER

Combinatorially-generated library of 6-fluoroquinolone analogs as potential novel antitubercular agents: a chemometric and molecular modeling assessment

Nikola Minovski · Andrej Perdih · Tom Solmajer

Received: 14 April 2011 / Accepted: 4 July 2011 / Published online: 12 August 2011

Abstract The virtual combinatorial chemistry approach as a methodology for generating chemical libraries of structurally-similar analogs in a virtual environment was employed for building a general mixed virtual combinatorial library with a total of 53.871 6-FQ structural analogs, introducing the real synthetic pathways of three well known 6-FQ inhibitors. The druggability properties of the generated combinatorial 6-FQs were assessed using an *in-house* developed drug-likeness filter integrating the Lipinski/Veber rule-sets. The compounds recognized as drug-like were used as an external set for prediction of the biological activity values using a neural-networks (NN) model based on an experimentally-determined set of active 6-FQs. Furthermore, a subset of compounds was extracted from the pool of drug-like 6-FQs, with predicted biological activity, and subsequently used in virtual screening (VS) campaign combining pharmacophore modeling and molecular docking studies. This complex scheme, a powerful combination of chemometric and molecular modeling approaches provided novel QSAR guidelines that could aid in the further lead development of 6-FQs agents.

Keywords Antibacterial agents · CombiChem · DNA gyrase · Fluoroquinolones · Molecular docking · Pharmacophore modeling · QSAR · Tuberculosis

Abbreviations

TB	Tuberculosis
ATP	Adenosine triphosphate
MIC	Minimal Inhibitory Concentration
6-FQs	6-Fluoroquinolones
SAR	Structure-Activity Relationships
QSAR	Quantitative Structure-Activity Relationships
CombiChem	Combinatorial Chemistry
SSS	Substructure Search
NN	Neural-Networks
KANN	Kohonen Artificial Neural Networks
CP ANN	Counter-Propagation Artificial Neural Networks
GHA	Global Hypothetical Activity
LBP	Ligand-Based Pharmacophore
SBP	Structure-Based Pharmacophore
VS	Virtual Screening

Introduction

Tuberculosis (TB), the ingeniously transferable bacterial infection, is still one of the global health concerns [1]. *Mycobaterium tuberculosis*, the accountant agent of tuberculosis, is a resistful pathogen microorganism responsible for infecting about one third (two billion people) of the human population and in the process causing around two million death cases each year worldwide (World Health Organization, 2003) [2]. TB is mainly caused by the pathogen *M. tuberculosis*, but in some cases the microorganisms such as *M. fortuitum*, *M. smegmatis* and *M. avium-intracellulare* complex (MAC) can also be involved in the disease development [3–5].

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1179-0) contains supplementary material, which is available to authorized users.

N. Minovski · A. Perdih · T. Solmajer (✉)
National Institute of Chemistry,
Hajdrihova 19,
1001, Ljubljana, Slovenia
e-mail: tom.solmajer@ki.si

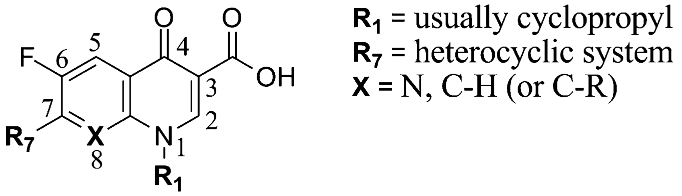
The tuberculosis treatment is mainly chemotherapeutically based, with the whole therapy requiring between 6 to 9 months or even longer to be successful. The problem of drug resistance and the continuous onset of multidrug resistant lethal TB strains in most cases are attributed to the potential toxicity of the chemotherapeutics, the durability of the whole treatment, as well as frequent poor patient compliance to the therapy regimen. The increased development of resistant TB mutants is one of the additional challenging factors to stimulate the design of novel chemotherapeutic agents which will be effective against resistant *Mycobacteria* [6].

One of the validated and well known molecular targets of quinolone antibiotics in *Mycobacteria* species is the DNA gyrase which belongs to the topoisomerase group of enzymes [7]. This enzyme catalyzes the ATP dependent process of introduction of negative supercoils into closed circular DNA as well as the relaxation of the supercoiled DNA molecule (ATP-independent catalysis) [8, 9]. DNA gyrase forms a functional heterodimer A_2B_2 consisting of two major subunits, GyrA and GyrB. The GyrA subunit is responsible for the process of breakage and reunion of the double-stranded DNA, i.e., activation of the process of DNA replication and elongation and together with GyrB is involved in the maintaining of the topological state of DNA molecule [10, 11]. Another structurally-similar enzyme which belongs to the topoisomerase group is the DNA topoisomerase IV. Like the DNA gyrase, this bacterial enzyme (a paralogue of DNA gyrase) also forms a functional heterodimer consisting of two subunits ParC and ParE (homologues of the GyrA and GyrB, respectively). Recent structural studies revealed that the quinolone antibiotics establish an interaction with the DNA breakage-reunion domain of the DNA gyrase and topoisomerase IV, stabilizing the covalent topoisomerase/DNA cleavage complex, leading to a blockade of DNA replication [12].

In the last few decades, tuberculosis chemotherapy was mainly based on the active agents belonging to two main categories of antibiotics, the bacterial cell wall inhibitors (isoniazide, ethambutol) and bacterial nucleic acid synthesis inhibitors (quinolones, rifampicin) [13]. The second category of antibiotics, especially the quinolone chemical class of chemotherapeutics is increasingly gaining importance in targeting *Mycobacteria*, because of their effective, strong and invasive mechanism of action. Fluoroquinolones belong to the quinolone's class of DNA gyrase inhibitors which have a fluorine atom attached to the main scaffold at the 6 position (Fig. 1) [14]. The mechanism of bactericidal action is based on the inhibition of the bacterial DNA synthesis process through a scission of the natural mycobacterial DNA molecule leading to a topological stress of DNA and bacterial cell death [15].

The structure-activity relationships (SAR) studies showed that the main quinolone core (1,4-dihydro-4-oxo-3-pyridinecarboxylic acid moiety) is of most significant importance for the anti-mycobacterial activity. Furthermore, unlike the cyclopropyl group at position 1 which is apparently optimal for biological activity, each substitution at positions 2, 3, and 4 will result in a significant loss of biological activity. Substitutions at positions 5 and 8 of the main quinolone core interfere with the required planarity of the system. Hydrogen and amino groups have been a good replacement for the fluorine atom at position 6 leading to an improved in vitro activity, but such modifications are not always followed by an improved in vivo activity. The substitutions at position 7 of the main scaffold are of significant importance for the biological activity as this position directly interacts with the DNA gyrase (5- or 6-membered N-hetero systems such as aminopyrrolidines and piperazines are optimal for activity) [16]. Today, the use of targeted synthetically-feasible chemical libraries significantly enhanced the hit identification as well as lead optimization phase leading to favorable novel drug candidates. Similarly, quantitative-structure activity relationship (QSAR) models proved to be useful in silico tools for the rationalization of the experimental SAR properties in the form of quantitative mathematical models which can be subsequently used for an efficient prediction of biological activity values for novel as well as unknown compounds. These tools can significantly enhance and enrich the screening process of the chemical libraries under investigation [17]. Furthermore, the three-dimensional assessment and screening of the generated molecules for optimal interaction with the binding site is of significant importance in the drug development processes [18]. For instance, ligand-based drug design approaches, such as pharmacophore modeling, which are taking into account the spatial orientation

Fig. 1 Generic structure of 6-fluoroquinolone antibiotics



of the ligand's functional groups and scaffold shape complementarity can be utilized for the construction of predictive pharmacophore models which can be more effective for virtual screening (VS). Such a model based on the similarity of the pharmacophoric features is particularly useful, if it is able to identify (recognize) active compounds among a pool of inactive molecules. Moreover, the availability of three-dimensional (3D) structure of the protein-ligand complex of interest enables the implementation of the structure-based approaches especially molecular-docking calculations of ligands into the defined binding site of the protein and the investigation of ligand-protein molecular interactions [19, 20].

The present study introduces an effective methodology for the in silico generation of novel unknown 6-fluoroquinolone (6-FQ) analogs using a combinatorial chemistry approach coupled with the prediction of values of biological activity by employing a previously derived neural-networks (NN) [21]. Furthermore, a three-dimensional pharmacophore approach and molecular docking calculations were used to assess the generated virtual library utilizing the available experimental data to select the most promising drug-like compounds. This scheme – a powerful combination of the chemometric and VS tools – is aimed to establish new possible SAR guidelines and trends in 6-fluoroquinolone optimization, which could be of importance in the on-going antibacterial lead discovery programs (Fig. 2).

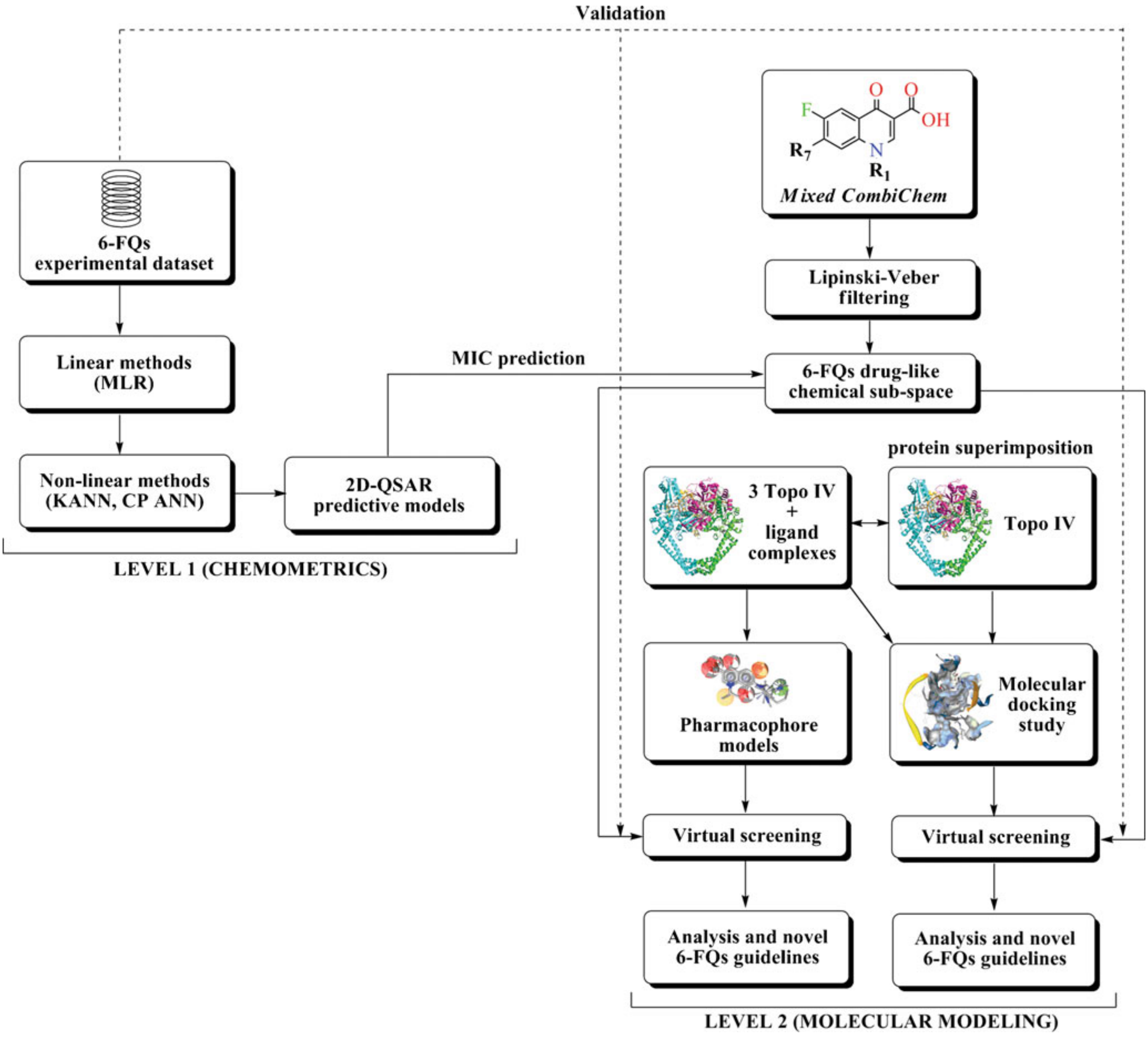

Validation
6-FQs experimental dataset
Linear methods (MLR)
Non-linear methods (KANN, CP ANN)
2D-QSAR predictive models
MIC prediction
LEVEL 1 (CHEMOMETRICS)
F
O
O
OH
N
R7
R1
Mixed CombiChem
Lipinski-Veber filtering
6-FQs drug-like chemical sub-space
protein superimposition
3 Topo IV + ligand complexes
Topo IV
Pharmacophore models
Molecular docking study
Virtual screening
Virtual screening
Analysis and novel 6-FQs guidelines
Analysis and novel 6-FQs guidelines
LEVEL 2 (MOLECULAR MODELING)

Fig. 2 The overall workflow depicting different stages of the performed chemometric and molecular modeling study

Computational details

Virtual combinatorial generation

One of the most popular strategies in modern drug design is the virtual synthesis of analogs of existing compounds for which the bioavailability and toxicity studies are already performed and have displayed activity and potency in human therapy. The implementation of SAR-based scaffold modifications to obtain novel molecules for the target under investigation is a well-known and advantageous strategy for analog design. There are several software tools available integrating the combinatorial algorithm for in silico virtual combinatorial generation. *CombiChem* add-on modul available in ChemBioOffice Ultra was used for in silico generation of 6-fluoroquinolone structural analogs [22]. For this purpose we employed the original synthetic pathways of three well known 6-FQ antibiotics: ciprofloxacin, moxifloxacin, and ofloxacin [23–25]. Using a building-blocks commercial dataset (Bionet Fragment Library of 6995 "Rule of 3" filtered lead-like fragments) [26], and SAR-based structural modifications at positions 1 and 7 of the main 6-FQ scaffold, we built six different subsets of combinatorially-generated 6-fluoroquinolone analogs.

The virtual environment enables a rational simplification of the real synthetic reactions taking into account only the crucial synthetic steps for derivation of the final product (structurally-similar 6-FQ analogs). Such a rational virtual simplification resulted in two-steps virtual synthetic pathways for ciprofloxacin and moxifloxacin (yielding R_1-monosubstituted intermediate product, as well as all possible R_1,R_7-disubstituted 6-FQ analogs as final products), and one-step virtual synthetic pathway for ofloxacin (obtaining all possible R_7-monosubstituted analogs as final products). Taking into account that position 7 of the main 6-FQ moiety directly interacts with the DNA gyrase and is of significant importance for activity [16], additionally we introduced several different substituents at this position of the main scaffold with non-amine attachment points. These substructural modifications enable a total of six virtual combinatorial synthetic mechanisms to be defined (Fig. 3):

(1) 7-amino substituted ciprofloxacin's structural analogs.
(2) 7-non-amino substituted ciprofloxacin's structural analogs.
(3) 7-amino substituted moxifloxacin's structural analogs.
(4) 7-non-amino substituted moxifloxacin's structural analogs.
(5) 7-amino substituted ofloxacin's structural analogs.
(6) 7-non-amino substituted ofloxacin's structural analogs.

Fragments selection for combinatorial enumeration

A building-blocks commercial library of 6995 substructural fragments was used as a source for the fragment selection for combinatorial enumeration [26]. Each fragment in such a library is defined by the "rule of 3" (MW≤300; *n*HBD, *n*HBA, *n*RB≤3), where MW is the molecular weight, while *n*HBD, *n*HBA, and *n*RB, are number of hydrogen bond donors, number of hydrogen bond acceptors, and number of rotatable bonds, respectively [27]. Fragments selection was performed by employing the substructure search (SSS) algorithm. The selection procedure resulted in the extraction of all possible substructural fragments (primary amines and secondary amines/non-amines) defined by the virtual synthetic pathways. Substructural fragmental subsets selected by the SSS algorithm for each virtual synthetic pathway were visually inspected for the presence of possible unwanted species (salt forms, mixtures, charged forms). These substructural fragments were excluded from the combinatorial procedure implementing a simple Boolean filtering algorithm (Y/N (Yes/No)). Furthermore, all substructural fragments containing more than one attachment point (two or more amino groups) were eliminated. The rest of the substructural fragments with (Y) output obtained using this filtering procedure (described in details in Table 1 for each virtual combinatorial definition), were used as a starting point for combinatorial enumeration.

Combinatorial enumeration

The combinatorial enumeration procedure was performed by implementing a standard methodology [28] for generating structural analogs in virtual environment (*CombiChem*) [22]. This procedure of statistical non-repetitive fragmental permutation of the selected building-blocks subsets (R_1: primary amines and R_7: secondary amines/non-amines)) in previously defined variable positions (1 and 7) within the main 6-FQ scaffold generated all possible virtual 1,7 substituted 6-FQ structural analogs. The total number of analogs (ρ_{tot}) obtained using this methodology (Table 1), mathematically can be estimated as a multiplication product between the total number of selected substructural fragments permutating at position 1 (R_1: primary amines, N_i) and total number of selected substructural fragments permutating at position 7 (R_7: secondary amines/non-amines, M_j) within the main 6-FQ core, using the following equation (Eq. 1) [29]:

$$\rho_{tot} = N_i \times M_j. \qquad (1)$$

The obtained products (ρ_{tot}=53.871) belong to three main categories of 6-FQ analogs: ciprofloxacin analogs (7-amino substituted analogs (12.296), 7-non-amino substituted analogs (21.965)), moxifloxacin analogs (7-amino substituted

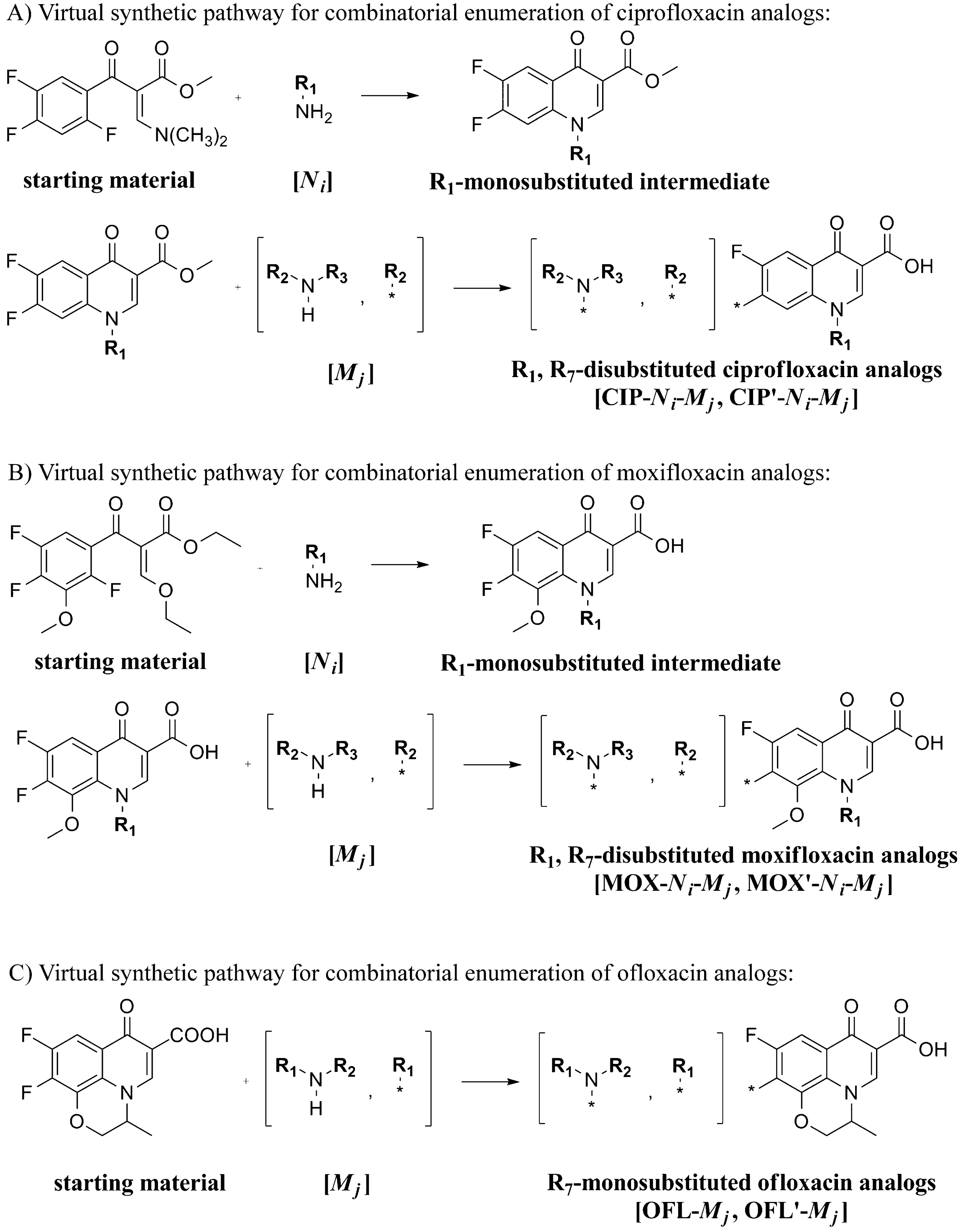

A) Virtual synthetic pathway for combinatorial enumeration of ciprofloxacin analogs:
starting material
[N_i]
R1-monosubstituted intermediate
[M_j]
R1, R7-disubstituted ciprofloxacin analogs
[CIP-N_i-M_j, CIP'-N_i-M_j]
B) Virtual synthetic pathway for combinatorial enumeration of moxifloxacin analogs:
starting material
[N_i]
R1-monosubstituted intermediate
[M_j]
R1, R7-disubstituted moxifloxacin analogs
[MOX-N_i-M_j, MOX'-N_i-M_j]
C) Virtual synthetic pathway for combinatorial enumeration of ofloxacin analogs:
starting material
[M_j]
R7-monosubstituted ofloxacin analogs
[OFL-M_j, OFL'-M_j]

Fig. 3 Generic virtual synthetic pathways for combinatorial enumeration of three different 6-FQs: (A) ciprofloxacin analogs (CIP-N_i-M_j, CIP′-N_i-M_j), (B) moxifloxacin analogs (MOX-N_i-M_j, MOX′-N_i-M_j), (C) ofloxacin analogs (OFL-M_j, OFL′-M_j)

analogs (8.510), 7-non-amino substituted analogs (10.731)), and ofloxacin analogs (7-amino substituted analogs (180), 7-non-amino substituted analogs (189)). The following designation scheme was used: ciprofloxacin analogs (7-amino substituted analogs (CIP-N_i-M_j), 7-non-amino substituted analogs (CIP′-N_i-M_j)), moxifloxacin analogs (7-amino substituted analogs (MOX-N_i-M_j), 7-non-amino substituted analogs (MOX′-N_i-M_j)), and ofloxacin analogs (7-amino substituted analogs (OFL-M_j), 7-non-amino substituted analogs (OFL′-M_j)). Combining all subsets of analogs

Table 1 Fragments selection details for combinatorial enumeration. The building-blocks (R_7-substructural fragments) with non-amino attachment point are signed with asterisk (*)

ID	Virtual combinatorial definition		Substructural fragments		Boolean output (Y/N)				ρ_{tot}
	R_1	R_7	R_1	R_7	R_1		R_7		
					Y (N_i)	N	Y (M_j)	N	
1.	R_1-NH_2	R_2R_3-NH	126	363	116	10	106	257	12.296
2.	R_1-NH_2	R_2-*	123	340	115	8	191	149	21.965
3.	R_1-NH_2	R_2R_3-NH	123	347	115	8	74	273	8.510
4.	R_1-NH_2	R_2-*	123	337	73	50	147	190	10.731
5.	N/A	R_1R_2-NH	N/A	348	N/A	N/A	180	168	180
6.	N/A	R_1-*	N/A	337	N/A	N/A	189	148	189

obtained by using the combinatorial algorithm, resulted in building a general virtual combinatorial library (*CombiTot*) with a total of 53.871 6-FQ structural analogs, which were subsequently used for assessing the drug-like properties (all structures are available as *.sdf file format in electronic supplementary material, online resource 1).

Drug-likeness assessment using chemometric approaches

Albeit this combinatorial library contains a considerable number of 6-FQ analogs (*CombiTot*, ρ_{tot}=53.871) it can be considered as a small sub-space in the available chemical space [30]. Since the combinatorial algorithm in first instance increased the molecular complexity, there is a low probability that each compound in such a virtual library will possess drug-like properties. Therefore, drug-likeness filtering was performed taking into consideration the rules of Lipinski [31] and Veber [32]. Using the ChemBioOffice pre-integrated cheminformatics functions [22], we developed a robust drug-likeness filtering tool integrating both rule-sets (Lipinski "rule of 5" and Veber rules) which was used for filtering of our general combinatorial library and defining the drug-like chemical sub-space of 6-FQ analogs. This filtering procedure yielded a list of 1.101 out of 53.871 virtual combinatorial 6-FQ analogs as promising compounds for further investigation. These compounds were further assessed for prediction of their unknown activity values (pMIC$_{pred\text{-}combi}$) using a pre-built QSAR model [21]. Combinatorially-generated 6-FQ analogs possessing drug-like properties (*CombiDL*, 1.101 analogs) are available as "electronic supplementary material" (*.sdf file format, online resource 2).

Prediction of the biological activity values using a derived non-linear neural networks (NN) model

Our previously published non-linear neural networks (NN) model (see [21] for details) was used for prediction of biological activity values (pMIC$_{pred\text{-}combi}$) for the external combinatorial library of 1.101 novel unknown 6-FQ structural analogs. The model was built employing counter-propagation artificial neural networks (CP ANN) methodology using a dataset of 145 structurally-similar 6-fluoroquinolones (all analogs are available as *.sdf file format in electronic supplementary material, online resource 3) with experimentally-determined biological activity values (pMIC$_{exp}$) against *M. tuberculosis* and an extensive set of nearly 600 calculated 2D molecular descriptors. Heuristic algorithm and intercorrelation matrix were used for selection of statistically-significant molecular descriptors for activity. The Heuristic algorithm which is a suitable method for pre-selection of molecular descriptors is first calculating the one-parameter correlation equations between molecular descriptors and activity, eliminating all descriptors that do not fulfill the pre-defined criteria: (1) The Fisher F-test value for the one-parameter correlation with the descriptor is less than 1.00, (2) The squared correlation coefficient (R^2) for the one-parameter equation is less than $R_{\min}{}^2$ (the default value is $R_{\min}{}^2$=0.1), (3) t-test value is less than $t1$ (where $R_{\min}{}^2$ and $t1$=1.5), and (4) the descriptor is highly intercorrelated with another descriptor (above r_{full}, where r_{full}=0.99), and this other descriptor has a higher squared correlation coefficient in the one-parameter equations based on these descriptors. With the remaining descriptors, the algorithm calculates all possible two- and more-parameter linear models [33]. Initially, several linear models with up to ten descriptors were developed. The frequency analysis of occurrence of the molecular descriptors between these models, resulted in selection of the most frequently occuring molecular descriptors which were used as input variables in neural-networks modeling part. Kohonen artificial neural networks (KANN) was employed for splitting the dataset into a training set (*Assay2*, 115 compounds) and an external validation set (*Assay2*, 30 compounds). The model was built on the training set using

CP ANN and internally validated employing the cross-validation leave-one-out procedure (R_{tr}=0.96, $R_{tr\text{-}cv}$=0.62, where R_{tr} designates the coefficient of correlation for the model, while $R_{tr\text{-}cv}$ is coefficient of correlation for cross-validation leave-one-out) as well as externally validated for its predictive performances using the external validation set (R_{val}=0.8454, where R_{val} is coefficient of correlation for the external validation set). Selection of a subset of combinatorially-generated 6-FQ analogs for pharmacophore analysis was performed by defining a global hypothetical activity (GHA) range 0.00≤MIC$_{pred\text{-}combi}$ [μg/mL]≤0.10 based on the determined experimental data for the following 6-FQs: *Structure2* (levofloxacin, MIC$_{exp}$=0.0115 μg/mL), *Structure21* (clinafloxacin, MIC$_{exp}$=0.01 μg/mL), and *Structure121* (moxifloxacin, MIC$_{exp}$=0.025 μg/mL) [21]. This selection resulted in the extraction of 427 out of 1.101 6-FQ analogs which had their activity predicted inside the pre-defined GHA range. These analogs were used as a drug-like combinatorial library (*CombiLib*) of novel 6-FQs for three-dimensional pharmacophore analysis and molecular docking study (all structures are available as *.sdf file format in electronic supplementary material, online resource 4).

Ligand-based and structure-based pharmacophore modeling of the 6-fluoroquinolone analogs

Although both DNA gyrase A (GyrA) and B (GyrB) subunits have been solved by the x-ray crystallography [34, 35], currently a full experimental atomistic picture of the GyrA$_2$GyrB$_2$ tetramer in complex with the DNA and 6-fluoroquinolone molecules remains unidentified. On the other hand, recent seminal studies on the closely related type II topoisomerase – topoisomerase IV – the complexes formed by the *Streptococcus pneumoniae* ParC (equivalent of GyrA subunit) and ParE TOPRIM domains of the topoisomerase IV together with 6-fluoroquinolones intercalated in the gap between nucleotides of the DNA were solved [36]. These data provided the first solid atomistic insights into the mechanism of action of the fluoroquinolone antibacterial agents. As both type II topoisomerases (DNA gyrase and topoisomerase IV) share a close structural and functional resemblance we used both available crystal structures to perform pharmacophore modeling and molecular docking calculations. Furthermore, for the 6-FQ chemical class a good correlation between the measured in vitro IC$_{50}$ values and in vivo MIC activities was observed. This enables utilization of the in vivo MIC data also for the interpretation of the atomistic pharmacophore and docking calculations [37].

Pharmacophore modeling approach is one of the widely used concepts in modern drug discovery [34]. Several software tools available today are able to provide predictive 3D pharmacophore models in an automated fashion. LigandScout [18], a software tool for automatic pharmacophore model generation from the available structural information was employed for the visualization and exploration of the topoisomerase IV binding site in complex with known 6-FQ active agents (structure-based design approach) as well as for constructing 3D-pharmacophore models based only on the 6-FQ ligand structures (ligand-based design approach). Three ligand-topoIV-DNA cleavage complexes from *S. pneumoniae*, available from the Protein Data Bank (PDB) repository [pdb codes: 3FOE, 3FOF, 3K9F], were used for the 3D-pharmacophore models generation [36, 38]. The models were obtained by automatic recognition of the 6-FQ co-crystallized ligand structures (clinafloxacin (3FOE), moxifloxacin (3FOF), and levofloxacin (3K9F)) and the surrounding amino acid residues of the 6-FQ analogs binding site and by analysis of possible ligand-protein interactions. Subsequently, the 3D structure-based pharmacophore (SBP) models were generated automatically (assignment of the pharmacophoric features for each ligand separately and ligand alignment generation) together with excluded volume spheres.

Therefore, two structure-based pharmacophore (SBP) models: shared (SBP$_{shared}$, interpolation of the identified pharmacophore features) and merged (SBP$_{merged}$, unification of the identified pharmacophore features) were constructed. Furthermore, the experimental conformations of the three bound 6-FQ analogs, were used for building of ligand-based pharmacophore (LBP) model. For each of the molecules, 250 unique conformations were calculated and aligned to yield a ligand-based pharmacophore (LBP) model. Finally, both combinatorial (*CombiLib*) and experimental library (*Assay 2*) were converted into multi-conformer libraries (maximum 25 conformations for each molecule) and subsequently screened with built-in LigandScout pharmacophore VS engine to evaluate the matching of the investigated compounds to the derived set of pharmacophore features.

Molecular docking of the 6-fluoroquiniolone analogs

Molecular docking of the 6-fluoroquinolone molecules into the type II topoisomerase crystal structure

The docking experiments were performed by using GOLD docking engine [39], and protein structure with the PDB code 3K9F was used to define the binding site located around the experimental coordinates of the bound 6-FQ inhibitor levofloxacin resulting in a cavity radius of 12.5 Å. Each investigated molecule was docked 10 times into the binding site by applying the following parameters of the GOLD genetic algorithm (GA) (population size=100, selection pressure=1.1, no. of operations=100000, no. of islands=5, niche size=2, migrate=10, mutate=95, crossover=95). Early termination was allowed if the top three solutions were within 1.5 Å of the root-mean-squared-deviation (RMSD) value. For the assessment of the binding affinity, GOLDscore scoring function was used [39].

The quality of the generated binding poses were validated by re-docking the levofloxacin into its binding site and exploring the positioning of the 145 6-FQ molecules with the available experimental MIC data [40]. Finally, all compounds of the generated reduced combinatorial library (*CombiLib*, 427 compounds, online resource 4), were docked into the defined binding site using the same procedure.

Results and discussion

Virtual combinatorial chemistry approach as a strategy for generating chemical libraries of structurally-similar analogs for a given compound of interest as well as their investigation implementing highly-sophisticated cheminformatics algorithms are extensively documented [41–45]. There are two crucial steps which must be taken into account for combinatorial generation of structural analogs for the entity under investigation: the synthetic procedure (generic synthetic reaction) for preparation of the compound as well as the introduction of substituents onto a scaffold [26]. In our case, a focused virtual combinatorial library of novel 6-FQ analogs (*CombiTot)* utilizing the established combinatorial algorithm for SAR-based scaffold permutation was obtained.

Druggability properties assessment and defining the drug-like chemical sub-space

The drug-likeness assessment of the combinatorially-generated library of 6-FQ structural analogs (*CombiTot*, 53.871 compounds) was carried out by calculating the druggability properties: MW, AlogP (Ghose-Crippen's atom-based model for logP, i.e., partition coefficient for *n*-octanol/water bi-phase system), *n*HBD, *n*HBA, MPSA (molecular polar surface area), and *n*RB.

The calculated druggability properties were analyzed using the property distribution comparison methodology (histogram-type of analysis, Table 2, Fig. 4) [46] between the dataset of known 6-FQs used in the development of the NN model (*Assay2*, 145 compounds, Fig. 4A) [21] and the combinatorial (*CombiTot*, 53.871 compounds, Fig. 4B).

The results of the statistical analysis show that only one of the calculated properties for the experimental compounds (Fig. 4A) follows normal Gaussian distribution ($n\mathrm{HBD}_{exp}$), whereas the rest of the calculated properties (MW_{exp}, MPSA_{exp}, AlogP_{exp}, $n\mathrm{HBA}_{exp}$, and $n\mathrm{RB}_{exp}$) are asymmetrically distributed. The peak-analysis (the top-point on the Gaussian curve where the distribution of the calculated property reaches 50%) showed acceptable values for drug-likeness: MW_{exp}=468, AlogP_{exp}=−0.07, $n\mathrm{HBD}_{exp}$=2, $n\mathrm{HBA}_{exp}$=8, MPSA_{exp}=103, and $n\mathrm{RB}_{exp}$=5. The corresponding mid-50% values (the interval between 25% and 75% of the distribution, i.e., first (Q1) and third (Q3) calculated quartile, respectively) were obtained employing the quartile calculation: $\mathrm{MW}_{exp\text{-}mid50\%}$=382–566, $\mathrm{AlogP}_{exp\text{-}mid50\%}$=(−0.848)-(+0.394), $n\mathrm{HBD}_{exp\text{-}mid50\%}$=2, $n\mathrm{HBA}_{exp\text{-}mid50\%}$=6–9, $\mathrm{MPSA}_{exp\text{-}mid50\%}$=74.38-124.77, and $n\mathrm{RB}_{exp\text{-}mid50\%}$=3–6 (Table 2A). These results are in accordance with the Lipinski and Veber rule-sets, respectively, and clearly define the domain in which the compounds possess drug-like characteristics [31, 32].

On the other hand, an obvious difference in the property distribution was observed for the compounds within the virtual combinatorial set (Fig. 4B). A normal Gaussian distribution was observed for five calculated properties (MW_{combi}, AlogP_{combi}, $n\mathrm{HBA}_{combi}$, MPSA_{combi}, and $n\mathrm{RB}_{combi}$), while the $n\mathrm{HBD}_{combi}$ parameter follows an asymmetric distribution. The peak-analysis for the combinatorial set resulted in following values: MW_{combi}=621, AlogP_{combi}=4.87, $n\mathrm{HBD}_{combi}$=2, $n\mathrm{HBA}_{combi}$=9, MPSA_{combi}=157, and $n\mathrm{RB}_{combi}$=7, whereas the corresponding mid-50% values were in the following boundaries: $\mathrm{MW}_{combi\text{-}mid50\%}$=578–667, $\mathrm{AlogP}_{combi\text{-}mid50\%}$=3.976-5.911, $n\mathrm{HBD}_{combi\text{-}mid50\%}$=1–2, $n\mathrm{HBA}_{combi\text{-}mid50\%}$=8–10, $\mathrm{MPSA}_{combi\text{-}mid50\%}$=137.36-174.89, and $n\mathrm{RB}_{combi\text{-}mid50\%}$=6–8 (Table 2B).

The increased values (MW_{combi} and MPSA_{combi}) pointed toward the increased molecular complexity [47] and a decreased probability for a good ligand-protein interaction [48]. Thus, a pre-filtering of the combinatorial library was performed by implementing a comprehensive Boolean-type (T/F (true/false)) drug-likeness filtering algorithm (CombiVL; MW<500, AlogP<5.0, nHBD≤5, nHBA≤10, MPSA≤140, nRB≤10) integrating both rule-sets (Lipinski "rule-of-five" and Veber rules) [31, 32].

Such a filtering procedure resulted in eliminating all 6-FQ analogs which do not satisfy the above criteria (in the *CombiTot* library 52.770 eliminated compounds were marked as false (F)). The retained 6-FQ structural analogs (total 1.101 compounds) define the drug-like chemical sub-space (*CombiDL*, online resource 2) which was subsequently used for prediction of the biological activity values ($p\mathrm{MIC}_{pred\text{-}combi}$) employing our derived NN model [21].

Prediction of the biological activity values for the novel combinatorially-generated 6-fluoroquinolone drug-like analogs and activity-based subset selection

A derived seven parameter neural-networks (NN) model (*Assay2*, 145 compounds, online resource 3) [21] was used for prediction of the biological activity values ($p\mathrm{MIC}_{pred\text{-}combi}$)

Table 2 Property distribution analysis of the experimental and the combinatorial set, respectively (No., number of compounds in the dataset; Prop., calculated property; Mean, the mean value; Min, minimum value; Q1, first quartile; Median, the median value; Q3, third quartile; Max, the maximum value). (A) the experimental set (145 compounds), (B) the combinatorial set (*CombiTot*, 53.871 compounds)

A)	No.	Prop.	Mean	Min	Q1	Median	Q3	Max
Experimental set	145	MW_{exp}	467.968	275.280	382.380	419.420	566.490	801.240
		AlogP_{exp}	−0.069	−2.453	−0.848	−0.337	0.394	6.251
		$n\mathrm{HBD}_{exp}$	1.952	1	2	2	2	3
		$n\mathrm{HBA}_{exp}$	7.621	4	6	7	9	14
		MPSA_{exp}	102.983	57.610	74.380	86.150	124.770	186.310
		$n\mathrm{RB}_{exp}$	4.959	1	3	5	6	11
B)	No.	Prop.	Mean	Min	Q1	Median	Q3	Max
Combinatorial set	53.871	MW_{combi}	621.099	329.280	578.410	623.540	666.620	814.440
		AlogP_{combi}	4.872	−3.735	3.976	5.002	5.911	9.265
		$n\mathrm{HBD}_{combi}$	1.958	1	1	2	2	6
		$n\mathrm{HBA}_{combi}$	8.868	4	8	9	10	13
		MPSA_{combi}	156.742	65.780	137.360	155.030	174.890	272.820
		$n\mathrm{RB}_{combi}$	6.807	2	6	7	8	12

for the previously filtered combinatorially-generated drug-like 6-FQ structural analogs (*CombiDL*, 1.101 compounds, online resource 2). Using this developed model the biological activity ($p\mathrm{MIC}_{pred\text{-}combi}$) in the series of novel unknown 6-FQ analogs is correlated to a set of seven constitutional, topological, and electrostatic parameters (Table 3).

As presented in Table 3, nR09 belongs to the class of pure constitutional parameters, GATS8v, YZS/YZR, X1A, and PW3, are molecular descriptors which belong to the class of topological parameters, while the parameters JGI2 and JGI3, belong to the class of pure electrostatic parameters. These molecular descriptors, which in general accentuate the importance of molecular shape can be linked to the accommodation of the main 6-FQ scaffold within the GyrA subunit binding pocket. Since the QSAR model was built employing a series of structurally-similar 6-FQ analogs with experimentally determined biological activity values, one would expect similar biological response of these compounds within the same protein target. In this respect, we believe that GATS8v parameter alone, as molecular descriptor describing the importance of the atomic van der Waals volumes of our 6-FQs, is of particular significance for the biological activity, mainly through establishing proper steric complementarity between the ligand and the enzyme. On the other hand, the X1A parameter is a typical topological molecular descriptor, i.e., a pharmacophore fingerprint that carries the connectivity information of 6-FQ analogs, probably through establishing π-π stacking interactions between the planar aromatic/heteroaromatic systems in the 6-FQs and the GyrA/DNA, respectively.

The electrostatic descriptors JGI2 and JGI3 suggest that anti-mycobacterial activity is potentially dependent on the charge indices for the oxygen of the carboxyl and carbonyl group within the main core. They also describe the possibility for establishing hydrogen-bonding interactions between these substituents and the amino acid residues within the GyrA binding pocket. Moreover, these parameters suggest the possibility of establishing an electrostatic interaction between the F atom at position 6 and the target protein which may result in an enhanced binding of the 6-FQ analogs to the complex.

Activity-based subset selection for molecular modeling calculations

The predicted biological activity values ($p\mathrm{MIC}_{pred\text{-}combi}$) for the combinatorially-generated drug-like 6-FQ analogs are in the range between $0.0125 < p\mathrm{MIC}_{pred\text{-}combi} < 0.9174$, while the corresponding $\mathrm{MIC}_{pred\text{-}combi}$ values, obtained after $p\mathrm{MIC}_{pred\text{-}combi}$ de-normalization and anti-logarithmization, are in the range between $0.0021 < \mathrm{MIC}_{pred\text{-}combi}$ [μg/mL]<6.3726. Using a GHA range described previously, a subset of 427 6-FQ analogs (*CombiLib*, online resource 4) was extracted from the pool of total 1.101 compounds (*CombiDL*). Such a selection procedure ensures that each 6-FQ analog in the MIC-based isolated subset (*CombiLib*) possess biological activity ($0.4809 < p\mathrm{MIC}_{pred\text{-}combi} < 0.9174$; $0.0021 < \mathrm{MIC}_{pred\text{-}combi}$ [μg/mL]<0.1000) against *M. tuberculosis.* The selected combinatorial subset (*CombiLib*, 427 compounds), was subsequently used as an external library in the three-dimensional pharmacophore analysis and molecular docking study.

6-FQs pharmacophore modeling study

Using LigandScout's the integrated automatic pharmacophore generation algorithm, three pharmacophore models

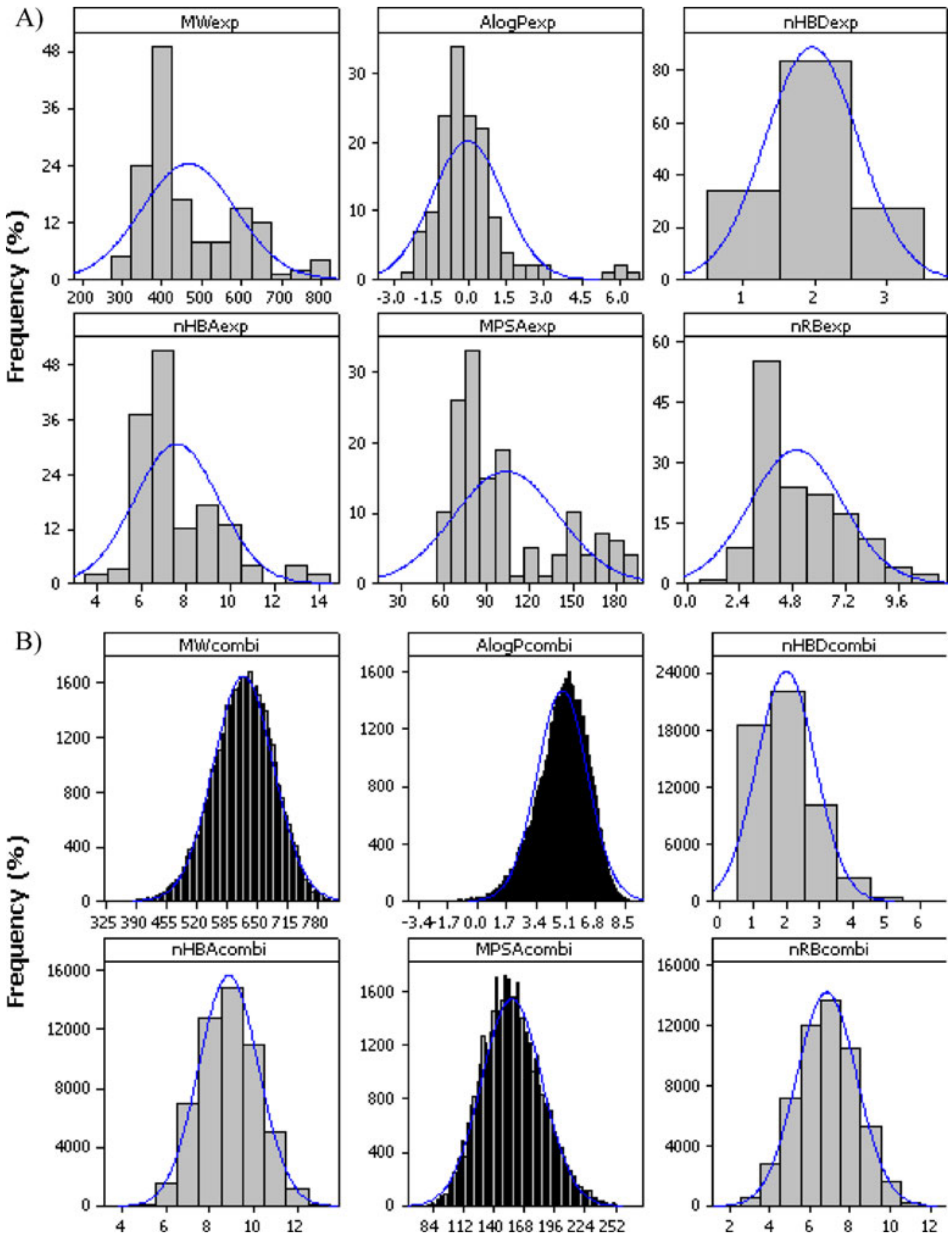


Fig. 4 Histogram-type of analysis for druggability properties assessment (property distribution). (A) the experimental set (*Assay2*, 145 compounds), (B) the combinatorial set (*CombiTot*, 53.871 compounds). The property distribution is fitted with normal distribution to the histogram of data

(LBP, SBP_{shared}, and SPB_{merged}) were constructed as described previously. Since the automatic pharmacophore generation yielded all possible pharmacophoric features, existing knowledge about the SAR of 6-FQs [16] was employed for pharmacophoric simplification of all models. Such a simplification approach resulted in a total of five pharmacophoric features per model (one aromatic ring, two hydrophobic features, one hydrogen bond donor, and one negative ionizable area). The obtained three-dimensional pharmacophore models (LBP, SBP_{shared}, and SPB_{merged}), served as highly effective in silico filtering tools for subsequent VS of the 6-FQ ligands (Fig. 5).

Since the core idea of the three-dimensional pharmacophoric concept is the selection of active compounds among a pool of inactive molecules [18], our generated

3D pharmacophore models (LBP, SBP$_{shared}$, and SPB$_{merged}$) were validated for their recognition performances as in silico filters in a VS experiment using a dataset of known 6-FQs with experimentally-measured biological activity values (*Assay2*, 145 compounds) [21]. Initially, the LBP model identified 58 active out of total 145 compounds, while the two SBP models (SBP$_{shared}$ and SPB$_{merged}$) identified 62 and 49 compounds, as active compounds respectively. The experimentally-determined biological activity values for the successfully filtered compounds are in the range between 0.001<MIC$_{exp}$ [μg/mL]<3.500, of which around 90% were determined as highly-active compounds with biological activity values in the range between 0.001<MIC$_{exp}$ [μg/mL]<0.900. The 6-FQs obtained by the pharmacophore-based VS procedure (LBP (58), SBP$_{shared}$ (62), and SBP$_{merged}$ (49)), were visually inspected for molecular fitness within each of the pre-defined pharmacophoric features of the models (one aromatic ring, two hydrophobic features, one hydrogen bond donor, and one negative ionizable area).

In order to assess the visual inspection more precisely, a Boolean-type of signing (T/F (true/false)) was introduced. Each visually-determined match between the 6-FQ's substructural elements and the pharmacophore-model features was marked as true (T). In addition, the biological activity values (MIC$_{exp}$) of the investigated compounds were implemented as a feature in the evaluation process using the GHA range. Thus, the compounds with biological activity values in the GHA range were marked as true (T), while the rest of the compounds were marked as false (F). Therefore, only the compounds with (T) outcome for all the pre-defined pharmacophoric-model features as well as lying within the GHA range (electronic supplementary material, online resource 5, experimental set), can be marked as highly favorable 6-FQs (LBP (20 highly active 6-FQs out of 58), SBP$_{shared}$ (19 highly active 6-FQs out of 62), and SPB$_{merged}$ (8 highly active 6-FQs out of 49). Online resource 5, shows that selected compounds from the experimental set (*Structure24*, *Structure39*, *Structure75*, and *Structure110*), which are all potent compounds, were identified as active compounds by all three pharmacophore models.

The structural analysis of the selected compounds (online resource 5, see "experimental set" sheet), shows that the cyclopropyl group is the most frequently used functional group at position 1 of the main 6-FQ scaffold, whereas position 7 can be successfully substituted by a range of substructural fragments, mainly heterocyclic systems of which 5- and 6-membered N-hetero systems (aminopyrolidines, piperazines), are the most optimal for anti-mycobacterial activity. These results suggest the SAR rules for optimal anti-mycobacterial activity of the 6-FQs [16] and that the three-dimensional pharmacophore concept can be successfully employed as a highly-effective in silico filtering tool [19].

Following the validation on the experimental set [21], the same 3D pharmacophore models were used for assessing the combinatorially-generated subset (*CombiLib*, 427 novel compounds). The LBP model initially identified 95 active out of a total of 427 compounds as hits, whereas the two structure-based pharmacophore models (SBP$_{shared}$ and SBP$_{merged}$) identified 95 and 77 compounds, respectively. Since the combinatorial subset was built defining a GHA range, all of the 6-FQ analogs in such a library are hypothetically active against *M. tuberculosis*, regarding the MIC$_{pred\text{-}combi}$ values (0.0021<MIC$_{pred\text{-}combi}$ [μg/mL]<0.1000). Therefore, a visual determination of the matches using (T/F) designation between the structural elements of the novel actively-recognized 6-FQs hits (LBP (95), SBP$_{shared}$ (95), and SBP$_{merged}$ (77)) and pharmacophoric features within the models was implemented to identify the most optimal structural features.

The selected compounds (LBP (32 out of 95), SBP$_{shared}$ (26 out of 95), and SBP$_{merged}$ (31 out of 77)), belong to three general classess of combinatorially-generated 6-FQ compounds: ciprofloxacin analogs (7-amino (CIP-N_i-M_j)/7-nonamino derivatives (CIP′-N_i-M_j)), moxifloxacin analogs (7-nonamino derivatives (MOX′-N_i-M_j)) and ofloxacin analogs (7-amino (OFL-M_j)/7-nonamino derivatives (OFL′-M_j)).

The frequency analysis of occurrence of the substructural fragments (throughout the models output,

Table 3 The most important 2D molecular descriptors (pharmacophore fingerprints) for the activity (T, topological; C, constitutional; E, electrostatic)

ID	Descriptor	Source	Definition	Class
1.	GATS8v	DRAGON	Geary autocorrelation-lag8/weighted by atomic van der Waals volumes	T
2.	nR09	DRAGON	Number of 9-membered rings	C
3.	JGI2	DRAGON	Mean topological charge index of order 2	E/T
4.	JGI3	DRAGON	Mean topological charge index of order 3	E/T
5.	YZS/YZR	CODESSA	YZ Shadow/YZ Rectangle	T
6.	X1A	DRAGON	Average connectivity index chi-1	T
7.	PW3	DRAGON	Path/Walk 3-Randic shape index	T

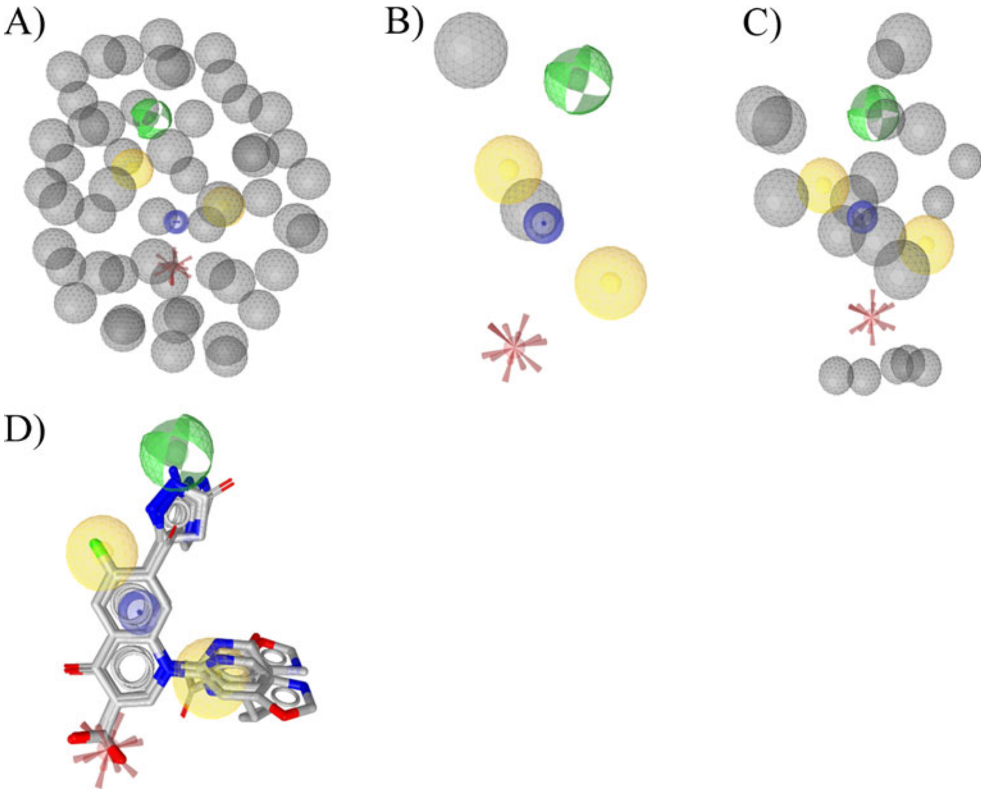


Fig. 5 The three pharmacophore models obtained after pharmacophoric features reduction (a total of five features per model). (A) ligand-based pharmacophore (LBP) model, (B) structure-based shared pharmacophore (SBP_{shared}) model, (C) structure-based merged pharmacophore (SBP_{merged}) model, (D) The most promising six combinatorial compounds obtained after pharmacophore models VS campaign and aligned into the SBP_{shared} pharmacophore model

electronic supplementary material, online resource 5, see "combinatorial set" sheet) attached at positions 1 and 7 of the main 6-FQ core - explicitly shows that the most frequently appearing fragment at position 1 (18 times) is the building-block benzo[*d*]oxazole marked as N_i=028 in the CIP′-analogs and N_i=016 in the MOX′-analogs, respectively, while the most frequently appearing fragments at position 7 of the main scaffold are the building-blocks: 1*H*-pyrazolo[3,4-*b*]pyridine (16 times) marked as M_j=148 in the CIP′-analogs, M_j=116 in the MOX′-analogs, and M_j=147 in the OFL′-analogs, respectively, and 1*H*-pyrazol-5(4*H*)-one (15 times) marked as M_j=186 in the CIP′-analogs and M_j=144 in the MOX′-analogs. In summary, six compounds were selected (CIP′-028-148, CIP′-028-186, MOX′-016-116, MOX′-016-144, OFL′-147, and OFL-148) as a result of pharmacophore modeling assessment (Table 4).

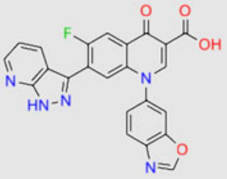
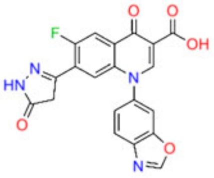
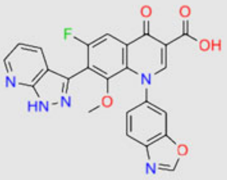
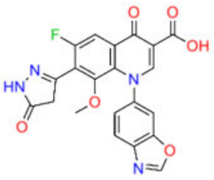
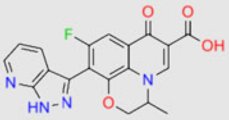
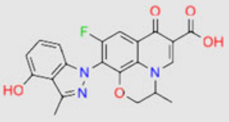
Interestingly, these compounds (except for OFL-148, i.e., an ofloxacin analog) belong to the 6-FQ analogs structurally different from ciprofloxacin and moxifloxacin with non-amino substructural fragments at position 7. The predicted biological activity values for these 6-FQ analogs are in the range $0.0035 < MIC_{pred-combi}$ [μg/mL]<0.1000. The selected combinatorially-generated 6-FQ analogs, were subsequently used for assessing the possible interactions with experimentally-bound 6-FQ protein structure, as an integrated part of the external combinatorial subset (*CombiLib*, 427 novel compounds) in the molecular docking study.

Molecular docking study

The molecular docking study on both, the experimental set (*Assay2*, 145 compounds) with experimentally-determined biological activity values and the combinatorial one (*CombiLib*, 427 compounds), was performed by using the experimentally-determined 6-FQs binding site of the type II topoisomerase protein (3K9F). In a previously published docking study, only GyrA protein subunit without full complex details (DNA and GyrB position), was used to assess the possible binding modes [49]. The co-crystallized 6-FQ inhibitor levofloxacin [38] was used for comparison as well as to analyze the geometric and structural properties of the docked 6-FQ compounds in the protein. The re-docking validation procedure [40] of levofloxacin into its binding site was successful as GOLD was able to reproduce the experimental bound conformation with high accuracy.

The post-docking VS analysis was divided into two levels. At the first level, the geometric properties of both sets of compounds were assessed by visually inspecting each 6-FQ dock position relative to the experimental conformation of levofloxacin [38]. The following geometric properties were compared: the visual

Table 4 The combinatorial compounds selected according to the frequency analysis of the sub-structural fragments (throughout the pharmacophore models output) attached at position R_1 and R_7 of the main 6-FQ scaffold, respectively, together with their corresponding pharmacophore fitness score (**PFS**). The selected compounds with highest predicted biological activity values (**MIC*pred-combi***), are highlighted in gray

ID	Chemical structure	R_1	R_7	MIC*pred-combi*	*p*MIC*pred-combi*	PFS
1.	 CIP'-028-148	028	148	0.0499	1.3022	54.39
2.	 CIP'-028-186	028	186	0.1000	1.0002	57.20
3.	 MOX'-016-116	016	116	0.0499	1.3022	57.21
4.	 MOX'-016-144	016	144	0.1000	1.0002	57.28
5.	 OFL'-147	N/A	147	0.0499	1.3022	57.15
6.	 OFL-148	N/A	148	0.0035	2.4580	57.06

orientation (how the screened 6-FQ compound is oriented relative to the levofloxacin position), visual fitness (how well the screened 6-FQ compound fits the experimental levofloxacin conformation), and the number of the matching pharmacophoric features (how many common pharmacophoric features, generated for each docked conformation, are shared by both screened 6-FQ compound and levofloxacin). Similarly to the pharmacophore analysis, a Boolean-type of signing (T/F (true/false)) was employed for assessing the visual orientation and visual fitness parameters, while the matching pharmacophoric features were defined numerically. The screened compounds which share 4–5 pharmacophoric features present in bound levofloxacin, were signed as true (T). Therefore, the result of the first level assessment can be defined as a sum of the obtained Boolean answers for all three investigated properties (electronic supplementary materials, online resource 6).

First level post-docking analysis of the experimental set of 6-FQ analogs docked into the 3K9F binding pocket (online resource 6, see “Exp-3K9F” sheet) resulted in 102 (T) marked compounds out of 145, of which 38 compounds have measured biological activity values within $0.001<\mathrm{MIC}_{exp}$ [μg/mL]<0.1. Moreover, as shown in Fig. 6A, a favorable positioning of these 6-FQ compounds into the 3K9F binding pocket could be observed. The structural analysis of these 38 compounds, once again showed that the cyclopropyl group is the most frequently appearing substituent at position 1 of the main 6-FQ scaffold (33 times), while the piperazinyl group attached at position 7 appeared 24 times. These results are again in accordance with the previous experimental SAR findings (online resource 6, “experimental set” sheet) [16].

The first level post-docking analysis of the 6-FQ analogs in the combinatorial set (*CombiLib*, 427 compounds) docked into the 3K9F binding pocket (electronic supplementary materials, online resource 6, see “Combi-3K9F” sheet) resulted in 166 (T) marked compounds out of 427.

Interestingly, comparing with the previous assessment using pharmacophore models filters where none of the MOX-N_i-M_j analogs was recognized as active for further investigation, the post-docking Boolean procedure performed here identified only one moxifloxacin analog (MOX-028-064) as geometrically suitable 6-FQ compound in 3K9F binding pocket. The 6-FQ combinatorial hit analogs belonging to the other five structural classes (CIP-N_i-M_j, CIP′-N_i-M_j, MOX′-N_i-M_j, OFL-M_j, OFL′-M_j) were successfully identified. According to the combinatorial 6-FQ (T) outcome obtained, 166 compounds from the combinatorial set have geometric properties, comparable to the co-crystallized levofloxacin (Fig. 6B).

The frequency analysis of occurrence of the substructural fragments attached at positions 1 and 7 of these 166 compounds, indicated that the most frequently appearing fragments at position 1 are the building-blocks: benzo[*d*]oxazole (46 times) marked as N_i=028 in the CIP′-analogs and in the CIP-analogs, and N_i=016 in the MOX′-analogs, respectively, and 1-(pyridin-3-yl)ethanone (11 times) marked as N_i=102. The most frequently appearing frag-

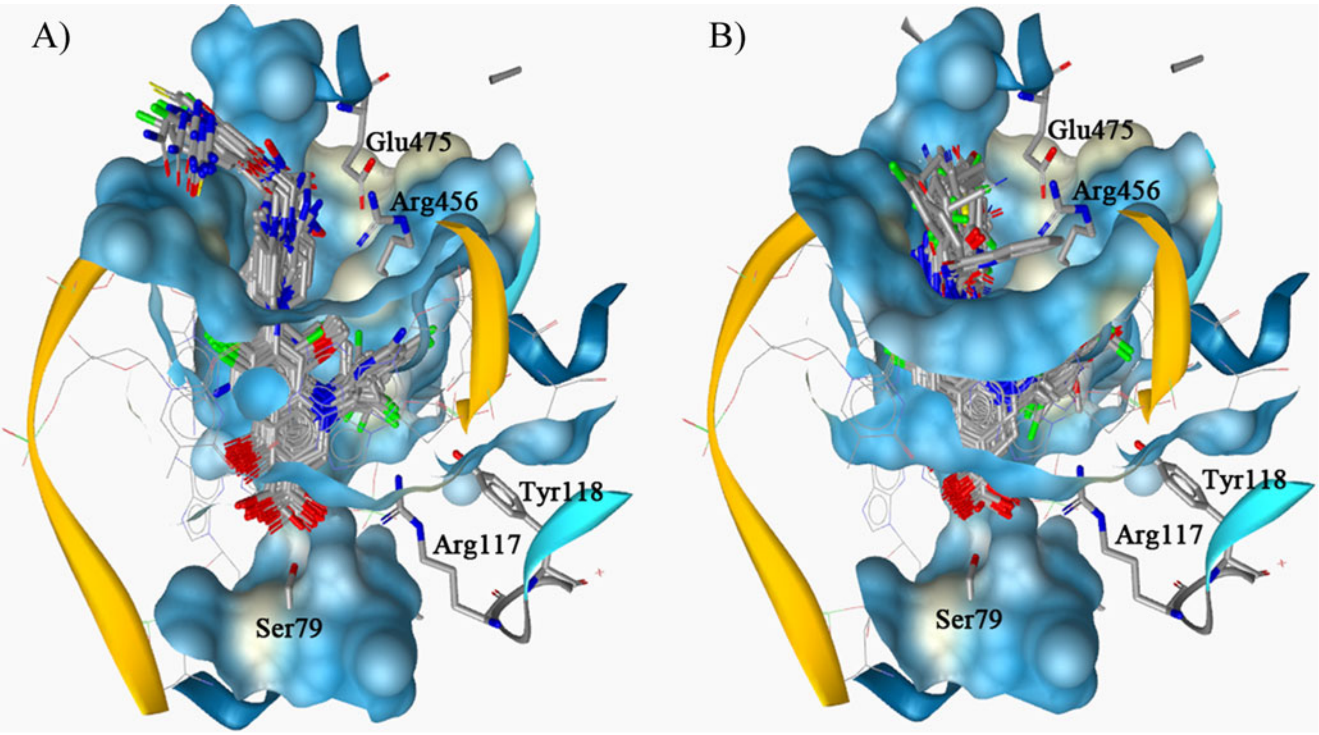


Fig. 6 Results of docking into the 3K9F protein binding pocket. (A) 38 compounds from the experimental set with acceptable geometric properties. (B) 166 compounds from the combinatorial set with acceptable geometric properties. Crucial amino acid residues from the 3K9F 6-FQs binding site are displayed [35]

ments at position 7 of the main scaffold are the building-blocks: 4*H*-furo[3,2-*c*]pyran-4-one (8 times) marked as M_j=006, 1*H*-pyrazolo[3,4-b]pyridine (7 times) marked as M_j=148, and 1*H*-pyrazole (7 times) marked as M_j=176.

Except compound OFL-148, all selected 6-FQs from the performed pharmacophore analysis (CIP′-028-148, CIP′-028-186, MOX′-016-116, MOX′-016-144, OFL′-147) were successfully identified as compounds with favorable three-dimensional binding geometry as well (online resource 6, see "Combi-3K9F" sheet). Therefore, all of the selected 6-FQ analogs from the experimental set (3K9F (102)) and the combinatorial one (3K9F (166)), respectively, were subsequently used for further investigation at the second level post-docking analysis.

At the second level of the post-docking analysis, the intermolecular interaction properties of the selected geometrically suitable 6-FQs from both experimental and combinatorial sets were investigated (electronic supplementary materials, online resource 8). The crystal structure of the *S. pneumoniae* topoisomerase IV protein in complex with the co-crystallized ligand levofloxacin (3K9F) was used as a starting point for measuring the interaction distances between the ligand's functional groups and the surrounding amino acid residues [38]. A set of five key interactions between the levofloxacin and the protein were used as standard measures: (R_3-COO$^-$)–(HO-Ser79) [$d_{3K9F\text{-}ref\text{-}1}$=3.18 Å], ($R_3$-COO$^-$)–(HO-Ser79) [$d_{3K9F\text{-}ref\text{-}2}$=3.35 Å], ($R_3$-COO$^-$)–(HO-Arg117) [$d_{3K9F\text{-}ref\text{-}3}$=3.01 Å], ($R_7$=N-CH$_3$)–(H$_2N^+$=Arg456) [$d_{3K9F\text{-}ref\text{-}4}$=3.30 Å], and ($R_7$=N-CH$_3$)–($^-$OOC-Glu475) [$d_{3K9F\text{-}ref\text{-}5}$=4.38 Å]. Since the biological systems are not static and are characterized with dynamic features, a distance tolerance of±1 Ångstrom unit around each of the standard interatomic distance values was employed: (R_3-COO$^-$)–(HO-Ser79) [$d_{3K9F\text{-}1}$=2.18-4.18 Å], (R_3-COO$^-$)–(HO-Ser79) [$d_{3K9F\text{-}2}$=2.35-4.35 Å], (R_3-COO$^-$)–(HO-Arg117) [$d_{3K9F\text{-}3}$=2.01-4.01 Å], (R_7=N-CH$_3$)–(H$_2$N$^+$=Arg456) [$d_{3K9F\text{-}4}$=2.30-4.30 Å], and (R_7=N-CH$_3$)–($^-$OOC-Glu475) [$d_{3K9F\text{-}5}$=3.38-5.38 Å].

The second level post-docking analysis in 3K9F (online resource 7, see "Exp-3K9F" sheet) selected 45 out of 102 compounds from the experimental set as optimal 6-FQ compounds (regarding the interatomic distances). The biological activity values for these compounds are in the range between 0.001<MIC$_{exp}$ [µg/mL]<6.700 of which approximately 38% have good measured biological activity (0.001<MIC$_{exp}$ [µg/mL]<0.100). The structural analysis once again showed that the most frequently occuring substituent at position 1 of the main 6-FQ scaffold is the cyclopropyl group (15 times), whereas the most frequent substituent attached at position 7 of the hit molecules is the piperazinyl group (12 times), whose results are in accordance with the established 6-FQs SAR rules (Fig. 1) [16].

The second level post-docking analysis of the combinatorial compounds resulted in selection of a total of 11 out of 166 compounds (online resource 7, see "Combi-3K9F" sheet). These compounds mainly belong to the class of 7-nonamino substituted ciprofloxacin analogs CIP′-N_i-M_j (nine compounds), one 7-amino substituted ciprofloxacin analog (CIP-N_i-M_j), and one 7-amino substituted ofloxacin analog (OFL-M_j). No moxifloxacin analogs (MOX-N_i-M_j and MOX′-N_i-M_j) and 7-nonamino substituted ofloxacin analogs (OFL′-M_j) were found that could be fitted in this set of defined interatomic distance boundaries [38].

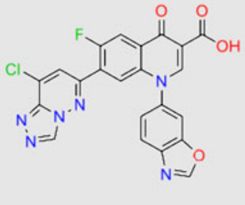
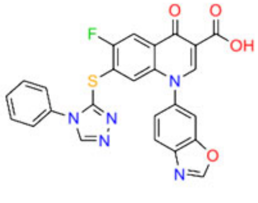
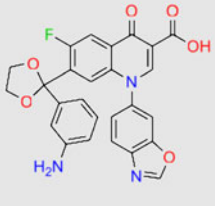
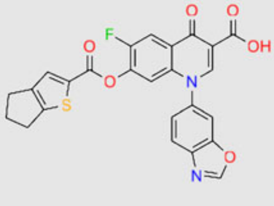
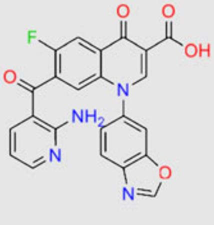
The frequency analysis of occurrence of the substituents attached at position 1 and 7 of the main scaffold, once again pinpointed the fragment marked as N_i=028 as the most frequently appearing substituent at position 1 (5 times), while position 7 can be successfully substituted with a range of different groups. Incorporation of these fragments thus forms novel 6-FQ compounds that could also serve as novel target compounds in the synthetically-driven lead optimization. The list of compounds that complements the previous selection in Table 4 is presented in Table 5.

Conclusions

In the present study a variety of chemometric and molecular modeling approaches were integrated into a powerful complex scheme capable of constructing as well as evaluating a virtual combinatorial library of 6-fluoroquinolone analogs (6-FQs). This 6-FQs library was designed by employing the synthetically-driven ligand generation rules feasible of yielding ligands which can be readily synthesized and have a predicted inhibitory activity toward GyrA.

The results can be summarized as follows: (1) a large number of virtual 6-fluoroquinolone analogs (53.871 compounds) were generated by a combinatorial generation of all substituted amine and non-amine compounds at positions 1 and 7, respectively. The selection of a drug-like set of 427 compounds from the library within the GHA range was performed by using drug-likeness filters (a combined Lipinski-Veber filtering tool based on the Lipinski's and Veber's rule-sets for drug-likeness) and our previously developed and validated neural-networks (NN) chemometric model (built on a dataset of structurally-similar 6-FQ compounds with experimentally-determined biological activity values by employing a combined QSAR modeling strategy, i.e., linear modeling as well as non-linear modeling using Kohonen and counter-propagation artificial neural networks) [21]. (2) Experimental data on the structurally-similar topoisomerase-IV enzyme in complex with levofloxacin was used to construct structure-based as well as ligand-based pharmacophore models to evaluate the most promising 6-FQs obtained by chemometric methods. The

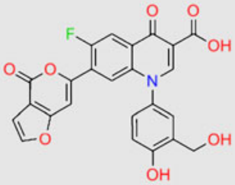
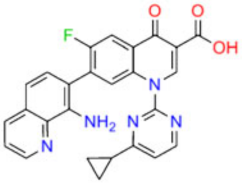
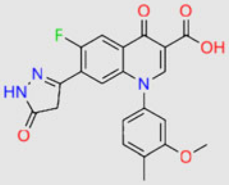
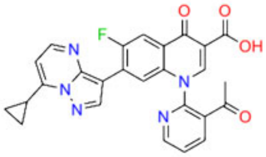
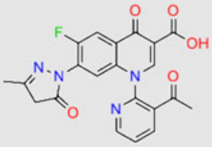
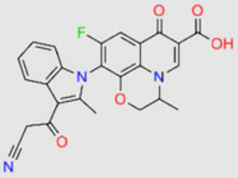
Table 5 The combinatorial compounds extracted as most suitable after the second level VS analysis into the 3K9F binding pocket, together with their GoldScore Fitness (**GSF**) scoring function and the corresponding inter-atomic distances (in Ångstrom units) between the 6-FQ sub-structural fragments and the crucial amino acid residues important for the biological activity. The compounds with highest predicted biological activity values (**MIC***pred-combi*), are highlighted in gray

ID	Chemical structure	$MIC_{pred\text{-}combi}$	$pMIC_{pred\text{-}combi}$	GSF	$d_{3K9F\text{-}1}$ [2.18-4.18]	$d_{3K9F\text{-}2}$ [2.35-4.35]	$d_{3K9F\text{-}3}$ [2.01-4.01]	$d_{3K9F\text{-}4}$ [2.30-4.30]	$d_{3K9F\text{-}5}$ [3.38-5.38]
1.	 CIP'-028-059	0.0499	1.3022	78.29	3.31	3.92	3.67	3.99	5.07
2.	 CIP'-028-065	0.1000	1.0002	87.09	2.88	3.27	3.29	4.25	4.81
3.	 CIP'-028-125	0.0021	2.6818	76.16	3.49	3.86	3.24	3.40	5.15
4.	 CIP'-028-154	0.0021	2.6818	80.90	3.16	3.51	2.80	4.19	5.37
5.	 CIP'-028-156	0.0783	1.1065	72.18	3.45	4.31	2.62	2.64	5.00

models were also validated using available experimental 6-FQs data from the literature. (3) Finally, these compounds were docked into the topo-IV binding pocket and the interactions of 6-FQ analogs from virtual library as well as experimental database with the surrounding amino acid residues were compared and analyzed.

The outcome of this study shows that promising compounds for further investigation originate from all three

Table 5 (continued)

6.	 CIP’-037-006	0.0130	1.8851	72.26	3.17	3.51	3.31	3.92	3.43
7.	 CIP’-087-073	0.1000	1.0002	71.96	2.96	3.84	3.92	4.08	3.96
8.	 CIP’-095-186	0.0401	1.3970	71.79	3.09	3.51	3.60	3.58	5.33
9.	 CIP’-102-113	0.1000	1.0002	83.74	3.94	4.33	2.89	4.12	4.91
10.	 CIP-103-102	0.0681	1.1666	74.36	3.71	4.17	2.83	4.29	5.29
11.	 OFL-127	0.0130	1.8851	72.53	3.34	3.65	3.00	3.01	3.73

chemical classes (ciprofloxacin, moxifloxacin, and ofloxacin analogs), whereas the ciprofloxacin chemical class yielded the highest number of hits (CIP′-028-059, CIP′-028-125, CIP′-028-148, CIP′-028-154, CIP′-028-156, CIP′-

037-006, CIP′-095-186, and CIP-103-102). Furthermore, the detailed analysis of occurrence of the substructural fragments present at positions R_1 and R_7 of the 6-FQ hit molecules revealed several novel attractive fragments, such as for R_1: [(N_i=028, benzo[*d*]oxazole), (N_i=037, 2-(hydroxymethyl)phenol), (N_i=095, 1-methoxy-2-methylbenzene), and (N_i=103, 1-(pyridin-3-yl)ethanone)] and for R_7: [(M_j= 006, 4*H*-furo[3,2-*c*]pyran-4-one), (M_j=059, 8-chloro-[1, 2, 4]triazolo[4,3-*b*]pyridazine), (M_j=102, 3-methyl-1*H*-pyrazol-5(4*H*)-one), (M_j=125, 3-(2-methyl-1,3-dioxolan-2-yl)aniline), (M_j =148, 1*H*-pyrazolo[3,4-*b*]pyridine), (M_j=154, methyl 5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-2-carboxylate), (M_j=156, 1-(2-aminopyridin-3-yl)ethanone), and (M_j=186, 1*H*-pyrazol-5(4*H*)-one], respectively, that satisfied the screening conditions at all levels. In conclusion, we hope that the results of our mixed chemometric-molecular modeling study will assist in providing new SAR guidelines to the lead optimization stage of the 6-FQ's drug design and thus enable the design of the novel - much-needed - antibacterial agents.

Acknowledgments Authors thank Agency of Research of R. Slovenia (ARRS) for the financial support through the Grants P1-0017 and 1000-07-310016. We are sincerely grateful to Dr. Marjana Novič for valuable insights, discussion and her continuing support of this research.

References

1. Munro SA, Lewin SA, Smith HJ, Engel ME, Feetheim A, Volmink J (2007) Patient adherence to tuberculosis treatment: a systematic review of qualitative research. PloS Med 4:1230–1245
2. Du Toit LC, Pillay V, Danckwerts MP (2006) Tuberculosis chemotherapy: current drug delivery spproaches. Respir Res 7:118–136
3. Bhanu NV, van Soolingen D, van Embden JDA, Seth P (2004) Two *Mycobacterium fortuitum* strains isolated from pulmonary tuberculosis patients in Delhi IS6110 harbour homologue. Diagn Micro Infec Dis 48:107–110
4. Dussurget O, Rodriguez M, Smith I (1998) Protective role of *Mycobacterium smegmatis* IdeR against reactive oxygen species and isoniazid toxicity. Tubercle Lung Dis 79:99–106
5. Field SK, Fisher D, Cowie RL (2004) *Mycobacterium avium* complex pulmonary disease in patients without HIV infection. Chest 126:566–581
6. Zhang Y, Martens KP, Denkin S (2006) New drug candidates and therapeutic targets for tuberculosis therapy. Drug Discovery Today 11:21–27
7. Wigley DB (1995) Structure and mechanism of DNA gyrase. In: Eckstein F, Lilley DMJ (eds) Nucleic Acids Molecular Biol. Springer, Berlin, pp 165–176
8. Barnard FM, Maxwell A (2001) Interaction between DNA gyrase and quinolones: effects of alanine mutations at GyrA subunit residues Ser^{83} and Asp^{87}. Antimicrob Agents Chemother 45:1994–2000
9. Peng H, Marians KJ (1993) Escherichia coli topoisomerase IV. Purification, characterization, subunit structure, and subunit interactions. J Biol Chem 268:24481–24490
10. Reece RJ, Maxwell A (1991) DNA gyrase: structure and function. Crit Rev Biochem Mol 26:335–375
11. Levine C, Hiasa H, Marians KJ (1998) DNA gyrase and topoisomerase IV: biochemical activities, physiological roles during the chromosome replication, and drug sensitivities. Biochim Biophys Acta 1400:29–43
12. Ostrov DA, Hernández Prada JA, Corsino PE, Finton KA, Le N, Rowe TC (2007) Discovery of novel DNA gyrase inhibitors by high-throughput virtual screening. Antimicrob Agents Chemother 51:3688–3698
13. Zhang Y (2005) The magic bullets and tuberculosis drug targets. Annu Rev Pharmacol Toxicol 45:529–564
14. Vashist J, Vishvanath KR, Kapil A, Yennamalli Y, Subbarao N, Rajeswari MR (2009) Interaction of nalidixic acid and ciprofloxacin with wild type and mutated quinolone-resistance-determining region of DNA gyrase A. Indian J Biochem Biophys 46:147–153
15. Hooper DC (1999) Mode of action of fluoroquinolones. Drugs 58 (suppl 2):6–10
16. Peterson LR (2001) Quinolone molecular structure-activity relationships: what we have learned about improving antimicrobial activity. Clin Infect Dis 33(Suppl 3):S180–S186
17. Ebalunode JO, Zheng W, Tropsha A (2011) Application of QSAR and shape pharmacophore modeling approaches for targeted chemical library design. Methods Mol Biol 685:111–133
18. Wolber G, Langer T (2005) LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J Chem Inf Comput Sci 45:160–169
19. Sippl W (2008) Pharmacophore identification and pseudo-receptor modeling. In: Wermuth CG (ed) The Practice of Medicinal Chemistry, 3rd edn. Academic Press/Elsevier, Amsterdam, pp 572–586
20. Perdih A, Kovač A, Wolber G, Blanot D, Gobec S, Solmajer T (2009) Discovery of novel benzene 1,3-dicarboxylic acid inhibitors of bacterial MurD and MurE ligases by structure-based virtual screening approach. Bioorg Med Chem Lett 19:2668–2673
21. Minovski N, Vračko M, Šolmajer T (2010) Quantitative structure-activity relationship study of antitubercular fluoroquinolones. Mol Div 15:417–426
22. http://www.cambridgesoft.com/software/ChemBioOffice
23. Schwalbe T, Kadzimirisz D, Jas G (2000) Synthesis of a library of ciprofloxacin analogues by means of sequential organic synthesis in microreactors. QSAR Comb Sci 24:758–768
24. Martel AM, Leeson PA, Castañer J (1997) BAY-12–8039: fluoroquinolone antibacterial. Drugs Fut 22:109–113
25. Serradell MN, Blancafort P, Castañer J (1983) DL-8280. Drugs Fut 8:395
26. Key Organics, Bionet Fragment Library "Rule of 3", http://www.keyorganics.ltd.uk
27. Congreve M, Carr R, Murray CW, Jhoti H (2003) A rule of three for fragment-based lead discovery? Drug Discovery Today 8:876–877
28. Minovski N, Šolmajer T (2010) Chemometrical exploration of combinatorially generated drug-like space of 6-fluoroquinolone analogs: a QSAR study. Acta Chim Slov 57:529–591
29. Wieland T (1997) Combinatorics of combinatorial chemistry. J Math Chem 21:141–157
30. Bohacek RS, mcMartin C, Guida WC (1996) The art and practice of structure-based drug design. Med Res Rev 16:3–50
31. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 23:3–25
32. Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD (2002) Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 45:2615–2623
33. Katritzky AR, Lobanov VS, Karelson M (1995) CODESSA. Reference manual, University of Florida, Gainsville

34. Tretter EM, Schoeffler AJ, Weisfield SR, Berger JM (2010) Protein Struct Funct Bioinf 78:492–495
35. Fu G, Wu J, Liu W, Zhu D, Hu Y, Deng J, En Zhang X, Bi L, Cheng Wang D (2009) Crystal structure of DNA gyrase B′ domain sheds lights on the mechanism for T-segment navigation. Nucleic Acids Res 37:5908–5916
36. Laponogov I, Sohi MK, Veselkov DA, Pan XS, Sawhney R, Thompson AW, McAuley KE, Fisher LM, Sanderson MR (2009) Structural insight into the quinolone-DNA cleavage complex of type IIA topoisomerases. Nat Struct Mol Biol 16:667–669
37. Kitamura A, Hoshino K, Kimura Y, Hayakawa I, Sato K (1995) Contribution of the C-8 substituent of DU-6859a, a new potent fluoroquinolone, to its activity against DNA gyrase mutants of Pseudomonas aeruginosa. Antimicrob Agents Chemoter 39:1467–1471
38. Laponogov I, Pan XS, Veselkov DA, McAuley KE, Fisher LM, Sanderson MR (2010) Structural basis of gate-DNA breakage and resealing by type II topoisomerases. PloS One 5:e11338(1–8)
39. Jones G, Willet P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748
40. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T (2008) Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection-What can we learn from earlier mistakes? J Comput Aided Mol Des 22:213–228
41. Huc I, Lehn J-M (1997) Virtual combinatorial libraries: Dynamic generation of molecular and supramolecular diversity by self-assembly. Proc Natl Acad Sci 94:2106–2110
42. Oprea TI, Gottfries J, Sherbukhin V, Svensson P, Kühler TC (2000) Chemical information management in drug discovery: Optimizing the computational and combinatorial chemistry interfaces. J Mol Graph Model 18:512–524
43. Langer T, Wolber G (2004) Virtual combinatorial chemistry and in silico screening: Efficient tools for lead structure discovery? Pur Appl Chem 76:991–996
44. Seneci P, Miertus S (2000) Combinatorial chemistry and high-throughput screening in drug discovery: different strategies and formats. Mol Diversity 5:75–89
45. Oprea TI (2002) Chemical space navigation in lead discovery. Curr Opin Chem Biol 6:384–389
46. Oprea TI (2000) Property distribution of drug-related chemical databases. J Comput Aided Mol Des 14:251–264
47. Allu TK, Oprea TI (2005) Rapid evaluation of synthetic and molecular complexity for in silico chemistry. J Chem Inf Model 45:1237–1243
48. Hann MM, Leach AR, Harper G (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. J Chem Inf Comput Sci 41:856–864
49. Madurga S, Sánchez-Céspedes J, Belda I, Vila J, Giralt E (2008) Mechanism of binding of fluoroquinolones to the quinolone resistance determining region of DNA gyrase: towards an understanding of the molecular basis of quinolone resistance. Chem Bio Chem 9:2081–2086

J Mol Model (2012) 18:1755–1766
DOI 10.1007/s00894-011-1206-1

ORIGINAL PAPER

Stacking interaction and its role in kynurenic acid binding to glutamate ionotropic receptors

Alexander V. Zhuravlev · Gennady A. Zakharov · Boris F. Shchegolev · Elena V. Savvateeva-Popova

Received: 22 November 2010 / Accepted: 26 July 2011 / Published online: 12 August 2011

Abstract Stacking interaction is known to play an important role in protein folding, enzyme-substrate and ligand-receptor complex formation. It has been shown to make a contribution into the aromatic antagonists binding with glutamate ionotropic receptors (iGluRs), in particular, the complex of NMDA receptor NR1 subunit with the kynurenic acid (KYNA) derivatives. The specificity of KYNA binding to the glutamate receptors subtypes might partially result from the differences in stacking interaction. We have calculated the optimal geometry and binding energy of KYNA dimers with the four types of aromatic amino acid residues in *Rattus* and *Drosophila* ionotropic iGluR subunits. All ab initio quantum chemical calculations were performed taking into account electron correlations at MP2 and MP4 perturbation theory levels. We have also investigated the potential energy surfaces (PES) of stacking and hydrogen bonds (HBs) within the receptor binding site and calculated the free energy of the ligand-receptor complex formation. The energy of stacking interaction depends both on the size of aromatic moieties and the electrostatic effects. The distribution of charges was shown to determine the geometry of polar aromatic ring dimers. Presumably, stacking interaction is important at the first stage of ligand binding when HBs are weak. The freedom of ligand movements and rotation within receptor site provides the precise tuning of the HBs pattern, while the incorrect stacking binding prohibits the ligand-receptor complex formation.

A. V. Zhuravlev (✉) · G. A. Zakharov · B. F. Shchegolev · E. V. Savvateeva-Popova
Pavlov Institute of Physiology,
Russian Acad. Sci, 6 Makarova nab,
199034 St.Petersburg, Russia
e-mail: beneor@mail.ru

Keywords Ionotropic glutamate receptors · Kynurenic acid · Möller-Plesset perturbation theory · Stacking interaction

Introduction

Noncovalent interactions are known to play a special role in formation of biomacromolecular structures [1]. Hydrogen bonds (HB), van der Waals (vdW), hydrophobic and electrostatic interactions can form and dissociate at physiological conditions: the magnitude of these interactions is at least 1–2 orders weaker than that of covalent ones. The spatial direction of HBs and steric correspondence of contacting chemical groups provide the basis for geometric specificity of protein and nucleic acid folding, as well as of ligand-receptor, enzyme-substrate and protein complexes formation. Thus, a proper understanding of biophysical conditions for the formation of noncovalent interactions is necessary for theoretical predictions of macromolecular structure and computer design of new drugs with specific pharmacological activity.

π-π Interactions are formed in proteins between the aromatic moieties of Phe, Tyr, Trp and His amino acid residues and also between the nucleic acids aromatic groups. They belong to the weak type of interactions with the binding energy (E_{BIND}) of ~5–50 kJ mol^{-1}. Together with other interactions they determine the specificity of folding [2] and protein-nucleic acid complex formation [3]. Stacking is a special type of π-π interaction that is characterized by parallel orientation of the π-electron moieties. T-shaped interaction is characterized by the perpendicular orientation of the moieties [4]. The dispersion and quadrupole-quadrupole interactions are believed to be of major contribution to the stacking E_{BIND} [1, 5, 6].

We will study the special role of stacking interactions in the formation of the aromatic ligands – NMDA and iGlu receptor complexes. KYNA, an intermediate of the kynurenine pathway of tryptophan metabolism, is a weak aromatic antagonist of iGluRs which preferentially interacts with the glicyne binding site of N-methyl-D-aspartate receptor (NMDAR) NR1 subunit. KYNA is the only known endogenous competitive inhibitor of mammalian iGluRs with neuroprotective properties. It is believed to be important for the design of neuroprotective medicines [7]. When binding to iGluRs, KYNA prevents the development of acute and chronic excitotoxicity [8, 9]. KYNA also modulates the neurodegenerative processes in *Drosophila* central nervous system [10], inhibiting glutamate and/or α7 nicotinic acetylcholine receptors [11].

5,7-di-Cl-KYNA (DCKA), a synthetic specific antagonist of NR1 glycine site, is presumed to form a stacking interaction with Phe^{92} of receptor binding pocket. This is supposed to be the common feature in the binding of aromatic antagonists to iGluRs [12]. DCKA forms several HBs with receptor pocket residues and mechanically stabilizes its “open” conformation (Fig. 1a): S1 and S2 subdomains are separated, that causes the inhibition of the receptor Ca^{2+} channel. The specific affinity of DCKA toward the NR1 glycine site could be explained by hydrophobic interactions [13] together with stacking interaction energy contributions [14].

K_i for KYNA is approximately 1.5^x10^{-5} M for the glycine site and 2^x10^{-4} M for the glutamate site (NR2 subunit) of NMDAR in rat telencephalon membranes [15]. KYNA is also a non-selective inhibitor of α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor (AMPAR, GluR1-GluR4 subunits) and kainate receptor [16]. KYNA affinity for kainate receptors (K_d $7.0\pm0.4^x10^{-5}$ M) is higher than that for AMPAR and lower than that for NR1 NMDAR ($3.5\pm0.6^x10^{-5}$ M) in rat spinal cord afferent C fibers [17]. Seemingly, the relative order of KYNA affinity to glutamate receptor subunits is the following: NR1>GluR>NR2.

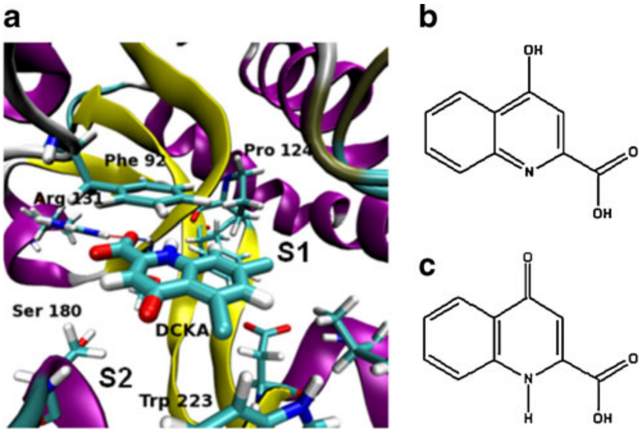


Fig. 1 The complex of antagonist with the binding site of glutamate receptor subunit. (**a**) DCKA in the binding site of NR1 subunit [12]; (**b**) KYNA in enol form; c. KYNA in oxo form

The conservative amino acid sequence and 3D structure of iGluRs subunits: NR1 [12], NR2 [18] and GluR2 [19] – allows us to propose a common mechanism for KYNA interaction with binding sites, both in vertebrate and invertebrate nervous system. The protein homologs of *Drosophila melanogaster*, dNR1, dNR2 and dGluR1, possess a similar conservative amino acid pattern within the putative binding sites (Table 1). Each subunit contains an aromatic residue at the same position as Phe^{92} in NR1.

Computer modeling of KYNA binding to iGluRs may help to reveal the physical basis of interaction specificity. Ab initio quantum chemical calculations of the geometry and E_{BIND} of stacking interaction performed on small model dimers are advantageous for ligand-receptor interaction modeling. The precise calculations of dispersion energy impact in π-π interactions could be performed only considering the electron correlations, for example, at MP2 perturbation theory level [20]. MP2 calculations are known to overestimate the E_{BIND} in benzene dimer by 30% for T-shaped interaction and by ~90% for stacking interaction [5]. This problem can be partly solved by using coupled cluster calculations (CCSD(T)) or, presumably, with the help of the forth order perturbation theory MP4. A parallel conformation with displaced centers of rings (PD) was shown to be energetically the most favorable for benzene [6, 14, 21] and pyridine [22] dimers. Both parallel and antiparallel PD conformers of pyridine dimer are possible, E_{BIND} of each being dependent on relative orientation of the polar N atoms [22]. In benzene dimers both stacking and T-shaped conformations are stabilized mainly by the correlation energy impact, electrostatic energy being attractive only in the T-shaped conformation [6]. A relative contribution of the correlation energy is variable due to different basis sets and computational methods [23]. Since PES of benzene-benzene is very flat, the small energy differences depending on the computational method may

Table 1 Sequence homology of mammalian (rat) and drosophila ionotropic glutamate receptor subunits

NR1 Phe 92	dNR1 Phe 76	NR2 His 88	dNR2 Trp 89	GluR2 Tyr 61	dGluR1 Tyt 98
Pro 124	Pro 112	Ser 114	Ser 115	Pro 89	Ala 127
Thr 126	Thr 114	Thr 116	Met 117	Thr 91	Thr 129
Arg 131	Arg 119	Arg 121	Arg 122	Arg 96	Arg 134
Ser 179	Ser 177	Gly 172	Ser 171	Gly 141	Gly 321
Ser 180	Ser 178	Ser 173	His 172	Ser 142	Ser 322
Val 181	Val 179	Thr 174	Thr 173	Thr 143	Thr 323
Trp 223	Trp 221	Tyr 214	Tyr 213	Leu 192	Val 371
Asp 224	Asp 222	Asp 215	Asp 214	Glu 193	Glu 372

change the stability status of computed PD conformation toward T-shaped or vice versa [1]. However, the existence of stable stacking conformations for aromatic dimers has been shown both empirically [24] and theoretically [25].

To investigate the possible role of stacking interaction in KYNA-iGluR complexes formation we have calculated the optimal geometry and E_{BIND} for KYNA dimers with benzene, imidazole, phenol and indole, considering their stacking interaction with Phe, His, Tyr and Trp residues of NR1/dNR1, NR2A, GluR2/dGluR1 and dNR2 receptor subunits, respectively. Then we docked the optimized dimers into the binding sites of the receptors and calculated the free energy of binding (E_F). The analysis of the binding energy components and geometry parameters of stacking in ligand-receptor complexes reveals the possible mechanisms of interaction specificity and the role of stacking interaction upon KYNA-iGluR complexes formation.

Methods

The coordinates of the receptor heavy atoms were taken from the RCSB Protein Data Bank (www.rcsb.org/pdb/home/home.do): 1pdq_A – NR1, 2a5s_A – NR2A, 1ftl_A – GluR2. The amino acid sequences for *R. norvegicus* and *D. melanogaster* receptors subunits were taken from NCBI data base (www.ncbi.nlm.gov). The automatic homology modeling of drosophila subunits structure in the "open" conformation was performed using Swiss-Model server [26–28]. The structures of rat NR1 and GluR2 were used as templates for drosophila dNR1/dNR2 and dGluR1 models, respectively. Swiss-PdbViewer 3.7 [27] and VegaZZ 2.0.8 [29] programs were used for the addition of hydrogen atoms, dimer construction, computer mutagenesis and manual docking of optimized dimer into the binding site.

The initial conformations of aromatic dimers were chosen to resemble the experimentally found geometry of DCKA – Phe^{484} complex [12], where the benzene ring of Phe and the ligand heteroaromatic ring are in PD conformation and the ligand carboxylic group electrostatically interacts with Arg^{131} residue. The optimal geometry of KYNA – benzene was calculated previously at MP2 level [14]. The starting interplanar spacing and parallel displacement of rings in KYNA – phenol and KYNA – imidazole corresponded to the optimized geometry of KYNA – benzene dimer. The starting geometry of KYNA – indole corresponded to KYNA – Trp complex after ligand docking into dNR2 binding site (Autodock 3.05 software). The following points served as the ring centroids: for benzene and phenol - the midpoints of length between two opposite C atoms, for imidazole – the midpoint of length between C2 atom and midpoint of adjoining C4 - C5 atoms, for indole – the midpoint of conjugated ring common C-C bond, for KYNA – the midpoint between C2 atom and the opposite ring atom. The monomers were neutrally charged during optimization.

Though KYNA can exist both in enol and in oxo tautomeric form (Fig 1b, c), we used the latter one in the majority of calculations, as it was shown to be pharmacologically active [13]. PC GAMESS 7.0 version of GAMESS software [30] was used for ab initio full gradient optimization of geometry (MP2 level of correlations) and final E_{BIND} calculations (MP4(SDQ) level of correlations). The basis set superposition error (BSSE) correction [31] and the permittivity-dependent energy calculations were performed using GAUSSIAN 03 program [32]. The IPCM method was used to model the environment with different permittivity values. We did not make BSSE correction for this type of calculations, as the ghost atoms added to the monomer would be centered in the solvent region, that interferes with the isodensity surface definition [33]. 6-31 G** basis set [34] was used in full optimization and E_{MP4} calculations, aug-cc-pVDZ basis set [35] was used in E_{MP2} calculations with BSSE correction. The full atomic system energy is a sum of Hartree-Fock (E_{HF}) and correlation components (E_{COR}). E_{BIND} was calculated as the difference between the optimized dimer and optimized monomers full energies. E_{MP2} and E_{MP4} are E_{BIND} calculated at MP2 or MP4 levels, respectively.

The computer programs Autodock 3.05, Autodock 4.0 [36–38] and Quantum 3.3.0 (St. Petersburg State University, the Department of Biochemistry, Russia) were used for automatic docking of KYNA into the receptor binding sites. Quantum 3.3.0 was used for E_F and IC50 calculations. The calculations of E_F were performed for the anionic form of KYNA (KYNA$_i$).

The molecular structures were visualized using VMD software [39].

Results

The calculations of dimers energy and optimal geometry

We performed the full gradient optimization for aromatic dimers structures at MP2 perturbation level. E_{BIND} and optimized geometry (Fig. 2, Table 2) was shown to depend on the chemical nature of KYNA aromatic partner. E_{MP2} for previously calculated aromatic dimer was −14.2 kJ mol^{-1} for benzene-benzene and −35.1 kJ mol^{-1} for KYNA – benzene in 6-31 G** basis set [14]. The absolute value of E_{BIND} decreased being revaluated at MP4 level. The addition of diffuse functions significantly increases the binding energy, BSSE correction decreases it almost twofold, as it has been previously shown for benzene dimer [6]. Since the absolute value of E_{BIND} strictly depends on

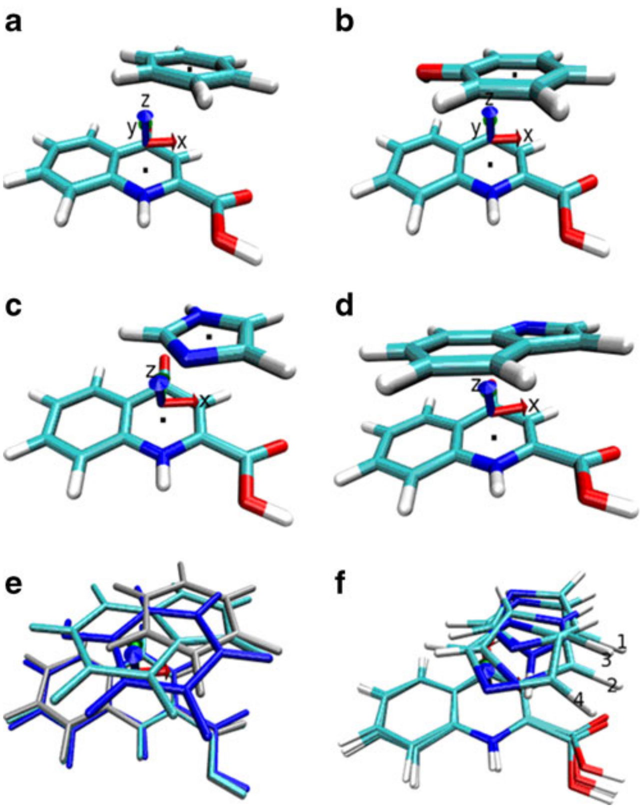


Fig. 2 The optimized structures of KYNA dimers with the aromatic amino acid rings. (**a**) KYNA – benzene; (**b**) KYNA – phenol; (**c**) KYNA – imidazole 3; (**d**) KYNA – indole; (**e**) superposition of dimers: 1. KYNA – benzene, 2. KYNA – phenol, 3. KYNA – indole; (f) superposition of dimers: 1. KYNA enol – imidazole 1, 2. KYNA oxo – imidazole 2; 3. KYNA oxo – imidazole 3; 4. KYNA oxo – imidazole 4. H is depicted by white color, C – by cyan, N – by blue, O – by red

the method of calculation, the relative differences should be considered in ligand-receptor relative affinity analysis. The relative order of E_{BIND} did not change upon the revaluation, except E_{BIND} non BSSE-corrected absolute value for KYNA – phenol in aug-cc-pVDZ basis set becoming higher than for KYNA – imidazole 3. E_{BIND} absolute value increases in the row: benzene – phenol – imidazole – indole. E_{HF} was positive in all cases. Thus, dimer stability was completely determined by E_{CORR} component.

Since the minimal absolute value of E_{BIND} corresponded to KYNA – benzene, the difference in the strength of aromatic interaction *per se* could not explain the highest KYNA affinity to NR1 glycine binding site. The position of benzene was symmetric relative to the long axis of KYNA aromatic moiety and moved half aside the heteroaromatic ring; this is a classical PD conformation [37]. The interplanar (dZ) and parallel (dX) displacements (3.17 and 1.24 Å, respectively) were smaller than those calculated for benzene-benzene using the same approach: 3.4 and 1.4 Å, respectively [14] or using CCSD(T) method at aug (d,p) 6-311 G** basis set: 3.5 and 1.8 Å, respectively [6]. Presumably, the difference between the calculated KYNA – benzene geometry and experimentally established location of DCKA in NR1 [12] resulted from the additional interactions with the receptor binding site residues.

E_{MP2} of KYNA – methylbenzene dimer increased up to −42.10 kJ mol^{-1} (6-31 G** basis set), however, its geometry was almost identical to that of KYNA – benzene. Thus, it can be concluded that unsubstituted benzene ring of Phe residue is sufficient for the correct calculation of KYNA – Phe stacking geometry. The increase of E_{BIND} upon methylation corresponded to the data obtained for substituted benzene [14, 21] and for halogen-substituted KYNA derivatives [14].

Compared to benzene, the phenol ring was displaced from the long axis of KYNA and was closer to its hydrocarbon ring (Fig. 2b, e). The hydrogen of hydroxyl group slightly rotated toward the aromatic ring center, reminding an improper HB formation between the polar hydrogen and π-electronic system [1]. This might additionally stabilize the dimer structure, its E_{BIND} being higher than that for nonpolar benzene (Table 2).

The optimal dimer conformation strictly depended on the relative orientation of rings polar atoms. Thus, in the case of two heteroaromatic residues (His, Trp) the electrostatic interaction might play an important role in the binding specificity. For the different KYNA enol/oxo and imidazole ND1/NE2 tautomers the formation of several energetically stable conformations were shown:KYNA – imidazole 1–4 (Fig. 2c, f). The polar hydrogen atoms were located under the negatively charged O or N atoms of partner ring, fixing the optimal rotational conformation for KYNA – imidazole. The N-H group of indole was located under the 4-oxo group of KYNA (Fig. 2d). The formation of classical HB between the monomers (N-H…O) is also possible, thus, the energy of stacking interaction must be large enough to provide the rings planes parallel orientation. At the same time, the interplane angles differed from zero: N-H side of the partner ring tended to rotate slightly toward N or O atom of KYNA. E_{BIND} was maximal for KYNA – indole, probably due to the superposition of two conjugated π - electron systems.

We also calculated the single-point E_{MP4} for the number of conformations derived from the optimized dimer through the step-by-step displacements of aromatic ring within the E_{BIND} minimum area (Table 3). The calculations using aug-cc-pVDZ basis set with BSSE correction did not significantly change the optimal dZ value (increased to about 0.1 Å). Thus the calculations in 6-31 G** basis set without BSSE correction seems to be appropriate for full geometry optimization of KYNA – benzene dimer. The differences between the MP2 and MP4 PES minima did not exceed 0.2 Å. E_{MP4} minimum for KYNA – benzene corresponded to

Table 2 Stacking energy and geometry of KYNA – aromatic monomers. E_{MP2_diff} – E_{MP2} in aug-cc-pVDZ basis set, in brackets – BSSE-corrected energy; OD – dimers with optimized geometry, MD – dimers after Quantum 3.3.0 MD simulation; dx, dy, dz – the parallel displacement of aromatic ring centroids. ^ X-ray structure data [12]. # KYNA in enol form, in all the rest cases KYNA is in oxo form

Monomer	Aromatic residue and receptor subunit	E_{MP2} kJ/ mol)	E_{MP4} (kJ/ mol)	E_{MP2_diff} (kJ/ mol)	dx (Å)	dy (Å)	dz (Å)	Plane angle (deg)
Benzene- DCKA^	Phe^{92} NR1 OD				0.89	−0.57	3.53	9.6
Benzene	Phe^{92} NR1 OD, Phe^{76} dNR1 OD	−36.86	−17.53	−68.70 (−36.15)	1.24	0.18	3.17	1.5
Benzene	Phe^{92} NR1 MD	−28.91	−17.53		0.68	−0.60	3.70	9.5
Benzene	Phe^{76} dNR1 MD	−27.87	−15.23		0.09	0.55	3.53	4.5
Phenol	Tyr^{61} GluR2 OD, Tyr ^{98} dGluR1 OD	−46.07	−25.02	−79.54 (−42.05)	0.74	−0.64	3.35	6.0
Phenol	Tyr^{61} GluR2 MD	−36.69	−22.30		0.61	−0.40	3.49	4.1
Phenol	Tyr^{98} dGluR1 MD	−38.33	−24.69		0.89	−0.63	3.52	1.7
Imidazole 1#	His^{88} NR2A OD	−42.76	−25.90		1.46	0.29	2.96	10.7
Imidazole 2	His^{88} NR2A OD	−44.73	−27.45		1.28	−0.36	3.10	7.7
Imidazole 3	His^{88} NR2A OD	−50.63	−31.76	−73.05 (−45.10)	1.29	0.60	2.93	6.5
Imidazole 4	His^{88} NR2A OD	−46.19	−30.12		0.77	−0.50	3.22	7.1
Imidazole	His^{92} NR1* MD	−31.92	−22.76					
Imidazole H+	His^{92} NR1* MD	−28.20	−19.04					
Indole	Trp^{89} dNR2 OD	−68.37	−34.77	−112.47 (−64.39)	0.83	0.18	3.09	6.3
Indole	Trp^{89} dNR2 MD	−41.71	−26.73		0.46	0.31	3.63	11.0

dZ of ~3.4 Å which was nearer to experimentally found value [12]. The same shift was observed in the cases of KYNA – phenol and KYNA – indole (not shown). For KYNA – phenol and KYNA – imidazole 3 there were no significant differences in optimal parallel displacements values: possibly, the less computation method-sensitive polar interactions mainly define the structure of dimers.

The full optimization of KYNA – benzene and KYNA – imidazole geometry without MP2/MP4 correlation calculations did not produce the stable stacking conformation. The optimization of KYNA – benzene structure achieved by semi-empirical calculations (AM1) gave the parallel orientation of the ring planes with the interplanar spacing greater then 5 Å, extremely differing from both experimental and calculated stacking geometry. Definitely, E_{COR} should be considered for the correct description of ligand-receptor stacking interaction.

The calculation of dimers potential energy surfaces

The flat form of stacking PES might define its functional role in ligand-receptor complex formation. This initial interaction provides a proper position for the ligand within the receptor binding site providing enough freedom for the ligand in the receptor pocket, enabling HBs to be formed. In order to check this assumption we had to compare PS curves for HB and stacking interaction.

The calculations of benzene – benzene, KYNA – benzene and KYNA – imidazole 3 PES were performed for the parallel displacements within the limits of ±1 Å (Fig. 3). MP2 level-optimized dX/dY and MP4 level-optimized dZ (3.70 Å - for dibenzene) were chosen as the initial displacements for each dimer. The benzene – benzene

Table 3 E_{MP4} potential surfaces of KYNA – aromatic rings dimers (kJ mol^{-1}). 1. Displacements; 2. KYNA – benzene; 3. KYNA – phenol; 4. KYNA – imidazole 3. The energy values for new local minima are printed by black. In brackets: the values of E_{BSSE} (MP2, aug-cc-pVDZ basis set)

1	2	3	4
0	−17.53 (−36.15)	−25.02	−31.80
dZ +0.1	−19.37 (−36.57)	−26.82	−33.18
dZ +0.2	−19.96 (−35.73)	−27.15	−33.10
dZ +0.3	−19.66 (−34.14)	−26.32	−31.97
dX −0.3	−13.85	−23.39	−29.92
dX −0.2	−15.40	−24.06	−30.54
dX −0.1	−16.65	−24.52	−31.17
dX +0.1	−18.03	−24.94	−31.88
dX +0.2	−18.16	−24.89	−31.80
dX +0.3	−18.03	−24.69	−31.46
dY −0.3	−14.64	−23.89	−31.34
dY −0.2	−15.86	−24.31	−31.84
dY −0.1	−16.78	−24.64	−31.97
dY +0.1	−17.95	−24.89	−30.96
dY +0.2	−18.16	−24.81	−29.79
dY +0.3	−18.12	−24.69	−28.07

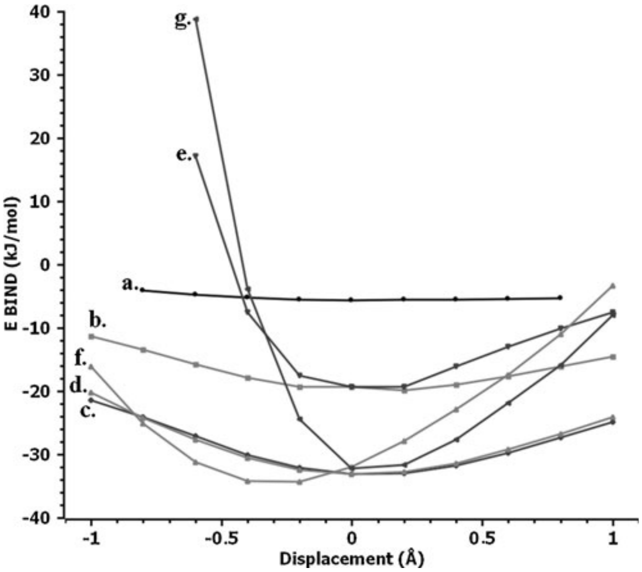


Fig. 3 E_{MP4} PES of aromatic dimers stacking interaction and KYNA – Pro[124] hydrogen bonds. (**a**) Benzene – benzene, dX; (**b**) KYNA – benzene, dX; (**c**) KYNA –imidazole 3, dX; (**d**) KYNA – imidazole 3, dY; (**e**) KYNA – benzene, dZ; (**f**) KYNA – Pro[124], dX; (**g**) KYNA – Pro[124], dZ

PES (a.) was extremely flat, E_{MP4} value being almost invariant within the given spatial limits. KYNA – benzene dX PES (b.) had a more marked minimum, but E_{MP4} value changes did not exceed the median thermal vibration energy (2.5 kJ mol^{-1} at 298 K) within ±0.5 Ǻ displacements. The positive dZ displacement (e.) led to significantly more pronounced loss of E_{BIND}, though for medium dZ value (about 0.3 Ǻ) it was also of the thermal vibration order. The left half of KYNA – benzene dZ PES curve was very steep due to the tight spatial approach of the aromatic rings. KYNA – imidazole 3 dX/dY PES curve (c., d.) resembled the right half of KYNA – benzene dZ curve, locating much lower than the last one. It can be assumed that at higher permittivity values it would approach the position and form of KYNA dX/dY PES.

E_{BIND} for KYNA – Pro[124] HB was calculated for the series of dX (f.) and dZ (g.) displacements starting with dynamically optimized ligand position within NR1 binding site (Quantum 3.3.0). Their PES, especially those of Pro dZ, were steeper than those of KYNA – benzene and KYNA – imidazole 3 (Fig. 3), crossing these curves at points of±0.5 – 1 Ǻ from the initial displacements. Thus, stacking interaction should play an important role at the initial stage of the complex formation when HBs are weak, allowing KYNA to optimize its HBs pattern without a significant loss of E_{BIND}. The vertical aromatic ring mobility is more limited compared to the parallel one.

The rotation of benzene ring beyond the plane parallel to KYNA (20/40° around X axis) led to the drastic fall of KYNA – benzene E_{MP4} absolute value. Probably, this was resulted from the steric overlapping of the ring electronic densities. dZ displacement increased the dimers stability: PES curve moved parallel to the area of higher dZ values (Fig. 4). Thus, non-planar rotation of KYNA in the binding site is impossible without a preliminary significant separation of the rings.

Dispersion and quadrupole interactions seem to define the stability of dimers in vacuum [1]. It is still obscure, whether π-π interaction possesses some special characteristics distinguishing it from the ordinary vdW interaction between two non-polar aliphatic groups. We calculated E_{MP4} PES (dX and dZ displacements) for the optimized dimer of Leu side chains (Fig. 5). Both E_{BIND} values and PES form resembled those for KYNA – benzene (6-31 G** basis set). As for KYNA – benzene, the impact of E_{COR} almost completely defined the dimer stability (E_{MP2}=−23.97 kJ mol^{-1}, E_{MP4}=−20.25 kJ mol^{-1}, E_{HF}=−0.88 kJ mol^{-1}).

The addition of diffuse functions on atoms is important for the precise estimation of aromatic E_{BIND} [37]. E_{MP2} of KYNA – benzene and Leu – Leu (6-31++G** basis set) was −57.65 kJ mol^{-1} and −26.40 kJ mol^{-1}, respectively: the addition of diffuse functions increased the absolute value of E_{MP2} by 36% and 10%. Thus, π-π interaction seemed to possess the specific properties compared to aliphatic one, and the precise value of its E_{BIND} could be calculated only in the basis set including the diffuse functions.

The electrostatic effects should greatly reduce in the environment with high permittivity (ε), varying in proteins from 3–4 in the hydrophobic core up to ~80 in the outer solvent layer. For nonpolar benzene E_{MP2} only slightly depended on ε value, the electrostatic forces minimally contributing to dimer stability (Fig. 6). On the contrary,

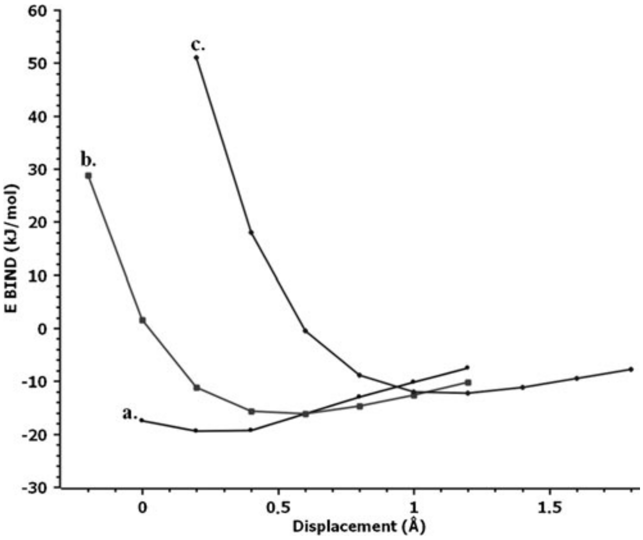


Fig. 4 E_{MP4} PES of KYNA – benzene dimers with the rotated benzene ring, dZ. (**a**) 0 deg.; (**b**) 20 deg.; (**c**) 40 deg

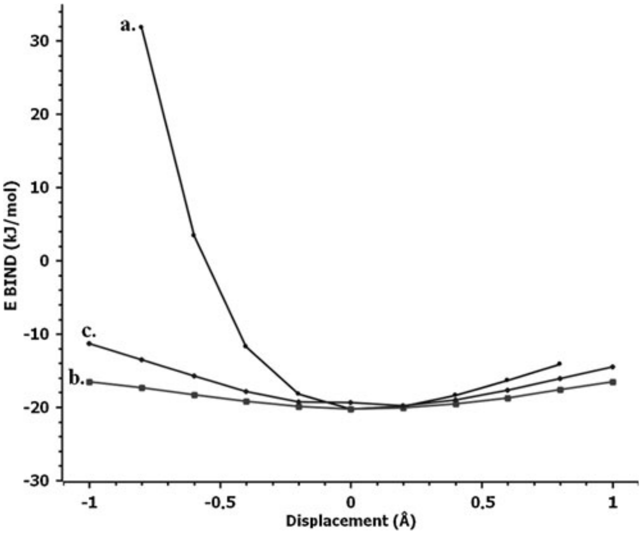


Fig. 5 E_{MP4} PES of noncovalent dimers. (**a**) Leu – Leu, dZ; (**b**) Leu – Leu, dX; (**c**) KYNA – benzene, dX

E_{BIND} of imidazole – imidazole decreased greatly with the growth of ε, up to the limit approximately equal to that of benzene – benzene. E_{MP2} for KYNA – benzene remained highly negative (−28.5 – -24.3 kcal mol^{-1}) at ε=4 – 10, the possible permittivity values for the receptor binding pocket. Obviously, vdW interaction is stronger in KYNA – benzene than in benzene – benzene and imidazole – imidazole. E_{MP2} for KYNA – imidazole 3 sharply grew, becoming positive at ε=20, possibly due to the suboptimal dZ value. Although the absolute energy values were apparently overestimated due to the lack of BSSE correction, we could observe the clear trend of the polar interactions weakening in the solution. The hydrophobic surroundings with low permittivity might be an important factor for the formation of stacking bond between the heteroaromatic rings.

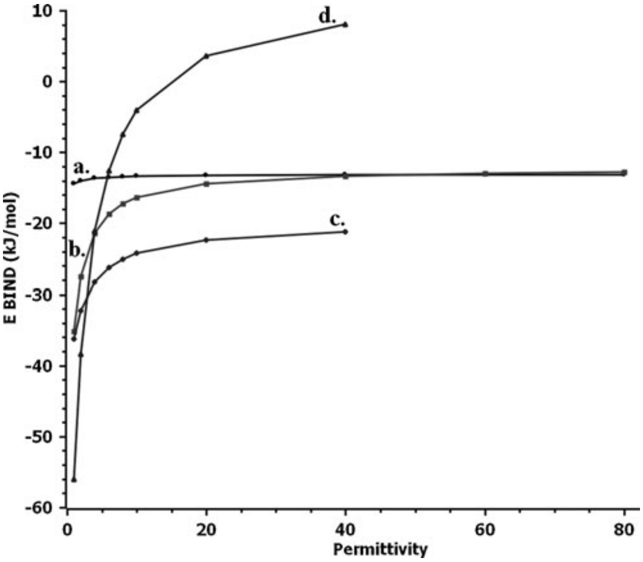


Fig. 6 The permittivity-E_{MP2} dependence. (**a**) benzene – benzene; (**b**) imidazole – imidazole; (**c**) KYNA – benzene; (**d**) KYNA – imidazole 3

Docking of optimized dimer structures into receptor binding sites

Manual docking was performed by superimposing the aromatic planes of monomer and receptor aromatic residue (Fig. 7). E_F and IC50 for KYNA docked into binding sites were calculated and then revaluated following the dynamical energy minimization (MD) of the complex using Quantum 3.3.0 software (Table 4). Since PDB structure of NR2 subunit was available only in the “closed” conformation, in order to model KYNA – His binding in the “open” conformation we introduced a computer mutation Phe^{92}/His into NR1 (NR1*). Using NR1 we also constructed the NR2 binding site (NR1**) via mutations of the following residues: Phe^{92}/His, Pro^{124}/Ser, Trp^{223}/Tyr, Phe^{250}/Tyr.

The values of E_F and IC50 for KYNA interactions with binding sites differed before and after MD simulation: the affinity diminished after MD for the majority of complexes. IC50 (2.51^{x}10^{-5} M for KYNA in NR1 prior to MD) was close to the experimental values: both to IC50 in the absence of glycine (1.5^{x}10^{-5} M [11]) and to K_i (1.5^{x}10^{-5} M [15]). The similar value was obtained for dNR1. The affinity of KYNA to GluR2 and NR1** was lower, especially after MD. NR1* (His in cationic form) had higher affinity to KYNA than NR1. Additional three mutations from aliphatic to polar residues in NR1** diminished KYNA interaction with NR2-like receptor pocket. The low affinity of KYNA for dNR2 and dGluR2 subunits rises after MD, mostly due to the electrostatic interactions optimization.

The MD-optimized KYNA – Phe^{92} orientations within *Rattus* and *Drosophila* NR1 binding sites resembled the orientations calculated for KYNA – benzene dimer, though all of them somewhat differed in the value of parallel displacements and interplanar angles (Table 2). Notably, KYNA position in NR1 binding site after MD was in agreement with the experimentally found DCKA position [12], revealing the same pattern of ligand-receptor interactions (Fig. 7a). HB was formed when the ligand was in oxo form: $KYNA_i$ NH – O=C Pro^{124} (E_{MP4}=−32.17 kJ mol^{-1}). The polar bond was formed between $KYNA_i$ COO group and NH group of Thr^{126} (E_{MP4}=−89.24 kJ mol^{-1}). Coulomb interaction was formed between $KYNA_i$ and Arg^{131} ($E_{MP2/MP4}$=−426.68 /-422.04 kJ mol^{-1}).

The MD-optimized space parameters of KYNA – Tyr$^{61/98}$ dimer were almost identical to those calculated for KYNA – phenol (Fig. 7b). The hydroxyl group had a tendency to approach KYNA aliphatic ring center, thereby confirming our hypothesis that the interaction of polar

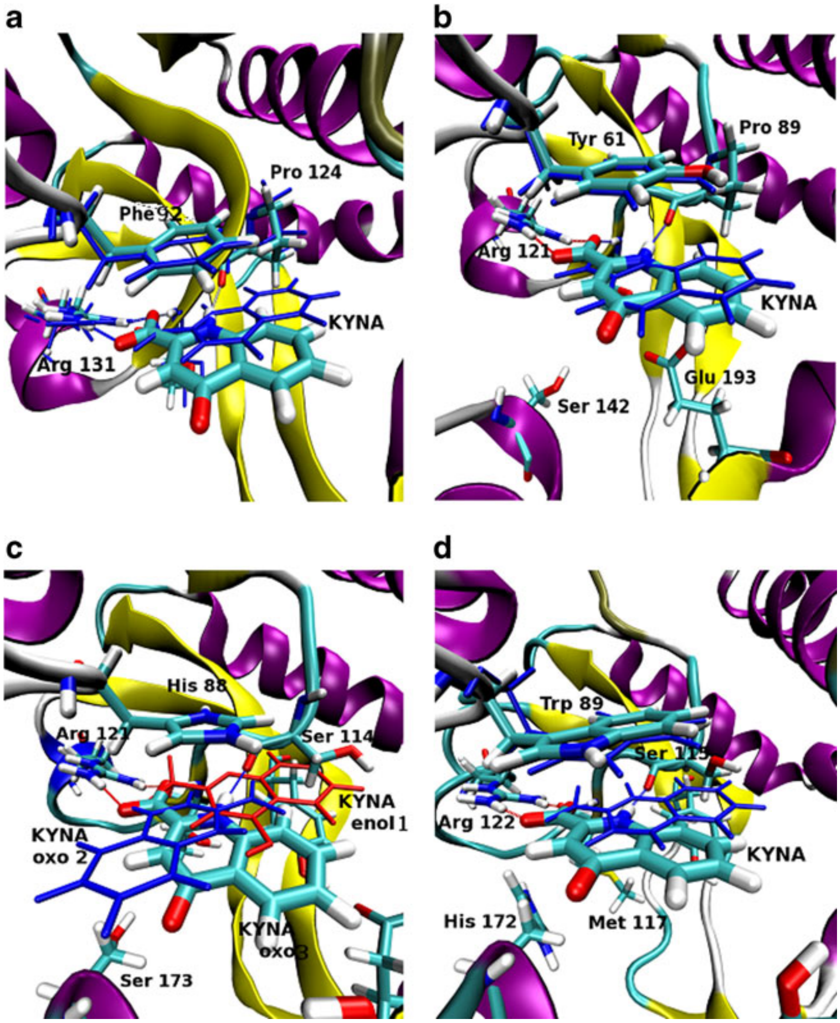


Fig. 7 The complexes of KYNA – glutamate receptor binding sites before and after MD simulation. (**a**) NR1; (**b**) GluR2; (**c**) NR1**; (**d**) dNR2. KYNA – binding sites complexes before MD are depicted by blue or red thin sticks, after MD – by cyan sticks. For NR1** the His88 aromatic ring is shown in the cationic form; the optimized complexes of imidazole with KYNA both in enol form (1) and in oxo form (2, 3) are depicteds

hydroxyl H with the aromatic ring stabilizes the dimer structure.

The position of KYNA within NR1* and NR1** binding pockets depended on its orientation relative to His imidazole group. The ligand binding was impossible for KYNA oxo – His ND1_H dimer (KYNA – imidazole 2), because the orientation of KYNA prohibited the formation of HBs and electrostatic bonds (Fig. 7c). Thus, to bind the physiologically active oxo form of KYNA His^{92} should be in NE2_H or in cationic $HisH^+$ (ND1_H, NE2_H) form. Hence the ligand-receptor affinity should depend on the extracellular pH. Importantly, E_{MP4} was −8.95 kJ mol^{-1} for KYNA – imidazole rotational position with two NH groups approaching each other: the binding energy diminished by 18.49 kJ mol^{-1} relative to the optimal conformation. It is known that the parallel rotation of rings in dibenzene is almost unrestricted: the energy value changes are about 0.04 kJ mol^{-1} [6]. In the case of KYNA – imidazole several directionally restricted energy minima were observed. Thus, the protonation character of the rings might predispose the ligand optimal orientation within the binding pocket.

Although quantum automatic docking failed to produce $NR1^*HisH^+$ – KYNA complex similar to NR1 – DCKA, the binding of KYNAi to $HisH^+$ form of NR1 was more energetically favorable than that to the uncharged form, mostly due to the additional electrostatic interaction. The similar complex was generated automatically for NR1* His NE2_H, providing evidence for the importance of proper protonation for biologically significant ligand-receptor docking.

The orientation of KYNA in dNR2 after MD somewhat differed from its initial position relative to indole. In that position KYNA could form the conservative pattern of ligand-receptor interactions (Fig. 7d). Probably, the interaction between NH (Trp ring) and KYNA carbonyl group affected the ligand orientation within the receptor site.

E_{MP4} of KYNA – aromatic group dimer was calculated for each ligand-receptor complex after MD simulation (Table 2). For benzene and phenol E_{MP4} did not differ significantly from that calculated for optimized dimers. Though the absolute value of E_{MP4} was lower in NR1* for $HisH^+$ form than for His NE2_H (−19.07 and −22.80 kJ mol^{-1}, respectively),

Table 4 The energies of ligand-receptor interaction (Quantum 3.3.0).1 – previous to MD simulation, 2 – after MD simulation. $ NR1 in complex with DCKA [12]

Receptor subunit	E_F (kJ/mol)		IC50 ($^{x}10^{-5}$ M)		$E_{ELECTROSTATIC}$ (kJ/mol)		E_{VDW} (kJ/mol)	
	1	2	1	2	1	2	1	2
NR1	−26.69	−22.80	2.57	10.1	−15.90	−32.43	−29.71	−21.46
NR1 (Phe$^{92/}$Ala)	−17.32	−17.49	106	98.5	−5.06	−36.40	−22.22	−14.23
NR1 – DCKA$^{\$}$	−30.46	−20.54	0.58	29.4	−17.03	−30.50	−34.52	−19.96
NR1 (Arg^{131}/Ala)	−22.72	−18.54	12.4	65.5	−7.24	4.94	−28.79	−25.23
dNR1	−26.07	−23.43	3.32	9.36	−13.43	−23.56	−30.17	−24.60
NR1* (His HE2)	−26.57	−22.18	2.69	15.5	−19.12	−24.69	−27.78	−22.34
NR1* (HisH+)	−27.61	−23.60	1.80	8.80	−19.08	−30.92	−29.25	−22.80
NR1** (HisH+)	−26.07	−21.51	3.30	20.3	−18.83	−47.45	−27.24	−19.25
GluR2	−25.82	−21.00	3.67	24.6	−12.47	−18.33	−30.33	−22.26
GluR2 (Tyr^{61}/Ala)	−22.05	−16.23	16.3	162	−12.01	−21.25	−25.15	−15.23
dGluR1	−25.40	−26.90	4.33	2.37	−12.34	−30.12	−29.83	−26.74
dNR2	−23.22	−24.73	10.2	5.62	−15.86	−30.08	−24.85	−26.78

it could provide the ligand-receptor binding. The interplanar distance increased after MD simulation, becoming closer to MP4 revaluated optimal dZ value. Since the geometry of ligand-receptor complex should also depend on HB formation and Coulomb interactions, we simulated MD in the absence of aromatic interaction, namely in the complexes of KYNA with mutant NR1 (Phe^{92}/Ala) and GluR2 (Tyr^{61}/Ala). KYNA changed its position within the binding sites and its E_F absolute value lowered, mostly due to vdW energy decrease (Table 4). In fact, π-π interaction determined KYNA position within the receptor binding site and significantly increased its affinity to receptor.

We calculated E_{MP4} values for several aromatic ring dimers with KYNA$_i$ after MD. It decreased moderately for benzene (−13.39 kJ mol^{-1}) and phenol (−18.37 kJ mol^{-1}), and more significantly for imidazole 3 (−15.02 kJ mol^{-1}) and indole (−17.70 kJ mol^{-1}) relative to the uncharged form of KYNA. E_{MP4} for the cationic form of imidazole was −291.83 kJ mol^{-1}: its interaction with KYNA$_i$ appeared to be Coulomb interaction. However, MP2 level calculations demonstrated that KYNA$_i$ – imidazole stacking conformation was not stable in vacuum and trended to form a classical HB. Since the true ε value was higher than 1, the electrostatic influence on geometry and E_{BIND} should be less severe.

KYNA flexible docking into NR1* and NR1** receptor binding sites

In order to confirm our assumption that the location of protons may influence the ligand binding to His residue, we performed Autodock 4.0 automatic docking of KYNA oxo into NR1* binding site with movable His^{92} residue in HD1_H or HE2_H tautomeric form. The majority of 100 structures generated using Lamarckian genetic algorithm was similar to the experimentally revealed NR1 – DCKA complex. Being in ND1_H form, His residue interacted with negatively charged KYNA carboxyl group, their aromatic rings tilting from parallel orientation and disrupting the stacking interaction. Surprisingly, His in NE2_H form and KYNA tended to acquire the equal directional orientation of NH groups, with the KYNA polar H atom located under the negatively charged N of His. Presumably, this conformation was stabilized by polar interactions, though KYNA – imidazole E_{MP4} for such a conformation was only −8.91 kcal mol^{-1}. KYNA position within the binding pocket seemed to be determined mostly by Coulomb and hydrogen bond patterns. Undoubtedly, the localization of protons affects the ligand-receptor binding, but its influence is rather complicated.

Discussion

Ab initio quantum-chemical calculations have shown that the relative spatial position of the ring polar atoms significantly influence the optimal geometry parameters of KYNA – aromatic residue stacking interaction. In KYNA – imidazole the imidazole polar hydrogen is located under the negatively charged N or O atoms of KYNA. The same has been observed in stacked imidazole dimer [25]. The protonation character of KYNA and imidazole rings defines the optimal rotation conformation of KYNA – imidazole, as in His complexes with the DNA nucleobases [40]. Compared to KYNA – Phe, the polar attraction additionally stabilizes KYNA – Tyr (an improper HB formation), KYNA – His and KYNA – Trp complexes, simultaneously restricting the ligand orientation freedom within the binding site.

KYNA was shown with the help of computer modeling to interact with rat and drosophila iGluRs, being an antagonist of both glycine and glutamate binding sites. The values of IC50 are closer to empirically found after the manual docking of the quantum-chemically optimized dimers into the binding sites, than after MD. The geometry of stacking interaction found by Quantum 3.3.0 resemble that being calculated ab initio, though addition interactions (Coulomb, HB and vdW) with receptor amino acid residues evidently influence KYNA position within the binding site. The ligand-receptor affinity decreases for NR1** and GluR2 compared to NR1, that corresponds to the experimental fact that KYNA is a more specific agonist of the glycine binding site. dNR1 has the antagonist binding properties similar to NR1, their binding sites being virtually identical (Table 1). dNR2 subunit may have the unusual binding properties, containing a large aromatic Trp residue within the binding site. KYNA affinity to dNR2 and dGluR1 becomes very high after MD. This may reflect some initial steric tension in the structure of binding sites obtained via an automatic homology modeling.

It is difficult to perform MD for ligand-receptor complex considering stacking effects, because this would require the correct parametrization of stacking interaction. The physical nature of stacking is in part the dispersion interaction, which E_{BIND} might be correctly estimated only using the basis set with diffuse functions. E_{MP4} of the fully optimized benzene dimer (6-31 G** basis set) is −5.69 kJ mol^{-1}, while the precisely calculated E_{BIND} of stacked benzene dimer is two-time larger: -11.30 kcal mol^{-1} [41]. Thus, we expect the further increase in E_{BIND} for KYNA – aromatic rings using the precise calculation methods.

The difference between vdW energy of KYNA – NR1 and KYNA – NR1 Phe^{92}/Ala (Quantum 3.3.0, after MD; Table 4) can be considered as the loss of stacking interaction resulting from Phe computer mutation. Hydrophobic and vdW interactions might be responsible for the increase in DCKA affinity to NR1 type of receptor subunit [13], for example, the interaction of ligand with large hydrophobic system of Trp^{223} residue. However, Trp^{223}/Ala computer mutation did not significantly change the E_{BIND} value for KYNA – NR1 (Table 4). One of the possible reasons for the increase in ligand specificity is the rise of stacking energy in halogenated KYNA, as well as in halogenated benzene dimers [14]. The affinity of DCKA to NR1 in the X-ray derived DCKA – NR1 complex [12] increases in comparison with KYNA: IC50 is 5.8^{x}10^{-6} M, though the experimentally found value is 0.06 – 0.1^{x}10^{-6} M [42]. Seemingly, NR1 – DCKA complex is additionally stabilized by stacking interaction.

Both vertical and horizontal parallel movements of KYNA inside the receptor site are accompanied by the small loss of E_{BIND} (the order of thermal vibration at T 298 K). This can be an important background for the formation of conservative HBs with backbone Thr N and Pro/Ser O atoms which are restricted in their movements. HB is a non-covalent interaction sensitive to the spatial position of atoms: the donor-acceptor distance and donor-hydrogen-acceptor angle should be 2.6 – 3.2 Å and 155 – 180°, respectively. Thus, HBs pattern is an important factor restricting the position of KYNA within the binding site. The following stages for KYNA – receptor interaction are proposed:

1. The initial formation of ligand-receptor stacking interaction, the starting ligand orientation within the receptor pocket;
2. The formation and tuning of HBs.

The improper location of KYNA and/or of His residue may prohibit the second stage, because KYNA carboxylic end may be directed out of the receptor pocket. Also, His tends to form HB with KYNA and this may disorientate the ligand within the binding pocket. Tyr OH group may slightly restrict the freedom of KYNA parallel movements within GluR2 and dGluR1 binding sites. Due to the possible existence of several energetic minima for KYNA – indole dimer, its PS was not evaluated in detail: only one physiologically active position for the ligand in dNR2 was found. The existence of two minima was shown for KYNA – benzene complex, but one of them is not realized in NR1 glycine binding site [14]. Therefore, in spite of higher E_{MP4} value for KYNA interaction with heteroaromatic rings its orientation freedom decreases compared to KYNA – benzene dimers. The absence of rotational preference for KYNA – benzene stacking interaction could speed up the first preliminary stage of KYNA – receptor interaction, while the incorrect pairing with imidazole due to the “improper” charges distribution must be followed by the stage of reassociation. In spite of being fast, these processes may decrease the whole time of ligand-receptor complex existence, thereby decreasing KYNA inhibitory properties. This may somewhat diminish KYNA affinity to NR2 subunits relatively to NR1.

KYNA – Arg Coulomb interaction must be important for ligand binding and orientation within receptor site. The absolute value of Coulomb energy at ε=4 (the approximate value for protein hydrophobic core) is still very high: E_{MP2} is −141.3 kJ mol^{-1}, compared to −438.2 kJ mol^{-1} at ε=1. Quantum 3.3.0 gives smaller values of electrostatic energy for ligand-receptor interaction (Table 4). The difference of electrostatic energy between NR1 – KYNA and NR1 Arg^{131}/Ala – KYNA is 37.32 kJ mol^{-1}: the total energy gain is close to stacking energy values. At least 2 HBs with water molecules can be formed by free ionic groups of both Arg and KYNA (E_{MP2} in vacuum is ~−84 and −113 kJ mol^{-1}, respectively) which are disrupted after the ligand-receptor complex formation. We propose that Quantum 3.3.0 value (~ −38 kJ mol^{-1}) is close to actual ligand-receptor electrostatic binding energy. Thus, the energy value of stacking within the receptor site is of the same order as that of Coulomb interaction. The electrostatic contribution to the E_{BIND} sharply decreases upon the permittivity growth.

KYNA and its synthetic derivatives are known to be pharmacologically active only in the 4-oxo tautomeric form [13]. The carbonyl group of DCKA forms a HB with water molecule in receptor binding site [12]. Another HB is formed between NH group of KYNA in oxo form and backbone C=O group of receptor Pro/Ser. In agonist glycine N atom is in sp^3-hybridization and hydrogen of NH3 group is directed toward carbonyl O of Pro. In KYNA NH group is located within the plane of aromatic system (Fig. 1). As shown via Quantum MD simulation, KYNA keeps rotating until N-H and C=O groups form the straight line.

We used a rather simple model, the complex of KYNA with iGluR subunit, to illustrate the impact of stacking interaction in ligand-receptor complex formation. These effects may be even more important for other molecular systems – receptors, izozymes or other structurally related proteins, where the binding specificity and activity depends on the local variations of functional aromatic amino acids. Stacking interaction, its E_{BIND} and geometry should be considered while developing new drugs with a specific action toward certain molecular targets, or when constructing new proteins with special binding functions. The insertion of polar atoms into the ring structure not only enhances its affinity to binding site residues, but also governs its rotational and spatial position. This might be crucial for fine tuning of the complex bonds pattern.

Acknowledgments We thank Dr. Alex A. Granovsky, Moscow State University for supply of PC GAMESS 7.0 software. We also thank the Department of Biochemistry, Biological and Soil Faculty of St. Petersburg State University for the help in computer docking accomplishment (Quantum 3.3.0 software) and Prof. O. E. Kvyatkovskii, Ioffe PTI, Russ. Acad. Sci. for the help with performance of quantum-chemical calculations (GAUSSIAN 03 software).

This work was supported by the Russian Foundation for Basic Research grant 09-04-01208.

References

1. Müller-Dethlefs K, Hobza P (2000) Noncovalent interactions: A challenge for experiment and theory. Chem Rev 100:143–167
2. Bhattacharyya R, Saha RP, Samanta U, Chakrabarti P (2003) Geometry of interaction of the histidine ring with other planar and basic residues. J Proteome Res 2:255–263
3. Biot C, Buisine E, Kwasigroch JM, Wintjens R, Rooman M (2002) Probing the energetic and structural role of amino acid/nucleobase cation-pi interactions in protein-ligand complexes. J Biol Chem 277:40816–40822
4. Waters ML (2000) Aromatic interactions in model systems. Curr Opin Cell Biol 6:736–741
5. Hobza P, Selzle HL, Schlag EW (1996) Potential energy surface for the benzene dimer. Results of ab initio CCSD(T) calculations show two nearly isoenergetic structures: T-shaped and parallel-displaced. J Phys Chem 100:18790–18794
6. Tsuzuki S, Honda K, Uchimaru T, Mikami M, Tanabe K (2002) Origin of attraction and diractionality of the pi / pi interaction: model chemistry calculations of benzene dimer interaction. J Am Chem Soc 124:104–112
7. Danysz W, Parsons CG (1996) Glycine and N-methyl-D-aspartate receptors: physiological significance and possible therapeutic applications. Pharmacol Rev 50:597–664
8. Foster AC, Vezzani A, French ED, Schwarcz R (1984) Kynurenic acid bloks neurotoxicity and seizures induced in rats by the related brain metabolite quinolinic acid. Neurosci Lett 48:273–278
9. Smith DH, Okiyama K, Thomas M, Mclntosh TK (1993) Effects of the excitatory amino acid receptor antagonists kynurenate and indole-2-carboxylic acid on behavioral and neurochemical outcome following experimental brain injuri. J Neurosci 13:5383–5392
10. Savvateeva E, Popov A, Kamyshev N, Bragina J, Heisenberg M, Senitz D, Kornhuber J, Riederer P (2000) Age-dependent memory loss, synaptic pathology and altered brain plasticity in the Drosophila mutant cardinal accumulating 3-hydroxykynurenine. J Neural Transm 107:581–601
11. Hilmas C, Pereira EF, Alkondon M, Rassoulpour A, Schwarcz R, Albuquerque EX (2001) The brain metabolite kynurenic acid inhibits alpha7 nicotinic receptor activity and increases non-alpha7 nicotinic receptor expression: physiopathological implications. J Neurosci 21:7463–7473
12. Furukawa H, Gouaux E (2003) Mechanisms of activation, inhibition and specificity: crystal structures of the NMDA receptor NR1 ligand-binding core. EMBO J 22:2873–2885
13. Leeson PD, Baker R, Carling RW, Curtis NR, Moore KW, Williams BJ, Foster AC, Donald AE, Kemp JA, Marshall GR (1991) Kynurenic acid derivatives. Structure-activity relathionship for excitatory amino acid antagonism and identification of potent and selective antagonists at the glycine site on the N-methyl-D-aspartate receptor. J Med Chem 34:1243–1252
14. Zakharov GA, Popov AV, Savvateeva-Popova EV, Shchegolev BF (2008) The role of stacking interactions in the mechanisms of binding of the glycine site of NMDA-receptor with antagonists and 3-hydroxykynurenine. Biofizika 53:22–29
15. Kessler M, Terramani T, Lynch G, Baudry M (1989) A glycine site assotiated with N-methyl-D-aspartic acid receptors: characterization and identification of a new class of antagonists. J Neurochem 52:1319–1328
16. Birch PJ, Grossman CJ, Hayes AG (1988) Kynurenate and FG9041 have both competitive and non-competitive antagonis actions at exitatory amino acid receptors. Eur J Pharmacol 151:313–315
17. Evans RH, Evans SJ, Pook PC, Sunter DC (1987) A comparison of excitatory amino acid antagonists acting at primary afferent C fibers and motoneurones of the isolated spinal cord of the rat. Br J Pharmacol 91:531–537
18. Furukawa H, Singh SK, Mancusso R, Gouaux E (2005) Subunit arrengement and functions in NMDA receptors. Nature 438:185–192

19. Armstrong N, Gouaux E (2000) Mechanisms for activation and antagonism of an AMPA-sensitive glutamate receptor: crystal structures of the GluR2 ligand binding core. Neuron 28:165–181
20. Möller C, Plesset MS (1934) Note on an approximation treatment for many-electron systems. Phys Rev 46:618–625
21. Lee EC, Kim D, Jurecka P, Tarakeshwar P, Hobza P, Kim KS (2007) Understanding of accembly phenomena by aromatic-aromatic interactions: benzene dimer and the substituted systems. J Phys Chem A 111:3446–3457
22. Mishra BK, Sathyamurthy N (2005) Pi-pi interaction in pyridine. J Phys Chem A 109:6–8
23. Tsuzuki S, Uchimaru T, Sugawa K, Mikami M (2002) Enegry profile of the interconversion path between T-shape and slipped-parallel benzene dimmers. J Chem Phys 117:11216–11221
24. McGaughey GB, Gagne M, Rappe AK (15458) π-Stacking interactions. Alive and well in proteins. J Biol Chem 273:15458–15463
25. Zhuravlev AV, Shchegolev BF, Savvateeva-Popova EV, Popov AV (2009) Molecular mechanisms of imidazole and benzene rings binding in protein. Biochemistry (Mosc) 74:1135–1144
26. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL Workspace: a web-based environment for protein structure homology modeling. Bioinformatics 22:195–201
27. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723
28. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 31:3381–3385
29. Pedretti A, Villa L, Vistoli G (2004) Vega – an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming. J Comput Aided Mol Des 18:167–173
30. Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH, Koseki S, Matsunaga N, Nguyen KA, Su SJ, Windus TL, Dupuis M, Montgomery JA (1993) General atomic and molecular electronic structure system. J Comput Chem 14:1347–1363
31. Boys SF, Bernardi F (1970) The calculation of small molecular interactions by the difference of separate total energies. Some procedure with reduced errors. Mol Phys 19:553–566
32. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2003) Gaussian 03, Revision B.05. Gaussian Inc, Pittsburgh, PA
33. Scherlis DA, Marzari N (2004) π-Stacking in charged thiophene oligomers. J Phys Chem B 108:17791–17795
34. Gordon MS (1980) The isomers of silacyclopropane. Chem Phys Lett 76:163–168
35. Kendall RA, Dunning TH, Jr H, Harrison RJ (1992) Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. J Chem Phys 96:6796–6806
36. Goodsell DS, Olson AJ (1990) Automated docking of substrates to protein by simulated annealing. Proteins 8:195–202
37. Morris GM, Goodsell DS, Huey R, Olson AJ (1996) Distributed automated docking of flexible ligands to proteins: parallel applications of Autodock 2.4. J Comput Aided Mol Des 10:293–304
38. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19:1639–1662
39. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14:33–38
40. Churchill CDM, Wetmore SD (2009) Noncovalent interactions involving histidine: the effect of charge on π-π stacking and T-shaped interaction with the DNA nucleobases. J Phys Chem B 113:16046–16058
41. Sinnokrot MO, Valeev EF, Sherrill CD (2002) Estimates of the ab initio limit for pi-pi interactions: the benzene dimer. J Am Chem Soc 124:10887–10893
42. Laurie DJ, Seeburg PH (1994) Ligand affinities at recombinant N-methyl-D-aspartate receptors depend on subunit compositions. Eur J Pharmacol 268:335–345

J Mol Model (2012) 18:1767–1777
DOI 10.1007/s00894-011-1208-z

ORIGINAL PAPER

The effect of anchoring group number on molecular structures and absorption spectra of triphenylamine sensitizers: a computational study

Jie Xu · Ligen Zhu · Lei Wang · Li Liu · Zikui Bai ·
Luoxin Wang · Weilin Xu

Received: 19 May 2011 / Accepted: 1 August 2011 / Published online: 12 August 2011

Abstract The molecular structures and absorption spectra of triphenylamine dyes containing different numbers of anchoring groups (S1-S3) were investigated by density functional theory (DFT) and time-dependent DFT. The calculated geometries indicate that strong conjugation is formed in the dyes. The interfacial charge transfer between the TiO_2 electrode and S1-S3 are electron injection processes from the excited dyes to the semiconductor conduction band. The simulated absorption bands are assigned to $\pi\rightarrow\pi^*$ transitions according to the qualitative agreement between the experimental and calculated results. The effect of anchoring group number on the molecular structures, absorption spectra and photovoltaic performance were comparatively discussed.

Keywords Absorption spectra · Density functional theory · Dye-sensitized solar cells · Molecular structures · Triphenylamine dyes

Introduction

Nanocrystalline dye-sensitized solar cells (DSSCs) have attracted much attention as promising alternatives for the photovoltaic conversion of solar energy owing to their potentially low fabrication costs, environmentally friendly components, and relatively high conversion efficiencies [1–6]. The DSSCs typically contain four components: a mesoporous semiconductor metal oxide film (typically TiO_2 nanoparticles); a sensitizer (dye); an electrolyte/hole transporter; and a counter electrode. In these components, the sensitizer is one of the key components for high solar-to-electric power conversion efficiencies. The most successful charge-transfer sensitizers used in DSSCs are ruthenium (Ru) polypyridyl complexes, yielding conversion efficiencies up to 11-12% under air mass (AM) 1.5 irradiation [6]. However, the Ru complexes are facing the problem of costs and environmental issues, which will limit the large-scale application of DSSCs [7]. In addition to the Ru complexes, metal-free organic dyes as sensitizers are also under intensive investigation due to their high molar extinction coefficients, flexible structural modifications and low costs; and so far some of them have reached good efficiencies [8–15].

In DSSCs, light-harvesting dyes first absorb visible or near infrared solar radiation, accompanied by the excitation of electrons to the excited states. The excited electrons subsequently are injected into the conduction band (CB) of the semiconductor, and then transported toward the counter electrode by electron diffusion through the disordered network of TiO_2. The oxidized dye molecules are regenerated by the iodide redox couple or hole-transporter with the positive charge being transported from the electrolyte to the platinum counter electrode. Therefore, the performance of DSSC strongly depends upon the following factors: (1) Absorption efficiency of the dye sensitizer for solar light spectrum; (2) Electron injection probability from the excited state of the dye sensitizer to TiO_2 (Efficiency of the charge separation); (3) Electron transfer probability from the electrolyte to the oxidized dye [3]. All these factors are closely associated with the ground and excited

J. Xu · L. Zhu · L. Wang · L. Liu · Z. Bai · L. Wang · W. Xu
College of Materials Science & Engineering,
Wuhan Textile University,
430073, Wuhan, China

J. Xu (✉)
Key Lab of Green Processing & Functional Textiles of New Textile Materials, Ministry of Education,
Wuhan Textile University,
430073, Wuhan, China
e-mail: xujie0@ustc.edu

electronic states of the sensitizer. From this point of view, it is imperative to investigate the electronic structures of both ground and excited states of the dye molecule for understanding the mechanism of the charge separation and transfer, which are the key processes in this type of solar cells. In order to design and synthesize more efficient dyes, it is also necessary to understand the electronic structures of the existing efficient sensitizers.

Donor-acceptor π-conjugated (D-π-A) dyes possessing both electron-donating (D) and electron-accepting (A) groups connected by covalent links (usually π conjugated), displaying broad and intense absorption spectral features, are one of the most promising classes of organic sensitizers. The photoabsorption properties of a D-π-A dye are associated with intramolecular charge transfer (ICT) excitation from the donor to the acceptor moiety of the dye, resulting in efficient electron transfer through the acceptor moiety (such as carboxyl or hydroxyl) from the excited dye into the semiconductor CB. The charge transfer or separation between the electron donor and acceptor moieties in the excited dye may facilitate rapid electron injection from the dye molecule into the semiconductor CB, so that it would be expected to separate the cationic charge effectively from the semiconductor surface and to restrict recombination between the photoelectron (the injected electron) and the oxidized dye sensitizer efficiently. Triphenylamine (TPA) has widely been used as an electron donor for organic sensitizers due to its excellent electron-donating capability and aggregation resistant nonplanar molecular configuration [16]. Aggregation can give rise to self-quenching, instability of the sensitizer, and reduce the electron injection efficiency, resulting in low conversion efficiency of the DSSCs. Recently, Shang et al. [17] synthesized three TPA-based dyes comprising different number of anchoring groups (S1-S3, as shown in Fig. 1), with power conversion efficiencies up to 7.38% under AM 1.5 irradiation.

In this work, to theoretically understand the effect of anchoring group numbers and sensitized mechanism at a molecular level, the geometrical and electronic structures of S1-S3 were studied using density functional theory (DFT), and the electronic absorption spectra were investigated based on the time-dependent DFT (TD-DFT) calculations. DFT has emerged as a reliable standard tool for the theoretical treatment of the structures as well as the electronic and absorption spectra. Its time-dependent extension called TD-DFT can give reliable values for valence excitation energies with the standard exchange-correlation functionals. The computational cost of TD-DFT calculation is comparative to that of a Hartree–Fock based single excitation theory, such as, configuration interaction singles (CIS) or time-dependent Hartree–Fock (TD-HF) method and maintains a uniform accuracy for open-shell and closed-shell systems. DFT has been extensively used to study the structures and absorption spectra of sensitizers for DSSCs [4, 18–34].

Computational method

All calculations were performed with the Gaussian 03 program package [35]. The ground-state geometries were fully optimized without any symmetry constrains at the DFT level of theory with Becke's [36] three parameters hybrid functional and Lee, Yang and Parr's correlational functional B3LYP [37] using a standard 6-31 g(d) basis set on all atoms. A full natural bond orbital (NBO) analysis was obtained by using the POP=NBO keyword, along with a second-order perturbation theory (SOPT) analysis. The excitation energies and oscillator strengths for the lowest 30 singlet-singlet transitions at the optimized geometry in the ground state were obtained by TD-DFT calculations using the same basis set as for the ground state and three kinds of hybrid functional PBE1PBE, MPW1B95 and BHandHLYP, respectively. From the calculated results, the UV-vis absorption spectra were simulated by means of the SWizard program (Revision 4.6) [38] using a Gaussian convolution with the full width at half-maximum of 3000 cm^{-1}. Solvation effects were introduced by the SCRF method, via the conductor polarizable continuum model (CPCM) [39, 40] implemented in the Gaussian program, for both geometry optimizations and TD-DFT calculations.

Results and discussion

Geometries

The sensitizers S1-S3 in this study comprise a triphenylamine donor, thiophene conjugated bridge and different number of cyanoacrylic acid anchoring groups. The optimized ground-state geometries of S1-S3 are shown in Fig. 2, and the selected bond lengths, bond angles and dihedral angles are listed in Table 1. Most of the corresponding parameters are in good agreement with the calculated results for the TPA1 dye (without two butoxy groups compared to S1) [33], indicating the reasonability of the present results. All C–C lengths in the thiophene and phenyl rings are between the distance of a single bonded C-C and a double bonded C=C, implying that there exists extensive delocalization throughout the molecule. The calculated distance between the C atom in carboxyl and the N atom in aniline is about 10.26Å for S1-S3, indicating that the distance between the electron donor and semiconductor surface is practically identical for these dyes.

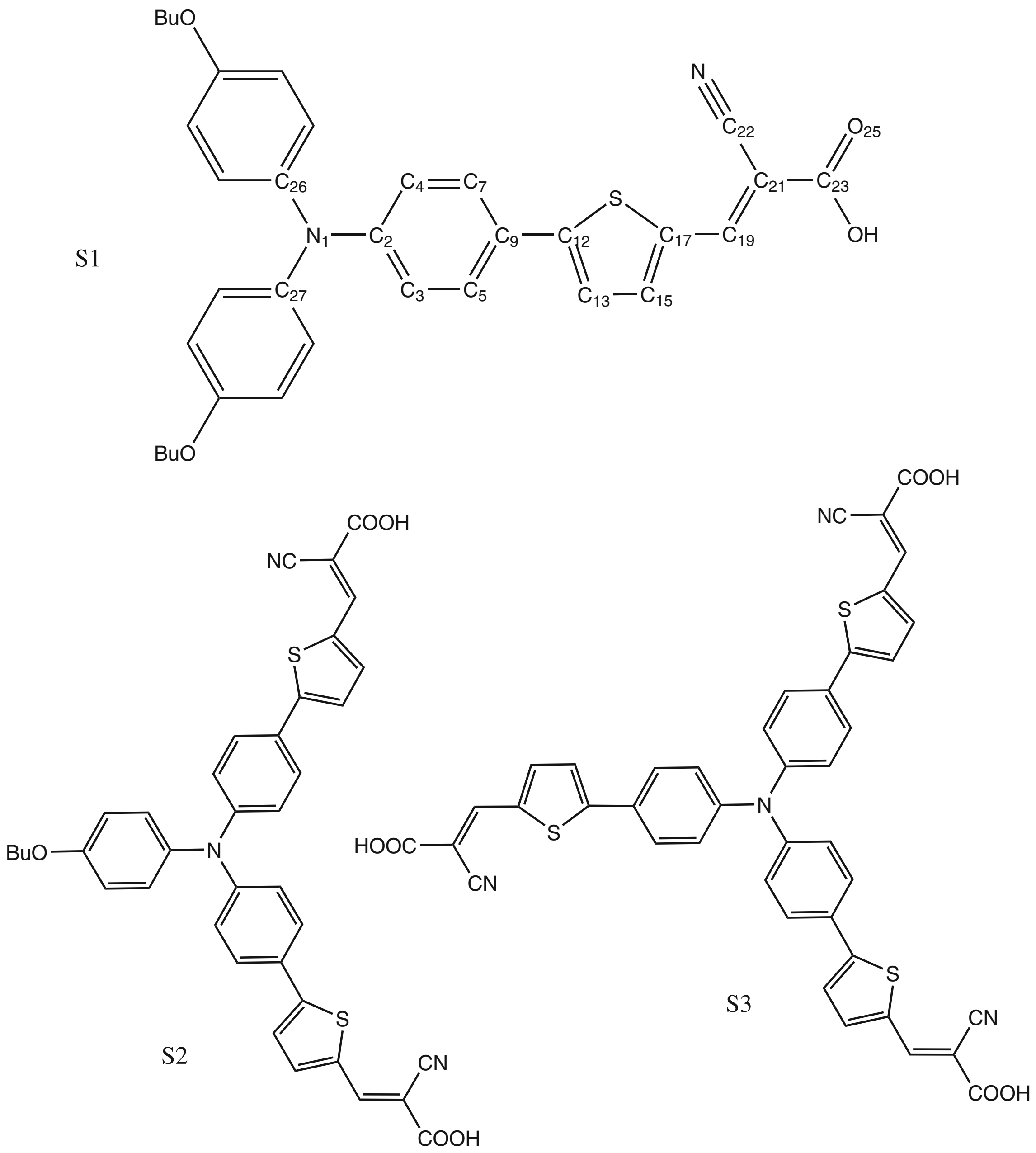
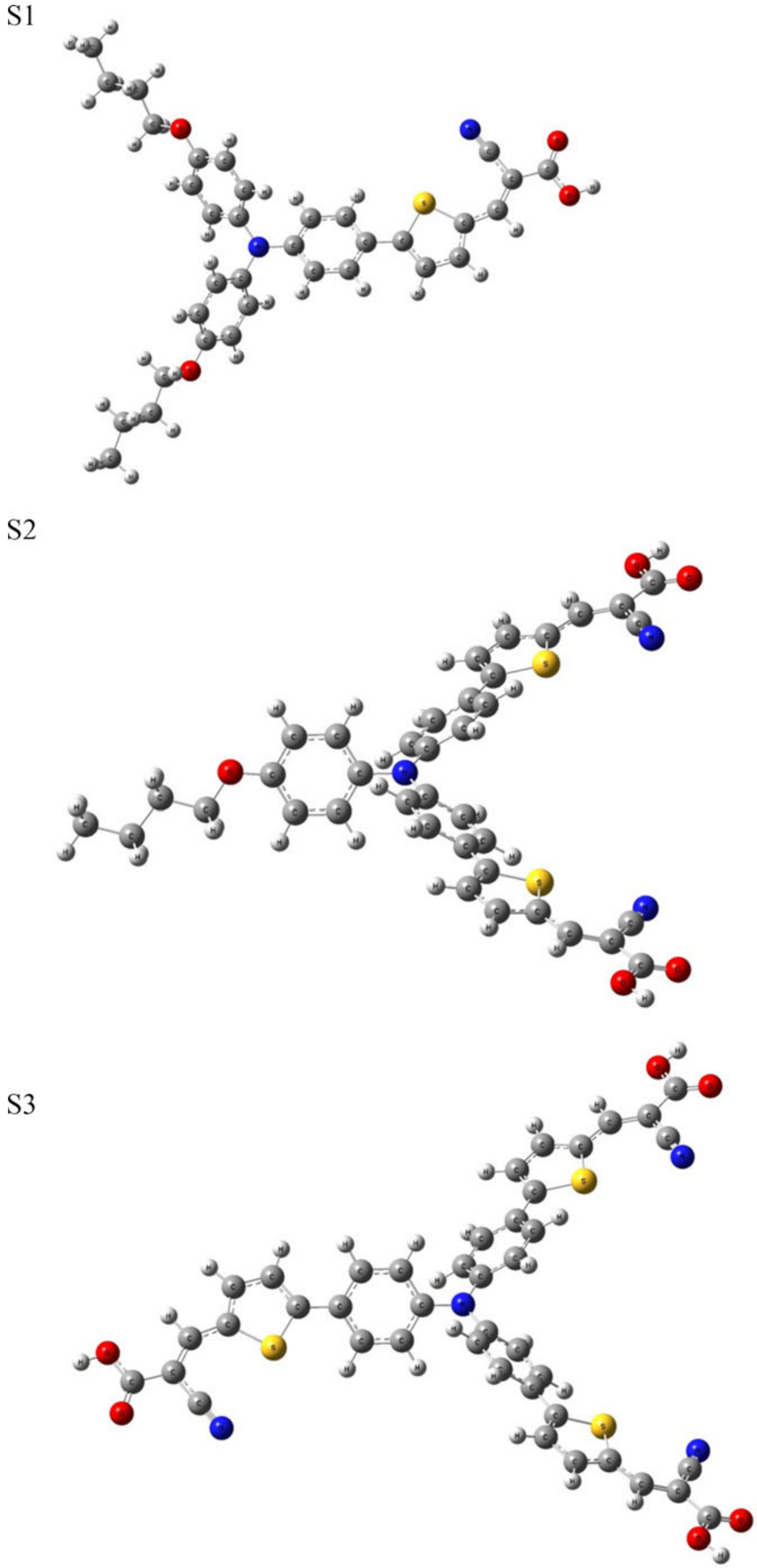


Fig. 1 Molecular structures of the three organic dyes S1-S3

The cyanoacrylic acid group (acceptor) is found to be fully coplanar with the thiophene bridge, as represented by the C15-C17-C19-C21 dihedral angle; while the coplanarity between the diphenylaniline group (donor) and the thiophene bridge is destroyed by about 20° due to steric repulsion between the diphenylaniline and thiophene groups, as shown by the C7-C9-C12-C13 angle. The distortion is increased slightly with the increased anchoring group number from S1 to S3. Thus, it can be concluded that the donor and acceptor moieties in these dyes are fully conjugated through the π-bridge. The delocalization in the conjugate bridge is beneficial to the intramolecular charge transfer and to the stability of the molecule. The values of C26-N1-C2-C4 and C27-N1-C2-C3 angles are increased gradually from S1 to S3, indicating that the anilino groups are distorted significantly in response to the increased anchoring group number. The anchoring carboxyl group (−COOH) is coplanar with the π extended system of the molecule, as demonstrated by the negligible C15-C17-C23-O25 values. Therefore, assuming a bidentate bind of the carboxyl group to the supporting semiconductor's surface [41], it could be inferred that the π system of these dyes most probably lay vertically to the surface, thus giving a denser package and coverage. In this way, the third



S1

S2

S3

Fig. 2 Optimized ground-state geometries of S1-S3

Table 1 Selected bond lengths (in angstrom), bond angles (in degree) and dihedral (in degree) of S1-S3

	S1	S2	S3
N1-C2	1.3967	1.4104	1.4186
C2-C4	1.4135	1.4086	1.4055
C4-C7	1.3850	1.3882	1.3882
C9-C12	1.4546	1.4590	1.4613
C12-C13	1.3934	1.3906	1.3892
C13-C15	1.3979	1.4002	1.4012
C17-C19	1.4215	1.4248	1.4265
C19-C21	1.3824	1.3797	1.3782
C26-N1-C2	120.8	119.2	120.0
C7-C9-C12	120.8	122.0	121.9
C9-C12-C13	127.9	127.9	127.7
C15-C17-C19	120.3	120.2	120.2
C17-C19-C21	137.3	137.2	137.2
C26-N1-C2-C4	−25.1	−34.0	−41.1
C27-N1-C2-C3	−25.4	−36.3	−41.7
C7-C9-C12-C13	−162.5	−160.3	−159.3
C15-C17-C19-C21	179.9	179.8	179.9
C15-C17-C21-C22	0.2	0.0	−0.1
C15-C17-C23-O25	3.1	1.1	0.4

cyanoacrylic acid anchoring group in S3 could not bind with the surface.

NBO analysis

NBO analysis was performed to characterize the intramolecular charge transfer in S1-S3. Table 2 shows the natural charges of cyanoacrylic acid, thiophene, diphenylaniline and butoxy groups in S1-S3. There are some charges transferred from the electron donor (diphenylaniline) to the electron acceptor (cyanoacrylic acid) through the chemical bonds of the π-bridge for these dyes. The charge of the donor shows a remarkable decrease with the increase of anchoring group number, indicating that the introduction of more anchoring groups reduces the electron-donating capability of the donor. The electron-drawing strength of each cyanoacrylic acid is decreased from S1 to S2 and S3, as revealed by the decreased charge values of this group.

Table 2 Natural charges (e) of different groups in S1-S3

Dyes	Cyanoacrylic acid	Thiophene	Diphenylaniline	BuO
S1	−0.179	0.081	0.498	−0.200 (−0.200)
S2	−0.152 (−0.152)	0.091 (0.090)	0.315	−0.193
S3	−0.136 (−0.136, -0.137)	0.097 (0.096, 0.097)	0.118	

In order to have a better insight into the nature and strength of the intramolecular resonance between the different parts of the molecule, particularly acceptor and donor moieties, the SOPT analysis of the Fock matrix within the NBO basis was performed. In the NBO procedure, the role of electronic delocalization can be quantitatively evaluated. The directly estimated approach of π-conjugative stable energies using NBO second order perturbation analysis will be very helpful to analyze the π-conjugation strength of S1-S3. The π-conjugation stabilization strength was evaluated by the NBO donor-acceptor interaction energies which are calculated on the basis of Lewis- and Pauling-like localized structural and hybridization theories and are presented with the classical π-conjugation concepts by a refinement of NBO analysis. This interaction energy corresponds to the charge delocalization due to the loss of electronic occupation from the localized Lewis molecular orbital to the non- Lewis molecular orbital leading to the distribution of electronic charge and therefore the perturbation from idealized Lewis structure description. For a characteristic conjugated π-bond network with two pairs of conjugated π bonds, the delocalized molecular orbitals can be pictured using the refined idealized Lewis structures by NBO donor-acceptor interaction of $\pi_a \rightarrow \pi_b$* and $\pi_a \rightarrow \pi_b$*. According to the perturbation theory, the lowering energy due to $\pi_a \rightarrow \pi_b$* interaction, which is also referred to as the quantum mechanical resonance energy (denoted as QMRE), is estimated as [42]:

$$\Delta E^{(2)}_{\psi \mathrm{donor} \rightarrow \psi \mathrm{acceptor}} \approx -2 \cdot \frac{\mathrm{F(acc,don)}^2}{\mathrm{E}_{acc} - \mathrm{E}_{don}}, \qquad (1)$$

where $\mathrm{E}_{acc} - \mathrm{E}_{don}$ is the energy difference of interacting NBOs and the matrix element and $\mathrm{F(acc,don)}$ is the off-diagonal element associated with the NBO Fock matrix. The strength of π-type conjugation and its variations by introducing resonance moieties can be conveniently visualized in terms of the NBO second order perturbation stabilization energies $\left(\Delta E^{(2)}_{\psi donor \rightarrow \psi acceptor}\right)$ and the charge transfer from π_a to π_b*. The sum of stabilization energies is chosen as an indicator of the degree of the π-conjugation. As mentioned above, to prove the reliability of QMRE as a criterion of π-conjugative strength, we studied the change of π-conjugation due to the introduction of conjugated functional groups and changes in physical properties. According to the NBO donor-acceptor interaction theory, the charge occupancy of the π^* NBO also indicates the strength of π-conjugation. The quantities of transferred charge from a given donor to a given acceptor orbital may be estimated again using elementary perturbation theory arguments, leading to the following approximate formula:

$$q_{\text{donor}\rightarrow\text{acceptor}} \approx 2\left(\frac{\mathrm{F(acc, don)}}{\mathrm{E}_{acc} - \mathrm{E}_{don}}\right)^2 \qquad (2)$$

All the NBO parameters from the SOPT analysis are summarized in Table 3. The charge transfer in S1-S3 is of chemical significance and essentially one directional, from the electron-donating to the electron-accepting moiety. The NBO parameters from the donors to the acceptors decrease with the increase of anchoring group number, further confirming that the introduction of more anchoring groups results in a decrease both in the donor and acceptor strength. The QMRE sums corresponding to second order perturbation energy for S1, S2 and S3 are 4770, 7596 and 9896 kcal mol^{-1}, which drastically rise with the π-conjugation increase from S1 to S3. As it can be easily observed, as a result of the introduction of more anchoring groups there is an increased resonance and stabilization in the molecule, verified by QMRE corresponding to second order perturbation.

Electronic structures

The dipole moments for S1, S2 and S3 are 18.98, 22.48 and 9.19 Debye, respectively. The polarity of S3 is much lower than that of S1 and S2, due to its centrosymmetry. The values of quadrupole moments for S1-S3 are listed in Table 4, where the average of the diagonal quadrupole moment tensor elements Q_{ii} and unique quadrupole moment Q are defined as follows:

$$Q_{ii} = (Q_{XX} + Q_{YY} + Q_{ZZ})/3$$

$$\mathrm{Q} = \mathrm{Q_{XX}} - \mathrm{Q_{YY}}$$

All the diagonal elements of the quadrupole moment tensor for S1-S3 are negative, indicating that the negative charge distribution is farther removed from the molecular center of the nuclear charges. The off-diagonal tensor elements Q_{ij} vanish whenever the molecule has a plane of symmetry perpendicular to either one of the coordinates i or j. The values of the off-diagonal elements Q_{XY} and Q_{XZ} of S1 are relatively lower, which can be attributed to its symmetric plane nearly perpendicular to the x-axis; while the symmetric plane of S2 is nearly perpendicular to the z-axis, as demonstrated by the low Q_{XZ} and Q_{YZ} values.

The frontier molecular orbital (MO) contribution is very important in determining the charge-separated states of sensitizers. To create an efficient charge-separated state, the highest occupied MO (HOMO) must be localized on the extended donor moiety and the lowest unoccupied MO (LUMO) on the acceptor moiety. The MO energies and isodensity plots of S1-S3 are shown in Figs. 3 and 4, respectively. For S1, the HOMO, lying at −4.90 eV, is dominated by a delocalized π orbital contribution over the cyanoacrylic acid group through the diphenylaniline group whereas the HOMO-1, lying 1.06 eV below the HOMO, is also delocalized over the entire molecule. The LUMO, lying at −2.50 eV, is a π* orbital localized from the cyanoacrylic acid to the nitrogen atom. The LUMO+1, lying 1.80 eV above the LUMO, is also a π* orbital localized in the cyanoacrylic acid, thiophene, and diphenylaniline groups. For S2, the HOMO, lying at −5.14 eV, is a π orbital delocalized over the entire molecule except the butyl group. The HOMO-1, lying 0.90 eV below the HOMO, is localized in the cyanoacrylic acid, thiophene, and phenyl groups connected to the bridge. The LUMO and LUMO+1, lying at −2.72 and −2.53 eV, respectively, are π* orbitals localized in the cyanoacrylic acid, thiophene, and phenyl groups connected to the bridge. As to S3, the

Table 3 Conjugative interaction energies (in kcal mol^{-1}) between the π and π* orbitals in S1-S3 from the second-order perturbation theory analysis of the Fock matrix within the NBO analysis

Dye	Donor orbital	Acceptor orbital	$\Delta E^{(2)}$/kcal mol^{-1}	$\mathrm{E_{acc}}$-$\mathrm{E_{don}}$/a.u.	F(acc, don)/a.u.	$q_{\text{donor}\rightarrow\text{acceptor}}$/e
S1	π(C2=C4)	π*(C7=C9)	61.42	0.46	0.151	0.216
	π(C7=C9)	π*(C12=C13)	28.97	0.46	0.105	0.104
	π(C12=C13)	π*(C15=C17)	49.89	0.49	0.143	0.170
	π(C15=C17)	π*(C19=C21)	49.08	0.53	0.146	0.152
S2	π(C2=C4)	π*(C7=C9)	54.27	0.47	0.144	0.188
	π(C7=C9)	π*(C12=C13)	25.13	0.47	0.098	0.087
	π(C12=C13)	π*(C15=C17)	46.55	0.50	0.140	0.157
	π(C15=C17)	π*(C19=C21)	45.91	0.54	0.142	0.138
S3	π(C2=C4)	π*(C7=C9)	49.86	0.48	0.139	0.168
	π(C7=C9)	π*(C12=C13)	23.22	0.47	0.095	0.082
	π(C12=C13)	π*(C15=C17)	44.57	0.50	0.137	0.150
	π(C15=C17)	π*(C19=C21)	44.06	0.54	0.139	0.133

Table 4 Quadrupole moments (in Debye·Å) of S1-S3

Dyes	Q_{XX}	Q_{YY}	Q_{ZZ}	Q_{XY}	Q_{XZ}	Q_{YZ}	Q_{ii}	Q
S1	−214.3	−353.3	−388.3	−10.9	−17.7	137.0	−318.6	139.0
S2	−396.9	−375.5	−473.5	−103.7	−1.7	−16.3	−415.3	−21.4
S3	−596.3	−454.2	−505.4	−186.4	70.5	13.1	−518.6	−142.1

HOMO, lying at −5.31 eV, is delocalized over the entire molecule; while the LUMO, lying at −2.78 eV, is a π* orbital localized in two of the three cyanoacrylic acid, thiophene and phenyl groups, indicating that the third anchoring group does not work during the HOMO-LUMO excitation. The LUMO+1, also lying at −2.78 eV, is π* with major contributions from all the cyanoacrylic acid, thiophene, and partial phenyl groups.

Apparently, in S1 and S2, the HOMO is quite delocalized on the extended donor moiety, while the LUMO is essentially centered on the respective acceptor groups. Thus, the HOMO-LUMO excitation on S1 and S2 induced by light irradiation could move the electron distribution from the donor moiety to the anchoring/acceptor moiety. This orientational spatial separation of HOMO and LUMO is ideal for DSSCs, as it facilitates rapid interfacial electron injection from the excited dyes to the TiO_2 conduction band and slows down the recombination of injected electrons in TiO_2 with oxidized sensitizers due to their remoteness. However, for S3, the electron density movement from the donor group to the acceptor groups is also significant during the HOMO-LUMO+1 transition. Both the HOMO and LUMO energies are computed to decrease with the increase of anchoring group number, pointing out that the introduction of more anchoring groups decreases the donor and acceptor strength (in agreement with the previously discussed natural charge results). The isodensity plots of LUMO exhibit a charge transfer to the anchoring acid group from S1 to S2, indicating that the introduction of more anchoring groups results in an increased contribution of the acid group to the LUMO. As a consequence of the variation of MOs energies, the HOMO-LUMO gap increases when going from S1 (2.39 eV) to S2 (2.42 eV) and S3 (2.53 eV), the latter molecules being expected to absorb at higher energies.

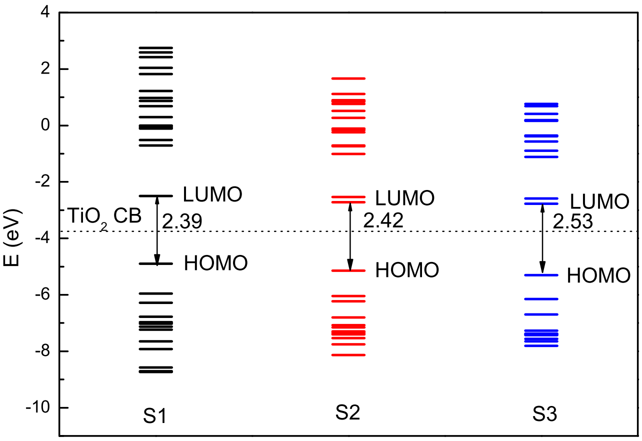


Fig. 3 Frontier molecular orbital energies of S1-S3 together with the TiO_2 anatase (101) conduction band

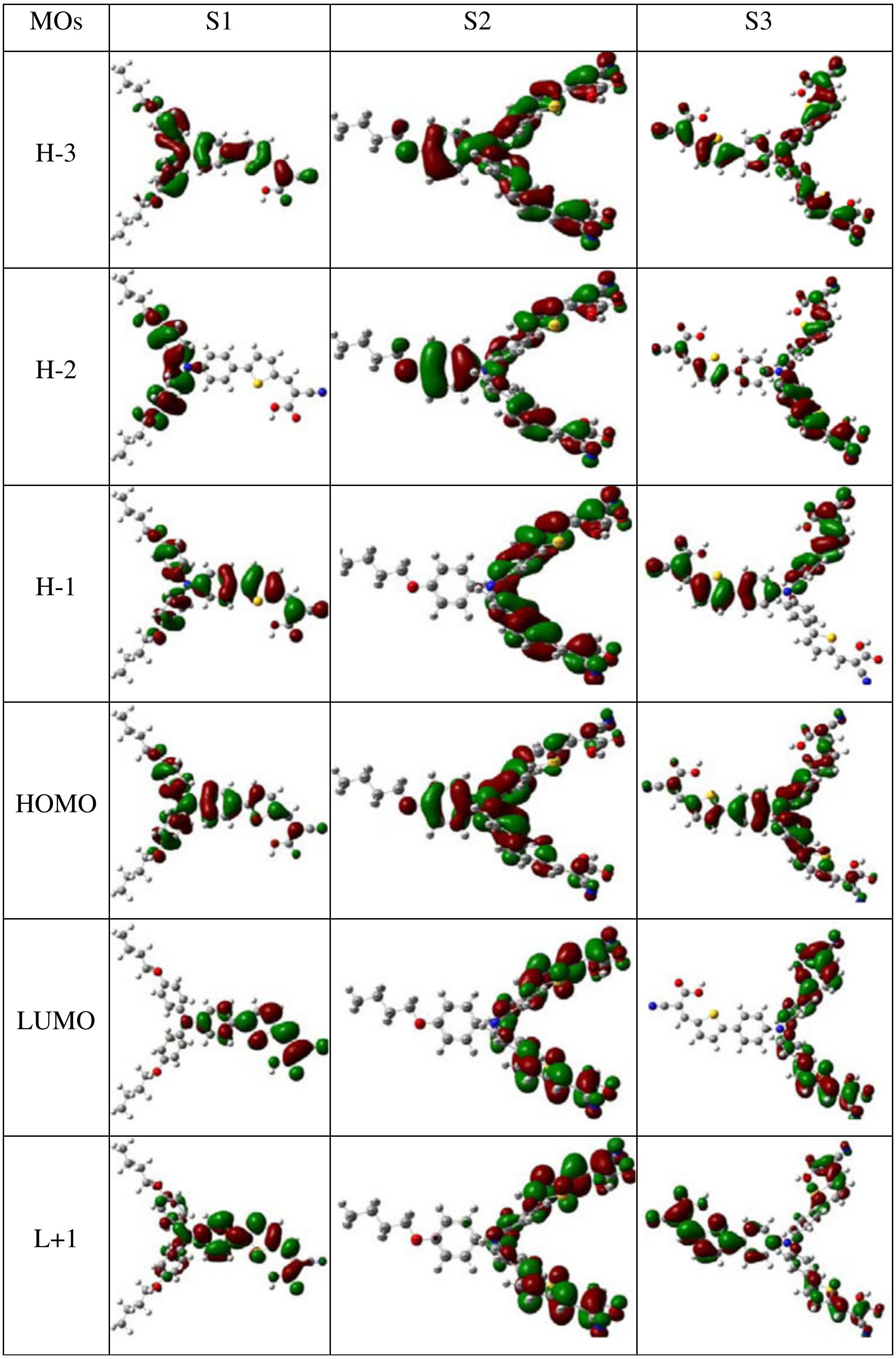
The calculated HOMO and LUMO energies of the bare $Ti_{38}O_{76}$ cluster as a model for nanocrystalline are −6.55 and −2.77 eV, respectively, resulting in a HOMO-LUMO gap of 3.78 eV; while the lowest transition is reduced to 3.20 eV according to TD-DFT, which is slightly smaller than the typical band gap of TiO_2 nanoparticles [4]. Furthermore, the HOMO, LUMO and HOMO-LUMO gap of $(TiO_2)_{60}$ cluster is −7.52, -2.97, and 4.55 eV (B3LYP/VDZ), respectively [43]. In addition, the edges of the valence band (VB) and CB of a TiO_2 anatase (101) surface are computed at −8.70 and −3.74 eV [44]. Usually an energy gap of more than 0.2 eV between the LUMO of the dye and the CB of the TiO_2 is necessary for effective electron injection [25]. Taking the above data into account, it can be found that the HOMO energies of these dyes lie above the VB of TiO_2 and the LUMO energies lie above the CB of TiO_2. The above data also reveal the sensitized mechanism: the interfacial electron transfer between the TiO_2 electrode and the dye sensitizers are electron injection processes from the excited dyes to the semiconductor CB. This is a kind of typical interfacial electron transfer reaction [45]. Relatively large energy gaps between the LUMO energies of these dyes and the semiconductor CB would be beneficial to the photovoltaic conversion efficiencies.

Absorption spectra

The UV-vis absorption spectra of S1-S3 were measured in $CHCl_3$ solution, and consist of a very intense and well isolated absorption band at 510 (33,000 M^{-1} cm^{-1}), 506 (46,000 M^{-1} cm^{-1}) and 484 nm (84,000 M^{-1} cm^{-1}) and of less intense bands in the UV region, respectively [17]. The absorption bands are blue-shifted and broadened when going from S1 to S2 and S3, which is attributed to the decreased donor and acceptor strength with the increased anchoring group number.

TD-DFT calculations in $CHCl_3$ solution were performed with three kinds of hybrid functional PEB1PBE, MPW1B95 and BHandHLYP based on the optimized ground-state geometries, taking the 30 lowest spin-

Fig. 4 Isodensity plots (isodensity contour=0.02 a.u.) of the frontier orbitals of S1-S3



allowed singlet-singlet transitions into account. The simulated UV-vis absorption spectra of S1 using the hybrid functional PEB1PBE, MPW1B95 and BHandHLYP are shown in Fig. 5 as representative. The hybrid functional MPW1B95 is more suitable than PBE1PBE and BHandHLYP for calculating absorption spectra of these dyes. The calculated vertical excitation energies and oscillator strengths along with the main excitation configurations of these dyes calculated by MPW1B95 are listed in Table 5. The simulated absorption spectra by MPW1B95 are shown in Fig. 6. The calculated line shapes and relative strengths are in satisfactory agreement with those of the experiment, and the overall spectral blue-shift when going from S1 to S2 and S3 is also correctly reproduced. The first band of S1

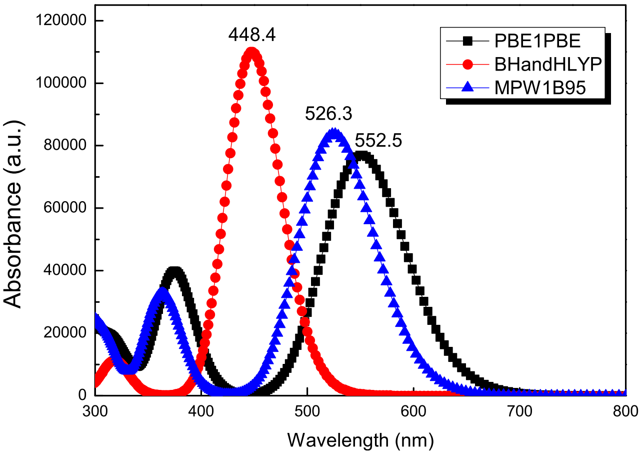


Fig. 5 Simulated absorption spectra of S1 using the hybrid functional PBE1PBE, BHandHLYP and MPW1B95

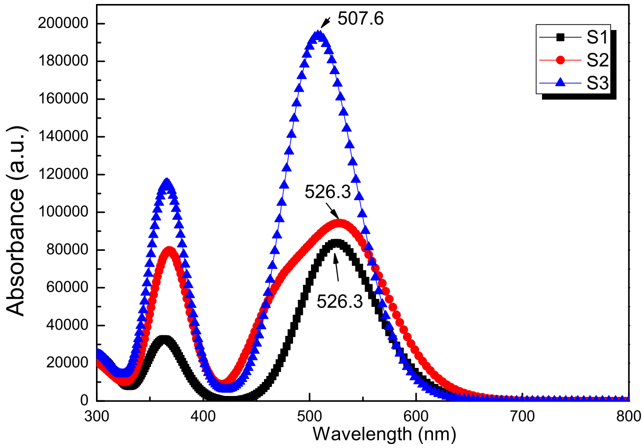


Fig. 6 Simulated absorption spectra of S1-S3 at the MPW1B95/6-31 g(d) level

and S2 clearly corresponds to the previously described HOMO-LUMO transition thus possessing high transition intensity, and is of charge transfer character as demonstrated by the electron distribution differences between the HOMO and LUMO levels in Fig. 4. Since the HOMO and the LUMO are of the π and π* type, the HOMO-LUMO transition can be classified as a π-π* intramolecular charge transfer. The second and less intense band of S1 and S2 also corresponds to a π-π* transition with a strong HOMO-1 to LUMO character. For S3, the first band is considered as the

Table 5 Electronic transition configurations, computed excitation energies and oscillator strengths (*f*) for the optical transitions with *f*>0.05 of the absorption bands in visible and near-UV region for S1-S3 in $CHCl_3$ at MPW1B95/6-31 g(d) level (H=HOMO, L=LUMO, L+1=LUMO+1, etc.)

Dye	Configuration	Excitation energy (eV/nm)	*f*	Assign.	Exp. (nm)
S1	H→L (+89%)	2.36/525.1	1.1564	π→π*	510
	H-1→L (+86%)	3.41/363.4	0.4525		
	H→L+1 (+78%); H→L+2 (+8%)	4.04/306.6	0.2028		
	H→L+3 (+35%); H-7→L (28%); H-3→L (+16%); H→L+5 (+5%)	4.37/283.5	0.2432		
	H→L+3 (+54%); H-7→L (+20%); H-4→L (+6%); H-3→L (6%)	4.39/282.6	0.1026		
	H→L+5 (+61%); H-4→L (+24%)	4.57/271.5	0.1343		
S2	H→L (+90%)	2.31/536.8	1.1714	π→π*	506
	H→L+1 (+91%)	2.62/472.9	0.7348		
	H-1→L (+82%)	3.34/371.3	0.8252		
	H-1→L +1(+53%); H-2→L (31%)	3.45/359.3	0.3306		
	H→L+2 (+54%); H→L+3 (16%); H→L+4 (+15%)	3.91/316.9	0.0668		
	H→L+2 (+32%); H→L+3 (+28%); H→L+4 (23%)	4.05/305.8	0.0906		
	H-6→L (+31%); H-5→L+1 (14%); H→L+4 (+10%); H-7→L+1 (+10%); H-4→L (+8%); H→L+3 (+8%)	4.21/294.8	0.0646		
	H→L+4 (+33%); H→L+3 (+28%); H-4→L (6%)	4.26/291.1	0.0887		
S3	H→L (+55%); H→L+1 (35%)	2.44/509.0	0.9784	π→π*	484
	H→L+1 (+55%); H→L (+35%)	2.44/507.8	1.6938		
	H-2→L (+32%); H-1→L+1 (+28%); H-1→L (17%); H-2→L+1 (+5%)	3.37/367.7	0.2246		
	H-2→L (+29%); H-1→L+1 (23%); H-2→L+1 (+11%); H-2→L+2 (11%); H-1→L (+9%)	3.39/365.7	1.0171		
	H-2→L+1 (+29%); H-1→L (+21%); H-1→L+1 (+16%); H-1→L+2 (+13%); H-2→L (6%)	3.40/365.1	0.3470		
	H-2→L+2 (+74%); H-1→L+1 (5%)	3.76/329.6	0.0675		
	H-3→L (+54%); H→L+3 (22%); H→L+4 (15%)	4.08/303.6	0.0892		
	H-3→L+1 (+38%); H→L+4 (+26%); H→L+3 (19%); H→L+5 (7%)	4.09/303.2	0.0828		

combination of HOMO to LUMO and LUMO+1 transition, which probably results from the same energy level of LUMO and LUMO+1.

The solar-to-electric conversion efficiency (η) of the DSSCs is calculated from short-circuit current (J_{SC}), the open-circuit photovoltage (V_{OC}), the fill factor (FF) and the intensity of the incident light (Pin) according to the following equation [46]:

$$\eta = \frac{[J_{sc}(mAcm^{-2})][V_{oc}(V)][FF]}{P_{in}(mWcm^{-2})} \tag{3}$$

The experimental J_{SC} of these dyes is in the following order: S1<S2<S3. Generally, the J_{SC} is determined by two processes: one is the rate of electron injection from the excited dyes to the semiconductor CB, and the other is the rate of redox between the excited dyes and electrolyte. The latter one is very complex and is not taken into account in the present calculations. On the basis of the analysis of excitation energies, electronic transition configurations, oscillator strengths and molecular orbitals of S1-S3 in UV-vis region, it can be found that the sensitizer with more anchoring groups have larger oscillator strengths for the most excited states with intramolecular charge transfer character, giving rise to the larger light-induced transition probability, the higher total electron injection rate and thus the larger J_{SC}. The experimental V_{OC} of these dyes decreases in order of S1>S2>S3. According to the sensitized mechanism (electron injected from the excited dyes to the semiconductor CB), the energy gap between the LUMO of the dyes and the CB edge of semiconductor E_{LUMO}- E_{CB} is denoted as the driving force of the electron injection, and larger driving force are desirable for higher V_{OC} [47, 48]. The sensitizer with less anchoring groups possesses higher E_{LUMO}, thus leading to larger V_{OC}.

Conclusions

A DFT study on the geometrical and electric structures of three TPA-based dyes with different anchoring group number (S1-S3) used for DSSC has been performed. The calculated geometries indicate that strong conjugation is formed in the dyes, which is of benefit to the intramolecular charge transfer. The NBO results suggest that there are some charges transferred from the electron donor (diphenylaniline group) to the electron acceptor (cyanoacrylic acid) through the chemical bonds of the conjugate thiophene bridge in the dyes. The introduction of more anchoring groups leads to a decrease both in the donor and acceptor strength, but an increased resonance and stabilization in the molecules. The HOMO energy levels are computed to be −4.90, -5.14 and −5.31 eV, while the LUMOs are −2.50, -2.72 and −2.78 eV for S1, S2 and S3, respectively, pointing out that the electron transfer from the excited dyes to the TiO_2 conduction band is possible. The UV-vis absorption spectra of the dyes have been simulated by TD-DFT calculations. The absorption bands of the dyes are assigned to π→π* transitions according to the qualitative agreement between the experimental and calculated results. The comparative analysis of electronic structures, spectra and photovoltaic properties point out that the larger J_{SC} of the sensitizer with more anchoring groups may be determined by the larger oscillator strengths for the most excited states with intramolecular charge transfer character; while the larger V_{OC} of the sensitizer with less anchoring groups may be deduced by the higher E_{LUMO}. Therefore, the TPA-based dyes with good performance for DSSC are required for good absorption properties with intramolecular charge transfer character in UV-vis region and higher E_{LUMO}. The incorporation of appropriate numbers of acceptors in the TPA structure is very important for the molecular design of new TPA-based dyes with improved performance.

Acknowledgments This work was supported by the Natural Science Foundation of China (No.51003082), and the Educational Commission of Hubei Province (Q20101606). The authors gratefully wish to express their thanks to the reviewers for critically reviewing the manuscript and making important suggestions.

References

1. O'Regan B, Grätzel M (1991) Nature 353:737–740
2. Nazeeruddin MK, Kay A, Rodicio, Humpbry-Baker R, Miiller E, Liska P, Vlachopoulos N, Grätzel M (1993) J Am Chem Soc 115:6382–6390
3. Nazeeruddin MK, Péchy P, Renouard T, Zakeeruddin SM, Humphry-Baker R, Comte P, Liska P, Cevey L, Costa E, Shklover V, Spiccia L, Deacon GB, Bignozzi CA, Grätzel M (2001) J Am Chem Soc 123:1613–1624
4. Nazeeruddin MK, Angelis FD, Fantacci S, Selloni A, Viscardi G, Liska P, Ito S, Takeru B, Grätzel M (2005) J Am Chem Soc 127:16835–16847
5. Gao F, Wang Y, Shi D, Zhang J, Wang M, Jing X, Humphry-Baker R, Wang P, Zakeeruddin SM, Grätzel M (2008) J Am Chem Soc 130:10720–10728
6. Chen CY, Wang M, Li JY, Pootrakulchote N, Alibabaei L, Ngoc-le CH, Decoppet JD, Tsai JH, Grätzel C, Wu CG, Zakeeruddin SM, Grätzel M (2009) ACS Nano 3:3103–3109
7. Amao Y, Komori T (2004) Biosensors Bioelectron 19:843–847
8. Horiuchi T, Miura H, Sumioka K, Uchida S (2004) J Am Chem Soc 126:12218–12219
9. Hara K, Kurashige M, Dan-oh Y, Kasada C, Shinpo A, Suga S, Sayama K, Arakawa H (2003) New J Chem 27:783–785
10. Horiuchi T, Miura H, Uchida S (2003) Chem Commun 3036–3037
11. Horiuchi T, Miura H, Uchida S (2004) J Photochem Photobiol A Chem 164:29–32
12. Kitamura T, Ikeda M, Shigaki K, Inoue T, Anderson NA, Ai X, Lian T, Yanagida S (2004) Chem Mater 16:1806–1812
13. Hagberg DP, Yum J-H, Lee H, Angelis FD, Marinado T, Karlsson KM, Humphry-Baker R, Sun L, Hagfeldt A, Grätzel M, Nazeeruddin MK (2008) J Am Chem Soc 130:6259–6266
14. Tian H, Yang X, Chen R, Zhang R, Hagfeldt A, Sun L (2008) J Phys Chem C 112:11023–11033

15. Mishra A, Fischer MKR, Bäuerle P (2009) Angew Chem Int Edn 48:2474–2499
16. Ning Z, Tian H (2009) Chem Commun 5483–5495
17. Shang H, Luo Y, Guo X, Huang X, Zhan X, Jiang K, Meng Q (2010) Dyes Pigm 87:249–256
18. Barolo C, Nazeeruddin MK, Fantacci S, Di Censo D, Comte P, Liska P, Viscardi G, Quagliotto P, De Angelis F, Ito S, Gratzel M (2006) Inorg Chem 45:4642–4653
19. Onozawa-Komatsuzaki N, Kitao O, Yanagida M, Himeda Y, Sugihara H, Kasuga K (2006) New J Chem 30:689–697
20. Monat JE, Rodriguez JH, McCusker JK (2002) J Phys Chem A 106:7399–7406
21. Fantacci S, De Angelis F, Selloni A (2003) J Am Chem Soc 125:4381–4387
22. Angelis FD, Fantacci S, Selloni A, Nazeeruddin MK (2005) Chem Phys Lett 415:115–120
23. Xu Y, Chen WK, Cao MJ, Liu SH, Li JQ, Philippopoulos AI, Falaras P (2006) Chem Phys 330:204–211
24. Kurashige Y, Nakajima T, Kurashige S, Hirao K, Nishikitani Y (2007) J Phys Chem A 111:5544–5548
25. Hara K, Sato T, Katoh R, Furube A, Ohga Y, Shinpo A, Suga S, Sayama K, Sugihara H, Arakawa H (2003) J Phys Chem B 107:597–606
26. Zhang X, Zhang JJ, Xia YY (2008) J Photochem Photobiol A Chem 194:167–172
27. Liu Z (2008) J Mol Struct Theochem 862:44–48
28. Alexander BD, Dines TJ, Longhurst RW (2008) Chem Phys 352:19–27
29. Lee C, Sohlberg K (2010) Chem Phys 367:7–19
30. Ma R, Guo P, Cui H, Zhang X, Nazeeruddin MK, Grätzel M (2009) J Phys Chem A 113:10119–10124
31. Gao Y, Sun S, Han K (2009) Spectrochim Acta A Mol Biomol Spectrosc 71:2016–2022
32. Kumar PS, Vasudevan K, Prakasam A, Geetha M, Anbarasan PM (2010) Spectrochim Acta A Mol Biomol Spectrosc 77:45–50
33. Xu J, Wang L, Liang G, Bai Z, Wang L, Xu W, Shen X (2011) Spectrochim Acta A Mol Biomol Spectrosc 78:287–293
34. Senthilkumar P, Anbarasan PM (2011) J Mol Model 17:49–58
35. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, M.A. Robb, Cheeseman JR, Montgomery JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima Y, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox IE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian Inc, Wallingford, CT
36. Becke AD (1993) J Chem Phys 98:5648
37. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785–789
38. Gorelsky SI (2010) University of Ottawa, Ottawa, Canada
39. Cossi M, Barone V, Cammi R, Tomasi J (1996) Chem Phys Lett 255:327–335
40. Barone V, Cossi M (1998) J Phys Chem A 102:1995–2001
41. Howie WH, Claeyssens F, Miura H, Peter LM (2008) J Am Chem Soc 130:1367–1375
42. King BF, Weinhold F (1995) J Chem Phys 103:333–347
43. Lundqvist MJ, Nilsing M, Persson P, Lunell S (2006) Int J Quantum Chem 106:3214–3234
44. Bahers TL, Pauporté T, Scalmani G, Adamo C, Ciofini I (2009) Phys Chem Chem Phys 11:11276–11284
45. Watson DF, Meyer GJ (2005) Annu Rev Phys Chem 56:119–156
46. Boschloo G, Hagfeldt A (2005) J Phys Chem B 109:12093–12098
47. Zhang X, Zhang J-J, Xia Y-Y (2007) J Photochem Photobiol A Chem 185:283–288
48. Asbury JB, Hao E, Wang Y, Ghosh HN, Lian T (2001) J Phys Chem B 105:4545–4557

J Mol Model (2012) 18:1779–1790
DOI 10.1007/s00894-011-1200-7

ORIGINAL PAPER

Conformational changes in 2-*trans*-enoyl-ACP (CoA) reductase (InhA) from *M. tuberculosis* induced by an inorganic complex: a molecular dynamics simulation study

André L. P. da Costa · Ivani Pauli · Márcio Dorn · Evelyn K. Schroeder · Chang-Guo Zhan · Osmar Norberto de Souza

Received: 8 July 2010 / Accepted: 20 July 2011 / Published online: 12 August 2011

Abstract InhA, the NADH-dependent 2-*trans*-enoyl-ACP reductase enzyme from *Mycobacterium tuberculosis* (MTB), is involved in the biosynthesis of mycolic acids, the hallmark of mycobacterial cell wall. InhA has been shown to be the primary target of isoniazid (INH), one of the oldest synthetic antitubercular drugs. INH is a prodrug which is biologically activated by the MTB catalase-peroxidase KatG enzyme. The activation reaction promotes the formation of an isonicotinyl-NAD adduct which inhibits the InhA enzyme, resulting in reduction of mycolic acid biosynthesis. As a result of rational drug design efforts to design alternative drugs capable of inhibiting MTB's InhA, the inorganic complex pentacyano(isoniazid)ferrate(II) (PIF) was developed. PIF inhibited both wild-type and INH-resistant Ile21Val mutants of InhA and this inactivation did not require activation by KatG. Since no three-dimensional structure of the InhA-PIF complex is available to confirm the binding mode and to assess the molecular interactions with the protein active site residues, here we report the results of molecular dynamics simulations of PIF interaction with InhA. We found that PIF strongly interacts with InhA and that these interactions lead to macromolecular instabilities reflected in the long time necessary for simulation convergence. These instabilities were mainly due to perturbation of the substrate binding loop, particularly the partial denaturation of helices α6 and α7. We were also able to correlate the changes in the SASAs of Trp residues with the recent spectrofluorimetric investigation of the InhA-PIF complex and confirm their suggestion that the changes in fluorescence are due to InhA conformational changes upon PIF binding. The InhA-PIF association is very strong in the first 20.0 ns, but becomes very week at the end of the simulation, suggesting that the PIF binding mode we simulated may not reflect that of the actual InhA-PIF complex.

Keywords Conformational changes · FAS-II pathway · InhA · Molecular dynamics simulations · *Mycobacterium tuberculosis* · Pentacyano(isoniazid)ferrate(II)

André L. P. da Costa and Ivani Pauli contributed equally to the work.

A. L. P. da Costa · I. Pauli · M. Dorn · E. K. Schroeder ·
O. Norberto de Souza
LABIO - Laboratório de Bioinformática,
Modelagem e Simulação de Biossistemas. PPGCC,
Faculdade de Informática, PUCRS,
Av. Ipiranga, 6681 – Prédio 32, Sala 602,
90619-900, Porto Alegre, RS, Brasil

A. L. P. da Costa · I. Pauli · O. Norberto de Souza
Programa de Pós-Graduação em Biologia Celular e Molecular,
Faculdade de Biociências, PUCRS,
Av. Ipiranga, 6681 – Prédio 12, Bloco A, Sala 204,
90619-00, Porto Alegre, RS, Brasil

I. Pauli · O. Norberto de Souza (✉)
INCT-TB - Instituto Nacional de Ciência
e Tecnologia em Tuberculose,
Av. Ipiranga 6681, Tecnopuc, Prédio 92 A, 90619-900, Partenon,
Porto Alegre, RS, Brasil
e-mail: osmar.norberto@pucrs.br

C.-G. Zhan
College of Pharmacy, University of Kentucky,
741 South Limestone, BBSRB Building 353,
Lexington, KY 40536, USA

Present Address:
M. Dorn
Instituto de Informática,
Universidade Federal do Rio Grande do Sul – UFRGS,
Av. Bento Gonçalves, 9500, Prédio 67, Sala 202,
Porto Alegre, RS, Brasil

Introduction

The 2-*trans*-enoyl-ACP (acyl carrier protein) reductase enzyme (InhA or ENR, EC number: 1.3.1.9) from *Mycobacterium tuberculosis* (MTB) is a member of type II dissociated fatty acid biosynthesis system (FAS-II) in MTB [1–4]. This pathway, consisting of monofunctional enzymes and ACP [1], elongates FAS-I acyl fatty acid precursors to produce the long carbon chains (50–60 carbons) [2, 5] of the meromycolate branch of mycolic acids, the hallmark of mycobacterial cell wall [6]. InhA, a NADH-dependent reductase, has specificity for long-chain substrates (12–24 carbons) [7], consistent with its involvement in mycolic acid biosynthesis [8]. InhA has been shown to be the primary target of isoniazid [9]. Isoniazid (INH, isonicotinic acid hydrazide), one of the oldest synthetic antitubercular drugs [10], is a prodrug [11] which is biologically activated by the MTB catalase-peroxidase KatG enzyme [12]. The activation reaction promotes the formation of an isonicotinyl-NAD adduct which inhibits the InhA enzyme, resulting in reduction of mycolic acid biosynthesis [4, 5]. Drugs such as INH, ethionamide (ETH), and pyrazinamide (PZA) require activation for activity against MTB. Interestingly, both activated forms of INH and ETH target the InhA enzyme, despite their different activation processes [13–15]. Based on the mechanism of activation proposed for INH, via electron transfer reaction [8, 16–18], an alternative to the self-activation route was proposed for the design of new drugs for the treatment of wild-type (WT) and INH-resistant tuberculosis, through the nonenzymatic INH activation method that mimic the isonicotinyl-NAD adduct [19]. Efforts to reproduce this mechanism of activation are in progress and new small molecule compounds have been suggested, for example: SQ109 diamine-based [20], alkyl diphenyl ethers [21], pyrrolidine carboxamide analogues [22], and others [23–25]. These compounds are not expected to be activated prior to interacting with its cellular target. Within this approach Basso, Moreira, Santos and collaborators [26–29] have proposed an INH analog, namely pentacyano (isoniazid)ferrate(II) that contains a cyanoferrate moiety, a metal center and the INH ligand [26]. This class of compounds constitutes a suitable model system for new perspectives of novel drug development for the treatment of MDR-TB [27].

The small molecule pentacyano(isoniazid)ferrate(II) (PIF) is the result of a rational drug design effort to find alternative drugs capable of inhibiting InhA [4, 5, 26]. PIF was found to inhibit both WT and INH-resistant Ile21Val mutants of InhA and this inactivation does not require activation by KatG [26]. Since crystal structures of the InhA-PIF binary complex are not available, we performed computational docking studies to predict the binding mode (s) of PIF in the InhA active site [27]. For that we used two crystal structures of InhA: the binary complex with the NADH coenzyme (PDB ID: 1ENY) [30] and the ternary complex with NAD+ and a substrate analog (PDB ID: 1BVR) [31]. We found that PIF preferentially occupies the pyrophosphate and nicotinamide sites in the NAD(H) binding pocket. However, we could not unambiguously assign a unique binding mode due to the distinct InhA active site conformations from the different PDB structures. We concluded that the flexibility of both, enzyme and inhibitor, should be taken into account to properly evaluate their interactions and to conform to the mechanism of slow-binding inhibition proposed for PIF based on WT InhA kinetic studies [27]. In a previous molecular dynamics (MD) simulation study of the NADH interaction with WT InhA and the mutants Ile21Val and Ile16Thr [31], both resistant to INH, we showed that InhA is a considerably flexible enzyme, capable of undergoing the conformational changes necessary to accommodate either substrate and inhibitor in an effective manner. Furthermore, we demonstrated that the mutations lead to conformational changes that reduced the affinity of the InhA-NADH complex. These results were soon after confirmed by X-ray crystallographic studies [32]. A recent characterization of PIF binding to MTB's InhA, using spectrofluorimetric techniques [28], hinted at the possibility that the quenching in protein fluorescence upon ligand binding, reported by the tryptophan amino acid (Trp) fluorescence, is due to conformational changes in the protein as previously suggested [27]. The identification of this enzyme's conformational changes requires a considerable experimental effort, highlighting the practical value of computer simulations in their prediction. Consequently there is current interest in the prediction of the three-dimensional (3-D) nature of InhA-PIF specificity, how the enzyme binds to the inorganic complex, which conformational changes takes place upon binding, the effect of these changes on the Trp residues solvent accessibility, and which amino acid residues are responsible for PIF binding in the enzyme active site. We address these issues using computational methods, including automated molecular docking and MD simulations. Classical MD simulations make possible the detailed analysis of the individual movements of the atoms in the molecules as a function of time, resulting in an ideal model for understanding the atomic and molecular mechanisms involved in the formation of non-covalent enzyme-ligand complex [33–36].

Methods

The initial structure of the complex

We obtained the initial structure of WT InhA from the first crystal structure of the InhA-NADH binary complex (PDB

ID: 1ENY) determined at 2.2 Å resolution [30]. The optimized structure of PIF and the Cartesian coordinates (x, y, z) of the InhA-PIF complex were taken from our previous molecular docking work, with the initial position of PIF in the active site of InhA taken from Oliveira et al. [27] (cluster 4 of Table 1 and Fig. 11e in reference 27). This cluster was top ranked and contained only one docked conformation. A simple translation and/or rotation could easily reproduce any of the other three top ranked clusters [27].

Although size exclusion chromatography analysis demonstrates that MTB's InhA biologically active structure is a homo-tetramer in solution, [37], each monomer binding cavity works independently. This is due to the fact that the active sites of the four monomers are about 40.0 Å apart from each other and are facing opposite sides in the quaternary structure. Hence, in this work, for all modeling and simulations, we use the InhA monomer (PDB ID: 1ENY). The all-atom model of the apo InhA enzyme contains 4,009 atoms with a net charge of −3 since His120 is protonated.

Force field and charges for PIF

Pentacyano(isoniazid)ferrate(II) is a new molecule [26] with 28 atoms and unknown charges and force field parameters. The partial atomic charges of PIF were determined as described by Oliveira et al. [27], and its force field parameters empirically derived by comparison with similar small molecules [38, 39]. To test these parameters, we run MD simulations of the PIF molecule fully solvated with TIP3P [39] water molecules at a temperature of 298.16 K. The initial cell dimensions containing the solute were 30.401×29.770×27.680 Å^3 with PIF solvated by a layer of water molecules of at least 10.0 Å in all orthogonal directions [40]. The test confirmed the PIF force field parameters we developed are adequate to be used in the MD simulation of the InhA-PIF complex. It is important to point out that the inorganic Fe^{+2} atom in the PIF molecule is covalently attached to its cyanide groups (Fig. 1a).

MD simulations

The main MD simulation started from the initial structure of InhA with the docked PIF inhibitor. InhA in complex with PIF (InhA-PIF) contains 4,037 atoms and a net molecular charge of −6, considering His120 protonated in the crystal structure. Hence, six sodium ions (Na$^+$) were added to neutralize the net negative charge density of the complex, which was then immersed in an orthorhombic box containing a total of 10,502 TIP3P water molecules [39]. The simulation cell dimensions were 77.725×73.328×77.345 Å^3 and the complex was solvated by a layer of water molecules of at least 10.0 Å in all orthogonal directions [40]. The simulation cell contained a total of 35,549 atoms. Energy minimization, equilibration and production phases of the MD simulations were performed as described earlier [40]. The simulation was computed in a NPT ensemble at 298.16 K with the Berendsen temperature coupling [41] and constant pressure of 1.0 atm, with isotropic molecule-based scaling [42]. The SHAKE algorithm [43] was applied, with a tolerance of 10^{-5} Å, to fix all bonds that contained a hydrogen atom, allowing the use of a 2.0 fs time step in the integration of the equations of motion. No extra restraints were applied after the equilibration phase. Periodic boundary conditions were applied, with electrostatic interactions between non-bonded atoms evaluated by the particle-mesh Ewald (PME) method. The Lennard-Jones interactions were evaluated using a 9.0 Å atom-based cutoff [44]. Four independent molecular systems' simulations were generated. The first one consisted of a 70.0 ns simulation of the PIF molecule alone in a neutral, with three Na$^+$ ions, aqueous solution. The apo InhA enzyme consisted of the second one and the third was composed by the binary complex InhA-PIF. For these two simulations, data were collected for 25.0 ns. From the InhA-PIF MD simulation we built the fourth simulation, named InhA-PIF$^{(-)}$. For this simulation we removed the PIF molecule and three Na$^+$ counter-ions "on-the-fly" from the third system (InhA-PIF) at 10.0 ns. The InhA-PIF$^{(-)}$ system was then allowed to relax for another 15.0 ns. "On-the-fly" [31] means that we modified the InhA-PIF molecular system and continued the MD simulation without reassignment of velocities.

Snapshots were collected at every 0.5 ps for analysis. All MD simulations were performed with the SANDER module of AMBER9 [42] using the ff99SB force field [45]. The stability of the simulations were analyzed in terms of the convergence of energy components, secondary structure content, solvent accessible surface area (SASA), radius of gyration (Rgyr), and the root-mean-squared deviation (RMSD) [46] from the initial, crystal structure (PDB ID: 1ENY). The tetramer structures for the InhA enzyme were generated using the symmetry operators available in the crystal structure (PDB ID: 1ENY) at the protein interfaces, surfaces and assemblies (PISA) web server of the EBI (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html) [47].

Analysis of the MD simulation trajectories

The MD simulation trajectories were visually monitored with the computer graphics software VMD [48]. Individual 3-D structures were further analyzed with Swiss-PdbViewer [49] and their illustrations prepared with the PyMOL molecular graphics system [50]. There are many different ways to evaluate the nature of intermolecular interactions or recognition, including making predictions of the estimated

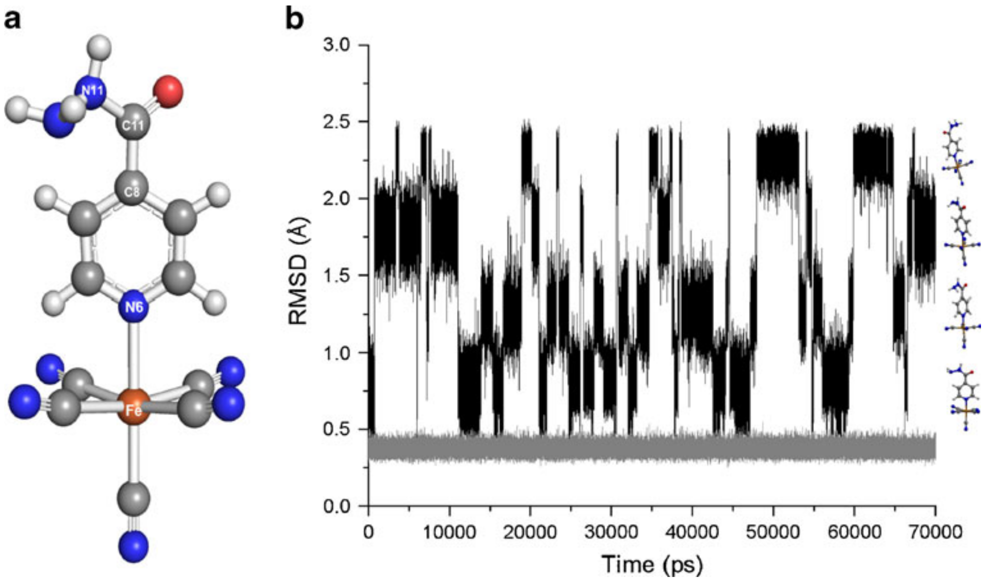


Fig. 1 (**a**) Ball-and-stick model of the 3-D structure of pentacyano(isoniazid)ferrate(II) (PIF) molecule. The atoms at the ends of a rotatable bond are highlighted. The first one is between the nitrogen atom (N6) of the pyridine ring and the iron atom (Fe) of the pentacyanoferrate moiety. The second one is delimited by atoms C8 and C11 and the last one by C11 and N11. (**b**) The RMSD as a function of time for the entire PIF molecule (black line) and the pentacyanoferrate moiety (gray line). The four molecular structures at the right illustrate each of the four sets of conformers adopted by PIF during the 70.0 ns MD simulation

free energy of binding, involved in protein-ligand affinity. In this work intermolecular recognition is evaluated by analyzing the total number of direct non-bonded interactions, i.e., hydrogen bonds (HB) and hydrophobic contacts (HC). Although waters play a major role in intermolecular recognition, here water-mediated H-bond interactions are not being considered. The total number of direct H-bonds in the InhA-PIF complex was calculated with the HBPLUS program [51] using a maximum donor-acceptor atoms' distance of 3.4 Å and a minimum angle of 90.0º. We used the program LIGPLOT [52] for plotting H-bonds and HCs. PROMOTIF [53] was used to evaluate the InhA secondary structure pattern along the MD simulation trajectories. NACCESS [54] was used to calculate the SASA parameters of the Trp amino acid residues. The RMSD and the radius of gyration (Rgyr) were calculated with the Ptraj module of AMBER9 [42]. For all comparative analyses we used as reference the initial crystal structure (PDB ID: 1ENY). Graphics and statistical analyses were performed with Origin 7 Scientific Graphing and Analysis Software (Microcal Software, OriginLab, Northampton, MA). We also developed in house Python-based software to automate the analysis of the 270,000 snapshots generated by the four MD simulations described in this work. In the analyses, we adopted the PDB [55] numbering scheme of the reference structure with amino acid residues represented by a three-letter code.

Results and discussion

Tests of the PIF parameters

To test the empirically derived PIF force field (FF) parameters, we first performed MD simulation in a water environment. No extraneous energy values and conformations were observed. Figure 1b shows the RMSD of the coordinates of PIF in water with respect to the initial structure [27]. The all-atoms RMSD fluctuates between 0.3 Å and 2.5 Å along the 70.0 ns simulation. During this time the PIF molecule reversibly explores four different sets of conformers (Fig. 1b). These sets are populated with 39,591, 42,076, 28,677, and 29,656 molecules, respectively. Their RMSD averages 0.5±0.3 Å, 1.3±0.2 Å, 1.8±0.2 Å, and 2.3±0.2 Å, correspondingly. The observed fluctuations are due to PIF's intrinsic flexibility mainly due to the ability of its pyridine moiety to flip about the N6-Fe bond, and the torsions about the other two rotatable bonds of PIF (Fig. 1a). The rigid, pentacyanoferrate(II) moiety of PIF has stable RMSD values during the entire simulation, converging to an insignificant change of about 0.4 Å. Altogether, these results demonstrate the stability of the PIF molecule when free in an aqueous solution. They showed that its FF parameters are adequate for PIF to be further explored in other simulation studies. Hence, we simulated the aqueous InhA-PIF complex. In this simulation the PIF movements about the N6-Fe and the other two rotatable bonds are severely restricted by interactions with the amino acids residues in its InhA's binding site. As a result the RMSD of PIF in the complex fluctuates much less, ranging from 0.4 Å to 1.4 Å (Fig. 2) whilst the rather rigid pentacyanoferrate moiety RMSD values converged to 0.4 Å, as expected.

InhA conformational features

After the warm-up phase of the MD simulation, the first 120 ps, the enzyme backbone RMSD with respect to the initial structure of the apo InhA increases slowly and monotonically to a value close to 2.1 Å about which it

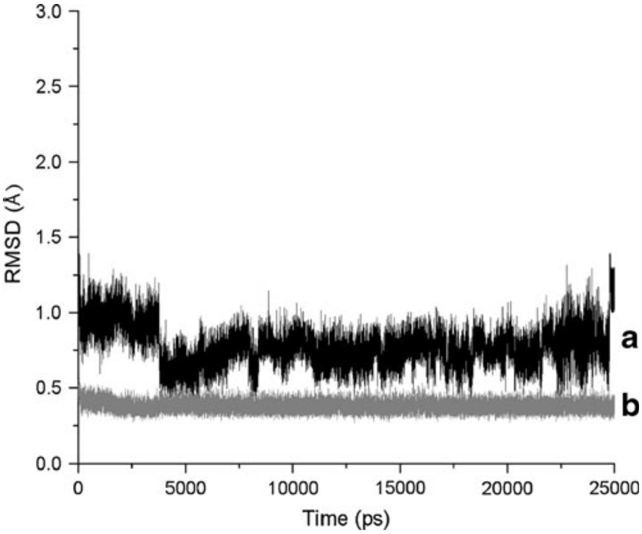


Fig. 2 Time dependence of the RMSD of (**a**) the entire PIF molecule in the binary InhA-PIF complex (black) and (**b**) the pentacyanoferrate moiety (gray). Compared to Fig. 1b, PIF movement is severely restricted when in the complex with the InhA enzyme

remains the entire simulation time with very small fluctuations (Fig. 3). The RMSD for the InhA-PIF complex (Fig. 3) ranges from 0.7 to 2.9 Å with a median value of 2.1 Å. After the thermalization phase it oscillates in the region of 1.7 Å remaining so for the next 2.0 ns. Then it rapidly decreases and stabilizes for another 1.0 ns. After that period the RMSD stably progresses to higher values reaching an average of 2.3±0.1 Å in the last 5.0 ns (interval from 20.0 to 25.0 ns). Because this analysis involved InhA backbone atoms only, the observed drifting of the RMSD to higher values suggests that some degree of conformational change took place in the enzyme structure during the simulation of the complex. To test whether these changes were caused by the presence of the ligand, at 10.0 ns PIF and three Na^+ ions were removed from the complex. The remaining apoenzyme, named InhA-PIF$^{(-)}$, in the simulation box, was allowed to relax for another 15.0 ns. During this simulation the enzyme backbone RMSD reached a plateau around 2.5±0.1 Å in the first 4.0 ns, followed by a gradual decrease over the remaining 11.0 ns, stabilizing at 2.4±0.1 Å in the last 5.0 ns of the simulation. The apo InhA and the InhA-PIF simulations suggested that the conformational changes observed in the enzyme were in fact due to the formation of a tight InhA-PIF complex during the first part of the simulation (0.0-10.0 ns of the InhA-PIF complex). It is interesting to notice that, even after 15.0 ns of simulation, the InhA-PIF$^{(-)}$ system did not recover from the changes caused by the interaction with PIF in the first 10.0 ns.

Inspection of other parameters such as the Rgyr and local secondary structure conservation, points to significant global and local changes in the InhA-PIF complex. The Rgyr is a measure of the protein dimension or compactness and, as the RMSD, is a measure of global structural changes in a protein. Figure 4 shows that, while for the apo InhA simulation Rgyr has an average of 17.9±0.1 Å and fluctuates about the experimental, crystal structure value of 17.9 Å, in the InhA-PIF and InhA-PIF$^{(-)}$ Rgyr averages 18.3±0.1 Å and 18.4±0.1 Å, respectively, for the last 5.0 ns of their MD trajectories. While these values are statistically identical, they are different from that of the apo InhA.

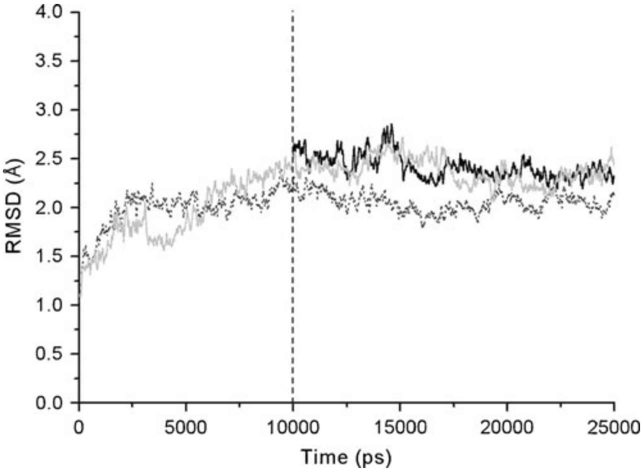


Fig. 3 RMSD of InhA backbone atoms (N, Cα, C, O) with respect to the initial, crystal structure (PDB ID: 1ENY) along the 25.0 ns MD simulation trajectory. The gray dashed line indicates the apo InhA simulation. The light gray represents the simulation of the InhA-PIF complex, and the black line indicates the InhA-PIF$^{(-)}$ simulation. The vertical dotted line at 10.0 ns indicates the time at which the PIF molecule was removed *on-the-fly* from the InhA-PIF complex. See the Methods section for details

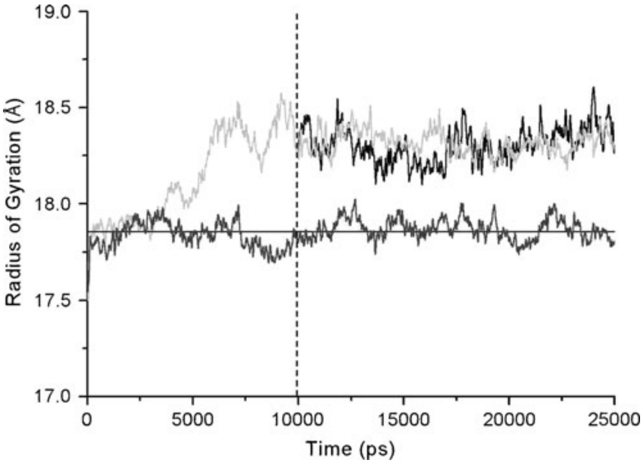


Fig. 4 Radius of gyration (Rgyr) for the backbone atoms of InhA as a function of time. The vertical dashed line at 10.0 ns indicates the time at which the PIF molecule was removed *on-the-fly* from the InhA-PIF complex. The horizontal black line indicates the radius of gyration for the initial, crystal structure (PDB ID: 1ENY). The dark gray line indicates the apo InhA simulation, the light gray represents the simulation of the InhA-PIF complex while the black line indicates the InhA-PIF$^{(-)}$ simulation

These differences become even clearer when we look at Fig. 4. From the InhA-PIF simulation, up to 10.0 ns, the Rgyr changes drastically.

In view of the fact that the RMSD as well as the Rgyr parameters measure global changes in the enzyme, it is possible that cooperation of local structural changes that led to the observed large deviations will take much longer than 15.0 ns to restore the RMSD values of the InhA-PIF$^{(-)}$ simulation to those of the apo InhA. Because Rgyr was calculated using the enzymes' backbone atoms, we hypothesize that these conformational changes could be the result of local disorder in regular (helices and sheets) and irregular (turns) secondary structure elements of the enzyme. These observations lead us to investigate the conservation of secondary structure in the InhA-PIF and InhA-PIF$^{(-)}$ systems.

To confirm this hypothesis and to detail the structural nature of the conformational changes, PROMOTIF [53] was used to measure the time dependence of InhA secondary structure (SS) content during the MD simulations (Fig. 5). These analyses suggest a considerable perturbation resulting in conversion of regular SS (α-helix mainly) in to irregular structures (coils) in the InhA-PIF complex. These alterations can be seen in the tertiary structure (Fig. 6). Comparing the MD snapshots at 10.0 ns (Fig. 6b) and 25.0 ns (Fig. 6c) with the initial enzyme structure (Fig. 6a) we notice that these structural changes mainly occur in the substrate binding loop [7], comprehending helices α6 and α7, and that the overall Rossmann fold [56–58] of the enzyme remained unchanged. This result is in agreement with the X-ray diffraction studies of Sullivan et al. [21]. These authors identified similar changes in MTB's InhA 3-D structure upon binding of alkyl diphenyl ethers or triclosan. Our observations are important as they support the hypothesis that PIF binding to InhA induces a conformational change [27, 28] albeit, in our models, it is not directly interacting with the substrate binging site. Up to 20.0 ns PIF interacts with a site that overlaps with the NADH binding site (see details below). This observation is in agreement with the experimental results by Oliveira et al. [32] which imply that the inhibition mechanism of PIF may involve interaction to both the NADH coenzyme and the substrate binding sites [27]. However, after 20.0 ns and in the last 5.0 ns, PIF abruptly dissociates from the complex (see details below).

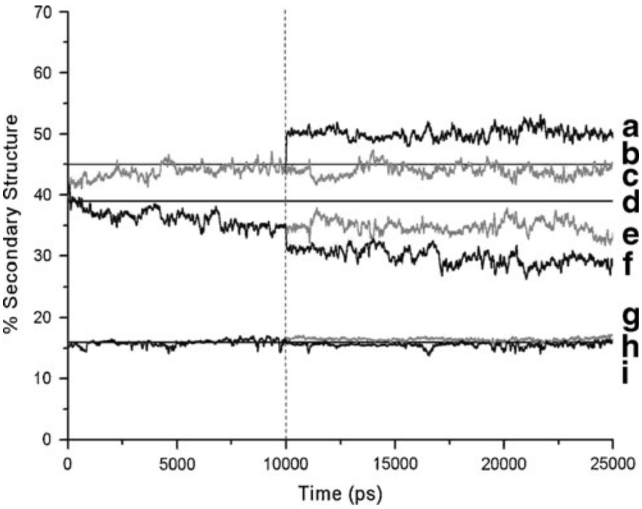


Fig. 5 InhA secondary structure content (in %) as a function of time along the MD simulations. The vertical dashed line at 10.0 ns indicates the time at which the PIF molecule was removed *on-the-fly* from the InhA-PIF complex. (**a**) Content of irregular structures and coils for the InhA-PIF$^{(-)}$ simulation. (**b**) The reference helical content based on the crystal structure (PDB ID: 1ENY). (**c**) Content of irregular structures and coils for the InhA-PIF simulation. (**d**) The reference coil and irregular structures' content based on the crystal structure (PDB ID: 1ENY). (**e**) Content of α-helices for the InhA-PIF complex. (**f**) Content of α-helices for the InhA-PIF$^{(-)}$ simulation. (**g**) Content of β-sheet for the InhA-PIF simulation. (h) The reference β-sheet content based on the crystal structure (PDB ID: 1ENY). (**i**) Content of β-sheet for InhA-PIF$^{(-)}$ simulation

Experimental data on intrinsic InhA fluorescence indicated that PIF binding to the enzyme active site triggers a conformational change, inducing the formation of a more stable enzyme-inhibitor complex [28]. This observation is important since protein fluorescence usually is related to the solvent accessibility of Trp amino acid side-chains and this is a suitable method to study protein conformational changes and interactions with others molecules [59]. To understand and possibly correlate our data with the experimental fluorescence results of the InhA-PIF interaction [28], we identified all Trp residues in InhA and measured their solvent accessible surface area (SASA) in the initial structure and along the MD simulation.

Changes in SASA of tryptophan residues correlate with the fluorescence spectra of InhA-PIF complex

MTB's InhA enzyme has four Trp residues. Trp160 is located in the A-loop, Trp222 in helix α7, Trp230 is in the loop between helices α7 and α8, and Trp249 is located in a loop at the C-terminus. As we are simulating the InhA monomer, and the InhA biological unit is a tetramer [7],

Table 1 Trp residues' solvent accessible surface area (SASA) analysis. The Trp SASA (in Å^2) for each monomer in the tetramer built from the experimental initial protein structure (PDB ID: 1ENY) and from the snapshot at 25.0 ns

Trp residue	Tetramer's SASA (Å^2) PDB ID: 1ENY	Tetramer's SASA (Å^2) for snapshot at 25.0 ns
160	4.62	12.42
222	11.33	0.02
230	17.72	28.43
249	39.51	93.7

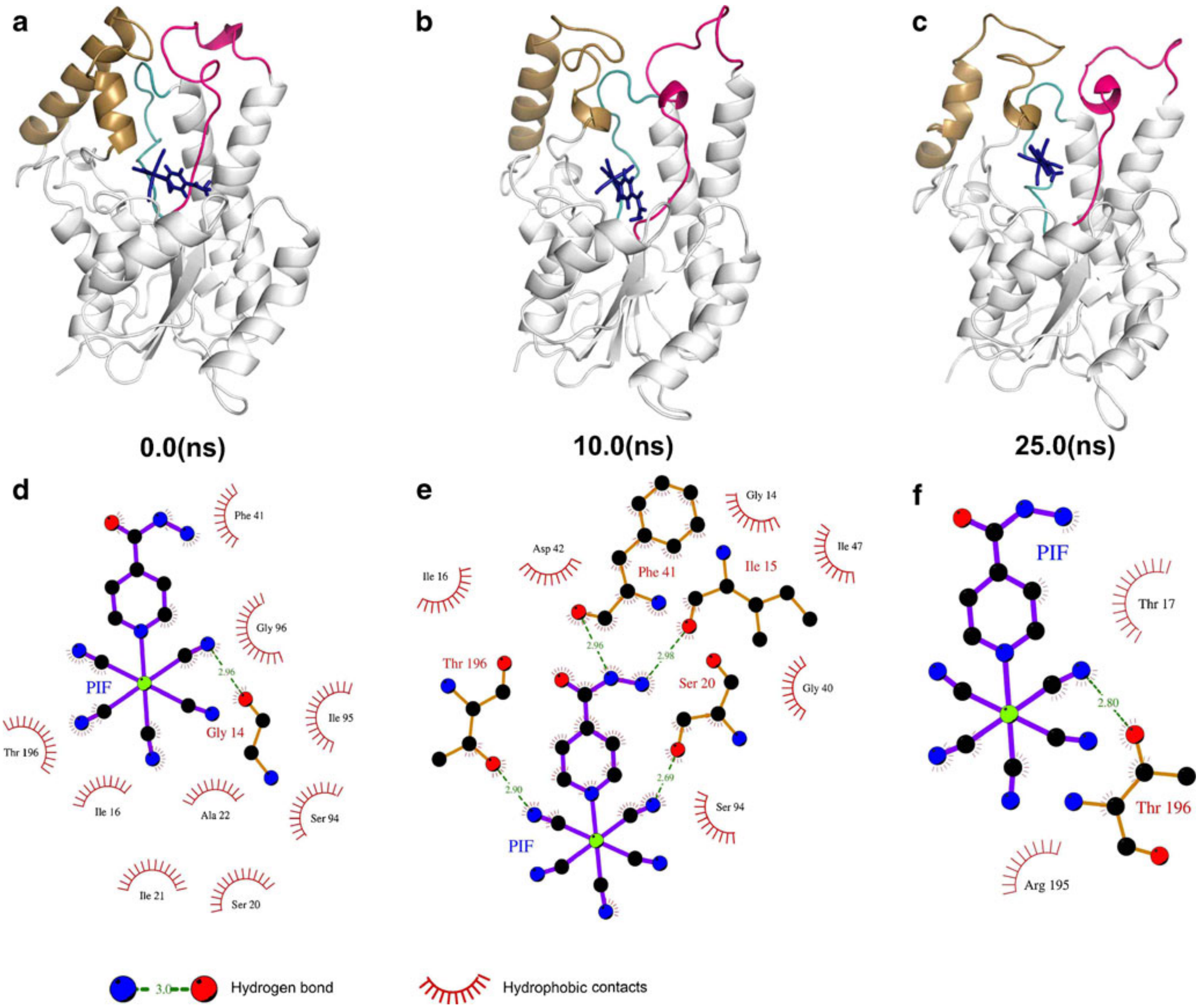


Fig. 6 Ribbon representation of the tertiary structure of the InhA-PIF complex (top) and their interactions calculated with HBPLUS and illustrated with LIGPLOT (bottom). (**a**) and (**d**) represent the initial simulation structure of the InhA-PIF complex; (**b**) and (**e**) the snapshot at 10.0 ns; and (**c**) and (**f**) the snapshot at 25.0 ns. The InhA motifs interacting with the inorganic PIF molecule (in dark blue stick model) are highlighted in different colors. The substrate binding loop (helices α6 and α7) is sand color, the A-loop is magenta, and the B-loop is cyan. At 10.0 ns we can notice a significant perturbation of the substrate binding loop and some disorder in helix α2, maintained until the end of the 25.0 ns simulation. In the initial structure PIF makes hydrophobic contacts and H-bond to nine and one residues of InhA, respectively. At the end of 10.0 ns PIF makes hydrophobic contacts and H-bonds to six and four residues of InhA, respectively. Finally, at 25.0 ns, only three residues makes hydrophobic contacts and only one makes a H-bond

SASA for those Trp residues located in the tetramer interfaces were not realistic. In a single subunit these residues are overexposed to the solvent (Fig. 7). Examples are Trp249 and Trp160, located in the interfaces of the tetramer subunits. To partially overcome this problem, we built two InhA tetramers models as described in the Methods section. One used the initial InhA structure (PDB ID: 1ENY) for reference purposes and the other the MD simulation snapshot at 25.0 ns. The total SASA for the Trp residues were calculated with NACCESS [54] and their values compared. Trp249 and Trp160, located internally in the tetramer, have a SASA remarkably reduced when compared to their values in the monomer (Trp160$_{\text{monomer}}$=53.02 Å^2, Trp160$_{\text{tetramer 0.0 ns}}$=4.62 Å^2, Trp249$_{\text{monomer}}$=189.91 Å^2, Trp249$_{\text{tetramer 0.0 ns}}$=39.51 Å^2). The SASA of Trp222 and Trp230 are closer independently of the form of calculation, (Trp222$_{\text{monomer}}$=15.55 Å^2 , Trp222$_{\text{tetramer 0.0 ns}}$=11.33 Å^2, Trp230$_{\text{monomer}}$=23.18 Å^2, Trp230$_{\text{tetramer 0.0 ns}}$=17.72 Å^2). Hence, we conclude that Trp222 and Trp230 residues are the ones likely to be related to the changes in the InhA-PIF fluorescence [28]. Table 1 shows the Trp SASA in the tetramers in the initial structure (PDB ID: 1ENY) and at the end of the InhA-PIF simulation at 25.0 ns. The conformational changes affect the SASA of the Trp residues

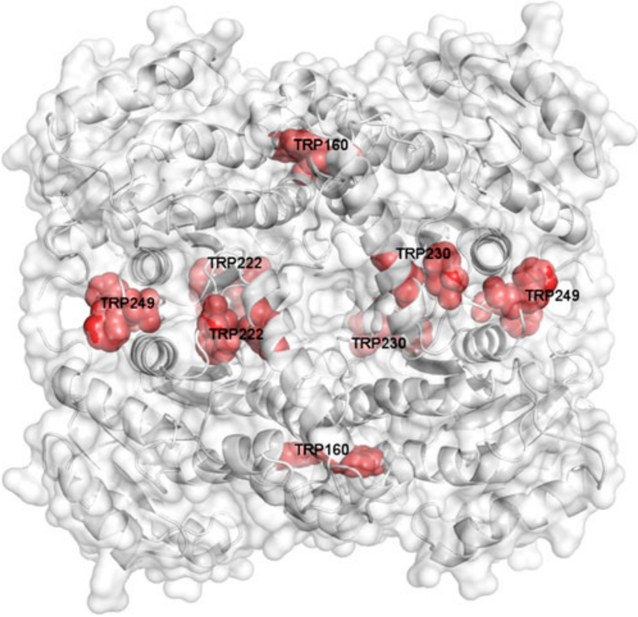


Fig. 7 Molecular surface representation of the InhA tetramer built from the initial structure (PDB ID: 1ENY). The ribbon representation of the monomers is white colored. The Trp residues are represented by van der Waals spheres colored in red

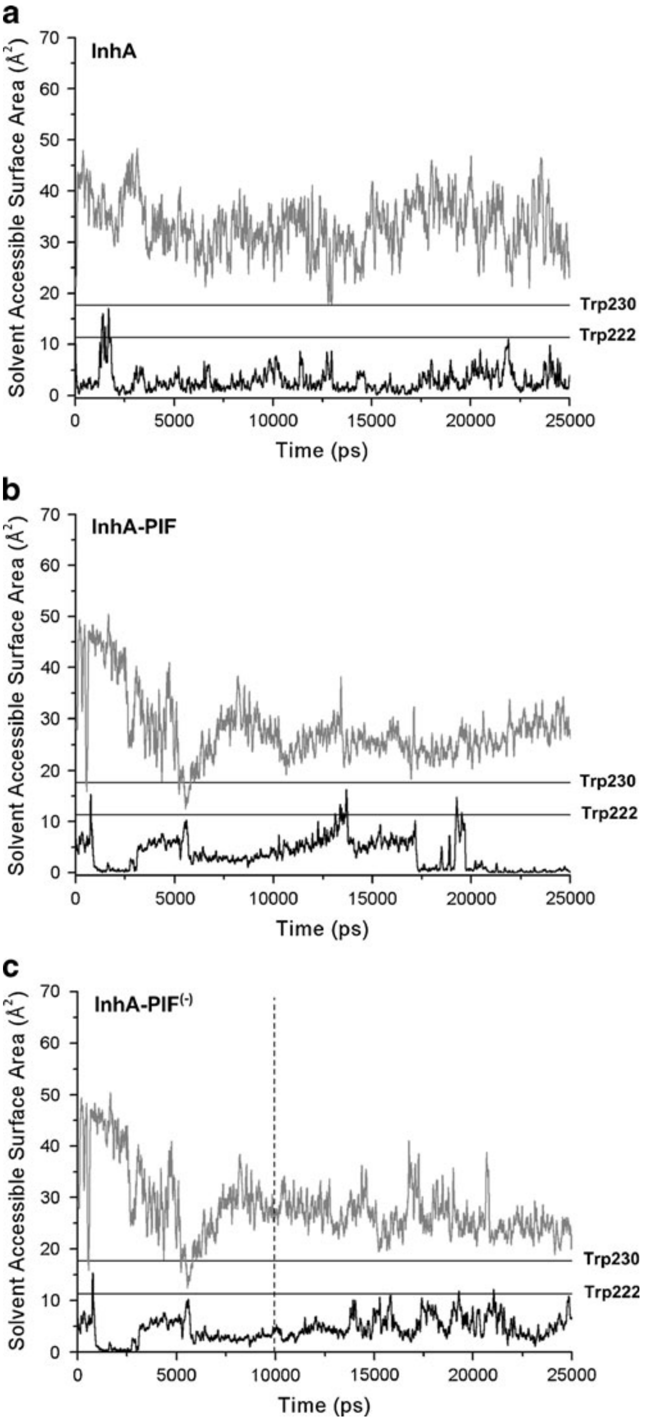


Fig. 8 The solvent accessible surface area (SASA) for Trp222 (black) and Trp230 (gray) as a function of simulation time for the three simulations: apo InhA, InhA-PIF complex and for InhA-PIF$^{(-)}$. The horizontal lines show the SASA values in reference tetramer for the initial structure. The vertical dashed line at 10.0 ns in the InhA-PIF$^{(-)}$ simulation indicates the time at which the PIF molecule was removed *on-the-fly* from the InhA-PIF complex

and consequently the fluorescence. After analyzing the data more closely we concluded that Trp222 is the one with the largest reduction in SASA when the tetramer's models are compared in the initial and final conformations of the InhA-PIF complex (Table 1). Figure 8 shows the SASA variation of Trp222 and Trp230 along the 25.0 ns MD simulations. Although there is a tendency of the Trp230 SASA to increase during the dynamics simulation, it is evident that there is a reduction of the SASA for Trp222, which is located in helix α7 that, together with helix α6, forms the substrate binding loop. Most of the structural changes occur in this motif. Together, these data corroborate the hypothesis that PIF causes 3-D conformational changes in InhA, thus agreeing with the findings of Sullivan et al. [21], and explaining from a structural dynamics standpoint the experimental studies of Vasconcelos et al. [28].

InhA-PIF association

The initial structure of the InhA-PIF complex was taken from our previous automated molecular docking studies, with the initial position of PIF in the active site of InhA obtained from cluster 4 of Table 1 and Fig. 11e from Oliveira et al. [27]. In this predicted complex the PIF ligand made 11 hydrophobic contacts (HCs) with InhA. To obtain further insight into the PIF-binding mechanism in InhA we explored their mode of interaction by atomistic MD simulations. Analysis of the InhA-PIF interactions along the simulation identified those residues important for PIF binding. We computed direct H-bonds as well as HCs with LIGPLOT [52]. Our analyses focused on the number and nature of the combined H-bonds and HCs interactions as described in the Methods section. Estimations of free-energy of binding, and its corresponding enthalpic and entropic contributions, are out of the scope of this investigation. Tables 2 and 3 show the percentage of time H-bonds and HCs, respectively, lasted during the whole simulation. The values indicate that 13 InhA amino acids residues make H-bonds to PIF in at least one snapshot during the 25.0 ns simulation. Residues Ile15, Ser20, Phe41, Arg43 and Thr196 interact with PIF over 50% of the time. Hence, they are considered as the most important residues for InhA-PIF interaction (Table 2). For instance, at 10.0 ns (Fig. 6e) only H-bonds from Arg43 are missing. Furthermore, 25 residues make HCs with PIF, with Gly14, Ile15, Ile16, Ser20, Gly40, Phe41, Arg43, Ile47 and Thr196 being the most important, interacting in more than 50% of the simulation time. These residues are located in the loop formed between helix α1 and strand β1 (Gly14, Ile15, Ile16, Ser20), in strand β2 (Gly40, Phe41), in the loop between strand β2 and helix α2 (Arg43), in helix α2 (Ile47), and in helix α6 (Thr196). Figures 9 and 10 show the total number of InhA residues making H-bonds and HCs with PIF as a function of the simulation time. Approximately five residues make H-bond interactions with PIF throughout almost the whole simulation time. Starting at about 20.0 ns this number drops sharply during the next 5.0 ns (Fig. 9), and at the end of the simulation there is only one residue (Thr196) making H-bond in the InhA-PIF complex (Fig. 6f).

The number of residues making HCs drops quickly from 17 to 7 in the first 10.0 ns (Fig. 10). Similarly to the dynamic behavior of H-bonding residues (Fig. 9), from

Table 2 Hydrogen bonds analysis. Amino acids residues H-bonded to PIF along the 25.0 ns and during the last 5.0 ns of the MD simulation expressed as percentage (%) of time

Amino acids that H-bonds to PIF	% of total time for 25.0 ns	% of total time for the last 5.0 ns
Gly14	47.97	19.64
Ile15	76.01	37.97
Ile16	0.05	0.24
Thr17	7.35	35.95
Ser19	0.85	4.24
Ser20	83.57	36.66
Phe41	76.48	45.10
Asp42	1.25	6.24
Arg43	56.84	35.82
Ser94	18.56	0.77
Gly96	0.24	0.09
Arg195	0.24	1.21
Thr196	73.58	87.21

Table 3 Hydrophobic contacts analysis. Amino acids residues making hydrophobic contacts to PIF along the 25.0 ns and during the last 5.0 ns of the MD simulation expressed as percentage (%) of time

Amino acids making HCs to PIF	% of total time for 25.0 ns	% of total time for the last 5.0 ns
Gly14	83.94	40.71
Ile15	86.48	39.20
Ile16	98.53	94.39
Thr17	10.65	45.86
Ser19	2.56	12.73
Ser20	89.81	52.40
Ile21	14.73	0.06
Ala22	15.11	0.00
Gly40	72.47	30.81
Phe41	93.31	66.54
Asp42	10.69	24.94
Arg43	70.81	70.97
Leu46	0.02	0.12
Ile47	86.42	39.88
Ser94	48.69	20.93
Ile95	6.99	1.14
Gly96	4.75	3.40
Phe97	0.16	0.04
Gln100	0.01	0.00
Met103	0.11	0.53
Met147	1.16	0.00
Arg195	3.11	13.60
Thr196	82.84	92.89
Leu197	21.04	29.39
Ala198	13.83	0.43

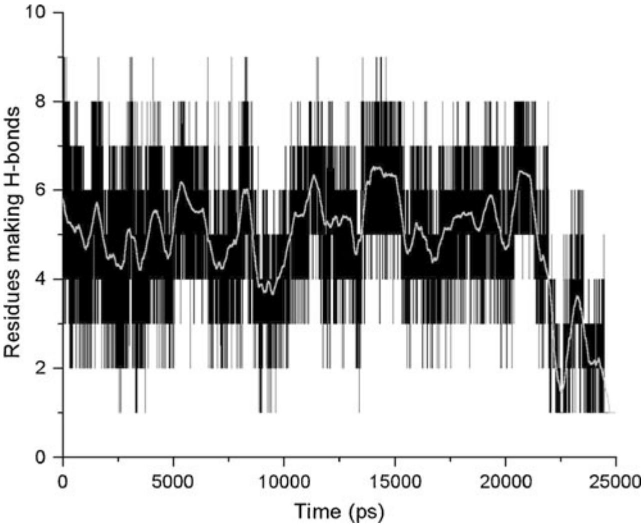


Fig. 9 Number of residues making H-bonds to PIF in the InhA-PIF complex as a function of time. The solid line represents smoothing of the data to facilitate the visualization

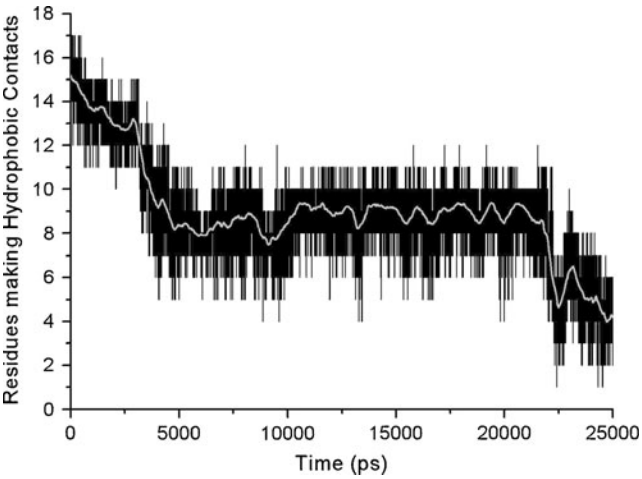


Fig. 10 Number of residues making HCs to InhA-PIF as a function of time. The solid line represents smoothing of the data to facilitate the visualization

20.0 to 25.0 ns, the number of HCs-making residues fell rapidly to a much lower value. At the end of the simulation time only three residues are making HCs to PIF in the InhA-PIF complex. This clearly indicates a dissociation of PIF from the initial InhA-PIF complex. These analyses and visual inspection of the InhA-PIF MD simulation trajectory with VMD [48] showed that, after forming a tight binding InhA-PIF complex during the first 20.0 ns, PIF starts to dissociate from the complex.

All this data supports the notion that PIF first strongly interacts with the InhA enzyme, causing large conformational changes that mostly affect, by diminishing, the content of regular secondary structures. The fact that the residues involved in PIF interaction with InhA include those involved in InhA mutations related to resistance to INH and other anti-tuberculosis drugs that targets InhA might explain PIF's ability to inhibit both wild type and I21V INH-resistant InhA [27]. In addition to that, PIF dissociation observed at the end of the InhA-PIF complex simulation strongly suggests that, despite the strong InhA-PIF association at the beginning of the simulation, the initial position of PIF in the binding site adopted for the initial structure might not be the one expected for this ligand. This binding site overlaps with that of InhA innate NADH coenzyme. This is encouraging since many important enzymes in humans contain the NAD(P)H coenzyme. If PIF does not compete with this site, it is likely to be a promising inhibitor candidate of MTB's InhA.

Conclusions

There is still no 3-D structure of the InhA-PIF binary complex to confirm the binding mode and to assess the molecular interactions PIF makes with the InhA active site residues which includes those that bind to the NADH coenzyme and those that make up the substrate binding cavity. While X-ray crystallography and nuclear magnetic resonance (NMR) are powerful and the preferred techniques to determine the 3-D structure of enzymes and their complexes, it is not always possible to obtain single crystals or NMR solution of macromolecular complexes. Meanwhile, alternative, computational methods, such as molecular docking and fully solvated all-atom MD simulations of the wild-type InhA enzyme in complex with PIF, can provide comparable and useful information to help elucidate the molecular nature of InhA inhibition by the inorganic PIF ligand. We also performed MD simulations of apo InhA enzyme and the InhA enzyme (InhA-PIF$^{(-)}$) after PIF removal from the initial InhA-PIF complex. These two simulations served as control simulations. They showed that the results observed for the InhA-PIF complex simulation are not an artifact of the method. Our approach for this work has provided important insights about the InhA-PIF interactions based on the analyses of the RMSD, Rgyr, SASA of Trp residues, H-bonds and HCs. We were capable of identifying all residues that interact with the inorganic PIF ligand along the simulation, the nature of these interactions, whether H-bonds and/or HCs, and details of the conformational changes undergone by InhA upon PIF binding. The PIF compound appears to be a promising candidate to further antitubercular drug development and may represent a new class of lead compounds since it needs no activation by KatG or other enzyme, and furthermore, there is no need for the formation of any kind of adduct with the coenzyme NADH to bind to its molecular target, the *M. tuberculosis* InhA enzyme [26]. From this work, we can conclude that PIF strongly interacts with InhA and that these interactions leads to macromolecular instabilities reflected in the long time necessary for simulation convergence. The instability caused by the PIF interaction with InhA is mainly due to perturbation of the substrate binding loop, particularly the partial denaturation of helices α6 and α7. We also found that residues Gly14, Ile15, Ile16, Ser20, Gly40, Phe41, Arg43, Ile47, and Thr196 are responsible for the strong InhA-PIF association in the first 20.0 ns of the simulation. During this period, PIF directly competes for the NADH binding site. However, because at the end of the simulation PIF is almost completely dissociated from the InhA-PIF, we conclude that the mode of InhA-PIF interaction we simulated in this work may not reflect the actual InhA-PIF binding mode. It is worth remembering that the initial position of PIF in the InhA-PIF complex investigated here overlapped with the NADH binding site. We have also been able to correlate the changes in the SASAs of the Trp residues with the recent spectrofluorimetric investigation of the InhA-PIF complex

[27] and confirm their suggestion that the changes in fluorescence are due to InhA conformational changes upon PIF binding. Work underway in our laboratory is now being able to explore other possible InhA-PIF binding modes, as well as with other ligands. One possibility is to have PIF directly bound in the substrate binding cavity. As future work we will perform MD simulation of such complexes in order to provide knowledge and support to further the use of PIF as a lead compound to develop alternative treatment for tuberculosis using MTB's InhA as a target. Finally, we believe that this work constitutes a relevant contribution to the field of drug design and development with the use of molecular docking and MD simulations to help elucidate the binding mode of ligands, or prospective lead compounds, to their target protein receptor.

Acknowledgments We would like to thank Luís Fernando Saraiva Macedo Timmers for his technical assistance. We thank the Laboratório de Alto Desempenho (LAD), at Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), for CPU time. This project was supported all or in part by grants from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) and Ministério da Ciência e Tecnologia (MCT) – Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – Departamento de Ciência e Tecnologia (DECIT), (Processes numbers 410505/2006-4, 554782/2008-1, 302641/2009-2, 551209/2010-0, 559917/2010-4) to Osmar Norberto de Souza and Instituto Nacional de Ciência e Tecnologia em Tuberculose (INCT-TB) – CNPq to Prof. Diógenes Santiago Santos and Programa de Apoio a Núcleos de Excelência 2009/FAPERGS to Prof. Luiz Augusto Basso. André L. P. da Costa was partially supported by a CAPES M.Sc. scholarship. Ivani Pauli was supported by a CAPES M.Sc scholarship. Márcio Dorn was supported by a CNPq M.Sc. scholarship. Osmar Norberto de Souza is a CNPq Research Fellow.

References

1. Schaeffer ML, Agnihotri G, Volke C, Kallender H, Brennan BJ, Lonsdale JTI (2001) Purification and biochemical characterization of the Mycobacterium tuberculosis beta-Ketoacyl-Acyl Carrier Protein Synthases KasA and KasB. J Biol Chem 276:47029–47037
2. Ratledge C (1982) The biology of mycobacteria. Academic Press, San Diego
3. Takayama K, Qureshi N (1982) In: Kubica GP, Wayne LG (eds) The mycobacteria: a sourcebook. Marcel Dekker, New York
4. Schroeder EK, Norberto de Souza O, Santos DS, Blanchard JS, Basso LA (2002) Drugs that inhibit mycolic acid biosynthesis in Mycobacterium tuberculosis. Curr Pharm Biotech 3:197–225
5. Mdluli K, Slayden RA, Zhu Y, Ramaswamy S, Pan X, Mead D, Crane DD, Musser JM, Barry CE III (1998) Inhibition of a Mycobacterium tuberculosis beta-Ketoacyl ACP Synthase by Isoniazid. Science 280:1607–1610
6. Chatterjee D (1997) The mycobacterial cell wall: structure, biosynthesis and sites of drug action. Curr Opin Chem Biol 4:579–588
7. Quémard A, Sacchettini JC, Dessen A, Vilcheze C, Bittman R, Jacobs WR, Blanchard JS (1995) Enzymic characterization of the target for Isoniazid in Mycobacterium tuberculosis. Biochemistry 34:8235–8241
8. Banerjee A, Dubnau E, Quemard A, Balasubramanian V, Um KS, Wilson T, Collins D, de Lisle G, Jacobs WR (1994) inhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis. Science 263:227–230
9. Lavender C, Globan M, Sievers A, Jacobe HB, Fyfe J (2005) Molecular characterization of Isoniazid-resistant Mycobacterium tuberculosis isolates collected in Australia. Antimicrob Agents Chemother 49:4068–4074
10. Johnsson K, King DS, Schultz PG (1995) Studies on the mechanism of action of Isoniazid and Ethionamide in the chemotherapy of tuberculosis. J Am Chem Soc 117:009–5010
11. Zhang Y, Heym B, Allen B, Young D, Cole S (1992) The catalase-peroxidase gene and isoniazid resistance of Mycobacterium tuberculosis. Nature 358:591–593
12. Baker LV, Brown TJ, Maxwell O, Gibson AL, Fang Z, Yates MD, Drobniewski FA (2005) Molecular analysis of Isoniazid-resistant Mycobacterium tuberculosis isolates from England and Wales reveals the phylogenetic significance of the ahpC-46A polymorphism. Antimicrob Agents Chemother 49:1455–1464
13. Zhang Y, Garcia MJ, Lathigra R, Allen B, Moreno C, van Embden JD, Young D (1992) Alterations in the superoxide dismutase gene of an isoniazid-resistant strain of Mycobacterium tuberculosis. Infect Immun 60:2160–2165
14. Scorpio A, Zhang Y (1996) Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. Nat Med 2:662–667
15. Baulard AR, Betts JC, Engohang-Ndong J, Quan S, McAdam RA, Brennan PJ, Locht C, Besra GS (2000) Activation of the pro-drug ethionamide is regulated in mycobacteria. J Biol Chem 275:28326–28331
16. DeBarber AE, Mdluli K, Bosman M, Bekker LG, Barry CE 3rd (2000) Ethionamide activation and sensitivity in multidrug-resistant Mycobacterium tuberculosis. Proc Natl Acad Sci USA 97:9677–9682
17. Vannelli TA, Dykman A, Ortiz de Montellano PR (2002) The antituberculosis drug ethionamide is activated by a flavoprotein monooxygenase. J Biol Chem 277:12824–12829
18. Vilcheze C, Wang F, Arai M, Hazbón MH, Colangeli R, Kremer L, Weisbrod TR, Alland D, Sacchettini JC, Jacobs WR (2006) Transfer of a point mutation in Mycobacterium tuberculosis inhA resolves the target of isoniazid. Nat Med 12:1027–1029
19. Nguyena M, Quemard A, Marrakchi H, Bernadou J, Meunier B (2001) The nonenzymatic activation of isoniazid by MnIII-pyrophosphate in the presence of NADH produces the inhibition of the enoyl-ACP reductase InhA from Mycobacterium tuberculosis. Comptes Rendus del'Académie des Sciences – Series IIC –. Chemistry 4:35–40
20. Jia L, Tomaszewski JE, Hanrahan C, Coward L, Noker P, Gorman G, Nikonenko B, Protopopova M (2005) Pharmacodynamics and pharmacokinetics of SQ109, a new diamine-based antitubercular drug. Br J Pharmacol 144:80–87
21. Sullivan TJ, Truglio JJ, Boyne ME, Novichenok P, Zhang X, Stratton CF, Li HJ, Kaur T, Amin A, Johnson F, Slayden RA, Kisker C, Tonge PJ (2006) High affinity InhA inhibitors with activity against drug-resistant strains of Mycobacterium tuberculosis. ACS Chem Biol 1:43–53
22. He X, Alian A, Ortiz de Montellano PR (2007) Inhibition of the Mycobacterium tuberculosis enoyl acyl carrier protein reductase InhA by arylamides. Bioorg Med Chem 15:6649–6658
23. Kuo MR, Morbidoni HR, Alland D, Sneddon SF, Gourlie BB, Staveski MM, Leonard M, Gregory JS, Janjigian AD, Yee C, Musser JM, Kreiswirth B, Iwamoto H, Perozzo R, Jacobs WR , Sacchettini JC, Fidock DA (2003) Targeting tuberculosis and malaria through inhibition of Enoyl Reductase: compound activity and structural data. J Biol Chem 278:20851–20859

24. Zhang Y, Post-Martens K, Denkin S (2006) New drug candidates and therapeutic targets for tuberculosis therapy. Drug Discov Today 11:21–27
25. Wang F, Langley R, Gulten G, Dover Lynn G, Besra GS, Jacobs WR, Sacchettini JC (2007) Mechanism of thioamide drug action against tuberculosis and leprosy. J Exp Med 204:73–78
26. Oliveira JS, Souza EHS, Basso LA, Palaci M, Dietze R, Santos DS, Moreira IS (2004) An inorganic iron complex that inhibits wild-type and an isoniazid-resistant mutant 2-trans-enoyl-ACP (CoA) reductase from Mycobacterium tuberculosis. Chem Commun 3:312–313
27. Oliveira JS, Souza EHS, Norberto de Souza O, Moreira IS, Santos DS, Basso LA (2006) Slow-Onset Inhibition of 2-trans-Enoyl-ACP (CoA) Reductase from Mycobacterium tuberculosis by an inorganic complex. Curr Pharm Design 12:2409–2424
28. Vasconcelos I, Meyer E, Sales FAM, Moreira IS, Santos DS (2008) The mode of inhibition of Mycobacterium tuberculosis wild-type and Isoniazid-resistant 2-trans-Enoyl-ACP(CoA) Reductase enzymes by an inorganic complex. Anti-Inf Ag Med Chem 7:50–62
29. Basso LA, Schneider CZ, dos Santos AJAB, dos Santos AA, Campos MM, Souto AA, Santos DS (2010) An inorganic complex that inhibits Mycobacterium tuberculosis enoyl reductase as a prototype of a new class of chemotherapeutic agents to treat tuberculosis. J Braz Chem Soc 00:1–6
30. Dessen A, Quémard A, Blanchard JS, Jacobs WR, Sacchettini JC (1995) Crystal structure and function of the isoniazid target of Mycobacterium tuberculosis. Science 267:1638–1641
31. Schroeder EK, Basso LA, Santos DS, Norberto de Souza O (2005) Molecular dynamics simulation studies of the wild-type, I21V, and I16T mutants of isoniazid-resistant Mycobacterium tuberculosis enoyl reductase (InhA) in complex with NADH: toward the understanding of NADH-InhA different affinities. Biophys J 89:876–884
32. Oliveira JS, Pereira JH, Canduri F, Rodrigues NC, Norberto de Souza O, de Azevedo WF, Basso LA, Santos DS (2006) Crystallographic and pre-steady-state kinetics studies on binding of NADH to wild-type and Isoniazid-resistant Enoyl-ACP(CoA) Reductase enzymes from Mycobacterium tuberculosis. J Mol Biol 359:646–666
33. Cheatham III TE, Brooks BR (1998) Recent advances in molecular dynamics simulation towards the realistic representation of biomolecules in solution. Theor Chem Acc 99:279–288
34. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. Nat Struct Biol 9:646–652
35. Karplus M, Kuriyan J (2005) Chemical theory and computation special feature: molecular dynamics and protein function. Proc Natl Acad Sci USA 102:6679–6685
36. Nyarady Z, Czompoly T, Bosze S, Nagy G, Petrohai A, Pál J, Hudecz F, Berki T, Németh P (2006) Validation of in silico prediction by in vitro immunoserological results of fine epitope mapping on citrate synthase specific autoantibodies. Mol Immunol 43:830–838
37. Quémard A, Sacchettini JC, Dessen A, Vilchèze C, Bittman R, Jacobs WR , Blanchard JS (1995) Enzymatic characterization of the target for isoniazid in Mycobacterium tuberculosis. Biochemistry 34:8235–8241
38. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A second generation Force Field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117:5179–5197
39. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
40. Norberto de Souza O, Ornstein RL (1997) Effect of periodic box size on aqueous molecular dynamics simulation of a DNA dodecamer with particle-mesh Ewald method. Biophys J 72:2395–2397
41. Berendsen HJC, Postma JPM, van Gunsteren WF, Dinola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:3684–3690
42. Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, Pearlman DA, Crowley M, Walker RC, Zhang W, Wang B, Hayik S, Roitberg A, Seabra G, Wong KF, Paesani F, Wu X, Brozell S, Tsui V, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Beroza P, Mathews DH, Schafmeister C, Ross WS, Kollman PA (2006) AMBER 9. University of California, San Francisco
43. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 23:327–341
44. Norberto de Souza O, Ornstein RL (1999) Molecular dynamics simulations of a protein-protein dimmer: particle-mesh Ewald electrostatic model yields far superior results to standard cutoff model. J Biomol Struct Dyn 16:1205–1218
45. Roe DR, Okur A, Wickstrom L, Hornak V, Simmerling C (2007) Secondary structure bias in generalized born solvent models: comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit salvation. J Phys Chem B 111:1846–1857
46. Maiorov VN, Crippen GM (1994) Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. J Mol Biol 235:625–634
47. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372:774–797
48. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14:33–38
49. Kaplan W, Littlejohn TG (2001) Swiss-PDB Viewer (Deep View). Brief Bioinform 2:195–197
50. DeLano WL (2002) The PyMOL molecular graphics system. DeLano Scientific, Palo Alto, CA, USA
51. McDonald IK, Thornton JM (1994) Satisfying hydrogen bonding potential in proteins. J Mol Biol 238:777–793
52. Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng 8:127–134
53. Hutchinson G, Thornton (1996) JM PROMOTIF-A program to identify and analyze structural motifs in proteins. Protein Sci 5:212–220
54. Hubbard SJ, Thornton JM (1993) NACCESS, Computer Program. Department of Biochemistry and Molecular Biology. University College London
55. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242
56. Rossmann MG, Liljas A, Branden CI, Banaszak LJ (1975) In: Boyer PD (ed) Evolutionary and structural relationships among dehydrogenases. The Enzymes, 3rd edn. Academic, New York, pp 61–102
57. Jornvall H, Persson B, Krook M, Atrian S, Gonzalez RD, Jeffery J, Ghosh D (1995) Shortchain dehydrogenases/reductases (SDR). Biochemistry 18:6003–6013
58. Oppermann U, Filling C, Hult M, Shafqat N, Wu X, Lindh M, Shafqat J, Nordling E, Kallberg Y, Persson B, Jörnvall H (2003) Short-chain dehydrogenase/reductases (SDR): the 2002 update. Chem Biol Interact 143:247–253
59. Chen Y, Barkley MD (1998) Toward understanding Tryptophan fluorescence in proteins. Biochemistry 37:9976–9982

J Mol Model (2012) 18:1791–1799
DOI 10.1007/s00894-011-1195-0

ORIGINAL PAPER

Novel insights into the structural requirements for the design of selective and specific aldose reductase inhibitors

Hirdesh Kumar · Anup Shah · M. Elizabeth Sobhia

Received: 10 March 2011 / Accepted: 19 July 2011 / Published online: 12 August 2011

Abstract Aldose reductase (ALR2) plays a vital role in the etiology of long-term diabetic microvascular complications (DMCs) such as retinopathy, nephropathy and neuropathy. It initializes the polyol pathway and under hyperglycemic conditions, catalyzes the conversion of glucose into sorbitol in the presence of NADPH. Many ALR2 inhibitors have been withdrawn from clinical trial studies due to their cross reactivity with other analogues enzymes or due to impairment with detoxification role of ALR2. To address these issues we characterized the possible rationalities behind the selectivity problem associated with the enzyme-inhibitor interactions. Novel molecules were designed for the induce fit cavity region of ALR2. Docking studies were carried out using Glide to analyze the binding affinity of the designed molecules for ALR2. The analysis showed that the designed ALR2 inhibitors are selective for ALR2 over its close analogs. These inhibitors are also specific for the induced cavity region of ALR2 and do not interfere with the detoxification role of ALR2.

Keywords ALR2 · Detoxification · Docking · Induced fit cavity · Selectivity · Specificity

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1195-0) contains supplementary material, which is available to authorized users.

H. Kumar · A. Shah · M. E. Sobhia (✉)
Department of Pharmacoinformatics, National Institute
of Pharmaceutical Education and Research (NIPER),
Sector 67, S.A.S. Nagar,
Punjab 160062, India
e-mail: mesophia@niper.ac.in

Introduction

Hyperglycemic condition causes diabetic microvascular complications (DMCs), majorly affecting nervous system, kidney and eyes [1]. One of the key mechanisms responsible for DMCs is the activation of Polyol pathway [2]. In normal physiological conditions glucose is phosphorylated by hexokinase, whereas in hyperglycemic conditions the excess glucose becomes the substrate to initiate the Polyol pathway. Polyol pathway comprises two consecutive steps. In the first rate limiting step, reduction of glucose to sorbitol takes place with the help of aldose reductase (ALR2), that requires NADPH as the cofactor. The second step is the reduction of sorbitol to fructose by sorbitol dehydrogenase with NAD+ as the cofactor [3] (Fig. S1). Kinetic and structural studies showed that glucose acts as the substrate for ALR2 under intracellular hyperglycemic condition. Accumulated sorbitol causes osmotic stress, the underlying mechanism of DMCs like neuropathy, nephropathy and retinopathy. Thus, the inhibition of ALR2 is an effective approach for prevention of DMCs. ALR2 has been gaining increasing attention in the last two decades as a promising therapeutic target, as it is involved in the etiology of a variety of pathologies that comprise major health problems of the 21st century [4–12]. Intense efforts have been directed toward the development of effective aldose reductase inhibitors (ARIs) that can be effective in addressing diabetes microvascular complications (DMCs) though with little success. Thus, it is very challenging to develop novel chemotypes with rich pharmacokinetic profile and fewest side-effects. Many ALR2 inhibitors have successfully reached various clinical trial phases but most of them were withdrawn either due to severe adverse effects produced by them or their inefficacy [13].

Aldose reductase (ALR2) is a monomeric (α/β)8-barrel protein of aldo-keto reductase (AKR) super family and a close paralog of aldehyde reductase (ALR1) [14]. ALR1 and ALR2 have similar molecular weight and substrate specificity. Moreover the two enzymes are supposed as isoenzymes [15, 16]. Both the enzymes share 65% sequence identity between them and the catalytic active site residues like Tyr48, His110 are also conserved for two enzymes [17]. Consequently, many ALR2 inhibitors like sorbinil and tolrestat also inhibit ALR1 [18]. Besides, ALR2 and ALR1, tolrestat is also present as a cocrystal complexed with AKR1B10 [19]. This non-selective binding seems the primary cause for the side effects produced by ALR2 inhibitors [20]. Hence the newly designed inhibitors must be selective for ALR2 over ALR1 and AKR1B10 to eliminate their undesired effect. In addition to sorbinil like non selective inhibitors, there are other ALR2 inhibitors, which are more selective toward ALR2. The adverse effect associated with the latter type of ALR2 inhibitors remains to be an obstacle in the drug development process. The withdrawal of selective ALR2 inhibitors from clinical trial studies encourages us to figure out the exact cause for toxicity associated with current ALR2 inhibitors.

The structural analysis of ALR2 revealed that the active site residues are more suitable for making hydrophobic interactions with the substrate rather than the H-bonding interactions which is characteristic of the majority of the sugar binding proteins [21, 22]. Literature reveals that ALR2 catalyzes the reduction of toxic metabolite products of the membrane lipids. The inhibition of ALR2 leads to the accumulation of lipid peroxidation products contributing to cytotoxicity [23]. Thus, in normal physiological conditions, the role of ALR2 may be detoxification of toxic metabolites instead of reduction of glucose. For example, methylglyoxal, a toxic 2-oxo-aldehyde is detoxified via reduction in the presence of ALR2. Hence the alteration of aldehyde-detoxifying role of ALR2 may be one of the reasons for withdrawal of current ALR2 inhibitors. Therefore, the aim of the current study is to focus on the selective inhibition of ALR2 to prevent binding of glucose to the active site region without interfering with the detoxification role of ALR1, ALR2 and AKR1B10.

To explore the above aspects further, we carried out preliminary comparative crystal structural analysis of ALR2 complexes. Studies of various high resolution ALR2 co-crystal structures enable us to look into the active site of ALR2 in a more detailed manner (Fig. 1). Thr111 residue demarcates the active site in two main regions. The rigid anion binding site consists of Tyr48 and His110, the charged catalytic residues, responsible for the detoxification mechanism of ALR2 [24]. The other region consists of highly flexible C–terminal loop, which extends when inhibitors like zopolrestat with an extra aromatic ring system bind to it. The latter leads to opening of an induced cavity region in ALR2. The comparative crystal structure analysis of human proteins similar to human ALR2 revealed that the induced cavity region is selective for ALR2 over other close paralogs like ALR1 and AKR1B10 [25].

In the present study, we have designed the new ALR2 inhibitors, which bind to the induced cavity region and leave the anion binding site free for detoxification. The cross docking analysis proves the concept that the newly designed inhibitors have better binding profile for ALR2 over ALR1 and AKR1B10. Moreover, the newly designed inhibitors prevent the binding of glucose to ALR2 thus avoiding its reduction to harmful sorbitol. This study supports the optimization of withdrawn ALR2 inhibitors and thus aids in drug development process. In the future, the newly designed inhibitors will be checked for their inhibitory activity and the lead molecules will further be optimized.

Materials and methods

Sequence based profile search

Human ALR2 protein sequence information was retrieved from Kyoto encyclopedia of genes and genomes) (KEGG)

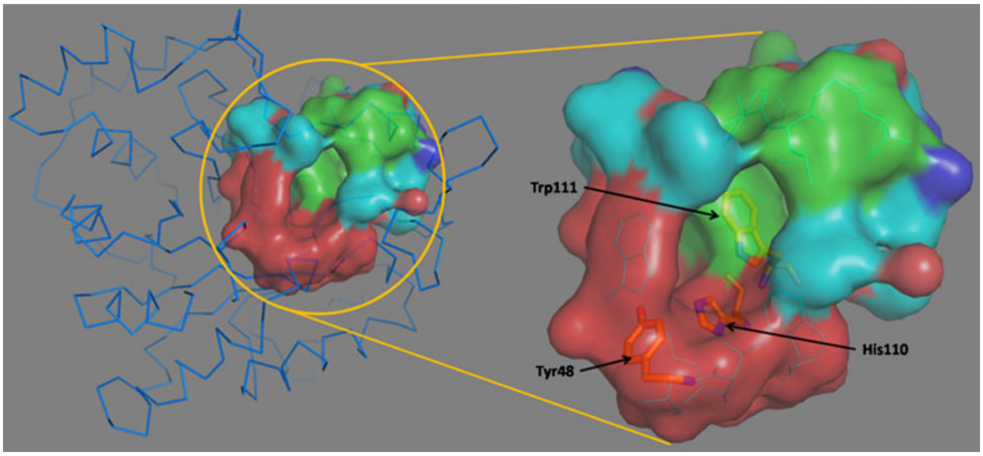


Fig. 1 Human ALR2 (ribbon model) showing the active site region in surface view. The red color corresponds to anion binding region and green color represents the induced cavity region (Left). On the right, enlarged view of active site region is displayed. The catalytic residues (*i.e.*, Tyr 48 and His 110) in anion binding region are displayed in stick model. The rest of the residues are shown in line model. Trp111 demarcating the two cavities is also shown in stick model

(KEGG I.D. → hsa:231) gene database [26]. Sequence similarity Database (SSDB) of KEGG was used to characterize the orthologs to human ALR2 in various organisms [27]. The orthologs from higher eukaryotes with greater than 70% sequence identity to ALR2 were exclusively considered for further analysis (Table S1). Thus screened 14 protein orthologs and human ALR2 were subjected to multiple sequence alignment (MSA) using ClustalX program [28]. The generated multiple sequence alignment was used to build Hidden Markov Model (HMM profile) using HMMER [29, 30]. A default E-value of 10 was taken to develop HMM profile. The human proteome was downloaded from the NCBI database and the HMM profile generated in the previous step, was used as query to search against human proteome. A total of 18 significant hits obtained in this study were used for further analysis. The hits include eight AKR super family members along with 10 hypothetical proteins. The hits were then categorized in various groups based on the sequence similarity to various isoforms of AKR family members (Table S2).

Comparative crystal structure analysis

The comparative crystal structure analysis involved two major aspects: the detailed study of human ALR2 structures bound to various inhibitors and the comparison of ALR2 with other AKR superfamily members (hits obtained in sequence based profile search). For the first study, the high resolution human ALR2 co-crystal structures were downloaded from PDB [31]. All structures were imported in SYBYL7.1 [32] and aligned to the holoenzyme (PDB code: 1ADS) structure. The RMSD was calculated with respect to Cα atoms. To analyze the induced cavity region, cavity volume analysis of various co-crystal structures was performed using Q-SiteFinder [33].

To examine the difference between ALR2 and hits obtained from sequence based profile search, the crystal structures of representative proteins (bolded in Table S2) were downloaded from the protein databank (PDB) [31] and were aligned to ALR2 complexed with zopolrestat (2FZ8) with respect to Cα atoms in SYBYL7.1 [32].

ALR2: rational drug design and selectivity analysis

The rational drug design for the ALR2 induced cavity region involved two major steps, *i.e.*, the design of the core region for induced cavity region to make the molecules selective for ALR2 followed by elongation with various linkers to optimally occupy the catalytic site region.

Various heteroaromatic ring systems were tried for binding to induced cavity region. The X-ray crystal structures of ALR2 bound with zopolrestat (PDB ID: 2FZ8) was retrieved from RCSB protein databank. The hydrogens were added and the protein was minimized using protein preparation wizard of Schrödinger. Charge on NADP was corrected manually. The designed molecules were sketched and minimized with LigPrep using the OPLS2005 force field. The molecules were docked into the active site of ALR2 using Glide (version 9.0, Schrödinger, Inc.) in standard precision mode (Glide SP) [34]. The binding region was defined by a 20Å box centered around the bound zopolrestat in the active site region to confine the centroid of the docked ligand. The van der Waals scaling factor for the nonpolar atoms was set to 0.8 to allow for some flexibility of the receptor. Default settings were used for the rest of the parameters. Moreover, no constraints were included during the grid generation. The top 20 poses were generated for each ligand. The docking poses were then energy minimized with Macromodel in the OPLS2005 force field, with flexible ligand and rigid receptor. Best pose was selected on the basis of Glide score and the interactions formed between the ligands and receptor's residues.

The benzothiazole ring portion of zopolrestat was first docked into the ALR2 followed by various modifications in it. The ring systems binding well in the induced cavity region, were further considered for elongation step. The linkers were designed keeping in mind that they should not interact in the anion binding region. The newly designed molecules with connected linkers were then docked in ALR2 using Glide SP (details mentioned previously).

For validation of toxicity issue, we did the cross-docking of newly designed ALR2 inhibitors in ALR1 and AKR1B10. The crystal structures for ALR1 (PDB ID: 2ALR) and AKR1B10 (PDB ID: 1ZUA) were downloaded from RCSB protein databank. The docking protocol was the same as in case of ALR2 docking. The proteins were prepared using protein preparation wizard of Schrödinger. The charge on NADP was corrected manually. The rest of the parameters were kept the same as in ALR2 docking to accurately compare the result. The best docked pose was selected on the basis of Gscore and the protein-ligand interactions.

Toxic metabolites: binding studies in ALR2-new inhibitors complex

The bound zopolrestat was extracted from the ALR co-crystal structure (PDB ID: 2FZ8). This structure devoid of zopolrestat (*i.e.*, the holoenzyme), and the one in which PI_1 was docked were taken for this study. The cofactor (NADP) was kept intact in both the cases. The binary (ALR2-NADP) and ternary (ALR2-NADP-PI_1) protein complexes were prepared for docking using protein preparation wizard of Schrödinger. The charge on NADP was corrected manually. Glucose-6-phosphate, and methylglyoxal molecules were considered as the representative of toxic metabolites. As the glyceraldehyde-3-phosphate is the substrate for ALR2 in the case of in vitro assay procedure, we also included this molecule for our study. The three molecules were sketched and minimized using LigPrep. The prepared molecules were docked into the active site of ALR2-NADP and ALR2-NADP-PI_1 complexes using Glide (version 9.0, Schrödinger, Inc.) in standard precision mode (Glide SP). The binding region was defined by a 20Å box centered around the active site residues, *i.e.* Tyr48, His110 and Lys77 to confine the centroid of anion binding site. No scaling factors were applied to the van der Waals radii. Default settings were used for all the remaining parameters. The top 20 poses were generated for each ligand. The docking poses were then energy minimized with Macromodel in the OPLS2005 force field, with flexible ligand and rigid receptor. The best pose was selected on the basis of Glide score and the interactions formed between the ligands and hinge region amino acids.

Novel ALR2 inhibitors: ADME prediction

The QikProp program was used to obtain the ADME properties of the newly designed inhibitors [35]. This predicts both physically significant descriptors and pharmaceutically relevant properties. The newly designed inhibitors were prepared by LigPrep and submitted to QikProp module of Schrödinger. The program was processed in normal mode, and predicted 44 properties for the molecules, consisting of principal descriptors and physiochemical properties like QPPCaco, #metab, % absorption and PSA etc. We have considered the MW, QPPCaco, #metab, % absorption, CNS activity and PSA for our analysis.

Results and discussion

Sequence based profile search

To get rid of the adverse effects associated with current ALR2 inhibitors, it is needful to ascertain the human proteins which are similar to ALR2. Sequence based profile search is an effective approach to enlist the proteins similar to ALR2 in human proteins. Total 14 enzymes were obtained in SSDB-KEGG search as ortholog sequences to human ALR2. The sequences are from eukaryotes, with more than 70% sequence identity to human ALR2. Most of the hit sequences correspond to AKR superfamily members from various organisms (Table S1). The best hit sequence was ALR2 from Rhesus monkey (denoted by mcc which stands for Macaca mulatta; 705957 represent the entry number) with identity of 0.98 (Table S1). Multiple sequence alignment (MSA) of human ALR2 protein and the rest of the 14 sequences are shown in Fig. S2. The active site residues are conserved among all sequences. The multiple sequence alignment was used to generate HMM profile. The HMM profiles are statistical models of multiple sequence alignments, which are a well-suited methods for searching databases that uses multiple sequence alignments instead of single query sequences. Thus developed HMM profile, was employed for HMM search against human proteome. The details of various human proteins obtained in HMM search are given in Table S2. The best scoring hit obtained was from aldo-keto reductase family 1 member, B10, *i.e.*, AKR1B10 with a score of 734.3 and it has the least E-value, *i.e.*, 3.10E-217. The HMM score obtained is related to the statistical significance of the alignment. A score of zero is marginal according to the model's statistics. The higher the score, the better the alignment. The other hits are AKR1A1 (*i.e.*, ALR1), AKR1D1, AKR1C1-2 etc. (Table S2). Thus screened, highlighted proteins from various groups represent that the human proteins have high similarity to human ALR2 sequence and thus may be responsible for toxicity caused by current ALR2 inhibitors. In the light of the above analysis, it is necessary to consider the structural properties of all representative proteins to neglect the toxicity issues associated with ALR2 inhibitors.

Comparative crystal structure analysis

The existence of induced fit phenomenon in ALR2 and its role in inhibitor selectivity encourages us to do the comparative crystal structure analysis for ALR2. ALR2 is a structurally well explored target and more than 100 crystal structures are available in the protein data bank (PDB). The comparative crystal structure analysis was performed to study the characteristics of induced fit cavity region and to choose the best structure out of a hundred ALR2 structures for further drug designing. The overall structure of the human ALR2 folds into an eight-stranded α/β-barrel, made up of 315 amino acids. The active site is located at the C terminal end of the barrel and cofactor (NADP+) binding site is located at the adjacent to the active site region. The active site is demarcated by Trp111 into the catalytic anion binding site and the induced cavity region. The induced cavity adopts multiple conformations based on the nature of bound ligand. The deeply buried catalytic site consists of Tyr48, Lys77, His110 and Trp111 residues.

Ligands with appropriate aromatic ring system frequently provoke “induced-fit” adaptations of the induced cavity making the Trp111 indole moiety to face the protein core and Ala299, Leu300, and Phe122, to adopt different rotameric states (Fig. 2). Interestingly, the induced cavity emerges due to the effect of conformational changes in loop A which is made up of residues 121–135 and another short

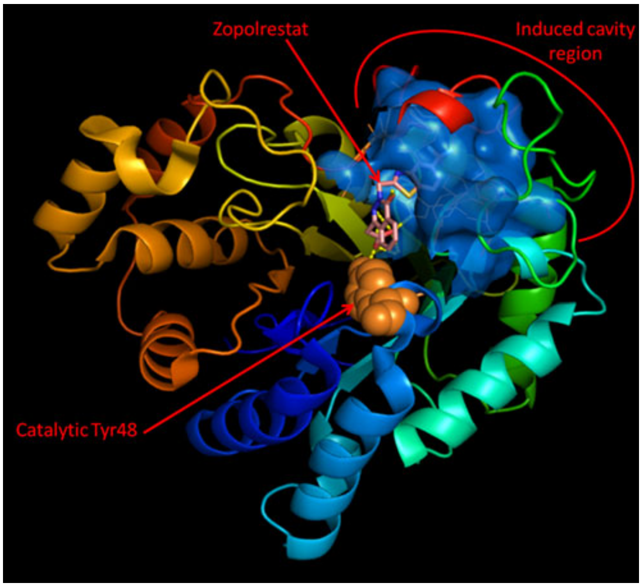


Fig. 2 Zopolrestat bound ALR2 structure: the eight α/β-barrel structure, catalytic Tyr48 (shown in sphere model), induced cavity region (semi-transparent marine blue surface) and bound zopolrestat (stick model) are displayed

segment of loop C which is made up of residues 298–303. Due to this, Phe122 participates in the ligand interactions and Leu300 acts as a gate keeper between the open and closed conformation. In complex crystal structure of ALR2 with fidarestat (PDB code: 1PWM) the amide group of fidalrestat interacts with Leu300 of the short segment which is susceptible to conformational change. This hydrogen bonding makes fidalrestat selective for ALR2. S-(1, 2-dicarboxyethyl) glutathione (DCEG) conjugated bound ALR2 structure (PDB code: 2F2K) revealed that carboxylate terminal of glutathione interacts with Tyr48 and His110 of anion binding region while the C-terminal of DCEG binds to the ALR2 loop C. The analysis of DCEG bound ALR2 suggests that we can prevent the interaction with catalytic residues by truncating the carboxylate and similar groups in newly designed ALR2 inhibitors. It is observed that the naphthalene group of tolrestat and the benzothiazole moiety of zopolrestat fit into the hydrophobic pocket of ALR2. It suggests that the aromatic ring system is crucial to induce the cavity in ALR2. The volume analysis of ALR2 co-crystal structure using Q-SiteFinder reveals that in the case of zopolrestat bound ternary structure, the volume is maximal, i.e., 191 Å^3 (in all other cases the maximal volume ranges from 120 Å^3 to 170 Å^3). Since the active site consists of a rigid catalytic portion and a flexible induced cavity region and hence, the difference among cavity volume is directly correlated to the volume of the induced cavity. In the case of zopolrestat bound ternary structure, the maximum volume of active site suggests that the induced cavity opens to the maximal extent in the case of zopolrestat bound ALR2 structure. Thus, zopolrestat bound structure is found to be the best structure to design specific ligands for ALR2 induced cavity.

Since zopolrestat induces maximum change in the induced cavity region, it is mandatory to study the structural features of zopolrestat as well. The analysis of zopolrestat bound ALR2 structure (PDB ID: 2FZ8) revealed the essential interactions with the induced cavity region. The pharmacophoric features of zopolrestat were analyzed using the phase module of Schrödinger [36] (Fig. 3). The presence of aromatic ring system in zopolrestat and complementary Trp111 conveys the importance of aromatic system for induced cavity binding. The terminal trifluoro- group and corresponding Thr113 in ALR2 suggests the importance of electronegative substituent for H-bonding with ALR2. The presence of electron withdrawing feature in zopolrestat and corresponding backbone NH- of Leu300 residue in ALR2 suggest the requirement of H-bonding of Leu300 for strong binding to ALR2 induced cavity. Besides the interactions with the residues of induced cavity region, the terminal carboxylic group of zopolrestat interacts with the catalytic residues of ALR2 thus altering the physiological detoxification role of ALR2. The newly designed ALR2 inhibitors must be devoid of any substituent corresponding to terminal carboxylic group of zopolrestat preventing the interaction with the catalytic site residues. The comparative analysis of various ALR2 cocrystal structures helps in exploring the induced cavity region in detail which was useful while designing new ALR2 inhibitors.

To eliminate the toxicity related issue, we also performed the structural analysis of hit proteins (obtained in sequence based profile search) *w.r.t.* ALR2. For all representative

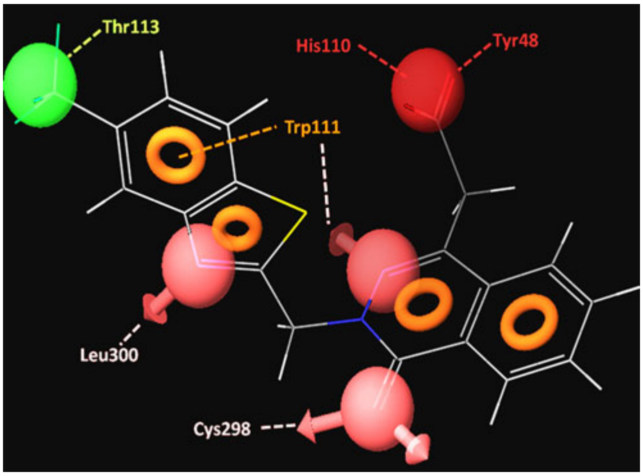


Fig. 3 Pharmacophoric features of zopolrestat. Red → hydrogen bond acceptor, blue → hydrogen bond donor, orange → aromatic and green → hydrophobic. Various interactions with protein residues are displayed as dotted lines

proteins of each group aligned to ALR2, RMSD value was less than one (Table S2). AKR6A was an exception with RMSD value of 2.348 thus not considered for further analysis. The significant difference in all aligned proteins was observed in C-terminal region (Fig. S3). The induced cavity of ALR2 lies in the same C-terminal region. The structural variation among all aligned proteins was observed in other portions of the enzymes as well. Since these locations were distant from active site region, these were not considered for further analysis. The structural analysis revealed that there is no cavity region in many of the aligned proteins corresponding to the induced cavity of ALR2. (Fig. S3). It is only in the case of AKR1B10 that the similar cavity was observed. Further alignment of AKR1B10 and ALR2 revealed that in AKR1B10, Gln114 occupies the position of Thr113 of ALR2 (Fig. S4). Besides Gln114, Leu300 andCys303 of ALR2 were replaced by Val300 andMet303 of AKR1B10 respectively. In light of the above discussion, it is clear that although AKR1B10 consists of the cavity corresponding to induced cavity region of ALR2, yet one can selectively target the ALR2 induced cavity due to structural difference in the cavity region of two enzymes.

As many ALR2 inhibitors are reported to show adverse effect as a result of non-selective inhibition of ALR1. Moreover, AKR2 inhibitors like sorbinil and tolrestat are also present as co crystallized ligands in ALR1. Thus, we took care of ALR1 and AKR1B10 both while designing new ALR2 inhibitors. The comparative structure analysis leads to the following conclusions which were considered during rational drug design: 1) ALR2 induced cavity region is selective for inhibitors binding and is hydrophobic in nature, 2) Trp111 is the aromatic residue in induced cavity which is in favorable orientation to make π- π interaction with the small molecules, 3) Hydrogen bonding with Thr113, Cys303 and with backbone NH- of Leu300 makes inhibitors selective to ALR2 over ALR1 and AKR1B10, thus minimizing toxicity related issues, 4) The anion binding site should be optimally occupied so that the linear toxic metabolites can bind to catalytic residues and at the same time preventing bulkier molecules like glucose binding to catalytic residues.

ALR2: rational drug design and selectivity analysis

The comparative volume analysis disclosed the volume difference in induced cavity region for various ALR2 co-crystals. The significance of induced cavity region for selectivity and specificity issues and the existence of maximum volume in induced cavity region for zopolrestat bound ALR2 (PDB code: 2FZ8) encourages us to use zopolrestat bound ALR2 structure for the purpose of designing new ALR2 inhibitors.

The newly designed inhibitors consist of heteroaryl ring systems connected to optimal linker via a carbonyl linkage (Table S3). These inhibitors have a high affinity toward the induced cavity region without any interaction with the catalytic residues, *i.e.*, with Tyr48 and His110. For most of the new ALR2 inhibitors, the binding affinity is in good range of −8 to −10 kcal mol^{-1} (Table 1). Interestingly the linker portion of designed inhibitors partly occupies the anion binding region (Fig. 4a). Thus partly occupied anion binding site is supposed to prevent the entry of bulky

Table 1 Newly designed ALR2 inhibitors, their docking scores and respective H-bonding residues in various proteins

Molecule	ALR2		AKR1B10		ALR1	
	Gscore	H-bonding residues	Gscore	H-binding residues	Gscore	H-binding residues
PI_1	−10.49	111,113,300	−6.13		−6.36	
PI_2	−10.13	113,113,300	−6.93	298	−7.03	299
PI_3	−9.72	111,113,300	−6.84	301	−6.63	122
PI_4	−9.51	113,300,303	−6.22	301,303	−7.03	80
PI_5	−9.43	111,113,300	−6.4	303	−6.48	299
PI_6	−9.18	113,300,303	−6.96	125	−6.7	122
PI_7	−8.92	111,113,300	−6.22		−6.29	299
PI_8	−8.89	113,300,303	−7.07	303	−6.64	299
PI_9	−8.88	111,113,300	−7.47	303	−7.49	80
PI_10	−8.62	113,300,303	−8.06	303	−7.22	299
PI_11	−8.61	113,300,303	−7.04		−7.15	300
PI_12	−8.54	111,113,300	−7.18	125	−7.09	122
PI_13	−8.53	113,300,303	−7.5	298	−7.42	300
PI_14	−8.39	111,113,300	−6.68		−6.5	122
PI_15	−8.22	111,113,300	−7.34	303	−6.99	122

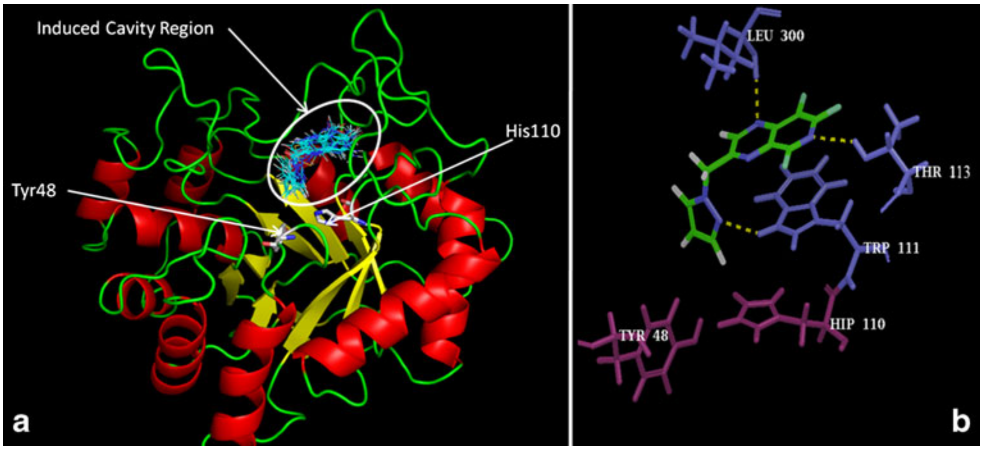


Fig. 4 ALR2 (PDB code: 2FZ8) (**a**) containing designed molecules docked in the induced cavity region leaving Tyr48 and His110 free for detoxification, (**b**) Docking pose of HI_1 in the induced cavity region

glucose and at the same time allow detoxification of linear toxic metabolites. The latter point is further validated in binding studies of toxic metabolite section.

The hydrogen boding residues correspond to Thr113, Leu300 and Cys303 in the induced cavity region. In addition to H-bonding, all inhibitors make strong π-π interactions with Trp111. In Fig. 4b, best docked pose of PI_1 inhibitor is shown. PI_1 has the docking Gscore of −10.49 kcal mol^{-1} and it makes H-bonds with Thr113, and with backbone nitrogen of Leu300. It also shows hydrogen bonding and π-π interaction with Trp111. None of the newly designed inhibitor interacts with the catalytic residues, *i.e.*, Tyr48 and His110 and hence detoxification role of ALR2 should not be interrupted.

Zopolrestat has better binding affinity to ALR2 (Gscore −14.31 kcal mol^{-1}) over our designed inhibitors. The study of zopolrestat bound ALR2 structure suggested the presence of an extra aromatic ring system in zopolrestat. Zopolrestat binds to anion binding region as well and interacts with catalytic Tyr48 and His110 residues thus altering the physiological detoxification mechanism of ALR2. Although the docking score of the designed molecules is less than that of zopolrestat, yet we have overcome the adverse effect issue by keeping catalytic residues free for detoxification. Moreover, the docking score for our designed molecules are in good range (−8 to −10 kcal mol^{-1}) which can further be optimized in future.

Besides keeping the anion binding site free for detoxification, the newly designed ALR2 inhibitors should also not bind to proteins which are highly similar to ALR2 to get rid of adverse effects. AKR1B10 and ALR1 are the close analogs of ALR2 predicted from sequence based profile search. The lower binding score of designed inhibitors for ALR1 (Gscore −6.3 to 7.5 kcal mol^{-1}) and AKR1B10 (Gscore −6.1 to 8.0 kcal mol^{-1}) suggests the lesser binding affinity of the newly designed inhibitors for close analogs of ALR2 thus further minimizing the risk of adverse effects.

Toxic metabolites: binding studies in ALR2-new inhibitors complex

All docking calculations were performed using the “standard precision” (SP) mode of Glide program and with OPLS-AA2005 force field. All three toxic compounds in the study were docked in the anion binding region of ALR2-NADP and ALR2-NADP-PI_1 complexes in separate cases and the binding interactions were calculated. The estimated docking scores (Gscore) by the algorithm for these compounds are listed in Table S4. As shown in Fig. 5c, glucose-6-phophate interacts in the anion binding site region of ALR2-NADP complex with a significant docking Gscore of −8.54 kcal mol^{-1}(Table S4). However, in the case of the ALR2-NADP-PI_1 ternary complex, glucose molecule could not access the catalytic pocket, instead it remains on the surface of the protein and makes irrelevant interaction with Ser302 (Fig. 5c). From the comparative docking analysis, it is clear that glucose-6-phosphate binds to ALR2-NADP complex with very high affinity and its affinity decreases significantly in the case of ALR2-NAPD-PI_1 ternary complex. (Table S4). Hence, our designed molecules prevent the reduction of glucose-6-phosphate to harmful sorbitol formation.

Furthermore, there is no significant difference for the binding affinity of glyceraldehyde-3-phosphate and methylglyoxal in ALR2-NADP and ALR2-NADP-PI_1 complexes (Table S4). As shown in Fig. 5, the molecules make interaction in the anion binding region in both cases and there is similar H-bonding pattern between these molecules and Tyr48 and His110 residues. The binding pose of the molecules was also the same in both cases (*i.e.*, the binary and ternary ALR2 complex).

Based on the above observation, we can conclude that our designed inhibitors do not alter the detoxification role of ALR2 and at the same time, prevent the catalytic reduction of glucose to sorbitol.

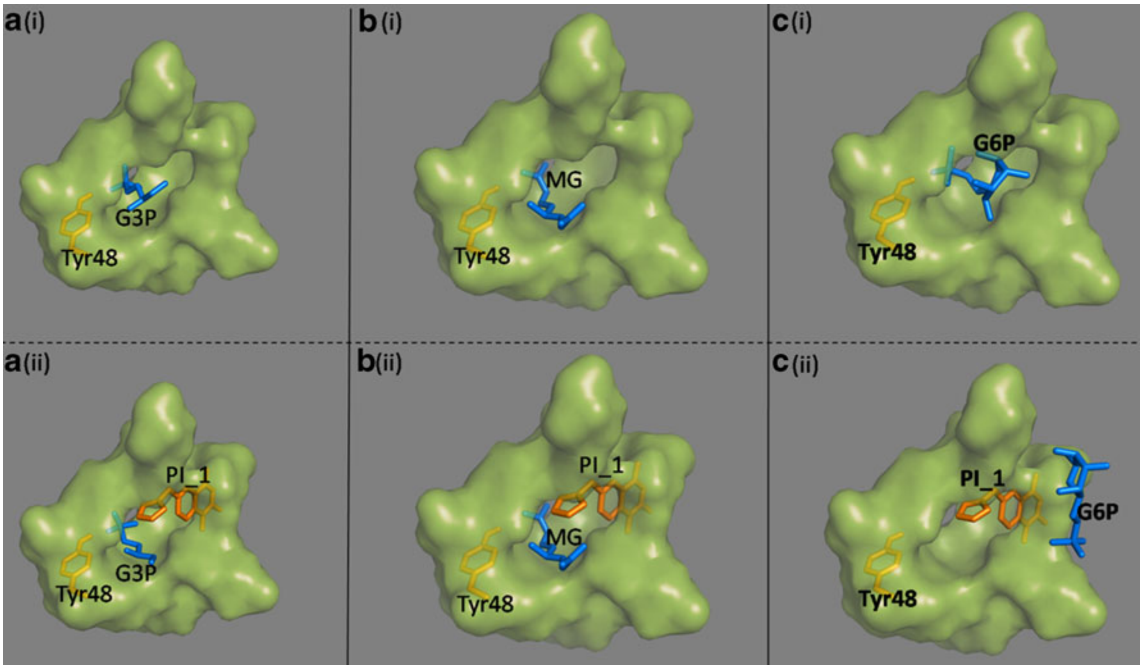


Fig. 5 GlideSP docking of (**a**) Glyceraldehyde-3-phosphate (G3P), (**b**) methylglyoxal (MG) and Glucose-6-Phosphate(G6P) in ALR2 [a (i) b(i) and c(i)] and in PI_1 bound ALR2 complex [a(ii), b(ii) and c (iii)]. Only the active site region of ALR2 is shown in surface view highlighting catalytic Tyr48 residue in stick model

Novel ALR2 inhibitors: ADME prediction

Potential drug candidates fail in the clinical trials due to their poor pharmacokinetic profile. Thus, the ability to predict the ADME (absorption, distribution, metabolism and excretion) profile would have a great impact on the drug discovery process. Our designed inhibitors show good ADME profile (Table S5). As far as the CNS activity is concerned, our designed molecules are presumably better than the already reported inhibitors and in few cases it is −2. Out of 15 molecules 10 molecules showed more than 70% absorption. The predicted polar surface area varies in acceptable range from 76 to 160, which indicates better water solubility for newly designed inhibitors. The passive diffusion representative QPPCaco value varies widely in our designed molecules from a very low value in the case of PI_11 to a very large value in case of PI_1. All other molecules except PI_11 lie within the satisfactory range. The designed molecules are also predicted to be metabolically stable as the #metab value in the case of the designed inhibitors ranges from 1 to 3. Thus, based on the above discussion, it is clear that our designed inhibitors have good ADME profile, which can further be optimized in future.

Conclusions

In the present study, we tried to understand the molecular mechanism behind the failure of current ALR2 inhibitors. We had carried out a sequence based profile search which revealed ALR1B10 and ALR1 as the close paralogs of ALR2, thus may be responsible for the current adverse effect associated with the current ALR2 inhibitors. ALR2 shows the induced fit phenomenon and this induce cavity region is selective for ALR2 over its close analogs. Thus new ALR2 inhibitors would be designed for induced fit cavity region leaving the catalytic site free for detoxification. The comparative crystal structure analysis of all human ALR2 structures revealed that in case of zopolrestat bound ALR2 the induced cavity opens to the maximal extent thus is the best structure to design specific inhibitors for ALR2.

The newly designed ALR2 inhibitors make interactions in the induced cavity region via H-bonding with Trp111, Thr113 and Leu300. Moreover these inhibitors do not bind in the anion binding region. The detailed docking analysis performed on ALR2, complexed with newly designed inhibitors revealed that they do not allow glucose to interact with catalytic residues, *i.e.*, Tyr48 and His110. Furthermore, the toxic metabolites make similar interaction with ALR2 holoenzyme as well as ALR2 bound with newly designed inhibitors. Thus, newly designed inhibitors are have a better binding profile for ALR2 and also they do not interfere with the detoxification mechanism of ALR2. In the future, the molecules will be checked for their inhibitory activity and will further be optimized.

References

1. Stratton IM, Adler AI, Neil HW, Matthews DR, Manley SE, Cull CA, Hadden D, Turner RC, Holman RR (2000) Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. BMJ 321:405–412
2. Gonzalez RG, Barnett P, Aguayo J, Cheng HM, Chylack LT (1984) Direct measurement of polyol pathway activity in the ocular lens. Diabetes 33:196–199
3. Finegold D, Lattimer SA, Nolle S, Bernstein M, Greene DA (1983) Polyol pathway activity and myo-inositol metabolism. A suggested relationship in the pathogenesis of diabetic neuropathy. Diabetes 32:988–992
4. Ramana KV, Willis MS, White MD, Horton JW, DiMaio JM, Srivastava D, Bhatnagar A, Srivastava SK (2006) Endotoxin-induced cardiomyopathy and systemic inflammation in mice is prevented by aldose reductase inhibition. Circulation 114:1838–1846
5. Regenold WT, Kling MA, Hauser P (2000) Elevated sorbitol concentration in the cerebrospinal fluid of patients with mood disorders. Psychoneuroendocrinology 25:593–606
6. Hasuike Y, Nakanishi T, Otaki Y, Nanami M, Tanimoto T, Taniguchi N, Takamitsu Y (2002) Plasma 3-deoxyglucosone elevation in chronic renal failure is associated with increased aldose reductase in erythrocytes. Am J Kidney Dis 40:464–471
7. Meyer WR, Doyle MB, Grifo JA, Lipetz KJ, Oates PJ, DeCherney AH, Diamond MP (1992) Aldose reductase inhibition prevents galactose-induced ovarian dysfunction in the Sprague-Dawley rat. Am J Obstet Gynecol 167:1837–1843
8. Lee KWY, Ko BCB, Jiang Z, Cao D, Chung SSM (2001) Overexpression of aldose reductase in liver cancers may contribute to drug resistance. Anticancer Drugs 12:129–132
9. Luo Y, Zhang J, Liu Y, Shaw AC, Wang X, Wu S, Zeng X, Chen J, Gao Y, Zheng D (2005) Comparative proteome analysis of breast cancer and normal breast. Mol Biotechnol 29:233–244
10. Zorn KK, Jazaeri AA, Awtrey CS, Gardner GJ, Mok SC, Boyd J, Michael J (2003) Choice of normal ovarian control influences determination of differentially expressed genes in ovarian cancer expression profiling studies. Birrer Clin Cancer Res AACR 9:4811–4818
11. Yim EK, Lee KH, Kim CJ, Park JS (2006) Analysis of differential protein expression by cisplatin treatment in cervical carcinoma cells. Int J Gynecol Cancer 16:690
12. Saraswat M, Mrudula T, Kumar PU, Suneetha A, Rao TS, Srinivasulu M, Reddy GB (2006) Overexpression of aldose reductase in human cancer tissues. Med Sci Monit 12:CR525–CR529
13. Alexiou P, Pegklidou K, Chatzopoulou M, Nicolaou I, Demopoulos VJ (2009) Aldose reductase enzyme and its implication to major health problems of the 21 (st) Century. Curr Med Chem 16:734–752
14. Jez JM, Penning TM (2001) The aldo-keto reductase (AKR) superfamily: an update. Chem Biol Interact 130:499–525
15. Wermuth B, Burgisser H, Bohren K, Eur WJP (1982) Purification and characterization of human-brain aldose reductase. J Biochem 127:279–284
16. Ris MM, Wartburg JP (1973) Heterogeneity of NADPH-dependent aldehyde reductase from human and rat brain. Eur J Biochem 37:69–77
17. Bohren KM, Bullock B, Wermuth B, Gabbay KH (1989) The aldo-keto reductase superfamily. cDNAs and deduced amino acid sequences of human aldehyde and aldose reductases. J Biol Chem 264:9547–9551
18. El-Kabbani O, Carbone V, Darmanin C, Oka M, Mitschler A, Podjarny A, Schulze-Briese C, PtC R (2005) Structure of aldehyde reductase holoenzyme in complex with the potent aldose reductase inhibitor fidarestat: implications for inhibitor binding and selectivity. J Med Chem 48:5536–5542
19. Gallego O, Ruiz F, Ardèvo A, Domínguez M, Alvarez R, de Lera A, Rovira C, Farrés J, Parés X (2007) Structural basis for the high all-trans-retinaldehyde reductase activity of the tumor marker AKR1B10. Proc Natl Acad Sci 104:20764–20769
20. Sheetz MJ, King GL (2002) Molecular understanding of hyperglycemia's adverse effects for diabetic complications. Am Med Assoc 288:2579–2588
21. Wilson DK, Bohren KM, Gabbay KH, Quiocho FA (1992) An unlikely sugar substrate site in the 1.65 A structure of the human aldose reductase holoenzyme implicated in diabetic complications. Science 257:81–84
22. Rondeau JM, Tete-Favier F, Podjarny A, Reymann JM, Barth P, Biellmann JF, Moras D (1992) Novel NADPH-binding domain revealed by the crystal structure of aldose reductase. 30:469–472
23. Srivastava S, Dixit BL, Cai J, Sharma S, Hurst HE, Bhatnagar A, Srivastava SK (2000) Metabolism of lipid peroxidation product, 4-hydroxynonenal (HNE) in rat erythrocytes: role of aldose reductase. Free Radical Biol Med 29:642–651
24. Tarle I, Borhani DW, Wilson DK, Quiocho FA, Petrash JM (1993) Probing the active site of human aldose reductase. Site-directed mutagenesis of Asp-43, Tyr-48, Lys-77, and His-110. J Biol Chem 268:25687–25693
25. Costantino L, Rastelli G, Vescovini K, Cignarella G, Vianello P, Del Corso A, Cappiello M, Mura U, Barlocco D (1996) Synthesis, activity, and molecular modeling of a new series of tricyclic pyridazinones as selective aldose reductase inhibitors. J Med Chem 39:4396–4405
26. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27
27. Sato Y, Nakaya A, Shiraishi K, Kawashima S, Goto S, Kanehisa M (2001) Genome Informatics Series 230–231
28. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci 23:403–405
29. Rabiner L, Juang B (1986) An introduction to hidden Markov models. IEEE ASSP Mag 3:4–16
30. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763
31. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S (2002) The protein data bank. Acta Crystallogr D Biol Crystallogr 58:899–907
32. (2005) SYBYL Molecular Modeling Package. Tripos Inc, St. Louis, MO
33. Laurie ATR, Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21:1908–1916
34. GLIDE, version 5.0, Schrödinger Inc, San Diego, CA, USA
35. QikProp, version 3.1,Schrödinger Inc, San Diego, CA, USA
36. Dixon SL, Smondyrev AM, Knoll EH, Rao SN, Shaw DE, Friesner RA (2006) PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. J Comput Aided Mol Des 20:647–671

J Mol Model (2012) 18:1801–1808
DOI 10.1007/s00894-011-1209-y

ORIGINAL PAPER

Assessing the reactivation efficacy of hydroxylamine anion towards VX-inhibited AChE: a computational study

Md Abdul Shafeeuulla Khan · Bishwajit Ganguly

Received: 6 May 2011 / Accepted: 1 August 2011 / Published online: 18 August 2011
© Springer-Verlag 2011

Abstract Oximate anions are used as potential reactivating agents for OP-inhibited AChE because of they possess enhanced nucleophilic reactivity due to the α-effect. We have demonstrated the process of reactivating the VX–AChE adduct with formoximate and hydroxylamine anions by applying the DFT approach at the B3LYP/6-311 G(d,p) level of theory. The calculated results suggest that the hydroxylamine anion is more efficient than the formoximate anion at reactivating VX-inhibited AChE. The reaction of formoximate anion and the VX–AChE adduct is a three-step process, while the reaction of hydroxylamine anion with the VX–AChE adduct seems to be a two-step process. The rate-determining step in the process is the initial attack on the VX of the VX–AChE adduct by the nucleophile. The subsequent steps are exergonic in nature. The potential energy surface (PES) for the reaction of the VX–AChE adduct with hydroxylamine anion reveals that the reactivation process is facilitated by the lower free energy of activation (by a factor of 1.7 kcal mol^{-1}) than that of the formoximate anion at the B3LYP/6-311 G(d,p) level of theory. The higher free energy of activation for the reverse reactivation reaction between hydroxylamine anion and the VX–serine adduct further suggests that the hydroxylamine anion is a very good antidote agent for the reactivation process. The activation barriers calculated in solvent using the polarizable continuum model (PCM) for the reactivation of the VX–AChE adduct with hydroxylamine anion were also found to be low. The calculated results suggest that V-series compounds can be more toxic than G-series compounds, which is in accord with earlier experimental observations.

Keywords Reactivation · OP-inhibited AChE · Alpha-nucleophiles · Density functional calculations · Hydroxylamine anion

Introduction

VX (*O*-ethyl *S*-[2-(diisopropylamino)ethyl] methylphosphonothiolate) is a member of the family of extremely toxic nerve agents known as the V-series. Due to its nonvolatile nature, VX is highly persistent compared to G-series nerve agents [1, 2]. The binding of VX to the enzyme acetylcholinesterase (AChE) leads to cause asphyxiation [3]. VX can enter the system through not only inhalation but also skin penetration [3]. AChE catalyzes the ester hydrolysis of the neurotransmitter acetylcholine (ACh) to end synaptic transmission [4–6]. Inhibition of AChE occurs as a consequence of the phosphylation of the active serine residue with organophosphorus compounds [7–9]. AChE inhibition results in acetylcholine accumulation at cholinergic receptor sites, thereby excessively stimulating the cholinergic receptors. This can lead to various clinical disorders, and occasionally death. Therefore, the reactivation of organophosphorus compound inhibited AChE is required in order to make its catalytically active in the hydrolysis of ACh again. The inhibited AChE may further undergo an "aging" process that normally involves dealkylation or deamidation, depending upon the nature of organophosphorus compounds attacked, and is irreversible in nature [10–12]. Therefore, there is a need to create efficient reactivating agents for OP-

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1209-y) contains supplementary material, which is available to authorized users.

M. A. S. Khan · B. Ganguly (✉)
Analytical Science Discipline, Central Salt & Marine Chemicals Research Institute (Council of Scientific and Industrial Research),
Bhavnagar, Gujarat, India 364 002
e-mail: ganguly@csmcri.org

inhibited AChE. Computational methods offer a way to discover such reactions [3, 12–24] while avoiding exposure to these deadly agents, and are thus very useful for proposing new nucleophiles with superior efficiency for inhibited AChE reactivation. It has been reported that α-nucleophiles such as oximes are capable of reactivating organophosphate–cholinesterase conjugates, giving rise to the free enzyme [22]. In our previous studies, we reported that hydroxylamine anion (NH_2O^-) is an efficient α-nucleophile for the detoxification of organophosphorus compounds such as VX and sarin [20, 21]. In this article, we report the reactivation efficacy of NH_2O^- towards the VX–serine adduct.

Computational methodology

All geometries were optimized using the B3LYP [25–27] density functional and the 6-311 G(d,p) basis set. Harmonic frequency calculations at the same level were used to confirm the stationary points and to calculate thermodynamic corrections. The Gibbs free energy is particularly relevant to calculations of activation energies, and can be obtained from the equation $G = H - TS$. Therefore, the Gibbs free-energy profiles were plotted against the reaction coordinates of geometries involved in the reactivation process at the B3LYP/6-311 G(d,p) level of theory. The reaction coordinates in the present study are bond making between the nucleophilic atom and the phosphorus center and bond breaking between the phosphorus center and the leaving group of VX-inhibited AChE (Scheme 1).

The key equation for calculating rate constants from the Gibbs free energy in the present study is $k = (k_B T/hc^\circ)\, e^{-\Delta^* G^\circ/RT}$, where $c^\circ = 1$ and the appropriate values are simply plugged into the other variables. Single-point calculations were performed at the B3LYP/6-311 + G(d,p) level to get accurate energies using B3LYP/6-311 G(d,p) geometries. Aqueous energies of solvation of the gas-phase structures were determined with the polarizable continuum model (PCM) [28–32]. Intrinsic reaction coordinate (IRC) calculations were performed to connect all of the transition states with their corresponding minima [33, 34]. Wiberg bond orders were obtained from NBO calculations. All quantum chemical calculations were performed using Gaussian 03, revision E.01 [35].

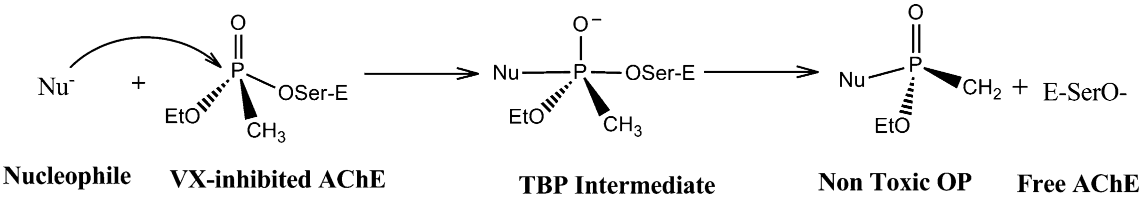
Results and discussion

Conformational analysis

Initially, an extensive conformational search was performed for the VX-serine adduct as a model for VX-inhibited AChE because of its flexibility. The conformational changes associated with the rotations of the C−O (SC1) and C−N (SC2) bonds were analyzed by constructing a two-dimensional potential energy scan using B3LYP/6-311 G** in the gas phase (Fig. 1). To construct the potential energy surface representing the effect of the internal rotations, the C−O and C−N bonds are allowed to rotate 180° in increments of 10°. The two unique lowest-energy conformers, adducts **1** and **2**, were chosen from the potential energy surface, and the energy difference between these two conformers was 2.7 kcal mol^{-1}.

The conformational difference between these two VX–serine adducts is due to the orientation of the –NHCHO group of the serine moiety. The greater stabilization of adduct **1** is due to the strong intramolecular hydrogen bonding between the hydrogen of the –NH group and the oxygen of the phosphonyl group (Fig. 2). This situation is similar to that for sarin-inhibited AChE, as we observed previously [36]. Our close analysis of the crystal structures (PDB IDs: 1VXO and 1VXR) of VX-inhibited AChE revealed that intramolecular hydrogen bonding is not present in any VX-inhibited AChE [37]. To examine the reactivation process of VX-inhibited AChE, we considered adduct **2**, which closely resembles the observed crystal structures.

Recently, hydroxylamine and its anionic form have been of considerable interest due to variations in their structural behavior and reactivity under different reaction conditions [38, 39]. Among α-nucleophiles, the superior reactivity of NH_2O^- with phosphate esters [40] makes it an important candidate for exploring the reactivation process of OP-inhibited AChE. The solvolysis of sarin and VX with hydroxylamine anion has been found to be a very efficient way to detoxify such OP compounds [20, 21]. Recent studies have shown that the reactivation of the sarin-inhibited AChE adduct with nucleophiles involves addition–elimination pathways [22, 24]. The reaction energy profiles generated for formoximate anion and hydroxylamine anion with the VX-inhibited AChE adduct also follow a



Scheme 1 Reaction pathway for the reactivation process along the reaction coordinates

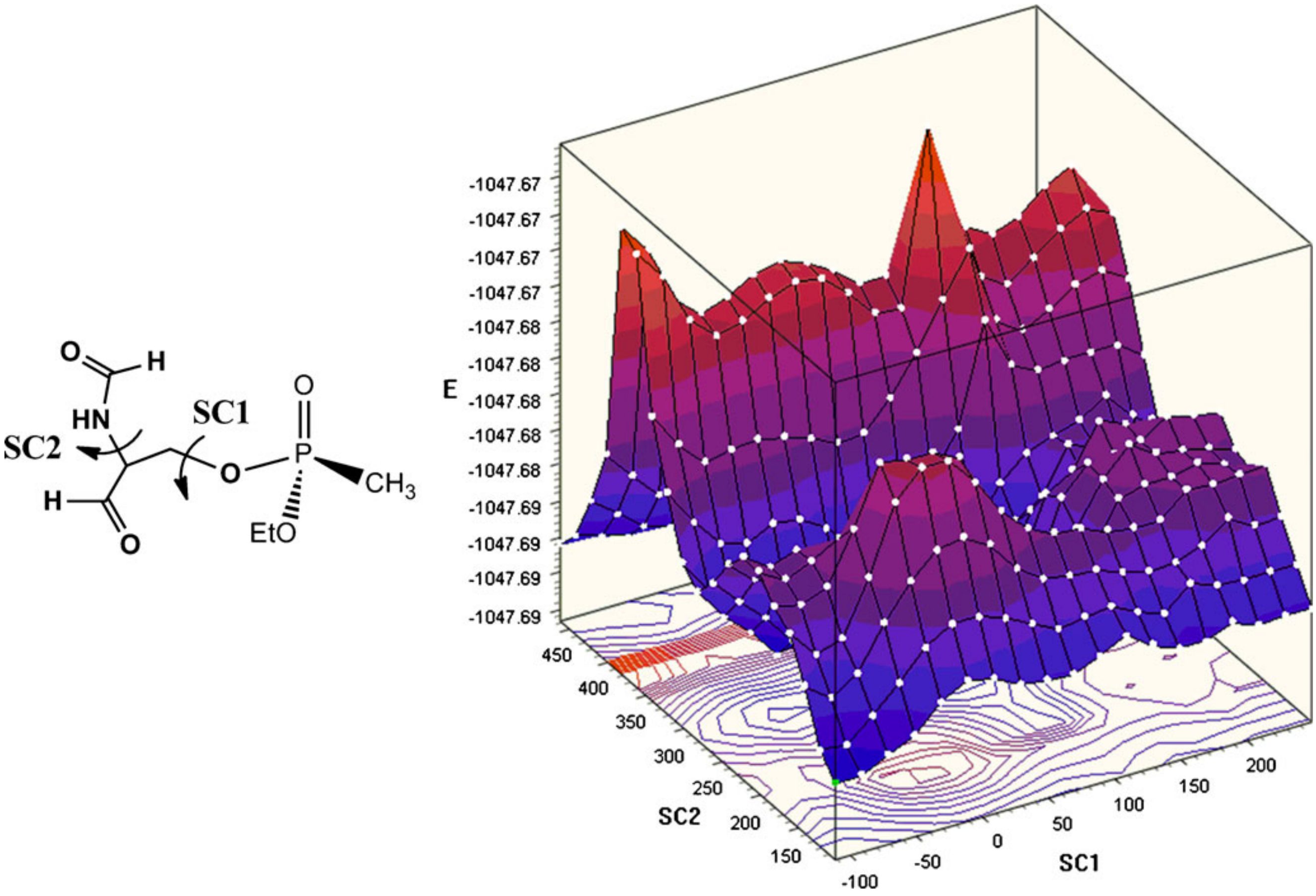


Fig. 1 Two-dimensional potential energy surface of the VX–serine adduct in the gas phase, as calculated at the B3LYP/6-311 G** level of theory

similar addition-elimination pathway involving a trigonal bipyramidal intermediate. In the VX–AChE adduct, we used a serine moiety to emulate AChE, as suggested in the literature [22, 24].

Reactivation with formoximate anion

The reaction energy profile in terms of the Gibbs free energy at the B3LYP/6-311 G(d,p) level for the reaction between the VX–serine adduct and formoximate anion is shown in Fig. 3, and the corresponding stationary points are depicted in Fig. 4. Two complexes and two intermediate structures were located as local minima on the potential energy surface. Three corresponding transition state structures that link these minima were also located as first-order saddle points. The intrinsic reaction coordinate (IRC) calculations connect the transition states to the respective minima.

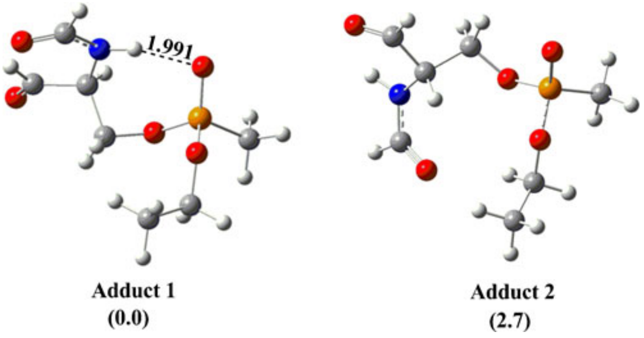


Fig. 2 Geometries of two unique conformers of the VX–serine adduct and their relative energies (kcal mol^{-1}) (*gray* carbon, *red* oxygen, *blue* nitrogen, *white* hydrogen, *orange* phosphorus), optimized at the B3LYP/6-311 G** level of theory

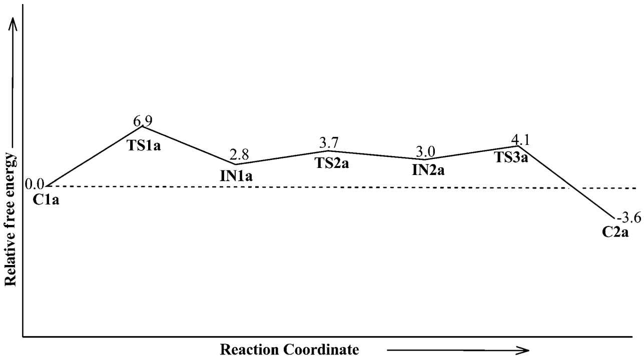


Fig. 3 Free-energy (kcal mol^{-1}) profile diagram for the reactivation of VX–serine adduct **2** with formoximate anion in the gas phase, as calculated at the B3LYP/6-311 G(d,p) level of theory

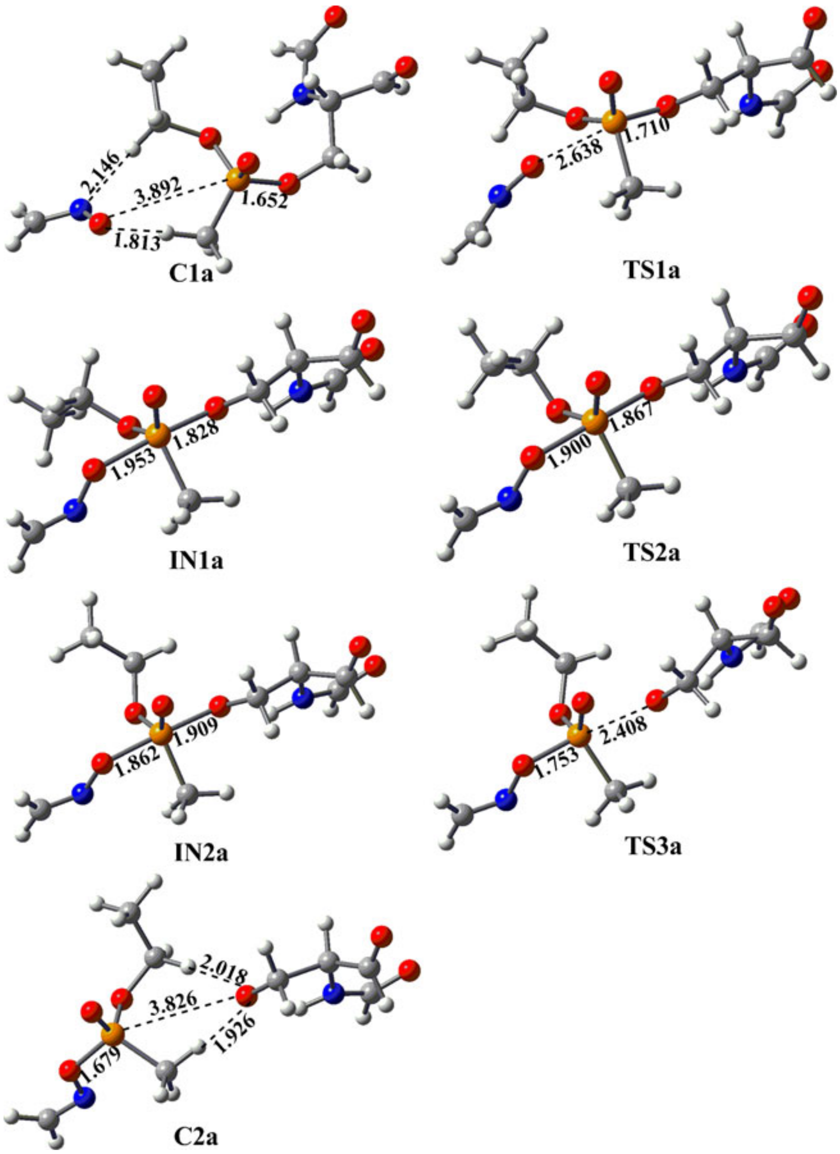


Fig. 4 Geometries optimized at the B3LYP/6-311 G(d,p) level of theory and selected bond distances (Å) for the modeled VX–serine adduct **2** involved in the reactivation process with formoximate anion in the gas phase. *Gray* carbon, *red* oxygen, *blue* nitrogen, *white* hydrogen, *orange* phosphorus

The VX–serine adduct and the formoximate anion form a complex, **C1a**. The anionic nucleophile and VX–serine adduct are stabilized through charge dipole type interactions as well as two C–H…O type hydrogen bonds in complex **C1a** (Fig. 4). The free energy of activation computed for the attack of the formoximate anion on the VX phosphorus atom is 6.9 kcal mol^{-1} compared to complex **C1a**. The formoximate anion approaches opposite to the oxygen atom of the serine moiety in a slightly nonlinear fashion (∠O–P–O=165.0°), and the P–ONCH$_2$ bond distance is 2.638 Å (**TS1a**) (Fig. 4). The Wiberg bond index calculated for the P–ONH$_2$ bond in **TS1a** was found to be 0.09 au, which is about 0.08 au higher than that of **C1a**, signifying a stronger interaction in the former case. After **TS1a**, the TBP intermediate **IN1a** is created, which is 4.1 kcal mol^{-1} stable than **TS1a**. The P–ONCH$_2$ bond distance and that between P and the oxygen atom of serine are 1.953 Å and 1.828 Å, respectively, in **IN1a**. The corresponding bond indices of 0.38 and 0.45 reveal the strengthening and weakening of the corresponding bonds compared to **TS1a**.

To stabilize the leaving group, the ethoxy group rotates toward the serine moiety of the VX–serine adduct through a rotational transition state **TS2a** of imaginary frequency 52i cm^{-1}, and forms another TBP intermediate, **IN2a** (Fig. 4). The elimination of the leaving serine group is exergonic in nature, and requires only 1.1 kcal mol^{-1} of free energy of activation (**TS3a**) from the intermediate **IN2a**. The Wiberg bond index of 0.50 au for P–ONH$_2$ further suggests a strong interaction, and the smaller bond index of 0.13 au for the distance between P and the oxygen atom of serine indicates the expulsion of the leaving group. The forward-direction intrinsic reaction coordinate (IRC) calculation for **TS3a** leads to a complex (**C2a**) between the OP moiety and the leaving group, at a distance of 3.826 Å. The reactivation process that occurs between the formoximate anion and the VX–serine adduct is governed by the first step, and the overall process is exergonic in nature at the B3LYP/6-311 G(d,p) level of theory.

Reactivation with hydroxylamine anion

The reaction energy profile in terms of the Gibbs free energy at the B3LYP/6-311 G(d,p) level for the reaction between the VX–serine adduct and hydroxylamine anion is shown in Fig. 5, and the corresponding stationary points are depicted in Fig. 6. Again, in this case, two complexes and two intermediate structures were located as local minima on the potential energy surface. Three corresponding transition state structures that link these minima were also located as first-order saddle points. The IRC calculations connect the transition states to the respective minima. The VX–serine adduct and the hydroxylamine anion forms a complex, **C1b**. The anionic nucleophile and the VX–serine adduct are stabilized through charge dipole type interactions as well as two C–H…O type hydrogen bonds in the complex **C1b** (Fig. 6). The free energy of activation computed for the attack of NH_2O^- on the VX phosphorus atom is 5.2 kcal mol^{-1} compared to the complex **C1b**. NH_2O^- approaches opposite to the oxygen atom of the serine moiety in a slightly nonlinear fashion (∠O–P–O=164.0°), and the P–ONH_2 bond distance is 2.990 Å (**TS1b**) (Fig. 6). The Wiberg bond index calculated for the P–ONH_2 distance in **TS1b** was found to be 0.05 au, which is about 0.03 au higher than that of **C1b**, signifying a stronger interaction in the former case. In this transition state, N–H…O hydrogen bonding (2.506 Å) was observed between the nucleophile and the –P=O bond of VX-inhibited AChE (Fig. 6) besides the C–H…O type interactions.

After **TS1b**, TBP intermediate **IN1b** was found to be 14.3 kcal mol^{-1} more stable than complex **C1b**. The P–ONH_2 bond distance and that between P and the oxygen atom of the serine are 1.804 Å and 1.833 Å, respectively, in **IN1b**. The corresponding bond indices of 0.52 and 0.46 reveal the strengthening and weakening of the corresponding bonds compared to **TS1b**. The hydrogen-bonding interaction between the nucleophile and the P=O bond of the VX–serine adduct becomes stronger as the H-bond distance shortens (1.961 Å) in **IN1b** (Fig. 6). To stabilize the leaving group, the ethoxy group rotates toward the serine moiety of the VX–serine adduct through a rotational transition state **TS2b** of imaginary frequency 60i cm^{-1}, and forms another TBP intermediate, **IN2b** (Fig. 6). The elimination of the leaving serine group is exergonic in nature and requires a free energy of activation of only 0.3 kcal mol^{-1} (**TS3b**) from the intermediate **IN2b**. The Wiberg bond index of 0.64 au for P–ONH_2 further suggests a strong interaction, and the smaller bond index of 0.11 au for the distance between P and the oxygen atom of the serine indicates the expulsion of the leaving group. The forward-direction IRC calculation for **TS3b** leads to a complex (**C2b**) between the VX moiety and the leaving group, at a distance of 3.817 Å. The reactivation process between the hydroxylamine anion and the VX–serine adduct is also governed by the first step, and the subsequent steps are downhill in nature. We consider that this reaction is mainly a two-step process, as the third step is almost barrierless.

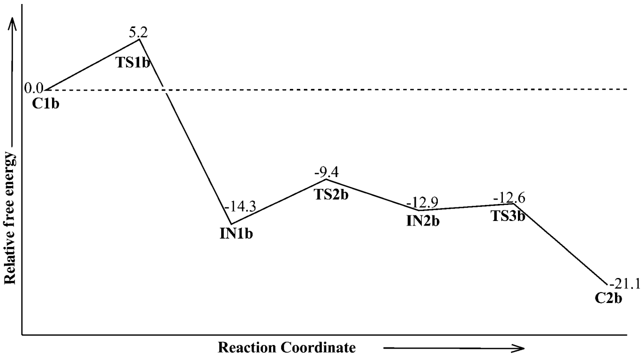


Fig. 5 Free-energy (kcal mol^{-1}) profile diagram for the reactivation of the VX–serine adduct **2** with hydroxylamine anion in the gas phase, as calculated at the B3LYP/6-311 G(d,p) level of theory

The calculated free energy of activation computed at the same level of theory for the reaction of hydroxylamine anion with the VX–serine adduct is 5.2 kcal mol^{-1}, which is 1.7 kcal mol^{-1} less than that of the formoximate anion. We have calculated the rate constants from the Gibbs free energies of activation for the rate-determining steps of the reactions involving the hydroxylamine anion and the formoximate anion with the VX–serine adduct. The first-order rate constants for the reactivation reactions involving the formoximate and hydroxylamine anions are $5.4 \times 10^7\ s^{-1}$and $9.6\times10^8\ s^{-1}$, respectively. The calculated rate constant results further indicate that NH_2O^- should give a reaction that is nearly 20 times faster than that given by the formoximate anion.

The potential energy profile further suggests that the free-energy activation barrier of the rate-determining step for inverse reactivation with NH_2O^- is 26.3 kcal mol^{-1}, which is much higher than the corresponding barrier for inverse reactivation with the formoximate anion (Table 1). These results also indicate that NH_2O^- is a better reactivating agent for the VX-inhibited AChE.

To examine the effect of solvent, and to better describe the anions for the reactions between the nucleophiles CH_2NO^- and NH_2O^- and the VX–serine adduct, single-point calculations at the B3LYP/6-311+G(d,p) level were performed using the polarizable continuum model (PCM) in aqueous solution. The energetics obtained at the B3LYP/6-311+G(d,p)// B3LYP/6-311 G(d,p) level for the reaction between CH_2NO^- or NH_2O^- and the VX–serine adduct

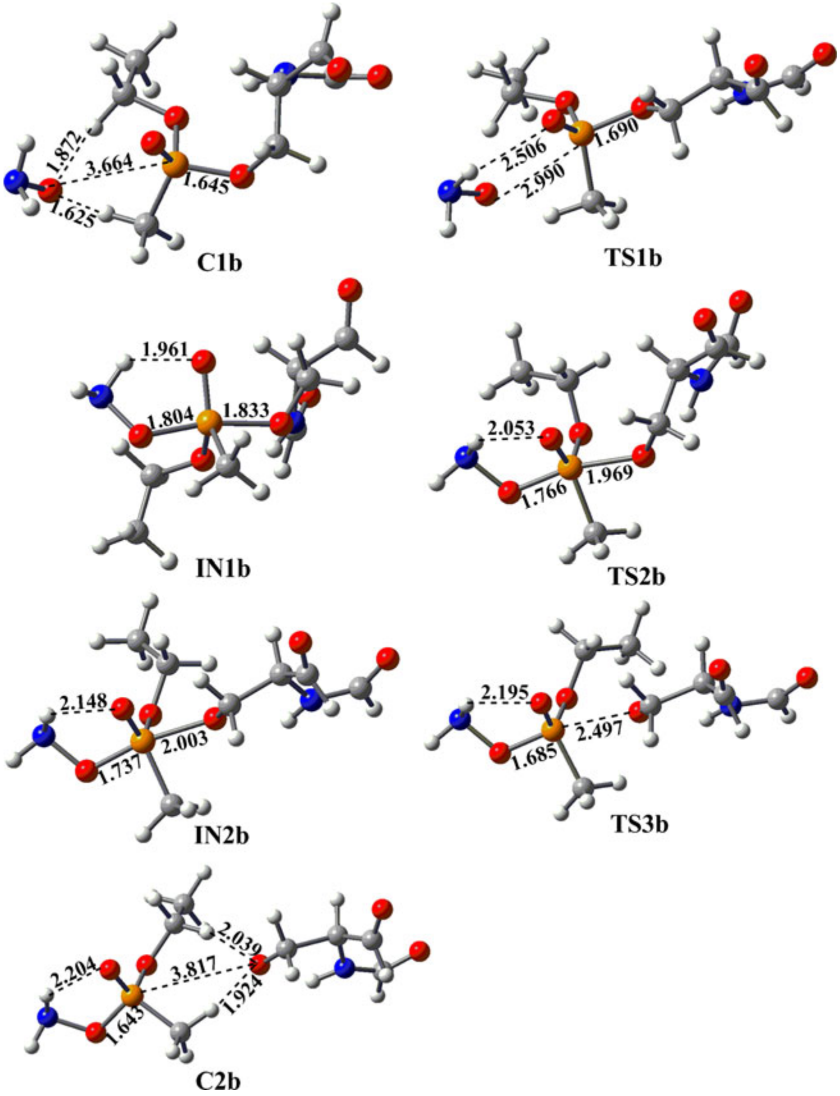


Fig. 6 Geometries optimized at the B3LYP/6-311 G(d,p) level of theory and selected bond distances (Å) for the modeled VX–serine adduct **2** involved in the reactivation process with hydroxylamine anion in the gas phase. *Gray* carbon, *red* oxygen; *blue* nitrogen; *white* hydrogen; *orange* phosphorus)

also suggests that the VX–serine adduct is more easily reactivated by the hydroxylamine anion than by the formoximate anion (Table 2).

The reactivation of the VX–serine adduct with formoximate anion and hydroxylamine anion showed that the hydroxylamine anion is a better reactivating agent. However, note that the calculated free energy of activation is 1.7 kcal mol^{-1} lower with the hydroxylamine anion compared to the formoximate anion. The reactivation of the sarin–serine adduct computed for both nucleophiles suggested similar trends, but reactivation with the hydroxylamine anion was found to be preferred by 4.1 kcal mol^{-1} over formoximate anion [36]. Overall, hydroxylamine

Table 1 Gibbs free-energy activation barriers at the B3LYP/6-311 G (d,p) level in the gas phase for the direct and inverse reactivation reactions in kcal mol^{-1}

Reactivation	NH_2O^-	Oximate
Direct	5.2	6.9
Inverse	26.3	10.5

Table 2 Electronic energies (kcal mol^{-1}) computed at the B3LYP/6-311+G(d,p) level of theory in the aqueous phase using B3LYP/6-311 G(d, p) gas phase optimized geometries for the reactivation of the VX–serine adduct with either the formoximate or hydroxylamine anion

B3LYP/6-311+G(d,p)// B3LYP/6-311 G(d,p)		
Structure	$\Delta E(CH_2NO^-)$	$\Delta E(NH_2O^-)$
C1	0.0	0.0
TS1	5.3	3.9
IN1	6.0	−3.4
TS2	6.2	−0.3
IN2	7.0	−3.3
TS3	10.6	−0.8
C2	4.9	−7.4

anion seems to be a better reactivating agent for both G- and V-series compounds, which are responsible for the inhibition of AChE. The higher free energy of activations obtained for the reactivation of the VX-serine adduct as compared to the sarin–serine adduct suggests that the reactivation of VX-inhibited AChE is more difficult than that of sarin-inhibited AChE [36]. These calculated results are in line with the toxic behavior of VX vs. that of sarin [41]. Results from our preliminary study of the mechanism of inhibition of a model of AChE by either VX or sarin (a similar model of AChE was used in both cases) also agreed with the experimental results reported in the literature [41]. The inhibition process with VX is preferred by a free energy of activation of 7.7 kcal mol^{-1} over the inhibition process with the sarin at the same level of theory. The higher toxicities of V-series compounds are also indicated by their nonvolatility. Our calculated results suggest that the inhibition and reactivation of VX may also be responsible for the higher toxicities of V-series compounds compared to G-series compounds.

Conclusions

In the present work, density functional calculations employing B3LYP/6-311 G(d,p) were performed for the reactivation of VX-inhibited AChE with either hydroxylamine anion or formoximate anion. The computed free energy of activation (5.2 kcal mol^{-1}) for the reactivation of VX-inhibited AChE is lower with the hydroxylamine anion than with the formoximate anion. The reactivation process with formaoximate anion follows a three-step mechanism, whereas it can be considered a two-step process with hydroxylamine anion. The activation barriers calculated in solvent using the polarizable continuum model (PCM) for the reactivation of the VX–AChE adduct with NH_2O^- were also found to be low. The calculated rate constants and free energies of activation for the reverse reactivation reactions of the VX–serine adduct with either formoximate or hydroxylamine anion suggest that hydroxylamine anion could be a very good antidote agent for the reactivation process. The reactivation of the VX–serine adduct with the formoximate anion or the hydroxylamine anion would be slower than the corresponding reactivation of the sarin–serine adduct. The inhibition process also indicates that VX is more toxic than the sarin compound [41].

Acknowledgments The authors thank the Department of Science and Technology (DST), New Delhi, India for financial support of this work. One of the authors, MASK, is thankful to the Council of Scientific and Industrial Research (CSIR), New Delhi, India for a fellowship (grant no. NWP-0010). We thank the reviewers for their comments and suggestions, which have helped us to improve the paper.

References

1. Yang YC, Baker JA, Ward JR (1992) Decontamination of chemical warfare agents. Chem Rev 92:1729–1743
2. Somani SM (1992) Chemical warfare agents. Academic, San Diego
3. Benschop HP, De Jong LPA (1988) Nerve agent stereoisomers: analysis, isolation and toxicology. Acc Chem Res 21:368–374
4. Kolb HC, Sharpless KB (2003) The growing impact of click chemistry on drug discovery. Drug Discov Today 8:1128–1137
5. Quinn DM (1987) Acetylcholinesterase: enzyme structure, reaction dynamics and virtual transition states. Chem Rev 87:955–979
6. Shafferman A, Kronman C, Flashner Y, Leitner M, Grosfeld H, Ordentlich A, Gozes Y, Cohen S, Ariel N, Barak D, Harel M, Silman I, Sussman JL, Velan B (1992) Mutagenesis of human acetylcholinesterase identification of residues involved in catalytic activity and in polypeptide folding. J Bio Chem 267:17640–17648
7. Wang J, Roszak S, Gu J, Leszczynski J (2005) Comprehensive global energy minimum modeling of the sarin–serine adduct. J Phys Chem B 109:1006–1014
8. Wang J, Gu J, Leszczynski J (2006) Phosphonylation mechanisms of sarin and acetylcholinesterase: a model DFT study. J Phys Chem B 110:7567–7573
9. Taylor P, Lappi S (1975) Interaction of fluorescence probes with acetylcholinesterase site and specificity of propidium binding. Biochemistry 14:1989–1997
10. Eddleston M, Szinicz L, Eyer P, Buckley N (2002) Oximes in acute organophosphorus pesticide poisoning: a systematic review of clinical trials. QJM-Ass Int J Med 95:275–283
11. Berends F, Posthumus CH, Sluys IVD, Deierkauf FA (1959) Biochim Biophys Acta 34:576–579
12. Wong L, Radić Z, Brüggemann RJM, Hosea N, Berman HA, Taylor P (2000) Mechanism of oxime reactivation of acetylcholinesterase analyzed by chirality, and mutagenesis. Biochemistry 39:5750–5757
13. Bermudez VM (2007) Computational study of the adsorption of trichlorophosphate, dimethyl methylphosphonate, and sarin on amorphous SiO_2. J Phys Chem C 111:9314–9323
14. Bandyopadhyay I, Kim MJ, Lee YS, Churchill DG (2006) Favorable pendant-amino metal chelation in VX nerve agent model systems. J Phys Chem A 110:3655–3661
15. Šečkutė J, Menke JL, Emnett RJ, Patterson EV, Cramer CJ (2005) Ab initio molecular orbital, and density functional studies on the solvolysis of sarin and O, S-dimethyl methylphosphonothiolate, a VX-like compound. J Org Chem 70:8649–8660
16. Zheng F, Zhan CG, Ornstein RL (2001) Theoretical studies of reaction pathways and energy barriers for alkaline hydrolysis of phosphotriesterase substrates paraoxon and related toxic phosphofluoridate nerve agents. J Chem Soc Perkin Trans 2:2355–2363
17. Patterson EV, Cramer CJ (1998) Molecular orbital calculations on the P–S bond cleavage step in the hydroperoxidolysis of nerve agent VX. J Phys Org Chem 11:232–240
18. Daniel KA, Kopff LA, Patterson EV (2008) Computational studies on the solvolysis of the chemical warfare agent VX. J Phys Org Chem 21:321–328
19. Menke JL, Patterson EV (2007) Quantum mechanical calculations on the reaction of ethoxide anion with *O*, *S*-dimethyl methylphosphonothiolate. J Mol Struct THEOCHEM 811:281–291
20. Khan MAS, Kesharwani MK, Bandyopadhyay T, Ganguly B (2009) Solvolysis of chemical warfare agent VX is more efficient with hydroxylamine anion: a computational study. J Mol Graph Model 28:177–182
21. Khan MAS, Kesharwani MK, Bandyopadhyay T, Ganguly B (2010) Remarkable effect of hydroxylamine anion towards the solvolysis of sarin: a DFT study. J Mol Struct THEOCHEM 944:132–136

22. Wang J, Gu J, Leszczynski J, Feliks M, Sokalski WA (2007) Oxime-induced reactivation of sarin-inhibited AChE: a theoretical mechanisms study. J Phys Chem B 111:2404–2408
23. Wang J, Gu J, Leszczynski JJ (2006) Theoretical modeling study for the phosphonylation mechanisms of the catalytic triad of acetylcholinesterase by sarin. Phys Chem B 110:7567–7573
24. Delfino RT, Figueroa-Villar JD (2009) Nucleophilic reactivation of sarin-inhibited acetylcholinesterase: a molecular modeling study. J Phys Chem B 113:8402–8411
25. Becke AD (1993) Density-functional thermo chemistry. III. The role of exact exchange. J Chem Phys 98:5648–5652
26. Lee C, Yang W, Parr RG (1988) Development of the Colle–Salvetti correlation-energy formula into a functional of the electron density. Phys Rev B 37:785–789
27. Beck JM, Hadad CM (2008) Hydrolysis of nerve agents by model nucleophiles: a computational study. Chem Biol Interact 175:200–203
28. Tomasi J, Persico M (1994) Molecular interactions in solution: an overview of methods based on continuous distributions of the solvent. Chem Rev 94:2027–2094
29. Cossi M, Barone V, Cammi R, Tomasi J (1996) Ab initio study of solvated molecules: a new implementation of the polarizable continuum model. Chem Phys Lett 255:327–335
30. Barone V, Cossi M, Tomasi J (1997) A new definition of cavities for the computation of solvation free energies by the polarizable continuum model. J Chem Phys 107:3210–3221
31. Barone V, Cossi M, Tomasi J (1998) Geometry optimization of molecular structures in solution by the polarizable continuum model. J Comput Chem 19:404–417
32. Cossi M, Barone V (1998) Analytical second derivatives of the free energy in solution by polarizable continuum models. J Chem Phys 109:6246–6254
33. González C, Schlegel HB (1990) Reaction path following in mass-weighted internal coordinates. J Phys Chem 94:5523–5527
34. González C, Schlegel HB (1991) Improved algorithms for reaction path following: higher–order implicit algorithms. J Chem Phys 95:5853–5860
35. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA Jr, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian 03, revision E 01. Gaussian Inc., Wallingford
36. Khan MAS, Lo R, Bandyopadhyay T, Ganguly B (2011) Probing the reactivation process of sarin-inhibited acetylcholinesterase with α-nucleophiles: hydroxylamine anion is predicted to be a better antidote with DFT calculations. J Mol Graph Model 29:1039–1046
37. Millard CB, Koellner G, Ordentlich A, Shafferman A, Silman I, Sussman JL (1999) Reaction products of acetylcholinesterase and VX reveal a mobile histidine in the catalytic triad. J Am Chem Soc 121:9883–9884
38. Kirby AJ, Tondo DW, Medeiros M, Souza BS, Priebe JP, Lima MF, Nome F (2009) Efficient intramolecular general-acid catalysis of the reactions of α-effect nucleophiles and ammonia oxide with a phosphate triester. J Am Chem Soc 131:2023–2028
39. Kirby AJ, Davies JE, Brandão TAS, da Silva PF, Rocha WR, Nome F (2006) Hydroxylamine as an oxygen nucleophile. Structure and reactivity of ammonia oxide. J Am Chem Soc 128:12374–123275
40. Kirby AJ, Manfredi AM, Souza BS, Medeiros M, Priebe JP, Brandão TAS, Nome F (2009) Reactions of alpha-nucleophiles with a model phosphate diester. ARKIVOC 3:28–38
41. Maxwell DM, Brecht KM, Koplovitz I, Sweeney RE (2006) Acetylcholinesterase inhibition: does it explain the toxicity of organophosphorus compounds? Arch Toxicol 80:756–760

J Mol Model (2012) 18:1809–1818
DOI 10.1007/s00894-011-1210-5

ORIGINAL PAPER

A comparative theoretical study of the catalytic activities of Au_2^- and $AuAg^-$ dimers for CO oxidation

Peng Liu · Ke Song · Dongju Zhang · Chengbu Liu

Received: 3 January 2011 / Accepted: 4 August 2011 / Published online: 18 August 2011

Abstract The detailed mechanisms of catalytic CO oxidation over Au_2^- and $AuAg^-$ dimers, which represent the simplest models for monometal Au and bimetallic Au-Ag nanoparticles, have been studied by performing density functional theory calculations. It is found that both Au_2^- and $AuAg^-$ dimers catalyze the reaction according to the similar mono-center Eley–Rideal mechanism. The catalytic reaction is of the multi-channel and multi-step characteristic, which can proceed along four possible pathways via two or three elementary steps. In $AuAg^-$, the Au site is more active than the Ag site, and the calculated energy barrier values for the rate-determining step of the Au-site catalytic reaction are remarkably smaller than those for both the Ag-site catalytic reaction and the Au_2^- catalytic reaction. The better catalytic activity of bimetallic $AuAg^-$ dimer is attributed to the synergistic effect between Au and Ag atom. The present results provide valuable information for understanding the higher catalytic activity of Au-Ag nanoparticles and nanoalloys for low-temperature CO oxidation than either pure metallic catalyst.

Keywords $AuAg^-$ · CO Oxidation · Au_2^- · DFT

Introduction

The supported [1–4] and unsupported [5–7] gold nanoparticles have attracted worldwide attention due to their unusual catalytic activity for oxidation reactions in contrast to bulk gold since the pioneering findings of Haruta et al. concerning low-temperature CO oxidation [8–10]. Researchers have shown the catalytic activity of the gold nanoparticles is sensitive to their size and shape [11–15], the nature of the support [16–18], and the preparation methods [19, 20]. During the past decades, extensive research interest has been paid to improve the catalytic activity of gold nanoparticles by tuning the particle morphology, modifying the substrate, and controlling the pretreatment conditions.

Recently, an alternative strategy to enhance the reactivity of gold nanoparticles, alloying gold nanoparticles with a second metal to form a bimetallic gold catalyst ("nanoalloy"), has attracted significant attention [21–23]. It has been shown that some Au-based binary-alloy catalysts, such as Au-Ag [24–27], Au-Pt [28, 29], Au-Pd [30], Au-Cu [31, 32], and Au-Sr [23], possess better activity and stability than the corresponding monometallic catalysts for many important processes, including CO oxidation [24–27], direct synthesis of H_2O_2 from H_2 and O_2 [30], and CH_3OH oxidation [29]. Among these Au-based bimetallic catalysts, Au-Ag nanoalloy is of particular interest because Au and Ag atoms are in intimate proximity to each other (Ag-Ag and Au-Au bond lengths in bulk materials are 2.889 and 2.884 Å, respectively) and thus easy form nanoalloy. So far, experimental studies [24–27] have established that the synergistic effect between Au and Ag leads to the higher activity of Au-Ag nanoalloy for low-temperature CO oxidation than either pure metallic catalyst. In particular, Wang et al. [27] found that the alloy nanoparticles with an Au/Ag ratio close to 1:1 have the highest activity. On the other hand, the theoretical studies on the geometrical and electronic structures [33–35] of Au-Ag binary clusters and on their reactivity toward CO and O_2 [36–38] have also received considerable interest in recent years. And relevant results have provided useful

First two authors contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1210-5) contains supplementary material, which is available to authorized users.

P. Liu · K. Song · D. Zhang (✉) · C. Liu
Key Lab of Colloid and Interface Chemistry,
Ministry of Education, Institute of Theoretical Chemistry,
Shandong University,
Jinan, 250100, People's Republic of China
e-mail: zhangdj@sdu.edu.cn

information in elucidating the synergistic effect of catalytic activity. However, the catalytic mechanism details of Au-Ag nanoalloy is still not well understood, and our knowledge about the origin of the exceptionally high activity of Au-Ag alloy nanoparticles is still far from complete. In recent years, density functional theory (DFT) has become a valuable tool for studying the properties of molecules and materials [39, 40] and for identifying reaction mechanisms [41, 42]. In this work, we present a comparative theoretical study of the catalytic activity of Au_2^- and $AuAg^-$ dimers for CO oxidation, from which we expect to provide understanding to some extent about the synergistic catalytic effect of Au-Ag nanoalloys. It is known that Au is the most electronegative metal and the polarization effect are usually observed in Au and its alloy nanoparticles supported on substrates, leading to the negatively charged nanoparticles. Thus the anionic clusters seem to be appropriate for mimicking the catalytic reactivities of Au and Au-Ag nanoparticles. In particular, previous studies [5, 43] show that Au_2^- is the smallest unit that can catalyze CO oxidation. So in the present study, we chose Au_2^- and $AuAg^-$ dimers as representative models of nanoparticles.

Computational details

The catalytic cycles studied in this work are summarized in Eq. 1. For $AuAg^-$ dimer, the reaction branches into the Au-site catalytic series and Ag-site catalytic series. So we actually need to perform theoretical calculations for three catalytic cycle systems.

$$2CO + O_2 \xrightarrow{AuX^-(X = Au, Ag)} CO_2 \qquad (1)$$

Calculations were carried out in the framework of density functional theory (DFT) by use of the hybrid B3LYP [44, 45] functional as implemented in the Gaussian 03 program package [46]. We chose Los Alamos LANL2DZ [47, 48] effective core pseudopotentials (ECPs) and valence double-ζ basis sets for gold and silver atoms, as well as 6-311+G(d) basis sets for carbon and oxygen atoms. The structures of the reactants, products, intermediates, and transition states were fully optimized without any symmetry constraints. Frequency calculations were carried out for each optimized structure at the same level to identify the natures of all the stationary points (minima or first-order saddle points) and to calculate the zero-point vibrational energies (ZPEs). Intrinsic reaction coordinate (IRC) [49] calculations were conducted in both directions (forward and reverse) from the transition states to the corresponding local minima to identify the minimum-energy paths. Stability tests of wave functions [50, 51] for all identified stationary points have been carried out to ensure that the lowest energy solutions in the SCF procedures are found. Charge delocalization has been carried out using natural bonding orbital (NBO) analysis. All calculations were carried out by resolving unrestricted Kohn-Sham equations.

Results and discussion

Previous investigations [37, 38, 52] showed that the B3LYP/LANL2DZ combination is sufficiently accurate for describing noble-metal systems. To further clarify the reliability of our calculations, we here provide benchmark calculations of the geometries, dissociation energies, and vertical ionization potentials (vertical electron detachment energies for anionic systems) for Au_2^-, $AuAg^-$, CO, O_2 and CO_2. As shown in Table 1, all calculated data are in fairly good agreement with the corresponding experimental results [53–56]. In addition, we have also calculated the adsorption energy of O_2 on Au_2^-, and it is found that the theoretical result, 0.97 eV, is in reasonable agreement with the corresponding experimental value (1.01±0.14 eV) [57]. Therefore, we believe that the level of theory selected in this work can describe the present systems with acceptable accuracy and precision.

Complexes of Au_2^- and $AuAg^-$ with O_2 and CO

The formation of complexes of Au_2^- and $AuAg^-$ dimers with O_2 and CO molecules is expected to be the initial step of CO oxidation. Our calculations considered various possible geometries where the O_2 or CO molecule approaches the dimers with different orientations, and different possible spin combinations between the dimers and O_2 molecule. The optimized most stable complexes are shown in Fig. 1, where the values in parenthesis denote natural charges of atoms calculated by performing NBO analysis. The ground states of all these complexes are found to

Table 1 Calculated and experimental bond lengths (d), dissociation energies (D), and vertical electron detachment energy (vDE) (I) or vertical ionization potentials (vIP) (II) for Au_2^- and $AuAg^-$ dimers and CO and O_2 molecules

	d (Å)	D (kcal mol^{-1})	vIP or vDE (kcal mol^{-1})
Au-Au$^-$	2.737 (2.58)^a	39.85 (44.27)^a	(I) 46.84 (46.35)^a
Au-Ag$^-$	2.772	26.99 (24.90)^b	(I) 33.01 (32.98)^b
C-O	1.128 (1.13)^c	249.04 (259.20)^c	(II) 327.40 (323.07)^d
O-O	1.206 (1.21)^c	117.30 (120.60)^c	(II) 292.34 (278.10)^d
O-C-O	1.157 (1.16)^c	376.25 (389.02)^c	(II) 318.47 (317.54)^d

The experimental data are shown in parentheses. [a] Ref. 50, [b] Ref. 51, [c] Ref. 52, [d] Ref. 53

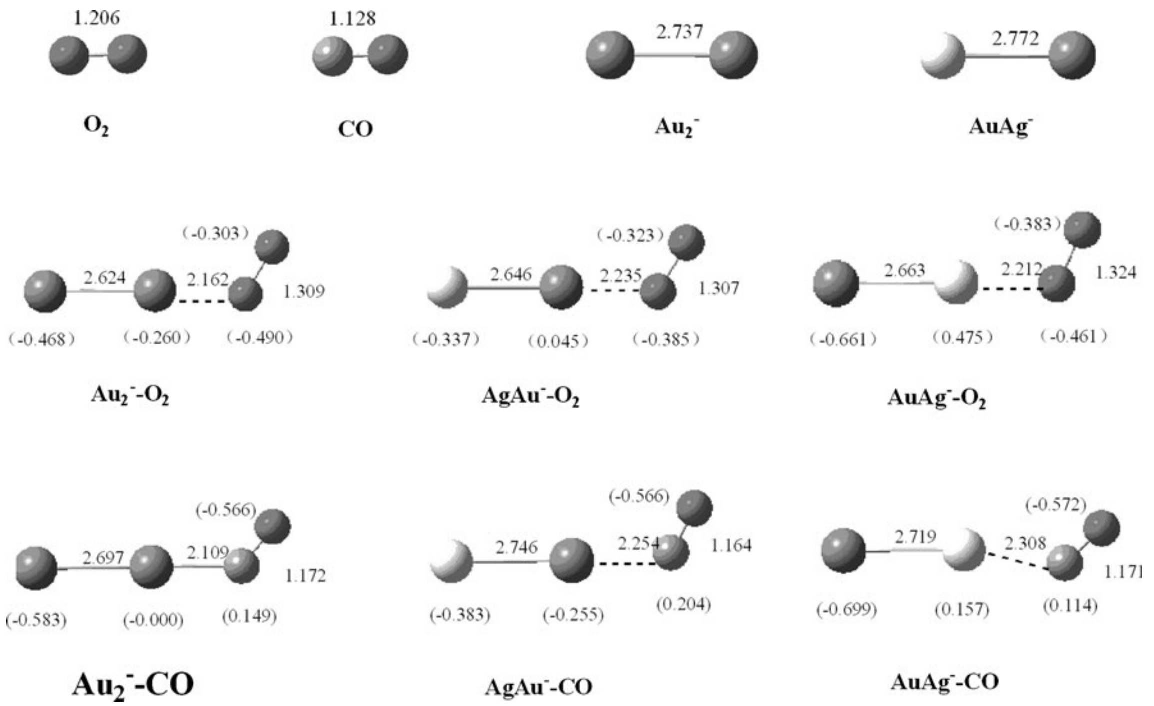
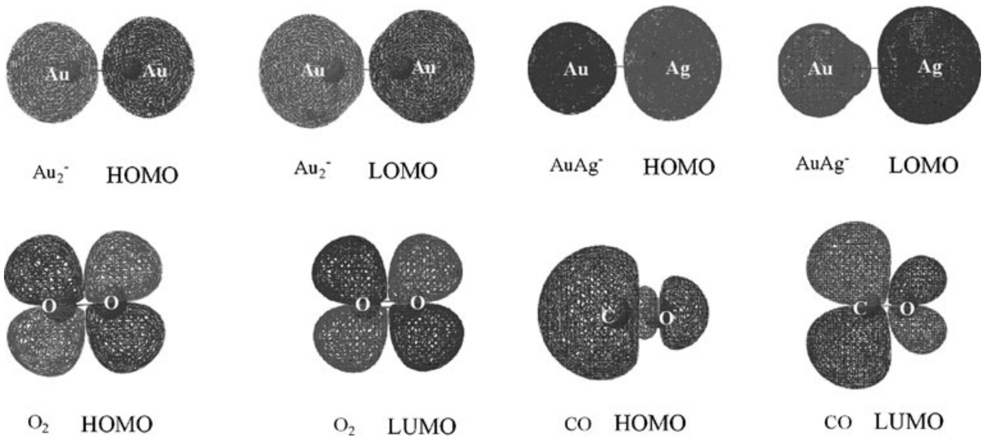


Fig. 1 Optimized geometries for isolated O_2 and CO molecules, and Au_2^- and $AuAg^-$ dimers, as well as the complexes between the dimers and molecules. The distances are in Å. The numbers in parentheses are calculated natural charges of atoms (in e)

be in their doublets. We see that in all complexes, the dimers interact with molecules in single-site binding mode (i.e., one atom of the molecules coordinates with one atom of the dimers) in a tilted manner, and net electron transfer occurs from the dimers to molecules. Such bonding mode can achieve the maximum orbital overlap between the HOMO of Au_2^- ($AuAg^-$) and the LUMO of O_2 (CO) to form the stable complexes. In Fig. 2, we show the HOMO and LUMO isosurfaces of Au_2^- and $AuAg^-$ dimmers as well as O_2 and CO molecules. Clearly, the shapes and symmetries of HOMOs (σ^* orbitals) of Au_2^- and $AuAg^-$ dimmers match the LUMOs (π^* orbitals) of O_2 and CO, which can successfully explain the binding orientations of O_2 and CO on Au_2^- and $AuAg^-$, i.e., the molecules must bind to the dimers in a tilted manner to form the maximum favorable overlay between their LUMOs and the HOMOs of the dimers.

For $AuAg^-$ dimer, O_2 can bind to Ag or Au site, forming complex $AuAg^-$-O_2 or $AgAu^-$-O_2. The calculated binding energies are 30.24 and 17.15 kcal mol^{-1} for these two complexes, respectively, indicating that O_2 prefers to bind to Ag-site of $AuAg^-$, where O_2 is activated in a larger extent, as confirmed by calculated longer O-O distance (1.324 Å) in $AuAg^-$-O_2 than that in $AgAu^-$-O_2 (1.307 Å). Compared to the binding energy of O_2 over Au_2^- (22.43 kcal mol^{-1}) and the O-O distance in Au_2^--O_2 (1.309Å), it is clear that $AuAg^-$ is more efficient for activating the

Fig. 2 The HOMO and LUMO isosurfaces for Au_2^- and $AuAg^-$ dimers and O_2 and CO molecules (isosurface value=0.02)



dissociation of O_2 molecule than Au_2^-. The amount of electron transfer from the dimer to O_2 in these three complexes is 0.844 *e* in $AuAg^-$-O_2, 0.793 *e* in Au_2^--O_2, and 0.708 e in $AgAu^-$-O_2. The transferred electrons enter the π^* orbital of O_2 to reduce O-O bond strength. While more electrons are transferred, the larger the O-O distance.

Similarly, for the complexes of $AuAg^-$ with CO, our calculations show that $AuAg^-$-CO is more stable than $AgAu^-$-CO (the binding energies of CO over the Ag and Au sites are 6.01 and 4.47 kcal mol^{-1}, respectively), i.e., the Ag-site of $AuAg^-$ is also more active toward CO binding than the Au-site. Compared to Au_2^-, however, $AuAg^-$ seems to be slightly more inert for CO binding, as indicated by the calculated larger binding energy (9.70 kcal mol^{-1}) of CO over Au_2^-.

When we compare the relative stability of the complexes of $AuAg^-$ with O_2 to those with CO, we see that the interaction of $AuAg^-$ with O_2 is stronger than that with CO. This is also confirmed by the geometrical parameters shown in Fig. 1. For example, the O-O distance in $AuAg^-$-O_2 (1.324 Å) is increased by 9.8 % compared to that in isolated O_2, while the C-O distance in $AuAg^-$-CO (1.171 Å) is only increased by 3.8 % compared to that in isolated CO, indicating the interaction of $AuAg^-$ with O_2 is stronger than that with CO. Thus we conjecture that the CO oxidation over $AuAg^-$ is initiated by activating O_2 rather than CO.

In the following sections, we discuss the detailed reaction mechanism of the CO oxidation over Au_2^- and $AuAg^-$, from which we expect to provide aid to some extent for understanding the superior property of binary Au-Ag alloy catalysts for CO oxidation compared to the corresponding pure metal catalysts.

The CO oxidation over Au_2^-

Early theoretical and experimental studies for the CO oxidation promoted by Au_2^- by Socaciu [5] and Hakkinen [43] have provided useful information in elucidating microscopic aspects of the CO oxidation mechanism. Our reexamination for this model system shows comprehensive potential energy surface details and some new results about the elementary mechanism. Figures 3, 4, 5 show the complete mechanism for CO oxidation over Au_2^- along four distinct reaction pathways, I-IV, where the crucial intermediates and transition states with selected geometrical parameters are given to easily see the structural transition process.

Pathways I and II start from the adsorption of O_2 on Au_2^- to form the common superoxo-like complex Au_2^--O_2. As shown in Fig. 3, pathway I involves the direct oxygen abstract by the CO molecule from the pre-adsorbed O_2. After the adsorption of O_2 on Au_2^-, the coming CO attacks the adsorbed O_2 to give a three-species metastable complex, $IM1_{Au_2^-}$, which involves the weak binding of CO to the adsorbed O_2 and lies below the reactants by 24.63 kcal mol^{-1}. Then the CO abstracts an O atom to produce a CO_2 via transition state $TS1_{Au_2^-}$. This process is exothermic by 51.15 kcal mol^{-1} with a barrier of 17.46 kcal mol^{-1}. Once the resultant CO_2 is released, the newly formed Au_2O^- can immediately trap a second CO molecule to lead to a very stable complex $IM2_{Au_2^-}$ with an energy release of 44.60 kcal mol^{-1}. In $IM2_{Au_2^-}$, a second CO_2 has almost formed, as shown by the geometry in Fig. 3. The subsequent process is to release the second CO_2 molecule via $TS2_{Au_2^-}$ with a small barrier (1.71 kcal mol^{-1}) and hence complete the catalytic cycle. The overall reaction is calculated to be exothermic by 137.11 kcal mol^{-1}.

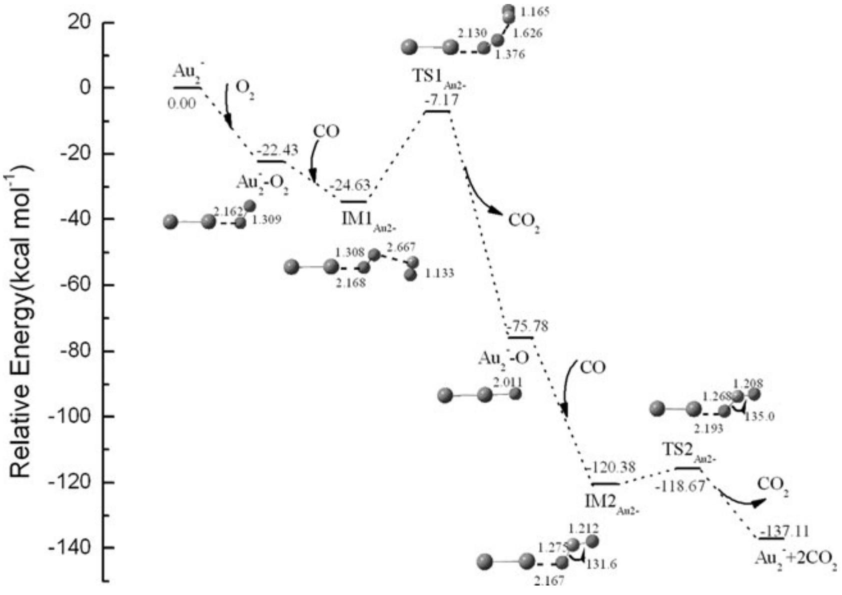
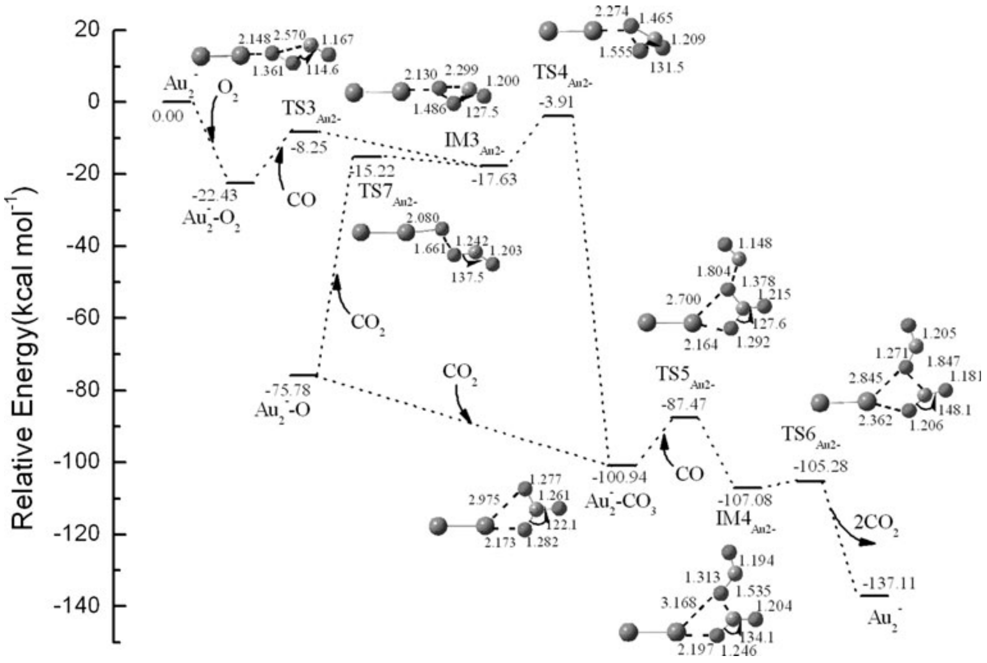


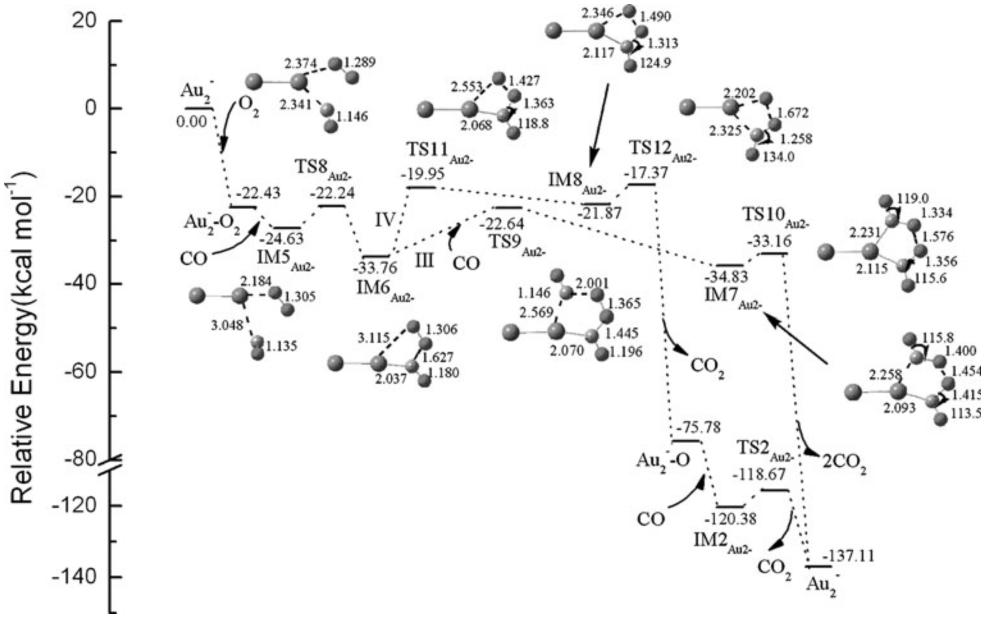
Fig. 3 Calculated energy profile for the CO oxidation over Au_2^- along the direct oxygen abstract pathway (pathway I). The total energy of reactants (Au_2^-+O_2+ 2CO) is taken as zero

Fig. 4 Calculated energy profile for the CO oxidation over Au_2^- along the carbonate intermediate path (path II). The total energy of reactants (Au_2^-+O_2+ 2CO) is taken as zero



For low-temperature CO oxidation, it is generally agreed that carbonate species are an important intermediate [58]. Several investigations have observed its formation on supported and unsupported gold clusters [57–61]. However, the relevant mechanism is still not fully understood. Here we studied the CO oxidization pathway involving carbonate intermediate in detail, as shown in Fig. 4 (pathway II). This path also starts from the superoxo-like complex Au_2^--O_2. By attaching the CO molecule to Au_2^--O_2, we locate another metastable complex $IM3_{Au_2^-}$, which lies below the reactants by 17.63 kcal mol^{-1}. The transition state structure involved is$TS3_{Au_2^-}$, and the barrier for this process is 14.18 kcal mol^{-1}. In $IM3_{Au_2^-}$, the C atom of CO coordinates to two O atoms, and the O-O distance in O_2 subunit has been elongated to 1.486 Å. Once this complex is formed, it can further evolve into a highly stable carbonate species, denoted as Au_2^--CO_3, via transition state $TS4_{Au_2^-}$. This process involves the insertion of CO molecule into the O-O bond in Au_2^--O_2, resulting in the breaking of the O-O bond and the simultaneous formation of two C-O bonds. The calculated barrier for the formation of the carbonate species is 13.72 kcal mol^{-1}. This process is similar to the early report by Liana et al. [5], where the relevant geometries for the intermediate and transition state were not presented.

Fig. 5 Calculated energy profile for the CO oxidation over Au_2^- along the energetically most favorable O-C-O-O-C-O group-involved pathway (pathway III) and the O-O-C-O group mediated oxygen abstract pathway (paths IV). The total energy of reactants (Au_2^-+O_2+ 2CO) is taken as zero



Due to the exoergic nature of the carbonate formation process, subsequent reaction barriers can be easily overcome. From Fig. 4, we see that the attack of a second CO molecule on the carbonate intermediate can result in the formation of two CO_2 molecules and hence completes the catalytic cycle. This involves two elementary steps, as indicated by $TS5_{Au_2^-}$ and $TS6_{Au_2^-}$. In $TS5_{Au_2^-}$, the second CO is abstracting an O atom in the carbonate to form $IM4_{Au_2^-}$, where two CO_2 subunits have emerged, while in $TS6_{Au_2^-}$, the two forming CO_2 subunits are being unbound from the catalyst. From calculated relative energies, it seems that pathway II is energetically comparable to pathway I.

Note that in a recent study [62] of ours, Au-carbonate species was found to originate from the effective collision between Au-oxides and newly formed CO_2, where the carbon atom of CO_2 directly attacks the O atom in the oxides of Au. In this sense, the AuO^- fragment formed in path I could also contribute to the formation of Au_2^--CO_3. Accordingly, in Fig. 4, we show this possible way, which is confirmed to be a barrierless process. In contrast, another way to from Au_2^--CO_3 from AuO^- and CO_2, where the O atom of CO_2 approaching AuO^- (see $TS7_{Au_2^-}$ in Fig. 4), is found to be energetically very demanding with a barrier of 60.56 kcal mol^{-1} and thus is not capable of competing with the barrierless process with the C atom of CO_2 approaching AuO^-.

Figure 5 shows another two branches of CO oxidation, pathways III and IV. The attack of a coming CO molecule on Au_2^--O_2 with the C atom approaching the end O in Au_2^--O_2 leads to intermediate $IM5_{Au_2^-}$. Such an intermediate can be converted into intermediate $IM6_{Au_2^-}$ via $TS8_{Au_2^-}$ with only a barrier of 2.39 kcal mol^{-1}. $IM6_{Au_2^-}$, where the C atom binds to Au_2^- and associates one O of O_2 to form a O-O-C-O group, lies below the reactants by 33.76 kcal mol^{-1}, from which we located two reaction branches to form CO_2, denoted as paths III and IV in Fig. 5. Path III involves the simultaneous formation of two CO_2 molecules, where a second CO is first attached to $IM6_{Au_2^-}$ to get $IM7_{Au_2^-}$ via $TS9_{Au_2^-}$ with a barrier of 11.12 kcal mol^{-1}, this step is the rate-determining step of pathway III. And then from $IM7_{Au_2^-}$ two CO_2 molecules are removed via $TS10_{Au_2^-}$ with an energy demand of only 1.67 kcal mol^{-1}. In contrast, along path IV, immediate $IM6_{Au_2^-}$ initially evolves into $IM8_{Au_2^-}$ via $TS11_{Au_2^-}$ with a barrier of 13.81 kcal mol^{-1}, and then $IM8_{Au_2^-}$ releases, via $TS12_{Au_2^-}$, the first CO_2 molecule and forms AuO^- fragment, which further reacts with the second CO molecule to release the second CO_2 molecule. So path IV follows the sequential formation of two CO_2 molecules and actually crosses into path I after forming AuO^- fragment. From the calculated energy profiles (Fig. 5), these two paths are expected to be slightly favorable in energywith pathway III . According to geometrical characteristics of intermediates and transition states, pathways III and IV can be described as the O-C-O-O-C-O group involved pathway and the O-O-C-O group mediated oxygen abstract pathway, respectively.

The results above show that the CO oxidation over Au_2^- follows mono-center Eley–Rideal mechanism and that the overall reaction is a highly exoergic process. The energy released from a complete catalytic cycle (137.11 kcal mol^{-1}) is much more than that required during the reaction (the calculated largest barrier along all four paths is smaller than 20 kcal mol^{-1}). Furthermore, it is found that the potential energy surface profile along every path lies below the reactants throughout the whole reaction. Therefore, from an energy point of view, all four paths located in the present work are feasible for the CO oxidation. In other words, the CO oxidation is characteristic of multi-channel and multi-step. In Table 2, we list calculated energy barrier values for each rate-determining step along each pathway. Clearly, pathway III is the most favorable, and thus is proposed to be the dominant pathway for the CO oxidation.

The CO oxidation over $AuAg^-$

Similarly, the CO oxidation over $AuAg^-$ is also studied along the four pathways discussed above: the direct oxygen pathway (pathway I), the carbonate intermediate pathway (pathway II), the energetically most favorable O-C-O-O-C-O group-involved pathway (pathway III), and the O-O-C-O group mediated oxygen abstract pathway (pathway IV). We have considered three possible situations by assuming that the catalytically active center is on (i) the Ag site, (ii) the Au site, and (iii) both the Ag and Au sites, where O_2 and CO molecules approach to Ag and Au sites, respectively. Our calculations show that the intermediates and transition states designed initially according to the last assumption always converge to those in either the first or second situation. In other words, the catalytic reaction is always found to proceed via a mono-atomic rather than double-atomic active center mechanism.

Figures 6, 7, 8 show the calculated Ag-site catalytic potential energy surfaces with geometries of intermediates and transition states, and Figs. S1-S3 in the supplementary

Table 2 Energy barrier values for each rate-determining step of each pathway studied. All values are given in kcal mol^{-1}

	Au_2^-	$AuAg^-$	
		Ag site	Au site
Pathway I	17.46	15.32	17.46
Pathway II	14.18	19.52	18.79
Pathway III	11.12	11.35	8.90
Pathway IV	13.81	13.94	10.32

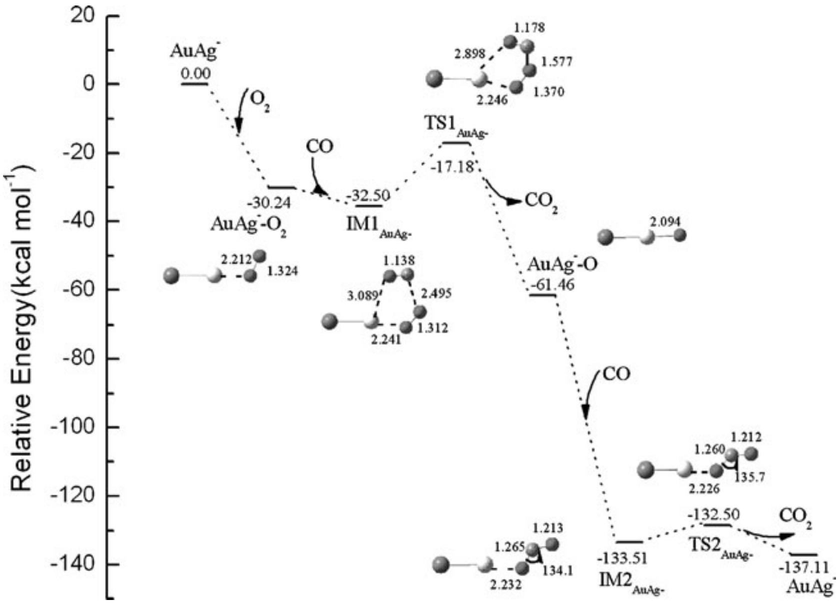


Fig. 6 Calculated energy profile for the CO oxidation over AuAg$^-$ along the direct oxygen abstract pathway (pathway I). The total energy of reactants (AuAg$^-$+O$_2$+ 2CO) is taken as zero

information are those for the Au-site catalytic reactions. It should be noted that in the two situations the geometrical characteristics for most species along pathway I-IV are generally similar to those for the Au$_2^-$ catalytic reaction, indicating the mechanistic details discussed above can also be applied to the AuAg$^-$ catalytic reaction. Thus, we conjecture that monometallic Au and bimetallic Au-Ag nanoparticles could work via the similar mono-center Eley–Rideal mechanism.

In the following sections we only summarize the main conclusions obtained from the present calculations, and the multi-channel and multi-step details of the reaction are not discussed again for brevity. From the potential energy surface profiles shown in Figs. 6–8 and Figs. S1-S3, it is noted that for both the Ag- and Au-site catalytic reactions, pathways III and IV are more favorable than pathways I and II, which is similar to the Au$_2^-$ catalytic reaction. Furthermore, we find that along pathways III and IV, the barriers of the rate-determining steps for the Au-site catalytic reactions are smaller than those for the Ag-site catalytic reactions. This fact indicates the reaction prefers to proceed on Au site to Ag site. Calculated energy barrier values of rate-determining steps along four pathways are also given in Table 2. From these data, it seems to be reasonable to draw out the following conclusions: (i) Bimetallic AuAg$^-$ can catalyze CO

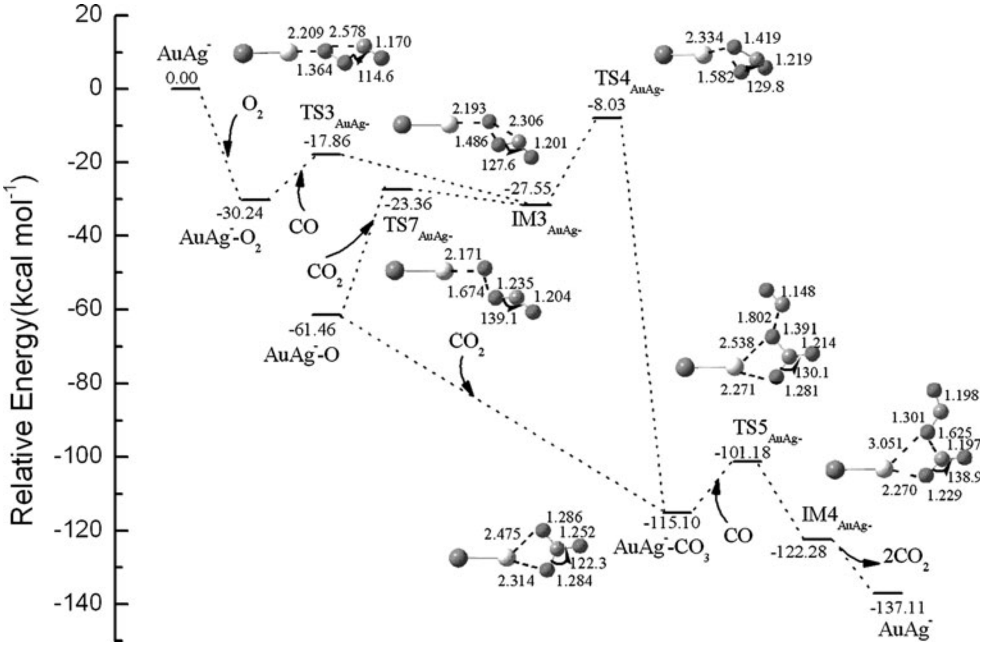


Fig. 7 Calculated energy profile for the CO oxidation over AuAg$^-$ along the carbonate intermediate path (path II). The total energy of reactants (AuAg$^-$+O$_2$+ 2CO) is taken as zero

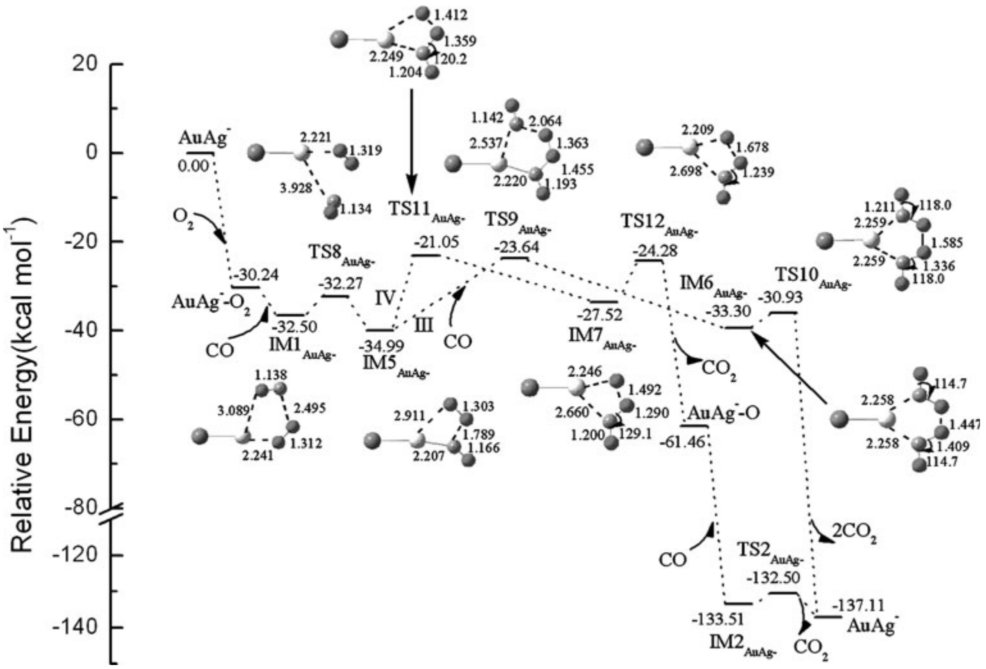


Fig. 8 Calculated energy profile for the CO oxidation over AuAg$^-$ along the energetically most favorable O-C-O-O-C-O group-involved pathway (pathway IIII) and the O-O-C-O group mediated oxygen abstract pathway (paths IV). The total energy of reactants (AuAg$^-$+ O$_2$+ 2CO) is taken as zero

oxidation than mono-metallic Au$_2^-$ more effectively, which explains the exceptionally high activity of Au-Ag bimetallic nanoparticles for low-temperature CO oxidation; and (ii) For dimer AuAg$^-$, the Au site is a catalytic active center, which is not consistent with the generally supposed mechanism picture of Au-Ag bimetallic catalysts, where both Ag and Au sites are proposed to participate in the reaction with Ag sites adsorbing and activating O$_2$ molecules and Au sites interacting with the coming CO molecules.

To clearly show the catalytic mechanism of dimer AuAg$^-$, in Fig. 9 we schematically depict the catalytic cycle along the energetically most favorable channel, pathway III. This cycle contains three elementary steps: (i) The formation of the crucial intermediate IM5$_{AgAu^-}$, where the O$_2$ molecule initially bound to the Au site is pushed away by the coming CO molecule, which binds now to the Au site and simultaneously interacts with the O$_2$ molecule to form the O-O-C-O group with a barrier of 1.33 kcal mol^{-1}. (ii) The second CO molecules approaches IM5$_{AgAu^-}$ with its C atom approaching the Au and the end O of the O-O-C-O group to form IM6$_{AgAu^-}$, where each CO molecule is ready to seize one atom of O$_2$ to form CO$_2$ product. This process requires surmounting the barrier of 8.90 kcal mol^{-1}. (iii) Two CO$_2$ molecules dissociate simultaneously from IM6$_{AgAu^-}$ with the barrier of only 1.76 kcal mol^{-1} to complete the catalytic cycle.

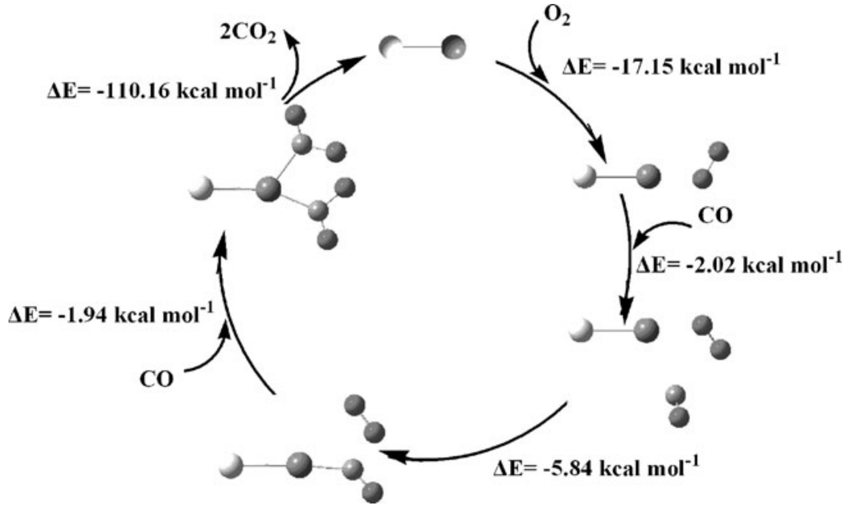

2CO$_2$
ΔE= -110.16 kcal mol^{-1}
O$_2$
ΔE= -17.15 kcal mol^{-1}
CO
ΔE= -2.02 kcal mol^{-1}
ΔE= -1.94 kcal mol^{-1}
CO
ΔE= -5.84 kcal mol^{-1}

Fig. 9 Schematic representation of the catalytic cycle over AgAu$^-$ along the energetically most favorable channel

Conclusions

To better understand the higher activity of Au-Ag bimetallic catalysts than pure metallic Au catalyst for low-temperature CO oxidation, we here presented a comparative theoretical study of the catalytic activity of Au_2^- and $AuAg^-$ dimers, which represents the simplest models for monometal Au and bimetallic Au-Ag nanoparticles. By performing DFT calculations, we have shown the mechanism details of CO oxidation over Au_2^- and $Au\text{-}Ag^-$ dimers, for which both the mono- and double-center catalytic mechanisms have been taken into account. It is found that Au_2^- and $AuAg^-$ catalyze CO oxidation according to the similar mono-center Eley–Rideal mechanism, which is different from the generally supposed mechanism picture of Au-Ag bimetallic catalysts, where both Ag and Au sites are proposed to participate in the reaction with Ag sites adsorbing and activating O_2 molecules and Au sites interacting with the coming CO molecules. The catalytic reaction is shown to be of the multi-channel and multi-step characteristic, which can proceed via two or three elementary steps along four possible pathways, including the direct oxygen pathway, the carbonate intermediate pathway, the energetically most favorable O-C-O-O-C-O group involved pathway, and the O-O-C-O group mediated oxygen abstract pathway. For $AuAg^-$ dimer, the Au site is more active than the Ag site, and the calculated energy barrier values for rate-determining step for the Au-site catalytic reaction are remarkably smaller than those for both the Ag-site catalytic reaction and the Au_2^- catalytic reaction. The present results provide assistance to some extent for understanding the experimentally observed exceptionally high catalytic activity of Au-Ag nanoparticles and nanoalloys for low-temperature CO oxidation.

Acknowledgments We are grateful for the financial support of the National Natural Science Foundations of China (No. 20873076), the Natural Science Foundation of Shandong Province (No. Z2008B02), and the Specialized Research Fund for the Doctoral Program of Higher Education (No. 200804220009).

References

1. Bond GC, Thompson DT (1999) Catal Rev Sci Eng 41:319–388
2. Comotti M, Li WC, Spliethoff B, Schuth F (2006) J Am Chem Soc 128:917–924
3. Hickey N, Larochette PA, Gentilini C, Sordelli L, Olivi L, Polizzi S, Montini T, Fornasiero P, Pasquato L, Graziani M (2007) Chem Mater 19:650–651
4. Zhang C, Yoon B, Landman U (2007) J Am Chem Soc 129:2228–2229
5. Socaciu LD, Hagen J, Bernhardt TM, Woste L, Heiz U, Hakkinen H, Landman U (2003) J Am Chem Soc 125:10437–10445
6. Xu CX, Su JX, Xu XH, Liu PP, Zhao HJ, Tian F, Ding Y (2007) J Am Chem Soc 129:42–43
7. Xu CX, Xu XH, Su JX, Ding Y (2007) J Catal 252:243–248
8. Haruta M, Kobayashi T et al. (1987) Chem Lett 405–408
9. Haruta M (1997) Catal Today 36:153–166
10. Haruta M (2003) Chem Record 3:75–87
11. Valden M, Lai X, Goodman DW (1998) Science 281:1647–1650
12. Wallace WT, Whetten RL (2000) J Phys Chem B 104:10964–10968
13. Meier DC, Goodman DW (2004) J Am Chem Soc 126:1892–1899
14. Lopez N, Janssens TVW, Clausen BS, Xu Y, Mavrikakis M, Bligaard T, Norskov JK (2004) J Catal 223:232–235
15. Chen M, Cai Y, Yan Z, Goodman DW (2006) J Am Chem Soc 128:6341–6346
16. Kozlov AI, Kozlova AP, Asakura K, Matsui Y, Kogure T, Shido T, Iwasawa Y (2000) J Catal 196:56–65
17. Chen YJ, Yeh CT (2001) J Catal 200:59–68
18. Moreau F, Bond GC et al. (2004) Chem Commun 1642–1643
19. Schubert MM, Hackenberg S, van Veen AC, Muhler M, Plzak V, Behm RJ (2001) J Catal 197:113–122
20. Arrii S, Morfin F, Renouprez AJ, Rousset JL (2004) J Am Chem Soc 126:1199–1205
21. Guczi L, Beck A, Horvath A, Koppany ZS, Stefler G, Frey K, Sajo I, Geszti O, Bazin D, Lynch J (2003) J Mol Catal A 204:545–552
22. Venezia AM, Liotta LF, Pantaleo G, La Parola V, Deganello G, Beck A, Koppany ZS, Frey K, Horvath D, Guczi L (2003) Appl Catal A 251:359–368
23. Hakkinen H, Abbet S, Sanchez A, Heiz U, Landman U (2003) Angew Chem Int Ed 42:1297–1300
24. Liu JH, Wang AQ, Lin HP, Mou CY (2005) J Phys Chem B 109:40–43
25. Wang AQ, Liu JH, Lin SD, Lin TS, Mou CY (2005) J Catal 233:186–197
26. Wang AQ, Chang CM, Mou CY (2005) J Phys Chem B 109:18860–18867
27. Wang C, Yin HF, Chan R, Peng S, Dai S, Sun SH (2009) Chem Mater 21:433–435
28. Chilukuri S, Joseph T, Malwadkar S, Damle C, Halligudi SB, Rao BS, Sastry M, Ratnasamy P (2003) Stud Surf Sci Catal 146:573–576
29. Lou YB, Maye MM et al. (2001) Chem Commun 473–474
30. Landon P, Collier PJ, Carley AF, Chadwick D, Papworth AJ, Burrows A, Kielyd CJ, Hutchings GJ (2003) Phys Chem Chem Phys 5:1917–1923
31. Rossi G, Rapallo A, Mottet C, Fortunelli A, Baletto F, Ferrando R (2004) Phys Rev Lett 93:105503
32. Barcaro G, Fortunelli A, Rossi G, Nita F, Ferrando R (2006) J Phys Chem B 110:23197–23203
33. Mitric R, Burgel C, Burda J, Bonacic-Koutecky V, Fantucci P (2003) Eur Phys J D 24:41–44
34. Bonacic-Koutecky V, Burda J, Mitric R, Ge M (2002) J Chem Phys 117:3120–3131
35. Lee HM, Ge M, Sahu BR, Tarakeshwar P, Kim KS (2003) J Phys Chem B 107:9994–10005
36. Lim DC, Lopez-Salido I, Dietsche R, Kim YD (2007) Surf Sci 601:5635–5642
37. Joshi AM, Tucker MH, Delgass WN, Thomson KT (2006) J Chem Phys 125:194707
38. Joshi AM, Delgass WN, Thomson KT (2006) J Phys Chem B 110:23373–23387
39. Stamenkovic V, Mun BS, Mayrhofer KJJ, Ross PN, Markovic NM, Rossmeisl J, Greeley J, Norskov JK (2006) Angew Chem Int Ed 45:2897–2901
40. Kalita B, Deka RC (2009) J Am Chem Soc 131:13252–13254
41. Qu XH, Zhang QZ, Shi XY, Xu F, Wang WX (2009) Environ Sci Technol 43:4068–4075

42. Zhang QZ, Qu XH, Xu F, Shi XY, Wang WX (2009) Environ Sci Technol 43:4105–4112
43. Hakkinen H, Landman U (2001) J Am Chem Soc 123:9704–9705
44. Becke AD (1992) J Chem Phys 96:2155–2160
45. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785–789
46. Frisch MJ, Trucks GW, Schlegel HB et al. (2004) Gaussian 03, Revision D. 01. Gaussian, Pittsburgh, PA
47. Hay PJ, Wadt WR (1985) J Chem Phys 82:270–283
48. Hay PJ, Wadt WR (1985) J Chem Phys 82:299–310
49. Fukui K (1970) J Phys Chem 74:4161–4163
50. Seeger R, Pople JA (1977) J Chem Phys 66:3045–3050
51. Bauernschmitt R, Ahlrichs R (1996) J Chem Phys 104:9047–9052
52. Wu DY, Ren B, Jiang YX, Xu X, Tian ZQ (2002) J Phys Chem A 106:9042–9052
53. Ho J, Ervin KM, Lineberger WC (1990) J Chem Phys 93:6987–7002
54. Negishi Y, Nakamura Y, Nakajima A (2001) J Chem Phys 115:3657–3663
55. Gray DE (1972) AIP Handbook, 3rd edn. McGraw-Hill, New York
56. Weast RC (1974) CRC Handbook of Chemistry and Physics, 55th edn. CRC, Cleveland
57. Wallace WT, Leavitt AJ, Whetten RL (2003) Chem Phys Lett 368:774–777
58. Ojifinni RA, Gong J, Froemming NS, Flaherty DW, Pan M, Henkelman G, Mullins CB (2008) J Am Chem Soc 130:11250–11251
59. Date M, Haruta M (2001) J Catal 201:221–224
60. Date M, Okumura M, Tsubota S, Haruta M (2004) Angew Chem Int Edn 43:2129–2132
61. Gong J, Mullins CB (2008) J Phys Chem C 112:17631–17634
62. Wang F, Zhang DJ, Xu XH, Ding Y (2009) J Phys Chem C 113:18032–18039

J Mol Model (2012) 18:1819–1840
DOI 10.1007/s00894-011-1198-x

ORIGINAL PAPER

Development of multiple QSAR models for consensus predictions and unified mechanistic interpretations of the free-radical scavenging activities of chromone derivatives

Indrani Mitra · Achintya Saha · Kunal Roy

Received: 7 June 2011 / Accepted: 20 July 2011 / Published online: 18 August 2011

Abstract Antioxidants are important defenders of the human body against nocive free radicals, which are the causative agents of most life-threatening diseases. The immense biomedicinal utility of antioxidants necessitates the development and design of new synthetic antioxidant molecules. The present report deals with the modeling of a series of chromone derivatives, which was done to provide detailed insight into the main structural fragments that impart antioxidant activity to these molecules. Four different quantitative structure–property relationship (QSAR) techniques, namely 3D pharmacophore mapping, comparative molecular similarity indices analysis (CoMSIA 3D-QSAR), hologram QSAR (HQSAR), and group-based QSAR (G-QSAR) techniques, were employed to obtain statistically significant models with encouraging external predictive potentials. Moreover, the visual contribution maps obtained for the different models signify the importance of different structural features in specific regions of the chromone nucleus. Additionally, the G-QSAR models determine the composite influence of pairs of substituent fragments on the overall antioxidant activity profiles of the molecules. Multiple models with different strategies for assessing structure–activity relationships were applied to reach a unified conclusion regarding the antioxidant mechanism and to provide consensus predictions, which are more reliable than values derived from a single model. The structural information obtained from the various QSAR models developed in the present work can thus be effectively utilized to design and predict the activities of new molecules belonging to the class of chromone derivatives.

Keyword Antioxidant · Hologram quantitative structure–activity relationship · Group-based quantitative structure–activity relationship · 3D pharmacophore · Comparative molecular similarity indices

Introduction

Free radicals pose a fatal threat to the healthy living of human beings. Evidence shows that free radicals and excited-state species play a key role in both normal biological function and the pathogeneses of certain human diseases [1], such as atherosclerosis [2], Alzheimer's disease [3], DNA mutations [4], and so on. Free radicals and other reactive oxygen species (ROS) are derived either from normal, essential metabolic processes in the human body or from external sources such as exposure to X-rays, ozone, passive uptake of cigarette smoke, air pollutants, and industrial chemicals [5]. The major source of ROS production within the human system is the mitochondrial respiratory chain [6]. These free radicals, which are highly unstable molecules, attempt to attain stability by reacting with reactive unsaturated molecules through four primary

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1198-x) contains supplementary material, which is available to authorized users.

I. Mitra · K. Roy (✉)
Drug Theoretics and Cheminformatics Laboratory,
Division of Medicinal and Pharmaceutical Chemistry,
Department of Pharmaceutical Technology, Jadavpur University,
Kolkata 700 032, India
e-mail: kunalroy_in@yahoo.com
URL: http://sites.google.com/site/kunalroyindia/

A. Saha
Department of Chemical Technology, University College
of Science and Technology, University of Calcutta,
92, A.P.C. Road,
Kolkata 700 009, India

types of chemical reaction: (a) hydrogen abstraction, (b) addition, (c) termination, and (d) disproportionation [7]. Proteins, phospholipids and polyunsaturated fatty acids are the molecules most susceptible to the free radical attack. Since the free radicals are produced naturally within the human system, the body has its own mechanism to combat these free radicals. Such detoxification processes employ a series of chemical entities referred to as antioxidants. Antioxidants primarily function by donating a hydrogen to the reactive radical, thereby terminating the chain reaction and affecting the rate of oxidation [8]. They may also chelate with the metal ions that catalyze the free-radical chain reactions.

The free radicals produced within the human system under normal conditions are either used up by the body's immune system to detect foreign invaders or damaged tissue, or they are detoxified by systemic antioxidant enzymes like superoxide dismutase, catalase, etc. [9]. Antioxidant activity is chiefly based on three molecular mechanisms: (a) hydrogen atom transfer (HAT), (b) single-electron transfer followed by proton transfer (SET-PT), and (c) sequential proton loss electron transfer (SPLET) [10–12]. However, under certain circumstances (such as acute or chronic alcohol exposure, or improper diet, etc.), either ROS production is enhanced or the level or activity of antioxidants is reduced. Such a state results in an imbalance between the production and the removal of the reactive radicals, which is followed by impairment of the body's ability to repair damaged complex molecules like proteins or DNA, leading to oxidative stress [13]. The normal defense mechanism of the body fails to repair the additional excessive changes caused by the ROS, leading to permanent changes or damage to the DNA [14], thus having potentially detrimental effects on the cell.

Free-radical production greatly outnumbers the systemic antioxidant supply, necessitating external antioxidant supplementation. Although fruits and vegetables serve as rich sources of antioxidants, several drugs (like vitamin C pills, probucol, etc.) that target the increased need for antioxidant supplementation are now also being marketed. The enhanced demand for antioxidants has led researchers to design and develop new chemical entities with improved antioxidant actions. The quantitative structure–activity relationship (QSAR) technique plays a crucial role in the design and screening of molecules with potent antioxidant activity. The QSAR technique attempts to correlate structural features with biological activity/toxicity/other physicochemical properties through the utilization of several descriptors [15, 16]. The descriptors are the numerical representations of the molecular properties defining the electronic, topological, physicochemical, and spatial features of the molecules. Similar molecules can exhibit large differences in their biological activities due to a minute difference in their structures. The QSAR technique focuses on these variations in biological activity with changes in molecular structure in a quantitative fashion. Initially developed by Hansch [17], and then advanced by several other researchers, the QSAR technique contributes efficiently to the screening of new chemical entities, thereby making the drug discovery pathway more cost-effective and concise.

In an attempt to design new chemical entities with efficient antioxidant activities, several researchers have performed QSAR analyses for different chemical classes of compounds with antioxidant activities. Multiple mechanisms underlying the reaction between hydroxyl radical and phenolic compounds have been studied by Cheng et al. [18] using the QSAR technique. Singh et al. [19] developed QSAR models using the quantitative topological molecular similarity (QTMS) method to compute the effects of substituents on the bond dissociation enthalpies (ΔBDEs) for a set of 39 phenolic derivatives that show antioxidant activity. Reis et al. [20] performed a theoretical study with 41 phenolic compounds that exhibit antioxidant properties, based on quantum chemical descriptors calculated at different levels of theory. Mitra et al. [21, 22] performed QSAR analyses for a variety of chemical classes (hydroxybenzalacetone [21], benzodioxoles [22], etc.) with efficient antioxidant activities using different categories of descriptors. Predictive pharmacophore models have also been developed by Mitra et al. [23] for a series of arylamino-substituted benzo[b]thiophenes that exhibit free-radical scavenging activity. Besides these, a variety of other QSAR models developed by different authors have been reviewed by Roy et al. [24]. In the present paper, a series of chromone derivatives reported by Samee et al. [25, 26] were modeled to determine their antioxidant activities using different QSAR techniques. The present work aimed at the development of QSAR models that can be used as query tools to search and screen large molecule databases for potent antioxidant molecules. The objective of this work was to reach a unified conclusion regarding the structure–activity relationship of antioxidant chromone derivatives, starting from multiple QSAR approaches and the consensus prediction of target response using robust statistical models. We also compared the statistical quality and observations of our models to the results achieved with the model developed by Samee et al. [25].

Methods and materials

The dataset

The model dataset used for the present work comprised of 36 synthetic chromone derivatives with antioxidant activity, as reported by Samee et al. [25, 26]. The antioxidant activities of the molecules were assessed [25, 26] based on their ability to scavenge 1,1-diphenyl-2-picryl hydrazyl (DPPH) free radicals. The 50% effective concentration (EC_{50}) of the molecules thus reported was converted to the nanomolar scale (nM) for the development of the 3D pharmacophore model. However, for the development of the CoMSIA and HQSAR models, EC_{50} values in millimolar units were converted to the negative logarithmic scale.

Splitting of the dataset into training and test sets

Training set selection plays an important role in the development of a statistically significant QSAR model. A QSAR model exhibits poor predictivity for test set molecules which are quite dissimilar from the training set ones, while good prediction results are obtained for molecules that are very similar to the training set molecules [27, 28]. Thus, the selection should be such that the test set molecules lie within the chemical space occupied by the training set molecules. In this study, the entire dataset was divided into training and test sets after activity ranking of the molecules under study. In this technique, all of the molecules were first ranked in ascending order of activity, and 25% of the compounds were then selected as the test set (n_{test}=9), while the remaining 75% (n_{training}=27) were used as the training set. The training set molecules were then utilized to develop the different QSAR models, while the predictive abilities of the models were assessed using the test set. In order to obtain a training set that captures the chemical features and range of activities of the entire dataset, the most active and least active compounds were placed in the training set. The training set thus selected spans the activity range of the entire dataset and hence yields unbiased results. In order to further ascertain the acceptability of the activity-based classification method for uniformly distributing the training and test set compounds, a principal component analysis (PCA) [29] of the descriptor matrix was performed using SPSS software [29]. The PCA score plot thus obtained shows the distribution of the training and test set compounds in the 3D space with respect to the first three principal components of the group-based QSAR (G-QSAR) descriptor matrix. The plot obtained (see Fig. S1 of the "Electronic supplementary material," ESM) shows that each of the test set compounds lies in close vicinity to at least one training set compound, indicating that the training set thus selected captures all of the essential features of the entire dataset of molecules.

Different methods employed for the development of QSAR models

The present work deals with the development of QSAR models for a set of chromone derivatives using four different techniques (Figs. 1, 2): (a) 3D pharmacophore generation, (b) a 3D-QSAR model developed using comparative molecular similarity indices analysis (CoMSIA), (c) fragment-based QSAR model developed using the hologram QSAR (HQSAR) technique, and (d) a group-based QSAR model (G-QSAR) based on fragment contributions. A 3D pharmacophore model quantitatively determines the features that are required to obtain the optimum activity of the molecules under study [30, 31]. For the present work, a 3D pharmacophore model was developed using conformers obtained from the BEST method of conformer generation,

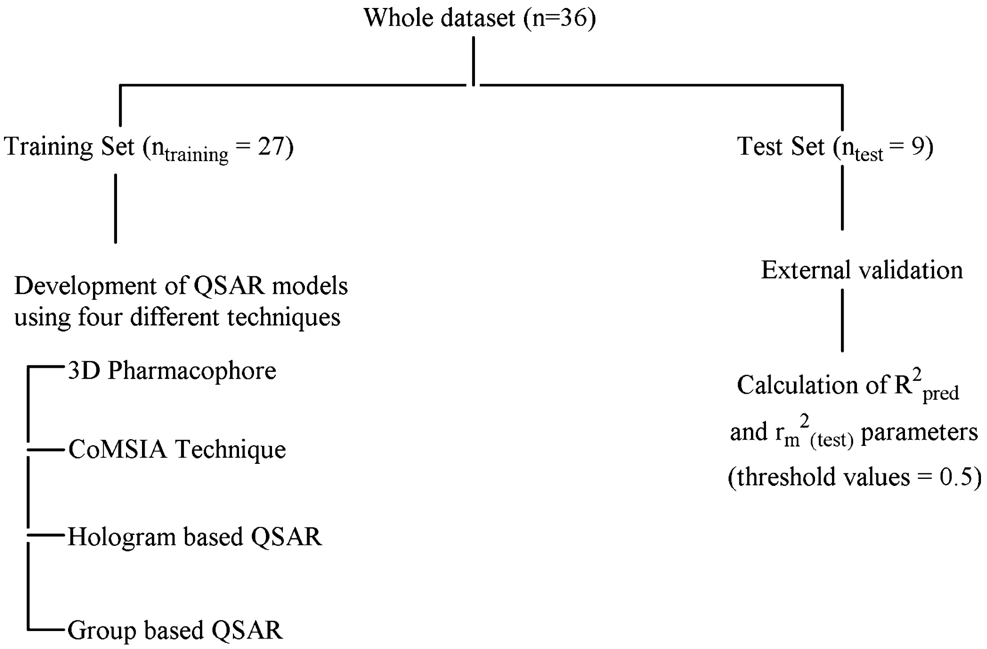
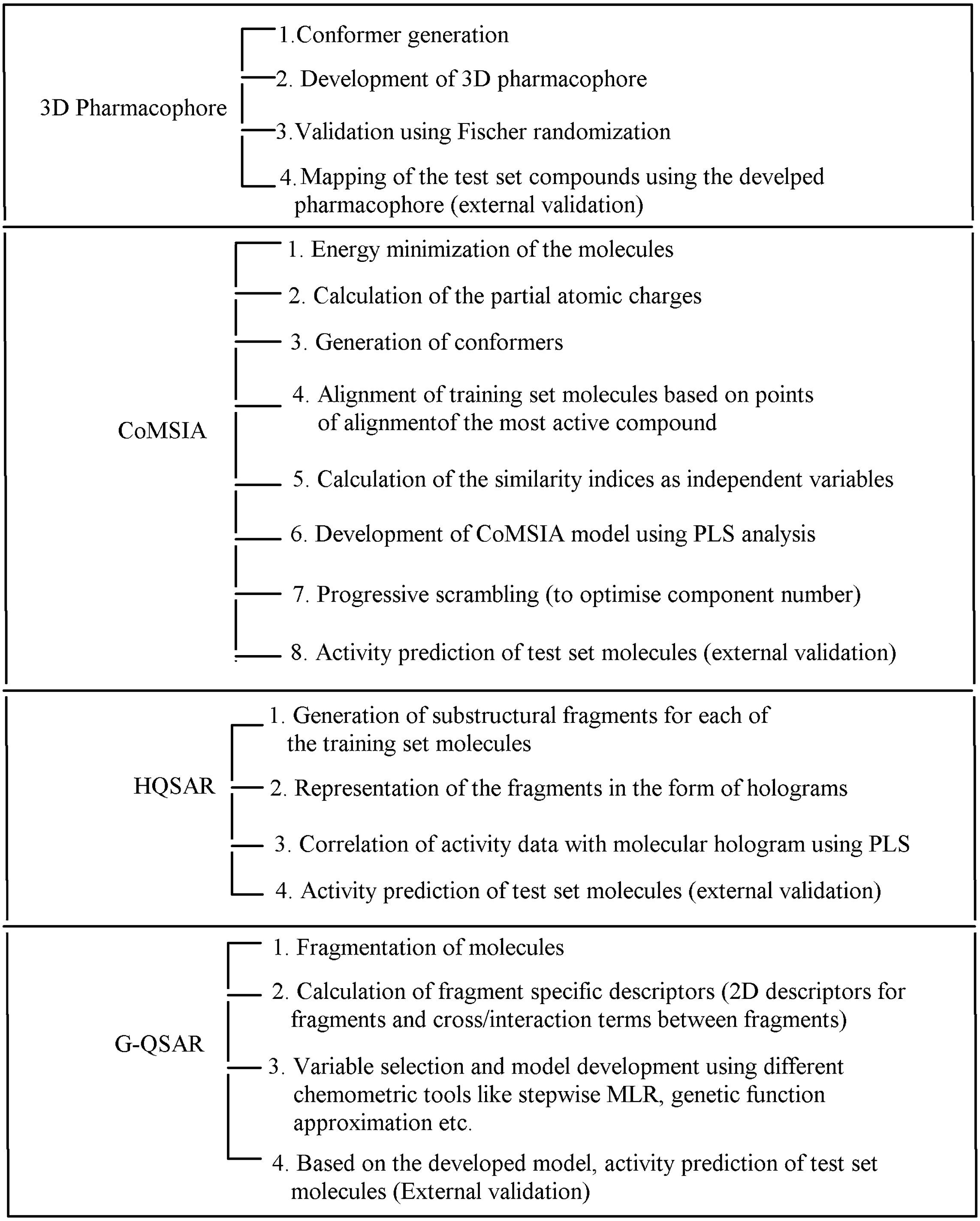

Whole dataset (n=36)
Training Set (n_{training} = 27)
Test Set (n_{test} = 9)
Development of QSAR models using four different techniques
3D Pharmacophore
CoMSIA Technique
Hologram based QSAR
Group based QSAR
External validation
Calculation of R^2_{pred} and $r_m^2{}_{(\mathrm{test})}$ parameters (threshold values = 0.5)

Fig. 1 Schematic representation of the QSAR methodology employed in the present work



3D Pharmacophore

1. Conformer generation
2. Development of 3D pharmacophore
3. Validation using Fischer randomization
4. Mapping of the test set compounds using the develped pharmacophore (external validation)

CoMSIA

1. Energy minimization of the molecules
2. Calculation of the partial atomic charges
3. Generation of conformers
4. Alignment of training set molecules based on points of alignmentof the most active compound
5. Calculation of the similarity indices as independent variables
6. Development of CoMSIA model using PLS analysis
7. Progressive scrambling (to optimise component number)
8. Activity prediction of test set molecules (external validation)

HQSAR

1. Generation of substructural fragments for each of the training set molecules
2. Representation of the fragments in the form of holograms
3. Correlation of activity data with molecular hologram using PLS
4. Activity prediction of test set molecules (external validation)

G-QSAR

1. Fragmentation of molecules
2. Calculation of fragment specific descriptors (2D descriptors for fragments and cross/interaction terms between fragments)
3. Variable selection and model development using different chemometric tools like stepwise MLR, genetic function approximation etc.
4. Based on the developed model, activity prediction of test set molecules (External validation)

Fig. 2 Steps associated with the different QSAR techniques performed in the present study

based on conformational analysis of the molecules using the poling algorithm [30]. To develop the model, the HypoGen module (see the ESM), implemented in Discovery Studio 2.1 [31], was employed. To assess the quality of the generated pharmacophore hypotheses, cost functions [32] (represented in bit units) were calculated during hypothesis generation. The statistical significance of the pharmacophore model was determined using a randomization test and external validation techniques. A pharmacophore model was considered to be generated by chance if the randomized dataset yielded better results than the original one. Calculation of the corrected R_p^2 value ($^cR_p^2$ statistic) [33] provides a quantitative approach to ascertain the existence of a chance correlation. Additionally, external validation of the model was performed using the test set compounds. The 3D pharmacophore developed using the training set compounds was used to map the test set compounds, and the activity of the test molecules was estimated based on the degree of mapping and the calculated value of the external predictive parameter (R^2_{pred}, with a threshold value of 0.5) [34]. Furthermore, to better determine the external predictive potential of the developed 3D pharmacophore model, the value of modified r^2 for the test set $\left[r^2_{\mathrm{m\ (test)}}\right]$ was also calculated [35–38].

Another method employed to determine the essential structural features of the chromone derivatives include the CoMSIA technique [39]. This technique provides a correlation between the differences in the biological activities of molecules and changes in molecular properties represented by differences in the shapes of the noncovalent fields surrounding the molecules. In the present work, CoMSIA models were developed based on the training set molecules and were subsequently validated using the external validation technique. Based on the favorable or unfavorable environments of the five fields (steric, electrostatic, hydrophobic, hydrogen-bond donor, and hydrogen-bond acceptor), molecular fragments and functional groups that make major contributions to the activity profiles of the molecules are determined [40]. Partial atomic charges of the molecules were calculated by the Gasteiger–Huckel method [41], and energy minimization was performed using the Tripos force field [42] method. The conformers were generated using the simulated annealing technique [43]. The training set molecules were then aligned based on the points of alignment of the most active compound (compound no. **29**), which was used as the template molecule (Fig. 3). The alignment of the molecules was performed using the database alignment technique (Fig. 4) implemented in the Sybyl software package [44]. Subsequently, the partial least squares (PLS) approach [45, 46] was used to derive the 3D-QSAR models using the similarity (CoMSIA) factors as independent variables and the antioxidant activity (pEC_{50}) as the dependent variable. The optimized CoMSIA model was evaluated based on progressive scrambling analyses [47]. The final model, derived with an optimum number of components, was subjected to external validation (calculation of the R^2_{pred} parameter) using the test set compounds.

The next approach employed to correlate the biological activity data of the chromone derivatives to structural fragments of molecules was the HQSAR methodology [48]. Molecular hologram is an extended form of the fingerprint encoding of all possible molecular fragments, including linear, branched, cyclic, and overlapping features of the molecules. In the present work, the HQSAR model was derived based on various combinations of fragment distinction and fragment generation parameters for each hologram length using the Sybyl software [44]. The selection of the statistically significant model at an optimum component number was performed based on the maximum value of Q^2 and the minimum value of the cross-validated standard error ($\mathrm{SE}_{\mathrm{cv}}$). The optimum number of components was also further checked based on the "5% rule," which permitted the addition of a latent variable only when the addition resulted in an increase in the value of Q^2 by 5% or more. However, the chances of overfitting the developed model were reduced by limiting the maximum number of components to $R/5$ (R is the number of training set compounds) [49]. The QSAR analysis was redone and the component number was selected. The final PLS model was obtained with the optimum component number based on the specific fragment distinction parameters, fragment size and bin length. The accuracy of prediction of the test set activity data using the developed HQSAR model was judged based on the value of the R^2_{pred} parameter. To penalize the overfitting nature of the external $\left(R^2_{\mathrm{pred}}\right)$ predictive parameter, the values of the r^2_{m} metric were also calculated.

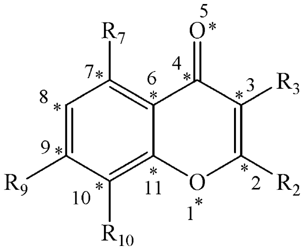


Fig. 3 Template molecule used for CoMSIA (atoms used for alignment are marked with an *asterisk*) and G-QSAR model development

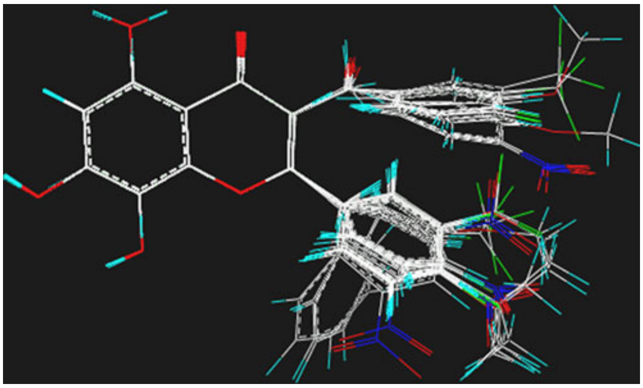


Fig. 4 Aligned geometries of the 36 chromone derivatives

Finally, the G-QSAR technique [50] was employed to derive a quantitative relationship between the activity and descriptors calculated for various molecular fragments of interest. The G-QSAR technique available within the VLife MDS 3.5 software package [51] begins with the fragmentation of the molecules under study, followed by the calculation of fragment-specific descriptors. Among the different variable selection techniques available within the program, the stepwise multiple linear regression (stepwise MLR) method based on forward selection and backward elimination techniques according to the "stepping criteria" [52, 53] (F=4 for inclusion and F=3.9 for exclusion) was employed for the present work. Finally, two different group-based QSAR models were developed, one with the fragment contribution descriptors only, and the other based on both the 2D descriptors and the cross-interaction terms. The models were subsequently validated externally using the R^2_{pred} and $r^2_{\mathrm{m\ (test)}}$ parameters. Since the G-QSAR-derived models could specifically determine the physicochemical and topological requirements of the different substituents, these emerged as the most acceptable ones in terms of both interpretability and statistical significance. Thus, the G-QSAR fragment contribution model was utilized to determine the normality of the distribution [54–56] of the residuals of the training set data and the corresponding applicability domain [57, 58] of the molecules. The applicability domain of a molecule determines its chemical space in terms of model descriptors and modeled response. In this work, the applicability domain of the molecules was checked using the leverage approach [57]. Predictions are considered unreliable for compounds with leverage values greater than the critical one ($h > h^*$, the critical value being $h{=}3p^{/}/n$, where $p^{/}$ is the number of model variables plus one, and n is the number of the compounds used to calculate the model). Compounds with cross-validated standardized residuals that are greater than three standard deviation units (>3σ) are response outliers.

Results and discussion

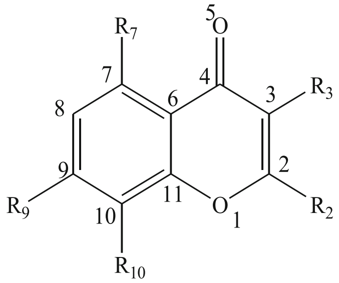
The four types of QSAR models thus developed were analyzed to determine their statistical significances and to ascertain the prime structural requirements for improved antioxidant activities of the chromone derivatives, with the aim being to reach a unified conclusion about the mechanism of antioxidant activity. The structures of all the dataset molecules, together with their observed and predicted/calculated activity data, are listed in Table 1. An overview of the results obtained from the four different methods employed in the present work for QSAR model development are detailed in Table 2. The observed vs. calculated/predicted activity data were plotted graphically, and the points obtained were minimally scattered about the diagonal of the scatter plot. This indicated that the models developed could calculate/predict the activity data of the molecules satisfactorily, and the calculated/predicted and observed activity data of the molecules lie in close proximity to each other.

Analysis of the 3D pharmacophore model

A set of ten pharmacophore hypotheses were developed using 27 training set compounds based on the conformers obtained using the BEST method of conformer generation. All the hypotheses are summarized in Table 3 together with their statistical parameters, such as cost functions, rms deviations, and correlation coefficients. Further, the fitnesses of the developed pharmacophore models were checked using the Fischer validation technique at the 95% confidence level. As the value of R_{r} for hypothesis 10 was much lower than the corresponding correlation coefficient (R) of the unrandomized matrix, model 10 (R_{r}=0.491, R=0.912) was selected for further analysis. Based on the results of the randomized data, the value of ${}^{c}R^2_{\mathrm{p}}$ was also calculated. Since the parameter penalizes the model R^2 for small differences in the values of R^2 and $\mathrm{R}^2_{\mathrm{r}}$, it provides a precise approach for selecting models that did not develop by chance. For hypothesis 10, the value of ${}^{c}R^2_{\mathrm{p}}$ (0.821) thus calculated was much higher than its stipulated threshold value of 0.5, implying that the model represents a true correlation and is not the outcome of mere chance. Moreover, the randomization results obtained for the cost functions showed that the total cost for hypothesis 10 was much closer to that of the null cost compared to the fixed cost, as calculated based on the scrambled activity data. Thus, the existence of true correlation for hypothesis 10 is reflected in the values of cost functions, the correlation coefficient, and the rms deviation for the model. Subsequently, hypothesis 10 (Fig. 5) was selected as the best-ranking pharmacophore and was analyzed further.

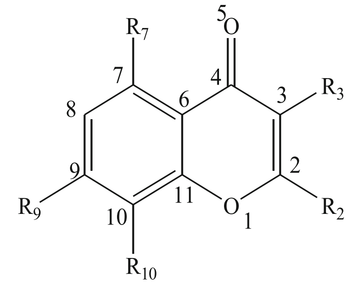
Four different chemical features are displayed in hypothesis 10: HBA, HBA, HY and RA. The vectors for the HBA features indicate the direction of formation of the hydrogen bond between the electronegative atom and the electropositive hydrogen atom of the free radical. Similarly, the vector for the ring aromatic feature indicates the direction of the π–π interaction between an electron rich and an electron deficient aromatic centers. The most active compound (compound no. **29**) was mapped using hypothesis 10 (Fig. 5b), which revealed that the ketonic oxygen of the parent choromone nucleus and the hydroxy substituent at the R10 position of the chromone moiety behave as hydrogen-bond acceptor groups. The presence of such hydrogen-bond acceptor groups means that a nucleophilic center is needed within the antioxidant molecules to obtain their activity profile. The hydrogen-bond acceptor groups contribute to the antioxidant's mechanism of action by transferring a single electron and then deprotonating [10]. Again, the presence of ring aromatic and hydrophobic features indicates that molecules bearing hydrophobic substituents develop an area of transient electron deficiency [59], which may in turn interact with nearby free radicals that have transient electron-rich areas. Thus the presence of a phenyl substituent at the R2 position of the chromone nucleus favors the antioxidant activity profiles of the molecules. In addition, the hydrophobic feature covering the substituent at the R3 position of the chromone moiety indicates that aliphatic or aromatic fragments with hydrophobic functions add to the activity profiles of the molecules. The pharmacophore obtained in hypothesis 10 was found to efficiently map

Table 1 Molecular structures of the 36 dataset compounds along with their observed and predicted activity data



Sl. No.	R_2	R_3	R_7	R_9	R_{10}	Observed activity [25]	Calculated/ Predicted activity[a]	Calculated/ Predicted activity[b]	Calculated/ Predicted activity[c]	Calculated/ Predicted activity[d]	Calculated/ Predicted activity[e]
1	Phenyl	H	H	H	OH	1.017	1.100	0.985	1.145	1.023	0.925
2	Phenyl	H	H	OH	H	0.901	0.912	0.802	0.905	0.803	0.944
3	Benzyl	H	H	OH	H	0.903	0.969	0.879	0.877	0.826	0.956
4	$4'$-(NO_2)-phenyl	H	H	OH	H	0.992	1.097	1.008	0.962	0.987	1.031
5	$3'$-(CF_3)-phenyl	H	H	OH	H	1.030	1.081	0.936	0.997	1.058	1.052
6*	$4'$-(F)-phenyl	H	H	OH	H	0.946	0.838	0.811	0.900	0.900	0.986
7	$3'$, $5'$-(diNO_2)-phenyl	H	H	OH	H	1.058	1.105	0.995	0.985	1.122	1.094
8	$3'$-(Cl)-phenyl	H	H	OH	H	0.932	0.976	0.885	0.879	0.946	0.983
9	$4'$-(*t*-butyl)-phenyl	H	H	OH	H	0.981	1.083	0.955	0.923	1.015	1.050
10*	Phenyl	CH_3	H	OH	H	0.906	0.589	0.847	0.965	0.787	0.959
11	Benzyl	CH_3	H	OH	H	0.908	0.615	0.936	0.971	0.810	0.971
12	$4'$-(NO_2)-phenyl	$4''$-(NO_2)-benzoyl		OH	H	1.227	1.339	1.232	1.266	1.258	1.146
13	$3'$-(CF_3)-phenyl	$3''$-(CF_3)-benzoyl		OH	H	1.265	1.353	1.398	1.412	1.438	1.139
14	$4'$-(F)-phenyl	$4''$-(F)-benzoyl		OH	H	1.139	1.383	1.278	1.210	1.210	1.087
15	$3'$, $4'$-(diF)-phenyl	$3''$, $4''$-(diF)-benzoyl		OH	H	1.201	1.388	1.262	1.265	1.285	1.114
16*	$4'$-(OCH_3)-phenyl	$4''$-(OCH_3)-benzoyl		OH	H	1.150	1.323	1.149	1.246	1.225	1.127
17	$3'$-(CF_3)-phenyl	H	OH	OH	H	1.068	1.059	1.095	1.044	1.059	1.053
18*	$4'$-(F)-phenyl	H	OH	OH	H	0.991	0.843	0.967	0.946	0.900	0.986
19	$3'$, $4'$-(diF)-phenyl	H	OH	OH	H	1.006	0.819	0.992	0.974	0.957	1.008
20*	$4'$-(*t*-butyl)-phenyl	H	OH	OH	H	1.059	1.074	1.109	0.970	1.015	1.050
21	$3'$-(Cl)-phenyl	H	OH	OH	H	0.982	0.984	1.043	0.926	0.946	0.983
22	$3'$, $4'$-(diCl)-phenyl	H	OH	OH	H	1.045	1.021	1.058	0.948	1.055	1.005
23	$4'$-(OCH_3)-phenyl	H	OH	OH	H	0.961	1.061	0.970	0.932	0.938	1.008
24	$3'$-(OCH_3)-phenyl	H	OH	OH	H	0.952	1.077	0.925	0.964	0.938	1.008
25	$3'$, $5'$-(diNO_2)-phenyl	H	OH	OH	H	1.098	1.071	1.179	1.039	1.122	1.094

Table 1 (continued)



Sl. No.	R_2	R_3	R_7	R_9	R_{10}	Observed activity [25]	Calculated/ Predicted activity[a]	Calculated/ Predicted activity[b]	Calculated/ Predicted activity[c]	Calculated/ Predicted activity[d]	Calculated/ Predicted activity[e]
26*	4′-(NO_2)-phenyl	4″-(NO_2)-benzoyl	OH	OH	H	1.240	1.268	1.363	1.370	1.258	1.146
27	Phenyl	H	H	OH	OH	1.496	1.235	1.660	1.586	1.689	1.507
28*	Benzyl	H	H	OH	OH	1.419	1.256	1.749	1.537	1.644	1.528
29	3′-(CF_3)-phenyl	3″-(CF_3)-benzoyl	H	OH	OH	2.588	2.359	2.167	2.522	2.394	2.651
30	4′-(F)-	4″-(F)-	H	OH	OH	2.406	2.354	0.645	2.316	2.183	2.386
	phenyl	benzoyl									
31	CH_3	H	H	OH	H	0.738	0.802	1.008	0.789	0.597	0.828
32*	3′, 4′-(diCl)-phenyl	H	H	OH	H	0.999	1.039	0.900	0.901	1.020	0.989
33	4′-(NO_2)-phenyl	H	H	OH	H	1.044	1.097	1.509	0.962	0.988	1.031
34	CH_3	H	H	OH	OH	1.385	0.823	1.142	1.383	1.609	1.295
35	3′-(OCH_3)-phenyl	3″-(OCH_3)-benzoyl	H	OH	H	1.153	1.393	2.537	1.297	1.223	1.127
36*	4′-(NO_2)-phenyl	4″-(NO_2)-benzoyl	H	OH	OH	2.472	2.279	2.11	2.367	2.212	2.589

[*] Test set compounds

[a] Activity calculated/ predicted based on the 3D pharmacophore model

[b] Activity calculated/ predicted based on the 3D-QSAR model developed using CoMSIA technique

[c] Activity calculated/ predicted based on the model developed from the HQSAR technique

[d] Activity calculated/ predicted based on the G-QSAR model

[e] Activity calculated/ predicted based on the G-QSAR_IT model

Table 2 Comparison of the statistical parameters of the three types of QSAR models developed in the present work

	3D pharmacophore model	3D-QSAR model (COMSIA analysis)	HQSAR analysis	G-QSAR analysis	
				G-QSAR	G-QSAR_IT
Leave-one-out cross-validation					
R^2	0.832	0.957	0.970	0.980	0.937
Q^2	-	0.834	0.932	0.851	0.963
Components/ descriptors	-	5	3	4	3
Progressive scrambling statistics at critical $r^2_{yy'}$ of 0.85					
Q^2	-	0.597	-	-	-
$dQ^2/dr^2_{yy'}$	-	0.944	-	-	-
Randomization test					
R^2_{r}	0.491	-	-	-	-
$^{c}R^2_{\mathrm{p}}$	0.821	-	-	-	-
Predictions of the test set					
R^2_{pred}	0.883	0.852	0.961	0.923	0.980
$r^2_{\mathrm{m\ (test)}}$	0.826	0.845	0.957	0.910	0.925

Table 3 Results for ten pharmacophore hypotheses generated using conformers developed based on the BEST method of conformer search

Hypothesis no.	Total cost	Error cost	rms	Correlation (*R*)	Configuration cost	Features [a]	R^2_{pred}	$r^2_{\mathrm{m\ (test)}}$	$^{c}R^2_{\mathrm{p}}$
1	84.596	67.732	0.532	0.973	14.950	HBA, HBD, RA	0.283	-	-
2	85.510	68.753	0.599	0.966	14.950	HBA, HBD, RA	0.182	-	-
3	86.102	69.190	0.626	0.963	14.950	HBA, HBD, RA	0.609	0.635	0.768
4	86.141	69.242	0.629	0.962	14.950	HBA, HBD, RA	0.648	0.645	0.784
5	91.771	73.395	0.838	0.933	14.950	HBA, HBD, RA	0.487	-	-
6	93.997	75.490	0.926	0.917	14.950	HBA, HBD, RA	0.625	0.584	0.699
7	94.329	76.938	0.983	0.906	14.950	HBD, HY, RA	0.439	-	-
8	95.088	77.055	0.987	0.905	14.950	HBD, HY, RA	0.843	0.857	0.671
9	96.805	78.391	1.036	0.895	14.950	HBD, HY, RA	0.500	0.429	0.680
10	97.133	76.903	0.981	0.907	14.950	HBA, HBA, HY, RA	0.883	0.826	0.821

Fixed cost: 79.979

Null cost: 136.211

[a] *HBD* hydrogen-bond donor, *HBA* hydrogen-bond acceptor, *HY* hydrophobic, *RA* ring aromatic

to the active compounds (compound nos. **29**, **30**, **12**, **13** and **15**) of the dataset employed. On the contrary, compounds that were unable to map to all of the features revealed a lack of the necessary substitutions, and hence exhibited reduced activity profiles.

Further, the pharmacophore was validated to assess its external predictive ability. The test set molecules were thus mapped using the pharmacophore obtained in hypothesis 10, and their activities were estimated based on their ability to capture the pharmacophoric features. The external predictive potential of the developed model was measured based on the value of the predictive R^2 $\left(R^2_{\mathrm{pred}}\right)$. A good overall correlation between the observed and predicted activity data was reflected in the value of the $R^2_{\mathrm{pred}}(0.883)$ parameter, which was much higher than the threshold value of 0.5. Thus, the compound that matched

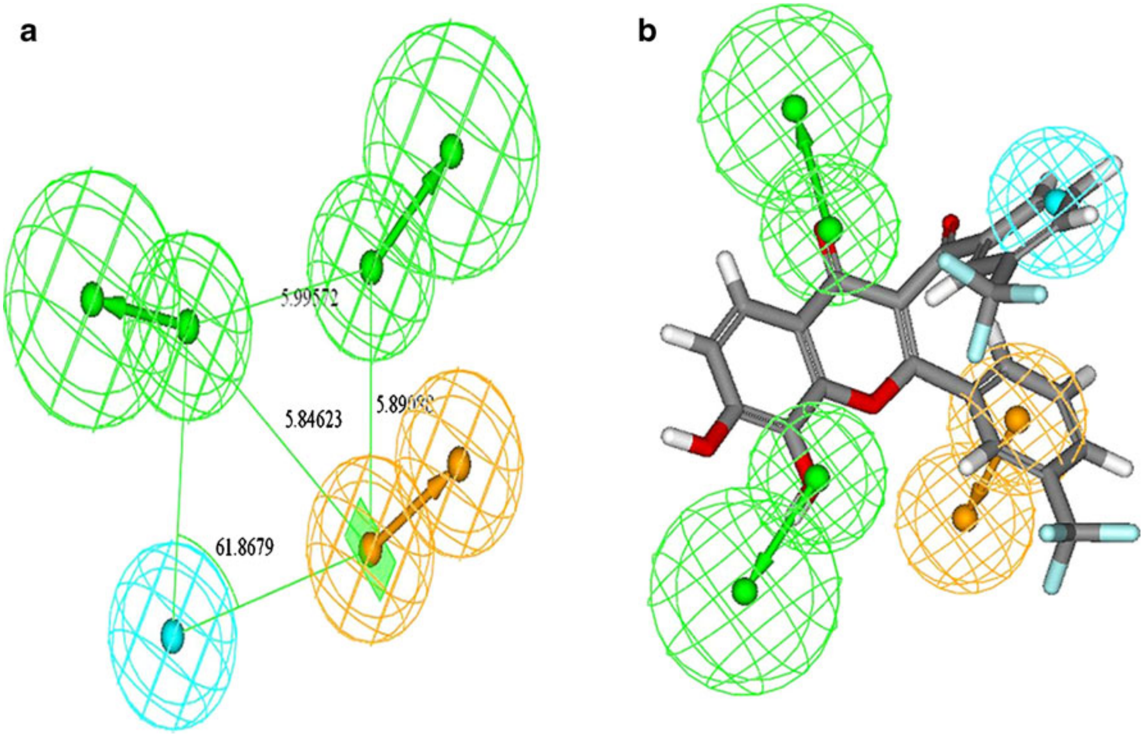


Fig. 5 **a**–**b** Pharmacophore obtained from hypothesis 10 by monitoring the positions of the different features (**a**), and the mapping of the most active compound (compound no. **29**) to the developed pharmacophore (**b**). Shown are ring aromatic sphere (*orange*), hydrophobic group (*cyan*) and hydrogen bond acceptor (*green*) features with vectors in the direction of the putative hydrogen bonds

all four features (compound no. **36**) exhibited a better activity profile than those of compounds (least active compounds, like compound nos. **6**, **10**, **18** and **32**) that are unable to map all of the essential features. This indicated that the lack of all of the necessary structural attributes by the poorly mapped compounds is responsible for their reduced activity profiles. Further, the proximity between the observed and the predicted activity data was checked based on the value of the r^2_{m} metric. A value of this parameter that is higher than 0.5 ensures closeness between the observed and predicted data more precisely than the traditional parameter, R^2_{pred}. A high value of $r^2_{\mathrm{m\ (test)}}$ parameter (0.826) for the selected hypothesis indicates that the pharmacophore obtained reflects the statistical significance and has enhanced external predictive potential. Figure 6 shows scatter plots for the observed vs. calculated/predicted values of different models. A good correlation between the observed and calculated/predicted activity data is revealed for the pharmacophore model by the scatter plot shown in Fig. 6a.

Results obtained for the 3D-QSAR study performed using the CoMSIA technique

The results of the CoMSIA study are summarized in Table 4. Various combinations of the five different intrinsic properties available in the CoMSIA technique were analyzed by mapping the training set compounds using the PLS method of regression and assessing the various statistical parameters. Based on the values of R^2 and Q^2, and the standard error of estimation (s), the best map (Fig. 7a and b) for the CoMSIA study was generated from four types of interactions (hydrogen bond acceptor, hydrogen bond donor, hydrophobic and steric) of the training set molecules with the probe atom. Combining these properties with the electrostatic interaction did not result in any further improvement in the result, and hence the model with these four intrinsic properties (a, d, p and s) was analyzed further. The contributions of these four properties to the best map were 15.5%, 40.6%, 32%, and 11.9%, respectively. The model yielded statistically significant values of the R^2 (0.957) and Q^2 (0.834) parameters at a component number of 5, signifying acceptable internal predictive ability and self-consistency of the developed model. Among the other parameters calculated, a high value of boot-strapped R^2 $\left(R^2_{\mathrm{bs}} = 0.963\right)$ and a low standard error of estimation (s=0.096) further support the acceptability of the developed model. In order to optimize the number of components used to develop the CoMSIA model, and to assess the sensitivity of the model to chance correlations, random progressive scrambling was performed for the best PLS analysis. The results further confirmed the consistency of the models as defined by the slope $\left(dQ^2/dr^2_{yy'}\right)$ and the optimum value of the Q^2 statistic obtained at the ends of different runs. At a critical $r^2_{yy'}$ (correlation of original and scrambled data) of 0.85, progressive scrambling resulted in derivatives of Q^2 (leave-one-out) with respect to $r^2_{yy'}$ that were close to 1 (0.944), and the maximum value of Q^2 was obtained for a five-component PLS model. Thus, it can be inferred that the degree of redundancy within the training set molecules is sufficiently low for five components. The model was additionally analyzed for its external predictive potential by mapping the test set molecules to the four essential features obtained using the training set molecules. A value of the $R^2_{\mathrm{pred}}(0.852)$ parameter that was much higher than the stipulated value of 0.5 indicated that the CoMSIA model was reproducible in terms of activity predictions for the test set molecules. Again, the r^2_{m} metric $\left(r^2_{\mathrm{m\ (test)}} = 0.845\right)$ was calculated to ensure that the predicted activity data lies in close vicinity to the corresponding observed activity data. The contour map thus obtained was analyzed further, and the scatter plot (Fig. 6b) obtained for the observed and predicted/calculated activity data revealed the existence of a significant correlation between them.

Figure 7 shows the contour map for the best PLS model obtained using the CoMSIA analysis, with the most active compound (compound no. **29**) mapped to the essential intrinsic features. The presence of hydrogen-bond acceptor groups is favored at regions indicated by the magenta contour, and disfavored at regions bearing the red contour. The presence of the magenta contour over the ketonic fragment of the benzoyl substituent at the R3 position and the hydroxyl fragment (−OH) at the R10 position of the chromone moiety indicates that these substituents are required for the activity profiles of the chromone derivatives. Although they all bear the benzoyl substituent, compound nos. **12**, **13**, **14**, **15**, **16**, **26** and **35** exhibit moderate activity profiles due to the absence of the hydroxyl group at the R10 position. Again, compound nos. **27** and **28**, despite having the –OH substitution at the R10 position, lie in the moderate activity range due to the absence of the other hydrogen-bond acceptor feature (i.e., the ketonic fragment). Thus, compound nos. **29**, **30** and **36**, which bear both of the hydrogen-bond acceptor features, exhibit the maximum activity profile. The cyan and purple contours reflect the favored and disfavored regions for the presence of hydrogen-bond donor groups, respectively. The presence of a purple contour over the ketonic fragment of the chromone nucleus signifies that hydrogen-bond donor groups are disfavored at the C4 position. The favored and disfavored regions for the hydrophobic feature are indicated by the yellow and white contours, respectively. The position of the yellow contour indicates that the presence of a hydrophobic substituent at the *meta* or *para* position of the aromatic substituent at the R2 position of the parent moiety is hydrophobically favored. Such a similar yellow contour is also present over the carbon fragment substituted onto the

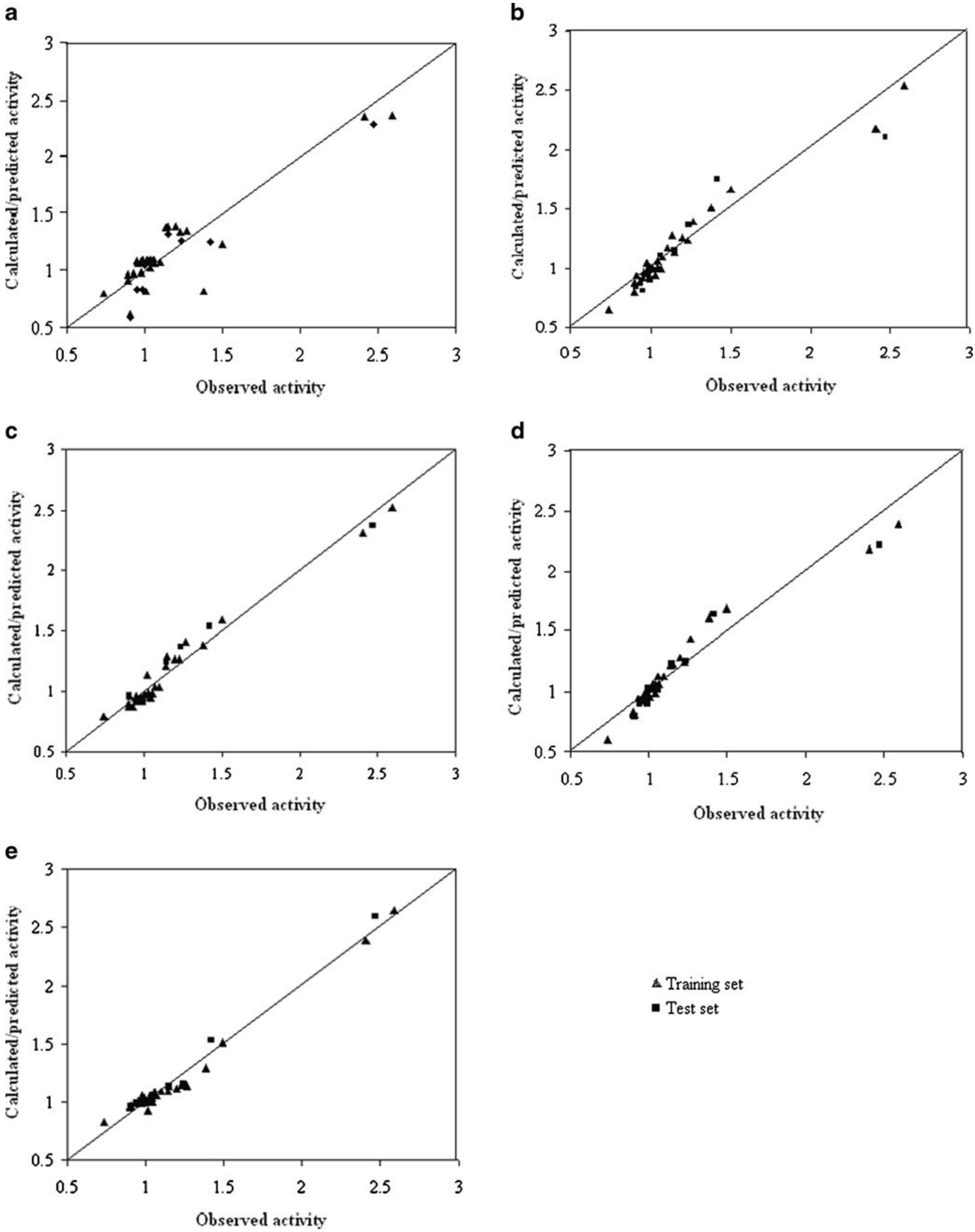


Fig. 6 **a**–**e** Scatter plots of the observed vs calculated/predicted activity values for the **a** 3D pharmacophore model, **b** 3D-QSAR model developed using the CoMSIA technique, **c** HQSAR model, **d** G-QSAR model, and **e** G-QSAR_IT model

Table 4 Summary of the CoMSIA study

Interactions	a*	d*	p*	s*	a+d+p	a+d+p+s
Components	4	5	5	5	5	5
n_{training}	27	27	27	27	27	27
R^2	0.886	0.711	0.900	0.853	0.953	0.957
s	0.156	0.249	0.147	0.178	0.101	0.096
F (df)	32.604 (4, 22)	10.356 (5, 21)	37.718 (5, 21)	24.320 (5, 21)	84.927 (5, 21)	93.398 (5, 21)
Q^2	0.670	0.472	0.420	0.494	0.818	0.834
Standard error of prediction (SEP)	0.260	0.337	0.353	0.329	0.198	0.188
R^2_{bs}	0.900	0.795	0.945	0.901	0.963	0.963
Standard deviation (s_{bs})	0.125	0.219	0.096	0.145	0.072	0.077
Contribution	a	d	p	s	a+d+p	a+d+p+s
a	1.000	-	-	-	0.186	0.155
d	-	1.000	-	-	0.422	0.406
p	-	-	1.000	-	0.392	0.320
s	-	-	-	1.000	-	0.119

* *a* hydrogen-bond acceptor feature, *d* hydrogen-bond donor feature, *p* hydrophobic feature, *s* steric feature

aromatic ring constituting the R3 position of the chromone nucleus. This indicates that these positions are hydrophobically favored, and the presence of substituents with transient electron density at these positions enhances the antioxidant activity profiles of these molecules, as seen for compound nos. **29**, **30** and **36**. Again, compound no. **25**, despite bearing the necessary substituent on the aromatic ring at the R2 position, has only a moderate activity profile due to its lack of a suitable substituent at the R3 position. Finally, the green and blue contours refer to the sterically favored and disfavored regions, respectively. The presence of the green contour over the benzoyl moiety indicates that the presence of such a substituent adds to the activity profiles of these molecules. Compared to other aromatic substituents, the benzoyl fragment increases the length of the molecule, enabling it to reach the sterically favored pocket, which in turn signifies an increase in the antioxidant activity profiles of the molecules. Again, since a bulky

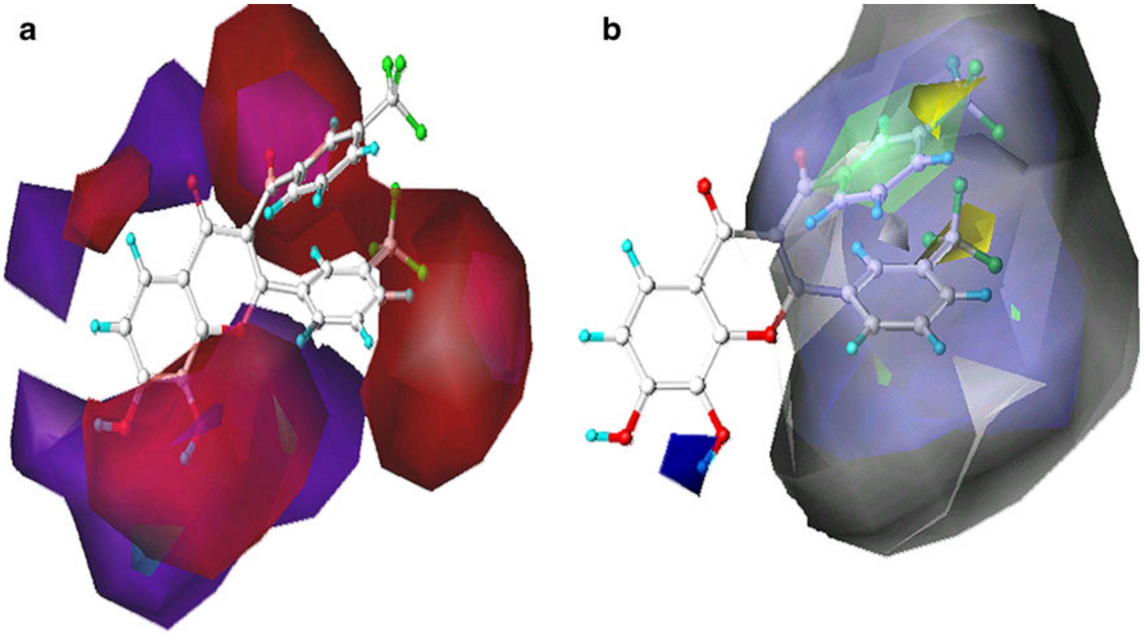


Fig. 7 **a**–**b** Space-mapped features from the CoMSIA study fitted to compound no. **29**. **a** Hydrogen-bond acceptor (*magenta*: favorable, *red*: unfavorable) and hydrogen-bond donor (*cyan*: favorable, *purple*: unfavorable) features. **b** Hydrophobic (*yellow*: favorable, *white*: unfavorable) and steric (*green*: favorable, *blue*: unfavorable) features

substituent is disfavored at the R2 position of the chromone nucleus (it has a blue contour), compound nos. **2**, **3**, **6**, **8**, **10** and **11**, which bear phenyl substituents at the R2 position, exhibit the lowest activity profiles. Compound no. **31**, which lacks all of the essential features, exhibits the lowest antioxidant activity. While compound nos. **12**, **13**, **14**, **15** and **35** have most of the intrinsic features of the CoMSIA contour map, these compounds exhibit moderate activity profiles due to a lack of the hydroxyl group at the R10 position. On the other hand, compound no. **27**, which bears the necessary hydrogen-bond acceptor substituent, has a reduced activity profile due to a lack of the hydrophobic substitution at the necessary position.

HQSAR method employed for model development

The results of the HQSAR analysis are reported in Tables 5, 6 and 7. The analyses were first performed based on the training set molecules using the default fragment length with different combinations of the six fragment distinction features. Based on the values of maximum Q^2 and minimum cross-validated standard error (SE_{cv}), the best combination of the fragment features was selected [A (atom type), C (connectivity) and D&A (donor and acceptor)]. The best fragment combination was then used to select the most suitable fragment size. The fragment size and the fragment combination thus optimized were utilized to select the significant hologram length. The fragment size, hologram length, and the optimum component number were selected based on the PLS analyses that yielded the lowest SE_{cv} and the highest Q^2. The final model was obtained by repeating the analysis using the specific fragment contribution, fragment size (5–10), hologram length (97), and optimum component number (5). Thereafter, the "5% rule" was employed to reduce the noise and obtain a more robust model. Thus, the model with the lowest component number (3) and highest Q^2 value (0.932) was selected as the best one according to the 5% rule. The model thus obtained was validated externally using the test set molecules. The activity predicted for the test set molecules was highly correlated with the observed activity data, and yielded a significantly high value for the R^2_{pred} parameter (0.961). A statistically significant value for the $r^2_{m\ (test)}$ (0.960) parameter further confirmed the close proximity of the observed and predicted activity data. Thus, the acceptable values for all internal and external predictive parameters imply that the model is robust and exhibits a high degree of external predictive potential. A significant correlation between the observed and predicted activity data was also revealed by the scatter plot, as shown in Fig. 6c.

The results of the HQSAR analysis are represented graphically in the form of a contribution map (Fig. 8), where the color of the atom or fragment determines its overall contribution to the activity profiles of the molecules under study. The maximum common substructure is colored cyan, and the contributions of the other colors are listed as follows: (i) a white color indicates an average contribution ranging from −0.097 to 0.102, (ii) a yellow color indicates a good contribution of 0.102 to 0.153, and (iii) a green color signifies the maximum contribution of 0.254 or above. For the purpose of discussion, the contributions of the different fragments with respect to the most active compound (compound no. **29**) are shown here. The chromone nucleus is present in all of the molecules—it is the common substructure—and is colored cyan. The fragments colored green, indicating maximum contributions, include: (i) the hydroxyl groups at the R9 and R10 positions of the chromone nucleus; (ii) the ketonic fragment of the benzoyl substituent at the R3 position of the parent moiety, and; (iii) the primary carbon fragment attached at the $3^{/}$ position of the aromatic substituent at R3. The fragments that contribute moderately (yellow colored) to the activity profile constitute the substituted aromatic carbon at the $3^{/}$ position of the benzoyl ring at R3, and the fluorine atom comprising the CF_3 fragment substituted at the same position on the benzoyl ring. The white coloration of the substituent at the R2 position indicates that it has minimal impact on the antioxidant activity profiles of the molecules.

Table 5 HQSAR analysis for various fragment distinctions using the default fragment size (4–7); $LV_{max}=5$

Fragment distinction	Q^2	SE_{cv}	R^2	SE	LV	Hologram length
A/B*	0.736	0.238	0.926	0.126	5	199
A/B/C*	0.735	0.238	0.928	0.124	5	83
A/B/H*	0.664	0.269	0.915	0.135	5	83
A/B/C/H*	0.698	0.254	0.923	0.128	5	307
A/C/D & A*	0.842	0.180	0.948	0.103	4	53
A/C/Ch*	0.722	0.244	0.914	0.136	5	97
A/B/Ch*	0.723	0.244	0.926	0.126	5	307
A/B/C/H/Ch*	0.702	0.253	0.924	0.127	5	307
A/B/C/ D & A*	0.833	0.185	0.952	0.100	4	401
A/B/C/H/Ch/ D & A*	0.756	0.224	0.924	0.125	4	353

SE non-cross-validated standard error

LV latent variable

* Fragment distinction: *A* atom type, *B* bond type, *C* connectivity, *H* hydrogens, *D & A* donor and acceptor, *Ch* chirality

Table 6 HQSAR analysis of the influence of fragment size when using the best fragment distinction (A/C/D & A)

Fragment size	Q^2	SE_{cv}	R^2	SE	LV	Hologram length
2–5	0.802	0.201	0.929	0.120	4	353
3–6	0.793	0.206	0.930	0.120	4	401
4–7	0.842	0.180	0.948	0.103	4	53
5–10	0.981	0.064	0.996	0.028	5	97
6–10	0.979	0.068	0.996	0.030	5	97
7–10	0.978	0.069	0.994	0.037	5	83

Compound nos. **29**, **30** and **36**, which bear all of the required substitutions, exhibit maximum antioxidant activity. Compound nos. **12**, **13**, **14**, **15**, and **16** (which bear all of the essential fragments except for the hydroxyl group at R10) as well as **27** and **28** (which lack the benzoyl fragment at R3) reveal moderate antioxidant activity profiles.

Models developed using the G-QSAR

In the G-QSAR method, every molecule of the data set is considered to bear a set of fragments based on the pattern of substitution. The descriptors are calculated for each fragment, and a relationship between these fragment descriptors and the activity of the whole molecule is developed. Thus, with the GQSAR method, it is possible to get important site-specific clues about where a particular descriptor (and hence a structural fragment) needs to be modified within a molecule.

The results obtained using the G-QSAR technique are summarized in Table 2. The two different models based on two different sets of descriptor matrices are reported in Table 8. The two models were developed based on: (a) the fragment descriptors only (G-QSAR), and (b) by considering the interaction terms for each pair of different fragments (G-QSAR_IT). The interaction terms take into consideration the impact of two consecutive fragment patterns on the overall activity profiles of the molecules. The scatter plots obtained for both the G-QSAR (Fig. 6d) and the G-QSAR_IT (Fig. 6e) models indicate that the activity data predicted based on the developed models closely match with the corresponding observed activity data.

The G-QSAR model (Eq. 1) was initially developed using only the fragment descriptors in order to determine their contribution to the overall antioxidant activity profiles of the molecules.

$$\begin{aligned}pC = {} & -0.4051 + 0.0070(\pm 0.0000) \times R10\mathrm{MomInertiaX} + 0.1685(\pm 0.0171) \times R3X \log P \\ & + 0.7663(\pm 0.1250) \times \mathrm{R9MomInertiaX} + 0.0032(\pm 0.0000) \times \mathrm{R2Mol.Wt.} \\ & n_{\mathrm{training}} = 27, F(df) = 81.774(4, 22), R^2 = 0.937, Q^2 = 0.851, n_{\mathrm{test}} = 9, \\ & R^2_{\mathrm{pred}} = 0.923, r^2_{\mathrm{m\ (test)}} = 0.910\end{aligned} \tag{1}$$

Based on the contributions of the different descriptors appearing in Eq. 1, they can be ranked as follows: (i) R10-MomInertiaX (48.18%), (ii) R3-XlogP (18.94%), (iii) R9-MomInertiaX (17.69%), and (iv) R2-Mol.Wt. (15.20%). The R10-MomInertiaX and the R9-MomInertiaX descriptors refer to the moment of inertia along the *x*-axis for the substituents at the R9 and R10 positions of the parent nucleus. The positive coefficients of these descriptors signify that they have a direct influence on the antioxidant activity profiles of the chromone derivatives. This observation aptly matches with the above models (3D pharmacophore and CoMSIA models), suggesting that proper orientation of the necessary substituent (−OH) at the R10 position enables the fragment to grab the position that favors the presence of a hydrogen-bond acceptor feature. Thus, the presence of a hydroxyl group at the R10 position is essential for increased antioxidant activity, as seen for compound nos. **27**, **29**, **30**, and **34** (all of these have maximum values for the R10-MomInertiaX descriptor). On the contrary, the required orientation of the hydroxyl substituent at the R9 position ensures that the fragment does not reach the area unfavorable to the activity profiles of the molecules. The XlogP descriptor is an atom-based evaluation of the partition coefficient (log*P*) [60], and it signifies the ratio of solute

Table 7 Selection of the best model with the least number of LVs using the 5% rule

LVs	Q^2	SE_{cv}	R^2	SE	Hologram length	Percentage increase in Q^2
5	0.981	0.064	0.996	0.028	97	2.08
4	0.961	0.089	0.989	0.047	97	3.11
3	0.932	0.116	0.970	0.076	97	-

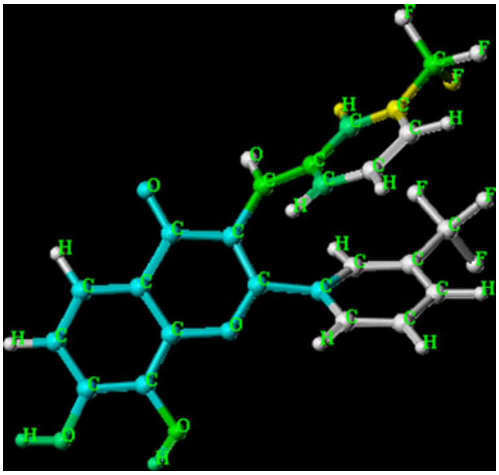


Fig. 8 Contribution map obtained using the HQSAR technique and based on compound no. **29** (see text for details)

concentrations in octanol and water. The R3-XlogP descriptor refers to the partition coefficient of the fragment at the R3 position. Since the value of the partition coefficient depends on the lipophilic character of the molecule, a positive coefficient for this descriptor refers to the fact that an increase in the hydrophobicity of the fragment at the R3 position adds to the antioxidant activity profiles of the molecules. This observation is strongly correlated with the 3D pharmacophore and the CoMSIA models, which also indicates the importance of having a hydrophobic feature at the R3 position of the molecules (Figs. 6 and 8), as observed for compound nos. **12**, **13**, **14**, **15**, **29**, **30**, and **35** (all these have a substituted benzoyl fragment at the R3 position). The R2-Mol.Wt descriptor refers to the molecular weight of the substituent at the R2 position of the chromone moiety, and a positive coefficient for this descriptor signifies that increasing the molecular weight of the substituent is conducive to the antioxidant activity profiles of the molecules. Compound no. **31**, which bears a low molecular weight group (methyl) at the R2 position, exhibits the lowest activity profile. On the contrary, although compound no. **34** lacks the necessary substitutions at the R2 and R3 positions, it shows a moderate activity profile due to the presence of the hydroxyl groups at the R9 and R10 positions. This implies that the moment of inertia descriptors are ranked higher than the R2-Mol.Wt descriptor, so the former exert a greater impact on the activity profiles of the molecules.

Further, the G-QSAR model was developed based on both the fragment descriptors and the interaction terms for two subsequent fragments, and this model was referred to as the "interaction-based G-QSAR model" (G-QSAR_IT). Thus, Eq. 2 was obtained, which takes into consideration both the group-based descriptors and their interaction terms.

$$\begin{aligned}
pC = {} & 0.350 + 0.0004(\pm 0.0000) \times \mathit{Mult}(\mathrm{R10}-\mathrm{Mol.Wt.}, \mathrm{R3}-\mathrm{Volume}) \\
& +0.473(\pm 0.040) \times \mathit{Mult}(\mathrm{R9}-\mathrm{MomInertiaX},\ \mathrm{R10}-\mathrm{polarizabilityAHP}) \\
& +0.029(\pm 0.002) \times \mathit{Mult}(\mathrm{R10}-\mathrm{polarizabilityAHP},\ \mathrm{R2}-\mathrm{0PathCount}) \\
& n_{\mathrm{training}} = 27, F(df) = 387.591(3,23), R^2 = 0.980, Q^2 = 0.963, n_{\mathrm{test}} = 9, \\
& R^2_{\mathrm{pred}} = 0.980, r^2_{\mathrm{m\ (test)}} = 0.925
\end{aligned} \tag{2}$$

Table 8 Summary of the models developed using the group-based QSAR technique

	G-QSAR	G-QSAR_IT
Descriptors	4	5
n (train/test)	27/9	27/9
F (df)	81.774 (5,22)	946.849 (5,22)
R^2	0.937	0.996
Q^2	0.851	0.990
R^2_{pred}	0.923	0.990
R^2_se	0.113	0.031
Q^2_se	0.174	0.047
R^2_{pred}_se	0.138	0.070
$r^2_{\mathrm{m\ (test)}}$	0.910	0.925
Descriptor_1	R10-MomInertiaX	R10-Mol.Wt.′ R3-Volume
Descriptor_2	R3-XlogP	R9-MomInertiaX′R10-polarizabilityAHP
Descriptor_3	R9-MomInertiaX	R10-polarizabilityAHP′ R2-0PathCount
Descriptor_4	R2-Mol.Wt.	-

All the cross-interaction terms in the above model refer to the products of the respective fragment descriptors, and signify the importance of each. The importance of the different interaction terms appearing in Eq. 2 is evaluated based on their contributions to the overall activity profiles of the molecules: (i) Mult(R10-Mol.Wt., R3-Volume) (54.09%), (ii) Mult(R9-MomInertiaX, R10-polarizabilityAHP) (30.51%), and (iii) Mult(R10-polarizabilityAHP, R2-0PathCount) (15.40%). The positive contributions of all of the descriptors indicate that increasing the values of these descriptors is conducive to the antioxidant activity profiles of the chromone derivatives. The interaction descriptor Mult(R10-Mol.Wt., R3-Volume) refers to the product of the R10-Mol.Wt. and R3-Volume descriptors, and indicates that increasing the molecular weight and volume for the substituents at the R10 and R3 positions, respectively, has a direct influence on the activity profiles of the molecules. Thus, as indicated by the previous models, molecules bearing hydroxyl substituents at the R10 position of the chromone nucleus exhibit enhanced antioxidant activity profiles compared to those lacking the necessary hydroxyl group. Similarly, as the R3 position is hydrophobically favored (as inferred from the CoMSIA and the 3D pharmacophore models), substituents with a larger volume (benzoyl group) are favored at this position. Compound nos. **27**, **29**, **30**, and **34** with hydroxyl substituents show higher values of the R10-Mol.Wt. descriptor and hence exhibit improved activity profiles. Compound nos. **12**, **13**, **14**, **15**, **29**, **30**, and **35**, which have large fragments at the R3 position, exhibit improved activities. The moderately contributing interaction variable Mult(R9-MomInertiaX, R10-polarizabilityAHP) signifies the positive influences of the R9-MomInertiaX and the R10-polarizabilityAHP descriptors on the activity profiles of the chromone derivatives. As mentioned earlier, in the G-QSAR-based model, the R9-MomInertiaX descriptor (referring to the moment of inertia of the R9 fragment along the *x*-axis) signifies that an increase in the value of this descriptor (as seen in the case of hydroxyl substitution) adds to the activity profiles of the molecules. This observation matches with the results obtained for the CoMSIA model, which shows that proper orientation of the R9 hydroxyl fragment enables it to escape from the unfavorable zone (Fig. 7a) and to exert a positive effect on the activity profile. Moreover, the model developed using the HQSAR technique shows green coloration of the R9 fragment (−OH) (Fig. 8), denoting that it provides the maximum contribution to the antioxidant activity profiles of the molecules. Again, the R10-polarizabilityAHP descriptor refers to the polarizability of the substituent at the R10 position using atom hybrid polarizability, and this indicates that easily polarizable fragments (−OH) are essential for substitution at the R10 position. This observation is in accordance with those obtained from the CoMSIA and the 3D pharmacophore models (which show hydrogen-bond acceptor features near the R10 fragment), as well as those obtained for the HQSAR model (the positive contribution of the hydroxyl substituent at the R10 position is indicated by its green color). Again, the hydroxyl group at the R10 position for compound nos. **27**, **29**, **30**, and **34** accounts for the polarizability of the R10 fragment and in turn adds to the activity profiles of the molecules. Finally, the interaction term Mult(R10-polarizabilityAHP, R2-0PathCount) once again imparts the significance of the R10-polarizabilityAHP descriptor in addition to the positive contribution of the R2-0PathCount descriptor. Zero path count simply refers to the number of skeletal atoms or vertices in the molecular graph, and the positive coefficient of the interaction term bearing the R2-0PathCount descriptor signifies that increasing the number of vertices for the substituent at the R2 position leads to an increase in the antioxidant activity profiles of the molecules. Thus, substituted aromatic fragments that fulfill the requirements of all of the developed QSAR models are the most suitable substituents for the R2 position. All of the compounds (compound nos. **27**, **29**, **30**, and **34**) in the higher activity range have aromatic substituents at the R2 position of the chromone nucleus.

Additional validation for the G-QSAR model

Amongst the different models developed in the present work, the G-QSAR models provide essential structural information regarding the substituent requirements in a more precise and quantitative manner. At the same time, the G-QSAR models show high predictive potential. Thus, because the G-QSAR model is the most acceptable one in terms of interpretability and statistical significance, it was analyzed further to ensure the statistical reliability of the developed model. The normality of the distribution of the residuals obtained from the training set data was checked using different statistical tests for normality. The Shapiro–Wilk test [54] for normality yielded a value of W=0.971 and p=0.639, while the Kolmororov–Smirnov test [55] yielded a value of d=0.113 at p>0.20. Besides these, the Lilliefors significance correction [56] performed for the residual data resulted in p>0.20. All of the above tests are performed in order to determine whether a population of data is normally distributed, and if the p value is greater than 0.05, the null hypothesis (that the population is normally distributed) is accepted. Thus, for the G-QSAR model, the computed values for the normality tests are much higher than the threshold value, so we can conclude that the residual values for the training set data are normally distributed. Additionally, the QUIK rule was performed in order to determine the predictor collinearity. The QUIK rule [61] allows models with high predictor collinearity, which often leads to chance correlation, to be rejected. The QUIK rule is based on the K multivariate correlation index [62], which measures the total correlation of a set of variables. According to this rule, for an acceptable model, the total correlation in the set given by the model predictors X plus the response Y (K_{XY}) should always be greater than that measured only with the set of predictors. For the present work, the value of K_{XY} (0.432) obtained for the G-QSAR model was higher than that of the K_X (0.285) parameter, and indicated that the developed G-QSAR model is statistically acceptable [K_{XY} (0.432)>K_X (0.285)]. Besides these, the model also satisfied all of the statistical validation parameters set forth by Golbraikh and Tropsha [34]. For the G-QSAR model, these statistical parameters yielded the following results (the threshold values are given inside parentheses):

$$\text{(i) } Q^2 = 0.851\ \left[Q^2 > 0.5\right];\ r^2 = 0.925\ \left[r^2 > 0.6\right]$$

$$\text{(ii) } \left(r^2 - r_0^2\right)/r^2 = 0.0002\ \left[\left(r^2 - r_0^2\right)/r^2 < 0.1\right];\ \left(r^2 - r_0^{/2}\right)/r^2 = 0.009\ \left[\left(r^2 - r_0^{/2}\right)/r^2 < 0.1\right]$$

$$\text{(iii) } k = 1.017\ [0.85 \le k \le 1.15];\ k^{/} = 0.974\ [0.85 \le k \le 1.15]$$

Model randomization was also performed for the G-QSAR model at the 99% confidence level. The lack of chance correlation in the G-QSAR model is well reflected in the value of ${}^{c}R_p^2$ (0.871) [33], which is much higher than the threshold value of 0.5.

Further, the applicability domain for the G-QSAR model was checked using the leverage approach [57, 58]. Figure 9 shows a plot of standardized residuals (y-axis) vs. leverage values (x-axis), which is referred to as the Williams plot. The Williams plot thus obtained for the G-QSAR model helps us to determine the domain of applicability of the developed model for a diverse set of untested molecules. The plot enables us to determine the poorly predicted molecules, as well those that have atypical characteristics. All the training set compounds (n_{training}=27) have standardized residual values within the limit of ±3σ, indicating that none of the compounds are prediction outliers. However, compound no. **1**, with a leverage value greater than the critical value (h>h*), although not a response outlier, behaves as an influential observation. The critical leverage value of the model is 0.556, and all of the test set (n_{test}=9) compounds were found to be within the applicability domain of the model (i.e., there were no structurally different chemicals).

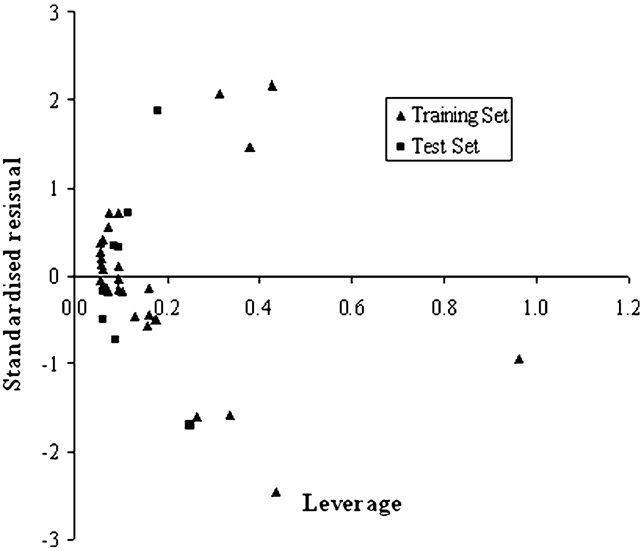


Fig. 9 Williams plot for the G-QSAR model

Further studies of the GQSAR model

In order to check the intercorrelation among the four different variables appearing in the G-QSAR model, the Pearson correlation matrix (Table 9) was developed. A maximum correlation (R) of −0.419 between the descriptors R10-MomInertiaX and R9-MomInertiaX indicated that the descriptors did not show significant intercorrelation. Moreover, the G-QSAR model was developed using four descriptors for 27 molecules, which satisfied the 1:5 rule (one descriptor for every five compounds) for QSAR analysis. However, to further reduce the descriptor set, MLR analysis was performed, and the least significant (R2-Mol.Wt) of the four descriptors in Eq. 1 was eliminated. Although the model obtained was acceptable in terms of both internal (Q^2=0.713) and external ($R^2_{\text{pred}} = 0.893$ and $r^2_{\text{m (test)}} = 0.829$) predictive parameters, the quality of the model deteriorated in terms of its prediction ability. Thus, partial least squares (PLS) analysis was performed to retain all four descriptors in Eq. 1 by generating fewer latent variables or components. Since the latent variables are functions of the input descriptors, they encode all of the information available within the descriptors, thus resulting in fewer variables. A three-component model was obtained based on the PLS technique, with significantly acceptable

Table 9 Pearson correlation matrix for the descriptors occurring in Eq. 5

Descriptors	R10-MomInertiaX	R3-XlogP	R9-MomInertiaX	R2-Mol.Wt.
R10-MomInertiaX	1.000	0.210	−0.419	−0.371
R3-XlogP	0.210	1.00	0.151	0.180
R9-MomInertiaX	−0.419	0.151	1.000	0.244
R2-Mol.Wt.	−0.371	0.180	0.244	1.000

values attained for all the internal and external predictive parameters.

$$pC = -0.415 + 0.007 \times \mathrm{R10 - MomInertiaX} + 0.169 \times \mathrm{R3 - XlogP} + 0.785 \times \mathrm{R9 - MomInertiaX} + 0.003 \times \mathrm{R2Mol.Wt.} \quad (3)$$

$$n_{\mathrm{training}} = 27, LVs = 3, F(df) = 113.74(3,23),$$

$$R^2 = 0.937, Q^2 = 0.837, n_{\mathrm{test}} = 9,$$

$$R^2{}_{\mathrm{pred}} = 0.923, r^2_{\mathrm{m\ (test)}} = 0.909.$$

The PLS model thus obtained was further validated based on the randomization technique using the Simca-P software package [63]. Each permutation of the data generated a new set of R^2 and Q^2 values, which were plotted against the correlation coefficient between the original Y values and the permuted Y values. A model is considered valid if the intercepts of R^2 and Q^2 are less than 0.4 and 0.05, respectively. The intercepts for the R^2 and Q^2 lines in the plot (Fig. S2 in the "Electronic supplementary material") are used to determine if the data are overfitted. The intercepts of R^2 and Q^2 for the plot obtained based on Eq. 3 are 0.040 for and −0.340, respectively. The values, which are much lower than the threshold limit, indicate that the model is robust.

Additionally, due to the uneven distribution of the activity data, there is a large gap in the activity range between 1.7 and 2.5, and most of the compounds lie in the range between 0.7 and 1.6. Thus, additional analysis was performed with the fragment-based descriptors after removing compounds with pIC_{50} values that are greater than 2.0. The PLS model thus developed with the reduced set of compounds also yielded a robust QSAR model with acceptable values for all of the statistical parameters (Q^2=0.881, $R^2_{\mathrm{pred}} = 0.956$ and $r^2_{\mathrm{m\ (test)}} = 0.897$).

Consensus model

Using input data from different sources and a variety of algorithms in the development of QSAR models increases the risk of model uncertainty. For QSAR models, such uncertainty increases for data relating to different chemical classes and end-points. This is where consensus modeling is found to be useful, as it helps to reduce the model uncertainty by averaging the model outputs. Consensus predictions are made based on the results generated by the multiple QSAR models. Each of the individual models may contain some noisy data. Consensus QSAR modeling diminishes the effects of these noisy data and hence provides more reliable predictions than the individual QSAR models. Since the consensus prediction is made based on the average of the results obtained using the different but comparable QSAR models, it is capable of capturing the relationship between the chemical structures of the molecules and the end-point more efficiently than a single model [64, 65]. In the present work, the consensus model was developed by averaging the predicted activity data for the test set molecules obtained using the five different QSAR models, with equal weights assigned to each of the models. Further, the external predictive potential of the consensus model was checked based on the R^2_{pred} and $r^2_{\mathrm{m\ (test)}}$ parameters. Moreover, the values of $R^2_{\mathrm{pred}}(0.969)$ and $r^2_{\mathrm{m\ (test)}}(0.957)$ obtained for the consensus model were higher than those obtained for all of the individual QSAR models. Thus, we can infer that the consensus model captures the features of all of the developed QSAR models, and so can predict the test set molecules more efficiently than the individual models.

Comparison with previous work

Samee et al. [25] reported a 3D-QSAR model based on the same dataset of chromone derivatives, which was obtained using the molecular field analysis (MFA) technique along with the genetic partial least squares method (G/PLS) as the chemometric tool. After deleting one compound as an outlier, they utilized a test set of five compounds to determine the external predictive ability of the developed model. However, in the present work, none of the molecules were removed as outliers, and a test set of nine compounds was used to assess the predictive potential of the developed model. A detailed comparison of the model developed by Samee et al. [25] with those developed in this study is shown in Table 10.

Table 10 Comparison of the present work with that reported by Samee et al. [25]

Source	Model development technique	n_{training}	R^2	Q^2	n_{test}	R^2_{pred}	$r^2_{\text{m (test)}}$
Samee et al. [25]	MFA QSAR	30	0.868	0.771	5	0.924	-
Present work	3D pharmacophore	27	0.832	-	9	0.883	0.826
	3D-QSAR (CoMSIA)	27	0.957	0.834	9	0.852	0.845
	HQSAR	27	0.970	0.932	9	0.961	0.957
	G-QSAR	27	0.937	0.851	9	0.923	0.910
	G-QSAR_IT	27	0.980	0.963	9	0.980	0.925
	Consensus model	-	-	-	9	0.969	0.957

The HQSAR and G-QSAR methodologies utilized in the present work are comparatively new approaches that estimate the structural requirements of the molecules in a more precise manner than conventional methods. Moreover, the 3D-QSAR models developed using the traditional (more commonly used) techniques determine the pharmacophoric features that are required for the antioxidant activity profiles of the chromone derivatives. The values obtained for all of the statistical parameters of our best models (G-QSAR models) are better than those reported by Samee et al. [25] (Table 10), indicating that the new models provide more reliable structural information on the features that are essential for improved antioxidant activity in this series of chromone derivatives. More specifically, the Q^2 value of the MFA model reported by Samee et al. [25] is 0.771, while those for the G-QSAR and G-QSAR_IT models reported by us are 0.851 and 0.963, respectively. Similarly, the R^2_{pred} value for the MFA model reported by Samee et al. [25] is 0.924 (for five test set compounds), while that for the G-QSAR_IT model reported by us (for nine test set compounds) is 0.980. Moreover, Samee et al. [25] deleted one of the active compounds, reporting it as an outlier. On the contrary, in the present work, all of the molecules were included in the QSAR analysis, since removing a molecule may result in the loss of essential chemical information, especially in the case of a small dataset. In the present work, we have used multiple strategies of model development. All of the developed models pointed to a similar type of structure-property relationship, suggesting the validity of the hypotheses. Samee et al. [25] reported the importance of steric and electrostatic features to the overall antioxidant activity profiles of the molecules, based on interactions with different probe atoms at specific points in the 3D MFA grid. The present work determines the importance of similar features based on the mapping of molecules to 3D contour maps. In addition to the steric and electrostatic interactions, the CoMSIA contour map analyses the importance of hydrogen-bond donor, hydrogen-bond acceptor and hydrophobic features in the antioxidant activity profiles of the chromone derivatives. Again, the HQSAR technique employed for the present work is a unique branch of the

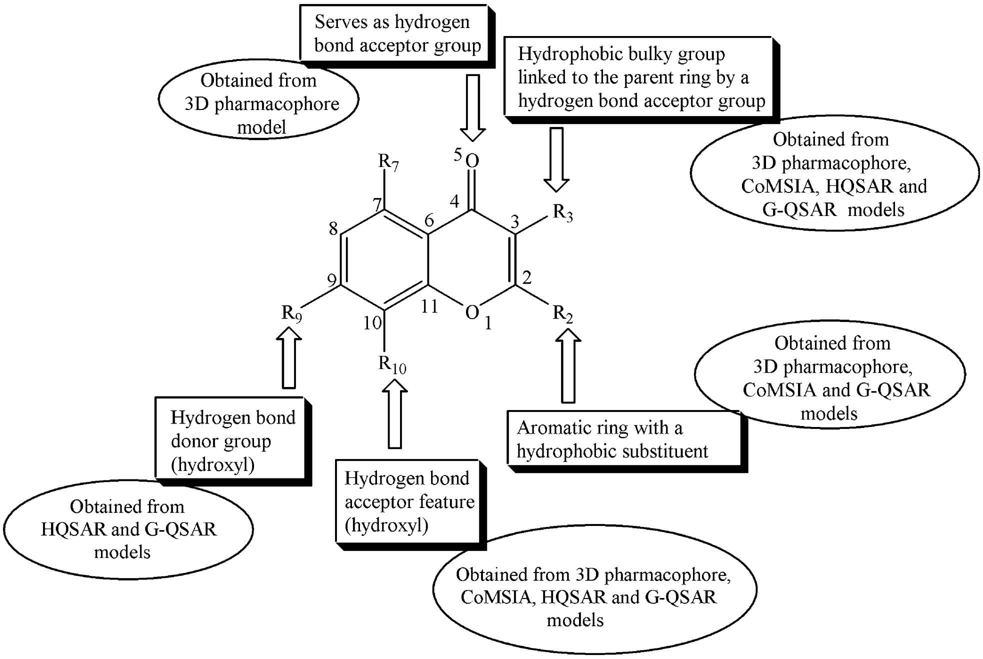

Serves as hydrogen bond acceptor group
Obtained from 3D pharmacophore model
Hydrophobic bulky group linked to the parent ring by a hydrogen bond acceptor group
Obtained from 3D pharmacophore, CoMSIA, HQSAR and G-QSAR models
R7
5 O
4
3
R3
7
8
6
9
2
R9
10
11
O 1
R2
R10
Obtained from 3D pharmacophore, CoMSIA and G-QSAR models
Hydrogen bond donor group (hydroxyl)
Aromatic ring with a hydrophobic substituent
Hydrogen bond acceptor feature (hydroxyl)
Obtained from HQSAR and G-QSAR models
Obtained from 3D pharmacophore, CoMSIA, HQSAR and G-QSAR models

Fig. 10 Schematic diagram showing different features at various positions favoring the antioxidant activity profiles of the chromone derivatives, as well as the different QSAR techniques adopted to reach these conclusions

traditional QSAR methodology, and deals with the contributions of different molecular fragments to the antioxidant activity profiles of the molecules. Additionally, the G-QSAR technique employed in the present work quantitatively determines the contributions of the different substituents to the overall antioxidant activity profiles of the chromone derivatives. Moreover, in the present work, besides the calculation of the R^2_{pred} parameter, the $r^2_{\mathrm{m\,(test)}}$ parameter was also determined, which means that the developed QSAR models are more reliable.

Overview and conclusions

In the present work, 36 chromone derivatives were modeled for their antioxidant activity profiles. The similarity of the features that occurred in all four types of model developed further supports the reliability and reproducibility of the developed models. Figure 10 shows a schematic representation of the various structural features that are crucial to the improved antioxidant activity profiles of the chromone derivatives, together with a note about the QSAR methodologies with which the corresponding results are obtained. The fact that the various QSAR methodologies point to similar structural requisites proves their unified mechanistic approach for modeling the antioxidant activity profiles of the chromone derivatives. The importance of the hydrogen-bond acceptor feature is revealed by all four models. Thus, the hydroxyl substituent at the R10 position and the benzoyl substituent at the R3 position of the chromone nucleus are essential fragments for improved antioxidant activity. Additionally, the ketonic group at C4 further enhances the abilities of the molecules to interact with the toxic free radicals through a mechanism of electron transfer followed by deprotonation [10]. The presence of the blue contour near the R2 position of the chromone moiety in the CoMSIA analysis indicates that bulky substituents are disfavored at this position. This observation matches with that reported by Samee et al. [25], since a bulky substituent may interfere with the radical delocalization of the chromone nucleus, affecting the electron density over the =O fragment. This in turn impairs the hydrogen abstraction mechanism [10] utilized by the hydrogen-bond acceptor fragments present in the antioxidant molecules. For the 3D pharmacophore model, the presence of the ring aromatic and the hydrophobic features over the substituents at the R2 and R3 positions, respectively, indicates that such groups separated by the specific distance of 5.890Å are essential for the enhanced activities of the molecules. Similar results were also obtained from the CoMSIA study, where these substituents map to the hydrophobically favored yellow contours. Moreover, the HQSAR contour study also revealed the importance of such fragments, with the green color for the substituent at R3 indicating its maximum contribution. The models developed here provide detailed information on the structural attributes required for optimum antioxidant activity of the chromone derivatives. Finally, the G-QSAR models developed for the present dataset provide a precise outline of the essential structural fragments based on the fragment-specific descriptors and the interaction terms. The inferences obtained from the G-QSAR models closely match with those of the remaining models, indicating the important impact of the hydroxyl substitution at the R10 position on the antioxidant activity profiles of the chromone derivatives, in addition to the remaining essential features, such as the presence of the substituted benzoyl fragment at the R3 position and the substituted aromatic fragment at the R2 position. Although all of the models developed here yield statistically significant results, the HQSAR model can be ranked as the best, based on the values of the external validation parameters $\left(R^2_{\mathrm{pred}} = 0.961, r^2_{\mathrm{m\,(test)}} = 0.957\right)$. All of the models developed here are statistically significant, and the observations made in each case are identical to those made using the other models. Thus, these models are reproducible in terms of both the results obtained for the essential structural attributes of the molecules as well as the statistical parameters. The results thus reflect the mechanistic interpretation of the free-radical scavenging activities of the chromone derivatives. Although the models were developed using different chemometric tools, the similar conclusions about the structure–activity relationship obtained from the models infer that these models are robust and highly predictive. Moreover, the consensus model developed here based on the five different QSAR models further highlights the potential predictive abilities of the models. All of the 2D- and 3D-QSAR approaches employed in the present work can be used as efficient query tools for designing as well as searching databases for chromone molecules with potent antioxidant activities. Thus, the models can be utilized to estimate the activity profiles of virtual libraries of newly designed antioxidant chromone molecules of this class prior to synthesis or biological testing.

Acknowledgments This research work is supported in the form of a major research project to K.R. and a senior research fellowship to I.M. by the Indian Council of Medical Research (ICMR), New Delhi. We thank the anonymous reviewers for useful comments. The authors thank VLife Sciences Technologies Pvt. Ltd., Pune for providing complimentary evaluation license of the software VLife MDS 3.5.

References

1. Proctor PH (1989) Free radicals and human disease. In: Miquel J (ed) CRC handbook of free radicals and antioxidants in biomedicine. CRC Press, Boca Raton, pp 209–221
2. Prasad K, Kalra J (1993) Oxygen free radicals and hypercholesterolemic atherosclerosis: effect of vitamin E. Am Heart J 125:958–973

3. Balazas L, Leon M (1994) Evidence of an oxidative challange in the Alzheimer's brain. Neurochem Res 19:1131–1137
4. Cooke MS, Evans MD, Dizdaroglu M, Lunec J (2003) Oxidative DNA damage: mechanisms, mutation, and disease. FASEB J 17:1195–1214
5. Langseth L (1996) Oxidants, antioxidants and disease prevention. International Life Science Institute, Brussels
6. Cadenas E, Davies KJ (2000) Mitochondrial free radical generation, oxidative stress, and aging. Free Radical Biol Med 29:222–230
7. McCord JM (1998) Iron, free radicals, and oxidative injury. Semin Hematol 35:5–12
8. Gordon MH (1990) The mechanism of antioxidant action in vitro. In: Hudson BJF (ed) Food antioxidants. Elsevier, New York, pp 1–18
9. Singh BK, Sharma SR, Singh B (2010) Antioxidant enzymes in cabbage: variability and inheritance of superoxide dismutase, peroxidase and catalase. Sci Hort 124:9–13
10. Wright JS, Johnson ER, DiLabio GA (2001) Predicting the activity of phenolic antioxidants: theoretical method, analysis of substituent effects, and application to major families of antioxidants. J Am Chem Soc 123:1173–1183
11. Vafiadis AP, Bakalbassis EG (2005) A DFT study on the deprotonation antioxidant mechanistic step of ortho-substituted phenolic cation radicals. Chem Phys 316:195–204
12. Musialik M, Litwinienko G (2005) Scavenging of DPPH• radicals by vitamin E is accelerated by its partial ionization: the role of sequential proton loss electron transfer. Org Lett 7:4951–4954
13. Genestra M (2007) Oxyl radicals, redox-sensitive signaling cascades and antioxidants. Cell Signal 19:1807–1819
14. Dizdaroglu M, Jaruga P, Birincioglu M, Rodriguez H (2002) Free radical-induced damage to DNA: mechanisms and measurement. Free Radic Biol Med 32:1102–1115
15. Helguera AM, Combes RD, Gonzalez MP, Cordeiro MN (2008) Applications of 2D descriptors in drug design: a DRAGON tale. Curr Top Med Chem 8:1628–1655
16. Gonzalez MP, Teran C, Saiz-Urra L, Teijeira M (2008) Variable selection methods in QSAR: an overview. Curr Top Med Chem 8:1606–1627
17. Hansch C, Maloney PP, Fujita T, Muir RM (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature 194:178–180
18. Cheng Z, Ren J, Li Y, Chang W, Chen Z (2002) Study on the multiple mechanisms underlying the reaction between hydroxyl radical and phenolic compounds by qualitative structure and activity relationship. Bioorg Med Chem 10:4067–4073
19. Singh N, Loader RJ, O'Malley PJ, Popelier PLA (2006) Computation of relative bond dissociation enthalpies (DBDE) of phenolic antioxidants from quantum topological molecular similarity (QTMS). J Phys Chem A 110:6498–6503
20. Reis M, Lobato B, Lameira J, Santos AS, Alves CN (2007) A theoretical study of phenolic compounds with antioxidant properties. Eur J Med Chem 42:440–446
21. Mitra I, Saha A, Roy K (2009) Quantitative structure–activity relationship modeling of antioxidant activities of hydroxybenzalacetones using quantum chemical, physicochemical and spatial descriptors. Chem Biol Drug Des 73:526–536
22. Mitra I, Roy K, Saha A (2009) QSAR of antilipid peroxidative activity of substituted benzodioxoles using chemometric tools. J Comput Chem 30:2712–2722
23. Mitra I, Saha A, Roy K (2010) Pharmacophore mapping of arylamino-substituted benzo[b]thiophenes as free radical scavengers. J Mol Model 16:1585–1596
24. Roy K, Mitra I (2009) Advances in quantitative structure–activity relationship models of antioxidants. Expert Opin Drug Discov 4:1157–1175
25. Samee W, Nunthanavanit P, Ungwitayatorn J (2008) 3D-QSAR investigation of synthetic antioxidant chromone derivatives by molecular field analysis. Int J Mol Sci 9:235–246
26. Samee W, Sae-Lee N, Ungwitayatorn J (2004) Structure-radical scavenging activity relationships of the synthesized chromone derivatives. J Pharm Sci 9:36–42
27. Leonard JT, Roy K (2006) On selection of training and test sets for the development of predictive QSAR models. QSAR Comb Sci 25:235–251
28. Roy PP, Leonard JT, Roy K (2008) Exploring the impact of the size of training sets for the development of predictive QSAR models. Chemom Intell Lab Sys 90:31–42
29. SPSS Inc. (2011) SPSS. SPSS Inc., Chicago. http://www.spss.com
30. Smellie A, Teig SL, Towbin P (1995) Poling: promoting conformational variation. J Comput Chem 16:171–187
31. Accelrys Inc (2010) Cerius 2, v.4.10. Accelrys Inc., San Diego
32. Sutter J, Guner OF, Hoffman R, Li H, Waldman M (2000) HypoGen: an automated system for generating 3D predictive pharmacophore models. In: Guner OF (ed) Pharmacophore perception, development, and use in drug design. International University Line, La Jolla, pp 501–511
33. Mitra I, Saha A, Roy K (2010) Exploring quantitative structure–activity relationship (QSAR) studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. Mol Simul 36:1067–1079
34. Golbraikh A, Tropsha A (2002) Beware of q^2! J Mol Graph Mod 20:269–276
35. Roy PP, Roy K (2008) On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci 27:302–313
36. Roy PP, Paul S, Mitra I, Roy K (2009) On two novel parameters for validation of predictive QSAR models. Molecules 14:1660–1701
37. Mitra I, Roy PP, Kar S, Ojha PK, Roy K (2010) On further application of r_m^2 as a metric for validation of QSAR models. J Chemometrics 24:22–33
38. Ojha PK, Mitra I, Das RN, Roy K (2011) Further exploring r_m^2 metrics for validation of QSPR models. Chemom Intell Lab Syst 107:194–205
39. Cramer RD III, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967
40. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
41. Streitwieser A (1961) Molecular orbital theory for organic chemists. Wiley, New York
42. White DNJ (1977) The principles and practice of molecular mechanics calculations. Comput Chem 1:225–233
43. Kirkpatrick S, Gelatt CD, Vecchi MP Jr (1983) Optimization by simulated annealing. Science 220:671–680
44. Tripos Inc. (2006) SYBYL 7.3. Tripos Inc., St. Louis. http://www.tripos.com
45. Wold S, Albano C, Dunn WJ III, Esbensen K, Hellberg S, Johansson E, Sjostrom M, Edlund U, Geladi P (1984) Multivariate data analysis in chemistry. In: Kowalski B (ed) Chemometrics: mathematics and statistics in chemistry. Reidel, Dordrecht
46. Hoskuldsson A (1987) PLS regression methods. J Chemometrics 2:211–228
47. Clark RD, Fox PC (2004) Statistical variation in progressive scrambling. J Comput Aided Mol Des 18:563–576
48. Doddareddy MR, Lee YJ, Cho YS, Choi KI, Koh HY, Pae AN (2004) Hologram quantitative structure activity relationship studies on 5-HT_6 antagonists. Bioorg Med Chem 12:3815–3824
49. Wold S, Johansson E, Cocchi M (1993) PLS: partial least squares projections to latent structures. In: Kubiniyi H (ed) 3D QSAR in drug design: theory, methods and applications. ESCOM, Leiden, pp 523–550

50. Ajmani S, Jadhav K, Kulkarni SA (2009) Group-based QSAR (G-QSAR): mitigating interpretation challenges in QSAR. QSAR Comb Sci 28:36–51
51. VLife Sciences Technologies Pvt. Ltd. (2007) VLife MDS 3.5. VLife Sciences Technologies Pvt. Ltd., Pune. http://www.vlifesciences.com
52. Darlington RB (1990) Regression and linear models. McGraw-Hill, New York
53. Snedecor GW, Cochran WG (1967) Statistical methods. Oxford & IBH, New Delhi
54. Stephens MA (1976) Asymptotic results for goodness-of-fit statistics with unknown parameters. Ann Stat 4:357–369
55. Massey FJ Jr (1951) The Kolmogorov–Smirnov test for goodness of fit. J Am Stat Assoc 46:68–78
56. Lilliefors HW (1967) On the Kolmogorov–Smirnov test for normality with mean and variance unknown. J Am Stat Assoc 64:399–402
57. Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26:694–701
58. Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 111:1361–1375
59. Patrick GL (2009) An introduction to medicinal chemistry. Oxford University Press, New York
60. Wang R, Gao Y, Lai L (2000) Calculating partition coefficient by atom-additive method. Perspect Drug Discov 19:47–66
61. Todeschini R, Consonni V, Maiocchi A (1999) The K correlation index: theory development and its applications in chemometrics. Chemom Intell Lab Syst 46:13–29
62. Todeschini R (1997) Data correlation, number of significant principal components and shape of molecules. The K correlation index. Anal Chim Acta 348:419–430
63. Umetrics AB (2002) SIMCA-P 10.0. Umetrics AB, Umea. http://www.umetrics.com
64. Golbraikh A, Shen M, Xiao ZY, Xiao YD, Lee KH, Tropsha A (2003) Rational selection of training and test sets for the development of validated QSAR models. Comput Aided Mol Des 17:241–253
65. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Oberg T, Dao P, Cherkasov A, Tetko IV (2008) Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. J Chem Inf Model 48:766–784

J Mol Model (2012) 18:1841–1854
DOI 10.1007/s00894-011-1205-2

ORIGINAL PAPER

Insights into the structural function of the complex of HIV-1 protease with TMC-126: molecular dynamics simulations and free-energy calculations

Dan Li · Ju-Guang Han · Hang Chen · Liang Li · Run-Ning Zhao · Guang Liu · Yuhua Duan

Received: 18 December 2010 / Accepted: 26 July 2011 / Published online: 18 August 2011

Abstract The binding properties of the protein–inhibitor complex of human immunodeficiency virus type 1 (HIV-1) protease with the inhibitor TMC-126 are investigated by combining computational alanine scanning (CAS) mutagenesis with binding free-energy decomposition (BFED). The calculated results demonstrate that the flap region (residues 38–58) and the active site region (residues 23–32) in HIV-1 protease contribute 63.72% of the protease to the binding of the inhibitor. In particular, the mechanisms for the interactions of key residues of these species are fully explored and analyzed. Interestingly, the regression analyses show that both CAS and BFED based on the generalized Born model yield similar results, with a correlation coefficient of 0.94. However, compared to CAS, BFED is faster and can decompose the per-residue binding free-energy contributions into backbone and side-chain contributions. The results obtained in this study are useful for studying the binding mechanism between receptor and ligand and for designing potent inhibitors that can combat diseases.

Keywords HIV-1 protease · TMC-126 · MM-PBSA/MM-GBSA · Free-energy decomposition · Computational alanine scanning

D. Li · J.-G. Han (✉) · H. Chen · L. Li · R.-N. Zhao · G. Liu
National Synchrotron Radiation Laboratory,
University of Science and Technology of China,
Hefei 230029, People's Republic of China
e-mail: jghan@ustc.edu.cn

Y. Duan
National Energy Technology Laboratory,
United States Department of Energy,
Pittsburgh, PA 15236, USA

Abbreviations

PR	Protease
PI	Protease inhibitors
MD	Molecular dynamics
PME	Particle mesh Ewald
MM	Molecular mechanics
GB	Generalized Born
PB	Poisson–Boltzmann
SA	Surface area
rmsd	Root-mean-square deviation

Introduction

Acquired immune deficiency syndrome (AIDS), which is induced by human immunodeficiency virus (HIV) infection, has become one of the major medical and humanitarian challenges. HIV-1 protease (PR), a member of the aspartyl protease family, is one of the most important enzymes targeted in research aimed at discovering new drugs to counter AIDS. It cleaves the nonfunctional polypeptide into viral structural (gag) and functional (pol) proteins, a process that is essential for the maturation of the infectious HIV particles [1, 2]. Repression of HIV-1 PR activity could prevent the production of mature and infectious HIV particles, blocking further HIV infection. HIV protease is a centrally symmetric homodimer containing two identical 99 amino acid monomers; the active residues Asp25 and Asp25′ are located at the interface between the two monomers [3–5]. The binding on an inhibitor to PR can lead to the inactivation of the enzyme and prevent the infection of the host cell. Thus, the dimeric HIV-1 protease is one of the most attractive targets in the development of antiviral therapeutics. Therefore, in order to design efficient inhibitors, it is critically important to investigate the mechanism of the interaction between PR and protease inhibitor (PI) in detail.

Currently, nine antiviral agents that can inhibit HIV-1 protease have been approved by the US Food and Drug Administration (FDA)—including saquinavir, ritonavir, lopinavir, atazanavir, indinavir, amprenavir, and tipranavir—with several others under clinical trial [6]. Due to the short-lived therapeutic benefits of these drugs and the rapid evolution of drug-resistant variants, there is an urgent need to develop antiretroviral drugs with minimal side effects and broad-spectrum activities for current and future wild-type and mutant strains of HIV protease [7, 8]. TMC-126 is an effective nonpeptide inhibitor of HIV-1 protease that is extremely potent against a wide spectrum of HIV protease variants (Fig. 1). Its structure is largely based on that of darunavir (TMC-114) [9], which contains a bistetrahydrofuranyl (bis-THF) urethane and an isostere of sulfonamide [10].

Since it is inconvenient to measure the binding affinities of different PR and inhibitors experimentally, molecular dynamics (MD) simulation can play an important role in investigations of the structural and functional characteristics of biological systems. Using MD simulations, kinetic and thermodynamic data on the simulated system can be obtained. The binding free energy—a very important thermodynamic quantity—can be used to evaluate the stability of a complex. Hence, accurately calculating the binding free energy is crucial when exploring the interactions between proteins and ligands [11]. In order to rapidly evaluate binding free energies, several semi-empirical methods such as the molecular mechanics Poisson–Boltzmann surface area (MM-PBSA) and the molecular mechanics generalized Born surface area (MM-GBSA) methods have been developed [12]. The MM-PBSA approach [13–16], which is based on the MD simulation of the protein–ligand complex of interest in explicit solvent, has been successfully used to describe the protein–ligand binding free energy in rational drug design [17–23]. In this method, the binding free energy is decomposed into the molecular mechanical free energy, the solvation free energy, and entropic contributions. The polar contribution to the solvation energy is evaluated using the Poisson–Boltzmann (PB) model, or calculated by the generalized Born (GB) model using the MM-GBSA method. The free-energy decomposition involved in the MM-GBSA method can elucidate the contribution of each protease residue to the overall protein–ligand binding free energy at the atomic level [24, 25]. Both MM-PBSA and MM-GBSA have been used to investigate the mechanisms of interaction between different protease inhibitors and different variants of the HIV-1 protease [17]. Previous studies of HIV-1 protease and inhibitors have mostly utilized the MM-PBSA method to obtain the binding free energy, whereas free-energy decomposition methods based on the GB model were selected to analyze the interaction mechanisms [17, 26]. The GB model is an attractive alternative to the PB model since it is significantly faster and can decompose the binding free energies on a per-residue basis. However, a comparison analysis of the PB and GB models was implemented on the same system that we study in this article. The results showed that the free-energy changes upon alanine mutation determined by the PB model are smaller and more accurate than those obtained by the GB model [27, 28]. Moreover, when the free-energy decomposition method was performed to investigate the per-residue contributions to the HIV-1 protease by the PB method, the dimerization between the two monomers was barely taken into account.

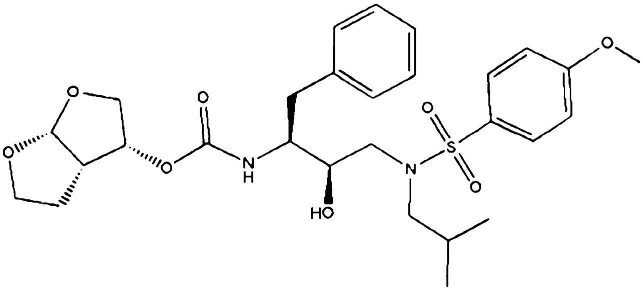


Fig. 1 Structure of TMC-126

An extension of the MM-PBSA/MM-GBSA approach based on both the GB and the PB models, computational alanine scanning (CAS) can estimate the free-energy consequences of PR mutations located at the active site, the flap region, and the binding interface from a single MD trajectory. Furthermore, the CAS method allows pairs of residues in the two monomers of PR to be mutated to alanine. Thus, it is a powerful tool for discovering hotspot residues [28–32].

In this work, the relative binding free energies in the complex of PR and TM-126 were obtained using 3 ns MD simulations. The per-residue interactions of the HIV-1 protease and TMC-126 were analyzed by the binding free energy decomposition methods mentioned above. Then the CAS method was implemented to discern 15 important residues of the HIV-1 protease. The results obtained by the CAS method based on the GB model and the PB model were compared and discussed. A full comparison of the decomposition method with the CAS method was also performed in this study.

Theoretical methods

Initial structure of the complex

The crystal structure of the HIV protease bound to the mutant resistant inhibitor UIC-98038 was obtained from the

Protein Data Bank (PDB; entry: 3I7E) [10]. The starting structures and force field parameters for the inhibitor were obtained as follows. Special attention was given to the protonation states of Asp25 and Asp25′ in the active site. In this work, monoprotonation was adopted, and a proton was added to the oxygen atom OD2 of Asp25 [33]. Considering the importance of water in the binding between PR and the inhibitor, water207 was included in the starting structure [34, 35]. Partial charges and force field parameters for the inhibitor were generated automatically using the Antechamber program in Amber10 [36, 37]. The atomic charges were derived with the AM1-BCC charge method [38]. The general force field (GAFF) [37] and the standard Amber force field (FF03) [39] were used to obtain parameters such as the Lennard–Jones, torsion, bond, and angle terms for small organic molecules and to describe the parameters of the protein, respectively. All missing hydrogens were added using tleap in Amber10 [36]. To neutralize the charge of the system, Cl^- counterions were placed in the grid regions with the largest positive Coulombic potentials around the protease, and then the whole system was soaked in an octahedral periodic box of TIP3P [40] waters. All solute atoms were 10 Å from the edge of the water box.

Molecular dynamics simulations

All molecular simulations presented in this work were carried out using the Amber10 simulation package and the force field parameters of Cornell et al. [41]. Periodic boundary conditions and a 10 Å cutoff for nonbonded van der Waals (VDW) interactions were applied in our simulations. The particle mesh Ewald (PME) [41] method was employed to account for the long-range electrostatic interactions under periodic boundary conditions. The SHAKE procedure [42] was applied to all atoms covalently bonded to a hydrogen atom. A time step of 2 fs was used to integrate the equations of motion.

In order to remove steric overlap, which produces bad effects between the complex and solvent, two stages of energy minimization were performed: 500 cycles of steepest descent and 2500 cycles of conjugate gradient minimization. First, the water molecules were minimized by keeping the solute fixed with a harmonic constraint of strength 100 kcal mol^{-1} Å^{-2}. Second, the entire system was minimized without restriction. Subsequently, before the actual MD simulations, the temperature of the system was gradually raised from 0 K to 300 K over 100 ps, followed by 100 ps of equilibration at 300 K. The initial velocities of atoms were assigned based on a Maxwellian distribution at the starting temperature. Finally, a 3 ns MD simulation was performed at a constant pressure of 1 atm and constant temperature, and controlled by Langevin dynamics with a collision frequency of 1.0 ps^{-1}. The resulting trajectories were analyzed using the ptraj module of Amber10. One snapshot was saved every 5 ps; 200 snapshots were collected from the previous 1000 ps of simulations for post-processing analysis.

MM-PBSA/MM-GBSA approach

The binding free energy between PR and PI was calculated by the MM-PBSA method according to the following equation:

$$\Delta G_{\text{bind}} = G_{\text{complex}} - \left(G_{\text{receptor}} + G_{\text{ligand}}\right), \tag{1}$$

where G_{complex},G_{receptor}, and G_{ligand} represent the free energies of the complex, receptor, and ligand averaged over snapshots taken from MD trajectories. The free energy of each reactant was estimated as the sum of the molecular mechanical free energy, the solvation free energy, and the contributions from the vibrational, rotational, and translational entropies:

$$G = E_{\text{MM}} + G_{\text{solvation}} - TS. \tag{2}$$

The molecular mechanical energy E_{MM} in Eq. 2 was determined with the Sander program from Amber10 software suite according to molecular mechanics with an empirical force field. The topology files thus obtained were further divided into the internal energy of the molecule (E_{int}), the electrostatic interactions (E_{ele}), and the van der Waals interactions (E_{vdW}):

$$E_{\text{MM}} = E_{\text{int}} + E_{\text{ele}} + E_{\text{vdW}} \tag{3}$$

$$E_{\text{int}} = E_{\text{bond}} + E_{\text{angle}} - E_{\text{torsion}}. \tag{4}$$

The internal energy E_{int} has three contributions: E_{bond}, E_{angle} and E_{torsion}, which represent the strain energies in bonds, angles, and torsion angles caused by deviations from their equilibrium values. The electrostatic and van der Waals energies were calculated using the Sander module. The solvation free energy contribution includes polar and nonpolar contributions:

$$\Delta G_{\text{sol}} = \Delta G_{\text{polar}} + \Delta G_{\text{nonpolar}}. \tag{5}$$

With the PB model, the polar portion (ΔG_{polar} in Eq. 5) was estimated by the pbsa program of Ambertools under the MM-PBSA approach.

In MM-PBSA calculations, the grid spacing was set to 0.5 Å, and the radii of the atoms were taken from the PARSE parameter set [42]. The values of the interior dielectric constant and the exterior dielectric constant were set to 1.0 and 80.0, respectively. The nonpolar contribution to the solvation free energy, $\Delta G_{\text{nonpolar}}$ in Eq. 5, was calculated from the solvent-accessible surface area (SASA) using the LCPO method [43] implemented within Sander, with a probe radius of 1.4 Å, according to the equation

$$\Delta G_{\text{nonpolar}} = \gamma SA + \beta, \tag{6}$$

where the surface tension γ and the offset β were set to the default values of 0.00542 kcal/(mol Å^2) and 0.92 kcal mol^{-1}, respectively.

Unlike the MM-PBSA method, in the MM-GBSA calculation, the polar contribution to the solvation free energy (ΔG_{polar} in Eq. 5) was calculated with the generalized Born (GB) model implemented in Sander, and the nonpolar contribution ($\Delta G_{\text{nonpolar}}$ in Eq. 5) was determined with the LCPO method based on the solvent-accessible surface area, as described in Eq. 6, in which the surface tension γ and the offset β were set to the default values of 0.0072 kcal/(mol Å^2) and 0.00 kcal mol^{-1}, respectively. Similar to MM-PBSA, the values of the interior dielectric constant and the exterior dielectric constant were set to 1.0 and 80.0, respectively [44].

For the calculations of E_{MM}, ΔG_{ELE}, and $\Delta G_{\text{nonpolar}}$, 200 snapshots from 2 ns to 3 ns were extracted from a single trajectory of the complex at time intervals of about 5 ps. In this study, we assumed that the entropy contributions were similar for different HIV protease variants and the ligand. When we calculated the relative binding free energies between them, the entropy contribution was neglected [29].

Binding free-energy decomposition (BFED)

Due to the time-consuming nature and the high computational demands of the PB calculation, the interactions between the inhibitor and each residue of HIV-1 protease were calculated with a decomposition process based GB model using the mm_pbsa program in Amber10. The per-residue contribution was further decomposed into two parts: one from side chains and the other from the backbone. The binding interactions of each inhibitor–residue pair ($\Delta G_{\text{inhibitor-residue}}$) were evaluated using the following equation:

$$\Delta G_{\text{inhibitor-residue}} = \Delta E_{\text{vdW}} + \Delta E_{\text{ele}} + \Delta G_{\text{polar}} + \Delta G_{\text{nonpolar}} \tag{7}$$

We did not take ΔE_{int} into account in Eq. 7 for per-atom decomposition because E_{int} is zero in a single trajectory and the entropy terms are neglected.

The van der Waals contribution (ΔE_{vdW}) and the electrostatic contribution (ΔE_{ele}) in Eq. 7 were computed by the Sander module in Amber10 [36]. In Eq. 7, ΔG_{polar} represents the polar interactions between the inhibitor and each protease residue during solvation, and was calculated using the GB model, with the charges taken from the Amber parameter set. The nonpolar solvation contribution ($\Delta G_{\text{nonpolar}}$ in Eq. 7) was obtained based on the corresponding SASA, as described in Eq. 6 [24].

By summing the atomic energy terms in Eq. 7 over each atom of a given residue, we obtained the contribution of this residue to the total binding free energy. The same snapshots were used to calculate all energy components as well as the total binding free energy.

Computational alanine scanning approach

The relative binding free-energy changes $\Delta\Delta G_{\text{bind}}$ of different HIV-1 protease variants and the inhibitor were calculated by CAS with the mm_pbsa.pl module in Amber10. $\Delta\Delta G_{\text{bind}}$ was estimated by comparing ΔG_{bind} of the alanine mutant to ΔG_{bind} of the wild type according to the following equation [29]:

$$\Delta\Delta G_{\text{bind}} = \Delta G_{\text{wildtype}} - \Delta G_{\text{mutant}}. \tag{8}$$

The key residues of HIV protease were chosen from the binding interface based upon the smallest ligand interaction distances. Since alanine scanning is not suitable for very small (such as glycine) or large residues (the backbone conformations of which differ significantly from that of alanine), prolines and glycines were not selected. The starting atomic coordinates in the alanine mutant structure were obtained by altering the coordinates of the last 1 ns trajectory. In this calculation, we assumed that the contribution to the change in the entropy of the mutant was not significant, so ΔG_{bind} derives from the changes in ΔE_{MM} and $\Delta G_{\text{solvate}}$. In computational alanine scanning, the entropy term can be removed because the entropies of the wild type and its mutants are similar for the same ligand and for similar receptors [29]. Here, the same set of snapshots obtained with the wild-type complex was used to calculate ΔG_{bind} for the mutants. The $\Delta\Delta G_{\text{bind}}$ values of the different HIV-1 protease variants and the inhibitor were obtained with the same snapshots of binding free-energy decomposition, according to Eq. 8.

Results and discussion

Stability and flexibility of the complex

For the complex of HIV-1 protease with inhibitor (TMC-126), MD simulations with the particle mesh Ewald (PME) method were performed in explicit water for 3 ns. In order to assess the dynamic stability of the protease complex, Fig. 2 shows the calculated root-mean-square displacement (RMSD) of the backbone atoms from the starting structure of the complex.

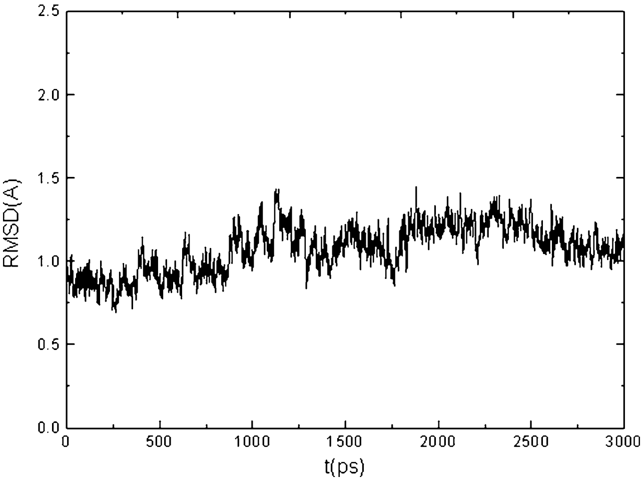


Fig. 2 The root-mean-square deviation (RMSD) of the backbone Cα atoms during the MD simulations of the complex with respect to the initial minimized structure of TMC-126

It is clear from Fig. 2 that a sharp rise is observed during the first 1200 ps. After that, the RMSD values fluctuate between 0.7 Å and 1.5 Å. After about 1.8 ns, the RMSD stabilizes and converges to a lower value of 1.1 Å, which indicates that the conformation of the complex has reached its equilibrium.

The initial structure and the superimposition of the average structure from the last 1 ns of snapshots of the complex are shown in Fig. 3. Our results showed that the average backbone RMSD value during the last 1 ns of MD trajectories was 0.92 Å, which indicates that the simulated structure is in good agreement with the experimental results. The last 1 ns of snapshots were used to calculate the binding free energy, free-energy decomposition, and computational alanine scanning, as described in the following subsections.

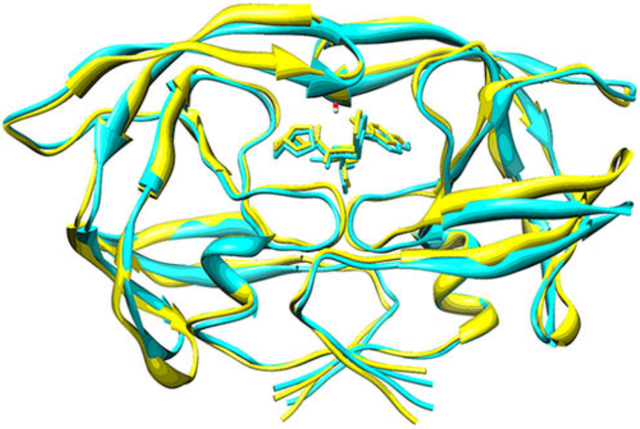


Fig. 3 The average structure from the last 1 ns of the MD trajectory of the complex of HIV-1 protease with TMC-126 superimposed on the initial structure via the protease's backbone atoms. The initial structure of the complex is shown in *cyan*, whereas the MD structure of the complex is shown in *yellow*. The figure was created using Chimera

A detailed analysis of the root-mean-square fluctuation (RMSF) versus the residue number in the complex is illustrated in Fig. 4. As seen from Fig. 4, the regions around Asp25 and Asp25′ show analogous RMSF behavior, with a minimum value of 0.3 Å. These results are in good agreement with experimental measurements [45], as well as other theoretical reports [46]. In addition to the N- and C-terminal residues, the regions around 17(17′), 41(41′), 52 (53′), 67(67′), and 81(81′) show the biggest dynamic fluctuations. Residues 1–37 and 59–99 in each monomer are defined as the core region, while residues 38–58 comprise the flap region. The flexibility of the flap region is crucial to the activity of the protease. As seen from Fig. 4, the flap region, especially the flap elbow region (residues 37–42), shows significant flexibility, which was also observed by Zhu et al. [47]. As described in "Theoretical methods," the crystallographic water that bridges the drug TMC-126 and Ile50/Ile50′ was included in the initial model. Our results showed that this bridging water was maintained throughout the whole MD simulation.

Binding free energy

In order to obtain the relative binding free energy and the VDW, electrostatic and solvation energy terms, the MM-PBSA and MM-GBSA methods were implemented using a single-trajectory protocol. As described in "Theoretical methods," when we calculated the relative binding free energy between HIV-1 and TMC-126, the entropy contribution was assumed to cancel out completely [48]. The values of the different energy terms shown in Eqs. 1–6 were obtained by averaging 200 snapshots taken from the last 1 ns of the MD simulation at 5 ps intervals.

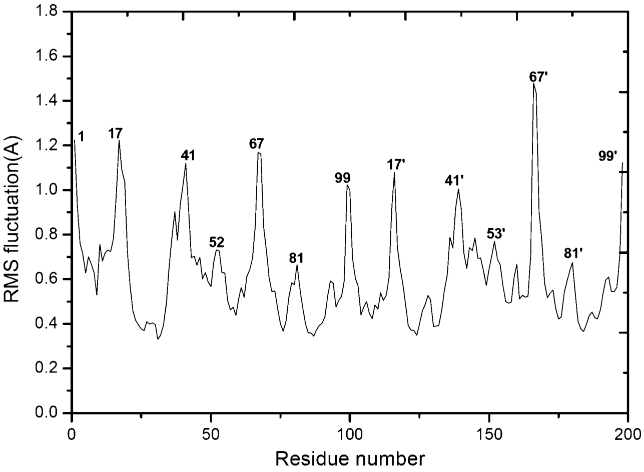


Fig. 4 The root-mean-square fluctuations of the backbone atoms versus the residue number of the HIV-1 protease and TMC-126 complex

Using Eqs. 1–6, the contributions of the energy components to the relative binding free energies of the receptor, ligand, and complex were calculated, and they are listed in Table 1. Using Eq. 7, with the MM-PBSA and MM-GBSA approaches, the calculated relative binding free energies of TMC-126 with HIV-1 are −37.01 and −62.37 kcal mol^{-1}, respectively (note that the entropy contribution was assumed to cancel out) [48]. Based on the contributions of the different energy components shown in Table 1, it is clear that the electrostatic interaction and the VDW interaction in the gas phase provide the driving forces for affinity binding. The nonpolar solvation energy provides a slightly favorable contribution to the binding of the inhibitor to PR, whereas the polar solvation energy provides an unfavorable energy component.

Free-energy decomposition

Analyzing the binding free-energy decomposition and the hydrogen bonds should lead to detailed insights into the binding mechanism for the PR–inhibitor complex. As described in "Theoretical methods," when the MM-GBSA approach is used, the binding free energy is decomposed into per-atom contributions that can be summed over atom groups to obtain the different energy contributions from residues, backbones and side chains [24, 25].

Table 2 reports the decomposition of ΔG_{bind} on a per-residue basis into the contributions from VDW, electrostatic interactions, polar solvation energy, and nonpolar solvation energy. From Table 2, it is apparent that the calculated protein–inhibitor binding energy (ΔG_{bind}) is greater than 0.12 kcal mol^{-1}, which is extremely helpful in elucidating the binding mechanism of TMC-126 to PR at the atomic level. The calculated binding energies between the asymmetric inhibitor TMC-126 and monomers A and B are 18.75 kcal mol^{-1} and 14.49 kcal mol^{-1}, respectively, which are very close to the values calculated by Zhang et al. [49]. The contributions of the flap region (38–58) and the active site region (residue 23–32) are 10.4 kcal mol^{-1} and 12.72 kcal mol^{-1}, respectively, which correspond to 31.39% and 38.39% of the contribution of PR to the binding; the flap elbow region makes a contribution of 0.02 kcal mol^{-1}, corresponding to 0.1% of the total binding.

Table 2 shows that the residues Ala28, Ile50, Ile84′, Gly27, Ile50′, Ala28′, Ile47′, and Gly49 contribute more than 1.5 kcal mol^{-1} to the binding. These residues are mostly from the flap region (residues 38–58) and the active site region (residues 23–32). As we know, the flap region and the active site region are important regions for the binding [47]. Ala28 contributes −4.65 kcal mol^{-1} to the binding affinity, most of which derives from its backbone (−3.96 kcal mol^{-1}). Ala28′ also makes a rather significant contribution of −1.77 kcal mol^{-1}, and its backbone contributes −1.11 kcal mol^{-1}. Ile50 and Ile50′ also provide large contributions to the binding: −2.63 kcal mol^{-1} and −2.08 kcal mol^{-1}, respectively. Unlike Ala28/Ala28′, instead of their backbones contributing most, their side chains (with VDW interactions) are the main contributors to the binding. Their nonpolar solvation energies (1.03 and 0.71 kcal mol^{-1}) appear to be unfavorable for binding. Similar to Ile50/Ile50′, the side-chain contribution dominates for the other two isoleucine residues, Ile84 and Ile27. As shown in Table 2, Ile47 has a greater contribution from its backbone than from its side chain, and its binding occurs mainly through the VDW interaction. More than a half of the contribution of Gly49′ (−1.61 kcal mol^{-1}) originates from the side chain (−0.91 kcal mol^{-1}), according to the calculated free-energy decomposition.

Table 1 The binding free energy components of the protein–inhibitor complex (HIV-1–TCM126), as calculated using MMPBSA methods (unit: kcal mol^{-1})

Component[a]	Complex		PR		TMC-126		Delta	
	Mean[b]	Std[c]	Mean	Std	Mean	Std	Mean	Std
E_{ele}	−4103.6	54.13	−3939.2	53.79	−131.6	2.19	−32.8	5.65
E_{vdW}	−833.22	20.58	−771.16	19.85	5.72	2.64	−67.77	3.59
E_{MM}	−539.47	64.38	−441.34	63.28	2.44	6.14	−100.57	5.6
$G_{\mathrm{nonpolar,PB}}$	56.83	0.77	58.83	0.78	5.29	0.05	−7.29	0.09
$G_{\mathrm{polar,PB}}$	−2380.55	45.03	−2411.06	45.05	−40.33	1.29	70.84	4.17
$G_{\mathrm{solvation,PB}}$	−2323.72	44.59	−2352.23	44.62	−35.04	1.28	63.55	4.15
$G_{\mathrm{subtotal,PB}}$	−2863.19	41.54	−2793.57	40.5	−32.61	5.97	−37.01	4.68
$G_{\mathrm{nonpolar,GB}}$	74.27	1.03	76.93	1.04	5.8	0.07	−8.46	0.12
$G_{\mathrm{polar,GB}}$	−2398.71	47.09	−2409.76	47.04	−35.61	1.58	46.66	4.42
$G_{\mathrm{solvation,GB}}$	−2324.44	46.5	−2332.83	46.45	−29.81	1.56	38.2	4.4
$G_{\mathrm{subtotal,GB}}$	−2863.91	39.26	−2774.17	38.56	−27.37	6.11	−62.37	4.46

[a] Components: E_{ele} Coulombic energy; E_{vdW} VDW energy; $E_{\mathrm{MM}}=E_{\mathrm{ele}}+E_{\mathrm{vdW}}$; $G_{\mathrm{polar,PB}}$ polar solvation energy; $G_{\mathrm{nonpolar,PB}}$ nonpolar solvation energy; $G_{\mathrm{solvation,PB}}=G_{\mathrm{polar,PB}}+G_{\mathrm{nonpolar,PB}}$; $G_{\mathrm{subtotal,PB}}=E_{\mathrm{MM}}+G_{\mathrm{solvation,PB}}$

[b] Average of 200 snapshots

[c] Standard error of the mean value

Table 2 Decomposition of ΔG_{bind} on a per-residue basis into contributions from the van der Waals energy (ΔE_{vdW}), electrostatic interaction energy (ΔE_{ele}), nonpolar solvation free energy ($\Delta G_{polar,GB}$) and polar free energy ($\Delta G_{nonpolar}$) (units: kcal mol^{-1})

Residue	ΔE_{vdW}	ΔE_{ele}	$\Delta G_{polar,GB}$	$\Delta G_{nonpolar}$	S $\Delta G_{subtotal}$	B $\Delta G_{subtotal}$	T $\Delta G_{subtotal}$
Ala28	−2.55	−2.01	0.09	−0.18	−0.69	−3.96	−4.65
Ile50	−2.3	−1.18	1.03	−0.18	−2.16	−0.47	−2.63
Ile84′	−1.85	−0.2	0.01	−0.17	−2.13	−0.09	−2.22
Gly27	−1.37	0.0	−0.66	−0.1	−0.71	−1.43	−2.14
Ile50′	−1.72	−0.9	0.71	−0.17	−1.83	−0.26	−2.08
Ala28′	−1.46	0.42	−0.55	−0.18	−0.66	−1.11	−1.77
Ile47′	−1.53	0.31	−0.3	−0.16	−1.56	−0.12	−1.69
Gly49′	−0.81	−1.67	0.96	−0.08	−0.91	−0.7	−1.61
Ile84	−1.15	0.22	−0.33	−0.13	−1.27	−0.11	−1.39
Gly49	−1.05	−1.29	1.09	−0.12	−0.75	−0.63	−1.38
Val82′	−0.97	0.06	−0.34	−0.09	−1.2	−0.15	−1.35
Val32′	−0.78	0.0	−0.43	−0.07	−1.15	−0.11	−1.27
Gly27′	−0.61	0.35	−0.67	−0.04	−0.41	−0.57	−0.98
Leu23′	−0.7	−0.14	−0.04	−0.04	−0.79	−0.13	−0.92
Arg87	−0.19	−2.19	1.58	0.0	−0.74	−0.07	−0.8
Ile47	−0.89	0.36	−0.14	−0.11	−0.74	−0.05	−0.79
Arg8′	−0.87	−1.24	1.48	−0.16	−0.73	−0.06	−0.79
Pro81′	−0.63	−0.17	0.11	−0.09	−0.61	−0.16	−0.78
Asp25	−1.1	1.44	−0.95	−0.07	−0.68	0.0	−0.68
Val32	−0.49	−0.03	−0.13	−0.03	−0.64	−0.04	−0.68
Val82	−0.38	−0.08	−0.05	−0.05	−0.52	−0.05	−0.57
Leu23	−0.38	0.12	−0.24	−0.04	−0.44	−0.09	−0.54
Asp29	−1.6	1.5	−0.3	−0.11	0.78	−1.3	−0.52
Pro81	−0.35	−0.1	0.04	−0.04	−0.4	−0.05	−0.46
Leu76′	−0.39	0.1	−0.11	−0.02	−0.4	−0.02	−0.42
Thr26	−0.2	0.3	−0.48	0.0	0.13	−0.52	−0.38
Asn83′	−0.09	0.05	−0.3	0.0	−0.03	−0.32	−0.35
Gly86	−0.09	−0.35	0.13	0.0	−0.23	−0.07	−0.31
Gly86′	−0.08	0.27	−0.4	0.0	−0.22	0.01	−0.21
Leu24′	−0.09	0.15	−0.27	0.0	−0.01	−0.19	−0.2
Val56′	−0.06	0.02	−0.13	0.0	−0.19	0.01	−0.17
Leu76	−0.14	−0.03	0.01	0.0	−0.15	−0.01	−0.16
Asn83	−0.05	−0.03	−0.08	0.0	−0.02	−0.14	−0.16
Arg87′	−0.07	1.82	−1.9	0.0	−0.08	−0.07	−0.15
Asp30	−1.27	1.99	−0.75	−0.08	0.23	−0.35	−0.12

S, B, and T represent the side-chain residue, monomer B, and total (monomers A and B) contributions, respectively

Computational alanine scanning

The computational alanine scanning (CAS) method was implemented to acquire $\Delta\Delta G_{bind}$ by mutating residues of proteins to alanine. HIV-1 is a homodimer protease consisting of two identical monomers. In terms of the binding decomposition method, the CAS method can mutate the same residue in each monomer simultaneously, so that we can gain insight into the function of the residue in the homodimer.

In order to perform CAS, we first need to select which residue pairs can be mutated to alanine. Our criteria for selecting these mutated pairs were that the chosen residues should make significant contributions to the binding, or that they should be located in a crucial area such as the binding interface, pocket, or flap region. As mentioned in "Theoretical methods," the CAS method does not work well for very small or very big residue mutations. The mutation of proline [50] to alanine sometimes leads to significant conformational changes [29]. In this study, based on our selection criteria, and by combining the residue locations and the results reported in Table 3, 15 pairs of residues were chosen for mutation to alanine in the CAS method: Arg8, Leu23, Leu24, Thr26, Asp29, Asp30, Val32, Ile47, Ile50, Val56, Leu76, Val82, Asn83, Ile84, and Arg87. In order to make this method practical, we assumed that local changes do not

Table 3 The computational alanine scanning results for the HIV-1 protease and TMC-126 complex

Contribution	Arg8Ala	Std	Leu23Ala	Std	Leu24Ala	Std	Thr26Ala	Std	Asp29Ala	Std
$\Delta\Delta E_{\mathrm{ele}}$	0.85	1.65	0.03	0.03	−0.03	0.01	−0.08	0.09	3.26	2.48
$\Delta\Delta E_{\mathrm{vdW}}$	−1.83	0.35	−1.73	0.43	−0.03	0.00	−0.07	0.01	−1.39	0.29
$\Delta\Delta E_{\mathrm{MM}}$	−0.98	1.68	−1.7	0.42	−0.06	0.01	−0.15	0.09	1.87	2.53
$\Delta\Delta G_{\mathrm{nonpolar,PB}}$	−0.13	0.04	0.12	0.03	0.0	0.01	0.0	0.01	−0.02	0.05
$\Delta\Delta G_{\mathrm{polar,PB}}$	0.77	1.68	−0.12	0.79	0.01	0.28	−0.07	0.31	−2.06	2.31
$\Delta\Delta G_{\mathrm{solvation,PB}}$	0.64	1.67	0.0	0.77	0.01	0.28	−0.07	0.31	−2.08	2.31
$\Delta\Delta G_{\mathrm{subtotal,PB}}$	−0.33	1.02	−1.69	0.85	−0.04	0.28	−0.21	0.28	−0.19	1.54
$\Delta\Delta G_{\mathrm{nonpolar,GB}}$	−0.17	0.05	0.16	0.04	0.0	0.01	0.0	0.01	−0.02	0.07
$\Delta\Delta G_{\mathrm{polar,GB}}$	−1.33	1.41	−1.02	0.14	−0.05	0.04	0.78	0.24	0.25	2.24
$\Delta\Delta G_{\mathrm{solvation,GB}}$	−1.5	1.42	−0.86	0.14	−0.05	0.04	0.78	0.24	0.25	2.24
$\Delta\Delta G_{\mathrm{solvation,GB}}$	−1.5	1.42	−0.86	0.14	−0.05	0.04	0.78	0.24	0.23	2.24
$\Delta\Delta G_{\mathrm{subtotal,GB}}$	−2.48	0.71	−2.57	0.46	−0.12	0.04	0.63	0.19	2.1	1.06
Contribution	Asp30Ala	std	Val32Ala	std	Ile47Ala	std	Ile50Ala	std	Val56Ala	Std
$\Delta\Delta E_{\mathrm{ele}}$	−1.11	2.37	0.01	0.24	0.25	0.20	−0.09	6.94	0.09	0.06
$\Delta\Delta E_{\mathrm{vdW}}$	−0.86	0.31	−1.4	0.46	−2.85	0.50	−3.84	4.51	−0.12	0.02
$\Delta\Delta E_{\mathrm{MM}}$	−1.97	2.34	−1.39	0.49	−2.6	0.56	−3.93	6.42	−0.03	0.06
$\Delta\Delta G_{\mathrm{nonpolar,PB}}$	−0.01	0.03	0.09	0.02	0.09	0.03	0.15	0.12	0.0	0.01
$\Delta\Delta G_{\mathrm{polar,PB}}$	4.35	2.08	−1.53	0.71	0.85	0.59	2.17	5.07	0.03	0.19
$\Delta\Delta G_{\mathrm{solvation,PB}}$	4.34	2.08	−1.44	0.71	0.94	0.58	2.32	5.07	0.03	0.19
$\Delta\Delta G_{\mathrm{subtotal,PB}}$	2.38	1.93	−2.82	0.80	−1.65	0.78	−1.6	5.54	0.01	0.19
$\Delta\Delta G_{\mathrm{nonpolar,GB}}$	−0.01	0.03	0.12	0.02	0.12	0.04	0.2	0.16	0.0	0.01
$\Delta\Delta G_{\mathrm{polar,GB}}$	3.59	2.00	−1.32	0.33	−0.24	0.17	−0.89	5.33	−0.58	0.11
$\Delta\Delta G_{\mathrm{solvation,GB}}$	3.58	2.01	−1.2	0.34	−0.12	0.18	−0.69	5.34	−0.58	0.11
$\Delta\Delta G_{\mathrm{subtotal,GB}}$	1.61	0.86	−2.59	0.49	−2.72	0.53	−4.62	5.57	−0.61	0.08
Contribution	Leu76Ala	std	Val82Ala	std	Asn83Ala	std	Ile84Ala	std	Arg87Ala	std
$\Delta\Delta E_{\mathrm{ele}}$	0.01	0.04	−0.38	0.25	0.12	0.11	0.26	0.28	−0.64	1.08
$\Delta\Delta E_{\mathrm{vdW}}$	−0.85	0.32	−1.61	0.41	−0.03	0.00	−4.6	0.82	−0.36	0.04
$\Delta\Delta E_{\mathrm{MM}}$	−0.84	0.30	−1.99	0.43	0.09	0.11	−4.34	0.88	−1.0	1.09
$\Delta\Delta G_{\mathrm{nonpolar,PB}}$	0.04	0.03	0.05	0.02	0.0	0.01	0.15	0.05	0.0	0.01
$\Delta\Delta G_{\mathrm{polar,PB}}$	0.8	0.90	0.35	0.35	−0.15	0.16	1.73	1.13	−0.77	1.07
$\Delta\Delta G_{\mathrm{solvation,PB}}$	0.83	0.90	0.4	0.34	−0.16	0.16	1.88	1.10	−0.77	1.07
$\Delta\Delta G_{\mathrm{subtotal,PB}}$	0.01	0.90	−1.57	0.53	−0.06	0.12	−2.45	1.49	−1.76	0.62
$\Delta\Delta G_{\mathrm{nonpolar,GB}}$	0.05	0.04	0.07	0.03	0.0	0.01	0.2	0.07	0.0	0.01
$\Delta\Delta G_{\mathrm{polar,GB}}$	−0.13	0.09	−0.9	0.31	−0.68	0.13	−1.15	0.31	−0.63	0.98
$\Delta\Delta G_{\mathrm{solvation,GB}}$	−0.08	0.11	−0.83	0.31	−0.69	0.13	−0.95	0.30	−0.64	0.98
$\Delta\Delta G_{\mathrm{subtotal,GB}}$	−0.92	0.27	−2.82	0.50	−0.6	0.10	−5.29	0.91	−1.63	0.40

Std standard error (units: kcal mol^{-1})

impact on the global conformation of the PR and the total binding modes for the binding of PR to inhibitor. Such an assumption has been proven to be applicable for most mutations according to various alanine scanning mutagenesis experiments [51].

The results obtained with the CAS approach for 15 pairs of residues of PR are shown in Table 3. The changes in the energy terms (VDW interactions, electrostatic interactions, the polar and nonpolar solvation free energies) upon alanine mutation are also listed in Table 3. According to Eq. 8, negative values of $\Delta\Delta G_{\mathrm{bind}}$ indicate unfavorable substitutions. In contrast, positive $\Delta\Delta G_{\mathrm{bind}}$ values suggest that the alanine residue at the mutated position is more favorable for binding.

As can be seen from Table 3, the binding free energies drop considerably when the six critical residues (Leu23, Val32, Ile47, Ile50, Val82, and Ile84) are mutated to alanine. On the contrary, the $\Delta\Delta G_{\mathrm{subtotal,GB}}$ values of Val26, Asp29, and Asp30 are all positive, which indicates that the binding between protease and TMC-126 is influenced when these residues are mutated to alanine. These results are in good agreement with those obtained from the free-energy decomposition method.

Interactions between HIV-1 protease and inhibitor

In this section, we analyze some PR residues that are key to the binding to the inhibitor, based on the results obtained using BFED and CAS and the hydrogen bond data listed in Table 4.

The interactions between PR and the inhibitor based on the average structure from the MD simulations are plotted in Fig. 5. In order to investigate the hydrogen bonds during MD simulations, the results of a dynamic analysis of hydrogen bonds based on the trajectories of the MD simulations are listed in Table 4. Some of the hydrogen bonds are also indicated in Fig. 5.

The BEFD-calculated results show that the favorable residues mainly come from six groups around Ala28/Ala28′, Ile50/Ile50′, and Ile84/Ile84′. Table 4 shows that the residues that make significant contributions to the binding are all hydrophobic amino acids. According to the results shown in Table 2, the major force that drives the inhibitor to most of the residues in PR is VDW interactions, especially for the essential residues. This conclusion is consistent with the binding free energies shown in Table 1.

As shown in Table 2, Ala28 contributes most of the binding affinity. The main driving forces for the binding of the inhibitor to Ala28 are the VDW energy (−2.55 kcal mol^{-1}) and electrostatic energy (−2.01 kcal mol^{-1}), which originate in the C–H…π interactions between the bis-tetrahydrofuran (THF) and the alkyl of Ala28, and in the C–H…O interactions of the side-chain atoms of Ala28 with the oxygen atoms of the bis-THF and inhibitor, respectively. The interaction between Ala28′ and the inhibitor is similar to that of Ala28, in that the main driving force for the binding of Ala28′ to the inhibitor is the van der Waals energy (−1.46 kcal mol^{-1}) from the C–H…π interactions between the phenoxymethyl of the inhibitor and the alkyl of Ala28′. In this case, the favorable polar solvation energy (−0.55 kcal mol^{-1}) is mostly countered by the unfavorable electrostatic energy (0.42 kcal mol^{-1}).

Table 4 Hydrogen bonds for the last 1000 ps of the trajectory

Hydrogen bonds		% Occupied	Distance
Donor	Receptor		
Asp25–OD1	Ala28–N–H	98.80	2.851 (0.10)
Asp25′–OD1	PI–O22–H	98.60	2.670 (0.12)
PI–O4	Asp29′–N–H	95.2	3.004 (0.17)
PI–O2	Asp30′–N–H	61.2	3.203 (0.17)
PI–O2	Asp29′–N–H	43.2	3.220 (0.17)
Hydrogen bonds between WAT and PR[b] and PI[a]			
WAT–O	Ile50–N–H	91.2	3.084 (0.18)
WAT–O	Ile50′–N–H	90.2	3.022 (0.17)
PI–O11	WAT–O–H1	86.60	2.787 (0.14)
PI–O26	WAT–O–H2	83.20	2.748 (0.15)

[a] HIV-1 protease

[b] HIV-1 protease inhibitor

Although Asp25 and Asp25′ are located at the active site as catalytic aspartic acids, monoprotonated Asp25 contributes only −0.68 kcal mol^{-1} with an unfavorable electrostatic interaction (1.44 kcal mol^{-1}) to the total binding affinity, and Asp25′ makes an unfavorable contribution (2.02 kcal mol^{-1}) to the binding affinity due to the combination of a strongly unfavorable polar solvation energy (11.48 kcal mol^{-1}) and a highly favorable electrostatic energy (−9.51 kcal mol^{-1}). As shown in Table 4, the OD1 of Asp25 and Asp25′ form hydrogen bonds with Ala28 and the inhibitor with occurrence rates of >98%, which suggests that the Ala28 pair make a favorable contribution to the binding that stabilizes the complex.

Although the PR is symmetrical, it is clear from Table 4 that the calculated contributions of Asp29 and Asp29′ are different: negative (−0.52 kcal mol^{-1}) and positive (0.14 kcal mol^{-1}), respectively. Similar behavior is also found for the residues Asp30 (−0.12 kcal mol^{-1}) and Asp30′ (0.36 kcal mol^{-1}). The results in Table 4 show that Asp29 and Asp30 form stable hydrogen bonds with the oxygen of the bis-THF in the inhibitor. The donor–acceptor distances for these pairs and their corresponding occupancies are 3.004 Å and 3.204 Å, 95.2% and 61.2%, respectively. On the other hand, Asp29′ and Asp30′ do not form hydrogen bonds with the inhibitor. Therefore, the hydrogen bonds make favorable contributions to the VDW interactions of Asp29 and Asp30 with the inhibitor, which suggests that the bis-THF group in TMC-126 plays an important role in the binding with the PR. In addition, as shown in Table 3, the calculated $\Delta\Delta G_{\text{subtotal,GB}}$ values of the residue pairs Asp29Ala and Asp30Ala are 2.1 and 1.61 kcal mol^{-1} respectively, which means that the side chains of Asp29 and Asp30 make unfavorable contributions to the binding affinity. The calculated $\Delta\Delta E_{\text{ele}}$ of Asp29Ala and the $\Delta\Delta G_{\text{polar,GB}}$ of Asp30Ala are 3.26 kcal mol^{-1} and 3.59 kcal mol^{-1}, respectively, indicating that the electrostatic interactions of their side chains have a significant effect on the binding to the inhibitor.

The calculated results listed in Table 3 show that there is a significant loss of binding free energy when the four critical residue pairs Ile47/Ile47′ Ile50/Ile50′, Ile84/Ile84′, and Val82/Val82′ are mutated to alanine. The CAS-calculated results show that the residues Ile50/Ile50′ have a significant effect on the binding, a −4.71 kcal mol^{-1} reduction in $\Delta\Delta G_{\text{subtotal,GB}}$, which is in fair agreement with the side-chain contribution (−3.99 kcal mol^{-1}) obtained

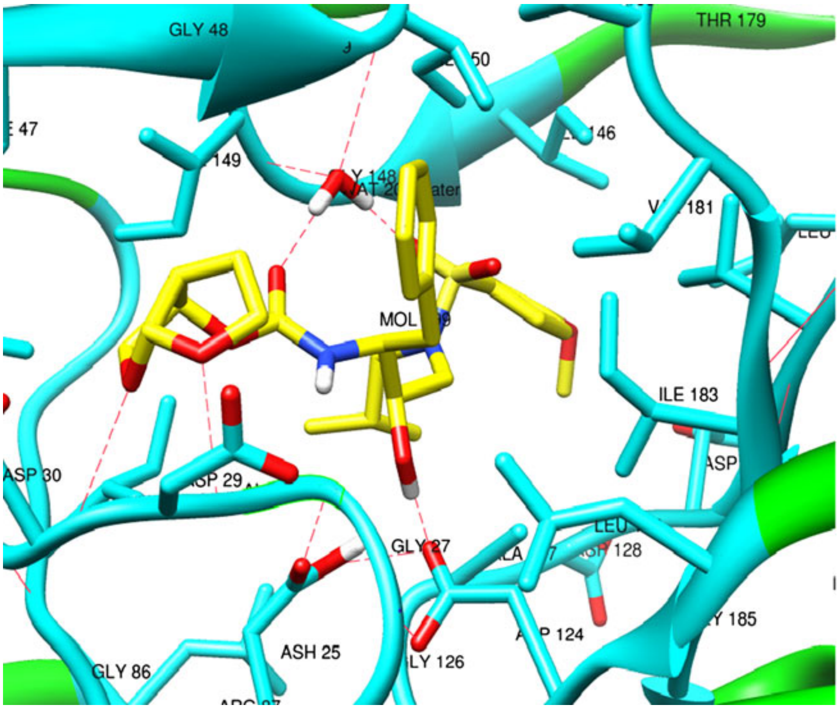


Fig. 5 Geometries of ten residues of HIV-1 protease that participate in some of the strongest interactions with TMC-126, based on the average structure from the last 1 ns of MD simulation

using the BFED method (Tables 2 and 3). As shown in Table 3, the favorable contributions of Ile50 and Ile50′ to the binding energy originate mainly from the VDW interaction (−3.84 kcal mol^{-1}) in Table 3. A possible reason for this is that the side-chain alkyls of Ile50 and Ile50′ form some C–H…H–C interactions with the inhibitor. In addition, the hydrogen atoms connect with the backbone nitrogen atoms of Ile50 and Ile50′, forming strong hydrogen bonds with the oxygen of the WAT207 (Table 4).

As can be seen from Tables 2 and 3, the interactions of Ile84/Ile84′ and Ile47′ with the inhibitor are similar to that of Ile50/Ile50′. Moreover, the contact of the alkyls of Ile47′ and Ile84′ with the phenyl group of the inhibitor, which results in C–H…π interactions, can also add strong van der Waals interactions to the binding. These results indicate that the VDW interactions significantly favor the binding of these isoleucine residues.

It is well known that the crystal water molecule WAT207 plays an important role in PR–inhibitor binding, since it can generally form four hydrogen bonds with Ile50/Ile50′ and TMC-126 [26]. As shown in Table 4, the occupancy rates of these four hydrogen bonds are higher than 80%, which suggest that these hydrogen bonds are extraordinarily stable during MD simulations, and that the WAT207 act as a conduit to connect the inhibitor with PR.

Comparisons between the CAS and the BFED methods

The calculated results obtained by computational alanine scanning using the MM-PBSA and MM-GBSA approaches are shown in Table 5. The regression between the calculated $\Delta\Delta G_{\text{subtotal.PB}}$ and $\Delta\Delta G_{\text{subtotal.GB}}$ values for the 15 pairs of residues is depicted in Fig. 6. From Fig. 6, the correlation coefficient between the calculated $\Delta\Delta G_{\text{subtotal.PB}}$ and $\Delta\Delta G_{\text{subtotal.GB}}$ values was found to be 0.76 for 15 mutations. As shown in Table 5, the values of $\Delta\Delta G_{\text{bind}}$ obtained by the PB model are clearly smaller than those yielded by the GB model, except for Val32Ala and Arg87Ala. Since the standard deviations obtained by the GB model are a little bit smaller than those given by the PB model, it appears that in this system, the PB model is more sensitive to the atomic coordinates than the GB model when calculating contributions to the solvation energy. These conclusions are consistent with the results reported by Li et al. [52].

In order to compare the computational alanine scanning (CAS) approach with the binding free-energy decomposition (BFED) approach, in Table 5 we sum the calculated $\Delta\Delta G_{\text{subtotal}}$ values of the same 15 pairs of residues corresponding to chains A and B using the BFED method, and the calculated $\Delta\Delta G_{\text{subtotal,GB}}$ results for the 15 pairs of residues obtained by the CAS method.

The per-residue contributions calculated using the BFED method include the contributions from the backbone and the side chain. As far as the CAS method is concerned, the backbone does not generally change, because only Cγ is replaced by a methyl in the mutated topology file. Consequently, $\Delta\Delta G_{\text{subtotal}}$ mainly reflects the contribution from the side chain of the residue. Thus, the side-chain contributions from the residues in monomers

Table 5 Summation of $\Delta G_{\text{subtotal}}$ obtained by free-energy decomposition for the same residues in chain A and chain B of the complex, and the $\Delta\Delta G_{\text{subtotal,GB}}$ values obtained by computational alanine scanning for the 15 pairs of residues (units: kcal mol^{-1})

Residue	A [a] S [c] $\Delta G_{\text{subtotal}}$	A $\Delta G_{\text{subtotal}}$	B [b]S $\Delta G_{\text{subtotal}}$	B $\Delta G_{\text{subtotal}}$	T [d]S $\Delta G_{\text{subtotal}}$	T $\Delta G_{\text{subtotal}}$	$\Delta\Delta G_{\text{subtotal,GB}}$
Arg8	0.01	−0.06	−0.73	−0.79	−0.72	−0.85	−2.48
Leu23	−0.44	−0.54	−0.79	−0.92	−1.23	−1.46	−2.57
Leu24	−0.01	−0.11	−0.01	−0.2	−0.02	−0.31	−0.12
Thr26	0.13	−0.38	0.11	−0.09	0.24	−0.47	0.63
Asp29	0.78	−0.52	0.13	0.14	0.91	−0.38	2.1
Asp30	0.23	−0.12	0.29	0.36	0.52	0.24	1.61
Val32	−0.64	−0.68	−1.15	−1.27	−1.79	−1.95	−2.59
Ile47	−0.74	−0.79	−1.56	−1.69	−2.3	−2.48	−2.72
Ile50	−2.16	−2.63	−1.83	−2.08	−3.99	−4.71	−4.62
Val56	−0.08	−0.07	−0.19	−0.17	−0.27	−0.24	−0.61
Leu76	−0.15	−0.16	−0.4	−0.42	−0.55	−0.58	−0.92
Val82	−0.52	−0.57	−1.2	−1.35	−1.72	−1.92	−2.82
Asn83	−0.02	−0.16	−0.03	−0.35	−0.05	−0.51	−0.6
Ile84	−1.27	−1.39	−2.13	−2.22	−3.4	−3.61	−5.29
Arg87	−0.74	−0.8	−0.08	−0.13	−0.82	−0.93	−1.64

[a] A represents $\Delta G_{\text{subtotal}}$ for the residues in chain A

[b] B represents $\Delta G_{\text{subtotal}}$ for the residues in chain B

[c] S represents $\Delta G_{\text{subtotal}}$ for the side chains

[d] T represents the summation $\Delta\Delta G_{\text{subtotal}}$ for the residues in chains A and B

A and B obtained using the BFED method and the $\Delta\Delta G_{\text{subtotal,GB}}$ results obtained by the CAS method are compared and shown in Fig. 7. The correlation coefficient between $\Delta\Delta G_{\text{subtotal}}$ for the side chain calculated by the BFED method and $\Delta\Delta G_{\text{subtotal,GB}}$ calculated by the CAS method is 0.94 for the 15 mutants of residues, which indicates that the CAS method allows useful insight into the contribution of the side chain. Zoete and Michielin studied different systems using both BFED and CAS, and concluded that both of these methods can achieve consistent results [32], which is in good agreement with our calculated results.

Our results showed that the correlation between BFED and CAS is better than that between GB and PB, as calculated by the CAS method. A possible reason for this is that both BEED and CAS are based upon the GB model in our study.

Both the BFED and the CAS methods have been widely used to identify the hotspots of receptor–ligand complexes reliably and to obtain further insight into their binding and related functional information [32]. Using the analysis of the BFED and CAS methods, it can be concluded that

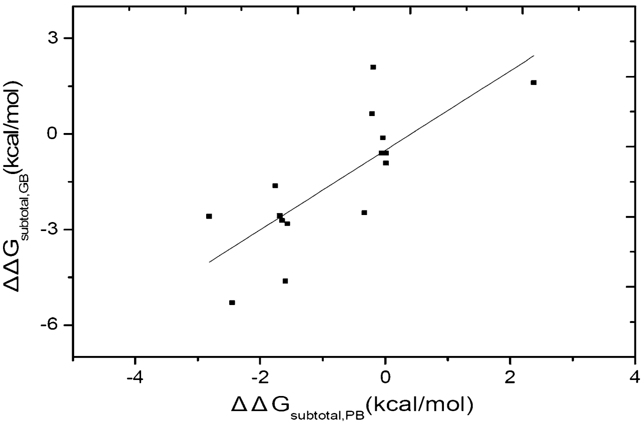


Fig. 6 Regression between the calculated $\Delta\Delta G_{\text{subtotal,PB}}$ and $\Delta\Delta G_{\text{subtotal,PB}}$ values obtained by computational alanine scanning for the 15 residues

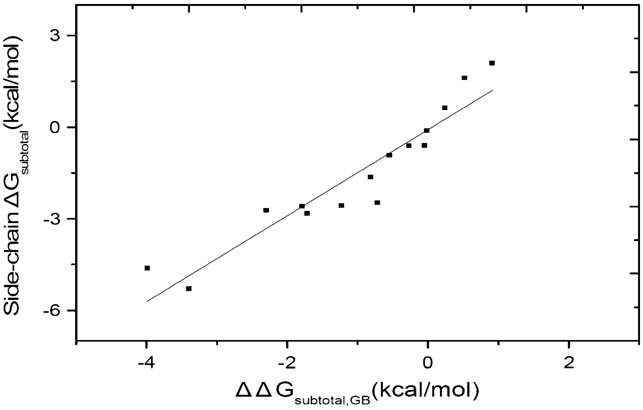


Fig. 7 Regression between the calculated $\Delta\Delta G_{\text{subtotal,GB}}$ obtained by computational alanine scanning and the $\Delta G_{\text{subtotal}}$ for the side chain, obtained by free-energy decomposition for the 15 pairs of residues of HIV-1 protease

although these two methods are different, they are complementary to some extent.

First, the BFED method is based upon the GB model; as far as the CAS method is concerned, both PB and GB models are applied to estimate the binding free-energy differences. Consequently, owing to the distinctness of the models' principles, the results calculated based on the PB model of CAS methods are more accurate than those calculated with the BFED methods. Compared to BFED, the CAS calculations are time-consuming. Second, by applying the BFED approach at an atomistic level, the per-atom contributions can be summed over atomic groups such as residues, backbones, and side chains, in order to obtain their contributions to the binding free energy of the receptor–ligand complex. Simultaneously, the BFED approach allows us to decompose each residue's binding free energy conveniently into different energetic components, such as VDW interactions, electrostatic interactions, and nonpolar solvation energy. On the other hand, the CAS approach requires separate, time-consuming calculations to perform a detailed study of protein–protein interactions at the residue level, and focuses on the impact of side chains on the binding affinity, since it is only employed to evaluate the side-chain contributions of the residues of intest that are mutated into alanine in the topology files. Unfortunately, the CAS method can only be applied to specific residues: not very small residues (such as glynine) or residues that would induce significant global conformational changes (such as proline and cysteine). Third, it is worth mentioning that the CAS approach provides a chance to investigate the influence of several mutants in the complex on the binding affinity, because it allows more than one residue to be mutated to alanine when the topology files are prepared. In contrast to the BFED method, the CAS method provides a preferable insight into systems in which several mutations exist simultaneously; in particular, it can be applied to the dimer and even more complicated systems, such as the HIV-1 protease homodimer and inhibitor complex. As the CAS method substitutes alanine for other residues, this approach can also be regarded as an effective tool to analyze the drug resistance caused by mutagenesis, which could result in improvements to drug design and better guidance for new experimental investigations. In addition, the results obtained by CAS can be compared directly to the experimental data on mutagenesis. However, the CAS approach is not suitable for detecting the binding mechanisms of complexes in which the binding affinity originates from the backbones.

Conclusions

The binding free energy of the complex of HIV-1 protease and TMC-126 was calculated using the MM-PBSA and MM-GBSA methods based on decomposition of the energy at an atomic level on the basis of the GB model. The computational alanine scanning method based on the GB and PB models was applied to this complex in order to investigate the different contributions of HIV-1 protease residues when binding to TMC-126.

The binding mechanism of PR with TMC-126 was investigated by structural analysis, by calculating the free energy and decomposing the inhibitor–residue interaction, and computational alanine scanning. The favorable interactions and the driving forces in the binding of the inhibitor to PR are the van der Waals and electrostatic forces, which mainly come from six groups around Ala28/Ala28′, Ile50/Ile50′, and Ile84/Ile84. Ala28 mainly contributes to the binding affinity. The contributions of Asp29 and Asp30 to the binding in chains A and B are significantly different because they form hydrogen bonds with bis-THF of the inhibitor, which suggests that the bis-THF group in TMC-126 plays an important role in the binding of the PR. The VDW energy significantly favors binding for isoleucine residues such as Ile47, Ile50 and Ile84. The flap region and the active site region of the PR are crucial to its binding affinity, contributing about 69.78% of the total binding affinity. The crystal water molecule acts as a bridging medium between the inhibitor and PR by forming four hydrogen bonds among the residues Ile50/Ile50′ and TMC-126.

Based on the correlation coefficients obtained from regression analyses relating to different theoretical methods and models, both the BFED and the CAS methods have particular advantages and weaknesses when investigating the binding mechanism. BFED is a rapid and convenient approach. It does not need to consider the global change and can prove the effects of both the backbone and side chains for each residue. However, the CAS method can provide preferable insight into the resistance of mutagenesis and the binding affinities between residues in dimic and multimeric proteins and inhibitors. To conclude, these two complementary methods provide a useful way to determine the hotspot residues and to investigate the binding affinity incisively. We expect that this work will provide some helpful insights into the future of drug design with potent inhibitors.

Acknowledgments This work is supported by the 973 Fund of the Chinese Ministry of Science and Technology (2010CB934504) and the Natural Science Fund of China.

References

1. Wlodawer A (2002) Rational approach to AIDS drug design through structural biology. Annu Rev Med 53:595–614
2. Navia MA, Fitzgerald PMD, Mckeever BM, Leu CT, Heimbach JC, Herber WK, Sigal IS, Darke PL, Springer JP (1989) 3-Dimensional structure of aspartyl protease from human immunodeficiency virus Hiv-1. Nature 337:615–620

3. Prabu-Jeyabalan M, Nalivaika E, Schiffer CA (2002) Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. Structure 10:369–381
4. Swain AL, Miller MM, Green J, Rich DH, Schneider J, Kent SBH, Wlodawer A (1990) X-ray crystallographic structure of a complex between a synthetic protease of human immunodeficiency virus-1 and a substrate-based hydroxyethylamine inhibitor. Proc Natl Acad Sci USA 87:8805–8809
5. Wlodawer A, Miller M, Jaskolski M, Sathyanarayana BK, Baldwin E, Weber IT, Selk LM, Clawson L, Schneider J, Kent SBH (1989) Conserved folding in retroviral proteases—crystal structure of a synthetic HIV-1 protease. Science 245:616–621
6. Barbaro G, Scozzafava A, Mastrolorenzo A, Supuran CT (2005) Highly active antiretroviral therapy: current state of the art, new agents and their pharmacological interactions useful for improving therapeutic outcome. Curr Pharm Des 11:1805–1843
7. Chen RX, Quinones-Mateu ME, Mansky LM (2004) Drug resistance, virus fitness and HIV-1 mutagenesis. Curr Pharm Des 10:4065–4070
8. Clavel F, Hance AJ (2004) Medical progress: HIV drug resistance. New Engl J Med 350:1023–1035
9. Surleraux DLNG, Tahri A, Verschueren WG, Pille GME, Kock HA, Jonckers THM, Peeters A, Meyer DS, Azijn H, Pauwels R, Bethune MP, King NM, Prabu-Jeyabalan M, Schiffer CA, Wigerinck PBTP (2005) Discovery and selection of TMC114, a next generation HIV-1 protease inhibitor. J Med Chem 48:1813–1822
10. Ghosh AK, Kulkarni S et al (2009) Design, synthesis, protein-ligand X-ray structure, and biological evaluation of a series of novel macrocyclic human immunodeficiency virus-1 protease inhibitors to combat drug resistance. J Med Chem 52:7689–7705
11. Stewart AAJ, Andrew M (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev 6:1589–1615
12. Kollman PA (1993) Free-energy calculations—applications to chemical and biochemical phenomena. Chem Rev 7:2395–2417
13. Kollman PA, Massova I, Reyes C, Kuhn B, Huo SH, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res 33:889–897
14. Lee MR, Duan Y, Kollman PA (2000) Use of MM-PB/SA in estimating the free energies of proteins: application to native, intermediates, and unfolded villin headpiece. Proteins 39:309–316
15. Wang W, Kollman PA (2000) Free energy calculations on dimer stability of the HIV protease using molecular dynamics and a continuum solvent model. J Mol Biol 303:567–582
16. Wang W, Kollman PA (2001) Computational study of protein specificity: the molecular basis of HIV-1 protease drug resistance. Proc Natl Acad Sci USA 98:14937–14942
17. Hou TJ, Yu R (2007) Molecular dynamics and free energy studies on the wild-type and double mutant HIV-1 protease complexed with amprenavir and two amprenavir-related inhibitors: mechanism for binding and drug resistance. J Med Chem 50:1177–1188
18. Swanson JM, Henchman RH, McCammon JA (2004) Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. Biophys J 86:67–74
19. Xu Y, Wang RX (2006) A computational analysis of the binding affinities of FKBP12 inhibitors using the MM-PB/SA method. Proteins 64:1058–1068
20. Wang J, Morin P, Wang W, Kollman PA (2001) Use of MM-PBSA in reproducing the binding free energies to HIV-1RT of TIBO derivates and predicting the binding mode to HIV-1 RT of Efavirenz by docking and MM-PBSA. J Am Chem Soc 123:5221–5230
21. Kuhn B, Gerber P, Schulz-Gasch T, Stahl M (2005) Validation and use of the MM-PBSA approach for drug discovery. J Med Chem 48:4040–4048
22. Luo C, Xu L, Zheng S, Luo X, Shen J, Jiang H, Liu X, Zhou M (2005) Computational analysis of molecular basis of 1:1 interactions of NRG-1beta wild-type and variants with ErbB3 and ErbB4. Proteins 59:742–756
23. Huo S, Wang J, Cieplak P, Kollman PA, Kuntz ID (2002) Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: insight into structure-based ligand design. J Med Chem 45:1412–1419
24. Gohlke H, Kiel C, Case DA (2003) Insights into protein–protein binding by binding free energy calculation and free energy decomposition for the Ras–Raf and Ras–RalGDS complexes. J Mol Biol 330:891–913
25. Zoete V, Meuwly M, Karplus M (2005) Study of the insulin dimerization: binding free energy calculations and per-residue free energy decomposition. Proteins 61:79–93
26. Chen JZ, Zhang SL, Liu XG, Zhang QG (2010) Insights into drug resistance of mutations D30N and I50V to HIV-1 protease inhibitor TMC-114: free energy calculation and molecular dynamic simulation. J Mol Model 16:459–468
27. Massova I, Kollman PA (2000) Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. Perspect Drug Disc Des 18:113–135
28. Huo S, Massova I, Kollman PA (2002) Computational alanine scanning of the 1:1 human growth hormone-receptor complex. J Comput Chem 23:15–27
29. Massova I, Kollman PA (1999) Computational alanine scanning to probe protein–protein interactions: a novel approach to evaluate binding free energies. J Am Chem Soc 121:8133–8143
30. Lillian TC, William CS, Jed WP, Vijay SP (2006) Kinetic computational alanine scanning: application to p53 oligomerization. J Mol Biol 357:1039–1049
31. Zoete V, Meuwly M (2006) Importance of individual side chains for the stability of a protein fold: computational alanine scanning of the insulin monomer. J Comput Chem 27:1843–1857
32. Zoete V, Michielin O (2007) Comparison between computational alanine scanning and per-residue binding free energy decomposition for protein–protein association using MM-GBSA: application to the TCR–p-MHC complex. Protein 67:1026–1047
33. Chen XN, Tropsha A (1995) Relative binding free energies of peptide inhibitors of HIV-1 protease: the influence of the active site protonation state. J Med Chem 38:42–48
34. Tie Y, Boross PI, Wang YF, Gaddis L, Hussain AK, Leshchenko S, Ghosh AK, Louis JM, Harrison RW, Weber IT (2004) High resolution crystal structures of HIV-1 protease with a potent non-peptide inhibitor [UIC-94017] active against multidrug-resistant clinical strains. J Mol Biol 338:341–352
35. Kovalevsky AY, Tie Y, Liu F, Boross PI, Wang YF, Leshchenko S, Ghosh AK, Harrison RW, Weber IT (2006) Effectiveness of nonpeptide clinical inhibitor TMC-114 on HIV-1 protease with highly drug resistant mutations D30N, I50V, and L90M. J Med Chem 49:1379–1387
36. Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossváry I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, Sagui C, Babin V, Kollman PA (2008) AMBER 10. University of California, San Francisco
37. Wang JM, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general Amber force field. J Comput Chem 25:1157–1174

38. Jakalian A, Bush BL, Jack DB, Bayly CI (2000) Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. J Comput Chem 21:132–146

39. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong GM, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang JM, Kollman PA (2003) Point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24:1999–2012

40. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935

41. Darden T, York D, Pedersen L (1993) Particle mesh Ewald—an N. log-[N] method for Ewald sums in large systems. J Chem Phys 98:10089–10092

42. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of Cartesian equations of motion of a system with constraints—molecular dynamics of n-alkanes. J Comput Phys 23:327–341

43. Weiser J, Shenkin PS, Still WC (1999) Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). J Comput Chem 20:217–230

44. Onufriev A, Bashford D, Case DA (2000) Modification of the generalized Born model suitable for macromolecules. J Phys Chem B 104:3712–3720

45. Freedberg DI, Wang YX, Stahl SJ, Kaufman JD, Wingfield PT, Kiso Y, Torchia DA (1998) Flexibility and function in HIV protease: dynamics of the HIV-1 protease bound to the asymmetric inhibitor kynostatin 272 [KNI-272]. J Am Chem Soc 120:7916–7923

46. Zoete V, Michielin O, Karplus M (2002) Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. J Mol Biol 315:21–52

47. Zhu ZW, Schuster DI, Tuckerman ME (2003) Molecular dynamics study of the connection between flap closing and binding of fullerene-based inhibitors of the HIV-1 protease. Biochemistry 42:1326–1333

48. Ishima RD, Freedberg I et al (1999) Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. Structure 7:1047–1055

49. Zhang DW, Zhang JZH (2005) Full quantum mechanical study of binding of HIV-1 protease drugs. Int J Quantum Chem 103:246–257

50. Clackson T, Ultsch MH, Wells JA, de Vos AM (1998) Structural and functional analysis of the 1:1 growth hormone: receptor complex reveals the molecular basis for receptor affinity. J Mol Biol 277:1111–1128

51. Cunningham BC, Wells JA (1989) High-resolution epitope mapping of hGH–receptor interactions by alanine-scanning mutagenesis. Science 244:1081–1085

52. Li T, Froeyen M, Herdewijn P (2008) Computational alanine scanning and free energy decomposition for *E. coli* type I signal peptidase with lipopeptide inhibitor complex. J Mol Graph Model 26:813–823

J Mol Model (2012) 18:1855–1866
DOI 10.1007/s00894-011-1204-3

ORIGINAL PAPER

In silico quest for putative drug targets in *Helicobacter pylori* HPAG1: molecular modeling of candidate enzymes from lipopolysaccharide biosynthesis pathway

Munmun Sarkar · Lakshmi Maganti · Nanda Ghoshal · Chitra Dutta

Received: 1 April 2011 / Accepted: 26 July 2011 / Published online: 18 August 2011

Abstract Aimed at identification and structural characterization of novel putative therapeutic targets in *H. pylori*, the etiological agent of numerous gastrointestinal diseases including peptic ulcer and gastric cancer, the present study comprised of three phases. First, through subtractive analysis of metabolic pathways of *Helicobacter pylori* HPAG1 and human, as documented in the KEGG database, 11 pathogen-specific pathways were identified. Next, all proteins involved in these pathogen-specific pathways were scrutinized in search of promising targets and the study yielded 25 candidate target proteins that are likely to be essential for the pathogen viability, but have no homolog in human. The lipopolysaccharide (LPS) biosynthesis pathway was found to be the largest contributor (nine proteins) to this list of candidate proteins. Considering the importance of LPS in *H. pylori* virulence, 3D structural models of three predicted target enzymes of this pathway, namely 2-dehydro-3-deoxy-phosphooctonate aldolase, UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase and Phosphoheptose isomerase, were then built up using the homology modeling approaches. Binding site analysis and docking of the known biological substrate PEP to 2-dehydro-3-deoxyphosphooctonate aldolase revealed the potential binding pocket present in the single monomeric form of the enzyme and identified 11 amino acid residues that might play the key roles in this protein-ligand interaction.

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1204-3) contains supplementary material, which is available to authorized users.

M. Sarkar · L. Maganti · N. Ghoshal · C. Dutta (✉)
Structural Biology & Bioinformatics Division,
Indian Institute of Chemical Biology (a unit of CSIR),
4, Raja S. C. Mullick Road,
Kolkata 700032, India
e-mail: cdutta@iicb.res.in

Keywords Binding site analysis · Differential pathway analysis · Essential proteins · Homology modeling · Molecular docking

Introduction

Helicobacter pylori is the first formally recognized bacterial carcinogen [1] and a major cause of various gastrointestinal diseases, ranging from chronic active gastritis without clinical symptoms to peptic ulceration, gastric adenocarcinoma, and gastric mucosa-associated lymphoid tissue (MALT) lymphoma [2–4]. More than 50% of the human population harbors this gram-negative microaerophilic microbe in their stomach [5], making it the most widespread infection in the world and the infection persists for life, if left untreated [3]. Multiple drug regimens have been proposed for the initial treatment of *H. pylori* infection [6–9]. However, rising antibiotic resistance and various side-effects of existing intervention strategies [10, 11] have increased the need of development of anti- *H. pylori* drugs and the crucial first step in designing a new drug is to identify one or more new therapeutic targets. To this end, the present endeavor attempts, through in silico approaches, to identify novel potential drug targets in *H. pylori* HPAG1. Among 20 different strains of *H. pylori*, for which complete genome sequences were available in the public domain at the time of initiation of the study, *H. pylori* HPAG1 was specifically selected for this analysis, since it was a clinical isolate from a patient with chronic atrophic gastritis, the precursor to gastric adenocarcinoma [12].

There are many computational approaches to identify potential targets such as identification of virulence genes or pathogen–specific essential genes, characterization of pathogen-specific unique metabolic pathways, elucidation of membrane localized drug targets etc. [13–16]. In most of these approaches, a set of candidate pathogen gene-products are identified on the basis of two major criteria- essentiality and selectivity [15–18]. A target protein must be indispensible for the growth, replication, viability or survival of the pathogen, but it should not have any homolog in the host genome. This would ensure that the inhibition of the identified targets would be detrimental to the pathogen, but have no undesired cross-reactivity with the host proteins. The present study employed the subtractive metabolic pathway strategy [15] for identification of pathogen-specific pathways and then performed the subtractive genome analysis [16] to sort out the enzymes present exclusively in the pathogen, but not in the host. Twenty five promising targets, participating in 11 pathogen-specific pathways have been identified - the lipopolysaccharide (LPS) biosynthesis pathway being the largest contributor to this list of candidate proteins. In the subsequent phase of the study, homology modeling of the candidate enzymes from the LPS biosynthesis pathway has been taken up depending upon the availability of the template structures. An attempt has also been made to predict the putative ligand binding cavities in the enzymes under study and to identify the key amino acid residues, potentially involved in the enzyme-substrate interactions.

Materials and methods

Figure 1 represents a flowchart of the work carried out in the present study. As shown in the flowchart, all annotated protein sequences of *Helicobacter pylori* HPAG1 were downloaded from National Center for Biotechnology Information (NCBI) [19]. All metabolic pathway identification numbers of *H. pylori* HPAG1 and *Homo sapiens* were also extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [20].

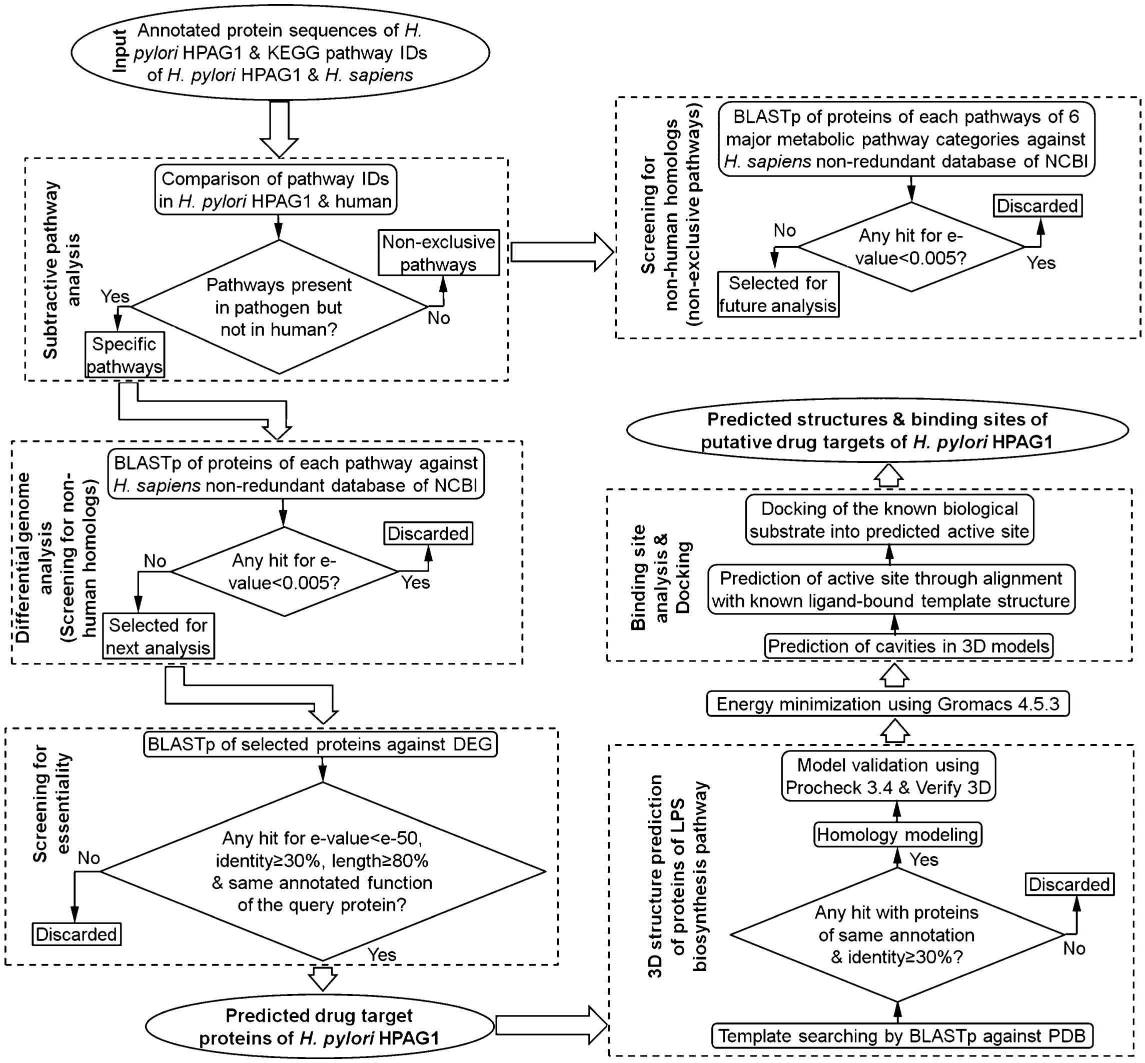

Input
Annotated protein sequences of H. pylori HPAG1 & KEGG pathway IDs of H. pylori HPAG1 & H. sapiens
Subtractive pathway analysis
Comparison of pathway IDs in H. pylori HPAG1 & human
Pathways present in pathogen but not in human?
Yes
No
Specific pathways
Non-exclusive pathways
Screening for non-human homologs (non-exclusive pathways)
BLASTp of proteins of each pathways of 6 major metabolic pathway categories against H. sapiens non-redundant database of NCBI
Any hit for e-value<0.005?
No
Yes
Discarded
Selected for future analysis
Differential genome analysis (Screening for non-human homologs)
BLASTp of proteins of each pathway against H. sapiens non-redundant database of NCBI
Any hit for e-value<0.005?
No
Yes
Discarded
Selected for next analysis
Screening for essentiality
BLASTp of selected proteins against DEG
Any hit for e-value<e-50, identity≥30%, length≥80% & same annotated function of the query protein?
No
Yes
Discarded
Predicted drug target proteins of H. pylori HPAG1
3D structure prediction of proteins of LPS biosynthesis pathway
Template searching by BLASTp against PDB
Any hit with proteins of same annotation & identity≥30%?
Yes
No
Discarded
Homology modeling
Model validation using Procheck 3.4 & Verify 3D
Energy minimization using Gromacs 4.5.3
Binding site analysis & Docking
Prediction of cavities in 3D models
Prediction of active site through alignment with known ligand-bound template structure
Docking of the known biological substrate into predicted active site
Predicted structures & binding sites of putative drug targets of H. pylori HPAG1

Fig. 1 Flowchart of the whole work

Subtractive pathway analysis

ID nos. of the pathways of *H. pylori* HPAG1 were compared with those of human. The pathways not found in human but present in *H. pylori* HPAG1 (according to KEGG annotation) were marked as the pathogen-specific pathways (Table 1) and proteins involved in these pathways were selected for downstream analysis.

Differential genome analysis

All proteins of the pathways enlisted in Table 1 were subjected to BLASTp [21] search against *H. sapiens* non-redundant database available at NCBI and sequences exhibiting no hit for e-value <0.005 were selected for further study.

Screening for essential proteins

BLASTp search was carried out individually for each of the selected *H. pylori* HPAG1 proteins against the Database of Essential Genes (DEG) [22], following the criteria shown in Fig. 1. If the annotated function of a query protein be same as that of its BLASTp hit in DEG, then the protein was considered as essential gene-product of *H. pylori* HPAG1. All non-human-homologous essential *H. pylori* proteins, identified this way, were considered as the potential therapeutic targets.

In an earlier study on *H. pylori* [18], the "essentiality" of the identified targets was assessed merely on the basis of their orthology to the DEG members. In the present study, the method of subtractive pathway analysis [15] was employed to find out *H. pylori*-specific pathways and then the proteins participating only in these pathways were screened for essential proteins with no human homologs.

Table 1 List of the metabolic pathways present in *H. pylori* HPAG1 but absent in Human

KEGG pathway ID	Metabolic pathway
00473	D-Alanine metabolism
00550	Peptidoglycan biosynthesis
00540	Lipopolysaccharide biosynthesis
00628	Fluorene degradation
00362	Benzoate degradation via hydroxylation
03090	Type II secretion system
03070	Type III secretion system
03080	Type IV secretion system
02020	Two-component system
02030	Bacterial chemotaxis
02040	Flagellar assembly

Prediction of membrane-localized targets

All the identified potential drug targets were submitted to the TMHMM [23] server 2.0 to identify the putative transmembrane helices, if any, in these proteins.

Prediction of potential 3D structures of candidate target proteins

The next step in the study was prediction of the potential tertiary structures of the identified target proteins using the approach of homology modeling. It was, however, not feasible to take up the task of homology modeling of a total of 25 identified target proteins in a single endeavor. Therefore, the present study focused only on the identified target proteins of the LPS biosynthesis pathway - the largest contributor to the list of identified candidate targets (Table 2).

Template selection

All candidate target proteins of the LPS biosynthesis pathways of *H. pylori* HPAG1 were individually subjected to BLASTp search against PDB [24]. A BLASTp hit was selected as a suitable template, if it had the same annotation as the query protein and they both shared an identity of at least 30%. However, following these criteria, suitable templates for homology modeling could be found only for three candidate enzymes of LPS biosynthesis pathway, namely 2-dehydro-3-deoxy-phosphooctonate aldolase (HPAG1_0003), UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase (HPAG1_0395) and Phosphoheptose isomerase (HPAG1_0840). Homology modeling was, therefore, carried out only for these three enzymes. Details of the template proteins selected for three query proteins are given in Table 3.

Model building

Commercial software package Discovery Studio 2.1 from Accelrys [25] was used for modeling, energy minimization, binding site analysis and docking purposes. To build up the models of the three proteins under study, the following procedure was followed. All templates for a specific target protein were aligned using the "align structures (MODELER)" module with gap open penalty set to 0.0 and gap extension penalty set to 2.0. Then the multiple alignment output was aligned with the target protein sequence, using the "align sequence with structure" module (scoring matrix: as1; gap open penalty: -900; gap extension penalty: -50). The final alignment was checked manually for further refinement. Models were built using the module "build homology models" with optimization level set to medium and cut

Table 2 List of essential gene-products from pathogen-specific pathways having no human homolog

Metabolic pathway (KEGG ID)	Product name	Length (aa)	Locus_tag	COG ID
Lipopolysaccharide biosynthesis (00540)	2-dehydro-3-deoxyphosphooctonate aldolase	276	HPAG1_0003	COG2877M
	3-deoxy-D-manno-octulosonic-acid transferase	393	HPAG1_0941	COG1519M
	ADP-heptose–LPS heptosyltransferase II	349	HPAG1_1132	COG0859M
	Hypothetical protein HPAG1_0843	173	HPAG1_0843	COG0241E
	Lipopolysaccharide heptosyltransferase-1	340	HPAG1_0281	COG0859M
	Phosphoheptose isomerase	192	HPAG1_0840	COG0279G
	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase	336	HPAG1_0190	COG1044M
	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase	295	HPAG1_0395	COG0774M
	UDP-N-acetylglucosamine acyltransferase	270	HPAG1_1321	COG1043M
Two-component system (02020)	Anthranilate phosphoribosyltransferase	335	HPAG1_1232	COG0547E
	Anthranilate synthase component 1	500	HPAG1_1230	COG0147EH
	C(4)-dicarboxylates and tricarboxylates/succinate antiporter	482	HPAG1_0141	COG0471P
	carbon storage regulator	76	HPAG1_1368	COG1551T
	Indole-3-glycerol phosphate synthase	181	HPAG1_0278	COG0134E
	Short-chain fatty acids transporter	454	HPAG1_0678	COG2031I
	Tryptophan synthase alpha chain	262	HPAG1_1235	COG0159E
	Tryptophan synthase beta chain	393	HPAG1_1234	COG0133E
Peptidoglycan biosynthesis (00550)	UDP-MurNac-pentapeptide presynthetase	493	HPAG1_0724	COG0770M
	UDP-N-acetylmuramate–L-alanine ligase	449	HPAG1_0606	COG0773M
	UDP-N-acetylmuramoylalanyl-D-glutamate–2, 6-diaminopimelate ligase	447	HPAG1_1419	COG0769M
	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase	422	HPAG1_0470	COG0771M
D-Alanine metabolism (00473)	Alanine racemase, biosynthetic	377	HPAG1_0924	
Type II secretion system (03090)	Hypothetical protein HPAG1_1440	191	HPAG1_1440	-
Peptidoglycan biosynthesis & D-Alanine metabolism (00550 & 00473)	D-alanyl-alanine synthetase A	347	HPAG1_0722	COG1181M
Two-component system & Bacterial chemotaxis (02020 & 02030)	Chemotaxis protein	124	HPAG1_0380	COG0784T

overhangs set to true. In each trial, 50 models were built with an option available. The models having least the discrete optimized protein energy (DOPE) score and lowest PDF energy were selected as the best model. Further correction of models in the loop regions were done using the “loop refinement (MODELER)” module with optimization level set to medium. Finally the side chain was adjusted using the “side-chain refinement” module applying CHARMm force field. Images of final models were generated using Discover studio 2.1 and the surface view was generated using UCSF Chimera [26] for the visualization of 3D models.

Model validation

To check the stereochemical quality as well as the overall and residue-by-residue geometry of the models, Ramachandran plots of the protein models were built using Procheck 3.4 [27]. The reliability of the predicted models were further assessed using Verify3D [28]. Models qualifying in all these tests were selected for further study.

Energy minimization of modeled proteins

Energy minimization was carried out in order to remove or reduce possible geometric problems in the bimolecular systems, such as improbable bond distances, bond angles and torsion angles. In the present study, energy minimization for all three modeled proteins was performed with GROMACS 4.5.3 software package [29] using the GROMOS96 43 a2 force field [30], with cubic cell geometry. The default simple point charge (SPC) water was added to the box and periodic boundary condition was applied. The distance between the grid box and the protein was set to 1.0 nm. In order to neutralize the total charge of the system counter ions were placed in the box. Energy minimization was carried out initially by steepest descent

Table 3 Details of templates used in Homology modeling

Protein		Template					
Name	Length (AA)	PDB id	Name	Length (AA)	Organism	Identity	Resolution
2-dehydro-3-deoxyphosphooctonate aldolase	276	1fx6	2-dehydro-3-deoxyphosphooctonate aldolase	267	*Aquifex aeolicus*	49%	2.06 Å
		1o60	2-dehydro-3-deoxyphosphooctonate aldolase	332	*Haemophilus influenzae*	43%	1.80 Å
		3fs2	2-dehydro-3-deoxyphosphooctonate aldolase	298	*Brucella melitensis*	46%	1.85 Å
Phosphoheptose isomerase	192	1tk9	Phosphoheptose isomerase 1	208	*Campylobacter jejuni*	61%	2.10 Å
		3bjz	Phosphoheptose isomerase	219	*Pseudomonas aeruginosa pao1*	50%	2.40 Å
		2i2w	Phosphoheptose isomerase	212	*Escherichia coli*	46%	1.95 Å
		1x94	putative Phosphoheptose isomerase	191	*Vibrio cholerae*	48%	2.50 Å
UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase	295	2ves	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase	299	*Pseudomonas aeruginosa*	44%	1.90 Å
		1yh8	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase	270	*Aquifex aeolicus*	36%	2.70 Å

method followed by conjugate gradient. Then all three systems were subjected to MD simulations for 5 ns. The trajectory stability was monitored by the analysis of energy (supplementary Fig. 10) and the backbone RMSD (supplementary Fig. 11) as a function of time for the three structures. Temperature plot showing the normal oscillation behavior of the temperature about the desired average (300 K) for all three structures are also given in supplementary information (supplementary Fig. 12).

Binding site analysis

In an attempt to identify the putative active sites of the modeled enzymes, their potential ligand binding sites (cavity) were identified using “find sites from receptor cavities” tool of Discovery Studio 2.1 [25], selecting the protein as receptor. The grid resolution and site opening were set to 0.5 Å and 5.00 Å respectively and the minimum site size was set to 100 grid points of solvent accessible surface. The proteins were then superimposed with their individual templates, having co-crystal substrate bound structure and their active sites were compared with the identified binding sites to find out the possible active site of the individual proteins.

Docking

To identify the active binding site of the enzyme 2-dehydro-3-deoxy-phosphooctonate aldolase and to predict the residues involved in the enzyme-substrate interactions, the biological substrate (PEP) of the enzyme was docked into the enzyme. The initial structure of the ligand was drawn in a new 3D window and a good 3D conformer was generated using the clean operation. Finally the ligand was optimized using “dreiding minimization” module (no. of iteration: 500) of Discovery Studio 2.1 [25]. The fully optimized ligand was docked into the respective predefined binding cavity using “Dock Ligands (LigandFit)” module, using a Monte Carlo docking protocol. This methodology allows for fully positional and conformational flexibility of the ligand inside the active binding cavity to generate different poses. In the docking study, ten poses were generated and the best pose was selected on the basis of various scoring functions (Ligscore1, Ligscore2, PLP1, PLP2, Jain, PMF) denoting the energy of interaction. The best geometry of ligand-receptor complex was subjected to energy minimization applying CHARMm force field and the interaction pattern was studied.

Proteins from non-exclusive pathways of the pathogen with no human homologs

Proteins involved in major metabolic pathways of *H. pylori* HPAG1 that are not pathogen-specific (also found in human), such as carbohydrate metabolism, energy metabolism, lipid metabolism, nucleotide metabolism, amino acid metabolism and metabolism of cofactors and vitamins, were also subjected to BLASTp search against *H. sapiens* non-redundant database available at NCBI. Proteins which do not have any hit below e-value 0.005 were cataloged in the supplementary Table. This list provides a shortlist of the *H. pylori* HPAG1 proteins that may be subjected in future to further screening in search of suitable drug targets.

Results and discussion

Identification of candidate target proteins in *H. pylori* HPAG1

Comparison of all annotated metabolic pathways in *H. pylori* HPAG1 with those in human, as documented in the KEGG database, revealed 11 pathogen-specific pathways, i.e., the pathways present in *H. pylori* HPAG1 but absent in human (Table 1). Proteins involved in these pathways were sorted out using the differential genome approach (Fig. 1). The shortlisted proteins (having no human homolog) were further screened for their essentiality and finally 25 *H. pylori* HPAG1 proteins (Table 2) satisfied all pre-requisites for promising drug-targets - they participate in pathogen-specific metabolic pathways, are likely to be essential for the pathogen but have no ortholog in the host. Hence, these 25 proteins may serve as candidate drug targets.

As can be seen from Table 2, among 11 pathogen-specific pathways identified in the present study, only six pathways contribute to the set of potential target components. The LPS biosynthesis pathway is the largest contributor to such candidate proteins – nine out of 25 potential targets belong exclusively to this pathway. The two-component system is the second largest contributor with eight candidate components exclusively belonging to this pathway and one candidate protein – a Chemotaxis protein - shared with bacterial chemotaxis. The peptidoglycan biosynthesis and D-alanine metabolism pathways, which exclusively possess four and one candidate proteins respectively, mutually share the candidate enzyme D-alanyl-alanine synthetase A and the type II secretion system offers a hypothetical protein as a potential target.

When these 25 proteins were subjected to TMHMM server, four proteins were found to have predicted trans-membrane helices, suggesting that these four proteins may be membrane-bound and hence, may serve as membrane localized drug targets. These proteins are C(4)-dicarboxylates & tricarboxylates/succinate antiporter, short-chain fatty acids transporter, 3-deoxy-D-manno-octulosonic-acid transferase, and UDP-MurNac-pentapeptide presynthetase.

Homology modeling of three candidate targets from the LPS biosysnthesis pathway

As already mentioned, the largest contributor to the list of identified candidate targets (Table 2) is the LPS biosynthesis pathway. Like all other gram negative bacteria, *H. pylori* also contains LPS as a major component of its outer membrane and there is an increasing body of evidence indicating that *H. pylori* LPS, owing to certain unique attributes, plays a vital role in host colonization and inflammatory response [31, 32]. The relatively low endotoxic activity of *H. pylori* LPS as compared to other bacteria facilitates persistence of inflammation in the gastric mucosa [31], while its O-specific chain that mimics Lewis blood group antigens in structure may signal the host immune system to down-regulate an inflammatory response [33], thereby aiding in effective colonization of host tissues. This molecular mimicry may also be instrumental in development of gastric autoimmunity [31, 34], leading to atrophic gastritis [35, 36] and likely contribute even to gastric lymphoma [4]. The core oligosaccharide of the LPS mediates the binding of the bacterium to laminin, and interferes with gastric cell receptor-laminin interaction, which, in turn, may trigger or exacerbate the mucosal degeneration [32].

Template searching

The observation that nine out of 25 candidate target proteins belong to LPS biosynthesis pathway and the fact that LPS is one of the key contributor to *H. pylori* virulence have prompted us to take up the task of homology modeling of the candidate target enzymes of this pathway. Among these nine candidate proteins, 3D structure of only one enzyme, UDP-N-acetylglucosamine acyltransferase, from another *H. pylori* strain was known experimentally (PDB id: 1j2z) at the time of initiation of this study. The other eight potential target proteins of this pathway were, therefore, subjected to BLASTp search against PDB for identification of templates for structural modeling. However, following the criteria described in the Materials & methods section, suitable templates could be found only for three candidate enzymes - 2-dehydro-3-deoxy-phosphooctonate aldolase (HPAG1_0003), UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase (HPAG1_0395), and Phospho-heptose isomerase (HPAG1_0840). In the present study, therefore, homology modeling was carried out for these three enzymes of *H. pylori* HPAG1.

Among the several X-ray crystal structures of the orthologs of these three proteins from different organisms, suitable templates for homology modeling were selected on the basis of the sequence identity of the orthologs with the target proteins and the quality of their alignment. Table 3 shows the detail description of the template sequences selected for three candidate proteins under consideration.

Homology models & their validation

2-dehydro-3-deoxy-phosphooctonate aldolase (HPAG1_0003) (EC: 2.5.1.55, formerly 4.1.2.16) This enzyme converts phosphoenolpyruvate (PEP) to 2-dehydro-3-deoxy-D-octonate 8-phosphate, which is a major step in the LPS biosynthesis. A recent study involving mutagenesis of its orthologous enzyme demonstrated the essentiality of this enzyme in *Pseudomonas*

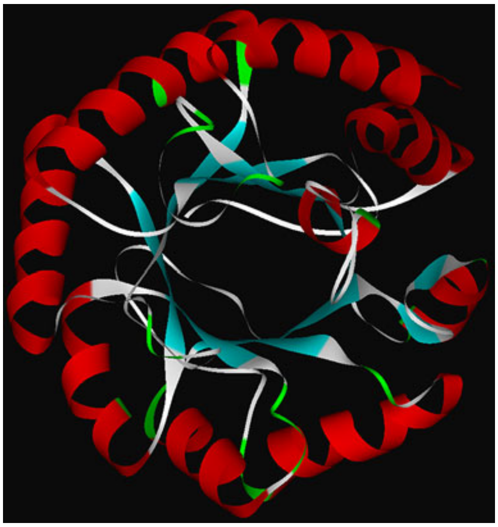


Fig. 2 Ribbon schematic representation of the homology model of 2-dehydro-3-deoxyphosphooctonate aldolase of *H. pylori* HPAG1. α-helices, β-strands, random loops and turns are represented as red, cyan blue, white and green respectively. Image is generated using Discovery studio visualizer

aeruginosa PAO1 [37]. The final model structure predicted for 2-dehydro-3-deoxy-phosphooctonate aldolase of *H. pylori* HPAG1 (Fig. 2) has eight β-strands and ten α-helices. The reliability of the model has been assured by its Ramachandran plot (Fig. 3a) showing 93% residues in the most favored region, 6.6% in additional allowed region, 0.4% in generously allowed region and 0% in disallowed region, as well as by the overall G-factor (0.0) that lies within the limit of the favored region for a valid 3D protein model. The reliability of the predicted model has also been substantiated by the assessment with Verify3D that shows 84.12% of the residues with an average 3D-1D score >0.2 (Fig. 4).

After energy minimization of the modeled structure the binding site analysis of the model has identified 12 possible ligand-binding cavities. In order to identify the potential ligand-binding site among these 12 cavities, the modeled structure was aligned with the ligand (PEP) bound structures of 2-dehydro-3-deoxy-phosphooctonate aldolase of *Aquifex aeolicus* (PDB id: 1fwn) and *Escherichia coli* (PDB id: 1q3n). The ligand-bound structure of *A. aeolicus* enzyme was selected as its ligand-free form served as one of the template during homology modeling of the target protein, while the ligand-bound structure of the *E. coli* ortholog was used as a control (i.e., its ligand-free form was not used as template). The largest cavity with grid point volume of 296.375 appeared to be the potential ligand binding pocket

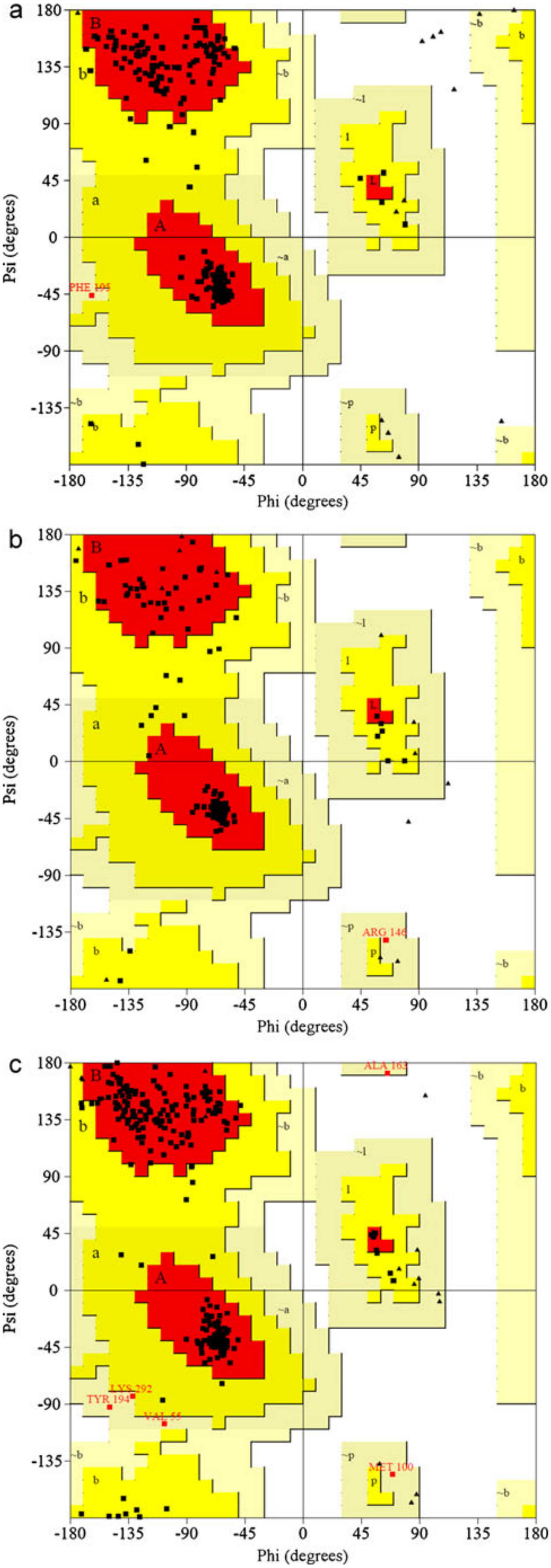


Fig. 3 Ramachandran plot obtained for the models of (**a**) 2-dehydro-3-deoxyphosphooctonate aldolase, (**b**) Phosphoheptose isomerase and (**c**) UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase. The most favored regions, additional allowed regions and generously allowed regions are represented by red, yellow and cream colors respectively

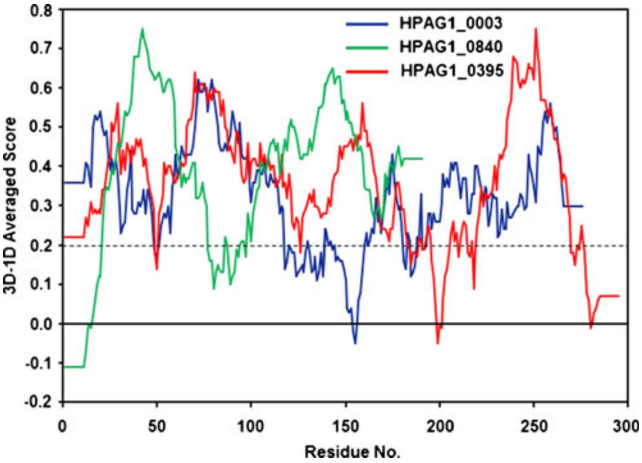


Fig. 4 Verify 3D plots of the three modeled structures built up in this study. In all three cases, more than 80% residues exhibit 3D-1D scores >0.2, endorsing the reliability of the models

of *H. pylori* HPAG1 2-dehydro-3-deoxy-phosphooctonate aldolase (Fig. 5). The predicted binding site is situated at the center of the protein at the opposite side of the two terminals of the amino acid chain and is surrounded by the β-strands (supplementary Fig. 2) and it is in good agreement with the reported binding site of its *A. aeolicus* ortholog (supplementary Fig. 3) [38].

It is worth mentioning at this point that in all the templates under study, the enzyme is found to be present in tetramer form, but the predicted binding sites remain the same as in their monomeric forms. So it was possible to study the protein-ligand interaction in a single monomer.

In the docking study using the substrate PEP (supplementary Fig. 1), 11 amino acid residues of *H. pylori* HPAG1 2-dehydro-3-deoxy-phosphooctonate aldolase were identified as the key residues, potentially involved in the protein-ligand interaction. The predicted protein-PEP complex is shown in Fig. 6. The identified key residues were Cys18, Ser49, Lys52, Asp87, Pro107, Lys130, Arg173, His204, Phe239 and Glu241 Asp252. In the docking study, PEP was found to interact with these amino acid residues and it formed a single H-bond of 1.1 Å length with Asp87, two H-bonds of 1.8 Å and 2.4 Å length with Lys130 and another single H-bond of 1.3 Å length with Glu241 amino acid residue.

These 11 amino acid residues were found to be conserved in all the templates used for building the model except for Cys18 which is replaced in *Haemophilus influenza* with Asn (Fig. 7). It is worth mentioning at this point that the reported ligand (PEP) bound structure of one of the template, i.e., 2-dehydro-3-deoxyphosphooctonate aldolase of *A. aeolicus* (PDB id: 1fwn) identified 14 amino acid residues to be involved in the protein ligand interaction [36]. Out of these 14 key residues, eight residues are identical to those predicted as the key residues in the docking study reported above (supplementary Fig. 4 and text).

Phosphoheptose isomerase (HPAG1_0840) (EC: 5.3.1.28) It converts D-sedoheptulose 7-phosphate to D-glycero-D-manno-heptose 7-phosphate [39] – another vital step of LPS biosynthesis. The final model structure of the protein, as shown in Fig. 8, has five β-strands and seven α-helices. The Ramachandran plot of the predicted model (Fig. 3b) with 90.1% residues in most favored region, 9.3% in additional allowed region, 0.6% in generously allowed region, and 0% in disallowed region endorses the high quality of the model. Overall G-factor is found to be 0.0, which represents the value of the favored region for an acceptable 3D protein model. The model structure also passes the study of compatibility of 3D-1D by Verify 3D that shows 81.25% of the residues having an average 3D-1D score >0.2 (Fig. 4).

In all templates the enzyme was found to be present as a tetramer. The observation of ligand (D-sedoheptulose 7-phosphate) bound conformation of the enzyme in *E. coli* (PDP id: 2i22) revealed that the binding site was shared by

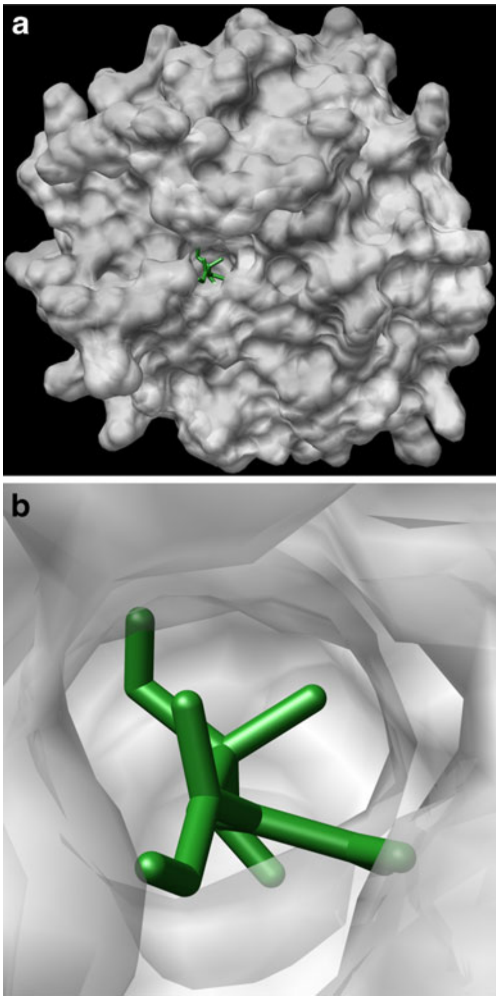


Fig. 5 (**a**) Surface view of the 3D structure of 2-dehydro-3-deoxyphosphooctonate aldolase of *H. pylori* HPAG1. Docked ligand is shown in green, (**b**) Closer view of the ligand binding cavity. Image is generated using UCSF Chimera

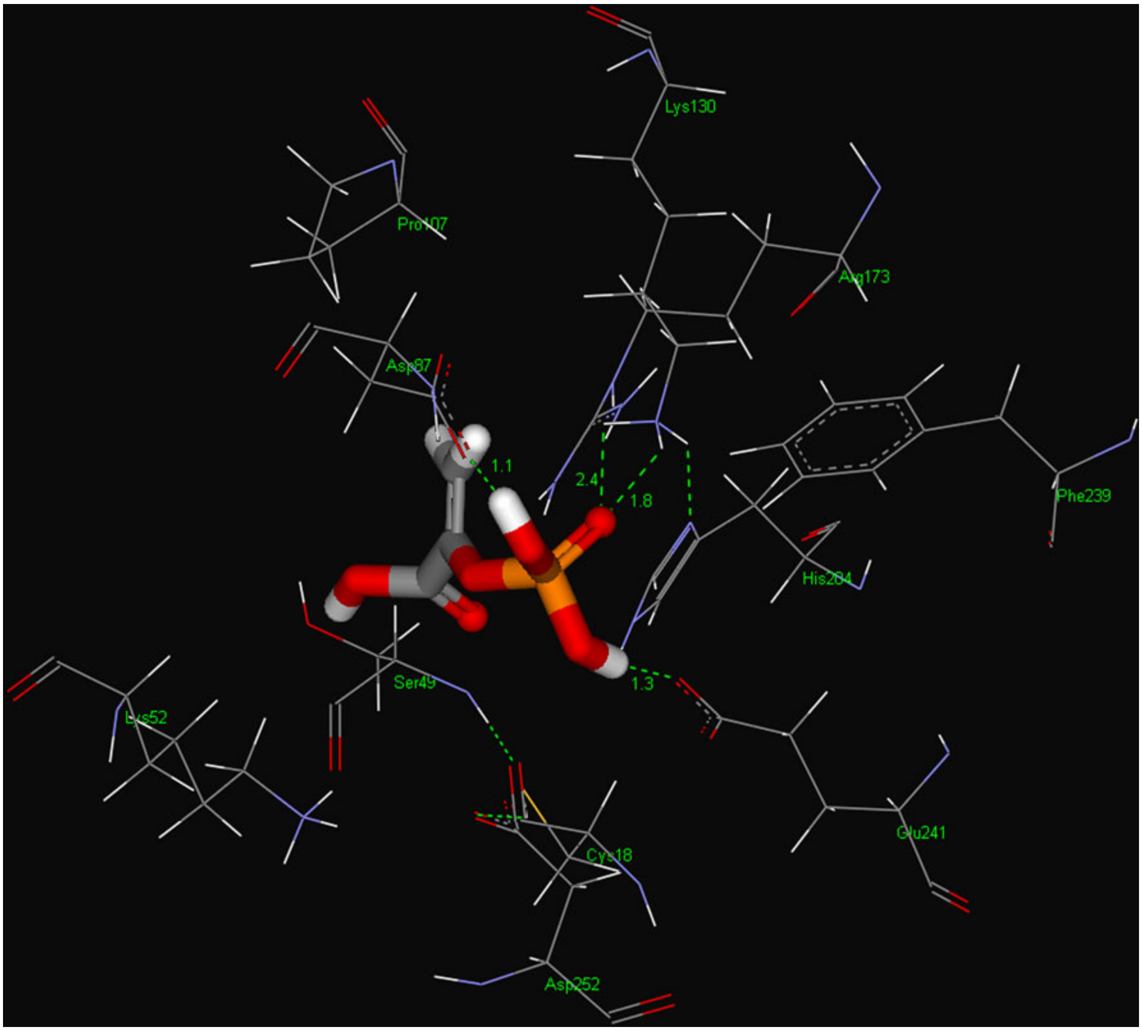


Fig. 6 Docked complex of the substrate phosphoenolpyruvate (PEP) in the active site of 2-dehydro-3-deoxyphosphooctonate aldolase of *H. pylori* HPAG1. Green dotted lines represent H-bonds. Image is generated using Discovery studio visualizer

two monomers, suggesting that the ligand binds to a cavity produced during dimerization of the protein. It was, therefore, not possible to identify the key amino acid residues contributing to in the protein-ligand interaction by simple docking studies on the modeled monomer structure of *H. pylori* HPAG1 Phosphoheptose isomerase.

UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase (HPAG1_0395) (EC: 3.5.1.108, formerly EC 3.5.1.n1) This enzyme not only participates in the LPS biosynthesis pathway but also in glycan metabolism and biosynthesis. It is a zinc-dependent enzyme that catalyzes the deacetylation of UDP-3-O-(3-hydroxytetradecanoyl)-N-acetylglucosamine to form UDP-3-O-(3-hydroxytetradecanoyl)-glucosamine and acetate [40] - a committed step in the biosynthesis of lipid A, which helps in anchoring LPS into membrane. Among the several X-ray crystal structures available from different organisms of the target protein, two orthologs from *P. aeruginosa* and *A. aeolicus* (Table 3) were selected as templates for homology modeling of this protein on the basis of their sequence identity and quality of optimal alignment with the query enzyme. Figure 9 represents the final model of the 3D structure of *H. pylori* HPAG1 UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase, which is rich in β-strands. The torsion angles of 88.6% of residues were within the most favored region and 1.9% in generously allowed region of the Ramachandran plot (Fig. 3c) and there was no residue in the disallowed region, assuring high quality of the predicted structure. Acceptability of the model was further endorsed by the overall G-factor (−0.1) that lies within the realm of the favored region of a 3D protein model and the average 3D-1D score >0.2 for 83.45% of the residues in 3D-1D compatibility test by Verify3D (Fig. 4).

As it is a zinc dependent enzyme where the metal zinc is also involved in the protein-ligand interaction, the interaction should not be studied by simple docking studies on the protein alone. The co-crystal ligand bound form of any of its selected templates was also not available, so the

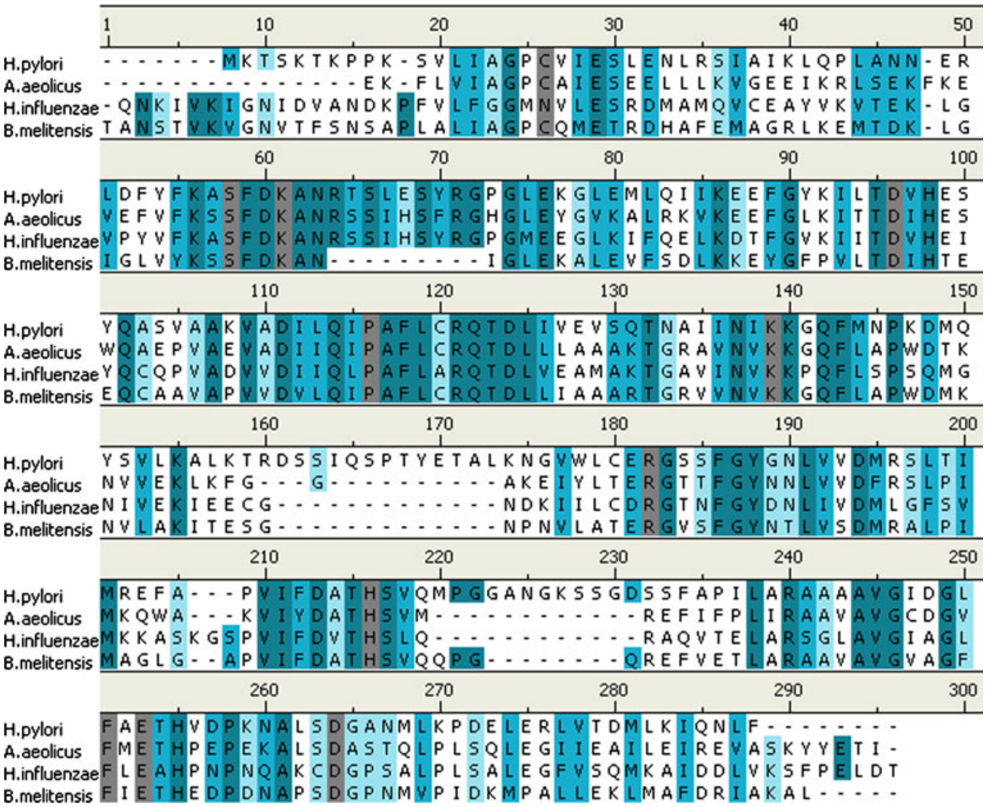

1 10 20 30 40 50
H.pylori -------MKTSKTKPPK-SVLIAGPCVIESLENLRSIAIKLQPLANN-ER
A.aeolicus ----------------EK-FLVIAGPCAIESEELLLKVGEEIKRLSEKFKE
H.influenzae -QNKIVKIGNIDVANDKPFVLFGGMNVLESRDMAMQVCEAYVKVTEK-LG
B.melitensis TANSTVKVGNVTFSNSAPLALIAGPCQMETRDHAFEMAGRLKEMTDK-LG
60 70 80 90 100
H.pylori LDFYFKASFDKANRTSLESYRGPGLEKGLEMLQIIKEEFGYKILTDVHES
A.aeolicus VEFVFKSSFDKANRSSIHSFRGHGLEYGVKALRKVKEEFGLKITTDIHES
H.influenzae VPYVFKASFDKANRSSIHSYRGPGMEEGLKIFQELKDTFGVKIITDVHEI
B.melitensis IGLVYKSSFDKAN---------IGLEKALEVFSDLKKEYGFPVLTDIHTE
110 120 130 140 150
H.pylori YQASVAAKVADILQIPAFLCRQTDLIVEVSQTNAIINIKKGQFMNPKDMQ
A.aeolicus WQAEPVAEVADIIQIPAFLCRQTDLLLAAAKTGRAVNVKKGQFLAPWDTK
H.influenzae YQCQPVADVVDIIQLPAFLARQTDLVEAMAKTGAVINVKKPQFLSPSQMG
B.melitensis EQCAAVAPVVDVLQIPAFLCRQTDLLIAAARTGRVVNVKKGQFLAPWDMK
160 170 180 190 200
H.pylori YSVLKALKTRDSSIQSPTYETALKNGVWLCERGSSFGYGNLVVDMRSLTI
A.aeolicus NVVEKLKFG---G-----------AKEIYLTERGTTFGYNNLVVDFRSLPI
H.influenzae NIVEKIEECG-------------NDKIILCDRGTNFGYDNLIVDMLGFSV
B.melitensis NVLAKITESG-------------NPNVLATERGVSFGYNTLVSDMRALPI
210 220 230 240 250
H.pylori MREFA---PVIFDATHSVQMPGGANGKSSGDSSFAPILARAAAVGIDGL
A.aeolicus MKQWA---KVIYDATHSVM------------REFIFPLIRAAVAVGCDGV
H.influenzae MKKASKGSPVIFDVTHSLQ-----------RAQVTELARSGLAVGIAGL
B.melitensis MAGLG--APVIFDATHSVQQPG--------QREFVETLARAAVAVGVAGF
260 270 280 290 300
H.pylori FAETHVDPKNALSDGANMLKPDELERLVTDMLKIQNLF--------
A.aeolicus FMETHPEPEKALSDASTQLPLSQLEGIIEAILEIREVASKYYETI-
H.influenzae FLEAHPNPNQAKCDGPSALPLSALEGFVSQMKAIDDLVKSFPELDT
B.melitensis FIETHEDPDNAPSDGPNMVPIDKMPALLEKLMAFDRIAKAL-----

Fig. 7 Sequence alignment of 2-dehydro-3-deoxyphosphooctonate aldolase of *H. pylori* HPAG1 with template proteins from *Aquifex aeolicus* (PDB id: 1fx6), *Haemophilus influenza* (PDB id: 1o60) and *Brucella melitensis* (PDB id: 3fs2). Identical matches, strong matches and weak matches are represented by dark cyan blue, deep sky blue and sky blue background respectively. The active amino acid residues for *H. pylori* HPAG1 are represented by gray background

prediction of active binding site of this enzyme was not possible.

It should be mentioned in this context that sequences of all three candidate target proteins selected for modeling in this study are highly conserved (~ 80% identity and >90% similarity) and the reported 15 active amino acid residues of 2-dehydro-3-deoxy-phosphooctonate aldolase are also conserved across all *H. pylori* strains (supplementary Figs. 7–9 show multiple alignment of these proteins with all *H. pylori* orthologs reported so far).

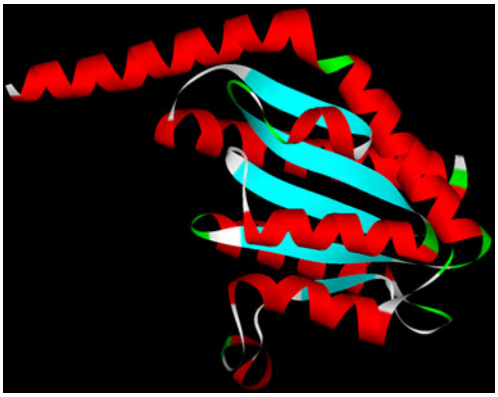


Fig. 8 Ribbon schematic representation of the modeled 3D structure of enzyme Phosphoheptose isomerase of *H. pylori* HPAG1. α-helices, β-strands, random loops and turns are represented as red, cyan blue, white and green respectively. Image is generated using Discovery studio visualizer

Conclusions

The present study aimed at identification of novel putative targets for therapeutic intervention in *H. pylori* HPAG1

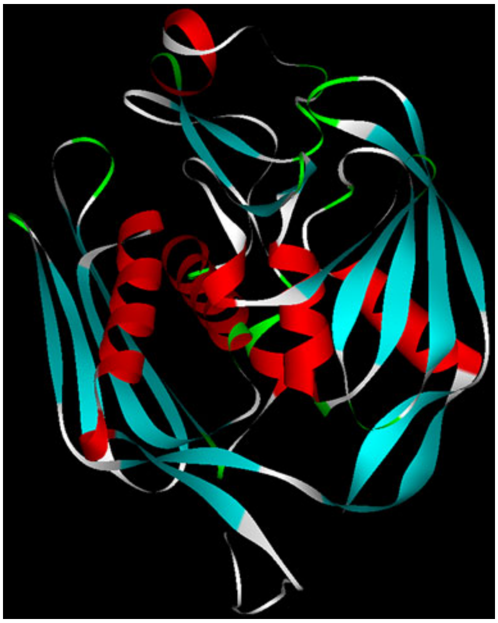


Fig. 9 Ribbon schematic representation of the 3D structure of enzyme UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase of *H. pylori* HPAG1. α-helices, β-strands, random loops and turns are represented as red, cyan blue, white and green respectively. Image is generated using Discovery studio visualizer

through in silico subtractive pathway analysis, differential genome approach and essentiality study. Subtractive analysis of the metabolic pathways of *H. pylori* HPAG1 and human, as documented in the KEGG database, revealed 11 pathogen-specific pathways and a search for potentially indispensible pathogen proteins having no human homolog yielded 25 candidate proteins as promising drug targets. All these candidate proteins are, however, found to have orthologs in other closely related / distant pathogens and hence, may be used for broad-spectrum antibiotics. Among the eleven identified pathogen-specific pathways, the LPS biosynthesis pathway has been found to be the largest contributor to such candidate proteins – nine out of 25 identified drug targets belong to LPS biosynthesis pathway. An attempt has, therefore, been made to build up 3D structural models of three candidate enzymes of this pathway, 2-dehydro-3-deoxy-phosphooctonate aldolase, UDP-3-O-[3-hydroxymyr-istoyl] N-acetylglucosamine deacetylase and Phosphoheptose isomerase, for which suitable template structures were available in PDB. All three proteins selected for modeling are known to play important roles in the LPS biosynthesis pathway and blocking the action of any of these essential enzymes with no human homolog would be detrimental to the pathogen, but not to the host. Binding site analysis and docking of the known biological substrate PEP to 2-dehydro-3-deoxy-phosphooctonate aldolase revealed the potential binding pocket present in the single monomeric form of the enzyme and the residues that may potentially be involved in the protein-ligand interaction. These findings are likely to open up avenues for designing and development of novel drugs against *H. pylori* and other related pathogenic microbes.

Acknowledgments This work was supported by the Council of Scientific and Industrial Research (Project no. CMM0017). MS is supported by Senior Research Fellowship and LM is supported by Project Assistantship from Council of Scientific and Industrial Research, India. Authors also acknowledge the infrastructural support received from Bioinformatics Centre, Indian Institute of Chemical Biology.

References

1. Schistosomes, liver flukes and *Helicobacter pylori*. “IARC Monographs on the evaluation of carcinogenic risks to humans” (1994) 61:1–241. Lyon, 7–14 June
2. Kelly DJ (1998) The physiology and metabolism of the human gastric pathogen *Helicobacter pylori*. Adv Microb Physiol 40:137–189
3. Kusters JG, van Vliet AHM, Kuipers EJ (2006) Pathogenesis of *Helicobacter pylori*. Infection Clin Microbiol Rev 19:449–490
4. Lehours P, Zheng Z, Skoglund A, Mégraud F, Engstrand L (2009) Is there a link between the lipopolysaccharide of *helicobacter pylori* gastric malt lymphoma associated strains and lymphoma pathogenesis? Plos One 4(10):e7297
5. Montecucco C, Rappuoli R (2001) Living dangerously: how *Helicobacter pylori* survives in the human stomach. Nat Rev Mol Cell Biol 2:457–466
6. Mirbagheri SA, Hasibi M, Abouzari M, Rashidi A (2006) Triple, standard quadruple and ampicillin-sulbactam-based quadruple therapies for H. pylori eradication: a comparative three-armed randomized clinical trial. World J Gastroenterol 12:4888–4891
7. Graham DY, Hoffman J, El-Zimaity HM, Graham DP, Osato M (1997) Twice a day quadruple therapy (bismuth subsalicylate, tetracycline, metronidazole plus lansoprazole) for treatment of Helicobacter pylori infection Aliment. Pharmacol Ther 11:935–938
8. Fischbach L, Evans EL (2007) Meta-analysis: the effect of antibiotic resistance status on the efficacy of triple and quadruple first-line therapies for Helicobacter pylori Aliment. Pharmacol Ther 26:343–357
9. Graham DY, Fischbach L (2010) *Helicobacter pylori* treatment in the era of increasing antibiotic resistance. Gut 59:1143–1153
10. Stenström B, Mendis A, Marshall B (2008) Helicobacter pylori - The latest in diagnosis and treatment. Aust Fam Physician 37:608–612
11. http://www.uptodate.com/patients/content/topic.do?topicKey=~gi0iITvAhVEvR5
12. Oh JD et al (2006) The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: evolution during progression. Proc Natl Acad Sci USA 103:9999–10004
13. Allsop AE (1998) Bacterial genome sequencing and drug discovery. Currt Opin Biotechnol 9:637–642
14. Galperin MY, Koonin EV (1999) Searching for drug targets in microbial genomes. Curr Opin Biotechnol 10:571–578
15. Morya VK, Dewaker V, Mecarty SD, Singh R (2010) In silico Analysis of Metabolic Pathways for Identification of Putative Drug Targets for *Staphylococcus aureus*. J Comput Sci Sys Biol 3:062–069
16. Perumal D, Lim CS, Sakharkar KR, Sakharkar MK (2007) Differential genome analyses of metabolic enzymes in *Pseudomonas aeruginosa* for drug target identification. In Silico Biol 7:453–465
17. Chong CE, Lim BS, Nathan S, Mohamed R (2006) In silico analysis of Burkholderia pseudomallei genome sequence for potential drug targets. In Silico Biol 6:341–346
18. Dutta A, Singh SK, Ghosh P, Mukherjee R, Mitter S, Bandyopadhyay D (2006) In silico identification of potential therapeutic targets in the human pathogen Helicobacter pylori. In Silico Biol 6:43–47
19. http://www.ncbi.nlm.nih.gov
20. Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucl Acids Res 30:42–46
21. Altschul SF, Thomas LM, Alejandro AS, Jinghui Z, Zheng Z, Webb M, David JL (1997) Gapped BLAST and PSI BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
22. Zhang R, Hong YO, Zhang CT (2004) DEG: a database of essential genes. Nucleic Acids Res 32:D271–D272
23. http://www.cbs.dtu.dk/services/TMHMM
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242
25. http://accelrys.com/
26. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera – a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612
27. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26:283–291
28. Eisenberg D, Lüthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol 277:396–404, http://nihserver.mbi.ucla.edu/Verify_3D

29. Van der Spoel D, Lindahl E, Hess B, van Buuren AR, Apol E, Meulenhoff PJ, Tieleman DP, Sijbers ALTM, Feenstra KA, Van Drunen R, Berendsen HJC (2010) Gromacs User Manual. www.gromacs.org
30. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) GROMACS: fast, flexible, and free. J Comput Chem 26:1701–1718
31. Moran AP (1996) The role of lipopolysaccharide in *Helicobacter pylori* pathogenesis. Aliment Pharmacol Ther 10:39–50
32. Moran AP (1999) Helicobacter pylori lipopolysaccharide-mediated gastric and extragastric pathology. J Physiol Pharmacol 50:787–805
33. Bergman MP, Engering A, Smits HH, van Vliet SJ, van Bodegraven AA et al (2004) Helicobacter pylori modulates the T helper cell 1/T helper cell 2 balance through phase-variable interaction between lipopolysaccharide and DC-SIGN. J Exp Med 200:979–990
34. Appelmelk BJ, Negrini R, Moran AP, Kuipers EJ (1997) Molecular mimicry between Helicobacter pylori and the host. Trends Microbiol 5:70–73
35. Piotrowski J, Piotrowski E, Skrodzka D, Slomiany A, Slomiany BL (1997) Induction of acute gastritis and epithelial apoptosis by Helicobacter pylori lipopolysaccharide. Scand J Gastroenterol 32:203–211
36. Sakagami T, Vella J, Dixon MF, O'Rourke J, Radcliff F, Sutton P, Shimoyama T, Beagley K, Lee A (1997) The endotoxin of Helicobacter pylori is a modulator of host-dependent gastritis. Infect Immun 65:3310–3316
37. Perumal D, Sakharkar KR, Tang TH, Chow VT, Lim CS, Samal A, Sugiura N, Sakharkar MK (2010) Cloning and targeted disruption of two lipopolysaccharide biosynthesis genes, kdsA and waaG, of *Pseudomonas aeruginosa* PAO1 by site-directed mutagenesis. J Mol Microbiol Biotechnol 19:169–179
38. Duewel HS, Radaev S, Wang J, Woodard RW, Gatti DL (2001) Substrate and metal complexes of 3-deoxy-D-manno-octulosonate-8-phosphate synthase from *Aquifex aeolicus* at 1.9-A resolution. Implications for the condensation mechanism. J Biol Chem 276:8393–8402
39. Taylor PL, Blakely KM, de Leon GP, Walker JR, McArthur F, Evdokimova E, Zhang K, Valvano MA, Wright GD, Junop MS (2008) Structure and function of sedoheptulose-7-phosphate isomerase, a critical enzyme for lipopolysaccharide biosynthesis and a target for antibiotic adjuvants. J Biol Chem 283:2835–2845
40. Coggins BE, McClerren AL, Jiang L, Li X, Rudolph J, Hindsgaul O, Raetz CR, Zhou P (2005) Refined solution structure of the LpxC-TU-514 complex and pKa analysis of an active site histidine: insights into the mechanism and inhibitor design. Biochemistry 44:1114–1126

J Mol Model (2012) 18:1867–1875
DOI 10.1007/s00894-011-1196-z

ORIGINAL PAPER

Molecular modeling approach to predict a binding mode for the complex methotrexate-carboxypeptidase G_2

Kely Medeiros Turra ·
Kerly Fernanda Mesquita Pasqualoto ·
Elizabeth Igne Ferreira · Daniela Gonçales Rando

Received: 11 January 2011 / Accepted: 19 July 2011 / Published online: 25 August 2011

Abstract Carboxypeptidase G_2 (CPG_2) is a zinc-metalloenzyme employed in a range of cancer chemotherapy strategies by activating selectively nontoxic prodrugs into cytotoxic drugs in tumor as well as in the treatment of intoxication caused by high-doses of the anticancer drug methotrexate (MTX). CPG_2 catalyzes the hydrolytic cleavage of C-terminal of glutamate moiety from folic acid and analogues. Regardless of its extensive application, its mechanism of catalysis has not yet been determined and, so far, no co-crystallized complex has been published. So, in this study, molecular docking and a short molecular dynamics (MD) simulation sampling scheme, as a function of temperature, were performed to investigate a possible binding mode for MTX, a recognized substrate of CPG_2. The findings suggested that MTX interacts possibly in quite specific points of the CPG_2 active site, which are probably responsible for the molecular recognition and cleavage procedures. The MTX substrate fits well in the catalytic site by accommodating the pteridine moiety in an adjacent pocket to the active site whereas a glutamate moiety is pointed toward the protein surface. Additionally, a glutamate residue can interact with a crystallization water molecule in the active site, supporting its activation as a nucleophilic group.

Keywords Binding mode · Carboxypetidase G_2 · Catalytic mechanism · Methotrexate · Molecular modeling

Introduction

Carboxypeptidase G_2 (CPG_2) is a zinc-dependent metalloenzyme employed in a range of cancer chemotherapy strategies such as antibody-directed enzyme prodrug therapy (ADEPT), gene-directed enzyme prodrug therapy (GDPET), as well as in the treatment of intoxication caused by high-doses of the anticancer drug methotrexate (MTX) [1, 2]. CPG_2 is a bacterial enzyme, produced as a homodimer by *Pseudomonas* sp. strain RS-16, which catalyzes the hydrolytic cleavage of C-terminal of glutamate moiety from folic acid and its analogues. Its X-ray crystal structure was determined at 2.5 Å of resolution [3]. Each subunit of the molecular dimer consists of a binding domain containing two zinc ions (oxidation state 2^+) and a dimerization domain of four anti-parallel β-sheets flanked by two α-helices [3–5]. The active site contains two His residues (His 112 and His 385), which coordinate each of the metal atoms separately in order to maintain them in an average distance of 3.3 Å. The Asp 141 residue coordinates simultaneously the both zinc ions while the Glu 175 residue forms a hydrogen bond with a crystallization water molecule (HOH_{36}), which bridges the metals. One of the

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1196-z) contains supplementary material, which is available to authorized users.

K. M. Turra · K. F. M. Pasqualoto · E. I. Ferreira ·
D. G. Rando (✉)
Laboratory of design and synthesis of chemotherapeutical agents potentially actives against tropical diseases – LAPEN,
Department of Pharmacy, Faculty of Pharmaceutical Sciences,
University of São Paulo- USP,
Av. Prof. Lineu Prestes, 580, Cidade Universitária,
05508900 São Paulo, SP, Brazil
e-mail: dgrando@unifesp.br

D. G. Rando
Departamento de Ciências Exatas e da Terra, Instituto de Ciências Ambientais, Químicas e Farmacêuticas,
Universidade Federal de São Paulo,
Campus Diadema, R. Prof. Arthur Riedel, 275, Eldorado,
09972-270 Diadema, SP, Brazil

zinc atoms (Zn2) is also coordinated by the Glu 200 residue whereas the other (Zn1) is coordinated by Glu 176 [3, 4].

A comparison of the CPG_2 tridimensional structure to other metallohydrolases revealed that it shares 56% of identity in the active site with an aminopeptidase from *Aeromonas proteolytica* (AMP) [3, 4, 6]. Despite its extensive application, the CPG_2 mechanism of catalysis has not yet been elucidated. Also, there is no co-crystallized structure of any of its known substrates reported [4, 7]. However, it is believed that the CPG_2 catalysis can happen via a general mechanism applied to all metallopeptidases containing similar co-catalytic metallo-active sites. For instance, the aminopeptidase from AMP bacteria and glutamate carboxypeptidase GII (GCPGII) have had their mechanism of catalysis suggested through the application of computational methods [3, 6, 8, 9].

In this regard, the first step in catalysis procedure is probably the recognition of the N- or C-terminal group from the substrate by an adjacent pocket to the active site, which can present hydrophobic or hydrophilic character. The zinc ion at the CPG_2 active site provides the orientation of the carbonyl group of substrate toward a crystallization water molecule to promote its polarization, and consequently making it more susceptible to a nucleophilic attack (see Fig. 1, step 1). The zinc bound-water molecule acts as a nucleophile, and the attack possibly occurs through a tetrahedral transition state (Fig. 1, step 2). Finally, an amino acid residue located near to the active site would act as a hydrogen bonding acceptor, aiding the deprotonation of the nucleophilic water molecule (Fig. 1, step 3) [4, 8].

The role of both zinc atoms is not yet completely understood but some authors believed that the Zn1 (Fig. 1), termed as catalytic zinc, could bind the substrate to orient the peptide whereas the Zn2 (Fig. 1), called co-catalytic zinc, should be important to stabilize the anionic tetrahedral intermediate [8]. This general proposed mechanism, how-

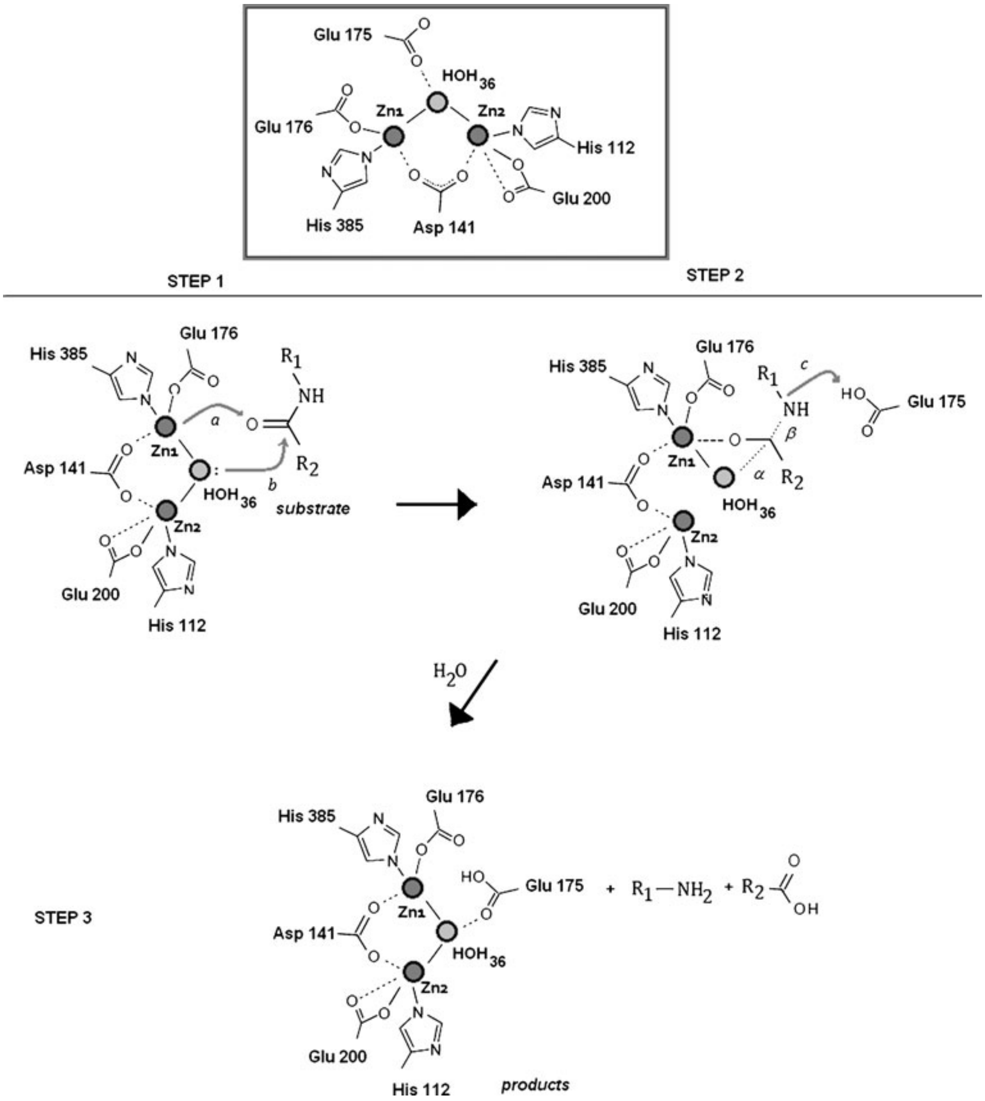


Fig. 1 Schematic view of the CPG_2 active site from *Pseudomonas* sp. based on X-ray crystallographic coordinates [3], and general mechanism proposed regarding the hydrolysis via metallopeptidases (α corresponds to the bond formation as result of a nucleophilic attack; β is related to the broken bond after a nucleophilic attack)

ever, has never been empirically or theoretically proved, at least regarding the CPG_2.

Thus, the present study reports a binding mode hypothesis established to the complex CPG_2-MTX by applying molecular docking and a short molecular dynamics (MD) simulation sampling scheme as a function of temperature (warming up scheme). Moreover, previous information related to binuclear zinc-metalloenzymes, such as AMP and GCPGII [6, 8, 9], was also considered for building up the binding hypothesis. The findings can be used as a starting point to the rational design of new substrates which can be applied in chemical drug delivery systems.

Methodology

Building up the three-dimensional models

The three-dimensional (3D) structure of MTX was constructed in its protonated state using the HyperChem 7.51 software [10]. The MTX co-crystallized structure bound to the dihydrofolate reductase from *Moritella profunda* was retrieved from Brookhaven Protein Data Bank (PDB) [11] (entry code 3IA4 at 1.70 Å resolution [12]) and used as the reference geometry for drawing the ligand. This crystallized structure was used to have a theoretical minimum closest to an experimental geometry. The geometry optimization was performed in MM + force field without any constraints and the partial atomic charges were computed using the AM1 [13] semiempirical method, also implemented in the HyperChem software [10].

Roswell et al. [3] deposited the Cartesian coordinates of CPG_2 in PDB [11] under the entry code 1CG2 at resolution of 2.5 Å. As already mentioned, each subunit of the molecular dimer consists of a larger catalytic domain containing two zinc ions at the active site, and a separate smaller domain that forms the dimer interface. This 3D structure was used as a receptor geometry reference, but in order to reduce the computational time consuming only one polypeptide chain was employed to build up the receptor model, and subsequently perform the MD simulations. The two active sites present in the dimer are 62 Å apart and are presumed to be independent. So, the analysis of only one chain can be consequently considered as biologically relevant.

The appropriate number of hydrogens was added on all atoms of the receptor model and methyl groups were used as blocking groups in the N- and C-terminal portions. AMBER atom types [14] and partial charges were assigned to all atoms, except to the blocking groups. All ionizable residues were assigned the charge state which normally present at the pH of the experimental conditions (7.2). Lone pair electrons were not modeled explicitly. Water molecules located in the crystal structure of CPG_2 were removed except to the one which probably participates in the ligand-receptor interactions at the enzyme active site (HOH_{36}). The final model was also visually inspected to be certain of its structural integrity.

Molecular docking and MD simulations

GOLD software, version 3.1 [15], was used to explore and derive the best binding interactions and conformation of MTX in the active site of the target protein, employing a genetic algorithm. The energy functions of the interactions are partly based on the conformational and non-bonded interactions. The fitness function used was GoldScore, which corresponds to the sum of four energy components, such as: protein-ligand hydrogen bond energy (external H-bond), protein-ligand van der Waals (vdW) energy (external vdW), ligand internal vdW energy (internal vdW), and ligand torsional strain energy (internal torsion) [16]. The zinc ions were parameterized as tetrahedral, and a cutoff radius of 10 Ǻ was created around the docked molecule. All the other options were kept as default. Ten docking runs were performed considering ten conformations each to totalize a hundred conformations. All conformations were analyzed according to the distance between protein-ligand atoms, which were most likely involved in the catalytic mechanism. The donor-acceptor bond distance limit was considered as 3.90 Å [8, 17, 18]. Then, the best binding model was chosen combining information from the energy rank position and the alignment/orientation in the active site regarding the intermolecular interactions (distance values).

The selected complex model was used as input to the energy-minimization procedure using the MOLSIM 3.2 program [19]. The steepest descent (500 iterations) and conjugated gradient (219 iterations) methods were applied and the convergence criterion established was 0.1 kcal mol^{-1}. A dielectric constant value of 3.5, which simulates the biological membrane environment, was considered in the analysis of each selected complex model [20, 21]. The energy-minimized output model was the initial structure for the MD simulations, also employing the MOLSIM 3.2 program [19].

A fictitious mass of 5000 u.m.a. was assigned to all main chain atoms (backbone) of the entire receptor/enzyme model in order to maintain the integrity of the model during the MD simulations. The use of fictitious masses is virtually the same as using Cartesian constraints, particularly when the masses are chosen to be very large [22].

A short MD simulation sampling scheme at progressively higher temperatures was repeated until a user-defined final temperature of evaluation was reached. This kind of MD simulation sampling scheme was applied here for generating the lowest energy state of the complex as fast as possible, allowing the crossing of energetic barriers through a warming up scheme of the system. Also, a better accommodation of the ligand MTX in the active site of CPG_2 can be provided by using that procedure, since at the higher simulated temperatures the docked ligand is allowed to find its optimal intermolecular alignment in the active site [22]. The lowest energy conformation of each simulation was selected and then used as the starting geometry for a subsequent MD simulation at a higher temperature. The MD simulation sampling scheme performed was the following: 20 ps (step size of 1 fs) at 50 K, 100 K, 200 K, and 300 K. Output trajectory files were saved at every 20 simulation steps, resulting 5000 conformations. The choice of short or relaxation MD simulations was found to be the best compromise in producing trajectory geometries and energies that remained close to the X-ray structure of CPG_2 for reasonable amounts of CPU time.

As soon as the higher temperature simulation was accomplished a longer slightly cooling down simulation was performed. So, the lowest energy conformation of the complex selected from the simulation at 300 K was gently cooling down in a MD simulation of 1 ns at 298 K (temperature of the CPG_2 experimental assay).

The hydration shell model proposed by Hopfinger [23] was employed to estimate the solvation energy contribution of the lowest energy conformation identified from each MD simulation sampling scheme, since the MOLSIM 3.2 software does not consider explicit water molecules during the MD simulations. The absence of explicit water molecules during the simulation analysis can lead to the generation and sampling of artifact states relative to the actual binding mode. Otherwise, the inclusion of explicit waters introduces questions regarding assignments of waters and the extent of sampling needed to generate equilibrium/steady ensembles. These unknowns can also lead to artifact states. The use of an implicit solvation model as a hydration shell scheme seems to be both a good compromise, and a diagnostic, to evaluate solvation effects [23].

The lowest energy conformation of the receptor/enzyme model selected from each MD simulation was compared to the crystal structure of CPG_2 through the root mean square deviation (RMSD), using HyperChem 7.51 [10], in order to verify if the integrity of the system was maintained during the MD simulations sampling scheme. RMSD values lower or equal to 1.5 Å were considered acceptable [22], indicating no significantly structural deviation.

Lipophilic potential and cavity depth properties mapping

Lipophilic potential (LP) and cavity depth (CD) properties of the CPG_2-MTX complex selected from MD simulation at 298 K were mapped onto Connolly partial surfaces, using the MOLCAD module of the Sybyl 8.0 package [24]. These procedures allow to visualize surface features and physical properties essential for molecular recognition, and also to characterize the size, shape, and physical properties of intramolecular cavities and channels. The colors coded for the LP map and CD property range from brown (lipophilic regions) to blue (hydrophilic regions) and from blue (shallow) to yellow (deep or buried), respectively.

Results and discussion

The molecular docking procedure was applied to find out the most likely alignments of MTX in the active site. The distance between the zinc atom (Zn1, 1) of the active site and the oxygen (2) of the carbonyl group of the amide from the substrate, and that involving the nucleophilic oxygen (3) of the crystallization water molecule and also the carbonyl carbon (4) of the amide from the substrate (see the dashed lines in Fig. 2) were considered for analyzing a set of 100 conformations.

Thus, a distance value of 3.90 Å was considered as the limit for the establishment of an electron donor-acceptor bond. This limit of distance was based upon reported distance values obtained from theoretical studies applied to the enzyme glutamate carboxypeptidase and a peptide substrate [8] as well as from geometry studies regarding metal-ligand interactions using crystallography [17].

As already mentioned, the values found for the energy fitness function were also evaluated for selecting the best conformations. Fifteen from a hundred conformations presented reasonable distance values but not all had a suitable energy scoring function (Goldscore). The findings are shown in Table 1. Only those conformations, which were sufficient in both criteria, were selected to further studies.

Conformation I, which was ranked in the first place regarding the energy fitness function (GoldScore), exhibited a distance value of 2.75 Å between atoms 1 and 2, and 2.92 Å between atoms 3 and 4, and consequently was chosen as the best complex model (see Fig. 3).

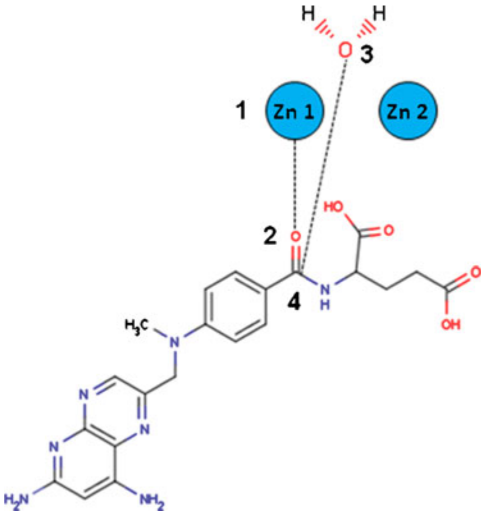


Fig. 2 Distances considered for selecting the best complex models. Carbon atoms are depicted in black, nitrogen in blue, oxygen in red, and zinc atoms are as light blue spheres called as Zn1 and Zn2, respectively

Table 1 The best 15 conformations from the molecular docking procedure

Conformation	Docking*	Goldscore**	Distance (Å) of atoms (1) and (2)	Distance (Å) of atoms (3) and (4)
I	1	1°	2.75	2.92
II	1	3°	3.33	3.22
III	1	8°	3.39	3.22
IV	2	1°	3.59	3.56
V	2	3°	3.36	3.25
VI	3	1°	3.44	3.22
VII	3	9°	2.79	3.67
VIII	4	10°	3.41	3.29
IX	5	6°	3.55	3.46
X	5	9°	3.49	3.37
XI	6	2°	3.26	3.12
XII	7	3°	3.49	3.53
XIII	8	3°	3.37	3.29
XIV	8	5°	3.65	3.58
XV	10	9°	3.42	3.30

* *Docking* column indicates from which docking run the complex was obtained

** *Goldscore* column points out the position of the complex among the pool obtained from a specific docking run. The scoring function applied here was based upon the energy value found for the complexes generated in each run. Although, conformations occupying the same position in the ranking, but resulting from distinct run, does not necessarily present the same energy

The complex model from molecular docking procedure showed that MTX probably interacts in specific points, providing a molecular recognition and cleavage by the target, CPG_2. MTX is aligned in the catalytic site by accommodating the pteridine moiety in an adjacent pocket (S1, Fig. 4) to the active site composed mainly by the amino acid residues Ser 210, Thr 213, Phe 327, Thr 361, Ala 363, Ile 374, and Glu 375. The glutamic acid side-chain is pointed towards the surface (S1', Fig. 4) and it is seemingly stabilized by the establishment of a hydrogen bonding interaction with the Arg 324 residue of the enzyme. A general view of the CPG_2 active site and the mentioned MTX orientation is depicted in Fig. 4a and b, respectively.

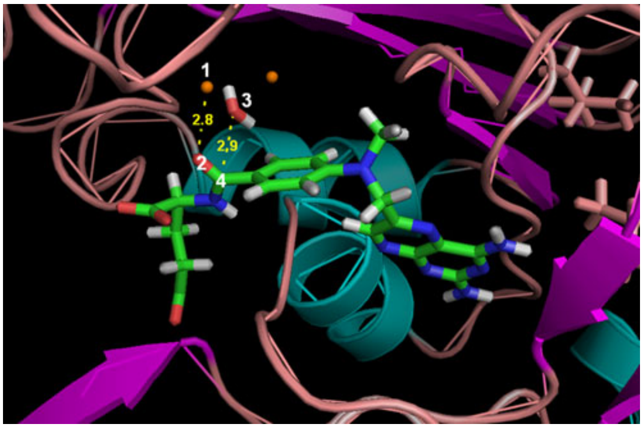


Fig. 3 Graphical representation of conformation I found for the complex CPG_2-MTX. MTX is presented as stick model where the carbon atoms are in light green, oxygen in red, nitrogen in blue, zinc ions in orange, and hydrogen atoms in white. The receptor model is shown as Richardson or cartoon scheme where β-sheets are as pink narrows and α-helices as cyan spiral ribbons (Pymol Viewer) [25]

The pteridine moiety fits well in the S1 pocket and it is held by hydrogen bonding interactions involving both the endo and exocyclic nitrogen atoms from pteridine and the carbonyl oxygen from the Lys 208 and Ser 210 residues, respectively. The amide nitrogen of Ser 210 also participates in an intermolecular interaction with MTX as hydrogen bonding donor. These molecular docking findings were validated by the MD simulation sampling scheme. The two approaches presented a similar profile regarding the distance values and atomic positions found for MTX and the amino acid residues of CPG_2 active site. That can be checked by comparing Figs. 4 and 5. However, it is noteworthy that the MD simulation sampling scheme provides a more reliable data in terms of interatomic distances and intermolecular interactions.

Rowsell et al. [3] reported a pocket in the CPG_2 crystal structure which could be an equivalent portion to a hydrophobic pocket present in a similar aminopeptidase from AMP, which accommodates the N-terminal phenylalanine residue of a hydroxamate inhibitor. However, in CPG_2, this pocket would be hydrophilic and it could accommodate the large pteroate moiety of folic acid. In this theoretical study, the pteroate moiety of MTX is indeed lying in the S1 pocket.

The glutamate moiety is pointed toward the surface of protein. Additionally, it establishes a hydrogen bonding interaction with the Arg 324 residue located in the S1' pocket, and probably interacts with adsorbed water molecules present at the CPG_2 active site (see Fig. 5). Again, the findings are in agreement with those published by Rowsell et al. [3]. They showed that the replacement of the Arg 324 residue by an Ala would provide a mutant enzyme with low activity

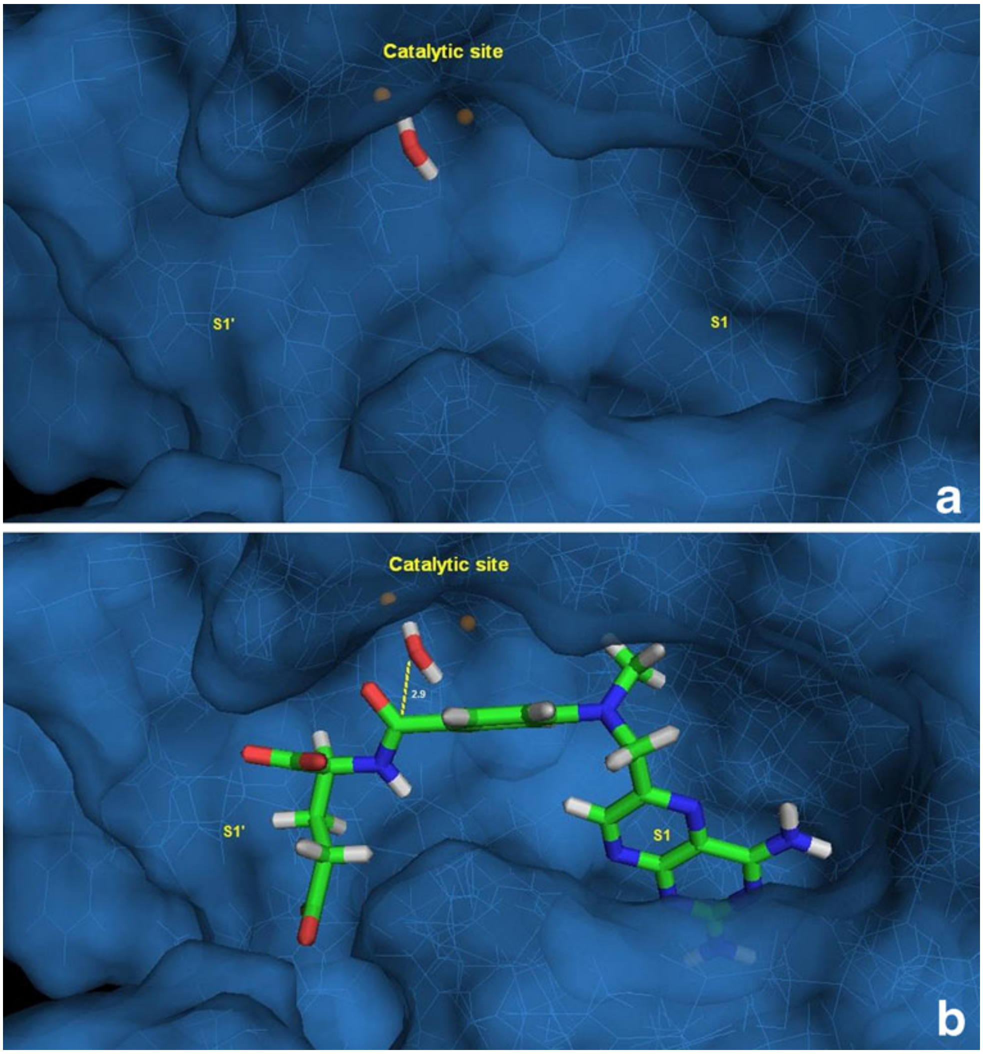


Fig. 4 General view of the CPG_2 active site showing the two possible pockets involved in the molecular substrates recognition (**a**). Alignment of the MTX in the active site displaying the pteridine ring lying at the S1 pocket (**b**). The protein molecular surface is colored in blue. MTX is presented as stick model where the carbon atoms are in green, oxygen in red, nitrogen in blue, zinc ions in orange, and hydrogen atoms in white (Pymol Viewer [25])

toward MTX. The results from the MD simulation sampling scheme also suggested that the Glu 175 residue of CPG_2 active site would act as a general acid/base during the catalysis. One of its carboxyl groups seems to be placed at a favorable distance (3.67 Å) to interact with the hydrogen of amide from MTX. The other Glu 175 carboxyl group would

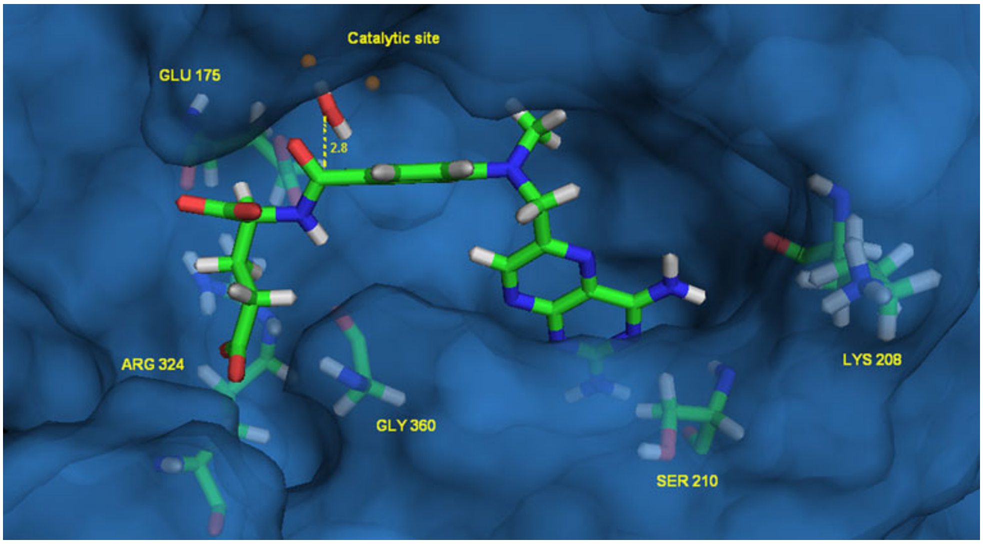


Fig. 5 Representation of the MTX binding mode in the CPG_2 active site found after the MD simulation sampling scheme. The protein molecular surface is colored in blue and the residues Glu 175, Arg 324, Gly 360, Ser 210, and Lys 208 are showed as stick model where the carbon atoms are in green, oxygen in red, nitrogen in blue, zinc ions in orange, and hydrogen in white. MTX and a water molecule are also presented as stick model whereas the zinc ions are as orange spheres (Pymol Viewer [25])

Table 2 Thermodynamic parameters found for the CPG_2-MTX lowest-energy conformations selected from the MD simulation sampling scheme

T (K)	$E_{stretch}$	E_{bend}	E_{tors}	$E_{1,4}$	E_{vdW}	E_{el}	$E_{intervdW+el}$	E_{solv}	E_{Hb}	E_T
50	2862.4	2755.6	2735.8	8690.3	−2254.8	−8641.1	588.6	−774.0	−57.5	−51.6
100	3111.7	3077.9	2793.8	8778.3	−2214.3	−8775.0	579.1	−705.0	−58.0	−51.3
200	3674.5	3687.8	2857.4	8850.9	−2011.8	−8837.9	581.9	−679.0	−56.6	−48.5
300	4123.2	4143.8	2976.3	8954.5	−1938.4	−8933.0	586.7	−564.0	−56.9	−47.6
298	3991.7	4037.2	2985.7	8936.8	−2022.3	−9418.5	584.9	−577.5	−57.2	−57.15

$E_{stretch}$ = stretching energy; E_{bend} = bending energy; E_{tors} = torsional energy; $E_{1,4}$ = 1–4 interaction energy; E_{vdW} = van der Waals energy; E_{el} = electrostatic energy; E_{vdW+el} = sum of van der Waals and electrostatic intermolecular energy contributions, E_{solv} = solvation energy; E_{Hb} = hydrogen bonding energy; E_T = total potential energy, which is the summation of all energy contributions (kcal mol^{-1})

interact with the water molecule in the active site, supporting the activation of this water as a nucleophilic agent as similarly proposed to the Glu 151 residue in the AMP aminopeptidase [3, 4]. Additional intermolecular interactions were also observed, for instance, involving the amide of the substrate and the carbonyl oxygen of the Gly 360 residue.

The plots of the results from the MD simulation sampling scheme found for the CPG_2-MTX complex selected in the molecular docking procedure are presented in Figs. S1 and S2 (Electronic supplementary material). The energy of the conformational ensemble profile (CEP) from the CPG_2-MTX complex was stabilizing as the temperature was increasing and it was more energetically stable only when the temperature of 300 K was reached (see Fig. S1). A longer and more stable CEP was generated at 298 K (Fig. S2).

The thermodynamic parameters found for the CPG_2-MTX lowest-energy conformations selected from the MD simulation sampling scheme are presented in Table 2. The total potential energy (E_T) is the summation of the following energy contributions: the stretching energy ($E_{stretch}$), the bending energy (E_{bend}), the torsional energy (E_{tors}), the 1–4 (Lennard-Jones) interaction energy ($E_{1,4}$), the van der Waals energy (E_{vdW}), the electrostatic energy (E_{el}), the sum of van der Waals and electrostatic intermolecular energy contributions (E_{vdW+el}), the solvation energy (E_{solv}), and the hydrogen bonding energy (E_{Hb}). It is

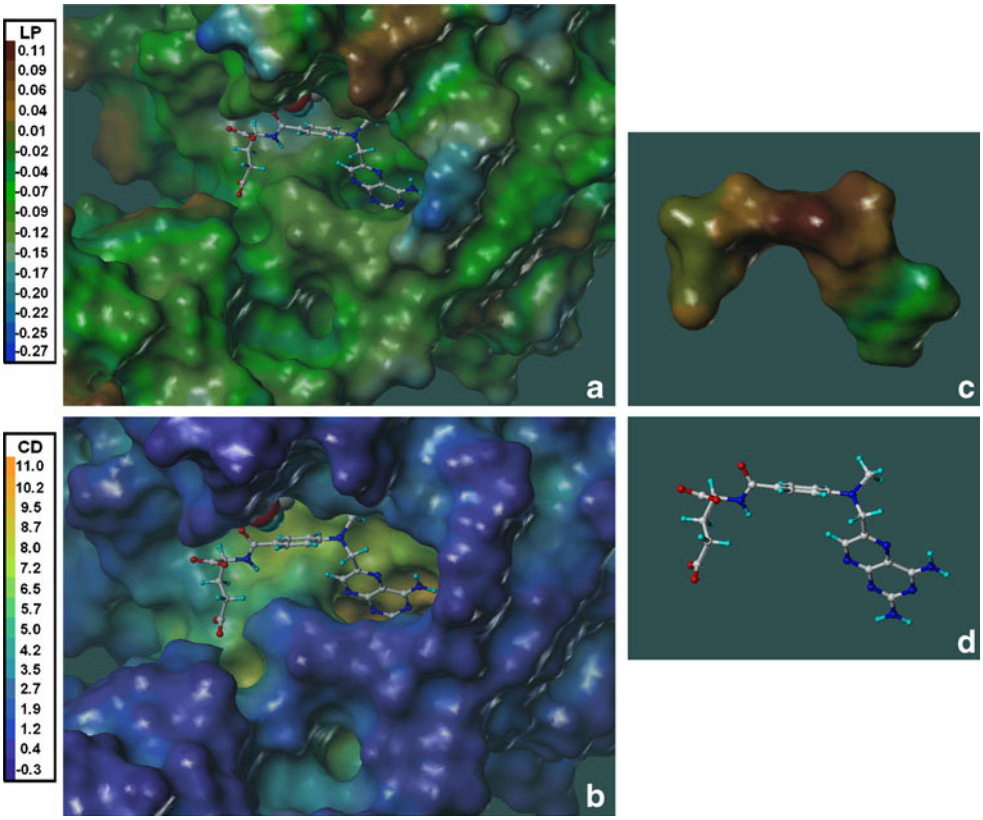


Fig. 6 Cavity depth, CD, (**a**) and lipophilic potential, LP, (**b**) maps calculated onto the Connolly partial surfaces of the CPG_2-MTX complex selected from the MD simulation at 298 K (MOLCAD, Sybyl 8.0 package [24]). The LP maps range from brown (lipophilic regions) to blue (hydrophilic regions) whereas the CD property ranges from blue (shallow) to yellow (deep or buried). (**c**) LP map of MTX; (**d**) Ball-stick model of MTX where carbon atoms are depicted in light gray, nitrogen in blue, oxygen in red and hydrogen in cyan

noteworthy that the CPG_2-MTX lowest-energy conformations selected from the MD sampling scheme did not present significant differences in terms of E_T values. The E_T values ranged from −57 to −47 kcal mol^{-1}.

Moreover, the CPG_2 models selected from the MD simulation sampling scheme presented RMSD values smaller than 1.5 Å when compared to the crystallographic enzyme (Fig. S3, Electronic supplementary material), which indicates that they maintained their structural integrity during the MD simulation procedure [22].

The molecular surfaces analysis for the CPG_2-MTX complex selected from MD simulation at 298 K was performed using the MOLCAD module (Sybyl 8.0 software [24]). The LP and CD properties were mapped onto Connolly partial surfaces as shown in Fig. 6.

The CD mapping (Fig. 6a) provided a clearer view regarding the tridimensional arrangement of the S1 pocket, which seems to be deeply buried in the 3D enzyme structure, and it would not be so available to establish molecular interactions as those observed to the zinc atoms and water molecule in the CPG_2 active site.

Observing Fig. 6(b), the pteroate moiety seems to anchor in the S1 pocket having as driving forces not only the already discussed hydrogen bonding interactions but also a complementarity of LP property with this part of the protein structure. The blue (to green) colored region in the CPG_2 LP map reveals a certain hydrophilic character of the S1 pocket. Interestingly, the most hydrophilic portion of the MTX LP map (Fig. 6c), colored in green (to blue), fits perfectly over that region of the active site. Otherwise, hydrophobic interactions would occur only in a portion above the active site, which is not quite relevant for the MTX positioning or enzymatic attack.

Finally, despite the presence of two carboxylic groups in the MTX glutamate portion, it did not present a highly hydrophilic potential in comparison to the whole system. Even so, the glutamate still points toward the most external part of CPG_2 where adsorbed water molecules would usually take placed. This arrangement can be well understood when a hydrogen bonding interaction with the Arg 324 residue is considered. The elimination of this amino acid residue from the enzyme structure, as mentioned above, leads to a decreasing of the catalytic activity. Thus, the hydrophilic character of the pteroate moiety would provide a better alignment of MTX toward the S1 pocket (also hydrophilic) in the CPG_2 molecular recognition process.

Conclusions

The evaluation of the findings from molecular docking and MD simulation sampling scheme to the complex CPG2-MTX indicated that MTX interacts in specific points regarding the recognition process by CPG_2, at a molecular level, and it probably is cleaved following a general mechanism proposed for metalloproteases [4, 8]. Moreover, some particularities should be pointed out, such as the S1 pocket arrangement and specific intermolecular interactions, which seem to keep MTX in a favorable orientation/alignment to suffer a nucleophilic attack.

Acknowledgments The authors thank the Brazilian scientific committee, CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), for the scholarship and financial support, and the Chem21 Group, for providing the academic license of MOLSIM 3.2 program.

References

1. Jamin Y, Gabellieri C, Smyth L, Reynolds S, Robinson SP, Springer CJ, Leach MO, Payne GS, Eykyn TR (2009) Hyperpolarized C magnetic resonance detection of carboxypeptidase G2 activity. Magn Reson Med 62:1300–1304
2. Hempel G, Lingg R, Boos J (2005) Interactions of carboxypeptidase G2 with 6S-leucovorin and 6R-leucovorin in vitro: implications for the application in case of methotrexate intoxications. Cancer Chemother Pharmacol 55:347–353
3. Rowsell S, Pauptit RA, Tucker AD, Melton RG, Blow DM, Brick P (1997) Crystal structure of carboxypeptidase G2, a bacterial enzyme with applications in cancer therapy. Structure 5:337–347
4. Holz RC, Bzymek KP, Swierczek SI (2003) Co-catalytic metallopeptidases as pharmaceutical targets. Curr Opin Chem Biol 7:197–206
5. Hedley D, Ogilvie L, Springer C (2007) Carboxypeptidase G2-based gene-directed enzyme–prodrug therapy: a new weapon in the GDEPT armoury. Nat Rev Cancer 7:870–879
6. Holz RC (2002) The aminopeptidase from Aeromonas proteolytica: structure and mechanism of co-catalytic metal centers involved in peptide hydrolysis. Coord Chem Rev 232:5–26
7. Wouters MA, Husain A (2001) Changes in Zinc ligation promote remodeling of the active site in the zinc hydrolase superfamily. J Mol Biol 314:1191–1207
8. Klusák V, Bařinka C, Plechanovová A, Mlčochová P, Konvalinka J, Rulíšek L, Lubkowski J (2009) Reaction mechanism of glutamate carboxypeptidase II revealed by mutagenesis, x-ray crystallography, and computational method. Biochemistry 48:4126–4138
9. Schürer G, Lanig H, Clark T (2004) Aeromonas proteolytica aminopeptidase: an investigation of the mode of action using a quantum mechanical/molecular mechanical approach. Biochemistry 43:5414–5427
10. HyperChem Program Release 7.05 for Windows (2005) Hybercube Inc,Gainesville, FL
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242
12. Hay S, Evans RM, Levy C, Loveridge EJ, Wang X, Leys D, Allemann RK, Scrutton NS (2009) Are the catalytic properties of enzymes from piezophilic organisms pressure adapted? ChemBioChem 10:2348–2353
13. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) AM1: a new general purpose quantum mechanical molecular model. J Am Chem Soc 107:3903–3909

14. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner PJ Jr (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 106:765–784
15. GOLD Suite Intuitive Protein-Ligand Docking Package, version 3.1 for Windows. The Cambridge Crystallographic Data Centre, UK
16. GOLD User Guide & Tutorials (2008) The Cambridge Crystallographic Data Centre, UK,.http://www.ccdc.cam.ac.uk/support/documentation/gold/3_2/doc/portable_html/gold_portable-3-246.html
17. Harding MM (2001) Geometry of metal ligand interactions in proteins. Acta Cryst D57:401–411
18. Mc Donald IK, Thornton JM (1994) Satisfying hydrogen-bonding potential in proteins. J Mol Biol 238:777
19. Doherty DC (2001) MOLSIM 3.2. The Chem21Group Inc, Lake Forest, IL
20. Tokarski JS, Hopfinger AJ (1997) Prediction of ligand-receptor binding thermodynamics by free energy force field (FEFF) 3D-QSAR analysis: application to a set of peptidometic renin inhibitors. J Chem Inf Comput Sci 37:792–811
21. Santos-Filho OA, Mishra RK, Hopfinger AJ (2001) Free energy force field (FEFF) 3D-QSAR analysis of a set of Plasmodium falciparum dihydrofolate reductase inhibitors. J Comput Aided Mol Des 15:787–810
22. Tokarski JS, Hopfinger AJ (1997) Constructing protein models for ligand-receptor binding thermodynamic simulations: an application to a set of peptidometic renin inhibitors. J Chem Inf Comput Sci 37:779–791
23. Forsythe KH, Hopfinger AJ (1973) The influence of solvent on the secondary structures of poly(L-alanine) and poly(L-proline). Macromolecules 6:423–437
24. SYBYL version 8.0, Molecular Modeling Software Packages (2007) Tripos Inc, St Louis, MO
25. DeLano WL (2004) The pymol molecular graphics system, version1.0. DelanoScientific LLC, Palo Alto, CA. http://www.pymol.org/

J Mol Model (2012) 18:1877–1883
DOI 10.1007/s00894-011-1207-0

ORIGINAL PAPER

Quantum chemical modeling of the kinetic isotope effect of the carboxylation step in RuBisCO

Jan Philipp Götze · Peter Saalfrank

Received: 27 June 2011 / Accepted: 31 July 2011 / Published online: 25 August 2011

Abstract Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), the most important enzyme for the assimilation of carbon into biomass, features a well-known isotope effect with regards to the CO_2 carbon atom. This kinetic isotope effect $\alpha = k_{12}/k_{13}$ for the carboxylation step of the RuBisCO reaction sequence, and its microscopic origin, was investigated with the help of cluster models and quantum chemical methods [B3LYP/6-31G(d,p)]. We use a recently proposed model for the RuBisCO active site, in which a water molecule remains close to the reaction center during carboxylation of ribulose-1,5-bisphosphate [B. Kannappan, J.E. Gready, J. Am. Chem. Soc. 130 (2008), 15063]. Alternative active-site models and/or computational approaches were also tested. An isotope effect alpha for carboxylation is found, which is reasonably close to the one measured for the overall reaction, and which originates from a simple frequency shift of the bending vibration of $^{12}CO_2$ compared to $^{13}CO_2$. The latter is the dominant mode for the product formation at the transition state.

Keywords Cluster model · Dark reactions · Densityfunctional theory · Isotope effect · Photosynthesis · Quantum chemistry · RuBisCO

J. P. Götze · P. Saalfrank
Theoretische Chemie, Institut für Chemie, Universität Potsdam,
Karl-Liebknecht-Straße 24-25,
D-14476 Potsdam-Golm, Germany

J. P. Götze (✉)
Max-Planck-Institut für Kohlenforschung,
Kaiser-Wilhelm-Platz 1,
D-45470 Mülheim an der Ruhr, Germany
e-mail: goetze@mpi-muelheim.mpg.de

Introduction

Plant growth depends to a large extent on the ability to fixate carbon dioxide. Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), is the core enzyme which adds carbon dioxide to ribulose-1,5-bisphosphate (RuBP) to form an enzyme-bound six-carbon intermediate, which is split by hydrolyzation into two molecules of 3-phosphoglyceric acid (3-PGA) [1–3]. During the reaction sequence leading from RuBP to 3-PGA, which is shown in Figure 1, an isotope effect is observed: There is a small kinetic preference to add $^{12}CO_2$ rather than $^{13}CO_2$. For RuBisCO from spinach (*Spinacia oleracea*) at pH=8 and room temperature, for example, one finds a ratio $\alpha = k_{12}/k_{13} = 1.030 \pm 0.001$ for the overall reaction rate constants [4, 5]. Note that Fig. 1 is an idealization, leaving, *e.g.*, the possibility that carboxylation and hydrolyzation are concerted [3].

In the present work, we study kinetic isotope effects with quantum chemical models. We focus on RuBisCO found in the leaves of spinach, for which several X-ray structures are available, *e.g.*, [6–8]. Our investigation considers the carboxylation step only, *i.e.*, the electrophilic attack of CO_2 to enolized RuBP. For more detailed models of the full RuBisCO reaction, see Refs. [4, 9–12].

The carboxylation in RuBisCO has been studied by quantum chemical approaches in the past, *e.g.*, in Refs. [10, 13–16]. In Ref. [10], a cluster model was developed based on crystal structures [6, 7] and a mechanism was proposed, in which an initially Mg^{2+}-coordinated water molecule is replaced by CO_2 during carboxylation. This model will be called the "conventional model" in what follows. The X-ray structures leave some room for mechanistic interpretation, however, and recently an alternative mechanism was suggested in Ref. [11]. In this "new model", it was assumed that the water molecule initially coordinated with Mg^{2+} is

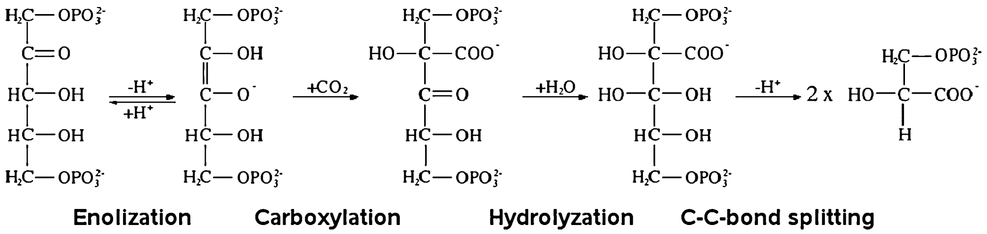


Fig. 1 Carboxylation reaction pathway of RuBisCO according to Refs. [1, 3, 7]. Individual steps: Enolization, followed by carboxylation and fast hydrolyzation, and finally C-C bond cleavage. Note that hydrolyzation and carboxylation are suspected to be concerted in some sources [1]

not replaced by CO_2 during the carboxylation reaction, but rather remains bound to the metal center before and during CO_2 addition. In this case, the water would be immediately available for the hydrolyzation step. In Ref. [11], several computational models were devised for this situation, the largest comprising 77 atoms, called the "FM20" cluster (see Figs. 2 and 3 below).

Here we will mostly consider the "new model" and the FM20 cluster also, now to study the kinetics of the carboxylation reaction. Test calculations are also done for the "conventional model". The precise composition and preparation of the model(s), along with a description of the computational methods is provided in Sect. System treatment and methods. The energetics and kinetics of the carboxylation step and the isotope effect will be considered in Sect. Results and discussion. The work concludes with a summary and outlook in Sect. Conclusions and outlook.

System treatment and methods

To set up a computational model for the "new mechanism" proposed by Kannappan and Gready [11], we used published coordinates [B3LYP/6-31G(d,p)] of the gas-phase transition state of their 77-atom FM20 model as a starting point. This model consists of a sixfold-coordinated Mg^{2+} ion, surrounded by CO_2, Lys175, Lys177, Lys201 (carboxylated, denoted "KCX201"), Asp203, Glu204, His294, Lys334, and a single water molecule bridging the Mg^{2+} and CO_2 units. In all amino acids backbone chains were replaced by single hydrogen atoms, and aliphatic chains by CH_3 groups, with the exception of Lys334, to remain faithful to the original model of Ref. [11]. By protonation / de-protonation, His and Lys are singly positively, and Glu, Asp are singly negatively charged. KCX was neutral. Further, the RuBP substrate molecule was cut between atoms C4 and C5, omitting one of the phosphate groups, and C5 was replaced by hydrogen. The other phosphate group was saturated with H and is neutral, but the RuBP model itself has –2 charge.

The overall charge of the cluster is +2. The FM20 transition state model is shown in Fig. 2.

Starting from the transition state structure, a fully relaxed scan along the distance of the C-atom of CO_2, and C-atom C2 of RuBP in steps of 0.1 Å was performed both in forward (to the product, carboxylated form) and backward

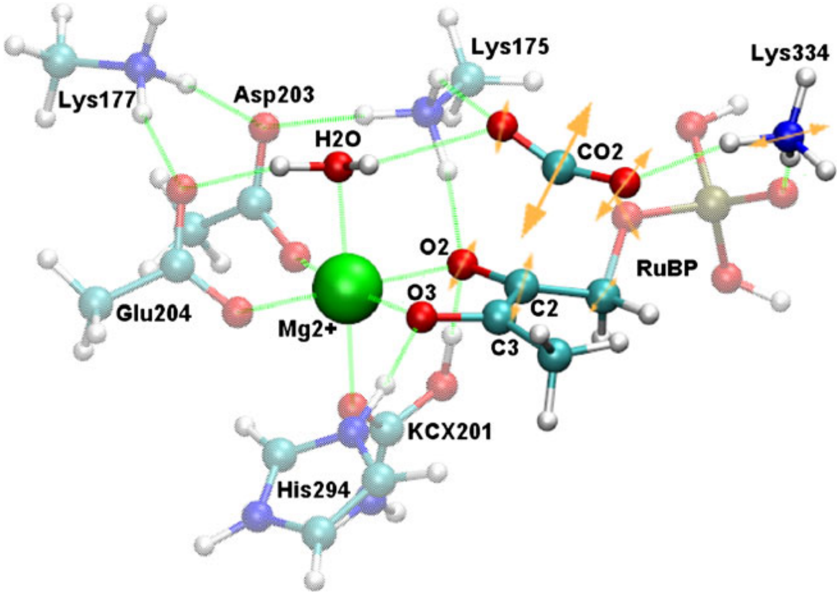


Fig. 2 Transition state of the 77-atom "new model" (model "FM20"), determined on the B3LYP/6-31G(d,p) level of theory in Ref. [11]. The arrows indicate atom motion along the reaction coordinate, for ^{12}C, as calculated in this work. Atom and group labeling as used in the text

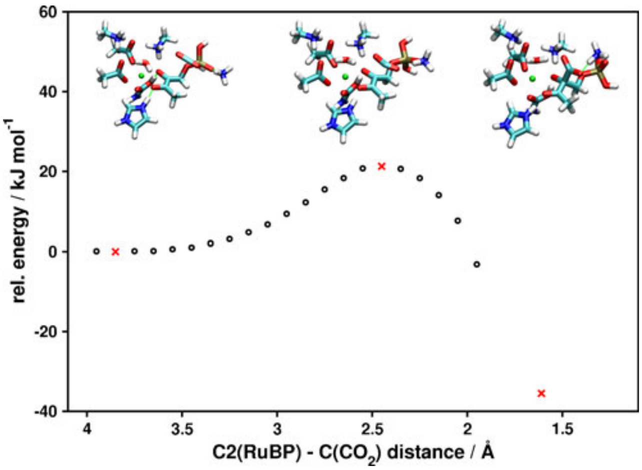


Fig. 3 Carboxylation reaction pathway of RuBisCO, in the form of a scan along the C(CO2)-C2(RuBP) distance, as obtained from a B3LYP/6-31G(d,p) calculation, using the 77-atom model of Fig. 2. The geometries of stationary points (reactant, transition, product state) are indicated by crosses. See text for details

(non-carboxylated form) directions. After obtaining the potential energy curve which is shown in Fig. 3, reoptimization of the structures closest to the two minima gave stable reactant and product geometries, as verified by normal mode analysis. Normal mode analysis at the maximum of the reaction path provided a single imaginary frequency if the same basis set and method [B3LYP/6-31G(d,p)] as in Ref. [11] was used, confirming the nature of this stationary point as transition state. In case of other methods / basis sets, the transition state had to be re-optimized using the QST3 method [17].

For all calculations, the Gaussian09 [18] program package was used. In addition to B3LYP [19] calculations with the 6-31G(d,p) basis [20], also other methods (Hartree-Fock), and basis sets [6-31G(d)] were tested. Besides gas phase (*in vacuo*) models as in Ref. [11], clusters embedded in a polarizable continuum (polarizable continuum model, PCM) [21] were also considered. Here, chloroform was used as a solvent (dielectric constant ε= 4.71) to roughly resemble the protein environment beyond those amino acid residues which are explicitly included in the quantum chemical model [22]. While an ε of 4 can be considered more close to the situation within a protein [23, 24], it has been previously reported that a slightly higher ε has almost no effect on the energies [25]. As there is no solvent defined in Gaussian09 having an exact ε of 4, we decided to use chloroform as a compromise in terms of dielectric constant and solvent probe radius.

Isotope effects were studied from normal modes at stationary points and by replacing the ^{12}C atom of reacting CO_2 with ^{13}C. Thermochemical properties were calculated in harmonic approximation, with B3LYP/6-31G(d,p) frequencies scaled by 0.983 [26]. A temperature of T=298.15 K and a pressure of 1 atm were chosen. In particular, free energy differences $\Delta G(T)=\Delta H(T)-T\ \Delta S$ were determined, where the enthalpy $\Delta H(T)=\Delta E_{el}+\Delta E_{ZPE}+\Delta\Delta E_{vib}(T)$ contains the electronic energy (and nuclear repulsion) contribution ΔE_{el}, a zero-point vibrational energy correction ΔE_{ZPE} , and the change of vibrational energy at temperature T, $\Delta\Delta E_{vib}(T)$. ΔS is an entropy difference.

Rate constants k were calculated from Eyring transition state theory as

$$k = \frac{k_B T}{h}\cdot e^{-\Delta G^{\ddagger}/k_B T}, \tag{1}$$

where k_B is Boltzmann's constant, and $\Delta G^{\ddagger}=G^{\ddagger}-$G (reactant) the activation free energy. For the carboxylation step, the reactant is the precarboxylation configuration (left cluster in Fig. 3 below). The kinetic isotope effect α is

$$\alpha = \frac{k_{12}}{k_{13}} = e^{(\Delta G^{\ddagger}_{13}-\Delta G^{\ddagger}_{12})/k_B T}, \tag{2}$$

where indices “12” and “13” refer to $^{12}CO_2$ and $^{13}CO_2$, respectively.

Tunneling contributions to the rate can be important and are estimated by a one-dimensional Wigner correction, predicting a rate enhancement factor [27]

$$\Gamma = 1+\frac{1}{24}\left(\frac{\hbar\omega^{\ddagger}}{k_B T}\right)^2 \tag{3}$$

Here, $\omega^{\ddagger}$ is the modulus of the (imaginary) TS frequency. The tunneling corrected kinetic isotope effect is $\alpha_t=\alpha\ (\Gamma_{12}/\Gamma_{13})$.

Results and discussion

Geometries, energetics and kinetics

In Fig. 3, besides the B3LYP/6-31G(d,p) reaction path for model FM20, the optimized geometries for reactant, transition, and product states of the carboxylation step are shown. Our geometries and energies (at T=0 K), are generally in good agreement with those already published [11]. (Small deviations arise from the slightly different computational protocols, and from the fact that we use frequency scaling.)

In the reactant state, the CO_2 molecule is linear, with a large distance to the C2 atom of RuBP, of 3.85 Å. In the transition state the C(CO_2)-C2(RuBP) distance is considerably shorter (2.45 Å), and CO_2 is already bent with a CO_2 bond angle of 153°. Finally, in the product configuration the C(CO_2)-C2(RuBP) bond has formed, with a C-C bond length of 1.61 Å, and a CO_2 bond angle of 127°.

In Table 1 we show energy-related data, namely energy differences ΔE_{el}, zero-point corrected energy differences $\Delta E_{el}+\Delta E_{ZPE}$, and finally, free energy differences ΔG, both for ^{12}C and ^{13}C. The differences refer to activation energies [*e.g.*, $\Delta G^{\ddagger}=G^{\ddagger}-G$(reactant)] and reaction energies [e.g., $\Delta G=G$(product)−G(reactant)], again for the "new" (FM20) model, gas phase, and B3LYP/6-31G(d,p). Also listed are computed rate constants k_{12} and k_{13}, and tunneling enhancement factors Γ. The latter were calculated from Eq. (3), with B3LYP/6-31G(d,p) transition state frequencies $\omega^{\ddagger}=2\pi c\tilde{v}^{\ddagger}$, of $\tilde{v}^{\ddagger}_{12}=122.8\text{cm}^{-1}$ and $\tilde{v}^{\ddagger}_{13}=121.0\text{cm}^{-1}$, respectively. Finally, the table shows isotope effects without (α) and with tunneling correction (α_t), and quantities related to more approximate treatments of the isotope effect (see below).

From Table 1, the following observations can be made:

(i) The *reaction free energies* ΔG are negative, indicating an exergonic, spontaneous reaction. Similarly, in Ref. [11] this reaction step was found to be exoenergetic at T=0 K within this model. Note that here we observe not only clear zero-point corrections as in Ref. [11], but considerable thermal corrections in addition.

(ii) The *activation free energies* $\Delta G^{\ddagger}$ are in the order of 40 kJ/mol, slightly higher for the heavier isotope. We find that the electronic contribution to the activation energy is about 21 kJ/mol. Zero-point energy corrections are non-negligible (in agreement with Ref. [11], more than 5 kJ/mol), but temperature and entropy contributions are more important, accounting for another 13 kJ/mol.

(iii) The larger activation free energy for the heavier isotope translates into a slightly lower reaction rate. The computed reaction rate for the carboxylation step alone is much larger than the experimental turnover rate of the RuBisCO active site, which is 1.75 s^{-1} [9]. There is of course no one-to-one correspondence between the carboxylation rate and the overall turnover, however, there are also great method and model dependencies of the carboxylation rate which can contribute to the disagreement. Nevertheless, the ratio $\alpha=k_{12}/k_{13}=1.036$ is surprisingly close to the experimental value of 1.030 for this system [4]. This value is much more insensitive to the specific reaction model and computational level (see below).

(iv) Tunneling corrections are small, in the order of one percent for absolute rates. They hardly show up in the isotope effect, and are therefore not further considered in the following.

Table 1 Quantities related to the energetics and kinetics of the carboxylation step for different isotopes (^{12}C and ^{13}C), using the "new" 77-atom model (FM20 [11]) and B3LYP/6-31G(d,p) *in vacuo*. The temperature was T=298.15 K

		$^{12}CO_2$		$^{13}CO_2$
		Energetics / (kJ/mol)		
ΔE_{el}		-35.50		-35.50
$\Delta E_{el}+\Delta E_{ZPE}$		-20.14		-20.18
ΔG		-7.06		-7.06
$\Delta E^{\ddagger}_{el}$		21.26		21.26
$\Delta E^{\ddagger}_{el}+\Delta E^{\ddagger}_{ZPE}$		26.72		26.80
$\Delta G^{\ddagger}$		39.98		40.06
		Kinetics		
k	(s^{-1})	6.16 10^5		5.95 10^5
Γ		1.0146		1.0142
$\alpha=\frac{k_{12}}{k_{13}}$			1.036	
$\alpha_t=\alpha\frac{\Gamma_{12}}{\Gamma_{13}}$			1.036	
$\tilde{v}_b$	(cm^{-1})	629.4		611.5
$\hbar({\omega_b}^{12}-{\omega_b}^{13})/2$	(J/mol)		107	
$e^{\hbar({\omega_b}^{12}-{\omega_b}^{13})/2k_BT}$			1.044	
$\frac{{\omega_b}^{12}}{{\omega_b}^{13}}$			1.029	

Microscopic origin and "robustness" of the isotope effect

Let us assume that the carboxylation step determines the isotope effect, because no other step involves the CO_2-unit directly. We can then analyze microscopic details of its origin, by referring to the eigenvector which corresponds to the imaginary frequency $i\omega^{\ddagger}$ at the transition state, as shown in Fig. 1. The arrows schematically indicate the motion of atoms along the reaction coordinate, toward reactant and product configurations. It is seen that this motion is dominated by a bending vibration of CO_2, which is needed to carboxylate RuBP.

In general, the isotope effect for the carboxylation depends on many factors. To gain basic physical insight, we note first of all that the zero-point energy of the reacting CO_2 bending mode contributes to G(reactant), but not to $G^{\ddagger}$, according to Eyring's transition state theory. Assuming in further approximation that (i) during the formation of the transition state the modes perpendicular to the reaction mode remain almost unchanged, and (ii) temperature effects cancel out when changing the isotope, one gets from Eq. (2)

$$\alpha=e^{(\Delta G^{\ddagger}_{13}-\Delta G^{\ddagger}_{12})/k_BT}\approx e^{\hbar({\omega_b}^{12}-{\omega_b}^{13})/2k_BT} \tag{4}$$

$$\approx 1+\frac{\hbar\left({\omega_b}^{12}-{\omega_b}^{13}\right)}{2k_BT}. \tag{5}$$

Here, $\hbar({\omega_b}^{12} - {\omega_b}^{13})/2$ is the difference in the zero-point energies of the $^{13}CO_2$ and $^{12}CO_2$ isotopomers in the reacting (bending) mode, which we can in further approximation set equal to the bending mode of the free CO_2 molecule.

Note that free CO_2 has two degenerate bending modes, of which only one transforms into the reaction coordinate. The second approximate Eq. (5), according to which the isotope effect depends linearly on the zero-point energy difference of the CO_2 bending mode, holds if $k_B\, T$ is much larger than $\Delta\Delta G^{\ddagger}_{13-12} = \Delta G^{\ddagger}_{13} - \Delta G^{\ddagger}_{12}$, which is the case at room temperature.

Table 1, lower part, shows the vibrational frequencies of the bending mode of free $^{12}CO_2$ and $^{13}CO_2$ obtained from a (scaled) normal mode analysis at B3LYP/6-31G(d,p) level, along with several quantities derived thereof. From the table we note that the ZPE difference reflects indeed reasonably well the changes in Gibbs free activation energies upon isotope substitution, $\Delta\Delta G^{\ddagger}_{13-12}$, as anticipated in Eq. (4). The ZPE difference of the free molecular bending mode is 107 J/mol, compared to $\Delta\Delta G^{\ddagger}_{13-12} = 86\ \mathrm{J/mol}$. The zero-point energy contribution to $\Delta\Delta G^{\ddagger}_{13-12}$ is 74 J/mol, the rest (12 J/mol) are due to temperature corrections.

As a consequence, the approximate relation for the kinetic isotope effect as suggested in Eq. (4) holds also, with α=1.044 as compared to the "exact" value of 1.036. In the high-temperature approximation (5), basically the same isotope effect (1.043) is obtained as with Eq. (4).

We finally note that an alternative picture can be used to explain the isotope effect which is based on simple Arrhenius expressions. In the Arrhenius model, rates for unimolecular reactions are given as:

$$k = \nu e^{-E_a/RT}, \tag{6}$$

where ν is the "attempt frequency", and E_a an effective activation energy.

Since in the Arrhenius model the reaction proceeds purely classical along a one-dimensional path with no perpendicular modes, E_a is the same for both isotopes and differences occur in attempt frequencies only. Assuming that the latter can be chosen as the harmonic frequency of the bending mode (of free CO_2), we get

$$\alpha \approx \frac{{v_b}^{12}}{{v_b}^{13}} = \frac{{\omega_b}^{12}}{{\omega_b}^{13}} \tag{7}$$

in this case. The result is α=1.029 (Table 1), in accidentally good agreement with the "exact" value (2).

The above analysis can - and should - be criticized in many respects, mostly because of the inaccuracy of our model. First of all, the choice of the cluster for the "new mechanism" suggested in Ref. [11] has a large impact on relative energies as demonstrated in that reference. Secondly, there is still the more established, "old" model according to which the water molecule at Mg^{2+} is replaced by CO_2 before carboxylation of RuBP. Third, method and basis set dependencies should be checked. Fourth, the carboxylation step may not be decisive alone, nor is it even fully clear that the carboxylated product is a true intermediate (*i.e.*, the reaction might be concerted with the hydration step). Last but not least a larger portion of the protein environment may be important and, related to that, it may be necessary to sample the thermal protein environment (*e.g.*, in a finite-*T* QM/MM setup) in order to arrive at reliable activation free energies.

To address some of these problems at least in an exploratory fashion, we have studied cluster models for the "old mechanism" with up to 102 atoms (see also Ref. [10]), which were based on crystal structures for the system at hand [6, 7]. Also, another basis set [6-31G(d)], another electronic structure method (Hartree-Fock, HF), and the embedding in a polarizable continuum as described above (for the "new mechanism") were tested. An outcome of all of these studies is that indeed the absolute activation energies for the carboxylation step depend critically on the computational protocol / method, and so do absolute rates. However, in all cases (where a true transition state could be found), it was observed that the isotope ratio α is remarkably robust, and that the CO_2 bending mode in the binding pocket plays a decisive role. For example, using (i) HF/6-31G(d,p) and the 77- atom "new" model, (ii) HF/6-31G(d,p) and the 102-atom "conventional" model, or (iii) B3LYP/6-31G(d,p)/PCM and the 77-atom "new" model, the computed isotope effects were 1.049, 1.042, and 1.049, respectively. At the same time, the activation free energies varied by up to a factor of two, and absolute rates by several orders of magnitude. The latter isotope effects are too large compared to experiment and the FM20 [B3LYP/6-31G(d,p)] model, but qualitatively correct and originating from differences in the CO_2 bending mode. This clearly shows that in order to compute the isotope effect for the turnover rate for RuBisCO in various systems as accurately as it can be measured [4, 28], it will certainly be necessary to go beyond the methodology / models proposed here: B3LYP activation energies are reported to be too low in general [29], and *ab initio* HF activation energies too high due to the lack of electron correlation. In the future, more advanced studies systematically exploring the effect of different computational methods should be performed, e.g., using DFT functionals such as M06-2X [30]. The prominent role of the of the CO_2 bending mode during C-C bond formation, however, seems to be quite generic.

Conclusions and outlook

We have studied the kinetic carbon isotope effect in the carboxylation step of the RuBisCO reaction. Absolute activation and reaction energies and reaction rates depend sensitively on model type, size, and computational method and require further refinement. The isotope effect itself is traced back to the CO_2 bending vibration, which is the reacting mode in the transition state during the formation of a C-C-bond. More precisely, the effect is due to a loss of zero-point energy for the vibrating CO_2 in the reactant configuration, leading to a slightly increased activation energy for $^{13}CO_2$. We believe that the essential physics is independent of details of the protein environment, while the exact magnitude of the kinetic isotope effect is clearly not [28].

For a complete assessment of isotope discrimination, the reactants' binding process to the protein and the product release should be considered. In general, for fully quantitative predictions, larger models of a thermal protein environment must be studied, and (thermo-)dynamic effects. For basic insight, however, the QM cluster models employed here seem to be valuable.

Acknowledgments We thank Dr. Tillmann Klamroth and Karl Zenichowski for constant technical and conceptional advice. J.G. was funded by the project GoFORSYS.

References

1. Cleland WW, Andrews TJ, Gutteridge S, Hartman FC, Lorimer GH (1998) Mechanism of Rubisco: The carbamate as general base. Chem Rev 98:549–561. doi:10.1021/cr970010r
2. Andersson I, Backlund A (2008) Structure and function of Rubisco. Plant Phys Biochem 46:275–291. doi:10.1016/j.plaphy.2008.01.001
3. Andersson I (2008) Catalysis and regulation in Rubisco. J Exp Bot 59:1555–1568. doi:10.1093/jxb/ern091
4. Roeske CA, O'Leary MH (1984) Carbon isotope effects on the enzyme catalyzed carboxylation of ribulose bisphosphate. Biochemistry 23:6275–6284. doi:10.1021/bi00320a058
5. McNevin DB, Badger MR, Kane HJ, Farquhar GD (2006) Measurement of (carbon) kinetic isotope effect by Rayleigh fractionation using membrane inlet mass spectrometry for CO (2)-consuming reactions. Func Plant Biol 33:1115–1128. doi:10.1071/FP06201
6. Andersson I (1996) Large structures at high resolution: The 1.6 angstrom crystal structure of spinach ribulose-1,5-bisphosphate carboxylase/oxygenase complexed with 2-carboxyarabinitol bisphosphate. J Mol Biol 259:160–174. doi:10.1006/jmbi.1996.0310
7. Taylor TC, Andersson I (1997) Structure of a Product Complex of Spinach Ribulose-1,5-bisphosphate Carboxylase/Oxygenase. Biochemistry 36:4041–4046. doi:10.1021/bi962818w
8. Mizohata E, Matsumura H, Okano Y, Kumei M, Takuma H, Onodera J, Kato K, Shibata N, Inoue T, Yokota A, Kai Y (2002) Crystal structure of activated ribulose-1,5-bisphosphate carboxylase/oxygenase from green alga Chlamydomonas reinhardtii complexed with 2-carboxyarabinitol-1,5-bisphosphate. J Mol Biol 316:679–691. doi:10.1006/jmbi.2001.5381
9. Farquhar GD (1979) Models describing the kinetics of ribulose biphos-phate carboxylase-oxygenase. Arch Biochem Biophys 193:456–468. doi:10.1016/0003-9861(79)90052-3
10. Mauser H, King WA, Gready JE (2001) CO2 fixation by Rubisco: Computational dissection of the key steps of carboxylation, hydration, and C-C bond cleavage. J Am Chem Soc 123:10821–10829. doi:10.1021/ja011362p
11. Kannappan B, Gready JE (2008) Redefinition of Rubisco carboxylase reaction reveals origin of water for hydration and new roles for active-site residues. J Am Chem Soc 130:15063–15080. doi:10.1021/ja803464a
12. Witzel F, Götze J, Ebenhöh O (2010) Slow deactivation of ribulose 1,5- bisphosphate carboxylase/oxygenase elucidated by mathematical models. FEBS J 277:931–950. doi:10.1111/j.1742-4658.2009.07541.x
13. Tapia O, Andres J, Safont VS (1995) Transition structures in vacuo and the theory of enzyme catalysis, Rubisco's catalytic mechanism: a paradigmatic case? J Mol Struc (THEOCHEM) 342:131–140
14. Safont VS, Oliva M, Andres J, Tapia O (1997) Transition structures of carbon dioxide fixation, hydration and C2 inversion for a model of Rubisco catalyzed reaction. Chem Phys Lett 278:291–296. doi:10.1016/S0009-2614(97)01001-4
15. Oliva M, Safont VS, Andres J, Tapia O (2001) Transition state structures and intermediates modeling carboxylation reactions catalyzed by rubisco. a quantum chemical study of the role of magnesium and its coordination sphere. J Phys Chem A 105:9243–9251. doi:10.1021/jp0113533
16. Tapia O, Fidder H, Safant VS, Oliva M, Andres J (2002) Enzyme catalysis: Transition structures and quantum dynamical aspects: Modeling rubisco's oxygenation and carboxylation mechanisms. Internat J Quant Chem 88:154–166. doi:10.1002/qua.10116
17. Peng C, Ayala PY, Schlegel HB (1996) Using redundant internal coordinates to optimize equilibrium geometries and transition states. J Comput Chem 17:49–56. doi:10.1002/(SICI)1096-987X(19960115)17:1<49::AID-JCC5>3.0.CO;2-0
18. Frisch MJ, Trucks GW, Schlegel HB et al (2004) Gaussian 09, revision a.02. Gaussian, Inc, Wallingford, CT
19. Becke AD (1988) Density-functional exchange-energy approximation with correct asymptotic-behavior. Phys Rev A 38:3098–3100. doi:10.1103/PhysRevA.38.3098
20. Ditchfield R, Hehre WJ, Pople JA (1971) Self-consistent molecular orbital methods.9. Extended Gaussian-type basis for molecular orbital studies of organic molecules. J Chem Phys 54:724–728. doi:10.1063/1.1674902
21. Tomasi J, Mennucci B, Cammi R (2005) Quantum mechanical continuum solvation models. Chem Rev 105:2999–3093. doi:10.1021/cr9904009
22. Sousa SF, Fernandes PA, Ramos MJ (2007) The carboxylate shift in zinc enzymes: A computational study. J Am Chem Soc 129:1378–1385. doi:10.1021/ja067103n
23. Liao RZ, Yu JG, Himo F (2009) Reaction mechanism of the dinuclear zinc enzyme N-acyl-L-homoserine lactone hydrolase: A quantum chemical study. Inorg Chem 48:1442–1448. doi:10.1021/ic801531n
24. Liao RZ, Yu JG, Himo F (2010) Reaction mechanism of the trinuclear zinc enzyme phospholipase C: A density functional theory study. J Phys Chem B 114:2533–2540. doi:10.1021/jp910992f
25. Blomberg MRA, Siegbahn PEM, Babcock GT (1998) Modeling electron transfer in biochemistry: A quantum chemical study of charge separation in Rhodobacter sphaeroides and photosystem II. J Am Chem Soc 120:8812–8824. doi:10.1021/ja9805268
26. Scott AP, Radom L (1996) Harmonic vibrational frequencies: An evaluation of Hartree-Fock, Moller-Plesset, quadratic configuration interaction, density functional theory, and semiempirical scale factors. J Phys Chem 100:16502–16513. doi:10.1021/jp960976r

27. Tanaka N, Xiao Y, Lasaga AC (1996) Ab initio study on carbon Kinetic Isotope Effect (KIE) in the reaction of CH4+Cl. J Atmosph Chem 23:37–49. doi:10.1007/BF00058703

28. Tcherkez GGE, Farquhar GD, Andrews TJ (2006) Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. Proc Natl Acad Sci 103:7246–7251. doi:10.1073/pnas.0600605103

29. Parthiban S, de Oliveira G, Martin JML (2001) Benchmark ab initio energy profiles for the gas-phase S(N)2 reactions Y-+ CH3X -> CH3Y+X- (X, Y=F, Cl, Br). Validation of hybrid DFT methods. J Phys Chem A 105:895–904. doi:10.1021/jp0031000

30. Zhao Y, Truhlar DG (2008) The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other functionals. Theor Chem Acc 120:215–241. doi:10.1007/s00214-007-0310-x

J Mol Model (2012) 18:1885–1906
DOI 10.1007/s00894-011-1201-6

ORIGINAL PAPER

Molecular dynamics simulations of the Bcl-2 protein to predict the structure of its unordered flexible loop domain

Pawan Kumar Raghav · Yogesh Kumar Verma · Gurudutta U. Gangenahalli

Received: 2 October 2010 / Accepted: 24 July 2011 / Published online: 25 August 2011

Abstract B-cell lymphoma (Bcl-2) protein is an anti-apoptotic member of the Bcl-2 family. It is functionally demarcated into four Bcl-2 homology (BH) domains: BH1, BH2, BH3, BH4, one flexible loop domain (FLD), a transmembrane domain (TM), and an X domain. Bcl-2's BH domains have clearly been elucidated from a structural perspective, whereas the conformation of FLD has not yet been predicted, despite its important role in regulating apoptosis through its interactions with JNK-1, PKC, PP2A phosphatase, caspase 3, MAP kinase, ubiquitin, PS1, and FKBP38. Many important residues that regulate Bcl-2 anti-apoptotic activity are present in this domain, for example Asp34, Thr56, Thr69, Ser70, Thr74, and Ser87. The structural elucidation of the FLD would likely help in attempts to accurately predict the effect of mutating these residues on the overall structure of the protein and the interactions of other proteins in this domain. Therefore, we have generated an increased quality model of the Bcl-2 protein including the FLD through modeling. Further, molecular dynamics (MD) simulations were used for FLD optimization, to predict the flexibility, and to determine the stability of the folded FLD. In addition, essential dynamics (ED) was used to predict the collective motions and the essential subspace relevant to Bcl-2 protein function. The predicted average structure and ensemble of MD-simulated structures were submitted to the Protein Model Database (PMDB), and the Bcl-2 structures obtained exhibited enhanced quality. This study should help to elucidate the structural basis for Bcl-2 anti-apoptotic activity regulation through its binding to other proteins via the FLD.

Keywords Bcl-2 · Loop modeling · MD simulations · Apoptosis · Flexible loop domain · Essential dynamics

Abbreviations

1GJH	Bcl-2 isoform 2 structure (from NMR)
2XA0	Model of complex of Bcl-2 and Bax peptide (from X-ray diffraction)
1G5J	Model of complex of Bcl-X_L and Bad peptide (from NMR)
SM	Swiss-Model generated model of Bcl-2 based on the 1GJH template
CW	ClustalW alignment based Bcl-2 model generated by MODELLER
MOD	MODELLER-generated model based on the 1GJH template and obtained using the same alignment as used by SM
MODLOPT	MODELLER loop-optimized model
PM0076467	PMDB ID of the MD-simulated average structure
PM0077081- PM0077103	23-Structure ensemble resulting from clustering
MD	Molecular dynamics
ED	Essential dynamics
PMDB	Protein Model Database
FLD	Flexible loop domain

Introduction

The FLD region (amino acids 34–92) of Bcl-2 (239 amino acids) lies between the BH4 and BH3 domains, lacks a defined structure, and is not structurally conserved among Bcl-2 family members (Fig. 1) [1]. The BH domains are responsible for Bcl-2's anti-apoptotic function, which is

P. K. Raghav · Y. K. Verma · G. U. Gangenahalli (✉)
Stem Cell and Gene Therapy Research Group,
Institute of Nuclear Medicine and Allied Sciences (INMAS),
Lucknow Road,
Timarpur, Delhi 110054, India
e-mail: gugdutta@rediffmail.com

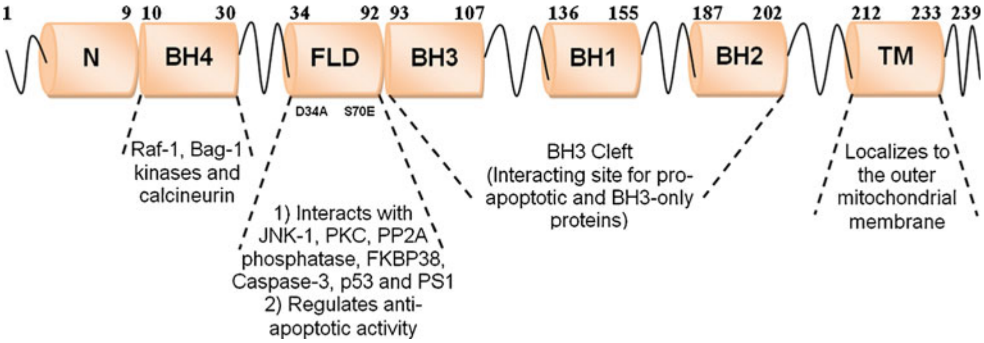

1
9 10
30
34
92 93
107
136
155
187
202
212
233 239
N
BH4
FLD
BH3
BH1
BH2
TM
D34A
S70E
Raf-1, Bag-1 kinases and calcineurin
1) Interacts with JNK-1, PKC, PP2A phosphatase, FKBP38, Caspase-3, p53 and PS1
2) Regulates anti-apoptotic activity
BH3 Cleft (Interacting site for pro-apoptotic and BH3-only proteins)
Localizes to the outer mitochondrial membrane

Fig. 1 Functional division of the Bcl-2 protein into domains (represented as *tubes*), *N-* N-terminal, *BH-* Bcl-2 homology, *FLD-* flexible loop domain and *TM-* transmembrane, connected through the flexible region represented as *wires*. The interacting proteins and functions of the domains of Bcl-2 are shown by *dotted lines*. Important residues such as Asp34 and Ser70 lie in the FLD region; the substitution of these residues (i.e., Asp34Ala and Ser70Glu) upregulates the anti-apoptotic activity of Bcl-2

linked with its homodimerization and its heterodimerization with its own family members and those of other families, whereas FLD regulates the activity of Bcl-2. The FLD contains phosphorylation sites—Thr56, Thr69, Ser70, Thr74, and Ser87—that are essential for the regulation of Bcl-2 activity. By using 2-D peptide mapping and sequencing, the residues Thr56, Thr74, and Ser87 were found to be phosphorylated in response to a microtubule damaging agent (paclitaxel) that also arrest T cells at G2/M phase of cell cycle. Changing these sites to Ala led to enhanced survival following death signal as well as paclitaxel treatment [2]. These residues also constitute a ubiquitin-dependent cleavage/MAP kinase site. It was discovered that there was cleavage of Bcl-2 at this site upon TNF-α induced cell death in endothelial cells due to a ubiquitin-dependent proteosome complex. The phosphorylation of these residues has been shown to abolish Bcl-2 degradation [3]. On the other hand, single phosphorylation of Bcl-2 at Ser70 is essential for full and potent cell survival activity. The residue Ser70 was shown to be phosphorylated by IL-3, PKC and EPO to maintain normal cellular homeostasis [4], whereas the Ser70Ala mutant is not phosphorylated after IL-3/Bryo stimulation, and is unable to prolong cell survival upon either IL-3 deprivation or etoposide treatment. Ser70Glu substitution also suppresses etoposide-induced apoptosis more potently than wild-type Bcl-2 [5]. Furthermore, the DNA damage induced by p53–Bcl-2 binding was shown to be associated with the weaker Bcl-2–Bax interaction and increased apoptotic cell death in a mechanism regulated by FLD. This demonstrates that FLD is also involved in regulating the binding of Bcl-2 with Bax and indirectly regulating such interactions with pro-apoptotic proteins [6]. Additionally, FLD contains a caspase 3 (apoptosis effecter protease) recognition and cleavage site at Asp34. Cleavage at Asp34 by caspase 3 renders Bcl-2 unable to inhibit apoptosis. The mutation of Asp34Ala makes Bcl-2 resistant to caspase 3 mediated cleavage and hence enhances anti-apoptotic activity [7]. The FLD-deleted mutant of Bcl-2 displays enhanced ability to inhibit apoptosis without impairing Bcl-2 heterodimerization with pro-apoptotic proteins. Full-length Bcl-2 was found to be ineffective at preventing anti-IgM induced cell death of an immature B cell line (WEHI-231). In contrast, this mutant protected WEHI-231 cells from death [8]. Moreover, a molecular interaction between FKBP38 and Bcl-2 has been shown to occur through the unstructured loop of Bcl-2, and this appears to regulate phosphorylation in the loop [9]. Normally proteins are degraded by cellular proteases; nonetheless, the flexible loop in Bcl-2 has been shown to shield or protect it from rapid degradation by cellular proteases due to the presence of a random coil structure that exhibits a long half-life [8, 10, 11].

All these studies indicate that the FLD plays an important role in providing stability to the Bcl-2 protein and regulation of its anti-apoptotic activity. Understanding the interactions of various proteins with FLD would likely help us to tweak apoptosis signaling and the treatment of various malignancies [12] in which Bcl-2 expression is not under control, such as acute myeloid leukemia (AML) [13]. Therefore, the FLD represents an attractive target for drugs that modulate protein–protein interactions, and studies of the FLD would likely facilitate drug discovery and improve our understanding of human diseases.

No data on the 3D conformation of the FLD are currently available; nevertheless, the NMR structure of Bcl-2, in which the FLD is replaced with residues of the Bcl-X_L protein, is available [14]. In our study, sequence-based results showed that the FLD is an extensively disordered region. In order to predict the conformation of the FLD, we elucidated the structure of the Bcl-2 protein using Swiss-Model [15] and MODELLER software [16]. Based on the stereochemical parameters, we observed that MODELLER generated a refined model. Further, the FLD (as generated by MODELLER) was optimized by the loop modeling method of MODELLER, and then energy minimization was performed. Since the accurate prediction of the secondary and tertiary stable structure and dynamics of the FLD cannot be achieved through experimental measurements [17], we used MD simulation to predict a putative conformation and the flexibility of the FLD. The MD simulation of the energy-minimized Bcl-2 model was carried out using GROMACS (Groningen Machine for Chemical Simulations) [18]. The average structure of the Bcl-2 protein was obtained from the 15 ns MD-simulated trajectory. Further, the predicted average model was energy minimized and validated by stereochemical and overall quality checks. Subsequently, the average model was checked for its ability to bind with pro-apoptotic peptides by docking. Clustering was performed to ensemble the structures based on the RMSD along the MD simulation trajectory. The electrostatic behavior (dipole moment) was predicted for the entire MD simulation trajectory, and was found to be in accordance with the X-ray structure. ED was used to reduce the dimensionality and to predict the essential subspaces for large collective motions that are relevant when generating a biologically functional Bcl-2 model. Our results show that the model obtained has enhanced quality and may be useful for studies focusing on, for example, mutagenesis and protein–protein docking. The average model and ensemble were submitted to the PMDB [19], and are available for further analysis.

Computational methods

Sequence analysis of Bcl-2 protein

The human Bcl-2 sequence (Genbank I.D. 231632; Swissprot I. D. P104152) containing 239 amino acids was used for sequence analysis and modeling. BLAST [20] identified the sequences homologous to Bcl-2 in different organisms. PDBblast was used to search for Bcl-2 homologs with solved 3D structures. The BLAST parameters of an expect value of 10, a hitlist size of 100, a threshold of 11, and a word size of 2 were used, and the BLOSUM62 matrix was employed. Unordered regions in the protein were predicted by the DISOPRED server [21]. The IUPred method [22] was used to obtain the specific amino acid composition of the disordered region of the FLD, which does not form a stable, well-defined structure.

Predicting the structure of Bcl-2 and FLD optimization

The Swiss-Model server (http://swissmodel.expasy.org) and the MODELLER program were used to generate the 3D model of Bcl-2. The Swiss-Model automated modeling mode constructed the Bcl-2 model (*SM*) based on the 1GJH template (73.171% sequence identity). MODELLER 9v7 generated the 3D structure of the Bcl-2 protein based on the 1GJH template by satisfying spatial restraints for the aligned regions.

Two models were generated using two different alignments between the Bcl-2 sequence and the 1GJH template by MODELLER. The first model (MOD) was obtained by manual alignment (similar to the alignment used to generate the SM model). The second model (CW) was obtained using ClustalW [23] PIR format alignment between the Bcl-2 and 1GJH template. This alignment was generated using the EBLOSUM 62 matrix [24] with a gap penalty of 10 and an extend penalty of 0.5. The first model (MOD) was subjected to refinement of the FLD using a loop optimization protocol (the loopmodel class of the DOPE-based method) followed by MD simulations (using the conjugate gradient optimization method) at the temperatures 150, 300, 400, 800 and 1000 K using MODELLER. An initial loop conformation (FLD) for residues 34–92 of the MOD model was generated using a fast loop refinement method by simply positioning the atoms of the FLD, uniformly spaced, on the line that connects the main-chain carbonyl oxygen and the amide nitrogen atoms of the N- and C-terminal anchor regions. This MODELLER loop-optimized (MODLOPT) structure was further energy minimized (globally) using the GROMOS96 43B1 force field [25] by Swiss-Pdb viewer [26]. Pymol [27] was employed to calculate structural alignment differences between the energy-minimized MODLOPT and SM models. The modeled structures were validated by ProSA-web [28] and SAVes server (http://nihserver.mbi.ucla.edu/SAVES/) using the PROCHECK [29], ERRAT [30] and WHAT CHECK [31] programs.

MD simulations

MD simulation of the energy-minimized MODLOPT model was performed in order to predict the native conformation of Bcl-2 and the structural changes induced by the FLD in its domains. We used GROMACS package 4.0 [32], which is a versatile collection of programs and libraries for simulating molecular dynamics and subsequently analyzing trajectory data. The simulations were carried out on a single PC (3.40 GHz Core 2 Duo processor, Pentium IV, 4 GB RAM; Hewlett Packard) running the Windows Vista operating system (using Cygwin). The GROMOS 96 53a6 [33, 34] force field including all hydrogens, along with a simple point-charge (SPC) water model [35], was used for energy minimization. The pre-equilibrated [36] SPC water was added to an octahedral box, and the protein was then placed in the center of the box. 16809 solvent molecules were embedded into the box, which extended at least 9 Å from the Bcl-2 protein to the edge of the box. Two Na^+ ions were added to the solvent to neutralize the charges on the Bcl-2 protein. The protein and nonprotein groups were energy minimized with a tolerance of 2000 kJ mol^{-1} nm^{-1} using the steepest descent method for 500 steps. All bonds were constrained using the LINCS algorithm [37], and the simulation was performed under NPT conditions, using the v-rescale coupling algorithm [38] and the Parrinello–Rahman coupling algorithm [39], which stabilized the temperature and pressure (P=1 bar, τ_P=0.1 ps; T=300 K, τ_T=0.1 ps). A smooth particle mesh Ewald (PME) method [40] was used with a cut-off of 1.4 nm for electrostatic [41] and van der Waals (vdW) [41–43] interactions. The electrostatic interactions were calculated with PME using a grid spacing of 0.12 nm. Periodic boundary conditions (PBC) were employed to eliminate surface effects [44]. The final MD simulations were carried out with a time step of 3 fs [45, 46] and without any position restraints [47]; 5,000,000 steps were performed for a total of 15 ns.

Analysis of the MD simulations

All analyses were carried out using programs included in GROMACS (version 4.0.7), VMD [48], and Pymol. Trajectories were subjected to energy analysis, global structural analysis (measuring the radius of gyration, R_g), and analyses of the RMSD (root mean square deviation) after least-squares fitting to the protein atoms except for hydrogens, secondary structure content, solvent accessibility, and intramolecular hydrogen bonding. The flexible regions and the stability of the Bcl-2 protein were predicted via the RMSF (root mean square fluctuation). The contact map was calculated using the minimum distance matrix in order to identify the native contacts. The solvent-accessible surface area (SASA) was calculated for the FLD residues. Salt bridges between oppositely charged residues in the FLD within a minimum cutoff distance of 0.5 nm were investigated. Secondary structure analysis was performed using DSSP [49]. The average structure was generated, energy minimized (using the steepest descent method), and validated using SAVes and ProSA-web before submitting it to the PMDB (id: PM0076467). ProSA-web was used to evaluate the stereochemical errors in and the quality of the model. The Z-score was measured to check the compatibility between the model's sequence and structure. The linkage method was used by the clustering tool (g_cluster) of GROMACS to generate an ensemble of structures. The secondary structures of the experimental and the PM0076467-superimposed models were predicted by the STRIDE [50] program. The electrostatic behavior between charged residues was calculated by g_dipoles.

Autodock 4.2 [51] was employed to generate the Bax (peptide)–Bcl-2 and Bad (peptide)–Bcl-2 complexes. AutoDockTools (ADT) was used to add hydrogen bonds (using AD4 atom types) to the peptides/proteins and to assign Gasteiger charges to 32 active torsions of the peptides. As our aim was to assess the binding efficacy of the BH3 receptor cleft, we made the search space large enough to include the whole BH3 cleft, and increased the exhaustiveness using affinity grids of 78×76×58 points and a spacing of 0.375 Å around the protein (via Autogrid). The Lamarckian genetic algorithm [52] was used for the conformational search. Each Lamarckian job consisted of 50 runs. The initial population consisted of 150 structures, and the maximum number of energy evaluations and generations was 2,500,000. The default values were used for the remaining parameters. The docking poses and hydrogen bonds were visualized via Accelrys Discovery Studio Visualizer 2.0 (http://accelrys.com/).

Essential dynamics analysis

The essential degrees of freedom (essential subspace) of Bcl-2 were extracted from the trajectories according to the ED method used (covariance analysis or principal component analysis) [53–56]. The ED method involves constructing the covariance matrix in order to observe the fluctuations in the coordinates of Bcl-2. Correlated motions were observed during the MD trajectories through the eigenvectors of the non-mass-weighted covariance matrix (C) for atomic position fluctuations. Before constructing C, the overall rotation and translation was removed to allow the visualization of internal motion. This was achieved by performing least squares fitting to the average structure based on the C_α coordinates. After the fitting procedure, the internal motions described by the trajectory $x(t)$ and the covariance matrix C were constructed from the coordinates of the positions of the C_α atoms:

$$C_{ij} = 1/S \sum_t \{x_i(t) - < x_i >\}\{x_j(t) - < x_j >\} \qquad (1)$$

where S is the total number of configurations, t=1, 2, … S; $x_i(t)$ are the position coordinates, with i=1, 2, ….3N; N is the number of atoms from which C is constructed, and$<x_i>$ is the average for coordinate i over all configurations [53].

The covariance matrix (621×621) was diagonalized to obtain the eigenvectors and eigenvalues that provide information about the correlated motions and overall flexibility throughout the Bcl-2 protein. The eigenvectors were then sorted according to their eigenvalues in descending order. Usually, the first ten eigenvectors were sufficient to describe almost all of the conformational subspace accessible to the protein. Principal component analysis was used to identify the essential subspaces explored by the simulations of the Bcl-2 protein that likely indicate changes in the BH3 cleft. The dimensionality of the essential subspace was monitored by noting the fraction of total motion described by the reduced subspace, and this was computed as the sum of the eigenvalues relative to the included eigenvectors.

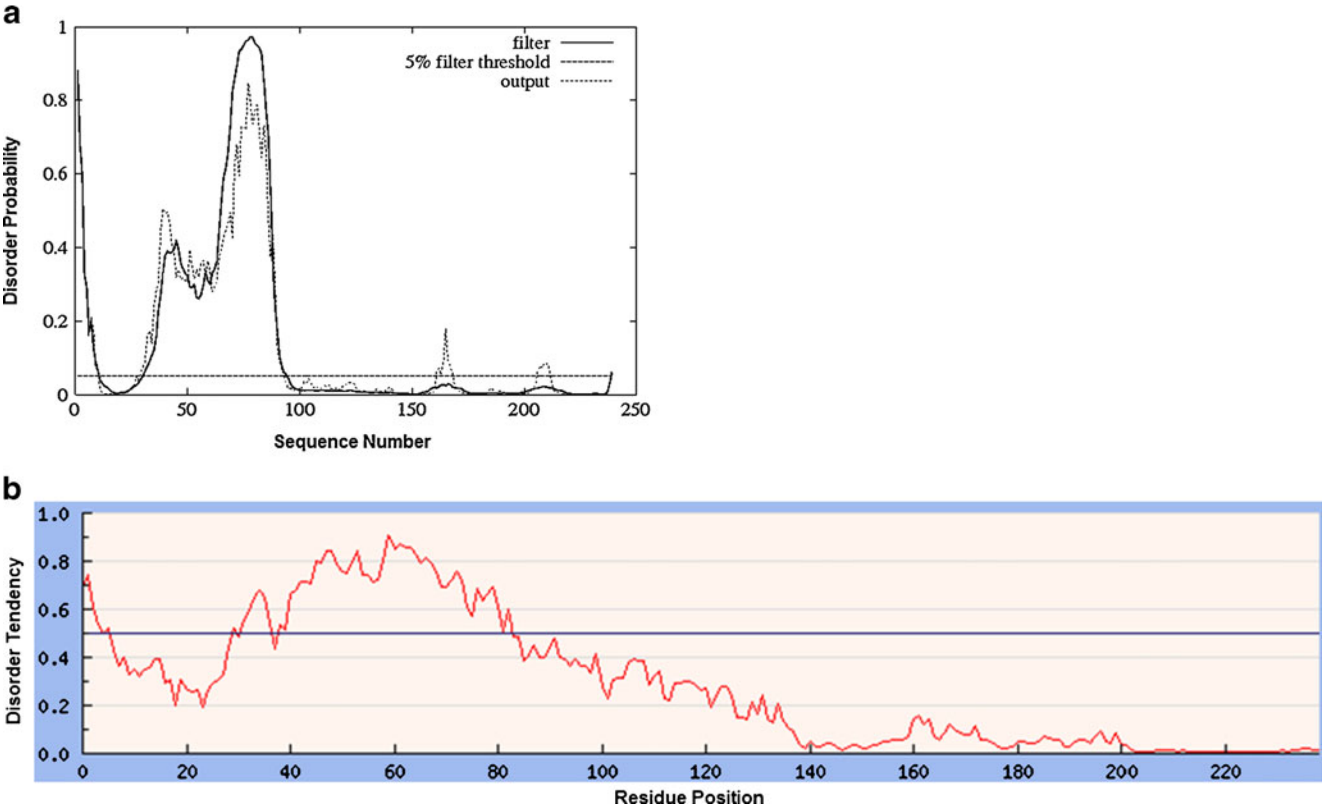


Fig. 2 **a** Plot of the disorder profile, which shows the probability of a disordered region in each residue, (*black line*). **b** The predicted tendency to be disordered (*red line*) shows large peaks for the FLD and N-terminal regions

Results and discussion

FLD disorder prediction

PDBblast gave 55 hits, among which the 1GJH (human Bcl-2, isoform 2, chain_A, NMR structure) sequence showed 73% sequence identity, 74% positives, and 19% gaps, with a score of 301 bits, an e-value of 2×10^{-82}, and a query coverage of 86%. We observed maximum sequence similarity to FLD only with the 1GJH (166 residues) template. No hits were obtained with individual pBLAST and PSI BLAST searches for the FLD (59 residues). In order to find a match for the FLD, we used the PS^2 [57] and HHpred server [58]. However, neither of these servers found a suitable template in their structure databases. These results revealed that the FLD is an unordered region. In addition, the FLD region was found to be largely unordered using two specific structural tests (DISOPRED and IUPred). The DISOPRED software predicted unordered regions, specifically in FLD (residues 34–92) and at the N-

Table 1 PROCHECK Ramachandran distributions, ERRAT overall quality factors, Verify3D scores, WHAT CHECK Z-scores, and ProSA-web Z-scores for the 1GJH, SW, CW, MOD, MODLOPT, and PM0076467 models

3D model	Ramachandran statistics				ERRAT (%)	Verify3D (%)	WHAT CHECK Z-score	ProSA-web Z-score
	Core (%)	Allowed (%)	Generous (%)	Disallowed (%)				
1GJH	73.6	24.3	1.4	0.7	79.355	96.36	−7.098	−6.61
SM	72.9	23.5	2.4	1.2	55.330	91.26	−5.909	−2.07
CW	84.5	10.1	4.2	1.2	42.347	97.34	−1.182	−5.97
MOD	88.7	9.5	1.2	0.6	36.788	99.04	−1.171	−6.52
MODLOPT	85.7	12.5	1.8	0.0	46.821	81.25	−2.732	−7.10
PM0076467*	81.5	16.7	1.2	0.6	80.233	98.08	−2.404	−6.63

* MD-simulated average structure

Fig. 3 **a** Manually aligned Bcl-2 (207 residues) and 1GJH (166 residues) sequences in PIR format. This alignment was used by MODELLER to generate MOD. **b** PIR-format alignment of Bcl-2 (207 residues) and 1GJH (166 residues), as generated using the Gonnet matrix in ClustalW

a

```
>P1;Bcl-2
MAHAGRTGYDNREIVMKYIHYKLSQRGYEWDAGDVGAAPPGAAPAPGIFSSQPGHTPHPA
ASRDPVARTSPLQTPAAPGAAAGPALSPVPPVVHLTLRQAGDDFSRRYRRDFAEMSSQLH
LTPFTARGRFATVVEELFRDGVNWGRIVAFFEFGGVMCVESVNREMSPLVDNIALWMTEY
LNRHLHTWIQDNGGWDAFVELYGPSMR

>P1;1GJH
--HAGRTGYDNREIVMKYIHYKLSQRGYEWDAGD---------------DVEENRTEAPE
GTESEV--------------------------VHLTLRQAGDDFSRRYRRDFAEMSSQLH
LTPFTARGRFATVVEELFRDGVNWGRIVAFFEFGGVMCVESVNREMSPLVDNIALWMTEY
LNRHLHTWIQDNGGWDAFVELYGPSMR
```

b

```
>P1;Bcl-2
MAHAGRTGYDNREIVMKYIHYKLSQRGYEWDAGDVGAAPPGAAPAPGIFSSQPGHTPHPA
ASRDPVARTSPLQTPAAPGAAAGPALSPVPPVVHLTLRQAGDDFSRRYRRDFAEMSSQLH
LTPFTARGRFATVVEELFRDGVNWGRIVAFFEFGGVMCVESVNREMSPLVDNIALWMTEY
LNRHLHTWIQDNGGWDAFVELYGPSMR

>P1;1GJH
--HAGRTGYDNREIVMKYIHYKLSQRGYEWDAG---------------------------
---DDVEEN---RTEAPEGTES--------EVVHLTLRQAGDDFSRRYRRDFAEMSSQLH
LTPFTARGRFATVVEELFRDGVNWGRIVAFFEFGGVMCVESVNREMSPLVDNIALWMTEY
LNRHLHTWIQDNGGWDAFVELYGPSMR
```

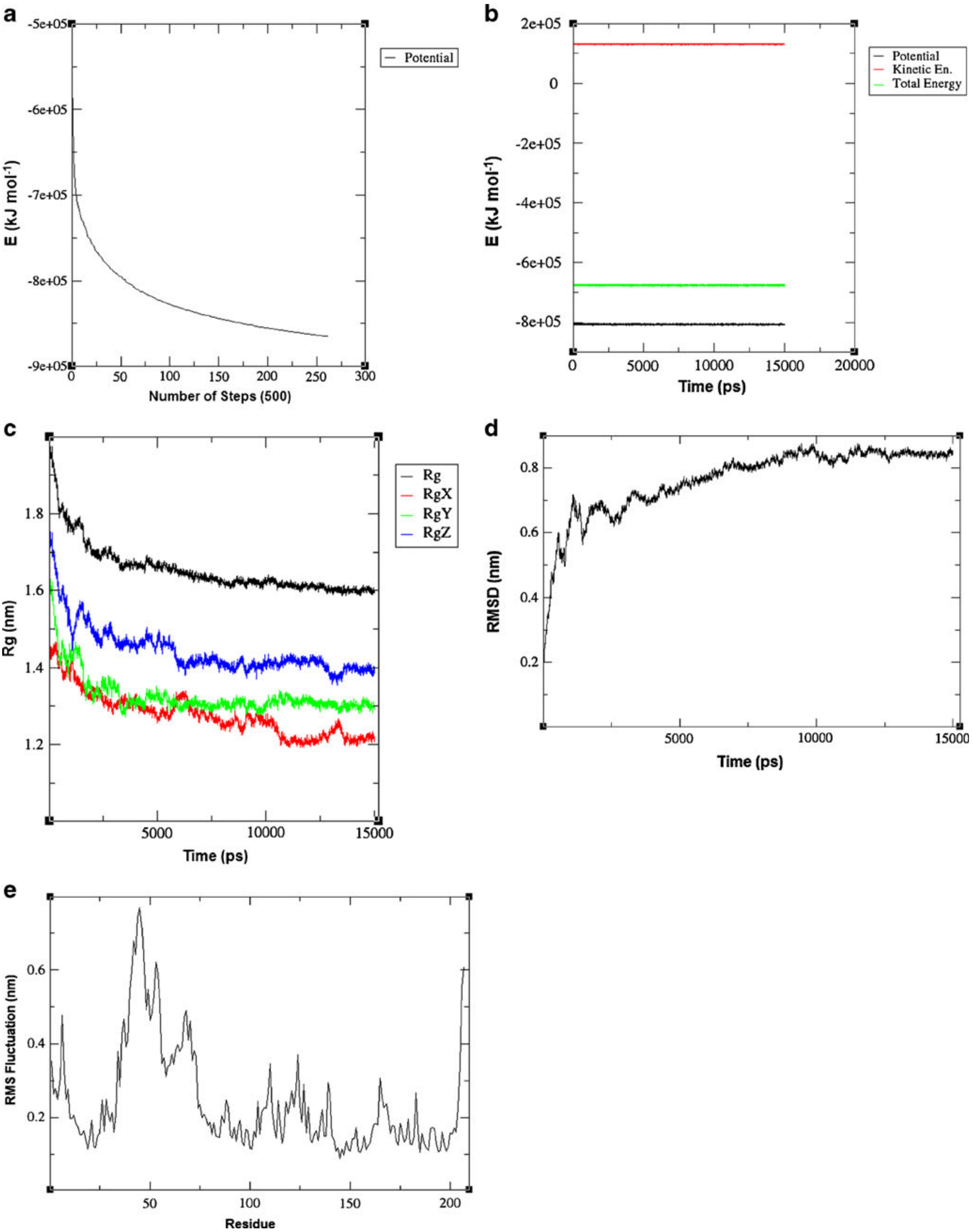
terminal domain (residues 1–9) (Fig. 2a). The DISOPRED program predicted that the Bcl-2 protein is unable to yield a good consensus for the FLD segment. The scattered output curve exhibits the presence of non-alpha and non-beta regions in the FLD zone. IUPred predicted the tendency for disorder in the Bcl-2 protein, and indicated that all four BH domains were well structured. The FLD residues show a high tendency to be disordered, with a probabilistic score of >0.5 (for residues 32–83) (Fig. 2b), but as this score is <1.0, the region does not appear to be completely disordered [59].

Molecular modeling and optimization

Upon aligning the 1GJH sequence with its PDB file, we noted differences in the FASTA format and coordinate files for the first two residues (i.e., Met1 and Ala2). So, we manually replaced these residues with gaps in the PIR-format alignment file. After comparing the two different models (MOD and CW) (Table 1), we observed that manual alignment (Fig. 3a) generates a better model (MOD) than that (CW) generated using default alignment with the ClustalW program (Fig. 3b). MOD showed fewer conformational strains as compared to CW, so we chose this model for further analysis. MOD (residues 1-207) showed N-terminal, BH4, BH3, BH1, BH2 and FLD (regulatory domain) regions. In pair-wise sequence alignment, we did not observe identity when comparing the complete FLD region of Bcl-2 with 1GJH, so it is likely that MOD would be constrained in the FLD. The accurate conformation of FLD was predicted by removing a bump, using the loop optimization protocol in the loop modeling method of MODELLER. Three loop models were outputted, among which the last model (MODLOPT) was chosen because this model had the best-refined loop. MODLOPT was then subjected to energy minimization for 22 energy iterations (total energy range: E=7888150.500 kJ mol^{-1} to −1671.042 kJ mol^{-1}). An RMSD of 4.189 Å for 1374 atoms was observed between the energy-minimized MODLOPT and SM models. We also identified two beta-strands in the FLD of SM, between residues 35 and 37 and residues 48 and 50.

A comparison of all of the structures (Table 1) showed that MODLOPT was the best. ERRAT was used to calculate the overall quality of each model, and this showed that MODLOPT has the highest overall quality (42.857) of any of the models (Table 1). The 1GJH model had an overall quality of 79.355, but when the FLD was added to this model, its quality was observed to decrease [14]. WHAT CHECK was employed to evaluate the

Fig. 4 **a** Potential energy of energy-minimized MODLOPT, which converged after 261 steps. **b** Stable potential, kinetic and total energy plot for the MD simulation. **c** Time evolution of the radius of gyration shows the compactness of structures with respect to time. The radius of gyration of a group of atoms was computed about the *x*- (*RgX*), *y*- (*RgY*) and *z*- (*RgZ*) axes, as shown by the three colored lines, which indicate the global shape of the molecule along the *x*-, *y*-, and *z*-coordinates. **d** Root mean square deviations (RMSD) of structures with respect to simulation time. **e** RMS fluctuations (in nm) of residues; greater fluctuation was observed in the loop region ▶


a
-5e+05
-6e+05
-7e+05
-8e+05
-9e+05
E (kJ mol-1)
0
50
100
150
200
250
300
Number of Steps (500)
Potential
b
2e+05
0
-2e+05
-4e+05
-6e+05
-8e+05
E (kJ mol-1)
0
5000
10000
15000
20000
Time (ps)
Potential
Kinetic En.
Total Energy
c
1.8
1.6
1.4
1.2
Rg (nm)
5000
10000
15000
Time (ps)
Rg
RgX
RgY
RgZ
d
0.8
0.6
0.4
0.2
RMSD (nm)
5000
10000
15000
Time (ps)
e
0.6
0.4
0.2
RMS Fluctuation (nm)
50
100
150
200
Residue

geometries of the models, and it showed that the Ramachandran Z-score (−7.098) was very low for 1GJH but normal (i.e., between −4 and +4) for the other 3D models. Furthermore, a very low chi-1/chi-2 correlation Z-score was observed for the 1GJH model (SD of Z-score= −6.113), but the corresponding values for the other structures were normal.

MD simulations

After validation, we realized that the energy-minimized structure of MODLOPT needs to be improved in terms of stereochemical and overall quality through optimization. MD simulations can improve the FLD structure by using a conformational search approach to study this type of unordered region [60, 61]. Thus, we employed MD to predict the folding of the FLD (i.e., the equilibrium between the folded and unfolded states of Bcl-2 containing the unstructured region). We observed conformational changes at 300 K on the minimization and optimization of the FLD domain. This observation encouraged us to attempt an extensive study of the conformational behavior of the Bcl-2 protein at 300 K over a time scale of 15 ns. The energy-minimized MODLOPT was solvated in an aqueous environment and then energy minimized. After that, the GROMACS force field was used for MD simulations. We chose several structural features along the MD trajectories in order to predict the near-native conformation [42], since discriminating models on the basis of free energy alone is not a reliable approach, due to the very small difference between the free energies of native and decoy structures [62–69].

Predicting Bcl-2 protein flexibility

We performed MD simulation analysis to predict the flexibility of FLD and its effects on the Bcl-2 protein. The energy coordinates were observed to converge after 261 steps, at which point the lowest potential energy (−865117.25 kJ mol^{-1}) was calculated (Fig. 4a). The average potential energy was observed to be −653040 kJ mol^{-1}. Finally, the potential, kinetic and total energies of the system throughout the 15,000 ps MD simulation run were calculated (Fig. 4b). An average potential energy of −806634 kJ mol^{-1}, an average kinetic energy of 130843 kJ mol^{-1}, and an average total energy of −675791 kJ mol^{-1}

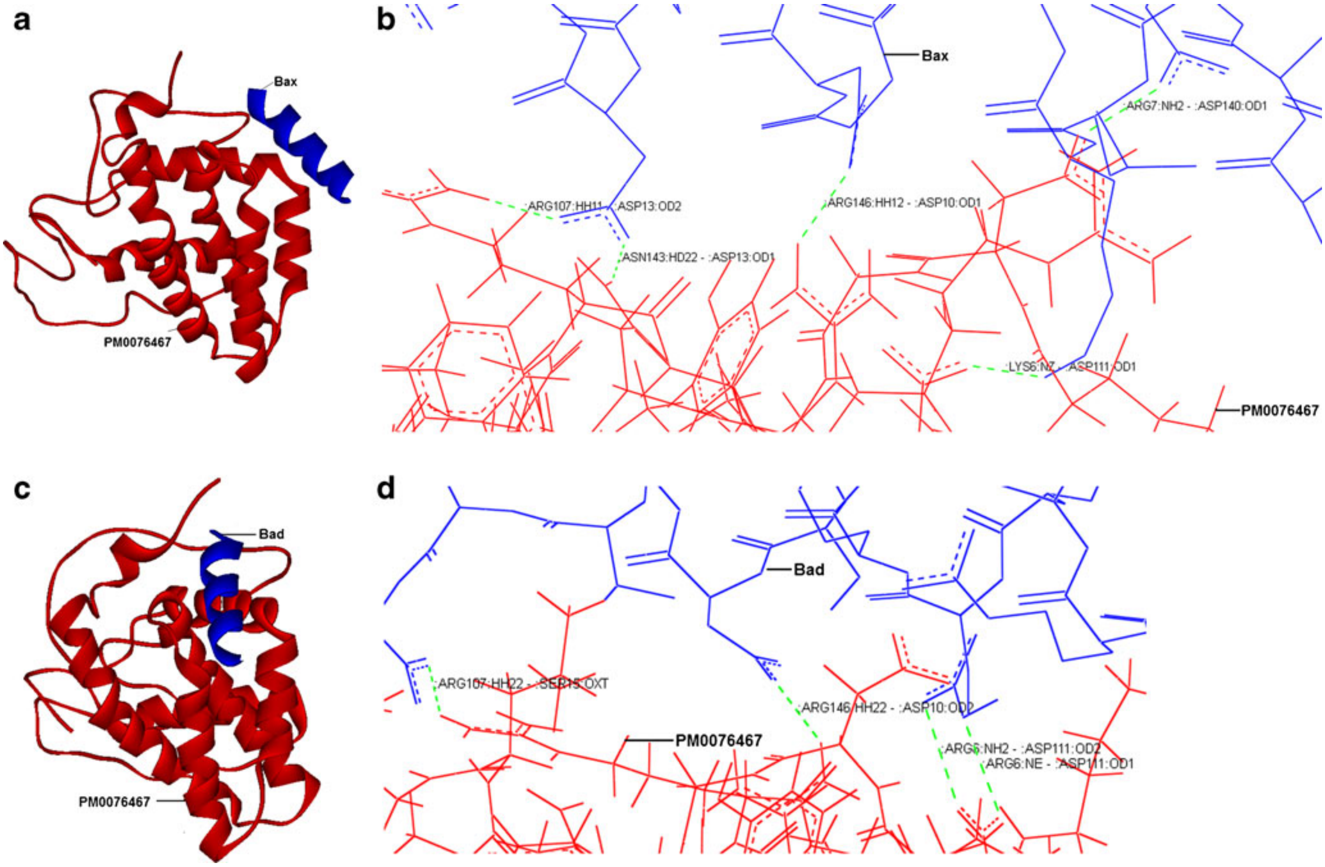


Fig. 5a–d Docking poses of PM0076467 (*red*) with **a** Bax (*blue*) peptide and with **c** Bad peptide; the hydrogen bonds observed between the complexes are shown in **b** and **d**, respectively

Table 2 Members of the 23 clusters, obtained on the basis of the RMSD; each cluster's middle member was submitted to the PMDB

Cluster	No. of members of the cluster	Mean RMSD (nm)	Middle structure	RMSD of middle structure (nm)	Members of the cluster (ps)	PMDB id
1	1	0	–	–	0	PM0077081
2	2	0.095	3	0.095	3, 6	PM0077082
3	2	0.097	9	0.097	9, 12	PM0077083
4	1	–	15	–	15	PM0077084
5	17	0.146	42	0.132	18–66	PM0077085
6	4	0.104	72	0.096	69–78	PM0077086
7	7	0.119	90	0.112	81–99	PM0077087
8	8	0.114	114	0.106	102–123	PM0077088
9	4	0.104	132	0.098	126–135	PM0077089
10	1	–	138	–	138	PM0077090
11	4	0.101	147	0.094	141–150	PM0077091
12	27	0.153	195	0.141	153–231	PM0077092
13	32	0.169	279	0.153	234–327	PM0077093
14	26	0.152	366	0.140	330–405	PM0077094
15	4	0.103	414	0.099	408–417	PM0077095
16	28	0.153	456	0.137	420–501	PM0077096
17	53	0.181	615	0.166	504–660	PM0077097
18	26	0.141	693	0.130	663–738	PM0077098
19	50	0.169	831	0.154	741–888	PM0077099
20	68	0.182	1002	0.161	891–1092	PM0077100
21	91	0.190	1257	0.173	1095–1365	PM0077101
22	148	0.198	1635	0.176	1368–1809	PM0077102
23	4397	0.291	7656	0.248	1812–15,000	PM0077103

were obtained. These three energies were observed to remain stable during the whole run.

The compactness, shape, and folding of the overall Bcl-2 structure at different time points during the trajectory can be seen in the plot of Rg (Fig. 4c). Rg was computed for atoms that were explicitly mass-weighted. 1.60 nm was the average converged value for Rg for 13,000–15,000 ps simulations. This converged and equilibrated Rg value showed that stability was obtained by the FLD. The evaluation of the Rg revealed that there was a difference between the initial and final value for the structure (it decreases from 1.9 nm to 1.8 nm). These results indicate that the final structure is more compact than the initial structure. The RMSD value of 0.84 nm was obtained after a least squares fit to the structure (except for hydrogens) of the protein at the equilibrium plateau from 12,714 ps to 15,000 ps (Fig. 4d). A decrease in Rg indicates FLD folding, whereas the attainment of an equilibrium state indicates stability of structures with similar RMSDs. The RMSF of the residues shows the overall quality of the model, the flexibility of the FLD region, and the displacement of the residues about the average position of the model. The RMSF plot (Fig. 4e) shows that the fluctuations of the FLD peak at residues Gly36 (0.4236 nm of fluctuation), Ala45 (0.7697 nm of fluctuation) and Arg68 (0.4904 nm of fluctuation). The RMSF plot is uniform, especially in the BH3 receptor cleft, indicating the importance of this region in Bcl-2 protein activity.

The energy-minimized average structure extracted from MD trajectory can be used to design new inhibitors or to identify a new binding site that could be useful in drug discovery [70]. Therefore, we computed the average structure from the MD simulations, which was then further energy minimized and validated by ProSA-web (Z-score of −6.63) (Table 1). The average structure obtained after the MD simulations displayed an overall quality factor of 80.233%, as predicted by the ERRAT program (Table 1); this shows improved model quality. This model was observed to have 81.5% of its residues in the core, 16.7% in the allowed, 1.2% in the generously allowed, and 0.6% in the disallowed regions. Verify3D analysis showed that 98.08% of the residues had average 3D-1D scores of >0.2 and G-factor dihedral values of −0.26 (acceptable values of the PROCHECK G-factor lie within the range of 0 to −0.5). Thus, all of the stereochemical and overall quality parameters were better than those of the other models (Table 1). The validated model was subsequently submitted to the PMDB (id: PM0076467).

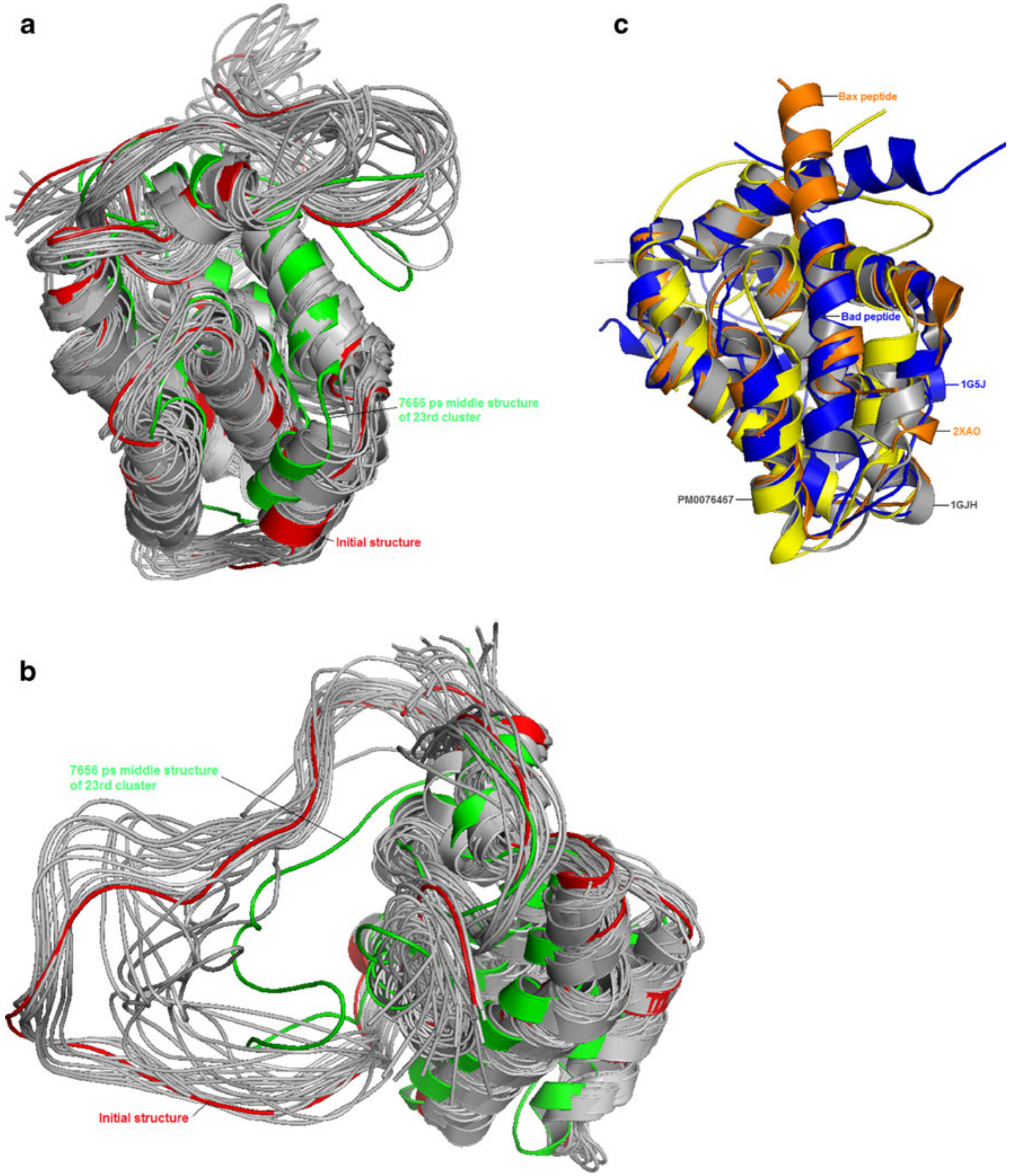

a
7656 ps middle structure of 23rd cluster
Initial structure
c
Bax peptide
Bad peptide
1G5J
2XAO
PM0076467
1GJH
b
7656 ps middle structure of 23rd cluster
Initial structure

Fig. 6 **a** Top view of the ensemble of 23 structures, showing the flexibility of the BH3 cleft. The initial structure is shown in *red*, and the structure at 7656 ps (middle structure of the 23rd cluster) is shown in *green*. **b** Side view of the ensemble of 23 structures, showing the structural changes in the FLD. The initial structure is shown in *red*, and the structure at 7656 ps (the middle structure of the 23rd cluster) is shown in *green*. **c** Superimposition of the models PM0076467 (*yellow*), 1GJH (*gray*), 2XA0 containing Bax peptide (*orange*) and 1G5J containing Bad peptide (*blue*), showing the variability of the BH3 cleft

In order to analyze the accuracy of the predicted average model, we docked it with pro-apoptotic peptides. The binding mode in the PM0076467 structure was predicted by docking this structure with the Bax peptide (residues 59–73, BH3 motif) (Fig. 5a) and the Bad peptide (residues 110–124, BH3 motif) (Fig. 5c). The docked complexes show that both of the peptides bind efficiently into the BH3 cleft. We identified the same hydrogen bonds in these complexes as those found in previously solved complex structures [71, 72]. Arg 107: HH11–Asp 13:OD2, Asn 143:HD22–Asp13:OD1, Arg146: HH12–Asp10:OD1, Asp111:OD1–Lys6:NZ and Asp140: OD1–Arg7:NH2 hydrogen bonds were observed between PM0076467 and Bax (Fig. 5b), and Arg107:HH22–Ser15: OXT, Arg146:HH22–Asp10:OD2, Arg6:NH2–Asp111:OD2

Table 3 Secondary structure contents of 1GJH, 2XA0, and PM0076467 for the most flexible region (residues 108–125) designated as H (alpha helix), G (310 helix), T (turn) and C (coil)

Residue	Number	Secondary structure contents of models		
		1GJH	2XA0	PM0076467
Tyr	108	T	H	H
Arg	109	T	H	H
Arg	110	T	H	H
Asp	111	T	H	H
Phe	112	H	C	H
Ala	113	H	C	H
Glu	114	H	C	H
Met	115	H	G	H
Ser	116	H	G	H
Ser	117	H	G	H
Gln	118	C	G	H
Leu	119	C	C	C
His	120	C	C	C
Leu	121	C	C	C
Thr	122	C	T	C
Pro	123	T	T	H
Phe	124	T	T	H
Thr	125	T	H	H

and Arg6:NE–Asp111:OD1 hydrogen bonds were noted between PM0076467 and Bad (Fig. 5d). These results indicate that FLD folding does not hinder or affect the binding of antagonistic peptides to the BH3 cleft of Bcl-2, and that the BH3 receptor cleft may play an important role in FLD folding.

Clustering of conformations

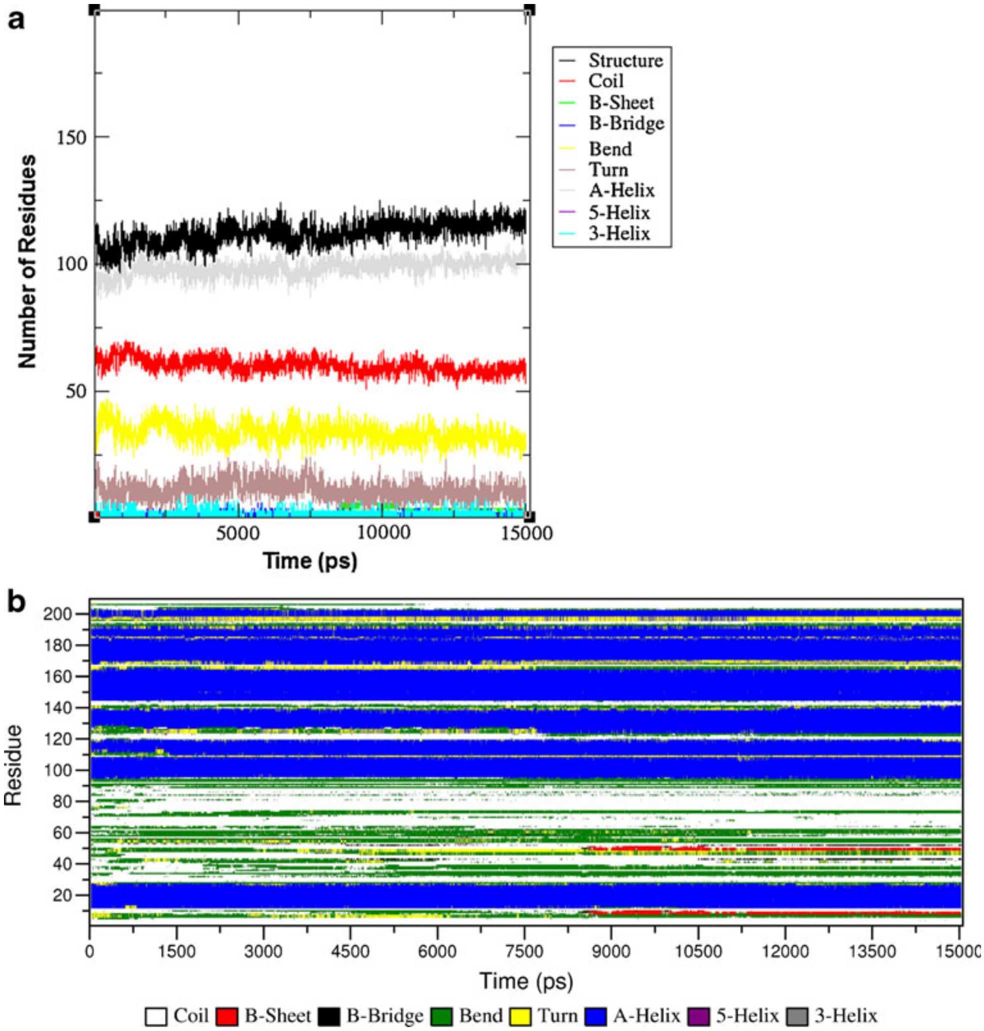
Several computationally designed and redesigned proteins have led to significant breakthroughs in biotechnological and biomedical applications, such as designing new biocatalysts [73–76], enhancing protein binding affinity [77] and redesigning protein binding specificity [78–80], and redesigning a protein in order to bind to cofactors [81]. MD simulation has an important role to play in elucidating the functional mechanisms attained by protein structures [82]. We have computed an average structure from a 15 ns simulation, but this does not necessarily mean that it is a physically meaningful structure. To characterize the behavior of the models along the trajectories, we computed an ensemble of structures based on the RMSD using the clustering tool of GROMACS. Here, we built 23 clusters using the linkage method, with all of the clusters corresponding to RMSD values ranging from 0 to 0.291 nm (Table 2). In this way, we obtained a coarse estimate for the conformational space along the trajectory explored by the same or different models. This indicates that all of the conformations that are in the same cluster have the same potential energy, because two clusters are formed when the system crosses a sufficiently high energy barrier to reach a new potential energy valley. The results shown in Table 2 allow improved discrimination between different types of models than the results shown in Fig. 4d in the form of a graph. The models obtained from 1812 ps to 15,000 ps belonged to the same cluster, with similar RMSDs. This implies that models in this cluster have the same energy and do not deviate from one another much.

To predict the flexibility of the clusters, we superimposed the middle structures of all 23 clusters. Superimposition of the 23 clusters (Fig. 6a and b) showed that the FLD folds due to changes that occur in the BH3 receptor cleft. We also observed that these changes occur mainly due to the highly flexible behavior and variability of the secondary structure content of residues 108–135, in the region connecting BH3 and BH1. These observations suggest a folded binding site for the FKBP38 protein, which is reported to bind with the FLD [9]. The changes in the BH3 cleft are also likely to regulate the binding of other proteins to the FLD. Subsequently, each cluster's middle structure was energy minimized by the Swiss-Pdb viewer program using the default parameters, and then submitted to the PMDB (Table 2). This ensemble of closely related structures should help us to predict the backbone flexibility based on computationally designed and experimentally obtained structures [83–85].

We observed the presence of the flexible region by superimposing the MD-simulated average structure (PM0076467) onto two experimental Bcl-2 structures (1GJH and 2XA0, the model of the complex of Bcl-2 with the Bax peptide) and one structure of Bcl-X_L (i.e., 1G5J, the structure of the complex of Bcl-X_L with the Bad peptide) (Fig. 6c). Stability was achieved by reducing the length of the FLD, thus stabilizing the secondary structure [86–88]. The BH3 cleft was observed to be flexible—which is essential if Bcl-2 is to optimally heterodimerize with pro-apoptotic proteins; the flexible nature of the native protein is essential to its function [89]. Substantial differences in the BH3 cleft were observed in the dynamics of the structural features of the backbones of the simulated and template structures. However, we identified the stabilized conformations (including that of the FLD) from the MD simulations.

Superimposing the models (2XA0, IGJH, 1G5J, and PM0076467) helped us to calculate the differences between them on the basis of the RMSD (Fig 6c). The observed RMSD between 2XA0 and 1GJH was 1.105 Å, that between 1G5J and 1GJH was 1.644 Å, that between 1G5J and 2XA0 was 1.234 Å, that between PM0076467 and

Fig. 7 a Average secondary structure contents along the trajectory of Bcl-2. Note that they increase from the contents of the corresponding starting structure. **b** Predicted secondary structure assignment for Bcl-2 residues along the 15,000 ps trajectory



1GJH was 2.084 Å, and that between PM0076467 and 2XA0 was 2.001 Å. We have observed that 2XA0 BH3 cleft is more inclined towards the Bcl-X_L structure. These results revealed that the 2XA0 and 1GJH models were associated with a transition state, as shown by their differences in the BH3 cleft. These BH3 cleft differences may be due to the large dynamic motion found in the region connecting BH3 and BH1 (residues 108–135). This region was found to be the most flexible one; it affects or constricts the BH3 cleft.

Since secondary structure content predicts the stability of a protein, we calculated the secondary structure content and observed transitions in this flexible region (Table 3). The region covering residues 108–111 was observed as a turn in 1GJH or a helix in 2XA0 and PM0076467, whereas the region covering residues 112–114 was found to be a helix in 1GJH and a coil in 2XA0. The structural features of the PM0076467 model show more of a resemblance to the structural features of 2XA0 than 1GJH. Residues 126–135 correspond to a helix in 1GJH, PM0076467, and 2XA0, indicating the stability of this region. Residues 108–116 represent the most important and flexible region, and the one that is likely to be responsible for most of the transitional behavior in the BH3 cleft. We did not perform any mutational studies, because in some cases it appeared that a residue mutation near or at the active site increases the flexibility, thereby decreasing the activity of the interacting site by disrupting its rigid active-site geometry [90].

Structure and stability analysis

The average numbers of residues involved in secondary structures were calculated by the DSSP program. The average structure was observed to contain an alpha-helix (94–104 residues), coils (57–68 residues), bends (24–44 residues), turns (2–19 residues), a beta-sheet (4–6 residues), a beta-bridge (2 residues) and a 3-helix (3–8 residues) (Fig. 7a). It is worth mentioning that the percentage of bends in the simulated conformation deviates greatly from the starting structure, but becomes stable after 12,000 ps and is maintained up to 15,000 ps (Fig. 7b). This may be

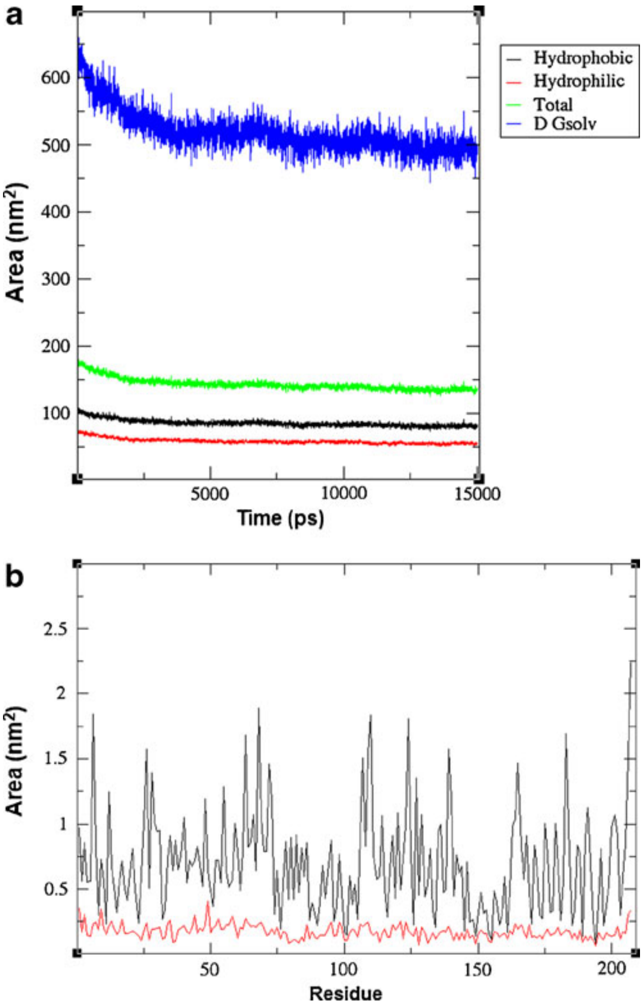


Fig. 8 **a** Calculated solvent-accessible surface area (SASA) during the whole simulation, showing the hydrophobic, hydrophilic, and total areas equilibrated at 15,000 ps. **b** The SASA per residue, displayed as *black peaks*; the *red line* indicates the standard deviations

due to the fact that several coils are converted into bends and beta-sheets in the FLD and the N-terminal region during simulations, but no alpha-helix was observed. This indicated that the FLD would probably only contain beta-sheets, as shown by the whole simulation. The structure assignment prediction for residues Ala45 to Ala60 in the FLD region shows stable beta-strands (in red) and bends (in green). This observation is in accord with the beta-strands observed in the SM model, and indicates that beta-strands are likely to exist in the native structure of the Bcl-2 FLD region.

The solvent-accessible surface area (SASA) was calculated (Fig. 8a and b), and this showed that both hydrophilicity and hydrophobicity decreased and then equilibrium was attained after 12,500 ps of simulation. The hydrophobic SASA was found to be 77–83 nm^2 and the hydrophilic SASA 52–56 nm^2. In addition, the free energy of solvation was found to cover a larger area than the total SASA (Fig. 8a). The calculation of the average SASA (over time, per residue) indicated that Pro40, His55, Pro59, Arg63, Arg68 and Gln73 had the greatest exposure to the solvent (Fig. 8b). The residues Arg63 and Arg68 covered areas of 1.67 nm^2 and 1.88 nm^2 of the solvent and showed the highest hydrophilicity. The hydrophobic residues Ala61, Ala76, Ala81 and Val89 covered areas of 0.489265, 0.193879, 0.423293, and 0.314359 nm^2; among these, Ala76 showed the highest hydrophobicity (Fig. 8b). These hydrophobic residues likely comprise a potent active site in the FLD for a ligand or another protein due to its hydrophobicity.

The secondary and tertiary structural changes that occurred during the whole simulation were predicted by the mean smallest minimum distance matrix (Fig. 9a), which accurately describes the native state of the protein. This can be obtained from the matrix of native contacts. The maps derived from these contacts contain essential geometric topological information, and describe structural interaction patterns in the protein [91, 92]. We observed the formation of contact maps between the backbone residues, and large changes in the FLD at different time scales (Fig. 9b, c and d). The formation of new contacts between residues 35 and 80 and residues 75 and 110 (in the average/mean distance matrix) was observed to stabilize the protein structure, and it helped to produce the secondary structure conformation of the FLD. This also suggests that the binding of FKBP38 and other interacting proteins to the FLD probably takes place at this region.

It appears that the hydrogen bonds stabilize the protein and have crucial role to play in protein folding [93–101]. The average number of intramolecular hydrogen bonds was calculated in different simulations. The intramolecular hydrogen bonds within the whole protein were found to involve 296 donors and 584 acceptors in total (within a cutoff distance of 0.35 nm and angle of 30°). The intramolecular hydrogen bonds that formed within the FLD (Table 4) show that Ala residues participate most frequently in hydrogen bonding. The intramolecular hydrogen bonds (within the FLD) were categorized according to whether they were between hydrophobic and hydrophilic, hydrophobic and neutral group, or hydrophilic and neutral group residues. Hydrogen bonding between hydrophobic and hydrophilic group residues was observed for the Val92N–Pro90O, Thr69N–Ala67O, Arg68N–Val66O, Ala67N–Pro65O, Val66N–Asp64O, Asp64N–Ala61O, Arg63N–Ala61O, Phe49N–Pro44O, Phe49N–Pro46O, Ile48N–Pro46O, Ala43N–Pro39O, Ala43N–Pro40O, Ala42N–Pro39O, and Ala42N–Pro40O bonds. Hydrogen bonds between hydrophobic and hydrophilic group residues were present inside the core and on the surface of the protein, and so they would probably participate in the stabilization of the FLD.

Besides intramolecular hydrogen bonding, salt bridges (Table 5) were observed between oppositely charged

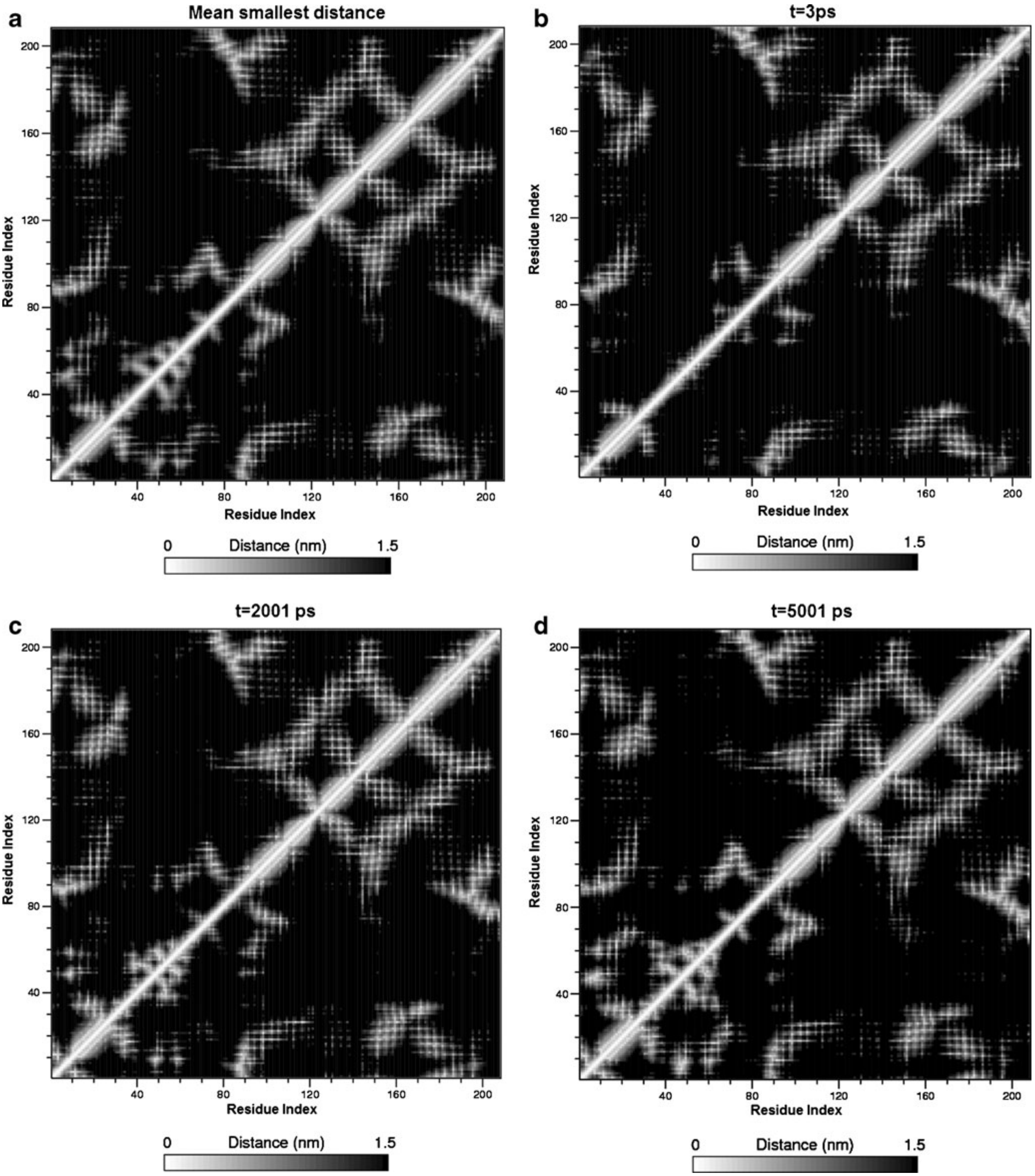


Fig. 9a–d Contact maps formed between the residues of Bcl-2, as calculated from the minimum distance matrix: **a** mean changes; **b** at 3 ps; **c** at 2001 ps; **d** at 5001 ps, for the whole simulation

residues. These salt bridges, may have a stabilizing effect on the overall structure of the Bcl-2 protein. This stabilizing effect is mainly attributed to the Arg and Asp residues present within the FLD. The residues Asp34, Asp64, Arg63, and Arg68 were noted to be particularly important, and are likely to be responsible for FLD stability. The mutation of these residues would understandably decrease the stability of the protein.

Table 4 Intramolecular hydrogen bonds within the FLD

Donor	Acceptors
Thr74N	Pro71O, Leu72O
Asp64N	His55O, Ala61O, Ser62O
Arg63N	His55O, Ala61O
Ser62N	Pro59O, Ala60O
Ala61N	His58O, His59O
Ala60N	Ile48O, His58O
His58N	Ser50O, Thr56O, Ala61O, Ser62O
Thr56N	Gln52O, Pro53O, Gly54O, Ser62O
His55N	Gln52O, Pro53O
Gly54N	Ala38O, Pro39O, Pro40O, Gln52O
Gln52N	Ser50O, His55O, Thr56O, Pro57O
Ser51N	Ala42O, Pro44O, Phe49O
Ser50N	Gly48O, Pro57O
Phe49N	Pro44O, Ala45O, Pro46O, Gly47O
IlE48N	Ala45O, Pro46O
Ala43N	Pro39O, Pro40O, Gly41O, Ser51O
Ala42N	Pro39O, Pro40O, Ser51O
Gly41N	Pro39O, Ser51O
Ala38N	Val35O, Gly36O

These results should help to enhance our knowledge of the interplay of forces that lead to the stability of the FLD through salt bridges and hydrogen bonding. All these results suggest that the FLD naturally adopts a folded state rather than a random coil structure.

Calculating the dipole moment in order to predict the electrostatic behavior

Permanent electric dipole moments contribute to the electrostatic forces (long-range forces) in biomolecules, and play an important role in determining biomolecule folding, structure, and properties. Alpha-helix conformations of proteins lead to large macro-dipoles that induce strong electric fields [102]. The fluctuations of a protein's polar groups in response to conformational changes play a key role in the folding of its structure and its binding properties [103]. The total dipole moment of Bcl-2 was evaluated (Fig. 10), and the result was found to be similar to that observed for 2XA0 (221.10 D). We observed that the differences in dipole moment direction (*x*, *y* and *z*) and the average fluctuations were in accordance with the X-ray structure (2XA0). Moreover, the 1GJH template displayed the highest dipole moment (635.91 D), which indicates that this structure has high overall rigidity and displays low flexible strength and kinetics in its associations with small molecules or proteins. This would also affect its anti-apoptotic activity, which is reported to decrease with structural rigidity [104]. This implies that 1GJH shows completely different electrostatic behavior to 2XA0 due to its large dipole moment from its charged residues, which in turn indicates unfavorable charge–macro-dipole interactions in the 1GJH template, although the interactions are favorable for simulated structures. The dipole moment analysis accurately characterized the active Bcl-2 models along the trajectory, which further validated the use of our MD simulations to obtain the near-native or native conformation.

Table 5 Residues that form salt bridges in the FLD region

Positive	Negative
Arg12	Asp34
Arg63	Asp102
Arg63	Asp64
Arg68	Asp102
Arg63	Asp34
Arg98	Asp34
Arg207	Asp64
Arg68	Asp64
Arg98	Asp64
Arg63	Glu29
Arg68	Glu29
Lys17	Asp34
Lys17	Asp64

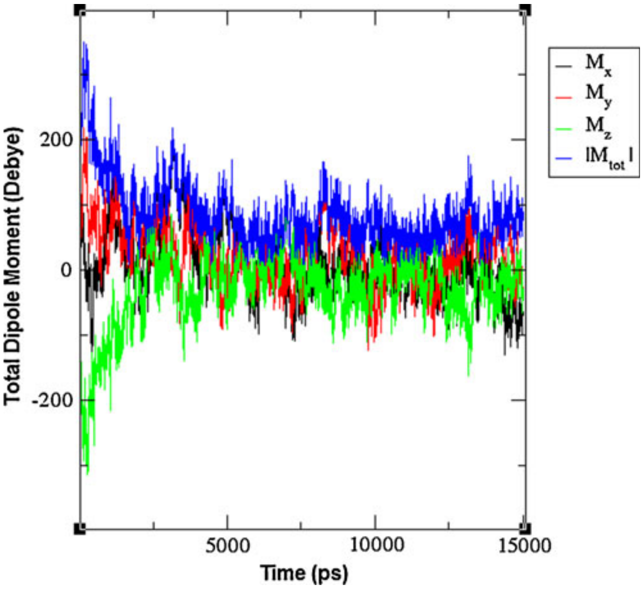


Fig. 10 Total dipole moment (*blue*) versus time, showing the electrostatic behavior during simulations

Essential dynamics

Molecular dynamics focuses on internal protein motions, and the correlation of the measured flexibility based on ordered parameters with the configuration entropy [44].

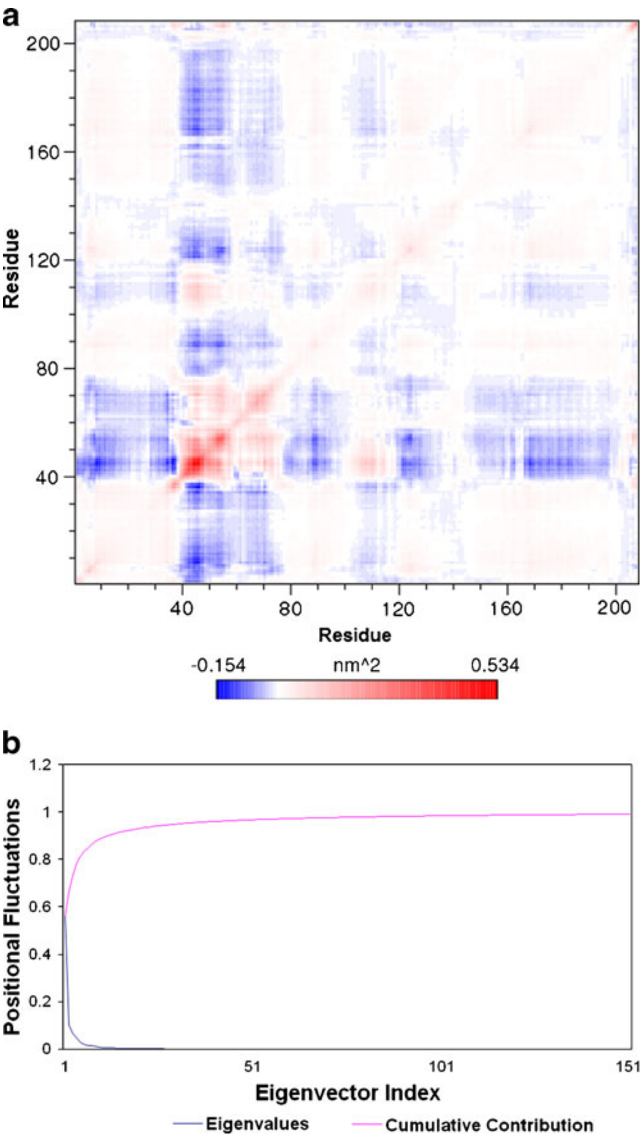


Fig. 11 **a** The covariance matrix shows anticorrelated and correlated motions between atoms. *Red* indicates that two atoms move together (correlated motions); *blue* indicates movement in opposite directions (anticorrelated motions). The *color intensity* indicates the amplitude of the RMS fluctuations (see the *color bar*). **b** Positional fluctuations, indicating the eigenvalues and the cumulative contributions along 149 eigenvectors. Eigenvalues fluctuations shown in decreasing order of magnitude and cumulative contribution fluctuation by increasing order of magnitude, as obtained from the C_α coordinate covariance matrix from a solvent simulation

Therefore, we applied principal component analysis to the Bcl-2 trajectory in order to identify large-scale collective motions of atoms and predict the flexible behavior of the FLD and the BH3 cleft.

The collective (correlated) motions of the atoms in the protein are key to its biological function [90, 105]. This study revealed that the structures underlying the atomic fluctuations (corresponding to B-factors) occur due to correlated interactions between Bcl-2 atoms. Correlated motions at the atomic level helped us to predict if the overall fluctuations of Bcl-2's C-α atoms in the system have functional or biophysical relevance. In addition, these atomic fluctuations describe the flexible behavior of the Bcl-2 protein. The covariance matrix captured the degree of collinearity in atomic motions for each pair among 207 residues. Cross-correlations between residue fluctuations helped us to identify highly correlated, moderately correlated, and anticorrelated regions (Fig. 11a). The covariance 621 × 621 symmetric matrix (Fig. 11a) shows that the large group of atoms in the FLD moves in an anticorrelated manner in relation to other domains. Using this matrix, we detected that the residues 40–75 (in the FLD) are highly correlated (red color), and that there is a weak atomic correlation (light red) between residues 105–118 (in-between the BH3 and BH1 domains) and residues 40–55 (in the FLD). This proved that the FLD and the BH3 cleft may play important roles in the regulation of Bcl-2 activity. These results indicate that any change in the FLD would likely have some effect on the BH3 cleft, as observed in the binding of p53 to FLD and the consequent decreased interaction of Bcl-2 with Bax through its BH3 cleft [6]. These results are in accordance with our superposition results (Table 3 and Fig. 6c). Hence, the results indicate that most of the internal motions of the C_α atoms of Bcl-2 protein are confined within a subspace with very small dimensions.

The diagonalization of this matrix leads to the generation of eigenvectors and eigenvalues (Fig. 11b). Each eigenvector describes a collective motion performed by particles, whereas the eigenvalues indicate how much a particular atom participated in the motion [106]. The calculated eigenvalues and cumulative contribution in the collective motion were calculated for the first 149 eigenvectors. 80% of the motion of the system was described by the first five eigenvectors, which explained the overall positional fluctuations contributing to the largest motions (Fig. 11b). Similarly, the C_α atom displacements showed that the largest motions are confined to the first five eigenvectors and producing large motions in the FLD and the BH3 cleft.

The eigenvectors describe the essential degrees of freedom (containing large-scale global motions), which are vital for protein function, and represent the part of conformational space called essential subspace, whereas near-constrained subspace refers to less interesting local fluctuations or Gaussian distributions [56, 105, 107–111]. The two-dimensional graphs of eigenvector 1 versus eigenvector 2, eigenvector 2 versus eigenvector 3, eigenvector 9 versus eigenvector 10, and eigenvector 20 versus eigenvector 25 show the protein motion in conformational space (Fig. 12a, b, c and d). No fluctuations or motions were observed along the 25th eigenvector subspace for the

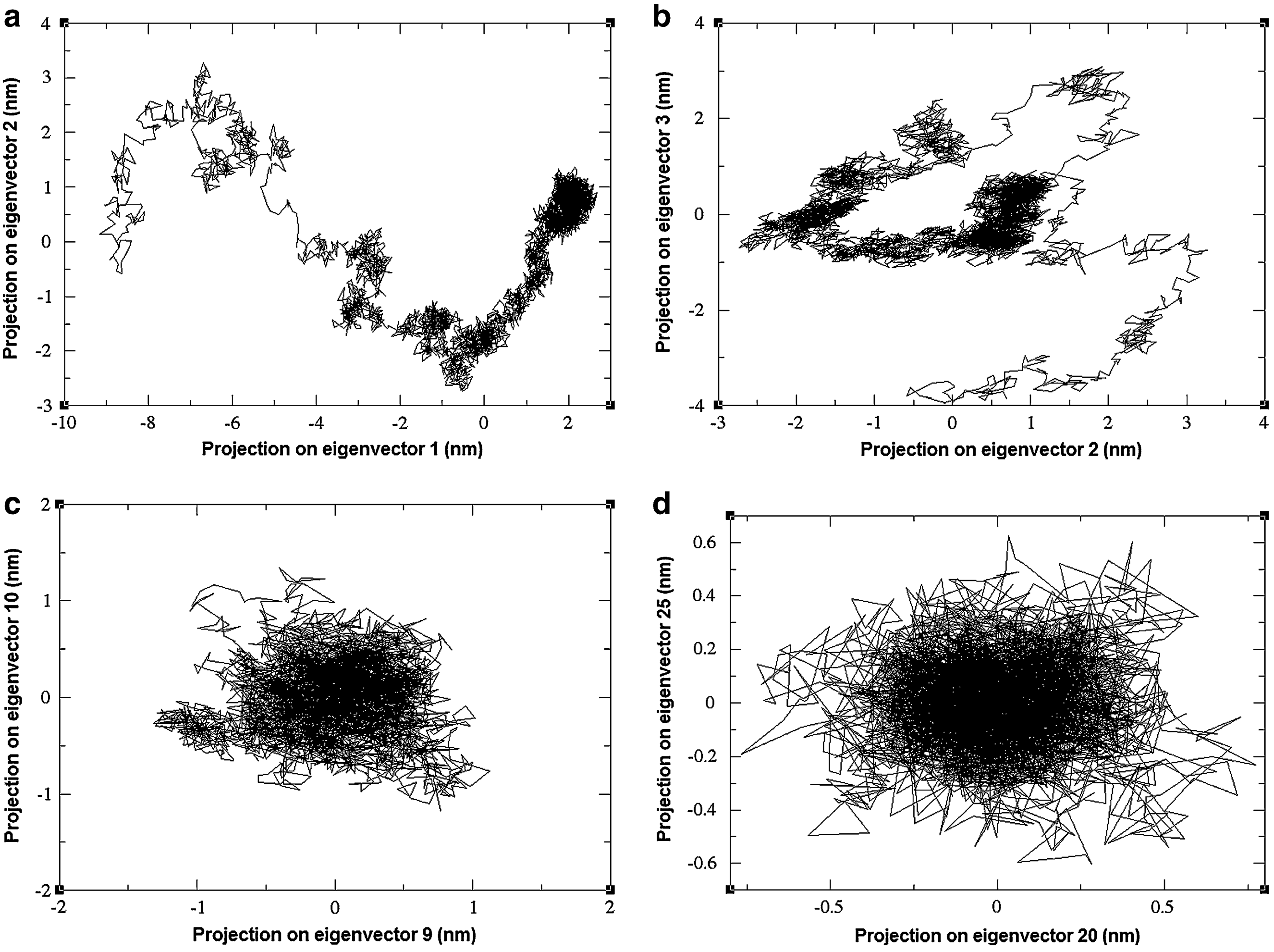


Fig. 12a–d Two-dimensional projections of the Bcl-2 trajectory on eigenvectors **a** 1 and 2, **b** 2 and 3, **c** 9 and 10, and **d** 20 and 25

Bcl-2 protein. The projections on eigenvectors 1 and 2 (Fig. 12a) show random walk and large collective motions in this essential subspace, which are observed to be relevant to protein function. The shapes of the projections were observed to be mutually independent (oval distribution) for projections on eigenvectors 9 and 10 (Fig. 12c) and 20 and 25 (Fig. 12d). These physically constrained subspaces are much less important for protein function, and are referred to as Gaussian fluctuations (irrelevant local fluctuations).

Table 6 Minimum and maximum energy scores of the projections along the ten eigenvectors, as calculated from the covariance matrix

Eigenvectors	Minimum		Maximum	
	Scores	Time (ps)	Scores	Time (ps)
1	−26.898708	99.0	8.427352	9450.0
2	−8.904247	4695.0	11.013029	558.0
3	−11.996984	51.0	7.409557	1824.0
4	−6.242445	2666.0	5.714304	1068.0
5	−10.428877	6.0	8.632933	1101.0
6	−4.668922	531.0	5.037412	78.0
7	−7.668024	45.0	5.500106	528.0
8	−4.987044	11406.0	3.829373	9309.0
9	−5.176743	465.0	5.057445	1131.0
10	−3.973459	8358.0	3.590667	582.0

The motions in the essential subspace are anharmonic diffusional motions of the protein, which randomly moves from one local minimum on the potential energy surface to the other. These motions were identified by the scores attained from extreme structure projections for eigenvectors 1–10 along the 15 ns MD trajectory (Table 6). The dynamic behavior of the protein is predicted by essential dynamics projections along the trajectory [112]. The projection of a trajectory onto the eigenvectors gives an indication of the sampling of conformational space. These spaces were obtained by the average local minimum and maximum scores describing the largest motions along the first eigenvector corresponding to extreme structure with respect to time. These scores, referred to as projections, show the motion along the axis, the total extent of the motion, and they predict the stability of the protein. The first eigenvec-

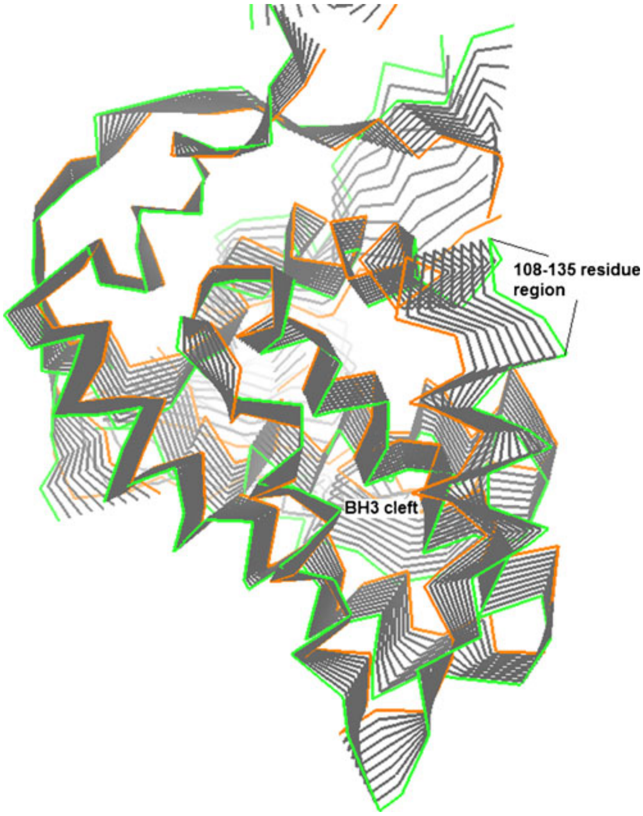


Fig. 13 BH3 cleft motion shown by ten extreme frames along the first eigenvector. All extreme structures were depicted using Pymol. The motions of atoms were observed to cause structural changes from the first frame (*green*) to the last frame (*orange*). *Arrows* indicate the direction of motion in the BH3 cleft; the region containing residues 108–135 shows the largest atomic motions

tor describes the largest motions, which take the longest time to converge during sampling. Our aim was to analyze the essential subspace motions and determine the residues that contribute to such motions, especially in the BH3 cleft. Therefore, the ten extreme structures extracted along the first eigenvector allow us to visualize the motion along the axis and the total extent of the motion, which are largely provided by the BH3 cleft (residues 108–135) (Fig. 13). These results provide a plausible reason for the difference in the BH3 clefts of the 1GJH and 2XA0 models.

Summary

In contrast to other members of the Bcl-2 family, the Bcl-2 and Bcl-X_L proteins include a large region known as the FLD. Our sequence and structural analysis results indicate that these proteins do not have the same FLD regions. The whole structure of the Bcl-2 protein has not yet been determined experimentally; only the isoform NMR structure (1GJH) has been solved (excluding the FLD). The structural features of the binding sites in the FLD are unknown, and no information is available on the structural changes to the Bcl-2 protein induced by the binding of other proteins to the FLD. Mutational and protein interaction studies have identified the importance of the FLD in the Bcl-2 protein, but the structural changes induced by FLD are still unknown. In view of the biological importance of the FLD, we focused our attention on creating a model of the complete Bcl-2 protein that would likely help us to get a better understanding of the regulation of Bcl-2 anti-apoptotic activity, which is mediated through FLD.

The structure of the complete Bcl-2 protein was elucidated by homology modeling. MODELLER 9v7 produced a better model (MOD) using the manually aligned 1GJH template than Swiss-Model. Obtaining an accurate Bcl-2 homology model is expected to reduce the errors that occur when performing protein–protein interaction (docking) and mutagenesis studies. Further, the energy-minimized 3D model of MODLOPT was observed to be an improved-quality model, and to have a better overall quality factor than the other models (CW and MOD). It could therefore be used for further computational studies. The presence of many coils in the FLD structure is accounted for by the presence of 19 Pro residues, which act as alpha-helix and beta-sheet breakers [113]. We also observed that the presence of a high number of Pro residues in the FLD disrupts the continuity of its structure, leading to an unstructured conformation. Thus, in order to predict the FLD folding, MD simulation of the energy-minimized model of MODLOPT was performed, which revealed that the Bcl-2 loop is very flexible, which may play a role in the structural diversity of FLD models. However, we observed that FLD stability was largely due by the presence of the large number of Pro residues in the FLD. These take part in hydrogen bonding between hydrophilic and hydrophobic residues, and were observed to form more stable bonds than those between hydrophilic and neutral group residues, hydrophobic and neutral group residues, or vice versa (Table 4). The residues Arg63 and Arg68 contributed the highest hydrophilicities to the SASA, and are also likely to play an important role in providing stability to the FLD by forming salt bridges. The average structure of Bcl-2 was extracted from the trajectories of MD simulations, and further validated by PROCHECK, WHAT CHECK, ERRAT and ProSA-web software. Our study indicates that the average structure can be considered a near-native conformation, as it shows good overall quality in comparison with other models. The MD-simulated average model was validated by several programs and was found to have the best 3D conformation, with improved folding of the unstructured flexible loop—a prerequisite for understanding the mechanism of mutation and the interaction of proteins with the FLD. The average structure resulting from the trajectory could vary greatly, so conformation clusters with the same RMSD were identified by g_cluster. The ensemble of structures of Bcl-2 can also be used to study the structural basis for its activation and regulation upon binding with other proteins at the FLD, considering its improved quality as compared to structures already available in the database. The electrostatic behavior of the whole MD-simulated trajectory was found to agree with the X-ray structure (2XA0), indicating the authenticity of the simulations. The quality of the model produced should aid attempts to make reliable predictions of the bioactive FLD's antagonists/agonists via in silico analysis.

Further, PCA analysis was used as a tool to obtain information about essential subspaces. PCA is very useful for predicting the directions of motion of MD-simulated structures in the essential subspace. This helped us to predict the unfolding mechanism behind the conformational and dynamic properties of the BH3 cleft, for which great conformational variability was observed. PCA also helped us to identify motions that are crucial to protein function (Fig. 13) in the essential subspace shown by the first eigenvector (Table 6). This study also indicates that the identification of representative subspaces from the PCA is useful for elucidating the structure–function relationship for the FLD.

Acknowledgments We are thankful to Director R. P. Tripathi, INMAS, Delhi, India for his support. This work is funded by the Defence Research and Development Organization (DRDO) (Govt. of India) and the Department of Science and Technology, India. We are also thankful to Meenakshi Anurag of IGIB, CSIR, Delhi, India for help with using DSSP in GROMACS, and Tsjerk A. Wassenaar of the University of Groningen, the Netherlands, for providing the second chapter on PCA from his Ph.D. thesis.

References

1. Gurudutta GU, Verma YK, Singh VK, Gupta P, Raj HG, Sharma RK, Chandra R (2005) Structural conservation of residues in BH1 and BH2 domains of Bcl-2 family proteins. FEBS Lett 579:3503–3507
2. Antonsson B, Martinou JC (2000) The Bcl-2 protein family. Exp Cell Res 256:50–57
3. Dimmler S, Breitschopf K, Haendeler J, Zeiher AM (1999) Dephosphorylation targets Bcl-2 for ubiquitin-dependent degradation: a link between the apoptosome and the proteosome pathway. J Exp Med 189:1815–1822
4. Ruvolo PP, Deng X, May WS Jr (2001) Phosphorylation of Bcl-2 and regulation of apoptosis. Leukemia 15:515–522
5. Meier P, Finch A, Evan G (2000) Apoptosis in development. Nature 407:796–801
6. Xiangming D, Fengquin G, Tammy F, Jessica A, May WS (2006) Bcl-2's flexible loop domain regulates p53 binding and survival. Mol Cell Biol 26:4421–4434
7. Nicholson DW, Ali A, Thornberry NA, Vaillancourt JP, Ding CK, Gallant M, Gareau Y, Griffin PR, Labelle M, Lazebnik YA, Munday NA, Raju SM, Smulson ME, Yamin TT, Yu VL, Miller DK (1995) Identification and inhibition of the ICE/CED-3 protease necessary for mammalian apoptosis. Nature 376:37–43
8. Chang BS, Minn AJ, Muchmore SW, Fesik SW, Thompson CB (1997) Identification of a novel regulatory domain in Bcl-X_L and Bcl-2. EMBO J 16:968–977
9. Kang CB, Tai J, Chia J, Yoon HS (2005) The flexible loop of Bcl-2 is required for molecular interaction with immunosuppressant FK-506 binding protein 38 (FKBP38). FEBS Lett 579:1469–1476
10. Ciechanover A (1994) The ubiquitin-proteosome proteolytic pathway. Cell 79:13–21
11. Hartl FU, Hayer-Hartl M (2002) Molecular cheparones in the cytosol: from nascent chain to folded protein. Science 295:1852–1858
12. Verma YK, Gangenahalli GU, Singh VK, Gupta P, Chandra R, Sharma RK, Raj HG (2006) Cell death regulation by B-cell lymphoma protein. Apoptosis 11:459–471
13. Deng X, Kornblau SM, Ruvolo PP, May WS Jr (2001) Regulation of Bcl2 phosphorylation and potential significance for leukemic cell chemoresistance. J Natl Cancer Inst Monogr 28:30–37
14. Petros AM, Medek A, Nettesheim DG, Kim DH, Yoon HS, Swift K, Matayoshi ED, Oltersdorf T, Fesik SW (2001) Solution structure of the antiapoptotic protein bcl-2. Proc Natl Acad Sci USA 98:3012–3017
15. Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T (2009) The SWISS-MODEL repository and associated resources. Nucleic Acids Res 37:D387–D392
16. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A (2006) Comparative protein structure modeling using MODELLER. Curr Protoc Bioinformatics 15:5.6.1–5.6.30. doi:10.1002/0471250953.bi0506s15
17. Wang J, Cao Z, Shuqiang L (2009) Molecular dynamics simulations of intrinsically disordered proteins in human diseases. Curr Comput Aided Drug Des 5:280–287
18. Lindahl E, Hess B, van der Spoel D (2001) Gromacs 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 7:306–317
19. Castrignano T, De Meo PD, Cozzetto D, Talamo IG, Tramontano A (2006) The PMDB Protein Model Database. Nucleic Acids Res 34:D306–D309
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
21. Ward JJ, Sodi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in protein. J Mol Biol 337:635–645
22. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 347:827–839
23. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680
24. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919
25. Van Gunsteren WF, Bakowies D, Baron R, Chandrasekhar I, Christen M (2006) Biomolecular modeling: goals, problems, perspectives. Angew Chem Int Edn Engl 45:4064–4092
26. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723
27. DeLano WL (2002) The PyMOL molecular graphics system. http://www.pymol.org

28. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35:W407–W410
29. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26:283–291. doi:10.1107/S0021889892009944
30. Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 2:1511–1519
31. Hooft RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 381:272–272
32. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theor Comput 4:435–447
33. Oostenbrink C, Soares TA, van der Veget NFA, van Gunsteren WF (2005) Validation of the 53A6 GROMOS force field. Eur Biophys J 34:273–284
34. Oostenbrink C, Villa A, Mark AE, van Gunsteren WF (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem 25:1656–1676
35. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1969) Interaction models for water in relation to protein hydration. Nature 224:175–177
36. Feenstra KA, Hofstetter K, Bosch R, Schmid A, Commandeur JN, Vermeulen NP (2006) Enatioselective substrate binding in a monooxygenase protein model by molecular dynamics and docking. Biophys J 91:3206–3216
37. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: a linear constraint solver for molecular simulations. J Comput Chem 18:1463–1472
38. Berendsen HJC, Postma JPM, Van der Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:3684–3690
39. Parrinello M, Rahman A (1980) Crystal structure and pair potentials: a molecular dynamics study. Phys Rev Lett 45:1196–1199
40. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh Ewald method. J Chem Phys 103:8577–8593
41. Marlino A, Mazzarella L, Carannante A, Difiore A, Di Donata A, Notomista E, Sica F (2005) The importance of dynamic effects on the enzyme activity. J Biol Chem 280:17953–17960
42. Taly JF, Marine A, Gibrat JF (2008) Can molecular dynamics simulations help in discriminating correct from errorneous protein 3D models? BMC Bioinforma 9:6
43. Malek K, Odijk T, Coppens MC (2005) Diffusion of water and sodium counter-ions in nanopores of a β-lactoglobulin crystal: a molecular dynamics study. Nanotechnology 16:S522–S530
44. Brooks CL, Karplus M, Pettitt BM (1988) Proteins. A theoretical perspective of dynamics, structure and thermodynamics (Advances in Chemical Physics). Wiley, New York
45. Seshasayee AS, Raghunathan K, Sivaraman K, Pennathur G (2006) Role of hydrophobic interactions and salt-bridges in β-hairpin folding. J Mol Model 12:197–204. doi:10.1007/S00894-005-0018-6
46. Zheng H, Wang S, Zhang Y (2009) Increasing the time step with mass scaling in Born–Oppenheimer ab initio QM/MM molecular dynamics simulations. J Comput Chem 30:2706–2711
47. Feenstra KA, Hess B, Berendsen HJC (1999) Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. J Comput Chem 20:786–798
48. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual Molecular Dynamics. J Mol Graph 14:33–38
49. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637
50. Heinig M, Frishman D (2004) STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res 32:W500–W502
51. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30:2785–2791
52. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function. J Comput Chem 19:1639–1662
53. Amadei A, Linssen AB, Berendsen HJC (1993) Essential dynamics of proteins. Proteins 17:412–425
54. Amadei A, Ceruso MA, Di Nola A (1999) On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. Proteins 36:419–424
55. Kitao A, Hirata F, Go N (1991) The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. J Chem Phys 158:447–472
56. Garcia AE (1992) Large-amplitude nonlinear motions in proteins. Phys Rev Lett 68:2696–2699
57. Chen CC, Hwang JK, Yang JM (2006) $(PS)^2$: protein structure prediction server. Nucleic Acids Res 34:W152–W157
58. Ding J, So BA, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction detection and structure prediction. Nucleic Acids Res 33:W244–W248
59. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21:3433–3434
60. Madhusudhan MS, Marti-Renom MA, Eswar N, John B, Pieper U, Karchin R, Yi Shen M, Sali A (2005) Comparative protein structure modeling. In: Walker JM (ed) The proteomics protocols handbook. Humana, Totowa, pp 831–860
61. Zhang H (2002) Protein tertiary structures: prediction from amino acid sequences. In: Encyclopedia of life sciences. Nature, London, pp 1–7. doi:10.1038/npg.els.0003040
62. Novotny J, Rashin AA, Bruccoleri RE (1988) Criteria that discriminate between native proteins and incorrectly folded models. Proteins 4:19–30
63. Vorobjev YN, Almagro JC, Hermans J (1998) Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. Proteins 32:339–413
64. Janardhan A, Vajda S (1998) Selecting near-native conformations in homology modeling: the role of molecular mechanics and solvation terms. Protein Sci 7:1772–1780
65. Lazaridis T, Karplus M (1999) Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 288:447–487
66. Gatchell DW, Dennis S, Vajda S (2000) Discrimination of near-native protein structure from misfolded models by empirical free energy functions. Proteins 41:518–534
67. Kinjo AR, Kidera A, Nakamura H, Nishikawa K (2001) Physicochemical evaluation of protein folds predicted by threading. Eur Biophys J 30:1–10
68. Dominy BN, Brooks CL III (2002) Identifying native-like protein structure using physics-based potentials. J Comput Chem 23:147–160

69. Ak F, Gallicchio E, Wallquist A, Levy RM (2002) Distinguishing native conformations of proteins from decoys with an effective free energy estimation based on the OPLS all-atoms force field and the surface generalized born solvent model. Protein 48:404–422
70. Azizian H, Rahrami H, Pasalar P, Amanlou M (2010) Molecular modeling of *Helicobactor pylori* arginase and the inhibitor coordination interactions. J Mol Graph Model 28:626–635
71. Pinto M, Perez JJ, Martinez JR (2004) Molecular dynamics study of peptide segment of the BH3 domain of the proapoptotic proteins Bak, Bax, Bid and Hrk bound to the Bcl-X_L and Bcl-2 proteins. J Comput Aided Mol Des 18:13–22
72. Ku B, Liang C, Jung JU, Oh BH (2010) Evidence that inhibition of Bax activation by Bcl-2 involves its tight and preferential interaction with the BH3 domain of Bax. Cell Res 21:627–641. doi:10.1038/cr.2010.149
73. Bolon DN, Mayo SL (2001) Enzyme-like proteins by computational design. Proc Natl Acad Sci USA 98:14274–14279
74. Kaplan J, DeGrado WF (2004) De novo design of catalytic proteins. Proc Natl Acad Sci USA 101:11566–11570
75. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, Hilvert D, Houk KN, Stoddard BL, Baker D (2008) De novo computational design of retero-aldol enzymes. Science 319:1387–1391
76. Rothlisberger D (2008) Kemp elimination catalysts by computational enzyme designing. Nature 453:190–195
77. Lazar GA, Dang W, Karki S, Vafa O, Peng JS, Hyun L, Chan C, Chung HS, Eivazi A, Yoder SC, Vielmetter J, Carmichael DF, Hayes RJ, Dahiyat BI (2006) Engineered antibody Fc variants with enhanced effector function. Proc Natl Acad Sci USA 103:4005–4010
78. Ogata K, Jaramillo A, Cohen W, Briand JP, Connan F, Choppin J, Muller S, Wodak SJ (2003) Automatic sequence design of major histocompatibility complex class I binding peptides imparing $CD8^+$ T-cell recognition. J Biol Chem 278:1281–1290
79. Joachimiak L, Koretemme T, Stoddard B, Baker D (2006) Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein–protein interface. J Mol Biol 361:195–208
80. Shifman JM, Choi MH, Mihalas S, Mayo SL, Kennedy MB (2006) Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII) is activated by calmodulin with two bound calciums. Proc Natl Acad Sci USA 103:13968–13973
81. Cochran FV, Wu SP, Wang W, Nanda V, Saven JG, Therien MJ, DeGrado WF (2005) computational de novo design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor. J Am Chem Soc 127: 1346–1347
82. Daggett V (2006) Protein folding-simulation. Chem Rev 106:1898–1916
83. Desjarlaris JR, Handel TM (1995) De novo design of the hydrophobic cores of proteins. Protein Sci 4:2006–2018
84. Desjarlais JR, Handel TM (1999) Side-chain and backbone flexibility in protein core design. J Mol Biol 290:305–318
85. Krasmer-Pecore CM, LeComte JT, Desjarbis JR (2003) A de novo redesign of the WW domain. Protein Sci 12:2194–2205
86. Vieille C, Zeikus JG (1996) Thermoenzymes: identifying molecular determinants of protein structural and functional stability. Trends Biotechnol 14:183–191
87. Russell RJ, Hough DW, Danson MJ, Taylor GL (1994) The crystal structure of citrate synthase from the thermophilic archaeon, *Thermoplasma acidophilum*. Structure 2:1157–1167
88. Nagi AD, Regan L (1997) An inverse correlation between loop length and stability in a four-helix-bundle protein. Fold Des 2:67–75
89. Carlson HA (2002) Protein flexibility is an important component of structure based drug discovery. Curr Pharm Des 8:1571–1578
90. Beer HD, Wohlfahrt G, McCarthy JE, Schomburg D, Schmid RD (1996) Analysis of the catalytic mechanism of a fungal lipase using computer-aided design and structural mutants. Protein Eng 9:507–517
91. Holm L, Sander C (1996) Mapping the protein universe. Science 273:595–603
92. Vendruscolo M, Kussell E, Domany E (1997) Recovery of protein structure from contact maps. Fold Des 2:295–306
93. Bulaj G, Goldenberg DP (2001) Mutational analysis of hydrogen bonding residues in the BPTI folding pathway. J Mol Biol 313:639–656
94. Chen YW, Fersht AR, Henrick K (1993) Contribution of buried hydrogen bonds to protein stability. The crystal structures of two barnase mutants. J Mol Biol 234:1158–1170
95. Byrne MP, Manuel RL, Lowe LG, Stites WE (1995) Energetic contribution of side chain hydrogen bonding to the stability of staphylococcal nuclease. Biochemistry 34:13949–13960
96. Pace CN, Shirley BA, McNutt M, Gajiwala K (1996) Forces contributing to the conformational stability of proteins. FASEB J 10:75–83
97. Yamagata Y, Kubota M, Sumikawa Y, Funahashi J, Takano K, Fujii S, Yutani K (1998) Contribution of hydrogen bonds to the conformational stability of human lysozyme: calorimetry and X-ray analysis of six tyrosine → phenylalanine mutants. Biochemistry 37:9355–9362
98. Takano K, Yamagata Y, Kubota M, Funahashi J, Fujii S, Yutani K (1999) Contribution of hydrogen bonds to the conformational stability of human lysozyme: calorimetry and X-ray analysis of six serine → alanine mutants. Biochemistry 38:6623–6629
99. Pace CN (2001) Polar group burial contributes more to protein stability than nonpolar group burial. Biochemistry 40:310–313
100. Grantcharova VP, Riddle DS, Santiago JV, Baker D (1998) Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. Nat Struct Biol 5:714–720
101. Krantz BA, Moran LB, Kentsis A, Sosnick TR (2000) D/H amide kinetic isotope effects reveal when hydrogen bonds form during protein folding. Nat Struct Biol 7:62–71
102. Antoine R, Compagnon I, Rayane D, Broyer M, Dugowd Ph, Breaux G, Hogemeister FC, Pippen D, Hudgins RR, Jarrold MF (2002) Electric dipole moments and conformations of isolated peptides. Eur Phys J D 20:583–587
103. Simonson T (1999) Dielectric relaxation in proteins: macroscopic and microscopic models. Int J Quantum Chem 73:45–57
104. Soufian S, Naderi-Manesh H, Alizadeh A, Sarbolouki MN (2009) Molecular dynamics and circular dichoism studies on Aurein 1.2 and retro analog. World Acad Sci Eng Technol 56:858–864
105. Van Aalten DMF, Amadei A, Linssen ABM, Eijsink VGH, Vriend G, Berendsen HJC (1995) The essential dynamics of thermolysin: confirmation of the hinge-bending motion and comparison of simulation in vacuum and water. Proteins 22:45–54
106. Spoel D van der, Lindahl E, Hess B, Kutzner C, Buuren AR van, Apol E, Meulenhoff PJ, Tieleman DP, Sijbers ALTM, Feenstra KA, van Drunen R, Berendsen HJC (2005) Gromacs user manual, version 4.0. http://www.gromacs.org
107. Scheek RM, Van Nuland NAJ, DeGroot BL, Linssen ABM, Amadei A (1995) Structure from NMR and molecular dynamics: distance restraining inhibits motion in the essential subspace. J Biomol NMR 6:106–111
108. Van Aalton DMF, Findley JBC, Amadei A, Berendsen HJC (1995) Essential dynamics of the cellular retinol-binding

protein—evidence for ligand-induced conformational changes. Protein Eng 8:1129–1135

109. Chillemi G, Falconi M, Amadei A, Zimatore G, Desideri A, DiNola A (1997) The essential dynamics of Cu, Zn superoxide dismutase: suggestion of intersubunit communication. Biophys J 73:1007–1018

110. Hayward S, Kitao A, Hirata F, Go N (1993) Effect of solvent on collective motions in globular proteins. J Mol Biol 234:1207–1217

111. Romo TD, Clarage JB, Sorensen DC, Phillips GN Jr (1995) Singular value decomposition analysis of time-average crystallographic refinement. Proteins 22:311–321

112. Hunenberger PH, Mark AE, van Gunsteran WF (1995) Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. J Mol Biol 252:492–503

113. Berg JM, Tymoczko JL, Stryer L (2002) Protein composition and structure. In: Biochemistry, 6th edn. WH Freeman and Co., New York, pp 40-53

J Mol Model (2012) 18:1907–1916
DOI 10.1007/s00894-011-1211-4

ORIGINAL PAPER

Characterization of molecular recognition of Phosphoinositide-3-kinase α inhibitor through molecular dynamics simulation

Yiping Li · Jiye Zhang · Delong He · Qi Liang · Yawen Wang

Received: 7 February 2011 / Accepted: 4 August 2011 / Published online: 26 August 2011

Abstract Phosphatidylinositol 3-kinase α (PI3Kα) is a promising target for anticancer drug discovery due to its overactivation in tumor cells. To systematically investigate the interactions between PI3Kα and PIK75 which is the most selective PI3Kα inhibitor reported to date, molecular docking, molecular dynamics simulation, and ensuing energetic analysis were utilized. The binding free energy between PI3Kα and PIK75 is −10.04 kcal•mol^{-1} using MMPBSA method, while −13.88 kcal•mol^{-1} using MMGBSA method, which is beneficial for the binding. The van der Waals/hydrophobic and electrostatic interactions play critical roles for the binding. The binding mode of PIK75 for PI3Kα is predicted. The conserved hydrophobic adenine region of PI3Kα made up of Ile800, Ile848, Val850, Val851, Met922, Phe930, and Ile932 accommodates the flat 6-bromine imidazo[1,2-a]pyridine ring of PIK75. The 2-methyl-5-nitrophenyl group of PIK75 extends to the P-loop region, and has four hydrogen-bond arms with the backbone and side chain of Ser773 and Ser774. And the distinct conformation of the P-loop induced by PIK75 is speculated to be responsible for the selectivity profile of PIK75. The predicted binding mode of PIK75 for PI3Kα presented in this study may help design high affinity and selective compounds to target PI3Kα.

Keywords Molecular dynamics simulation · Phosphatidylinositol 3-kinase α

Introduction

Phosphatidylinositol 3-kinases (PI3Ks) are lipid kinases that phosphorylate the 3-hydroxyl of phosphatidylinositol, generating phosphatidylinositol 3-phosphate, phosphatidylinositol 3,4-bisphosphate, and phosphatidylinositol 3,4,5-trisphosphate that act as second messengers. The resulting second messengers interact with pleckstrin-homology- (PH-) domain-containing proteins, such as the Akt serine-threonine kinases, eliciting a series of signal transduction events that lead to DNA synthesis and cell proliferation via the activation of the MDM2 and mTOR (mammalian target of rapamycin) pathways [1, 2]. There are three major classes of PI3Ks, namely Class I, II and III, based on their sequences and substrate specificities. Among three distinct PI3K subfamilies, only the class I PI3Ks are capable of catalyzing the conversion of phosphatidylinositol 4,5-bisphosphate (PIP2) to phosphatidylinositol 3,4,5-trisphosphate (PIP3). Class I PI3Ks can be further divided into IA and IB subclasses. The class IA subfamily contains three isoforms, namely, PI3Kα, β and δ activated by tyrosine kinases, antigen and cytokine receptors. The class IB subfamily contains only one isoform, namely, PI3Kγ activated by G-protein-coupled receptors [3, 4]. PI3Kα is a heterodimeric protein consisting of a catalytic p110α subunit and a p85 regulatory subunit. The p110α subunit contains N-terminal adaptor-binding (ABD), Ras-binding, C2, helical and catalytic kinase domains. The ABD domain was proposed to be responsible for p85α binding, and

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1211-4) contains supplementary material, which is available to authorized users.

Y. Li · J. Zhang · D. He · Q. Liang
Department of Pharmacy, College of Medicine,
Xi'an Jiaotong University,
Xi'an 710061, Peoples Republic of China

Y. Wang (✉)
First Affiliated Hospital, College of Medicine,
Xi'an Jiaotong University,
Xi'an 710061, Peoples Republic of China
e-mail: Wanglxy628@sina.com

the C2 domain for cellular membrane binding [5]. PI3Kα is activated by receptor tyrosine kinases (RTKs) such as endothelial growth factor receptor (EGFR), human epidermal growth factor receptor 2 (HER2), and vascular endothelial growth factor receptor (VEGFR). The activated p110α catalytic subunit catalyzes the conversion of the PIP2 to PIP3 [1, 2].

The implication of PI3Kα in cancer was confirmed by the observation that PI3Kα is frequently mutated in some human cancers [6, 7]. Recently, Liu et al. [8] reported the incidence of tumors with PI3Kα mutations in a much larger population: breast, 27% (468/1766); endometrial, 24% (102/429); colon, 15% (448/3024); upper digestive tract, 11% (38/352); stomach, 8% (29/362); pancreas, 8% (29/362); and ovarian, 8% (61/787). And the mutations constitutively confer a marked increase in its kinase activity. In addition, under normal physiological conditions, PIP3 levels are tightly regulated by the phosphatase and tensin homologue protein (PTEN). The inactivation of PTEN by mutations in tumors leads to the accumulation of PIP3 [9]. Therefore, PI3Kα has become a potential and attractive target for anti-tumor therapy and hence spark great interest in the discovery and development of inhibitors. However, Due to the PI3Kα and PI3Kγ isoforms share~35% sequence identity, and the catalytic kinase domains sequence identity is~43.5%, the sequence identity makes it more challenging to find inhibitors with high selectivity PI3Kα and PI3Kγ. In Hayakawa et al.'s work [10], PIK75, sulfonylhydrazone substituted imidazo[1,2-a]pyridines derivative as shown in Fig. 1, inhibits PI3Kα and PI3Kγ with IC_{50} values of 0.0003 and 0.040 μM, respectively, which is a selective PI3Kα inhibitor up to now.

In our previous work, we have reported the pharmacophore and docking study of PI3Kα inhibitors based on X-ray crystal structure of human PI3Kα/p85α complex [11]. Several other studies based on X-ray crystal structures or homology models built on the PI3Kγ structure using homology modeling have been conducted [12–14]. All these studies are based on one conformation from the X-ray crystal structure or one homology model. However, protein flexibility plays an important role in molecular recognition. Recently, in Han and Zhang.'s work [15] the residue Trp780 and Asn782 in PI3Kα were suggested to confer the isoform-specific selectivity between PI3Kα and PI3Kγ to PIK75 based on the combination of docking and molecular dynamics simulation. Sabbah et al. [16] reported the PI3K inhibitor interactions with the PI3Kα H1047R mutant. However, in these two studies, to study the effect of protein flexibility on ligand docking, they carried out molecular dynamics (MD) simulations on the wild-type and mutant PI3Kα and then docked the ligands to the protein conformations built from molecular dynamics simulations at the well-equilibrated snapshot, but no molecular dynamics simulations on the complex of the ligands and PI3Kα were directly carried out. However, protein-ligand recognition is an induced fit process. It has become increasingly clear that it is critical to accurately model ligand-induced protein movement in order to obtain high enrichment factors. Therefore, molecular dynamics simulation on the complex of ligand and protein becomes very valuable since it takes into account of molecular flexibility and induced fit.

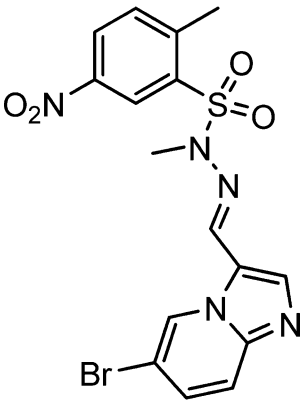


Fig. 1 Chemical structure of PIK75

To identify the interactions between PI3Kα and its inhibitor, this work is to systematically evaluate the interactions between PI3Kα and PIK75 computationally through molecular docking and molecular dynamics simulation. Based on the MD ensemble structures, we calculated the binding free energy and binding energy decomposition over the course of the trajectory and by residue by means of molecular mechanics (MM)-Poisson-Boltzmann (generalized Born) surface area (PB(GB)SA) approach. We also analyzed the accumulated hydrogen bond distribution and binding energy decomposition by residue, which revealed the hotspot residues of PI3Kα-PIK75 binding. Finally, according to our results, the binding mode of PIK75 for PI3Kα was predicted, which will help design new PI3Kα inhibitor.

Materials and methods

Preparation of PI3Kα and its inhibitor PIK75

The X-ray crystal structures of PI3Kα (PDB id: 2RD0) was retrieved from the RCSB Protein Data Bank. The lost loop zones of the PI3Kα X-ray structure were generated and refined by *ab initio* refinement of the loop in the loop refine module of Modeler 9v5. The overall 2RD0 was subsequently subjected to 500 iterations of energy minimization with backbone atoms being restrained using the OPLS2005 force field in the MacroModel module in the Schrodinger software suite. The PI3Kα inhibitor PIK75 was built using the Maestro Build panel and minimized by the MacroModel program using the OPLS2005 force field.

Docking simulations

First, the Gasteiger charges were applied to the PI3Kα and the PIK75 structures. Then their non-polar hydrogens were merged so that these hydrogen structures were not considered in the docking calculations. AutoDock Tools 1.5.4 [17] was used to set up rotatable bonds of PIK75. Second, energy affinity maps for PIK75's atom types, desolvation energies, and electrostatic potentials were pre-calculated using AutoGrid4. Third, the binding site on the PI3Kα was defined by a grid system of (x, y, z)=(46-point, 46-point, 52-point) with a grid Spacing of 0.375 Å that originated at the center of the catalytic kinase domains. Finally, docking simulations were carried out via Autodock4 [18] with a rigid receptor structure, which allowed for flexibility in the ligand structure using a Lamarckian genetic algorithm (LGA) in combination with a hybrid local and global search for new docking conformations. The Lamarckian genetic algorithm was applied to the following protocol: trials of 100 runs, energy evaluations of 50 000 000, maximum number of generations of 30 000, population size of 200, a mutation rate of 0.02, a crossover rate of 0.8, and an elitism value of 1. The docking results were evaluated by sorting the binding free energy predicted by docking conformations. Docked conformations were clustered using a tolerance of 1.5 Å root-mean-square deviations (rmsd).

MD simulations of the PI3Kα/PIK75 complex

The PI3Kα receptor coordinate was concatenated with the docked coordinates of PIK75 taken from the docking simulation. To generate the topology and parameter files for PIK75, firstly *ab initio* calculation was carried out at the Hartree-Fock level of theory with 6–31+G* basis set using Gaussian03 suite [19]. The computed electrostatic potential (ESP) was then read in by the ANTECHAMBER [20] protocol of the Amber9 suit [21] for the RESP charge fitting and the atom equivalence treatment in conjunction with the generalized Amber force field (GAFF) [22], subsequently topology and parameter files were generated for PIK75. The topology and parameter files of PIK75 were included in the Supporting information.

All simulations were conducted by using the Amber9 program. Two parameter sets were used, the biomolecular force field ff03 for the protein and general AMBER force field (GAFF) for the organic small molecule. The PI3Kα/PIK75 complex was soaked in a truncated octahedron box of TIP3P water molecules with a margin of 15 Å along each dimension. Ten Na^+ ions were added to neutralize the system. In summary, the system consists of PI3Kα, PIK75, 10 Na^+ ions and 51082 water molecules. The covalent bonds involving hydrogen atoms of the complex system were constrained using the SHAKE option, and the particle mesh Ewald (PME) method [23] was used to model the long-range electrostatic interactions using the parallel sander protocol on 16 cores of the IBM opteron cluster in National High Performance Computing Center (Xi'an). The system was then energy minimized with a 100 cycle steepest descent method, which was followed by a 1900 cycle conjugate gradient method. The temperature of the system was elevated from 100 K to 300 K over 50 ps via the Berendsen temperature coupling schemes in Amber using a TAUTP of 2.0 ps (time constant for heat bath coupling). The pressure of the system was equilibrated for 200 ps using the Berendsen pressure coupling schemes in Amber using a TAUP 2.0 ps (pressure relaxation time). Finally, a 20 ns production run was carried out and the trajectory of the complex structure was written out every 10 ps in order to collect 2000 snapshots.

Binding free energy calculations

The binding free energies were calculated using the MMPB(GB)SA method as implemented in Amber9. MMPB(GB)SA computes the binding free energy by using a thermodynamic cycle that combines the molecular mechanical energies with the continuum solvent approaches [24]. The binding free energy was calculated according to the equation:

$$\Delta G_{\mathrm{bind}} = G_{\mathrm{complex}} - G_{\mathrm{PI3K}\alpha} - G_{\mathrm{PIK75}}, \tag{1}$$

where G_{complex}, $G_{\mathrm{PI3K}\alpha}$ and G_{PIK75} are the free energies of the complex, the protein PI3Kα and the ligand PIK75, respectively. The free energy of each term was calculated as a sum of the three terms:

$$G = E_{\mathrm{MM}} + G_{\mathrm{sol}} - TS, \tag{2}$$

where E_{MM} is the molecular mechanics energy of the molecule expressed as the sum of the internal energy (bonds, angles and dihedrals) (E_{int}), electrostatic energy (E_{ele}) and van der waals term (E_{vdw}) computed using an Amber9 force field:

$$E_{\mathrm{MM}} = E_{\mathrm{int}} + E_{\mathrm{ele}} + E_{\mathrm{vdw}}. \tag{3}$$

G_{sol} accounts for the solvation energy which can divided into the polar ($G_{\mathrm{PB(GB)}}$) and nonpolar part (G_{NP}).

$$G_{\mathrm{sol}} = G_{\mathrm{PB(GB)}} + G_{\mathrm{NP}} \tag{4}$$

The polar part ($G_{\mathrm{PB(GB)}}$) accounts for the electrostatic contribution to solvation and was calculated using a Poisson-Boltzmann (PB) model and a generalized-Boltzmann (GB) model at igb=5 [25] via Amber9's pbsa protocol [26] with a PARSE charge/radii set, a 1.4 Å solvent probe radius, and a 0.5 Å grid spacing. The solvent's dielectric constant was set to 80, while the dielectric constant was set to 1 in the protein's interior.

The nonpolar part (G_{NP}) accounts for the nonpolar contribution to solvation and was approximated by relating it to the solvent accessible surface area (SASA) with coefficient of 0.0072 [27].

The entropy contribution (– TS) arising from changes in the degrees of freedom (translational, rotational, and vibrational) of the solute molecules was included applying classical statistical thermodynamics. Entropy contribution was calculated using an nmode protocol with a distance dependent dielectric constant [28].

After including all the energetic terms for PI3Kα, PIK75 and the complex Eq. 1 can be reorganized and expressed as:

$$\Delta G_{\mathrm{bind}} = \Delta E_{\mathrm{MM}} + \Delta G_{\mathrm{sol}} - T\Delta S, \tag{5}$$

where ΔE_{MM}, ΔG_{sol} and ΔS are simply the change in the internal energy, the solvation energy and the entropy between PI3Kα, PIK75 and the complex. Binding free energy was calculated using 1400 snapshots sampled with ptraj program every 10 ps; these snapshots cover the last 14 ns of the MD trajectory. Due to the high computational demand, the entropy calculations were performed only for every tenth one of the 1400 snapshots (140 snapshots in total) described above.

Free energy decomposition

To provide further insight into the changes that occur in the energetic profile of the interaction over the course of the trajectory, we extracted and plotted the components of the binding energy with respect to time. Notice that this energy will be called hereafter as binding energy (ΔE_{bind}) and not binding free energy (ΔG_{bind}) since it does not compute average values but just single decomposition in a conformation. Additionally the entropy was not considered because the entropy calculations were performed only for every tenth one of the 1400 snapshots. So the binding energy (the enthalpy, ΔE_{bind}) was calculated according to the equation:

$$\Delta E_{\mathrm{bind}} = \Delta E_{\mathrm{MM}} + \Delta E_{\mathrm{sol}}. \tag{6}$$

In addition, in order to identify the residues that contribute the most to the calculated overall binding energy, we used a residue-by-residue decomposition protocol embedded in the GB solvent model based in MMGBSA. The GB model is an alternative to the PB solvation model that uses a pair-wise analytical approximation of the PB model. Using this model the calculated energies can be further broken down into individual residue's contributions.

Results and discussion

Docking PIK75 to the crystal structure of PI3Kα

Due to no inhibitor-bound PI3Kα crystal has been solved, we applied docking experiments to obtain the inhibitor-bound complex for further study. Like other typical kinases, the ATP-binding pocket of PI3Kα is located between a mostly-helical carboxy-terminal (C) lobe and an amino-terminal (N) lobe of the kinase domain. All known PI3K inhibitors have extensive hydrophobic contacts with the residues around the adenine-binding region, and make backbone hydrogen-bond interactions with the hinge region, which is the short polypeptide connecting the C- and N-terminal lobes. Therefore, according to Knight's work [29], in our docking experiments, we define the poses with hydrophobic contacts with the residues around the adenine-binding region and the hydrogen bond between the nitrogen atom of the imidazole ring of PIK75 and the backbone of the hinge region as the correct poses.

In our docking experiment, 100 docked conformations of PIK75 for PI3Kα were obtained and clustered to 36 clusters using a tolerance of 1.5 Å rmsd. The lowest binding free energy among 100 docked conformations is −9.25 kcal•mol^{-1}, but this cluster only includes two docked conformations. The most populated cluster has 20 conformations, and its lowest binding free energy among 20 conformations is −8.71 kcal•mol^{-1}. The two poses of PIK75 for PI3Kα, named as pose-A and pose-B respectively, are shown in Fig. 2. As seen from Fig. 2, the orientations of the imidazo[1,2-a]pyridine moiety of pose-A and pose-B are very different. The imidazo [1,2-a]pyridine moiety of pose-A is out of the cavity which is very well conserved in all PI3K isoforms and coincides with the adenine-binding region. Pose-B has the critical hydrogen-bonding interaction between the nitrogen atom of the imidazole and the nitrogen atom of backbone of Val851 in PI3Kα, and the imidazo[1,2-a]pyridine moiety inserts deeply into the cavity coinciding with the adenine-binding region. This binding orientation of the imidazo[1,2-a]pyridine moiety is consistent with the results of Han and Zhang [15]. Additionally, the cluster including pose-B is the most populated cluster, 20 out of 100 conformations. Therefore, pose-B was selected as the initial conformation of PIK75 for molecular dynamics simulation of PIK75-bound PI3Kα.

Molecular dynamics simulation of PIK75-bound PI3Kα

To assess the stability of the MD trajectories, the backbone atoms root-mean-square deviation (rmsd) of catalytic kinase domain of PI3Kα and the heavy atoms rmsd of PIK75 from the starting structure of PI3Kα and PIK75 obtained from molecular docking above have been plotted in Fig. 3. As seen from Fig. 3, during the first 3.5 ns, a sharp rise is observed and then the function keeps stable and the rmsd values of PI3Kα converge to a lower value about 2.3 Å. The rmsd of PIK75 has an about 2 Å fluctuation at about 6.0 ns of the simulation then is stable in the rest of the simulation. As can be seen from Fig. 3, the simulation reaches equilibrium within 6.0 ns.

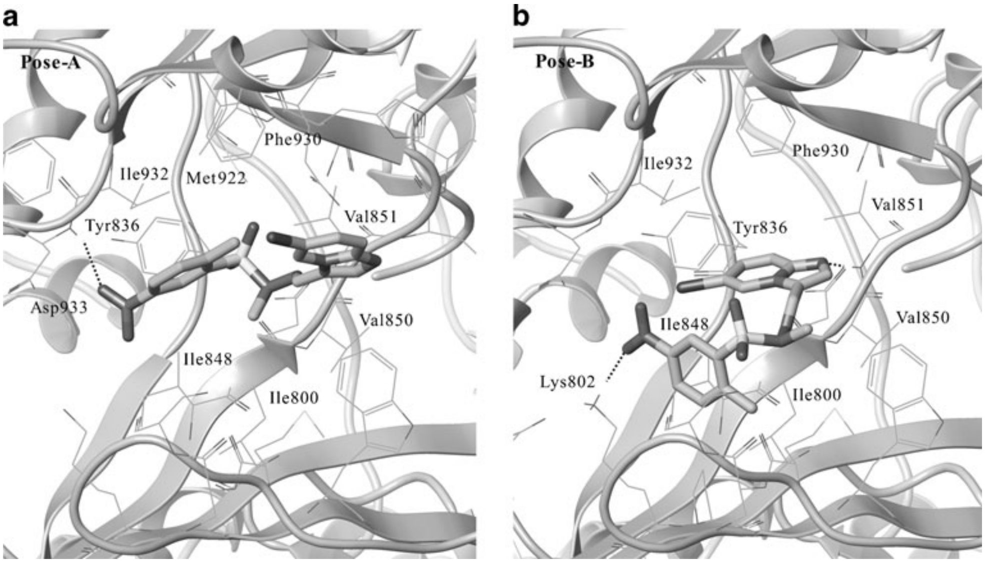


Fig. 2 Two docked conformations of PIK75 for the X-ray crystal structures of PI3Kα. Hydrogen bonds are dashes

The calculated relative binding free energy and contributions of van der Walls, electrostatic interaction and solvation energy using the single trajectory MMPB(GB)SA method are listed in Table 1. As seen from Table 1, the contributions of the molecular mechanics part (ΔE_{MM}) and the solvation part (ΔG_{pb_sol}, ΔG_{gb_sol}) are calculated to be −55.52 kcal•mol^{-1}, 22.11 kcal•mol^{-1} and 18.27 kcal•mol^{-1}, respectively. Adding the entropy contribution (TΔS, −23.37kcal•mol^{-1}) calculated by nmode protocol, the binding free energy (ΔG_{bind}) between PI3Kα and PIK75 is −10.04 kcal•mol^{-1} using MMPBSA method, while −13.88 kcal•mol^{-1} using MMGBSA method, which is beneficial for binding and is good agreement with Han and Zhang's value −11.24 kcal•mol^{-1} (or −12.99 kcal•mol^{-1}) calculated by the formula $\Delta G=-2.303RT \log k_i$ [15]. And the agreement does support the physical relevance of the model and suggest that the decomposition analysis below is meaningful. Therefore, this PI3Kα and PIK75 complex formation exemplifies a classical favorable reaction in solution where the increase of the stability produced by the formation of the complex overcomes the cost of the entropy and desolvation of protein and ligand.

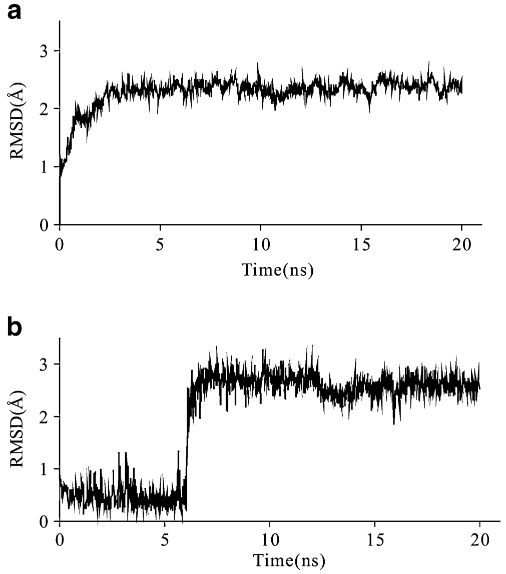


Fig. 3 RMSD of the backbone atoms of the catalytic kinase domain of PI3Kα (**a**) and the heavy atoms of PIK75 (**b**)

From an energy component point of view, the PI3Kα/PIK75 complex formation leads to strongly favorable Coulombic interactions (ΔE_{ele}, −16.49kcal•mol^{-1}), opposed by disfavorable contributions due to the polar part of solvation free energy (ΔG_{pb}, 27.82kcal•mol^{-1}; ΔG_{gb}, 23.98kcal•mol^{-1}). So the total electrostatic contribution is 11.33 kcal•mol^{-1} using PB model, while 7.49 kcal•mol^{-1} using GB model, and thus disfavors complex formation. And the complex

Table 1 Energy components and binding free energy for the PI3Kα/PIK75 complex

	Mean(kcal·mol^{-1})	Std(kcal·mol^{-1})
ΔE_{ele}	−16.49	5.63
ΔE_{vdw}	−39.03	3.62
ΔE_{MM}	−55.52	6.69
ΔG_{pb_sur}	−5.71	0.37
ΔG_{pb}	27.82	6.02
ΔG_{pb_sol}	22.11	5.77
ΔG_{gb_sur}	−5.71	0.37
ΔG_{gb}	23.98	6.35
ΔG_{gb_sol}	18.27	6.10
ΔH_{pb}	−33.41	9.60
ΔH_{gb}	−37.25	8.89
TΔS	−23.37	14.53
$\Delta G_{bind(pb)}$	−10.04	
$\Delta G_{bind(gb)}$	−13.88	

formation also leads to favorable van der Wales interactions (ΔE_{vdw}, −39.03 kcal•mol^{-1}), added by favorable contributions due to the non-polar part of solvation free energy (ΔG_{pb_sur}, ΔG_{gb_sur}, −5.71kcal•mol^{-1}). So the total hydrophobic interaction contribution is −44.74kcal•mol^{-1} and thus favors complex formation. Therefore, it is concluded that both the electrostatic and van der Waals/hydrophobic interactions are important for binding. These results suggest that a potential PI3Kα inhibitor should be designed to interact with PI3Kα by the stronger electrostatic and van der Waals/hydrophobic interactions which can increase the contribution of molecular mechanics. Additionally, this inhibitor should be a more rigid one, because it might reduce the entropy lost and improve the affinity.

Decomposition of the binding energy, ΔE_{bind}, over the course of the trajectory

In order to gain an insight into the factors that may drive the formation of the PI3Kα/PIK75 complex, the plot of the decomposition of the binding energy (ΔE_{bind}) into the molecular mechanics (ΔE_{MM}) and solvation (ΔE_{sol}) parts is undertaken, see Fig. 4a-b. In the energy plot, three stages can be identified as a consequence of the different energy patterns of ΔE_{bind}, ΔE_{MM} and ΔE_{sol} at specific trajectory regions. Stage I, which is called preparation stage, is characterized by an overall stabilization of all parts and it lasts approximately until 6.0 ns. In stage II, from 6.0 ns to 12.5 ns, although the molecular mechanics energy (ΔE_{MM}) decreases and the solvation energy (ΔE_{sol}) increases, the binding energy (ΔE_{bind}) is stable and approximately equal to that of stage I. In stage III, the last 7.5 ns, the binding energy is lower than that of stage I and II, which is the global minimum of the binding energy over the course of the trajectory. As seen from Fig. 4a-b, the favorable molecular mechanics energy component and the unfavorable solvation energy component offset each other in all of stages. And the molecular mechanics energy makes the prominent contribution to the binding energy. Especially, from stage II to III, the drop of the molecular mechanics energy where the solvation energy is relatively stable makes the drop of the binding energy.

Therefore, the decomposition of the molecular mechanics energy (ΔE_{MM}) as a crucial part of ΔE_{bind} into the electrostatic (ΔE_{ele}), van der Waals (ΔE_{vdw}) and internal (ΔE_{int}) can contribute to clarify which kind of interaction causes the shift of the molecular mechanics energy. As seen from Fig. 5, van der Waals interaction steadily contributes to the molecular mechanics energy over the course of the trajectory, while electrostatic interaction is positively correlated with the molecular mechanics energy. So this result suggests that the electrostatic term makes the prominent contribution to the molecular mechanics energy.

As well known, in most docking programs, the number of H-bonds is explicitly taken into account in the scoring function. However, in MD simulation this H-bond contribution is not explicitly calculated by default because H-bond is included to electrostatic interaction as a particular aspect of electrostatic interaction.

So to investigate the hydrogen bonding interaction between PI3Kα and PIK75 in detail, the hydrogen bonding interactions were clustered based on PIK75, see Table 2. As seen from Table 2, the hydrogen bond between the nitrogen atom of the imidazole of PIK75 and the nitrogen atom of backbone of Val851 in PI3Kα is very stable in all of three

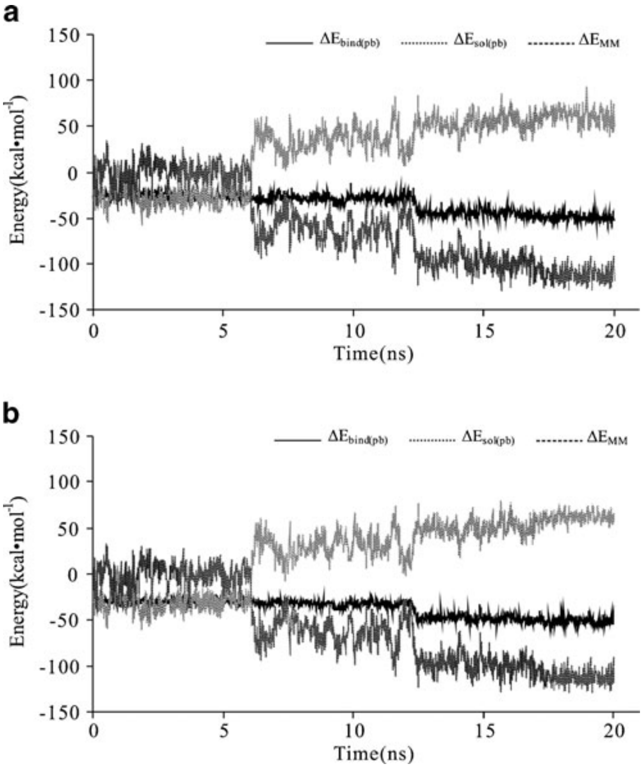


Fig. 4 Evolution of ΔE_{bind} decomposition according to ΔE_{bind}= ΔE_{MM}+Δ_{Esol} (**a**) MMPBSA (**b**) MMGBSA

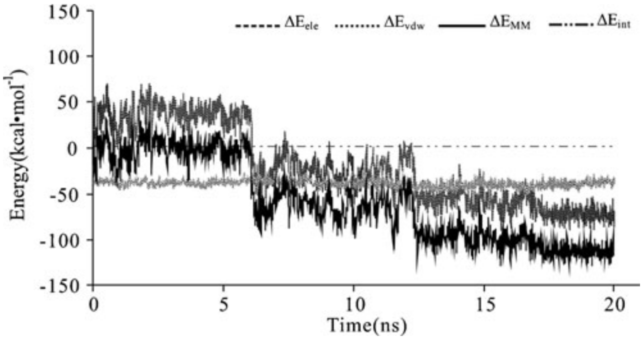


Fig. 5 Evolution of ΔE_{MM} decomposition according to ΔE_{MM}= ΔE_{int}+ΔE_{ele}+ΔE_{vdw}.

Table 2 Hydrogen bonds of all of trajectories

Hydrogen bond		Occupancy			Distance(Å)
PIK75	PI3Kα	Stage I	Stage II	Stage III	
Imidazole ring-N	NH-Val851	97.8	99.2	99.4	2.918(0.105)
Nitro-O1	OH-Ser773	0	0	33.8	2.792(0.207)
Nitro-O1	NH-Ser773	0	0	92.6	2.941(0.176)
Nitro-O2	OH-Ser774	0	0	99.2	2.684(0.143)
Nitro-O2	NH-Ser774	0	0	98.6	2.866(0.119)
Sulfonyl-O	NH-Gln859	0	0	17.4	3.016(0.213)
Sulfonyl-O	NH-Thr856	0	0	46.3	3.138(0.144)

Hydrogen bonds were defined by acceptor···donor atom distances of<3.2 Å and acceptor···H-donor angles of>120°. Hydrogen bonds are reported only if they exist for >10% of the investigated time period.Occupancy is in units of percentage of the investigated time period

stages, which agrees with Ming Han and John Z. H. Zhang's work [15] and suggests that this hydrogen bond plays a very crucial role in the binding of PI3Kα and PIK75. Three hydrogen bonds, which are between two oxygen atoms of nitro of PIK75 and the oxygen atom of side chain of Ser774, the nitrogen atom of backbone of Ser773 and the nitrogen atom of backbone of Ser774, are not observed in stage I and II but are formed in all of stage III, and the oxygen atom of side chain of Ser773 forms a hydrogen bond with one oxygen atom of nitro of PIK75 in the last about 3.0 ns of stage III, which can account for the drop of the molecular mechanics energy and the binding energy from stage II to stage III and is very well beneficial for the binding. Additionally, the oxygen atom of sulfuryl of PIK75 forms a hydrogen bond with the nitrogen atom of Gln859 in the starting of stage III, then the nitrogen atom of Thr856 in the last of 5.5 ns of stage III, which further helps the binding.

Therefore, these results above suggest that both the van der Waals/hydrophobic and electrostatic interactions are important for the binding. Especially hydrogen bond interaction is a crucial factor for the binding and responsible for the change of the binding energy directly.

Decomposition of binding energy on a per-residue basis

In order to examine the residues contribution to the whole binding, the binding energy (the binding enthalpy) decomposition method by residue was used. As seen from Fig. 6, the contribution of individual residue to binding varies in the range of +0.3 to −6.0 kcal•mol^{-1}. The significant residues to binding are mainly located in two regions, the conserved hydrophobic region of PI3Kα made up of Ile800, Tyr836, Ile848, Val850, Val851, Met922, Phe930, and Ile932 and the P-loop (residues Ile771-Leu779) region at the active site cleft of PI3Kα. The contribution of the conserved hydrophobic region is −0.83 kcal•mol^{-1} of Ile800, −1.03 kcal•mol^{-1} of Tyr836, −0.83 kcal•mol^{-1} of Ile848, −2.47 kcal•mol^{-1} of Val850, −4.31 kcal•mol^{-1} of Val851, −1.51 kcal•mol^{-1} of Met922, −0.23 kcal•mol^{-1} of Phe930 and −2.66 kcal•mol^{-1} of Ile932, respectively, which occupies about 37.1% of the binding enthalpy (ΔH_{gb}, −37.25 kcal•mol^{-1}) and suggests that the residues of the conserved hydrophobic region are very important to the binding, especially Tyr836, Val850, Val851, Met922 and Ile932. The contribution of the P-loop region is −5.66 kcal•mol^{-1} of Met772, −3.36 kcal•mol^{-1} of Ser773 and −2.91 kcal•mol^{-1} of Ser774, which occupies about 32.0% of the binding enthalpy. Additionally the contribution of Trp780, His855 and Thr856 is −1.49,−1.24 and −1.10 kcal•mol^{-1} respectively, which shows that these residues are very important to the binding also. Therefore, PI3Kα possesses two binding "hot spots": the conserved hydrophobic adenine region and the P-loop region which includes Ile771-Leu779.

Dynamics analysis of the interactions between PI3Kα and PIK75

The trajectory file of 20ns molecular dynamics simulation of PIK75-bound PI3Kα was clustered to six clusters by the average-linkage clustering algorithm and the representative structures were extracted from every cluster. The cluster distribution along the simulation time 20ns is as follow: the

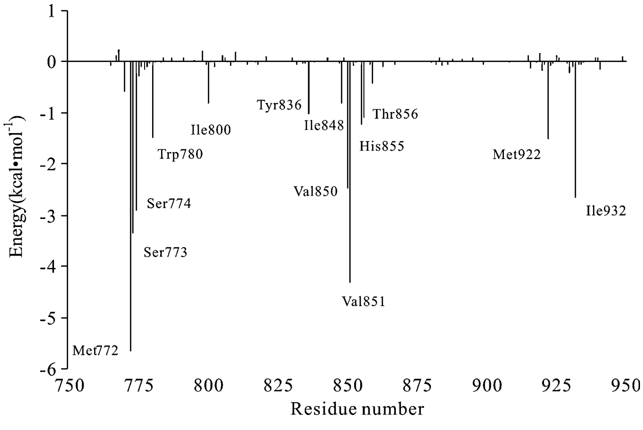


Fig. 6 Decomposition of $\Delta E_{MM}+\Delta G_{sol}$ (the binding enthalpy) on a per-residue for residues of the catalytic kinase domains of PI3Kα

six cluster respectively include the snapshots of 0~4.0 ns, 4.0~6.0 ns, 6.0~9.0 ns, 9.0~12.5 ns, 12.5~13.5 ns and 13.5~20.0 ns, which account for 20%, 10%, 15%, 17.5%, 5% and 32.5% of the ensemble of PIK75-bound PI3Kα respectively. The representative conformations of PIK75-bound PI3Kα of every cluster are shown in Fig. 7.

In the simulation of PIK75-bound PI3Kα, as seen from Fig. 7a-f, the conformation of PIK75 keeps changing within 0~13.5 ns (the first cluster to the fifth cluster), but is stable within the last 6.5 ns simulation, which is the last clusters and accounts for 32.5% of the ensemble of PIK75-bound PI3Kα. The flat 6-bromine imidazo[1,2-a]pyridine moiety of PIK75 is stable within 20 ns simulation, which inserts deeply in the conserved hydrophobic region of PI3Kα coinciding with the adenine-binding region, and accommodates with this region. The key hinge hydrogen bond

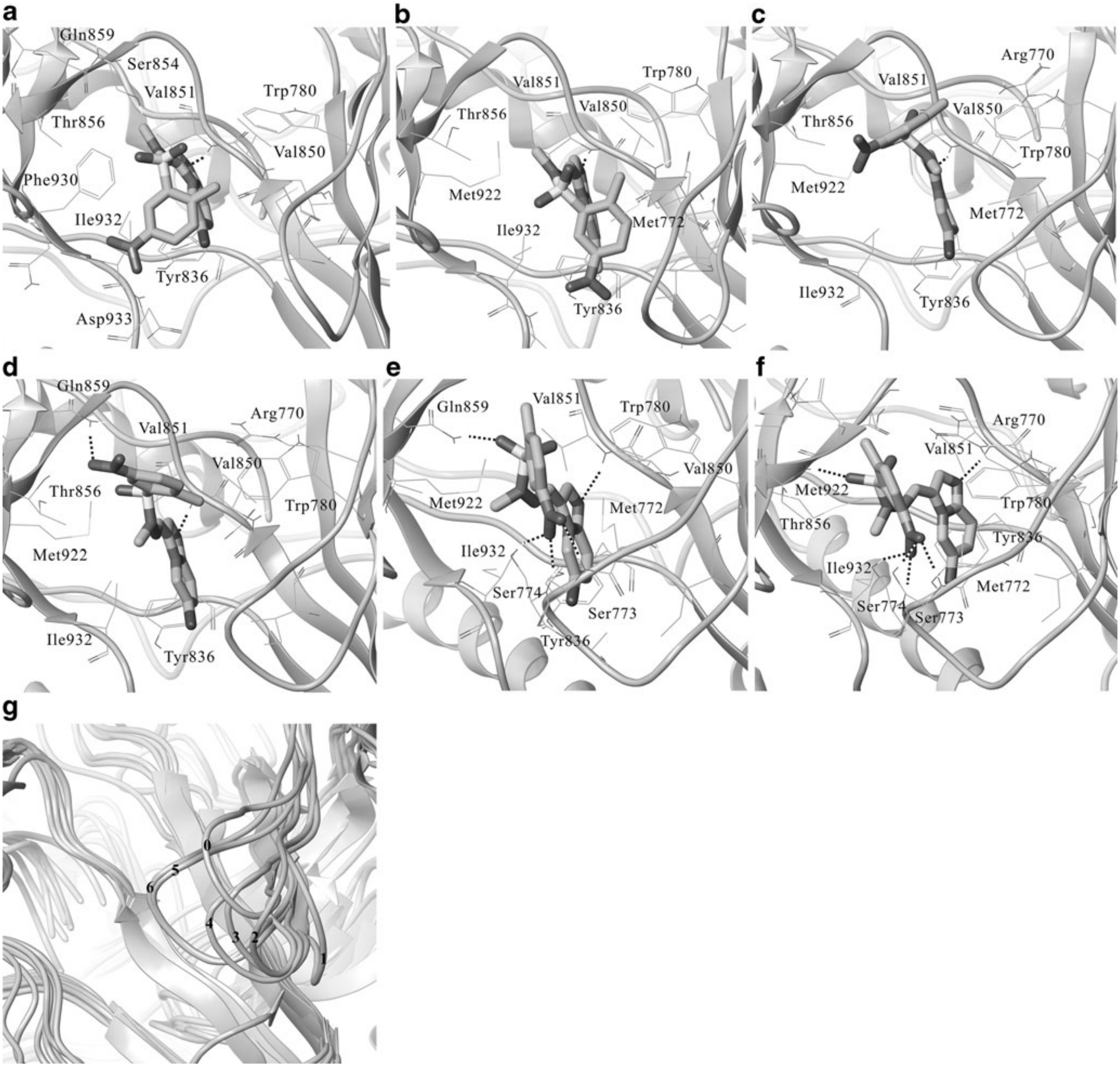


Fig. 7 Six representative structures from clustering the molecular dynamics simulation of PIK75-bound PI3Kα by the average-linkage clustering algorithm (**a**-**f**) and six representative structures superimposed onto the docked structure in PI3Kα obtained from docking PIK75 to the apo X-ray structure of PI3Kα (**g**). (**a**) from the first cluster, (**b**) from the second cluster, (**c**) from the thirst cluster, (**d**) from the fourth cluster, (**e**) from the fifth cluster, (**f**) from the sixth cluster, (**g**) 0, 1, 2, 3, 4, 5, and 6 represent P-loop region of the PI3Kα from the docked structure obtained from docking PIK75 to the apo X-ray structure of PI3Kα and the representative structure from the first cluster, the second cluster, the third cluster, the fourth cluster, the fifth cluster and the sixth cluster, respectively. Hydrogen bonds are dashes

between the nitrogen atom of the imidazo[1,2-a]pyridine moiety of PIK75 and the residue Val851 is stable during all of 20 ns simulation, and there is a T-shaped contact between the aromatic rings of Tyr836 and imidazo[1,2-a]pyridine ring of PIK75, which agreed with Han and Zhang's observation [15] and can account for the important contribution of the conserved hydrophobic region to the binding using the free energy decomposition method by residue above.

However, the 2-methyl-5-nitrophenyl group of PIK75 is observed to show significant conformational flexibility within 0~13.5 ns simulation (the first cluster to the fifth cluster) as seen from Fig. 7a-e. In the last of 6.5 ns simulation, it is very interesting that the nitro moiety of PIK75 extends to the P-loop region and has four hydrogen-bond arms with the backbone and side chain of Ser773 and Ser774 as shown in Fig. 7f. Additionally, in this stage, the oxygen atom of the sulfonyl group of PIK75 forms hydrogen bond with Thr856, and the phenyl group of PIK75 has a T-shaped contact with Trp780. These interactions help lock PIK75 in the ATP-binding pocket of PI3Kα and make the binding between PI3Kα and PIK75 more stable, which can explain the important contribution of P-loop region, Thr856, and Trp780 to the binding using the free energy decomposition method by residue above.

In the simulation of PIK75-bound PI3Kα, as seen from Fig. 7g, the catalytic kinase domain of is stable after 4.0 ns, except for the P-loop region. The P-loop moves toward the outside of the ATP-binding pocket within 0~6.0 ns as seen from Fig. 7a, b, g, then moves toward the inside of the ATP-binding pocket within 6.0~12.5 ns as seen from Fig. 7c, d, g. The movement made the P-loop to adopt a more close conformation than that in the ligand-free structure of PI3Kα, which agrees with the knowledge that the ligand-free structures are in the most open form. Then due to the PIK75-induced effect, the P-loop further moves toward the inside of the ATP-binding pocket within the last 7.5 ns as seen from Fig. 7e, f, g and forms stable interactions with the 2-methyl-5-nitrophenyl moiety of PIK75 by hydrogen bond and hydrophobic interactions. Therefore, our MD simulation of PIK75-bound PI3Kα reflects the course of induced fit effect of PIK75 for PI3Kα.

Additionally, six clusters of PIK75-bound PI3Kα simulation generated by the average-linkage clustering algorithm are consistent with the change in the energy profile above (Figs. 4 and 5). The binding energy of the first and second clusters is approximately equal to that of the third and fourth clusters. Due to the PIK75-induced effect, PIK75 forms more stable interactions with PI3Kα in the fifth and sixth clusters. The binding energy of the fifth and sixth clusters is lower than that of the first to fourth clusters, which is the global minimum of the binding energy over the course of the trajectory. Therefore, the representative conformation extracted from the sixth cluster should be the stable binding mode of PIK75 for PI3Kα.

Hence, according to the MD simulation, the binding mode of PIK75 for PI3Kα is predicted. The conserved hydrophobic adenine region of PI3Kα made up of Ile800, Ile848, Val850, Val851, Met922, Phe930, and Ile932 accommodates the flat 6-bromine imidazo[1,2-a]pyridine ring of PIK75. The NH of Val851 forms the conserved hydrogen with the nitrogen atom of the imidazole in PIK75, and there is a T-shaped contact between the aromatic ring of Tyr836 in PI3Kα and imidazo[1,2-a]pyridine ring of PIK75. The 2-methyl-5-nitrophenyl group of PIK75 extends to the P-loop region, and has four hydrogen-bond arms with the backbone and side chain of Ser773 and Ser774. There is a T-shaped contact between the aromatic ring of Trp780 and the benzene ring of PIK75.

PI3Kα shares~35% sequence identity with PI3Kγ isoforms, and the kinase domain of PI3Kα shares~43.5% sequence identity with that of PI3Kγ, which makes it more challenging to find inhibitors with high selectivity between PI3Kα and PI3Kγ. According to Hayakawa [10], PIK75 is the most selective PI3Kα inhibitor reported to date, which inhibits PI3Kα and PI3Kγ with IC_{50} values of 0.0003 and 0.040 μM, respectively. In the binding mode of PIK75 for PI3Kα predicted in our study PIK75 forms stable interactions with the conserved hydrophobic adenine region and P-loop region of PI3Kα. The residues lining the ATP-binding pockets of PI3Kα and PI3Kγ isoforms are highly conserved, but the residues of P-loop region of PI3Kα and PI3Kγ are obviously different. The P-loop in PI3Kα is IMSSAKRPL (residues Ile771-Leu779), while the corresponding loop in PI3Kγ is VMASKKKPL (residues Val803-Leu811). So it is suggested the P-loop plays an important role in the selectivity profile of PIK75. Amzel et al. [12] compared the X-ray crystal structures of PI3Kα and PI3Kγ and speculated that the different conformations of the P-loops of PI3Kα and PI3Kγ could be exploited for the design of the isoform-specific PI3Kα inhibitors. So it is speculated that the critical difference between the P-loop sequence of PI3Kα and PI3Kγ causes the different conformation of the two P-loops and the different conformation of the two P-loops causes specific interactions with PIK75 when they bind to PIK75.

Conclusions

The prevalence of PI3Kα signaling abnormalities in human cancer cells has made PI3Kα an attractive target for anticancer drug discovery. This work is to systematically investigate the interactions between PI3Kα and PIK75 which is the most selective PI3Kα inhibitor reported to date via combined molecular docking and molecular dynamics simulation. The binding free energy (ΔG_{bind}) between PI3Kα and PIK75 is −10.04 kcal•mol^{-1} using MMPBSA method, while −13.88 kcal•mol^{-1} using MMGBSA method, which is beneficial for the binding. And this complex formation exemplifies a classical favorable reaction in solution where the increase of the stability produced by the formation of the complex overcomes the cost of the entropy and desolvation of protein and ligand. PI3Kα possesses two binding "hot spots": the conserved hydrophobic adenine region which is made up of Ile800, Ile848, Val850, Val851, Met922, Phe930, and Ile932; the P-loop region which includes Ile771-Leu779. And the P-loop region is speculated to be responsible for the selectivity profile of PIK75. The predicted binding mode of PIK75 for PI3Kα presented in this study could be very useful for the discovery of more promising compounds to target PI3Kα.

Acknowledgments This work was financially supported by Natural Science Foundation of Shaanxi Province (NO. SJ08C207) and the Fundamental Research Funds for the Central Universities.

References

1. Bader AG, Kang S, Zhao Li, Vogt PK (2005) Nat Rev Cancer 5:921–929
2. Vivanco I, Sawyers CL (2002) Nat Rev Cancer 2:489–501
3. Vanhaesebroeck B, Waterfield MD (1999) Exp Cell Res 253:239–254
4. Domin J, Waterfield MD (1997) FEBS Lett 410:91–95
5. Huang CH, Mandelker D, Schmidt-Kittler O, Samuels Y, Velculescu VE, Kinzler KW, Vogelstein B, Gabelli SB, Amzel LM (2007) Science 318:1744–1748
6. Perrone F, Lampis A, Orsenigo M, Bartolomeo MD, Gevorgyan A, Losa M, Frattini M, Riva C, Andreola S, Bajetta E, Bertario L, Leo E, Pierotti MA, Pilotti S (2009) Ann Oncol 20:84–90
7. Samuels Y, Wang Z, Bardelli A, Silliman N, Ptak J, Szabo S, Yan H, Gazdar A, Powell SM, Riggins GJ, Willson JK, Markowitz S, Kinzler KW, Vogelstein B, Velculescu VE (2004) Science 304:554–554
8. Liu I, Cheng H, Roberts TM, Zhao JJ (2009) Nat Rev Drug Discov 8:627–644
9. Chang HW, Aoki M, Fruman D, Auger KR, Bellacosa A, Tsichlis PN, Cantley LC, Roberts TM (1997) Vogt PK Science 276:1848–1850
10. Hayakawa M, Kawaguchi K, Kaizawa H, Koizumi T, Ohishi T, Yamano M, Okada M, Ohta M, Tsukamoto S, Raynaud FI, Parker P, Workman P, Waterfield MD (2007) Bioorg Med Chem 15:5837–5844
11. Li YP, Wang YW, Zhang FQ (2010) J Mol Model 16:1449–1460
12. Amzel LM, Huang CH, Mandelker D, Lengauer C, Gabelli SB, Vogelstein B (2008) Nat Rev Cancer 8:665–669
13. Zvelebil MJ, Waterfield MD, Shuttleworth SJ (2008) Arch Biochem Biophys 477:404–410
14. Frederick R, Denny WA (2008) J Chem Inf Model 48:629–638
15. Han M, Zhang JZH (2010) J Chem Inf Model 50:136–145
16. Sabbah DA, Vennerstrom JL, Zhong H (2010) J Chem Inf Model 50:1887–1898
17. Sanner MF (1999) J Mol Graph Model 17:57–61
18. Huey R, Morris GM, Olson AJ, Goodsell DS (2007) J Comput Chem 28:1145–1152
19. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W (2004) Gaussian 03. Gaussian Inc, Wallingford, CT
20. Wang J, Wang W, Kollman PA, Case DA (2006) J Mol Graph Model 25:247–260
21. Case DA, Darden T, Cheatham T III, Simmerling C, Wang J, Duke R, Luo R, Crowley MW, Zhang RCW, Merz K, Wang B, Hayik S, Roitberg A, Seabra G, Kolossvary I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell S, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews D, Seetin M, Sagui C, Babin V, Kollman P (2008) AMBER 10. University of California, San Francisco
22. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) J Comput Chem 25:1157–1174
23. Darden T, York D, Pedersen L (1993) J Chem Phys 98:10089–10092
24. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE (2000) Accounts Chem Res 33:889–897
25. Onufriev A, Bashford D, Case DA (2004) Proteins 55:383–394
26. Luo R, David L, Gilson MK (2002) J Comput Chem 23:1244–1253
27. Sitkoff D, Sharp KA, Honig B (1994) J Phys Chem 98:1978–1988
28. Case DA (1994) Curr Opin Struct Biol 4:285–290
29. Knight ZA, Gonzalez B, Feldman ME, Zunder ER, Goldenberg DD, Williams O, Loewith R, Stokoe D, Balla A, Toth B, Balla T, Weiss WA, Williams RL, Shokat KM (2006) Cell 125:733–747

J Mol Model (2012) 18:1917–1925
DOI 10.1007/s00894-011-1212-3

ORIGINAL PAPER

Structure-based functional inference of hypothetical proteins from *Mycoplasma hyopneumoniae*

Marbella Maria da Fonsêca · Arnaldo Zaha · Ernesto R. Caffarena · Ana Tereza Ribeiro Vasconcelos

Received: 20 June 2011 / Accepted: 5 August 2011 / Published online: 26 August 2011

Abstract Enzootic pneumonia caused by *Mycoplasma hyopneumoniae* is a major constraint to efficient pork production throughout the world. This pathogen has a small genome with 716 coding sequences, of which 418 are homologous to proteins with known functions. However, almost 42% of the 716 coding sequences are annotated as hypothetical proteins. Alternative methodologies such as threading and comparative modeling can be used to predict structures and functions of such hypothetical proteins. Often, these alternative methods can answer questions about the properties of a model system faster than experiments. In this study, we predicted the structures of seven proteins annotated as hypothetical in *M. hyopneumoniae*, using the structure-based approaches mentioned above. Three proteins were predicted to be involved in metabolic processes, two proteins in transcription and two proteins where no function could be assigned. However, the modeled structures of the last two proteins suggested experimental designs to identify their functions. Our findings are important in diminishing the gap between the lack of annotation of important metabolic pathways and the great number of hypothetical proteins in the *M. hyopneumoniae* genome.

Keywords Comparative modeling · Known function · Modeller · Mollicutes · Threading

Introduction

Mycoplasmas belong to the class Mollicutes and number approximately 200 species, among which are obligate parasites of humans and commercially important mammals [1] such as pigs. Mycoplasmas are wall-less bacteria distinguished by small genomes of low G+C content. The parasitism, the reduced genome, and the close association of these bacteria with their hosts have contributed to the absence of enzymes involved in important biosynthetic pathways in mycoplasma [2].

Enzootic pneumonia caused by *Mycoplasma hyopneumoniae* is a major constraint to efficient pork production worldwide. The *M. hyopneumoniae* genome contains 920,079 base pairs and 716 protein-coding genes, of which 418 encode proteins that are homologous to proteins with known functions. Currently, there are nearly 1,500 complete genome sequences in GenBank, and half of all of the predicted genes encode proteins having no inferable functions. Similarly, almost 42 % of predicted *M. hyopneumoniae* genes correspond to proteins annotated as hypothetical [3]. This lack of annotation is a particularly intriguing and unsolved issue because, as mentioned above, components of important and essential metabolic pathways present in other organisms have not been identified in mycoplasmas [4, 5].

M. M. da Fonsêca
Universidade Federal do Rio de Janeiro,
Rio de Janeiro, RJ, Brazil

A. Zaha
Laboratório de Genômica Estrutural e Funcional,
Centro de Biotecnologia, UFRGS,
Porto Alegre, RS, Brazil

E. R. Caffarena
Programa de Computação Científica, Fundação Oswaldo Cruz,
Rio de Janeiro, RJ, Brazil

M. M. da Fonsêca · A. T. R. Vasconcelos (✉)
Laboratório Nacional de Computação Científica,
Laboratório de Bioinformática,
Petrópolis 25651-075 RJ, Brazil
e-mail: atrv@lncc.br

The BLAST program [6] has contributed significantly to the analysis of nucleotide and amino acid sequences, allowing the prediction of biological functions and evolutionary relationships of genes and proteins [7]. However, this tool can be used with a high degree of confidence only when the sequences are evolutionarily close to each other and the identity between them is over 50%. To overcome these limitations, alternative methodologies such as threading and homology modeling have been used to answer questions about protein properties. These methods are possible because biological processes such as gene duplication and evolutionary divergence occur in many distantly related organisms [8], giving rise to structurally and functionally similar families of proteins. When one or more proteins in a family have experimentally determined structures, it is feasible to model the structures of many other members with reasonable accuracy. This condition is particularly true when the sequence identity between protein domains is ≥30% and larger than 100 residues.

Threading and homology modeling can identify domains and active sites, aiding in placing their locations within a 3D structure (i.e.,surface or buried). Because the determination of a crystal structure is an arduous and sometimes impractical task for some proteins, the homology modeling methodology is a helpful approach that can guide further experimental assays to investigate protein function [9–11]. The rapid growth of structural genomics is producing a considerable number of templates that can be used for homology modeling. The availability of more templates increases the quality of new models, thereby diminishing the gap between computationally derived models and experimental outcomes.

Thus far, mycoplasma genome sequences have not been annotated for activities related to the utilization of ATP, NAD and NADH and amino acid synthesis derived from pyruvate. However, genes corresponding to these activities must exist, otherwise their enzymatic activities would not have been found [12]. This discrepancy suggests that sequence-based methodologies for identifying protein function may not be suitable for mycoplasmas in some cases.

In this study, using structure-based approaches, we were able to predict the function of seven proteins annotated as hypothetical in the *M. hyopneumoniae* genome. Three of the proteins are involved in metabolic processes, a finding that may enhance further studies concerning the metabolism of this bacterium. Another two proteins are involved in transcription, controlling gene expression based on cellular or environmental signals, an important characteristic of pathogenic bacteria such as *M. hyopneumoniae*. Functions for the other two proteins could not be assigned, but their modeled structures suggest experimental designs, which will allow future investigation concerning their function.

Materials and methods

The sequences of 298 proteins belonging to *M. hyopneumoniae* strain 7448, currently annotated as hypothetical in the Genesul database (http://www.genesul.lncc.br/finalMP/), were submitted to two threading programs, GenThreader [13] and Prospect-PSPP [14]. Additionally, these data were analyzed by InterProScan [15] and COG [16], and the functional predictions of these four programs were compared. Thirty-four sequences with the same functional predictions given by at least two of the mentioned programs were selected for manual analysis, resulting in the further selection of seven targets for structural investigation. Firstly, the sequences of these seven proteins were submitted to a PSI-BLAST search at http://blast.ncbi.nlm.nih.gov/Blast.cgi against the Protein Data Bank (PDB). To guide the functional inference of uncharacterized proteins, other bioinformatics tools were used as described elsewhere [17]. These other tools suggested scans against sequence pattern, domain, and family classification databases, as well as structural family databases, to identify conserved, functional residues and to extract homologs for post-hoc comparative modeling.

The local alignment between sequences of the seven selected proteins and their templates provided by threading results was performed using the EMBL/EBI software MAFFT [18] with little manual editing. Sequences were retrieved from NCBI and GeneSul. The BLOSUM30 matrix was used with gap and extension penalties of 1.0 and 0.123, respectively. Afterward, the alignment was used to model the selected proteins with the Modeller program [19] (version 9v8). The overall geometric and stereochemical qualities of the structures were assessed using PROCHECK through the PDBsum server [20] and PROSA-web [21] and are listed in Table 1.

Results and discussion

Threading is based on sequence-to-structure alignment. The target sequence is “threaded” through each template present in databases that contain all known protein folds. Threading is performed by using measures for fitness for each type of amino acid in local structural environments and defined in terms of solvent accessibility and protein secondary structure. If a sequence fits well with a given fold, conserved residues are likely shared suggesting similar functions [22].

The PROSPECT-PSPP threading pipeline showed that 27 (9.06%) of 298 target proteins gave PSI-BLAST hits against the PDB with an E-value$<$0.0001, indicating the existence of homologs. Additionally, 83 (27.85%) of the proteins had hits against PDB with a Z-score >20, indicating that the fold recognition confidence level was >99%; the remainder of the

Table 1 Sequence and structure information of the selected proteins and their templates

Protein ID	Template[a]	Identity	Evaluation		Proposed function
			PROSA[b]	Ramachandran[c]	
YP_287866			−7.6	97.9 %	Nicotinic acid mononucleotide adenylyltransferase
	2H29	35 %	−8.66		
	2O08	23 %	−7.28		Putative metal-dependent phosphohydrolase (HD domain)
	2OGI	24 %	−8.03		
YP_287786			−6.17	97 %	Nicotinic acid Phosphoribosyltransferase (NAPRTase)
	1YTK	25 %	−8,93		
YP_287675			−7.22	96.5 %	Flavin Adenine Dinucleotide (FAD) synthetase
	1S4M	20 %	−8.68		
YP_287559			−4.52	95.1 %	Participates in the antitermination process (NusB)
	1EY1	21 %	−4.59		
	2JR0	23 %	−5.45		
	1TZT	23 %	−6.52		
	1Q8C	15 %	−6.75		
	1EYV	18 %	−4.6		
YP_288024			−4.22	96.2 %	Key regulator of bacterial transcription initiation (SigE, Sigma-28)
	2Z2S	18 %	−6.33		
	1RP3	20 %	−7.73		
YP_287971			−5.92	100 %	Unknown function. Likely binds to nucleic acids (YlxR)
	1G2R	29 %	−5.47		
YP_288034			−5.9	96.6 %	Unknown function. Likely binds to nucleic acids (YrdC)
	1HRU	18 %	−7.4		

[a] PDB ID

[b] Favored and allowed regions

[c] Z- score templates

proteins had hits with confidence levels between 85 and 99%. The GenThreader results had high confidence levels (certain) for 84 (32.43%) of 259 proteins (total number of hypothetical sequences available in 2005). Detailed information analysis obtained by threading provided interesting and consistent results, which helped us to select seven proteins having the same prediction by the both mentioned programs. In addition, we followed the protocol suggested by Mazumder and Vasudevan [17], as mentioned in Materials and methods. The results proposed homologs with 3D structures available, thereby providing new knowledge to be applied for comparative modeling.

In the following sections, we will discuss the 3D structures and functions predicted for the seven proteins (YP_287866, YP_287786, YP_287675, YP_287559, YP_288024, YP_287971 and YP_288034). Table 1 lists the templates used to obtain the 3D structures and information about the selected protein models.

Completing the NAD biosynthesis pathway

The 3D structure of hypothetical protein YP_287866 exhibits similarity to portions of two different proteins, i.e., the N-terminal region of nicotinate-nucleotide adenylyltransferase (NadD) and the C-terminal region of an uncharacterized histidine-aspartate (HD) domain. Although the steps in NAD biosynthesis and recycling can vary between species, the enzymes involved in these pathways are generally the following: 1) nicotinate phosphoribosyltransferase (NAPRTase) (EC 2.4.2.11), 2) nicotinate mononucleotide adenylyltransferase (NaMNAT or NadD) (EC 2.7.7.1), and 3) NAD synthetase (NadE) (EC 6.3.1.5) (Fig. 1). These enzymes are encoded by the conserved genes *pncB*, *nadD* and *nadE*, respectively. Enzymes involved in NAD biosynthesis have been considered as promising drug targets because they are essential for the viability of most bacteria [23, 24]; however, only *nadE* is annotated in *M. hyopneumoniae*. Because NadD is likely essential, characterization of this enzyme using a structure-based approach for *M. hyopneumoniae* will improve its annotation and add this enzyme to the list of potential therapeutic targets.

The sequence similarity between the YP_287866 N-terminal region and other nicotinate-nucleotide adenylyltransferases is low (approximately 30%); however, the proteins share two highly conserved ATP-binding motifs,

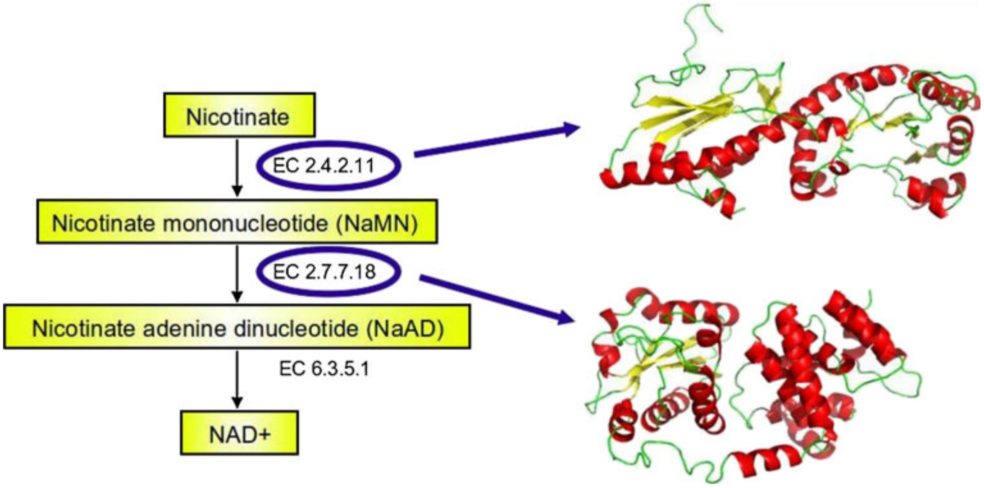

Nicotinate
EC 2.4.2.11
Nicotinate mononucleotide (NaMN)
EC 2.7.7.18
Nicotinate adenine dinucleotide (NaAD)
EC 6.3.5.1
NAD+

Fig. 1 Simplified NAD biosynthesis pathway proposed for *M. hyopneumoniae*. Highlighted in blue circles are the EC numbers of the enzymes whose 3D structure was predicted in this study. YP_287786 is proposed to be EC 2.4.2.11, a nicotinate phosphoribosyltransferase. YP_287866 (N-terminal region) is suggested to be a nicotinate-nucleotide adenylyltransferase, EC 2.7.7.18. EC 6.3.5.1 is the enzyme NadE, already annotated in *M. hyopneumoniae*. The 3D structures were obtained using comparative modeling methodology, and the structures were rendered with Pymol (www.pymol.org)

GXXXPX(T/H)XX and SX(T/S)XXR. The crystal structures of many NaMNAT proteins [25–29] reveal the residues involved in their function, such as the following: 1) His20, Ser162, Arg167 and the essential His17 in the enzymes from *Pseudomonas aeruginosa* [30], *Escherichia coli* [31] and *B. subtilis* [28], located in the ATP binding site, 2) Thr87 and Trp117 that interact with the substrate nicotinic acidyl, and 3) Arg134 that interacts with the adenosine.

The template selected to obtain the 3D structure of the YP_287866 N-terminal region was the crystal structure of nicotinic acid mononucleotide adenylyltransferase from *Staphylococcus aureus* [26] (PDB ID: 2H29). The sequence identity between these two proteins is 35%; however, they share similar topologies, being composed of eight α-helices, a six-stranded parallel β-sheet and an additional β-strand.

The model obtained for the YP_287866 C-terminal region adopted a similar conformation to proteins belonging to the metal-dependent phosphohydrolase superfamily. These proteins possess a variety of uncharacterized domains associated with nucleotidyltransferases from bacteria, archaea and eukaryotes; YP_287866 also appears to possess one of these domain architectures. The limitation of low sequence identity (~ 25%) between YP_287866 and these proteins was circumvented by the presence of a metal-binding HD motif [32] in YP_287866. Crystal structures of HD-domain proteins have been solved for *Bacillus halodurans* (PDB ID: 2O08) and *Streptococcus agalactiae* (PDB ID: 2OGI); however, a large number of the HD-domain proteins remains uncharacterized [33].

Concerning the C-terminal region of YP_287866 (YP_287866C), the template used was the crystal structure of the putative metal-dependent phosphohydrolase from *S. agalactiae* (PDB ID: 2OGI). The resulting model consisted of an all-alpha structure formed by 13 helices.

YP_287866 is encoded by only one gene; however, it comprises two distinct domains with different functions. The complete model showed both domains linked by a disulfide bond between Cys74 and Cys275 within the N-terminal and C-terminal regions, respectively. This domain architecture was also found in another HD-domain protein fused to a nucleotidyltransferase domain [32]. Because the binding sites in both domains are not spatially superimposed, and the templates form dimers (2H29 and 2OGI), we can conclude that this architecture is likely to exist. Moreover, the model has 97.9% of its residues in preferred and allowed regions of the Ramachandran plot, indicating good stereochemical quality.

As mentioned above, some enzymes of the NAD biosynthetic and recycling pathways have not been identified in *M. hyopneumoniae*. However, based on structural information, we propose that one of the YP_287866 domains is NadD, and we also suggest that YP_287786 functions in this same metabolic pathway, thereby completing the NAD biosynthetic pathway.

The threading programs suggested the crystal structure of nicotinate phosphoribosyltransferase from *Thremoplasma acidophilum* (TmNAPRTase) [34] (PDB ID: 1YTK) as the best hit for the YP_287786 sequence. Further structural analysis suggested another homolog with a solved 3D structure, i.e., NAPRTase (EC 2.4.2.11) from *Enterococcus faecalis* (EfNAPRTase) (PDB ID: 2F7F). This enzyme catalyzes the synthesis of nicotinic acid mononucleotide (NAMN) from adenine and phosphoribosyl pyrophosphate (PRPP), regardless of the presence of ATP.

Although the sequence similarities between YP_287786 and its structural homologs TmNAPRTase and EfNAPRTase showed low overall identity (~ 25%), many residues were found conserved, among which were TmNAPRTase residues Arg224, Asp226, Glu273 and Glu292 involved in NAMN binding [34]. Two other residues also implicated in NAMN binding are found in TmNAPRTase and substituted in YP_287786, i.e., Thr179/Ser166 and Thr293/Val294. The first substitution, between amino acids having a similar physicochemical property, may not affect the function of YP_287786 because NAMN binds TmNAPRTase through a hydroxyl group.

To transfer the phosphoribosyl group, PRPP must bind to NAPRTase. Two conserved motifs, 275hSGGh279 (h stands for hydrophobic residue) and 298GVG301, are responsible for accommodating the phosphate group of PRPP. Both motifs are conserved in YP_287786 except for a glycine residue being replaced by a serine at position 277. The stereochemical quality of the YP_287786 model was verified by the Ramachandran plot calculated using PROCHECK, which showed 97% of the residues in preferred or allowed positions.

Filling gaps in *M. hyopneumoniae* pathways

The biosynthesis of flavin adenine dinucleotide (FAD) in prokaryotes involves bifunctional proteins belonging to the FAD synthetase family that catalyze both riboflavin (RF) phosphorylation and flavin mononucleotide (FMN) adenylylation. In our study, the sequence of YP_287675 showed similarities to the crystal structure of FAD synthetase (TM379) from *T. maritime* [35] (PDB ID: 1S4M) and the *in silico* model of FAD synthetase from *Corynebacterium ammoniagenes* [36] (*Ca*FADS) (PDB ID: 2X0K). Using the comparative genome tool from Genesul, we noticed that FAD synthetase was annotated in other mycoplasma genomes and YP_287675 also belongs to this cluster.

The 3D structure obtained for YP_287675 showed an overall topology similar to its template 1S4M. As expected, these proteins are folded in two domains. The N-terminal domain contains the FMN adenylylation function, catalyzing the reaction between ATP and FMN to form pyrophosphate and FAD (EC 2.7.7.2). Structurally, this domain consists of a typical nucleotide-binding fold (Rossmann fold) containing an ATP-binding site. The motif V/IXGX$_{1\text{-}2}$GXXGXXXG/A associated with the Rossmann fold and FMN binding is present in YP_287675 with a few amino acid substitutions, i.e., VX$_3$GGX$_2$AX$_3$GX$_7$A. This motif was important in assigning biological function to proteins with unknown function from fully sequenced genomes [37]. Moreover, these residues are located in conserved positions allowing substrate binding. Similarly, the residues believed to be involved in ATP-binding are conserved between YP_287675 and its template, except for Glu25 and Phe100 (replaced by aspartate and tyrosine, respectively, in 1S4M and 2X0K).

The second domain of YP_287675, the C-terminal domain, folds into a six-stranded, antiparallel β-barrel architecture, implicated in RF binding. This interaction also involves a long α-helix and a conserved histidine at position 233. RF phosphorylation by *Ca*FADS involves three important residues, Thr208, Asn210 and Asp268 [36]. With respect to sequence, none of these residues are at the same positions in YP_287675; however, the asparagine is maintained at the same structural location. Despite lacking structural information for some regions, the 3D structure of YP_287675 revealed that 96.5% of the residues are in favored and allowed regions.

The understanding of mycoplasma metabolism requires adequate annotation of its proteome. Our structure-based annotation of the proteins YP_287866, YP_287786 involved in NAD biosynthesis and YP_287675 implicated in FAD biosynthesis fills gaps in this annotation. Furthermore, proteins required in these biosynthetic pathways are being considered as antimicrobial drug targets.

Two important proteins implicated in transcription may not be absent from *M. hyopneumoniae*

The hypothetical protein YP_287559 exhibited structural similarities to the prokaryotic transcription factor NusB. NusB participates in the antitermination process, in which RNA polymerase is prevented from reading specific RNA secondary structures that usually terminate transcription. In *E. coli*, antitermination involves at least three Nus proteins: NusB, NusE (identical to the ribosomal protein S10), and NusG [38]. NusB, in association with these other proteins, is believed to bind an RNA motif, *boxA*, present in *E. coli rrn* operons. Mutations in NusB lower growth rate, which is an evidence for its role in rRNA synthesis [39]. *E. coli* has seven *rrn* operons whereas *M. tuberculosis* [40] and *M. hyopneumoniae* have only one such operon. Therefore, an efficient antitermination mechanism is particularly important in these pathogenic bacteria to ensure the expression of the entire single *rrn* operon [41]. Except for NusB, all other proteins required for efficient antitermination, such as NusA, NusG and S10, have been annotated in *M. hyopneumoniae*.

YP_287559 has only 133 residues (of 216) that align with the NusB sequence annotated in other bacterial genomes, including other species of mycoplasma. The remaining sequence (residues 1–82) possesses similarities to a transposase. As no suitable template was found to build the 3D structure of this part of the protein, only its C-terminal region was modeled.

The three dimensional structures of *E. coli* NusB [42] (PDB ID: 1EY1) and *Aquifex aeolicus* NusB [43] (PBD ID:

2JR0) derived from NMR experiments and the crystal structures of NusB from *Thermotoga maritime* [44] (PDB ID: 1TZT), *M. genitalium* [45] (PDB ID: 1Q8C), and *M. tuberculosis* [46] (PDB ID: 1EYV) were used as templates to model YP_287559.

The C-terminal portion of YP_287559 displays a topology composed of only alpha helices. Its structure can be divided into two subdomains, α1-α3 forming the N-terminal region and α4-α7 encompassing the C-terminal subdomain. In the N-terminal region, YP_287559 contains the conserved, positively charged residues Lys83, Arg84, Arg85 and Arg88, forming an arginine-rich motif with a high probability of being the RNA binding site of this protein. Also, interactions between nucleic acid bases and RNA binding proteins often involve aromatic residues essential for stacking [47]. As found in other NusB proteins, the YP_287559 sequence contains the following aromatic residues: Tyr96, Trp98, Phe101, Tyr114, Phe115, Phe127, Tyr132, Phe134, Trp147, Trp149, Phe168, Phe169, Phe176, Phe186, Phe194, Phe196, Tyr207, Tyr208, and Phe214 (Fig. 2). These amino acids located on the surface of the protein are believed to participate in recognition processes, whereas the remaining residues are probably involved in protein fold stabilization.

Previous studies have determined that NusB exists as a homodimer in *M. tuberculosis* (*mtu*NusB) [46], as a monomer in *E. coli* (*eco*NusB) [42], *M. genitalium* (*mge*NusB) [45], and *A. aeolicus* (*aq*NuB) [43], and as a monomer/dimer equilibrium with a preference for the monomeric form [44] in *Thermotoga maritima* (*tma*NusB). We searched the YP_287559 structure for amino acids important for *mtu*NusB dimerization. However, two key residues in mtuNusB, alanine and phenylalanine, are replaced by serine and tyrosine, respectively, in both *M. hyopneumoniae* and *E. coli*. In *mtu*NusB, the dimer interface overlaps the region involved in RNA binding, which may allow *mtu*NusB to remain inactive until needed for transcriptional regulation [46].

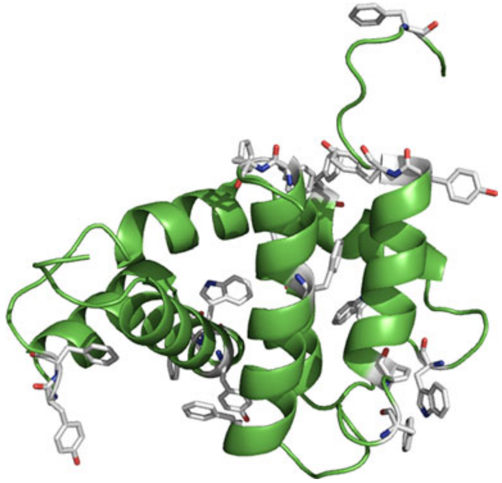


Fig. 2 The 3D structure of YP_287559. Highlighted in green are α-helices and loops; sticks represent aromatic residues likely involved in substrate recognition

We concluded that YP_287559 is composed of two domains, one similar to a transposase and the other to NusB. The Ramachandran plot analysis of the model structure from this last region showed that 95.1% of the residues are in favored and allowed regions.

The *M. hyopneumoniae* habitat is the porcine mucosal surface where amino acids, purines, and pyrimidines are acquired to compensate for the lack of important metabolic pathways. Studies suggested that, in mycoplasmas, genes involved in replication, transcription and translation are constitutively expressed in constant environments, eliminating the need for sophisticated genetic control mechanisms [1]. Moreover, *M. hyopneumoniae* has only one annotated sigma factor, RpoD [3], a key regulator of bacterial transcription initiation that is responsible for promoter recognition and melting [48]. However, the −35 regions of *M. hyopneumoniae* promoters have low sequence conservation, suggesting the presence of more than one sigma factor to respond rapidly to environmental changes.

In our structure-based analysis, we found similarities between the YP_288024 structure and the crystal structures of *Rhodobacter sphaeroides* SigE [49] (PDB ID: 2Z2S) and the flagellar Sigma-28 of *A. aeolicus* [50] (PDB ID: 1RP3). These similarities could indicate that mycoplasmas have a regulatory system not yet identified by traditional tools. Although gene expression in mycoplasma is not well characterized, recent work investigating transcriptional changes has shown that *M. hyopneumoniae* regulates its genes in response to environmental changes [51–54], and 93% of its intergenic regions are transcribed [55].

The sequence alignment of the sigma -70 family revealed the conservation of four regions, divided into subregions. Highly conserved among all members of this family are subregions two and four that compose the sigma factor binding site for the −10 and −35 promoter elements [56]. Conserved only in a highly related sigma factor, subregion one is apparently involved in an antagonistic DNA-binding activity. Subregion three is absent from YP_288024 and from extracytoplasmic function sigma factors that allow bacteria to adapt rapidly to environmental changes. Furthermore, subregion three of extracytoplasmic function sigma factors interacts with the −10 element of promoters lacking a −35 element.

The structural alignment between these proteins showed the complete lack of α-helices four and five and a portion of α-helix six corresponding to the subregion three. All the other α-helices are conserved in YP_288024, suggesting their interaction with the −10 and −35 promoter elements. This functional prediction was based on a model where

96.2% of the residues lie in the most favorable and allowed regions.

High homology to protein with unknown function

The hypothetical protein YP_287971 exhibited structural homology to YlxR from *S. pneumoniae* [57] (PDB ID: 1G2R), a small protein with unknown function, although the YlxR gene is probably in an operon with the other well-studied genes *nus*A, *inf*B, and *rbf*A. The protein encoded by *rbf*A (RbfA) binds to the 30S ribosomal subunit, perhaps promoting subunit maturation [58]. Crucial for translation initiation, IF2 (the product of *inf*B) also functions by binding the 30S subunit [59]. NusA is a highly conserved, essential elongation factor that binds RNA polymerase as part of the transcriptional antitermination complex in many organisms [60]. The YlxR-containing operon has also been studied in *E. coli* and *B. subtilis* [61]. The latter presents two additional genes (Ylx-R and Ylx-Q) between *nus*A and *inf*B; this order was not found in *E. coli* nor in *M. hyopneumoniae* wherein these genes are adjacent.

The 3D structure of YP_287971 showed a similar topology to YlxR of *S. pneumoniae*. Besides a short 3_{10}-helix, no regular secondary structure was found in the N-terminal region. The central core of the model was comprised of three antiparallel β-strands followed by two α-helices, one of which bends at Lys61. The YP_287971 sequence also possesses highly conserved residues, such as the GRGA(Y/W) motif present in the hydrophobic core together with Val10, Leu20, Leu24, Ile32, Ile47, Phe63 and Leu79. At the protein surface several positively charged residues are conserved (Arg6, Arg22, Asp27, Arg43, Lys60, Lys61 and Arg65), forming a patch typical of nucleic acid-binding proteins, as shown in Fig. 3. This region is proposed to be related in YlxR function, which may involve an RNA-binding activity found in proteins encoded by the genes in the *nus*A/*inf*B operon [57].

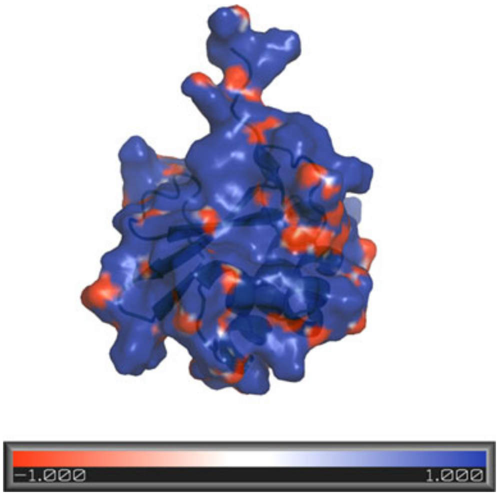


Fig. 3 Probable nucleotide binding site of YP_287971. The electrostatic potential surface distribution shows an extensive positively charged region (blue) typical of nucleic acid-binding proteins

YP_287971 is probably a member of a highly conserved family (DUF448) of unknown function, distributed in many organisms, including 14 species of mycoplasmas for which complete genome sequences are available. The stereochemical quality of YP_287971 was evaluated, resulting in 93.3% of the residues located in favored regions and 6.7% in additional allowed regions of the Ramachandran plot. Because it is of high quality and shows a significant structural resemblance to YlxR of *S. pneumoniae*, the model suggests the same function for YP_287971 and YlxR, and it will aid in the design of future experiments to verify the function.

Finally, the YP_288034 protein showed structural similarities to the crystal structure of YrdC from *E. coli* [62] (PDB ID: 1HRU). Members of the *yrd*C family code for proteins that fold into a single domain, as in the case of 1HRU, or as a domain in proteins implicated in regulation process. YP_288034 is probably an example of the latter because its alignment with *E. coli* YrdC involves only 164 amino acids out of the YP_288034 total of 287 residues. Searching for homologs within mycoplasmas, we observed that this protein clusters with a Sua5-like translation factor found in six other species. Thus, YP_288034 constitutes a two-domain protein containing a YrdC domain as found in *E. coli* and in Sua5 members such as that from *Saccharomyces cerevisiae.*

The function of *E. coli* YrdC is unknown, but its crystal structure suggested that it possesses a double-stranded RNA-binding capacity [62]. The Sua5 protein, containing an YrdC homolog domain in yeast, has been implicated in the re-initiation of translation [63]. This function is consistent with the large concave surface of Sua5; this surface has a positive electrostatic potential akin to that of the YrdC binding surface, which resembles other nucleic acid-binding proteins. The geometry of our model shows 96.6% of the residues in the most favored and additionally allowed regions of the Ramachandran plot.

Conclusions

One of the key challenges in the post-genomic era is the prediction of function for proteins annotated as hypothetical proteins. A combination of bioinformatic tools, focused not only on sequence analysis but also on structural information, guided us to suggest functions for seven hypothetical proteins in the *M. hyopneumoniae* genome. NadD, NAPRTase and FAD synthetase involved in metabolic processes; NusB and SigE in transcription; and for YrdC and YlxR, no conclusive functions were assigned; however, the results obtained helped us design rational experimental strategies for future works. Our results suggest that this structure-based approach provides significant improvements to domain and function prediction, especially for minimal genomes having poorly annotated metabolic pathways. Mycoplasma metabolism requires an adequate annotation of its proteome, and our results fill significant gaps in this annotation. Each target protein used in this work was approached from a unique perspective, taking into account the genomic localization/organization of its open reading frame, its conserved structural features, and any biological evidence available in the literature, even if such evidence was for remote homologs. The annotation of each target required an intense effort. However, our results proved to be important for both structural and biochemical genomics.

Acknowledgments MMF thanks Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for a PhD fellowship.

References

1. Razin S, Yogev D, Naot Y (1998) Molecular biology and pathogenicity of mycoplasmas. Microbiol Mol Biol Rev 62:1094–1156
2. Yus E, Maier T, Michalodimitrakis K, van Noort V, Yamada T, Chen W-H, JaH W, Güell M, Martínez S, Bourgeois R, Kühner S, Raineri E, Letunic I, Kalinina OV, Rode M, Herrmann R, Gutiérrez-Gallego R, Russell RB, Gavin A-C, Bork P, Serrano L (2009) Impact of genome reduction on bacterial metabolism and its regulation. Science 326:1263–1268
3. Vasconcelos ATR, Ferreira HB, Bizarro CV, Bonatto SL, Carvalho MO, Pinto PM, Almeida DF, Almeida LGP, Almeida R, Alves-filho L, Assunc EN, Azevedo VAC, Brigido MM, Brocchi M, Burity HA, Camargo AA, Camargo SS, Carepo MS, Carraro DM, Castro LA, Cavalcanti G, Chemale G, Collevatti RG, Cunha CW, Dallagiovanna B, Dambro BP, Dellagostin OA, Falca C, Fantinatti-garboggini F, Felipe MSS, Fiorentin L, Franco GR, Freitas NSA, Grangeiro TB, Grisard EC, Guimara CT, Hungria M, Krieger MA, Laurino JP, Lima LFA, Lopes MI, Madeira HMF, Manfio GP, Maranha AQ, Martinkovics CT, Moreira MAM, Ramalho-neto CE, Nicola MF, Oliveira SC, Paixa RFC, Pereira M, Pereira-ferrari L, Piffer I, Pinto LS, Potrich DP, Salim ACM, Schmitt R, Schneider MPC, Schrank A, Schrank IS, Schuck AF, Seuanez HN, Silva DW, Silva R, Souza KRL, Souza RC, Staats CC, Steffens MBR, Teixeira SMR, Urmenyi TP, Vainstein MH, Zuccherato LW, Simpson AJG, Zaha A (2005) Swine and poultry pathogens: the complete genome sequences of two strains of. J Bacteriol 187:5568–5577
4. Razin S, Hayflick L (2010) Highlights of mycoplasma research–An historical perspective. Biologicals 38:183–190. doi:10.1016/j.biologicals.2009.11.008
5. Hutchison C, Montague M (2002) Mycoplasmas and the minimal genome concept. In: Herrmann R, Razin S (eds) Molecular Biology and Pathogenicity of Mycoplasmas. Kluwer, New York, pp 221–254
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
7. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL (2008) NCBI BLAST: a better web interface. Nucleic Acids Res 36(suppl 2):W5–W9. doi:10.1093/nar/gkn201
8. Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJP, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36 (Database issue):D419–D425
9. Pal D, Eisenberg D (2005) Inference of protein function from protein structure. Structure 13:121–130
10. Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nat Rev Mol Cell Bio 8:995–1005
11. Erdin S, Ward RM, Venner E, Lichtarge O (2010) Evolutionary trace annotation of protein function in the structural proteome. J Mol Biol 396:1451–1473
12. Pollack DJ (1997) Mycoplasma genes: a case for reflective annotation. Trends Microbiol 5:413–418
13. Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287:797–815. doi:10.1006/jmbi.1999.2583
14. Guo JT, Ellrott K, Chung WJ, Xu D, Passovets S, Xu Y (2004) PROSPECT-PSPP: an automatic computational pipeline for protein structure prediction. Nucleic Acids Res 32(suppl 2): W522–W525. doi:10.1093/nar/gkh414
15. Zdobnov EM, Apweiler R (2001) InterProScan - an integration plataform for the signature-recognition methods in InterPro. Bioinformatics 17:847–848
16. Tatusov RL, Galperin MY, Da N, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36
17. Mazumder R, Vasudevan S (2008) Structure-guided comparative analysis of proteins: principles, tools, and applications for predicting function. PLoS Comput Biol 4:e1000151
18. Katoh K, Misawa K, K-i K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066
19. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A (2002) Comparative protein structure modeling using modeller. Curr Protoc Bioinformatics
20. Laskowski RA (2009) PDBsum new things. Nucleic Acids Res 37 (Database issue):D355–D359
21. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35(suppl 2):W407–W410. doi:10.1093/nar/gkm290
22. Zhang Y (2008) Progress and challenges in protein structure prediction. Curr Opin Struct Biol 18:342–348
23. Bi J, Wang H, Xie J (2011) Comparative genomics of NAD(P) biosynthesis and novel antibiotic drug targets. J Cell Physiol 226:331–340
24. Sorci L, Pan Y, Eyobo Y, Rodionova I, Huang N, Kurnasov O, Zhong S, MacKerell AD, Zhang H, Osterman AL (2009) Targeting NAD biosynthesis in bacterial pathogens: structure-based development of inhibitors of nicotinate mononucleotide adenylyltransferase NadD. Chem Biol 16:849–861
25. Lu S, Smith CD, Yang Z, Pruett PS, Nagy L, McCombs D, Delucas LJ, Brouillette WJ, Brouillette CG (2008) Structure of nicotinic acid mononucleotide adenylyltransferase from *Bacillus anthracis*. Acta Crystallogr F 64(Pt 10):893–898
26. Han S, Forman MD, Loulakis P, Rosner MH, Xie Z, Wang H, Danley DE, Yuan W, Schafer J, Xu Z (2006) Crystal structure of nicotinic acid mononucleotide adenylyltransferase from *Staphyloccocus aureus*: structural basis for NaAD iInteraction in functional dimer. J Mol Biol 360:814–825. doi:10.1016/j.jmb.2006.05.055
27. Kim MK, Kim YS, Rho SH, Im YJ, Lee JH, Kang GB, Eom SH (2003) Crystallization and preliminary X-ray crystallographic

analysis of quinolinate phosphoribosyltransferase of *Helicobacter pylori*. Acta Crystallogr D 59:1265–1266

28. Olland AM, Underwood KW, Czerwinski RM, Lo MC, Aulabaugh A, Bard J, Stahl ML, Somers WS, Sullivan FX, Chopra R (2002) Identification, characterization, and crystal structure of *Bacillus subtilis* nicotinic acid mononucleotide adenylyltransferase. J Biol Chem 277:3698–3707
29. Singh SK, Kurnasov OV, Chen B, Robinson H, Grishin NV, Osterman AL, Zhang H (2002) Crystal structure of *Haemophilus influenzae* NadR protein. A bifunctional enzyme endowed with NMN adenyltransferase and ribosylnicotinimide kinase activities. J Biol Chem 277:33291–33299
30. Yoon HJ, Kim HL, Mikami B, Suh SW (2005) Crystal structure of nicotinic acid mononucleotide adenylyltransferase from Pseudomonas aeruginosa in its Apo and substrate-complexed forms reveals a fully open conformation. J Mol Biol 351:258–265
31. Zhang H, Zhou T, Kurnasov O, Cheek S, Grishin NV, Osterman A (2002) Crystal structures of *Escherichia coli* nicotinate mononucleotide adenylyltransferase and its complex with deamido-NAD. Structure 10:69–79
32. Aravind L, Koonin EV (1998) The HD domain defines a new superfamily of metal-dependent phosphohydrolases. Trends Biochem Sci 23:469–472
33. Zimmerman MD, Proudfoot M, Yakunin A, Minor W (2008) Structural insight into the mechanism of substrate specificity and catalytic activity of an HD-domain phosphohydrolase: the 5′-deoxyribonucleotidase YfbR from *Escherichia coli*. J Mol Biol 378:215–226
34. Shin DH, Oganesyan N, Jancarik J, Yokota H, Kim R, Kim S-H (2005) Crystal structure of a nicotinate phosphoribosyltransferase from *Thermoplasma acidophilum*. J Biol Chem 280:18326–18335
35. Wang W, Kim R, Yokota H, Kim SH (2005) Crystal structure of flavin binding to FAD synthetase of *Thermotoga maritima*. Proteins 58:246–248
36. Frago S, Martínez-Júlvez M, Serrano A, Medina M (2008) Structural analysis of FAD synthetase from *Corynebacterium ammoniagenes*. BMC Microbiol 8:160–175
37. Kleiger G, Eisenberg D (2002) GXXXG and GXXXA motifs stabilize FAD and NAD(P)-binding Rossmann folds through Cα–HO hydrogen bonds and van der Waals interactions. J Mol Biol 323:69–76
38. Zellars M, Squires CL (1999) Antiterminator-dependent modulation of transcription elongation rates by NusB and NusG. Mol Microbiol 32:1296–1304
39. Quan S, Zhang N, French S, Squires CL (2005) Transcriptional polarity in rRNA operons of *Escherichia coli* nusA and nusB mutant strains. J Bacteriol 187:1632–1638
40. Verma A, Sampla AK, Tyagi JS (1999) *Mycobacterium tuberculosis* rrn promoters: differential usage and growth rate-dependent control. J Bacteriol 181:4326–4333
41. Arnvig KB, Zeng S, Quan S, Papageorge A, Zhang N, Villapakkam AC, Squires CL (2008) Evolutionary comparison of ribosomal operon antitermination function. J Bacteriol 190:7251–7257
42. Altieri AS, Mazzulla MJ, Horita DA, Heath Coats R, Wingfield PT, Das A, Court DL, Andrew Byrd R (2000) The structure of the transcriptional antiterminator NusB from *Escherichia coli*. Nat Struct Biol 7:470–474
43. Das R, Loss S, Li J, Waugh DS, Tarasov S, Wingfield PT, Byrd RA, Altieri AS (2008) Structural biophysics of the NusB:NusE antitermination complex. J Mol Biol 376:705–720
44. Bonin I, Robelek R, Benecke H, Urlaub H, Bacher A, Richter G, Wahl MC (2004) Crystal structures of the antitermination factor NusB from Thermotoga maritima and implications for RNA binding. Biochem J 383:419–428
45. Liu J, Yokota H, Kim R, Kim SH (2004) A conserved hypothetical protein from *Mycoplasma genitalium* shows structural homology to Nusb proteins. Proteins 55:1082–1086
46. Gopal B, Haire LF, Cox RA, Colston MJ, Major S, Brannigan JA, Smerdon SJ, Dodson G (2000) The crystal structure of NusB from *Mycoplasma tuberculosis*. Nature 7:475–478
47. Oubridge C, Ito N, Evans PR, Teo CH, Nagai K (1994) Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. Nature 372:432–438
48. Koo B-M, Rhodius VA, Nonaka G, deHaseth PL, Gross CA (2009) Reduced capacity of alternative σs to melt promoters ensures stringent promoter recognition. Genes Dev 23:2426–2436
49. Ea C, Greenwell R, Anthony JR, Wang S, Lim L, Das K, Sofia HJ, Donohue TJ, Sa D (2007) A conserved structural module regulates transcriptional responses to diverse stress signals in bacteria. Molecular Cell 27:793–805
50. Brown PN, Mathews MA, Joss LA, Hill CP, Blair DF (2005) Crystal structure of the flagellar rotor protein FliN from *Thermotoga maritima*. J Bacteriol 187:2890–2902
51. Madsen ML, Nettleton D, Thacker EL, Minion FC (2006) Transcriptional profiling of *Mycoplasma hyopneumoniae* during iron depletion using microarrays. Microbiology (Reading, England) 152(Pt 4):937–944
52. Oneal MJ, Schafer ER, Madsen ML, Minion FC (2008) Global transcriptional analysis of *Mycoplasma hyopneumoniae* following exposure to norepinephrine. Microbiology (Reading, England) 154(Pt 9):2581–2588
53. Hwang MH, Damte D, Lee JS, Gebru E, Chang ZQ, Cheng H, Jung BY, Rhee MH, Park SC (2011) *Mycoplasma hyopneumoniae* induces pro-inflammatory cytokine and nitric oxide production through NFκB and MAPK pathways in RAW264.7 cells. Veterinary Res Commun 35:21–34
54. Schafer ER, Oneal MJ, Madsen ML, Minion FC (2007) Global transcriptional analysis of *Mycoplasma hyopneumoniae* following exposure to hydrogen peroxide. Microbiology (Reading, England) 153(Pt 11):3785–3790
55. Gardner SW, Minion FC (2010) Detection and quantification of intergenic transcription in *Mycoplasma hyopneumoniae*. Microbiology 156(Pt 8):2305–2315
56. Campbell EA, Muzzin O, Chlenov M, Sun JL, Olson CA, Weinman O, Trester-Zedlitz ML, Darst SA (2002) Structure of the bacterial RNA polymerase promoter specificity sigma subunit. Molecular Cell 9:527–539
57. Osipiuk J, Górnicki P, Maj L, Dementieva I, Laskowski R, Joachimiak A (2001) {\it Streptococcus pneumonia} YlxR at 1.35{Å} shows a putative new fold. Acta Crystallograph Sec D 57:1747–1751
58. Goto S, Kato S, Kimura T, Muto A, Himeno H (2011) RsgA releases RbfA from 30S ribosome during a late stage of ribosome biosynthesis. EMBO J 30:104–114
59. Caserta E, Tomsic J, Spurio R, Anna PCL, Gualerzi CO (2006) Translation initiation factor IF2 interacts with the 30 S ribosomal subunit via two separate binding sites. J Mol Biol 362:787–799
60. Yang X, Lewis PJ (2010) The interaction between bacterial transcription factors and RNA polymerase during the transition from initiation to elongation. Transcription 1:66–69
61. Shazand K, Tucker J, Stansmore K, Leighton T (1993) Similar organization of the nusA-infB operon in *Bacillus subtilis* and *Escherichia coli*. J Bacteriol 175:2880–2887
62. Teplova M, Tereshko V, Sanishvili R, Joachimiak A, Bushueva T, Anderson WF, Egli M (2000) The structure of the yrdC gene product from *Escherichia coli* reveals a new fold and suggests a role in RNA binding. Protein Sci 9:2557–2566
63. Na J, Pinto I, Hampsey M (1992) Isolation and characterization of SUA5, a novel gene required for normal growth in Saccharomyces cerevisiae. Genetics 131:791–801

J Mol Model (2012) 18:1927–1934
DOI 10.1007/s00894-011-1216-z

ORIGINAL PAPER

Theoretical study of BN_4: potential precursors of high energy density materials (HEDMs)

Li Ping Cheng · Yu Qi Xu · Gen Li Wang · Hui Hong He

Received: 16 June 2011 / Accepted: 9 August 2011 / Published online: 26 August 2011

Abstract Ab initio (MP2) and density functional theory (DFT) methods were used to examine nine isomers of the doublet BN_4 species with the 6-311 + G(d) basis set. To our knowledge, these nine structures are all first reported here. Energy analysis indicates that the C_{2v} branched structure is the global minimum of potential energy surface. Research results show that the C_{2v} branched, the *cis*-linear, the C_{4v} pyramidal, and the C_S five-membered ring structures are likely to be stable and to be observed experimentally. Among these four kinetically stable species, the last three are suitable to be used as potential precursors of HEDMs due to their high dissociation energies. However, the C_{2v} bent, the *trans*-linear, the D_2 bicyclic, the C_{2v} four-membered ring, and the C_{2v} cage structures are kinetically unstable due to their low dissociation or isomerization barriers. Two synthesis pathways of the C_{2v} branched isomer were located. It seems more feasible to synthesize this species by linear NBN and N_2.

Keywords Ab initio · BN_4 · Boron nitrides · HEDMs · Potential energy surface

Introduction

Polynitrogen compounds have received considerable attention for more than 20 years due to their potential use as prime candidates for “green” high energy density materials (HEDMs). However, generally, in the search for polynitrogen candidates for HEDMs, one faces an apparent dilemma, i.e., the instability of these species; the low dissociation barrier in particular appears to be a major hindrance for these molecules to be useful in a wider variety of applications. Nevertheless, the recent experimental progress in the synthesis of nitrogen-rich compounds has been very encouraging with the N_5^+ and N_5^- ions produced [1, 2] in the laboratory. More recently, several salts containing the CN_7^- anion were prepared by deprotonation of 5-azido-1*H*-tetrazole using common bases like ammonia, hydrazine, or alkali as well as alkaline earth metal salts [3]. Experimental successes have stimulated theoretical studies on other potential nitrogen-rich compounds. To search for states/structures that may be suitable candidates as precursors of HEDMs, Lee et al.[4] have studied various low-lying, high- and low-spin electronic states of Al_2N_4 and AlN_n (n=4 to 7) clusters. They concluded that AlN_n systems are potential precursors of HEDMs. The accurate enthalpies of formation of gas-phase N_3, N_3^-, N_5^+ and N_5^- have been calculated by ab initio molecular orbital theory. The calculations show that neither $N_5^+N_3^-$ nor $N_5^+N_5^-$salt could be stabilized and this conclusion had also been experimentally confirmed by low-temperature metathetical reactions between N_5SbF_6 and alkali metal azides in different solvents [5].

In recent years, the group III nitrides have been studied extensively due to their high-energies, distinctive properties as precursors for bulk semiconductors and high-power applications. The wurtzite polytypes of gallium nitride (GaN), aluminum nitride (AlN), and indium nitride (InN) are excellent materials for bandgap engineering [6]. Al_xN_y clusters have received considerable attention from computational chemists because of the importance of aluminum nitride (solid and thin-film AlN) in various industrial applications [7]. Boron nitride has attracted considerable

L. P. Cheng (✉) · Y. Q. Xu · G. L. Wang · H. H. He
School of Chemical and Environmental Engineering,
Shanghai Institute of Technology,
Shanghai 200235, People’s Republic of China
e-mail: chengliping@sit.edu.cn

interest in materials science. Boron nitride and carbon, being isoelectronic, tend to form similar compounds or materials. The sphalerite-type (or β-crystalline phase) of boron nitride, isostructural and isoelectronic to cubic diamond, displays excellent physicochemical properties (e.g., mechanical hardness, excellent thermal and chemical stability, and conductivity) [8, 9]. As it has proven to be very difficult to obtain pure β-BN as a solid film, mixed and larger B_xN_y clusters have attracted much attention as precursors in the growth of β-BN thin films using chemical vapor deposition or plasma techniques [8]. In addition, the existence of boron-nitrogen clusters with fullerene geometries was postulated. The envisaged boron nitride fullerenes have been synthesized [10].

Information on the geometry and electronic structure of boron-nitrogen clusters is essential for their applications. Several theoretical studies on some small boron-nitrogen clusters such as BN, BN_2, B_2N_2, B_3N, BN_3, B_3N_2 and B_2N_3 had been previously performed [11–14]. These studies mainly focus on cluster's structures, vibrational, thermochemical, dissociation, and spectroscopic properties. Furthermore, some molecules, such as the B_2N and BN_2, have been characterized experimentally in pulsed laser evaporation experiments combined with matrix infrared spectroscopy. [11–15]. The structure and vibrations of B_nN_n (n=3–10) species have been studied with the density functional theory (DFT) method. The results show that the B_nN_n (n=3–10) clusters have D_{nh} cumulenic monocyclic structures with θ_{NBN} the largest and θ_{BNB} the sharpest angle [16]. The structural, rotational, and vibrational properties of $B_nN_n^+$ (n=3–10) have also been investigated using DFT method [17]. It has been found that the $B_nN_n^+$ clusters display different behaviors depending on whether they have an even (n=4, 6, 8, 10) or odd (n=3, 5, 7, 9) number of BN pairs. The $B_nN_n^+$ cations with n even display a D_{nh} symmetry as the neutral systems; while the $B_nN_n^+$ clusters with n odd show a lowered symmetry compared to the neutral form.

In the present report, we extended the study of boron-nitrogen clusters to BN_4 system. One aim of the present study is to explore the geometries and electronic structures of BN_4 and determinate the global minimum; another is to explore whether these species are suitable candidates as precursors of HEDMs. To our knowledge, no theoretical study has been devoted to the structures, especially kinetic stabilities of the BN_4 species.

Computational methods

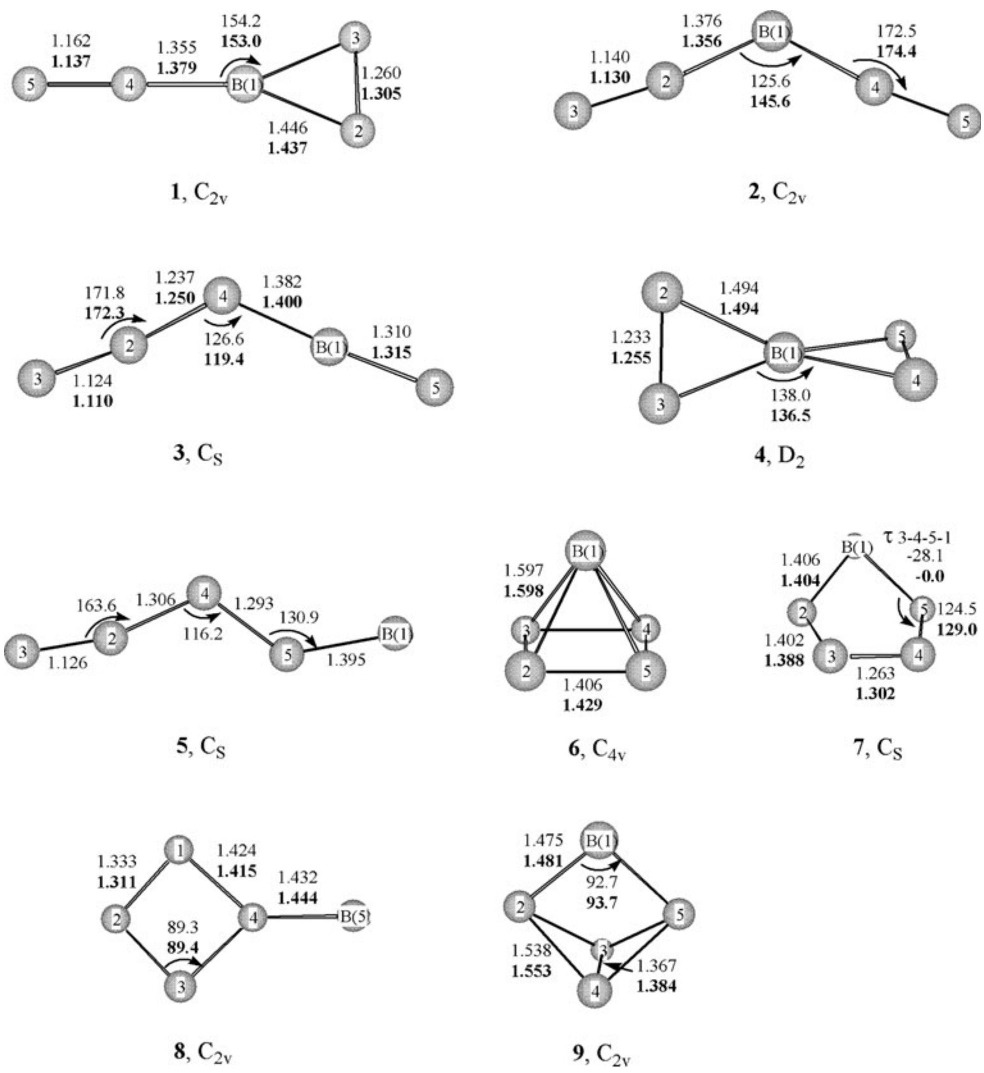
All calculations were performed using the Gaussian 03 program package [18]. We initially optimized geometries and calculated the harmonic vibrational frequencies for BN_4 at the B3LYP/6-311+G* level of theory, where B3LYP is the DFT method using Becke's three-parameter gradient-corrected functional [19] with the gradient-corrected correlation of Lee, Yang, and Parr [20] and 6-311+G* is the split-valence triple-ζ plus polarization basis set augmented with diffuse functions [21]. Then, the geometries were refined and the vibrational frequencies were calculated at the level of second-order Møller-Plesset perturbation theory (MP2) [22] with the 6-311+G* basis set. Stationary points were characterized as minima without any imaginary vibrational frequency and a first-order saddle point with only one imaginary vibrational frequency. For transition states, the minimum energy pathways connecting the reactants and products were confirmed using the intrinsic reaction coordinate (IRC) method with the Gonzalez-Schlegel second-order algorithm [23, 24]. Final energies were refined at the CCSD (T) [25]/6-311+G*//B3LYP/6-311+G*+ZPE (B3LYP/6-311+G*) level of theory.

Throughout this paper, bond lengths are given in Ångströms, bond angles in degrees, total energies in Hartrees, relative and zero-point vibrational energies, unless otherwise stated, in kJ mol^{-1}.

Results and discussion

Our optimized structures for nine BN_4 species are illustrated in Fig. 1. Their total energies, ZPE, relative energies (with ZPE corrections), and number of imaginary frequencies are listed in Table 1. Note that the symbol "τ" on the figures stands for the dihedral angle. As is seen, these BN_4 isomers except the C_S chain **5** are all local minima on their potential energy surfaces (PES) at the above-mentioned two levels. Regarding structure **5**, it is a local minimum at the B3LYP level of theory, but the optimization directly lead to dissociation into linear BN_2 and one N_2 molecule at the MP2 level of theory. The optimized structures for nine transition states are shown in Fig. 2. Their total energies, ZPE, and lowest vibrational frequencies are listed in Table 2. The energy differences between the minima and their corresponding transition states are tabulated in Table 3. The reaction energies for dissociation of the BN_4 isomers to B + 2 N_2 and linear BN_2 + N_2 molecules are shown in Table 4 and Table 5.

We performed ab initio calculations on a wide variety of doublet structures of BN_4 by using two different and sophisticated theoretical methods. As exhibited in Fig. 1, nine structures were located. To our knowledge, these nine structures are all first reported here. As seen from Table 1, according to our calculation, the energetic stability ordering of the nine isomers is **1** > **2** > **3** > **4** > **5** > **6** > **7** > **8** > **9**. It should be noted that the C_{2v} branched **1** is



1, C_{2v} **2**, C_{2v} **3**, C_S **4**, D_2 **5**, C_S **6**, C_{4v} **7**, C_S **8**, C_{2v} **9**, C_{2v}

Fig. 1 Optimized geometries for nine BN_4 species at the B3LYP/6-311+G* and MP2/6-311+G* (bold font) levels of theory

energetically higher than the C_{2v} bent **2** by 13.4 kJ mol^{-1} at the B3LYP/6-311 + G* level of theory, but it is energetically lower than **2** by 71.1 and 6.2 kJ mol^{-1} at the MP2 and CCSD(T) levels, respectively. Results from CCSD(T) level are generally most reliable among the three used levels. Accordingly, isomer **1** should be

Table 1 Total energies (E), zero-point energies (ZPE), and relative energies (RE) for BN_4 species

Species	B3LYP/6-311+G*			MP2/6-311+G*			CCSD(T)/6-311+G* //B3LYP/6-311+G*	
	E[a]	ZPE[b]	RE[c]	E[a]	ZPE[b]	RE[c]	E[a]	RE[c]
1(C_{2v})	−243.82970	46.5 (0)	0.0	−243.16818	57.8 (0)	0.0	−243.22645	0.0
2(C_{2v})	−243.83554	48.6 (0)	−13.4	−243.16517	121.0 (0)	71.1	−243.22490	6.2
3(C_S)	−243.80040	50.2(0)	80.8	−243.14848	67.8(0)	62.0	−243.19311	91.3
4(D_2)	−243.79135	39.8 (0)	94.2	−243.14665	229.4 (0)	228.1	−243.19590	73.7
5(C_S)	−243.69537	42.4 (0)	348.7	-	-	-	−243.09530	340.3
6(C_{4v})	−243.65365	47.3(0)	463.4	−243.03829	47.7(0)	331.1	−243.06969	412.7
7(C_S)	−243.65305	44.8(0)	462.1	−243.03129	51.5(0)	353.3	−243.06472	423.2
8(C_{2v})	−243.62951	36.8 (0)	516.1	−242.95075	40.2 (0)	553.4	−243.04293	472.6
9(C_{2v})	−243.62784	43.5(0)	527.4	−243.00701	44.8(0)	410.2	−243.04531	473.0

[a]Total energies in Hartree. [b] Zero-point energies in kJ mol^{-1}. The integers in parentheses are the number of imaginary frequencies (NIMAG). [c] The relative energies with ZPE corrections in kJ mol^{-1}

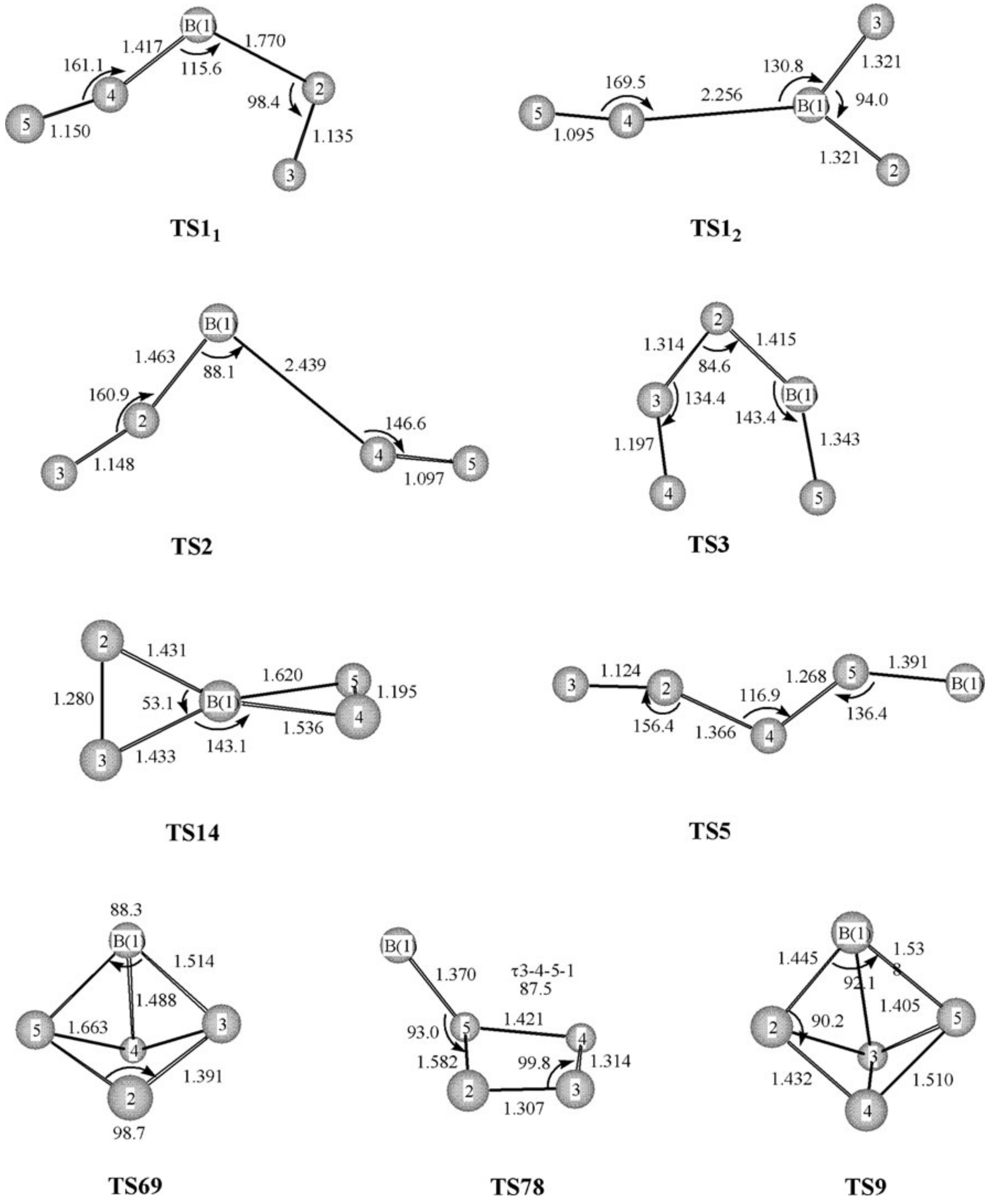


Fig. 2 Optimized geometries for nine BN_4 transition states at the B3LYP/6-311+G* level

regarded as the most energetically favored for all BN_4 species considered here.

As shown in Fig. 1, the C_{2v} branched **1** (N_2BNN) seems to be a complex formed by a BN_2 ring and a N_2 molecule. The bond length of N4-N5, 1.137-1.162 Å, is indeed close to the experimental N≡N triple-bond length 1.098 Å for the nitrogen molecule N_2 [26], but the bond length of N2-N3, 1.260-1.305 Å, is closer to the double-bond length 1.252 Å of HN=NH [26]. We predicted that the lengthening in the N4-N5 bond from N_2 molecule may be attributed to the induced polarization from the boron atom. The D_2 bicyclic **4** seems to be a complex between the fragments a boron atom and two equivalent dinitrogen molecules, but the bond lengths of N2-N3 and N4-N5, 1.233-1.255 Å, are closer to the N=N double-bond length. Like the case in **1**, the B atom induces the lengthening in the N-N bond lengths from

Table 2 Total energies (E) and zero-point energies (ZPE) for the BN$_4$ transition states

Species	B3LYP/6-311+G*		CCSD(T)/6-311+G *//B3LYP/6-311+G*
	E	ZPE	E
TS1$_1$(C$_1$)	−243.77925	37.7(284*i*)[d]	−243.18516
TS1$_2$(C$_S$)	−243.75670	36.8(124*i*)	−243.17250
TS14(C$_1$)	−243.79091	41.9(*254i*)	−243.19954
TS2(C$_S$)	−243.79560	36.8 (89*i*)	−243.19761
TS3(C$_1$)	−243.68425	42.3(537*i*)	−243.08323
TS5(C$_S$)	−243.69502	38.9 (387*i*)	−243.09502
TS69(C$_S$)	−243.59958	40.6(550*i*)	−243.01477
TS78(C$_1$)	−243.62086	39.3(283*i*)	−243.02879
TS9(C$_1$)	−243.61847	42.3(347*i*)	−243.03719

d The values in parentheses are the lowest vibrational frequencies [υl(cm-1)]

Table 3 Energy differences (kJ mol^{-1}) of transition states relative to BN$_4$ isomers (Including ZPE corrections at the B3LYP/6-311+G* level of theory

Species	B3LYP/6-311+G*	CCSD(T)/6-311+G* //B3LYP/6-311+G*
1(C$_{2v}$)	0.0	0.0
TS1$_1$(C$_1$)	123.9	99.6
TS1$_2$(C$_S$)	182.1	132.3
TS14(C$_1$)	97.1	66.1
2(C$_{2v}$)	0.0	0.0
TS2(C$_S$)	93.3	59.9
3(C$_S$)	0.0	0.0
TS3(C$_1$)	297.2	280.9
4(D$_2$)	0.0	0.0
TS14(C$_1$)	3.3	−7.5
5(C$_S$)	0.0	0.0
TS5(C$_S$)	−2.5	−2.5
6(C$_{4v}$)	0.0	0.0
TS69(C$_S$)	135.2	137.7
7(C$_S$)	0.0	0.0
TS78(C$_1$)	79.1	88.7
8(C$_{2v}$)	0.0	0.0
TS78(C$_1$)	25.1	39.8
9(C$_{2v}$)	0.0	0.0
TS69(C$_S$)	71.1	77.4
TS9(C$_1$)	23.4	20.1

molecular nitrogen. The covalent radius for nitrogen is 0.70 Å [27], the corresponding value for B is 0.88 Å [27]. Obviously, in structure **1**, the B1-N2 (N3) bond distances (1.437-1.446 Å) and B1-N4 bond distances (1.355-1.379 Å) are all slightly shorter than the sum of covalent radii of the corresponding B atom and nitrogen atom. In structure **4**, the B-N bond distances (1.494 Å) are also slightly shorter than the sum of covalent radii of the corresponding B atom and nitrogen atom. To study the kinetic stabilities of these two isomers, their dissociation and isomerization reactions have been investigated. The schematic potential energy surfaces (PESs) for isomers **1** and **4** are depicted in Fig. 3. Structure **TS1$_1$** (seen in Fig. 2) is a transition state (TS) of the dissociation of **1** characterized to be a saddle point of index 1 by vibrational frequency analysis. IRC calculations performed at the B3LYP/6-311+G* level directly lead to dissociation into a linear BN$_2$ and one N$_2$ molecule. The barrier for the decomposition reaction **1** → **TS1$_1$** → BN$_2$ (linear) + N$_2$ is predicted to be 99.6 kJ mol^{-1} at the CCSD(T) level of theory, indicating the high kinetic stability toward decomposition. **TS1$_2$** is another transition state (TS) of the dissociation of **1**. IRC calculations performed at the B3LYP level directly lead to dissociation into linear NBN atom and one N$_2$ molecule. The barrier for the decomposition reaction **1** → **TS1$_2$** → NBN + N$_2$ is predicted to be 132.3 kJ mol^{-1} at the CCSD(T) level of theory, indicating the high kinetic stability toward decomposition. In addition, Fig. 3 shows that the energies of linear BN$_2$ + N$_2$ and linear NBN + N$_2$ molecules are all higher than **1**, and **TS1$_1$**, **TS1$_2$** are virtually transition structures of synthesis isomer **1**. The synthesis energy barrier heights via **TS1$_1$**, **TS1$_2$** corrected by ZPE were predicted to be only 86.2 and 35.6 kJ mol^{-1} at the CCSD(T) level, respectively. Therefore, the experimental synthesis of **1** via **TS1$_1$**, **TS1$_2$** seems possible theoretically. Furthermore, it seems more feasible to synthesize **1** by linear NBN and N$_2$ via **TS1$_2$**. The possible isomerization from **1** to **4** was also studied. The two conformers interconvert through a transition structure **TS14** (seen in Fig. 3). Conformer **1** converts to **4** with a barrier of 66.1 kJ mol^{-1}, and conformer **4** converts to **1** with a barrier of only −7.5 kJ mol^{-1}. Therefore, conformer **4** is not likely to be stable, and if it is formed in any process, it will transform into the C_{2v} branched **1**.

The C_{2v} bent **2**, the *cis*-C_S linear **3** and the *trans*-C_S linear **5** are three chain structures. As tabulated in Table 1, they lie above **1** by 6.2, 91.3 and 340.3 kJ mol^{-1} at the CCSD(T) level of theory, respectively. As shown in Fig. 1, the two terminal N-N bonds (1.130-1.140 Å) in structure **2** are close to the N≡N triple bond. The bond length of N2-N3 (1.110-1.124 Å) in structure **3** is also close to that of N≡N triple-bond, but the bond length of N2-N4 (1.237-1.250 Å) is closer to that of N=N double-bond. The bond length of N2-N3 (1.126 Å) in structure **5** is also close to that of N≡N triple-bond, but the bond lengths of N2-N4 (1.306 Å) and N4-N5 (1.293 Å) are closer to that of N=N double-bond. Structure **2** is

Table 4 Reaction energies (kJ mol^{-1}) for dissociation of the BN_4 isomers to B + 2 N_2

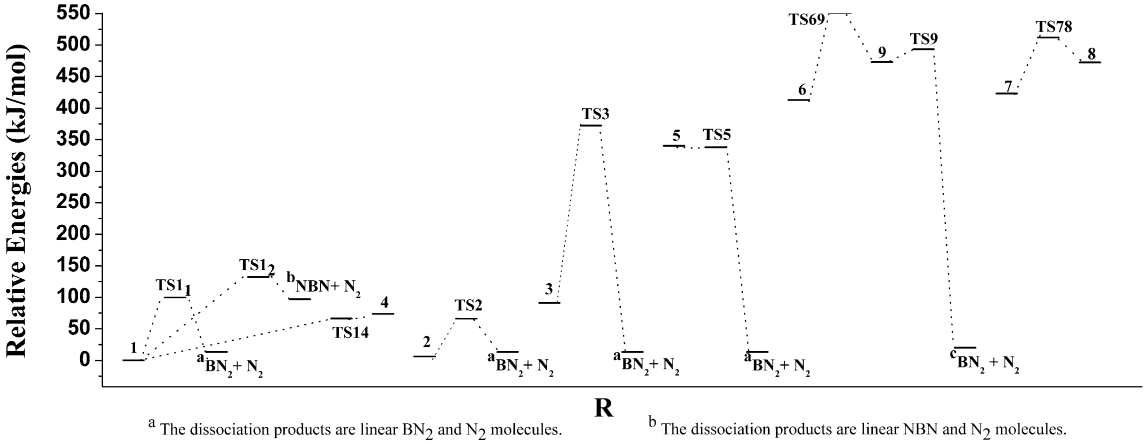
Species	B3LYP/6-311+G*	MP2/6-311+G*	CCSD(T)/6-311+G* //B3LYP/6-311+G*
1(C_{2v})	−108.4	44.0	4.2
2(C_{2v})	−121.8	115.1	10.3
3(C_S)	−27.6	105.9	95.4
4(D_2)	−14.7	272.1	77.9
5(C_S)	240.3	-	344.5
6(C_{4v})	354.6	375.1	416.9
7(C_S)	353.7	397.3	427.4
8(C_{2v})	407.7	597.8	476.8
9(C_{2v})	418.6	454.6	477.2

energetically lower than **3** and **5** probably due to having more N≡N triple-bonds. The B-N distances in **2**, **3** and **5** are all slightly shorter than the sum of covalent radii of the corresponding B atom and nitrogen atom. To further analyze their kinetic stabilities, we have investigated their decomposition pathways. The dissociation of **2** proceeds in a straightforward manner with simple bond fission. The transition state **TS2** (C_S) was located on the PES. As shown in Fig. 2, we can note that, compared with structure **2**, the bond length of B1-N4 in the transition state is stretched to eliminate one N_2 molecule whereas that of N4-N5 is actually compressed. The barrier for dissociation is 93.3 kJ mol^{-1} at the B3LYP level but only 59.9 kJ mol^{-1} at the CCSD(T) level, indicating that it is not very stable toward decomposition. Structure **TS3** (C_S) is a dissociation transition structure of **3**. IRC calculation performed at the B3LYP/6-311+G* level directly leads to dissociation into a linear BN_2 and one N_2 molecule. The barriers for the decomposition reaction **3** → **TS3** → BN_2 + N_2 are predicted to be 297.2 and 280.9 kJ mol^{-1} at the B3LYP and CCSD(T) levels of theory, respectively. The high dissociation barriers suggest that species **3** is highly stable kinetically. Similarly, structure **TS5** (C_S) is a dissociation transition structure of **5**. IRC calculation performed at the B3LYP/6-311+G* level directly leads to dissociation into linear BN_2 and one N_2 molecule. The barriers for the decomposition reaction **5** → **TS5** → BN_2 + N_2 are even negative (−2.5 kJ mol^{-1}) at the B3LYP and CCSD(T) levels. Such low barriers imply that structure **5** is highly unstable toward decomposition.

The C_{4v} pyramidal **6**, the C_S five-membered ring **7**, the C_{2v} four-membered ring **8**, and the C_{2v} cage **9** are all high-energy species. They are higher in energy than the most stable **1** by 412.7, 423.2, 472.6, and 473.0 kJ mol^{-1} at the CCSD(T) level of theory, respectively. As shown in Fig. 1, in structure **6**, the N-N bond distances are all close to that of N-N single-bond (1.449 Å) [26] and the B-N distances are slightly longer than the sum of covalent radii of the corresponding B atom and nitrogen atom, but in structure **7**, the case is different. The bond distances between nitrogen and nitrogen are all between that of N-N single-bond and N=N double-bond, and the B-N distances are slightly shorter than the sum of covalent radii of the corresponding B atom and nitrogen atom. In structure **3**,

Table 5 Reaction energies (kJ mol^{-1}) for dissociation of the BN_4 isomers to BN_2 (linear) + N_2

Species	B3LYP/6-311+G*	MP2/6-311+G*	CCSD(T)/6-311+G* //B3LYP/6-311+G*
1(C_{2v})	−77.4	−15.9	−13.4
2(C_{2v})	−90.4	55.3	−7.1
3(C_S)	3.3	46.0	77.9
4(D_2)	16.7	212.2	60.3
5(C_S)	271.3	-	326.9
6(C_{4v})	385.9	315.2	399.3
7(C_S)	384.7	337.4	409.8
8(C_{2v})	438.7	537.5	459.2
9(C_{2v})	450.0	394.3	459.6



[a] The dissociation products are linear BN_2 and N_2 molecules.

[b] The dissociation products are linear NBN and N_2 molecules.

[c] The dissociation products are cyclic BN_2 and N_2 molecules.

Fig. 3 Schematic potential energy surfaces for doublet BN_4 isomers

the bond lengths of N1-N4 and N2-N3, 1.415-1.424 Å, are all close to that of N-N single-bond, whereas the bond lengths of N1-N2 and N3-N4, 1.311-1.333 Å, are all between that of N-N single-bond and N=N double-bond. The B-N distance is also slightly shorter than the sum of covalent radii of the corresponding B atom and nitrogen atom. Structure **9** is interesting. Its lowest frequency, 234 and 203 cm^{-1} at the B3LYP and MP2 levels, respectively, is high enough to prove the minimum. As shown in Fig. 1, the N-N bond distances in this structure are all either close to an N-N single bond or slightly longer, which is consistent with those in the tetrahedral N_4 (where all N-N linkages are single bonds). In structure **9**, the B1-N3 (N4) distances (2.010-2.011 Å) are far longer than the sum of covalent radii of the corresponding B atom and nitrogen atom. While the B1-N2(N5) distances (1.475-1.481 Å) are slightly shorter than the sum of covalent radii of the corresponding B atom and nitrogen atom. The schematic potential energy surfaces for isomers **6**, **7**, **8** and **9** are also depicted in Fig. 3, indeed, on the basis of B3LYP geometries, two transition structures (**TS69** and **TS78**) have been located connecting **6** and **9**, **7** and **8** on the PES, respectively. The barrier going from **6** to **9** is 137.7 kJ mol^{-1} and from **9** to **6** is 77.4 kJ mol^{-1} at the CCSD(T)/6-311+G*// B3LYP/6-311+G*+ZPE (B3LYP/6-311+G*) level of theory. The corresponding barrier from **7** to **8** is 88.7 kJ mol^{-1} and from **8** to **7** is only 39.8 kJ mol^{-1}. Therefore, the conversion reactions of **6** to **9**, and **7** to **8** are difficult to occur and structures **6**, **7** are stable on kinetics, but species **9** and **8** are not likely to be stable and if they are formed in any process, they will transform into **6** and **7**, respectively. In addition, we have located one dissociation transition structure for isomer **9**. The barriers for the reaction **9** → **TS9** → cyclic $BN_2 + N_2$ are predicted to be 23.4 and 20.1 kJ mol^{-1} at the B3LYP and CCSD(T) levels of theory, respectively. The low dissociation barriers also suggest that species **9** is highly unstable kinetically.

The reaction energies for dissociation of the BN_4 isomers to B + 2 N_2 molecules are listed in Table 4. On the other hand, kinetic analysis indicates that it seems very difficult for BN_4 isomers to dissociation into one B atom and two N_2 molecules, while the dissociation products are generally a linear BN_2 and one N_2 molecules. Accordingly, the reaction energies for dissociation of the BN_4 isomers to linear BN_2 + N_2 molecules are also listed (see Table 5). Table 4 and Table 5 show that most dissociation reactions are exothermic. Furthermore, structures **3**, **6** and **7** have high dissociation energies as well as significant dissociation or isomerization barriers and therefore should be regarded as suitable precursors of high energy density materials.

Summary

We have examined nine nitrogen-rich BN_4 compounds in the present study. Among them, the C_{2v} branched structure is the global minimum. Kinetic analysis shows that the C_{2v} branched, the *cis*-linear, the C_{4v} pyramidal, and the C_S five-membered ring structures are all likely to be stable and to be observed experimentally. However, the C_{2v} bent structure and the *trans*-linear are kinetically unstable due to their low dissociation barriers. The D_2 bicyclic, the C_{2v} four-membered ring, and the C_{2v} cage structures are also kinetically unstable, and if they are formed in any process, they will transform into other structures. Therefore, among the nine BN_4 isomers, the *cis*-linear, the C_{4v} pyramidal , and the C_S five-membered ring structures may be possibly used as precursors of HEDMs because of their high dissociation energies and significant stabilities. Two potential synthesis pathways of the C_{2v} branched isomer were located. It seems more feasible to synthesize this species by linear NBN and N_2.

Acknowledgments This work was financially supported by the Shanghai Education Committee (11YZ224) and Shanghai Leading Academic Discipline Project (J51503).

References

1. Christe KO, Wilson WW, Sheehy JA, Boatz JA (1999) N_5^+: A novel homoleptic polynitrogen ion as a high energy density material. Angew Chem Int Edn Engl 38:2004–2009. doi:10.1002/(SICI)1521-3773(19990712)38:13/14<2004
2. Vij A, Pavlovich JG, Wilson WW, Vij V, Christe KO (2002) Experimental Detection of the Pentaazacyclopentadienide (Pentazolate) Anion, *cyclo*-N_5^-. Angew Chem Int Edn 41:3051–3054. doi:10.1002/1521-3773(20020816)41:16<3051
3. Klapötke TM, Stierstorfer J (2009) The CN_7^- anion. J Am Chem Soc 131:1122–1134. doi:10.1021/ja8077522
4. Lee EPF, Dyke JM, Claridge RP (2002) Ab initio calculations on Al_2N_4 and AlN_n (n =4 to 7): potential precursors of high energy density materials. J Phys Chem A 106:8680–8695. doi:10.1021/jp021059q
5. Dixon DA, Feller D, Christe KO et al. (2004) Enthalpies of f of gas-phase N_3, N_3^-, N_5^+, and N_5^- from Ab initio molecular orbital theory, stability predictions for $N_5^+N_3^-$ and N_5 + N_5^-, and Experimental Evidence for the instability of $N_5^+N^{3-}$. J Am Chem Soc 126:834–843. doi:10.1021/ja0303182
6. Neumayer DA, Ekerdt JG (1996) Growth of group III nitrides. a review of precursors and techniques. Chem Mater 8:9–25. doi:10.1021/cm950108r
7. Chang C, Patzer ABC, Sedlmayr E, Steinke T, Sülzle D (2001) A density functional study of small $(AlN)_x$ clusters: structures, energies, and frequencies. Chem Phys 271:283–292. doi:10.1016/S0301-0104(01)00439-6
8. Seyferth D (1984) Gmelin Handbook of Inorganic Chemistry. 8th edn. B. Boron Compounds. Organometallics 3:139. doi:10.1021/om00080a901
9. Demazeau G (1995) High pressure diamond and cubic boron nitride synthesis. Diamond Relat Mater 4:284–287. doi:10.1016/0925-9635(94)05281-6
10. Golberg D, Bando Y, Stéphan O, Kurashima K (1998) Octahedral boron nitride fullerenes formed by electron beam irradiation. Appl Phys Lett 73:2441–2443. doi:10.1063/1.122475
11. Andrews L, Hassanzadeh P, Burkholder TR, Martin JLM (1993) Reactions of pulsed laser produced boron and nitrogen atoms in a condensing argon stream. J Chem Phys 98:922–931. doi:10.1063/1.464256
12. Martin JML, Slanina Z, Francüois JP, Gijbels R (1994) The structure, energetics, and harmonic vibrations of B_3N and BN_3. Mol Phys 82:155–164. doi:10.1080/00268979400100114
13. Martin JML, Taylor PR, Francüois JP, Gijbels R (1994) Ab initio Study of the spectroscopy, kinetics, and thermochemistry of the BN_2 molecule. Chem Phys Lett 222:517–523. doi:10.1016/0009-2614(94)00378-5
14. Peterson KA (1995) Accurate multireference configuration interaction calculations on the lowest $^1\Sigma^+$ and $^3\Pi$ electronic states of C_2, CN^+, BN, and BO^+. J Chem Phys 102:262–277. doi:10.1063/1.469399
15. Hassanzadeh P, Andrews L (1992) Pulsed laser-assisted reactions of boron and nitrogen atoms in a condensing nitrogen stream. J Phys Chem 96:9177–9182. doi:10.1021/j100202a020
16. Martin JML, El-Yazal J, Francois JP (1996) Structure and vibrations of B_nN_n (n=3–10). Chem Phys Lett 248:95–101. doi:10.1016/0009-2614(95)01302-4
17. Giuffreda MG, Deleuze MS, Francois JP (2000) Structural, rotational, and vibrational properties of mixed ionized boron−nitrogen clusters $B_nN_n^+$ (n=3–10). J Phys Chem A 104:5855–5860. doi:10.1021/jp994450t
18. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2003) Gaussian 03, Revision B.04. Gaussian Inc, Pittsburgh, PA
19. Becke AD (1993) A new mixing of Hartree–Fock and local density–functional theories. J Chem Phys 98:1372–1377. doi:10.1063/1.464304
20. Lee C, Yang W, Parr RG (1988) Development of the colle-salvetti correlation-energy formula into a functional of the electron density. Phys Rev B 37:785–789. doi:10.1103/PhysRevB.37.785
21. Hehre WJ, Radom L, PvR S, Pople JA (1986) Ab initio molecular orbital theory. Wiley, New York
22. Møller C, Plesset MS (1934) Note on an approximation treatment for many-electron systems. Phys Rev 46:618–622. doi:10.1103/PhysRev.46.618
23. Gonzalez C, Schlegel HB (1989) An improved algorithm for reaction path following. J Chem Phys 90:2154–2161. doi:10.1063/1.456010
24. Gonzalez C, Schlegel HB (1990) Reaction path following in mass-weighted internal coordinates. J Phys Chem 94:5523–5527. doi:10.1021/j100377a021
25. Raghavachari K, Trucks GW, Pople JA, Head-Gordon M (1989) A fifth-order perturbation comparison of electron correlation theories. Chem Phys Lett 157:479–483. doi:10.1016/S0009-2614(89)87395-6
26. Lide DR (1992) CRC Handbook of Chemistry and Physics, 73rd edn. CRC Press, Boca Raton, FL
27. http://chemed.chem.purdue.edu/genchem/topicreview/bp/ch7/size.html#cov

J Mol Model (2012) 18:1935–1949
DOI 10.1007/s00894-011-1220-3

ORIGINAL PAPER

Computational studies on full-length Ku70 with DNA duplexes: base interactions and a helical path

Shaowen Hu · Francis A. Cucinotta

Received: 23 November 2010 / Accepted: 9 August 2011 / Published online: 26 August 2011

Abstract The Ku70/80 heterodimer is among the first responding proteins to recognize and bind the DNA double strand breaks (DSBs). Once Ku is loaded at the DSB, it works as a scaffold to recruit other repair factors in non-homologous end joining thereby facilitates the following repair processes. In this work, we characterized the detailed interactions and binding free energies between a Ku70 subunit and several DNA duplexes, by using some well-established computational methods. The results reveal that the structure of the protein may suffer certain contractions without the company of Ku80, and may experience large conformational changes in the presence of different DNA duplexes. Notably, we observe the closest interactions between Ku70 and DNA can be easily strengthened to form H-bonds with the bases in the minor groove, which is unexpected. However, this finding is supported by the presence of a similar bond between Ku80 and DNA in the published crystal structure (PDB code 1JEY). We suggest that these interactions are responsible for the observed pausing sites when Ku translocates along DNA and the subtle difference in binding with AT- and GC-rich DNA ends. Additionally, simulations indicate the inner surface of the ring encircling the DNA is not flat, but contains a delicate clamp like structure, which is ideal to grip the two strands of DNA in the minor groove and confine the movement of the duplex in a unique helical path.

Keywords Binding free energy · DNA repair · Ku70 · Ku-DNA binding · Molecular dynamics simulation

Introduction

Non-homologous end joining (NHEJ) is the primary DNA double strand break (DSB) repair pathway in multi-cellular eukaryotes [1, 2]. The first step in NHEJ is the specific recognition and tethering of the DNA ends at the site of the lesion. This is carried out by Ku, a heterodimer protein consisting of the two highly related subunits Ku70 and Ku80 [3]. Once bound at the DNA ends, Ku works as a scaffold protein to recruit other repair factors that are required in NHEJ, which in mammalian cells, include DNA-PKcs (DNA-dependent protein kinase catalytic subunit), Artemis, polymerase μ and λ, and a complex of XLF (Cernunnos), XRCC4, and DNA ligase IV, etc. [4]. These proteins act together in a highly coordinated way to cleave the incompatible section, fill the gap, and ligate the strands of DNA [1]. It was reported recently that the recruitment of these enzymes is not necessarily in the exact order of nuclease-polymerase-ligation, but can have a wide range of flexibility disregarding the exact structure of the DNA broken ends [5]. This observation underscores the key mediation role that Ku plays in NHEJ pathway.

The crystallographic structure of human Ku heterodimer reveals that, despite a low level of sequence identity (~15%), the two subunits share a common core structure consisting of an N-terminal α/β domain, a central β-barrel domain, and a helical C-terminal arm, which together form a pseudo-

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1220-3) contains supplementary material, which is available to authorized users.

S. Hu (✉)
Division of Space Life Sciences,
Universities Space Research Association,
Houston, TX 77058, USA
e-mail: Shaowen.Hu-1@nasa.gov

F. A. Cucinotta
NASA, Lyndon B. Johnson Space Center,
2101 NASA Parkway,
Houston, TX 77058, USA

symmetrical ring-like channel that is just large enough to encircle a DNA duplex [6]. This finding explains many previous works attempting to identify the interfaces of Ku subunits in dimerization, DNA binding and repair components interaction [7]. However, there are certain segments of each subunit that are undetermined presumably due to experimental difficulty. The missing parts in Ku70 include residues 1–34, 224–230, and 539–558 [6]. Several studies have indicated these residues are important in DNA-binding and protein-protein interactions [8–11]. In Ku80, the most prominent missing part is the C-terminal region, which extends from residue 543 to 732 and is about 40% of the total sequence of this 86 kDa polypeptide [6]. Two separate experiments have identified the isolated structure of this region, which is characterized by a hex-helical globular domain flanked with a long unstructured loop at the N-terminus and a short loop at the extreme C-terminus [12, 13]. The extreme C-terminal loop was found to be responsible for the association of Ku with DNA-PKcs [14]. Though structural modeling of the full length Ku70/80 aided with advanced detecting technologies were reported recently [15, 16], due to their low resolutions and the apparently discrepant results, the structural knowledge of this domain in Ku heterodimer as well as its precise functions in the context of Ku protein-DNA and protein-protein interactions has been limited.

In addition to DNA repair, several investigations suggest that Ku70/80 play important roles in a number of other fundamental cellular processes such as telomere maintenance, transcription, and apoptosis [1]. Since no computational modeling studies have been reported for this system, it would be of general interest to utilize the well-established computational tools to examine the detailed interactions of the various domains of Ku with DNA ends. Such studies could consider the conformational and functional implications of these domains in the monomer and heterodimer forms as well as in the presence of DNA and other NHEJ factors. In this study, as a first step of this effort, we singled out the Ku70 subunit from the crystal structures of Ku heterodimer and its complex with DNA, and applied classical molecular dynamics (MD) simulations and binding affinity analysis on the full-length Ku70 monomer and several Ku70-DNA complexes. Our goals are: (i) mapping out the flexible versus rigid regions of Ku70 and monitoring the conformational alterations without the support of Ku80 and in the presence of different DNA duplexes; (ii) quantifying the energetic contributions of different domains of Ku70 in the binding affinity of this prominent protein-DNA complex; and (iii) examining the processes of Ku70-DNA association and dissociation in atomic details and checking the possible forces that drive the translocation of Ku70 along the DNA duplex.

This paper is divided into three major parts plus Methods and Conclusions. We first identify the differences between the structures with and without the support of Ku80, and trace the time evolution of key interactions between the domains of Ku70 and DNA duplexes along multi-nanosecond trajectories. Evidences from this analysis indicate that the structural integrity of Ku70/80 heterodimer is important if not essential for its role in the initial recognition and binding of DNA DSBs. Second, we calculate the binding affinity of several Ku70-DNA complexes and decompose them into each domain of Ku70, with the help of the well-established MM-GBSA method. This energetic analysis helps to shed light on the functional implications of each domain of Ku70 in its interactions with DNA and with other repair factors. Third, we present results of two targeted molecular dynamics simulations that mimic the loading and unloading processes of DNA duplex with respect to the channel of Ku70; these simulations, along with the illuminated uneven inner surface of the channel, support the proposed concept that the movement of Ku along DNA duplex is constrained to a unique helical path [6].

Methods

Modeling

Coordinates of the Ku70 and Ku70-DNA complex were extracted from crystal structures with PDB codes 1JEQ and 1JEY [6], as the initial structures for nine simulation systems to build up (Table S1 of Supporting information). Missing residues 1–34, 224–230, and 539–558 were added by using Swiss-PdbViewer [17] (Fig. 1a). For the two complexes with Ku70 located at the internal site of DNA (Ku70AT2L and Ku70AT2R), the loop 1–34 was initially arranged at the left side and right side of the duplex, respectively, when viewed from the N-terminus to C-terminus of Ku70 (Fig. 1b). These two models were built in an effort to examine the possible effect of this loop to the translocation of Ku70 along a DNA duplex.

After the three-way junction was removed, the crystal structure of Ku70-DNA complex contains only a 14-bp DNA duplex, which is of the minimal length of oligonucleotides that Ku-DNA association could be detected in vitro [18, 19]. To investigate the interaction of Ku70 with a longer DNA helix, two 10-bp DNA duplexes were added to the ends of 14-bp DNA in Ku70-DNA complex in different ways (Fig. 1c). One duplex (5′-GCGTTAACGC-3′) was obtained with PDB code 1CQO [20]. Another duplex was constructed by the nucgen facility of the AMBER10 software package [21], with sequence 5′-GCGCGCGCGC-3′. All DNA duplexes are blunt-ended and in the canonical B-form.

Relaxation of the loops

To help the model of missing residues find their optimal positions in the monomer and in the presence of different

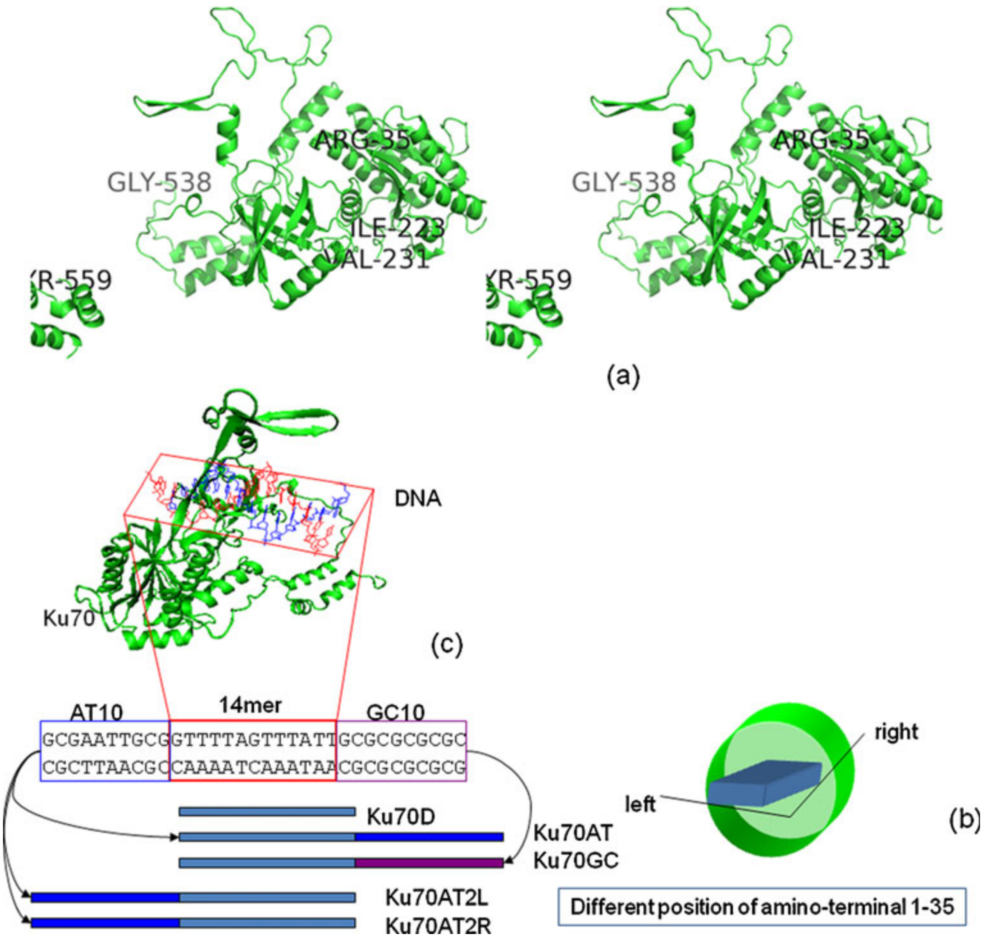


Fig. 1 Modeling scheme of full length Ku70 and its complexes with DNA. (**a**) The stereo structure of the crystal structure of Ku70 (PDB code: 1JEQ). The residues at the ends of three missing loops are labeled. (**b**) The schematic diagram of Ku70AT2L and Ku70AT2R, viewed from N-terminal side of Ku70 to C-terminal side. (**c**) Five Ku70-DNA complexes investigated in this study

DNA duplexes, an extensive relaxation protocol was applied to each system before putting them into explicit water boxes. This includes 10,000 steps of constrained energy minimization and five rounds of 100 ps simulated annealing. During this stage, the coordinates of the experimentally determined residues of Ku70 as well as all nucleotides were restrained with a force constant of 2.0 kcal (mol Å)$^{-1}$, while the modeled residues of Ku70 were kept free. An implicit solvent model of Hawkins-Cramer-Truhlar [22] was used to represent the electrostatics of aqueous solution. GB/SA methodology (igb=1, cut=16.0 Å, rgbmax=12.0 Å, and surften=0.005 kcal (mol·Å^2)$^{-1}$) was used with the salt concentration set to 0.2 M, to represent the solvent and ionic effects, respectively. Each simulated annealing step was performed by heating the system from 0 K to 600 K for 10 ps using a heat bath time coupling constant of 0.2 ps, maintaining the system at 600 K for another 15 ps. During the remaining 75 ps, the system's heat bath was softened with time coupling constant to 4.0 ps and slowly tightened to 0.05 ps, and the temperature was simultaneously reduced from 600 K to 0 K. These and the following MD simulations were performed using the software package AMBER10 [21], with all-hydrogen ff99SB protein [23] and ff99bsc0 nucleic acid [24] force fields.

MD simulations

The final structures of relaxation were placed into periodic boxes of TIP3P water molecules, with counter ions to neutralize the total charge. The distances between the edges of the water box and the closest atom of the solute were at least 10 Å in all cases (Table S1 of Supporting information). The particle mesh Ewald method was used to treat the long-range electrostatic interactions, and bond lengths involving bonds to hydrogen atoms were constrained using SHAKE. The time-step for all MD simulations was 2.0 fs, with a direct-space, non-bonded cutoff of 9.0 Å. Translational center-of-mass motions were removed every 1000 steps. The systems were minimized by 500 steps of EM with restraints of 2.0 kcal (mol·Å^2)$^{-1}$, followed by 35 ps canonical ensemble (NVT)-MD. Then five rounds of 600 steps of minimization were conducted to reduce the solute restraints gradually; 2.0 kcal (mol·Å^2)$^{-1}$ restraints were again used while heating the entire system to 300 K. Then, with a time constant of 2.0 ps for heat-bath coupling, solute restraints were reduced gradually over 50 ps, and the systems underwent isothermal isobaric ensemble (NPT)-MD simulations to adjust the solvent density. After the equilibration phase, the production phase without any constraint was followed at 300 K and 1 atm for 20 ns. Structural figures were generated using VMD [25] and PyMOL [26].

Targeted MD simulations

To examine the dynamical process of DNA loading to and unloading from Ku70, the 14-bp DNA duplex in Ku70D was manually translated to a place close to the C-terminal loop, where the center of the DNA duplex is about 60 Å away from the fully loaded position, with all atomic RMSD 66.9 Å from the crystal structure. Two 1-ns targeted MD (TMD) simulations were conducted, for loading and unloading, respectively. The bound structure was set as the target and the unbound Ku70-DNA system as the starting system for loading process, and vice versa for the unloading process. The bound and unbound systems were placed into periodic boxes of TIP3P water molecules, with counter ions to neutralize the total charge, and buffer size were adjusted to ensure the numbers of water molecules of the two systems are the same. As a result, the least distance between the edges of the water box and the closest atom of the solute was 6.82 Å for unbound system and 13.80 Å for bound system. The two initial systems went through the same minimization, heating, and equilibration steps as the above MD procedure. The time steps for TMD simulations were all set to be 1.0 fs. During the TMD steps, the backbone atoms of residues 35–250 of Ku70 (i.e., the α/β domain) were fixed in space with constraint 2.0 kcal (mol·Å^2)$^{-1}$, while all atoms in the 14-bp DNA duplex were driven by a force constant of 0.01 kcal (mol·Å^2)$^{-1}$, to adjust the RMSD from 66.9 Å to around 0.0 Å. In both systems the direction of the dragging force is approximately along the helical axis of the DNA duplex. The magnitude of force was chosen via a trial and error procedure. Larger or smaller forces were found to render either very rugged trajectories or much deviated ending structures.

Binding energy calculation

The MM-GBSA approach to calculate the binding free energy of complexes formation A + B → AB usually uses the following thermodynamic cycle [27]:

$$\begin{tikzcd}
A_{aqu} \arrow[d, "-\Delta G^{A}_{solv}"] & + B_{aqu} \arrow[r, "\Delta G_{binding}"] \arrow[d, "-\Delta G^{A}_{solv}"] & AB_{aqu} \arrow[d, "-\Delta G^{A}_{solv}"] \\
A_{gas} & + B_{gas} \arrow[r, "\Delta G_{gas}"] & AB_{gas}
\end{tikzcd} \tag{1}$$

$$\Delta \mathrm{G}_{\mathrm{bind}} = \Delta \mathrm{G}_{\mathrm{gas}} - \Delta \mathrm{G}^{A}_{\mathrm{solv}} - \Delta \mathrm{G}^{\mathrm{B}}_{\mathrm{solv}} + \Delta \mathrm{G}^{\mathrm{AB}}_{\mathrm{solve}}$$

$$= \Delta H_{gas} - T\Delta S - \Delta G^{A}_{GBSA} - \Delta G^{B}_{GBSA} + \Delta G^{AB}_{GBSA}$$

$$= \Delta H_{gas} - T\Delta S + \Delta\Delta G_{GB} + \Delta\Delta G_{SA}$$

$$\Delta H_{gas} \approx \Delta E_{gas} = \Delta E_{\mathrm{intra}} + \Delta E_{ele} + \Delta E_{vdW} \tag{2}$$

$$\Delta\Delta G_{GB} = \Delta G^{AB}_{GB} - (\Delta G^{A}_{GB} + \Delta G^{B}_{GB}) \tag{3}$$

$$\Delta\Delta G_{SA} = \Delta G^{AB}_{SA} - (\Delta G^{A}_{SA} + \Delta G^{B}_{SA}). \tag{4}$$

In Eq. 1 $\Delta \mathrm{G}_{\mathrm{gas}}$ is the interaction energy between A and B in the gas phase. The enthalpy part ($\Delta E_{\mathrm{intra}} + \Delta E_{ele} + \Delta E_{vdW}$) was approximated by the Sander module of AMBER 10 using an infinite cutoff for nonbonded interactions. As we applied single trajectory approach in which the protein and DNA structures were taken from the complex simulation, the $\Delta E_{\mathrm{intra}}$ term in all systems is consistently zero. $\Delta \mathrm{G}^{A}_{\mathrm{solv}}$, $\Delta \mathrm{G}^{\mathrm{B}}_{\mathrm{solv}}$, $\Delta \mathrm{G}^{\mathrm{AB}}_{\mathrm{solve}}$ are the solvation free energies of A, B and AB, which include the electrostatic and non-polar component [27]. The electrostatic portion was calculated using the GB/SA method with 0.1 M salt [22]. The non-polar portion of the solvation energy (i.e., due to cavity formation and hydrophobicity) was calculated using the Still equation, $G_{SA} = \gamma SA$, where γ is an empirical atomic solvation parameter, 7.2 cal Å^{-2}, and SA is the solvent accessible surface area calculated with a solvent probe radius of 1.4 Å. The entropic part of the energy, TΔS, is calculated through normal-mode analysis with the NAB program of Ambertools [28]. Each solute structure was minimized using the conjugate-gradient method to an RMS gradient of 10^{-4} kcal (molÅ)$^{-1}$ with a constant dielectric 1.0. Only five snapshots of each trajectory were considered, due to the high demanding of such calculations [21].

In order to understand the contributions of the various domains of Ku70 to the binding affinity of complex, differences in the energy terms upon formation of the complexes were calculated. This typically involved calculation of the difference between a selected term for the complex and those of the individual molecules comprising the complex.

Results and discussion

Structural contraction and conformational changes

To examine in detail the biologically critical functional interactions of Ku70 with DNA broken ends, we carried out extensive molecular dynamics simulations with the Ku70 monomer as well as several Ku70-DNA complexes. The simulations reveal significant contraction of the overall structure of Ku70, particularly the ring structure and the C-terminal arm, and large conformational shifts exemplified by the C-terminal SAP domain. Figure 2 shows the structures of snapshots taken from the end of the simu-

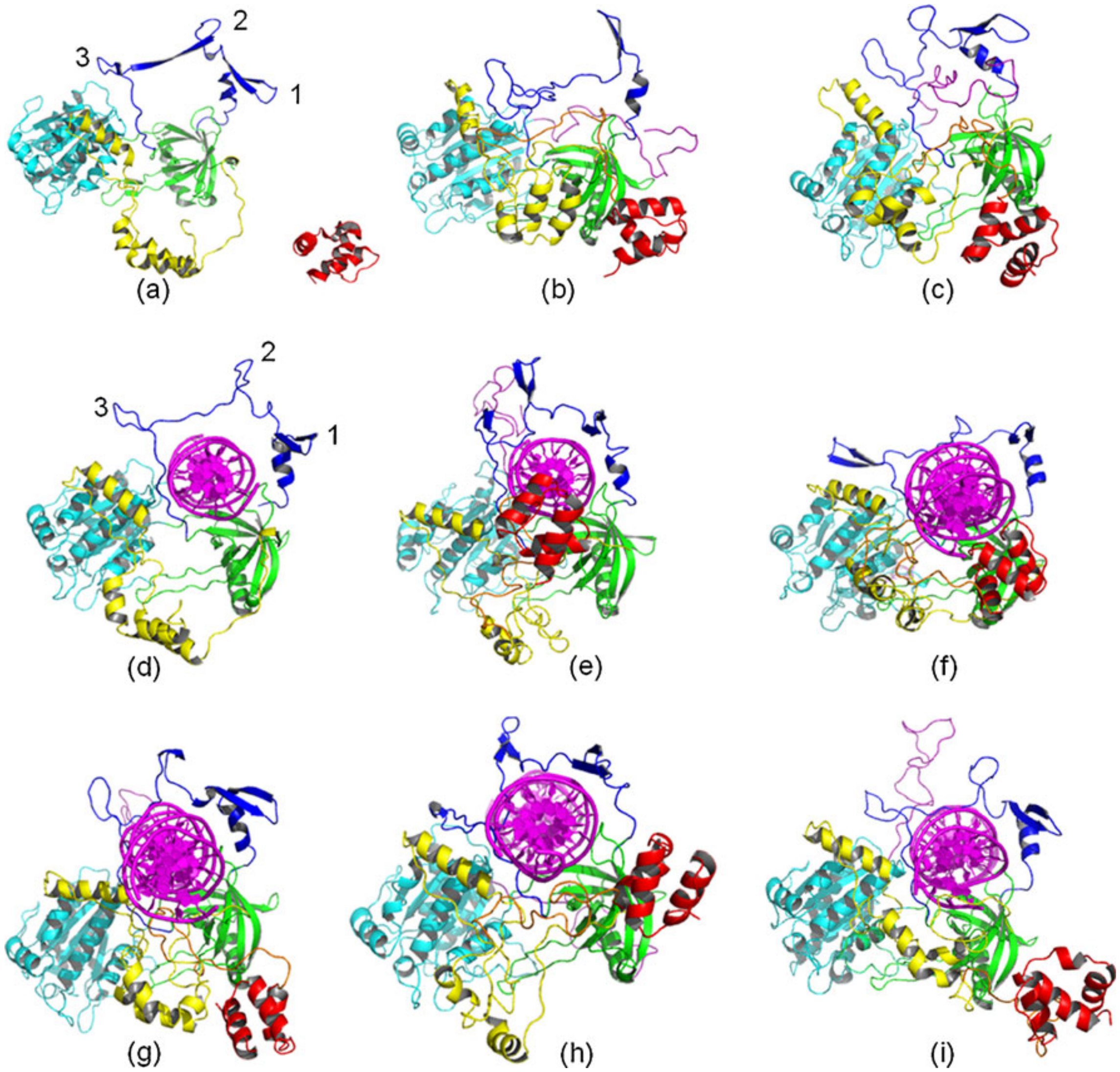


Fig. 2 Structural contraction and conformational shift of Ku70 without the support of Ku80. (**a**) Experimental structure of Ku70; (**b**) snapshot of Ku70 with free N-loop at 20 ns; (**c**) snapshot of Ku70 with trapped N-loop at 20 ns; (**d**) experimental structure of Ku70D; snapshot at 20 ns of (**e**) Ku70D, (**f**) Ku70AT, (**g**) Ku70GC, (**h**) Ku70AT2L, and (**i**) Ku70AT2R. The color scheme of the cartoons is as follows: SAP domain (560–609), red; C-terminal loop (536–559), orange; C-terminal arm (440–535), yellow; β-barrel (251–276, 343–439), green; Bridge (277–342), blue; α/β domain (36–250), cyan; N-terminal loop (1–35), purple; DNA, magenta. All structures are viewed from the C-terminus to N-terminus of Ku70, with DNA helix in the complexes approximately perpendicular to the plane of the ring. Hairpin 1, 2, 3 discussed in the text are labeled in the crystal structures (**a**) and (**d**)

lations, with comparison of the crystal structures. Though this is not surprising as a result of losing the support from Ku80, questions are raised as to how in some experiments Ku70 could bind DNA in the absence of Ku80 [29], and how the collapsed ring and arm of Ku70 could be expanded so as to form heterodimer with Ku80. In the following, we describe the major structural features of these systems and detailed interactions of the key residues of Ku70 with DNA during simulation, and discuss the possible mechanisms of Ku70-DNA binding and Ku70/80 heterodimerization.

Table 1 shows the average root-mean-squared-deviations (RMSD) of the various domains of Ku70 as well as DNA during the last 5 ns MD simulations, as compared with the crystal structures. In all seven systems, the structures of the α/β domain, the β-barrel, and the SAP domain (residues in each domain are denoted in the caption of Fig. 2) are very stable, while the bridge and C-terminal arm are rather flexible. The extensions of the 14-bp DNA to 24-bp in all four cases have increased the flexibility of the duplexes, reflecting the stronger interplaying between DNA and the solvent in these systems. If comparisons of the last 5 ns trajectories with the snapshots at 15 ns are made, much smaller deviations of the bridge, the C-terminal arm, and the DNA duplexes can be observed (Supporting informa-

Table 1 The average RMSDs of the backbone heavy atoms of distinct regions in Ku70 and Ku70-DNA complexes with respect to their crystal structures, determined for snapshots over the last 5 ns of simulation. For residues in each domain refer to the legend of Fig. 2. The RMSDs are reported in Å, with standard deviation followed. In each complex system only the 14-bp DNA the same as in the crystal structure is considered

System	α/β domain	β-barrel domain	Bridge	C-terminal arm	SAP domain	DNA
Ku70(trapped)	1.2±0.1	1.3±0.1	6.3±0.4	7.0±0.3	1.0±0.1	
Ku70(free)	1.3±0.1	1.8±0.2	12.1±0.4	10.2±0.5	1.5±0.2	
Ku70D	1.8±0.3	1.5±0.2	7.0±0.3	5.9±0.4	1.3±0.2	2.9±0.2
Ku70AT	1.6±0.1	1.4±0.1	6.7±0.5	7.2±0.5	0.9±0.1	6.5±0.6
Ku70GC	1.3±0.1	1.4±0.1	5.4±0.5	6.6±0.7	1.1±0.2	4.4±0.6
Ku70AT2L	2.3±0.1	1.4±0.1	5.8±0.6	7.4±0.2	1.1±0.2	6.2±0.7
Ku70AT2R	1.5±0.1	1.4±0.1	4.1±0.4	9.2±0.3	1.0±0.1	6.8±0.7

tion Table S2). This indicates the overall structures of Ku70 and DNA are well equilibrated and maintained in the later phase of simulations. Except in the Ku70 monomer with trapped N-terminal loop (discussed below), the modeled missing N-loop is as flexible as the bridge and C-terminal arm. Unexpectedly, the C-terminal loop is relatively stable in four of seven systems (Supporting information Table S2), indicating the existence of some persistent contact of this region with other parts of the systems. In addition, we found the missing loop 224–230 shares the same stability as the α/β domain. It appears that the weak electron density experimentally detected in these regions is not directly associated with their flexibility [6].

The observed contraction of the overall structure of Ku70 is caused by the collapse of the ring structure and the C-terminal arm (Fig. 2). In all systems, while the three hairpins which sustain the most flexible bridge part of the ring are well maintained, their relative positions are significantly shifted. In fact, the collapse/contraction of the bridge is consistently caused by the positional shift of hairpin 2 in all systems, as hairpin 1 is supported by its short helix neighbor and hairpin 3 buttressed by the α/β domain and hence both are relatively stable (Fig. 2). For the C-terminal arm, we observe its contraction is correlated with the position of the SAP domain and the extension of the C-terminal loop. This stems from the easily formed electrostatic association between the loop 440–455 of the arm and the proximate C-terminal loop and the SAP domain, as there are alternative positively and negatively charged patches of residues along these two mobile regions. Such interactions are noted in six systems, except for Ku70AT2R in which the SAP domain diffuses far away from the core of the system (Fig. 2). After dimerization with Ku80, the C-terminal arm of Ku70 is known to function as a holder for the β-barrel domain of Ku80, and the SAP domain is positioned against the base of the α/β domain of Ku80 [6]. Therefore the diffusion and translocation of the SAP domain can presumably help to open the collapsed arm.

Diverse results were obtained for the locations as well as the interaction modes of the N-terminal loop and the SAP domain in different systems, as in simulations both regions display large conformational change from their initial positions (Fig. 2). We find, even in monomer simulations, the N-terminal loop 1–34 can have two different locations. It can either stay freely at the N-terminal side, or extend most of itself into the ring (Fig. 2b and c). The attraction is obviously due to the overall negatively charges this loop has (12 out of 35 residues are aspartate or glutamate), a feature similar to DNA. In the three complexes with DNA duplexes ending at the N-terminus of Ku70 (Ku70D, Ku70AT, and Ku70GC), this loop clearly shows an obstructive effect on the further translocation of DNA. This is accomplished by the stacking residues 27–35 of the loop with the end of DNA. This pose is consistent with a previous proposal for its function in directing the binding orientation of Ku with DNA [6]. In Ku70AT2L and Ku70AT2R where such a barrier is overcome, there is evidence of interactions between the basic residues of this loop and DNA backbones. The right side binding appears to be more favored as more residues are involved in this mode than in the left side binding mode.

The SAP domain was proposed to be a putative DNA binding domain [8, 30]. In all complex systems, the SAP domain is initially located far from the DNA (about 30 Å in Ku70AT and Ku70GC, and 37 Å in the other three). In Ku70D and Ku70AT, we found this domain migrates in solution and forms close contact with DNA toward the end of the simulation (Fig. 2e and f). In Ku70D, the interactions involve the ending bases and the basic residues at loop La which links the first and second α helices of this domain [10]. In Ku70AT, the interface of contact is composed of residues along La and Lb of SAP and the minor groove of DNA. These results are consistent with previous studies of this region, such as the chemical shift perturbation experiment [10] and the chemical modification and mass spectrometry experiment [11]. However, the binding affinity of this domain with DNA may not be strong enough to overcome all the hurdles along the path, which is demonstrated in three other complexes (Fig. 2g, h, and i). Though started from similar positions as in Ku70D and Ku70AT, the SAP domain fails to form contacts with DNA in these systems. It is possible that the interactions of the C-terminal loop or the C-terminal arm with DNA hamper the association of SAP with DNA, which will be discussed below in energetic analysis.

Because of the contraction that occurred for the ring structure, we observe the association of Ku70 with DNA is significantly tightened in all complex systems. This is demonstrated not only from the closer interaction of the bridge with the DNA duplexes (Fig. 2), but also from the simulation identified direct interactions of nucleotide bases with the residues of the β-barrel domain. Figure 3 depicts the close contacts of R403 and R254 to the bases in Ku70D at the end of simulation. Conformational analysis indicates some of them are actually strong H-bonds during the most time of simulation (Table 2). These base touches are rather dynamic in the process of breaking and formation, with competition between interactions with neighboring bases and with the phosphodiester groups of DNA. Since these features were identified in all five complex systems (Table 2 and Supporting information Table S3A-D), we curiously examined the reported crystal structure of Ku-DNA complex as Ku-DNA binding is widely recognized as in a sequence-independent manner. Unexpectedly, we found an overlooked close contact between R400 of Ku80 with the base of adenine group of DNA at level −4, with P-N distance of 2.59 Å (PDB code 1JEY [6]), which is presumably a strong H-bond. However, in the crystal structure, the R254 and R403 groups are clearly associated with the phosphodiester groups of DNA, as their distances to the closest phosphodiester groups are significantly shorter than the distances to the bases. As computational simulations are supposed to reflect a situation closer to solution phase dynamics while the crystallographic experiments are done under different condition and on a vastly different timescale, at this point we cannot attribute the formation of such base touches solely to the structural contraction of the protein. Further work is needed to address this issue possibly by carrying out simulations of Ku70/80 bound to DNA, or developing new detecting methods to identify the presence of such base interactions experimentally. Whether these interactions may affect the prevailing concept of sequence non-specificity for Ku-DNA binding will be discussed in following energetic analysis.

The seven systems that we simulated in detail indicate that, without the support of Ku80, the core channel structure of Ku70 is not stable. This means the celebrated bolt-nut mode for Ku-DNA binding may not be formed for Ku70 alone with DNA duplex segments. It is however highly possible that other binding modes may exist, as has been demonstrated in a previous protease digestion experiments with Ku binding with different forms of DNA [31]. In these experiments, the authors showed that Ku could adopt multiple conformations on DNA, each specific for a particular DNA form. Particularly, there were evidences that Ku can bind with sequence-specific closed DNA microcircles that are without any breaks or nicks [31]. The structure of Ku70 reveals that the inner surface of the channel dominated with charged residues is not fully covered by the narrow bridge but mostly exposed to the solvent. These residues are therefore still capable of forming strong association with DNA from the side of the ring. In addition, other domains of Ku70 such as the SAP domain and the C-terminal loop may also contribute to the alternative modes. The controversial issue of whether individual Ku subunits could bind DNA might be related to the sensitivity of the detecting methods. It has been reported that, in DNA immunoprecipitation and southwestern blot assays, Ku70 binds DNA in the absence of Ku80 [29, 32, 33], but in gel shift assays, only the dimer can bind [8, 34, 35]. In experiments that Ku80-independent binding were observed, the binding affinities were apparently reduced as compared to that of the dimer [29]. Nevertheless, the bolt-nut mode on which we investigated can yield a set of peptide-nucleotide interactions that are ideally complemented to the topological structure of the interfaces

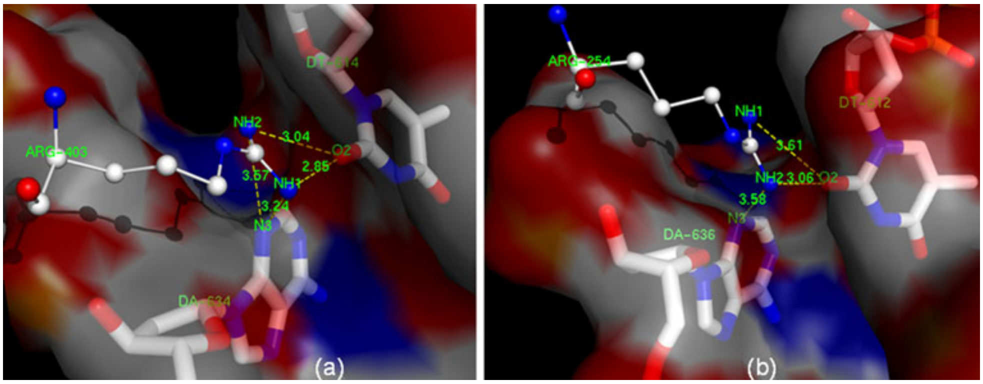


Fig. 3 H-bond network of R403 (**a**) and R254 (**b**) formed with nucleotide bases in the minor groove. Produced from the snapshot of Ku70D at 20 ns with PyMOL [26]. The minor groove of DNA is displayed in semi-transparent surface representation, while the involved nucleotides are in stick form. The same color scheme is applied for the heavy atoms of protein and nucleic acid. The unit of labeled bond length is Å

Table 2 H-bonds between R254, R403 and the bases of DNA in Ku70D during the last 5 ns MD simulation. Donors and acceptors are reported as res@atom. Data is sorted according to occupancy (%). Distance cutoff is 3.00 Å, angle cutoff is 120.0 degrees. The distances (Å) are those between the donor and the acceptor in each pair, angles (degree) are 180.0-<(donor, acceptor-H, and acceptor), and standard deviations are in the brackets

Donor	Acceptor-H	Acceptor	Occupancy	Distance	Angle
T614@O2	403@HH12	403@NH1	69.2	2.84 (0.08)	29.5 (12.4)
T612@O2	254@HH22	254@NH2	48.4	2.84 (0.09)	30.7 (12.7)
T614@O2	403@HH22	403@NH2	37.6	2.86 (0.08)	33.7 (11.0)
A636@N3	254@HH22	254@NH2	16.4	2.90 (0.08)	37.8 (13.2)
A634@N3	403@HH12	403@NH1	15.6	2.91 (0.06)	36.3 (14.7)
T612@O2	254@HH12	254@NH1	14.4	2.88 (0.09)	30.2 (12.0)
A636@N3	254@HH12	254@NH1	4.8	2.90 (0.06)	33.7 (10.4)
A634@N3	403@HH22	403@NH2	0.8	2.86 (0.11)	42.6 (8.0)
A636@N3	254@HH21	254@NH2	0.4	2.87 (0.00)	55.7 (0.0)
T612@O2	254@HH21	254@NH2	0.4	2.95 (0.00)	48.0 (0.0)

of protein and DNA. The maintenance of the channel structure, particularly of the narrow bridge, is probably essential to confine the DNA free ends along a path that further processes of DNA repair can be facilitated. Our data support the notion that integrity of Ku heterodimer is important for its function in NHEJ pathway [11].

Binding energies analysis

The dissociation constant of Ku-DNA evaluated in vitro is in the range of 2.4×10^{-9} to 5.0×10^{-10} M [3], indicating a very strong interaction among the reported protein-DNA complexes. Considerable progress has been achieved in recent years using the continuum electrostatics models to estimate the binding affinities and free energies of macromolecular binding processes [36]. In the following, based on the MM-GBSA method [27], the energetic analysis on the overall binding as well as contributions of different regions of Ku70 involved in DNA binding are reported. This is followed by analyzing the role of key individual amino acids located at the inner surface of the channel structure that form close contacts with DNA.

Table 3 lists the free energies of binding and their major components for the various complexes, among which $\Delta G'_{bind}$ refers to the binding free energy without the contribution of the configuration entropy TΔS while ΔG_{bind} includes such contribution. The calculated binding energies for all systems are extremely exothermic. It should be noted all these calculations start with structures where the DNA duplexes are already bound in the channel of the protein. This scenario is vastly different from the way that macromolecules encounter, interact, and form association with each other in the in vitro or in vivo experimental settings. What Table 3 shows are actually the energetic contributions of the Ku70 subunit to the bolt-nut mode interactions of Ku heterodimer with different DNA duplexes. Previous studies indicated Ku70 is the major contributor to Ku-DNA binding [19, 37]. As the magnitudes of the contribution of the electrostatic and van der Waals as well as the entropy are also consistent with similar protein-DNA systems in the literature [38, 39], and the emphasis in this work are placed on the differences in energies and structures between the systems rather than the absolute energies, these results allow for confidence to the following elucidation of atomic details of events driving Ku70-DNA binding.

In five Ku70-DNA complexes the major components of binding free energies are stunningly different (Table 3). The combined electrostatic term ($\Delta E_{ele}+\Delta\Delta G_{GB}$) contributes unfavorably to the binding in Ku70D, but favorably in the other four systems. This term is usually reported as destabilizing the protein-DNA complexation in the literature [39, 40]. How the extension of DNA changes this trend is an interesting issue, however it is consistent with the observation that Ku-DNA binding is more efficient for nucleic acids with bp>20 [41]. In the systems we considered, the van der Waals interaction (ΔE_{vdW}) is the

Table 3 Binding free energy components of Ku70-DNA complexes. All values with standard deviations followed are given in kcal mol^{-1}, which are averaged over 250 (5 in the case of entropy contribution) snapshots during the last 5 ns of simulations

System	$\Delta E_{ele}+\Delta\Delta G_{GB}$	ΔE_{vdW}	$\Delta\Delta G_{SA}$	$\Delta G'_{bind}$	TΔS	ΔG_{bind}
Ku70D	19.6±13.1	−181.5±12.7	−30.4±1.6	−192.3±14.3	−113.2±12.3	−79.1±18.9
Ku70AT	−9.8±18.1	−137.5±12.2	−25.8±1.2	−173.1±11.2	−105.9±19.8	−67.2±22.7
Ku70GC	−8.6±15.9	−140.2±12.1	−25.1±1.5	−173.9±11.1	−88.5±17.8	−85.4±21.0
Ku70AT2L	−20.1±15.2	−146.7±8.8	−28.0±1.0	−194.8±14.2	−95.2±13.2	−99.6±19.4
Ku70AT2R	−17.9±15.0	−169.7±10.3	−29.6±1.3	−217.2±16.0	−100.4±18.8	−116.8±24.7

major contributor to the binding energies, contributing the most in Ku70D, followed by Ku70AT2R. The extensions of the 14mer DNA duplex systematically diminish such contributions, probably caused by averaging the extended surface of DNA that is not involved in direct interaction of the protein in these four systems. The difference of ΔE_{vdW} between Ku70AT2L and Ku70AT2R is also significant (≈23.0 kcal mol^{-1}), indicating the N-terminal loop is energetically more favorable to locate at the right side of DNA when Ku70 binds at internal site of DNA (Fig. 1). This is supported by the fact that this loop in Ku70AT2R maintains much more direct contacts with DNA than in Ku70AT2L, as discussed above. The non-polar portion of the solvation energy $\Delta\Delta G_{SA}$ follows a similar trend as ΔE_{vdW}, while the counter contribution of the entropy term is significantly large in all the systems. Overall, the magnitudes of the binding free energies of Ku70AT2L and Ku70AT2R are significantly larger than Ku70AT and Ku70GC, though the bound DNA duplexes are of the same length. This suggests that Ku70 sliding to internal positions of the DNA molecule is an energetically advantageous process, which may provide a mechanistic basis to rationalize the well-known energy-independent feature of the translocation of Ku along DNA [3].

Table 4 shows the contribution of each domain of Ku70 to the binding free energies of five complexes. Though lacking entropy contributions due to the computational difficulty, this analysis can help in understanding the detailed interaction and function of each domain in the process of DNA binding. The primary contributions apparently come from the β-barrel domain and the bridge part, including residues 251–439 that form the channel. The contributions of β-barrel domain are persistently the largest and are almost of the same magnitude in five systems. The contributions from the bridge part show some variations, which are possibly caused by its flexibility and the different configurations in different systems (Table 1 and Fig. 2). For the three systems with DNA duplexes ending at the N-terminal of Ku70 (Ku70D, Ku70AT, and Ku70GC), the energetic contribution of the N-terminal loop and α/β domain are very small. The extension of DNA at this site (Ku70AT2L and Ku70AT2R) enhances their contributions. However, the magnitude of the overall contributions of these two domains suggests that they just play auxiliary roles in DNA binding [8, 35, 42]. The three domains at the C-terminal side of Ku70 altogether show more contribution to DNA binding, but with dependence of their location with respect to DNA (Fig. 2 and Table 4). The C-terminal arm, which functions as a holder for the β-barrel domain of Ku80 in heterodimer, is observed to form an interaction with the backbone of DNA at the positively charged patch from K443 to T449 in Ku70D and Ku70AT2R, while in other systems almost no contribution can be detected (Table 4). In the C-terminal loop, there are two positively charged patches that can participate in DNA binding, one with residues from K539 to H545, the other from K553 to K556. The contribution of this loop can be as large as −23.0 kcal mol^{-1} in the calculation of Ku70GC. The contributions of the SAP domain are also system-dependent. This domain has been found in biochemical studies to be responsible for the Ku heterodimer's high affinity binding to DNA [8, 35], and has been proposed to be a DNA-binding motif in different proteins [30]. Our analysis indicates, however, this domain is not a major contributor to bind DNA as compared with the ring part of Ku70, and competition may exist between this domain and the proximate C-terminal arm and C-terminal loop (Table 4).

Binding energy partitioning analysis also indicates that the primary contribution of Ku70-DNA binding comes from the positively charged residues along the inner surface of the ring (Table 5). In each of the five Ku70-DNA systems, five arginine groups R254, R258, R363, R403, and R404 are consistently among the groups that contribute the most to the binding affinity, each in the range of −7.1∼−11.2 kcal mol^{-1} (Table 5). All of them belong to the β-barrel domain and are located at the inner surface of the cradle. Another arginine group R252 of this surface demonstrates enhanced contributions in Ku70GC and Ku70AT2R only (−5.9 and −8.0 kcal mol^{-1}, respectively). Two arginine groups R301 and R318 of the bridge region also show very strong interactions with DNA in some systems (Table 5). Analysis indicates the contributions from the lysine groups are systematically smaller than those of the

Table 4 Domain contributions of Ku70 to the Ku70-DNA binding energies. All values with standard deviation followed are given in kcal mol^{-1}, which are averaged over 250 snapshots during the last 5 ns. Residues in each region of protein are the same as in Fig. 2

System	N-terminal loop	α/β domain	β-barrel domain	Bridge	C-terminal arm	C-terminal loop	SAP domain
Ku70D	−4.0±1.1	−0.5±1.7	−64.3±4.6	−41.2±4.2	−7.9±2.7	−7.3±2.8	−11.3±2.9
Ku70AT	−2.2±1.7	0.6±1.2	−67.1±5.1	−26.8±5.8	−0.9±0.4	−9.3±4.1	−21.2±4.2
Ku70GC	−8.0±1.4	−3.2±2.3	−60.5±4.8	−29.2±5.9	−2.8±2.3	−23.0±2.8	0.0±0.0
Ku70AT2L	−8.4±2.0	−7.3±3.1	−64.3±3.3	−45.7±5.8	−1.4±2.8	−14.3±4.2	0.3±0.2
Ku70AT2R	−11.2±3.7	−7.8±2.4	−65.0±4.6	−44.9±5.6	−14.4±3.9	−0.1±0.0	0.0±0.0

Table 5 The contributions to the binding free energy of Ku70-DNA complexes from residues of the inner channel with the largest contributions. All values with standard deviation followed are given in kcal mol^{-1}, which are averaged over 250 snapshots during the last 5 ns. Residues of the bridge part are indicated with "*"

Residue	Ku70D	Ku70AT	Ku70GC	Ku70AT2L	Ku70AT2R
R252	−0.2±0.0	−0.7±0.3	−5.9±1.3	−0.8±0.1	−8.0±1.8
R254	−9.7±1.6	−7.7±1.2	−8.1±1.2	−8.1±1.4	−9.1±1.8
R258	−9.5±2.7	−11.2±0.9	−10.2±1.2	−11.0±1.1	−9.0±1.8
K297*	−0.5±0.7	−0.7±0.5	−3.1±2.1	−4.8±1.6	−6.2±2.1
R301*	−0.4±0.0	−5.7±2.5	−10.3±2.3	−6.0±2.4	−4.6±2.8
R318*	−8.3±0.9	−6.3±2.1	−4.6±2.4	−11.8±1.5	−12.0±1.9
K331*	−6.3±1.2	−3.0±2.8	−1.4±1.1	−4.2±2.9	−8.8±1.8
R363	−10.2±1.9	−10.8±3.6	−7.1±1.2	−7.6±1.0	−9.6±2.3
R403	−10.2±1.1	−10.6±1.0	−8.4±2.6	−7.2±1.2	−10.0±1.1
R404	−8.9±0.8	−9.7±1.7	−9.0±2.4	−9.6±1.0	−8.8±1.6

arginine groups. Though both residues bear one positive charge in physiological conditions, arginine has three nitrogen atoms at the tip of its side chain, and therefore is more advantageous to form direct electrostatic/H-bonding interactions with neighboring nucleotides. This difference may have not been fully aware of in current biochemical studies. The large-scale mutational analysis based on site-specific mutagenesis conducted on Ku70 did not consider any change of these spots [35]. In the recent chemical modification experiment, K331 was identified as a contact point to DNA [11], which is in agreement with our simulations (Table 5). However, this study also focused on the lysine groups and provided no information of the arginine groups.

According to Table 5, the contributions of R254 and R403 which interact directly with the nucleotide bases are not significantly enhanced, as compared to those of three other arginine groups (R258, R363, and R404) which associate exclusively with the phosphodiester groups during simulations. However, for this pair of positively charged residues, there is a clear correlation of the magnitude of the calculated contribution to the H-bonding occupancy identified in simulations (Table 2 and Supporting information Table S3A-D). The energetic contributions of this pair are consistently higher in systems with higher H-bonding occupancy (Ku70D and Ku70AT2R), though the difference is not very significant (~1-3 kcal mol^{-1}; see Table 5). Therefore these base interactions may not significantly affect the sequence-independence feature of Ku-DNA binding. However it is possibly related to the early biochemical study of the preference of Ku's binding with AT-rich DNA ends to GC-rich ones [18], and the observed pausing of Ku at specific DNA sequences [43].

The presence of H-bonds between these two arginine groups and the DNA bases supports the notion that DNA is constrained to move inward with a helical path through the ring of Ku [6]. A close examination of the inner surface of the cradle tells that the surface is not flat but equipped with extruded contour perfectly complimented to the minor groove of DNA (Fig. 4). Along with the two arginine groups which are deeply wedged into the groove, several other proximate residues (R258, R363, R404, etc.) also form strong interactions with the phosphodiester groups of both strands. The persistent presence of these interactions indicates that they may work together to confine the movement of DNA duplex, which could be essential to structurally support and align DNA ends to maintain the correct setting of thermodynamically weak base pairing and stacking interactions [6].

Mimicking the loading and unloading processes

After the introduction of DNA breaks, it has been found that Ku can load onto the DNA ends very rapidly (within seconds) in living cells [44]. This is probably due to the extraordinarily high affinity of Ku for DNA termini [3] and its abundant presence in the cell nucleus (approximate 5×10^5 per nucleus) [45]. The subsequent recruitment of other repair factors such as DNA-PKcs in higher eukaryotes generally results in further translocation to the internal site of DNA [3]. After

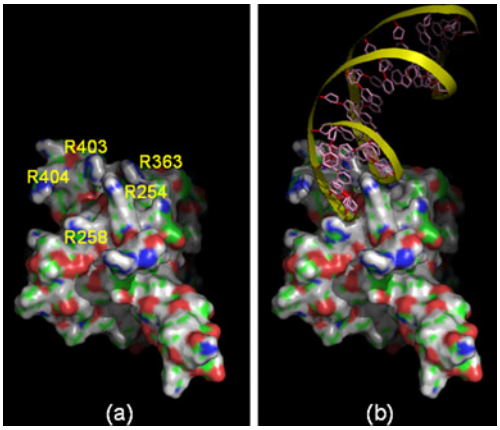


Fig. 4 Extruded contour of the inner surface of the Ku70 β-barrel (**a**) which is perfectly complemented to the minor groove of DNA (**b**). Produced from the final snapshot of Ku70D simulation with PyMOL [26]

ligation of the break, if no degradation occurs, the Ku70/80 will probably be trapped on the DNA helix, and will present a challenge to the progression of the replication fork [6]. To understand in more details the interactions between Ku70 and DNA duplex during the processes of association and dissociation, two TMD simulations were conducted to imitate the initial loading of Ku70 onto the DNA duplex and the unloading process before their departure.

Due to the large size of the systems and short time steps (see Methods), the two TMD simulations we conducted were confined to a time frame of 1 ns. This limitation has shown both pros and cons for our purpose. First, we observe the Ku70 subunit in these simulations can maintain their core structure without significant contraction. Therefore the systems allow the loading and unloading of DNA duplex with respect to the channel to occur with less steric block. In the meantime, the effect of losing the support of Ku80 on Ku70-DNA interaction can be mitigated. However, from the MM-GBSA energetic profile depicted in Fig. 5a, the binding free energies of the bound structures (the starting structure of unloading and the ending structure of loading) are significantly smaller than that calculated from the last 5 ns simulation of Ku70D (Table 4). This is apparently due to the absence of the effect of structural contraction. In addition, we observe smaller association energies for the conformations at the end of loading simulation as compared to those when the unloading simulation starts (Fig. 5a). This indicates the loading process is not fully completed. A careful examination on the final structures of loading simulation indicates that the optimal binding pattern observed at the end of 20 ns simulation is not fully formed. Although the dragging force applied to the DNA duplex accelerates the diffusion of the segment to a location in the channel of Ku70, the final configuration is probably just close to a transition state and will be subject to further adjustment [46]. The following results thus only highlight the energetic and structural features during the courses of loading and unloading.

The results of energetic analysis are consistent with the observed structural features of Ku70, especially the inner surface of the ring. For the loading simulation, the system starts to gain binding energy at around 600 ps (Fig. 5a), which is the moment that one end of the DNA duplex reaches the opening of the ring. From this point on we also observe continuous shortening of the DNA length (Fig. 5b), and more and more associations formed between the DNA end and Ku70 residues at the opening of the channel. The snapshot at 872 ps is a turning point for this process. At this point, the system experiences the largest gain in binding free energy and the shortest DNA length (Fig. 5a and b). At this moment, the DNA end is still at the opening of the ring, but has formed many strong associations with the basic residues on the bridge and the inner surface of the cradle (Fig. 6a). It takes about 172 ps for the DNA end to process from the first contact at the opening of the ring to the fully associated configuration. After this snapshot, the DNA end enters the ring, and the system starts to destabilize for more than 20.0 kcal mol^{-1} and cannot regain the loss even at the end of the simulation (Fig. 5a). This is highly indicative that it is not an energy favored process to pull the DNA end from one side of the ring to the other side by the force applied along the DNA axis. The shape of the cradle surface suggests that, if the applied dragging forces are not along the DNA axis but in a direction along the helical path of DNA, the association would encounter less resistance, especially from the two extruded arginine groups (Fig. 4).

The helical path of Ku70-DNA translocation is also suggested from the analysis of the unloading trajectory.

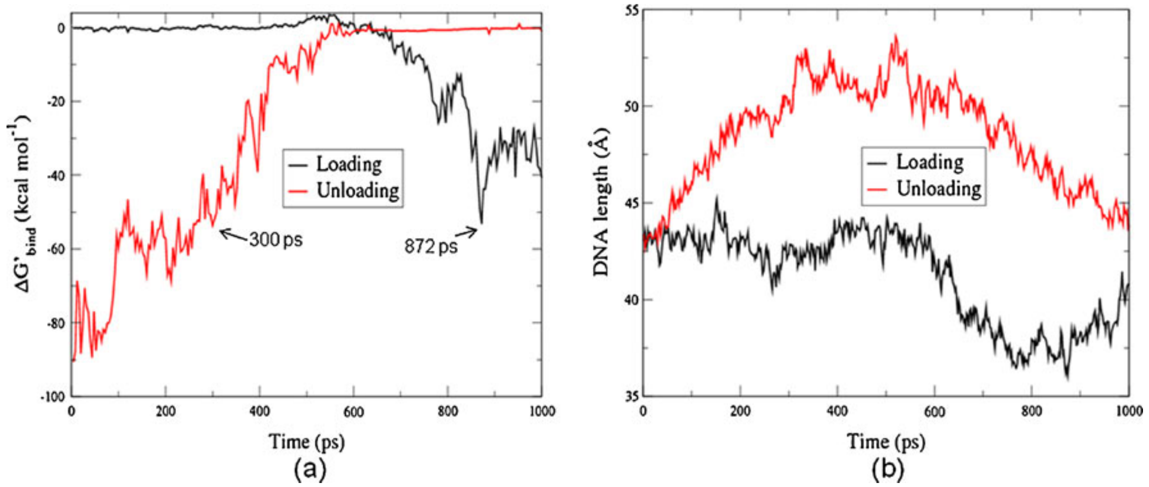


Fig. 5 Energetic profile (**a**) and the changes of DNA length (**b**) during the loading and unloading processes of Ku70D. The binding free energies are calculated from 250 snapshots extracted evenly from the two 1 ns TMD simulations with MM-GBSA method, without the entropy contributions. The lengths of DNA duplex are calculated with the distance function of the ptraj module of Ambertool [28]

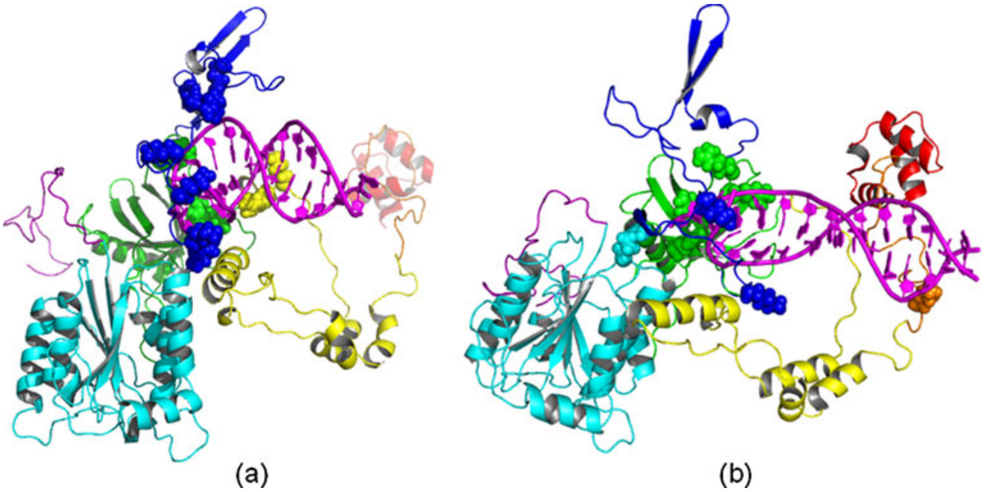


Fig. 6 Structures of Ku70D at 872 ps of loading simulation (**a**) and 300 ps of unloading simulation (**b**), which are the turning points of Ku70-DNA association and dissociation (see text). The calculated binding free energies are (**a**) -53.25 and (**b**) -53.62 kcal mol^{-1}, respectively. Also the basic groups (arginine and lysine) of Ku70 within 5 Å of the DNA duplex are shown in sphere representation. The color scheme is the same as in Fig. 2

During the first 100 ps of this trajectory, the strong interactions between the DNA and the residues along the inner surface of the Ku70 ring are almost unaltered, and the DNA end is kept bound in the starting position. Therefore, the binding free energies do not change much during this period (Fig. 5a). However, it is observed that the DNA duplex is continuously elongated (Fig. 5b) and its B-form structure is gradually deformed. The sudden loss of binding energy at about 100 ps (Fig. 5a) is most likely caused by the breakage/adjustment of the strong interactions of DNA with the residues on the inner surface of the channel, as no abrupt conformational alterations can be observed for the protein as well as the DNA duplex at this stage. Figure 5b shows that, starting from the beginning of the unloading simulation, the length of the duplex is increased almost linearly with time, and reaches a plateau after 200 ps. Visual examination of the trajectory indicates the duplex has finished its deformation at this point, which leads to a shrunken diameter, enabling the pulling out of the duplex to process smoothly. During the following 100 ps simulation, although there is a significant loss of the binding free energy, the length of the DNA is almost unchanged (Fig. 5a and b). Apparently the deformation and the shrinking of DNA are caused by the combination of the force applied along the DNA axis and the uneven inner surface of the narrow channel. A helical path which the Ku protein rotates around DNA helical structure is therefore a reasonable concept for the system to maintain the delicate forces of DNA to base pairing and stacking interactions while allowing the Ku protein to translocate along the duplex with minimal expense of energy [6].

Similar to the snapshot at 872 ps of the loading simulation, the snapshot at 300 ps of unloading simulation also represents a turning point for this process. At this stage, the DNA end has already been pulled out from the channel, but there are still many strong interactions available between the DNA end and the basic residues of the inner surface of the channel (Fig. 6b). The calculated binding affinity is as strong as the configurations just after 100 ps (Fig. 5a), in which the end of the DNA duplex is still almost fully trapped in the channel. The following departure from this position causes the duplex to elongate another 2.5 Å (Fig. 5b), leading to a most disrupted structure along the whole trajectory. Interestingly, the calculated binding free energies of these two structures are very close to each other (Fig. 6). Though in both structures the end of the DNA is similarly located at the opening of the ring, the patterns of protein-DNA interaction are very different. The bridge part apparently plays a more important role in the loading process, as there are seven basic groups of this region forming close contacts with DNA (within 5 Å). These include R301, R318, K279, K282, K287, K297, and K331 (Fig. 6a). By contrast, only two of them (K279 and K297) closely interact with DNA in the unloading structure (Fig. 6b). The β-barrel only makes a minor contribution at this stage of loading, as only two residues (R363 and R403) form close interactions with DNA; two basic groups from the C-terminal arm (R444 and K445) also join these interactions (Fig. 6a). Therefore, the binding pattern of Ku70-DNA in the loading structure is very different from the complex analyzed above (Table 5). On the contrary, despite the significant deformation of the DNA duplex and a certain translocation of the DNA end in the unloading structure, the binding network between Ku70 and DNA at this stage is still similar to the original structure. The basic groups from the β-barrel (R254, R258, R403, and R404) are still among the groups maintaining closest contacts with DNA. This implies that, once the interactions in Ku-DNA are formed, it is very hard to break them, which is consistent with the experimentally measured high affinity of Ku-DNA binding [3], and the observed vast difference between the association rate and dissociation rate [37].

Conclusions

The Ku protein has undergone many experimental studies since it was initially identified about three decades ago, however details of its many functions in DNA repair as well as in other fundamental cellular processes are still the subject of much debate. In this study, by utilizing some well-established computational tools, we analyzed the unique interactions between a full-length human Ku70 subunit and several DNA duplexes in atomic details. The results reveal a number of features not previously characterized by using ordinary biochemical or biophysical techniques, and thereby shed light on some controversial issues of this protein and its interaction with DNA.

Many previous experiments implied that the two subunits of Ku need to be dimerized so as to function properly in DNA repair [47–49]. Our simulations reveal that, without the support of Ku80, the ring structure and the C-terminal arm of Ku70 can easily be contracted, and the N-terminal loop may be attracted into the positively charged ring due to its overall negative charges. All these conformational changes most likely preclude the binding of Ku70 to DNA ends in the bolt-nut mode as identified experimentally for the Ku heterodimer. However, as the opening of the ring and other domains such as the C-terminal loop and the SAP domain also demonstrate DNA-binding capability in this study, it is possible that some different modes may exist for Ku70 monomers to bind DNA. The energetic analysis indicates the binding of DNA with these regions are significantly weaker than that of the β-barrel and the bridge domains in the bolt-nut binding mode, which may provide a basis to resolve the issue of subunit independent DNA binding of Ku70 or Ku80. Whether such bindings can be detected is probably dependent upon the sensitivity of the methods.

It is widely recognized that Ku interacts with DNA in a sequence-independent manner [3]. However, there have been reports that, when Ku translocates along DNA, it can be observed to pause at specific locations [6, 43]. Such sequence specific interactions are extremely interesting, as Ku is known to play important roles in the transcription and other chromosomal regulatory processes. Previous works proposed the SAP domain, which shows DNA binding capability, might cause such sequence preference [6]. However to date it seems that no experiment has been carried out to verify this proposal. Alternatively, we propose this type of interaction could be related to the base contacts identified in this study. Interestingly, there is a strong H-bond between R400 of Ku80 with the nucleotide base in the published crystal structure, which was overlooked in the corresponding reference [6]. In our simulations, all these base interactions are located in AT region. The AT pair have two donor atoms (N3 of adenine and O2 of thymine) that are unpaired in the minor groove, which presumably make it easier for the acceptor atom of the arginine to form H-bonds. By contrast, it is known that no such atom exists in GC pair along the minor groove. This analysis is therefore consistent with the observation that Ku's binding with AT-rich DNA ends is preferred to with GC-rich ones [18].

The structural analysis also indicates that the inner surface of the channel, particularly the surface of the cradle, is not flat, but is equipped with an uneven contour composed of positively charged residues, which is perfectly complemented with the shape of the minor groove of DNA. The residues that contribute most to the binding affinity of the complexes form a clamp like shape, being able to grip both strands of DNA at this spot. As these interactions are unlikely dissolved during the translocation of Ku along DNA, the strong association of Ku70-DNA may work together with other domains to confine the movement of Ku along DNA in a unique helical path. Two targeted MD simulations conducted to imitate the initial association and the final dissociation of Ku70 with DNA also indicate that a direct translation of the DNA duplex along its axis results in deformation and corruption of the canonical structure. This finding suggests a helical path for the Ku protein to rotate around DNA duplex while translocating internally is essential for the system to maintain the delicate forces of DNA to base pairing and stacking interactions.

From this study, all loops and domains of Ku70 appear to be well designed to work together to carry out its unusual mode of DNA binding. The C-terminal region contains one flexible loop with DNA-binding capability and one special DNA-binding motif (SAP domain) that has also been identified in other protein families. Since the Ku protein approaches DNA free ends from this side, it is highly possible that this region may work as the first functional group to interact with DNA. In the loading and unloading simulations, we found, when the DNA duplex is close to Ku70, the opening of the ring can form interactions with the DNA end that are as strong as the nearly loaded conformation. The ring along with the cradle which serve as the core of this protein contribute the most to the binding affinity to associate with DNA; its inner surface contains a delicate clamp like structure to grip the strands of the minor groove of DNA that could confine the duplex to move in a unique helical path. At the N-terminal side, some residues at the acidic loop can stack with the DNA end, thus providing some hindrance to the further translocation of DNA. Though the α/β domain apparently does not participate in DNA-binding, its presence is important to buttress the narrow bridge of the channel. A previous study indicated that, after deleting either some or all amino acids of this domain, Ku70 can still heterodimerize with Ku80, but none of the truncated mutants maintains the capability of binding DNA ends [32]. Due to the structural similarity, we believe the counterpart Ku80 may contain some similar or different structural features that are also suitable for DNA binding, enabling the heterodimer to play its unique role in DNA repair and other prominent cellular processes. In future work, we will use the results of the current analysis to develop approaches to further dissect the interactions of Ku70/80 heterodimer with DNA, which will be an extremely challenging computational problem.

Acknowledgments This work was supported by the NASA Space Radiation Program, Risk Assessment Project.

References

1. Mahaney BL, Meek K, Lees-Miller SP (2009) Repair of ionizing radiation-induced DNA double-strand breaks by non-homologous end-joining. Biochem J 417:639–650
2. Weterings E, Chen DJ (2008) The endless tale of non-homologous end-joining. Cell Res 18:114–124
3. Downs JA, Jacksons SP (2004) A means to a DNA end: the many roles of Ku. Nature Rev Mol Cell Biol 5:367–378
4. Lieber MR (2008) The mechanism of human nonhomologous DNA end joining. J Biol Chem 283:1–5
5. Lieber MR, Lu H, Gu J, Schwarz K (2008) Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and ligase of vertebrate non-homologous DNA end joining: relevance to cancer, aging, and the immune system. Cell Res 18:125–133
6. Walker JR, Corpina RA, Goldberg J (2001) Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. Nature 412:607–614
7. Doherty AJ, Jackson SP (2001) DNA repair: how Ku makes ends meet. Curr Biol 11:R920–R924
8. Wu X, Lieber MR (1996) Protein-protein and protein-DNA interaction regions within the DNA end- binding protein Ku70-Ku86. Mol Cell Biol 16:5186–5193
9. Chan DW, Ye R, Veillette CJ, Lees-Miller SP (1999) DNA-dependent protein kinase phosphorylation sites in Ku 70/80 heterodimer. Biochemistry 38:1819–1828
10. Zhang Z, Zhu L, Lin D, Chen F et al. (2001) The Three-dimensional Structure of the C-terminal DNA-binding Domain of Human Ku70. J Biol Chem 276:38231–38236
11. Lehman JA, Hoelz DJ, Turchi JJ (2008) DNA-dependent conformational changes in the Ku heterodimer. Biochemistry 47:4359–4368
12. Harris R, Esposito D, Sankar A et al. (2004) The 3D solution structure of the C-terminal region of Ku86 (Ku86CTR). J Mol Biol 335:573–582
13. Zhang Z, Hu W, Cano L et al. (2004) Solution structure of the C-terminal domain of Ku80 suggests important sites for protein-protein interactions. Structure 12:495–502
14. Singleton BK, Torres-Arzayus MI, Rottinghaus ST et al. (1999) The C terminus of Ku80 activates the DNA-dependent protein kinase catalytic subunit. Mol Cell Biol 19:3267–3277
15. Rivera-Calzada A, Spagnolo L, Pearl LH, Llorca O (2007) Structural model of full-length human Ku70-Ku80 heterodimer and its recognition of DNA and DNA-PKcs. EMBO Rep 8:56–62
16. Hammel M, Yu Y, Mahaney BL et al. (2010) Ku and DNA-dependent protein kinase dynamic conformations and assembly regulate DNA binding and the initial non-homologous end joining complex. J Biol Chem 285:1414–1423
17. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723
18. Falzon M, Fewell JW, Kuff EL (1993) EBP-80, a transcription factor closely resembling the human autoantigen Ku, recognizes single- to double-strand transitions in DNA. J Biol Chem 268:10546–10552
19. Yaneva M, Kowalewski T, Lieber MR (1997) Interaction of DNA-dependent protein kinase with DNA and with Ku: biochemical and atomic-force microscopy studies. EMBO J 16:5098–5112
20. Smith JA, Tsui VT, Chazin WJ, Case DA (1999) NMR Structure of the Palindromic DNA Decamer d(GCGTTAACGC)$_2$. http://www.rcsb.org/pdb/explore/explore.do?structureId=1cqo. Accessed November 23, 2010
21. Case DA, Darden TA, Cheatham TE et al. (2008) AMBER 10. University of California, San Francisco
22. Hawkins GD, Cramer CJ, Truhlar DG (1995) Pairwise solute descreening of solute charges from a dielectric continuum. Chem Phys Lett 246:122–129
23. Hornak V, Abel R, Okur A et al. (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins 65:712–725
24. Perez A, Marchan I, Svozil D et al. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophys J 92:3817–3829
25. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14:33–38
26. DeLano WL (2002) The PyMOL molecular graphics system. http://www.pymol.org. Accessed November 23, 2010
27. Cheatham TE III, Srinivasan J, Case DA, Kollman PA (1998) Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution. J Biomol Struct Dynam 16:265–280
28. Macke TJ, Svrcek-Seiler WA, Brown RA et al. (2008) Ambertools. University of California, San Francisco
29. Chou CH, Wang J, Knuth MW, Reeves WH (1992) Role of a major autoepitope in forming the DNA binding site of the p70 (Ku) antigen. J Exp Med 175:1677–1684
30. Aravind L, Koonin EV (2001) Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. Genome Res 11:1365–1374
31. Giffin W, Gong W, Schild-Poulter C, Haché RJG (1999) Ku antigen-DNA conformation determines the activation of DNA-dependent protein kinase and DNA sequence-directed repression of mouse mammary tumor virus transcription. Mol Cell Biol 19:4065–4078
32. Mimori T, Hardin JA (1986) Mechanism of interaction between Ku protein and DNA. J Biol Chem 261:10375–10379
33. Allaway GP, Vivino AA, Kohn LD et al. (1989) Characterization of the 70 kDa component of the human Ku autoantigen expressed in insect cell nuclei using a recombinant baculovirus vector. Biochem Biophys Res Commun 168:747–755
34. Griffith AJ, Blier PR, Mimori T, Hardin JA (1992) Ku polypeptides synthesized in vitro assemble into complexes which recognize ends of double-stranded DNA. J Biol Chem 267:331–338
35. Jin S, Weaver DT (1997) Double-strand break repair by Ku70 requires heterodimerization with Ku80 and DNA binding functions. EMBO J 16:6874–6885
36. Gilson MK, Zhou HX (2007) Calculation of protein-ligand binding affinities. Annu Rev Biophys Biomol Struct 36:21–42
37. Torrance H, Giffin W, Rodda DJ et al. (1998) Sequence-specific binding of Ku autoantigen to single-stranded DNA. J Biol Chem 273:20810–20819
38. Habtemariam B, Anisimov VM, MacKerell AD Jr (2005) Cooperative binding of DNA and CBFß to the Runt domain of

the CBFα studied via MD simulations. Nucleic Acids Res 33:4212–4222

39. Jayaram B, McConnell KJ, Dixit SB, Beveridge DL (1999) Free energy analysis of protein–DNA binding: the EcoRI Endonuclease–DNA complex. J Comput Phys 151:333–357

40. Jayaram B, Mcconnell K, Dixit SB et al. (2002) Free-energy component analysis of 40 protein–DNA complexes: a consensus view on the thermodynamics of binding at the molecular level. J Comput Chem 23:1–14

41. Arosio D, Costantini S, Kong Y, Vindigni A (2004) Fluorescence anisotropy studies on the Ku-DNA interaction: anion and cation effects. J Biol Chem 279:42826–42835

42. Wang J, Dong X, Reeves WH (1998) A model for Ku heterodimer assembly and interaction with DNA. Implications for the function of Ku antigen. J Biol Chem 273:31068–31074

43. de Vries E, van Driel W, Bergsma WG et al. (1989) HeLa nuclear protein recognizing DNA termini and translocating on DNA forming a regular DNA-multimeric protein complex. J Mol Biol 208:65–78

44. Mari PO, Florea BI, Persengiev SP et al. (2006) Dynamic assembly of end-joining complexes requires interaction between Ku70/80 and XRCC4. Proc Natl Acad Sci USA 103:18597–18602

45. Anderson CW, Carter TH (1996) The DNA-activated protein kinase-DNA-PK. In: Jessberger R, Lieber MR (eds) Molecular Analysis of DNA Rearrangements in the Immune System. Springer, Heidelberg, pp 91–112

46. Schreiber G, Haran G, Zhou HX (2009) Fundamental aspects of protein-protein association kinetics. Chem Rev 109:839–860

47. Gu Y, Jin S, Gao Y et al. (1997) Ku70-deficient embryonic stem cells have increased ionizing radiosensitivity, defective DNA end-binding activity, and inability to support V(D)J recombination. Proc Natl Acad Sci USA 94:8076–8081

48. Singleton BK, Priestley A, Steingrimsdottir H et al. (1997) Molecular and biochemical characterization of xrs mutants defective in Ku80. Mol Cell Biol 17:1264–1273

49. Errami A, Smider V, Rathmell WK et al. (1996) Ku86 defines the genetic defect and restores X-ray resistance and V(D)J recombination to complementation group 5 hamster cell mutants. Mol Cell Biol 16:1519–1526

J Mol Model (2012) 18:1951–1960
DOI 10.1007/s00894-011-1222-1

ORIGINAL PAPER

Modeling translocation dynamics of strand displacement DNA synthesis by DNA polymerase I

Ping Xie

Received: 21 June 2011 / Accepted: 11 August 2011 / Published online: 26 August 2011

Abstract A model is presented for the translocation dynamics of the strand displacement DNA synthesis by DNA polymerases such as polymerase I family. (i) The model gives an explanation to the experimental results which showed that the rate of strand displacement DNA synthesis is nearly consistent with that of single stranded primer extension synthesis, although the two are expected to have substantial differences in their energetics. (ii) During strand displacement DNA synthesis, the pausing at the specific sequence is considered to be due to an affinity of the fingers subdomain for the specific sequence of dsDNA downstream of the single strand. The theoretical results on the sequence-dependent pausing dynamics such as the mean pausing lifetimes and the distribution of the pausing lifetime are consistent with the experimental data. Moreover, predicted results are presented for the binding affinity of the fingers subdomain for the specific sequence of dsDNA and the dependence of the mean sequence-dependent pausing lifetime on the external force acting on the polymerase.

Keywords DNA replication · Model · Molecular motor · Sequence-dependent pausing · Translocation

Introduction

Replicative DNA polymerases exhibit high speed and high fidelity in replicating DNA to ensure the fast and faithful transfer of genetic information to daughter cells. During the replication, the polymerases are able to move processively along the DNA template accompanied by a processivity factor. In the case of the DNA polymerase I family, the enzyme can also displace the template strand as a helicase as it translocates down the template strand. This strand displacement DNA synthesis is an essential process in the removal and replacement of RNA primer moieties of Okazaki fragments [20]. Although this function requires both the polymerase domain and the 5′-nuclease domain of polymerase I, it has been known that the 5′-3′ polymerase activity and the strand separation activity resides in the polymerase domain [20, 27], with the O1-helix present in the fingers subdomain playing a role in the strand separation activity [36].

It has been known that the DNA replication by the DNA polymerases is not a uniform process. The replication can be slowed or paused by template tension [28, 42]. The DNA-binding ligands [37], stalled RNA polymerase [22] or DNA-bound proteins [4, 39] can block the replication elongation. The DNA polymerases can also pause due to specific DNA sequences such as palindromic DNA capable of forming hairpin secondary structure [1, 23, 24], slow zones [2], trinucleotide repeats of $(CGG)_n$ $(CCG)_n$ or $(CTG)_n$ $(CAG)_n$ [16, 32] and novel sites such as Pyr-G-C [29]. Recently, by using single-molecule techniques, the non-uniform polymerase activity and sequence-dependent pausing during the strand displacement DNA synthesis were demonstrated directly and their dynamics was studied quantitatively [35]. The pauses at the specific sequences have been proposed to be caused by difficulties in the polymerase fingers-closing conformational change, since this transition was thought to be rate-limiting and the most sensitive to changes in temperature [29]. However, a later experiment [15] showed that the slow prechemistry step is not the fingers-closing transition. Thus, another proposal was that the pauses are associated with an earlier DNA template rearrangement step that might be sequence dependent [35].

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1222-1) contains supplementary material, which is available to authorized users.

P. Xie (✉)
Key Laboratory of Soft Matter Physics and Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences,
Beijing 100190, China
e-mail: pxie@aphy.iphy.ac.cn

In this work, on the basis of available structures [7–11, 17–19, 21, 25, 26, 38, 49], a model is presented to describe the translocation dynamics of strand displacement DNA synthesis by DNA polymerases such as polymerase I family. The model is built up by modifying the previous model for the single stranded primer extension synthesis by replicative DNA polymerases [43, 45]. With the present model, the calculated moving times of the polymerase along the DNA during an incorporation cycle for both the strand displacement synthesis and the single stranded primer extension synthesis are much shorter than the chemical reaction time of the phosphodiester bond formation. This thus gives a good explanation to the experimental results which showed that the DNA synthesis rates for both cases are nearly consistent although they have substantial differences in energetics. The sequence-dependent pausing is considered to be due to an affinity of the fingers subdomain for the specific sequence of double-stranded DNA (dsDNA) downstream of the active site. The theoretical results on the sequence-dependent pausing dynamics such as the mean pausing lifetimes and the distribution of the pausing lifetime are in agreement with the experimental data. Moreover, we present predicted results for the binding affinity of the fingers subdomain for the specific sequence of dsDNA and the dependence of the mean sequence-dependent pausing lifetime on the external force acting on the polymerase.

Methods

Interaction of polymerase with DNA substrate

Before dNTP binding

Previous experiments on Klenow fragment and bacteriophage T4 DNAP showed that their fingers subdomains show a high binding affinity for 5′-single-stranded DNA (5′-ssDNA) overhang [5, 6, 40]. In addition, it was shown that the fingers subdomains of DNA polymerase I from *Bacillus subtilis* and *E. coli* have significant structural similarity with a novel DNA-binding motif found in transcription factor Mrf-2 [49]. On the other hand, the Mrf-2 DNA-binding domain can bind with a high affinity to specific sequences of dsDNA [49]. Thus, it is inferred that the fingers subdomain of polymerase I may also have an affinity for the specific sequences of dsDNA[1] while it shows very weak binding affinity for the nonspecific sequence of dsDNA. It is generally believed that the site-specific binding is mainly driven by the hydrogen bonding interactions between the protein and the specific site of the DNA lattice. Based on the above, we make the following hypothesis for the interaction of polymerase I with the DNA substrate. The fingers subdomain shows a high affinity for ssDNA template, with the interacting zones on the fingers drawn in pink in Fig. 1a. Although having very weak binding affinity for the nonspecific sequence of dsDNA, the fingers subdomain shows an affinity for the specific sequences of dsDNA, with the interacting zones on the fingers drawn in gray in Fig. 1a. Besides the interactions of the fingers subdomain with DNA substrate, it is hypothesized that other subdomains in the polymerase domain such as palm and thumb subdomains have a high affinity for dsDNA, with the interaction independent of the DNA sequence.

After dNTP binding

The available experimental evidence indicated that the dNTP binding to the active site involves (at least) two substeps, $E \cdot DNA + dNTP \rightarrow E \cdot DNA \cdot dNTP \rightarrow E^{*} \cdot \quad DNA \cdot dNTP$, where E represents the DNA polymerase [14]. The transition from the unactivated $E \cdot DNA \cdot dNTP$ ternary complex to activated $E^{*} \cdot DNA \cdot dNTP$ ternary complex induces the closing of the fingers subdomain, activating the chemical reaction of nucleotide incorporation. Similarly, the PPi releasing from the active site also involves (at least) two substeps, $E^{*} \cdot DNA \cdot PPi \rightarrow E \cdot DNA \cdot PPi \rightarrow E \cdot DNA +$ PPi, where the transition from the activated $E^{*} \cdot DNA \cdot PPi$ ternary complex to unactivated $E \cdot DNA \cdot PPi$ ternary complex results in the opening of the fingers subdomain. Here it is hypothesized that the closing of the fingers subdomain enhances the interactions of the polymerase with both DNA substrate and dNTP, while the opening of the fingers subdomain reduces the interactions of the polymerase with both DNA substrate and PPi. Since in the activated $E^{*} \cdot DNA \cdot dNTP$ (or $E^{*} \cdot DNA \cdot PPi$) complex the polymerase has a stronger interaction with its DNA substrate and nucleotide than in the unactivated $E \cdot DNA \cdot dNTP$ ($E \cdot DNA \cdot PPi$) complex, for simplicity of analysis, it is considered that in the activated state the polymerase is unable to move relative to the DNA substrate and the dNTP or PPi bound to the active site has a negligible probability to release.

Model for polymerase translocation

Based on the hypotheses for the interaction between the polymerase and the DNA substrate, as presented above, the model for the movement of the polymerase along DNA template during one incorporation cycle is described as follows.

We begin with the fingers subdomain of the polymerase binding the ssDNA template near the replication fork, as shown in Fig. 1a, with the active site being nucleotide free. In the nonspecific sequence, the fingers subdomain has no

[1] Due to slight structural differences, the fingers subdomain of DNA polymerase I and the novel DNA-binding motif of Mrf-2 could bind to different specific sequences of dsDNA

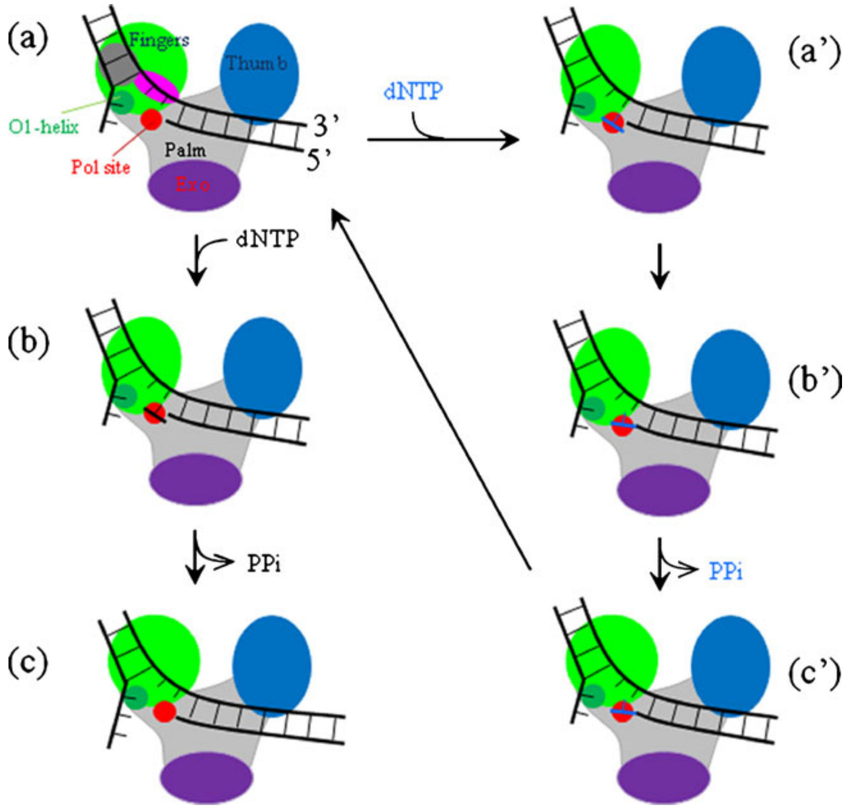


Fig. 1 Schematic illustrations of the model for processive translocation of DNA polymerase during strand displacement DNA synthesis (see text for detailed description). The green circles in (**a**), (**c**) and (**c'**) represent the open conformation of the fingers subdomain, while the ellipsoids (**b**), (**a'**) and (**b'**) represent the closed conformation. For clarity, the mismatched dNTP is drawn in blue while the matched dNTP in black

affinity for the downstream dsDNA, whereas in the specific sequence the fingers subdomain can also bind the downstream dsDNA. In this nucleotide-free state, either a matched or a mismatched dNTP can bind to the active site, although a matched dNTP has a much larger probability to bind. Thus we consider the two cases separately.

Incorporation of a matched base

After a matched dNTP binds to the active site, the transition from the unactivated E · DNA · dNTP ternary complex to activated E* · DNA · dNTP ternary complex induces the closing of the fingers subdomain, enhancing the interactions of the polymerase with both DNA substrate and dNTP and activating the chemical reaction of nucleotide incorporation (Fig. 1b). After the completion of the nucleotide incorporation, the transition from the activated E* · DNA · PPi ternary complex to unactivated E · DNA · PPi ternary complex results in the opening of the fingers subdomain, reducing the interactions of the polymerase with both DNA substrate and PPi. Then, the fingers subdomain will bind to new nearest unpaired base of the ssDNA template because the previous base where the fingers subdomain has just bound has disappeared due to base pairing and, at the same time, PPi is released from the active site (Fig. 1c). Note that, as the polymerase moves downstream to the next position, the O1-helix (refer to Ref. [36]) present in the fingers subdomain pushes on the non-template strand of the downstream dsDNA, thus breaking the base pair. Besides, the movement requires overcoming the binding affinity of the palm and thumb subdomains for the upstream dsDNA in the nonspecific sequence. In the specific sequence, it is required to overcome both the binding affinity of the palm and thumb subdomains for the upstream dsDNA and that of the fingers subdomain for the downstream dsDNA. Note here that, during either the closing or opening of the fingers subdomain, the polymerase has not moved relative to the DNA. The binding of the fingers subdomain to the new unpaired base on the ssDNA template makes the polymerase move forward relative to the DNA by one base pair.

Incorporation of a mismatched base

We still begin with Fig. 1a. After a mismatched dNTP binds, the transition from the unactivated E · DNA · dNTP ternary complex to activated E* · DNA · dNTP ternary complex induces the closing of the fingers subdomain, enhancing the interactions of the polymerase with both DNA substrate and dNTP and activating the chemical reaction of nucleotide incorporation (Fig. 1a'). After the completion of the nucleotide incorporation (Fig. 1b'), the transition from the activated E* · DNA · PPi ternary complex to unactivated E · DNA · PPi ternary complex results in the opening of the fingers subdomain, reducing the interactions of the polymerase with both DNA substrate and PPi. Then PPi is released (Fig. 1c’). However, although the sugar-phosphate backbone of the mismatched dNTP was connected to the backbone of the already formed dsDNA, the mismatched base is not paired with the sterically corresponding base on the ssDNA template. Thus, the fingers subdomain is still binding to the same unpaired base of the ssDNA template and the polymerization cannot proceed (Fig. 1c’). In other words, the polymerase becomes stalled. In Fig. 1c’, after the mismatched base is excised by the 3′-5′ exonuclease active site [45], the polymerization proceeds again (Fig. 1a).

Equations for polymerase translocation

We denote V_1 the interaction potential of the palm and thumb subdomains with the dsDNA segment upstream of the polymerase active site, V_2 the interaction potential of the fingers subdomain with the ssDNA template, and V_3 the interaction potential of the fingers subdomain with the specific sequence of dsDNA segment downstream of the active site.

Considering that the residues interacting with dsDNA on palm and thumb subdomains cover N_1 base pairs, the interaction potential $V_1(x_1)$ can be approximately shown in Fig. 2a (middle).[2] Here, E_1 is the binding affinity for the sugar-phosphate backbones connecting N_1 base pairs on the dsDNA, E'_1 is the binding affinity for the backbones connecting only (N_1-1) base pairs on the dsDNA, and x_1 represents the coordinate of the point on the dsDNA-binding residues that is nearest to the active site along the x_1 direction. This point is assumed to be $L_1=mp$ away from the active site, where p=0.34 nm is the distance between two successive base pairs (here we draw m=2). Note that the binding affinity E'_1 that corresponds to binding (N_1-1) base pairs is smaller than E_1 that corresponds to binding N_1 base pairs. Considering that the interaction is mainly via the electrostatic force, with the interaction distance approximately equal to the Debye length (of about 1 nm) in solution larger than p=0.34 nm, it is thus expected that the value at maxima of $V_1(x)$ increases as the dsDNA-binding site on the palm and thumb subdomains deviates away from the upstream dsDNA segment along the x_1 direction. Moreover, as the available structures of the polymerase complexed with the DNA substrate indicated, the primer 3′ terminus, due to the structural restriction, is not allowed to move forward relative to the polymerase when its active site is located at the primer 3′ terminus, implying that the polymerase is not allowed to move backward in this case. This feature is also represented in the form of the potential $V_1(x_1)$ [as seen below, it is also represented in the form of the potential $V_2(x_2)$].

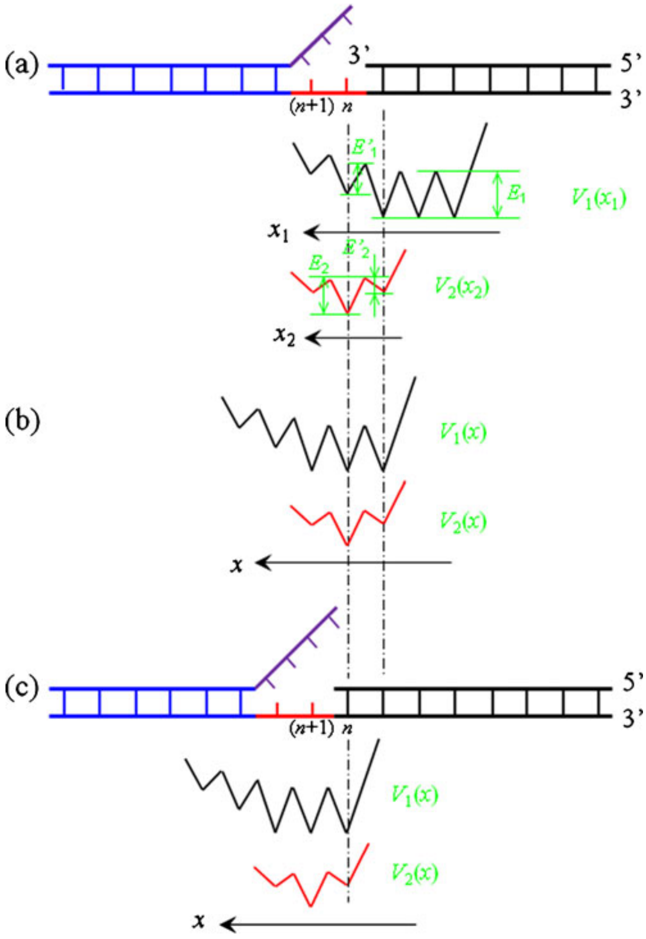


Fig. 2 Interaction potentials of polymerase with DNA substrate at nonspecific sequence (see text for detailed description). (**a**) Potentials $V_1(x_1)$ and $V_2(x_2)$ as a function of different coordinates on the polymerase. The potentials represent forms at the moment after the incorporation of the (n-1)th base but before the incorporation of the nth base, with the corresponding DNA substrate shown in the top of (**a**). (**b**) The potentials $V_1(x)$ and $V_2(x)$ shown in (**a**) as a function of the fixed coordinate x on the polymerase. (**c**) The potentials $V_1(x)$ and $V_2(x)$ as a function of the fixed coordinate x on the polymerase after the incorporation of the nth base but before the incorporation of the (n+1) th base, with the corresponding DNA substrate shown in the top of (**c**)

Considering that the residues interacting with the ssDNA template on the fingers subdomain covers N_2 bases, the interaction potential $V_2(x_2)$ can be approximately shown in Fig. 2a (bottom). Here E_2 is the binding affinity for all N unpaired bases on the ssDNA template ($N<N_2$), E'_2 is the binding affinity for only (N-1) unpaired bases, and x_2 represents the coordinate of the point on the ssDNA-binding residues that is nearest to the active site along the x_2 direction. The point is coincident with the active site along the x_2 direction. Note that different points on the ssDNA-binding residues should have different binding affinities for the unpaired base (see Supporting information).

[2] As it is known, the dynamics for a Brownian particle to escape from one potential well depends mainly on the well depth of the potential while it is insensitive to the form of the potential (see. e.g., Ref. [12]). Thus, the calculated results presented in this work depend mainly on values of the well depth of the potential while forms of the potential are not important.

In our analysis, we represent the position, x, of the polymerase by that of its active site along the template. Then, after the incorporation of the nucleotide complementary to the (n-1)th base but before the incorporation of the nucleotide complementary to the nth base, the potential $V_1(x)$ is obtained by shifting $V_1(x_1)$ toward the x direction by L_1=mp, as shown in Fig. 2b (top), while the potentials $V_2(x)$ is in the same position of $V_2(x_2)$ (bottom of Fig. 2b). After the incorporation of the nucleotide complementary to the nth base but before the incorporation of the nucleotide complementary to the (n+1)th base, $V_1(x)$ and $V_2(x)$ become those shown in Fig. 2c.

In the nonspecific sequence, the movement of the polymerase along the x direction in the over-damped environment can be described by the following Langevin equation

$$\Gamma\frac{dx}{dt}=-\frac{\partial V(x)}{\partial x}-F_U+\xi(t), \tag{1}$$

where $V(x)$=$V_1(x)$+$V_2(x)$, with $V_1(x)$ and $V_2(x)$ being shown in Fig. 2b after the incorporation of the nucleotide complementary to the (n-1)th base but before the incorporation of the nucleotide complementary to the nth base. F_U is the force resulting from the unwinding of the (n+2)th base pair as shown in the top of Fig. 2a, which is approximately calculated by $F_U=E_{bp}/p$, where E_{bp} is the free energy change required to unwind one base pair. Using parameters for the nearest-neighboring thermodynamic model for DNA-DNA duplex stability [33, 41], it is estimated that the mean free energy change is about E_{bp}=3k_BT, which gives F_U=36.3 pN. Γ is the frictional drag coefficient on the polymerase and $\xi(t)$ is the fluctuating Langevin force with $\langle\xi(t)\rangle=0$ and $\langle\xi(t)\xi(t')\rangle=2k_BT\Gamma\delta(t-t')$. The drag coefficient $\Gamma=6\pi\eta R=9.4\times10^{-11}\mathrm{kg}\cdot\mathrm{s}^{-1}$, where the viscosity of the aqueous medium is η=0.01 g · cm^{-1} · s^{-1} and the polymerase is considered as a sphere with radius R=5 nm. The Fokker-Planck equation corresponding to Langevin equation (Eq. 1) has the form

$$\frac{\partial P(x,t)}{\partial t}=\frac{1}{\Gamma}\frac{\partial}{\partial x}\left[\frac{\partial(V(x)+F_Ux)}{\partial x}P(x,t)\right]+D\times\frac{\partial^2P(x,t)}{\partial^2x}, \tag{2}$$

where $P(x,t)$ represents the probability of the polymerase positioned at x at time t and $D=k_BT/\Gamma$.

From Eq. 2, the mean moving time T_m, i.e., the mean first-passage time, for the polymerase to move from the (n-1)th site (Fig. 2a) at position x=0 to the next nth site at position x=p=2l=0.34 nm can be calculated by [12]

$$T_m=\frac{1}{D}\int_0^{2l}\exp\left[\frac{(V(y)+F_Uy)}{\Gamma D}\right]dy\int_0^{y}\exp\left[-\frac{(V(z)+F_Uz)}{\Gamma D}\right]dz. \tag{3}$$

From Eq. 3, the mean moving time T_m is obtained as follows

$$\begin{aligned}T_m=&\frac{l^2\Gamma k_BT}{(E'+F_Ul)}\cdot\left[\exp\left(\frac{E'+F_Ul}{k_BT}\right)-1\right]-\frac{l^2\Gamma}{E'+F_Ul}+\frac{l^2\Gamma k_BT}{(E'+F_Ul)(E-F_Ul)}\\&\cdot\left[\exp\left(\frac{E'+F_Ul}{k_BT}\right)-\exp\left(\frac{E'-E+2F_Ul}{k_BT}\right)\right]\cdot\left[1-\exp\left(-\frac{E'+F_Ul}{k_BT}\right)\left(1+\frac{E'+F_Ul}{E-F_Ul}\right)\right]\\&+\frac{l^2\Gamma}{E-F_Ul},\end{aligned} \tag{4}$$

where $E'=E_1+E_2'$ and $E=E_1+E_2$ (see Fig. 2b). Note that, for the single stranded primer extension synthesis, the mean moving time T_m is still calculated by using Eq. 4 but with F_U=0.

Results

Moving time

From Eq. 4 it is noted that the mean moving time T_m is insensitive to the value of E, which can be seen from Fig. 3a, where we show the results of T_m versus E for a fixed value of E*'*. The results of T_m versus E*'* with a fixed value of E are shown in Fig. 3b. From Fig. 3b, it is seen that, although the strand displacement synthesis and the single stranded primer extension synthesis have substantial differences in their energetics, the difference of the mean moving time T_m between the two cases is not very large. Even for a very large value of E*'*=20k_BT (see Supporting information), which is equivalent to an equilibrium dissociation constant K$_\mathrm{d}\approx$ 2 nM, the mean moving time T_m is only 5.96 ms and 1.46 ms for the strand displacement synthesis and single stranded primer extension synthesis,

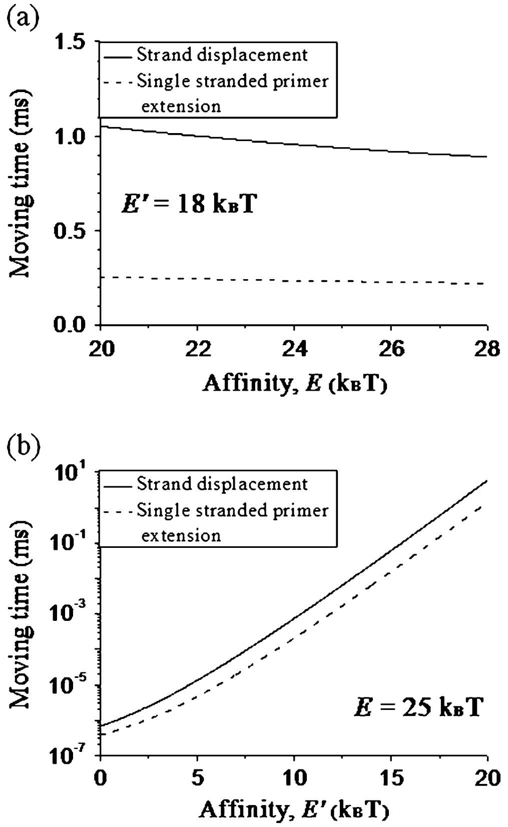


Fig. 3 Calculated results of mean moving time T_m at nonspecific sequence of dsDNA. (**a**) T_m versus $E(E = E_1 + E_2)$ for a fixed value of $E'(E' = E_1 + E_2')$. (**b**) T_m versus E' with a fixed value of E

respectively. These values of T_m are much smaller than the mean dwell time, $T_d \approx 71.43$ ms, between two successive nucleotide incorporations, which is obtained from the incorporation rate of about 14 s^{-1} for polymerase I at saturating concentration of dNTP [35]. This implies that the incorporation rate is mainly determined by the chemical reaction rate of the nucleotide incorporation or phosphoryl transfer rather than the moving time. Thus, it is expected that the incorporation rate for the strand displacement synthesis is nearly consistent with that for the single stranded primer extension synthesis, because after the polymerase is moved to the potential well of deeper depth E (i.e., the nth site at Fig. 2a) the polymerase has the same chemical reaction rate at saturating concentration of dNTP for the two cases. This is in agreement with the experimental data [3, 35, 36].

Sequence-dependent pausing

At specific sequences of dsDNA, besides the presence of the interactions with potentials $V_1(x)$ and $V_2(x)$, another interaction between the polymerase and dsDNA is present. This site-specific interaction potential $V_3(x)$ is shown in Fig. 4a,

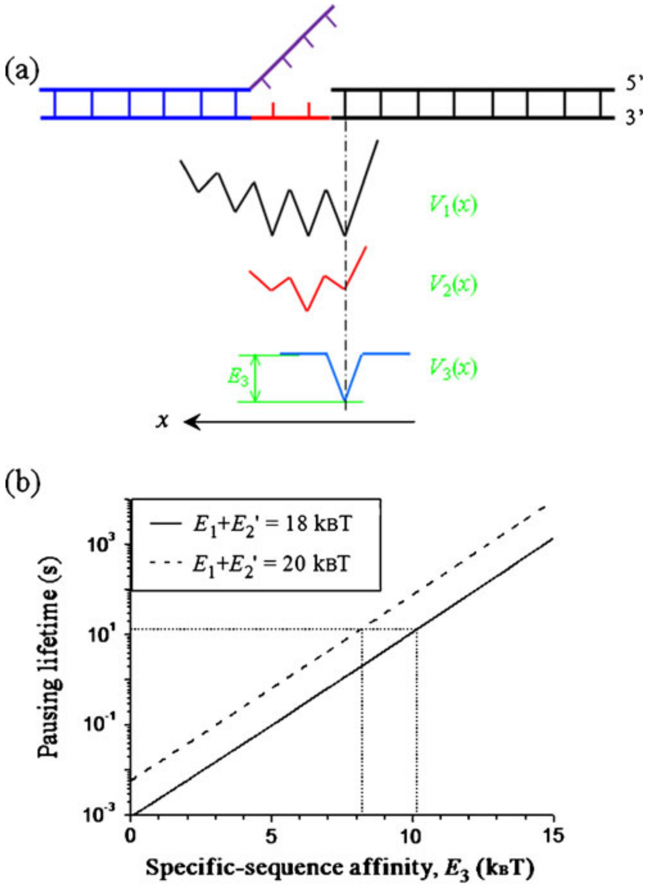


Fig. 4 Calculated results of mean moving time (or mean pausing lifetime) T_m at specific sequences of dsDNA. (**a**) Potentials $V_1(x)$, $V_2(x)$ and $V_3(x)$ as a function of the fixed coordinate x on the polymerase. (**b**) T_m versus E_3 for two different values of $E_1 + E_2'$

with E_3 being the binding affinity. Since the site-specific interaction is mainly driven by the short-ranged hydrogen bonding, we take the interaction distance equal to p. The movement of the polymerase along the x direction is still described by Langevin Eq. 1 but with $V(x) = V_1(x) + V_2(x)$ being replaced by $V(x) = V_1(x) + V_2(x) + V_3(x)$. Thus the mean moving time T_m at specific sequences is still calculated by Eq. 4 but with $E' = E_1 + E_2' + E_3$ and $E = E_1 + E_2$.

The results of T_m versus E_3 for two values of $E_1 + E_2'$ are shown in Fig. 4b. It is seen that, for a given value of $E_1 + E_2'$, T_m increases exponentially with the increase of E_3. The curve of T_m versus E_3 for $E_1 + E_2' = 18k_BT$ can be obtained by shifting the curve for $E_1 + E_2' = 20k_BT$ along the horizontal axis by $2k_BT$ ($20k_BT$ - $18k_BT$). From the figure it is also seen that, for $E_1 + E_2' = 18k_BT(20k_BT)$, at E_3=$10.18k_BT$ ($8.18k_BT$), T_m=13.2 s, which is consistent with the measured value of polymerase I for the sample DNA substrate at 23 °C in Schwartz and Quake [35]. Note that the experiment data showed that the pausing position occurs at +15 or +16 bp (see Fig. S1 in Supporting information), implying that the specific sequence corresponds to several downstream base pairs starting from +16 or +17 bp. Our above calculated results indicate that the total affinity of the polymerase I for the sample dsDNA substrate at the specific sequence is $E_1 + E_2' + E_3 = 28.18k_BT$ (equivalent to $K_d\approx5.8\times10^{-4}$nM) and the binding affinity of the fingers subdomain for the specific sequence of the sample dsDNA substrate is around $10k_BT$ (equivalent to $K_d\approx45$ μM).[3] It is interesting to note that the value of mean moving time T_m=13.2 s at the specific sequence is much larger than the mean dwell time $T_d\approx$ 71.43 ms. Thus, the sequence-dependent pausing lifetime is mainly determined by the moving time T_m rather than the nucleotide-incorporation time, in contrast to the case at the nonspecific sequence, as shown in the above section.

Moreover, it is seen from Fig. 4b that, for $E_1 + E_2' = 18k_BT(20k_BT)$, at E_3=10k_BT (8k_BT), T_m=11.1 s, which is consistent with the measured value for the sample DNA substrate at 37 °C in Schwartz and Quake [35]. This indicates that the increase of the heat from 23 °C to 37 °C reduces the total affinity for the sample dsDNA substrate at the specific sequence from 28.18k_BT (equivalent to $K_d\approx$ 5.8×10^{-4}nM) to 28k_BT (equivalent to $K_d\approx6.9\times10^{-4}$nM) or reduces the site-specific binding affinity by only 0.18k_BT, implying that the increase of the heat melts the dsDNA slightly. Similarly, from the comparison of the theoretical results (Fig. 4b) with the experimental data [35], we obtain that the addition of 1 M betaine at 23 °C reduces the total affinity at the specific sequence from 28.18k_BT (equivalent to $K_d\approx5.8\times10^{-4}$nM) to 27.8$k_BT$ (equivalent to $K_d\approx8.5\times10^{-4}$nM) that gives T_m=9.2 s. In other words, the addition of 1 M betaine reduces the site-specific binding affinity by 28.18k_BT-27.8k_BT=0.38k_BT. Betaine is a zwitteronic osmoprotectant that has been found to alter dsDNA stability so that GC-rich regions melt at temperatures more similar to AT-rich regions [31]. This implies that the addition of 1 M betaine melts the dsDNA slightly at GC-rich sequences. The melting effect by adding 1 M betaine at 23 °C is stronger than by heating the solution to 37 °C but without the addition of betaine.

In addition, it is expected that different specific sequences should have slightly different binding affinities. From Fig. 4b, it is seen that, at $E_1 + E_2' + E_3 = 27.62k_BT$ (equivalent to $K_d\approx$ 1×10^{-3}nM), T_m=7.7 s, which is consistent with the measured value for the controlled DNA substrate at 23 °C [35]. This indicates that the site-specific binding affinity difference between the sampled and the controlled DNA substrates (see Fig. S1 in Supporting information) is about 0.56k_BT. Furthermore, it is expected that the pausing efficiency (or probability) for the specific sequence with a small E_3 that gives a short mean lifetime should be smaller than that with a large E_3 that gives a long mean lifetime. This is also consistent with the experimental data for polymerase I [35].

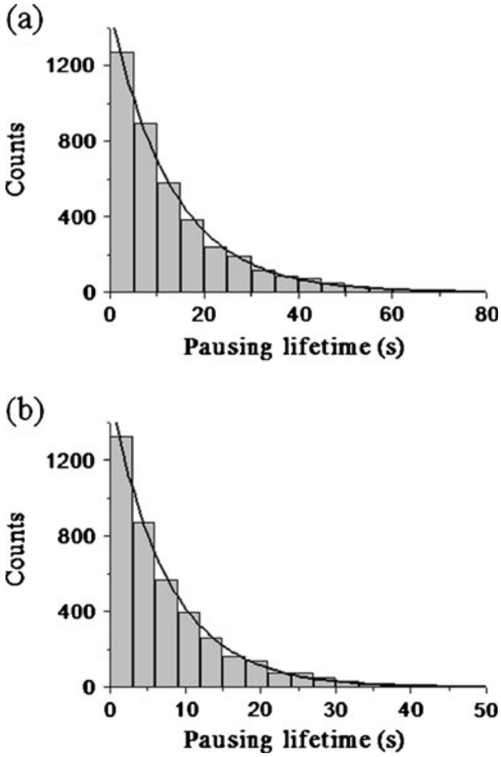


Fig. 5 Calculated results for distributions of pausing lifetime at specific sequences of dsDNA. Lines are fittings to $C\exp(-t/\tau)$, where C is constant. (**a**) $E' = E_1 + E_2' + E_3 = 28.18k_BT$ and τ= 13.2 s. (b) E'=27.62k_BT and τ=7.7 s

To see the distribution of pausing lifetimes, we resort to the numerical solution of Eq. 1, with the Stochastic Runge-Kutta

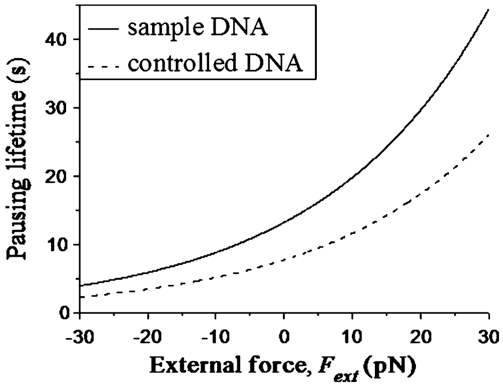


Fig. 6 Predicted results of mean pausing lifetime T_m at specific sequences of dsDNA as a function of external force F_{ext} acting on polymerase

[3] Besides different specific dsDNA sequences for the fingers subdomain of polymerase I and for Mrf-2, the binding affinities of the two proteins for their specific sequences are also very different, with an equilibrium dissociation constant $K_d\approx10$ nM for Mrf-2 bound to its target DNA sequence [49].

algorithm as used elsewhere [44, 48]. In Figs. 5(a) and (b) we show the calculated distributions of pausing lifetime for $E' = E_1 + E_2^{'} + E_3 = 28.18k_BT$ (equivalent to $K_d \approx 5.8 \times 10^{-4}$nM) and $27.62k_BT$ (equivalent to $K_d \approx 1 \times 10^{-3}$nM), respectively. It is seen that the lifetime distribution for a given E' has the single-exponential form, which is in agreement with the experimental results of Schwartz and Quake [35].

Next, we give some predicted results on sequence-dependent pausing lifetimes at saturating concentration of dNTP. Consider an external force, F_{ext}, acting on the polymerase for the fixed DNA substrates with the same specific sequences as used in Schwartz and Quake [35]. Here it is defined that the backward force has the positive value while the forward force has the negative value. Then, in Eq. 4, F_U is replaced by F_U+F_{ext}. As shown above, at 23 °C, we have $E' = E_1 + E_2^{'} + E_3 = 28.18k_BT$ (equivalent to $K_d \approx 5.8 \times 10^{-4}$nM) for the specific sequence of the sample DNA substrate, which gives T_m=13.2 s, and E'=27.62k_BT (equivalent to $K_d \approx 1 \times 10^{-3}$nM) for that of the controlled DNA substrate, which gives T_m=7.7 s. The results of T_m versus F_{ext} for E'=28.18k_BT and E'=27.62k_BT are shown in Fig. 6. It is seen that the mean pausing lifetime increases with the increase of the backward force while decreases with the increase of the forward force. For example, under the backward force of 20 pN, the mean pausing lifetimes increase to 29.6 s and 17.4 s for the sample and controlled DNA substrates, respectively; while under the forward force of 20 pN, the mean pausing lifetimes decrease to 2.9 s and 3.7 s for the sample and controlled DNA substrates, respectively.

Discussion

In previous works [43, 45], we have presented the model for translocation of replicative DNA polymerases along the DNA template during single stranded primer extension synthesis. In the current work, we modify the previous model to describe the strand displacement synthesis for polymerases such as polymerase I. In the present model, the downstream base pair is unwound by the O1-helix present in the fingers subdomain as the polymerase moves forward. The sequence-dependent pausing is considered to be due to a binding affinity of the fingers subdomain for the specific sequences of dsDNA downstream of the active site. With the model, the available experimental results on the sequence-dependent pausing dynamics such as the mean pausing lifetimes and the distribution of the pausing lifetime are well explained. Moreover, due to the very high binding affinity $E' = E_1 + E_2^{'} + E_3$ at specific sequences, the polymerase has a very small probability to detach from the DNA substrate. Thus, the polymerase can pause at the specific sequences for a very long time (e.g., 13.2 s) without detaching, which is also in agreement with the experimental data. A speculation for the biological function of the sequence-dependent pausing during strand displacement replication might imagine that it is purposed to increase the probability to cleave the displaced non-template strand by the 5' nuclease domain [48].

Similar sequence-dependent pausing behavior has also been observed for other enzymes such as RNA polymerase [13] and lambda exonuclease (λ-exo) [30]. It is thus expected that similar pausing behaviors for the DNA polymerase, RNA polymerase and λ-exo may share the same mechanism, i.e., the short pauses result from the sequence-dependent binding affinities for DNA substrates [44, 46]. Moreover, different dsDNA sequences should have slightly different binding affinities, resulting in different short pausing lifetimes.

It is argued here that the polymerases capable of strand displacement replication have the binding affinity for specific sequences on dsDNA downstream of the active site. It is interesting to note that the X-family polymerases, which are mainly involved in base excision repair and repair of double-strand breaks, also have the binding affinity for the downstream dsDNA [34, 47]. However, the two types of the polymerases show the following distinctions. (i) The former polymerase only shows the binding affinity for specific sequences of the downstream dsDNA, while it shows no or very weak binding affinity for the nonspecific sequence of the downstream dsDNA. The latter polymerase always shows the binding affinity for the downstream dsDNA, independent of the sequence. (ii) For the former polymerase, the sequence-dependent interaction is via the fingers subdomain. For the latter polymerase, the sequence-independent interaction is via another domain – the 8-kDa domain.

Finally, we mention that, to verify the present model, it is hoped to test the following predictions: (i) the about $10k_BT$ difference in the binding affinity (equivalent to $K_d \approx 45$ μM) of the polymerase I for DNA substrate used in Schwartz and Quake [35] at the specific sequence and that at the nonspecific sequence of downstream dsDNA difference, (ii) the dependence of the mean sequence-dependent pausing lifetime on the external force (see Fig. 6).

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grant No. 10974248).

References

1. Bedinger P, Munn M, Alberts BM (1989) Sequence-specific pausing during in vitro DNA replication on double-stranded DNA templates. J Biol Chem 264:16880–16886

2. Cha RS, Kleckner N (2002) ATR homolog Mec1 promotes fork progression, thus averting breaks in replication slow zones. Science 297:602–606
3. Dahlberg ME, Benkovic SJ (1991) Kinetic mechanism of DNA polymerase I (Klenow fragment): Identification of a second conformational change and evaluation of the internal equilibrium constant. Biochemistry 30:4835–4843
4. Dalgaard JZ, Klar AJS (2000) swi1 and swi3 perform imprinting, pausing, and termination of DNA replication in S pombe. Cell 102:745–751
5. Datta K, Wowor AJ, Richard AJ, LiCata VJ (2006) Temperature dependence and thermodynamics of Klenow polymerase binding to primed-template DNA. Biophys J 90:1739–1751
6. Delagoutte E, von Hippel PH (2003) Function and assembly of the Bacteriophage T4 DNA replication complex. J Biol Chem 278:25435–25447
7. Doublie S, Ellenberger T (1998) The mechanism of action of T7 DNA polymerase. Curr Opin Struct Biol 8:704–712
8. Doublie S, Sawaya MR, Ellenberger T (1999) An open and closed case for all polymerases. Structure 7:R31–R35
9. Doublie S, Tabor S, Long AM, Richardson CC, Ellenberger T (1998) Crystal structure of a bacteriophage T7 DNA replication complex at 22Å resolution. Nature 391:251–258
10. Eom SH, Wang J, Steitz TA (1996) Structure of Taq polymerase with DNA at the polymerase active site. Nature 382:278–281
11. Franklin MC, Wang J, Steitz TA (2001) Structure of the replicating complex of a Pol α family DNA polymerase. Cell 105:657–667
12. Gardiner CW (1983) Handbook of stochastic methods for physics, chemistry and the natural sciences. Springer, Berlin
13. Herbert KM, La Porta A, Wong BJ, Mooney RA, Neuman KC, Landick R, Block SM (2006) Sequence-resolved detection of pausing by single RNA polymerase molecules. Cell 125:1083–1094
14. Johnson KA (1993) Conformational coupling in DNA polymerase fidelity. Annu Rev Biochem 62:685–713
15. Joyce CM, Potapova O, DeLucia AM, Huang X, Basu VP, Grindley NDF (2008) Fingers-closing and other rapid conformational changes in DNA polymerase I (Klenow fragment) and their role in nucleotide selectivity. Biochemistr 47:6103–6116
16. Kang S, Ohshima K, Shimizu M, Amirhaeri S, Wells RD (1995) Pausing of DNA synthesis in vitro at specific loci in CTG and CGG triplet repeats from human hereditary disease genes. J Biol Chem 270:27014–27021
17. Kiefer JR, Mao C, Braman JC, Beese LS (1998) Visualizing DNA replication in a catalytically active Bacillus DNA polymerase crystal. Nature 391:304–307
18. Kiefer JR, Mao C, Hansen CJ, Basehore SL, Hogrefe HH, Braman JC, Beese LS (1997) Crystal structure of a thermostable Bacillus DNA polymerase I large fragment at 21Å resolution. Structure 5:95–108
19. Kim Y, Eom SH, Wang J, Lee DS, Suh SW, Steitz TA (1995) Crystal structure of Thermus aquaticus DNA polymerase. Nature 376:612–616
20. Kornberg A, Baker T (1992) DNA replication, 2nd edn. Freeman, New York
21. Korolev S, Nayal M, Barnes WM, Di Cera E, Waksman G (1995) Crystal structure of the large fragment of Thermus aquaticus DNA polymerase I at 25-Å resolution: structural basis for thermostability. Proc Natl Acad Sci USA 92:9264–9268
22. Krasilnikova MM, Samadashwily GM, Krasilnikov AS, Mirkin SM (1998) Transcription through a simple DNA repeat blocks replication elongation. EMBO J 17:5095–5102
23. LaDuca RJ, Fay PJ, Chuang C, McHenry CS, Bambara RA (1983) Site-specific pausing of deoxyribonucleic acid synthesis catalyzed by 4 forms of Escherichia coli DNA polymerase III. Biochemistry 22:5177–5188
24. Lemoine FJ, Degtyareva NP, Lobachev K, Petes TD (2005) Chromosomal translocations in yeast induced by low levels of DNA polymerase: A model for chromosome fragile sites. Cell 120:587–598
25. Li Y, Korolev S, Waksman G (1998) Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of thermus aquaticus DNA polymerase I: structural basis for nucleotide incorporation. EMBO J 17:7514–7525
26. Li Y, Mitaxov V, Waksman G (1999) Structure-based design of Taq DNA polymerases with improved properties of dideoxynucleotide incorporation. Proc Natl Acad Sci USA 96:9491–9496
27. Lyamichev V, Brow MA, Dahlberg JE (1993) Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases. Science 260:778–783
28. Maier B, Bensimon D, Croquette V (2000) Replication by a single DNA polymerase of a stretched single-stranded DNA. Proc Natl Acad Sci USA 97:12002–12007
29. Mytelka DS, Chamberlin MJ (1996) Analysis and suppression of DNApolymerase pauses associated with a trinucleotide consensus. Nucl Acids Res 24:2774–2781
30. Perkins TT, Dalal RV, Mitsis PG, Block SM (2003) Sequence-dependent pausing of single lambda exonuclease molecules. Science 301:1914–1718
31. Rees WA, Yager TD, Korte J, von Hippel PH (1993) Betaine can eliminate the base pair composition dependence of DNA melting. Biochemistry 32:137–144
32. Samadashwily GM, Raca G, Mirkin SM (1997) Trinucleotide repeats affect DNA replication in vivo. Nat Genet 17:298–304
33. SantaLucia J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl Acad Sci USA 95:1460–1465
34. Sawaya MR, Prasad R, Wilson SH, Kraut J, Pelletier H (1997) Crystal structures of human DNA polymerase β complexed with gapped and nicked DNA: evidence for an induced fit mechanism. Biochemistry 36:11205–11215
35. Schwartz JJ, Quake SR (2009) Single molecule measurement of the “speed limit” of DNA polymerase. Proc Natl Acad Sci USA 106:20294–20299
36. Singh K, Srivastava A, Patel SS, Modak MJ (2007) Participation of the fingers subdomain of Escherichia coli DNA polymerase I in the strand displacement synthesis of DNA. J Biol Chem 282:10594–10604
37. Smolina IV, Demidov VV, Frank-Kamenetskii MD (2003) Pausing of DNA polymerases on duplex DNA templates due to ligand binding in vitro. J Mol Biol 326:1113–1125
38. Steitz TA, Yin YW (2004) Accuracy, lesion bypass, strand displacement and translocation by DNA polymerases. Phil Trans R Soc Lond B 359:17–23
39. Takeuchi Y, Horiuchi T, Kobayashi T (2003) Transcription-dependent recombination and the role of fork collision in yeast rDNA. Genes Dev 17:1497–1506
40. Turner RM, Grindley NDF, Joyce CM (2003) Interaction of DNA polymerase I (Klenow fragment) with the single-stranded template beyond the site of synthesis. Biochemistry 42:2373–2385
41. Wu P, Nakano S, Sugimoto N (2002) Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation. Eur J Biochem 269:2821–2830
42. Wuite GJL, Smith SB, Young M, Keller D, Bustamante C (2000) Single-molecule studies of the effect of template tension on T7 DNA polymerase activity. Nature 404:103–106

43. Xie P (2007) Model for forward polymerization and switching transition between polymerase and exonuclease sites by DNA polymerase molecular motors. Arch Biochem Biophys 457:73–84
44. Xie P (2008) A dynamic model for transcription elongation and sequence-dependent short pauses by RNA Polymerase. Biosystems 93:199–210
45. Xie P (2009) A possible mechanism for the dynamics of transition between polymerase and exonuclease sites in a high-fidelity DNA polymerase. J Theor Biol 259:434–439
46. Xie P (2009) Molecular motors that digest their track to rectify Brownian motion: processive movement of exonuclease enzymes. J Phys Condens Matter 21:375108
47. Xie P (2011) A model for the dynamics of mammalian family X DNA polymerases. J Theor Biol 277:111–122
48. Xie P, Sayers JR (2011) A model for transition of 5′-nuclease domain of DNA polymerase I from inert to active modes. PLoS One 6:e16213
49. Yuan YC, Whitson RH, Liu Q, Itakura K, Chen Y (1998) A novel DNA-binding motif shares structural homology to DNA replication and repair nucleases and polymerases. Nat Struct Biol 5:959–964

J Mol Model (2012) 18:1961–1968
DOI 10.1007/s00894-011-1225-y

ORIGINAL PAPER

Metal–metal interactions in linear tri-, penta-, hepta-, and nona-nuclear ruthenium string complexes

Mika Niskanen · Pipsa Hirva · Matti Haukka

Received: 21 March 2011 / Accepted: 11 August 2011 / Published online: 26 August 2011

Abstract Density functional theory (DFT) methodology was used to examine the structural properties of linear metal string complexes: $[Ru_3(dpa)_4X_2]$ (X = Cl^-, CN^-, NCS^-, dpa = dipyridylamine$^-$), $[Ru_5(tpda)_4Cl_2]$, and hypothetical, not yet synthesized complexes $[Ru_7(tpta)_4Cl_2]$ and $[Ru_9(ppta)_4Cl_2]$ (tpda = tri-α-pyridyldiamine^{2-}, tpta = tetra-α-pyridyltriamine^{3-}, ppta = penta-α-pyridyltetraamine^{4-}). Our specific focus was on the two longest structures and on comparison of the string complexes and unsupported ruthenium backboned chain complexes, which have weaker ruthenium–ruthenium interactions. The electronic structures were studied with the aid of visualized frontier molecular orbitals, and Bader's quantum theory of atoms in molecules (QTAIM) was used to study the interactions between ruthenium atoms. The electron density was found to be highest and distributed most evenly between the ruthenium atoms in the hypothetical $[Ru_7(tpta)_4Cl_2]$ and $[Ru_9(ppta)_4Cl_2]$ string complexes.

Keywords Ruthenium · Density functional theory · Quantum theory of atoms in molecules · Extended metal atom chain · Linear metal string complex

Introduction

Structures that incorporate one-dimensional metal atom chains attract interest for a variety of reasons, such as their conductivity [1, 2], luminescence [3–5], vapochromism [6–8], and magnetic [9–11], and catalytical [12–15] properties. Some of these properties are linked directly to the interacting metal atoms in the metal chain, while others can be attributed to metal–ligand interactions of single units in the chain.

One-dimensional metal atom chains are found in many different kinds of structures. Square planar metal complexes can form stacks where the metal atoms are lined up, as is the case with Magnus' Green salt, Vauquelin's salt, Krogmann's salt, and derivatives thereof [16–18]. $[Ru(CO)_4]_n$, $[Ru(bpy)(CO)_2]_n$, and $\{[Rh(MeCN)_4](BF_4)_{1.5}]\}_n$ are examples of polymeric unsupported chains, where the chain is formed without the aid of supporting ligands [19–21]. Platinum, iridium, and rhodium blues are formed from ligand-supported dimers that combine into tetranuclear or longer chain structures [22]. Finally, in extended metal atom chains (EMACs), which are also known as linear metal string complexes, the metal chains are formed with the aid of surrounding ligands [23].

In a typical metal string complex, the linear chain of transition metals is surrounded and supported by four oligo-α-pyridylamine^{n-} ligands [24] in a helix (Fig. 1). Other surrounding ligands, such as the oligo-α-naphthyridylpyridylamine^{n-} ligands that result in mixed-valence complexes [25, 26] can be used as well. Axial ligands (X in Fig. 1a) vary from small anions, such as Cl^-, CN^-, and NCS^- to larger arylacetylide ligands [27]. Overall, the versatile ligand structure provides an easy tool for tuning the structures. The benefits of the metal string complexes include the relative ease with which

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1225-y) contains supplementary material, which is available to authorized users.

M. Niskanen · P. Hirva (✉) · M. Haukka
Department of Chemistry, University of Eastern Finland,
PO Box 111, 80101 Joensuu, Finland
e-mail: pipsa.hirva@uef.fi

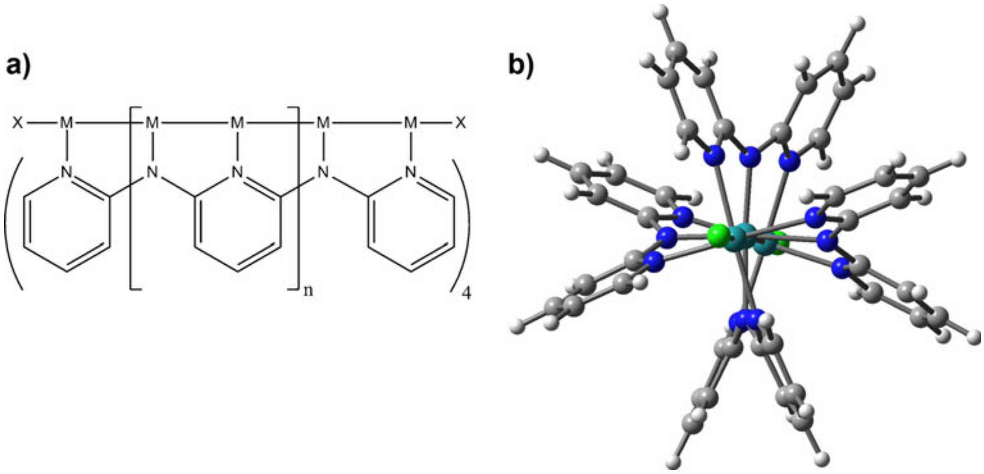


Fig. 1 **a** Schematic illustration of a typical metal string complex. **b** Structure of Ru_3(dpa)$_4$$Cl_2$, showing the helical ligands

different transition metals can be trapped inside the helical surrounding ligands, and also the short metal–metal distance that is forced by these same ligands. The challenges posed by metal string complexes lie in the synthesis of longer complexes. Metal string complexes up to a size of nine metal atoms have been synthesized with first row transition metals [28], and up to five metal atoms with second row transition metal atoms [29, 30]. Metal string complexes provide a promising approach for the creation of nano-scale electrical wires [1]. The electric conductivity of metal string complexes correlates qualitatively with the strength of the metal–metal interaction in the complex, and can be tuned by the choice of transition metal, its oxidation state, and the ligands used [27]. Metal string complexes also have interesting magnetic properties, such as the spin-crossover behavior of $[Co_3(dpa)_4Cl_2]^+$ [9, 31].

Studies of ruthenium EMACs began from the synthesis and structure of a triruthenium metal string complex published in 1996 [32]. Later, a synthesis with a better yield [33] and triruthenium metal string complexes with different axial ligands [27, 33, 34] were reported. Moreover, it was noted that the terminal ligands affected the electronic configurations on triruthenium metal string complexes, causing, for example, a singlet state in $[Ru_3(dpa)_4Cl_2]$ and a triplet state in $[Ru_3(dpa)_4(CN)_2]$ [33]. The first pentaruthenium metal string complex was reported in 2008 [30].

Bader's quantum theory of atoms in molecules (QTAIM) [35] has seen increasing use in bonding studies of various molecules. The method is based on a topological study of the electron density, ρ(r), where a search is made for critical points. The critical points are located where the gradient of electron density, $\nabla\rho$(r), vanishes, and are classified as (3,−3) critical points, i.e., the local maxima of electron density found in nuclear positions; (3,−1) critical points, i.e., the bond critical points that are saddle points of the electron density between the nuclei that share an interatomic surface; and subsequently into (3,+1) ring critical points and (3,+3) cage critical points.

Information about bonding is obtained through the properties pertaining at the bond critical points (bcp). Commonly reported values are the electron density and its Laplacian. A negative Laplacian means that the charge is concentrated locally and is typical in shared shell interactions, whereas a positive Laplacian suggests a locally depleted charge and is typical of closed shell interactions. The bonding can be further assessed with the aid of the potential energy density $V(r_{bcp})$ and kinetic energy density $G(r_{bcp})$ at the bcp. If $|V(r_{bcp})|/G(r_{bcp}) > 2$, the interaction is classified as a pure shared shell interaction between the nuclei; if $|V(r_{bcp})|/G(r_{bcp}) < 1$, the interaction is classified as a pure closed shell interaction; and when $1 < |V(r_{bcp})|/G(r_{bcp}) < 2$, the interaction is classified as a closed shell interaction with some covalent nature. Additional information may also be gained from the bond degree, which is the relationship between the total energy density $H(r_{bcp})$ and the electron density $\rho(r_{bcp})$ at the bond critical point, $H(r_{bcp})/\rho(r_{bcp})$ [36].

Our group has previously studied unsupported ruthenium chains, such as $[Ru(CO)_4]_n$ and $[Ru(bpy)(CO)_2]_n$ which are formed via direct Ru–Ru bonding [37, 38]. In the current study, our aim was to investigate the geometry and M–M interactions of ruthenium string complexes, especially the as yet not synthesized $[Ru_7(tpta)_4Cl_2]$ and $[Ru_9(ppta)_4Cl_2]$, and to compare the string complexes and unsupported ruthenium chains. The metal–metal interactions were examined using the QTAIM approach.

Computational methods

The calculations were conducted using the Gaussian 03 program package [39]. The density functional theory (DFT) methodology was used with the PBE0 hybrid density functional [40]. The standard all-electron basis set 6-31 G (d) was used for non-metal atoms, while Huzinaga's all-electron basis set [41] with an additional p-polarization function (433321/4331/421) was used for ruthenium.

Table 1 Selected geometrical data for $Ru_3(dpa)_4X_2$ ($X = Cl^-$, CN^-, or NCS^-). Values given are averages

	Cl^-		CN^-		NCS^-	
	Computed	Experimental [32]	Computed	Experimental [33]	Computed	Experimental [27]
Ru-Ru	2.303	2.254	2.445	2.377	2.314	2.282
Ru-X	2.498	2.596	1.995	2.057	2.103	2.240
$Ru_{terminal}$-N	2.147	2.108	2.124	2.108	2.138	2.080
$Ru_{central}$-N	2.091	2.066	2.032	2.066	2.091	2.120
angle Ru1-Ru2-Ru3	165	171	172	171	165	166
dihedral N-Ru-Ru-N	21.0	21.7	16.2	21.7	20.0	22.2

Huzinaga's all-electron basis set has been found in earlier studies to be stable and also to describe ligand-unsupported ruthenium systems reliably [37, 38, 42]. However, to verify the reliability of the AE basis set, we performed tests with a relativistic ECP basis set (LANL2DZ). The results for the structures and molecular orbitals in triruthenium and pentaruthenium complexes were similar with both basis sets (see Supplementary material), suggesting that relativistic effects do not play an important role in these ruthenium string complexes.

Where possible, we took the initial geometries of the metal string complexes under study from crystal structures and optimized the structures in C_1 symmetry. All of the complexes were calculated in singlet state in order to facilitate comparison of the different models with different lengths and axial ligands. Since the experimentally obtained spin states for Ru_3 complexes have been found to vary with the nature of the axial ligands [27, 33], we optimized the spin state for the trimetallic complexes. Triplet ground state was predicted for all the Ru_3 complexes regardless of axial ligand. However, the effect of the spin state on the properties of the metal–metal bonding and the appearance of the molecular orbitals was found to be minimal and therefore we chose to conduct all succeeding calculations for the singlet state only. Frequencies were calculated for up to the pentametal complexes to ensure that all of the optimizations yielded minima. Bonding and electron density were studied using Bader's QTAIM [35], as implemented in AIM2000 [43]. The bonding interactions were further studied by calculating Mayer bond orders [44] for selected bonds.

Results and discussion

Structures and geometry

To verify the performance of the selected modeling method in the context of ruthenium metal string complexes, the structures of $Ru_3(dpa)_4X_2$ ($X = Cl^-$, CN^-, or NCS^-) were computed. The geometries obtained were compared with the experimental X-ray data. The comparison is presented in Table 1. The computed Ru–Ru and Ru–N distances are only slightly longer than had been experimentally determined, but the bonding of axial ligands is overestimated. Some of the differences may arise from the effects of crystal packing. The trend for the Ru–Ru distances is the same in our molecular calculations as in the X-ray data, showing that Cl^- as the axial ligand causes the shortest Ru–Ru distances, while CN^- leads to the longest Ru–Ru distances. A similar trend in Ru–Ru distances was also obtained in a previous computational study of $Ru_3(dpa)_4X_2$ ($X = Cl^-$, CN^-) [33], and $X = (NCS^-)$ [27]. The longer Ru–Ru distance for $X = CN^-$ was explained as originating from the

Table 2 Ru–Ru distances for $Ru_5(tpda)_4Cl_2$, $Ru_7(tpta)_4Cl_2$, $Ru_9(hpta)_4Cl_2$ and unsupported chains $[Ru(CO)_4]_8H_2$, $Ru(bpy)(CO)_2]_8H_2$ (bpy = 2,2′-bipyridine)[a]

	$Ru_5(tpda)_4Cl_2$		$Ru_7(tpta)_4Cl_2$	$Ru_9(hpta)_4Cl_2$	$[Ru(CO)_4]_8H_2$	$[Ru(bpy)(CO)_2]_8H_2$
	Computed	Experimental [30]	Computed	Computed	Computed [38]	Computed [38]
Ru1-Ru2	2.288	2.283	2.259	2.263	2.877	2.919
Ru2-Ru3	2.253	2.276	2.230	2.250	2.848	2.878
Ru3-Ru4	-	-	2.259	2.245	2.849	2.888
Ru4-Ru5	-	-	-	2.232	2.849	2.893

[a] The atoms are labeled inwards from the terminal Ru; Ru1 is connected to the axial ligand

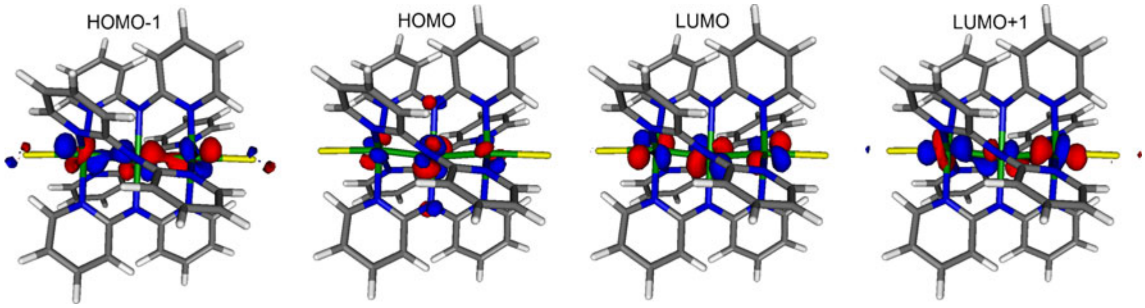


Fig. 2 Frontier orbitals of $Ru_3(dpa)_4Cl_2$. The frontier orbitals of $Ru_3(dpa)_4(NCS)_2$ are almost identical

better ability of the cyanide ligand to form a π back-bonding interaction with the ruthenium centers, therefore reducing the strength of the metal–metal interaction.

On the other hand, it is likely that the effect of axial ligands on the properties of the metal chain decreases as the chain length increases. Therefore, we chose to investigate the effect of the chain length on the geometry of the ruthenium metal string complexes by computing longer complexes with chloride axial ligands: experimentally known $Ru_5(tpda)_4Cl_2$ and hypothetical $Ru_7(tpta)_4Cl_2$ and $Ru_9(hpta)_4Cl_2$. Table 2 shows the ruthenium–ruthenium distances of the optimized complexes, and also the distances for selected optimized unsupported ruthenium chains.

For $Ru_5(tpda)_4Cl_2$, the computed and experimentally obtained Ru–Ru distances are closer than in the case of the corresponding triruthenium complex, although the Ru–Cl bonding is still overestimated. The Ru–Ru distances inside the Ru-chains are slightly shorter than at their edges. The ruthenium backbone forms a linear chain, although the terminal ruthenium atoms were displaced slightly from linearity to form a ~172° Ru–Ru–Ru angle at the ends of the complex, which can also be observed in the experimental structures of the triruthenium complexes. Moreover, the helical ligands bonded unevenly, as has also been noted with respect to the experimental X-ray structures of triruthenium complexes. Typically, two of the nitrogen atoms from two of the helical ligands had a longer Ru–N bond distance, while the other two in the different ligands had shorter Ru–N distances in alternating sequence along the chain.

Compared to the unsupported chains, the metal string complexes displayed substantially shorter Ru–Ru distances. The differences can be explained in terms of the effect of the supporting ligands and also the formal oxidation states of the ruthenium atoms, which are Ru(II) in the metal string complexes and Ru(0) in the unsupported chains.

Electronic structures

Attention was also paid to the effect of the axial ligand and chain length on the electron structure of the ruthenium string complexes. The qualitative molecular orbital (MO) and MO diagrams for idealized $Ru_3(dpa)_4Cl_2$ in D_4 symmetry have been studied previously [27, 33, 45]. However, the ruthenium atoms do not form a perfectly linear trimetal system in either the crystal structures or in our optimized geometry, as can be seen from the Ru1–Ru2–Ru3 angles in Table 1. This causes slight changes in the MOs compared to the idealized case.

Presuming the direction of the *z*-axis along the ruthenium atoms in the complex, the d_{xz} and d_{yz} orbitals are no longer equal in energy. Moreover, atomic orbital combinations resembling a combination of $d_z^2 - d_{xz} - d_z^2$ were found in the triruthenium complexes. Selected frontier orbitals for the triruthenium complexes can be seen in Figs. 2 and 3.

The frontier molecular orbitals of $Ru_5(tpda)_4Cl_2$, $Ru_7(tpta)_4Cl_2$, and $Ru_9(hpta)_4Cl_2$ were also visualized. The HOMO and LUMO for these complexes are presented in Fig. 4. In all cases, both LUMO and LUMO + 1 were

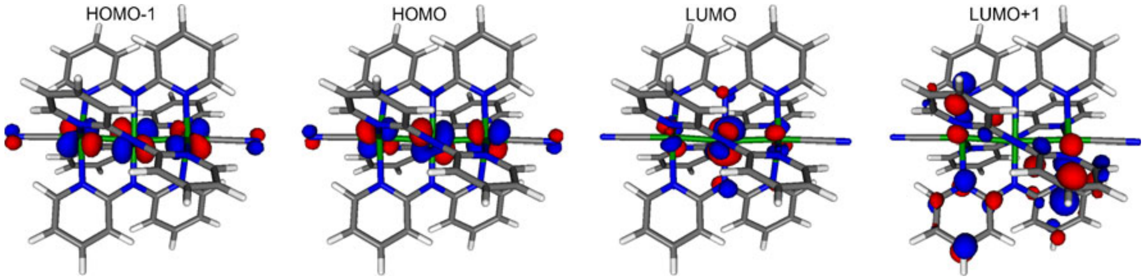


Fig. 3 Frontier orbitals of $Ru_3(dpa)_4(CN)_2$

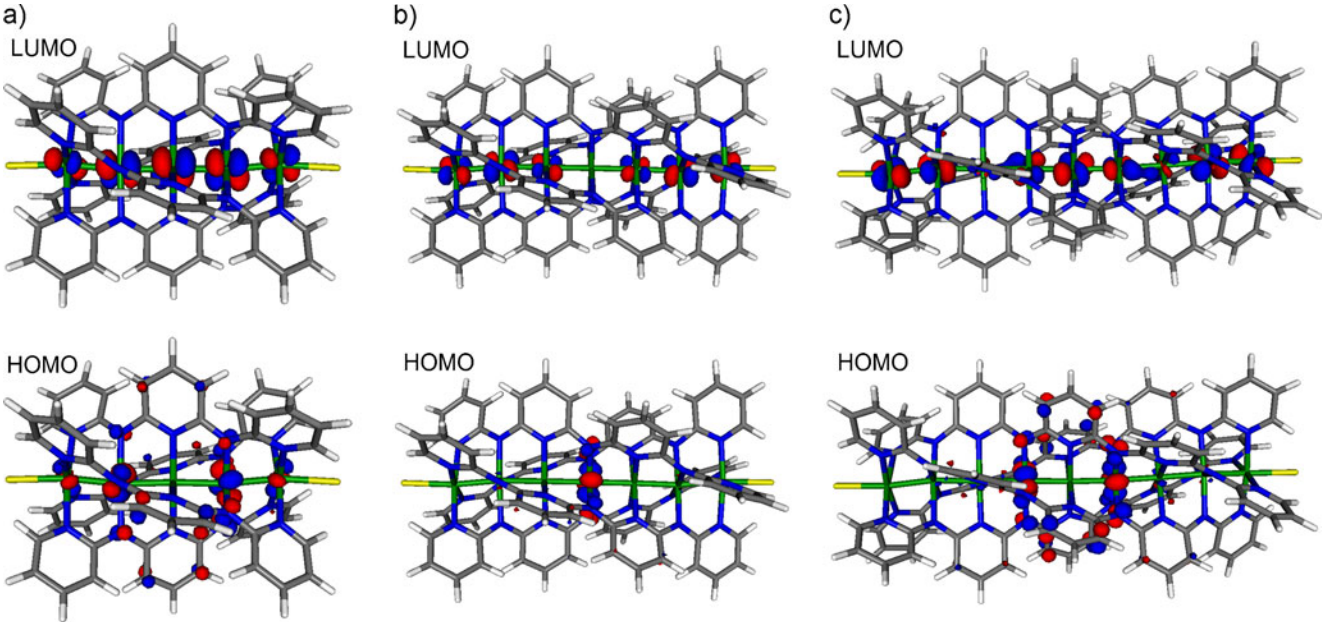


Fig. 4 The HOMO and LUMO obtained for **a** $Ru_5(tpda)_4Cl_2$, **b** $Ru_7(tpta)_4Cl_2$, and **c** $Ru_9(hpta)_4Cl_2$. Enlarged figures are available as Supplementary material

antibonding d_{xz} or d_{xy} (π^*) orbitals, while both HOMO and HOMO-1 were antibonding d_{xy} (δ^*) orbitals, similar to those in $Ru_3(dpa)_4Cl_2$. Slight differences can be observed due to alternation of the pyridyl and amine segments as the ligand part that bonds to the centermost ruthenium as the chain size increases. Differences in the nature of the frontier MOs were observed when the string complexes were compared with unsupported ruthenium chains. While the frontier MOs in string complexes exist more clearly either on metal atoms or ligands, in the case of unsupported ruthenium chains, the orbitals are often delocalized both to the ligands and to the metals (Fig. 5). Lengthening of the string complex did not have a noticeable effect on the filling order of the MOs. However, longer complexes resulted in a sharply decreasing HOMO–LUMO gap, as shown in Table 3. Compared to the unsupported chains, the HOMO–LUMO gaps were smaller in the metal string complexes.

Electron density and QTAIM studies

To gain additional information about metal–metal interactions in the string complexes, the topology of the electron density was studied using Bader's QTAIM. Bond critical points were found between all of the adjacent ruthenium atoms, and the properties of the electron density at bcps were analyzed to assess M–M interactions. Mayer bond order [45] was also calculated to aid further in the assessment. The results can be seen in Table 3.

In $Ru_3(dpa)_4X_2$ ($X = Cl^-, CN^-, NCS^-$), the Cl^- and NCS^- complexes had similar electron densities at the Ru–Ru bcps, while in the CN^- complex the Ru–Ru bonding was considerably weaker, as could be expected from the longer Ru–Ru distance. The strong π back-bonding ability of the CN^- ligand is reflected by the smaller electron density between Ru atoms in the $Ru_3(dpa)_4(CN)_2$ string complex compared to that of the Cl^- and NCS^- substituted

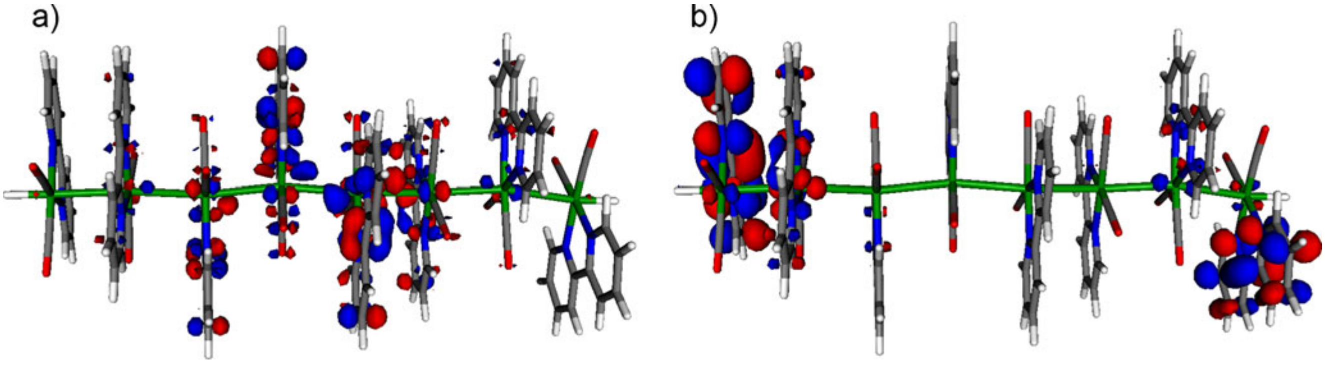


Fig. 5 **a** HOMO, **b** LUMO obtained for $[Ru(bpy)(CO)_2]_8H_2$

Table 3 Electron density at Ru–Ru bond critical points (e Å^{-3}) and HOMO–LUMO gaps[a]

Ruthenium string complex[a]	Ru1-Ru2	Ru2-Ru3	Ru3-Ru4	Ru4-Ru5	MBO[b]	HOMO-LUMO gap [eV]
$Ru_3(dpa)_4Cl_2$	0.692				1.22	1.88
$Ru_3(dpa)_4(CN)_2$	0.470				0.70	1.76
$Ru_3(dpa)_4(NCS)_2$	0.668				1.09	1.76
$Ru_5(tpda)_4Cl_2$	0.696	0.756			0.99	1.73
$Ru_7(tpta)_4Cl_2$	0.765	0.775	0.744		0.99	1.18
$Ru_9(hpta)_4Cl_2$	0.753	0.737	0.772	0.777	0.96	0.98
$[Ru(CO)_4]_8H_2$	0.265	0.287	0.287	0.287	0.70	3.75
$[Ru(bpy)(CO)_4]_8H_2$	0.241	0.271	0.269	0.266	0.65	1.52

[a] Atoms are labeled inwards from terminal Ru; Ru1 is connected to the axial ligand

[b] Average Mayer's bond order between the ruthenium atoms

complexes. This can also be seen in the substantially smaller Ru–Ru Mayer bond order with the triruthenium complex with axial CN^- ligands. In the case of the longer Ru metal string complexes, the total electron density at bcps between ruthenium atoms slightly increased and was distributed more evenly, even to terminal ruthenium and its neighbor in string complexes of seven and nine ruthenium atoms. This suggests strong metal–metal interactions along the ruthenium backbone, and also the reduction of the effect of axial ligands on the electron density between Ru atoms. Furthermore, the Mayer bond orders converge to a value around 1 with the longer chain lengths—again an indication of evenly distributed metal-metal interactions. Compared to the unsupported chains, the metal string complexes have a much higher electron density at the Ru–Ru bcps as well as higher MBO values.

At bcps, metal–metal bonds typically have a low electron density, $\rho(r_{bcp})$, a low and positive Laplacian of electron density, $\nabla^2\rho(r_{bcp})$, and negative and close to zero total energy density, $H(r_{bcp})$. The metal–metal bonds can be further assessed on the basis of the ratio of potential and kinetic energy densities ($|V(r_{bcp})|/G(r_{bcp})$), which is typically between 1 and 2 for transition metals, and also on the basis of the bond degree $H(r_{bcp})/\rho(r_{bcp})$, which indicates the strength of the interaction: a greater negative value means a stronger interaction [36].

A comparison between Ru–Ru bcps in $[Ru(bpy)(CO)_2]_8H_2$ and $Ru_7(tpta)_4Cl_2$ is presented in Table 4. The metal string complex has a higher electron density at M–M bcps and also a surprisingly high Laplacian compared to bcps of unsupported chain. We can also see that the bcps differ in $|V(r_{bcp})|/G(r_{bcp})$ and $H(r_{bcp})/\rho(r_{bcp})$, the former meaning that the string complex is topologically closer to the border between the pure closed shell and the transit closed shell regions ($|V(r_{bcp})|/G(r_{bcp}) = 1$), but still well within the transit-closed shell region. The $H(r_{bcp})/\rho(r_{bcp})$ bond degree suggests that the string complex has stronger Ru–Ru interactions. The reasons for the differences can probably be found from the different oxidation states and the bond distances of ruthenium atoms in their corresponding complexes.

It has been suggested that, in linear metal string complexes, the conductivity and bond order correlate qualitatively [1, 27, 30]. Following this idea, it can be

Table 4 Properties of electron density at selected bond critical points (bcp) in $Ru_7(tpta)_4Cl_2$ and $[Ru(bpy)(CO)_2]_8H_2$

Ruthenium string complex[a]	Distance (Å)	e-density (e Å^{-3})	Laplacian (e Å^{-5})	|λ1|/λ3 (-)	H, E_{total} density (Hartree Bohr^{-3})	|V|/G (-)	H/ρ (kJ mol^{-1})
$[Ru(bpy)(CO)_2]_8H_2$							
bcp1: Ru1-Ru2	2.919	0.2410	1.029	0.22	−0.00823	1.44	−604.7
bcp2: Ru2-Ru3	2.877	0.2709	1.037	0.25	−0.00927	1.46	−606.5
bcp3: Ru3-Ru4	2.889	0.2685	0.948	0.26	−0.00937	1.49	−618.4
bcp4: Ru4-Ru5	2.893	0.2659	0.936	0.26	−0.00929	1.49	−618.9
$Ru_7(tpta)_4Cl_2$							
bcp1: Ru1-Ru2	2.260	0.7648	9.921	0.18	−0.03142	1.23	−727.8
bcp2: Ru2-Ru3	2.229	0.7759	12.78	0.15	−0.03163	1.19	−722.3
bcp3: Ru3-Ru4	2.258	0.7456	10.67	0.17	−0.02953	1.21	−701.8

[a] Atoms are labeled inwards from terminal Ru; Ru1 is connected to the axial ligand

expected that the conductivity and electron density will also correlate. The experimental conductances of Ru_3(dpa)$_4$(CN)$_2$ and Ru_3(dpa)$_4$(NCS)$_2$ have been reported to be approximately 2.04×10^{-3} G_0 and 9.81×10^{-3} G_0, respectively ($G_0 = 2e^2/h$), which means that the CN axial ligands reduces the conductivity [27]. In the longer string complexes, the higher electron density between the metal atoms suggests good conductivity along the Ru chain. The conductivities and calculated electron density are in agreement, although additional experimental data would be needed to see how well the electron density and conductivity correlate.

Conclusions

Ruthenium metal string complexes were studied using DFT methodology and QTAIM. The experimental geometries were reproduced reliably with the methodology used. In the absence of symmetry restrictions, we observed a slight displacement of the terminal ruthenium atoms from the idealized 180° Ru–Ru–Ru angle and also uneven bonding of surrounding ligands to the ruthenium atoms. Both effects are also visible in the experimental crystal structures of triruthenium complexes. The electronic structures of the complexes under study were similar, except for Ru_3(dpa)$_4$(CN)$_2$, where the nature of HOMO and LUMO were interchanged. QTAIM studies revealed that, in the longest studied complexes, the electron density between the ruthenium atoms was both the highest and most evenly distributed, suggesting strong ruthenium–ruthenium interactions. Unsupported ruthenium backboned chain complexes, such as [Ru(bpy)(CO)$_4$]$_n$, have a much lower electron density and weaker ruthenium–ruthenium interactions than the ruthenium metal string complexes under study. The structures of hypothetical [Ru_7(tpta)$_4$$Cl_2$] and [$Ru_9$(ppta)$_4$$Cl_2$] seem feasible to synthesize, although the synthesis itself would no doubt have its difficulties.

Acknowledgment We are grateful for financial support (grant 129772) provided by the Academy of Finland.

References

1. Tsai TW, Huang QR, Peng SM, Jin BY (2010) Smallest electrical wire based on extended metal-atom chains. J Phys Chem C 114:3641–3644. doi:10.1021/jp907893q
2. Anderson BM, Hurst SK, Spangler L, Abbott EH, Martellaro P, Pinhero PJ, Peterson ES (2006) Growth and characterization of partially oxidized platinum polymers in nanoscale templates. J Mater Sci 41:4251–4258. doi:10.1007/s10853-005-5429-3
3. Heyduk AF, Krodel DJ, Meyer EE, Nocera DG (2002) A luminescent heterometallic dirhodium–silver chain. Inorg Chem 41:634–636. doi:10.1021/ic015562d
4. Stender M, White-Morris RL, Olmstead MM, Balch AL (2003) New structural features of unsupported chains of metal ions in luminescent [$(NH_3)_4$Pt][Au(CN)$_2$]$_2$•1.5(H_2O) and related salts. Inorg Chem 42:4504–4506. doi:10.1021/ic034383o
5. Yeh TT, Wu JY, Wen YS, Liu YH, Twu J, Tao YT, Lu KL (2005) Luminescent silver metal chains with unusual μ_4-bonded 2,2′-bipyrazine. Dalton Trans 2005:656–658. doi:10.1039/b416703a
6. Buss CE, Anderson CE, Pomije MK, Lutz CM, Britton D, Mann KR (1998) Structural investigations of Vapochromic Behavior. X-ray single crystal and powder diffraction studies of [Pt(CN-iso-C_3H_7)$_4$][M(CN)$_4$] for M = Pt or Pd. J Am Chem Soc 120:7783–7790. doi:10.1021/ja981218c
7. Grate JW, Moore LK, Janzen DE, Veltkamp DJ, Kaganove S, Drew SM, Mann KR (2002) Steplike response behavior of a new vapochromic platinum complex observed with simultaneous acoustic wave sensor and optical reflectance measurements. Chem Mater 14:1058–1066. doi:10.1021/cm0104506
8. Drew SM, Smith LI, McGee KA, Mann KR (2009) A platinum(II) extended linear chain material that selectively uptakes benzene. Chem Mater 21:3117–3124. doi:10.1021/cm900401u
9. Pantazis DA, Murillo CA, McGrady JE (2008) A re-evaluation of the two-step spin crossover in the trinuclear cation [Co_3(dipyridylamido)$_4$$Cl_2$]$^+$. Dalton Trans 2008:608–614. doi:10.1039/b715021k
10. Cotton FA, Murillo CA, Wang Q, Young MD (2008) Unusual magnetism of an unsymmetrical trinickel chain. Eur J Inorg Chem 5257–5262. doi:10.1002/ejic.200800808
11. López X, Rohmer MM, Bénard M (2008) DFT modeling of the [M-Pd-M]$^{6+}$ metal atom chains (M = Ni, Pd); structural electronic and magnetic issues. J Mol Struct 890:18–23. doi:10.1016/j.molstruc.2007.12.007
12. Chardon-Noblat S, Deronzier A, Ziessel R, Zsoldos D (1998) Electroreduction of CO_2 catalyzed by polymeric [Ru(bpy)(CO)$_2$]$_n$ films in aqueous media: parameters influencing the reaction selectivity. J Electroanal Chem 444:253–260. doi:10.1016/S0022-0728(97)00584-6
13. Chardon-Noblat S, Deronzier A, Hartl F, van Slageren J, Mahabiersing T (2001) A novel organometallic polymer of osmium(0), [Os(2,2′-bipyridine)(CO)$_2$]$_n$: its electrosynthesis and electrocatalytic properties towards CO_2 reduction. Eur J Inorg Chem 613–617. doi:10.1002/1099-0682(200103
14. Oresmaa L, Moreno MA, Jakonen M, Suvanto S, Haukka M (2009) Catalytic activity of linear chain ruthenium carbonyl polymer [Ru(CO)$_4$]$_n$ in 1-hexene hydroformylation. Appl Catal A 353:113–116. doi:10.1016/j.apcata.2008.10.028
15. Kontkanen ML, Oresmaa L, Moreno MA, Jänis J, Laurila E, Haukka M (2009) One-dimensional metal atom chain [Ru(CO)4]n as a catalyst precursor—hydroformulation of 1-hexene using carbon dioxide as a reactant. Appl Catal A 365:130–134. doi:10.1016/j.apcata.2009.06.006
16. Caseri W (2004) Derivatives of Magnus' green salt; from intractable materials to solution-processed transistors. Platinum Metals Rev 48:91–100. doi:10.1595/147106704X1504
17. Bremi J, Brovelli D, Caseri W, Hähner G, Smith P, Tervoort T (1999) From Vauguelin's and Magnus's salts to gels, uniaxially oriented films, and fibers: synthesis, characterization, and properties of tetrakis(1-aminoalkane)metal(II) tetrachlorometalates(II). Chem Mater 11:977–994. doi:10.1021/cm9806376
18. Krogmann K (1969) Planare Komplexe mit Metall-Metall-Bindungen. Angew Chem 81:10–17. doi:10.1002/ange.19690810103
19. Masciocchi N, Moret M, Cairati P, Ragaini F, Sironi A (1993) Solving simple organometallic structures solely from X-ray powder diffraction data: the case of polymeric [{Ru(CO)$_4$}$_n$]. J Chem Soc Dalton Trans 1993:471–475. doi:10.1039/DT9930000471
20. Maschiocchi N, Sironi A, Chardon-Noblat S, Deronzier A (2002) X-ray powder diffraction study of organometallic polymers: [Ru

(L)(CO)$_2$]$_n$ (L=2,2′-bipyridine or 1,10-phenantroline). Organometallics 21:4009–4012. doi:10.1021/om020298x

21. Finniss GM, Canadell E, Campana C, Dunbar KR (1996) Unprecedented conversion of a compound with metal–metal bonding into a solvated molecular wire. Angew Chem Int Edn Engl 35:2771–2774. doi:10.1002/anie.199627721
22. Tejel C, Ciriano MA, Oro LA (1999) From platinum blues to rhodium and iridium blues. Chem Eur J 5:1131–1135. doi: 10.1002/(SICI)1521-3765(19990401)5:4<1131::AID-CHEM1131>3.0.CO;2-3
23. Berry JF (2010) Metal–metal bonds in chains of three or more metal atoms: from homometallic to heterometallic chains. Struct Bond 136:1–28. doi:10.1007/978-3-642-05243-9_1
24. Yang MH, Chou CC, Lee HC, Lee GH, Leung MK, Peng SM (1997) New oligo-α-pyridylamido ligands and their metal complexes. Chem Commun 1997:2279–2280. doi:10.1039/A706439J
25. Liu IPC, Wang WZ, Peng SM (2009) New generation of metal string complexes: strengthening metal–metal interaction via naphthyridyl group modulated oligo-α-pyridylamido ligands. Chem Commun 2009:4323–4331. doi:10.1039/b904719k
26. Hua SA, Liu IPC, Hasanov H, Huang GC, Ismayilov RH, Chiu CL, Yeh CY, Lee GH, Peng SM (2010) Probing the electronic communication of linear heptanickel and nonanickel string complexes by utilizing two redox-active [Ni$_2$(napy)$_4$]$^{3+}$ moieties. Dalton Trans 39:3890–3896. doi:10.1039/b923125k
27. Shih KN, Huang MJ, Lu HC, Fu MD, Kuo CK, Huang GC, Lee GH, Chen CH, Peng SM (2010) On the tuning of electric conductance of extended metal atom chains via axial ligands for [Ru$_3$(μ$_3$-dpa)$_4$(X)$_2$]$^{0/+}$ (X = NCS$^-$, CN$^-$). Chem Commun 46:1338–1340. doi:10.1039/b916677g
28. Peng SM, Wang CC, Jang YL, Chen YH, Li FY, Mou CY, Leung MK (2000) One-dimensional metal string complexes. J Magn Magn Mater 209:80–83. doi:10.1016/S0304-8853(99)00650-2
29. Huang GC, Liu IPC, Kuo JH, Huang YL, Yeh CY, Lee GH, Peng SM (2009) Further investigations of linear trirhodium complexes: experimental and theoretical studies of [Rh$_3$(dpa)$_4$Cl$_2$] and Rh$_3$(dpa)$_4$Cl$_2$]BF$_4$ [dpa = bis(2-pyridyl)amido anion]. Dalton Trans 2009:2623–2629. doi:10.1039/b820060b
30. Yin C, Huang GC, Kuo CK, Fu MD, Lu HC, Ke JH, Shih KN, Huang YL, Lee GH, Yeh CY, Chen CH, Peng SM (2008) Extended metal-atom chains with and inert second row transition metal: [Ru$_5$(μ$_5$-tpda)$_4$X$_2$] (tpda^{2-} = tripyridylamido dianion, X = Cl and NCS). J Am Chem Soc 130:10090–10092. doi:10.1021/ja8016818
31. Clérac R, Cotton FA, Dunbar KR, Lu T, Murillo CA, Wang X (2000) A new linear tricobalt combound with di(2-pyridyl)amide (dpa) ligands: two-step spin crossover of [Co$_3$(dpa)$_4$Cl$_2$][BF$_4$]. J Am Chem Soc 122:2272–2278. doi:10.1021/ja994051b
32. Sheu JT, Lin CC, Chao I, Wang CC, Peng SM (1996) Linear trinuclear three-centered metal–metal bonds: synthesis and crystal structure of [M$_3$(dpa)$_4$Cl$_2$] [M = Ru^{II} or Rh^{II}, dpa = bis(2-pyridyl) amido anion]. Chem Commun 1996:315–316. doi:10.1039/CC9960000315
33. Kuo CK, Liu IPC, Yeh CY, Chou CH, Tsao TB, Lee GH, Peng SM (2007) Oxidation of linear trinuclear ruthenium complexes [Ru$_3$(dpa)$_4$(CN)$_2$]: synthesis, structures, electrochemical and magnetic properties. Chem Eur J 13:1442–1451. doi:10.1002/chem.200601219
34. Kuo CK, Chang JC, Yeh CY, Lee GH, Wang CC, Peng SM (2005) Synthesis, structures, magnetism and electrochemical properties of triruthenium–acetylide complexes. Dalton Trans 2005:3696–3701. doi:10.1039/b506267e
35. Bader RFW (1990) Atoms in molecules: a quantum theory. Clarendon, Oxford
36. Gervasio G, Bianchi R, Marabello D (2004) About the topological classification of the metal–metal bond. Chem Phys Lett 387:481–484. doi:10.1016/j.cplett.2004.02.043
37. Niskanen M, Hirva P, Haukka M (2009) Computational DFT study of ruthenium tetracarbonyl polymer. J Chem Theor Comput 5:1084–1090. doi:10.1021/ct800407h
38. Niskanen M, Hirva P, Haukka M (2010) The effect of N-ligands on the geometry, bonding, and electronic absorption properties of ruthenium carbonyl chains. Phys Chem Chem Phys 12:9777–9782. doi:10.1039/c0cp00189a
39. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann E, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzales C, Pople JA (2004) Gaussian 03, Revision C.02. Gaussian Inc, Wallingford CT
40. Adamo C, Barone V (1999) Toward reliable density functional methods without adjustable parameters: the PBE0 model. J Chem Phys 110:6158–6170. doi:10.1063/1.478522
41. Huzinaga S (ed) (1984) Gaussian basis sets for molecular calculations, physical sciences data 16. Elsevier, Amsterdam
42. Hirva P, Haukka M, Jakonen M, Moreno MA (2008) DFT tests for group 8 transition metal carbonyl complexes. J Mol Model 14:171–181. doi:10.1007/s00894-007-0259-7
43. Biegler-König F, Schönbohm J (2002) Update of the AIM2000-program for atoms in molecules. J Comput Chem 42:1489–1494. doi:10.1002/jcc.10085
44. Bridgeman AJ, Cavigliasso G, Ireland LR, Rothert J (2001) The Mayer bond order as a tool in inorganic chemistry. J Chem Soc Dalton Trans 2001:2095–2108. doi:10.1039/B102094N
45. Berry JF, Cotton FA, Daniels LM, Murillo CA, Wang X (2003) Oxidation of Ni$_3$(dpa)$_4$Cl$_2$ and Cu$_3$(dpa)$_4$Cl$_2$: nickel–nickel bonding interaction, but no copper–copper bonds. Inorg Chem 42:2418–2427. doi:10.1021/ic0262740

ORIGINAL PAPER

Combinatorial screening of polymer precursors for preparation of benzo[α] pyrene imprinted polymer: an ab initio computational approach

Muntazir S. Khan · Prateek S. Wate · Reddithota J. Krupadam

Received: 28 May 2011 / Accepted: 9 August 2011 / Published online: 30 August 2011

Abstract A combinatorial screening procedure was used for the selection of polymer precursors in the preparation of molecularly imprinted polymer (MIP), which is useful in the detection of the air pollution marker molecule benzo[a] pyrene (BAP). Molecular imprinting is a technique for the preparation of polymer materials with specific molecular recognition receptors. The preparation of imprinted polymers requires polymer precursors such as functional monomer, cross-linking monomer, solvent, an initiator of polymerization and thermal or UV radiation. A virtual library of functional monomers was prepared based on interaction binding scores computed using HyperChem Release 8.0 software. Initially, the possible minimum energy conformation of the monomers and BAP were optimized using the semi-empirical (PM3) quantum method. The binding energy between the functional monomer and the template (BAP) was computed using the Hartree-Fock (HF) method with 6-31 G basis set, which is an ab initio approach based on Moller-Plesset second order perturbation theory (MP2). From the computations, methacrylic acid (MAA) and ethylene glycol dimethacrylate (EGDMA) were selected for preparation of BAP imprinted polymer. The larger interaction energy (ΔE) represents possibility of more affinity binding sites formation in the polymer, which provides high binding capacity. The theoretical predictions were complimented through adsorption experiments. There is a good agreement between experimental binding results and theoretical computations, which provides further evidence of the validity of the usefulness of computational screening procedures in the selection of appropriate MIP precursors in an experiment-free way.

Keywords Molecularly imprinted polymer · Computer simulation · Combinatorial screening · Polymer precursors · Molecular recognition

Introduction

The development of polymer receptors capable of recognizing traces of environmental toxins represents a challenge in chemical/physical science today. Many approaches are used currently to produce synthetic receptors; however, molecular imprinting has been studied intensively in recent years [1–3]. The concept of molecular imprinting was put forward by Wulff in 1972 [4–6] and has developed rapidly since 1997 when a molecularly imprinted polymer (MIP) using theophylline as template was reported [7]. In general, the synthesis of MIP can be summarized as follows. Firstly, the template and the monomer are mixed in a rational ratio to form a pre-polymerization complex. Cross-linking monomer and an initiator are then added to the mixture. Polymerization is carried out by heat treatment or UV radiation. After polymerization, the template molecule is removed, leaving cavities in the polymer matrix that are complementary both in size and functional arrangement to those in the template molecules (Fig. 1). Therefore, the MIP

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1218-x) contains supplementary material, which is available to authorized users.

M. S. Khan · R. J. Krupadam (✉)
National Environmental Engineering Research Institute,
Nagpur 440 020, India
e-mail: rj_krupadam@neeri.res.in

P. S. Wate
Department of Materials Science and Engineering,
University of Florida,
Gainesville, FL 32608, USA

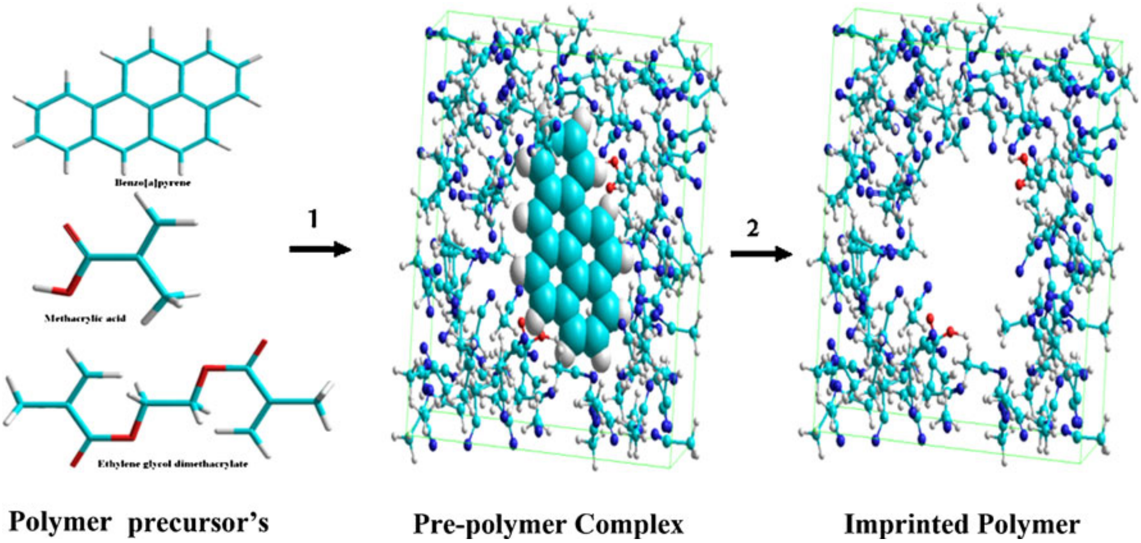


Fig. 1 Schematic representation of the molecular imprinting of benzo [α]pyrene. specific binding sites are generated using methacrylic acid as a functional monomer. After polymerization (*step 1*) and removal of the template (*step 2*), binding sites containing template-specific shapes are left in the polymer

can selectively recognize the template molecule from structurally similar compounds. A vast number of papers have been published to describe MIP preparation formats and to present new application areas for these materials [8–16]. MIPs have advantages such as an excellent predetermined selectivity and easy preparation. These advantages over other methods have drawn extensive attention in recent years in the field of the preparation of synthetic receptors for sensing explosives, and biological and environmental toxins.

Although the synthesis of MIPs is easy, a large library of functional monomers and the presence of cross-linking monomers make the task of screening out the best polymer precursors quite difficult. In practice, standard formulations using chemical intuition are usually employed, and attempts aiming at modifying the properties of polymers are based mainly on trial-and-error methods. The selection of appropriate functional monomers using simulated annealing (molecular dynamics) was reported for the preparation of ephedrine and cyanotoxin-imprinted polymers [17–21]. In an interesting study, Takeuchi et al. [22] demonstrated the stability of the pre-polymer complex formed between template (biotin) and functional monomer (MAA) using Monte Carlo simulations. A library of functional monomers was prepared for template (cyanotoxin; microcystin-LR) with the aid of computer simulation using SYBYL 6.7 software [23–25]. Dumitru et al. [26] applied a state-of-the-art computational tool (Cerius2 simulation tool) to achieve an understanding of intermolecular interactions in molecular imprinting of theophylline into complex polymeric systems. Dong et al. [27] used high level density functional theory (DFT) to calculate the binding energy and interaction energy (ΔE) between a template and functional monomers as a measure of their interaction, which facilitated the selection of monomers for MIP synthesis. The above cited research findings demonstrate that a combinatorial approach can provide experiment-free (trial-and-error methods) selection of the best polymer precursors for MIP preparation. However, there are no reports aimed at understanding the weak interactions in a template with no functional groups (such as benzo[α] pyrene, BAP) during pre-polymerization in MIP preparation, or the subsequent adsorption properties of MIPs.

In this paper, the authors prepared a virtual library of functional monomers for template-BAP (this template has no functional groups) using HyperChem Release 8.0 software [28]. BAP is an air pollution marker molecule and a probable human carcinogenic pollutant [15]. Due to the lack of functional groups in the template, it is quite difficult to generate interaction energy scoring of a template–functional monomer complex in a given solvent system. The interaction energy computed between functional monomers and the template, followed by combinatorial screening, was used to prepare selected MIPs. The adsorption capacity was determined experimentally and compared with theoretically predicted internal energy scoring. These computer simulations aided in the selection of the most appropriate polymer precursors for MIP preparation. The approach presented in this article shows that computer simulations reduce experimental time and also provide useful information about functional monomers that can express strong imprinting effects with the targeted molecule.

Materials and methods

Computational methods

Hardware and software

The workstation used to simulate functional monomer–template interactions was an Intel (R) Pentium IV running a Windows XP operating system, CPU 2.80 GHz, and 1 GB of RAM (memory), and 160 GB hard disk. This system was used to execute the software package HyperChem Release 8.0 [28] (http://www.hyper.com/).

Geometry optimization and energy calculation

Geometric optimization In the first step, 2-D chemical structures of the functional monomers (a virtual library of 24 monomers is shown in Table 1), template, and cross-linking monomers were prepared using HyperChem Release 8.0 software. Then, using the molecular builder option, the 2-D structure was converted to a 3-D structure. A schematic of the computational steps followed is given in Fig. 2. Geometric optimization was then carried out by the semi-empirical (SE) quantum mechanical approach (PM3 method) to obtain minimum energy structures. This SE method is a quantum mechanical method that uses approximations to solve the Schrödinger equation (see HyperChem user's manual). SE method uses only pre-calculated data from empirical studies, ignoring the core electrons and considering only valence electrons, which are of special interest to the chemistry of the molecule. SE also neglects or parameterizes two-electron integrals. The SE method generates relatively standard information about molecules and executes faster than other quantum mechanical methods. Hence, the SE method was used for geometrical optimization of polymer precursors. In this method, the Polak-Ribiere algorithm was chosen, which is a conjugate gradient method used for specially aromatic or conjugated organic compounds. The molecular structures were optimized using the Polak-Ribiere algorithm until the root mean square gradient was 0.01.

Interaction energy calculations The binding scores of the monomer, and interaction energies of the monomer–template complex were then calculated using Hartree-Fock (HF) method with 6-31 G basis set and Moller-Plesset second order perturbation theory (MP2), which is an ab initio method. Binding score is defined as the amount of energy required to disassemble a molecule into its atoms. In this method, it is possible to apply an electronic structure package capable of predicting molecular properties such as stable conformation, vibrational frequencies of atoms, molecules, and reactive systems. The ab initio calculation has the additional advantage of high accuracy level of information, reliability, and provides better results for weak interaction systems. It is a two-electron integral system, which gives more accurate binding scores. The binding score calculations between functional/cross-linking monomers and template were performed using HF method with 6-31 G basis set and MP2 levels of theory.

The minimum binding energies between the optimized conformations of 1: N ratio of template–monomer complexes are listed in Table 2. Using conformation optimization, the most stable template–monomer complexes were screened based on interaction energy, ΔE. The ΔE values were calculated using the following equation:

$$\Delta E = E(\text{template} - \text{monomer}) - E(\text{template}) - \sum E(\text{monomer}) \tag{1}$$

QSAR (quantitative structure-activity relationships) are used to correlate molecular structure, or properties derived from molecular structure, with a particular kind of chemical activity. The QSAR calculations of the template and functional monomers studied are presented in the Electronic supplementary material (S-I.1). The calculations are empirical and generally faster than other methods. The QSAR Properties calculated were: atomic partial charges (Gasteiger-Marsili method), van der Waals and solvent-accessible surface areas, hydration energy, van der Waals-surface-bounded molecular volume, solvent-accessible surface-bounded molecular volume, log P (the log of the octanol-water partition coefficient), molar refractivity, polarizability and molecular mass.

Experimental methods

Benzo[a]pyrene (BAP), methacrylic acid (MAA) and 4-vinyl pyridine (4-VP) were purchased from Sigma-Aldrich (Taufkirchen, Germany). Ethylene glycol dimethacrylate (EGDMA) (washed successively with 15% NaOH, saturated with $NaHCO_3$ and NaCl solutions, dried with CaH_2 and distilled before use) was obtained from Fluka (Steinhiem, Germany). 2, 2′-Azobis (2-isobutyronitrile) (AIBN), and acetonitrile (ACN) for synthesis were obtained from Merck (Darmstadt, Germany). In addition to these two polymers (MAA-EGDMA and 4-VP-EGDMA), 74 polymers were prepared using different polymer precursors for BAP and the functional and cross-linking monomers of analytical grade were purchased from different vendors and used as procured. All solutions were prepared using ultrapure water, obtained by reserved osmosis including UV treatment (Milli-RO 5 Plus, Millipore, Singapore).

Table 1 A virtual library of functional monomers for preparation of benzo[a]pyrene imprinted polymer

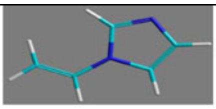
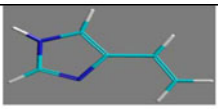
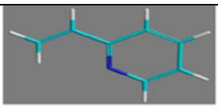
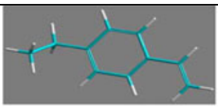
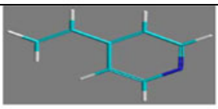
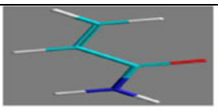
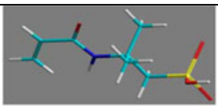
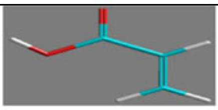
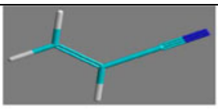
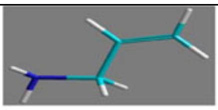
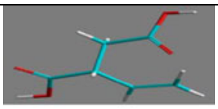
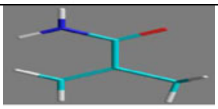
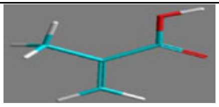
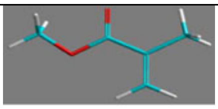
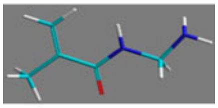
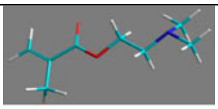
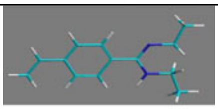
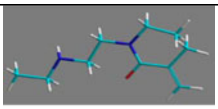
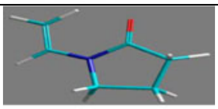
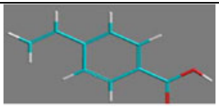
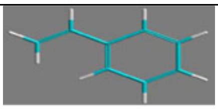
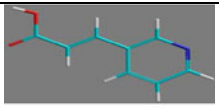
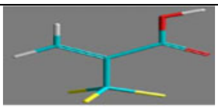
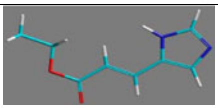
S.No	Functional monomers (FM)	Simulated structures	Acidic or basic property of FM	Binding score ΔE(kcalmol^{-1})
1	1-Vinylimidazole		Basic monomer	−9.21
2	2 (5)-Vinylimidazole		Basic monomer	−8.36
3	2-Vinylpyridine		Basic monomer	−23.43
4	4-Ethylstyrene		Neutral monomer	2.37
5	4-Vinylpyridine		Basic monomer	−27.84
6	Acrylamide		Neutral monomer	−7.94
7	Acrylamido-2-methyl-1-propane- sulfonic acid		Acidic monomer	−1.96
8	Acrylic acid		Acidic monomer	−7.65
9	Acrylonitrile		Neutral monomer	−5.92
10	Allylamine		Basic monomer	−12.31
11	Itaconic acid		Acidic monomer	−19.99
12	Methacrylamide		Neutral monomer	−11.78
13	Methacrylic acid		Acidic monomer	−33.14

Table 1 (continued)

S.No	Functional monomers (FM)	Simulated structures	Acidic or basic property of FM	Binding score ΔE(kcalmol^{-1})
14	Methyl Methacrylic acid		Neutral monomer	−10.11
15	N,N'-diethyl aminoethyl methacrylamide (DEAEM)		Basic monomer	−2.33
16	N,N'-diethyl-4-styrylamidine		Basic monomer	−7.32
17	N,N,N,-trimethyl aminoethylmethacrylate		Basic monomer	5.63
18	N-(2-aminethyl)-methacrylamide		Basic monomer	−4.23
19	N-vinylpyrrolidone (NVP)		Basic monomer	−10.78
20	p-Vinylbenzoic acid		Acidic monomer	−16.87
21	Styrene		Neutral monomer	1.67
22	Trans-3-(3-pyridyl)-acrylic acid		Neutral monomer	−6.78
23	Trifluoro methacrylic acid		Acidic monomer	−17.84
24	Urocanic ethyl ester		Basic monomer	−3.41

Preparation of molecularly imprinted polymers

The procedure followed for preparation of MIP was as follows: in a 30 mL glass vial, the template BAP (1 mmol) was dissolved in 10 mL ACN; the functional monomer MAA (4 mmol) was added and the contents mixed in a shaker for 5 min. Next, 4 mmol of the functional monomer MAA was added, and the solution was mixed well for 5 min by placing it in a shaker. Then, 40 mmol cross-linking monomer EGDMA and 0.5 mg AIBN were added to the

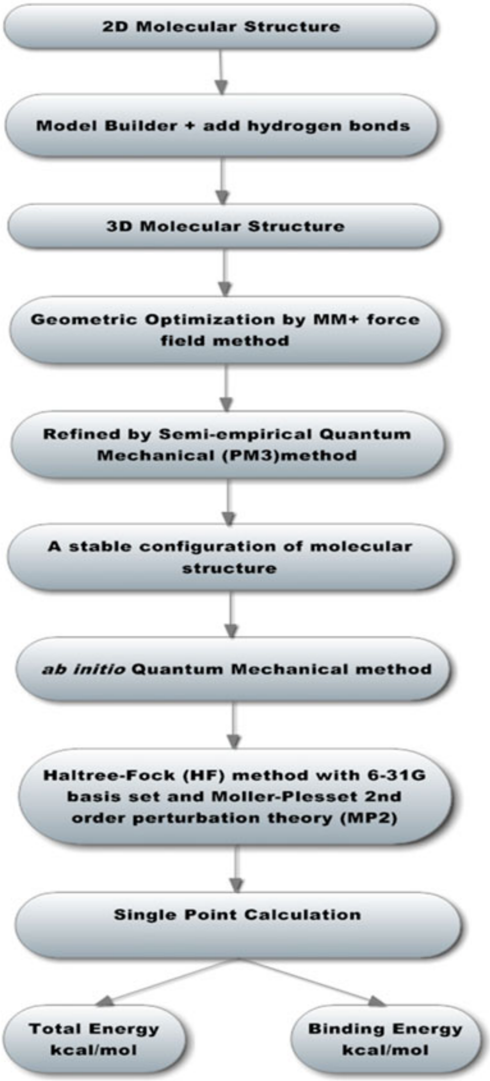

2D Molecular Structure
Model Builder + add hydrogen bonds
3D Molecular Structure
Geometric Optimization by MM+ force field method
Refined by Semi-empirical Quantum Mechanical (PM3)method
A stable configuration of molecular structure
ab initio Quantum Mechanical method
Haltree-Fock (HF) method with 6-31G basis set and Moller-Plesset 2nd order perturbation theory (MP2)
Single Point Calculation
Total Energy kcal/mol
Binding Energy kcal/mol

Fig. 2 Flow-chart representing the steps followed for molecular energy computations using HyperChem software [28]

solution. The sealed glass vial containing the reaction mixture was freeze-thaw degassed by submerging the tube in liquid nitrogen and holding the frozen tube under a vacuum of 100 mTorr for a period of 5 min. The tube was then sonicated for 10 min in order to mix the solution uniformly, then placed in water bath at 60°C for 24 h. Upon completion of polymerization, the tube was taken out of the water bath and crushed. The polymer monolith was ground in a ball mill to polymer particles sized 75 μm or less (200 mesh). The BAP was extracted in batch mode, using acetonitrile on a Soxlet distillation assembly for 24 h. The washing procedure was repeated (10 times) until BAP in the extraction solvent could not be detected by GC-MS. Finally, the particles were dried under vacuum for further use. MIPs with different polymer precursors were prepared [76 and free (F) guest molecules in heterogeneous systems and free (F) guest molecules in heterogeneous systems MIPs] using various combinations, and then their binding capacity determined using equilibrium adsorption studies.

The adsorption capacity of BAP onto MIPs was determined by contacting 10 mg polymer with 5 mL standard BAP solutions of different concentrations (1–10 mgL^{-1}). The samples were then kept in a shaker at 25°C for 3 h. After sedimentation of the adsorbents, the supernatant was decanted and the concentration of BAP was measured using gas chromatography/mass spectrometry (GC/MS). The analytical protocol followed is presented in the Electronic supplementary material. The amount of BAP adsorbed was calculated by subtraction using a calibrating curve obtained from the same experiment leaving out the adsorbent. The experiment was repeated at least twice for each adsorbent. The imprinting factor was calculated as the ratio between the adsorption capacity of MIP and its corresponding NIP.

Batch rebinding experiments

BAP adsorption studies were performed in batch mode. The dry adsorbents (MIP or NIP) were weighed in 5 mL glass vials, and 0.1, 0.2, 0.3, 0.4, and 0.5 ml BAP standard solution (B) was added followed by addition of acetonitrile to a final volume of 5 mL. The samples were then stirred in a circularly shaking water-bath at 25°C for 2 h. After sedimentation of the adsorbents, the concentration of BAP was measured using GC/MS. The amount of BAP adsorbed was calculated by subtraction using a calibrating curve obtained from the same experiment leaving out the adsorbent. The experiment was repeated at least twice for each adsorbent. Three adsorption isotherm models were chosen to represent experimental data. Model parameters were determined by nonlinear, least-squares regression of these data. Regression was done using the solver function

Table 2 Computationally derived binding energies for fixing template/monomer (T/M) molar ratio. *MAA* Methacrylic acid, *4-VP* 4-vinyl pyridine, *BAP* benzo[a]pyrene

Template	Functional monomer (MAA)	ΔE kcal mol^{-1}	Functional monomer (4-VP)	ΔE kcal mol^{-1}
BAP	1	−87.77	1	−74.84
	2	−93.87	2	−81.34
	3	−134.23	3	−117.42
	4	−178.39	4	−147.80
	5	−171.31	5	−148.21
	6	−164.87	6	−143.71
	7	−160.37	7	−150.33
	8	−175.07	8	−157.07
	9	−177.12	9	−161.74
	10	−180.54	10	−172.93

in Microsoft Excel 2000 to minimize the standard error by varying the model parameters. The experimental data were then calculated to obtain the corresponding standard deviations.

Results and discussion

Selection of the proper functional monomer

The selection of suitable functional monomers is a key factor in the preparation of MIPs. An interesting aspect of this study is the use of a template having no functional groups. The template BAP contains electron-rich polycyclic aromatic hydrocarbons with five condensed benzene structures. Based on the interaction energy between BAP and the functional monomer, a virtual library of BAP-functional monomer complexes was prepared. From the simulation results, it was found that the functional monomers MAA and 4-VP showed the highest interaction energy (ΔE) scoring with BAP to form the most stable complexes in the equilibrium state. The functional monomers, namely 2-vinylpyridine, itaconic acid, p-vinylbenzoic acid, and trifluoro methacrylic acid, are also fairly good candidates for imprinting monomers based on interaction energy scoring. N-(2-aminethyl)-methacrylamide, urocanic ethyl ester and 2-acrylamido-2-methyl-1-propanesulfonic acid forms the least stable structures based on lowest interaction energy scoring. In fact, 4-ethylstyrene, styrene and N,N,N, trimethyl aminoethyl methacrylate also showed low interaction energy with BAP.

Identification of interaction type between template and monomer

The monomer simulations were analyzed further to determine which part of the monomer comes closest to the template (molecule orientation), and the magnitude of these distances in vacuum and in a virtual solvent box were determined. The results of this analysis are presented in Fig. 3. In most cases, it was found that the functional group of monomers interacting with template tends to be –COOH or CH_2=CH–. The two most stable complexes of BAP were found to be MAA and 4-VP, the simulated closest distance of approach was approximately 2.8 and 4.5 Å, and the binding was predominantly with –COOH (in the case of methacrylic acid, –OH was also involved). This indicates clearly that the presence of weak H-bonding and π–π interaction is involved between the template and the functional monomer. The binding distances or simulated closest distance between the template and the monomer ranged from 2 to 10 Å in solvent, while in a vacuum the distance was between 5 and 13 Å for all monomers. A closer view of the possible optimized configuration in vacuum and solvent is presented in Fig. 4.

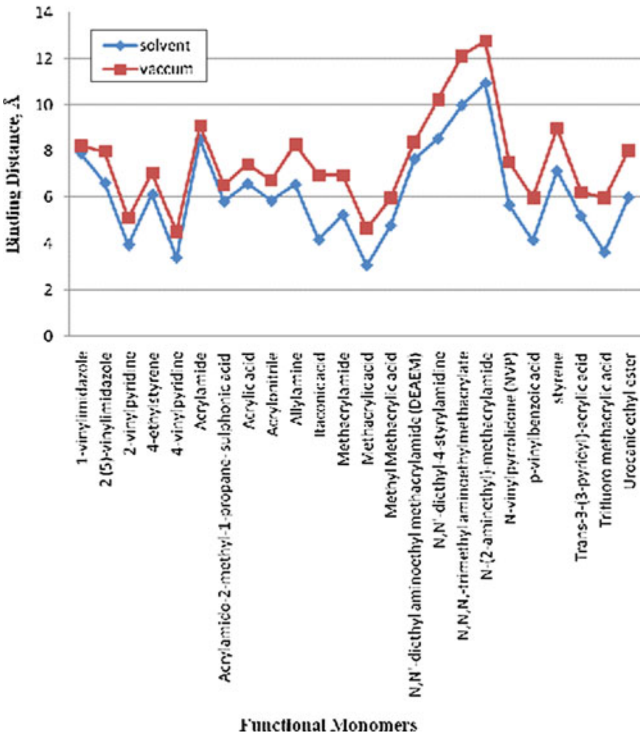


Fig. 3 Binding distance between benzo[a]pyrene (BAP) and functional monomers computed in vacuum and in a virtual solvent box

From the observations, a weak correlation can be noted, namely distances between 3 and 5 Å tend to correspond to greater negative interaction energies, indicating that this is roughly the distance of closest approach required for the formation of stable complexes between BAP-monomers. In vacuum, the binding distance between monomer and template is more than that of the solvent. These distances are consistent with the 3-D structures of the template and the monomers as determined from their total electronic charge distributions (densities). The total electronic charge distribution data provides information about the possible interactions between the monomer and the template. Generally, total electronic charge density plots represent the electron density function for the molecular valence electrons, in units of e/a_0^3. This property is associated with the surface of a molecule. It describes the probability of finding an electron at a point in space. The value is the sum for each electron of Ψi^2, where Ψ is the molecular orbital occupied by the *i*th electron. For a closed-shell system this is 2 Ψi^2, summed over the occupied orbitals. In the present case, functional monomer and template, BAP interactions were computed for total charge density; the results are depicted in Fig. 5. The simulated electronic charge distribution of BAP electronic charge distribution extends 1–2 Å outside the O or H nucleus. A similar case applies to

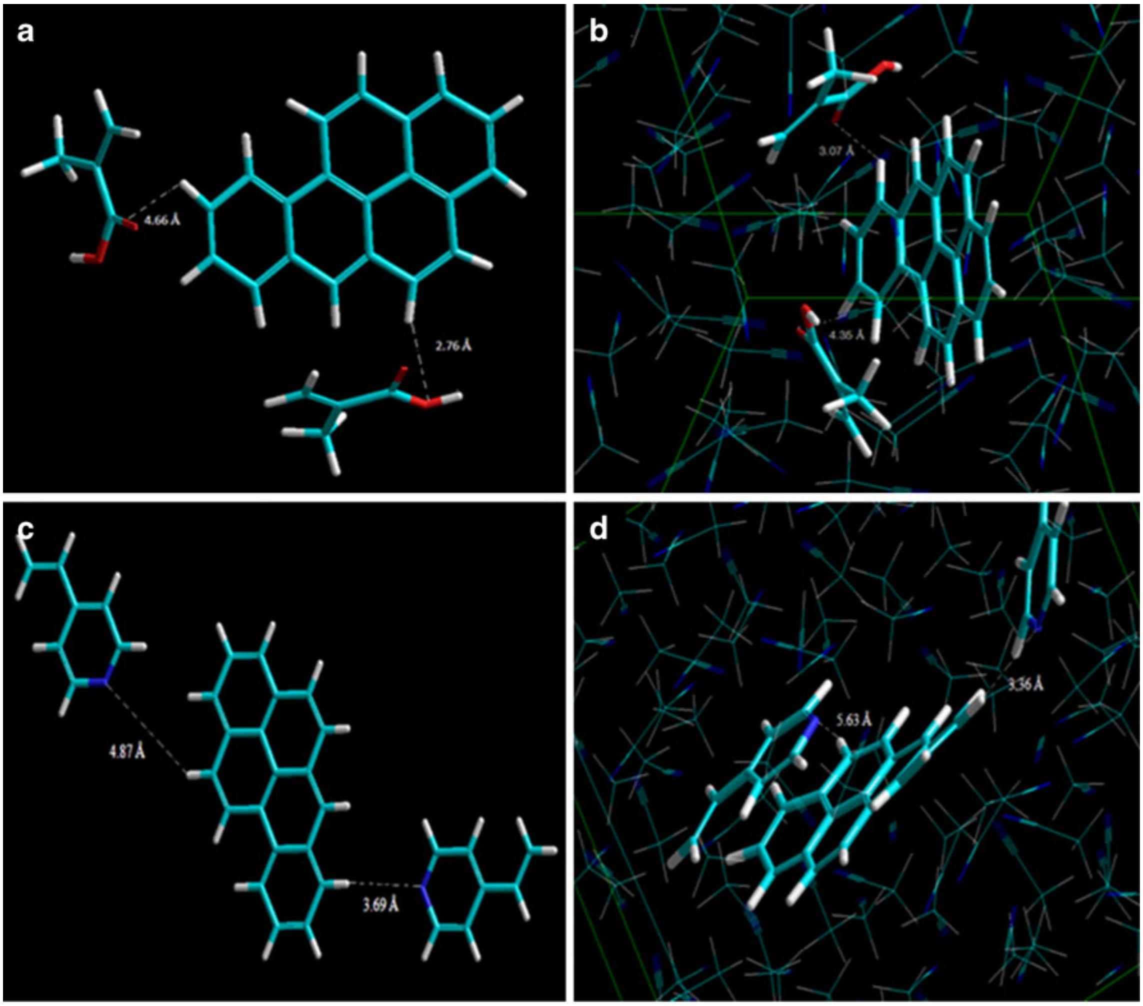


Fig. 4 Examples of two possible optimized configurations (**a**, **c**) and (**b**, **d**) for two molecules of methacrylic acid (MAA) and 4 vinyl pyridine (4-VP) with one molecule of BAP in vacuum and in acetonitrile solvent box, respectively

the other monomers, hence the minimal contact distance must fall between 3.0 and 5.0 Å for an appropriate and attractive interaction to occur.

Template-monomer mole ratio

The interaction energies of the stable complexes of BAP with each of the monomers at a molar ratio of 1: 1, 1: 2, 1: 3……., 1: N were calculated by applying the conformation optimization to these complexes (Table 2). In theory, a monomer giving a high binding energy with the template molecule would interact strongly with template molecules. In addition, when ΔE is higher, more active sites are expected to form in MIPs and the sites are more regular, which help MIPs to recognize the template, leading to good molecular recognition. Therefore, monomers with high binding energy are the best candidates for the preparation of MIPs. The binding energy of one MAA with BAP was calculated as −87.77 kcal mol^{-1}, while that of four MAA with BAP was calculated as −178.39 kcal mol^{-1}. This suggests that, in BAP–MIP, binding sites of high affinity should have four MAA molecules interacting with BAP, while binding sites of low affinity should have one MAA interacting with BAP. The results indicate clearly that at least four functional monomers are required to saturate all the binding sites of BAP. Therefore, a 1:4 ratio of template–monomer (T/M) was used for synthesis of MIP. The same ratio was found in the case of the BAP–4-VP complex.

To complement computational predictions, a series of MIPs was prepared by changing the amount of functional monomer (MAA) and keeping 1 mmol of template (BAP). The binding capacity of the resulting MIPs is given in Fig. 6. As can be seen, with increasing monomer portion of the T/M ratio, the binding capacity of MIPs increased gradually up to a ratio of 1:4. Thereafter, further increase in the functional monomer portion showed a downward trend. This could be because the functional monomer has

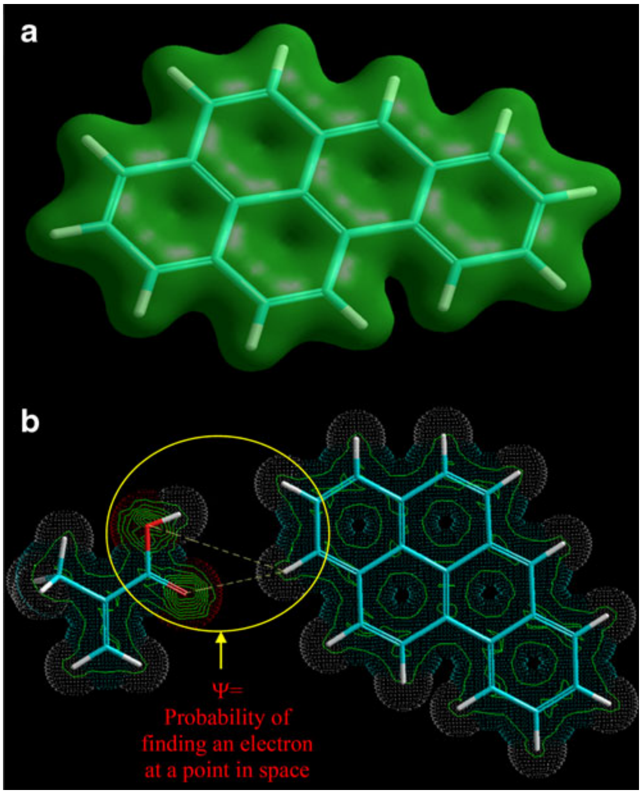


Fig. 5 Electronic charge density of **a** benzo[a]pyrene molecule, and **b** MAA and BAP

insufficient template to form stable pre-polymer complex at low and high template ratio. In addition to this, there could be interactions between monomer–monomer (self-association) resulting in the formation of non-specific binding sites. These two factors are critical for preparation of selective molecular recognition sites in MIP. Hence, selection of the optimal T/M ratio plays a dominant role in preparation of MIPs. The optimum T/M ratio chosen from the experimental results is 1:4, and same was confirmed from the computational predictions.

Selection of cross-linking monomer

The cross-linking monomer in MIP provides mechanical stability and formation of the appropriate template architecture. In the present study, computer simulations were performed using three cross-linking monomers, namely N, N-methylene bisacrylamide (NNMB), trimethylolpropane trimethacrylate (TRIM), and ethylene glycol dimethacrylate (EGDMA). From the computational results it was found that the cross-linker ethylene glycol dimethacrylate (EGDMA) showed the highest interaction energy towards many T, and M forms the most stable pre polymer complex in the equilibrium state in the polymer matrix. In summary, the template (BAP) interaction energy with 24 functional monomers was computed, and then the best 5 functional monomers were selected for further computations using three cross-linking monomers. Pre-polymer complex stability in different solvents is also reported. The data obtained from computations is presented in Christmas tree form (Fig. 7).

Adsorption capacity

The adsorption capacity of all 15 MIPs prepared in this study was also determined following batch adsorption experiments (Fig. 7). The MIPs demonstrated the same correlation between the calculated binding energy and the binding affinity for the few polymers shown in Fig. 8. MIP–BAP and its NIP were synthesized using assisted computational design. The Langmuir-Freundlich isotherm was used to determine the binding capacity of MIPs. As shown by Eq. 2, the LF model describes the relationship between the equilibrium concentrations of bound (B) and free (F) guest molecules in heterogeneous systems [29].

$$B = \frac{N_t \alpha F^m}{1 + \alpha F^m} \qquad (2)$$

where N_t is the binding capacity (μg mL^{-1} solute/mass polymer) in the polymer matrix, which is related to the binding affinity constant (K_0) via K_0=$a^{1/m}$, and m is the heterogeneity index. The heterogeneity index values varies from 0 to 1 (m=1 means the mediaumis homogeneous). Equation 2 was used to fit the adsorption isotherm by non-linear least square fitting, and the results are shown in Table 3. The adsorption isotherms of MIP and NIP represent a more homogeneous distribution of binding sites for BAP than NIP (Fig. 9). This would be due to the shape-specific memory of the template in the polymer formed during imprinting. In the case of NIP, the polymerization is performed without the template BAP and thus there are no

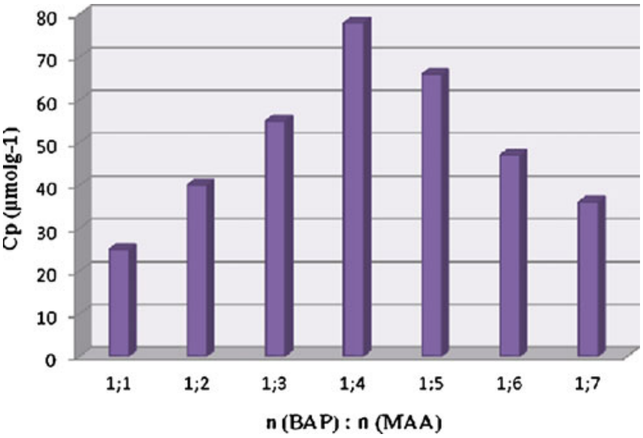


Fig. 6 Binding capacity of MIPs with different ratio of BAP:MAA

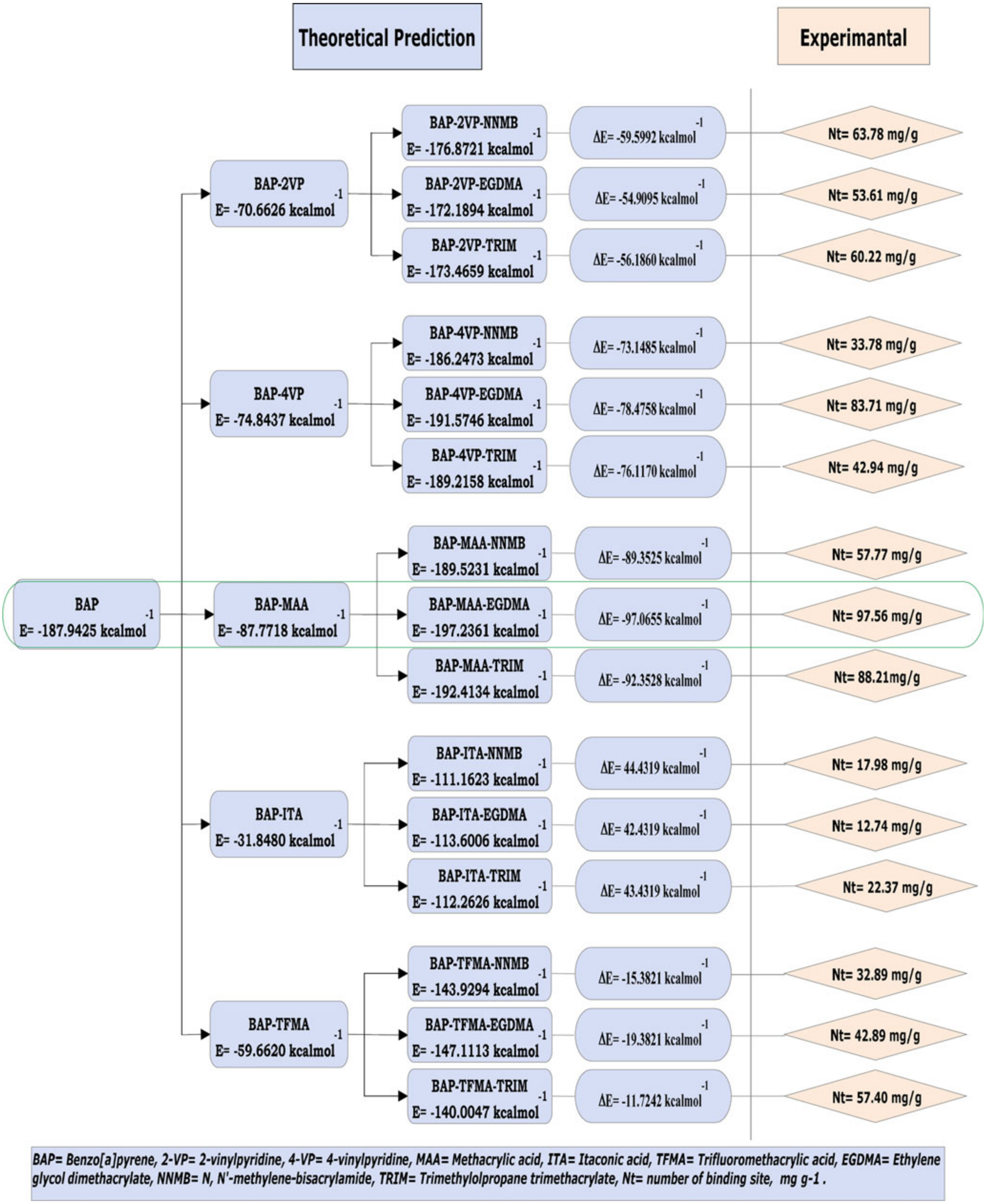
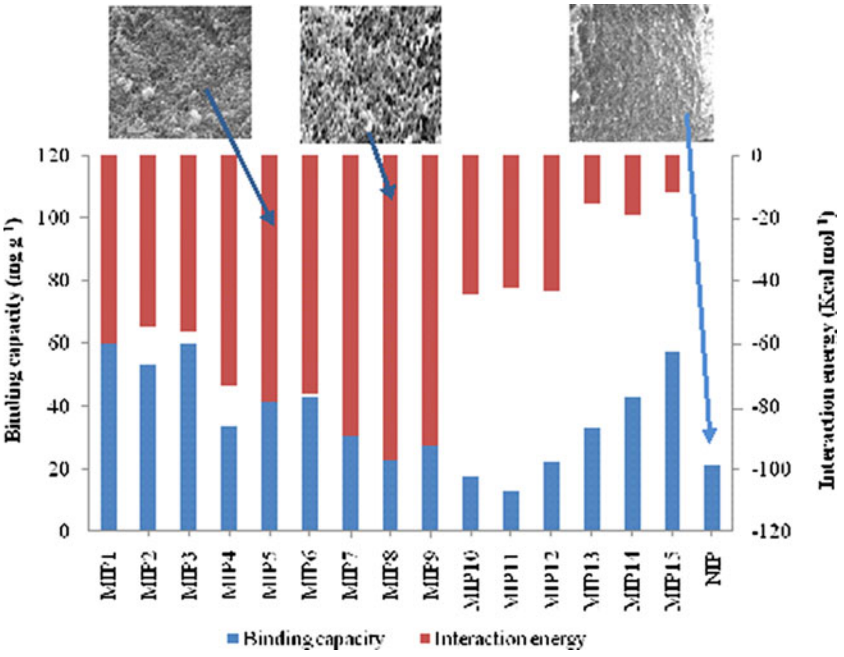

Theoretical Prediction
Experimantal
BAP E= -187.9425 kcalmol^{-1}
BAP-2VP E= -70.6626 kcalmol^{-1}
BAP-2VP-NNMB E= -176.8721 kcalmol^{-1}
ΔE= -59.5992 kcalmol^{-1}
Nt= 63.78 mg/g
BAP-2VP-EGDMA E= -172.1894 kcalmol^{-1}
ΔE= -54.9095 kcalmol^{-1}
Nt= 53.61 mg/g
BAP-2VP-TRIM E= -173.4659 kcalmol^{-1}
ΔE= -56.1860 kcalmol^{-1}
Nt= 60.22 mg/g
BAP-4VP E= -74.8437 kcalmol^{-1}
BAP-4VP-NNMB E= -186.2473 kcalmol^{-1}
ΔE= -73.1485 kcalmol^{-1}
Nt= 33.78 mg/g
BAP-4VP-EGDMA E= -191.5746 kcalmol^{-1}
ΔE= -78.4758 kcalmol^{-1}
Nt= 83.71 mg/g
BAP-4VP-TRIM E= -189.2158 kcalmol^{-1}
ΔE= -76.1170 kcalmol^{-1}
Nt= 42.94 mg/g
BAP-MAA E= -87.7718 kcalmol^{-1}
BAP-MAA-NNMB E= -189.5231 kcalmol^{-1}
ΔE= -89.3525 kcalmol^{-1}
Nt= 57.77 mg/g
BAP-MAA-EGDMA E= -197.2361 kcalmol^{-1}
ΔE= -97.0655 kcalmol^{-1}
Nt= 97.56 mg/g
BAP-MAA-TRIM E= -192.4134 kcalmol^{-1}
ΔE= -92.3528 kcalmol^{-1}
Nt= 88.21mg/g
BAP-ITA E= -31.8480 kcalmol^{-1}
BAP-ITA-NNMB E= -111.1623 kcalmol^{-1}
ΔE= 44.4319 kcalmol^{-1}
Nt= 17.98 mg/g
BAP-ITA-EGDMA E= -113.6006 kcalmol^{-1}
ΔE= 42.4319 kcalmol^{-1}
Nt= 12.74 mg/g
BAP-ITA-TRIM E= -112.2626 kcalmol^{-1}
ΔE= 43.4319 kcalmol^{-1}
Nt= 22.37 mg/g
BAP-TFMA E= -59.6620 kcalmol^{-1}
BAP-TFMA-NNMB E= -143.9294 kcalmol^{-1}
ΔE= -15.3821 kcalmol^{-1}
Nt= 32.89 mg/g
BAP-TFMA-EGDMA E= -147.1113 kcalmol^{-1}
ΔE= -19.3821 kcalmol^{-1}
Nt= 42.89 mg/g
BAP-TFMA-TRIM E= -140.0047 kcalmol^{-1}
ΔE= -11.7242 kcalmol^{-1}
Nt= 57.40 mg/g
BAP= Benzo[a]pyrene, 2-VP= 2-vinylpyridine, 4-VP= 4-vinylpyridine, MAA= Methacrylic acid, ITA= Itaconic acid, TFMA= Trifluoromethacrylic acid, EGDMA= Ethylene glycol dimethacrylate, NNMB= N, N'-methylene-bisacrylamide, TRIM= Trimethylolpropane trimethacrylate, Nt= number of binding site, mg g-1 .

Fig. 7 Christmas tree representing the combinatorial screening procedure for selection of appropriate polymer precursor; *right panel* experimentally derived binding capacities of the MIPs

Fig. 8 Correlation between experimental binding capacity and interaction energy



specific binding sites in the NIP. The adsorption capacity of MIP is N_t=97 μg mg^{-1} whereas for NIP, N_t is 37 μg mg^{-1}, clearly indicating the formation of a greater number of binding site for BAP. It is clear from Table 3 that MIP has a more homogeneous distribution of binding sites for BAP than NIP. This may be due to the formation of cavities for BAP, which were 80–90% homogeneous in size. In contrast, when polymerization was performed in the absence of BAP, no specific binding sites for BAP were formed. The binding affinity (a) of MIP is 1.99 whereas for NIP the value of a is 0.52. The nonlinear regression (R^2) values of the LF model for MIP and NIP are 0.997 and 0.941, respectively.

The MIPs, namely, MIP8 and MIP5, prepared using polymer precursors (MAA and 4-VP) were screened from the virtual library of functional monomers based on interaction energy criteria. MAA and 4-VP were the best functional monomers on ΔE value. The adsorption capacity of the MAA-MIP was higher than that of 4VP-MIP. The computational predictions and experimental results were in good agreement based on the parameters ΔE and N_t derived from computation and experiment, respectively.

Table 3 Langmuir–Freundlich fitted coefficients for MIP1 and NIP1. N_t Binding capacity (μg mg^{-1}), m heterogeneity index, K_o binding parameter in Langmuir isotherm (mL mg^{-1}), a affinity for corresponding model systems (MIP and NIP)

Adsorption model-parameters	MIP8	NIP8
N_t	97	37
a	1.99	0.52
m	0.781	1.13
R^2	0.997	0.941

Simulations in solvents

The pre-polymer complex formed between BAP-MAA and BAP-4-VP is influenced significantly by the nature of the solvent [30]. Therefore, the stability, ΔE_{sol}, of BAP, MAA and 4VP in the different solvents, namely acetonitrile

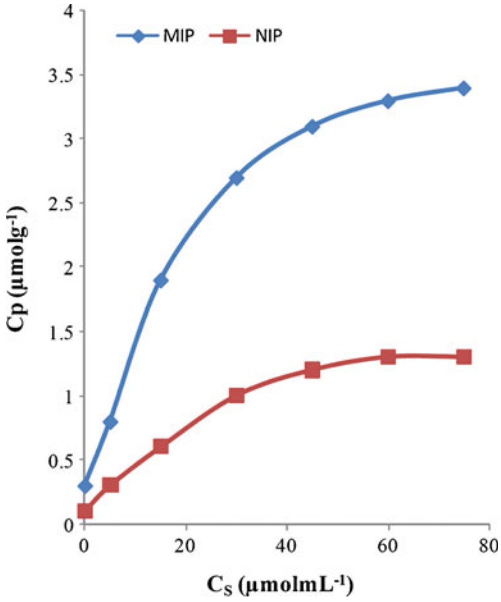


Fig. 9 Langmuir-Freundlich adsorption isotherm of MIP and NIP

(ACN), chloroform (CHCl3), dichloromethane (DCM) and toluene (TUL), was computed to determine the energy scoring, where ΔE_{sol} is defined as:

$$\Delta E_{sol} = E(solution) - E(vacuum) \tag{3}$$

The ΔE_{sol} of BAP in the four different solvents, ACN, CHCl3, DCM and TUL, was then computed according to Eq. 3 by means of the PBC (periodic box calculation) method implemented in HyperChem software [28]. The ΔE_{sol} of BAP in TUL was the largest, indicating the strongest interaction between BAP and functional monomers, while ACN resulted in the smallest values of ΔE_{sol}. The energy scoring of two MIPs in different solvents is given in Table 4. The decreasing order of energy scoring between BAP and the functional monomers is as follows: $|\Delta E_{sol}\ (TUL)| > |\Delta E_{sol}(CHCl3)| > |\Delta E_{sol}\ (DCM)| > |\Delta E_{sol}(ACN)|$. TUL is expected to have the highest affinity to the template molecule and the monomers. As has been noted in the literature [31], this affinity could acts as a shield or may reduce the interaction between BAP and MAA or BAP and 4-VP that is required in the formation of the pre-polymerization complexes. ACN has the least effect on complex formation, as indicated by its smallest ΔE_{sol}. Therefore, MIP synthesized in ACN will have the highest binding capacity because of the minimal interference of the solvent with the interaction between BAP and MAA or 4-VP.

In order to evaluate the adsorption capacity of the synthesized MIPs, a parameter known as imprinting factor, α, was calculated as follows:

$$K_D = \frac{C_P}{C_S} \tag{4}$$

$$\alpha = \frac{K_D(MIP)}{K_D(NIP)} \tag{5}$$

where C_p (μmol g^{-1}) is the concentration of template molecule on the polymer, C_S (μmol mL^{-1}) the equilibrium concentration of template molecule in solution, and K_D the partition coefficient of template molecule between polymer and solution. The imprinting factor (α) is the ratio between the binding capacity of imprinted and non-imprinted polymers. This factor represents the precise template imprints formed in the MIPs. Imprinting factor values depend primarily on the solvent system used in the preparation of MIPs. For example, in the present study, four solvents were used and the binding capacity and imprinting factors are given in Table 4. It can be seen that the MIP synthesized in acetonitrile has the best imprinting factor to BAP, while the MIP synthesized in toluene showed the least. The reasons for such imprinting factor variations can be explained based on the stabilization energy of both the template and the functional monomer in solvent systems. The polarity of the solvent plays an important role in molecular imprinting. The interaction between solvent and BAP/MAA increased with increasing polarity of solvent. MIP prepared in solvents with the highest dielectric constant showed high imprinting factors (Fig. 10). The reason for this could be that the polarity of the solvent favors the formation of H-bonds or electrostatic interactions between the template and the monomer. In other words, polar solvents facilitate stronger interaction between the template (BAP) and the monomer (MAA or 4-VP), and prevent self-association of monomers. This leads to the formation of imprinting sites in MIP and increases the MIP's adsorption selectivity and capacity.

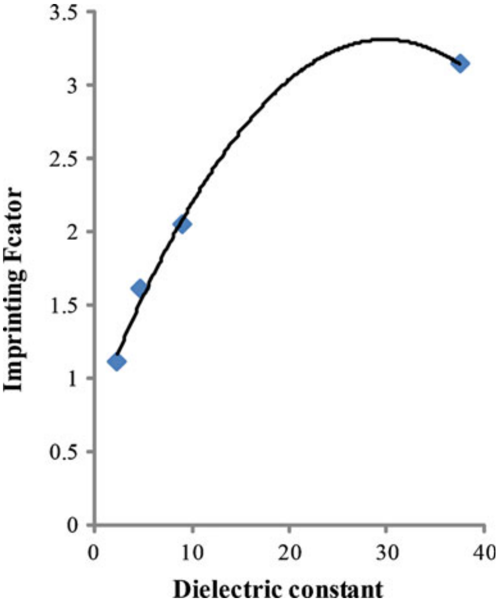


Fig. 10 Imprinting factors (α) of MIPs prepared in different solvents

Table 4 ΔE_{sol} of BAP, MAA, and 4-VP in different solvents

Solvent	ΔE of BAP (kcal mol^{-1})	ΔE of MAA (kcal mol^{-1})	ΔE of 4-VP (kcal mol^{-1})	$K_{D\ (MIP)}$ (μg g^{-1})	$K_{D\ (NIP)}$ (μg g^{-1})	Imprinting factor (α)	Dielectric constant
ACN	−10.63	−8.51	−6.47	126	40	3.15	37.5
CHCl3	−14.12	−8.97	−7.65	160	60	2.66	9.1
DCM	−17.34	−9.45	−8.83	137	63	2.17	4.8
TUL	−19.21	−12.99	−10.60	121	89	1.36	2.4

Conclusions

A computational (ab-initio quantum mechanical) approach developed for the rational design of MIPs, predicts that MAA/acetonitrile is the combination of functional monomer/solvent that leads to the most stable pre-polymerization adducts with BAP as template. A library of 25 monomers and their corresponding polymers has been established, and the interaction energies and closest approach distances computed. The simulated functional monomers and polymers with template indicated that the functional groups interacting with template tends to be either –COOH or CH2=CH– with π–π interaction, and the binding distances between the ligand and the monomer or polymer in the most stable cases are between 3.0 and 5.0 Å. To validate the computational procedure, the adsorption capacity of the MIPs was determined experimentally and compared with theoretically predicted interaction energies. The computational screening procedure described in this paper will be useful in the selection of appropriate MIP precursors for BAP.

Acknowledgments This work was supported by the Council of Scientific and Industrial Research (CSIR), New Delhi and the Planning Commission, Government of India under Supra Institutional Project: Environmental Molecular Sciences (SIP-16/3.3). M.S.K. acknowledges a CSIR Senior Research Fellowship.

References

1. Haupt K, Mosbach K (2000) Chem Rev 100:2495–2504
2. Henry OYF, Cullen DC, Piletsky SA (2005) Anal Bioanal Chem 382:947–956
3. Sellergren B (2001) Molecularly imprinted polymers: man-made mimics of antibodies and their applications in analytical chemistry. Elsevier, Amsterdam
4. Wulff G, Sarhan A, Zabrocki K (1973) Tetrahedron Lett 4329–4332
5. Wulff G, Vesper W, Grobe-Einsler W, Sarhan A (1977) Macromol Chem 178:2799–2816
6. Wulff G, Poll HG (1987) Macromol Chem 188:741–748
7. Vlatakis G, Andersson LI, Muller R, Mosbach K (1993) Nature 361:645–647
8. Yu C, Mosbach K (2000) J Chromatogr A 888:63–72
9. Chen WY, Chen CS, Lin FY (2001) J Chromatogr A 923:1–6
10. Kirsch N, Alexander C, Lubke M, Whitcombe MJ, Vulfson EN (2000) Polymer 41:5583–5590
11. Andersson HS, Koch-Schmidt AC, Ohlson S, Mosbach K (1996) J Mol Recognit 9:675–682
12. Katz A, Davis ME (1999) Macromolecules 32:4113–4121
13. Andersson HS, Karlesson JG, Piletsky SA, KouchSchmidt AC, Mosbach K, Nicholls IA (1999) J Chromatogr A 848: 39–49
14. Nicholls IA (1995) Chem Lett 11:1035–1036
15. Krupadam RJ, Khan MS, Wate SR (2010) Water Res 44:681–688
16. McNiven S, Yokobayashi Y, Cheong SH, Karube I (1997) Chem Lett 12:1297–1298
17. Lanza F, Sellergren B (1999) Anal Chem 71:2092–2096
18. Takeuchi T, Fukuma D, Matsui J (1999) Anal Chem 71:285–290
19. Subrahmanyam S, Piletsky SA, Piletska EV, Chen B, Karim K, Turner APF (2001) Biosens Bioelectron 16:631–637
20. Wu L, Sun B, Li Y, Chang W (2003) Analyst 128:944–949
21. Meng Z, Yamazaki T, Sode K (2004) Biosens Bioelectron 20:1068–1075
22. Takeuchi T, Dobashi A, Kimura K (2000) Anal Chem 72:2418–2422
23. Piletsky S, Karim K, Piletska EV, Day CJ, Freebairn KW, Legge C, Turner APF (2001) Analyst 126:1826–1830
24. Chianella I, Lotierzo M, Piletksy SA, Tothill IE, Chen BN, Karim K, Turner APF (2002) Anal Chem 74:1288–1293
25. Chianella I, Karim K, Piletska EV, Preston C, Piletsky SA (2006) Anal Chim Acta 559:73–78
26. Dumitru P, Jolanta L (2005) Polymer 46:7543–7556
27. Dong WG, Yan M, Zhang ML, Liu Z, Li YM (2005) Anal Chim Acta 542:186–193
28. HyperChem® for Windows. Hypercube. HyperChem Release version 8.0. http://www.hyper.com/
29. Baggiani C, Giraudi G, Giovannoli C, Tozzi C, Anfossi L (2004) Anal Chim Acta 504:43–52
30. Bhagat B, Krupadam RJ (2010) Adsorpt Sci Technol 28:79–88
31. Whitcombe MJ, Martin L, Vulfson EN (1998) Chromatographia 47:457–464

J Mol Model (2012) 18:1983–1991
DOI 10.1007/s00894-011-1163-8

ORIGINAL PAPER

Density functional theory study on the interaction between keto-9H guanine and aspartic acid

Patrina Thompson Harris · Glake A. Hill

Received: 31 March 2011 / Accepted: 24 June 2011 / Published online: 30 August 2011

Abstract A theoretical study was performed using density functional theory (DFT) to investigate hydrogen bonding interactions in signature complexes formed between keto-9H guanine (Gua) and aspartic acid (Asp) at neutral *pH*. Optimized geometries, binding energies and the theoretical IR spectra of guanine, aspartic acid and their corresponding complexes (Gua-Asp) were calculated using the B3LYP method and the 6-31+G(d) basis set. Stationary points found to be at local minima on the potential energy surface were verified by second derivative harmonic vibrational frequency calculations at the same level of theory. AIM theory was used to analyze the hydrogen bonding characteristics of these DNA base complex systems. Our results show that the binding motif for the most stable complex is strikingly similar to a Watson-Crick motif observed in the guanine-cytosine base pair. We have found a range of hydrogen bonding interactions between guanine and aspartic acid in the six complexes. This was further verified by theoretical IR spectra of ω(C-H—O-H) cm^{-1} stretches for the Gua-Asp complexes. The electron density plot indicates strong hydrogen bonding as shown by the 2p_z dominant HOMO orbital character.

Keywords Aspartic acid · Guanine · Hydrogen bonding · IR spectrum

P. T. Harris · G. A. Hill (✉)
Interdisciplinary Nanotoxicity Center, Department of Chemistry,
Jackson State University,
1400 J. R. Lynch Street,
Jackson, MS 39217, USA
e-mail: glakeh@ccmsi.us

Introduction

Interactions between proteins and DNA occur regularly in biological systems and they are a fundamental part of the ecological process. However, at the molecular level the recognition protein and DNA mechanisms are not fully understood in great detail. Therefore, the study of DNA bases and proteins are still the subject of many theoretical and experimental investigations [1–6]. In a recent experimental study conducted by deVries et al. [7], it was found that the binding motif in the guanine-aspartic acid (Gua-Asp) complex closely resembles a Watson-Crick motif displayed in the guanine-cytosine base pair.

This is significant because, although the keto form was not observed in experiment, it is theorized that the specific bonding pattern facilitates the observation of the keto form of guanine. Free guanine is completely unobservable in its keto form due to a shortened excited state lifetime [7]. Because of these factors, protein-DNA complexes, and the observation of the base-amino acid interaction are essential to the understanding of specific recognition of DNA target sites by regulatory proteins [8, 9] because they control complex spatial and temporal patterns of gene expression in higher organism.

Furthermore, studies have shown that the base-amino acid interactions are formed primarily through hydrogen bonds and hydrogen bonding interactions are used heavily to predict the folding of biological complexes such as proteins [10]. Its importance stems from it directionality and modest bonding energies midway between strong and weak Van der Waals bond. For this reason, the hydrogen bond is characterized by a certain amount of charge transfer interaction, which could be determined and measured in a particular complex of DNA base pairs. Undeniably, hydrogen bonding plays a crucial role in maintaining the three-dimensional structure of proteins and is equally central in numerous aspects of biological functions. These hydrogen bonding interactions make substantial contributions to the specificity of protein-nucleic acid complexes [11]. Achieving a substantial understanding of the underlying chemistry that contributes to hydrogen's abilities to bond and exploiting this theoretically can lead to future understanding of sophisticated protein-DNA complex studies.

Since many prior studies of protein-base interactions have been completed in order to determine how certain sequences of amino acids can recognize DNA target sequences, the complex of Gua-Asp seems to be a synthetic depiction of a real model system interaction occurring between base pairs and amino acids that can be modeled theoretically. In our findings, aspartic acid is shown to bind to the major groove of the Watson-Crick or the edge of the guanine. At the same time, the carboxylic acid and NH_2 groups of the aspartic acid remain accessible to form peptide bonds with other amino acids. This is a unique contradiction in that the protein data bank (www.pdb.org) shows 2,318 structures containing a protein and a nucleic acid, 208 of the structures have contacts between guanine and aspartic acid within a distance of 3 Å. However, 34.6% of the cases show aspartic acid is bound to the Watson-Crick edge of guanine; 47.1% of the contacts are between aspartic acid and phosphate groups, 11.5% are between aspartic acid and sugar, and the remaining 5.3% show aspartic acid being bound to the sugar edge of guanine.

In our study, we have evaluated possible complex structures of keto-9H guanine and the zwitterionic form of aspartic acid with the $-NH_3$ group where the R groups are negatively charged, (at neutral *pH*) to establish optimal hydrogen bonding patterns between the two interacting molecules. The guanine has six potential binding sites for aspartic acid where hydrogen bonding can form between the two monomers. We have obtained the most stable complex and verified hydrogen bond formation by relative energetic, IR spectrum analysis, and thermochemical data of Gua-Asp complex formation. The aim of this work is to: (1) obtain the most stable complex of Gua-Asp formed by hydrogen bonds, (2) show that our artificial optimized complexes are similar to the Watson-Crick motif that can be displayed in the guanine-cytosine base pair, and (3) provide insight into the underlying chemistry of nucleic acids and proteins through hydrogen binding sites by analyzing thermochemical properties ΔH and ΔG and the theoretical IR spectra.

Theoretical methods and computational details

The Gua-Asp complexes investigated were modeled with GaussView 3.0 [12] and Chemcraft [13] visualization programs. Using the Gaussian 03 program package [14], the electron correlation effects were described by density functional theory (DFT) [15] using the B3LYP method [16–18] in combination with the 6-31+G(d) basis set. Geometries were optimized followed by harmonic vibrational frequency calculations at the same level of theory to confirm all structures were at local minima on the potential energy surface. B3LYP/6-31+G(d) level calculations were performed because it has been cited for its reliability and accuracy of predicting hydrogen bonding interactions at minimal computational cost [19–22].

Interaction energies [23] including effects of basis set superposition error [24] (BSSE) were calculated to correct for any significant major energy differences caused by the basis functions of the complex formed by hydrogen bond interacting with monomers that are different from those employed for the isolated systems. Single point calculations of the optimized complexes were carried out in aqueous medium utilizing the polarizable continuum model (PCM) [25, 26] to observe solvent impact on our Gua-Asp complexes. To understand the nature of the interactions involved in our study, the "atoms in molecules" (AIM) program [27] was used in analyzing the bonding characteristics of the six systems studied in this work. Analyses of the HOMO-LUMO ($\Delta E_{H\text{-}L}$) gaps reveal the energy difference from one orbital to another in each of the six Gua-Asp systems (Fig. 1).

Results and discussion

Geometry and selected vibrational frequencies

Figure 2a-f shows the optimized geometries of the guanine-Asp systems obtained by modeling all possible hydrogen-bonding interaction between the carboxylic acid group of aspartic acid and the functional groups of the keto-9H guanine. No symmetry constraints were imposed during the optimization of each complex. A conformational analysis of the guanine-Asp complexes illustrates that the interactions between the two species contribute four functional groups, two from each molecule, to the hydrogen bond interactions. The distance between the C=O—OH group and N-H-–C=O group of the Gua-Asp complex differ between 1.6 and 2.8 Å, respectively. As one can see, our results clearly show formation of hydrogen bonding between guanine and aspartic acid. The hydrogen bond interaction distances are listed in Table 1. Further confirmation of this concept is illustrated by two coexisting systems with an energy difference of ~0.5 kcal mol^{-1}, complexes (b) and (a), as shown in Table 2. The most stabilized complex (b) has three hydrogen bond interactions, C=O—H-O with a bond length of 1.658 Å, N-H-–O=C with a bond length of 2.238 Å , and

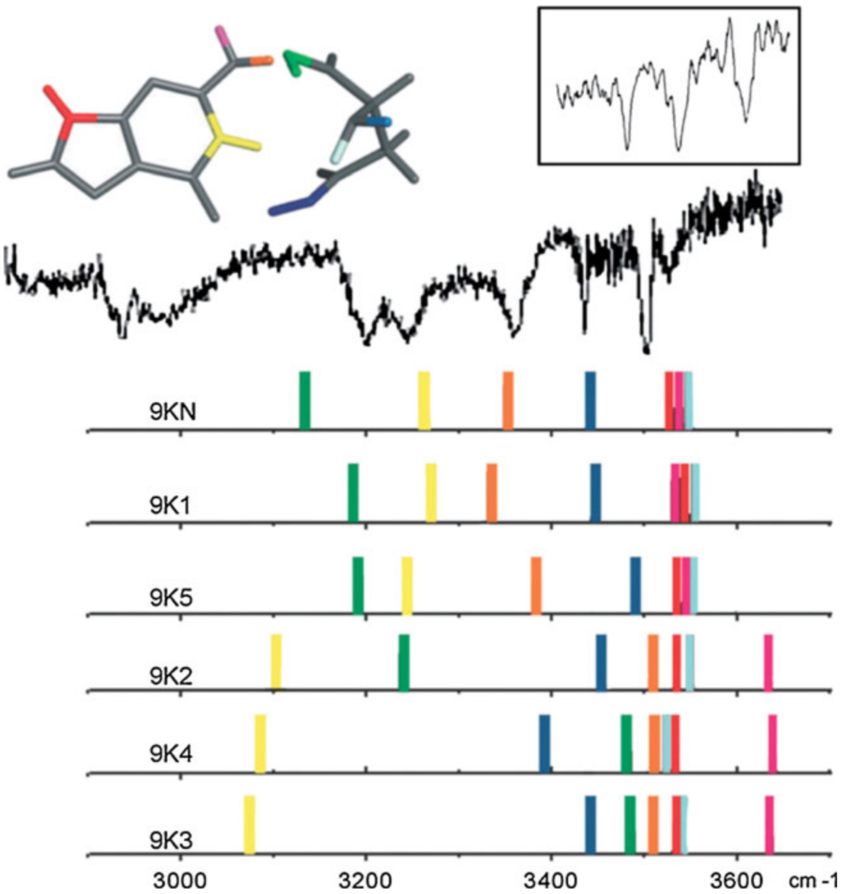


Fig. 1 IR hole-burning spectrum of isomer I, 9KN. Stick spectra show anharmonic frequencies calculated for the lowest energy keto structures. Color coding indicates modes as shown in the 9KN structure at the top. The spectrum in the inset is from protonated aspartic acid fragments [7]

N-H--O=C with a bond length of 2.008 Å. Complex (a) has two hydrogen bond interactions, C=O—H-O with a bond length of 1.633 Å and N-H--O=C with a bond length of 1.814 Å, respectively. Although the bond lengths in complex (b) are much longer than complex (a), the three hydrogen bond interactions occurring in the system actually make the complex a more stabilized unit since there is one more hydrogen bond interaction occurring in the system, which greatly contributes to the overall stability of complex (b). This is very significant because formal DNA structures display hydrogen bonds that contribute to the overall stability of the double helix structure, and three hydrogen bonds occur between the guanine-cytosine base pairs [28]. Likewise, our predicted structures bear a resemblance to the guanine-cytosine base pair.

IR spectra

To validate our study and for future interpretation of experimental work the theoretical IR spectra reveals that there is substantial interaction and complex formation occurring between guanine and aspartic acid via specific significant shifts. Figure 3a corresponds to complex (a) and Fig. 3b corresponds to complex (b), our lowest energy complex. The theoretical IR spectra exhibit both monomers and the complex formed. As seen in both of the complexes' spectra, there are additional peaks in the area between 2800 cm^{-1} and 3300 cm^{-1}, and no peaks are found in this area for guanine or aspartic acid. In the hydrogen bonding formation of complex (a), the frequency of ω(C=O) stretching shifted slightly to the left in the spectrum to 1758 cm^{-1} with an increase of intensity from 737 km mol^{-1} in pure guanine to 1251 km mol^{-1} in complex (a) as well as a strong intensity of 2255 km mol^{-1} of ω(C-O—H-O) stretching occurring at 2940 cm^{-1}. The ω(C=O—H-O) stretching of complex (a) is blue shifted about 500 cm^{-1} when compared to pure Asp theoretical IR spectrum. Formation of ω(N-H—O) stretching occurring at 3279 cm^{-1} is confirmed by a higher intensity at 1516 km mol^{-1}, which is a blue shifted peak of ω(N-H) stretching at 3580 cm^{-1} in the pure Gua region which has a weak intensity of 44 km mol^{-1}. In complex (b) the frequency of ω(C=O) stretching shifted to 1743 cm^{-1} with an intensity of 807 km mol^{-1}, while the ω(C=O—H-O) stretching occurred at 3040 cm^{-1} with a high intensity of 1921 km mol^{-1}, giving some indication of formation. In the ω(N-H—O) stretch occurring at 3544 cm^{-1}, the intensity drastically decreased to 574 km mol^{-1} as compared to complex(a) ω(N-H—O) stretching that occurs at 3279 cm^{-1}, with a strong intensity of 1516 km mol^{-1}.

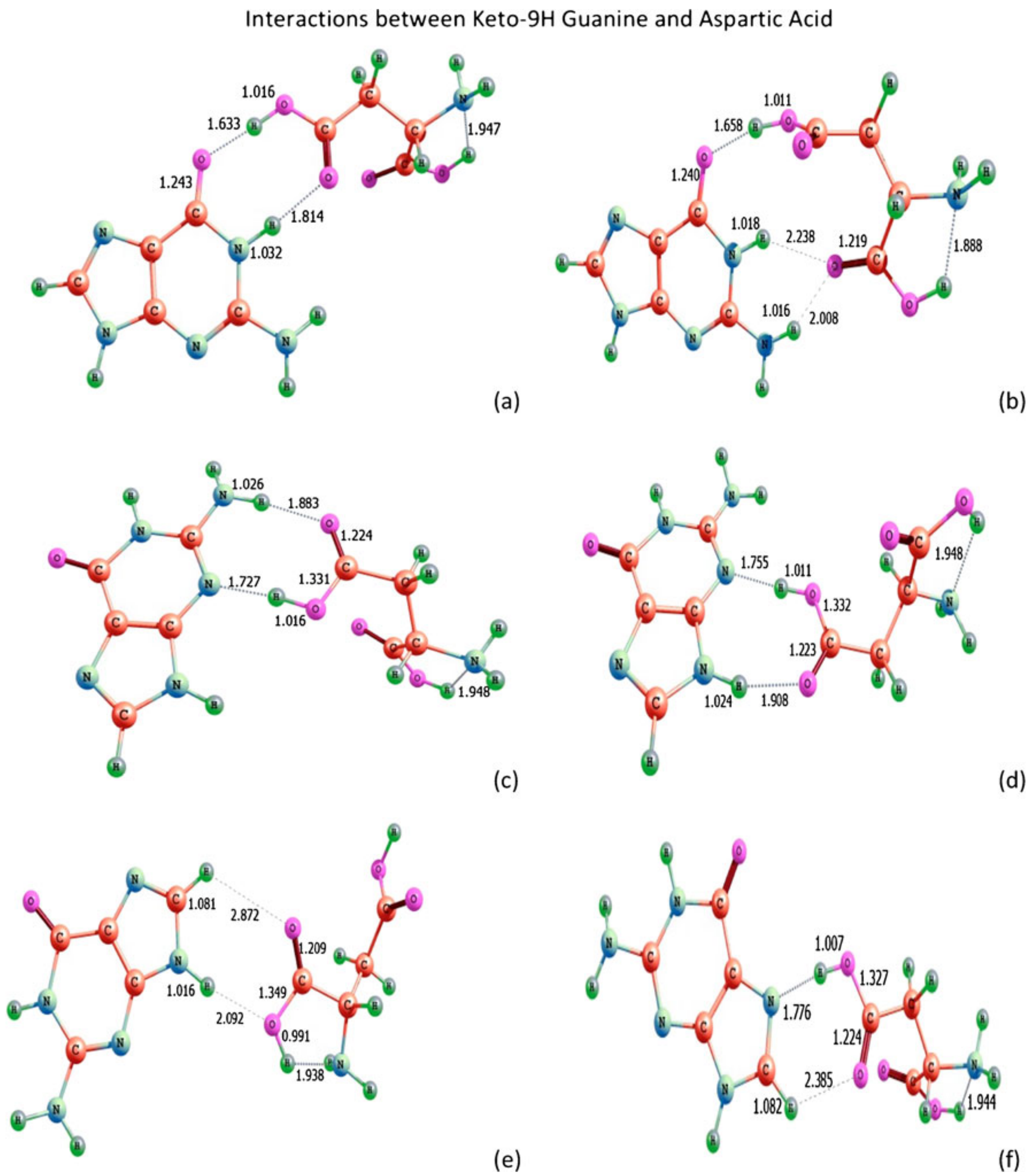


Fig. 2 Complexes (**a**-**f**) of keto-9H guanine-aspartic acid at B3LYP/6-31+G(d) interactions between keto-9H guanine and aspartic acid

There is also a blue shifted peak of the ω(N-H) stretching in the pure Gua theoretical IR spectrum indicating a decrease in frequency for complex (b) for ω(N-H—O) stretching and the same for complex (a).

Relative energies

The relative energies are shown in Table 2. The energy difference between our two lowest energy complexes (b) and (a) is only 0.5 kcal mol^{-1}. The energy difference between the two lowest energy complexes and the third complex is significantly higher, almost 6 kcal mol^{-1}. On the other hand, complex (e) exhibits an extremely higher energy relative to complex (b) by 13 kcal mol^{-1}. This can be rationalized by acknowledging that this area of binding is the particular site where the five-carbon sugar binds to the guanine and requires a substantial amount of energy. Therefore, since this is the area where the sugar binds, there will be no hydrogen bond interaction between the base and the amino acid in this vicinity of the guanine.

When the solvent effect is introduced, this changes the order of relative stability, as shown in Table 2. Complex (c) is lowest in energy, while complex (f) is the second lower energy, by 0.3 kcal mol^{-1}. However, complex (e), as in the

Table 1 Hydrogen bond distances, electron density, and HOMO-LUMO gap

Complex	HB distance(Å)	ρ (BCP) (a.u.)	ΔE HOMO-LUMO (eV)
A	d(O···H-O)=1.633 d(N-H···O)=1.814	0.156 0.112	5.18
B	d(O···H-O)=1.658 d(N-H···O)=2.238	0.155 0.044	4.62
	d(H-N-H···O)=2.008	0.075	
C	d(H-N-H···O)=1.883 d(N···H-O)=1.727	0.095 −1.42	5.08
D	d(N···H-O)=1.755 d(N-H···O)=1.908	0.116 0.092	5.05
E	d(C-H···O)=2.872 d(N-H···O)=2.092	0.017 0.602	4.35
F	d(N···H-O)=1.776 d(C-H···O)=2.385	0.116 0.042	5.37

gas phase has an extremely higher energy of 3.98 kcal mol^{-1}. Overall, the energies decreased drastically when solvent effect is introduced to the systems, with the exception of complex (e). Water readily competes with itself for hydrogen bonding. As a result, these hydrogen bonded systems in the solvent make it more difficult for the existing hydrogens to bond to the Gua-Asp complexes. Therefore, according to the relative energies, the solvent's effect does have an adverse effect of stabilizing the formation of complexes. This is a very important finding because now our results show that the hydrogen bond is not the only force that contributes to stability, since many of these reactions occur in our bodies in solvent conditions.

Interaction energies

The complex of k9H-guanine and aspartic acid indicate a few binding sites as previously discussed. The calculated interaction energies of the six modeled interactions are presented in Table 3. The energy of interaction are defined as $\Delta E = E_{AB} - E_A + E_B$, where E_{AB} is the energy of the complex (gua-Asp) and E_A and E_B are the energies of the monomers, keto9H-guanine and aspartic acid. According to our calculated interaction energies, the strongest level of interaction occurs in complex (b) with interaction energy of −19.50 kcal mol^{-1}. The next strongest interaction occurs in complex (a) with interaction energy of −19.00 kcal mol^{-1} and the energy differs by 0.50 kcal mol^{-1}. The energies of these complexes are in very close proximity of each other. The explanation for this strikingly small energy difference between complexes (a) and (b) is that both complexes are formed with aspartic acid bonding to the C=O and N-H bonds of guanine and this makes sense given that oxygen and nitrogen are more electronegative, thus pulling electron density away from the hydrogen.

Interestingly enough, we found that there were no significant proton transfers between the interacting monomers of Gua-Asp. There was only proton transfer in the zwitterionic form of aspartic acid, the NH_3 donated a proton to the negatively charged R group, C=O, which resulted in the aspartic acid transforming from the zwitterionic state to an unionized state.

The BSSE corrected energies (calculated as the difference between the total energy of the dimer minus the sum of the total energies of the two monomers), were calculated to account for energy difference. Based on our calculations, there was no significant energy loss/gain. For example, in complex (a) the interaction energy calculated without BSSE is −19.00 kcal mol^{-1} and with BSSE counterpoise corrections, the interaction energy is −17.81 kcal mol^{-1}, which is 1.19 kcal mol^{-1} difference, and therefore, the BSSE corrections are negligible for each system given the small ΔE_{BSSE}.

Table 2 Relative energies of Gua-Asp complexes (a-f) (kcal mol^{-1}) at B3LYP/6-31+G(d)

	a	b	c	d	e	f
ΔE(gas)	0.49	0.00	5.72	5.61	13.58	6.82
ΔE(solv)	0.71	1.90	0.00	1.59	3.98	0.27

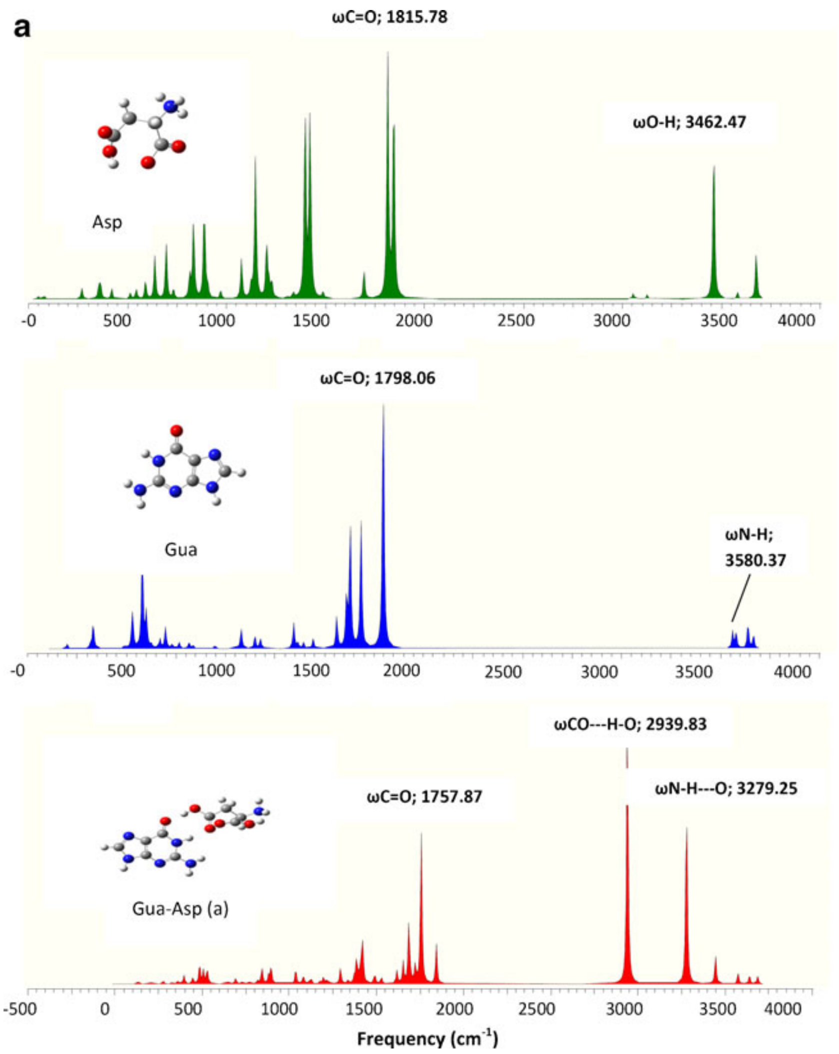
*Relative energies have 0.00 for gas phase

*Relative energies have 0.00 for solvent phase

Electron density and HOMO-LUMO gap

In order to understand the nature of the hydrogen-bonding interactions occurring in our systems of interest, we utilized the atoms in molecules (AIM) theory to evaluate the bonding characteristics. This theory allows one to quantitatively evaluate the nature of bonding in a molecule on the basis of the topological analysis of the electron charge density. Table 1 shows the hydrogen bond distances, the electron density, ρ(BCP), which is the Rho for the bond critical points and tells where the electrons are concentrated and the HOMO-LUMO gaps analyzed. According to our

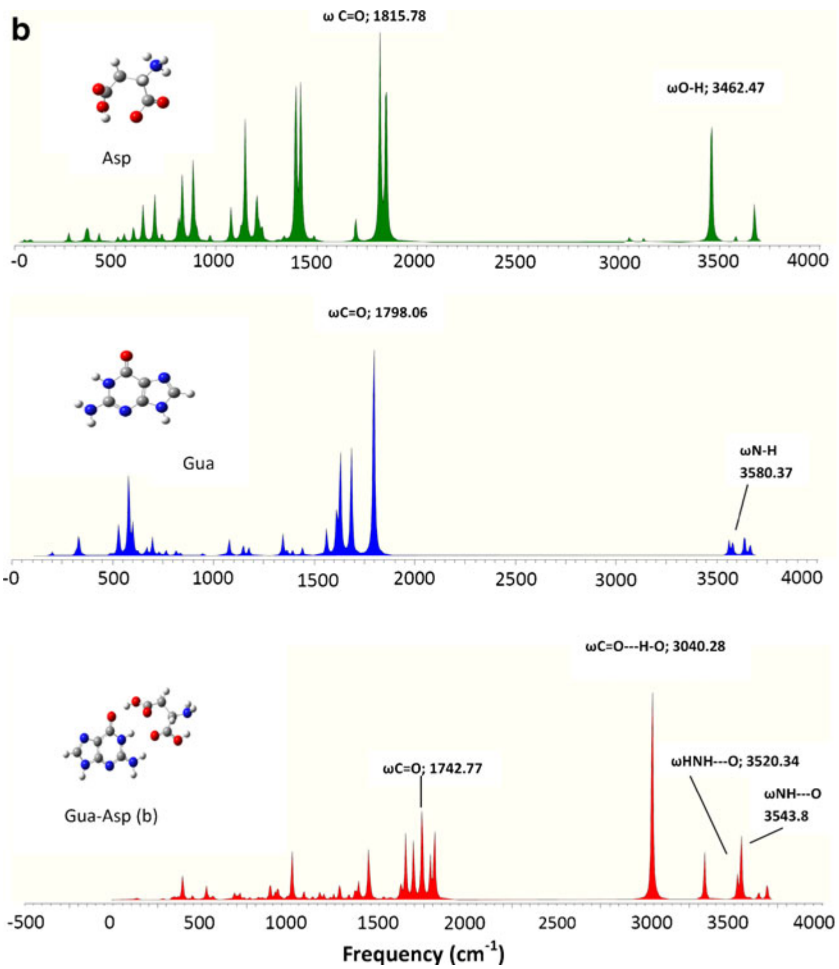
Fig. 3 (**a**) Complex A theoretical IR spectra of aspartic acid, guanine, Gua-Asp at B3LYP/6-31+G(d). (**b**) Complex B Theoretical IR spectra of aspartic acid, guanine, Gua-Asp at B3LYP/6-31+G(d)



hydrogen bond distances, the shorter bond lengths indicate that there are more electrons centrally located at that particular bonding site, indicating a strong interaction occurring between guanine and aspartic acid. Regarding the characteristics of the electron density, in complex (a) for the O-–H-O bond, the electron density is 0.156 a.u. indicating that there are more electrons concentrated at this bonding site. The hydrogen bond length is shorter, and the interaction is stronger than the N-H-–O bonding site of complex (a), with an electron density of 0.112 a.u. In complex (b), for the O-–H-O bond, the electron density is 0.155 a.u. indicates more electrons are concentrated at this bond critical point, as well as a shorter bond length, and stronger interaction than the N-H-–O bond. This was found to be the weakest interaction compared to the O-–H-O bond and H-N-H—O bond, which has an electron density of 0.075 a.u. of complex (b). In complex (e) the N-H-–O bond has the strongest interaction and more electrons concentrated at its' bond critical point than any of the N-H-–O bond interactions which occur in complexes (a) and (b). The AIM calculations revealed that our interactions are not covalent in nature and that they exhibit more van der Waals bonding properties. More specifically, the complexes favor an electrostatic type of interaction. The suitable balance of hydrogen bonding and van der Waals interactions is needed for the creation of three-dimensional structures and other hydrogen bonded systems [29]. Major advances in both theoretical and experimental methods for studying van der Waals in the last two decades have been reviewed extensively.

Figure 4 shows pictures of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) models of the Gua-Asp complexes (a) and (b). The HOMO is the part of the molecule that is capable of absorbing a photon and it is also the σ (sigma) and bonding molecular orbital which is located on the guanine. Once the photon is absorbed the electron density migrates to other parts of the complex, normally to the LUMO which is the σ* (sigma star) and antibonding orbital located on the aspartic acid. According to the

Fig. 3 (continued)



HOMO-LUMO analysis in Table 1, complex (E) required the least amount of energy for the electron to transfer back and forth between the highest occupied molecular orbital, and the lowest unoccupied molecular orbital. According to our calculated interaction energies in Table 3, complex (E) has the weakest interaction and therefore it requires the least amount of energy for hydrogen bonding to occur according to the $\Delta E_{\text{HOMO-LUMO}}$ difference. At the same time, complex (B) also requires a small amount of energy for the electron to transfer back and forth between the two orbitals, hence our lowest energy complex. However, complex (F) at 0.199 a.u., requires more energy than any of the complexes for an electron to transfer between the HOMO and LUMO orbitals, indicating a larger gap and electrostatic interaction.

Conclusions

In the present article, we have shown that our artificial complexes are similar to the Watson-Crick motif displayed in the guanine-cytosine base pair and we have obtained the most stable complexes of Gua-Asp that are verified hydrogen bond formations by relative energetic, theoretical IR spectra analysis and thermochemical data. DFT calculations provide reliable means of determining the strength of the hydrogen-bonding interactions and to demonstrate that our complexes of guanine and aspartic acid exhibit hydrogen bond activity. The electron density at the bond critical points of each system determined strong hydrogen bonding interactions. Our results show that the interactions

Table 3 Interaction energies, BSSE corrections, enthalpies, and free energies of complexes (a-f) (kcal mol^{-1}) at B3LYP/6-31+G(d)

	a	b	c	d	e	f
E_{int}(gas)	−19.00	−19.50	−13.77	−13.89	−5.91	−12.67
BSSE (kcal/mol)	1.19	1.54	1.21	1.23	0.68	0.98
E_{int}(solv)	−2.23	−1.04	−2.95	1.35	1.02	−2.67
ΔH(gas)	−0.39	0.00	5.44	5.49	−13.55	−6.81
ΔG(gas)	0.00	−0.42	−4.66	−4.85	−3.28	−5.47

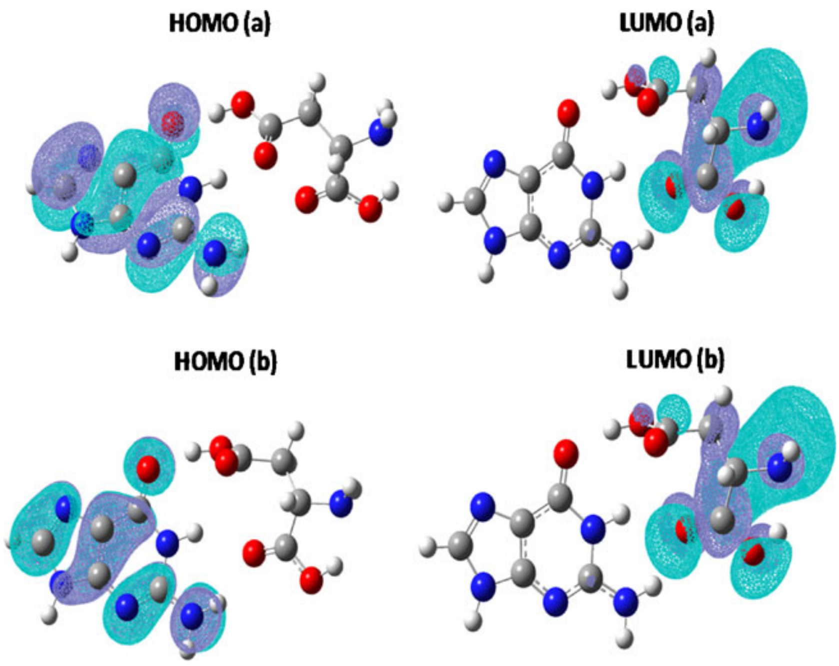


Fig. 4 HOMO-LUMO of the guanine-aspartic acid complexes (**a**), (**b**)

are more van der Waals type bonding rather than a strong dipole-dipole interaction. Major advances in both theoretical and experimental methods for studying van der Waals in the last two decades have been reviewed extensively [30–32]. Significant hydrogen bonding has been observed and proven to be an interaction that can stabilize these systems as well as the reorientation of the functional groups. B3LYP/6-31+G(d) results predict complex (b) and complex (a) are our lowest energy systems in the gas phase and they both bind the guanine at the C=O and N-H sites. The electronegativity of the oxygen contributes to both of these complexes having relatively low energies. At the same level of theory, single point calculations performed in the solvent phase exhibit no effect on stabilization energies. Moreover, solvent phase calculations contributed to a decrease in energy and changed the order of stability for the complexes, making (c) the lowest energy complex, rather than (b) as predicted in the gas phase. This was very notable to highlight in our study because our results now show that the hydrogen bond is not the only force that contributes to stability. The BSSE counterpoise correction had no discernible effect on the calculated interaction energies for the systems because the differences in all complexes were very minute. The optimized structures of complexes (b) and (a) resembles the guanine-cytosine Watson Crick base pairing in DNA very closely. The hydrogen bond patterns in complexes (b) and (a) are similar to the hydrogen bonding pattern in the G-C base pair. The aspartic acid is making contact on the Watson-Crick edge of guanine. As much as 34.6% of these contacts where the aspartic acid binds to the edge of the guanine accounts for this particular contact. The overall stability of complexes (b) and (a) can be of some contribution from the electronegativity of the oxygen on the guanine since it is the most electronegative atom and the best proton acceptor. Calculations of interaction energies revealed that complexes (b) and (a) were our lowest energy systems, although there was no significant proton transfer in any of the two systems. The theoretical IR spectra showed that there is electrostatic type of interaction occurring in the complexes. The HOMO-LUMO difference investigation allowed us to analyze the size of the gaps for each system to see how much energy was being utilized for the transfer of the electron from one orbital to another and complex (E) used the least amount of energy to transfer. Most importantly, our observations show that the binding motif in our complexes is similar to a Watson-Crick motif that can be seen in a guanine-cytosine base pair and other molecules such as amino acids are able to form complexes with DNA bases. In the absence of experimental data, a study such as this can aid in future interpretation experimental works. Additionally, this work may also have some future medicinal applications that can help gain insight into the mechanisms of an amino acid formation on a particular base as it may provide insight on ways to detect, prevent, and treat modern diseases.

Acknowledgments This work is supported by the Office of Naval Research (ONR) #N000 14-08-1-0324/P00002, the National Science Foundation awards Experimental Program to Stimulate Competitive Research (EPS) 0903787 and Center for Research Excellence in Science and Technology (CREST) HRD 0833178. The authors gratefully acknowledge and appreciate the University of Mississippi, Mississippi Center for Supercomputing Research for supercomputers Redwood, Mimosa and Sequoia.

References

1. Yanson IK, Teplitsky AB, Sukhodub LF (1997) Biopolymers 18:1149–1170
2. Fodor SP, Rava RP, Copeland RA, Spiro TG (1986) J Raman Spectrosc 17:471–475
3. Urabe H, Hayashi H, Tominaga Y, Nishimura Y, Kubota K, Zsuboi M (1985) J Chem Phys 82:531–536
4. Cocco S, Monasson R (2000) J Chem Phys 112:10017
5. Hobza P, Sponer JH (1999) Chem Rev 99:3247–3276
6. Wesolowski TA (2004) J Am Chem Soc 126:11444
7. Crews BO, Abo-Riziq A, Pluhackova K, Thompson P, Hill G, Hobza P, de Vries MS (2010) J Phys Chem Chem Phys 12:3597–3605
8. Luscombe NM, Laskowski RA, Thornton J (2001) Nucleic Acids Res 29:4294–4309
9. Mandel-Gutfreund Y, Margalit H, Jernigan RL, Zhurkin VB (1998) J Mol Biol 277:1129–1140
10. Monajjemi M, Mollaamin F, Karimkeshteh T (2005) J Mex Chem Soc 49:344–352
11. Cheng AC, Frankel AD (2004) J Am Chem Soc 126:434–435
12. www.gaussian.com (2000–2003) Gaussian Inc, Pittsburgh, PA
13. Zhurko GA, Zhurko DA. Chemcraft http://www.chemcraftprog.com
14. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JR, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian 03, Revision C.02. Gaussian Inc, Wallingford
15. Parr RG, Wang W (1984) Density functional theory of atoms and molecules. Oxford University Press, Oxford
16. Becke AD (1993) J Chem Phys 98:5648–5652
17. Vosko SH, Wilk L, Nusair M (1980) Can J Phys 58:1200–1211
18. Lee C, Wang W, Parr RG (1988) Phys Rev B 37:785–789
19. Sponer J, Leszczynski J, Hobza P (2002) Biopolymers 61:3–31
20. Murashov VV, Leszczynski J (2000) J Mol Struct 529:1–14
21. Holmes TM, Doskocz J, Wright T, Hill GA (2008) Int J Quantum Chem 109:119–123
22. Sponer J, Leszczynski J, Hobza P (1996) J Phys Chem 100:1965–1974
23. Jaguar Software, Version 4.1 (2001) Schrödinger Inc, Portland, OR
24. Boys SB, Bernardi F (1970) Mol Phys 19:553–559
25. Miertus S, Scrocco E, Tomasi J (1981) Chem Phys 55:117–129
26. Cammi R, Tomasi J (1995) J Comput Chem 16:1449–1458
27. Bader RFW (1990) Atoms in molecules. A quantum theory. Oxford University Press, Oxford
28. Voet D, Voet JG, Pratt CW (1999) Fundamentals of biochemistry. Biochemical interactions CD-ROM-J. Wiley and Sons
29. Suzuki M, Nakajima Y, Yumoto M, Kimura M, Shirai H, Hanabusa K (2003) Effects of hydrogen bonding and van der Waals interactions on organogelation using designed low-molecular-weight gelators & gel formation at room temperature. Langmuir 19:8622–8624
30. Buckingham AD, Fowler PW, Hutson JM (1988) Chem Rev 88:963–988
31. Wormer PES, Avoird AVD (2000) Chem Rev 100:4109–4143
32. Muller-Dethlefs K, Hobza P (2000) Chem Rev 100:143–167

J Mol Model (2012) 18:1993–2001
DOI 10.1007/s00894-011-1214-1

ORIGINAL PAPER

Studies on molecular structure and tautomerism of a vitamin B_6 analog with density functional theory

Suban K. Sahoo · Darshna Sharma · Rati Kanta Bera

Received: 24 June 2011 / Accepted: 7 August 2011 / Published online: 30 August 2011

Abstract This work presents a computational study on the molecular structure and tautomeric equilibria of a novel Schiff base **L** derived from pyridoxal (PL) and *o*-phenylenediamine by using the density functional method B3LYP with basis sets 6-31 G(d,p), 6-31++G(d,p), 6-311 G(d,p) and 6-311++G(d,p). The optimized geometrical parameters obtained by B3LYP/6-31 G(d,p) method showed the best agreement with the experimental values. Tautomeric stability study of **L** inferred that the enolimine form is more stable than its ketoenamine form in both gas phase and solution. However, protonation of the pyridoxal nitrogen atom (**LH**) have accelerated the formation of ketoenamine form, and therefore, both ketoenamine and enolimine forms could be present in acidic media.

Keywords DFT · Pyridoxal · Schiff base · Tautomeric equilibria · Vitamin B_6

Introduction

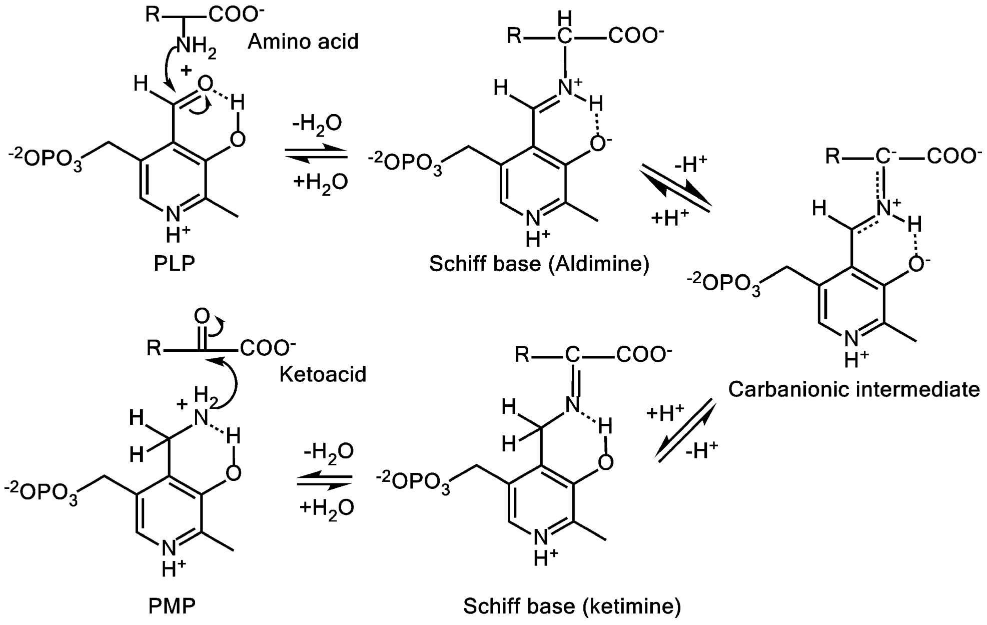
The vitamin B_6 cofactor pyridoxal-5′-phosphate (PLP) plays an important role in various enzymatic transformations of amino acids such as racemization, decarboxylation and transamination [1–6]. In such transformations, PLP initially forms a Schiff base with the ε-amino group of a lysine residue of the enzyme that subsequently evolves to the end-products. During the process, the Schiff base of PLP undergoes deprotonation of its Cα atom to give a carbanionic intermediate and on further protonation gives a ketoimine that is hydrolyzed to pyridoxamine-5′-phosphate (PMP) and the corresponding ketoacid (Scheme 1). This process finishes with the condensation of another ketoacid with PMP to form a carbinolamine which undergoes dehydration to a new Schiff base (a ketoimine). Finally, the Schiff base release a new amino acid and PLP is recovered [1–8]. Apart from the above mentioned general functions, vitamin B6 can also play a crucial role in protecting cells from oxidative stress because the vitamin has been shown to exhibit antioxidant activity [9, 10]. Furthermore, the antioxidant activity of vitamin B6 is found to be greater than that of vitamin C and E, though it is not classified as an antioxidant compound [9, 10].

Based on the above crucial biological processes, there is burgeoning interest for both theoretical [11–20] and experimental [21–23] chemists to explore the tautomerism in vitamin B_6 cofactors. On the theoretical aspect, Munoz and coworkers [11–13] have studied the Schiff base formation of vitamin B_6 analogues with B3LYP/6-31+G* method. Based on their DFT calculations on the electron charge distribution and '*electron-sink*' effect, they proposed that protonation of the pyridoxal nitrogen atom promotes the conversion of enolimine to ketoenamine [12], and also the Schiff base between PLP and an amine or amino acid requires a contribution of external water molecule in order to facilitate the transfer of proton [11, 13]. Kibura and Wong [19] explored the tautomeric equilibria of a series of 3-hydroxypyridine derivatives and reported that the neutral

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1214-1) contains supplementary material, which is available to authorized users.

S. K. Sahoo (✉) · D. Sharma
Department of Applied Chemistry,
SV National Institute of Technology (SVNIT),
Surat 395 007, Gujrat, India
e-mail: suban_sahoo@rediffmail.com

R. K. Bera
Department of Chemistry, Sant Longowal Institute of Engineering & Technology (SLIET),
Longowal, Punjab, India



Scheme 1 PLP to PMP interconversion

hydroxyl form is more stable than the zwitterionic oxo form in gas phase, but the stability influenced significantly by the effect of solvent polarity. Furthermore, recently some DFT studies have also been performed to elucidate the spectroscopic [15, 18] and antioxidant [17, 20] properties of vitamin B_6.

In this communication, we have performed a density functional study on molecular structure and tautomeric equilibria of a Schiff base (imine) **L** derived from pyridoxal and *o*-phenylenediamine at the B3LYP/6-31 G(d,p) level of theory in both gas phase and solution. Calculated properties such as relative energies, tautomeric equilibrium constant, atomic charges, and the highest-occupied molecular orbital (HOMO) and the lowest-unoccupied molecular orbital (LUMO) energies have been examined to explain the tautomerism.

Computational details

All calculations were performed with the GAUSS VIEW 5.0 visualization program and the GAUSSIAN 09 software packages [24]. It is well known that the B3LYP method has been successfully applied to elucidate the proton transfer reactions, particularly for tautomeric conversions [25–32]. Therefore, all DFT calculations were performed with a hybrid functional B3LYP [33] (Becke's three parameter hybrid functional using the LYP correlation functional) using the basis sets 6-31 G(d,p), 6-31++G(d,p), 6-311 G(d,p) and 6-311++G(d,p). The vibrational frequencies calculations showed no imaginary

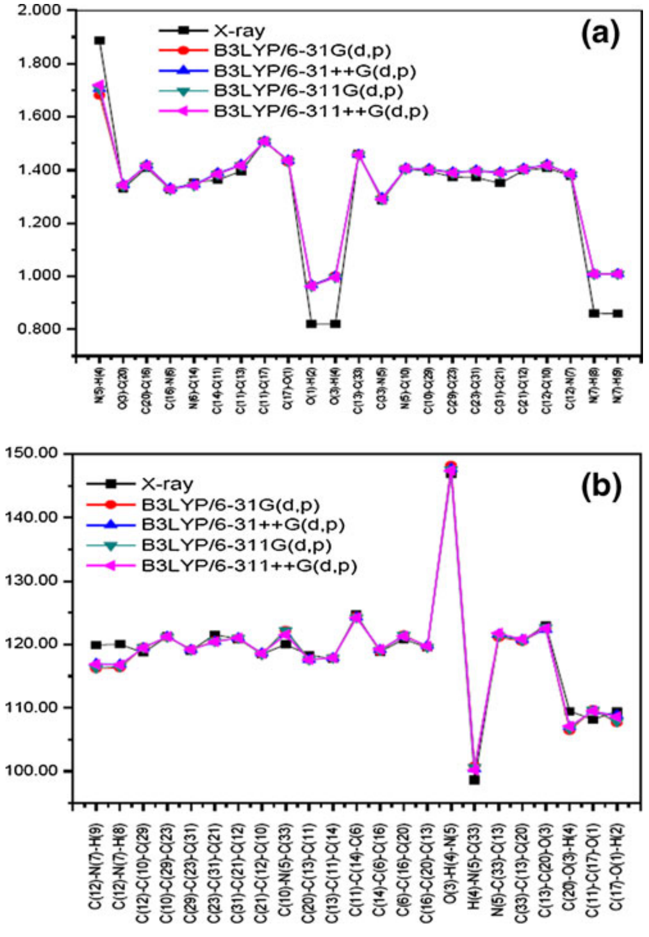


Fig. 1 The bond lengths (**a**) and the bond angles (**b**) differences from the theoretical values

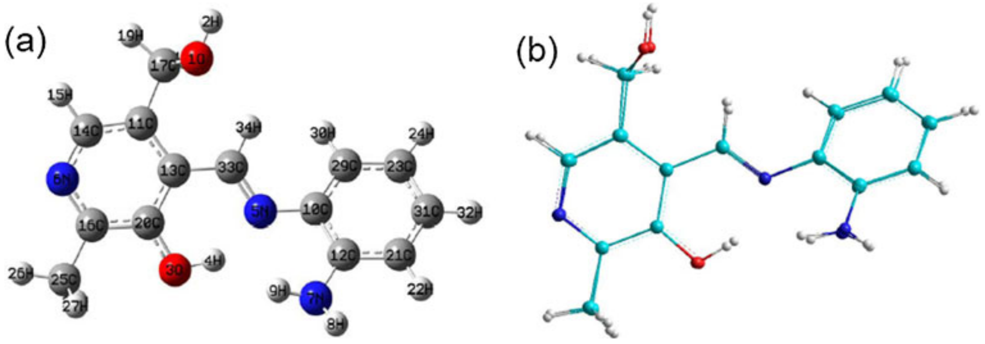


Fig. 2 (**a**) The optimized geometrical structure of the title compound **L** and (**b**) superimposition of the X-ray structure of **L** and its B3LYP/6-31 G(d,p) optimized counterpart

frequencies that ascertained the optimized structure were stable, and also provide various thermodynamic parameters to investigate the tautomeric stability. Some other properties such as total energy, HOMO and LUMO energies, Mulliken's atomic charges, and the chemical hardness [34] for the enolimine and ketoenamine forms of the compounds **L** and **LH** were obtained at B3LYP/6-31 G(d,p) level. These properties were also examined in solvent media with three different kinds of solvents (chloroform, ethanol and water) by using the conductor-like polarized continuum model (CPCM) [35, 36].

Results and discussion

Geometrical structure

The experimentally determined molecular structure of the compound **L** was obtained from Cambridge Crystallographic Data Centre (CCDC), and was used as the initial structure for different theoretical calculations. The crystal structure of the compound **L** is a monoclinic and space group is $P2_{1/n}$. The crystal structure parameters of **L** are a=12.9277(9) Å, b=13.4080(1) Å, c=14.7206(11) Å and

Table 1 Selected molecular structure parameters of **L** and **LH** obtained at B3LYP/6-31 G(d,p) level

Bond lengths (°)	Expt.	**L**	**LH**	Bond angles (Å)	Expt.	**L**	**LH**
O(3)-C(20)	1.331	1.342	1.328	C(12)-N(7)-H(9)	119.93	116.19	119.43
C(20)-C(16)	1.409	1.418	1.404	C(12)-N(7)-H(8)	120.04	116.33	118.19
C(16)-N(6)	1.325	1.330	1.352	C(12)-C(10)-C(29)	118.72	119.55	119.36
N(6)-C(14)	1.353	1.345	1.358	C(10)-C(29)-C(23)	121.26	121.17	121.06
C(14)-C(11)	1.363	1.387	1.377	C(29)-C(23)-C(31)	119.05	119.21	119.37
C(11)-C(13)	1.396	1.419	1.423	C(23)-C(31)-C(21)	121.55	120.54	120.99
C(11)-C(17)	1.508	1.508	1.513	C(31)-C(21)-C(12)	120.86	120.96	120.75
C(17)-O(1)	1.434	1.430	1.420	C(21)-C(12)-C(10)	118.51	118.54	118.45
O(1)-H(2)	0.819	0.967	0.967	C(10)-N(5)-C(33)	119.99	122.25	124.55
O(3)-H(4)	0.820	1.002	1.019	C(20)-C(13)-C(11)	118.34	117.62	119.29
C(13)-C(33)	1.461	1.456	1.450	C(13)-C(11)-C(14)	117.71	117.87	119.00
C(33)-N(5)	1.285	1.295	1.304	C(11)-C(14)-C(6)	124.70	124.39	120.32
N(5)-C(10)	1.405	1.404	1.383	C(14)-N(6)-C(16)	118.87	118.98	124.34
C(10)-C(29)	1.395	1.403	1.416	C(6)-C(16)-C(20)	120.83	121.52	117.44
C(29)-C(23)	1.374	1.391	1.379	C(16)-C(20)-C(13)	119.53	119.61	120.61
C(23)-C(31)	1.372	1.398	1.410	O(3)-H(4)-N(5)	146.82	148.19	149.70
C(31)-C(21)	1.350	1.391	1.382	H(4)-N(5)-C(33)	98.64	100.81	102.06
C(21)-C(12)	1.399	1.405	1.412	N(5)-C(33)-C(13)	121.37	121.10	119.67
C(12)-C(10)	1.408	1.420	1.431	C(33)-C(13)-C(20)	120.61	120.49	119.99
C(12)-N(7)	1.379	1.385	1.367	C(13)-C(20)-O(3)	123.03	122.52	122.46
N(7)-H(8)	0.861	1.009	1.009	C(20)-O(3)-H(4)	109.47	106.41	105.66
N(7)-H(9)	0.859	1.009	1.007	C(11)-C(17)-O(1)	108.18	109.72	108.40
N(5)-H(4)	1.886	1.681	1.608	C(17)-O(1)-H(2)	109.48	107.66	108.50
O(3)-N(5)	2.611	2.587	2.541				

V=2518.0(3) Å^3 [37]. The optimized parameters (bond lengths and bond angles, Table 1S) of **L** were obtained by applying exchange-correlation energy function B3LYP and different basis sets, 6-31 G(d,p), 6-31++G(d,p), 6-311 G(d,p) and 6-311++G(d,p). The differences of computed bond lengths as well as selected bond angles with corresponding experimental values are shown in Fig. 1. The calculated bond lengths with the basis sets 6-311 G(d,p) and 6-311++G(d,p) were found to be slightly shorter than the 6-31 G(d,p) and 6-31++G(d,p), and the reduction of bond lengths was more pronounced at B3LYP/6-311 G(d,p) level as compared to others. Further, in order to account for the accuracy of the different theoretical approaches, the optimized structures were superimposed with that obtained from X-ray crystallography that resulted in root-mean-square error (RMSE) of 0.191Å, 0.221Å, 0.197Å and 0.221Å with respect to structure obtained at B3LYP/6-31 G(d,p), B3LYP/6-31++G (d,p), B3LYP/6-311 G(d,p) and B3LYP/6-311++G(d,p) levels. The RMSE revealed that the optimization at B3LYP/6-31 G(d,p) level reproduces the geometry of the compound **L** (Fig. 2). For this reason, later in this work, to keep a reasonable computational time, all the calculations were performed at the B3LYP/6-31 G(d,p) level of theory.

The calculated geometrical parameters obtained for **L** and its protonated form **LH** at B3LYP/6-31 G(d,p) level are listed in Table 1 and compared with the experimental data of **L**. Except the hydrogen-bond length between N (5)·····H(4), the result in Table 1 inferred that the optimized bond lengths are slightly longer than the experimental values. The most probable reason for this is that the theoretical calculations are performed for an isolated molecule in gaseous phase, and the experimental results belong to the solid phase. In the solid phase, the existence of the crystal field along with the inter-molecular interactions such as van der waals interactions that connected the molecules together resulted in the difference of bond parameters between the calculated and the experimental values [38]. Such intermolecular interactions can be identified from the experimentally determined crystal packing diagram of the compound **L** (Fig. 3) that showed the presence of intermolecular H-bonding between pyridoxal alcoholic-OH of one molecule with the pyridoxal nitrogen

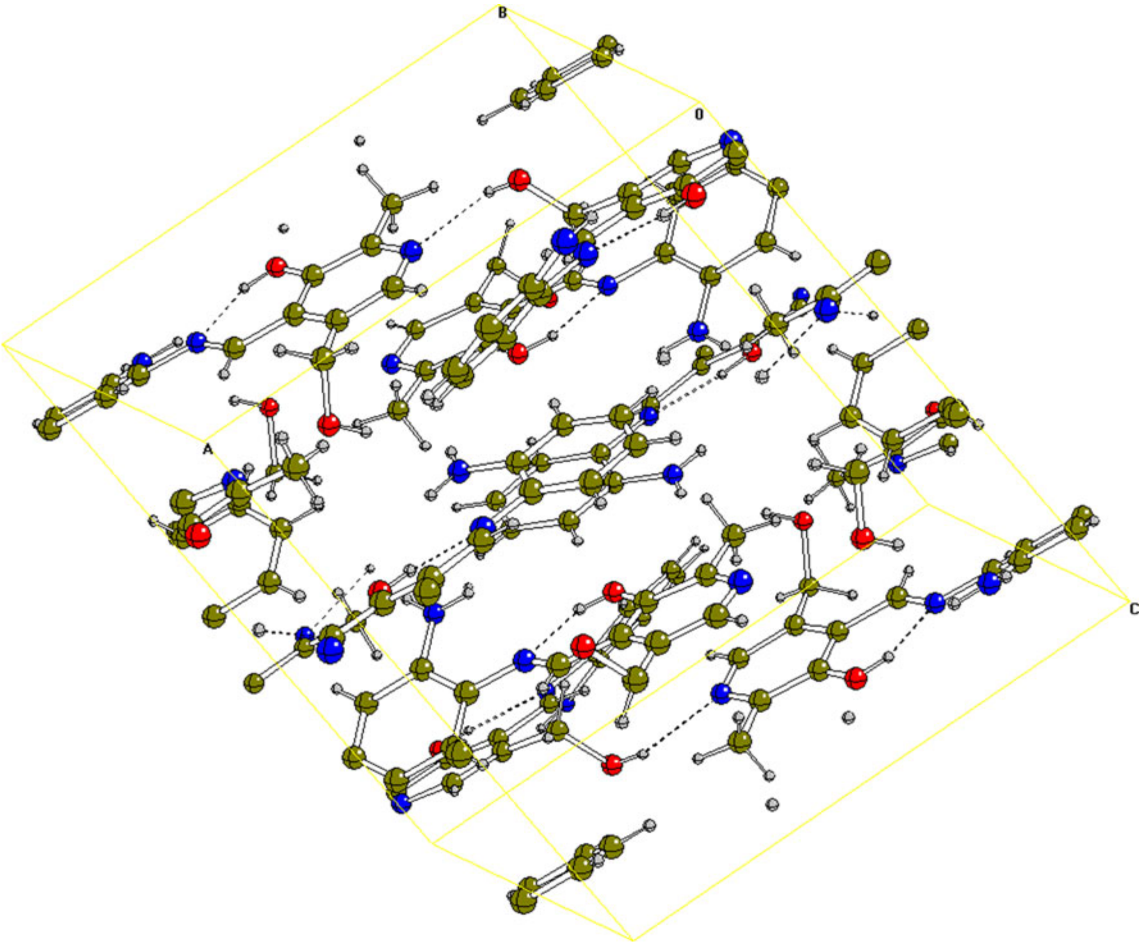


Fig. 3 Packing diagram of the title compound **L** showing H-bonding [37] and visualized by using the program MOLDRAW [39]

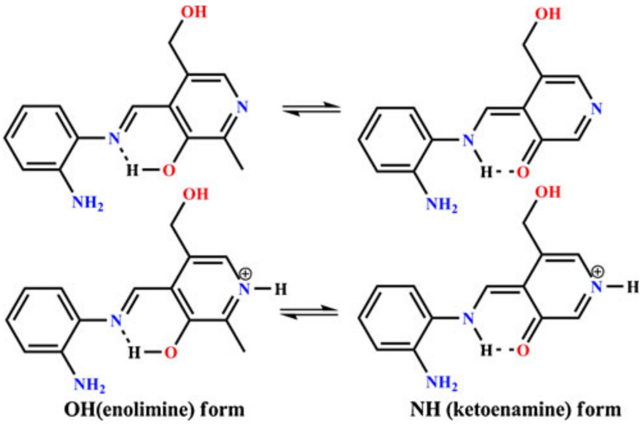


Fig. 4 The enolimine-ketoenamine tautomerism for the neutral **L** and protonated **LH** form of the title compound

atom of another, and intramolecular H-bond between N(5) and H(4). On the other hand, comparison between calculated bond angles and experimental data in Table 1 shows good agreement between then.

The optimized structural parameters of **LH** revealed that in the intramolecular O(3)-H(4)·····N(5) hydrogen bridge, the proton H(4) is located near the imine-nitrogen atom (H (4)·····N(5)=1.608 Å) and the O(3)·····N(5) distance 2.541 Å is substantially more compressed than the **L** (Table 1). The increase in the hydrogen bond strength and the decrease in O(3)·····N(5) distance in **LH** indicates the favourable formation of ketoenamine form. This can further be manifested from the shortening of the C(20)-O(3) distance of 1.328 Å and the increase of the C(33)-N(5) distance of 1.304 Å for **LH** compared to **L** that corresponds to an increased double-bond character of the first and an increased single-bond character of the second (Table 1).

Relative stability of tautomers

The enolimine (O-H···N) and ketoenamine (N-H···O) tautomers for the compounds **L** and **LH** are shown in Fig. 4. To investigate the tautomeric stability, optimization at B3LYP/6-31 G(d,p) level for both enolimine and ketoenamine forms of **L** and **LH** was performed in gas phase as well as solutions. In order to evaluate the solvent effect on ketoenamine-enolimine tautomerism, optimization calculations in three different solvents (water, ethanol and chloroform) were also performed at the same level of theory using the CPCM model. Some calculated physicochemical properties such as total energies, relative energies, HOMO and LUMO energies, and chemical hardness (η) are given in Table 2. The chemical hardness is quite useful to rationalize the relative stability and reactivity of the chemical species. The chemical hardness is approximated in terms of the energies of the HOMO and LUMO frontier molecular orbitals by applying equation, $\eta = (E_{LUMO} - E_{HOMO})/2$. Hard species having a large HOMO-LUMO gap will be more stable and less reactive than soft species having a small HOMO-LUMO gap [40].

Table 2 Calculated total energies, frontier orbital energies, relative energies and chemical hardness at B3LYP/6-31(d,p) level for the enolimine and ketoenamine form of the compounds **L** and **LH** (values of **LH** are in parenthesis)

	Gas phase (ε=1)	Chloroform (ε=4.9)	Ethanol (ε=24.55)	Water (ε=78.39)
	Enolimine form			
E_{TOTAL} (Hartree)	–857.24596933	–857.25723330	–857.26007987	–857.26055090
	(–857.64681492)	(–857.70246282)	(–857.71545928)	(–857.71757459)
E_{HOMO} (eV)	−0.19956	−0.20251	−0.20322	−0.20331
	(–0.30959)	(–0.23173)	(–0.21576)	(–0.21326)
E_{LUMO} (eV)	–0.07188	–0.07576	–0.07678	–0.07694
	(–0.22357)	(–0.13652)	(–0.11693)	(–0.11377)
η (eV)	0.064	0.063	0.063	0.063
	(0.043)	(0.048)	(0.049)	(0.050)
	Ketoenamine form			
E_{TOTAL} (Hartree)	–857.24001483	–857.25276933	–857.25619160	–857.25676803
	(–857.64695022)	(–857.70225080)	(–857.71507400)	(–857.71715407)
E_{HOMO} (eV)	–0.19453	–0.19948	–0.20085	–0.20107
	(–0.31538)	(–0.23836)	(–0.22199)	(–0.21939)
E_{LUMO} (eV)	–0.07836	–0.08296	–0.08427	–0.08449
	(–0.22717)	(–0.14118)	(–0.12193)	(–0.11884)
η (eV)	0.058	0.058	0.058	0.058
	(0.044)	(0.049)	(0.050)	(0.050)
ΔE(kcal/mol) [a]	3.74	2.80	2.44	2.37
	(–0.09)	(0.13)	(0.24)	(0.26)

[a] $\Delta E(\text{kcal mol}^{-1})=[E_{TOTAL}$ (ketoenamine form) - E_{TOTAL} (enolimine form)] X 627.5095

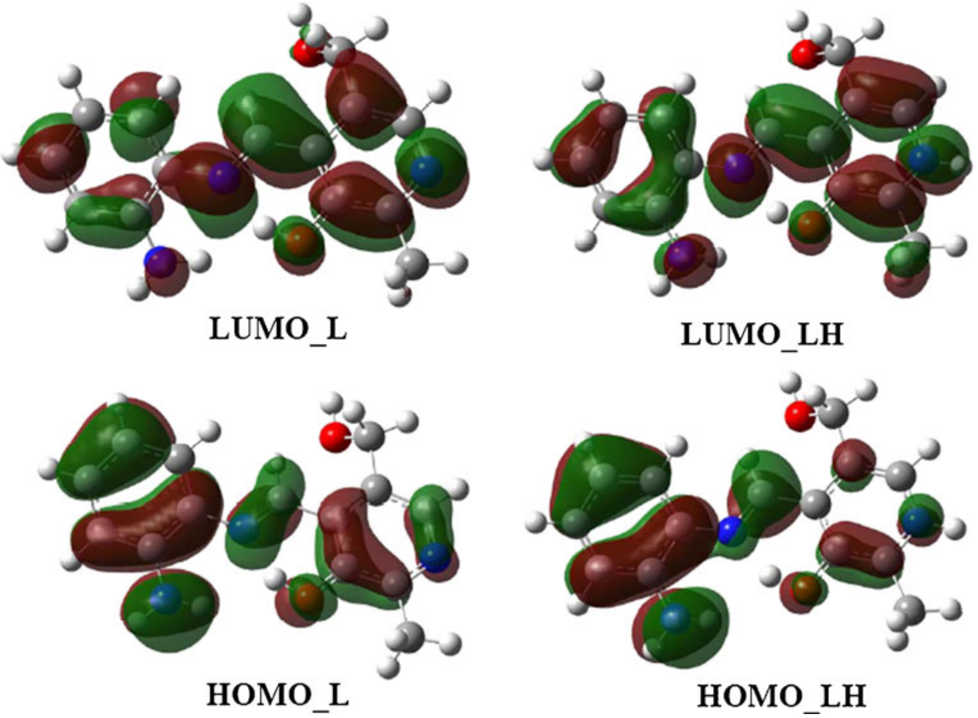


Fig. 5 HOMO and LUMO orbital pictures of **L** and **LH** computed at B3LYP/6-31 G(d,p) level in gas phase

According to Table 2, the total energy in gas phase calculated for the enolimine form of **L** is lower than the ketoenamine form while the chemical hardness of the enolimine form is greater than the ketoenamine one, which indicates that the enolimine form of **L** is more stable than its ketoenamine form. In solvent phase, the total molecular energies for both enolimine and ketoenamine form slightly decreases. The relative energies between the enolimine and ketoenamine forms decrease with the increase in solvent polarity, but the hardness of the enolimine form remains higher than the hardness of the ketoenamine form in all solvents. Therefore, the enolimine form of compound **L** is preferred over the ketoenamine form in both, gas phase and solution. However, in the case of **LH,** the relative energies between the enolimine and ketoenamine forms are reduced substantially as compared to **L** in both, gas phase and solution (Table 2). In gas phase, the ketoenamine form was energetically more stable by −0.085 kcal mol^{-1} and also the hardness of the ketoenamine form was found to be higher than the enolimine form. In solution, the hardness of the enolimine form was comparable with that of the ketoenamine and the relative energies increase from 0.13 kcal mol^{-1} to 0.26 kcal mol^{-1} with the increase in the solvent polarity. However, the difference in relative energies between the enolimine and ketoenamine forms for the protonated compound **LH** is much less than that computed for **L**.

Further, the 3D plots of the frontier orbitals HOMO and LUMO of **L** and **LH** calculated at B3LYP/6-31 G(d,p) in

Table 3 Calculated atomic charges of some important atoms of **L** and **LH** at B3LYP/6-31 G(d,p) level in gas phase and water[a]

Atom no.	L		LH	
	Gas phase	Water	Gas phase	Water
N(7)	–0.663	–0.676	–0.658	–0.671
N(6)	–0.480	–0.516	–0.550	–0.539
N(5)	–0.635	–0.631	–0.643	–0.634
O(1)	–0.577	–0.558	–0.519	–0.549
O(3)	–0.526	–0.593	–0.564	–0.576
H(4)	0.359	0.361	0.382	0.378
H(9)	0.270	0.281	0.278	0.288
H(8)	0.260	0.284	0.289	0.293
H(2)	0.311	0.335	0.334	0.344
C(10)	0.262	0.248	0.275	0.259
C(12)	0.276	0.258	0.308	0.276
C(29)	–0.099	–0.116	–0.084	–0.105
C(23)	–0.111	–0.130	–0.107	–0.128
C(31)	–0.084	–0.100	–0.072	–0.091
C(21)	–0.123	–0.139	–0.114	–0.133
C(13)	0.042	0.037	0.084	0.086
C(11)	0.071	0.059	0.040	0.031
C(14)	0.030	0.018	0.101	0.120
C(20)	0.275	0.276	0.273	0.274
C(16)	0.249	0.248	0.328	0.360
C(17)	–0.041	–0.035	–0.030	–0.028
C(25)	–0.351	–0.360	–0.361	–0.362

[a] See Fig. 2 for atoms numbering for the compounds

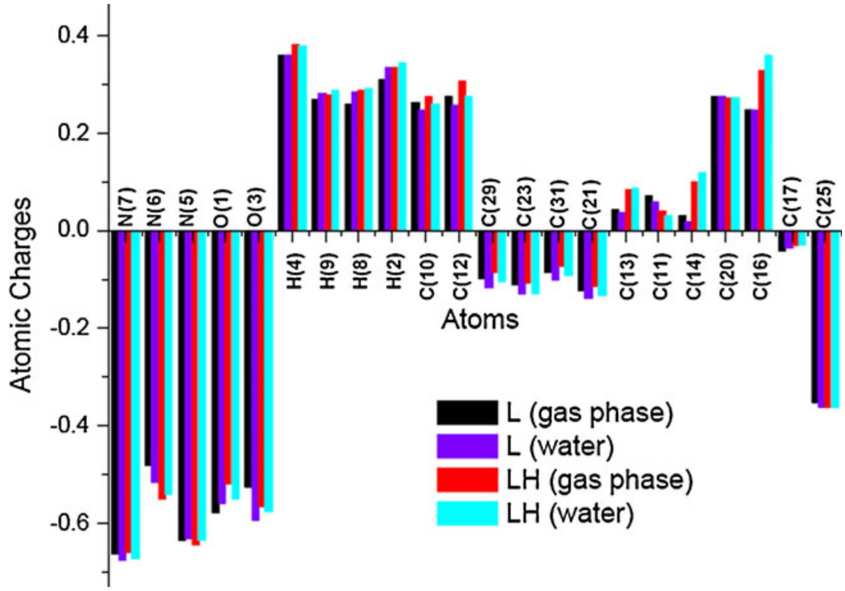


Fig. 6 Comparison for calculated atomic charges of **L** and **LH** at B3LYP/6-31 G(d,p) level

gas phase are examined (Fig. 5). The HOMO is the orbital that primarily acts as an electron donor and the LUMO is the orbital that largely acts as the electron acceptor [41]. The LUMO plots of both **L** and **LH** look alike. However, it can be seen from the figure that the HOMO in **L** is distributed uniformly between the two aromatic rings, whereas the HOMO in **LH** is found to be located more towards the *o*-phenylenediamine ring. Therefore, it can be concluded that transfer of electron density occurs from the pyridoxal ring to the *o*-phenylenediamine moiety upon protonation, *i.e.* intra-molecular charge transfer takes place within the molecule. The increase of electron density in *o*-phenylenediamine moiety in **LH** may favour the transfer of a proton towards the imine-nitrogen atom and facilitate the formation of ketoenamine form.

Mulliken's atomic charges

Atomic charges are used to describe the processes of electronegativity equalization and charge transfer in chemical reactions [42, 43], to model the electrostatic potential outside the molecular surfaces and for the relocation of the electron density of a compound [44]. Also, the local concentration and local depletion of electron charge density allows us to determine whether the nucleophile or electrophile can be attracted. Furthermore, the charge distribution can also play a vital role in tautomerism and therefore, the calculated Mulliken's atomic charges at the B3LYP/6-31 G(d,p) level were examined for the compounds **L** and **LH** both in gas phase and water. The calculated atomic charges are listed in Table 3 and represented in the graphical form in Fig. 6.

From Table 3, the gas phase results inferred that the atomic charges on the intramolecular H-bond bridge (O-H….N) between H(4), O(3) and N(5) atoms in **L** increases from 0.359, –0.526 and −0.635 to 0.382, –0.564 and −0.643 respectively on the protonation of the pyridoxal-nitrogen atom (**LH**). The increased positive charge of H(4) and the net negative charge of N(5) clearly demonstrated the favourable path for the transfer of a proton from O(3) to N(5). In water, the charge distributions are influenced due to dielectric effect and mostly showed higher values than in gas phase. The charges on H(4) and N(5) atoms increases from 0.361 and −0.631 to 0.378 and −0.534 respectively on protonation,

Table 4 Calculated thermodynamic parameters (E, H and G) in Hartrees for the compounds **L** and **LH** at B3LYP/6-31 G(d,p) level in gas phase and water[a]

Compounds		E	H	G
L	Enolimine form	–856.955262	–856.954318	–857.017994
		(–856.970200)	(–856.969255)	(–857.032927)
	Ketoenamine form	−856.949529	−856.948584	−857.012159
		(−856.966212)	(−856.965268)	(−857.028991)
LH	Enolimine form	−857.342595	−857.341651	−857.405653
		(−857.413235)	(−857.412291)	(−857.476330)
	Ketoenamine form	−857.342348	−857.341404	−857.405609
		(−857.412319)	(−857.411374)	(−857.475326)

[a] Values are in parenthesis

whereas the charges at O(3) slightly decreases from −0.593 to −0.576. The magnitude of the five carbon atomic charges of the pyridoxal ring of the compounds **L** and **LH** are found to be positive due to the more negative charge at the pyridoxal-nitrogen atom. Excepting the atoms C(10) and C(12) connected directly to the amine-N in in the *o*-phenylenediamine ring, the atoms C(29), C(23), C(31) and C(21) showed net negative charges. Furthermore, all the hydrogen atoms have a net positive charge (Table 3); in particular, the hydrogen atoms H(2) and H(4). The presence of large amounts of negative charge on the oxygen and nitrogen atoms, and the net positive charge on the hydrogen atoms H(2) and H(4) indicate the presence of both inter-molecular as well as intra-molecular hydrogen bonding in crystalline phase (Fig. 1).

Equilibrium constant of tautomers

To estimate the relative stabilities of tautomers, the vibrational analyses of the compounds **L** and **LH** are performed at B3LYP/6-31 G(d,p) level to get various thermodynamic properties (Table 4) and are given according to the formulas:

$$E_0 = E_{elec} + ZPE$$

$$E = E_0 + E_{vib} + E_{rot} + E_{transl}$$

$$H = E + RT$$

$$G = H - TS$$

Where, E_{elec}, *ZPE*, E_0, E, H, and G represent the total energy, zero point energy, corrected energy with *ZPE*, thermal energy, enthalpy and Gibb's free energy of the compound respectively.

One of the ways of getting the most stable tautomer is to calculate the tautomeric equilibrium constant. The calculated Gibb's free energies (Table 4) are used to calculate the equilibrium constant (K_T) between the tautomers by applying the equations $K_T = \exp(-\Delta G/RT)$ and $pK_T = -\log K_T$, where ΔG is the Gibb's free energy difference between the tautomers (enolimine ↔ ketoenamine) at the temperature 298.15 K and R is the gas constant. In gas phase, the ΔG value is calculated to be 3.66 kcal mol^{-1} and 0.03 kcal mol^{-1} between the enolimine and ketoenamine tautomers of **L** and its protonated form **LH**, respectively. In water, the ΔG values for **L** and **LH** are computed as 2.47 kcal mol^{-1} and 0.63 kcal mol^{-1} respectively. The lowering of ΔG values on the protonation of the pyridoxal-nitrogen atom is an indication for the favourable formation of the ketoenamine form. Further, the pK_T was calculated to determine the privileged direction of equilibrium. If the pK_T was positive, equilibrium moved from right towards the left and when it was negative, equilibrium moved from left towards the right. The calculated pK_T for **L** and **LH** in gas phase are 2.68 and 0.02, whereas in water are 1.81 and 0.46, respectively. The calculated pK_T for both **L** and **LH** are positive and favour the equilibrium to shift towards the left. However, the protonated compound **LH** is showing pK_T value near to zero and also appreciably lower than **L**, which inferred that the compound **LH** is favouring a ketoenamine form over **L**.

Conclusions

DFT study of the vitamin B_6 Schiff base analog **L** inferred that the enolimine form is more stable than its ketoenamine form both in gas phase and solution. However, protonation of the pyridoxal nitrogen influenced the tautomeric equilibria and accelerate the formation of the ketoenamine form. In gas phase, **LH** preferred the ketoenamine form over the enolimine, but with the increase in solvent polarity the relative stability of the enolimine form increases slightly from the ketoenamine form. Therefore, both the enolimine and ketoenamine forms of **LH** could be present in solvents. Other properties such as relative energies, the tautomeric equilibrium constant, atomic charges, and the HOMO and LUMO energies provide essential evidences for the favourable formation of the ketoenamine form on the protonation of the pyridoxal-nitrogen atom.

Acknowledgments This work was made possible by a grant from the Department of Science and Technology, New Delhi, India under Fast Track Young Scientist Scheme. The authors thank Prof. BK Kanungo (Sant Longowal Institute of Engineering & Technology, Longowal, India) for helpful discussions.

References

1. Christen P, Metzler DE (1985) Transaminases. Wiley, New York
2. Jansonius JN (1998) Curr Opin Struct Biol 8:759–769
3. Percudani R, Peracchi A (2003) EMBO Rep 4:850–854
4. Andrew CE, Kirch JF (2004) Annu Rev Biochem 73:383–415
5. Dolphin D, Poulson R, Avramovic O (1986) Vitamin B6 pyridoxal phosphate, chemical, biochemical, and medicinal aspects. Part A, Wiley, New York
6. Martell AE (1989) Acc Chem Res 22:115–124
7. Metzler DE, Ikawa M, Snell EE (1954) J Am Chem Soc 76:648–652
8. Lim YH, Yoshimura T, Kurokawa Y, Esaki N, Soda K (1998) J Biol Chem 273:4001–4005
9. Mooney S, Leuendorf JE, Hendrickson C, Hellmann H (2009) Molecules 14:329–351
10. Cabrini L, Bergami R, Fiorentini D, Marchetti M, Landi L, Tolomelli B (1998) Biochem Mol Biol Int 46:689–697
11. Salva A, Donoso J, Frau J, Munoz F (2003) J Phys Chem A 107:9409–9414
12. Casasnovas R, Salva A, Frau J, Donoso J, Munoz F (2009) Chem Phys 355:149–156

13. Ortega-Castro J, Adrover M, Frau J, Salva A, Donoso J, Munoz F (2010) J Phys Chem A 114:4634–4640
14. Adrover M, Vilanova B, Munoz F, Donoso J (2009) Bioorg Chem 37:26–32
15. Salva A, Frau J, Munoz F, Vilanova B, Donoso J (2003) Biochim Et Biophys Acta 1647:83–87
16. Zhang Y, Yang J, Fan H, Li C (2010) J Mol Struct 951:21–27
17. Matxain JM, Ristila M, Strid A, Eriksson LA (2006) J Phys Chem A 110:13068–13072
18. Ristila M, Matxain JM, Strid A, Eriksson LA (2006) J Phys Chem B 110:16774–16780
19. Kibura GSM, Wong MW (2003) J Org Chem 68:2874–2881
20. Mohajeri A, Asemani SS (2009) J Mol Struct 930:15–20
21. Vazquez MA, Munoz F, Donoso J, Blanco FG (1992) J Phys Org Chem 5:142–154
22. Witherup A, Abbptt EH (1975) J Org Chem 40:2229–2233
23. Sharif S, Powel DR, Schagen D, Steiner T, Toney MD, Fogle E, Limbach HH (2006) Acta Cryst B62:480–487
24. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA, Peralta Jr JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth G A, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas O, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ (2009) Gaussian 09, Revision A.1. Gaussian Inc, Wallingford, CT
25. Buemi G, Zuccarello F, Venuvanalingam P, Ramalingam M (2000) Theor Chem Acc 104:226–234
26. Sheikhshoaie I, Fabian WMF (2009) Curr Org Chem 13:149–171
27. Filarowski A, Koll A, Sobczyk L (2009) Curr Org Chem 13:172–193
28. Musin RN, Mariam YH (2006) J Phys Org Chem 19:425–444
29. Raissi H, Moshfeghi E, Jalbout AF, Hosseini MS, Fazli M (2007) Int J Quantum Chem 107:1835–1845
30. Lenain P, Mandado M, Mosquera RA, Bultinck P (2008) J Phys Chem A 112:10689–10696
31. Rybarczyk-Pirek J, Grabowski SJ, Malecka M, Nawrot-Modranka J (2002) J Phys Chem A 106:11956–11962
32. Nowroozi A, Raissi H, Farzad F (2005) J Mol Struct theochem 730:161–169
33. Becke AD (1998) Phys Rev A 38:3098–3100
34. Parr RG, Pearson RG(1983) J Am Chem Soc 105:7512–7516
35. Barone V, Cossi M (1998) J Phys Chem A 102:1995–2001
36. Cossi M, Rega N, Scalmani G, Barone V (2003) J Comput Chem 24:669–681
37. Back DF, de Oliveira GM, Lang ES, Vargas JP (2008) Polyhedron 27:2551–2556
38. Jian FF, Zhao PS, Bai ZS, Zhang L (2005) Struct Chem 16:635–639
39. Ugliengo P, Viterbo D, Chiari G (1993) Z Kristallogr 207:9–23
40. Ozbek N, Kavak G, Ozcan Y, Ide S, Karacan N (2009) J Mol Struct 919:154–159
41. Fukui K (1982) Science 218:747–754
42. Maksic ZB (1991) Theoretical Model of chemical bonding, Part 3. Springer, Berlin
43. Fliszar S (1983) Charge distributions and chemical effects. Springer, New York
44. Smith PE (1991) J Am Chem Soc 113:6029–6037

J Mol Model (2012) 18:2003–2011
DOI 10.1007/s00894-011-1221-2

ORIGINAL PAPER

Characteristics and nature of the intermolecular interactions in boron-bonded complexes with carbene as electron donor: an *ab initio*, SAPT and QTAIM study

Mehdi D. Esrafili

Received: 1 July 2011 / Accepted: 10 August 2011 / Published online: 30 August 2011

Abstract We report geometries, stabilization energies, symmetry adapted perturbation theory (SAPT) and quantum theory of atoms in molecules (QTAIM) analyses of a series of carbene–BX_3 complexes, where X=H, OH, NH_2, CH_3, CN, NC, F, Cl, and Br. The stabilization energies were calculated at HF, B3LYP, MP2, MP4 and CCSD(T)/aug-cc-pVDZ levels of theory using optimized geometries of all the complexes obtained from B3LYP/aug-cc-pVTZ. Quantitatively, all the complexes indicate the presence of B–$C_{carbene}$ interaction due to the short B–$C_{carbene}$ distances. Inspection of stabilization energies reveals that the interaction energies increase in the order NH_2 > OH > CH_3 > F > H > Cl > Br > NC > CN, which is the opposite trend shown in the binding distances. Considering the SAPT results, it is found that electrostatic effects account for about 50% of the overall attraction of the studied complexes. By comparison, the induction components of these interactions represent about 40% of the total attractive forces. Despite falling in a region of charge depletion with $\nabla^2\rho_{BCP}>0$, the B–$C_{carbene}$ bond critical points (BCPs) are characterized by a reasonably large value of the electron density (ρ_{BCP}) and $H_{BCP}<0$, indicating that the potential energy overcomes the kinetic energy density at BCP and the B–$C_{carbene}$ bond is a polar covalent bond.

Keywords *Ab initio* · Carbene · QTAIM · Symmetry-adapted perturbation theory

M. D. Esrafili (✉)
Laboratory of Theoretical Chemistry, Department of Chemistry,
University of Maragheh,
Maragheh, Iran
e-mail: esrafili@maragheh.ac.ir

Introduction

It is well-known that intermolecular interactions are very important in understanding organic, organometallic, and biomolecular structures, supramolecular assembly, crystal packing, reaction selectivity specificity, and drug-receptor interactions [1–4]. On the basis of these interaction forces, not only theoretical design but also experimental realization of novel functional molecules, nanomaterials, and molecular devices has become possible [5, 6]. Thus, the study of the fundamental intermolecular interactions and new types of interactions are very important for aiding self-assembly synthesis and nanomaterials design as well as for understanding molecular cluster formation [7, 8].

Carbenes are neutral compounds featuring a divalent carbon atom with only six electrons in its valence shell. Carbenes are in general highly reactive species with short lifetimes; consequently, very few examples of carbenes stable at room temperature are known [9]. In general, carbenes are classified as either singlets or triplets depending upon their electronic structure there [10]. Carbenes play important roles both as reactive intermediates and also as ligands; consequently, considerable effort has been devoted to understand their molecular and electronic structures [11–15]. Special interest is associated with carbenes that feature the attachment of donor groups to the carbenic carbon since they behave as nucleophiles and, in some instances, can be isolated. Whereas triplet carbenes exhibit radical-like reactivity, singlet carbenes are expected to show nucleophilic as well as electrophilic behavior because of the lone pair and vacant orbital. Hydrogen bond with singlet carbene as an electron donor has been confirmed due to presence of a free electron pair in the singlet carbene [16]. Among the most typical reactions of singlet carbenes are the rearrangements resulting from 1,2-shifts, dimerizations, [1+2]-cycloadditions to carbon-carbon double bonds, and insertions into C-H bonds [17–19]. The reactivity of transient singlet carbenes has recently enabled a wide variety of new carbon-carbon and carbon–metal bond-forming reactions to be developed [20–22]. Many other reactions involving singlet carbenes have been reported, including the formation of ylides with Lewis bases [23, 24]. Pioneering work on nucleophilic carbenes was carried out by Wanzlick and co-workers [25], who, in the early 1970s, predicted that imidazol-2-ylidene carbenes would possess enhanced stability due to the possibility of aromatic resonance and examined the reactivity patterns of these species as nucleophilic carbenes. Recently, a new kind of carbene-lithium binding in H_2C-LiX (X=H, OH, NH_2, CN, NC, CH_3, F, Cl, Br, C_2H_3, C_2H) complexes was predicted and characterized by Li et al. [26]. However, to the best of our knowledge, the study of the boron bond with carbene as an electron donor is rare.

Computational chemistry provides numerous methods to investigate the nature of intermolecular interaction, allowing, in many cases, estimation of the stabilization energy or other quantities related to this term. A very interesting tool for studies of intermolecular interaction is the decomposition of energy into particular contributions. One of the first successful schemes was that of Kitaura-Morokuma [27]. Symmetry-adapted intermolecular perturbation theory (SAPT) presents a viable alternative to the supermolecular approach [28–32]. In SAPT the interaction energy is calculated as the sum of terms of distinct physical origin, i.e., the first-order electrostatic and the second-order induction and dispersion energies, each of these terms being accompanied by a corresponding exchange correction due to the simultaneous exchange of electrons between the monomers. In the many-body version of SAPT the interacting monomers are described through Møller–Plesset or even coupled cluster theory, depending on the accuracy required for each individual interaction term. The quality of the total interaction energies compares with that obtained from CCSD(T). The interaction energy partitioning techniques are, however, usually global. From the set of theories with the capability of describing local variations of bonding, the quantum theory of atoms in molecules (QTAIM) approach [33] was also selected for the current study.

The aim of the present study is to analyze B–$C_{carbene}$ interactions for a wider spectrum of imidazol-2-ylidene carbene–BX_3 complexes, where X=H, OH, CN, NC, NH_2, CH_3, F, Cl, and Br (Fig. 1). Although carbene has two classes: singlet and triplet, we only consider the singlet carbene due to presence of a free electron pair in singlet carbene. The energy decomposition scheme is applied to gain more detailed insight into the nature of the interactions. In addition, the Bader theory has also been applied. One of the aims of this study is to answer the following questions: What is the nature of B–$C_{carbene}$ interactions in carbene–BX_3 complexes? Are there any sharply defined differences in the physical nature of these complexes? What is the substitution effect in these complexes? And are SAPT energies consistent with *ab initio* and DFT binding energies?

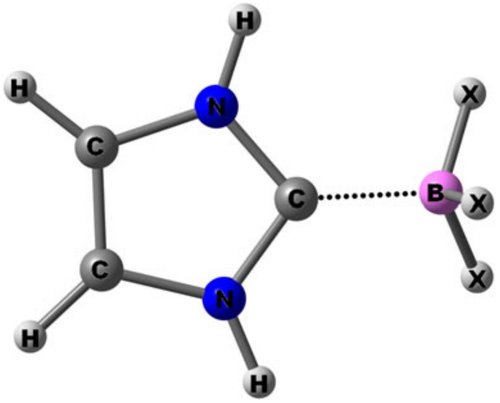


Fig. 1 Optimized structure of imidazol-2-ylidene carbene–BX_3 complex (X=H, OH, CN, NC, NH_2, CH_3, F, CL, Br)

Computational details

All molecular orbital calculations were performed using GAMESS suite of programs [34]. The geometry of the investigated various carbene–BX_3 complexes was optimized at the B3LYP level [35, 36] employing aug-cc-pVTZ basis set. Then corresponding frequency calculations were carried out at the same level to ensure that the optimized structures are true minima. The interaction energy for each cluster was calculated at the HF, B3LYP, MP2, and MP4 and CCSD(T) levels of theory by using the supermolecule method [37] which defines it as the difference between the energy of the cluster and those of the individual molecules in isolation:

$$E_{\text{int}} = E_{ijk\ldots}(ijk\ldots) - \sum_i E_i(i), \qquad (1)$$

where the terms in brackets denote the basis sets to be used. The results of Eq. 1 are subject to the basis superposition error, BSSE, as each molecule uses the basis set of the others in the cluster, decreasing the energy and resulting in overestimated interaction energies. This problem is usually overcome by using the counterpoise method of Boys and Bernardi [38], where all energies are calculated by using the basis set for the whole cluster, and the geometries of the monomers correspond to those they adopt in the cluster.

In this study, the DFT-SAPT calculations were carried out using the optimized geometries of all the complexes obtained from B3LYP/aug-cc-pVTZ method. Calculations were performed with aug-cc-pVDZ standard basis set. DFT-SAPT uses monomer properties and electronic densities from DFT in order to compute interaction energies using the SAPT [39–41]. This is the only variant of the SAPT methods that can be practically used for systems containing more than a few atoms and is, thus, the most useful for computations on biomolecular systems. Based on DFT-SAPT energy decomposition scheme [30], the two-body binding energy can be decomposed as:

$$E_{\text{int}}^{SAPT} = E_{pol}^{1} + E_{ex}^{1} + E_{ex-ind}^{2} + E_{ind}^{2} + E_{ex-disp}^{2} + E_{disp}^{2} \tag{2}$$

some of these terms can be combined in order to define values that correspond to commonly understood physical quantities. The terms are commonly combined as such:

$$\begin{aligned} E_{\text{elec}} &= E_{pol}^{1} \\ E_{\text{exch}} &= E_{ex}^{1} \\ E_{\text{ind}} &= E_{ind}^{2} + E_{ex-ind}^{2} \\ E_{\text{disp}} &= E_{disp}^{2} + E_{ex-disp}^{2} \end{aligned}$$

where E_{elec} is the first-order electrostatic term describing the classical columbic interaction of the occupied orbitals of one monomer with those of another monomer, E_{exch} is the repulsive first-order exchange component resulting from the antisymmetrization (symmetry adaption) of wave function, E_{ind} and E_{disp} correspond to induction and dispersion effects, respectively. The induction component is the energy of interaction of the permanent multipole moments of one monomer and the induced multipole moments on the other, whereas the dispersion part comes from the correlation of electron motions on one monomer with those on the other monomer. For binding energy decomposition analysis, molecular integrals were first obtained with the DALTON 2.0 package [42]; SAPT partitioning was then performed using the SAPT2008 program [43].

The QTAIM methodology [33] has been used to analyze the electron density of the systems considered at the B3LYP/cc-pVTZ computational level using AIM2000 program [44]. For atom–atom interactions such as intermolecular contacts or valence bonds, the characteristics of the corresponding bond critical point (BCP) of molecular charge density, ρ_{BCP} are very important. These are points where the electron density gradient $\nabla\rho_{BCP}$ vanishes and additional characterization is done using the corresponding Hessian matrix (a 3×3 matrix of second derivatives). Diagonalization of this matrix yields the coordinate invariant eigenvalues: $\lambda_1 \leq \lambda_2 \leq \lambda_3$. The quantities Laplacian, $\nabla^2\rho_{BCP}$, of charge density at the bond critical point is defined as:

$$\nabla^2\rho_{BCP} = \sum_{i=1}^{3} \lambda_i. \tag{3}$$

There are well–known relationships between energetic topological parameters and the Laplacian of electron density at BCP:

$$\frac{1}{4}\nabla^2\rho_{BCP} = 2G_{\text{BCP}} + V_{\text{BCP}} \tag{4}$$

$$H_{\text{BCP}} = G_{\text{BCP}} + V_{\text{BCP}}, \tag{5}$$

where G_{BCP}, V_{BCP}, and H_{BCP} are the kinetic, potential, and total electronic energy densities at critical point, respectively. G_{BCP} is a positive value, whereas V_{BCP} is a negative one.

Results and discussion

Geometries

The graphical illustration of the complexes under consideration is depicted in Fig. 1. Table 1 also presents the evaluated geometrical parameters for various carbene–BX_3 complexes (X=H, OH, CN, NC, NH_2, CH_3, F, Cl, and Br). Quantitatively, all the complexes indicate the presence of B–C_{carbene} interaction due to the short B–C_{carbene} distances. From Table 1, it is apparent that the estimated B–C_{carbene} distances are in a range of 1.531–1.633 Å which is much smaller than the sum of Van der Waals radii for carbon and boron (about 3.7 Å). The binding distance is calculated to be 1.587 Å in the carbene–BH_3 complex. However, the presence of the electron-donating groups makes an increase of binding distance. More especially, the substitution of electron-donating groups (OH and NH_2) in the BX_3 molecule makes a 0.040 and 0.046 Å increase of the binding distance, respectively, whereas the electron-withdrawing groups (F, CN and NC) result in a 0.020, 0.056 and 0.046 Å decrease of the binding distance. The calculated binding distance in methyl substituted complex is 1.614 Å, which is 0.027 Å longer than that of carbene–BH_3. An interesting aspect of the results presented in Table 1 is the fact that the binding distance of the systems tends to decrease as the size of the halogen increases, which corresponds to a decreasing value of the halogen atom electronegativity.

From the data in Table 1, it is also evident that all B–X bonds are systematically lengthened upon complexation. These results reveal that the binding between carbene and the BX_3 molecules weakens the B–X bond. The elongation of the B–X bond varies from 0.0273 to 0.1279 Å. It should be noted that this elongation is larger than that in the H_2C–LiX lithium bond [26]. The B–H bond elongation is 0.0273 Å in carbene–BH_3 complex. However, both the electron-donating (OH and NH_2) and electron-withdrawing groups (F, CN and NC) result in a significant increase of the B–X bond elongation. An interesting finding is that the amount of elongation in $B(NC)_3$ is 0.02 Å greater than that of B

Table 1 B–$C_{carbene}$ and B–X bond distances, dipole moments and principal components of electronic quadrupole moment in carbene-BX_3 compounds

X	$r_{B\text{-}Ccarbene}$ (Å)	$r_{B\text{-}X}$(Å) [a]	μ(D) [b]	Q_{XX}:Q_{YY}:Q_{ZZ} (D-Å) [c]
H	1.587	1.215 (0.027)	6.79 (0.75)	−50.23: −28.40: −41.86
OH	1.626	1.461 (0.090)	5.77 (1.15)	−44.69: −45.49: −57.73
CN	1.531	1.585 (0.064)	10.99 (2.21)	−62.00: −73.60: −87.12
NC	1.541	1.510 (0.084)	10.90 (2.2)	−63.04: −73.69: −87.11
NH_2	1.633	1.538 (0.056)	4.54 (2.48)	−45.21: −45.07: −65.09
CH_3	1.614	1.643 (0.069)	6.71 (0.26)	−56.08: −52.19: −64.84
F	1.567	1.394 (0.078)	9.11 (1.80)	−49.71: −41.25: −55.51
Cl	1.560	1.867 (0.119)	9.15 (0.73)	−62.96: −67.52: −81.12
Br	1.549	2.038 (0.128)	8.93 (0.33)	−70.52: −85.87: −99.00

[a] Data in parentheses are the difference between the complex and the monomer. [b] Data in parentheses are for BX_3 moieties in the complex. [c] The principal components of quadrupole moments

$(CN)_3$. The estimated B–X bond distance in -CH_3 substitution is 1.6426 Å, which is 0.0692 Å shorter than free B$(CH_3)_3$ molecule. The B–X bond elongation in the three halogen-containing complexes increases in the following order: carbene–BF_3<carbene–BCl_3<carbene–BBr_3. According to the above analyses, it is seen that the B–X bond elongation is not completely consistent with the change of the binding distance.

It is expected that the carbene–BF_3 complex formation is associated with a dipole moment enhancement due to the charge transfer and electron polarization of the molecules involved in the interaction. Dipole moments (*μ*) and quadrupole moments (*Q*) of the carbene–BX_3 systems are listed in Table 1. For the all complexes studied here, the largest dipole moment component is directed along the BX_3–carbene bond. One can see that the substitution of X atoms into the carbene–BH_3 complex has a significant influence on the dipole and quadrupole moments. As also evident from Table 1, the electron-donating groups tend to decrease the total dipole moments of the studied species, while a reverse trend is found for the electron- withdrawing groups. Figure 2 shows the relationship of the binding distance with the dipole moment of the carbene–BX_3 complexes. As can be seen from Fig. 2, there is a linear relationship between the binding distance and the dipole moment of the complexes (R^2=0.931).

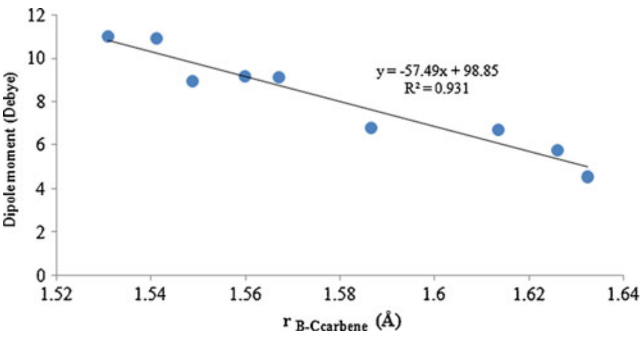


Fig. 2 Correlation between dipole moments and binding distances of carbene–BX_3 complexes

Binding energies

The interaction energy provides a measure of the strength of the interaction between carbene and BX_3 moieties in carbene–BX_3 complexes. Table 2 presents the interaction energies for these complexes at HF, B3LYP, MP2, MP4 and CCSD(T) levels of theory. Estimation of the BSSE for all of the structures presented here was performed by the full counterpoise method [38].

Considering the results listed in Table 2, it can be seen that all the methods indicate the presence of a relatively strong carben–BX_3 interaction due to the interaction energies between −42 and −101 kcal mol^{-1}. We could not find any theoretical information in the literature regarding the binding energies of the carbene–BX_3 complexes. However, we can compare our estimates with the binding energies of $(R–BH)_2$, where R=imidazol-2-ylidene. The B3LYP-estimated complexation energy for the carbene–BH_3 complex is −71.96 kcal mol^{-1}, which is −25 kcal mol^{-1} smaller than that for the $(R–BH)_2$ [45]. It should be noted that the estimated binding energies in the carbene–BX_3 complexes are also much larger than that for the H-bonded

Table 2 Calculated binding energies by supermolecule HF, B3LYP, MP2, MP4 and CCSD(T) methods carbene–BX_3 complexes[a]

X	E_{int}^{HF}	E_{int}^{B3LYP}	E_{int}^{MP2}	E_{int}^{MP4}	$E_{int}^{CCSD(T)}$
H	−60.80	−71.96	−72.05	−69.68	−68.97
OH	−50.48	−51.72	−55.81	−53.81	−53.08
CN	−92.57	−92.52	−101.57	−97.60	−96.09
NC	−80.34	−84.97	−97.41	−92.33	−90.12
NH_2	−42.54	−46.68	−53.36	−50.22	−42.66
CH_3	−44.12	−52.23	−59.43	−56.32	−54.02
F	−62.61	−69.87	−71.22	−69.33	−69.72
Cl	−78.68	−77.75	−85.47	−80.09	−79.03
Br	−80.64	−79.52	−88.60	−82.43	−81.10

[a] All calculated binding energies in kcal mol^{-1} and BSSE corrected

H_2C–HY (Y=F, CN, OH, and NH_2) [46] and H_2C–LiX [26] complexes, consistently with the fact that the BX_3 molecules are better Lewis acid than HY and LiX. Table 2 shows that, in general, the MP2 method provides larger stabilization energy than the others. Moreover, the inclusion of correlation (MP2) produces a stabilization of the complexes as high as 5–15 kcal mol^{-1}. Higher order perturbation (MP4) also produces significant variation in the binding energies of the B–$C_{carbene}$ complexes. As can be seen, the HF level of theory extremely underestimates the interaction energies of the complexes. In a previous study [47], it was demonstrated that the difference between the MP2 and HF energies is mainly assigned to the effects of high-order electrostatic interaction such as a dispersion interaction. Consequently, due to the large gain of the attraction by electron correlation (5–15 kcal mol^{-1}), dispersion force plays an important role in the stability of the complexes.

The interaction energies calculated at the CCSD(T) level are smaller by about 2–7 kcal mol^{-1} than those at the MP2 and MP4 levels. Even so, the change in interaction energy in the different systems is similar for the different levels of theory. However, the interaction energies at the MP4 level are close to those at the CCSD(T). Inspection of Table 2 reveals that the CCSD(T) interaction energies increase in the order NH_2>OH>CH_3>F>H>Cl>Br>NC>CN, which is the opposite trend shown in the binding distances. That is, upon complexation with carbene, electron-withdrawing groups form the strongest B–$C_{carbene}$ bond. This result is consistent with that obtained for Li–$C_{carbene}$ bonds [26]. Focusing on CCSD(T) results, which are available for each type of system considered, it can be seen that BBr_3 moiety is bound about 3% more strongly than BCl_3, which binds about 13% more strongly than BF_3. Figure 3 shows the correlation between $E_{int}^{CCSD(T)}$ interaction energies and dipole moments of the complexes. As evident, there is a linear relationship between the interaction energy and the dipole moment of the complexes (R^2=0.918). This suggests that the electrostatic interaction may contribute significantly to the formation of the complexes.

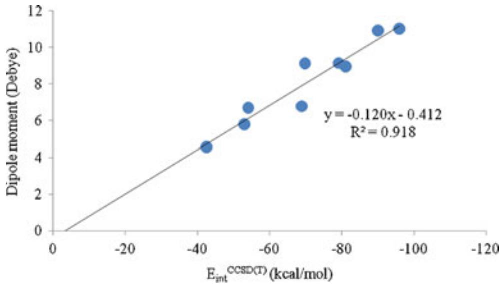


Fig. 3 Correlation between dipole moment and CCSD(T) binding energies of carbene–BX_3 complexes

Application of SAPT

To further understand the nature of the B–$C_{carbene}$ bonds in these complexes, the interaction energies of complexes were decomposed into four parts: electrostatic interaction energy (E_{elect}), Pauli exchange repulsion energy (E_{exch}), induction energy (E_{ind}) and dispersion energy (E_{disp}). The results are given in Table 3. The SAPT interaction energy (E_{int}^{SAPT}) is the sum of E_{elect}, E_{exch}, E_{ind} and E_{disp}. One of the most striking features of these data is the fact that the stabilities of the B–$C_{carbene}$ interactions are predicted to be attributable mainly to electrostatic and induction effects, while dispersion forces, which have been widely believed to be responsible for these types of interactions, play a smaller role in stabilizing these complexes. Considering the SAPT results, it is also found that electrostatic effects account for about 52% of the overall attraction in the carbene–BH_3 complex. By comparison, the induction component of this interaction represents about 39% of the total attractive forces, while dispersion contributes 9% to the stability of this complex. Thus it can be said that the carbene–BH_3 interaction is remarkably dependent on both electrostatic and induction forces, with electrostatic playing the largest role in their stability. Thus, the character of the B–$C_{carbene}$ bond is almost equally due to covalency and ionicity.

Unlike H–bonding interactions [48–50], the exchange energy term outweighs the electrostatic term for each complex studied here (Table 3). Based on our SAPT results, the electrostatic contribution to the overall attraction energies of X=OH, CN, NC, NH_2, CH_3, F, Cl, and Br are 56%, 51%, 52%, 56%, 53%, 56%, 51%, 50 %, respectively. Clearly, the electrostatic contribution is largest for the electron-withdrawing groups (F, CN and NC) and smallest for electron-donating groups (OH and NH_2). It is interesting to note that, although these types of interactions are largely dependent on electrostatic forces, the induction interaction

Table 3 DFT-SAPT energy decomposition analysis for carbene–BX_3 complexes[a]

X	E_{elect}	E_{exch}	E_{ind}	E_{disp}	E_{int}^{SAPT}
H	−115.71	156.88	−86.42	−20.89	−66.14
OH	−126.65	180.14	−77.99	−20.74	−45.25
CN	−145.14	195.11	−114.22	−26.47	−90.73
NC	−143.78	198.19	−113.54	−24.82	−83.94
NH2	−125.53	186.59	−75.79	−23.73	−38.45
CH3	−122.36	178.43	−82.18	−26.02	−52.13
F	−129.82	168.90	−83.66	−17.49	−62.07
Cl	−158.46	227.99	−122.87	−28.50	−81.82
Br	−166.73	243.50	−132.19	−31.37	−86.79

[a] All calculated SAPT components and SAPT energies in kcal mol^{-1}

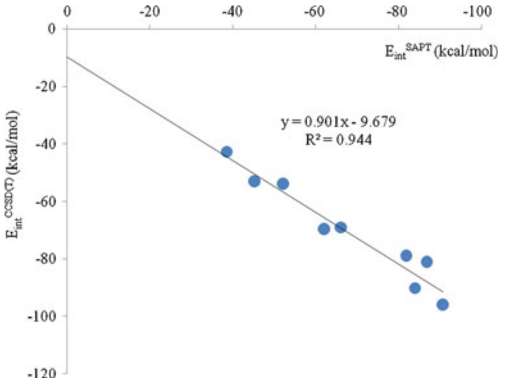


Fig. 4 Correlation between SAPT and CCSD(T) binding energies of carbene–BX_3 complexes

between the carbene and BX_3 moiety seems to play a significant role in determining the geometric structures of these complexes. For the CN and NC electron-withdrawing groups, the estimated induction energy contributions to the total stabilization energy are 40% and 39%, respectively. As the size of the halogen substituent increases the electrostatic interaction would be expected to decrease. Comparing the data for the fluorine, chlorine, and bromine substituted carbene-BX_3 systems, it can be seen that both the dispersion and induction components of the interaction energy increase with increasing halogen size. Interestingly, there is a larger increase in the induction interaction, going from fluorine to bromine, than in the dispersion interaction.

Table 3 and Fig. 4 indicate that the SAPT interaction energies for the carben–BX_3 complexes are generally in good agreement with those obtained using the correlated B3LYP, MP2, MP4 and CCSD(T) methods. The calculated $E_{\mathrm{int}}^{\mathrm{SAPT}}$energy for carbene–$BH_3$ complex is −66.14 kcal mol^{-1} which underestimates CCSD(T) and MP4 energies by about 2.8 and 3.5 kcal mol^{-1}, respectively. The SAPT result for the carbene–$B(OH)_3$ complex compare particularly poorly to CCSD(T), with the SAPT binding energy being 8 kcal mol^{-1} lower than that calculated using the CCSD(T). It is interesting to note that, in the case of the Cl and Br substitution, SAPT binding energies are underestimated in relation to CCSD(T), while binding energies for the BF_3 is overestimated.

Application of QTAIM

A great deal of information about the nature of B···C_{carbene} intermolecular interactions in the carbene–BX_3 complexes can be obtained from topological analysis of its electron density. Based on the QTAIM [33], properties of BCPs serve to summarize the nature of the interaction between two atoms as shared (covalent) or closed–shell (ionic) interaction. For a set of H–bonded complexes, Koch and Popelier [51] found the correlation between the HB energy and ρ_{BCP} to be linear as long as the acceptor atom remained unchanged.

Table 4 shows the λ_i, ρ_{BCP}, $\nabla^2\rho_{BCP}$, and the energy components G_{BCP}, V_{BCP} and H_{BCP} values for all of the complexes examined in this work. The molecular graph of the carbene-BH_3 complex is also displayed as Fig. 5, where the positions of all critical points are indicated as well as the bond paths between attractors. As seen in the molecular graph, there is BCP for B–C_{carbene} and one ring critical point (RCP) within the five-member ring. Ring-bond paths connecting BCPs and RCPs have also been found. The QTAIM analysis of B–C_{carbene} bonding has been studied in a previous study [45]. Liu indicated that the polar B–C_{carbene} bond, possessing a high degree of covalency in the bonding character, contributes positive net charge to atom B, and simultaneously, negative net charge to atom C_{carbene}. The author concluded that this means that more π-backdonation than σ-donation occurs from the center to the ligand in R(H)B=B(H)R, where R=imidazol-2-ylidene. On the basis of the fact that there is a BCP between the donor C_{carbene} and the B atom in the carben-BX_3 (X=H, OH, CN, NC, NH_2, CH_3, F, Cl, Br) complex, a topological analysis of the electron density further validates the existence of B–C_{carbene} bonds in all of the complexes.

From the results in Table 4, despite falling in a region of charge depletion with $\nabla^2\rho_{BCP}$ >0, all B–C_{carbene} BCPs are

Table 4 QTAIM analysis at the B3LYP/aug-cc-pVTZ level of theory for carbene–BX_3 complexes $\nabla^2\rho_{BCP}$[a]

X	λ_1	λ_2	λ_3	ρ_{BCP}	$\nabla^2\rho_{BCP}$	G_{BCP}	V_{BCP}	H_{BCP}
H	−0.224	−0.216	0.821	0.144	0.381	0.212	−0.339	−0.127
OH	−0.205	−0.203	0.513	0.131	0.104	0.144	−0.262	−0.118
CN	−0.236	−0.232	0.577	0.157	0.109	0.165	−0.314	−0.149
NC	−0.257	−0.253	0.535	0.149	0.026	0.151	−0.296	−0.145
NH_2	−0.186	−0.182	0.512	0.126	0.143	0.146	−0.256	−0.110
CH_3	−0.202	−0.194	0.714	0.135	0.318	0.190	−0.31	−0.122
F	−0.229	−0.226	0.520	0.138	0.065	0.144	−0.272	−0.128
Cl	−0.285	−0.277	0.659	0.158	0.096	0.176	−0.328	−0.152
Br	−0.293	−0.283	0.684	0.162	0.107	0.182	−0.337	−0.155

[a] All QTAIM parameters in au

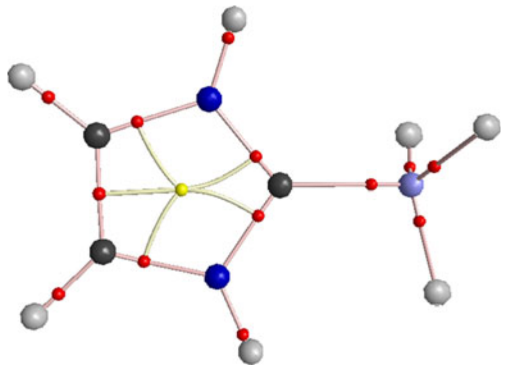


Fig. 5 Molecular graph of carbene–BH_3 complex. The graph was obtained at the B3LYP/aug-cc-pVTZ level. Big circles correspond to attractors and small red and yellow circles are bond and ring critical points, respectively. The lines are bond paths

characterized by a reasonably large value of the electron density ρ_{BCP} and H_{BCP} <0, indicating that the potential energy overcomes the kinetic energy density at BCP and the B–$C_{carbene}$ bond is a polar covalent bond. The estimated values of ρ_{BCP} in B-C_{carben} BCPs are in the range 0.126-0.162 au, whereas the values of $\nabla^2\rho_{BCP}$ are between 0.026-0.381 au. For carbene-BH_3 complex, the calculated ρ_{BCP} and $\nabla^2\rho_{BCP}$ value is 0.144 and 0.381 au, respectively. These values decrease to 0.131, 0.126 and 0.104, 0.143 au for the OH, and NH_2 electron-donating groups, respectively, which is in accordance with the evidence for small destabilization of the B–$C_{carbene}$. The average values of ρ_{BCP} ($\nabla^2\rho_{BCP}$) in au for CN and NC electron-withdrawing groups are 0.147 (0.109) and 0.149 (0.026), respectively. For the halogen substitution, QTAIM analyses indicate the capacity of the carbene-BX_3 complexes to concentrate electrons at the B–$C_{carbene}$ BCPs enhance considerably with the size of halogen atom. This conclusion is completely the same as that drawn from B–$C_{carbene}$ energies.

Two topological parameters both the total electron energy density H_{BCP} and Laplacian at BCP ($\nabla^2\rho_{BCP}$) may be useful in characterization of the strength of the B–$C_{carbene}$ interactions. According to Rozas et al. [52], the character of X–Y interaction could be classified as a function of the total electron energy density H_{BCP} with Laplacian of the electron density at X–Y BCP ($\nabla^2\rho_{BCP}$). It means that for strong X–Y interactions ($\nabla^2\rho_{BCP}<0$ and $H_{BCP}<0$) the covalent character is established, for medium strength X–Y ($\nabla^2\rho_{BCP}>0$ and $H_{BCP}<0$) their partially covalent character is defined, and weak X–Y ($\nabla^2\rho_{BCP}>0$ and $H_{BCP}>0$) are mainly electrostatic [53]. Therefore, B–$C_{carbene}$ interactions for the complexes studied here have a partially covalent nature. Figure 6 shows the correlation between SAPT electrostatic energies with H_{BCP} values. This correlation (R^2=0.89) indicates that the two representations of B–$C_{carbene}$ characteristics, based upon QTAIM and SAPT, are approximately equivalent. In terms of Espinosal's proposal [54], the local electronic potential energy density (V_{BCP}) can represent the capacity of the complexes in concentrating electrons at the B–$C_{carbene}$ BCP and gives an approach to describing the B–$C_{carbene}$ strength. Abramov [55] has proposed the evaluation of the local electronic potential energy density from the experimental electron density distribution. The Abramov's local electronic potential energy density $V_{BCP,A}$ can be evaluated according to the following expression:

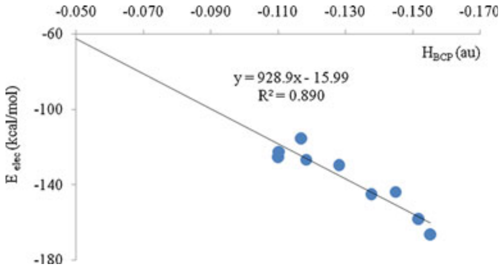


Fig. 6 Correlation of SAPT electrostatic (E_{elect}) term with total electronic density (H_{BCP}) at B–$C_{carbene}$ BCPs

$$V_{BCP,A} = -\left(\frac{3}{5}\left(3\pi^2\right)^{2/3}\rho^{5/3} + \frac{1}{12}\nabla^2\rho\right). \qquad (6)$$

Figure 7 compares the calculated values of $V_{BCP,A}$ from Eq. 6 with the V_{BCP} data from the topological analysis for the nine B–$C_{carbene}$ bonds. The distribution of the points exhibits a very good linear relationship between V_{BCP} and $V_{BCP,A}$. The graphical representation indicates that Eq. 6 can equivalently evaluate the capacity of the carbene-BX_3 complexes in concentrating electrons at the B–$C_{carbene}$ BCP and therefore the B–$C_{carbene}$ strength.

Conclusions

Within this study, we used *ab initio*, DFT, SAPT and QTAIM theories to investigate the nature of B–$C_{carbene}$ interaction in a series of imidazol-2-ylidene carbene-BX_3 complexes, where X=H, OH, NH_2, CH_3, CN, NC, F, Cl,

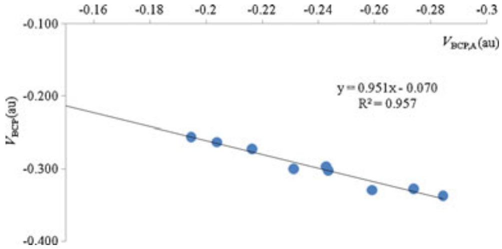


Fig. 7 Correlation between calculated potential energy density (V_{BCP}) and Abramov potential energy density ($V_{BCP,A}$)

and Br. Based on the results found in this study, it is concluded that the imidazol-2-ylidene carbene makes a relatively strong interaction with BX_3 moieties. All B–X bonds are systematically lengthened upon complexation. These results reveal that the binding between carbene and the BX_3 moieties weakens the C–X bond. A more detailed analysis of calculated binding energies shows that, in general, the MP2 method provides larger stabilization energy than the other. Focusing on CCSD(T) results, which are available for each type of system considered, it can be seen that BBr_3 moiety is bound about 3% more strongly than BCl_3, which binds about 13% more strongly than BF3. The SAPT interaction energies for the carbene–BX_3 complexes are generally in good agreement with those obtained using the supermolecule MP4 and CCSD(T) levels of theory. Based on QTAIM results, it is evident that the B–$C_{carbene}$ interactions for the complexes studied here have a partially covalent nature. Moreover, this work should be interesting for future theoretical investigations and experimental works, which can also enrich the theory of donor-acceptor interaction. Since imidazol-2-ylidene carbene is an intermediate in some chemical reactions, thus the study of the interaction between this carbene and BX_3 moieties is helpful to understand the mechanism in carbene–boron chemical reactions. The results of SAPT analysis can also lead to deeper understanding of carbene–boron interactions, and the quantitative results may be used to guide development of empirical, yet computationally fast force fields for biomolecular simulation and modeling. We believe that this study enriches the knowledge of carbene-boron interactions and need some support from future experimental work.

References

1. Woo HK, Wang XB, Wang LS, Lu KC (2005) Probing the low-barrier hydrogen bond in hydrogen maleate in the gas phase: a. photoelectron spectroscopy and *ab initio* study. J Phys Chem A 109:10633–10637
2. Hobza P, Havlas Z (2000) Blue-shifting hydrogen bonds. Chem Rev 100:4253–4264
3. Scheiner S (1997) Hydrogen bonding: a theoretical prospective. Oxford University Press, Oxford, UK
4. Thar J, Kirchner B (2006) Hydrogen bond detection. J Phys Chem A 110:4229–4237
5. Hong BH, Lee JY, Lee CW, Kim JC, Bae SC, Kim KS (2001) Self-assembled arrays of organic nanotubes with infinitely long one-dimensional H-bond chains. J Am Chem Soc 123:10748–10749
6. Hong BH, Bae SC, Lee CW, Jeong S, Kim KS (2001) Ultrathin single-crystalline silver nanowire arrays formed in an ambient solution phase. Science 294:348–351
7. Tarakeshwar P, Kim KS (2004) In: Encyclopedia of nanoscience and nanotechnology, vol 7. American Science Publishers, California, pp 367–404
8. Bandyopadhyay I, Lee HM, Kim KS (2005) Phenol vs water molecule interacting with various molecules: σ-type, π-type, and X-type hydrogen bonds, interaction energies, and their energy components. J Phys Chem A 109:1720–1728
9. Regitz M (1996) Nucleophilic carbenes: an incredible renaissance. Angew Chem 35:725–728
10. Bourissou D, Guerret O, Gabbaï FP, Bertrand G (2000) Stable carbenes. Chem Rev 100:39–91
11. Couzijn EPA, Zocher E, Bach A, Chen P (2010) Gas-phase energetics of reductive elimination from a palladium (II) N-heterocyclic carbene complex. Chem Eur J 16:5408–5415
12. Benitez D, Shapiro ND, Tkatchouk E, Wang Y, Goddard WA III, Toste FD (2009) A bonding model for gold(I) carbene complexes. Nature Chem 1:482–486
13. Yao S, Xiong Y, Driess M (2010) N-heterocyclic carbene (NHC)-stabilized silanechalcogenones: NHC→Si(R2)=E (E=O, S, Se, Te). Chem Eur J 16:1281–1288
14. Liu Z (2009) Chemical bonding in silicon−carbene complexes. J Phys Chem A 113:6410–6414
15. Standard JM, Steidl RJ, Beecher MC, Quandt RW (2011) Multireference configuration interaction study of bromocarbenes. J Phys Chem A 115:1243–1249
16. Alkorta I, Rozas I, Elguero J (1998) Non-conventional hydrogen bonds. Chem Soc Rev 27:163–170
17. Nolan SP (2006) N-Heterocyclic carbenes in synthesis. Wiley, New York
18. Enders D, Breuer K, Raabe G, Runsink J, Teles JH, Melder JP, Ebel K, Brode S (1995) Preparation, structure, and reactivity of 1,3,4-Triphenyl-4,5-dihydro-1H-1,2,4-triazol-5-ylidene, a new stable carbene. Angew Chem Int Edn Eng 34:1021–1023
19. Grasa GA, Kissling RM, Nolan SP (2002) N-Heterocyclic carbenes as versatile nucleophilic catalysts for transesterification/acylation reactions. Org Lett 4:3583–3586
20. Glorius F (2006) N-Heterocyclic carbenes in transition metal catalysis (Topics in Organometallic Chemistry). Springer, Heidelberg
21. Ullah F, Kindermann MK, Jones PG, Heinicke J (2009) Annulated N-Heterocyclic carbenes: 1,3-ditolylphenanthreno [9,10-d]imidazol-2-ylidene and transition metal complexes thereof. Organometallics 28:2441–2449
22. Khramov DM, Lynch VM, Bielawski CW (2007) N-Heterocyclic carbene−transition metal complexes: spectroscopic and crystallographic analyses of π-back-bonding interactions. Organometallics 26:6042–6049
23. Moya-Barrios R, Fregeau BM, Cozens FL (2009) Reactivity of halo(pyridinium) carbenes. J Org Chem 74:9126–9131
24. Padwa A, Hornbuckle SF (1991) Ylide formation from the reaction of carbenes and carbenoids with heteroatom lone pairs. Chem Rev 91:263–309
25. Wanzlick HW, Schönherr HJ (1970) Liebigs Ann Chem 731:176
26. Li Q, Wang H, Liu Z, Li W, Cheng J, Gong B, Sun J (2009) *Ab initio* study of lithium-bonded complexes with carbene as an electron donor. J Phys Chem A 113:14156–14160
27. Kitaura K, Morokuma K (1976) A new energy decomposition scheme for molecular interactions within the Hartree-Fock approximation. Int J Quantum Chem 10:325–340
28. Jeziorski B, Moszynski R, Szalewicz K (1994) Perturbation theory approach to intermolecular potential energy surfaces of van der Waals Complexes. Chem Rev 94:1887–1930
29. Williams HL, Szalewicz K, Jeziorski B, Moszynski R, Rybak S (1993) Symmetry–adapted perturbation theory calculation of the Ar–H_2 intermolecular potential energy surface. J Chem Phys 98:1279–1292
30. Misquitta AJ, Podeszwa R, Jeziorski B, Szalewicz K (2005) Intermolecular potentials based on symmetry-adapted perturbation theory with dispersion energies from time-dependent density-functional calculations. J Chem Phys 123:214103–214116

31. Misquitta AJ, Szalewicz K (2005) Symmetry-adapted perturbation-theory calculations of intermolecular forces employing density-functional description of monomers. J Chem Phys 122:214109–214127
32. Williams HL, Chabalowski CF (2001) Using Kohn-Sham orbitals in symmetry-adapted perturbation theory to investigate intermolecular interactions. J Phys Chem A 105:646–659
33. Bader RFW (1990) Atoms in molecules-a quantum theory. Oxford University Press, New York
34. Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH, Koseki S, Matsunaga N, Nguyen KA, Su SJ, Windus TL, Dupuis M, Montgomery JA (1993) General atomic and molecular electronic structure system. J Comput Chem 14:1347–1363
35. Becke AD (1988) Density-functional exchange-energy approximation with correct asymptotic behavior. Phys Rev A 38:3098–3100
36. Lee C, Yang W, Parr RG (1988) Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Phys Rev B 37:785–789
37. Hobza R, Zahradnik R (1988) Intermolecular Interactions between Medium-Sized Systems. Nonempirical and empirical calculations of interaction energies: successes and failures. Chem Rev 88:871–897
38. Boys SF, Bernardi F (1970) The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. Mol Phys 19:553–566
39. Hesselmann A, Jansen G (2003) The helium dimer potential from a combined density functional theory and symmetry-adapted perturbation theory approach using an exact exchange–correlation. Phys Chem Chem Phys 5:5010–5014
40. Hesselmann A, Jansen G (2003) Intermolecular dispersion energies from time-dependent density functional theory. Chem Phys Lett 367:778–784
41. Jansen G, Hesselmann A (2001) Comment on: using Kohn-Sham orbitals in symmetry-adapted perturbation theory to investigate intermolecular interactions. J Phys Chem A 105:11156–11157
42. Dalton, a molecular electronic structure program, Release 2.0 (2005) see http://www.kjemi.uio.no/software/dalton/dalton.html
43. Bukowski R, Cencek W et al. (2008) SAPT 2008: an *Ab Initio* program for many-body symmetry-adapted perturbation theory calculations of intermolecular interaction energies. Sequential and Parallel Versions, University of Delaware and University of Warsaw, 2008; see http://www.physics.udel.edu/_szalewic/SAPT/SAPT.html
44. Biegler-Konig F, Schonbohm J, Bayles D (2001) AIM 2000. J Comput Chem 22:545–559
45. Liu Z (2010) On the nature of B-C$_{carbene}$ bonding in a stable neutral diborene. J Chem Phys 132:084303–084309
46. Alkorta I, Elguero J (1996) carbenes and silylenes as hydrogen bond acceptors. J Phys Chem 100:19367–19370
47. Jaffe RL, Smith GD (1996) A quantum chemistry study of benzene dimer. J Chem Phys 105:2780–2788
48. Esrafili MD, Beheshtian J, Hadipour NL (2011) Computational study on the characteristics of the interaction in linear urea clusters. Int J Quantum Chem 111:3184–3195
49. Cybulski H, Sadlej J (2008) Symmetry-adapted perturbation-theory interaction-energy decomposition for hydrogen-bonded and stacking structures. J Chem Theor Comput 4:892–897
50. Panek JJ, Jezierska A (2007) Symmetry-adapted perturbation theory analysis of the N···HX hydrogen Bonds. J Phys Chem 111:650–655
51. Koch U, Popelier PLA (1995) Characterization of C-H···O hydrogen bonds on the basis of the charge density. J Phys Chem 99:9747–9754
52. Rozas I, Alkorta I, Elguero (2000) Behaviour of ylides containing N, O and C atoms as hydrogen bond acceptors. J Am Chem Soc 122:11154–11161
53. Lu Y, Zou J, Wang Y, Jiang Y, Yu Q (2007) *Ab initio* investigation of the complexes between bromobenzene and several electron donors: some insights into the magnitude and nature of halogen bonding. J Phys Chem A 111:10781–10788
54. Espinosa E, Molins E, Lecomte C (1998) Hydrogen bond strengths revealed by topological analyses of experimentally observed electron densities. Chem Phys Lett 285:170–173
55. Abramov YA (1997) On the possibility of kinetic energy density evaluation from the experimental electron-density distribution. Acta Crystallogr Sect A Found Crystallogr 53:264–272

J Mol Model (2012) 18:2013–2019
DOI 10.1007/s00894-011-1215-0

ORIGINAL PAPER

In silico analysis of *Pycnoporus cinnabarinus* laccase active site with toxic industrial dyes

Nirmal K. Prasad · Vaibhav Vindal · Siva Lakshmi Narayana · Ramakrishna V. · Swaraj Priyaranjan Kunal · Srinivas M.

Received: 27 April 2011 / Accepted: 7 August 2011 / Published online: 30 August 2011

Abstract Laccases belong to multicopper oxidases, a widespread class of enzymes implicated in many oxidative functions in various industrial oxidative processes like production of fine chemicals to bioremediation of contaminated soil and water. In order to understand the mechanisms of substrate binding and interaction between substrates and *Pycnoporus cinnabarinus* laccase, a homology model was generated. The resulted model was further validated and used for docking studies with toxic industrial dyes- acid blue 74, reactive black 5 and reactive blue 19. Interactions of chemical mediators with the laccase was also examined. The docking analysis showed that the active site always cannot accommodate the dye molecules, due to constricted nature of the active site pocket and steric hindrance of the residues whereas mediators are relatively small and can easily be accommodated into the active site pocket, which, thereafter leads to the productive binding. The binding properties of these compounds along with identification of critical active site residues can be used for further site-directed mutagenesis experiments in order to identify their role in activity and substrate specificity, ultimately leading to improved mutants for degradation of these toxic compounds.

Keywords Bioremediation · Homology modeling · Laccase · Mediators · Toxic dyes

Abbreviations

ABTS	2,2′-azino-bis(3-ethylbenzthiazoline-6-sulfonic acid)
PROSA	Protein structure analysis
NAMD	Nanoscale molecular dynamics
PDB	Protein data bank
RMSD	Root mean square deviation
CASTp	Computed atlas of surface topography of proteins
GOLD	Genetic optimization for ligand docking
MD	Molecular dynamics

Introduction

Laccases (EC 1.10.3.2) are oxidoreductases constituting a class of blue multicopper enzymes, first described in 1883 [1]. These are present ubiquitously in bacteria [2], fungi [3] and plants [4] and are classified into high and low redox potential oxidoreductases [5, 6]. Variations in protein structure are responsible for variable redox potentials and consequently, varied rates of one-electron transfer [7–10]. This electron transfer is mediated through Cu2+ ion. Three copper binding sites (T1-T3) contain four Cu2+ ions. T1

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1215-0) contains supplementary material, which is available to authorized users.

N. K. Prasad (✉) · S. P. Kunal
Department of Bioinformatics, Institute of In silico Biology,
Tirupati 517501, India
e-mail: nimmynirmal@gmail.com

V. Vindal
Department of Biotechnology and Bioinformatics
Infrastructure Facility, University of Hyderabad,
Hyderabad 500046, India

S. L. Narayana · S. M.
Mother Teresa College of Pharmacy,
Ghatkesar,
Hyderabad 501301, India

R. V.
Department of Biotechnology & Bioinformatics,
Yogi Vemana University,
Kadapa 516 003, India

site, being a mononuclear center, possesses one copper atom and T2&T3, being trinuclear centers, possess three copper atoms. Two conserved channels allow the passage of di-oxygen and release of water molecule from the T2/T3 center. Two β-turns close to the T1 Cu atom constitute the substrate binding pocket. The geometry of this pocket defines the substrate specificity of the enzyme [11]. Laccases utilize a one-electron transfer mechanism to oxidize a variety of substrates including biphenyls, polyphenols, aromatic amines, diamines and ascorbic acid [12]. Owing to its low substrate specificity, laccase from *Pycnoporus cinnabarinus* finds use in treatment of toxic dyes which are present in the effluent of the textile industry [13]. A general mode of laccase action involves the one electron oxidation of hydroxylated aromatic substrates, coupled to the reduction of dioxygen to water, converting the substrate to a free radical [5].

$$O_2 + 4e^- + 4H^+ \rightarrow 2H_2O.$$

The structural studies on several fungal and bacterial laccases have further allowed the rationalization of important structural and functional aspects of multi-copper oxidases such as the positioning of the T1 site, the electron transfer pathways between T1 and the T2/T3 cluster, the oxygen and water channels [8, 10, 11, 14–18].

Since 1990, when ABTS (2,2′-azino-bis(3-ethylbenzthiazoline-6-sulfonic acid)) was found to serve as a laccase substrate mediating or enhancing the enzyme action [19], the range of compounds that can be transformed by laccases has increased. An ideal mediator must be a low-molecular-weight laccase substrate whose enzymatic oxidation gives rise to stable high-potential intermediates. These intermediates act as reactive species which take part in chemical reactions with other compounds. Their oxidized and reduced forms should be stable and at the same time must not inhibit the enzymatic reaction. Ideally, a mediator can perform many cycles without degradation. In particular the biotechnological application of laccases, aiming at the development of various industrial oxidative processes is to produce fine chemicals to bioremediation of contaminated soil and water [20].

In the present study, the homology model of *Pycnoporus cinnabarinus* laccase has been constructed in order to get an in depth knowledge of its structural and functional aspects. To analyze its structural integrity, the constructed model was validated using structure analysis tools like PROSA and PROCHECK. Further, the docking studies were performed to understand the mechanism of laccase catalyzed enzymatic reactions as well as the role of chemical mediator interaction with active site residues. The approach is applicable in engineering 3D structures of enzymatic models, and studying interactions of active site residues with substrates.

Materials and methods

Homology modeling

The amino acid sequence of *Pycnoporus cinnabarinus* laccase was obtained from the NCBI protein database (Accession number: AAF13052) (http://www.ncbi.nlm.nih.gov/protein). Crystal structure of *Trametes hirsuta* laccase was taken from the protein data bank (PDB ID: 3FPX) [21] and used as the template for building the initial 3D model. The sequence alignment of laccase with the template was accomplished using ClustalW 2.0 (http://www.ebi.ac.uk/Tools/clustalw2/index.html). The Modeller 9v7 program [22] was employed to generate the initial 3D models of laccase. Modeller generates the 3D models by optimization of molecular probability density functions. The optimization process consists of applying the variable target function as well as conjugated gradients and molecular dynamics with simulated annealing. A set of 20 models of laccase were produced based on the resulting alignment obtained above. The outcomes were ranked based on the internal scoring function of Modeller.

Modeling copper atoms

All the models were reinitialized as quires and by setting the input output HETAM function in true mode. The function was read in HETAM records from template PDB and the copper atoms were inserted in the query models. These models containing copper atoms were further validated and refined.

Homology models validation

The top five models with high scores were validated by the Procheck [23], ProSA [24] and VADAR [25]. After validation, a model was finally chosen for further refinement by energy minimization. The energy minimization was performed using the NAMD package [26]. The optimized model was subjected to quality assessment with respect to its geometry and energy and was then subjected to molecular docking. Procheck was utilized for geometric evaluation. ProSA program was employed to evaluate the quality of consistency between the native fold and the sequence and examine the energy of residue–residue interactions using a distance-based pair potential. The substrate molecules acid blue 74, reactive black 5, reactive blue 19, ABTS, acetosyringone and syrangaldehyde were downloaded from Pubchem database of NCBI [27], and converted to 3D structure with VEGA ZZ software [28]. These substrates were geometrically optimized for further use in docking.

Structural analysis

Both the template and homology model C alpha and back bone atoms RMSD were calculated by magic fit program [29]. Packing architecture of the modeled protein was calculated by VADAR program. It analyzes mean hydrogen bonds distances, mean dihedral angles, accessible surface area and packing volume of the model.

Active site analysis

The substrate accessible pockets and active sites of laccase were identified by computed atlas of surface topography of proteins (CASTp) calculation [30] and GOLD software [31–33]. CASTp program uses the weighted Delaunay triangulation and the alpha complex for shape measurements. It provides identification and measurements of surface accessible pockets as well as interior inaccessible cavities of proteins. The program measures analytically the area and volume of each pocket and cavity, both in solvent accessible surface and molecular surface. The identified active sites were analyzed for amino acid cluster groups based on the solvent exposed active site atoms and bonding capacity of the polar groups. All the active site pockets were further evaluated by docking to test their capacity of accommodating substrates.

Molecular docking

Substrate molecules were docked to the binding sites using GOLD software [31–33]. One-hundred genetic algorithm (GA) runs were performed for each compound, and 10 ligand bumps were allowed in an attempt to account for mutual ligand/target fit. The binding region for the docking study was defined as a 20 Å radius sphere centered on the active site. For each of the GA run a maximum number of 100,000 operations were performed on a population of 100 individuals with a selection pressure of 1.1. The number of islands was set to 5 with a niche size of 2. The weights of crossover, mutation and migration were set to 95, 95 and 10 respectively. The scoring function Gold Score implemented in GOLD was used to rank the docking positions of the molecules, which were clustered together when differing by more than 2 Å rmsd [34, 35]. The best ranking clusters for each of the molecules were selected. Hydrogen bonds, bond lengths and close contacts between enzyme active site and substrate atoms were analyzed.

Results and discussion

Pycnoporus cinnabarinus laccase modeling

The first important step in homology modeling is to select an appropriate template structure for constructing the target model. To date, several crystal structures of fungal and bacterial laccases have been determined. A BLASTp search against protein data bank confirmed that several fungal laccase crystal structures could serve as the potential template for building the model. *Pycnoporus cinnabarinus* laccase has 82% sequence identity with *Trametes hirsuta* laccase (PDB ID: 3FPX). The template was chosen based on sequence similarity, residue completeness, and crystal resolution. The majority of the structure is considered conserved except for the limited gap inserts. The resulting alignment was used as input file for Modeller to generate the initial 3D models using the fast simulated annealing procedure implemented in the Modeller program. Since the N-terminus does not affect the substrate binding, the corresponding first 18 residues were not modeled in the

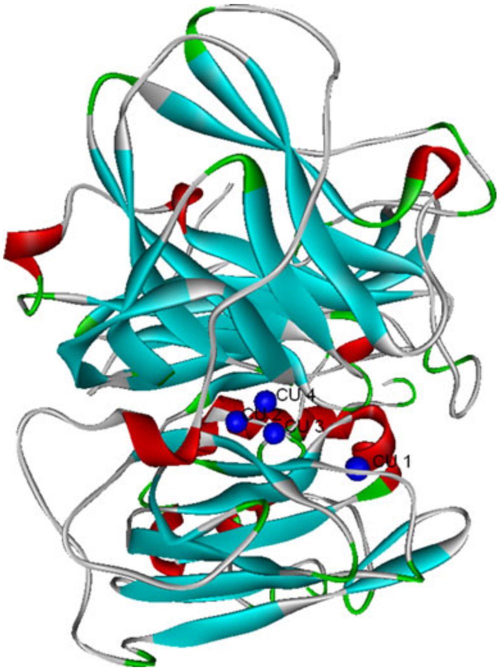


Fig. 1 Final three dimensional model of *Pycnoporus cinnabarinus* laccase in ribbon display mode showing α- helices in red, β-sheets in cyan, β-turns in green and copper atoms in blue color

Table 1 Copper atom distances in modeled Laccase and known 3D crystal laccases

Protein	Cu1-Cu2	Cu1-Cu3	Cu1-Cu4	Cu2-Cu3	Cu2-Cu4	Cu3-Cu4
MODEL	14.80	12.81	12.30	4.12	3.80	3.85
3FPX	14.84	12.86	12.18	4.10	3.79	3.84
1GWO	14.50	12.93	12.42	4.15	3.92	4.91
3CG8	14.51	13.08	12.64	4.13	3.91	4.97
1GYC	14.74	12.90	12.31	3.81	3.82	3.91
2H5U	15.12	13.26	12.43	4.40	4.08	4.77
2VDS	14.97	12.98	12.17	4.57	3.86	4.70
2HZH	14.74	13.12	12.02	4.33	3.93	4.80
2HRH	14.59	12.69	12.20	3.56	3.31	3.19

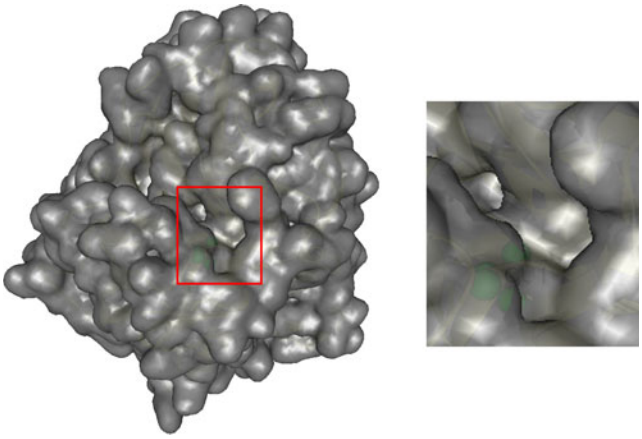


Fig. 2 Surface representation of model and close up view of active site pocket

model. The initial 3D models of laccase were energy-minimized to release the bad atomic contact and tune unreasonable local structural conformations.

Validation of homology models

Validation of a 3D model is an essential step that can be performed at different levels of structural organization to check the stereochemical parameters and accuracy of the overall packing. The assessment of chosen model for stereochemical properties of main-chain and side-chain residues was performed using Procheck-Ramachandran plot analysis. Procheck analysis, showed 91.0% of the residues were in the core region, 8.5% residues in the allowed regions and 0.5% in disallowed region. The overall main-chain and side-chain parameters are favorable for further analysis.

In order to investigate whether the interaction energy of each residue with the remainder of the protein is negative, a test was carried out to apply energy criteria using ProSA II energy plot. The ProSA analysis of the model showed maximum residues to have negative interaction energy with few residues displaying positive interaction energy. The overall interaction energy of the model was −7.34 kcal mol^{-1}, which is quite similar to the template 3FPX.pdb (−7.93 kcal mol^{-1}). Hence, the final model which proved to be well validated in terms of geometry and energy profiles suggests that the model is good enough to be a starting point for our next phase of docking studies. The final model structure of laccase is displayed in Fig. 1.

Structural analysis of model

Cα atoms and back bone atoms RMSD of both the model and template are 0.14 Å and 0.20 Å respectively. The mean residue volume and total packing volume of the model are 132.8 Å^3 and 68801.3 Å^3 respectively. VADAR analysis of the model showed, the mean helix phi, psi and omega angles are −63.5, -35.3 and −179.5 respectively, which is promising residue packing when compared to the crystal structure information. The accuracy of position of the modeled four copper atoms in three copper binding sites was checked by comparing the distances between the copper atoms with known 3D laccase copper binding site information (Table 1). The distances between copper atoms are quite similar to the distances between copper atoms in known 3D laccase structures. The catalytic active site is present close to the T1 copper binding site and residues at the active site are closely packed together forming a narrow cavity (Fig. 2). Therefore relatively small solvent accessible area, large substrate molecules cannot freely approach into the active site.

Docking conformations of ligands in the active site

Docking interaction was carried between all the solvent accessible atoms and ligands. Fruitful binding docking

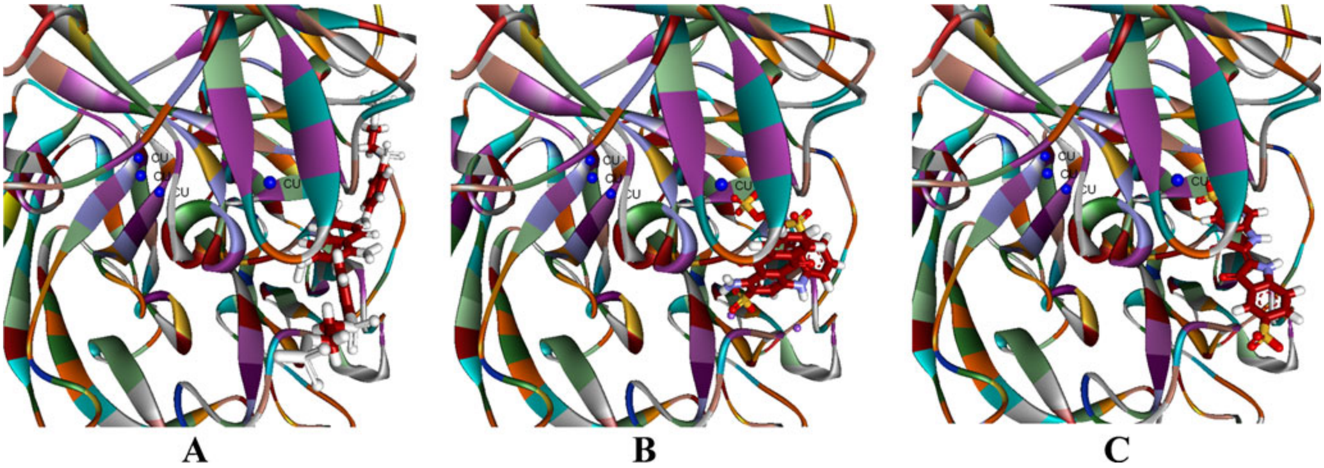


Fig. 3 Docked conformations of (**A**) Reactive black 5, (**B**) Reactive blue 19 and (**C**) Acid blue 74 in the active site

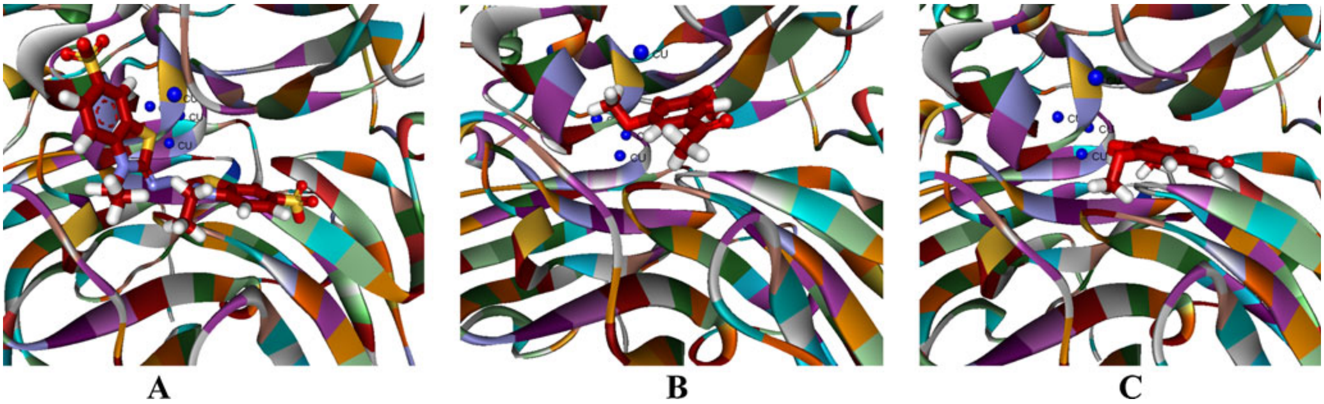


Fig. 4 Docked conformations of (**A**) ABTS, (**B**) Acetosyringone and (**C**) Syrangaldehyde in the active site

interactions were selected for further analysis. The selected binding site is near to the tri copper centers and has 394.5 Å^3 volume. The solvent accessible region at active site cavity comprises of residues ARG182, PRO184, PHE185, ASP227, PRO228, ASN229, ALA261, PRO412, GLY413, PHE414, PRO415, GLN451, HIS473, ILE474, ASP475, PHE476 and HIS477. The topological polar surface area of dyes is ranging from 189 Å^2 to 458 Å^2. Docking of reactive black 5, reactive blue 19, acid blue 74 azure B, amido black and aniline blue with active site residue atoms showed that the dye molecules are producing only some fruitful binding conformations at active site. This was due to steric hindrance caused by amino acid side chains near and around the active site cavity and bulky planar ring conformations of the dyes. Figure 3 depicts the productive binding conformations of the dyes in the active site. In the current picture, the narrow active site cavity could not always accommodate the bulky dye molecules for productive binding.

Mediator - enzyme docking studies reveal that mediator molecules easily approach the active site cavity and are able to rotate in the active site to bind in close proximity to the Cu1 atom (Fig. 4), which leads to productive binding and further electron transfer. The topological polar surface area of mediators is ranging from 55 Å^2 to 215 Å^2. 2-D structural details of ligands were presented in supplementary data. Hydrogen bonding profile of dyes and mediators with active site atoms are depicted in Table 2. The hydrogen bonds range between 1.700 Å to 2.471 Å. A mediator can perform many cycles without degradation due to its highly stable oxidized and reduced forms. In particular the biotechnological application of laccases, aiming at the development of various industrial oxidative processes is to produce fine chemicals to bioremediation of contaminated soil and water. Thus, mediator approach to laccase activity is preferred in degradation of industrial dyes over direct laccase-substrate reactions.

Conclusions

The modeled laccase exhibits 91.0% of residues falling in the most favorable region of the Ramachandran's plot, which showed the proper modeling of laccase. Further model overall quality analysis, geometrical and packing architecture analysis of the model shows that the modeled protein structure is suitable for docking studies. Active site

Table 2 Docking statistics of laccase with dyes and mediators

Substrate/mediator	Residue atom	Ligand atom	H.Bond distance
Reactive black 5	GLN 451:H	O25	1.272
	ASN 229:H	O22	2.500
	ASN 229:1HD2	O23	2.470
	ASN 229:2HD2	O23	2.338
	ASN 229:2HD2	O21	2.334
	THR 356:HG1	O29	2.312
Reactive blue 19	ASN 229:2HD2	O16	2.195
	PHE 185:O	O42	2.676
Acid blue 74	ASN 229:2HD2	O8	1.807
Pontamine sky blue	ALA261:H	O14	1.458
	ALA261:H	O15	1.897
	ALA261:H	O16	2.686
Amido black	ASN229:H	O6	1.990
	ASN229:2HD2	O6	2.335
	ASN229:2HD2	O8	2.346
	PHE414:O	H45	2.671
Aniline blue	ARG182:H	O11	1.536
	ASN229:2HD2	O5	1.582
	ALA261:H	O9	2.590
	PRO415:O	H79	2.342
	HIS477:NE2	H66	2.136
ABTS	ASN 229:1HD2	O6	2.584
	ASN 229:1HD2	O7	1.741
Acetosyringone	ASP 227:OD1	H20	1.548
	ASP 227:OD2	H20	2.691
Syringaldehyde	ASP 227:OD1	H17	1.668

analysis revealed closely packed narrow shape active site conformation, which lies close to the T1 copper binding site. The surface of the active site is composed of residues ARG182, PRO184, PHE185, ASP227, PRO228, ASN229, ALA261, PRO412, GLY413, PHE414, PRO415, GLN451, HIS473, ILE474, ASP475, PHE476, and HIS477. The docking conformations of substrates at active site showed that substrate molecules cannot freely access the active site pocket due to their bulky planar ring structure and steric hindrance of the active site residues. Mediator conformations at active site showed that mediator molecules are entering into the active site without any steric repulsion and interacting with residues. This is because mediator molecules are relatively small and can easily approach residues in the narrow active site as well as bind in close proximity to the T1 copper site. This leads to enzymatic oxidation and gives rise to stable high-potential intermediates. These intermediates oxidize substrate molecules and they themselves get reduced to their original state. The above mentioned enzymatic and chemical interactions are thus carried out in a cyclic manner without any degradation of the mediator.

Acknowledgments Research in VV's laboratory is supported by XI OBC plan grant of University of Hyderabad. The authors wish to thank Dr. A.K. Verma for his valuable suggestions.

References

1. Yoshida H (1883) Chemistry of lacquer (urichi). J Chem Soc Trans 43:472–486
2. Claus H (2003) Laccases and their occurrence in prokaryotes. Arch Microbiol 179:145–150
3. Mayer AM, Staples RC (2002) Laccase: new functions for an old enzyme. Phytochemistry 60:551–565
4. Caparros-Ruiz D, Fornale S, Civardi L, Puigdomenech P, Rigau J (2006) Isolation and characterisation of a family of laccases in maize. Plant Sci 171:217–225
5. Xu F, Shin W, Brown SH, Wahleithner JA, Sundaram UM, Solomon EI (1996) A study of a series of recombinant fungal laccases and bilirubin oxidase that exhibit significant differences in redox potential, substrate specificity, and stability. Biochim Biophys Acta 1292:303–311
6. Eggert C, Temp U, Eriksson KE (1996) The lignolytic system of the white rot fungus *Pycnoporus cinnabarinus*: purification and characterization of the Laccase. Appl Environ Microbiol 62:1151–1158
7. Kanbi LD, Antonyuk S, Hough MA, Hall JF, Dodd FE, Hasnain SS (2002) Crystal structures of the Met148Leu and Ser86Asp mutants of rusticyanin from *Thiobacillus ferrooxidans*: insights into the structural relationship with the cupredoxins and the multi copper proteins. J Mol Biol 320:263–275
8. Solomon EI, Chen P, Metz M, Lee SK, Palmer AE (2001) Oxygen binding, activation, and reduction to water by copper proteins. Angew Chem Int Edn 40:4570–4590
9. Ducros V, Davies JG, Lawson DM, Brown SH, Østergaard P, Pedersen AH, Schneider P, Yaver DS, Brzozowski AM (1987) Crystallisation and preliminary X-ray analysis of the laccase from *Coprinus cinereus*. Acta Crystallogr 53:605–607
10. Ducros V, Brzozowski AM, Wilson KS, Brown SH, Østergaard P, Schneider P, Yaver AH, Pederson AH, Davies GJ (1998) Crystal structure of the type-2 Cu depleted laccase from *Coprinus cinereus* at 2.2 Å resolution. Nat Struct Biol 5:310–316
11. Piontek K, Antorini M, Choinowski T (2002) Crystal structure of a laccase from the fungus *Trametes versicolor* at 1.90-angstrom resolution containing a full complement of coppers. J Biol Chem 277:37663–37669
12. Thurston CF (1994) The structure and function of fungal Laccase. Microbiology 140:19–26
13. Camarero S, Ibarra D, Martínez MJ, Martínez AT (2005) Lignin-derived compounds as efficient laccase mediators for decolorization of different types of recalcitrant dyes. Appl Environ Microbiol 71:1775–1784
14. Enguita FJ, Martins LO, Henriques AO, Carrondo MA (2003) Crystal structure of a bacterial endospore coat component. A laccase with enhanced thermostability properties. J Biol Chem 278:19416–19425
15. Hakulinen N, Kiiskinen LL, Kruus K, Saloheimo M, Paananen A, Koivula A, Rouvinen J (2002) Crystal structure of a laccase from *Melanocarpus albomyces* with an intact trinuclear copper site. Nat Struct Biol 9:601–605
16. Lyashenko AV, Bento I, Zaitsev VN, Zhukhlistova NE, Zhukova YN, Gabdoulkhakov AG, Morgunova EY, Voelter W, Kachalova GS, Stepanova EV, Koroleva OG, Lamzin VS, Tishkov VI, Betzel C, Lindley PF, Mikhailov AB (2006) X-ray structural studies of the fungal laccase from *Cerrena maxima*. J Biol Inorg Chem 11:963–973
17. Messerschmidt A, Huber R (1990) The blue oxidases, ascorbate oxidase, laccase and ceruloplasmin. Modelling and structural relationships. Eur J Biochem 187:341–352
18. Solomon EI, Sundaram UM, Machonkin TE (1996) Multicopper oxidases and oxygenases. Chem Rev 96:2563–2606
19. Bourbonnais R, Paice MG (1990) Oxidation of non-phenolic substrates. An expanded role for laccase in lignin biodegradation. FEBS Lett 267:99–102
20. Verma AK, Raghukumar C, Verma P, Shouche YS, Naik CG (2010) Four marine-derived fungi for bioremediation of raw textile mill effluents. Biodegradation 2:217–233
21. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242
22. Sali A, Blundell TL (1993) Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 234:779–815
23. Laskowski RA, Macarthur MW, Moss DS, Thornton JM (1993) Procheck: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26:283–291
24. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in threedimensional structures of proteins. Nucleic Acids Res 35:407–410
25. Willard L, Ranjan A, Zhang H, Monzavi H, Boyko RF, Sykes BD, Wishart DS (2003) VADAR: a web server for quantitative evaluation of protein structure quality. Nucleic Acids Res 31:3316–3319
26. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26:1781–1802
27. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH (2009) PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 6:1–11
28. Pedretti A, Villa L, Vistoli G (2004) VEGA - An open platform to develop chemo-bioinformatics applications, using plug-in architecture and script" programming. J Comput Aided Mater Des 18:167–173
29. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723

30. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 34:116–118
31. Jones G, Willett P, Glen RC (1995) Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J Mol Biol 245:43–53
32. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748
33. Verdonk ML, Cole JC, Hartshorn MJC, Murray W, Taylor RD (2003) Improved protein-ligand docking using GOLD. Proteins 52:609–623
34. Phogat N, Vindal V, Kumar V, Krishna KI, Nirmal KP (2010) Sequence analysis, *in silico* modeling and docking studies of Caffeoyl CoA-O-methyltransferase of *Populus trichopora*. J Mol Model 16:1461–1471
35. Nirmal KP, Vindal V, Kumar V, Ashish K, Phogat N, Kumar M (2011) Structural and docking studies of *Leucaena leucocephala* Cinnamoyl CoA reductase. J Mol Model 17:533–541

ORIGINAL PAPER

Theoretical study of the local reactivity of electrophiles of the type MPR_3^+ (M=Cu, Ag, Au ;R=−H, -Me, -Ph)

Darwin Burgos · Claudio Olea-Azar · Fernando Mendizabal

Received: 21 April 2011 / Accepted: 5 August 2011 / Published online: 30 August 2011

Abstract Reactivity prediction in the series of MPR_3^+ fragments (M=Au, Ag, Cu; R=−H, -Me, -Ph) has been achieved at the *ab initio* (HF and MP2) and density functional theory (B3LYP and PBE) levels. We have used global and local descriptors based on conceptual DFT such as hardness, Fukui function and electrophilicity index. For all methods and fragments, we have found an equal trend in reactivity using both the global and local electrophilicity index: QR-$AuPR_3^+$>$CuPR_3^+$≈$AgPR_3^+$>NR-$AuPR_3^+$. It is also found that the electrophilicity power decreases as the volume of R increases.

Keywords Electrophile fragments · Quasi-Relativistic effects · Reactivity

Introduction

The coin metal phosphine fragment $[M(PR_3)]^+$ (M=Au, Ag, Cu) has been extensively used in organometallic and inorganic clusters [1–6]. The $[M(PR_3)]^+$ fragments are Lewis acids, still they are always found as terminal ligands in a large number of complexes. In particular, triphenylphosphines (PPh_3) are used in many synthesized coin clusters because it acts as a stabilizing agent [4–6]. In the literature it is possible to find several examples that illustrate this situation. For example, clusters of the $[X(AuPR_3)_n]^{+m}$ type with X=C, N, O and their analogous from the rows further down the periodic system have been studied experimental and theoretically [7–9]. Compounds of the $[Pt_3(\mu\text{-}L)_3(L')_3]$ (L=PR_3 (phosphine); SO_2, CNR (isocyanides); L'=PR_3; CNR) type act as Lewis bases toward the Lewis acids $[M(PR_3)]^+$ [10–12]. Other examples are the compounds of the type $[M(PR_3)_n]^+$ (M=Cu, Ag, Au; n=1–4) which have been known and characterized [13, 14]. These are three examples among many cases.

It is customary for computational chemists to replace the triphenylphosphine (PPh_3) ligand with trimethylphosphine (PMe_3) or phosphine (PH_3) ligands [15, 16] when the cluster models are built. The goal is a reduction of the computational cost. The influence of the phosphine ligands on the structural properties turned out to be moderate. However, the dipole moment, the first ionization potential, electron affinity, and the binding energy are described only approximately [17, 18]. Replacement of the original ligands leads to changes in the reactivity properties of the clusters. This last has been shown by Rösch and co-workers for the $MeAuPR_3$ system (R=H, Me, Ph) [17, 18]. The structural properties the PH_3 and PMe_3 ligands provide satisfactory models of the full PPh_3 ligand. However, the phosphine and trimethylphosphine ligated models tend to only approximate for energy properties and for the dipole moments [17].

The aim of the current study is to computationally predict the reactivity of the MPR_3^+ (M=Au, Ag, Cu) fragment using three types of phosphines (PR_3): triphenyl-

D. Burgos · F. Mendizabal (✉)
Departamento de Química, Facultad de Ciencias,
Universidad de Chile,
Casilla 653,
Santiago, Chile
e-mail: hagua@uchile.cl

C. Olea-Azar
Departamento de Química Inorgánica y Analítica, Facultad de
Ciencias Químicas y Farmacéuticas, Universidad de Chile,
Casilla 233,
Santiago 1, Chile

F. Mendizabal
Center for the Development of Nanoscience and Nanotechnology,
CEDENNA,
Santiago, Chile

phosphines (PPh_3), trimethylphosphine (PMe_3), and pure phosphine (PH_3). This contribution focuses on the structure and properties of the MPR_3^+ fragments by reactivity indices as introduced through conceptual density functional theory (CDFT) [19–22].

Models and computational methods

The model of the electrophiles of the type MPH_3^+, $MPMe_3^+$ and $MPPh_3^+$ (M=Au, Ag, Cu) are depicted in Fig. 1. The MPH_3^+ models assume a C_{3v} point symmetry, while $MPMe_3^+$ and $MPPh_3^+$ have a C_1 point symmetry. We first fully optimized the geometries of the fragments at the Hartree-Fock (HF), second-order Møller-Plesset perturbation theory (MP2) [23], B3LYP and PBE [24] levels.

The theoretical studies have been carried out by *ab initio* calculations available in the Gaussian03 program [25]. For the heavy elements Au, Ag and Cu, we have used pseudopotentials (PP). For gold 19-valence electron (VE) Schwerdtfeger non- and quasi-relativistic (NR and QR) PP have been used [26]. The silver and copper atoms were treated by a 19-VE Stuttgart quasi-relativistic pseudopotentials [27]. Two *f*-type polarization functions were added: Au (α_f=0.20, 1.19), Ag (α_f=0.22, 1.72), and Cu (α_f=0.24, 3.70) [28]. The C and P atoms were also treated with PP, using a double-zeta basis set and adding one d-type polarization function [29]. For hydrogen, a valence-double-zeta basis set with one p-polarization function was used [30].

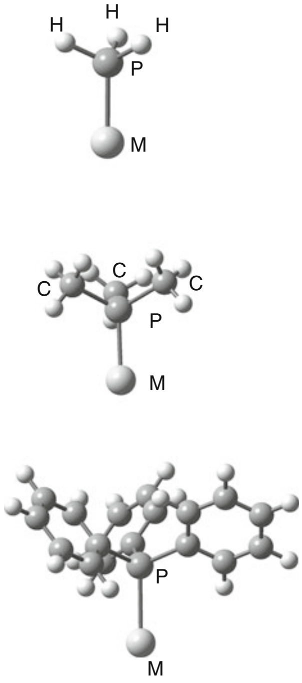


Fig. 1 The MPH_3^+, $MPMe_3^+$ and $MPPh_3^+$ (M=Au, Ag, Cu) models

With the aim of understanding the properties of the electrophilic MPH_3^+, $MPMe_3^+$ and $MPPh_3^+$ (M=Au, Ag, Cu) fragments we have used the CDFT. The chemical potential (μ) and chemical hardness (η) from operational DFT [31–35], which are defined as:

$$\mu \approx -\frac{(IP + EA)}{2} \tag{1}$$

$$\eta \approx \frac{(IP - EA)}{2}, \tag{2}$$

where IP is the ionization potential and EA is the electron affinity. These two quantities can also be defined based on the frontier molecular orbitals eigenvalues; on the basis of Koopmans' theorem as $IP \approx -E_{HOMO}$ and $EA \approx -E_{LUMO}$, where E_{HOMO} and E_{LUMO} are the energies of the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), respectively. Moreover, this definition of chemical potential is related to Mulliken's definition of electronegativity (χ) called absolute electronegativity: $\chi = -\mu$ [31].

On the other hand, the electrophilicity index (ω) is defined as [34]

$$\omega = \frac{\mu^2}{2\eta} . \tag{3}$$

It is a measure of the electrophilicity of the fragment. The ω is called the "electrophilicity index" and from a classical electrostatic point of view is considered to be a measure of electrophilic power as was defined by Parr and co-workers [34]. Also, they have shown that ω measures the second-order energy of an electrophile when it gets saturated with electrons. The higher its value, the greater its electrophilicity. In addition, to see reactive sites, the orbital Fukui local function [36, 37] for electrophilic fragments was determined from its frontier orbital density at atom *M,* where *M* represents a metal atom. The orbital Fukui function at atom *M* for nucleophilic attack is given as:

$$f_M^{\alpha} = \sum_{\nu \in M}^{AO} C_{\nu\alpha}^2 + \sum_{\chi \neq \nu}^{AO} C_{\chi\alpha} C_{\nu\chi} S_{\nu\chi} \quad , \tag{4}$$

where α=+ for LUMO, $C_{\nu\alpha}$ are the molecular orbital frontier expansion coefficients (LUMO) and $S_{\nu\chi}$ are the atomic orbital overlap matrix elements. This definition of the orbital Fukui function has been used in several studies yielding reliable results [38].

Moreover, a local counterpart of electrophilicity has been introduced to analyze the electrophile-nucleophile reactions [39]. It is defined as

$$\omega_M^+ = \omega f_M^+ \ . \tag{5}$$

The MPR_3^+ metal electrophile fragments studied in this work act as soft acids. Therefore, they are in the category of orbital controlled reactions. An analysis of local electrophilicity (ω_k^+) provides the information of a particular atomic site in a molecule being attacked by a nucleophile. This local property has been proposed as a better intermolecular reactivity index than the Fukui function itself for analyzing electrophile-nucleophile interactions becuase the Fukui function allows to compare the sites selectivity within a molecule, while for a comparison of the reactivity of a specific site on different molecules, it is more appropriate to use a property that includes intrinsic information of the systems studied [19–22]. In this context, the local softness and the local electrophilicity descriptors are suitable to explain hard-soft and electrophile-nucleophile interactions, respectively. Thus, the Fukui function is used as a distribution function which allows to map any global property within a molecule. In the literature some of these aspects have been verified [40].

Table 1 Main geometric parameters of the MPH_3^+ electrophiles (distances in pm and angles in degrees) at different levels of calculation

Electrophile	Method	MP	PH	MPHº
$AuPH_3^+$	HF-QR	239.2	140.2	114.47º
	MP2-QR	225.1	140.8	113.05º
	B3LYP-QR	229.5	142.0	113.47
	PBE-QR	226.7	143.5	113.28º
	MP2-QR [13]	229.1	139.6	114.30º
$AuPH_3^+$	HF-NR	282.8	140.8	117.30º
	MP2-NR	260.5	141.1	117.00º
	B3LYP-NR	265.2	142.3	117.09º
	PBE-NR	260.3	143.7	117.22º
	MP2-NR [13]	269.6	140.0	117.40º
$AgPH_3^+$	HF	260.9	140.6	116.45º
	MP2	240.5	141.1	115.82º
	B3LYP	243.9	142.2	115.85º
	PBE	239.2	143.6	115.88º
	MP2 [13]	247.5	139.8	116.40º
$CuPH_3^+$	HF	237.2	140.5	115.98º
	MP2	217.7	141.0	115.40º
	B3LYP	221.3	142.2	115.30º
	PBE	217.6	143.6	115.33º
	MP2 [13]	220.6	139.8	115.70º

Table 2 Main geometric parameters of the $MPMe_3^+$ electrophiles (distances in pm and angles in degrees) at different levels of calculation

Electrophile	Method	MP	PC	MPCº
$AuPMe_3^+$	HF-QR	237.5	183.2	111.63
	MP2-QR	223.9	182.5	110.11
	B3LYP-QR	230.2	183.7	110.83
	PBE-QR	227.7	183.7	110.48
$AuPMe_3^+$	HF-NR	276.2	184.4	113.73
	MP2-NR	255.8	184.1	113.27
	B3LYP-NR	262.1	184.9	114.09
	PBE-NR	257.9	185.1	113.43
$AgPMe_3^+$	HF	256.0	184.1	113.13
	MP2	237.1	183.7	112.44
	B3LYP	242.2	184.5	112.78
	PBE	238.3	184.6	112.30
$CuPMe_3^+$	HF	234.2	183.9	112.95
	MP2	216.1	183.6	112.59
	B3LYP	220.9	184.4	112.43
	PBE	218.2	184.5	112.08

Results and discussion

Structural description

Table 3 Main geometric parameters of the $MPPh_3^+$ electrophiles (distances in pm and angles in degrees) at different levels of calculation

Electrophile	Method	MP	PC	MPCº
$AuPPh_3^+$	HF-QR	238.2	183.0	110.48º
	MP2-QR	223.2	180.8	108.95
	B3LYP-QR	231.7	183.0	110.27º
	PBE-QR	229.3	183.0	109.97º
$AuPPh_3^+$	HF-NR	275.2	184.0	112.19º
	MP2-NR	253.2	182.6	112.25º
	B3LYP-NR	262.8	183.9	111.68º
	PBE-NR	257.9	184.3	110.86º
$AgPPh_3^+$	HF	255.7	183.7	111.65º
	MP2	235.4	182.2	111.35º
	B3LYP	242.9	184.0	110.59º
	PBE	239.2	184.0	110.32º
$CuPPh_3^+$	HF	233.6	183.6	111.32º
	MP2	215.1	182.2	111.41º
	B3LYP	221.3	183.6	111.22º
	PBE	218.7	183.9	109.84º

All models were assumed as a singlet ground state. Tables 1, 2, 3 summarize the main geometric parameters at several theoretical levels. The models are shown in Fig. 1. In

Table 4 Ionization potential (I), electron affinity (A), electronic chemical potential (μ), chemical hardness (η), global electrophilicity index (ω). All values are in eV

Electrophile	Method	I	A	-μ	η	ω
$AuPH_3^+$	HF-QR	16.23	4.89	10.56	5.66	9.84
	MP2-QR	16.27	4.77	10.52	5.75	9.62
	B3LYP-QR	13.56	8.64	11.10	2.46	25.04
	PBE-QR	12.47	9.19	10.83	1.64	35.76
$AuPH_3^+$	HF-NR	15.58	4.39	9.99	5.60	8.91
	MP2-NR	15.88	4.28	10.08	5.80	8.76
	B3LYP-NR	13.20	7.75	10.48	2.73	20.12
	PBE-NR	12.30	8.28	10.29	2.01	26.34
$AgPH_3^+$	HF	16.03	4.53	10.28	5.75	9.19
	MP2	16.26	4.42	10.34	5.92	9.03
	B3LYP	13.47	8.06	10.77	2.71	21.42
	PBE	15.99	5.92	10.96	5.04	28.93
$CuPH_3^+$	HF	16.59	4.47	10.53	6.06	9.14
	MP2	16.72	4.33	10.53	6.20	8.94
	B3LYP	13.27	8.04	10.66	2.61	21.73
	PBE	16.13	5.92	11.03	5.11	32.62
$AuPMe_3^+$	HF-QR	14.57	4.10	9.33	5.24	8.31
	MP2-QR	14.52	3.92	9.22	5.30	8.02
	B3LYP-QR	12.12	7.46	9.79	2.33	20.57
	PBE-QR	11.20	7.97	9.59	1.62	28.47
$AuPMe_3^+$	HF-NR	13.72	3.88	8.80	4.92	7.87
	MP2-NR	13.86	3.76	8.81	5.05	7.68
	B3LYP-NR	11.47	6.97	9.22	2.25	18.89
	PBE-NR	10.63	7.41	9.02	1.61	25.27
$AgPMe_3^+$	HF	14.13	3.95	9.04	5.09	8.03
	MP2	14.22	3.82	9.02	5.20	7.82
	B3LYP	11.81	7.13	9.47	2.34	19.16
	PBE	10.90	7.61	9.26	1.65	25.98
$CuPMe_3^+$	HF	14.60	3.87	9.24	5.37	7.95
	MP2	14.75	3.76	9.26	5.49	7.81
	B3LYP	11.87	7.07	9.47	2.40	18.68
	PBE	10.63	7.59	9.11	1.52	27.30
$AuPPh_3^+$	HF-QR	11.96	3.67	7.82	4.15	7.37
	MP2-QR	11.87	3.43	7.65	4.22	6.93
	B3LYP-QR	10.25	6.77	8.51	1.74	20.81
	PBE-QR	9.59	7.22	8.41	1.19	29.72
$AuPPh_3^+$	HF-NR	11.52	3.57	7.55	3.98	7.16
	MP2-NR	11.53	3.39	7.46	4.10	6.25
	B3LYP-NR	9.77	6.46	8.12	1.66	19.85
	PBE-NR	9.12	6.82	7.97	1.15	27.62
$AgPPh_3^+$	HF	11.69	3.59	7.64	4.05	7.21
	MP2	11.66	3.40	7.53	4.13	6.86
	B3LYP	9.84	6.54	8.19	1.65	20.33
	PBE	9.30	7.00	8.15	1.15	28.88
$CuPPh_3^+$	HF	11.81	3.47	7.64	4.17	7.00
	MP2	11.81	3.32	7.57	4.25	6.74
	B3LYP	10.05	6.45	8.25	1.80	18.91
	PBE	9.33	6.94	8.14	1.20	27.61

the literature, there are theoretical models of the $[M(PH_3)]^+$ type described by Schwerdtfeger and co-workers at the MP2 level [13]. Small deviations are due to pseudopotential type and size of the basis sets (see Table 1). The geometric magnitudes of MP2 are the shortest, followed by PBE, B3LYP, and finally HF, which are always higher. If we take as reference the M-P distance in the different models, it is clear that electronic correlation effects play an important role in the stability of the system. The M-P distances obtained with all methods are close to those of a typical single bond, with the shortest distance obtained with the MP2 method [13].

Taking into account the non-relativistic case, the distance of the M-P bond is Cu>Ag>Au, but this changes when quasi-relativistic effects are introduced, getting the order Cu<Au<Ag. As explained in the literature, the relativistic effects increase the electronegativity of the gold atom (from ca. 1.9 to 2.4) [41], which enhances the possibility for sigma-charge donation from the PR_3 lone pair [13]. This is because the relativistic effects decrease the size of the gold atom and change the energy patterns of the frontier orbitals. Also, we can see that when the ligand goes from PH_3 to PMe_3 and finally PPh_3, the M-P distance decreases as the volume of the substituent R increases. This shows the nature of the R group in the PR_3 ligand. This has an important influence on the bonding and stability of these phosphine complexes, as was demonstrated by Rösch and co-workers for several gold phosphine complexes [2, 17, 18].

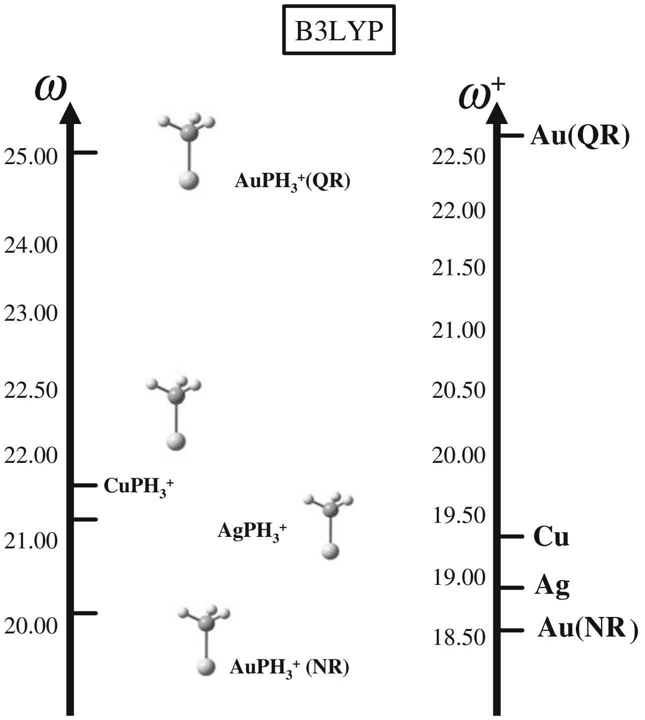


Fig. 2 Molecular and atomic representations of the electrophiles with global and local electrophilicity indices (ω, ω^+) at the B3LYP level

Table 5 NBO analysis of the MP2 density for MPH_3^+, $MPMe_3^+$ and $MPPh_3^+$ complexes(M=Au, Ag, Cu)

System	Method	M	P	R	Natural electron configuration on metal
$AuPH_3^+$	QR	0.4918	0.2507	0.0858	$6S^{0.67}5\,d^{9.67}6p^{0.04}5f^{0.09}6\,d^{0.05}$
	NR	0.8707	−0.0850	0.0714	$6S^{0.15}5\,d^{9.85}6p^{0.03}5f^{0.08}6\,d^{0.04}7p^{0.01}$
$AgPH_3^+$	QR	0.7910	−0.0282	0.0790	$5S^{0.25}4\,d^{9.83}5p^{0.03}4f^{0.07}5\,d^{0.05}5f^{0.01}$
$CuPH_3^+$	QR	0.7559	0.0006	0.0801	$4S^{0.30}3\,d^{9.78}4p^{0.03}4\,d^{0.13}5p^{0.01}4f^{0.02}$
$AuPMe_3^+$	QR	0.2999	1.1881	−0.9501	$6S^{0.89}5\,d^{9.65}6p^{0.04}5f^{0.09}6\,d^{0.05}$
	NR	0.7889	0.7796	−0.9310	$6S^{0.24}5\,d^{9.83}6p^{0.03}5f^{0.09}6\,d^{0.04}7p^{0.01}$
$AgPMe_3^+$	QR	0.6726	0.8642	−0.9335	$5S^{0.38}4\,d^{9.81}5p^{0.03}4^{0.07}5\,d^{0.05}5f^{0.01}$
$CuPMe_3^+$	QR	0.6445	0.8776	−0.9335	$4S^{0.43}3\,d^{9.76}4p^{0.03}4\,d^{0.13}5p^{0.01}4f^{0.02}5f^{0.01}$
$AuPPh_3^+$	QR	0.2523	1.3158	−0.4446	$6S^{0.94}5\,d^{9.64}6p^{0.04}5f^{0.09}6\,d^{0.05}$
	NR	0.8808	0.8816	−0.4400	$6S^{0.14}5\,d^{9.98}$
$AgPPh_3^+$	QR	0.6650	0.9566	−0.4114	$5S^{0.39}4\,d^{9.81}5p^{0.02}4f^{0.07}5\,d^{0.05}5f^{0.01}$
$CuPPh_3^+$	QR	0.6452	0.9597	−0.4103	$4S^{0.43}3\,d^{9.76}4p^{0.03}4\,d^{0.13}5p^{0.01}4f^{0.02}5f^{0.01}$

Global properties

Using CDFT, the most stable species in a group of similar complexes is the one with the greatest chemical hardness (according to the principal of maximum hardness, PMH [33, 42]). The global properties were obtained from Eqs. 1, 2 and 3, see Table 4. The chemical potential and hardness are computed from ionization energy and electron affinity, with those at the *ab initio* level (HF and MP2) better than those at the DFT level (B3LYP and PBE). It is know that in calculations with DFT the electronic density decays faster than at the *ab initio* level [43]. This is manifested in the μ and η values in Table 4. This effect is seen in all MPR_3^+ fragments. Anyway, regardless of the fragment, the trend is maintained. Also, this is obtained for the electrophilicity index (ω) values.

When we analyze the first set of MPH_3^+ fragments $CuPH_3^+$ is found as the hardest group. This behavior is the most stable of the series regardless of the method. Then either $AgPH_3^+$ or NR-$AuPH_3^+$ and finally QR-$AuPH_3^+$ show the minimum chemical hardness, or in other words, the softer fragment is predicted for $AuPH_3^+$ when the relativistic effects are described. This situation changes when we use $MPMe_3^+$ and $MPPh_3^+$ fragments. The $CuPR_3^+$ is still the hardest, the second is QR-$AuPR_3^+$, followed by $AgPR_3^+$ and finally NR-$AuPR_3^+$. As can be noted the size of R groups alter the reactivity trends in MPR_3^+ fragments by decreasing η as the size increases.

Table 4 shows the global electrophilicity index (ω) of the electrophile fragments. As Chattaraj and Roy have described [40] "*During an electrophile and nucleophile interaction process, when two reactants approach each other from a large distance, they feel only the effect of the global electrophilicity of each other and not its local counterpart. The molecule with the large ω value will act as an electrophile, and the other will behave as the nucleophile*". This effect is seen in the fragments analyzed, the most and least electrophilic are predicted for QR-$AuPR_3^+$ and NR-$AuPR_3^+$, respectively. This shows the importance of relativistic effects in the reactivity indices based on the

Table 6 Condensed Fukui function on metal center (f_M^+)

Electrophile	Method	f_M^+ (HF)	f_M^+ (MP2)	f_M^+ (B3LYP)	f_M^+ (PBE)
$AuPH_3^+$	QR	0.7940	0.9094	0.9085	0.7588
	NR	0.9799	0.9809	0.9224	0.9045
$AgPH_3^+$	QR	0.9693	0.9694	0.8860	0.8626
$CuPH_3^+$	QR	0.9774	0.9838	0.8913	0.8683
$AuPMe_3^+$	QR	0.9699	0.9729	0.8827	0.8590
	NR	0.8889	0.8883	0.7414	0.7061
$AgPMe_3^+$	QR	0.9572	0.9604	0.8453	0.8157
$CuPMe_3^+$	QR	0.9673	0.9635	0.8572	0.8295
$AuPPh_3^+$	QR	0.9588	0.9706	0.8545	0.8240
	NR	0.8634	0.8548	0.6816	0.6395
$AgPPh_3^+$	QR	0.9428	0.9504	0.8096	0.7708
$CuPPh_3^+$	QR	0.9558	0.9558	0.8154	0.7749

Fig. 3 The LUMO orbital of MPH_3^+ (M=Au, Ag, Cu) models at the B3LYP levels

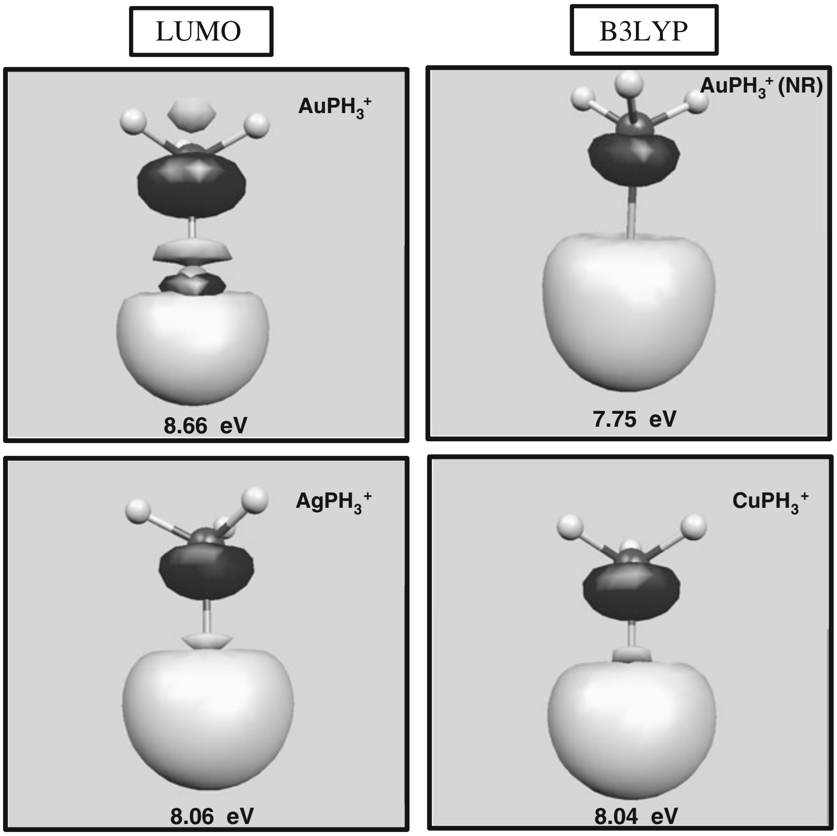
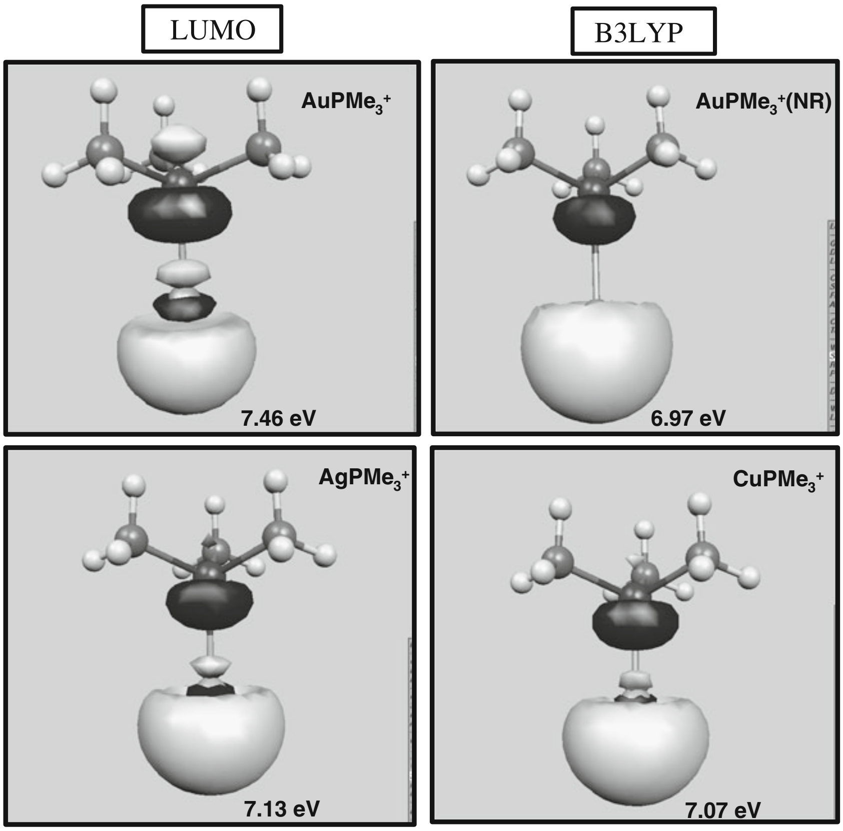


Fig. 4 The LUMO orbital of $MPMe_3^+$ (M=Au, Ag, Cu) models at the B3LYP levels



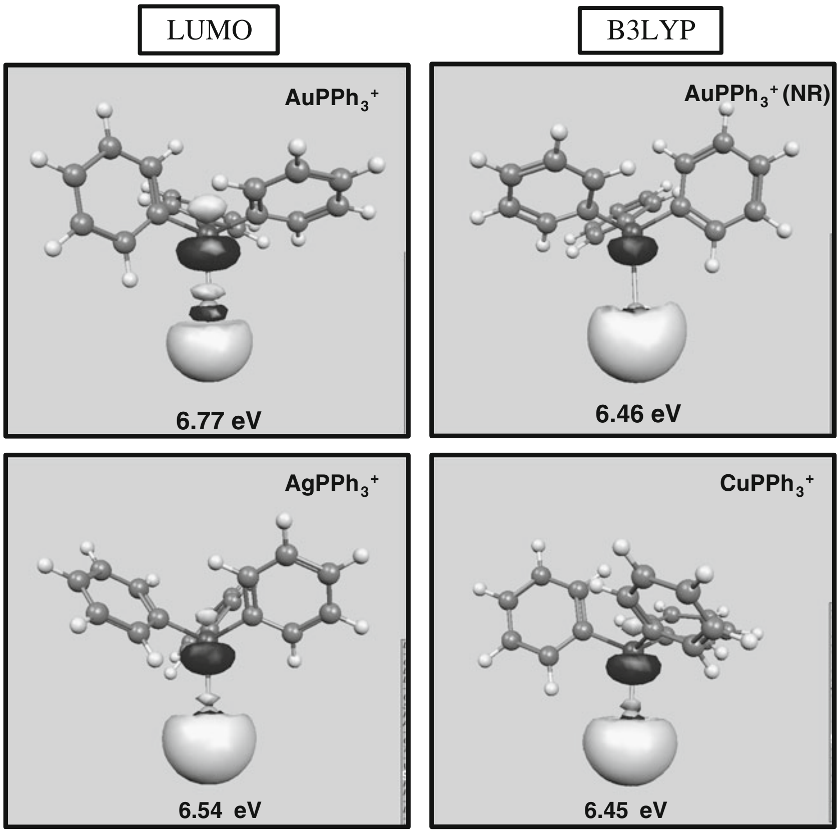


Fig. 5 The LUMO orbital of $MPPh_3^+$ (M=Au, Ag, Cu) models at the B3LYP levels

conceptual DFT. $CuPR_3^+$ and $AgPR_3^+$ are in an intermediate position (see Fig. 2). On the other hand, the effect of the R group is noticeable because when going from -H to –Ph the ω values decrease. The Ph group linked to phosphorus decreases the reactivity of all the fragments possibly because the positive charge is more delocalized. In general, the electrophilicity index also shows that the QR-Au fragment is the most electrophilic among the systems studied, indicating that it is a better electron acceptor.

When electronegativity is used as the negative of chemical potential, in Table 4, we can seen that such magnitude always increases in fragments of gold with relativistic effects (QR) in comparison to the non-relativistic (NR). The $CuPR_3^+$ and $AgPR_3^+$ are in an intermediate position. The effect of the R group is noticeable because when going from -H to –Ph the electronegativity values decrease according to all fragments.

Local properties

In this section, the local properties centered on the metal of the fragment are analyzed. In this sense, we have performed

Table 7 Local electrophilicity index on the metal center (ω^+). The values are in eV

Electrophile	Method	ω^+ (HF)	ω^+ (MP2)	ω^+ (B3LYP)	ω^+ (PBE)
$AuPH_3^+$	QR	7.81	8.75	22.76	27.13
	NR	8.73	8.59	18.56	23.82
$AgPH_3^+$	QR	8.91	8.75	18.98	24.96
$CuPH_3^+$	QR	8.93	8.80	19.37	28.32
$AuPMe_3^+$	QR	8.06	7.61	18.16	24.46
	NR	6.99	6.82	14.01	17.84
$AgPMe_3^+$	QR	7.69	7.51	16.20	21.19
$CuPMe_3^+$	QR	7.69	7.53	16.01	22.65
$AuPPh_3^+$	QR	7.07	6.73	17.78	24.49
	NR	6.18	5.34	13.53	17.66
$AgPPh_3^+$	QR	6.80	6.52	16.46	22.26
$CuPPh_3^+$	QR	6.69	6.44	15.42	21.39

the natural bond orbital (NBO) population analysis by different methods (HF, MP2, B3LYP and PBE), with similar results. Table 5 shows a summary at the MP2 level. The initial addition of the PR_3 ligand to M^+ decreases the charge on the metal center, but even more so for QR-Au. The relativistic effects have a strong influence on the coordination of gold. For MPH_3^+ fragments similar results were obtained by Schwerdtfeger and co-workers [13]. By going from PH_3 to PMe_3 to PPh_3, the charge on the metal center decreases and the charge on the phosphorus atom increases in the same proportion. The charge flow comes from phosphorus and goes toward both the metal and R group. It is greater when the electrophile is the QR-Au fragment, followed by Cu, Ag, and finally NR-Au.

The local reactivity of electrophile fragments has been studied using the local orbital nucleophilic Fukui function on the metal center (f_M^+). In all the fragments the highest value for the nucleophilic Fukui function is expected at the metal center. Low values are expected at the other atoms. For clarity, we have not included the values of f_k^+ for the other atoms of each fragment. The results are summarized in Table 6. For comparison, for all methods the same trend is found: QR-Au>Cu>Ag>NR-Au with PMe_3 and PPh_3. If PH_3 is used, the values of QR-Au and NR-Au are similar. We are using the approximate relationship for the nucleophilic Fukui function expressed in terms of the frontier LUMO (Φ_{LUMO}), as: $f+\approx-\Phi_{LUMO}|^2$ [36]. Figures 3, 4, 5 show the LUMO level in the MPR_3^+ fragments at the B3LYP level. Similar results are obtained for the other methods used in this work. Regardless of the fragment, the stabilization of the LUMO energy follows the same trend: $QR\text{-}AuPR_3^+>CuPR_3^+\approx AgPR_3^+>NR\text{-}AuPR_3^+$.

On the other hand, the local electrophilicity index on the metal (ω_M^+) gives the final trend in Table 7. We can emphasize that the global variation in the electrophilicity index is modulated through the local variations being mapped in the more reactive site, as indicated through the Fukui function [39] using the ω_M^+. This means that the variation of the electrophilic power is directed to the sites where the Fukui function for nucleophilic attacks is important. For comparison, the same trend is found for all methods and fragments: QR-Au>Cu≈Ag>NR-Au. This behavior is shown in Fig. 2.

In a previous work, we studied the formation of complexes of type $[Pt_3(\mu\text{-}CO)_3(PH_3)_3]\text{-}MPH_3^+$ [44]. The energy of formation of the complex follows the Au>Cu>Ag trend. Experimentally, the order of formation is the same, Au>Cu>Ag, which indicates that gold is the most stable complex with the $[Pt_3(\mu\text{-}CO)_3(PH_3)_3]$ nucleophile cluster. The local reactivity of the electrophile fragments (MPH_3^+) is correlated with the interaction energy of $[Pt_3(\mu\text{-}CO)_3(PH_3)_3]\text{-}MPH_3^+$.

Conclusions

According to the orbital Fukui and electrophilicity local indices on the metal, reactivity in the $[MPR_3]^+$ series increases from NR-Au<Ag≈Cu<QR-Au. The effect of the R group is noticeable since when going from -H to –Ph the global and local indices decrease. The -Ph group does decrease the reactivity of all the fragments. The local electrophilicity index indicates that the QR-Au fragment is the most electrophilic of the three, indicating that it is a better electron acceptor. The results show the importance of relativistic effects on the reliability of the reactivity indices based on the conceptual DFT.

Acknowledgments This research was financed by Fondo Nacional de Ciencia y Tecnología (FONDECYT) under Project 1100162 (Conicyt-Chile), Project Millennium P07-006-F and Basal Financing Program Comisión Nacional de Ciencia y Tecnología (CONICYT)-FB0807 ([3]Center for the Development of Nanoscience and Nanotechnology, CEDENNA). D.B. has a National Doctoral Fellowship (CONICYT).

References

1. Imhof D, Venanzi LM (1994) Chem Soc Rev 23:185–193
2. Häberlen OD, Schmidbaur H, Rösch N (1994) J Am Chem Soc 116:8241–8248
3. Burdett JK, Eisensteis O, Schweizer WB (1994) Inorg Chem 33:3261–3268
4. Pyykkö P (2004) Angew Chem Int Edn 43:4412–4456
5. Gimeno MC, Laguna A (2008) 37:1952–1966
6. Schmidbaur H, Schier A (2008) 37:1931–1951
7. Schmidbaur H, Hofreiter S, Paul M (1995) Nature 377:503–504
8. Angermair K, Schmidbaur H (1994) Chem Ber 127:2387–2391
9. Canales F, Gimeno MC, Jones PG, Laguna A (1994) Angew Chem Int Ed Engl 33:769–773
10. Evans DG, Hallam MF, Mingos DMP, Wardle RWM (1987) J Chem Soc Dalton Trans 9:1889–1895
11. Moor A, Pregosin PS, Venanzi LM (1982) Inorg Chim Acta 61:135–140
12. Braunstein P, Freyburger S, Bars O (1988) J Organomet Chem 352:C29–C33
13. Schwerdtfeger P, Hermann HL, Schmidbaur H (2003) Inorg Chem 42:1334–1342
14. Melnik M, Parish RV (1986) Coord Chem Rev 70:157–185
15. Li J, Pyykkö P (1993) Inorg Chem 32:2630–2634
16. Canales S, Crespo O, Gimeno MC, Jones P, Laguna A, Mendizabal F (2001) Organometallics 20:4812–4848
17. Häberlen OD, Chung SC, Rösch N (1994) Int J Quantum Chem 28:595–603
18. Häberlen O, Rösch N (1993) J Phys Chem 97:4970–4973
19. Geerlings P, De Proft F, Langenaeker W (2003) Chem Rev 103:1793–1873
20. Torrent-Sucarrat M, De Proft F, Geerlings P, Ayers PW (2008) Chem Eur J 14:8652–8660
21. Sablon N, Mastalerz R, De Proft F, Geerlings P, Reiher M (2010) Theor Chem Acc 127:195–202
22. De Vleeschouwer F, Jaque P, Geerlings P, Toro-Labbé A, De Proft F (2010) J Org Chem 75:4964–4974

23. Møller C, Plesset MS (1934) Phys Rev 46:618–622
24. Perdew JP, Burke K, Ernzerhof M (1996) Phys Rev Letter 77:3865–3868
25. Frisch MJ, Trucks GW, Schlegel HB, Gill PMW, Johnson BG, Robb MA, Cheeseman JR, Keith KT, Petersson GA, Montgomery JA, Raghavachari K, Al-Laham MA, Zakrzewski VG, Ortiz JV, Foresman JB, Cioslowski J, Stefanov BB, Nanayakkara A, Challacombe M, Peng CY, Ayala PY, Chen W, Wong MW, Andres JL, Replogle ES, Gomperts R, Martin RL, Fox DJ, Binkley JS, Defrees DJ, Baker J, Stewart JP, Head-Gordon M, Gonzalez C, Pople JA (2003) Gaussian 03. Gaussian Inc, Pittsburgh, PA
26. Schwerdtfeger P, Dolg M, Schwarz WHE, Bowmaker GA, Boyd PDW (1989) J Chem Phys 91:1762–1774
27. Andrae M, Heisserman M, Dolg M, Stoll H, Preuss H (1990) Theor Chim Acta 77:123–141
28. Pyykkö P, Runeberg N, Mendizabal F (1997) Chem Eur J 3:1451–1457
29. Bergner A, Dolg M, Küchle W, Stoll H, Preuss H (1993) Mol Phys 80:1431–1441
30. Huzinaga S (1965) J Chem Phys 42:1293–1301
31. Parr RG, Donnelly RA, Levy M, Palke WE (1978) J Chem Phys 68:3801–3808
32. Parr RG, Yang W (1989) Density Functional Theory for atoms and molecules. Oxford Press, New York
33. Parr RG, Pearson RG (1983) J Am Chem Soc 105:7512–7516
34. Parr RG, Von Szentpaty L, Liu S (1999) J Am Chem Soc 121:1922–1924
35. Parr RG, Pearson R (1984) J Am Soc 106:4049–4050
36. Pérez P, Contreras R (1998) Chem Phys Letters 293:239–244
37. Contreras R, Fuentealba P, Galván M, Pérez P (1999) Chem Phys Letters 304:405–413
38. Fuentealba P, Pérez P, Contreras R (2000) J Chem Phys 113:2544–2551
39. Pérez P, Toro-Labbé A, Aizman A, Contreras R (2002) J Org Chem 67:4747–4752
40. Chattaraj PK, Roy DR (2006) Chem Rev 106:2065–2091
41. Schwerdtfeger P (1991) Chem Phys Lett 183:457–463
42. Parr RG, Gázquez JL (1993) J Phys Chem 97:3939–3948
43. Muñiz J, Sansores LE, Pyykkö P, Martínez A, Salcedo R (2009) J Mol Model 15:1165–1173
44. Donoso, Mendizabal F (2011) Theor Chem Acc 129:381–387

ORIGINAL PAPER

Transient pockets on XIAP-BIR2: toward the characterization of putative binding sites of small-molecule XIAP inhibitors

Susanne Eyrisch · Jose L. Medina-Franco ·
Volkhard Helms

Received: 12 May 2011 / Accepted: 9 August 2011 / Published online: 30 August 2011

Abstract Protein-protein interactions are abundant in signal transduction pathways and thus of crucial importance in the regulation of apoptosis. However, designing small-molecule inhibitors for these potential drug targets is very challenging as such proteins often lack well-defined binding pockets. An example for such an interaction is the binding of the anti-apoptotic BIR2 domain of XIAP to the pro-apoptotic caspase-3 that results in the survival of damaged cells. Although small-molecule inhibitors of this interaction have been identified, their exact binding sites on XIAP are not known as its crystal structures reveal no suitable pockets. Here, we apply our previously developed protocol for identifying transient binding pockets to XIAP-BIR2. Transient pockets were identified in snapshots taken during four different molecular dynamics simulations that started from the caspase-3:BIR2 complex or from the unbound BIR2 structure and used water or methanol as solvent. Clustering of these pockets revealed that surprisingly many pockets opened in the flexible linker region that is involved in caspase-3 binding. We docked three known inhibitors into these transient pockets and so determined five putative binding sites. In addition, by docking two inactive compounds of the same series, we show that this protocol is also able to distinguish between binders and nonbinders which was not possible when docking to the crystal structures. These findings represent a first step toward the understanding of the binding of small-molecule XIAP-BIR2 inhibitors on a molecular level and further highlight the importance of considering protein flexibility when designing small-molecule protein-protein interaction inhibitors.

Keywords Apoptosis · Docking · Inhibitors · Molecular dynamics simulation · Pocket detection · Protein-protein interaction · Transient pockets

Introduction

Apoptosis is a process that enables multi-cellular organisms to preserve their viability by selectively inducing the death of damaged cells. The decision whether a cell divides or undergoes apoptosis is controlled by cell signals that may originate either on the cell inside (intrinsic inducer) or outside (extrinsic inducer). However, in many diseases like cancer, the cell's inherent capability to kill itself is disturbed and the damaged cell proliferates in an uncontrolled way. Thus, a promising therapeutic strategy is modulating the involved signal transduction to activate the apoptosis pathway. The challenge in this approach is that signal transduction pathways mainly involve interactions between proteins and such drug targets are commonly considered as "high-hanging fruits" [1].

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1217-y) contains supplementary material, which is available to authorized users.

S. Eyrisch · V. Helms (✉)
Center for Bioinformatics,
Building E2 1, P.O. Box 151150, 66041 Saarbruecken, Germany
e-mail: volkhard.helms@bioinformatik.uni-saarland.de

J. L. Medina-Franco
Torrey Pines Institute for Molecular Studies,
11350 SW Village Parkway,
Port St. Lucie, FL 34987, USA

Present Address:
S. Eyrisch
Priaxon AG,
Gmunder Str. 37-37a,
81379 München, Germany

Inhibitors of Apoptosis proteins (IAPs) are endogenous caspase inhibitors [2, 3] that share a conserved structure, the BIR domain [4]. As caspases are responsible for apoptosis, their inhibition leads to the survival of damaged cells and, thus, to tumor proliferation [5, 6]. Not surprisingly, some IAP family proteins are commonly overexpressed in human cancers [7] and are therefore important drug targets. X-chromosome linked inhibitor of apoptosis (XIAP) is the best-characterized member of the IAP family. It is composed of three BIR domains (called BIR1 to BIR3) and a RING zinc-finger motif. BIR2 and the linker region connecting BIR2 to BIR1 bind and inhibit caspase-3 and -7, while BIR3 suppresses caspase-9 [8, 9]. The activity of XIAP is regulated by inhibitory proteins like Smac that disrupt XIAP-caspase complexes and thus reconstitute caspase activity [10]. While the molecular details of the interactions with caspase-3 [11], -7 [12], and -9 [13], as well as the interaction of the BIR3 domain with Smac [14, 15] have been resolved, it is still unclear whether Smac also binds to the BIR2 domain.

The interaction between Smac and XIAP-BIR3 has been used as a template to design small-molecules inhibiting caspase-9 binding. The X-ray structures of XIAP-BIR3 complexed with these compounds confirm that they bind into the Smac pocket [16–20]. Likewise, several classes of small-molecule compounds have been identified that selectively target the interaction between XIAP-BIR2 and caspase-3 [21–23], but in contrast to the XIAP-BIR3 antagonists, the structural basis for this inhibition is not yet revealed. The X-ray structure of the BIR2 - caspase-3 complex [11] shown in Fig. 1 and site-directed mutagenesis studies [24] reveal that the interaction interface involves mainly the linker region (residues 124-168) of BIR2 and that the XIAP residues Leu140, Leu141, Val146, Asp148, Ile149, Ile153, Tyr154, Arg156, Phe228, and Arg233 essentially contribute to the interaction. Other publications [21, 23] reported a series of polyphenylurea-based compounds and studied their mechanism of action using biochemical, molecular biological, and genetic methods. Since the inhibitors did not compete with Smac, it was hypothesized that they bind to the linker region. However, the exact binding site and mode was unknown. Moreover, the NMR structure of unbound XIAP-BIR2 reveals that the linker region is highly flexible [25]. This impeded structure-based drug design attempts using the apo NMR structure or the X-ray structure of the XIAP-BIR2 - caspase-3 complex.

We have previously presented a computational protocol for identifying transient pockets that is able to provide starting points for structure-based drug design especially for such challenging systems [26, 27]. For the three protein systems MDM2, BCL-X_L, and interleukin-2 (IL-2), we found that large pockets not detectable in the apo crystal/NMR structures opened frequently on the protein surfaces during standard molecular dynamics (MD) simulations of 10 nanoseconds length at room temperature. At the native binding site, pockets of similar size as with a known inhibitor bound could indeed be observed for all three systems. Docking known inhibitors with AutoDock3 [28] into these transient pockets resulted in docking results with smaller than 2 Å root mean square deviation (RMSD) from the crystal structures.

In a subsequent study, we could show that, when the water solvent was replaced by methanol, the transient pockets opening in the MD simulations tended to be larger and less polar [27]. Moreover, the docking results improved significantly for two of the three systems. In a subsequent study, the pocket detection protocol was applied to the

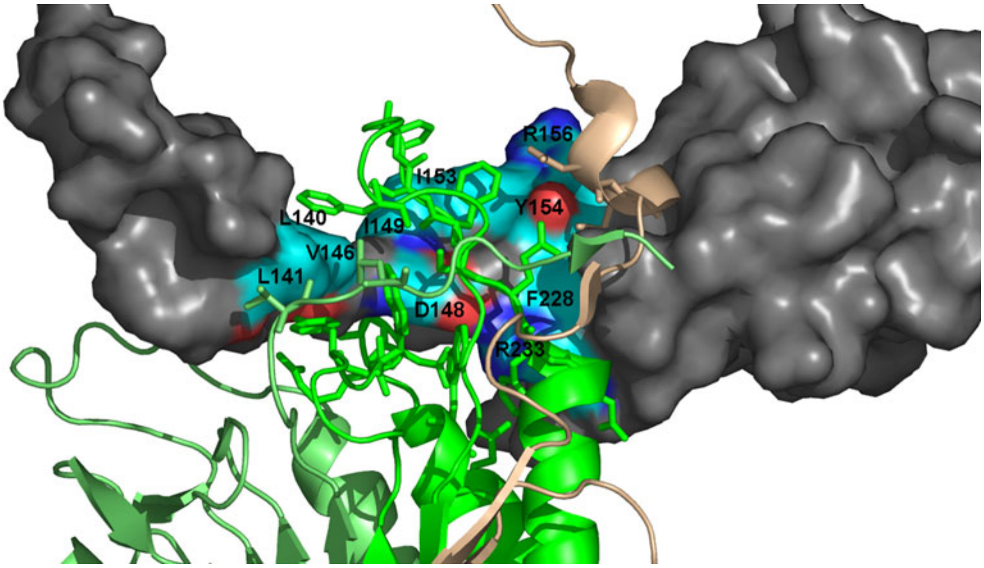


Fig. 1 The interaction between XIAP-BIR2 (shown in gray surface representation) and caspase-3 (shown as cartoon, the main interacting caspase is colored green, a second caspase mainly interacting with a second XIAP is colored beige) as revealed by the complex crystal structure 1I3O [20]. The caspase-3 residues involved in the interaction are shown as sticks and those of XIAP-XIR2 are labeled and colored by element

adrenodoxine protein for which polyamine binding was measured experimentally, but the binding sites were unknown. By docking the polyamines into the predicted transient pockets we were able to suggest favorable binding sites that were validated by site-directed mutagenesis studies [29]. It is a justified concern that incorporating protein flexibility may result in a lack of specificity and increase the number of false positive binders. However, Carlson and co-workers showed that this is not the case with the related multiple protein structures (MPS) approach that uses multiple protein conformations either taken from MD simulations or X-ray structures for building pharmacophore models to identify inhibitors over non-inhibitors [30, 31].

By comparing the binding energies for a dataset of residues from heterodimeric protein-protein interfaces using alanine scanning it turned out that only a fraction of residues named *hot-spots* account for the majority of the binding energy [32, 33]. Hot-spot residues are generally located in the tightly packed interior of interfaces where they are excluded from solvent. Several methods were developed for predicting hot spots from structural information using energy-based calculations [34–39]. Some of these achieve accuracies of up to 70 percent for the prediction of hot spot residues [39].

The goal of this work was to explore the putative interaction of small-molecule XIAP inhibitors with the BIR2 domain. In particular we address the following two questions: (1) Is the pocket detection protocol able to distinguish between binders and non-binders? (2) Where (in the linker region) do the known inhibitors bind? This study is based on the assumption that the experimentally identified polyphenylurea-based compounds bind to the region connecting the BIR1 and BIR2 domain where they would directly affect its ability to bind caspase-3 simultaneously. The other possible alternative where the polyphenylurea-based compounds would act as allosteric inhibitors appeared less likely. Thus, we used our pocket detection protocol for suggesting reasonable binding sites located within this region for the three previously published potent inhibitors, 1396-11, 1396-34, and 1540-14 [21–23], here referred to as **A1**, **A2**, and **A3**, respectively. The concern mentioned before about the possible loss of specificity by incorporating protein flexibility was met by also including in this study the inactive compounds 1540-20 (**I1**) [23] and 1396-28 (**I2**) [21]. The chemical structures of the five compounds are depicted in Fig. 2.

In order to ensure a thorough sampling of the pocket space, four different conformational ensembles were generated: XIAP-BIR2 was simulated in water as well as in methanol and two different starting structures were used namely the unbound NMR structure [25]), or XIAP-BIR2 extracted from the X-ray structure of its complex with caspase-3 [11].

Materials and methods

Parameterization of the Cys_3His-Zinc finger

The force field parameterization of the Cys_3His-Zinc finger unit was based on the energy minimized average NMR structure of the unbound XIAP-BIR2 (PDB ID 1C9Q [25]) as well as on the X-ray structure of XIAP-BIR2 bound to caspase-3 (PDB ID 1I3O [11], chain E). Geometry optimizations were performed using NWChem 4.7 [40] whereby the ligating cysteines were modeled as CH_3S^- and the histidine as imidazole. Thus, the resulting system contained 24 atoms and had a total charge of -1e. The geometries were optimized without constraints by the density functional theory (DFT) module using the B3LYP exchange-correlation functional [41] and the 6-31G* basis set. The number of iterations was set to 500 and the default convergence criteria were used for the optimization. The optimized geometry was then used for calculating the restrained electrostatic potential fit (RESP) charges [42] using the Hartree Fock (HF) method with the same basis set. Both input geometries converged to almost identical minimum energies with an RMSD of 0.8 Å on the heavy atoms and the calculated RESP charges differed by at most 0.018 e. As the optimized geometry based on the X-ray structure was closer to the conformation in either experimental structures than the one based on the NMR structure (0.7 and 0.6 Å instead of 0.8 and 0.9 Å), the former was used for the parameterization of the Cys_3His-Zinc finger in the OPLS-all atom force field [43].

The RESP charges obtained from the HF calculation (listed in Table S1, Supplementary material) were used for Coulombic interactions and the van-der-Waals parameters for the zinc ion were taken from [44]. The interactions between the Cys:S_γ or the His:$N_{\varepsilon 2}$ and the Zn^{2+} were modeled as bonded interactions and the equilibrium values for the respective bond lengths, angles, and dihedrals were taken from the optimized geometry. The force constants were assigned in analogy to similar groups in the OPLS force field.

Molecular dynamics simulations

Molecular dynamics (MD) simulations in water and in methanol were performed starting from the X-ray structure 1I3O and the NMR structure 1C9Q. All MD simulations, energy minimizations, and analysis were performed with the Gromacs 3.3.1 package [45]. The proteins were parameterized using the OPLS-AA force field [43] and placed in cubic boxes filled with TIP4P water [46] or methanol (using parameters from the OPLS-AA force field). The box dimensions were 90 and 99 Å. Periodic boundary conditions were applied. The system was then relaxed by 500 steps of

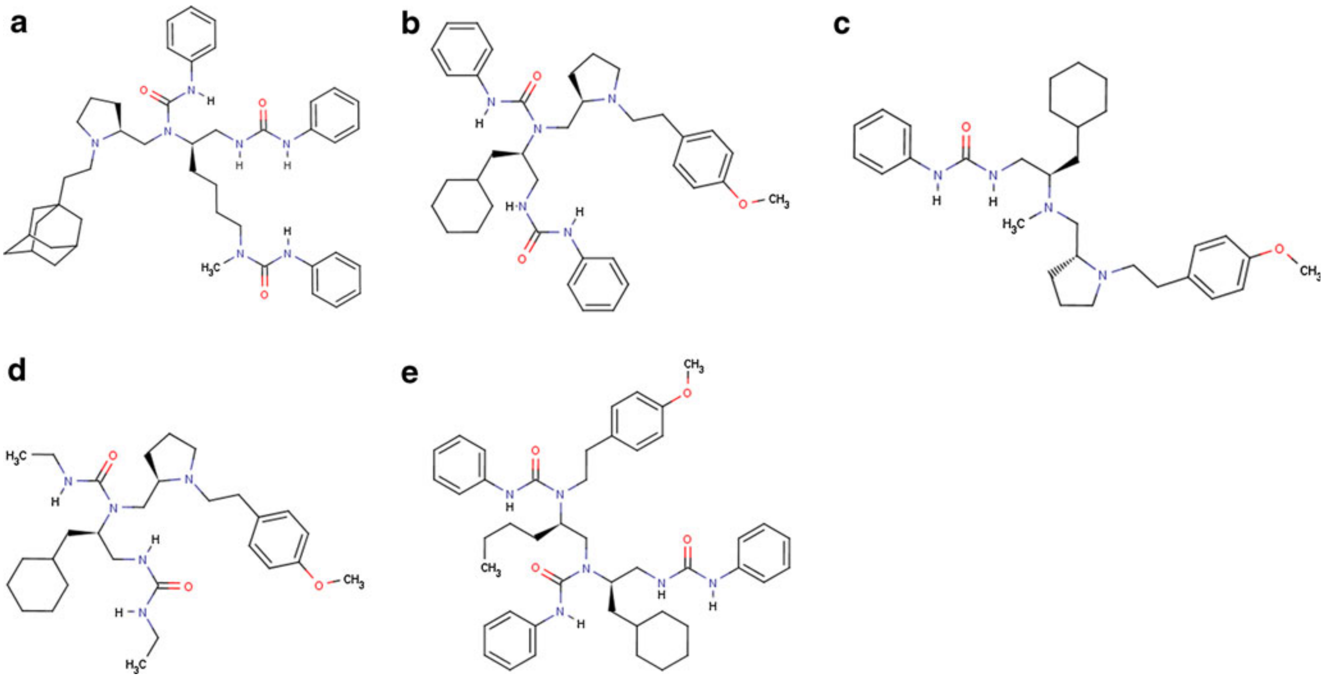


Fig. 2 The three active ((**a**) – (**c**)) and the two inactive ((**d**), (**e**)) compounds used in this study

steepest-descent energy minimization while keeping the positions of the heavy protein atoms harmonically restrained (force constant of 2.39 kcal·mol^{-1}·Å^{-2}). Two sodium counter ions were added to ensure a net neutral charge of the systems and the pre-equilibrating run was repeated. For the simulation in water, the equilibration continued during a 100 ps simulation with harmonical restraints on the heavy atoms in the NPT ensemble at a temperature of 300 K. For the simulation in methanol, the equilibration was extended to 500 ps followed by a 1 ns MD run in which all restraints were removed. Thereafter, simulation snapshots were collected during a subsequent 10 ns simulation and saved every 2.5 ps. Electrostatic interactions beyond the short-range cut-off of 9Å were treated by the particle-mesh-Ewald method [47] and Van der Waals interactions were computed within a 9Å cut-off. Temperature and pressure were kept constant at standard conditions (10^5 Pa, 300 K) by weak coupling to a temperature and pressure bath [48] with coupling constants of 0.1 ps for the temperature coupling and 1 ps for the pressure coupling. Protein, solvent, and counter ions were coupled to separate baths. The LINCS procedure [49] was used to constrain all covalent bonds.

Detection of transient pockets

The MD simulations yielded four conformational ensembles, each consisting of 4001 snapshots. Additionally, we included the experimental structures in the pocket detection step. The ensembles of pockets were identified with our in-house program EPOS^{BP} as described in [27] that is freely available at http://gepard.bioinformatik.uni-saarland.de/software/epos-bp. In that software, pockets are detected with an implementation of the PASS algorithm [50] with the BALL library. In the clustering step we reduced the similarity threshold defining two pockets as states of the same transient pocket from 85% to 75%.

Docking into transient pockets found in MD snapshots

The ligands were set as neutral and Gasteiger atom charges [51] as calculated by molecular operating environment (MOE), version 2007 were assigned. Rotatable bonds were defined using AutoTors. The number of flexible torsions was 16 for **A1**, 13 for **A2**, 12 for **A3** and **I1**, and 18 for **I2**.

For preparing the MD snapshots for the docking with AutoDock3, the nonpolar hydrogen atoms of the MD snapshots were removed and Kollman united-atom partial charges and solvation parameters were assigned using the AutoDockTools modules (ADT 1.4.6) of the Python Molecular Viewer software [52]. The pockets detected in the individual MD snapshots were used to define the putative binding region for which grid maps were calculated with AutoGrid3. The center of the pocket was used as center and the grid dimensions were set to 26.25Å×26.25Å×26.25 Å allowing the ligands to place only a terminal moiety into the transient pocket.

For the docking procedure, the standard Lamarckian genetic algorithm protocol was used with an initial population of 150 randomly placed individuals, a maximum number of 2,500,000 energy evaluations, a mutation rate of 0.02, a crossover rate of 0.80, and an elitism value of 1. The probability of performing a local search on an individual was set to 0.06, and the maximum number of consecutive successes or failures before doubling or halving the local search step size was 4. The Solis and Wets algorithm [53] was applied for these local searches with a maximum of 300 iterations. Twenty independent docking runs were carried out for each MD snapshot.

The best docking poses were then docked again with AutoDock4.2 [54]. For this purpose, the ligand files were converted into AutoDock4 format. As before, the MD snapshots were prepared for the docking with AutoDock4.2 with AutoDockTools. Gasteiger atom charges and AutoDock4 atom types were assigned and nonpolar hydrogens were removed. As this docking step was used for refining the docking poses predicted by AutoDock3, the center of the ligand pose obtained from AutoDock3 was used to define the center of the grid calculated by AutoGrid4.2. The grid dimensions were set to 15Å×15Å×15Å. For the AutoDock4.2 runs the standard Lamarckian genetic algorithm protocol was used with the same parameters and ten independent docking runs were carried out for each docking pose.

Clustering of the docking poses

For each ligand, the docking poses were ranked by their predicted binding free energy and the best 10% of the ranks were selected for clustering. Note that by picking the best 10% of the ranks, the number of selected docking poses differs among the ligands as those poses having the same predicted binding free energy also have the same rank.

The similarities of the docking poses for the five different ligands were calculated by comparing the heavy protein atoms (pose lining atoms, PLA) found within 5Å of the docked ligand. These similarities were then used as input for a single linkage clustering where clusters of docking poses A and B were merged, if

$$\min\{|PLA(a)\cap PLA(b)| : a\in A, b\in B\}\geq 30. \qquad (1)$$

Results

Stability of XIAP-BIR2 during the MD simulations in water and methanol

The molecular dynamics simulations of XIAP-BIR2 revealed that the Zinc finger motif remained very close to its optimized geometry and, thus, did not distort the overall protein structure. The core of the protein underwent only minor conformational changes during the simulations in water (Fig. S1, Supplementary material) as reflected by the relatively small RMS deviation of the protein backbone from the starting structure of about 1.5Å. Slightly larger structural transitions with a RMSD of 2.0 – 2.5Å were observed when methanol was used as solvent. Still, these values compare well to typical RMSD values observed in other MD simulations on stably folded proteins. We note that although the XIAP-BIR2 protein may not be stably folded when studied experimentally in methanol, it is extremely unlikely to observe such unfolding events during relatively short MD simulations at room temperature. As expected, the N-terminal linker region as well as the C-terminus was quite mobile (Fig. S2, Supplementary material). In both simulation setups, the protein conformation taken from the complex X-ray structure showed slightly smaller RMS deviations from the starting conformation than the one taken from the NMR structure.

Overall, the secondary structure remained stable throughout the simulation. Solely the third β-sheet between residues 205 and 207 partly unfolded in three of the four different MD simulations (Fig. S3, Supplementary material).

Detection of transient pockets

Our protocol identified 32 transient pockets (after removal of pockets detected in only one MD snapshot) that are spread over the whole protein surface. However, as structural studies suggested that caspase-3 mainly interacts with the linker region of XIAP-BIR2 (residues 124-168) [24], we focused our analysis on the pockets involving these residues. Indeed, a surprisingly high number of 45% of all pockets involved the residues of the linker region. They are listed in Table 1 (for an overview of all transient pockets see Table S2, Supplementary material). These pockets are all overlapping but were not assigned to the same cluster because their lining protein atoms vary more than the given threshold depending on the MD simulation setup. Note that even transient pockets assigned to the same cluster substantially differ in their frequencies and volume. This highlights the influence of the starting structure and the used solvent on the pocket openings as discussed in a previous study [27] and further suggests that these pockets are highly mobile and adaptable. From our previous experience on the BCL-X$_L$, IL-2, and MDM2 systems, individual snapshots with pocket volumes larger than 200 Å^3 appear promising candidates for docking studies. Thus, all pockets located in the linker region (PIDs 0, 13-16, 18, 20, 23, 25-27, 29, 30, 32) having a pocket volume ≥200Å^3 were selected as putative binding sites for the five ligands. This resulted in the selection of 6,662 pockets from the four different MD simulations (1,624 in the simulation that was

Table 1 The mean properties and the frequencies of the transient pockets opening in the linker region that are detected in more than one snapshot of the four MD simulations

Binding Site (PID)	Residues	NMR in methanol		NMR in water		X-ray in methanol		X-ray in water	
		Freq. [%]	Vol. [Å^3]	Freq. [%]	Vol. [Å^3]	Freq. [%]	Vol. [Å^3]	Freq. [%]	Vol. [Å^3]
0	137, 138, 140, 141	0.87	199.2	35.5	495.0	1.4	264.3	0.3	172.2
13	151-153	1.8	182.4	18.0	367.0			18.0	221.6
14	132-136			2.3	201.7	30.2	194.5	2.8	140.9
15	127, 131, 134, 135	18.8	234.6	51.4	497.9			0.1	134.8
16	148, 151, 228, 233-235			21.3	453.8	65.7	353.1	21.2	183.8
18	153-155, 157, 158, 161	6.3	282.8			7.8	338.3	2.2	211.6
20	141-147	6.4	182.8	20					
23	128, 136, 139, 140, 143, 145							5.4	150.4
25	129, 131, 133-135, 138, 141, 142	52.3	237.0						
26	145-149, 151-157	27.3	218.2						
27	161-163, 166, 186, 201, 229, 233-235			90.0	279.7				
29	124-127, 131, 134, 135, 137, 141, 145,			4.8	300.6				
30	141, 146, 148, 151, 233					23.4	290.2		
32	161-163, 166, 201, 229, 232-235							78.4	281.2

started from the NMR structure in water, 137 from the NMR structure in methanol, 418 from the X-ray structure in water, and 4,483 from the X-ray structure in methanol).

Can the protocol for detecting transients pockets distinguish between active and inactive compounds?

So far, we have only applied this protocol for predicting the binding modes of known binders. However, in virtual screening experiments the goal is to identify putatively active compounds among a large number of putatively inactive compounds. Therefore, a reliable discrimination between binders and non-binders is crucial. In order to explore whether this protocol predicts binding affinities that are reliable enough for such an application, three active as well as two inactive compounds were docked into the transient pockets. Note that in this analysis it is difficult to separate the quality of this protocol from that of the docking program itself which calculates and scores the binding poses. In order to estimate the impact of the pocket detection protocol on the discrimination between binders and non-binders, the docking scores and predicted binding free energies obtained from docking into the X-ray or NMR structure (Table 2) were compared to those obtained from docking into the transient pockets (Table 3). Table 2 clearly indicates that when using the X-ray or NMR structure as receptor for docking, the docking score and the predicted binding free energy do not allow inferring the activity of the compounds. Although the most active compound **A1** obtained the best docking score, the other two active compounds were scored similar or worse than the inactive ones. When focusing on the predicted binding free energy alone, it is at least possible to classify one of the two inactive compounds correctly.

Table 3 reveals that combining the pocket detection protocol with the predicted binding free energy significantly improves the discrimination between active and inactive compounds. Here, the difference in the predicted binding free energy between the “best” inactive compound with the “worst” active compound is 1.05 kcal mol^{-1} and thus larger than the difference observed among the three active compounds (0.9 kcal mol^{-1}). Moreover, this table also reveals that when trying to discriminate between active and inactive compounds, the predicted binding free energy outperforms the docking score. Taken together, these results suggest that

Table 2 Comparison between the best docking scores and $\Delta G_{binding}$ per ligand obtained from docking into the X-ray and NMR structure with AutoDock3

	X-ray structure		NMR structure	
	Best docking score [kcal/mol]	Best $\Delta G_{binding}$ [kcal/mol]	Best docking score [kcal/mol]	Best $\Delta G_{binding}$ [kcal/mol]
A1	−14.46	−7.05	−14.77	−6.94
A2	−11.21	−5.05	−12.52	−6.26
A3	−10.93	−6.03	−11.74	−7.08
I1	−12.57	−6.52	−13.14	−7.02
I2	−11.47	−2.46	−14.18	−4.75

Table 3 Comparison between the mean and best docking scores and $\Delta G_{\mathrm{binding}}$ per ligand obtained when docking into the MD snapshots with AutoDock3

	Mean docking score [kcal/mol]	Best docking score [kcal/mol]	Mean $\Delta G_{\mathrm{binding}}$ [kcal/mol]	Best $\Delta G_{\mathrm{binding}}$ [kcal/mol]
A1	−12.83±1.58	−19.59	−4.92±1.83	−12.65
A2	−11.69±1.28	−17.85	−5.64±1.45	−11.85
A3	−10.60±1.19	−16.31	−6.33±1.35	−11.75
I1	−11.06±1.12	−16.29	−5.09±1.25	−10.70
I2	−11.58±1.51	−18.71	−2.64±1.67	−10.16

more reliable binding affinities can be predicted when dynamic properties of the binding site (like the existence of transient pockets into which moieties of the ligands can bind) are combined with predicted binding free energies.

Favorable binding sites

Table 3 indicates that the best predicted binding free energy is the most reliable parameter for distinguishing between active and inactive compounds. Therefore, this measure was used as the basis to suggest favorable binding sites. When selecting the best 10% of docking results with respect to the predicted binding free energy, 302 poses were kept for compound **A1**, 187 for **A2**, 191 for **A3**, 178 for **I1**, and 224 for **I2**. Interestingly, all these docking results started from transient pockets either detected in the MD snapshots of the simulation of the NMR structure in water or from the simulation of the X-ray structure in methanol. Clustering of these poses and keeping only those clusters with at least 15 members resulted in the 11 favorable binding sites listed in Table 4. At each binding site, the inactive compounds have a worse predicted binding free energy than the three actives. Note that many binding sites involve almost the same residues (for example binding sites 1, 4, 5, 7, 8, and 11) but not the same atoms. This subdivision of very similar binding sites emphasizes that the docking pose clustering is very sensitive toward different conformational substates of the protein.

All docking poses listed in Table 4 were subsequently refined by re-docking them with the AutoDock4.2 program. This new version of AutoDock incorporates a new charge-based desolvation method [54]. As all binding pockets considered in this study are relatively flat and solvent exposed, the resulting docking poses and binding free energies from AutoDock4.2 are expected to be more accurate than those from AutoDock3. In this setup, the ligands were restrained to the same binding site by using smaller grid dimensions. As before, the best 10% of docking poses with respect to the predicted binding free energy were chosen and clustered. The results are compiled in Table 5. The clustering yielded five overlapping binding sites. Although the distinction between active and inactive compounds in each individual binding site is not as clear as with AutoDock3, the best predicted binding free energies per compound allow a reliable identification of the three binders with an energy difference of 1.05 kcal mol^{-1} between the "worst" active and the "best" inactive. Moreover, the refined binding energies per binding site listed in Table 5 suggest that compounds **A1** and **A3** prefer the same binding site. Considering that **A3** is structurally more similar to **A2** than to **A1**, this observation may be surprising. However, one should keep in mind that the clustering of the binding sites is based on the similarity of the heavy protein atoms found within a certain distance of

Table 4 The most favorable binding sites for the five compounds as predicted by docking them with AutoDock3 into the transient pockets, selecting the best 10% of docking poses (w.r.t. binding free energy) and clustering them based on the protein atoms located within 5Å from any heavy ligand atom. Shown are the best binding free energies per binding site and compound in (kcal mol^{-1}). The most favorable binding free energies per ligand are highlighted

Binding site /no. of poses	Residues	**A1**	**A2**	**A3**	**I1**	**I2**
1 (18)	148, 152, 228, 231, 234, 235, 236	**−12.65**	−10.37	−10.41	−8.93	−7.84
2 (178)	125, 126, 129, 131, 140, 141, 145, 146, 236	−11.52	−11.46	**−11.75**	−10.34	**−10.16**
3 (234)	147, 148, 225-227, 232-234	−12.05	−11.00	−11.68	**−10.70**	−9.66
4 (261)	147, 148, 151-153, 228, 231, 234-237	−12.02	**−11.85**	−11.22	−10.01	−9.12
5 (56)	148, 151, 228, 231, 234-237	−11.48	−10.79	−11.56	−9.35	−8.71
6 (51)	148, 226, 232-234	−11.32	−10.63	−10.97	−9.86	−8.08
7 (30)	148, 151, 228, 231, 234-236	−10.85	−10.20	−10.56	−9.18	−9.75
8 (22)	148, 151, 152, 228, 234-236	−10.90	−11.04	−11.14	−8.69	−7.83
9 (19)	226, 232-234	−10.92	−10.64	−10.81	−9.46	−8.26
10 (68)	126, 129-131, 140, 141, 146, 236, 237	−11.32	−11.20	−11.21	−9.83	−9.03
11 (15)	148, 151, 152, 228, 231, 234-236	−10.61	−11.69	−10.33	−9.74	−7.42

Table 5 The most favorable binding sites for the five compounds as predicted by re-docking them with AutoDock4.2 to the positions of the docking poses listed in Table 4. Shown are the best binding free energies per binding site and compound in (kcal mol^{-1}). The most favorable binding free energies per ligand are highlighted

Binding site/ no. of poses	Residues	A1	A2	A3	I1	I2
1 (95)	125, 126, 129-131, 140, 141, 145, 146, 236, 237	**−10.54**	−10.29	**−10.48**	−8.51	−9.20
2 (92)	147, 148, 151-153, 228, 231, 234-237	−10.22	**−10.56**	−9.16	−8.79	**−9.43**
3 (120)	151, 228, 231, 234-237	−9.55	−10.40	−10.03	**−8.84**	−8.79
4 (43)	147, 148, 225-227, 232-234	−10.37	−10.02	−9.49	−8.06	−8.17
5 (15)	125, 126, 129-131, 140, 141, 145, 146, 236	−8.95	−9.73	−9.21	−8.64	−8.8

the predicted binding pose and, thus, does not consider individual protein conformations or ligand orientations explicitly. The suggested binding modes of all five compounds in binding sites 1 to 3 are shown in Fig. 3.

For comparison, we also used the hotspot prediction server HOTPOINT [39] on the complex structure 1I3O using default parameters. Within a few seconds HOTPOINT predicted residues M176, H237, and L141 at the interface between chains A and E as hotspot residues and residues M176 and H237 at the interface between chains C and F. Of these, residue L141 belongs to the XIAP protein and the two other residues belong to Caspase3. Residue L141 belongs to the best binding pockets 1 and 5 in Table 5, but not to the other binding pockets. This emphasizes the nature of the expensive molecular dynamics simulations that yield a much more detailed view of the shape modularity of this highly flexible system. We suggest that hotspot predictions may be more useful for characterizing essential binding residues at preformed binding pockets.

Discussion

In this work, the ability of the transient pocket detection protocol to distinguish between active and inactive compounds was explored. Comparing the AutoDock3 docking scores and predicted binding free energies for the known binders and nonbinders identified the predicted binding free energy as a more reliable indicator. The docking score is calculated as the sum of the intermolecular interaction energy and the intramolecular ligand energy. The predicted binding free energy, on the other hand, also takes the approximated entropy loss into account. It is calculated as the sum of the intermolecular interaction energy and the weighted number of flexible ligand torsions. Although this disregards the impact of highly strained ligand conformations, it emphasizes the influence of the entropic contribution. This may be particularly relevant for the set of floppy, extended ligands studied here. Admittedly, the number of known binding and non-binding molecules is very small, so that the energy difference separating the two classes is not statistically significant yet. Another limitation of this study is that MD simulations of 10 ns length clearly cannot sample the entire conformational ensemble accessible to this extremely floppy linker region. However, the results suggest that MD simulations of this length are well able to generate a large number of suitable binding conformations that may be representative for many of the alternative binding modes of the five ligands if they were to bind to the XIAP-BIR2 domain in the assumed region.

A further interesting, though not unexpected, observation is that the best predicted binding free energies outperform the mean binding free energies per ligand listed in Table 3. If the dominance of a certain protein conformation was a prerequisite for ligand binding, one can expect the prediction power of the mean predicted binding free energies to be much better. Our findings for the XIAP system are thus in accordance with the conformational selection model of ligand binding [55]. This model implies that a protein in the unbound state exists in a myriad of different conformations. Although the unbound state is predominant, a small percentage of conformations also exist resembling the bound state. A ligand can then selectively bind to such a conformation and even though this conformation will be of higher energy when considering the protein alone, the binding event shifts the equilibrium toward the bound state in which the ligand bound conformation then becomes predominant. In the context of our pocket detection protocol, this underpins that pockets opening less frequently are not necessarily non-binding pockets. Or analogously, when docking multiple ligands into multiple protein conformations, the ligand with the best mean score is not necessarily the most active one.

However, one should always keep in mind that the predicted binding free energies or docking scores are very error-prone. For example, the authors of a publication describing the testing of AutoDock4 [56] stated that the method has a standard error of about 2-3 kcal mol^{-1} in predicting binding free energies. They further mentioned

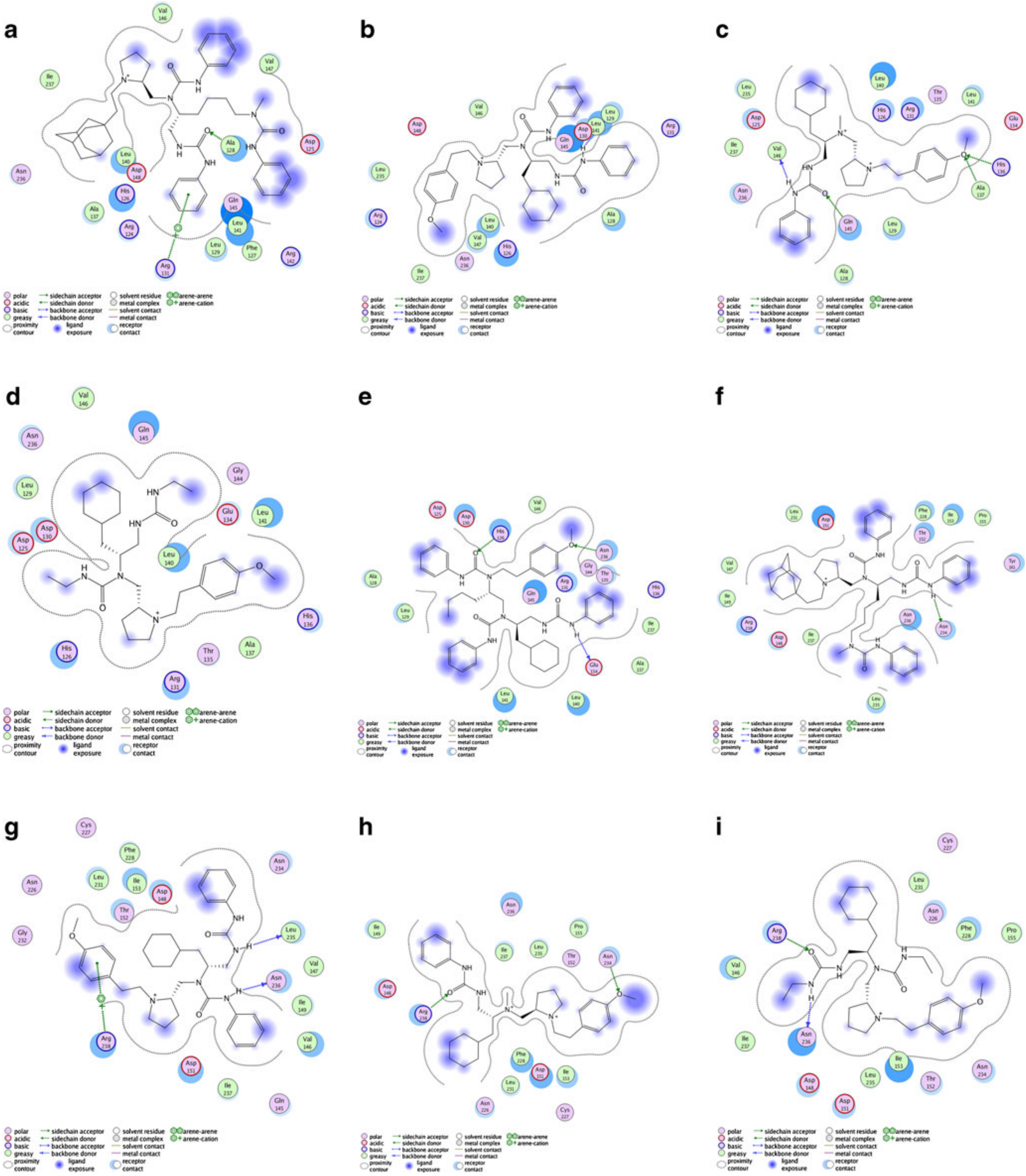


Fig. 3 The best scored binding modes of the five ligands in binding sites 1 to 3 as listed in Table 5: (**a**) ligand **A1** in binding site 1, (**b**) ligand **A2** in binding site 1, (**c**) ligand **A3** in binding site 1, (**d**) ligand **I1** in binding site 1, (**e**) ligand **I2** in binding site 1, (**f**) ligand **A1** in binding site 2, (**g**) ligand **A2** in binding site 2, (**h**) ligand **A3** in binding site 2, (**i**) ligand **I1** in binding site 2, (**j**) ligand **I2** in binding site 2, (**k**) ligand **A1** in binding site 3, (**l**) ligand **A2** in binding site 3, (**m**) ligand **A3** in binding site 3, (**n**) ligand **I1** in binding site 3, (**o**) ligand **I2** in binding site 3

that AutoDock4 successfully re-docked complexes with ten or fewer flexible torsions while re-docking failed for most ligands of higher flexibility and re-docking the same complexes with AutoDock3 performed even a bit worse.

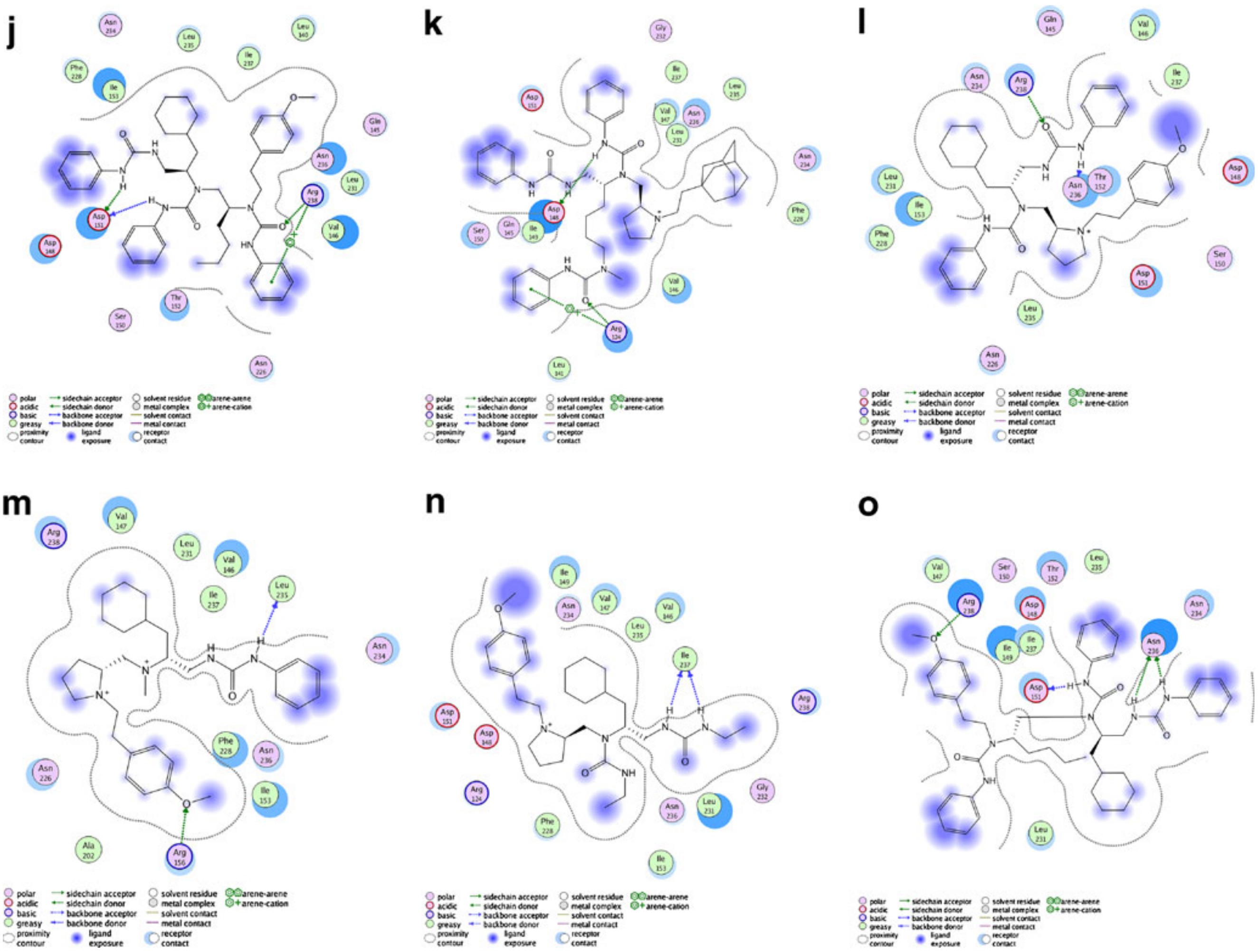


Fig. 3 (continued)

Considering that the number of flexible torsions of the ligands docked in this study was between 13 and 18, the binding modes presented in this study have to be regarded as what they are: *suggested* binding poses. This is also the reason why we are rather focusing on binding sites than on binding modes. From this point of view, it is encouraging that refining the clustered AutoDock3 docking poses by re-docking them with AutoDock4 and clustering the best results reduced the number of favorable binding sites from 11 to 5. The predicted binding free energies yet suggest that binding sites 1 to 3 (Table 5) are the most likely binding sites for the five ligands tested in this study. One could argue that binding site 1 is the most reasonable binding site because the predicted binding free energies allow for a clear discrimination between the three binders and the two nonbinders. But on the other hand, Table 5 suggests that the preferred binding site may differ for the investigated ligands and the nonbinders can only be winnowed from the binders by focusing on their largest predicted binding free energies regardless of the binding site. However, to arrive at a reliable conclusion, one would have to investigate a larger number of known active and inactive compounds using the protocol.

To our knowledge, this is the first publication reporting a MD simulation of the BIR2 domain of XIAP. Obiol-Pardo *et al.* also used MD simulations for analyzing the protein-protein interactions appearing in the Smac/Diablo – XIAP complex but they only studied the BIR3 domain [57]. They used the cationic dummy approach [58] for maintaining the tetrahedral coordination of the zinc ion during the simulation while in our simulation, the tetrahedral coordination was preserved through bonded interactions with equilibrium values for angles, dihedrals, and bond lengths taken from a optimized geometry. Their reported average simulation distances of the zinc ion to the coordinating Cys:S$_{\gamma}$ and the His:N$_{\varepsilon 2}$ atoms are in the same order of magnitude as in our simulations (data not shown).

Conclusions

In this work, the putative binding sites of previously reported small-molecule XIAP-BIR2 inhibitors were explored using the transient pocket detection protocol. A surprisingly large number of small pockets opened in the linker region during molecular dynamics simulation in water and in methanol. This does not only highlight the flexibility of the domain but also the sensitivity of the pocket detection protocol toward different conformational substates of the protein. The detected cavities were rather small in volume suggesting that the binding site is composed of multiple subpockets that accommodate different parts of the large ligands. When disregarding these transient (sub)pockets and docking into the X-ray or NMR structures of the BIR2 domain, it was impossible to distinguish between binders and nonbinders. In contrast, the three binders could successfully be identified when docking into transient pockets detected by the protocol when the best calculated binding free energy was considered. Furthermore, clustering the most favorable binding modes resulted in five putative binding sites. To the best of our knowledge, this is the first MD simulation study of the BIR2 domain of XIAP. Although this study was conducted with a small set of active and inactive XIAP inhibitors, this work represents a first step toward understanding at the molecular level the mode of action of protein-protein interaction inhibitors targeting the XIAP-BIR2 domain.

Acknowledgments We thank Jan Fuhrmann and Dirk Neumann (Center for Bioinformatics, Saarland University) for making available their BALLPass implementation to us.

References

1. Wells AL, McClendon CL (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature 450:1001–1009
2. Roy N, Deveraux QL, Takahashi R, Salvesen GS, Reed JC (1997) The c-IAP-1 and c-IAP-2 proteins are direct inhibitors of specific caspases. EMBO J 16:6914–6925
3. Deveraux QL, Roy N, Stennicke HR, Van Arsdale T, Zhou Q, Srinivasula SM, Alnemri ES, Salvesen GS, Reed JC (1998) IAPs block apoptotic events induced by caspase-8 and cytochrome c by direct inhibition of distinct caspases. EMBO J 17:2215–2223
4. Reed JC (2001) The survivin saga goes in vivo. J Clin Invest 108:965–969
5. Cohen GM (1997) Caspases: the executioners of apoptosis. Biochem J 326:1–16
6. Thornberry NA, Lazebnik Y (1998) Caspases: enemies within. Science 281:1312–1316
7. Tamm I, Kornblau SM, Segall H, Krajewski S, Welsh K, Kitada S, Scudiero DA, Tudor G, Qui YH, Monks A, Andreeff M, Reed JC (2000) Expression and prognostic significance of IAP-family genes in human cancers and myeloid leukemias. Clin Cancer Res 6:1796–1803
8. Takahashi R, Deveraux Q, Tamm I, Welsh K, Assa-Munt N, Salvesen GS, Reed JC (1998) A single BIR domain of XIAP sufficient for inhibiting caspases. J Biol Chem 14:7787–7790
9. Deveraux QL, Leo E, Stennicke HR, Welsh K, Salvesen GS, Reed JC (1999) Cleavage of human inhibitor of apoptosis protein XIAP results in fragments with distinct specificities for caspases. EMBO J 18:5242–5251
10. Du C, Fang M, Li Y, Li L, Wang X (2000) Smac, a mitochondrial protein that promotes cytochrome c-dependent caspase activation by eliminating IAP inhibition. Cell 102:33–42
11. Riedl SJ, Renatus M, Schwarzenbacher R, Zhou Q, Sun C, Fesik SW, Liddington RC, Salvesen GS (2001) Structural basis for the inhibition of caspase-3 by XIAP. Cell 104:791–800
12. Chai J, Shiozaki E, Srinivasula SM, Wu Q, Datta P, Alnemri ES, Shi Y (2001) Structural basis of caspase-7 inhibition by XIAP. Cell 104:769–780
13. Shiozaki EN, Chai J, Rigotti DJ, Riedl SJ, Li P, Srinivasula SM, Alnemri ES, Fairman R, Shi Y (2003) Mechanism of XIAP-mediated inhibition of caspase-9. Mol Cell 11:519–527
14. Liu Z, Sun C, Olejniczak ET, Meadows RP, Betz SF, Oost T, Herrmann J, Wu JC, Fesik SW (2000) Structural basis for binding of Smac/Diablo to the XIAP BIR3 domain. Nature 408:1004–1008
15. Wu G, Chai J, Suber TL, Wu JW, Du C, Wang X, Shi Y (2000) Structural basis of IAP recognition by Smac/Diablo. Nature 408:1008–1012
16. Wist AD, Gu L, Riedl SJ, Shi Y, McLendon GL (2007) Structure-activity based study of the Smac-binding pocket within the BIR3 domain of XIAP. Bioorg Med Chem 15:2935–2943
17. Sun H, Stuckey JA, Nikolovska-Coleska Z, Qin D, Meagher JL, Qiu S, Lu J, Yang CY, Saito NG, Wang S (2008) Structure-based design, synthesis, evaluation, and crystallographic studies of conformationally constrained Smac mimetics as inhibitors of the X-linked inhibitor of apoptosis protein (XIAP). J Med Chem 51:7169–7180
18. Mastrangelo E, Cossu F, Milani M, Sorrentino G, Lecis D, Delia D, Manzoni L, Drago C, Seneci P, Scolastico C, Rizzo V, Bolognesi M (2008) Targeting the X-linked inhibitor of apoptosis protein through 4-substituted azabicyclo[5.3.0]alkane Smac mimetics. Structure, activity, and recognition principles. J Mol Biol 384:673–689
19. Nikolovska-Coleska Z, Meagher JL, Jiang S, Yang CY, Qiu S, Roller PP, Stuckey JA, Wang S (2008) Interaction of a cyclic, bivalent Smac mimetic with the x-linked inhibitor of apoptosis protein. Biochemistry 47:9811–9824
20. Cossu F, Milani M, Mastrangelo E, Vachette P, Servida F, Lecis D, Canevari G, Delia D, Drago C, Rizzo V, Manzoni L, Seneci P, Scolastico C, Bolognesi M (2009) Structural basis for bivalent Smac-mimetics recognition in the IAP protein family. J Mol Biol 392:630–644
21. Schimmer AD, Welsh K, Pinilla C, Wang Z, Krajewska M, Bonneau MJ, Pedersen IM, Kitada S, Scott FL, Bailly-Maitre B, Glinsky G, Scudiero D, Sausville E, Salvesen G, Nefzi A, Ostresh JM, Houghten RA, Reed JC (2004) Small-molecule antagonists of apoptosis suppressor XIAP exhibit broad antitumor activity. Cancer Cell 5:25–35
22. Wang Z, Cuddy M, Samuel T, Welsh K, Schimmer A, Hanaii F, Houghten R, Pinilla C, Reed JC (2004) Cellular, biochemical, and genetic analysis of mechanism of small molecule IAP inhibitors. J Biol Chem 279:48168–48176
23. Kater AP, Dicker F, Mangiola M, Welsh K, Houghten R, Ostresh J, Nefzi A, Reed JC, Pinilla C, Kipps TJ (2005) Inhibitors of XIAP sensitize CD40-activated chronic lymphocytic leukemia cells to CD95-mediated apoptosis. Blood 106:1742–1748

24. Scott FL, Denault JB, Riedl SJ, Shin H, Renatus M, Salvesen GS (2005) XIAP inhibits caspase-3 and -7 using two binding sites: evolutionarily conserved mechanism of IAPs. EMBO J 24:645–655
25. Sun C, Cai M, Gunasekera AH, Meadows RP, Wang H, Chen J, Zhang H, Wu W, Xu N, Ng SC, Fesik SW (1999) NMR structure and mutagenesis of the inhibitor-of-apoptosis protein XIAP. Nature 401:818–822
26. Eyrisch S, Helms V (2007) Transient pockets on protein surfaces involved in protein-protein interaction. J Med Chem 50:3457–3464
27. Eyrisch S, Helms V (2009) What induces pocket openings on protein surface patches involved in protein-protein interactions? J Comput Aid Mol Des 23:73–86
28. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19:1639–1662
29. Berwanger A, Eyrisch S, Schuster I, Helms V, Bernhardt R (2010) Polyamines: naturally occurring small molecule modulators of electrostatic protein-protein interactions. J Inorg Biochem 104:118–125
30. Lerner MG, Bowman AL, Carlson HA (2007) Incorporating dynamics in E. coli dihydrofolate reductase enhances structure-based drug discovery. J Chem Inf Model 47:2358–2365
31. Bowman AL, Lerner MG, Carlson HA (2007) Protein flexibility and species specificity in structure-based drug discovery: dihydrofolate reductase as a test system. J Am Chem Soc 129:3634–3640
32. Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. J Mol Biol 280:1
33. Clackson T, Wells J (1995) A hot-spot of binding energy in a hormone-receptor interface. Science 267:383–386
34. Moreira IS, Fernandes PA, Ramos MJ (2007) Hot spots-a review of the protein-protein interface determinant amino-acid residues. Proteins 68:803–812
35. Halperin I, Wolfson H, Nussinov R (2004) Protein-protein interactions: coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. Structure 12:1027–1038
36. Keskin O, Ma B, Nussinov R (2005) Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol 345:1281–1294
37. Kortemme T, Baker D (2002) A simple physical model for binding energy hot spots in protein-protein complexes. Proc Natl Acad Sci USA 99:14116–14121
38. Guerois R, Nielsen J, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320:369–387
39. Tuncbag N, Gursoy A, Keskin O (2009) Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. Bioinformatics 25:1513–1520
40. Bylaska EJ, de Jong WA, Kowalski K, Straatsma TP, Valiev M et al (2006) NWChem, a computational chemistry package for parallel computers, version 5.0, Pacific Northwest National Laboratory, Richland, Washington 99352-0999, USA
41. Becke AD (1997) Density-functional thermochemistry. V. Systematic optimization of exchange-correlation functionals. J Chem Phys 107:8554
42. Bayly CI, Cieplak P, Cornell W, Kollman PA (1993) A well-behaved electrostatic potential based method using charge restraints for determining atom-centered charges: the RESP model. J Phys Chem 97:10269–10280
43. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118:11225–11236
44. Merz KM (1991) Carbon dioxide binding to human carbonic anhydrase ii. J Am Chem Soc 113:406–411
45. Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 7:306–317
46. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
47. Darden T, York D, Pedersen L (1993) Particle mesh ewald: an n log(n) method for ewald sums in large systems. J Chem Phys 98:10089–10092
48. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:3684–3690
49. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: a linear constraint solver for molecular simulations. J Comput Chem 18:1463–1472
50. Brady GP Jr, Stouten PF (2000) Fast prediction and visualization of protein binding pockets with pass. J Comput Aided Mol Des 14:383–401
51. Gasteiger J, Marsili M (1980) Iterative partial equilibration of orbital electronegativity – a rapid access to atomic charges. Tetrahedron 36:3219–3228
52. Sanner MF (1999) Python: a programming language for software integration and development. J Mol Graph Model 17:57–61
53. Solis FJ, Wets JB (1981) Minimization by random search techniques. Math Oper Res 6:19–30
54. Huey R, Morris GM, Olson AJ, Goodsell DS (2007) A semiempirical free energy force field with charge-based desolvation. J Comput Chem 28:1145–1152
55. Boehr DD, Wright PE (2008) How do proteins interact? Science 320:1429–1430
56. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30:2785–2791
57. Obiol-Pardo C, Granadino-Roldan JM, Rubio-Martinez J (2008) Protein-protein recognition as a first step towards the inhibition of XIAP and Survivin anti-apoptotic proteins. J Mol Recognit 21:190–204
58. Pang YP, Xu K, El Yazal J, Prendergast PG (2000) Successful molecular dynamics simulation of the zinc-bound farnesyltransferase using the cationic dummy atom approach. Protein Sci 9:1857–1865

ORIGINAL PAPER

Si-doped graphene: an ideal sensor for NO- or NO_2-detection and metal-free catalyst for N_2O-reduction

Ying Chen · Bo Gao · Jing-Xiang Zhao · Qing-Hai Cai ·
Hong-Gang Fu

Received: 14 June 2011 / Accepted: 15 August 2011 / Published online: 1 September 2011

Abstract Exploring and evaluating the potential applications of two-dimensional graphene is an increasingly hot topic in graphene research. In this paper, by studying the adsorption of NO, N_2O, and NO_2 on pristine and silicon (Si)-doped graphene with density functional theory methods, we evaluated the possibility of using Si-doped graphene as a candidate to detect or reduce harmful nitrogen oxides. The results indicate that, while adsorption of the three molecules on pristine graphene is very weak, Si-doping enhances the interaction of these molecules with graphene sheet in various ways: (1) two NO molecules can be adsorbed on Si-doped graphene in a paired arrangement, while up to four NO_2 molecules attach to the doped graphene with an average adsorption energy of −0.329 eV; (2) the N_2O molecule can be reduced easily to the N_2 molecule, leaving an O-atom on the Si-doped graphene. Moreover, we find that adsorption of NO and NO_2 leads to large changes in the electronic properties of Si-doped graphene. On the basis of these results, Si-doped graphene can be expected to be a good sensor for NO and NO_2 detection, as well as a metal-free catalyst for N_2O reduction.

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1226-x) contains supplementary material, which is available to authorized users.

Y. Chen · B. Gao · J.-X. Zhao (✉) · Q.-H. Cai
Key Laboratory for Design and Synthesis of Functionalized materials and Green Catalysis, School of Chemistry and Chemical Engineering, Harbin Normal University,
Harbin 150025, People's Republic of China
e-mail: xjz_hmily@yahoo.com.cn

H.-G. Fu
Key Laboratory of Functional Inorganic Material Chemistry, Ministry of Education of the People's Republic of China, Heilongjiang University,
150080, Harbin, People's Republic of China

Keywords Graphene · Nitride oxides · Density functional theory

Introduction

Since its discovery in 2004 [1], graphene—a rapidly rising star on the horizon of materials science and technology—has attracted tremendous attention and holds great promise in various fields [2–14]. Especially in gas detecting or sensing [8, 9], the emergence of graphene has opened new avenues for utilizing two-dimensional planer carbon materials as solid-state sensors due to excellent properties such fewer crystal defects [5, 15, 16], and their semimetallic nature (low Johnson noise) [5, 15–17]. Graphene can act as a single atomic layer, which can maximize its interaction with adsorbate. Recently, mechanically exfoliated graphene sheets and reduced graphene oxide (GO) have been shown to exhibit high sensitivity towards some gas molecules, such as NO_2, NH_3, H_2O, and CO [8, 18]. The gas sensing mechanism is based on changes in the electrical conductivity of graphene due to charge transfer between graphene and the adsorbate. Other studies, however, have shown that the above molecules can only be "physisorbed" on pristine graphene [19–21]. Many experimental and theoretical studies have focused on improving the sensing performance of graphene to various desired molecules by functionalizing or doping graphenes [22–26]. For example, Ural [22] and Ramaprabhu [23] reported independently that graphenes functionalized with Pd and Pt nanoparticles become effective H_2 sensors, whereas Dai et al. [24]. demonstrated that B- and S-doped graphene could be a good sensor for NO and NO_2. Additionally, Zhang et al. [25] suggested that the sensitivity and selectivity of graphene-based sensors could be improved greatly by introducing dopants or defects into graphenes.

It is well known that nitrogen oxides, such as NO, N_2O, and NO_2, are major components of air pollution, contributing to acid rain formation, photochemical smog, and depletion of the ozone layer [27–30]. The greatest anthropogenic source of these pollutants is the combustion of solid and liquid fuel sources, encompassing both static and portable applications. Hence, monitoring of, or control of exposure to, nitrogen oxides is of special interest in both industrial and residential settings. We note that NO or NO_2 adsorption on doped or defective graphene has been investigated recently. However, prior reports considered only an individual NO or NO_2 molecule adsorbed on graphene [23, 24]. Several questions remain to be addressed: (1) can N_2O (a highly potent greenhouse gas) be adsorbed on pristine or doped graphene? (2) What would happen if more than one NO or NO_2 molecule is attached to a graphene sheet? (3) What are the effects of adsorption of nitrogen oxides on the properties of graphene? To address these questions, this study took Si-doped graphene—formed by its growth on the (0001) surface of a 6H-SiC substrate [31]—as an example to study the adsorption of NO, N_2O, and NO_2 on Si-doped graphene using density functional theory (DFT) calculations.

Computational details

We carried out all-electron ab initio DFT calculations using the double numerical basis set with polarization function (the DNP basis set), implemented in the DMol3 package [32, 33]. The Perdew-Burke-Ernzerhof (PBE) functional [34] within the generalized gradient approximation (GGA) was chosen for the spin-unrestricted DFT calculations. The GGA/PBE method used has also been employed successfully to study graphene systems in previous theoretical studies, including their functionalization with various functional groups [24, 35–40]. A tetragonal supercell with dimension 17.22×17.22× 30 Å^3 (seven periodic lengths of graphene layer in both *a* and *b* directions) was chosen as the benchmark model system, which consists of 98 carbon atoms. The size of this employed supercell in this work is large enough to completely avoid image interactions. For geometry optimizations, the Brillouin zone was sampled by 2×2×1 special *k*-points using the Monkhorst-Pack scheme [41], while a 15×15×1*k*-point was used to calculate the electronic properties. The Hirshfeld method was adopted in the present work. It should be pointed out that this method is based on the deformation electron density, which is less

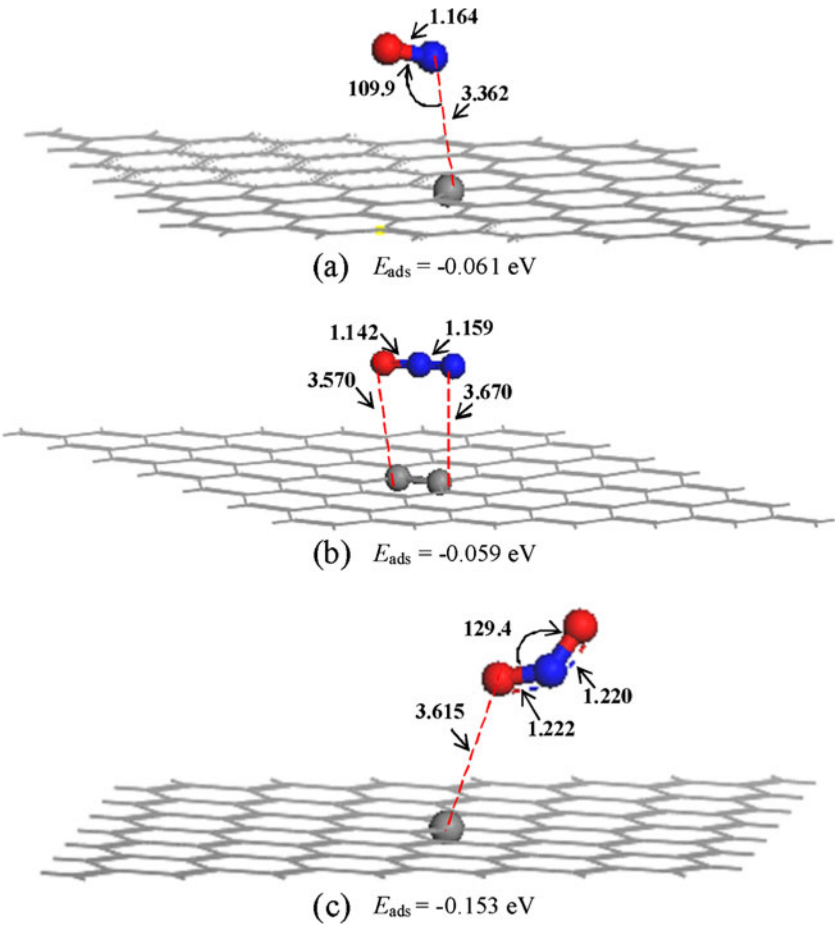

1.164
109.9
3.362
(a) E_{ads} = -0.061 eV
1.142
1.159
3.570
3.670
(b) E_{ads} = -0.059 eV
129.4
3.615
1.222
1.220
(c) E_{ads} = -0.153 eV

Fig. 1 Optimized geometrical configurations of **a** NO, **b** N_2O, and **c** NO_2 molecules adsorbed on the pristine graphene. The bond distances and angles are in Ångstroms and degrees, respectively

sensitive to the chosen basis sets than Mulliken charge analysis, although Hirshfeld charge analysis generally underestimates atomic charges, according to many results in the literature [42–48]. We did not consider correction for basis set superposition error (BSSE) [49] to calculate the adsorption energy because a recent study has proven that the numerical basis sets implemented in Dmol3 can minimize or even eliminate BSSE [50].

Results and discussion

NO, N_2O, and NO_2 adsorption on pristine graphene sheet

First, we study adsorption of NO, N_2O, and NO_2 on pristine graphene sheet. The three molecules were placed initially on various sites on the graphene sheet (e.g., on-top of a carbon site, the center of a hexagonal ring, or a C–C bond) with different orientations (adsorbed molecule perpendicular or parallel to the graphene sheet). After geometrical optimization, the most stable adsorption configurations of the three molecules on pristine graphene are shown in Fig. 1a–c. The calculated E_{ads} values for NO, N_2O, and NO_2 on pristine graphene are −0.061, −0.059, and −0.153 eV, respectively. Moreover, the shortest distances between the three adsorbed molecules and graphene sheet are 3.362, 3.570, and 3.615 Å, respectively. The small adsorption energies and long distances indicate the three molecules are only adsorbed *physically* onto the sheet of pristine graphene, which is in good agreement with Leenaerts' study [20]. We should point out that GGA in DFT is not capable of describing physisorption, while local density approximation (LDA) has been shown to be a reliable functional to study systems involving van der Waals interactions [51–53] and can give an adsorption energy much closer to the MP2 calculation [54–56]. Thus, we also calculated adsorption energies of NO, N_2O, and NO_2 on perfect graphene through LDA with the Perdew-Wang (PWC) functional [57]. The results indicate that the interactions of the three molecules with perfect graphene are indeed weak, with adsorption energies of −0.082 (for NO), −0.074 (for N_2O), and −0.202 eV (for NO_2), respectively. The weak interaction is further confirmed by the negligible charge transfer between pristine graphene and these adsorbates (< 0.01 e), which is not enough to change the electronic properties of the intrinsic graphene. In other words, the intrinsic graphene is not sensitive to NO, N_2O, and NO_2 molecules. Thus, from a theoretical point of view, pristine graphene is not a suitable candidate for detecting these three gases.

For doped graphene with one carbon atom substituted by one silicon atom in a supercell, drastic changes in the geometric structure of the graphene sheet have been observed, as shown in Fig. 2a: the silicon atom preserves its sp^3 character and bonds with pyramidal-like configurations, with bond angles close to 105°. The Si–C bond length is 1.765 Å, which is quite large compared to 1.420 Å for C–C sp^2 bonds. The 24% increase in the bond length combined with the difference in bond angles forces Si to protrude from the graphene plane, also displacing the positions of the first-, second-, and third-out-of-plane neighbors. This can be explained as a corrugation induced by the presence of the Si atom. Moreover, we also explored the effects of Si-doping on the electronic properties of graphene by analysis of the calculated band structures (Fig. 2c). Compared to the electronic structure of perfect graphene (Fig. 2b), the minimum of the conduction band edge (CBM) of doped graphene is found to be shifted up slightly, forming a small band gap of 0.054 eV (Fig. 2c).

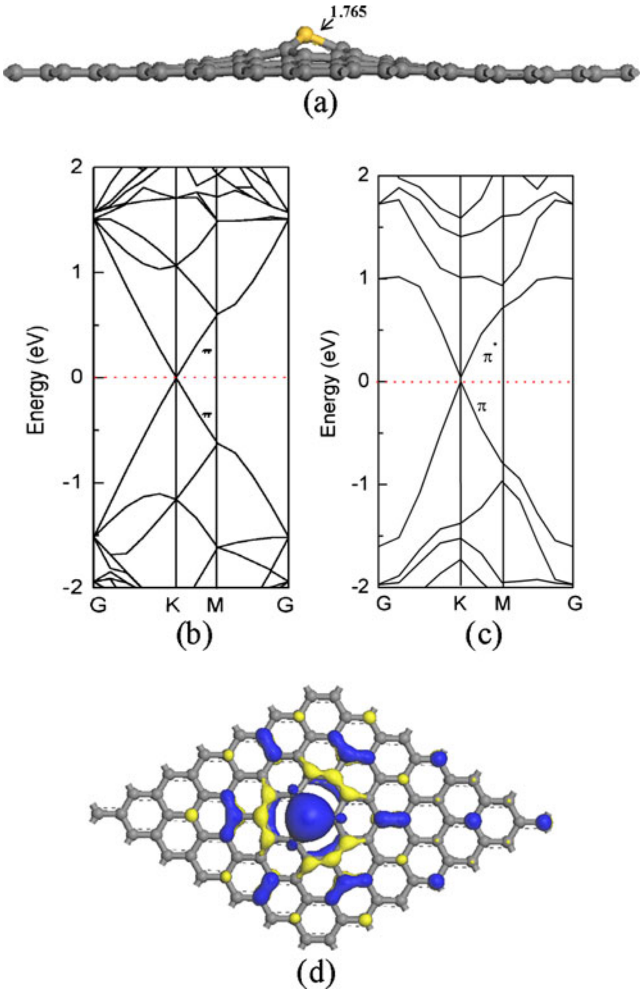


Fig. 2 **a** Optimized geometrical structure of Si-doped graphene. *Gray* and *yellow balls* represent carbon and silicon atoms, respectively. Bond distances are in Ångstroms. **b** Band structure of the Si-doped graphene. *Red dotted lines* denote Fermi levels. **c** Isosurface (isovalue is 0.025 au) of the highest occupied molecular orbitals (HOMOs) of Si-doped graphene. *Blue* and *yellow* regions denote positive and negative sign of wave functions, respectively

This is because silicon has four electrons in its valence shell, but it binds with sp^3 hybridization, following a trigonal pyramidal coordination, thus creating a localized state when bonded to a graphitic network, which would have little effect on the semi-metallic character of the graphene. This behavior of Si-doping of graphene is similar to a recent report of phosphorus-doped graphene [58]. Additionally, it is known that the Si atom works as a donor when incorporated into graphene. Therefore, the highest occupied level (highest occupied molecular orbital, HOMO) for Si-doped graphene is contributed mainly to the excess electrons of the Si atom. This can be reflected by its HOMOs (Fig. 2d): most states of the HOMOs are localized around the dopant, indicating that the Si atom has much higher reactivity than other atoms. Thus, the Si-dopant acts as the active site for foreign adsorbates, as will be testified by the following results.

NO, N_2O, and NO_2 adsorption on Si-doped graphene

In this section, we explore mainly the effects of Si-doping on the adsorption of NO, N_2O, and NO_2 on graphene. The most stable adsorption configurations of the three molecules on Si-doped graphene are listed in Figs. 3, 4, 5, 6. We find that interaction of the adsorbate with graphene is greatly enhanced due to introduction of the Si-dopant into the graphene sheet.

NO adsorption

Two stable configurations are obtained on an individual NO molecule on Si-doped graphene, i.e., the Si-atom of Si-doped graphene is close to the N- and O-atoms of NO, respectively, as shown in Fig. 3 (labeled as configurations I and II). The E_{ads} values for configuration I (Fig. 3a) and II (Fig. 3b) are −0.816 and −0.209 eV, respectively, which are higher than that of intrinsic graphene (−0.061 eV). Obviously, configuration I, in which the N-atom in NO is bound to the Si-atom in graphene, is the most stable, and its E_{ads} increases by 0.755 eV, compared with the intrinsic graphene system. This indicates that Si-doped graphene is more sensitive to the NO molecule. Moreover, NO adsorption causes a change in the geometrical structure of Si-doped graphene, resulting in an expansion of the Si–C

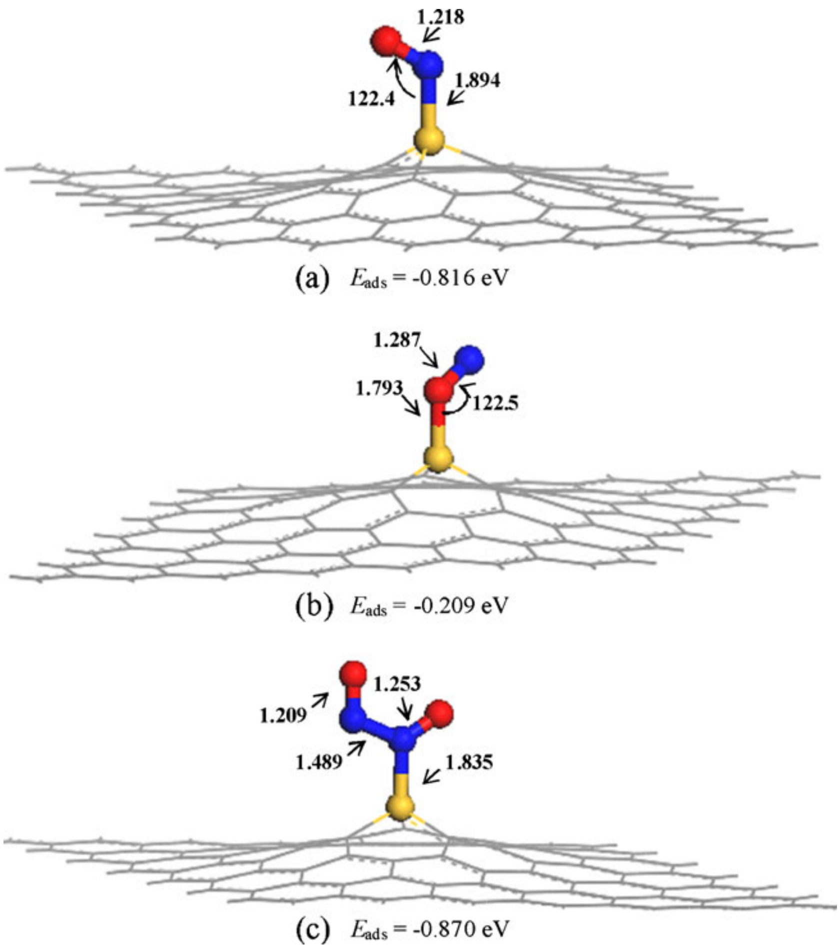

1.218
1.894
122.4
(a) E_{ads} = -0.816 eV
1.287
1.793
122.5
(b) E_{ads} = -0.209 eV
1.253
1.209
1.489
1.835
(c) E_{ads} = -0.870 eV

Fig. 3 a, **b** Optimized structures of NO molecule adsorbed onto Si-doped graphene: the NO-graphene systems labeled by configuration I (**a**) and II (**b**) with the N and O atom of a NO molecule close to the Si atom of Si-doped graphene, respectively. **c** Configuration of two NO molecules on Si-doped graphene. The *gray*, *yellow*, *blue*, and *red balls* represent carbon, silicon, nitrogen, and oxygen atoms, respectively. Bond distances and angles are in Ångstroms and degrees, respectively

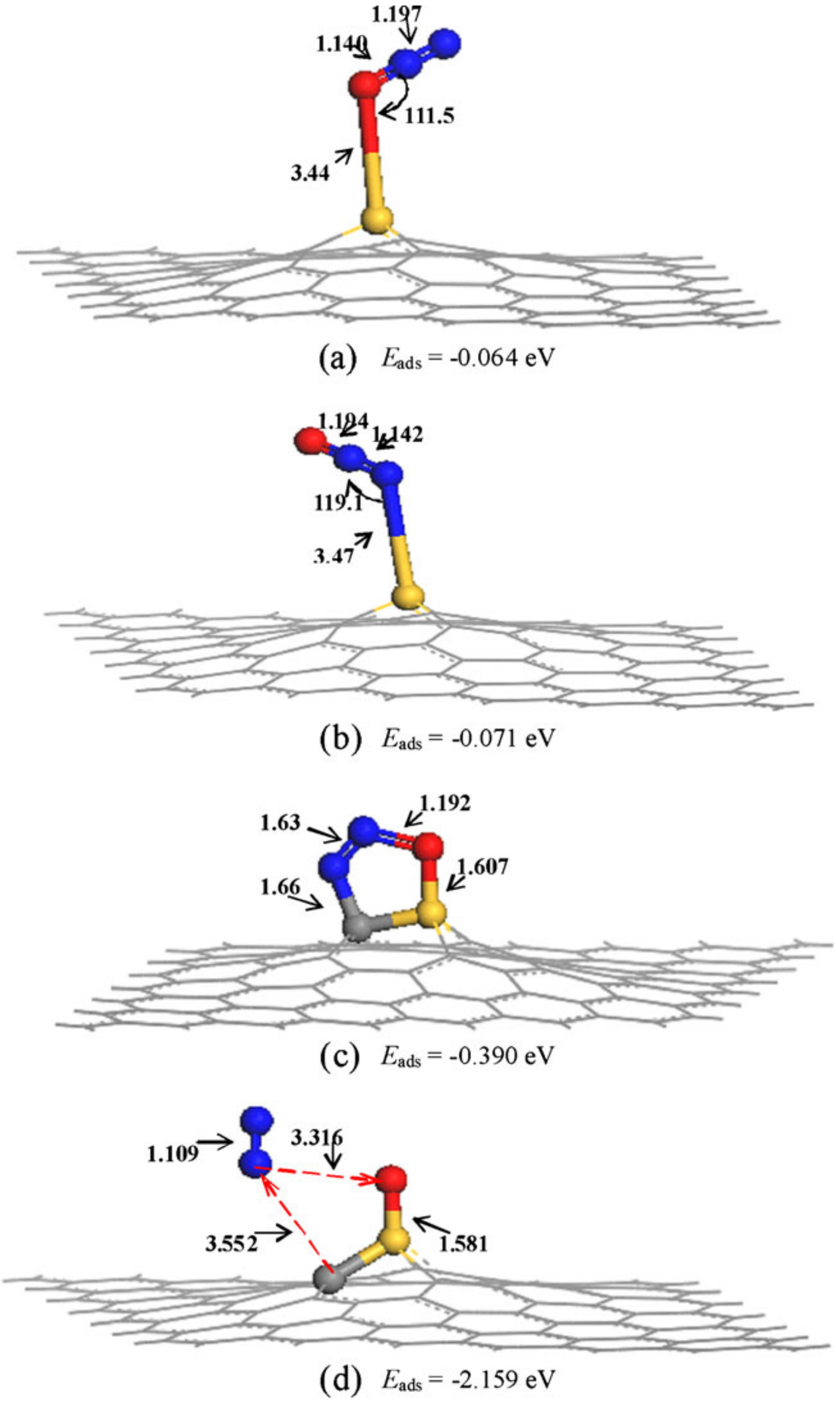

1.197
1.140
111.5
3.44
(a) E_{ads} = -0.064 eV
1.194
1.142
119.1
3.47
(b) E_{ads} = -0.071 eV
1.63
1.192
1.607
1.66
(c) E_{ads} = -0.390 eV
1.109
3.316
3.552
1.581
(d) E_{ads} = -2.159 eV

Fig. 4 a–d Optimized structures of a single N_2O adsorbed on Si-doped graphene, showing various adsorption configurations. Bond distances and angles in Ångstroms and degrees, respectively

bond from 1.765 Å to 1.781 Å, while the N–O bond is increased to 1.218 Å from the isolated bond length of 1.164 Å. The distance between the adsorbed NO and Si-doped graphene sheet is 1.894 Å (Fig. 3a). Because the ground state of Si-doped graphene is nonmagnetic, the net spin of configuration I with an unpaired electron should originate from the magnetism of the adsorbed NO molecule (Fig. S1).

On the basis of the adsorption of one NO molecule, we further studied adsorption of a second NO on Si-doped graphene. Three kinds of initial adsorption configurations were considered, i.e., a second NO adsorbed on (1) the C-atoms nearest to the Si-dopant, and (2) the O-, or (3) N-atoms of the first NO molecule. Each initial configuration was fully optimized. Adsorption of the second NO onto Si-graphene of types (1) and (2) were found to be unstable and to collapse to type (3), i.e., the N-atom of the second NO is attached to the N-atom of the first NO with a distance of 1.489 Å (Fig. 3c). The Si–N distance is 1.835 Å,

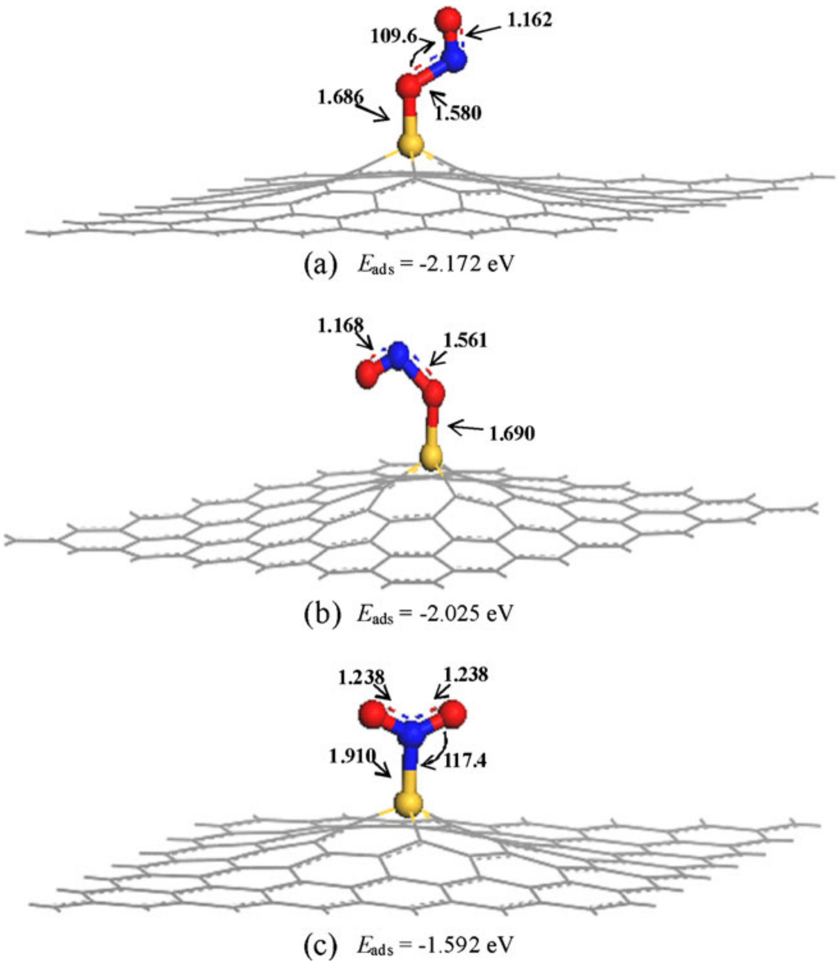

109.6
1.162
1.686
1.580
(a) E_{ads} = -2.172 eV
1.168
1.561
1.690
(b) E_{ads} = -2.025 eV
1.238
1.238
1.910
117.4
(c) E_{ads} = -1.592 eV

Fig. 5 **a**–**c** Optimized structures of a single NO_2 adsorbed on Si-doped graphene, showing various adsorption configurations. Bond distances and angles in Ångstroms and degrees, respectively

which is smaller than that of adsorption of an individual NO (1.894Å). The adsorption energy[1] of two NO molecules (−0.870 eV) is slightly larger than that of the first NO (−0.816 eV), indicating that NO molecules prefer the pair arrangement (or dimerization) on Si-doped graphene. This can be attributed to the following: (1) the net spins of the NO-Si-doped-graphene system and the isolated NO are derived mainly from the contributions of their respective N atoms (Fig. S1). Therefore, when a second NO molecule is adsorbed on Si-doped graphene sheet at the position wehre the first NO molecule is located, their net spins prefer to couple spontaneously with each other. (2) The HOMO of the Si-doped graphene functionalized by one NO molecule (Fig. S2) locates mainly on the N-atom—possibly the most reactive site of this whole system toward a second NO molecule. This mechanism of NO dimerization on Si-doped graphene is very similar to that of metal-based catalysts reported by Sojka [59, 60], which have also been confirmed experimentally by IR spectroscopy [61]. We also tried adding more NO molecules to Si-doped graphene. The results show that the structure is unstable and the third NO molecule is seen to fly off.

N_2O adsorption

When a single N_2O molecule is attached to Si-doped graphene sheet, we initially consider three typical possible configurations, as shown in Scheme 1: (1) O- or N-attacking, (2) [3+2]-cycloaddition, and (3) [2+2]-cycloaddition. In type (1), the linear N_2O molecule is attached vertically to the active sites (i.e., Si- or its nearest C-atoms). In type (2), N_2O uses its N- and O-atoms to bond with the Si- and C-atoms of Si-doped graphene, forming a five-membered ring, while a four-membered ring, in which the N–N or N–O bond of the N_2O molecule attacks the Si–C bond of Si-doped graphene, will be obtained via type (3).

In more detail, after full structural optimization for type (1), two stable configurations are obtained (Fig. 4a,b): the N_2O molecule is shown to be adsorbed only *physically* onto the Si-doped graphene with a small adsorption energy (a few tens meV), stemming from van der Waal's attraction.

[1] The adsorption energy of n adsorbate on Si-doped graphene is defined as: E_{ads}=[E_{total} (nadsorbate-Si-doped graphene)] - n[E_{total} (adsorbate)] - [E_{total} Si-doped graphene)]/n, where E_{total} is the total energy of the studied system and n is the number of the adsorbate

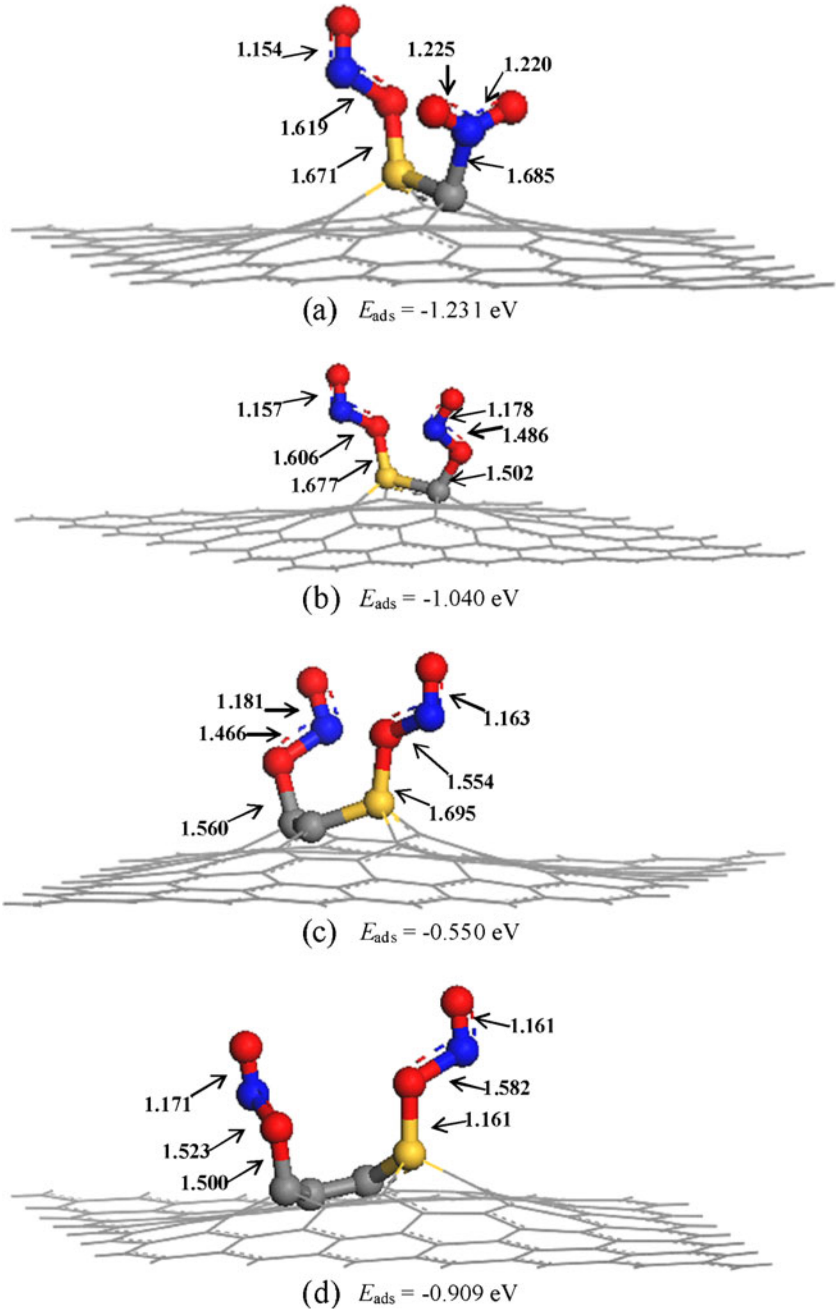
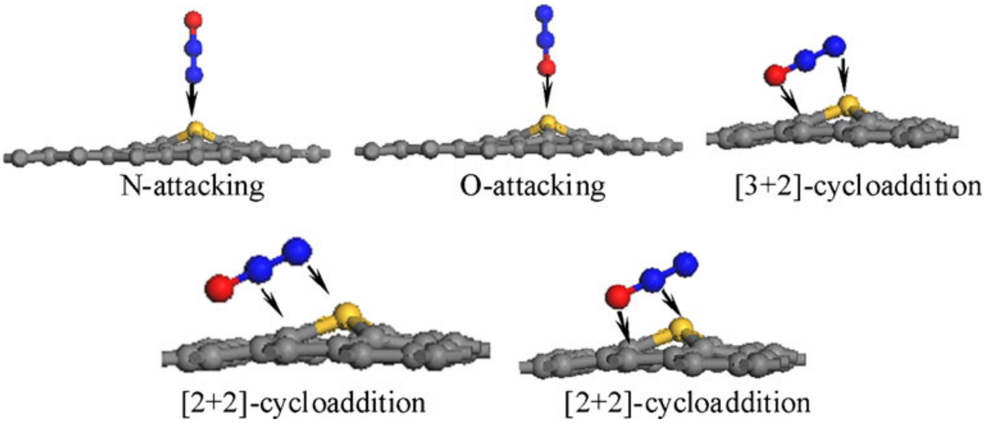


Fig. 6 a–**d** Optimized structures of a second NO_2 adsorbed on Si-doped graphene where the first NO is located, showing various adsorption configurations. Bond distances and angles in Ångstroms and degrees, respectively

The distances between the O- or N-atom in N_2O and the Si-atom of Si-doped graphene are 3.443 and 3.471 Å, respectively. The weak physisorption of the N_2O molecule is thought to be because no local structural deformation is observed for the Si-doped graphene sheet. Moreover, when the N_2O molecule is adsorbed onto the Si-doped graphene in type (2), i.e., [3+2]-cycloaddition, another stable configuration (Fig. 4c) is obtained, with adsorption energy of −0.390 eV and 0.226 electrons being transferred from graphene to N_2O. The bond lengths of the newly formed O–Si and N–C bonds are 1.607 and 1.665 Å, respectively. N_2O [3+2] cycloaddition on Si-doped graphene induces local structural deformation to both N_2O and graphene: (1) the bond angle of O–N–N of N_2O is decreased greatly from 180° in free N_2O to 115.2° in the adsorbed form; (2) the N_2O-adsorbed Si–C bond is increased from 1.765 to 1.841 Å. Another stable configuration (Fig. 4d) is obtained when the N_2O molecule is adsorbed on Si-doped graphene in type (3). Of particular interest, a N_2 molecule in this configuration is found to escape from the sheet of Si-doped graphene, leaving an O-atom attached to the Si-atom of graphene. The O–Si bond length is 1.581 Å, while the N–O and N–C distances are further apart by 3.316 and 3.552 Å, respectively. This



Scheme 1 Typical possible configurations for attaching a single N_2O molecule to Si-doped graphene sheet

configuration is the most stable of all obtained adsorption configurations, with an adsorption energy of −2.159 eV. This leads to the suggestion that the N_2O molecule could be reduced into the N_2 molecule on the Si-doped graphene sheet. Hence, Si-doped graphene might be an ideal candidate as a metal-free catalyst for N_2O-reduction.

NO_2 adsorption

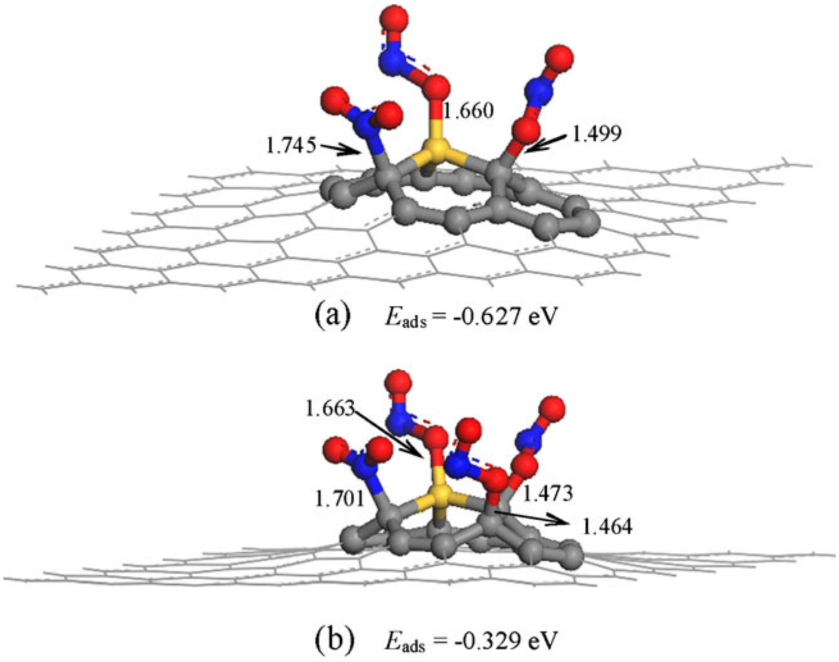
For NO_2 adsorption on Si-doped graphene, the most stable configuration is one in which the Si-atom in Si-doped graphene is bound with one O-atom in a NO_2 molecule with a nitrite configuration. The Si–O bond length is 1.686 Å as shown in Fig. 5a. As with NO and N_2O adsorption, the structures of Si–doped graphene and NO_2 are also deformed due to NO_2 adsorption. For example, the C–Si bond length of graphene expands to 1.799 Å, while the O–N bond of the NO_2 on the side of the Si-atom is elongated from 1.210 Å in isolated NO_2 to 1.580 Å. The calculated adsorption energy is about −2.172 eV. It can be expected that the adsorbed Si-doped graphene becomes a magnetic material, with a magnetic moment of 1 μ_B upon adsorption of a single NO_2 molecule. In contrast to NO adsorption on Si-doped graphene (Fig. S1), however, the spin densities of this graphene system are located mainly on a few C-atoms around the Si-, N-, and O-atoms in the adsorbed NO_2 (Fig. S3). In addition, we also obtained two meta-stable adsorption configurations as presented in Fig. 5b,c. Their adsorption energies are −2.025 and −1.592 eV, respectively, which are smaller than that of the most stable one.

Based on the adsorption of a single NO_2 molecule, we further studied the attachment of a second NO_2 molecule to Si-doped graphene on which the first NO_2 has been located. After geometrical optimization, the most energetically favored configuration is one in which the second NO_2 molecule is located at the ortho position of the Si-atom with a nitro configuration (Fig. 6a). The most stable configuration can be rationalized thus: (1) the adsorption of a single NO_2 molecule with an unpaired electron activates those carbon atoms near adsorption sites (Fig. S2). Hence, these “activated” carbon sites can be considered as the “default state” of the graphene, which initiates adsorption of a second NO_2 molecule onto it. (2) About 0.374 electrons are shown to be transferred from the HOMO of Si-doped graphene to the LUMO of a second NO_2 molecule. Thus, the second NO molecule prefers to be adsorbed on the site that makes the largest contribution to the HOMO of Si-doped graphene (Fig. 3d). The length of the newly formed C–N bond is 1.685 Å, which is larger than the typical C–N distance of 1.500 Å. The Si–C bonds involving the adsorption NO_2 molecules are 1.802, 1.802, and 1.863 Å, respectively. Because steric repulsion exists between the two NO_2 molecules, and the Si–O binding energy (798 kJ mol^{-1}) is slightly larger than that of N–C (770 kJ mol^{-1}) [62], it is not surprising that the adsorption energy[2] of the second NO_2 molecule (−1.231 eV) is significantly smaller than that of the first (−2.172 eV). As far as the lowest-energy configuration is concerned, three metastable adsorption configurations are obtained. For these configurations (Fig. 6b–d), we find that the second NO_2 molecule is in the nitrite configuration, and located at the ortho-, meta-, and para-sites of the same six-membered ring on which the first NO_2 molecule is located. Moreover, the calculated adsorption energies are −1.040 (ortho-site, Fig. 6b), −0.550 (meta-site, Fig. 6c), and −0.909 eV (para-site, Fig. 6d), respectively.

The next and most important question is: what is the maximum number of NO_2 molecules that can be bound to Si-doped graphene? To answer this question, we gradually added adsorbed NO_2 molecules up to the number five. The results show that Si-doped graphene functionalized with five NO_2 molecules is unstable, and one NO_2 molecule leaves the sheet of Si-doped graphene. In other words, the maximum number of NO_2 molecules that can bind to doped graphene with one carbon atom substituted by one silicon atom is four (Fig. 7). The average adsorption energy for

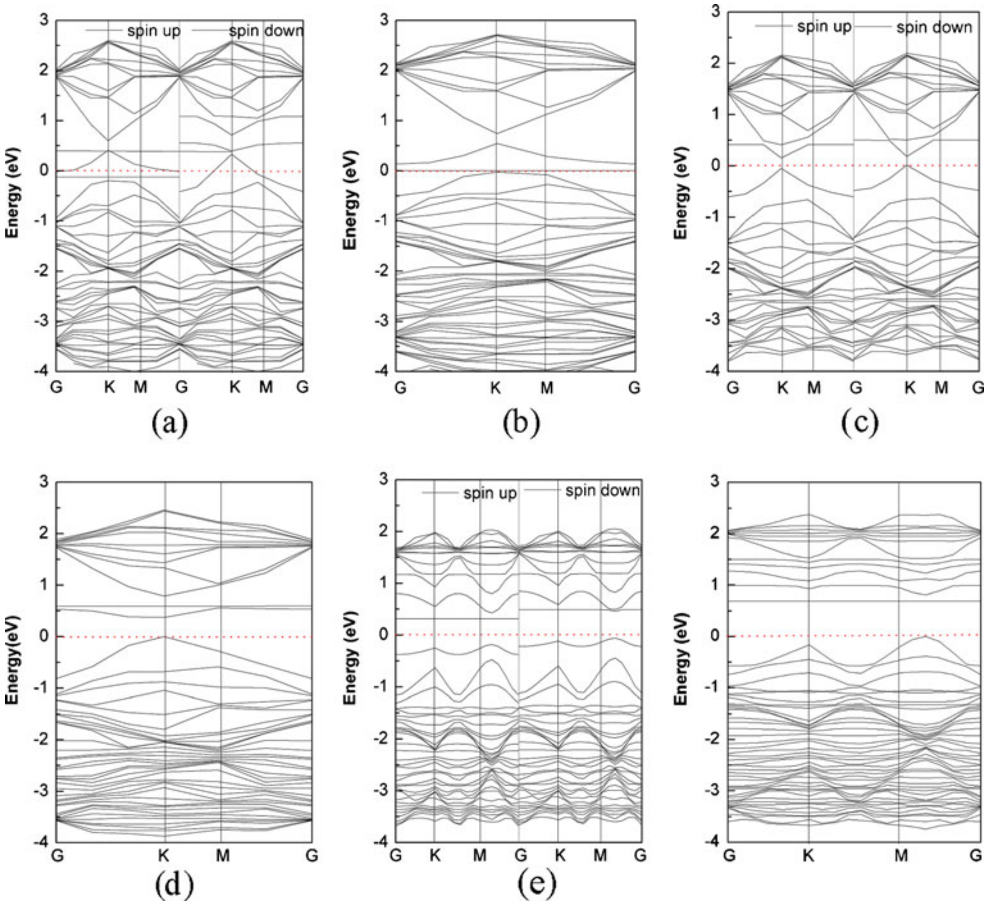
[2] see footnote 1

Fig. 7 Optimized structures of **a** three and **b** four NO_2 molecules adsorbed onto Si-doped graphene. Bond distances in Ångstroms



four NO_2 molecules is −0.329 eV. The large surface:volume ratio of graphene plays a decisive role in device engineering, as a large number of Si atoms may be doped onto each graphene sheet.

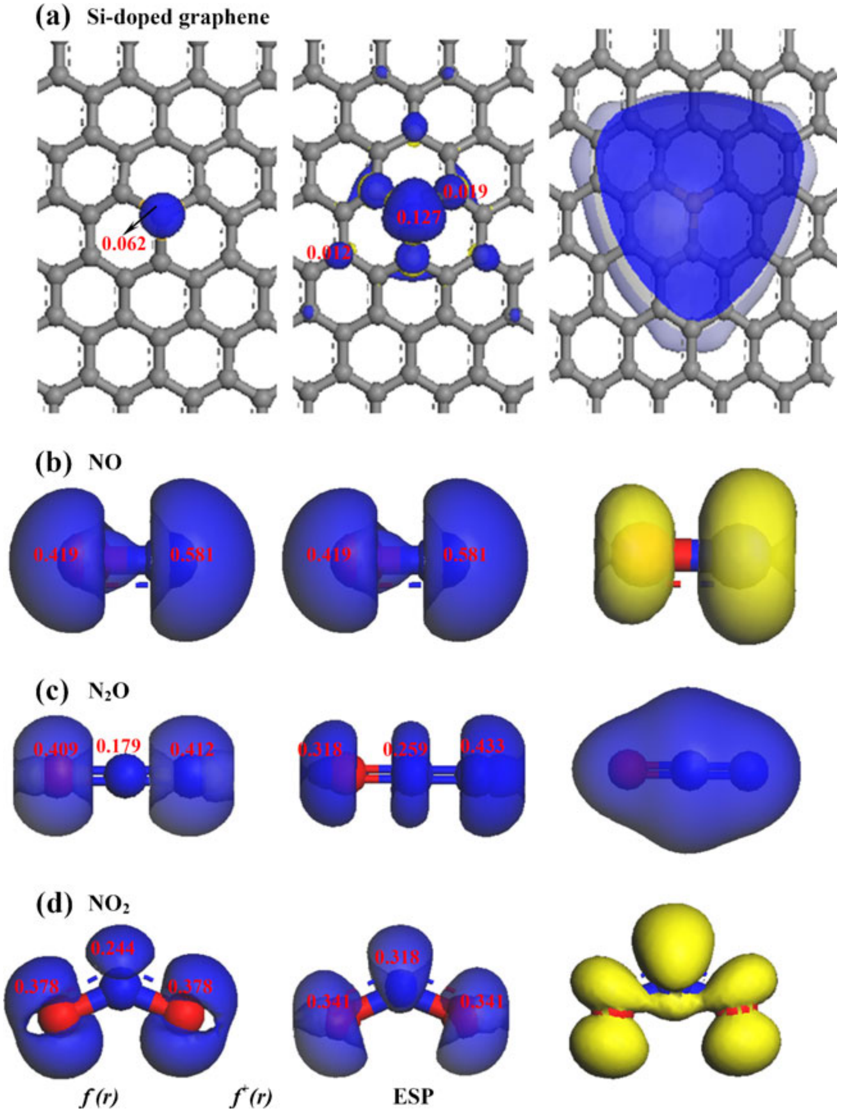
Fig. 8 Band structure of Si-doped graphene after adsorption of **a** one NO, **b** two NO, or **c** one, **d** two, **e** three, or **f** four NO_2 molecules. *Red dotted lines* denote the Fermi level



Effects of NO and NO_2 adsorption on the electronic properties of Si-doped graphene

Any change in electronic properties is an important factor in evaluating the potential application of Si-doped graphene in gas sensing. As discussed above, Si-doped graphene is a semiconductor with a small band gap of ~0.054 eV. Adsorption of a single NO or NO_2 molecule onto Si-doped graphene introduces spin polarization into the whole system. It can be seen from Fig. 8a that Si-doped graphene exhibits half-metallic behaviors after adsorbing one NO. The spin-up state opens a band gap of 0.186 eV, and the top valence band lies below the Fermi level, while the spin-down state has densities up to 0.065 eV above Fermi level. For adsorption of two NO molecules, the spin polarization of Si-doped graphene disappears and a band gap with a width of ~0.137 eV is opened (Fig. 8b). On the other hand, the CBM of Si-doped graphene is shifted slightly, resulting in an increase of its band gap for NO_2-adsorption, which is dependent on the number of the adsorbed NO_2 molecules. For example, when one or two NO_2 molecules are adsorbed, the band gaps of Si-doped graphene are increased to 0.204 (one NO_2) and 0.374 eV (two NO_2), respectively, while the band gaps are increased to 0.490 and 0.680 eV, respectively, upon adsorption of three and four NO_2 molecules. The changes in band structures of Si-doped graphene due to NO or NO_2 adsorption are further confirmed by the charge transfer between graphene and adsorbate. Analysis of Hirshfeld charges shows that there is a charge transfer of about 0.125 e from Si-doped graphene to NO molecules, and of about 0.334 e to NO_2 molecules, suggesting that both NO and NO_2 can work as an acceptor. In short, the appreciable adsorption energy and large charge transfer render Si-doped graphene an excellent sensor for

Fig. 9 Iso-surface of Fukui functional $f^-(r)$, $f^+(r)$, and electrostatic potential (ESP) of Si-doped graphene, NO, N_2O, and NO_2. *Blue* and *yellow regions* denote positive and negative sign of wave functions, respectively



the detection of NO and NO_2 molecules. The former (appreciable adsorption energy) allows reversible adsorption to be accomplished easily without destroying the host materials, while the latter, which is larger than the well-established experimental case [63], is expected to induce sizeable changes in the conductivity of the system.

From the above results and analysis, it is obvious that the substituted doping of a Si atom into graphene can greatly enhance the chemical reactivity of graphene towards NO, N_2O, and NO_2 molecules. The good chemical reactivity of Si-doped graphene can be further comfirmed by computed Fukui functions [64]. Fukui functions measure the sensitivity of the charge density ρ(r) with respect to the loss or gain of electrons via the following definitions:

$$\mathrm{f}^{+}(\mathrm{r}) = (\rho\mathrm{N} + \Delta\mathrm{N}(\mathrm{r}) - \rho_{\mathrm{N}}(\mathrm{r}))/\Delta\mathrm{N} \tag{1}$$

$$\mathrm{f}^{-}(\mathrm{r}) = (\rho_{\mathrm{N}}(\mathrm{r}) - \rho\mathrm{N} - \Delta\mathrm{N}(\mathrm{r})))/\Delta\mathrm{N} \tag{2}$$

$$\mathrm{f}^{0}(\mathrm{r}) = (\mathrm{f}^{+}(\mathrm{r}) +)\mathrm{f}^{-}(\mathrm{r}))/2 \tag{3}$$

in which $f^{+}(r)$ measures changes in electron density when the molecule (or cluster) gains electrons, thereby providing a description of reactivity with respect to nucleophilic attack. In contrast, f(r) measures the reactivity with respect to electrophilic attack (loss of electrons). The $f^{0}(r)$, which is the average of $f^{+}(r)$ and $f^{\,-}(r)$, describes radical attack. Figure 9 plots the iso-surface of $f^{+}(r)$ and $f^{\,-}(r)$ of optimized Si-doped graphene, NO, N_2O, and NO_2, respectively. As shown in Fig. 9, the Si atom of Si-doped graphene exhibits a fairly large contour of f^{+}(r) iso-surface, compared to its $f(r)$ iso-surface. Thus, Si-doped graphene possesses relatively high reactivity with respect to nucleophilic attack, and its Si atom is the most reactive site for nucleophilic guest molecule adsorption. This fact is in good agreement with the HOMO (Fig. 2d) and electrostatic potential (ESP, Fig. 9): most states of the HOMO or ESP are localized around Si atom. For NO and NO_2, their highest reactivity sites are the N atom (f_N^{-}= 0.581) of NO and O atom (f_N^{-}=0.581) of NO_2 derived from Hirshfield scheme, respectively. Hence, it is clear that the N atom of NO and the O atom of NO_2 will most favor being attached to the Si atom of Si-doped graphene (Fig. 3a, 5a). On the other hand, it is also clear that the $f(r)$ iso-surface of N_2O is located mainly on the terminal O and N atoms. In other words, the two atoms represent the reactivity sites for adsorption on Si-doped graphene (Fig. 4c).

Conclusions

Using DFT calculations, we have studied the adsorptions of three nitrogen oxides—NO, N_2O, and NO_2 molecules—onto pristine and Si-doped graphene. It was found that Si-doped graphene exhibited completely different behavior when exposed to the three gaseous molecules. Specifically, the moderate adsorption strength and obvious changes in electronic structure produced by NO and NO_2 adsorption make Si-doped graphene a suitable candidate fpr a NO or NO_2 sensor. Interestingly, N_2O, a greenhouse gas, can be easily reduced to benign N_2 on a sheet of Si-doped graphene. This suggests that Si-doped graphene may be used as a metal-free catalyst for N_2O reduction. Finally, by exploring the calculated band structures, we find that the electronic properties are modified significantly after adsorption of NO and NO_2 molecules, which is dependent on the coverage of the adsorbate. The present work is useful not only in deepening our understanding of the properties of graphene, but also to further widen its fields of application.

Acknowledgments We gratefully acknowledge support of this research by the Key Program Projects of the National Natural Science Foundation of China (No 21031001), the National Natural Science Foundation of China (No 20971040), the Cultivation Fund of the Key Scientific and Technical Innovation Project, Ministry of Education of China (No 708029), the Key Program Projects of the Province Natural Science Foundation of Heilongjiang Province (No. ZJG0602-01), the National Basic Research Priority Program (No. 2007CB914104), the Committee of Education of Heilongjiang Province (No. 11541095), the Scientific Research Foundation for Doctor of Harbin Normal University (08XKYL38), and Heilong Jiang Postdoctoral Funds for scientific research initiation. The authors would like to express their gratitude to the reviewers for raising invaluable comments and suggestions.

References

1. Novoselov KS, Geim AK, Morozov SV, Jiang D, Zhang Y, Dubonos SV, Grigorieva IV, Firsov AA (2004) Science 306:666–669
2. Geim AK (2009) Science 324:1530–1534
3. Rao CNR, Sood AK, Subrahmanyam KS, Govindaraj A (2009) Angew Chem Int Edn 48:7752–7777
4. Rao CNR, Biswas K, Subrahmanyam KS, Govindaraj A (2009) J Mater Chem 19:2457–2469
5. Neto AHC, Guinea F, Peres NMR, Novoselov KS, Geim AK (2009) Rev Mod Phys 81:109–162
6. Taghioskoui M (2009) Mater Today 12:34–37
7. Zhu Y, Murali S, Cai W, Li X, Suk JW, Potts JR, Ruoff RS (2010) Adv Mater 22:3906–3924
8. Schedin F, Geim AK, Moeozov SV, Hill EW, Blake P, Katsnelson MI, Novoselov KS (2007) Nat Mater 6:652–655
9. Barbolina II, Novoselov KS, Morozov SV, Dubonos SV, Missous M, Volkov AO, Christian DA, Grigorieva IV, Geim AK (2006) Appl Phys Lett 88:013901
10. Allen MJ, Tung VC, Kaner RB (2010) Chem Rev 110:132–145
11. Loh KP, Bao Q, Ang PK, Yang J (2010) J Mater Chem 20:2277–2289
12. Terronesa M, Botello-Méndez AR, Campos-Delgado J, López-Urías F, Vega-Cantú YI, Rodríguez-Macías FJ, Elías AL, Muñoz-Sandoval E, Cano-Márquez AG, Charlier J-C, Terrones H (2010) Nanotoday 5:351–372

13. Abergel DSL, Apalkov V, Berashevich J, Ziegler K, Chakraborty T (2010) Adv Phys 59:261–482
14. Choi W, Lahiri L, Seelaboyina R, Kang YS (2010) Crit Rev Solid State Mater Sci 35:52–71
15. Novoselov KS, Geim AK, Morozov SV, Jiang D, Katsnelson MI, Grigorieva IV, Dubonos SV (2005) Firsov AA. Nature 438:197–200
16. Zhang Y, Tan Y, Stormer HL, Kim P (2005) Nature 438:201–204
17. Danneau R, Wu F, Craciun MF, Russo S, Tomi MY, Salmilehto J, Morpurgo AF, Hakonen P (2008) J Phys Rev Lett 100:196802
18. Wehling TO, Noveselov KS, Morozov SV, Vdovin EE, Katsnelson MI, Geim AK, Lichtenstein AI (2008) Nano Lett 8:173–177
19. Goldoni A, Larciprete R, Petaccia L, Lizzit S (2003) J Am Chem Soc 125:11329–11333
20. Leenaerts O, Partoens B, Peeters FM (2008) Phys Rev B 77:125416
21. Dai J, Giannozzi P, Yuan J (2009) Surf Sci 603:3234–3238
22. Johnson JJ, Behnam A, Pearton SJ, Ural A (2010) Adv Mater 22:4877–4880
23. Kaniyoor A, Jafri RI, Arokiadoss T, Ramaprabhu S (2009) Nanoscale 1:382–386
24. Dai J, Yuan J, Giannozzi P (2009) Appl Phys Lett 95:232105
25. Zhang YH, Chen YB, Zhou KG, Liu CH, Zeng J, Zhang HL, Peng Y (2009) Nanotechnology 20:185504
26. Ao ZM, Yang J, Li S, Jiang Q (2008) Chem Phys Lett 461:276–279
27. Francesco FAP, Kelly FJ, Holgate ST (2005) Air Quality Guidelines Global Update; World Health Organization
28. Trogler WC (1999) Coord Chem Rev 187:303–327
29. Duce R et al (2008) Science 320:893–897
30. Ravishankara AR, Daniel JS, Portmann RW (2009) Science 326:123–125
31. Berger C, Song ZM, Li TB, Li XB, Ogbazghi AY, Feng R, Dai ZT, Marchenkov AN, Conrad EH, First PN, de Heer WA (2004) J Phys Chem B 108:19912–19916
32. Delley B (1990) J Chem Phys 92:508–517
33. Delley B (2000) J Chem Phys 113:7756–7764
34. Perdew JP, Burke K, Ernzerhof M (1996) Phys Rev Lett 77:3865
35. Wu X, Zeng XC (2009) Nano Lett 9:250–256
36. Jiang D, Sumpter BG, Dai S (2006) J Phys Chem B 110:23628–23632
37. Choi WI, Jhi SH, Kim K, Kim YH (2010) Phys Rev B 81:085441
38. Manna AK, Pati SK (2009) Chem Asian J 4:855–860
39. Dai J, Yuan J, Giannozzi P (2010) Phys Rev B 81:165414
40. Suggs K, Reuven D, Wang XQ (2011) J Phys Chem C 115:3313–3317
41. Monkhorst HJ, Pack JD (1976) Phys Rev B 13:5188–5192
42. Hirshfeld FL (1977) Theor Chim Acta 44:129–138
43. Davidson ER, Chakravorty S (1992) Theor Chim Acta 83:319–330
44. Meister J, Schwarz WHE (1994) J Phys Chem 98:8245–8252
45. Wiberg KB, Rablen PR (1993) J Comput Chem 14:1504–1518
46. Verstraete M, Gonze X (2003) Phys Rev B 68:195123
47. Wu XJ, Zeng XC (2006) J Chem Phys 125:044711
48. Bultinck P, Alsenoy VC, Ayers PW, Carbó-Dorca R (2007) J Chem Phys 126:144111
49. Boys SF, Bernardi F (1970) Mol Phys 19:553
50. Inada Y, Orita H (2008) J Comput Chem 29:225–232
51. Cobian M, Iniguez J (2008) J Phys Condens Matter 20:285212
52. Ataca C, Aktürk E, Ciraci S, Ustunel H (2008) Appl Phys Lett 93:043123
53. de Andres PL, Ramírez R, Vergés JA (2008) Phys Rev B 77:045403
54. Cabria I, López MJ, Alonso JA (2005) J Chem Phys 123:204721
55. Okamoto Y, Miyamoto Y (2001) J Phys Chem B 105:3470–3474
56. Wu XJ, Gao Y, Zeng XC (2008) J Phys Chem C 112:8458–8463
57. Perdew JP, Wang Y (1992) Phys Rev B 45:13244–13249
58. Cruz-Silva E, López-Urías F, Muñoz-Sandoval E, Sumpter BG, Terrones H, Charlier J-C, Meunier V, Terrones M (2009) ACS Nano 3:1913–1921
59. Pietrzyk P, Zasada F, Piskorz W, Kotarba A, Sojka Z (2007) Catal Today 119:219–227
60. Pietrzyk P, Gil B, Sojka Z (2007) Catal Today 126:103–111
61. Hadjiivanov K (2000) Catal Rev Sci Eng 42:71–144
62. Dean JA (1992) Lange's Chemistry Handbook, 15th edn. New York, McGraw-Hill
63. Jhi SH, Louie SG, Cohen ML (2000) Phys Rev Lett 85:1710–1713
64. Parr RG, Yang W (1989) Density-functional theory of atoms and molecules. Oxford University Press, New York

J Mol Model (2012) 18:2055–2064
DOI 10.1007/s00894-011-1224-z

ORIGINAL PAPER

Molecular modeling of *Trypanosoma cruzi* glutamate cysteine ligase and investigation of its interactions with glutathione

Carlos F. Lagos · Raul Araya-Secchi · Pablo Thomas · Tomás Pérez-Acle · Ricardo A. Tapia · Cristian O. Salas

Received: 17 June 2011 / Accepted: 11 August 2011 / Published online: 6 September 2011

Abstract *Trypanosoma cruzi* glutamate cysteine ligase (TcGCL) is considered a potential drug target to develop novel antichagasic drugs. We have used a variety of computational methods to investigate the interactions between TcGCL with Glutathione (GSH). The three-dimensional structure of TcGCL was constructed by comparative modeling methods using the *Saccharomyces cerevisiae* glutamate cysteine ligase as template. Molecular dynamics simulations were used to validate the TcGCL model and to analyze the molecular interactions with GSH. Using RMSD clustering, the most prevalent GSH binding modes were identified paying attention to the residues involved in the molecular interactions. The GSH binding modes were used to propose pharmacophore models that can be exploited in further studies to identify novel antichagasic compounds.

Keywords Comparative modeling · Glutamate cysteine ligase · Molecular dynamics · Pharmacophore · *Trypanosoma cruzi*

Introduction

Chagas disease represents the leading cause of cardiac lesions in young, economically productive adults in Latin American countries where this disease is endemic [1]. *Trypanosoma cruzi*, the eukaryotic protozoan responsible for Chagas disease, has a redox metabolism based on trypanothione, a glutathionyl spermidine derivative. In *Trypanosoma cruzi*, glutathione (GSH) is synthesized from its constituent amino acids by the consecutive actions of

Carlos F. Lagos and Raul Araya-Secchi have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1224-z) contains supplementary material, which is available to authorized users.

C. F. Lagos
Departamento de Farmacia, Facultad de Quimica,
P. Universidad Catolica de Chile,
Av Vicuña Mackenna 4860,
Macul-Santiago, Chile

C. F. Lagos · R. Araya-Secchi
Departamento de Ciencias Biologicas, Facultad de Ciencias
Biologicas, Universidad Andres Bello,
Av Republica 217,
Santiago, Chile

R. Araya-Secchi · T. Pérez-Acle
Computational Biology Lab (DLab), Centro de Modelamiento
Matematico (CMM), Facultad de Ciencias Fisicas y Matematicas,
Universidad de Chile,
Santiago, Chile

T. Pérez-Acle
Fundacion Ciencia para la Vida,
Zañartu 1482, Ñuñoa Santiago, Chile

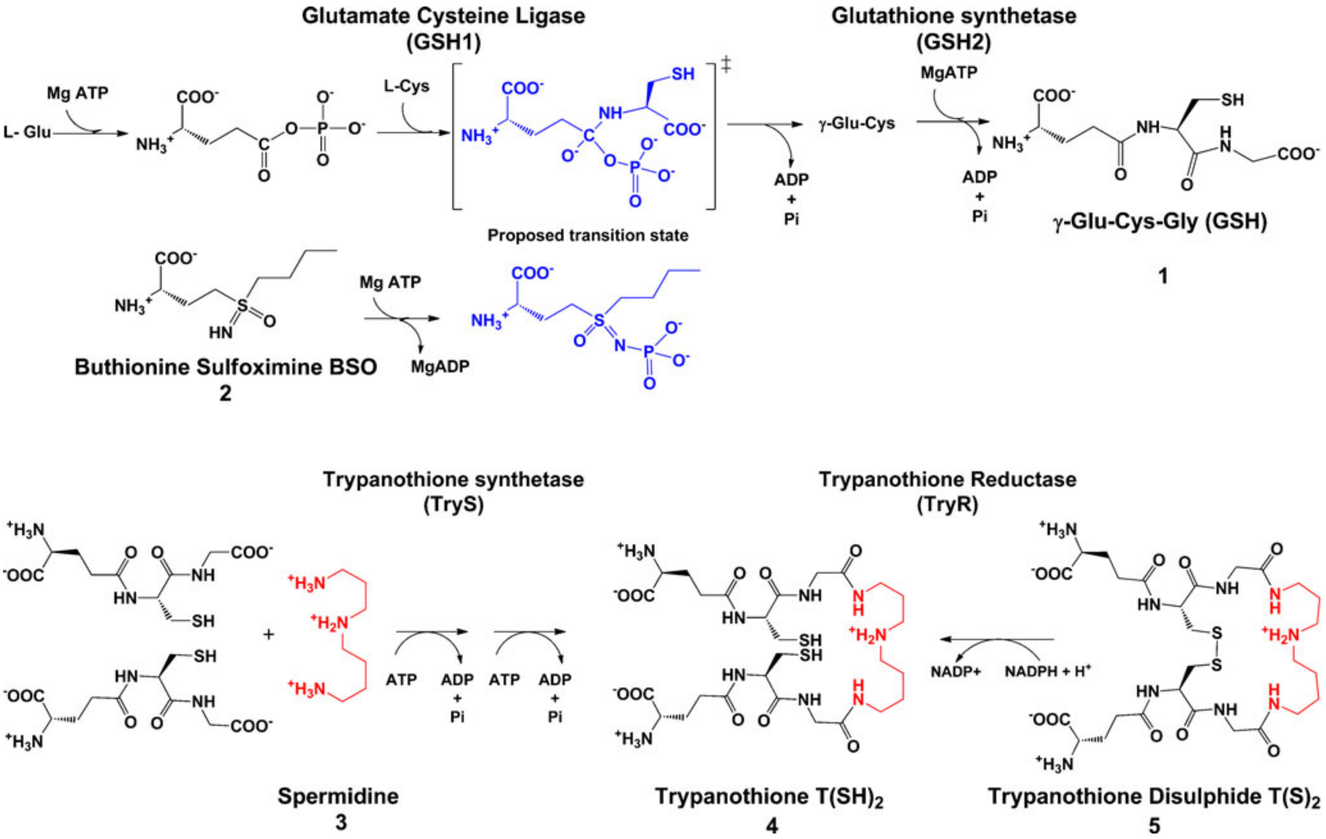
T. Pérez-Acle
Centro Interdisciplinario de Neurociencias de Valparaiso (CINV),
Pasaje Harrington 287, Playa Ancha Valparaiso, Chile

P. Thomas · R. A. Tapia · C. O. Salas (✉)
Departamento de Quimica Organica, Facultad de Quimica,
P. Universidad Catolica de Chile,
Av Vicuña Mackenna 4860,
Macul- Santiago, Chile
e-mail: cosalas@uc.cl

two related ATP-dependent peptide ligases (Scheme 1): i) glutamate cysteine ligase (GCL or GSH1, EC 6.3.2.2) and ii) glutathione synthetase (GSH2, EC 6.3.2.3) [2, 3]. The GCL reaction is rate limiting and essential for the parasite as shown by RNAi experiments [4]. TcGCL activity is precisely controlled by non-allosteric feedback inhibition by glutathione, the limited availability of cellular L-Cys and the transcriptional and post-transcriptional regulation of the enzyme's expression under various physiological conditions [5, 6]. Generated glutathione, is then conjugated with spermidine by trypanothione synthetase (TryS, EC 6.1.1.9) to form trypanothione (T(SH)2), the central thiol that delivers electrons for the synthesis of DNA precursors, the detoxification of hydroperoxides and other trypanothione-dependent pathways [7–9].

GCLs sequences can be classified in three groups: i) sequences primarily from gamma-proteobacteria; ii) sequences from non-plant eukaryotes; and iii) sequences primarily from alpha-proteobacteria and plants. Although sequence identities between groups are insignificant, some conserved sequence motifs are found, suggesting distant phylogenetic relationship [10]. Recently, the crystal structures of *Saccharomyces cerevisiae* GCL (ScGCL) in complex with BSO and GSH were solved at 2.20 Å and 2.50 Å resolution (pdb ids 3LVV and 3LVW, respectively) [11]. Despite their low sequence identity (<10 %), ScGCL shares significant structural similarity with *Brassica juncea* (BjGCL) and *Escherichia coli* GCL (EcGCL). Both proteins were solved in complex with BSO (pdb id 2GWC and pdb id 1VA6, respectively) [12–14]. All these enzymes provide a source of structural insights to identify main protein-ligand interactions for buthionine sulfoximine (BSO) and other analogs. In all these cases, the inhibitors bind on the bottom of the catalytic domain comprised by six anti-parallel β-strands that form a partial barrel with a funnel-like shaped internal cavity (Fig. 1). The sequence of the small variable domain changes widely among GCL family members; sequence analysis revealed that there are no conserved regions corresponding to these sequences among family members of mechanistically related glutamine synthetase [14].

In light of its central role in the essential glutathione and trypanothione metabolism, GCL has been studied as a target for the design and identification of novel analogs of BSO [15–18], an inhibitor of GCL and effective GSH-depleting agent, which have shown to prolongs survival of mice infected with *Trypanosoma brucei* and increases the trypanocidal activity



Scheme 1 Biosynthetic pathways for Glutathione and Trypanothione in *Trypanosoma cruzi*

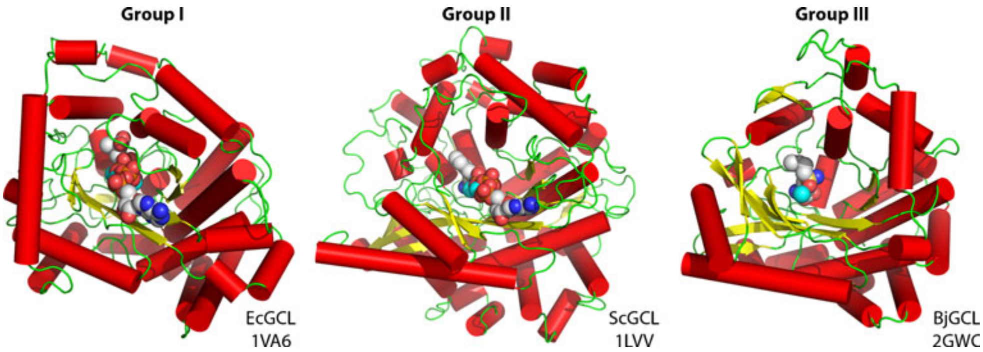


Fig. 1 Topology of GCLs proteins. Class I GCLs are mainly bacterial, class II eukaryotic and class III mostly plant homologues. Co-crystallized ligands and cofactor metals are shown as spheres for comparison

of Nifurtimox and Benznidazole against *Trypanosoma cruzi* [19–22]. Unfortunately, structure-activity relationships considering GSH or sulfoximine-based transition state analogues have not been assessed for *T. cruzi* GCL. However, some efforts have been conducted for GCL homologues present in *E. coli*, *T. brucei*, rat and human [23–29].

In this work, we report the comparative modeling of TcGCL and the investigation of its interaction with GSH, the identification of the preferred binding modes, and the development of structure-based pharmacophore models suitable for the design and identification of novel chemical entities targeting TcGCL.

Materials and methods

Molecular modeling

The molecular model of TcGCL was constructed using Modeller [30] as implemented in the Protein Modeling module of Discovery Studio 2.1 (Accelrys Inc., San Diego, CA). The sequence of TcGCL (Uniprot entry O77252) [31] was aligned with that of the class II *Saccharomyces cerevisiae* GCL in complex with GSH, solved at 2.5Å of resolution (pdb id 3LVW and Uniprot entry P32477) [11]. Due to the lack of an appropriate template, residues 214–283 of TcGCL were not included in the model. The copy ligand function was used to model the crystallographic position of GSH. Secondary structure restrains were applied in the following segments according to PCI-SS secondary structure prediction [32]: W13-T40 helix, K81-D86 helix, S113-S130 helix, S227-A238 helix, P444-K463 helix, Y492-R496 strand, P507-S511 strand, K522-V533 helix, and E587-E623 helix. A total of 100 models were constructed and the best model according to Modeller internal DOPE score was subjected to a loop refinement protocol that was applied to the P240-H248 zone. Twenty-five different loop conformations were constructed and the best-generated loop variant model according to CHARMM energies was subjected to a molecular minimization protocol using the CHARMM 22 force-field [33]. The protocol consisted of 5000 steps of steepest descent method followed by 10,000 steps of conjugate gradient method to reach a final root-mean square (RMS) gradient of 0.001 kcal mol^{-1} $Å^{-1}$. The overall quality of the final model was assessed by Ramachandran plot analysis using the RAMPAGE server [34] and Profiles-3D analysis [35]. Additional quality model assessments were performed using the ProSA-web [36], QProt [37] and SAVES (http://nihserver.mbi.ucla.edu/) servers. Binding site search was performed in Discovery Studio 2.1 with default parameters. DelPhi software was used to calculate the spatial distribution of electrostatic potential on protein atoms, using a two-dielectric implicit solvent model and the finite difference method to solve the Poisson-Boltzmann equation. The dielectric constant used for protein was 2 and 80 for the solvent [38].

Molecular dynamics (MD) simulation

The simulated periodic cells were constructed using VMD v1.9 [39] and comprised the TcGCL model alone and the TcGCL model plus the GSH molecule. The models were solvated in a water box, keeping at least 18 Å between every protein atom and the cell boundaries. Both systems were neutralized by randomly placing 16 and 17 Na^+ ions for TcGCL and TcGCL-GSH respectively. The final dimensions of the periodic cells were 95x106x91 Å comprising a total of ~86,300 atoms. The systems were minimized and subjected to MD for 0.2 ns with the protein fixed. The proteins were then released keeping alpha carbons and side-chain non-hydrogen atoms constrained with a force constant of 20 and 5 kcal mol^{-1} $Å^{-2}$ respectively. The full systems were then minimized and a slow relaxation procedure was performed in which the constraint applied to alpha carbon atoms and side-chain non-hydrogen atoms of the proteins were decreased at a rate of 0.5 kcal mol^{-1} ps^{-1} until no constraints were applied. Subsequently, 12 ns of unconstrained NPT-MD simulation were performed with the first 2 ns considered as equilibration and the last 10 ns considered for analysis. The MD program NAMD [40] with CHARMM22 force field corrected by CMAP for proteins [41, 42] and TIP3P for water were used for the simulation [43]. Periodic boundary conditions were imposed and the particle mesh Ewald method [44, 45] was used for electrostatic forces calculation. Constant temperature (300K) and pressure (1 atm) were maintained by using Langevin dynamics [46]. The SHAKE algorithm [47] was applied to constrain the lengths of all bonds that involve hydrogen allowing the use of a 2 fs integration timestep. MD trajectory analyses were performed with VMD v1.9 to calculate the Cα root-mean square deviation (Cα−RMSD) and the residue wise Cα root-mean square fluctuation (Cα−RMSF), for each system.

RMSD clustering and binding mode selection

To generate a reduced set of structures that represent the dynamical behavior of the GSH binding site of TcGCL, a root-mean-square deviation (RMSD) conformational clustering was performed to reduce the structural redundancy in the MD ensemble. Two hundred receptor conformations were extracted from the 10 ns MD trajectory, one every 50 ps. The structures of the trajectory were superimposed using all α-carbon atoms to remove overall rotation and translation in order that subsequent RMSD calculations could focus on the internal conformational variability of the protein. A hierarchical clustering procedure was performed with the WORDOM (v0.21) software [48] based on a subset of 15 residues located within a 5 Å radius from the center of mass of the GSH molecule during the MD-simulation (Residues: 55 93 94 179 180 183 256 260 262 264 311 335 412 and 415). These residues were grouped into clusters of similar configurations using the atom-positional RMSD for all atoms (including side chains and hydrogen) as the similarity criterion. A cutoff of 1.5 Å was chosen. The central member within each cluster, i.e., the structure having the smallest RMSD to all other structures, was chosen as the representative structure.

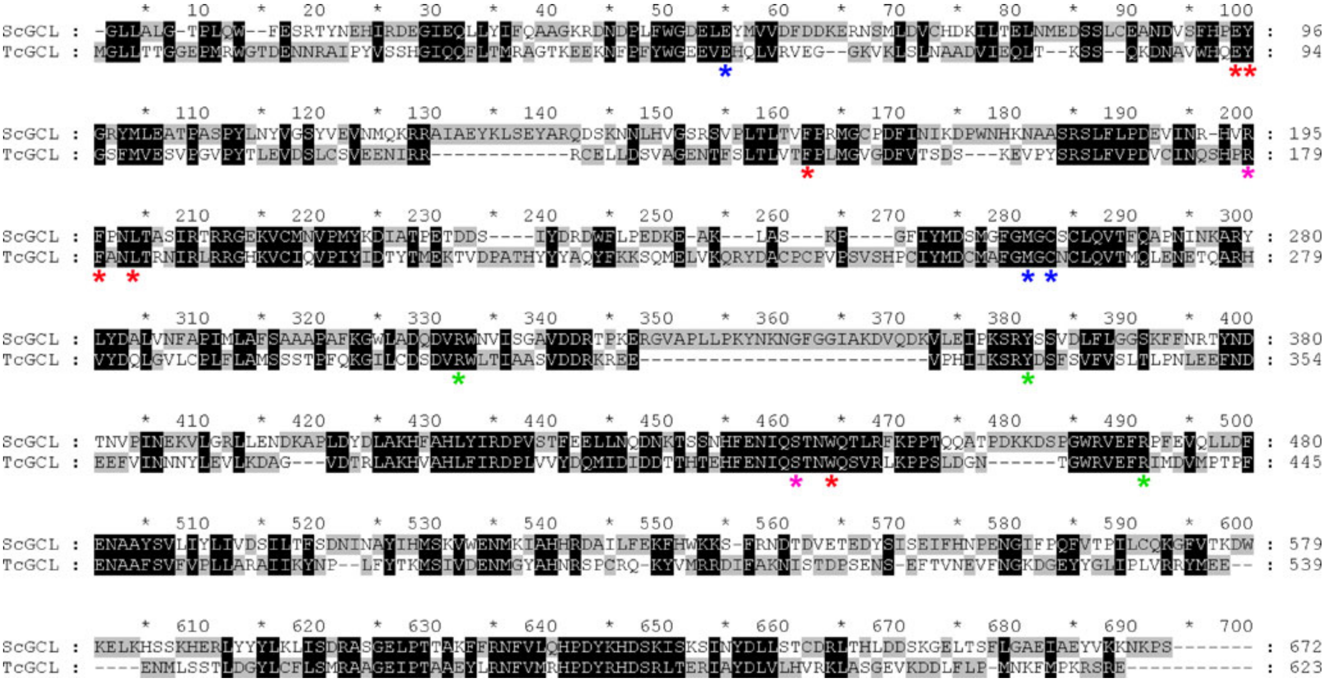
Binding energy calculations and pharmacophore hypotheses generation

Each cluster representative structure was minimized to convergence using 5000 steps of steepest descent method followed by 10,000 steps of conjugate gradient method to reach a final root-mean square (RMS) gradient of 0.001 kcal mol^{-1} Å. The Ludi 2 empirical scoring function was used to estimate the binding energy and the individual energy descriptors that contribute to the score. For each minimized structure, a structure-based pharmacophore hypothesis was generated with LigandScout v3.01 (Inteligand GmbH, Vienna, Austria). Chemical features perception was performed using default parameters.

Results and discussion

Comparative modeling

The *Saccharomyces cerevisiae* glutamate cysteine ligase ScGCL in complex with glutathione (3LVW) was identified as a suitable template for modeling using the sequence search facility in the Protein Data Bank [49]. For residues 214–283 of TcGCL, no suitable template structure could be identified. The initial alignment obtained from the PDB-BLAST search was manually corrected by inserting or removing gaps. After several rounds of comparative modeling, where we evaluated the impact of the alignment corrections on the secondary structure, a set of constrains were used to preserve the length of predicted secondary structure elements (see methods). The final alignment between TcGCL and ScGCL shows a 32.7% and 52.2% sequence identity and sequence similarity respectively (Fig. 2). According to STRIDE web-server [50], the secondary structure is composed of 18 α-helices, 12 extended β-sheets and two 3–10 helices. In contrast with other GCLS such as EcGCL (1 disulfide bridge) and BjGCL (3 disulfide bridges), ScGCL has no disulfide bridges. It is important to note that *Trypanosoma brucei* GCL, the best-characterized trypanosome GCL, has no reported disulfides bridges. For these reasons, no disulfide bridges were included in our model. Although DiANNA server [51] predicts 2 disulfide bridges between cysteines 72–196 (score 0.95909) and 241–255 (Score 0.82679), analysis of TcGCL model shows that the first pair is far away in the structure (distance ~18 Å) and the second pair is not present in our model, which lacks the portion containing these residues. The model is in agreement with mutagenesis data for *Trypanosoma brucei* GCL (TbGCL), to which TcGCL has a 67.7% and 81.6% sequence identity and sequence similarity respectively [2, 24] (Fig. S1). The resulting TcGCL minimized model was superimposed with ScGCL and a root-mean square deviation (RMSD) of 0.8 Å based on alpha carbon atoms (Cα-RMSD) of 600 equivalent residues was obtained (Fig. 3a). The final TcGCL structure contains 623 residues, of which 97.3% of them were found in allowed region and 2.7% in the outlier region of the Ramachandran plot according to the RAMPAGE server evaluation (Fig. 3b). The Profile-3D score was also computed to measure the compatibility of the protein model with its sequence, using a 3D profile. A ten residues window size for smoothing and the Kabsch-Sander secondary structure were used. A compatibility score of 203.35 with a maximum expected score of 286.07 was obtained for the model. These values compare well with the profile and score of 277.88 for the template structure ScGCL (Fig. 3c). The results obtained from the SAVES server for the Ramachandran Plot



```
          *    10      *    20      *    30      *    40      *    50      *    60      *    70      *    80      *    90      *   100
ScGCL : -GLLALG-TPLQW--FESRTYNEHIRDEGIEQLLYIFQAAGKRDNDPLFWGDELEYMVVDFDDKERNSMLDVCHDKILTELNMEDSSLCEANDVSFHPEY :  96
TcGCL : MGLLTTGGEPMRWGTDENNRAIPYVSSHGIQQFLTMRAGTKEEKNFPFYWGEEVEHQLVRVEG--GKVKLSLNAADVIEQLT--KSS--QKDNAVWHQEY :  94

          *   110      *   120      *   130      *   140      *   150      *   160      *   170      *   180      *   190      *   200
ScGCL : GRYMLEATPASPYLNYVGSYVEVNMQKRRAIAEYKLSEYARQDSKNNLHVGSRSVPLTLTVFFRMGCPDFINIKDPWNHKNAASRSLFLPDEVINR-HVR : 195
TcGCL : GSFMVESVPGVPYTLEVDSLCSVEENIRR------------RCELLDSVAGENTFSLTLVTFPLMGVGDFVTSDS---KEVPYSRSLFVPDVCINQSHPR : 179

          *   210      *   220      *   230      *   240      *   250      *   260      *   270      *   280      *   290      *   300
ScGCL : FPNLTASIRTRRGEKVCMNVPMYKDIATPETDDS----IYDRDWFLPEDKE-AK---LAS---KP----GFIYMDSMGFGMGCSCLQVTFQAPNINKARY : 280
TcGCL : FANLTRNIRLRRGHKVCIQVPIYIDTYTMEKTVDPATHYYYAQYFKKSQMELVKQRYDACPCPVPSVSHPCIYMDCMAFGMGCNCLQVTMQLENETQARH : 279

          *   310      *   320      *   330      *   340      *   350      *   360      *   370      *   380      *   390      *   400
ScGCL : LYDALVNFAPIMLAFSAAAPAFKGWLADQDVRWNVISGAVDDRTPKERGVAPLLPKYNKNGFGGIAKDVQDKVLEIPKSRYSSVDLFLGGSKFFNRTYND : 380
TcGCL : VYDQLGVLCPLFLAMSSSTPFQKGILCDSDVRWLTIAASVDDRKREE------------------------VPHIIKSRYDSFSVFVSLTLPNLEEFND : 354

          *   410      *   420      *   430      *   440      *   450      *   460      *   470      *   480      *   490      *   500
ScGCL : TNVPINEKVLGRLLENDKAPLDYDLAKHFAHLYIRDPVSTFEELLNQDNKTSSNHFENIQSTNWQTLRFKPPTQQATPDKKDSPGWRVEFRPFEVQLLDF : 480
TcGCL : EEFVINNNYLEVLKDAG---VDTRLAKHVAHLFIRDPLVVYDQMIDIDDTTHTEHFENIQSTNWQSVRLKPPSLDGN------TGWRVEFRIMDVMPTPF : 445

          *   510      *   520      *   530      *   540      *   550      *   560      *   570      *   580      *   590      *   600
ScGCL : ENAAYSVLIYLIVDSILTFSDNINAYIHMSKVWENMKIAHHRDAILFEKFHWKKS-FRNDTDVETEDYSISEIFHNPENGIFPQFVTPILCQKGFVTKDW : 579
TcGCL : ENAAFSVFVPLLARAIIKYNP--LFYTKMSIVDENMGYAHNRSPCRQ-KYVMRRDIFAKNISTDPSENS-EFTVNEVFNGKDGEYYGLIPLVRRYMEE-- : 539

          *   610      *   620      *   630      *   640      *   650      *   660      *   670      *   680      *   690      *   700
ScGCL : KELKHSSKHERLYYYLKLISDRASGELPTTAKFFRNFVLQHPDYKHDSKISKSINYDLLSTCDRLTHLDDSKGELTSFLGAEIAEYVKKNKPS------- : 672
TcGCL : ----ENMLSSTLDGYLCFLSMRAAGEIPTAAEYLRNFVMRHPDYRHDSRLTERIAYDLVLHVRKLASGEVKDDLFLP-MNKFMPKRSRE----------- : 623
```

Fig. 2 Sequence alignment between TcGCL and template ScGCL proteins. Residues interacting with GSH are marked with blue and red, green and magenta asterisks for Glu-COO^{-}, Glu-NH$_3$$^{+}$, Cys and Gly binding sites respectively

(PROCHECK) and Verify 3D profiles using a window size of 21 residues for smoothing compared well with those obtained with Rampage and Profile3D respectively (Fig. S2). The ProSA-web server quality assessment provides a Z-score of −10.73 and −8.43 for the template and model structure respectively (Fig. S3). This result indicates that the TcGCL

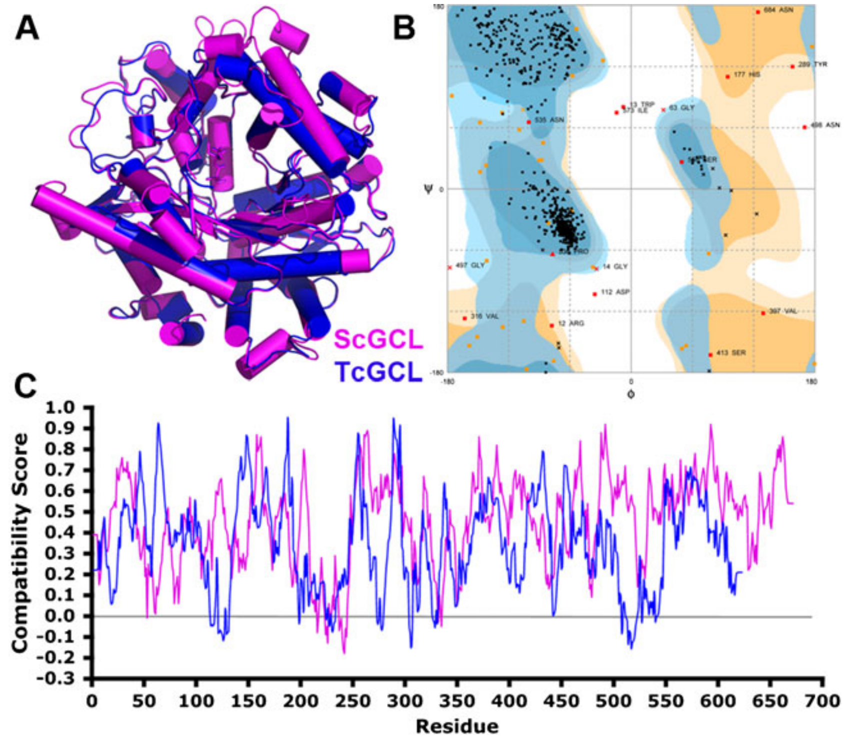


Fig. 3 Comparative modeling of TcGCL (**a**) Structural superposition between TcGCL model and template. (**b**) Ramachandran plot from RAMPAGE, (**c**) Profiles-3D plot of TcGCL model and template structure

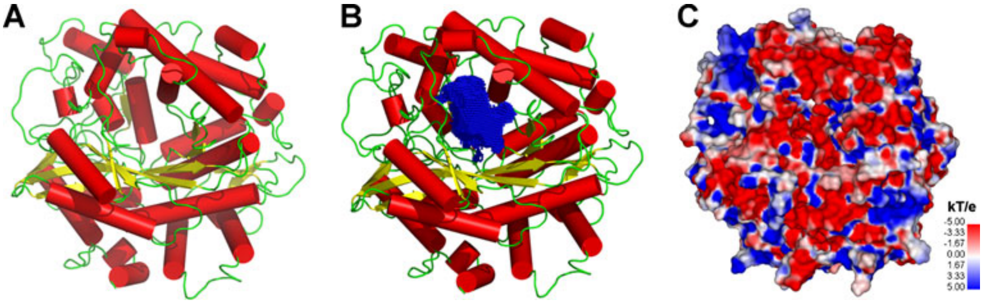


Fig. 4 From left to right. Topology, binding site cavity and electrostatic potential surface of the obtained TcGCL model

model is of similar quality as equivalent sized X-ray structures. The ProQ neural network provided an LGscore of 5.726 and 6.464 for the TcGCL model and the template structure respectively. This method considers that models with LGscore>4 are extremely good. All these results suggest that the fold of the TcGCL model is reliable and the model is suitable for further analysis and studies. Figure 4a shows the 3D-structure of the minimized TcGCL-GSH model. A 726 Å^3 cavity was identified with the binding site analysis module of DS2.1 (Fig. 4b) and a highly electronegative

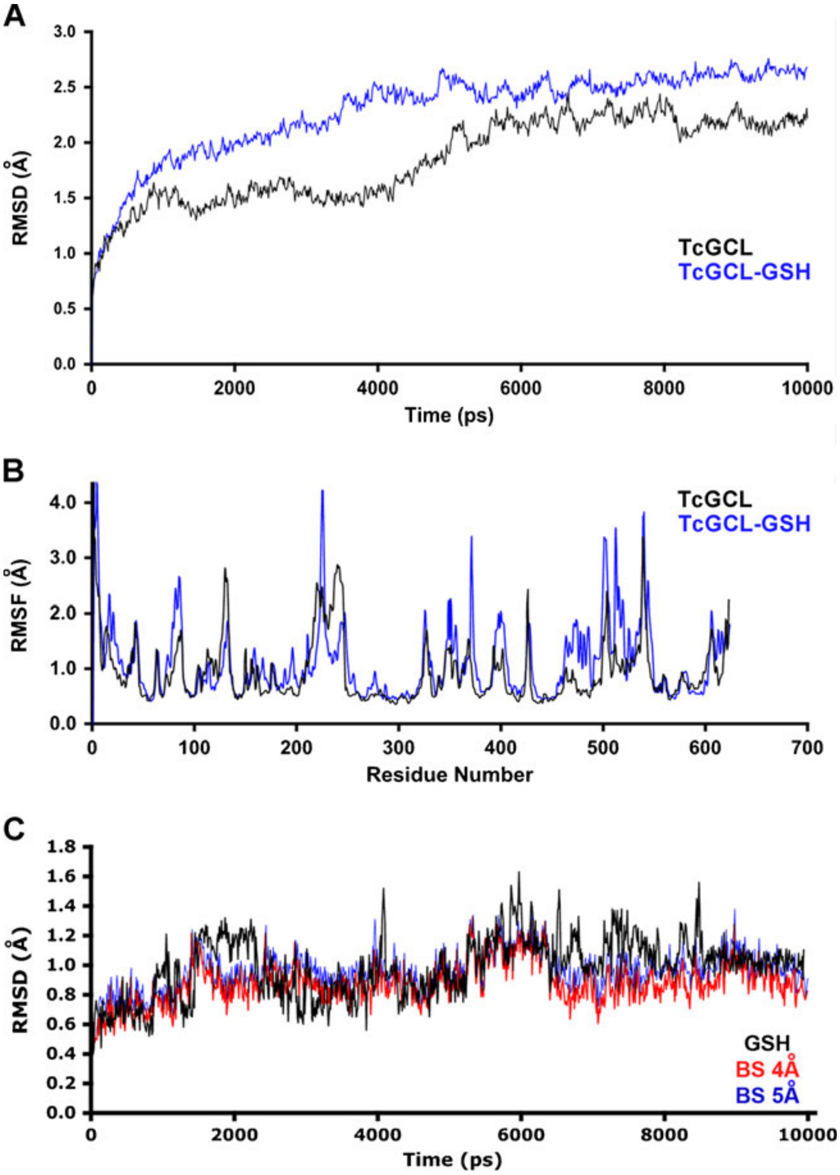


Fig. 5 (**a**) Cα-RMSD of TcGCL and TcGCL-GSH complex during the MD, (**b**) Cα-RMSF of TcGCL and TcGCL-GSH complex during the MD. (**c**) RMSD of GSH, and α-carbons from residues within 4 and 5 Å from GSH during MD

GSH binding site was identified by the Delphi spectrum (Fig. 4c).

Molecular dynamics simulations

TcGCL is thought to be biologically functional as a monomer [52], and can be feedback inhibited by GSH. In order to study the TcGCL-GSH complex stability, a 10 ns molecular dynamics simulation was performed. Figure 5a shows the alpha carbon RMSD plot using the first frame of the production stage of the simulation as reference for the TcGCL-GSH and using the TcGCL alone as a control. The proteins reach a stable equilibrated state after 4 ns of simulation, with an average Cα-RMSD value of 2.27±0.38 Å. The Cα root mean square fluctuation (RMSF) plot for each residue during the simulation was calculated to identify the most mobile residues, all of which are located far from the GSH binding site (Fig. 5b). The calculated RMSDs for GSH (average 0.99±0.20 Å) and for the residues surrounding it at 4 and 5Å (average 0.89±0.14 and 0.96±0.13 Å, respectively) show that the binding site is fairly rigid but some movement can be detected. As expected, correlated deviations are observed between GSH and the binding site (Fig. 5c). To further characterize this behavior, the binding pattern between GSH and relevant conserved binding site residues was followed during simulation and analyzed. The H-bond pattern for which distances were measured is shown in Fig. 6a. As shown in Fig. 6b, for the glutamate carboxylate group or motif of GSH (Glu:COO^-) distances between R381:NE/Glu:OT2, Y405:OH/Glu:OT1 and R506:NH2/Glu:OT1 are within hydrogen bond distance, except for the C334:SG/Glu:OT2 interaction, with average values of 2.82±0.13, 2.60±0.09, 2.73±0.11 and 4.09±0.36 Å, respectively. This result is in agreement with mutagenesis data that shows that the equivalent residue in TbGCL enzyme (C319) has no significant effect on the specific activity of the enzyme [53]. For the charged amino moiety of GSH (Glu:NH_3^+), the distance to E55:OE1, C332:O and M330:O are within hydrogen bond distances with average values of 2.68±0.09, 2.74±0.12 and 3.00±0.26 Å, respectively. In particular the interaction with the main-chain carbonyl of M330 is non-water mediated in contrast to the corresponding interaction with the M262 residue in ScGCL. For the cysteine residue of GSH distances E93:OE1/Cys:NH, E93:OE2/Cys:NH and W485:NE1/Cys:O are within hydrogen bond distance, with average values of 3.32±0.28, 2.84±0.17 and 3.20±0.37 Å, respectively. Finally, for the glycine residue of GSH the measured distances E93:OE2/Gly:NH, R179:NH/Gly:OT2, R179:NH2/Gly:OT1 and R179:CZ/Gly:C, with average distances of 3.17±0.52, 3.63±1.00, 4.04±1.21 and 4.31±0.69 Å, respectively, are consistent with a hydrogen bond with E93, and a transient double salt bridge between R179 and the terminal carboxylate observed around 1.5-2.5 ns of

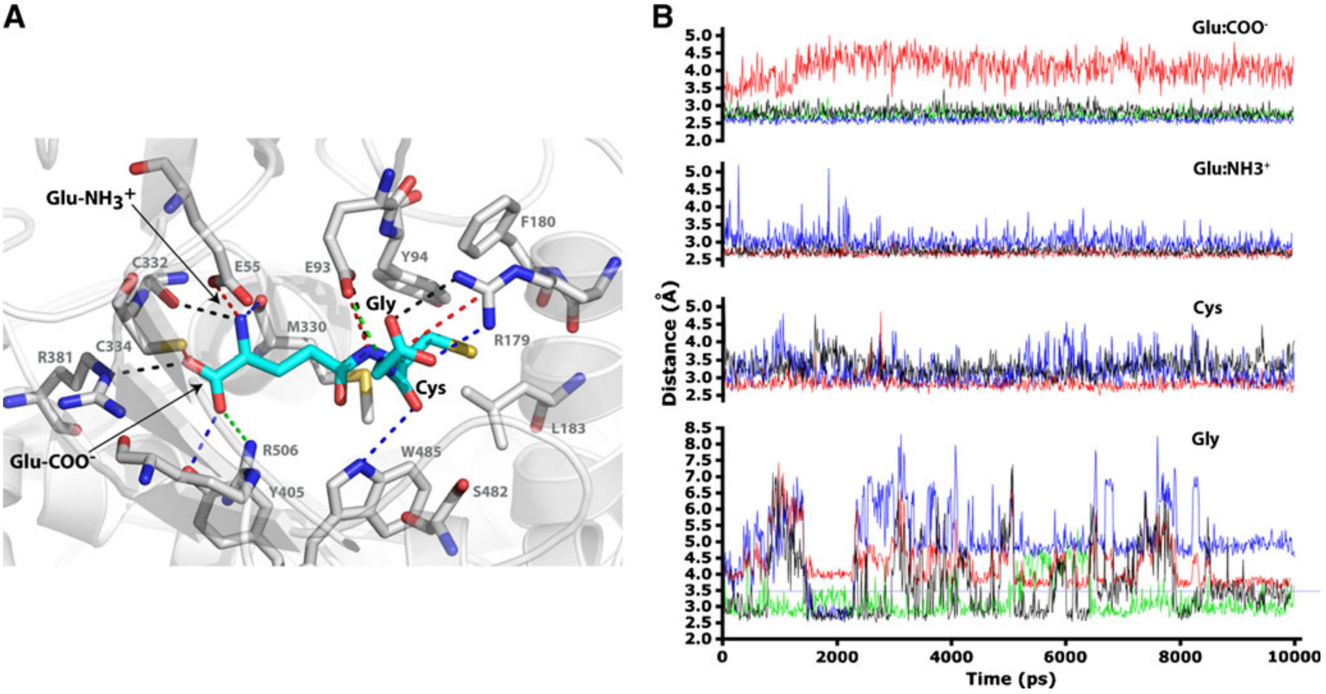


Fig. 6 (**a**) Snapshot of the averaged TcGCL-GSH complex during simulation, water molecules are not shown for clarity. (**b**) Binding site H-bond interaction distances between several GSH atoms and the Glu, Cys and Gly binding sites residues during MD. Distance colors are shown as in panel A for each zone

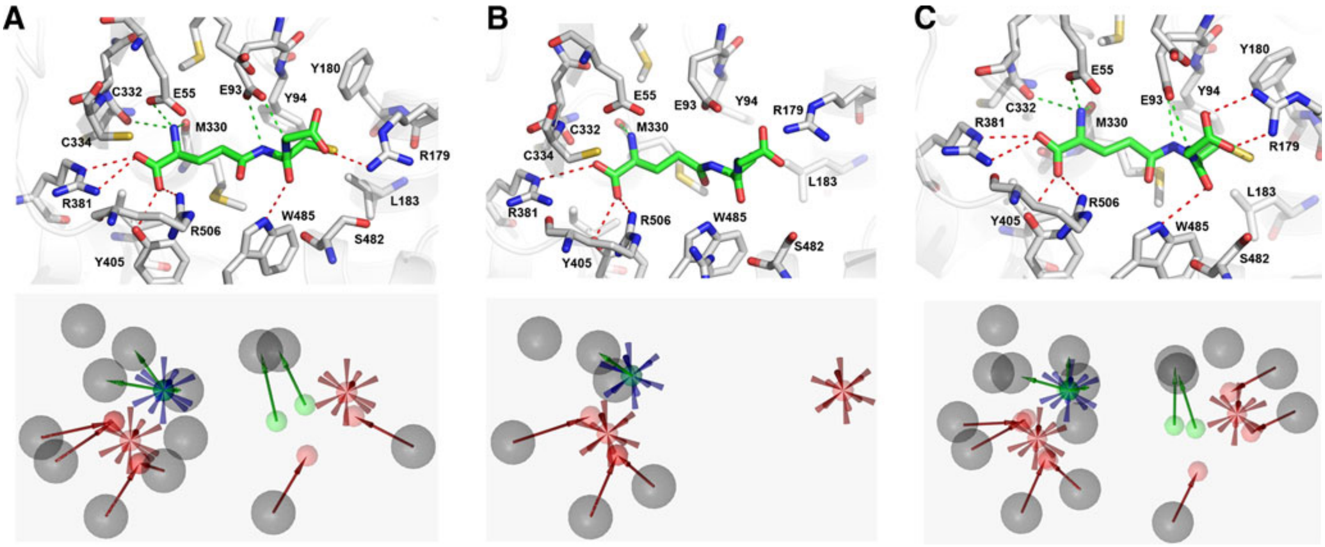


Fig. 7 Representative binding modes for GSH during simulation obtained by RMSD clustering of the binding site residues and their corresponding pharmacophore models for GSH interaction binding mode. (**a**) Cluster 1 (24%), (**b**) Cluster 2 (7%), and (**c**) Cluster 3 (69%). Gray exclusion spheres represent the sterical circumference of the protein, red vectors are H-bond acceptor and green vectors H-bond donors. Spiked red and blue spheres are negative and positive ionizable groups

simulation, which can be explained by the high flexibility of the terminal glycine residue of GSH. These results reveal that the GSH binding site has a rigid and a flexible zone. The rigid zone comprises the glutamate zwitterion binding residues and the cysteine counterparts, and the flexible zone is mainly located within the terminal carboxy group of glycine.

RMSD clustering and binding mode selection

To identify the preferred GSH binding modes suggested by trajectory analysis, an RMSD conformational clustering based on the GSH binding site was performed on a set of 200 snapshots separated by 50 ps taken from the trajectory, in order to reduce structural redundancy in the MD ensemble. Using a 1.5 Å cut-off, we identified three clusters from which the structure, from each cluster having the smallest RMSD to all other structures within the cluster was selected as the representative structure as shown in Fig. 7. Cluster 1 (Fig. 7a) contains 48 structures and represents a 24% of occurrence during the MD, cluster 2 (Fig. 7b) contains only 14 structures representing the 7%, and cluster 3 (Fig. 7c) contains 138 structures and account for the last 69%.

Binding energy calculations and pharmacophore hypotheses generation

Individual components of the binding energy estimated using the Ludi 2 empirical scoring function for each representative structure show that ionic, lipophilic, and H-bond interactions display similar energy contribution (Table 1). A comparison of the structure based pharmacophore models suggest that drop in binding affinity can be explained by the loss of key H-bond and ionic interactions between GSH and TcGCL residues E55, E93, R179, C332 and W485. Several of these residues have been reported as essential for the binding and stabilization of metal and L-Glu positioning for further metal-dependent complexation with L-Cys [2]. In particular Glu93, a metal binding determinant, plays an important role by anchoring both GSH-Cys and GSH-Gly amino groups, making this residue less available to metal binding. Aliphatic interactions between the carbon chain of GSH-Glu and the side-chain of I385, and the interactions of the GSH-Cys side-chain with the Cys binding site composed of the side-chain of residues L173, R179, F180, M330 and W485, appears to account for the Lipo component in binding energy. R179 is highly conserved between GCLs and R179 to Ala alters the

Table 1 Energy contribution of GSH binding modes clusters, according to Ludi 2 scoring function

Cluster	ΔG (kcal/mol)	ΔG H-bond	ΔG Lipo	ΔG Ionic	ΔG Rot
1	−8.237	−2.550	−3.137	−3.832	1.964
2	−5.100	−1.146	−2.946	−2.291	1.964
3	−8.673	−3.327	−3.027	−3.600	1.964

active site and effects the substrate dependencies [24]. In our model R179 plays a role highly relevant for GSH binding, acting as a lid that allow the transition between a high and a low affinity state through the formation of a double salt-bridge with the carboxylate group of GSH-Gly. Thus, targeting the residues involved in these relevant interactions seems a reasonable strategy for the development of GCLs modulators.

Conclusions

In this study, we present the first molecular model for *Trypanosoma cruzi* glutamate cysteine ligase TcGCL. Analysis of the results of the comparative modeling procedure and MD simulations indicated that the theoretical predictions and obtained fold is consistent with the known set of experimental results available for *Trypanosoma* GCL and other homologues. Molecular dynamics simulations and binding energy contribution analyses highlight the relevant forces involved in the GSH binding, and identify E93 and R179 with putative roles as anchor and switch key residues that could explain the differential binding mode of GSH within the TcGCL active site. The prevalent binding modes and their pharmacophore-derived models provide a source for structure-based design of new GCL inhibitors.

Acknowledgments The authors would like to thank Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT, Chile) project Nº 11085027 to COS for financial support. RAT acknowledges Programa Bicentenario de Ciencia y Tecnología del Proyecto de Inserción (PSD70). TPA acknowledge to Proyectos de Financiamiento Basal PFB03 and PFB16, and to Millenium Institute Centro Interdisciplinario de Neurociencias de Valparaíso (CINV). Molecular dynamics simulations were performed at the National Laboratory for High Performance Computing (NLHPC ECM-02) supercomputing infrastructure: Powered@NLHPC. CFL, RAS and PT are PhD fellows from Comisión Nacional de Investigación Científica y Tecnológica (CONICYT, Chile).

References

1. Moncayo A, Silveira AC (2009) Current epidemiological trends for Chagas disease in Latin America and future challenges in epidemiology, surveillance and health policy. Mem Inst Oswaldo Cruz 104(Suppl 1):17–30
2. Abbott JJ, Pei J, Ford JL, Qi Y, Grishin VN, Pitcher LA, Phillips MA, Grishin NV (2001) Structure prediction and active site analysis of the metal binding determinants in gamma -glutamylcysteine synthetase. J Biol Chem 276:42099–42107
3. Brekken DL, Phillips MA (1998) Trypanosoma brucei gamma-glutamylcysteine synthetase. Characterization of the kinetic mechanism and the role of Cys-319 in cystamine inactivation. J Biol Chem 273:26317–26322
4. Huynh TT, Huynh VT, Harmon MA, Phillips MA (2003) Gene Knockdown of γ-Glutamylcysteine Synthetase by RNAi in the Parasitic Protozoa Trypanosoma brucei Demonstrates That It Is an Essential Enzyme. J Biol Chem 278:39794–39800
5. Krzywanski DM, Dickinson DA, Iles KE, Wigley AF, Franklin CC, Liu RM, Kavanagh TJ, Forman HJ (2004) Variable regulation of glutamate cysteine ligase subunit proteins affects glutathione biosynthesis in response to oxidative stress. Arch Biochem Biophys 423:116–125
6. Hiratake J (2005) Enzyme inhibitors as chemical tools to study enzyme catalysis: rational design, synthesis, and applications. Chem Rec 5:209–228
7. Ariyanayagam MR, Oza SL, Mehlert A, Fairlamb AH (2003) Bis (glutathionyl)spermine and other novel trypanothione analogues in Trypanosoma cruzi. J Biol Chem 278:27612–27619
8. Oza SL, Tetaud E, Ariyanayagam MR, Warnon SS, Fairlamb AH (2002) A single enzyme catalyses formation of Trypanothione from glutathione and spermidine in Trypanosoma cruzi. J Biol Chem 277:35853–35861
9. Wyllie S, Oza SL, Patterson S, Spinks D, Thompson S, Fairlamb AH (2009) Dissecting the essentiality of the bifunctional trypanothione synthetase-amidase in Trypanosoma brucei using chemical and genetic methods. Mol Microbiol 74:529–540
10. Copley SD, Dhillon JK (2002) Lateral gene transfer and parallel evolution in the history of glutathione biosynthesis genes. Genome Biol 3:research0025
11. Biterova EI, Barycki JJ (2010) Structural basis for feedback and pharmacological inhibition of Saccharomyces cerevisiae glutamate cysteine ligase. J Biol Chem 285:14459–14466
12. May MJ, Leaver CJ (1994) Arabidopsis thaliana gamma-glutamylcysteine synthetase is structurally unrelated to mammalian, yeast, and Escherichia coli homologs. Proc Natl Acad Sci USA 91:10059–10063
13. Hothorn M, Wachter A, Gromes R, Stuwe T, Rausch T, Scheffzek K (2006) Structural basis for the redox control of plant glutamate cysteine ligase. J Biol Chem 281:27557–27565
14. Hibi T, Nii H, Nakatsu T, Kimura A, Kato H, Hiratake J, Oda J (2004) Crystal structure of gamma-glutamylcysteine synthetase: insights into the mechanism of catalysis by a key enzyme for glutathione homeostasis. Proc Natl Acad Sci USA 101:15052–15057
15. Hamilton D, Wu JH, Batist G (2007) Structure-based identification of novel human gamma-glutamylcysteine synthetase inhibitors. Mol Pharmacol 71:1140–1147
16. Griffith OW (1982) Mechanism of action, metabolism, and toxicity of buthionine sulfoximine and its higher homologs, potent inhibitors of glutathione synthesis. J Biol Chem 257:13704–13712
17. Griffith OW, Meister A (1979) Potent and specific inhibition of glutathione synthesis by buthionine sulfoximine (S-n-butyl homocysteine sulfoximine). J Biol Chem 254:7558–7560
18. Griffith OW, Anderson ME, Meister A (1979) Inhibition of glutathione biosynthesis by prothionine sulfoximine (S-n-propyl homocysteine sulfoximine), a selective inhibitor of gamma-glutamylcysteine synthetase. J Biol Chem 254:1205–1210
19. Faundez M, Pino L, Letelier P, Ortiz C, Lopez R, Seguel C, Ferreira J, Pavani M, Morello A, Maya JD (2005) Buthionine sulfoximine increases the toxicity of nifurtimox and benznidazole to Trypanosoma cruzi. Antimicrob Agents Chemother 49:126–130
20. Moncada C, Repetto Y, Aldunate J, Letelier ME, Morello A (1989) Role of glutathione in the susceptibility of Trypanosoma cruzi to drugs. Comp Biochem Physiol C 94:87–91
21. Repetto Y, Opazo E, Maya JD, Agosin M, Morello A (1996) Glutathione and trypanothione in several strains of Trypanosoma cruzi: effect of drugs. Comp Biochem Physiol B Biochem Mol Biol 115:281–285
22. Maya JD, Cassels BK, Iturriaga-Vasquez P, Ferreira J, Faundez M, Galanti N, Ferreira A, Morello A (2007) Mode of action of natural and synthetic drugs against Trypanosoma cruzi and their interac-

tion with the mammalian host. Comp Biochem Physiol A Mol Integr Physiol 146:601–620

23. Harvison PJ, Kalman TI (1992) Synthesis and biological activity of novel folic acid analogues: pteroyl-S-alkylhomocysteine sulfoximines. J Med Chem 35:1227–1233
24. Abbott JJ, Ford JL, Phillips MA (2002) Substrate binding determinants of Trypanosoma brucei gamma-glutamylcysteine synthetase. Biochemistry 41:2741–2750
25. Janowiak BE, Griffith OW (2005) Glutathione synthesis in Streptococcus agalactiae. One protein accounts for gamma-glutamylcysteine synthetase and glutathione synthetase activities. J Biol Chem 280:11829–11839
26. Jez JM, Cahoon RE, Chen S (2004) Arabidopsis thaliana glutamate-cysteine ligase: functional properties, kinetic mechanism, and regulation of activity. J Biol Chem 279:33463–33470
27. Huynh TT, Huynh VT, Harmon MA, Phillips MA (2003) Gene knockdown of gamma-glutamylcysteine synthetase by RNAi in the parasitic protozoa Trypanosoma brucei demonstrates that it is an essential enzyme. J Biol Chem 278:39794–39800
28. Ashida H, Sawa Y, Shibata H (2005) Cloning, biochemical and phylogenetic characterizations of gamma-glutamylcysteine synthetase from Anabaena sp. PCC 7120. Plant Cell Physiol 46:557–562
29. Drew R, Miners JO (1984) The effects of buthionine sulphoximine (BSO) on glutathione depletion and xenobiotic biotransformation. Biochem Pharmacol 33:2989–2994
30. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815
31. The UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. Nucl Acids Res 37:D169–D174
32. Green JR, Korenberg MJ, Aboul-Magd MO (2009) PCI-SS: MISO dynamic nonlinear protein secondary structure prediction. BMC Bioinforma 10:222
33. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4:187–217
34. Lovell SC, Davis IW, Arendall WB 3rd, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by Calpha geometry: phi, psi and Cbeta deviation. Proteins 50:437–450
35. Luthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356:83–85
36. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35:W407–W410
37. Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12:1073–1086
38. Honig B, Nicholls A (1995) Classical electrostatics in biology and chemistry. Science 268:1144–1149
39. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. J Mol Graph 14:33–38
40. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26:1781–1802
41. Mackerell AD Jr (2004) Empirical force fields for biological macromolecules: overview and issues. J Comput Chem 25:1584–1604
42. MacKerell AD, Bashford D, Bellott DRL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M (1998) All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J Phys Chem B 102:3586–3616
43. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
44. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An N [center-dot] log(N) method for Ewald sums in large systems. J Chem Phys 98:10089–10092
45. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh Ewald method. J Chem Phys 103:8577–8593
46. Feller SE, Zhang Y, Pastor RW, Brooks BR (1995) Constant pressure molecular dynamics simulation: The Langevin piston method. J Chem Phys 103:4613–4621
47. Ryckaert J-P, Ciccotti G, Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Chem Phys 23:327–341
48. Seeber M, Cecchini M, Rao F, Settanni G, Caflisch A (2007) Wordom: a program for efficient analysis of molecular dynamics simulations. Bioinformatics 23:2625–2627
49. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242
50. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23:566–579
51. Ferre F, Clote P (2005) DiANNA: a web server for disulfide connectivity prediction. Nucleic Acids Res 33:W230–W232
52. Lueder DV, Phillips MA (1996) Characterization of Trypanosoma brucei γ-Glutamylcysteine Synthetase, an Essential Enzyme in the Biosynthesis of Trypanothione (Diglutathionylspermidine). J Biol Chem 271:17485–17490
53. Brekken DL, Phillips MA (1998) Trypanosoma brucei Gamma-Glutamylcysteine Synthetase. J Biol Chem 273:26317–26322

J Mol Model (2012) 18:2065–2078
DOI 10.1007/s00894-011-1219-9

ORIGINAL PAPER

Combination of docking, molecular dynamics and quantum mechanical calculations for metabolism prediction of 3,4-methylenedioxybenzoyl-2-thienylhydrazone

Rodolpho C. Braga · Vinícius M. Alves · Carlos A. M. Fraga · Eliezer J. Barreiro · Valéria de Oliveira · Carolina H. Andrade

Received: 10 February 2011 / Accepted: 9 August 2011 / Published online: 8 September 2011

Abstract In modern drug discovery process, ADME/Tox properties should be determined as early as possible in the test cascade to allow a timely assessment of their property profiles. To help medicinal chemists in designing new compounds with improved pharmacokinetics, the knowledge of the soft spot position or the site of metabolism (SOM) is needed. In silico methods based on docking, molecular dynamics and quantum chemical calculations can bring us closer to understand drug metabolism and predict drug–drug interactions. We report herein on a combined methodology to explore the site of metabolism prediction of a new cardioactive drug prototype, LASSBio-294 (**1**), using MetaPrint2D to predict the most likely metabolites, combined with structure-based tools using docking, molecular dynamics and quantum mechanical calculations to predict the binding of the substrate to CYP2C9 enzyme, to estimate the binding free energy and to study the energy profiles for the oxidation of (**1**). Additionally, the computational study was correlated with a metabolic fingerprint profiling using LC-MS analysis. The results obtained using the computational methods gave valuable information about the probable metabolites of (**1**) (qualitatively) and also about the important interactions of this lead compound with the amino acid residues of the active site of CYP2C9. Moreover, using a combination of different levels of theory sheds light on the understanding of (**1**) metabolism by CYP2C9 and its mechanisms. The metabolic fingerprint profiling of (**1**) has shown that the metabolites founded in highest concentration in different species were metabolites **M1**, **M2** and **M3**, whereas **M8** was found to be a minor metabolite. Therefore, our computational study allowed a qualitative prediction for the metabolism of (**1**). The approach presented here has afforded new opportunities to improve metabolite identification strategies, mediated by not only CYP2C9 but also other CYP450 family enzymes.

Keywords Cytochrome P450 · Docking · Metabolism prediction · Molecular dynamics · QM calculations

Introduction

More than half of drug candidates fail during clinical trials due to an unsuitable pharmacokinetic profile. For this reasons, the study of ADME/Tox properties (absorption, distribution, metabolism, excretion, along with toxicity) has become an essential task to reduce the attrition rate at the late stages of the drug development process [1]. Cytochrome P450s (CYPs) form a superfamily of heme-containing proteins, which plays a crucial role in the metabolism of endogenous and exogenous compounds [2]. CYP enzymes participate in Phase I metabolism of 90% of all drugs [3]. The most important isoforms are CYP1A2 (~5% of current drugs), CYP2C9 and CYP2C19 (~25%), CYP2D6 (~20%) and CYP3A4 (~48%) [4]. Since CYP-450 enzymes metabolize the majority of xenobiotics, it is necessary to know the CYP450-mediated metabolic profiles of compounds during drug discovery and development. As the importance of CYPs became clear, the interest

R. C. Braga · V. M. Alves · V. de Oliveira · C. H. Andrade (✉)
Faculty of Pharmacy, Federal University of Goiás,
1a. Av. com Praça Universitária, PO Box 131, Goiânia, GO
74605-220, Brazil
e-mail: carolina@farmacia.ufg.br

C. A. M. Fraga · E. J. Barreiro
LASSBio, Faculty of Pharmacy,
Federal University of Rio de Janeiro,
PO Box 68023, Rio de Janeiro, RJ 21941-902, Brazil

in studying these protein systems both in vitro and in silico increased [5–7].

CYP2C9 is the predominant member of the 2C family with a major contribution to human drug metabolism [8, 9]. CYP2C9 exhibits selectivity for the oxidation of relatively small, lipophilic neutral or acidic compounds [10]. In particular, anti-inflammatory agents (diclofenac, flurbiprofen, ibuprofen, naproxen, piroxicam), anticoagulant compounds (S-warfarin), hypoglycaemic agents (tolbutamide), anticonvulsants (phenytoin) and loop diuretics (torsemide) as well as progesterone are CYP2C9 substrates. However, they represent just a few of the structurally diverse range of compounds that are oxidized by CYP2C9 and, more recently, the diversity of CYP2C9 substrates has widened with phosphorus-containing thioether pesticides shown to have significant CYP2C9 activity [11]. The importance of CYP2C9 in drug metabolism has led the enzyme to be one of the "standard" enzymes screened during the in vitro investigation of the hepatic metabolism of xenobiotics, particularly newly discovered drugs [9].

A detailed understanding of the metabolism mechanisms and prediction of metabolites is thus a major challenge being crucial to screen drugs in the early stage of lead development [12]. Since experimental investigation of the catalytically competent species in the metabolism already requires the presence of a substrate to initiate the reaction cycle, computational methods are very important to accomplish this task. Such techniques involve docking in the active site, pharmacophore modeling, molecular dynamics (MD) simulations, quantitative structure activity relationship (QSAR) and/or quantum mechanical and molecular mechanical (QM/MM) studies [13].

In this paper we focus on the compound 3,4-methylenedioxybenzoyl-2-thienylhydrazone (LASSBio-294, **1**), a novel cardioactive compound of the *N*-acylhydrazone class [14] that was found to improve intracellular Ca^{2+} regulation [15] and prevent myocardial infarction induced by cardiac dysfunction, which could potentially prevent heart failure. In addition, (**1**) also promoted vasodilation in aortic rings, mediated by the guanylate cyclase/cyclic guanylate monophosphate pathway [16].

Herein, we show a successful application of a combined computational approach to explore the site of metabolism (SOM) prediction of a new drug candidate, using ligand- and structure-based metabolism prediction tools and the correlation with a metabolic fingerprint profiling using LC-MS analysis. The study was subdivided into four parts: (a) application of MetaPrint2D, a ligand-based tool to predict the sites of metabolism (SOM) of (**1**); (b) docking studies to address the prediction of the SOM based on the most likely poses of substrate within the active site of CYP2C9 followed by molecular dynamics (MD) simulations of some docked complexes to verify their stability; (c) QM calculations to study the energy profiles for the oxidation of (**1**); and (d) comparison with experimental results. Apart from the ligand-based tool, all other theoretical calculations aimed at predicting the CYP2C9 metabolism of LASSBio-294 (**1**). Therefore, our goals were to combine methods to improve the predictivity of SOM of (**1**), which have not been widely explored, and to correlate with experimental assays.

Material and methods

MetaPrint2D

MetaPrint2D is a tool for predicting the sites of a molecule that are most likely to undergo Phase I metabolism, based on their similarity to known and unknown sites of metabolism to be metabolized [17]. The method builds on a database of atom environments found in molecules known to undergo metabolic transformation, such as the data found in the Symyx(R) (previously MDL) Metabolite database http://www.symyx.com/, which contains over 80,000 metabolic transformations of xenobiotics, curated from reports in scientific literature. The software was used on the web platform (http://www-metaprint2d.ch.cam.ac.uk/metaprint2d/), by uploading the SMILES string of LASSBio-294.

Preparation of protein and ligand for docking studies

The crystal structure of LASSBio-294 (**1**), obtained from Cambridge Crystallographic Data Centre (CCDC code 707596), was energy-minimized using force field MMFF94x, and the partial atomic charges were computed using the AM1 [18] semiempirical method implemented in the Molecular Operating Environment (MOE) version 2008.10 software (Chemical Computing Group Inc., Montreal, Canada). The crystal structure of the human cytochrome P450 2C9 in complex with flurbiprofen was obtained from the Protein Data Bank (PDB code: 1R9O, resolution 2.0Å) [10]. Hydrogens atoms were added and minimized using the AMBER99 [19] force field and AMBER99 atomic charges [20] until the RMS force was <0.01 kcal mol-1Å-1 with the truncated Newton method. For the heme partial charges of CYP2C9, RESP charges determined by quantum chemical calculations were used [20]. The iron metal atom from the heme group was set to a + 3 charge. Protonation states according to a pH of 7 were assigned using the "Protonate 3D" option in MOE. Protonation states of histidines were assigned according to their H bonding environment. The net charge of the protein was 6.0. For the docking studies, the enzyme was prepared where: (i) ligand molecule was removed from the enzyme active site; (ii) Water molecules were removed, keeping the active site water molecules (wat600, wat819 and wat842, as described in ref [10] to be important for CYP2C9 substrate binding); (iii) MOE Alpha Site Finder was used for the active sites search in the enzyme structure and dummy atoms were created from the obtained alpha spheres.

Docking studies

Docking studies were performed using the MOE-Dock software [21] allowing side chains flexibility. Ligand placement was performed using alpha PMI method, with Affinity dG scoring function. The Alpha PMI placement method generates poses by aligning ligand conformations' principal moments of inertia to a randomly chosen subset of alpha sphere dummies in the receptor site. The Affinity dG is a scoring function that estimates the enthalpy contribution to the free energy of binding using a linear function of hydrophobic, ionic, hydrogen bond and metal binding terms (kcal mol^{-1} of total estimated binding energy) [22, 23]. The top 30 poses were retained and refined using MMFF94x force field energy minimization with Generalized Born solvation model [22], allowing the receptor side chain residues within 6 Å to relax around the mobile ligand. The receptor side chains were tethered with a force constant of 1.0 kcal mol^{-1} $Å^{-2}$). Energy minimization was stopped when the root-mean-square (RMS) gradient cutoff of 0.01 kcal mol^{-1} $Å^{-2}$ was reached. Final poses were ranked using the Affinity dG scoring method to calculate the free energy of binding. In order to determine the possible metabolic sites of the substrate, the distances between the heme iron of CYP2C9 and the atoms of (**1**) were measured for all docking results. A catalytically reactive distance from the heme iron of CYP450 is generally known to be within 5 Å, thus, the atom sites within a catalytically reactive distance from the heme iron were selected as possible metabolic sites.

MD simulations

The most favorable docking results for the CYP2C9·LASSBio-294 complexes were further optimized by molecular dynamics (MD) simulations. All MD simulations were carried out using the Desmond MD package version 2.2 [24] and the OPLS-AA 2005 force field [25]. MD simulations were carried out to provide additional structural relaxation and establishment of reasonable model hydrogen bonding patterns. The system was solvated with SPC water molecules generated via an orthorhombic box. The number of solvated water molecules in each system is about 17,000 and the initial MD simulation cell dimension was about 95 Å × 80 Å × 100 Å and involved the complex being solvated by a layer of water molecules of at least 10 Å in all directions. By assuming normal charge states of ionizable groups corresponding to pH 7, sodium (Na^{+}) and chloride (Cl^{-}) counter-ions at physiological concentration of 0.015 mol/L were added in the box in random positions to ensure the global charge neutrality. The structures were relaxed by performing equilibration dynamics at constant temperature (300 K) and constant pressure (1 bar). The constant pressure and temperature were controlled via Langevin dynamics method [26]. The simulation time was 120 ps and the coordinates were stored every 1.2 ps.

Protein-bound ligand entropy calculation

SZYBKI (OpenEye Scientific Software Inc.) [27, 28] was applied to predict the protein-bound ligand entropy for the metabolism of (**1**) carried out by the active site of CYP2C9 using our MD results as starting point for the calculations.

Quantum-mechanical (QM) calculations

Starting structures were taken from the MD simulations described above. The QM region comprised heme group without side chains, the SH group of the cysteinyl ligand, and the substrate (**1**). Jaguar 7.6 software [29] was applied in order to study the energy profiles for the oxidation of (**1**). The QM calculations were performed using the density functional theory (DFT) with the spin-unrestricted UB3LYP. Geometry optimizations (without constraints) were performed with the LACV3P basis set on iron and 6-31 G* on the rest of the atoms (basis set BSI). Subsequently, single point calculations were performed on the optimized geometries using BSII, which corresponds to LACV3P(Fe)/6-311 + G** (rest).

Biological data

The details of the methodology regarding the biological data to study the metabolism of LASSBio-294 (**1**) have been published in previous research articles [30, 31] and are only summarized herein. For the in vivo metabolism evaluation of (**1**), 12 plasma samples from beagle dogs and rats were collected before and after 1.5 h of administration of (**1**), and analyzed by LC-MS. Six urine samples from dogs and rats treated with (**1**) were also collected after 4 h of oral administration. The animal handling protocol of this study had been reviewed and approved by the Institutional Animal Care and Use Committee of the Federal University of Goias (UFG). The in vitro metabolism study of (**1**) was conducted using the fungus *Beauveria bassiana* ATCC 7159. Twelve whole-cell microorganism media samples were taken every 24 hours, up to 96 hours of incubation with (**1**) and analyzed by LC-MS.

Metabolic fingerprinting using LC-MS

We have studied the metabolic fingerprint profiling of (**1**) between different species. A general workflow simplified in three stages was implemented for the metabolic fingerprinting using LC-MS analysis. The first stage was concerned in preparing the samples to be analyzed by LC-MS. The second one was brought up to analyze and treat all samples by MS's software using the MassHunter software (Agilent Technologies, Inc., USA) for targeted metabolites and MZmine 2 [32] for un-targeted metabolites. At the third stage, we were able to identify and quantify all detectable metabolites produced in vitro by filamentous fungi. Furthermore, using the data produced by un-targeted metabolites from dogs, rats and filamentous fungi samples, we have tried to identify the similarity across these species taking advantage of the principal component analysis (PCA) analysis. Therefore, we have performed post-processing statistical analysis of all the experimental data (in vitro and in vivo), using multivariate analysis (PCA) to extract information from the complex data to obtain metabolic fingerprints. High-resolution MS fingerprints were acquired on a mass spectrometer (Agilent 6520 Accurate-Mass QTOF mass) in both positive and negative electrospray ionization mode. The source temperature was kept at 325 °C, Cap voltage (2100 V), fragmentor voltage (175 V) and drying gas (5 L/min). The m/z ranges were acquired (300–2000 for mass ranges, with an acquisition rate 1.02 spectra/sec) for plasma, urine and whole-cell microorganism samples analysis, respectively.

Results and discussion

Metabolite identification studies are performed relatively late in the compound optimization process because they are work intensive and generally aimed to understand the metabolic pathway (generally in vivo) of an already potent and optimized drug candidate. In modern drug discovery process, ADME/Tox properties should be determined as early as possible in the test cascade to allow a timely assessment of their property profiles [33]. To help medicinal chemists in designing new compounds with improved pharmacokinetics, the knowledge of the soft spot position or the site of metabolism (SOM) is needed. In recent years, driven by the development of new software and advances in hardware technology, it has become evident that the incorporation of quantum mechanical (QM) methods in combination with standard classical approaches, in certain stages of in silico drug metabolism studies [34, 35], leads to many improvements.

The approach described herein is new in that it introduces the combination of MetaPrint2D, a improved algorithm for site of metabolism prediction, to predict the most likely metabolites, with structure-based methodologies using docking, MD simulations and QM calculations to improve predictivity of SOM of a new drug candidate (**1**), which has not been widely explored, and the correlation with a metabolic fingerprint profiling using LC-MS analysis. Figure 1 shows a schematic diagram of the approach described in this paper aiming at improving drug metabolism studies using different levels of theory and the correlation with experimental assays for complex matrix analysis from different species.

Panel A from Fig. 1 shows the energy changes from substrate binding to product formation in CYP450-catalyzed drug metabolism. Therefore, the integration of computational methods such as MetaPrint2D, docking, molecular dynamics and QM calculations can bring us closer to understand drug metabolism and predict drug–drug interactions. Panel B shows the metabolic fingerprint workflow for complex matrix analysis from different species using LC-MS.

MetaPrint2D is a fast, efficient and accurate predictor of both the sites and products of metabolism in small molecule drugs. The approach adopted is a development of the method of Boyer and co-workers [36] using circular fingerprints and substrate/product ratios. The method has been completely developed from scratch, fast algorithms and extensive testing employed to maximize the performance of the approach [17]. The sites of metabolism prediction of LASSBio-294 (**1**) from MetaPrint2D are shown in Fig. 2. The results are visualized so that atoms are colored according to the likelihood of a metabolic site being centered on this atom. This method has been compared with other rule-based methods or based on expert systems and is a promising option to use in combination with other methods [37].

Accordingly to Fig. 2, MetaPrint2D predicted that the sulfur (S14) in thiophene ring and the carbon atom (C1) of the benzodioxolyl ring of (**1**) were most likely to be metabolized (colored in red), followed by the group colored in orange (N7) and then by the groups marked in green (C13, O2, O2' and C9). The structures of the possible metabolites predicted by MetaPrint2D are presented in Fig. 3.

The process of the metabolic reaction of xenobiotic consists of a series of processes, including substrate binding to the enzyme, catalytic reaction of a substrate by the enzyme, and release of a metabolite from the enzyme. At first, the substrate must bind in close proximity between the metabolic reaction atom within the substrate and the catalytic site of the CYP enzyme (*i.e.*, heme oxygen). Force field based docking techniques and molecular dynamics (MD) simulations can mimic this complex formation process and the dynamic motion of the substrate-enzyme complex [38]. Substrate

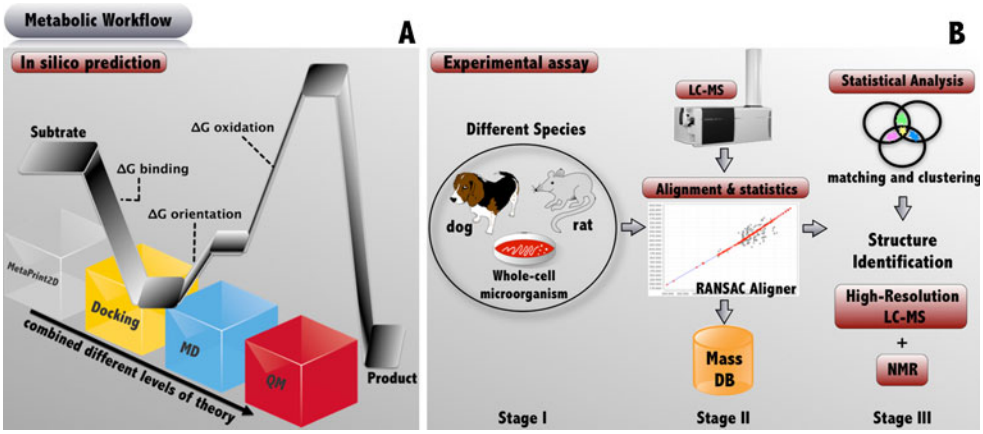


Fig. 1 Workflow for metabolic investigation proposed in our work. (**a**) Proposed computational methods to improve drug metabolism studies using different levels of theory, showing the energy changes from substrate binding to product formation in CYP450-catalyzed drug metabolism. (**b**) Metabolic fingerprint workflow for complex matrix analysis from different species. **Stage I**: the preparation of the samples to be analyzed by LC-MS; **Stage II**: analyze the samples into LC-MS and treat the data; and **Stage III**: identify and quantify all detectable metabolites produced in vitro by filamentous fungi and discover the similarity across the species using PCA analysis

orientation within the active site of CYP450s is a crucial factor for CYP-mediated metabolism. Therefore, docking studies can be particularly useful for gaining selectivity and steric information about potential compounds, which can be used to predict their sites of metabolism and possible toxic metabolites. Previous studies with the principal human CYP isoforms indicated by the decision tree for evaluating P450 specificity developed by Lewis [39] have pointed out that CYP2C9 possess important molecular recognition pattern to (**1**) [40].

As a measure of docking reliability, the root-mean-square deviation (RMSD) was used to compare differences between the atomic distances of the docked poses and the co-crystallized structure. The CYP2C9-flurbiprofen complex (PDB: 1R9O) [10] was used for the initial validation run where flurbiprofen was docked into CYP2C9. Docking using the MOE-Dock software [21] allowing side chains flexibility, accurately predicted the crystallographic placement of flurbiprofen in the crystal structure of human CYP2C9, with a RMSD of 1.21 Å.

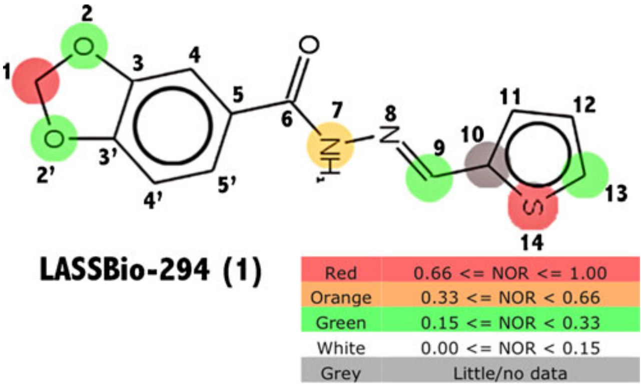


Fig. 2 Plot of MetaPrint2D predictions. Site of metabolism: the atoms in (**1**) that most will be metabolized are colored according to the likelihood of a metabolic site: High: red, Medium: orange, Low: green, Very low is not colored, and No data: gray. NOR indicates the normalized occurrence ratio; a high NOR indicates a more frequently reported site of metabolism in the metabolite database

The drug candidate (**1**) was docked into the active site of CYP2C9 with three structural water molecules, which were kept due its proximity with the binding site and its importance for CYP2C9 substrate binding [10]. To score our docking results for predictions of the sites of metabolism, we considered any conformation in which a group or an atom on (**1**) moiety was within 5 Å of the heme iron a successful prediction for (**1**) metabolism. Of the 30 conformations obtained, only five docking solutions had the substrate in a favorable distance to the heme iron for metabolism (less than 5 Å) [10, 41]. These results are shown in Table 1.

Table 1 also depicts the calculated binding energies of the docked poses (ΔG_{calc}), using the affinity dG scoring function, presented in the methods section [22]. One should note that, given the difficulties of scoring functions to approximate the binding energies, the ΔG differences between the docked poses are not large enough to provide specificity for a certain reaction site in the substrate. The last column of Table 1 shows the calculated values of entropy for the binding of the complexes protein-ligand, using vibrational modes, based on the Hessian matrix from the minimization, available on SZYBKI software, which will be discussed further.

Figure 3 is our final prediction model, which summarizes all the predicted metabolites for (1) in our study, using a combination of computational methods, *i.e.*, MetaPrint2D, docking, MD simulations and QM calculations.

In Fig. 3, all predicted metabolites were included, even those chemically labile compounds. As we can see from

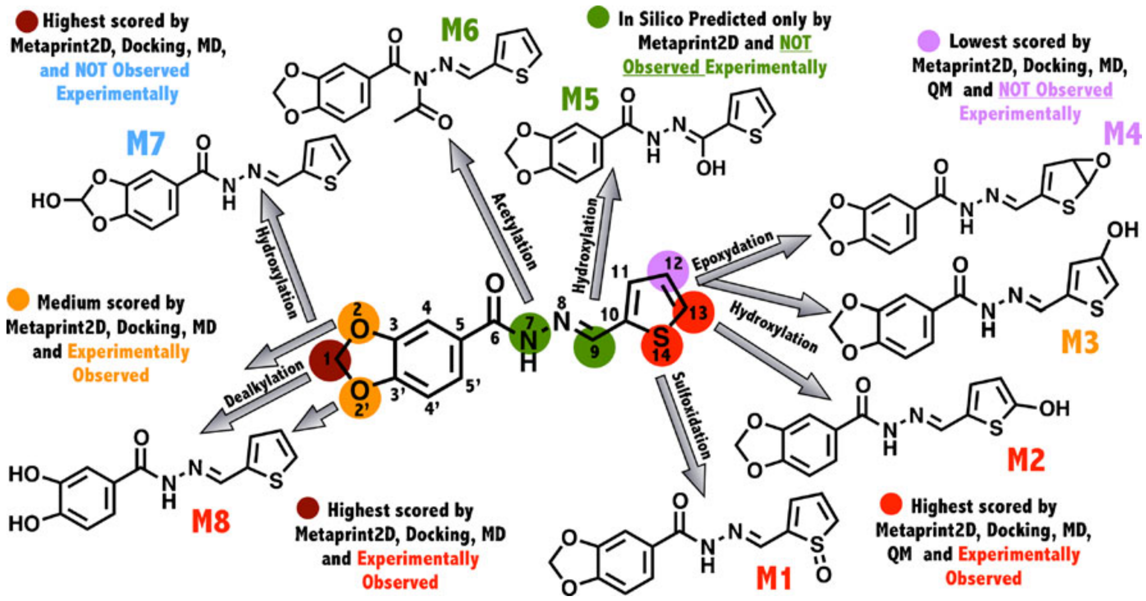

● Highest scored by Metaprint2D, Docking, MD, and NOT Observed Experimentally
M7
M6
● In Silico Predicted only by Metaprint2D and NOT Observed Experimentally
M5
● Lowest scored by Metaprint2D, Docking, MD, QM and NOT Observed Experimentally
M4
Hydroxylation
Acetylation
Hydroxylation
Epoxydation
Hydroxylation
M3
● Medium scored by Metaprint2D, Docking, MD and Experimentally Observed
Dealkylation
Sulfoxidation
M2
M8
● Highest scored by Metaprint2D, Docking, MD and Experimentally Observed
M1
● Highest scored by Metaprint2D, Docking, MD, QM and Experimentally Observed

Fig. 3 Predicted metabolites for (1) using a combination of computational methods: MetaPrint2D, docking, MD simulations and QM calculations. The sites that are predicted both by MetaPrint2D and docking are color-coded by likeness and binding energy. The metabolites that were found experimentally are colored in red and orange

Fig. 3, the highest scored metabolites predicted by all computational methods were **M1**, **M2** and **M8**. The metabolites **M5** and **M6** (Fig. 3) were only predicted by Metaprint2D.

The docked complexes that predicted metabolically active states were submitted to coupled energy minimization and MD relaxation studies, to provide additional structural relaxation and establishment of reasonable model hydrogen bonding patterns. The relaxed complexes after MD simulations were then visually analyzed to identify active site residues that could potentially position (**1**) for metabolism and/or stabilize transition states. The distances of the substrate with key residues in the binding pocket of CYP2C9 were calculated and shown in Table 2.

The association reaction of two molecules to form a single complex must overcome a large entropic barrier, due to the loss of translational and rotational degrees of freedom. Therefore, to better estimate the protein-ligand binding entropy we have used the Hessian matrix available at SZYBKI software, using the complex LASSBio-294-CYP2C9.

This methodology has shown that the Hessian matrix of second derivatives built by a quasi-Newton optimizer during geometry optimization of a molecule with a classical molecular potential in the protein-receptor environment can be used to predict vibrational entropies, being updated by analyzing successive gradient vectors [28].

The five CYP2C9·LASSBio-294 complexes previously optimized by MD simulations were used to estimate the values for the ligand-protein of each docked pose, during binding equilibrium protein + ligand $\rightleftharpoons$ protein·ligand. These results are shown in Table 1. Remarkably, the lowest values of protein-ligand binding entropy were found for the SOM reaction types of *S*-oxidation (**M1**), aromatic hydroxylation (**M2**) and dealkylation (**M8**) (Table 1), which corroborate with the results achieved from MetaPrint2D

Table 1 Metabolism site prediction from docking of LASSBio-294 into CYP2C9

Pose #	Atom site of metabolism	Metabolite ID	d^a (Å)	SOM reaction type	$\Delta G_{calc}{}^b$ (kcal/mol)	$-T\Delta S_{cal}{}^c$ (kcal/mol)
1	S tiophene ring	M1	3.39	*S*-oxydation	−4.89	21.0380
2	C_{13} tiophene ring	M2	3.15	aromatic hydroxylation	−4.81	21.2623
3	CH_2 benzodioxolyl	M8	3.06	*CH_2* dealkylation	−4.77	21.3993
4	C_{12} tiophene ring	M3	3.32	aromatic hydroxylation	−4.69	21.5901
5	O benzodioxolyl	M8	3.46	*O*-dealkylation	−4.47	21.4010

[a] Indicates the distance between the site of metabolism and the heme iron of CYP2C9

[b] Calculated free energy of binding using the Affinity dG method

[c] Calculated values of entropy for the binding equilibria Protein + Ligand $\rightleftharpoons$ Protein-Ligand

Table 2 Distances of the poses of LASSBio-294 submitted to MD simulations with key residues in the binding pocket of CYP2C9

		Docking solutions of LASSBio-294[a]				
		1 (M1)	2 (M2)	3 (M8)	4 (M3)	5 (M8)
Arg108	d[b] (Å)	2.49±0.06	3.78±0.10	4.81±0.10	2.30±0.34	3.01±0.87
Asn204	d (Å)	2.78±0.15	1.88±0.32	5.37±0.22	2.35±0.07	3.63±0.43
Phe114	d (Å)	3.77±0.55	3.85±0.16	3.57±0.09	3.75±0.25	5.16±0.21

[a] The five docking solutions (and the metabolites ID) that had the substrate in a favorable distance to the heme iron for metabolism

[b] Distances calculated after MD simulations, with the standard deviations

and docking, showing that these metabolites are the most probable to be found experimentally.

The three docked poses of (**1**) that predicted the metabolism of thiophene ring moiety (metabolites **M1**, **M2** and **M3**) are shown in Fig. 4. Inspection of the docked structures (Fig. 4) suggests that the oxygen atoms of the benzodioxolyl ring of (**1**) may be able to form hydrogen bonds with Arg108 and Asn204 residues, which are well recognized as important residues in the CYP2C9 active site [10, 42]. This leaves the thiophene ring between 3.15 and 3.39 Å from the iron heme (Table 1). This may reflect the fact that the thiophene ring is an energetically favorable site of metabolism, and may lead to sulfoxidated (**M1**) and hydroxylated (**M2** and **M3**) metabolites. The NH groups on the Arg108 and Asn204 side chains were within hydrogen bonding distance, approximately 2.3 - 3.8 and 1.9 - 2.8 Å, respectively, of the oxygen atoms of the benzodioxolyl ring moiety, indicating that Arg108 and Asn204 residues could potentially stabilize the binding mode of the ligand and also the transition state during LASSBio-294 (**1**) metabolism. Hence, Arg108 and Asn204 appeared to play a key role in positioning (**1**) the thiophene ring moiety for metabolism (Fig. 4).

In addition, Phe114, which is one of the most important hydrophobic/aromatic complimentary site of CYP2C9 ligands, located at the entrance of the active site [43–45], also appeared to play a role in positioning (**1**) for thiophene ring moiety metabolism. The steric interactions between (**1**) and the Phe114 residue come out to provide a hydrophobic pocket. In addition to the steric interactions, LASSBio-294's benzene moiety was positioned to enable a π-bond stacking with the Arg108 residue. These steric and electronic interactions are shown to position the thiophene ring moiety of (**1**) toward the heme, supporting the sulfoxidation and hydroxylation pathways.

The other two metabolically active poses according to the docking and MD simulations of (**1**) in the active site of CYP2C9, predicted the metabolism of benzodioxolyl ring moiety and are shown in Fig. 5. These orientations are consistent with (A) *CH_2*-dealkylation; and (B) *O*-dealkylation of the benzodioxolyl ring, both for the formation of metabolite **M8**.

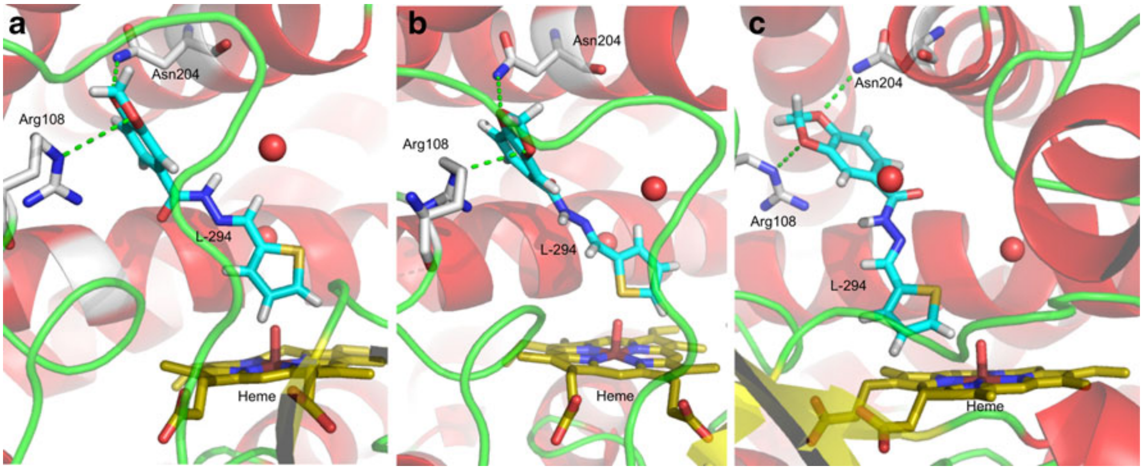


Fig. 4 Binding poses of substrate in the active site of CYP2C9 (1R9O) predicted by docking and MD simulations. Orientations consistent with (**a**) and (**c**) hydroxylation of thiophene ring, to generate metabolites M3 and M2, respectively; (**b**) sulfoxidation, to form metabolite M1. 3D representation including the heme group (carbon atoms are shown in yellow), the active site water molecules (red spheres) and LASSBio-294 (1) (cyan). (1) is shown in color-coded sticks: carbon = cyan, nitrogen = blue, oxygen = red, sulfur = yellow, and hydrogen = white. Green dot lines denote hydrogen bonds

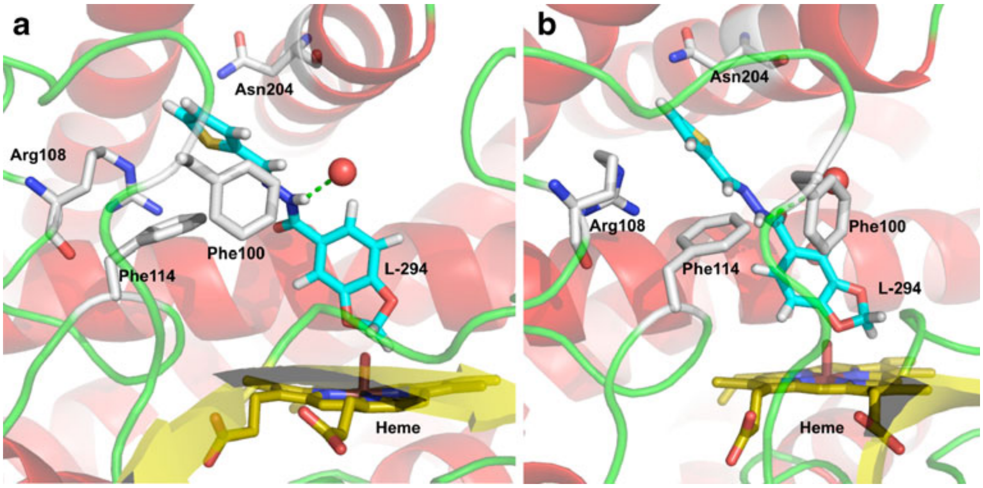


Fig. 5 Binding poses of (1) in the active site of CYP2C9. 3D representation including the heme group (carbon atoms are shown in yellow), the active site water molecules (red spheres) and (1) (cyan). In both panels, the water molecule Wat842 is involved in hydrogen bonding with (1), are shown as green dotted lines. These orientations are consistent with (**a**) *CH_2*-dealkylation; and (**b**) *O*-dealkylation of the benzodioxolyl ring, both for the formation of metabolite M8

Active-site water molecules play an important role in biological systems, facilitating promiscuous binding or an increase in specificity and affinity. Studies of the water molecules in the ligand-binding cavities of cytochromes P450 indicate that their high mobility facilitates the movement of the substrates and products into and out of the active site [46, 47]. Therefore, we included three active-site water molecules in molecular docking simulations of the CYP2C9 enzyme. In Fig. 5(a and b), the hydrazone nitrogen of (**1**) is involved in hydrogen bonding with one active-site water molecule (Water 842), showing that this water molecule could potentially stabilize the transition state during benzodioxolyl ring moiety metabolism of (**1**).

As we can see from Fig. 5, the binding poses to generate the metabolite **M8** from both *CH_2*-dealkylation (A) and *O*-dealkylation (B) are not in a favorable position to form hydrogen bonds with Arg108 and Asn204. Although these poses have shown the lack of H-bonds, they have other favorable interactions, such as a π-bond stacking between the Arg108 residue and the thiophene ring of (**1**). Moreover, the molecular recognition is reinforced by favorable hydrophobic π-π stacking interactions among two phenylalanine amino acids residues - Phe100 and Phe114 - with thiophene ring of (**1**), as illustrated in Fig. 5. All these interactions, are shown to support the dealkylation metabolism of (**1**), to form the metabolite **M8**.

Thus, for the five docked poses metabolically active obtained, in all of them Arg108 residue was flagged as potential substrate recognition moieties, in agreement with previous studies [10, 42], providing solid evidence for the preference of relatively small lipophilic anionic substrates for CYP2C9.

Quantum chemical calculation is a major tool for predicting CYP450 catalysis. From the calculated energy barrier value, we can tell the absolute or relative oxidation potential in xenobiotic metabolism [48–50]. The identification of the active oxidant in the reaction process is fundamental to understand the formation of products catalyzed by cytochromes P450. According to experimental and computational evidence, the major part of the large variety of reactions catalyzed by P450 enzymes all involve the same active species, a high-valent iron-oxo derivative of the active site heme group, known as compound I (Cpd I) [34, 49]. Its ground state has three unpaired electrons, two in Fe–O π^* orbitals, and one in a π-orbital of the porphyrin. Due to the weak coupling between the Fe–O based and porphyrin orbitals, the energy difference between the resulting quartet and open-shell doublet states of Cpd I is very small, giving rise to two-state reactivity (TSR) of P450 enzymes [51]. This species reacts with substrates via oxygen atom transfer to give oxygenated products. Indirect evidence of Cpd I through kinetic isotope effects and product distributions has implicated it to be the key oxidant involved in monooxygenase reactions with substrates [52]. More recent low pressure mass spectrometric studies [53, 54] on biomimetic iron-porphyrins and computational modeling [49, 55] have shown it to be a very efficient oxidant of substrate hydroxylation and epoxidation reactions. The spin-unrestricted UB3LYP hybrid DFT method was chosen as it has been shown to predict accurately structures and energetics for bioinorganic systems such as CYP450 Cpd I and other transition-metal complexes [34, 49, 51].

In silico methods based on a combination of docking, molecular dynamics and QM reactivity calculations can bring us closer to understand drug metabolism and predict drug–drug interactions. Therefore, we have carried out QM calculations on two of the possible metabolic routes of (**1**), aimed at an understanding of the mechanistic level of the reactions to form the metabolites **M1** and **M2**. These metabolites were chosen to perform QM calculations to get insight into how aromatic substrates are oxidized by human P450 isoforms.

Figure 6 shows the potential energy profile and critical species for the formation of **M1** by Cpd I, in their lowest

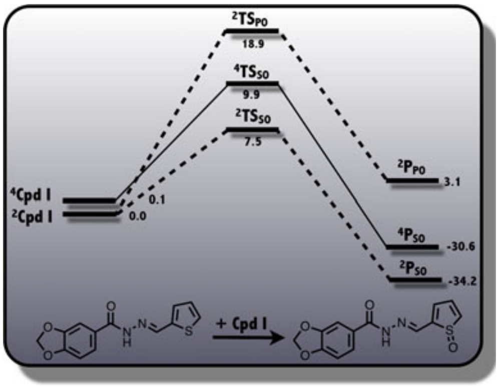


Fig. 6 Potential energy profiles for the sulfoxidation of (1) by Cpd I, to form the metabolite M1. Energies are in kcal mol^{-1} relative to the reactant complex in the doublet ($^2TS_{SO;}$ dashed line) and quartet ($^4TS_{SO;}$ plain line) spin states and contain zero-point corrections. All data were obtained with DFT (UB3LYP) single-point calculations on the optimized geometries using LACV3P(Fe)/6-311 + G**(rest)

doublet ($^2TS_{SO}$) and quartet ($^4TS_{SO}$) spin states as obtained from QM modeling. The difference between doublet (low-spin) and quartet (high-spin) states is mainly manifested in the spin of the porphyrin ring; whether or not the porphyrin ring's spin is parallel to the unpaired spin on Fe and oxo determines the multiplicity of the complex. The observed mechanism is seen to involve an *O*-transfer step via transition states $^{4,2}TS_{SO}$ that lead to the product complexes $^{4,2}P_{SO}$. Although the doublet and quartet spin states are degenerate for Cpd I, in the transition states, the doublet spin state ($^2TS_{SO}$) is lower in energy than the quartet ($^4TS_{SO}$) by 2.4 kcal mol^{-1} for CYP2C9. The $^2TS_{PO}$ is found to lead to the porphyrin self-oxidation (PO) product $^2P_{PO}$, and this specie was previously reported [56]. The label $^2TS_{PO}$ signifies that this is a transition state for porphyrin oxidation, leading to N-O porphyrin adducts found experimentally [57]. The sulfoxidation reaction of (**1**) by Cpd I of P450 is a concerted reaction via a transition state ($^2TS_{SO}$) leading to sulfoxide product complex ($^2P_{SO}$). Thus, Cpd I will carry out a fast sulfoxidation [58] of thiophene ring of (**1**) with an energy barrier of 7.5 kcal mol^{-1}, to form metabolite **M1**.

The mechanism of aromatic hydroxylation of was investigated for (**1**) hydroxylation using DFT calculations of the whole reaction profile. Figure 7 shows the mechanistic scheme for the aromatic hydroxylation of (**1**) by Cpd

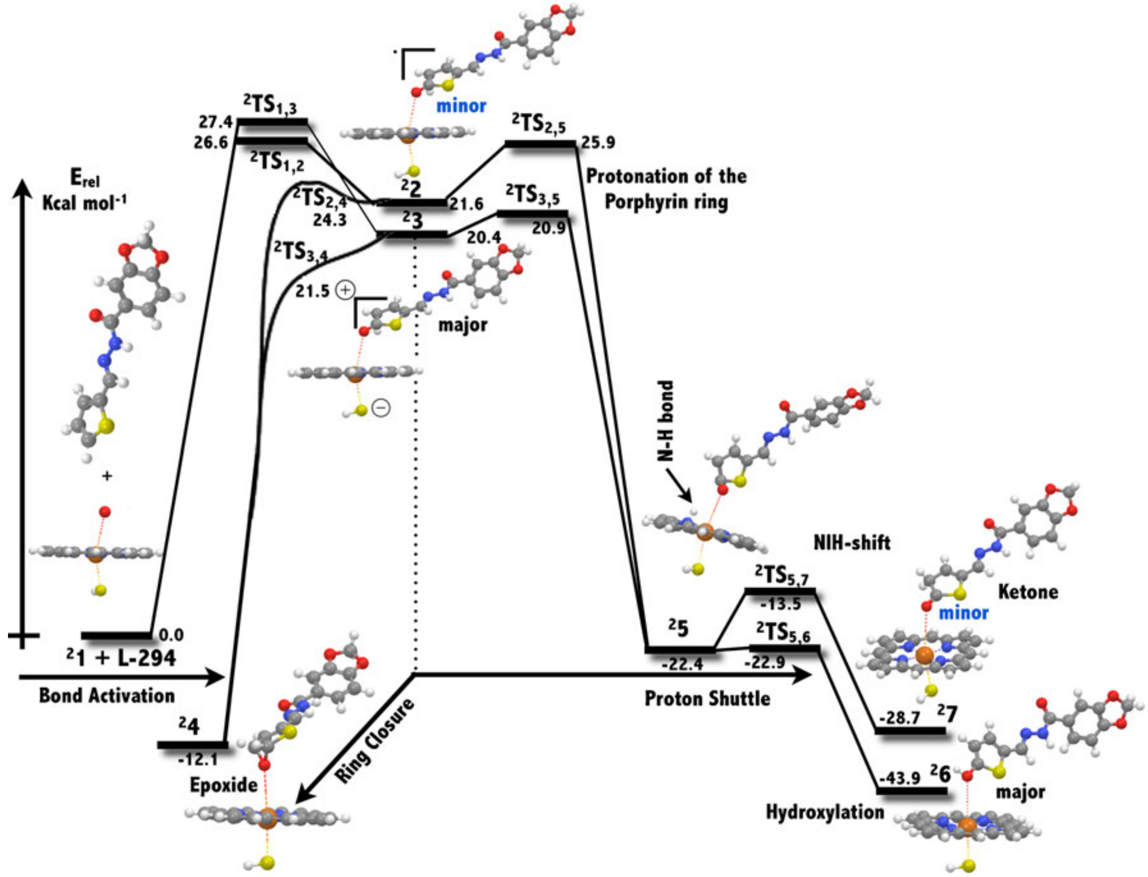


Fig. 7 Mechanistic scheme for the aromatic oxidation of (**1**) by Cpd I. Only the doublet low-spin (LS) mechanism is shown. The relative energies (kcal mol^{-1}) were taken from DFT (UB3LYP) single-point calculations and were done on the optimized geometries using LACV3P(Fe)/6-311 + G**(rest)

I. The QM results show that during the reaction, thiophene ring activation occurs by an initial attack on π-system of the thiophene, which preferentially takes place via the lower-energy doublet state (LS) to produce the transition states that were found to have a hybrid nature with radical ($^2TS_{1,2}$) and cationic ($^2TS_{1,3}$) characters [37, 59, 60]. This hybrid character is retained in the tetrahedral intermediates, which is neither fully cationic nor radicalar [59]. The levels of the high-spin (HS) species are not shown, since the HS transition states (TSs) are much higher in energy than the LS state. The free energy of activation barrier for H abstraction is 26.6 kcal mol^{-1} in the doublet state radical ($^2TS_{1,2}$) and 27.4 kcal mol^{-1} in the doublet state cationic ($^2TS_{1,3}$).

The main reaction path is electrophilic leading to the cationic σ-complex, ^{2}**3**, while a minor path involves the radical σ-complex, ^{2}**2**, and it is revealed to be the rate determining step for all mechanism (Fig. 7). Ring closure in these intermediates produces the epoxide product ^{2}**4**. The epoxide product ^{2}**4** does not rearrange to hydroxylated (^{2}**6**) or ketone (^{2}**7**) at the thiophene moiety and requires an appropriate media to protonate the epoxide or to catalyze its ring opening. The computational study implicates that the active species of the enzyme play a role as an internal base and catalyzes directly the production of ^{2}**6** e ^{2}**7**. This enzymatic mechanism involves proton-shuttle mechanism yielding to the protonated porphyrin intermediate ^{2}**5**. The protonated porphyrin intermediate subsequently transfers the proton back to the oxygen or to epoxy carbon that accounts to hydroxylated or ketone by NIH shift mechanism. The formation of **M2** (^{2}**6**) via proton-shuttle mechanism is predictive to be fast to compete with the post-enzymatic conversion of epoxide (^{2}**4**) to the correspondent hydroxylation of thiophene by external acid catalysis.

As a result from our QM calculations, we suggest that the formation of **M1** catalyzed by the oxyferryl active site of CYP2C9 is preferred in comparison to the thiophene ring hydroxylation (**M2**), as the sulfoxidation pathway involves a single step that occurs from the lowest-energy $^2TS_{SO}$ species with a barrier of merely 7.5 kcal mol^{-1}. In contrast, the aromatic hydroxylation process has a higher energetic barrier (26.6 kcal mol^{-}1) established by the multi-step proton-shuttle mechanism (Fig. 7).

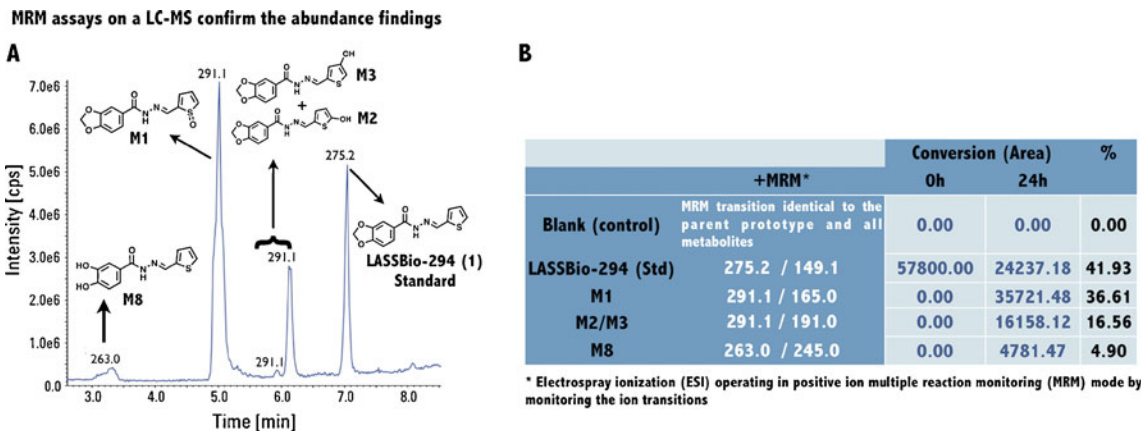
Our laboratory has been performing a variety of in vitro metabolism studies, using filamentous fungi, and in vivo using rats and dogs [31, 61, 62]. Here, we have carried out a metabolic fingerprint study, aimed to identify and quantify all detectable metabolites produced in vitro by filamentous fungi and to compare them with the mammalian metabolites detected in dogs and rats. To achieve this goal, our purpose was to use metabolic profiling for clustering the similarity between the three studied species, taking advantage of the principal component analysis (PCA) analysis.

In order to give a better understanding of the metabolism of (**1**), a two-step LC-MS approach was employed due to its wide dynamic range, reproducible quantitative and quantitative analysis, and its ability to analyze complex biological matrix. The samples were analyzed using time-of-flight (TOF) mass spectrometry (MS) followed by targeted identification of differentially produced metabolites using quadrupole time-of-flight (Q-TOF) MS/MS. An accurate-mass Q-TOF was applied to preform targeted MS-MS analysis of a metabolite and produce fragmentation information rule out candidate identities generated previously by molecular formula generation (MFG), to produce a list of possible molecular formulas based on accurate-mass data and isotope patterns for search at the METLIN personal metabolite database [63]. According to the MS analysis, we proposed the formation of metabolites **M1**, **M2**, **M3** and **M8**. The extracted ion chromatograms (EIC) for metabolites **M1**, **M2**, **M3** and **M8** and the relative formation of each product are shown in Fig. 8.

The compounds identified between each sample set using molecular feature extraction (MFE), and then were aligned for comparison using PCA. The objective was to discover new components (variables), which account for the majority of the differences in the data. The PCA plot finds the relationships beyond pair-wise comparison and enables biological interpretation through pathway analysis clustering the data samples in distinct groups. Here, we have clustered the data into four major groups (groups I to IV) (Fig. 9). As we can see from Fig. 9, group IV shows that the majority of samples could be found in all matrices, *i.e.*, dog, rat (plasma and urine) and whole-cell microorganism media, and they produced frequently the same metabolites.

In total, three different classes of biological matrix were compared, and PCA analysis clearly distinguished the common metabolites in the three different species. Minimal differences on production of metabolites were founded among them. MFA analysis accomplish to PCA pointed out group IV as the major group of produced metabolites, including the **M1**, **M2**, **M3** and **M8,** in all mammalian species and filamentous fungi strains studied here.

It is noteworthy that our final prediction model, presented in Fig. 3, summarizes all the predicted metabolites for (**1**) in our study, using a combination of computational methods, *i.e.*, MetaPrint2D, docking, MD simulations and QM calculations. Moreover, the metabolites that were found experimentally were marked in this figure (as red and orange). The proposal of the integration of different levels of theory for in silico prediction of drug metabolism reported herein qualitatively predicted the metabolites of (**1**), which was supported by the experimental assays.



	+MRM*	Conversion (Area) 0h	24h	%
Blank (control)	MRM transition identical to the parent prototype and all metabolites	0.00	0.00	0.00
LASSBio-294 (Std)	275.2 / 149.1	57800.00	24237.18	41.93
M1	291.1 / 165.0	0.00	35721.48	36.61
M2/M3	291.1 / 191.0	0.00	16158.12	16.56
M8	263.0 / 245.0	0.00	4781.47	4.90

* Electrospray ionization (ESI) operating in positive ion multiple reaction monitoring (MRM) mode by monitoring the ion transitions

Fig. 8 A two-step LC-MS approach employed in the metabolic fingerprint profiling. The samples were analyzed using time-of-flight (TOF) mass spectrometry (MS) followed by targeted identification of differentially produced metabolites using quadrupole time-of-flight (Q-TOF) MS/MS. (**a**) shows the EIC for metabolites **M1**, **M2**, **M3** and **M8**. (**b**) gives the relative proportion of formation of monitored the ion transitions of the metabolites above by multiple reaction monitoring (+MRM)

Conclusions

Drug metabolism in the context of drug discovery is a complex process that includes issues relating to metabolic stability, enzyme identification, metabolite identification, reactive metabolites, and enzyme inhibition properties. All of these parameters are interrelated and need to be considered in parallel in the development of new therapeutic agents. Novel technologies that increase the probability of making the right choice early save resources and promote safety, efficacy and profitability.

In this work, we described the application of a combined methodology to explore the site of metabolism prediction of a new cardioactive drug prototype, (LASSBio-294, **1**), using MetaPrint2D, an improved algorithm for ligand-based site of metabolism prediction, to predict the most likely metabolites, combined with structure-based methodologies using docking, molecular dynamics and quantum mechanical calculations, to predict the binding of the substrate to CYP2C9 enzyme, to estimate the binding free of energy and to study the energy profiles for the oxidation of (**1**), along with comparison to a metabolic fingerprint

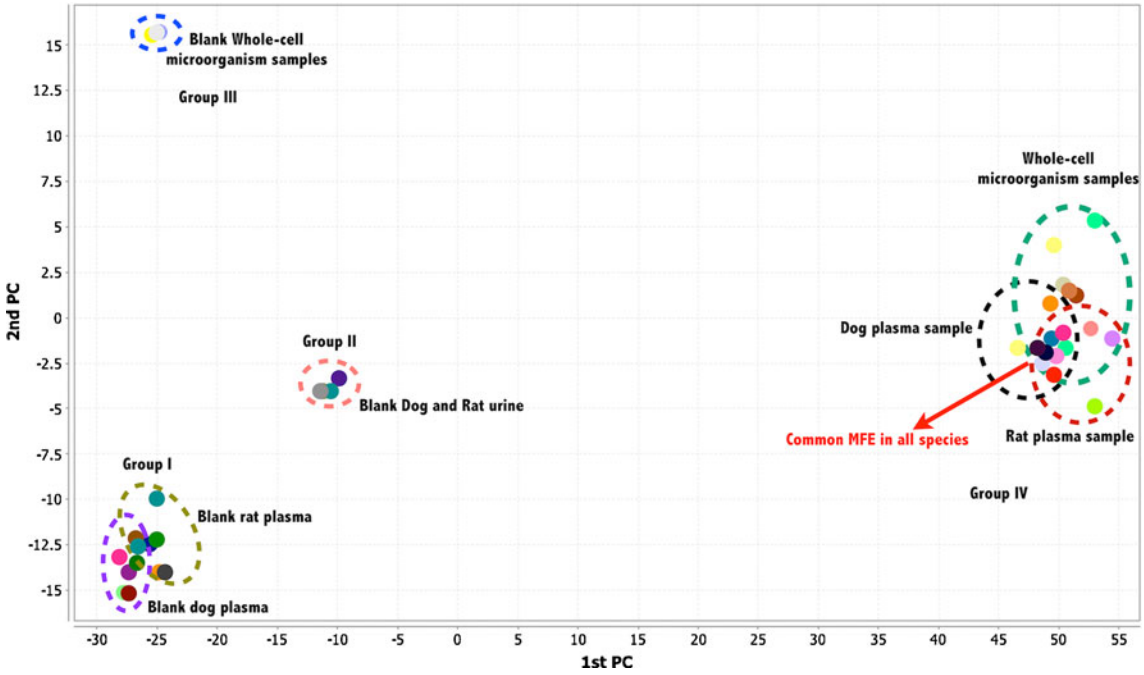


Fig. 9 PCA plot of samples for 16 h LC-MS positive runs after the statistical analysis. PCA plot finds the relationships beyond pair-wise comparison in complex experimental data and enables biological interpretation through pathway analysis clustering the data samples in distinct groups (I to IV)

profiling using LC-MS analysis. The results obtained using the computational methods gave valuable information about the probable metabolites of (**1**) (qualitatively) and also about the important interactions of this lead compound with the amino acid residues of the active site of CYP2C9. Moreover, using a combination of different levels of theory sheds light on the understanding of (**1**) metabolism by CYP2C9 and its mechanisms.

The metabolic fingerprint profiling of (**1**) has shown that three major metabolites were founded both in vitro and in vivo studies in highest concentration (**M1**, **M2**, **M3**) and one in lower concentration (**M8**). Therefore, our computational study presented that Metaprint2D predicted M1 and M8 as major metabolites. The docking predicted M1, M2, M3, and M8 and QM calculations predicted the most favorable energy barriers during the formation of M1. So, in conclusion, M1 was correctly predicted, whereas M2 and M3 would be expected to be minor products, and M8 should be seen in small quantities.

The approach presented here has afforded new opportunities to improve metabolite identification strategies. This method can be also applicable for the qualitative prediction of drug metabolism mediated by not only CYP2C9 but also other CYP450 family enzymes.

Acknowledgments The authors would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) for financial support and fellowships.

References

1. Jones BC, Middleton DS, Youdim K (2009) Cytochrome P450 metabolism and inhibition: analysis for drug discovery. Prog Med Chem 47:239–263. doi:10.1016/S0079-6468(08)00206-3
2. Vistoli G, Pedretti A, Testa B (2008) Assessing drug-likeness–what are we missing? Drug Discov Today 13:285–294. doi:10.1016/j.drudis.2007.11.007
3. Lewis DFV, Ito Y (2008) Human cytochromes P450 in the metabolism of drugs: new molecular models of enzyme-substrate interactions. Expert Opin Drug Metab Toxicol 4:1181–1186. doi:10.1517/17425250802352412
4. de Montellano PRO (2010) Cytochrome P450: structure, mechanism, and biochemistry, 3rd edn. Springer, New York
5. Baranczewski P, Stanczak A, Sundberg K, Svensson R, Wallin A, Jansson J, Garberg P, Postlind H (2006) Introduction to in vitro estimation of metabolic stability and drug interactions of new chemical entities in drug discovery and development. Pharmacol Rep 58:453–472
6. de Graaf C, Pospisil P, Pos W, Folkers G, Vermeulen NPE (2005) Binding mode prediction of cytochrome p450 and thymidine kinase protein-ligand complexes by consideration of water and rescoring in automated docking. J Med Chem 48:2308–2318. doi:10.1021/jm049650u
7. Stjernschantz E, Vermeulen NPE, Oostenbrink C (2008) Computational prediction of drug binding and rationalisation of selectivity towards cytochromes P450. Expert Opin Drug Metab Toxicol 4:513–527. doi:10.1517/17425255.4.5.513
8. Miners JO, Birkett DJ (1998) Cytochrome P4502C9: an enzyme of major importance in human drug metabolism. Br J Clin Pharmacol 45:525–538. doi:10.1046/j.1365-2125.1998.00721.x
9. Sykes MJ, McKinnon RA, Miners JO (2008) Prediction of metabolism by cytochrome P450 2C9: alignment and docking studies of a validated database of substrates. J Med Chem 51:780–791. doi:10.1021/jm7009793
10. Wester MR, Yano JK, Schoch GA, Yang C, Griffin KJ, Stout CD, Johnson EF (2004) The structure of human cytochrome P450 2C9 complexed with flurbiprofen at 2.0-A resolution. J Biol Chem 279:35630–35637. doi:10.1074/jbc.M405427200
11. Usmani KA, Karoly ED, Hodgson E, Rose RL (2004) In vitro sulfoxidation of thioether compounds by human cytochrome P450 and flavin-containing monooxygenase isoforms with particular reference to the CYP2C subfamily. Drug Metab Dispos 32:333–339. doi:10.1124/dmd.32.3.333
12. Khan MTH (2010) Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. Curr Drug Metab 11(4):285–295. doi:10.2174/138920010791514306
13. Sun H, Scott DO (2010) Structure-based drug metabolism predictions for drug design. Chem Biol Drug Des 75:3–17. doi:10.1111/j.1747-0285.2009.00899.x
14. Figueiredo JM, Camara CD, Amarante EG, Miranda ALP, Santos FM, Rodrigues CR, Fraga CAM, Barreiro EJ (2000) Design and synthesis of novel potent antinociceptive agents: methyl-imidazolyl N-acylhydrazone derivatives. Bioorg Med Chem 8:2243–2248. doi:10.1016/S0968-0896(00)00152-8
15. Zapata-Sudo G, Sudo RT, Maronas PA, Silva GLM, Moreira OR, Aguiar MIS, Barreiro EJ (2003) Thienylhydrazone derivative increases sarcoplasmic reticulum Ca2+ release in mammalian skeletal muscle. Eur J Pharmacol 470:79–85. doi:10.1016/S0014-2999(03)01757-6
16. Costa DG, da Silva JS, Kummerle AE, Sudo RT, Landgraf SS, Caruso-Neves C, Fraga CAM, de Lacerda Barreiro EJ, Zapata-Sudo G (2010) LASSBio-294, A compound with inotropic and lusitropic activity, decreases cardiac remodeling and improves Ca2(+) influx into sarcoplasmic reticulum after myocardial infarction. Am J Hypertens 23:1220–1227. doi:10.1038/ajh.2010.157
17. Carlsson L, Spjuth O, Adams S, Glen RC, Boyer S (2010) Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse. BMC Bioinforma 11:362–362. doi:10.1186/1471-2105-11-362
18. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) The development and use of quantum-mechanical molecular-models.76. Am1 - a new general-purpose quantum-mechanical molecular-model. J Am Chem Soc 107:3902–3909. doi:10.1021/ja00299a024
19. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1996) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 118:2309–2309. doi:10.1021/ja955032e
20. Oda A, Yamaotsu N, Hirono S (2005) New AMBER force field parameters of heme iron for cytochrome P450s determined by quantum chemical calculations of simplified models. J Comput Chem 26:818–826. doi:10.1002/jcc.20221
21. Clark AM, Labute P (2007) 2D depiction of protein-ligand complexes. J Chem Inf Model 47:1933–1944. doi:10.1021/ci7001473
22. Labute P (2008) The generalized Born/volume integral implicit solvent model: estimation of the free energy of hydration using London dispersion instead of atomic surface area. J Comput Chem 29:1693–1698. doi:10.1002/jcc.20933

23. young dc (2009) computational drug design: a guide for computational and medicinal chemists 1 Har/Cdr edn. Wiley-Interscience, Hoboken. N.J
24. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W (2010) Atomic-level characterization of the structural dynamics of proteins. Science 330:341–346. doi:10.1126/science.1187409
25. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118:11225–11236. doi:10.1021/ja9621760
26. Paterlini MG, Ferguson DM (1998) Constant temperature simulations using the Langevin equation with velocity Verlet integration. Chem Phys 236:243–252. doi:10.1016/S0301-0104(98)00214-6
27. SZYBKI (2010). 1.5.1 edn. OpenEye Scientific Software Inc, Santa Fe, NM
28. Wlodek S, Skillman AG, Nicholls A (2010) Ligand entropy in Gas-Phase. Upon solvation and protein complexation. Fast estimation with Quasi-Newton Hessian. J Chem Theory Comput 6:2140–2152. doi:10.1021/ct100095p
29. Jaguar (2009) 7.6 edn. Schrodinger, LLC, New York
30. Braga RC, Tôrres ACB, Persiano CB, Alves RO, Fraga CAM, Barreiro EJ, de Oliveira V (2011) Determination of the cardioactive prototype LASSBio-294 and its metabolites in dog plasma by LC-MS/MS: Application for a pharmacokinetic study. J Pharm Biomed Anal 55:1024–1030. doi:10.1016/j.jpba.2011.02.031
31. Carneiro EO, Andrade CH, Braga RC, Torres ACB, Alves RO, Liao LM, Fraga CAM, Barreiro EJ, de Oliveira V (2010) Structure-based prediction and biosynthesis of the major mammalian metabolite of the cardioactive prototype LASSBio-294. Bioorg Med Chem Lett 20:3734–3736. doi:10.1016/j.bmcl.2010.04.073
32. Ts P, Castillo S, Villar-Briones A, Oresic M (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinforma 11:395. doi:10.1186/1471-2105-11-395
33. Castro-Perez JM (2007) Current and future trends in the application of HPLC-MS to metabolite-identification studies. Drug Discov Today 12:249–256. doi:10.1016/j.drudis.2007.01.007
34. Bathelt CM, Zurek J, Mulholland AJ, Harvey JN (2005) Electronic structure of compound I in human isoforms of cytochrome P450 from QM/MM modeling. J Am Chem Soc 127:12900–12908. doi:10.1021/ja0520924
35. Czodrowski P, Kriegl JM, Scheuerer S, Fox T (2009) Computational approaches to predict drug metabolism. Expert Opin Drug Metab Toxicol 5:15–27. doi:10.1517/17425250802568009
36. Boyer S, Arnby CH, Carlsson L, Smith J, Stein V, Glen RC (2007) Reaction site mapping of xenobiotic biotransformations. J Chem Inf Model 47:583–590
37. Rydberg P, Vasanthanathan P, Oostenbrink C, Olsen L (2009) Fast prediction of cytochrome P450 mediated drug metabolism. ChemMedChem 4:2070–2079. doi:10.1002/cmdc.200900363
38. Park J-Y, Harris D (2003) Construction and assessment of models of CYP2E1: predictions of metabolism from docking, molecular dynamics, and density functional theoretical calculations. J Med Chem 46:1645–1660. doi:10.1021/jm020538a
39. Lewis DFV (2000) On the recognition of mammalian microsomal cytochrome P450 substrates and their characteristics - Towards the prediction of human P450 substrate specificity and metabolism. Biochem Pharmacol 60:293–306. doi:10.1016/S0006-2952(00)00335-X
40. Fraga AGM, Fraga CAM, Barreiro EJ, Romeiro NC Perfil metabolico in silico de Prototipo N-Acilidrazonico Cardioativo. In: 30th Reunião Anual da Sociedade Brasileira de Química, Águas de Lindóia - SP, 2007. Abstracts of Papers. Sociedade Brasileira de Química, pp MD-053
41. Tarcsay A, Rb K, GrM K (2010) Site of metabolism prediction on cytochrome P450 2C9: a knowledge-based docking approach. J Comput-Aided Mol Des 24:399–408. doi:10.1007/s10822-010-9347-3
42. de Groot MJ, Alex AA, Jones BC (2002) Development of a combined protein and pharmacophore model for cytochrome P450 2C9. J Med Chem 45:1983–1993. doi:jm0110791[pii]
43. Clodfelter KH, Waxman DJ, Vajda S (2006) Computational solvent mapping reveals the importance of local conformational changes for broad substrate specificity in mammalian cytochromes P450. Biochemistry 45:9393–9407. doi:10.1021/bi060343v
44. Polgar T, Menyhard DK, Keseru GM (2007) Effective virtual screening protocol for CYP2C9 ligands using a screening site constructed from flurbiprofen and S-warfarin pockets. J Comput-Aided Mol Des 21:539–548. doi:10.1007/S10822-007-9137-8
45. Williams PA, Cosme J, Ward A, Angove HC, Matak Vinkovi*fá* D, Jhoti H (2003) Crystal structure of human cytochrome P450 2C9 with bound warfarin. Nature 424:464–468. doi:10.1038/nature01862
46. Rydberg P, Rod TH, Olsen L, Ryde U (2007) Dynamics of water molecules in the active-site cavity of human cytochromes P450. J Phys Chem B 111:5445–5457. doi:10.1021/jp070390c
47. Santos R, Hritz J, Oostenbrink C (2010) Role of water in molecular docking simulations of cytochrome P450 2D6. J Chem Inf Model 50:146–154. doi:10.1021/ci900293e
48. Mulholland AJ (2005) Modelling enzyme reaction mechanisms, specificity and catalysis. Drug Discov Today 10:1393–1402. doi:10.1016/S1359-6446(05)03611-1
49. Shaik S, Cohen S, Wang Y, Chen H, Kumar D, Thiel W (2010) P450 enzymes: their structure, reactivity, and selectivity-modeled by QM/MM calculations. Chem Rev 110:949–1017. doi:10.1021/cr900121s
50. Shaik S, Milko P, Schyman P, Usharani D, Chen H (2011) Trends in aromatic oxidation reactions catalyzed by Cytochrome P450 Enzymes: a valence bond modeling. J Chem Theory Comput 7:327–339. doi:10.1021/ct100554g
51. Schoneboom JC, Lin H, Reuter N, Thiel W, Cohen S, Ogliaro F, Shaik S (2002) The elusive oxidant species of cytochrome P450 enzymes: characterization by combined quantum mechanical/molecular mechanical (QM/MM) calculations. J Am Chem Soc 124:8142–8151. doi:ja026279w[pii]
52. Kellner DG, Hung SC, Weiss KE, Sligar SG (2002) Kinetic characterization of compound I formation in the thermostable cytochrome P450 CYP119. J Biol Chem 277:9641–9644. doi:10.1074/jbc.C100745200
53. Crestoni ME, Fornarini S, Lanucara F (2009) Oxygen-atom transfer by a naked manganese(V)-Oxo-Porphyrin complex reveals axial ligand effect. Chem-Eur J 15:7863–7866. doi:10.1002/Chem.200901361
54. Chiavarino B, Cipollini R, Crestoni ME, Fornarini S, Lanucara F, Lapi A (2008) Probing the Compound I-like reactivity of a bare high-valent oxo iron porphyrin complex: the oxidation of tertiary amines. J Am Chem Soc 130:3208–3217. doi:10.1021/ja077286t
55. Lonsdale R, Ranaghan KE, Mulholland AJ (2010) Computational enzymology. Chem Commun 46:2354–2372. doi:10.1039/B925647d
56. Rydberg P, Ryde U, Olsen L (2008) Sulfoxide, sulfur, and nitrogen oxidation and dealkylation by Cytochrome P450. J Chem Theor Comput 4:1369–1377. doi:10.1021/ct800101v
57. Watanabe Y (2001) Alternatives to the oxoferryl porphyrin cation radical as the proposed reactive intermediate of cytochrome P450: two-electron oxidized Fe(III) porphyrin derivatives. J Biol Inorg Chem 6:846–856. doi:10.1007/s007750100278
58. Porro CS, Sutcliffe MJ, de Visser SP (2009) Quantum mechanics/molecular mechanics studies on the sulfoxidation of dimethyl sulfide by compound I and compound 0 of cytochrome P450:

which is the better oxidant? J Phys Chem A 113:11635–11642. doi:10.1021/jp9023926

59. Bathelt CM, Mulholland AJ, Harvey JN (2008) QM/MM modeling of benzene hydroxylation in human cytochrome P450 2C9. J Phys Chem A 112:13149–13156. doi:10.1021/jp8016908
60. de Visser SP, Shaik S (2003) A proton-shuttle mechanism mediated by the porphyrin in benzene hydroxylation by cytochrome P450 enzymes. J Am Chem Soc 125:7413–7424. doi:10.1021/Ja034142f
61. Costa EMDB, Pimenta FC, Luz WC, de Oliveira V (2008) Selection of filamentous fungi of the Beauveria genus able to metabolize quercetin like mammalian cells. Braz J Microbiol 39:405–408. doi:10.1590/S1517-83822008000200036
62. Pazini F, Menegatti R, Sabino JR, Andrade CH, Neves G, Rates SMK, Noel F, Fraga CAM, Barreiro EJ, de Oliveira V (2010) Design of new dopamine D2 receptor ligands: biosynthesis and pharmacological evaluation of the hydroxylated metabolite of LASSBio-581. Bioorg Med Chem Lett 20:2888–2891. doi:10.1016/J.Bmcl.2010.03.034
63. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G (2005) METLIN: a metabolite mass spectral database. Ther Drug Monit 27:747–751

J Mol Model (2012) 18:2079–2098
DOI 10.1007/s00894-011-1232-z

ORIGINAL PAPER

A systematical comparison of DFT methods in reproducing the interaction energies of halide series with protein moieties

Xiuhong Liu · Peng Zhou · Zhicai Shang

Received: 13 March 2011 / Accepted: 24 August 2011 / Published online: 9 September 2011

Abstract A systematic theoretical investigation on the interaction energies of halogen-ionic bridges formed between halide ions and the polar H atoms bonded to N of protein moieties has been carried out by employing a variety of density functional methods. In this procedure, full geometry optimizations are performed at the Møller-Plesset second-order perturbation (MP2) level of theory in conjunction with the Dunning's augmented correlation-consistent basis set, aug-cc-pVDZ. Subsequently, two distinct basis sets, i.e. 6-311++G(df,pd) and aug-cc-pVTZ, are employed in the following single-point calculations so as to check the stability of the results obtained at the different levels of DFT. The performance of DFT methods has been evaluated by comparing the results with those obtained from the rigorous MP2 theory. It is shown that the B98, B97-1, and M05 give the lowest root-mean-square error (RMSE) for predicting fluoride-binding energies, M05-2X, MPW1B95, and MPW1PW91 have the best performance in reproducing chloride-binding energies, B97-1, PBEKCIS, and PBE1KCIS present the optimal result for bromide-binding energies, while B97-1, MPW1PW91, and TPSS perform most well on iodide-binding energies. The popular B3LYP functional seems to be quite modest for studying halide-protein moiety interactions. In addition, the PBE1KCIS functional provide accuracies close to the computationally expensive MP2 method for the calculation of interaction energies of all halide-binding systems.

X. Liu · P. Zhou · Z. Shang (✉)
Department of Chemistry, Zhejiang University,
Hangzhou 310027, China
e-mail: shangzc@zju.edu.cn

Keywords DFT functional · Halide-binding energy · Halide ion · Halide motif · Protein · Quantum-mechanics

Introduction

Water is the most common solvent and many of its unique chemical and physical properties are determined by the hydrogen-bonded network. Halide ions, especially chloride ion, are among the most common anions present in nature, and consequently numerous works have been addressed for ascertaining the nature of halide ion-water interactions [1]. It is widely believed that the addition of halide ions to water engenders structural changes in the hydrogen-bond network well beyond the adjacent shell of solvating molecules, which could affected the physicochemical properties of aqueous solutions in viscosity, osmotic pressure, activity coefficient, lowering of freezing point, refractive index and optical rotation [2–4]. Apart from this, in recent years, there have been a number of experimental and theoretical results showing that halide ions are of fundamental importance in chemical and biological systems when in studies of protein stability and unfolding, enzymatic activity, membrane permeability, in molecular forces in colloid science, ion binding to micelles, ionic microemulsions, and so on [5–11]. Originally, it was thought that ion's influence on water's or protein's properties was caused at least in part by continually forming and breaking hydrogen bonds through concerted hydrogen bond rearrangements [12]. Both Ninham et al. and Jungwirth et al. have done many works about the specific ion effects on glycan, protein and colloid. They are convinced that the dispersion forces are likely the foremost driving forces for ion-specific surface phenomena [13–16]. Ninham et al. found that the majority of the stabilization energy between ions and protein charge groups or between ions and macroions stems from electrostatic force, and the specific ion effects (also known as the Hofmeister effect [17]) on protein stability could be explained by incorporating the ionic dispersion potentials into classical double-layer theory [14, 18, 19]. In addition, Jungwirth et al. pointed out that the polarizability is probably an important aspect for describing ion's behavior [20, 21]. Recent time-resolved and thermodynamic studies of water molecules in salt solution, however, suggested that, instead of remodeling water structure through ions, direct ion-protein interactions as well as the ionic interactions with water molecules that are bound to the proteins seem to be also responsible for these effects [5].

Recently, by exhaustively surveying all high-quality protein crystal structures deposited in the current Protein Data Bank (PDB), we have found a striking magnitude (>10 000) of halide ions located in the interior or attached at the surface of proteins, from which we identified more than 6000 halide motifs which we named halogen-ionic bridges could show a potential role in conferring stability and specificity for the structure of proteins and their complexes with small ligands and nucleic acids [22]. In these halide motifs, the halogen ions can bridge between the spatially vicinal moieties in biomolecules, thus the role of a water molecule in mediating the hydrogen-bond network in biomolecules can be functionally replaced by a halogen ion. This replacement is feasible because the halogen-ionic bridge stabilization energy is estimated to be generally more than 100 kcal·mol^{-1} for gas-phase states or about 20 kcal·mol^{-1} for solution conditions, which is much greater than that found in sophisticated water-mediated (< 10 kcal·mol^{-1}) and salt (~ 3.66 kcal·mol^{-1}) bridges [22]. In addition, we also observed that most structured halide motifs packed in protein crystals show a substantially stabilizing effect on the protein architecture through direct noncovalent interactions with their context [23].

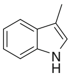
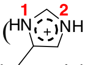
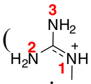
Several specific intermolecular forces involved in ligand recognition and binding by protein receptors have been investigated in detail by means of the hybrid QM/MM methodology [24–26]. These works confirmed that, if reasonably collocated with a MM context, it is possible to apply the expensive QM method to treat the nonbonding interactions of interest in the whole biomacromolecular framework. In our previous studies, we have applied a two-layer ONIOM-based QM/MM methodology to investigate the role and significance of the protein-ligand complexes with structured halogen-ionic bridges in biological context [22, 23]. These works have been done with some very empirical rules for allocating seemingly appropriate DFT theories to the QM layer, leading to the significant difficulty in assessing the reliability of obtained results. This is because in biomolecular systems, beyond the electrostatic effect, the non-electrostatic factors, especially the dispersion potential is also a critical factor governing the interaction behavior of halides with protein moieties. In other words, dispersion forces play an important role in halide ion-protein interactions, but it cannot be treated properly using some popular functionals, such as the B3LYP [27].

The density functional theory (DFT) has become the one of most popular methods in computational chemistry community. Because of its dramatic savings in computational effort, DFT can be easily applied to considerably large molecular systems. Nevertheless, it was also noted that, owing to the local-corrected functional (LCF) used, DFT normally underappreciated the dispersion force [28] which has been recognized as an important factor affecting the physicochemical behavior of the diffuse, polarizable anions [14, 15, 19]. Dispersion is an intermolecular electron correlation effect, and the simplest quantum mechanical (QM) method for electron correlation, second-order Møller-Plesset perturbation theory (MP2), describes dispersion well. Unfortunately, MP2 calculations are far more expensive than DFT calculations. Its steep (N^5) scaling behavior prevents the application of MP2 to relatively large biochemical entities. In the current condition, it is difficult to calculate the effect of halide motifs on biomacromolecular systems solely using the *ab initio* MP2 method, not to mention the much higher-level correlated CCSD(T) method. Hence, it is significant to explore strategies that are computationally less demanding but describe these interactions with a similar accuracy as MP2 or higher levels of theory. A reasonable alternative is offered by DFT methods, albeit the DFT approaches are less reliable due to the lack of an appropriate description of the dispersion aspect. Several previous studies have pointed out that the DFT methods are able to provide the optimum compromise between the accuracy and cost of computation [29–33]. Moreover, considering the vital significance of halide-binding in proteins, it would be worth making large biologically relevant systems associated with halide-binding interactions tractable at a relatively reliable quantum-mechanical level, which could reproduce the intermolecular dispersion potentials best for a series of halide motifs in biological context as that obtained at the MP2 level.

To better understand the significance of halide motifs in functioning to protein and other biosystems, in the current work, we launched a comprehensive investigation on the interaction energetic properties of well-characterized halide motifs forming by halide anions and electrophilic groups in protein within the whole framework of protein-halide complexes with particularly diverse DFT methods. These halide motifs were found to ubiquitously exist in the interior of proteins by exhaustively surveying all the protein crystal structures deposited in the current Protein Data Bank (PDB). The MP2 method is used as the reference because it can well describe the long-range correlation effects that are usually missing from the popular DFT functionals. Apart from this, we evaluated the performance of 31 sophisticated DFT methods, one noncorrelation *ab initio* theory (Hartree-Fock), two semi-empirical methods (AM1 and PM3) and one mechanical force field (UFF) in comparison with the results obtained from the accurate but expensive correlation *ab initio* MP2 theory. Furthermore, several effective DFT functionals from the comparison is applied to ONIOM-based QM/MM calculations on real systems to render its feasibility.

Materials and methods

The geometry of halide motifs

To inspect the interaction profile of halogen ions with the protein moieties of interest, a thorough search for all the low-lying energy structures of halogen series (F^-, Cl^-, Br^-, and I^-) binding to electrophilic hydrogen atoms of seven protein groups, respectively, modeled by methanol (CH_3OH) (for hydroxyl group), N-methylformamide ($HCONHCH_3$) (for main chain's amide), acetamide (CH_3CONH_2) (for side chain's amide), 3-methyl-1H-indole () (for main chain's tryptophan), methylammonium ($CH_3NH_3^+$) (for lysine's ammonium), 4-methylimidazolium () (for histidine's imidazolium), and N-methylguanidinium () (for arginine's guanidinium) were studied. In our previous studies, we got the statistics about the distribution states of halogen ions around different protein moieties retrieved from the PDB [19], as shown in the first row in Fig. 1, and the low-lying energy structures of Cl^- in complex with corresponding protein moieties obtained using a thorough MP2/aug-cc-pVDZ search were depicted in the second row in Fig. 1. In all cases, the halogen ions in the ligands were observed to participate in halogen-ionic bridges and interact with the hydrogen atoms of functional groups of proteins. In geometry optimization procedure, the complex model systems were fully optimized at the MP2/aug-cc-pVDZ (or MP2/Lanl2DZ + (df) for iodine) level to avoid secondary interactions between halogen ions and other hydrogen atoms in these systems.

Quantum-mechanical (QM) calculations

All QM calculations were carried out using a locally modified Gaussian suite of program [33, 34]. In this article, we tested 31 practical DFT methods as follows: (a) six GGAs: BP86 [35, 36], BLYP [36, 37], BPW91 [36, 38], PW91 [38], HCTH [39], MPWLYP [36, 40], (b) two meta GGA methods: PBEKCIS [41–44], TPSS [45, 46]; (c) twelve hybrid GGA methods: B3LYP [37, 47, 48], B3P86 [35, 47], B3PW91 [38, 47], BH&HLYP [49], B97-1 [39], B98 [50], MPW1PW91 [40], MPW1K [51], MPW3LYP [37, 40, 52], O3LYP [53, 54], PBE1PBE [41], and X3LYP [55]; (d) ten hybrid meta GGA methods: MPW1B95 [52], MPWB1K [52], MPW1KCIS [56], MPWKCIS1K [56], TPSS1KCIS [42–46, 57], PBE1KCIS [27, 41, 43], M05 [58], M05-2X [29], M06 [59], and M06-HF [60]. In particular, we assessed one LSDA: SVWN5 [61, 62]. Since the theory behind the various DFT functionals was clarified fairly well in the original literature, we herein refer the readers to these references for those details. Except two more recent DFT methods, M06 and M06-HF, these others tested density functionals that selected here for evaluation were based on at least one of the following reasons: (i) has been used to study the nonbonding interactions in biomolecule systems [25, 27, 31, 52, 55, 63, 64]; (ii) has been used to investigate the bondings involving halogens or halides [65–69], such as halogen-water-hydrogen bridges [70], halogen bondings [71], fluorine bondings [24]; (iii) has been used to determine the intermolecular interaction potentials [72]; (iv) has been used to study the hydrogen bonding systems [73–77]; (v)

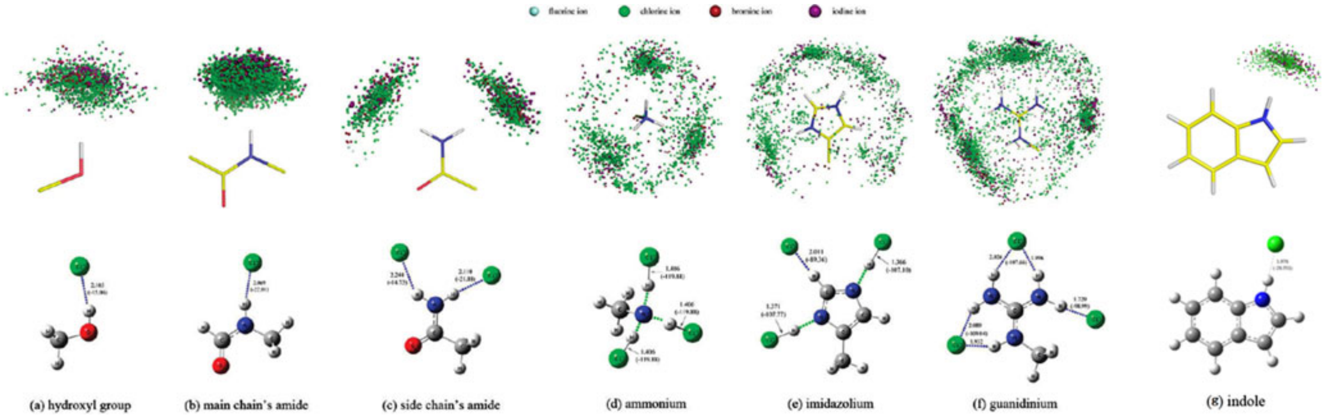


Fig. 1 This figure is modified from our recent publication[23]

has been used to predict the binding energies of some dispersion-bound complexes [78].

Two basis sets, 6-311++G(df,pd) and the Dunning's augmented correlation consistent basis set, aug-cc-pVTZ, were applied in the calculations so as to check the stability of the results obtained at the DFT levels of theory. Since either 6-311++G(df,pd) or aug-cc-pVTZ is unavailable for iodine, the Lanl2DZ basis set, augmented by a set of *d* and *f* polarization functions (exponents 0.292 and 0.441, respectively) and *s* and *p* diffuse functions (exponents 0.0569 and 0.0330, respectively), *abbr*. Lanl2DZ + (df), was used for I^-. This large version of a valence electron orbit seems to be necessary for reliably describing the outer electronic structure of diffuse anions, and previous theoretical calculations which used this modified effective core potential (ECP) basis set have been shown to give reasonably good results for the I^--participating S_N2 reactions [79] and the OCS ··· I^- van der Waals complexes [80].

Accurate estimation of nonbonded intermolecular potential energies has long been a challenge in the computational chemistry area. On the basis of a systematic study on a set of nonbonded complexes, Rappe and Bernstein [81] concluded that low levels of correlation theory such as the second-order Møller-Plesset perturbation theory (MP2) can account for the full range of intermolecular interactions, and the accuracy mainly lies in the convergence with respect to the basis set expansion. According to this claim, and also in consideration of the size of the model complexes and the available computer resources, we used MP2 theory to account for the correlation energy and focused on the convergence. In addition, a detailed examination of energetic profile of the simplest model systems, water molecule (H_2O) in complex with four kinds of halogen ions (F^-, Cl^-, Br^-, and I^-) were carried out by using the MP2 and CCSD(T) methods with two distinct basis sets, 6-311++G(df, pd) and aug-cc-pVTZ (Table 1). As can be seen, the results shown that the calculated interaction energies, ΔE_{int}, are very close to the experimental interaction energies (in parentheses). Because the CCSD(T) calculations even with the smaller 6-31G* basis set are extremely time-consuming, this particularly stringent method was not considered here as the reference method.

In order to comprehensively investigate the performance of lower-level DFT calculations in reproducing halogen-ionic bridges energies obtained at the expensive MP2 level, other methods, including one noncorrelation *ab initio* theory (HF) and two semi-empirical methods (AM1 and PM3) were tested by performing calculations on seven types of short halide-binding observed in crystal structures of protein-ligand complexes deposited in the PDB. The binding energies (ΔE_{int}) calculated using QM methods were obtained under the indirect supermolecule approach [82], this method considers the difference between the total energy of the complex and the sum energies of isolated monomers, viz. $\Delta E_{\mathrm{int}} = E_{\mathrm{complex}} - (E_{\mathrm{monomer1}} + E_{\mathrm{monomer2}})$, and the associated basis set superposition error (BSSE) calculated by MP2, DFT, and HF methods was eliminated by means of the counterpoise strategy [83].

Database survey

Up to January 2010, there were 3391 protein records and 133 nucleic acid entries (solved at 3 Å or better) deposited in the PDB in which at least one nonbonded halogen ion is contained. All selected complex structures were subjected to a pretreatment procedure, that is, (i) remove water molecules, metal ions, and other cofactors, except halogen ions and small organic ligands; (ii) using the newly released SCWRL4 program [84] to repair the missing side chains of protein residues; (iii) according to the dictionary secondary structure of proteins (DSSP) protocol [85] to assign the secondary structure class for protein residues; (iv) using the REDUCE program [86] to add hydrogen atoms for all protein and nucleic acid heavy

Table 1 Energetic parameters for the complexes of halogen ions with H_2O serving as hydrogen donor

Complex	ΔE_{int} [a] (kcal·mol^{-1})			
	Calculated			Experimental
	MP2/6-311++G(df, pd)	MP2/aug-cc-PVTZ	CCSD(T) /aug-cc-PVTZ	
water-F^-	−21.87	−22.47	−26.61	−23.32 [c]
water-Cl^-	−13.90	−14.94	−14.11	−14.71 [d]
water-Br^-	−12.03	−12.70	−11.99	−11.71 [d]
water-I^- [b]	−9.47	−10.40	−9.80	−10.30 [d]

[a] ΔE_{int} (kcal·mol^{-1}) calculated at the MP2/6-311++G(df,pd) and MP2/aug-cc-pvtz theory and, if exist, experimental interaction energies

[b] Lanl2DZ + (df) basis set for iodine

[c] From gas-phase equilibrium measurements by high-pressure mass spectrometry [94].

[d] From gas-phase equilibrium measurements by pulsed electron beam high-pressure mass spectrometry [95]

atoms (REDUCE was adopted here because this program was tested in our previous study to be capable of precisely reproducing the neutron diffraction-determined hydrogen's positions [87]); (v) using the PROPKA 2.0 program [88] to define the protonation state of all charged residues at PH=7.0, and (vi) using the I-INTERPRET program [89] to interpret the structural information of small ligands, which are marked by header 'HETATM' in the PDB files. This program reads an assembly of ligands in standard PDB format and writes a MOL2 file in which the atomic states, connection manners, and neutral/charged hydrogen's positions are assigned in a considerable accuracy for these ligands. After that, the following criteria were defined to describe the effective biological interactions involving halogen ions: (i) for an uncharged polar group, an ellipsoid with its center at the polar H atom and its semi-minor/semi-major axis of 3.0/3.5 Å was constructed. Only those halogen ions occurring within the ellipsoidal space and with the forming angle θ>120° were considered; and (ii) for a charged basic group, the halogen ions with their distances, D, to any one of the heavy atoms in the group less than 4.5 Å were considered. In this way, a halogen-ionic bridge can be readily defined as the entity in which a halogen ion effectively interacts with two or more biomoleuclar groups simultaneously; the number of the groups participating in bridging was called the branch degree of this halogen-ionic bridge.

ONIOM-based QM/MM calculations on real systems

In this work, we also implemented the hybrid quantum mechanical/molecular mechanical (QM/MM) calculations with the help of the Gaussian 03 suite of programs [34] to examine the structural and energetic properties for several well-characterized halide motifs formed within the whole framework of real protein-halide complexes. The structured halide ion and the protein residues that are directly bound to the halide ion were included in the QM layer and treated with a high level of density functional theory (DFT/Lanl2DZ for I^- or DFT/6-31 + G* for other atoms/ions, from F^- to I^-, the DFT functional is B98, M05-2X, B97-1, and MPW1PW91, respectively, which were obtained from Tables 7, 8, 9, and 10), while the of the rest atoms of protein in MM layer were modeled by a low level of molecular force field (AMBER parm96) [90]. The generalized AMBER force field (GAFF) package was used for parameters not found in the AMBER force field [91]. In addition, to mimic the real environment of protein-ligand interactions, all of the water molecules in the crystal structures were retained.

After QM/MM optimization, the QM layer of the model was selected for higher-level single-point energy calculations, which were performed by using the DFT (from F^- to I^-, the DFT functional is B98, M05-2X, B97-1, and MPW1PW91, respectively) and MP2 methods with two basis sets, 6-311++G(df,pd) and aug-cc-pVTZ (or Lanl2DZ + (df) for iodine). The interaction energy (ΔE_{int}) was calculated as described in the section of Quantum-Mechanical (QM) calculation.

Results and discussion

Halide series with protein moieties

Using the criteria described in the section of database survey, we selected the halide-binding with the polar hydrogen atoms and positively charged groups of proteins to conduct a systematical comparison of the DFT methods in reproducing the interaction energies of halide series with protein moieties (Fig. 1). From Fig. 1 it is seen that, as for polar moieties like hydroxyl group, main chain's amide, and main chain's tryptophan, each of them hold only one hydrogen site to accommodate halogen ions, while the congeneric side chain's amide provides two hydrogen sites for halogen ions. In addition, three charged moieties, *i.e.* ammonium, imidazolium, and guanidinium, have a larger surface to contact with surrounding halogen ions. In this work, 10 kinds of low-lying energy structures of halogen ions (F^-, Cl^-, Br^-, and I^-) binding to the electrophilic hydrogen atoms of seven protein groups were calculated (three kinds of low-lying energy structures for guanidinium group of arginine and two for imidazolium group of histidine, shown in Table 2). As can be seen from Table 2, as might be anticipated, the mean intermolecular distances for halogen-ionic bonds in biological systems increase with the radius or polarizability of halide anions, namely $F^-\cdots H \ll Cl^-\cdots H < Br^-\cdots H < I^-\cdots H$. Note that the interaction strength for these studied systems exhibits an opposite tendency as that in intermolecular distances.

Interaction energy analysis

In this paper, we have assessed the ability of various DFT methods for an accurate description of halide-binding complexes. The interaction energies of these complexes (ΔE_{int}) were calculated by using the 31 appointed DFT methods in conjunction with 6-311++G(df,pd) and Lanl2DZ + (df) basis sets at respective optimized geometries. ΔE_{int} was defined as the minimum interaction energy between the halide ions and their interacting partners. All model systems were arranged in low-energy conformations (see Table 2). ΔE_{int} calculated using MP2/6-311++G(df,pd) theory as well as the other 34 QM methods and 1 MM method are tabulated in Tables 3, 4, 5 and 6. Subsequently, the MP2/6-311++G(df,pd) (or MP2/Lanl2DZ + (df) for iodine) energies were used as the "standard" values to evaluate the performance of all other methods on the basis of these halide adducts.

Table 2 Structure models of protein moieties in complex with different halide ions fully optimized at the MP2/aug-cc-PVDZ level

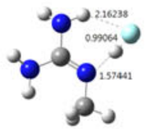
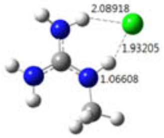
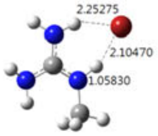
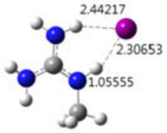
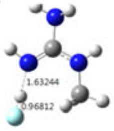
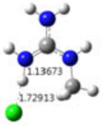
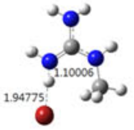
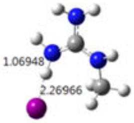
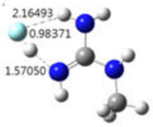
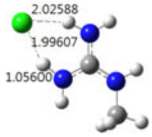
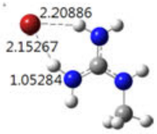
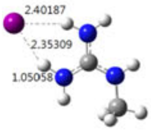
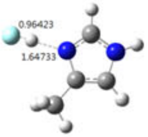
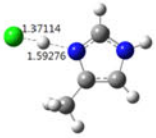
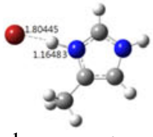
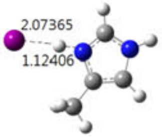
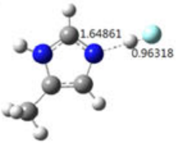
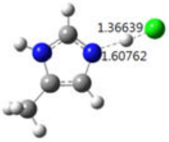
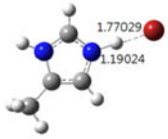
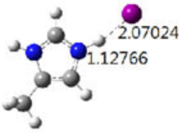
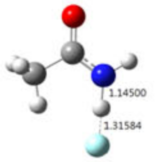
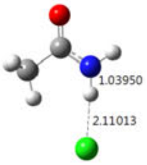
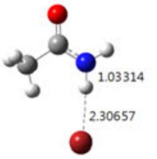
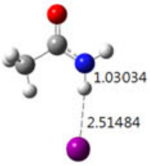
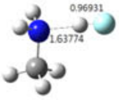
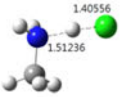
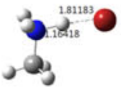
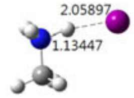
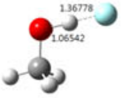
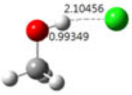
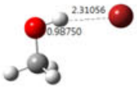
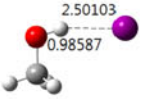
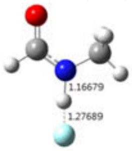
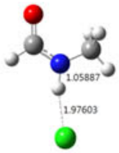
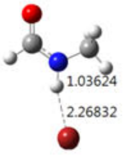
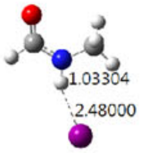
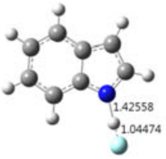
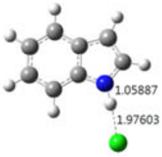
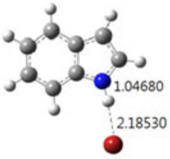
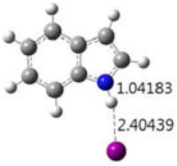
NO.	F^{-}	Cl^{-}	Br^{-}	I^{-}
1	 N^{1}-H ··· F^{-} (**1 F**$^{-}$)	 N^{13}-H ··· Cl^{-} (**1 Cl**$^{-}$)	 N^{13}-H ··· Br^{-} (**1 Br**$^{-}$)	 N^{13}-H ··· I^{-} (**1 I**$^{-}$)
2	 N^{2}-H ··· F^{-} (**2 F**$^{-}$)	 N^{2}-H ··· Cl^{-} (**2 Cl**$^{-}$)	 N^{2}-H ··· Br^{-} (**2 Br**$^{-}$)	 N^{2}-H ··· I^{-} (**2 I**$^{-}$)
3	 N^{23}-H ··· F^{-} (**3 F**$^{-}$)	 N^{23}-H ··· Cl^{-} (**3 Cl**$^{-}$)	 N^{23}-H ··· Br^{-} (**3 Br**$^{-}$)	 N^{23}-H ··· I^{-} (**3 I**$^{-}$)
4	 N^{1}-H ··· F^{-} (**4 F**$^{-}$)	 N^{1}-H ··· Cl^{-} (**4 Cl**$^{-}$)	 N^{1}-H ··· Br^{-} (**4 Br**$^{-}$)	 N^{1}-H ··· I^{-} (**4 I**$^{-}$)
5	 N^{2}-H ··· F^{-} (**5 F**$^{-}$)	 N^{2}-H ··· Cl^{-} (**5 Cl**$^{-}$)	 N^{2}-H ··· Br^{-} (**5 Br**$^{-}$)	 N^{2}-H ··· I^{-} (**5 I**$^{-}$)
6	 N-H··· F^{-} (**6 F**$^{-}$)	 N-H ··· Cl $^{-}$(**6 Cl**$^{-}$)	 N-H ··· Br^{-} (**6 Br**$^{-}$)	 N-H ··· I^{-} (**6 I**$^{-}$)
7	 N-H ··· F^{-} (**7 F**$^{-}$)	 N-H ··· Cl^{-} (**7 Cl**$^{-}$)	 N-H ··· Br^{-} (**7 Br**$^{-}$)	 N-H ··· I^{-} (**7 I**$^{-}$)
8	 O-H ··· F^{-} (**8 F**$^{-}$)	 O-H··· Cl $^{-}$(**8 Cl**$^{-}$)	 O-H ··· Br^{-} (**8 Br**$^{-}$)	 O-H ··· I^{-} (**8 I**$^{-}$)
9	 N-H ··· F^{-} (**9 F**$^{-}$)	 N-H ··· Cl $^{-}$(**9 Cl**$^{-}$)	 N-H ··· Br^{-} (**9 Br**$^{-}$)	 N-H ··· I^{-} (**9 I**$^{-}$)
10	 N-H ··· F^{-} (**10 F**$^{-}$)	 N-H ··· Cl $^{-}$(**10 Cl**$^{-}$)	 N-H ··· Br^{-} (**10 Br**$^{-}$)	 N-H ··· I^{-} (**10 I**$^{-}$)

Table 3 Comparison of the interaction energies (ΔE_{int}, in kcal mol^{-1}) of the 10 fluoride-binding types that observed in protein-ligand interactions calculated using different QM methods with the 6-311++G(df,pd) basis set

QM [a]	Fluoride-binding energies of model protein moieties										R^2 [b]	RMSE [b]	MSE [b]
	1 F$^-$	2 F$^-$	3 F$^-$	4 F$^-$	5 F$^-$	6 F$^-$	7 F$^-$	8 F$^-$	9 F$^-$	10 F$^-$			
MP2	−210.970	−224.215	−218.513	−226.100	−225.863	−47.230	−230.877	−33.440	−49.788	−93.916	——	——	——
B3LYP	−207.279	−219.607	−214.767	−222.523	−222.070	−48.319	−228.584	−34.788	−50.648	−91.172	0.999	3.043	2.116
B3P86	−212.054	−223.483	−218.832	−226.527	−226.234	−50.870	−232.249	−37.057	−53.508	−95.274	0.999	2.146	−1.518
B3PW91	−210.195	−221.788	−216.976	−225.027	−224.790	−49.122	−230.810	−35.525	−51.880	−93.597	1.000	1.536	0.120
B98	−210.101	−222.694	−217.641	−225.729	−225.350	−49.620	−231.243	−35.740	−52.022	−93.702	1.000	1.429	−0.293
B97-1	−210.008	−222.554	−217.546	−225.578	−225.043	−49.680	−231.072	−35.789	−52.016	−93.628	1.000	1.489	−0.200
BH&HLYP	−215.678	−229.614	−223.577	−231.738	−231.401	−50.372	−236.807	−36.461	−52.727	−96.801	0.997	4.588	−4.426
BLYP	−199.661	−210.744	−206.909	−213.957	−213.700	−45.771	−220.670	−33.190	−48.017	−85.312	0.987	9.607	8.298
BP86	−203.959	−214.468	−210.901	−217.914	−217.644	−48.307	−224.131	−34.998	−50.687	−89.513	0.994	6.372	4.839
BPW91	−203.427	−214.078	−210.282	−217.753	−217.556	−47.096	−223.948	−33.901	−49.607	−88.724	0.994	6.643	5.454
HCTH	−204.196	−215.929	−211.390	−219.922	−219.652	−46.269	−225.823	−33.512	−49.333	−89.620	0.996	5.359	4.527
M05	−210.922	−222.624	−217.519	−225.763	−225.445	−47.526	−231.511	−35.168	−50.052	−93.292	1.000	0.880	0.109
M05-2X	−216.171	−228.773	−223.520	−231.114	−230.692	−52.733	−235.778	−38.043	−55.094	−99.021	0.996	5.011	−5.003
M06	−211.868	−224.333	−219.363	−226.973	−226.428	−49.026	−232.917	−36.007	−51.097	−94.276	1.000	1.358	−1.138
M06-HF	−217.774	−230.318	−225.068	−232.255	−231.949	−55.689	−236.327	−39.305	−57.726	−100.879	0.993	6.698	−6.638
MPW1B95	−212.326	−224.984	−219.808	−227.739	−227.480	−49.953	−233.296	−35.648	−52.431	−95.089	0.999	1.895	−1.784
MPW1K	−217.328	−230.632	−224.934	−233.204	−232.927	−51.481	−238.140	−37.226	−54.147	−99.071	0.995	5.949	−5.818
MPW1KCIS	−207.697	−219.471	−214.885	−222.951	−222.670	−48.787	−228.784	−35.206	−51.268	−92.405	0.999	2.844	1.679
MPW1PW91	−212.582	−224.295	−219.355	−227.518	−227.241	−50.440	−233.040	−36.536	−53.013	−95.489	0.999	2.114	−1.860
MPW3LYP	−208.149	−220.388	−215.578	−223.195	−222.689	−49.296	−229.471	−35.723	−51.561	−92.020	0.999	2.606	1.284
MPWB1K	−216.757	−229.198	−223.582	−231.963	−231.705	−50.812	−237.070	−36.677	−53.238	−97.761	0.997	4.908	−4.785
MPWKCIS1k	−216.049	−228.243	−223.425	−231.636	−231.339	−50.624	−236.866	−36.620	−53.407	−97.942	0.997	4.621	−4.524
MPWLYP	−200.871	−211.764	−208.065	−214.899	−214.597	−47.130	−221.539	−34.501	−49.246	−86.485	0.989	8.732	7.182
O3LYP	−205.462	−217.852	−212.702	−221.415	−221.253	−45.503	−227.202	−32.281	−48.412	−89.838	0.997	4.291	3.899
PBE1KCIS	−209.584	−222.016	−217.012	−225.164	−224.869	−49.143	−230.650	−35.549	−51.603	−93.181	1.000	1.512	0.214
PBE1PBE	−213.065	−224.694	−219.814	−227.849	−227.530	−51.038	−233.333	−37.051	−53.369	−96.066	0.999	2.513	−2.290
PBEKCIS	−202.663	−214.083	−210.035	−217.235	−216.733	−47.457	−223.283	−34.691	−49.977	−88.233	0.993	7.051	5.652
PBEPBE	−205.204	−215.593	−211.962	−219.013	−218.722	−49.229	−225.147	−35.961	−51.586	−90.575	0.996	5.562	3.792
PW91PW91	−205.382	−216.165	−212.602	−219.551	−219.268	−49.789	−225.735	−36.612	−52.233	−91.175	0.996	5.234	3.240
SVWN5	−213.449	−222.315	−220.129	−225.248	−224.706	−58.038	−231.334	−43.814	−60.183	−99.162	0.995	6.121	−3.747
TPSS1KCIS	−208.620	−220.387	−215.705	−223.499	−223.136	−48.776	−229.214	−35.121	−51.401	−92.143	0.999	2.366	1.291
TPSS	−205.090	−216.667	−212.720	−219.991	−219.764	−48.182	−225.566	−35.027	−50.328	−89.873	0.996	4.979	3.770

Table 3 (continued)

QM [a]	Fluoride-binding energies of model protein moieties										R^2 [b]	RMSE [b]	MSE [b]
	1F⁻	2F⁻	3F⁻	4F⁻	5F⁻	6F⁻	7F⁻	8F⁻	9F⁻	10F⁻			
X3LYP	−208.255	−220.585	−215.695	−223.377	−222.908	−49.015	−229.352	−35.498	−51.316	−91.981	0.999	2.459	1.293
HF	−217.842	−234.063	−226.208	−237.434	−237.114	−47.053	−241.306	−33.044	−49.396	−97.309	0.992	7.610	−5.986
AM1	−258.354	−273.428	−262.436	−277.477	−278.653	−52.915	−280.338	−35.096	−58.285	−129.017	0.771	39.760	−34.509
PM3	−241.707	−255.663	−243.923	−272.732	−273.368	−48.955	−251.519	−27.671	−55.007	−126.056	0.876	29.212	−23.569
UFF	5.110	1.370	4.986	1.178	1.306	1444.413	1.273	884.314	884.314	2074.121	−113.432	888.401	598.313

[a] In conjugation with the 6-311++G(df,pd) basis set

[b] As the measures against the results derived from the MP2/6-311++G(df,pd) level of theory

To statistically evaluate the performance of the examined methods, correlation coefficient (R^2), root-mean-square error (RMSE) and mean signed errors (MSEs) were calculated for each of the method/basis set. Among them, the MSE is taken as the difference between the values calculated with the method tested and the correspondent "true" value. In this case, the "true" value was calculated based on the reference method (MP2). A negative MSE indicates that the application of given methodology to the type of halide-binding complexes considered overestimates the value of interaction energies ΔE_{int}, whereas a positive MSE indicates that the value is underestimated.

Fluoride-binding energies

As seen from Table 3, the ΔE_{int} based on MP2/6-311++G (df,pd) shows that the strength of fluoride interactions with polar moieties are quite modest, with their ΔE_{int} falling into the range of −33.440~−93.916 kcal mol^{-1}. In contrast, fluoride contacting with charged species, *i.e.* ammonium, imidazolium, and guanidinium cations, giving rise to a noticeably strong interaction energy (> −200 kcal mol^{-1}), which is much greater than the fluoride-binding energy between the fluoride and the polar moieties. These phenomena indicate that fluoride-binding in protein-ligand interactions are mainly derived from electrostatic force, which is more like the ionic bonding and much greater than the hydrogen bonding (hydrogen bonding is imparted more covalent and polar components) [92]. Fluoride interacting with polar hydrogen atoms in acetamide ($CH_3CONH_2 \cdots F^-$) has received a slighter attraction (−47.230 kcal mol^{-1} as by MP2) compared with the interacting with polar hydrogen atom in main chain's tryptophan ($C_8H_8NH \cdots F^-$), of which the attractive energy is more than −90 kcal mol^{-1} as determined by MP2. This difference could be owing to the presence of aromatic ring in tryptophan, suggesting that the fluorides in ligands are involved in nonbonding interactions with the π-cloud of aromatic residues in the protein matrix. Besides, by natural bond orbital (NBO) [93] analysis of charge transfers (CTs) between the fluoride and the polar hydrogen atom, we also found that a lot of electrons are transferred to hydrogen from fluoride upon the bonding.

The statistics of ΔE_{int} calculated using the 34 lower-level QM methods, 1 MM method and the corresponding R^2, RMSE and MSEs as the measures against that derived from the rigorous MP2/6-311++G(df,pd) are also listed in Table 3. As can be seen, most of the DFT methods perform very well in reproducing MP2-based ΔE_{int}; the greatest RMSE is only 0.880 kcal mol^{-1}, as obtained by the M05 functional. While other DFT methods like SVWN5, BP86, BPW91, M06, PBEKCIS, MPWLYP and BLYP are incapable of reproducing well the MP2-based ΔE_{int} with corresponding

Table 4 Comparison of the interaction energies (ΔE_{int}, in kcal mol^{-1}) of the 10 chloride-binding types that observed in protein-ligand interactions calculated using different QM methods with the 6-311++G(df,pd) basis set

QM [a]	Chloride-binding energies of model protein moieties										R^2 [b]	RMSE [b]	MSE [b]
	1 Cl^-	2 Cl^-	3 Cl^-	4 Cl^-	5 Cl^-	6 Cl^-	7 Cl^-	8 Cl^-	9 Cl^-	10 Cl^-			
MP2	−112.220	−107.073	−109.917	−177.369	−178.880	−21.904	−173.769	−15.168	−22.530	−24.896	——	——	——
B3LYP	−111.731	−105.772	−109.990	−171.951	−173.448	−22.202	−170.851	−15.501	−22.568	−23.404	0.998	2.679	1.631
B3P86	−114.493	−108.650	−112.715	−175.416	−176.841	−23.548	−173.910	−16.663	−23.760	−25.014	0.999	1.730	−0.728
B3PW91	−112.817	−107.037	−110.999	−174.248	−175.781	−22.334	−172.757	−15.595	−22.672	−23.893	0.999	1.526	0.559
B98	−112.824	−107.086	−111.036	−173.984	−175.373	−23.195	−172.189	−16.133	−23.480	−24.457	0.999	1.777	0.397
B97-1	−113.249	−107.545	−111.458	−174.402	−175.751	−23.634	−172.588	−16.447	−23.796	−24.826	0.999	1.729	0.003
BH&HLYP	−111.086	−105.674	−109.430	−176.151	−177.821	−22.748	−173.173	−15.857	−23.009	−23.461	1.000	0.996	0.532
BLYP	−110.822	−104.930	−109.603	−167.423	−168.780	−21.169	−167.489	−14.930	−21.861	−22.824	0.994	5.023	3.390
BP86	−113.630	−107.588	−112.265	−170.458	−171.732	−22.556	−170.610	−15.850	−23.383	−24.349	0.997	3.443	1.131
BPW91	−112.595	−106.675	−111.199	−170.757	−172.170	−21.598	−170.821	−14.972	−22.617	−23.536	0.998	3.184	1.679
HCTH	−110.806	−105.905	−109.177	−172.709	−174.058	−21.395	−171.989	−15.065	−22.292	−23.703	0.999	2.320	1.663
M05	−113.097	−107.453	−111.110	−176.010	−177.257	−23.636	−174.749	−17.015	−24.009	−24.707	1.000	1.281	−0.532
M05-2X	−113.070	−108.260	−111.277	−175.431	−176.688	−24.059	−172.914	−16.929	−24.219	−24.748	0.999	1.546	−0.387
M06	−114.127	−108.121	−112.499	−174.676	−175.551	−24.274	−173.338	−17.435	−24.205	−24.889	0.999	2.085	−0.539
M06-HF	−114.695	−111.125	−112.924	−178.091	−179.477	−25.274	−174.755	−17.835	−25.251	−25.921	0.999	2.457	−2.162
MPW1B95	−113.281	−108.450	−111.506	−177.250	−177.047	−23.237	−172.318	−16.134	−23.317	−24.357	1.000	1.276	−0.317
MPW1K	−113.471	−108.138	−111.694	−178.685	−180.333	−23.458	−175.791	−15.359	−23.722	−24.746	1.000	1.330	−1.167
MPW1KCIS	−113.316	−107.489	−111.455	−173.897	−175.276	−22.646	−173.071	−14.660	−23.130	−24.446	0.999	1.750	0.434
MPW1PW91	−113.808	−108.132	−111.981	−176.050	−177.571	−23.216	−174.226	−16.367	−23.489	−24.733	1.000	1.252	−0.585
MPW3LYP	−112.886	−106.889	−111.125	−172.818	−174.209	−23.143	−171.747	−16.411	−23.506	−24.367	0.999	2.299	0.663
MPWB1K	−113.184	−107.948	−111.324	−177.639	−179.188	−23.410	−171.627	−16.360	−23.494	−24.381	1.000	1.154	−0.483
MPWKCIS1k	−112.768	−107.495	−110.949	−177.943	−179.544	−23.003	−175.463	−16.105	−23.388	−24.440	1.000	0.907	−0.737
MPWLYP	−112.553	−106.524	−111.385	−168.658	−169.879	−22.640	−168.709	−16.175	−23.158	−24.007	0.996	4.334	2.004
O3LYP	−109.551	−104.562	−107.779	−172.219	−173.879	−19.968	−170.878	−13.608	−20.785	−21.958	0.998	3.094	2.854
PBE1KCIS	−112.618	−106.396	−110.339	−174.234	−175.616	−22.988	−172.271	−16.230	−23.335	−24.477	0.999	1.632	0.522
PBE1PBE	−114.576	−108.888	−112.737	−176.401	−177.831	−23.862	−174.567	−16.974	−24.099	−25.432	0.999	1.712	−1.164
PBEKCIS	−113.370	−107.173	−111.577	−171.070	−172.303	−23.139	−170.409	−16.515	−23.548	−24.788	0.998	3.205	0.983
PBEPBE	−115.181	−109.345	−113.744	−172.465	−173.662	−23.808	−172.565	−17.082	−24.640	−25.796	0.998	3.064	−0.456
PW91PW91	−115.886	−110.077	−114.483	−173.205	−173.915	−24.333	−173.351	−17.586	−24.697	−26.273	0.998	3.222	−1.008
SVWN5	−123.844	−117.316	−122.658	−177.716	−178.476	−29.445	−177.643	−22.098	−29.838	−31.091	0.986	7.838	−6.640
TPSS1KCIS	−113.340	−107.376	−112.464	−174.026	−175.512	−23.042	−172.421	−16.202	−23.431	−24.444	0.999	1.886	0.147
TPSS	−113.656	−107.717	−112.325	−172.236	−173.690	−22.883	−171.281	−15.417	−23.517	−24.584	0.998	2.643	0.642

Table 4 (continued)

QM [a]	Chloride-binding energies of model protein moieties										R^2 [b]	RMSE [b]	MSE [b]
	1 Cl^-	2 Cl^-	3 Cl^-	4 Cl^-	5 Cl^-	6 Cl^-	7 Cl^-	8 Cl^-	9 Cl^-	10 Cl^-			
X3LYP	−112.346	−106.382	−110.599	−172.581	−174.033	−22.889	−171.390	−16.017	−23.085	−23.954	0.999	2.365	1.045
HF	−103.244	−98.498	−101.785	−174.514	−176.518	−19.551	−168.834	−12.514	−19.850	−19.532	0.993	5.539	4.889
AM1	−119.167	−118.003	−116.811	−193.956	−196.898	−22.652	−188.034	−13.152	−23.387	−27.943	0.976	10.164	−7.628
PM3	−114.362	−116.941	−111.810	−180.948	−184.006	−21.972	−180.347	−14.005	−22.301	−25.585	0.996	4.356	−2.855
UFF	105.012	296.965	93.452	0.573	0.771	24.930	1.515	25.640	31.851	56.783	−7.456	189.722	158.122

[a] In conjugation with the 6-311++G(df,pd) basis set

[b] As the measures against the results derived from the MP2/6-311++G(df,pd) level of theory

RMSE of 6.121, 6.372, 6.443, 6.698, 7.051, 8.732 and 9.607 kcal mol^{-1}, respectively.

Chloride-binding energies

As expected, the ΔE_{int} obtained from MP2/6-311++G(df, pd) shows that three charged moieties possess a much stronger chlophilicity than the polar counterparts (Table 4). The energies of chloride interactions with the three charged moieties are in the range from −107.073 to −173.769 kcal mol^{-1}, which is about 60~100 kcal mol^{-1} higher than the corresponding fluoride-binding energies. The differences in the ΔE_{int} between the chloride interacting with hydroxyl group ($CH_3OH \cdots Cl^-$) and its interactions with polar hydrogen in main chain's tryptophan ($C_8H_8NH \cdots Cl^-$) is about 10 kcal mol^{-1} (by MP2). Note that this value is significantly smaller than the interaction energy of the corresponding fluoride-involved bonding, which is about 60 kcal mol^{-1} (by MP2), indicating that the changes in complex affinity are not only contributed by the form of halide bonds but also result from the indirect effects of fluoride altering the electron distribution of protein moieties.

It is evident from Table 4 that most of the DFT methods are much capable of reproducing chloride-ionic bridging energies obtained at the MP2/6-311++G(df,pd) level, such as MPWKCIS1K, BH&HLYP, MPW1PW91, and M05 functionals, with the corresponding RMSE are as follows: 0.907, 0.996, 1.252, and 1.281 kcal mol^{-1}, respectively. Both the SVWN5 and BLYP methods perform not very well in reproducing MP2-based ΔE_{int}.

Bromide-binding energies

The interaction energy ΔE_{int} relative to the formation of the bromide-ionic complexes are summarized in Table 5. As can be seen, the ΔE_{int} value from MP2/6-311++G(df,pd) shows that the binding strengths of the ammonium $\cdots Br^-$ and imidazolium $\cdots Br^-$ complexes are increased by about 50~70 kcal mol^{-1} when compared with the corresponding chloride-binding counterparts as given in Table 4, while increased by only 2~8 kcal mol^{-1} for the remaining species. In addition, when reproducing the bromide-binding energies obtained at the MP2/6-311++G(df,pd) level, most of the DFT methods perform as well. 27 out of the 31 density functionals gave the RMSE<1.5 kcal mol^{-1}. The worst performance functional is SVWN5 with a particularly significant RMSE value of 8.315 kcal mol^{-1}.

Iodide-binding energies

The ΔE_{int} values of the iodide-ionic adducts are tabulated in Table 6, from which the three charged complexes stand more stably (ranging from −92.375 to −111.300 kcal mol^{-1})

Table 5 Comparison of the interaction energies (ΔE_{int}, in kcal mol^{-1}) of the 10 bromide-binding types that observed in protein-ligand interactions calculated using different QM methods with the 6-311++G(df,pd) basis set

QM [a]	Bromide-binding energies of model protein moieties										R^2 [b]	RMSE [b]	MSE [b]
	1 Br$^-$	2 Br$^-$	3 Br$^-$	4 Br$^-$	5 Br$^-$	6 Br$^-$	7 Br$^-$	8 Br$^-$	9 Br$^-$	10 Br$^-$			
MP2	−106.328	−99.494	−104.490	−109.870	−111.727	−19.430	−121.715	−13.263	−20.061	−21.621	——	——	——
B3LYP	−105.315	−98.155	−104.404	−107.374	−109.062	−19.319	−121.369	−13.455	−19.767	−19.618	0.999	1.429	1.016
B3P86	−108.005	−100.755	−107.028	−110.240	−111.934	−20.610	−123.874	−14.305	−21.053	−21.246	0.999	1.389	−1.105
B3PW91	−106.389	−99.173	−105.398	−108.870	−110.715	−19.581	−122.559	−13.304	−20.021	−20.157	1.000	0.764	0.183
B98	−106.439	−99.670	−105.526	−109.103	−110.773	−20.310	−122.611	−13.921	−20.653	−20.786	1.000	0.753	−0.179
B97-1	−107.219	−100.068	−105.914	−109.904	−111.149	−20.658	−123.006	−14.229	−20.941	−21.157	1.000	0.926	−0.625
BH&HLYP	−104.483	−97.668	−103.591	−106.851	−108.717	−19.648	−120.396	−13.466	−19.921	−19.723	0.998	1.766	1.354
BLYP	−104.563	−97.089	−103.558	−106.407	−108.171	−18.371	−120.947	−12.549	−19.491	−19.057	0.998	2.086	1.780
BP86	−107.361	−99.795	−106.683	−109.163	−110.873	−19.560	−123.258	−13.489	−20.392	−20.491	0.999	1.050	−0.307
BPW91	−106.245	−98.779	−105.644	−108.603	−110.530	−18.598	−122.682	−12.623	−19.598	−19.682	0.999	1.044	0.502
HCTH	−104.632	−97.493	−103.377	−108.304	−110.317	−18.312	−121.949	−12.925	−19.635	−19.665	0.999	1.341	1.139
M05	−106.978	−99.904	−105.526	−109.686	−111.623	−20.766	−123.055	−15.210	−21.198	−21.194	0.999	1.027	−0.714
M05-2X	−106.694	−100.156	−105.682	−108.681	−110.072	−21.304	−122.233	−14.759	−21.548	−21.320	0.999	1.203	−0.445
M06	−108.313	−101.233	−107.411	−110.229	−111.253	−21.539	−123.994	−15.368	−21.609	−21.777	0.998	1.798	−1.473
M06-HF	−107.535	−101.843	−106.482	−110.167	−111.505	−22.035	−122.705	−15.340	−22.087	−21.899	0.999	1.654	−1.360
MPW1B95	−106.890	−100.097	−105.957	−108.950	−110.644	−20.366	−122.559	−14.150	−20.680	−20.571	1.000	0.933	−0.286
MPW1K	−106.840	−99.895	−105.854	−109.518	−111.485	−20.385	−122.856	−13.974	−20.635	−20.822	1.000	0.784	−0.426
MPW1KCIS	−107.037	−99.874	−106.017	−109.532	−111.470	−19.589	−123.234	−13.443	−20.079	−20.879	1.000	0.779	−0.315
MPW1PW91	−107.292	−100.142	−106.283	−109.636	−111.570	−20.305	−123.334	−14.034	−20.746	−20.949	1.000	0.977	−0.629
MPW3LYP	−106.444	−99.063	−105.266	−108.366	−109.943	−20.253	−122.358	−14.380	−20.645	−20.737	1.000	0.985	0.054
MPWB1K	−106.565	−99.896	−105.624	−108.892	−110.609	−20.400	−122.223	−14.075	−20.525	−20.598	1.000	0.827	−0.141
MPWKCIS1k	−106.191	−99.322	−105.191	−109.214	−111.224	−20.064	−122.747	−13.760	−20.510	−20.501	1.000	0.663	−0.073
MPWLYP	−106.029	−98.719	−105.033	−107.813	−109.364	−19.613	−122.296	−13.734	−20.179	−20.236	0.999	1.154	0.498
O3LYP	−102.551	−95.414	−101.897	−106.632	−108.985	−16.828	−120.473	−11.122	−17.630	−17.992	0.996	2.960	2.848
PBE1KCIS	−106.129	−99.114	−105.068	−108.899	−110.387	−20.043	−122.341	−13.846	−20.438	−20.669	1.000	0.736	0.107
PBE1PBE	−108.030	−100.864	−106.984	−110.308	−112.129	−20.892	−123.913	−14.603	−21.307	−21.521	0.999	1.472	−1.255
PBEKCIS	−106.869	−99.774	−105.878	−109.367	−110.922	−20.168	−123.179	−14.202	−20.578	−20.989	1.000	0.862	−0.393
PBEPBE	−108.881	−101.365	−108.103	−110.811	−112.426	−20.810	−124.860	−14.740	−21.656	−21.937	0.998	2.027	−1.759
PW91PW91	−109.512	−101.898	−108.288	−111.344	−112.951	−21.473	−125.471	−15.213	−22.086	−22.355	0.997	2.464	−2.259
SVWN5	−117.076	−109.009	−115.949	−117.709	−118.759	−26.023	−131.664	−19.377	−25.937	−26.993	0.965	8.315	−8.050
TPSS1KCIS	−106.447	−99.474	−105.388	−108.640	−110.500	−19.781	−122.150	−13.546	−20.143	−20.513	1.000	0.739	0.142
TPSS	−107.198	−99.756	−106.620	−108.787	−110.572	−19.849	−122.483	−13.707	−20.195	−20.696	1.000	0.985	−0.186

Table 5 (continued)

QM [a]	Bromide-binding energies of model protein moieties										R^2 [b]	RMSE [b]	MSE [b]
	1 Br⁻	2 Br⁻	3 Br⁻	4 Br⁻	5 Br⁻	6 Br⁻	7 Br⁻	8 Br⁻	9 Br⁻	10 Br⁻			
X3LYP	−105.884	−98.730	−104.967	−107.687	−109.492	−19.840	−121.814	−13.944	−20.408	−20.151	0.999	1.170	0.508
HF	-96.625	-90.436	-95.927	-99.755	−101.651	−16.334	−112.490	−10.281	−16.708	−15.946	0.970	7.748	7.185
AM1	−114.369	−111.663	−112.840	−126.658	−131.207	−20.269	−135.905	−11.303	−20.874	−24.625	0.942	10.767	−8.171
PM3	−100.824	−101.546	-99.520	−111.223	−114.631	−17.538	−122.788	-9.619	−17.071	−18.261	0.995	3.282	1.498
UFF	56.609	104.728	51.751	265.962	336.002	11.970	253.523	11.609	14.756	23.957	−28.125	240.847	185.887

[a] In conjugation with the 6-311++G(df,pd) basis set

[b] As the measures against the results derived from the MP2/6-311++G(df,pd) level of theory

as compared to those of remaining polar complexes (ranging from −10.712 to −17.954 kcal mol^{-1}). This tendency is clearly consistent with the ΔE_{int} of the other halide-binding moieties in Tables 3, 4 and 5, indicating that, compared to polar halide-bindings, charged are more long-range and hold a considerable strength. In addition, the fundamental difference between polar and charged halide-bindings in long-range interaction behavior renders their natures of ionic hydrogen bonding and ionic bonding, respectively.

From the values of RMSE in Table 6, it is found that the lowest RMSE with a value of 0.646 kcal mol^{-1} is obtained from the PBE1KCIS functional, while the largest RMSE of 8.036 kcal mol^{-1} is associated with the SVWN5 method. Other DFT methods like B98, B97-1, M05, MPW1K, MPW1B95, MPW1KCIS, MPWB1K, MPW3LYP, MPW1PW91, MPWKCIS1K, MPWLYP, and TPSS seem to be capable of effectively reproducing the ΔE_{int} calculated at the MP2/6-311++G(df,pd) level.

From the data in Tables 3, 4, 5 and 6, it is evident that the binding energies are basically consistent with the complex geometries (Table 2), from I⁻ to F⁻ ion, the complexes with shorter N-H ··· X⁻ bonds have larger binding energies, which reflects the decreasing tendency of the size of halide ions. It is worth noting that the fluoride-binding energies are much greater than that of other halide adducts, probably because of the electro-negativity of fluorine element is much stronger than other halogens. As can be seen from Tables 3, 4, 5 and 6, the capability of 31 density functionals in determination of halide-bridging energies was evaluated on a representative database of seven protein moieties. In this study, most of these functionals performed with the 6-311++G(df,pd) basis set are thought to be good candidates of the stringent MP2/6-311++G(df,pd) methods, such as PBE1KCIS, B98, B97-1, B3PW91, M05, MPW1B95, MPW1K, MPW1PW91, MPWB1K, and MPWKCIS1K. However, the SVWN5 is a rather poor method with its RMSE>6.0 kcal mol^{-1}, which may be due to the fact that the SVWN5 functional fail to account for important dispersion components. Overall, the hybrid functionals generally yield deviations that are smaller than the corresponding pure ones. Notably, good performance of 6-311++G(df,pd), a relatively small basis set considered here, is inspiring, specifically with regard to bromide and iodide adducts, since one of the attractive features of DFT is its application to large systems for which larger basis sets can be very demanding in routine calculations. The other QM methods employed here, such as the *ab initio* HF theory and the semi-empirical AM1 and PM3, all have a noticeable feature to largely underestimate binding energies of all halide complexes, especially for fluoride complexes, given by the great RMSE of 7.610, 39.760,

Table 6 Comparison of the interaction energies (ΔE_{int}, in kcal mol^{-1}) of the 10 iodide-binding types that observed in protein-ligand interactions calculated using different QM methods with the 6-311++G(df,pd) basis set

QM [a]	Iodide-binding energies of model protein moieties										R^2 [b]	RMSE [b]	MSE [b]
	1 I^-	2 I^-	3 I^-	4 I^-	5 I^-	6 I^-	7 I^-	8 I^-	9 I^-	10 I^-			
MP2	−98.746	−92.375	−97.054	−98.208	−98.064	−16.206	−111.300	−10.712	−16.987	−17.954	——	——	——
B3LYP	−97.681	−91.064	−96.560	−95.935	−95.979	−15.717	−111.232	−10.464	−16.333	−15.667	0.999	1.363	1.097
B3P86	−100.188	−93.385	−98.978	−98.337	−98.255	−16.987	−113.323	−11.595	−17.617	−17.069	0.999	1.162	−0.813
B3PW91	−98.571	−97.239	−97.350	−96.919	−96.981	−15.984	−111.948	−10.585	−16.586	−15.981	0.998	1.765	−0.054
B98	−98.191	−92.169	−97.382	−97.238	−97.095	−16.751	−112.094	−11.220	−17.197	−16.728	1.000	0.712	0.154
B97-1	−98.727	−92.786	−97.953	−97.816	−97.605	−17.157	−112.638	−11.592	−17.563	−17.185	1.000	0.757	−0.342
BH&HLYP	−96.251	−90.220	−95.331	−94.856	−95.171	−16.047	−109.796	−10.640	−16.455	−15.547	0.998	2.045	1.729
BLYP	−97.462	−90.546	−96.167	−95.695	−95.746	−15.075	−111.466	−10.019	−15.760	−15.185	0.998	1.686	1.449
BP86	−99.555	−92.721	−98.465	−97.859	−97.767	−16.103	−113.195	−10.904	−17.130	−16.538	0.999	0.929	−0.263
BPW91	−98.659	−91.662	−97.520	−97.200	−97.262	−15.447	−112.589	−10.118	−16.447	−15.763	0.999	1.004	0.494
HCTH	−96.912	−90.037	−95.671	−96.975	−96.942	−15.253	−112.051	−10.149	−16.143	−15.934	0.999	1.418	1.154
M05	−98.360	−92.278	−96.966	−97.215	−97.000	−16.879	−111.858	−11.590	−17.507	−16.781	1.000	0.739	0.117
M05-2X	−99.278	−93.897	−98.022	−97.234	−97.240	−17.775	−112.248	−12.185	−18.248	−17.352	0.999	1.125	−0.587
M06	−100.350	−94.918	−99.251	−98.387	−97.709	−17.963	−113.430	−12.620	−18.290	−17.730	0.998	1.733	−1.304
M06-HF	−99.400	−94.640	−98.128	−98.038	−97.840	−18.599	−112.003	−12.815	−18.837	−17.818	0.999	1.523	−1.051
MPW1B95	−98.814	−92.910	−97.755	−97.140	−96.895	−16.563	−111.984	−11.124	−17.252	−16.629	1.000	0.767	0.054
MPW1K	−98.597	−92.194	−97.530	−96.943	−97.295	−16.759	−111.851	−11.223	−17.174	−16.666	1.000	0.710	0.137
MPW1KCIS	−99.398	−92.539	−98.163	−97.888	−98.167	−16.346	−113.165	−10.913	−16.921	−16.678	1.000	0.835	−0.257
MPW1PW91	−99.384	−92.660	−98.165	−97.701	−97.714	−16.731	−112.627	−11.314	−17.312	−16.774	1.000	0.775	−0.278
MPW3LYP	−98.815	−92.245	−97.652	−96.930	−96.869	−16.645	−112.247	−11.352	−17.200	−16.640	1.000	0.819	0.101
MPWB1K	−98.336	−92.566	−97.304	−96.756	−96.636	−16.779	−111.382	−11.319	−17.105	−16.945	1.000	0.784	0.248
MPWKCIS1k	−98.179	−91.730	−96.994	−96.885	−97.238	−16.461	−111.950	−11.038	−16.911	−16.470	1.000	0.773	0.375
MPWLYP	−99.009	−92.197	−97.685	−97.089	−96.903	−16.327	−112.863	−11.254	−16.990	−16.493	1.000	0.894	0.080
O3LYP	−95.161	−87.931	−93.879	−94.792	−95.151	−13.504	−110.050	-8.462	−14.330	−13.971	0.994	3.157	3.038
PBE1KCIS	−98.565	−92.015	−97.334	−97.467	−97.320	−16.688	−112.212	−11.263	−17.261	−16.779	1.000	0.646	0.070
PBE1PBE	−100.189	−93.503	−98.940	−98.446	−98.355	−17.366	−113.301	−11.930	−17.924	−17.483	0.999	1.224	−0.983
PBEKCIS	−99.723	−93.052	−98.533	−98.378	−98.031	−16.955	−113.589	−11.700	−17.454	−17.275	0.999	1.053	−0.708
PBEPBE	−101.208	−94.429	−100.037	−99.555	−99.307	−17.556	−114.827	−12.219	−18.475	−18.071	0.998	2.032	−1.808
PW91PW91	−102.033	−95.036	−100.615	−100.033	−99.697	−17.928	−115.372	−12.670	−18.510	−18.473	0.996	2.501	−2.276
SVWN5	−109.126	−102.116	−107.574	−105.794	−104.990	−22.603	−121.068	−16.715	−22.614	−22.704	0.962	8.034	−7.770
TPSS1KCIS	−98.975	−99.723	−97.759	−97.147	−97.079	−16.322	−111.989	−10.954	−17.053	−16.545	0.997	2.433	−0.594
TPSS	−99.283	−92.635	−98.116	−97.354	−97.229	−16.470	−112.231	−11.056	−16.919	−16.654	1.000	0.752	−0.034

Table 6 (continued)

QM [a]	Iodide-binding energies of model protein moieties										R^2 [b]	RMSE [b]	MSE [b]
	1 I^-	2 I^-	3 I^-	4 I^-	5 I^-	6 I^-	7 I^-	8 I^-	9 I^-	10 I^-			
X3LYP	−98.206	−91.645	−97.071	−96.384	−96.390	−16.232	−111.668	−10.938	−16.959	−16.182	0.999	1.014	0.593
HF	−88.173	−83.035	−87.240	−87.788	−88.296	−12.778	−101.643	−7.261	−13.296	−11.708	0.961	8.179	7.639
AM1	−107.686	−102.734	−106.105	−114.932	−116.357	−17.727	−125.973	−9.550	−18.088	−21.020	0.935	10.550	−8.257
PM3	−82.811	−82.643	−82.019	−86.481	−88.220	−9.954	−99.054	−2.336	−9.491	−9.595	0.922	10.929	10.500
UFF	33.250	28.682	30.905	86.361	88.680	6.746	94.930	7.314	8.112	12.292	−8.494	127.404	105.488

[a] In conjugation with the 6-311++G(df,pd) basis set

[b] As the measures against the results derived from the MP2/6-311++G(df,pd) level of theory

and 29.212 kcal mol^{-1}, respectively. This is because, first, the electron correlation is completely neglected by *ab initio* HF theory, and second, the core electrons are not included in the calculation using semi-empirical method and only a minimal basis set was used. In addition, the results obtained by UFF method are unsatisfactory.

To evaluate the dependence on basis set effect, we performed additional calculations by using the 10 best functionals in conjunction with the aug-cc-pVTZ basis set for each halide-binding complex to reproduce the ΔE_{int} of halide complexes obtained at the more rigorous MP2/aug-cc-pVTZ level of theory, the results are shown in Tables 7, 8, 9 and 10. These 10 DFT methods were selected from Tables 3, 4, 5 and 6 with the lowest root-mean-square error (RMSE) and highest correlation coefficient (R^2). Attention should be paid that the more widely used three-parameter function, B3LYP, is not among the 10 best functionals, since its average absolute deviations amount to 2.129 kcal mol^{-1} for halide-binding energy. The poor performance of B3LYP is not only undergone with the halide-bound complexes, but also encountered in hydrogen-bonded systems [27]. As can be seen, when the basis set size increases from 6-311++G (df,pd) to aug-cc-pVTZ, the MP2-based ΔE_{int} values decrease a litter (by 2.0~4.0 kcal mol^{-1} or less) in cases of bromide and iodide adducts (Tables 5, 6 and Tables 9, 10), while for the other two halide-bound systems in Tables 7, 8, the ΔE_{int} at the MP2/aug-cc-pVTZ level are very close to those of MP2/6-311++G(df,pd). This could be attributed to that the size of electron clouds increases from F^- to I^-, as for Br^- and I^-, the electron diffusion are relatively larger than that of F^- and Cl^-, the aug-cc-pVTZ basis set could describe the large electron diffusion system more precisely and therefore could obtain more stable energies. While for the F^- and Cl^- cases, 6-311++G(df,pd) is a good compromise between the accuracy and efficiency of computation. In these tested DFT functionals, the B98, B97-1, and M05 gave the lowest RMSE for fluoride-binding energies, the optimal performances of chloride-binding energies were obtained with M05-2X, MPW1B95, and MPW1PW91, the best results of bromide-binding energies were obtained with the B97-1, PBEKCIS, and PBE1KCIS, meanwhile, B97-1, MPW1PW91, and TPSS gave the lowest RMSE for iodide interaction energies. Overall, the statistics listed in Tables 7, 8, 9 and 10 showed that the PBE1KCIS functional is a good candidate for the general purpose of analyzing interaction behavior of biological adducts involving halide ions.

QM/MM analysis of halide-bindings in protein-ligand complexes

To better understand the effects of halide-binding and further prove the feasibility of DFT functionals obtained by above analysis based on small model systems, the

Table 7 Comparison of the interaction energies (ΔE_{int}, in kcal mol^{-1}) of the 10 fluoride-binding types that observed in protein-ligand interactions calculated using 10 best performance DFT methods in Table 3 with the aug-cc-PVTZ basis set

QM [a]	Fluoride-binding energies of model protein moieties										R^2 [b]	RMSE [b]	MSE [b]
	1 F$^-$	2 F$^-$	3 F$^-$	4 F$^-$	5 F$^-$	6 F$^-$	7 F$^-$	8 F$^-$	9 F$^-$	10 F$^-$			
MP2	−210.825	−223.795	−218.288	−225.370	−224.922	−48.980	−229.879	−34.648	−51.154	−94.235	——	——	——
B3P86	−211.121	−223.522	−218.534	−226.458	−225.086	−50.678	−232.082	−36.505	−53.219	−95.931	1.000	1.406	−1.104
B3PW91	−209.274	−221.825	−216.670	−224.969	−223.646	−48.699	−230.629	−35.294	−51.532	−94.207	1.000	1.091	0.535
B98	−210.088	−222.815	−217.572	−225.509	−225.273	−49.630	−231.101	−35.635	−51.988	−94.766	1.000	0.776	−0.228
B97-1	−209.958	−222.641	−217.441	−225.324	−225.065	−49.671	−230.878	−35.668	−51.929	−94.907	1.000	0.799	−0.139
M05	−211.977	−225.245	−219.554	−226.635	−226.128	−47.608	−232.772	−35.550	−50.113	−94.218	1.000	0.799	−0.770
M06	−211.406	−224.111	−218.943	−226.571	−225.980	−48.784	−232.696	−35.662	−50.756	−93.629	1.000	1.138	−0.644
MPW1B95	−211.773	−224.895	−219.380	−229.015	−226.991	−49.623	−233.061	−35.954	−52.087	−95.515	0.999	1.887	−1.620
MPW1PW91	−211.516	−224.261	−218.986	−226.904	−227.092	−50.082	−232.900	−36.200	−52.715	−95.955	1.000	1.624	−1.452
PBE1KCIS	−209.687	−222.012	−216.887	−225.236	−223.415	−50.106	−230.965	−36.050	−52.093	−93.978	1.000	1.187	0.167
TPSS1KCIS	−207.199	−219.857	−214.866	−223.311	−222.801	−48.403	−228.505	−35.250	−50.963	−92.587	0.999	2.333	1.835

[a] In conjugation with the Dunning's augmented basis set, aug-cc-pVTZ

[b] As the measures against the results derived from the MP2/ aug-cc-PVTZ level of theory

Table 8 Comparison of the interaction energies (ΔE_{int}, in kcal mol^{-1}) of the 10 chloride-binding types that observed in protein-ligand interactions calculated using 10 best performance DFT methods in Table 4 with the aug-cc-PVTZ basis set

QM [a]	Chloride-binding energies of model protein moieties										R^2 [b]	RMSE [b]	MSE [b]
	1 Cl$^-$	2 Cl$^-$	3 Cl$^-$	4 Cl$^-$	5 Cl$^-$	6 Cl$^-$	7 Cl$^-$	8 Cl$^-$	9 Cl$^-$	10 Cl$^-$			
MP2	−115.932	−110.114	−114.110	−176.899	−178.067	−24.326	−173.385	−16.935	−24.608	−26.951	——	——	——
B3PW91	−113.851	−107.274	−112.233	−175.730	−177.200	−22.680	−173.304	−15.365	−23.083	−24.244	0.999	1.813	1.636
BH&HLYP	−112.349	−106.381	−110.957	−176.325	−178.172	−22.033	−173.762	−15.651	−22.231	−24.023	0.999	2.415	1.944
M05	−113.394	−106.530	−111.625	−175.912	−177.214	−23.294	−175.039	−16.753	−23.876	−24.729	0.999	1.910	1.296
M05-2X	−114.275	−108.521	−112.811	−176.705	−178.013	−24.579	−173.862	−17.838	−24.602	−25.315	1.000	1.039	0.481
MPW1B95	−114.527	−108.469	−112.728	−177.671	−178.878	−23.703	−173.948	−15.840	−23.657	−24.738	1.000	1.247	0.717
MPW1K	−114.754	−108.681	−113.230	−178.697	−180.513	−22.769	−176.279	−16.135	−24.087	−23.511	0.999	1.923	0.267
MPW1PW91	−114.777	−108.213	−113.157	−177.572	−178.982	−23.524	−174.588	−16.012	−23.831	−25.135	1.000	1.183	0.554
MPWKCIS1K	−113.903	−107.897	−112.333	−179.547	−179.617	−23.342	−175.862	−15.747	−23.733	−24.867	0.999	1.877	0.448
MPWB1K	−114.309	−108.481	−112.881	−177.490	−179.279	−23.875	−175.255	−16.004	−23.868	−24.803	1.000	1.355	0.508
PBE1KCIS	−114.645	−107.970	−112.953	−177.225	−177.506	−23.850	−174.334	−16.275	−23.902	−25.492	1.000	1.105	0.718

[a] In conjugation with the Dunning's augmented basis set, aug-cc-pVTZ

[b] As the measures against the results derived from the MP2/ aug-cc-PVTZ level of theory

Table 9 Comparison of the interaction energies (ΔE_{int}, in kcal mol^{-1}) of the 10 bromide-binding types that observed in protein-ligand interactions calculated using 10 best performance DFT methods in Table 5 with the aug-cc-PVTZ basis set

QM [a]	Bromide-binding energies of model protein moieties										R^2 [b]	RMSE [b]	MSE [b]
	1 Br$^-$	2 Br$^-$	3 Br$^-$	4 Br$^-$	5 Br$^-$	6 Br$^-$	7 Br$^-$	8 Br$^-$	9 Br$^-$	10 Br$^-$			
MP2	−109.002	−101.935	−107.768	−111.468	−112.899	−21.317	−123.207	−14.565	−21.699	−23.268			
B3PW91	−106.689	−98.928	−105.713	−109.244	−109.848	−19.456	−123.016	−12.549	−19.906	−20.112	0.998	2.315	2.167
B98	−107.597	−100.377	−106.648	−109.599	−111.303	−20.736	−122.430	−13.711	−21.004	−21.094	0.999	1.363	1.263
B97-1	−108.032	−100.826	−107.069	−110.035	−111.686	−21.219	−122.860	−14.034	−21.197	−21.483	0.999	1.002	0.869
MPW1K	−107.457	−100.147	−106.586	−109.545	−111.811	−19.219	−123.414	−13.563	−20.824	−19.445	0.998	1.808	1.512
MPWKCIS1K	−106.650	−99.408	−105.735	−109.069	−111.422	−20.061	−123.220	−13.177	−20.487	−20.681	0.998	1.888	1.722
MPW1KCIS	−106.715	−98.848	−105.722	−109.803	−111.688	−19.702	−123.775	−13.751	−20.312	−20.649	0.998	1.886	1.616
MPWB1K	−107.160	−100.066	−106.352	−108.779	−110.044	−20.542	−122.713	−12.772	−20.704	−20.790	0.998	1.883	1.721
PBEKCIS	−107.526	−101.546	−106.602	−110.597	−112.379	−20.429	−124.996	−14.751	−21.532	−21.575	0.999	1.079	0.520
PBE1KCIS	−107.527	−99.854	−106.510	−111.488	−111.610	−20.654	−124.179	−13.566	−20.953	−21.404	0.999	1.271	0.938
TPSS1KCIS	−107.362	−100.032	−107.355	−109.088	−110.987	−19.933	−122.680	−13.677	−20.511	−20.605	0.998	1.652	1.490

[a] In conjugation with the Dunning's augmented basis set, aug-cc-pVTZ

[b] As the measures against the results derived from the MP2/ aug-cc-PVTZ level of theory

Table 10 Comparison of the interaction energies (ΔE_{int}, in kcal mol^{-1}) of the 10 iodide-binding types that observed in protein-ligand interactions calculated using 10 best performance DFT methods in Table 6 with the aug-cc-PVTZ basis set

QM [a]	Iodide-binding energies of model protein moieties										R^2 [b]	RMSE [b]	MSE [b]
	1 I$^-$	2 I$^-$	3 I$^-$	4 I$^-$	5 I$^-$	6 I$^-$	7 I$^-$	8 I$^-$	9 I$^-$	10 I$^-$			
MP2	−102.171	−95.386	−100.747	−100.800	−100.499	−18.173	−113.992	−12.176	−18.758	−19.720	——	——	——
B98	−98.744	−92.358	−97.761	−97.729	−97.580	−16.875	−112.678	−11.201	−16.975	−16.937	0.996	2.511	2.358
B97-1	−99.282	−93.022	−98.361	−98.330	−98.109	−17.284	−113.242	−11.586	−17.681	−17.410	0.998	1.991	1.812
M05	−98.684	−92.205	−97.155	−97.152	−97.130	−16.582	−112.166	−11.592	−17.279	−16.672	0.996	2.784	2.581
MPW1B95	−99.482	−93.188	−98.548	−98.326	−98.602	−16.650	−112.535	−11.384	−16.886	−16.800	0.998	2.091	2.002
MPW1K	−99.246	−91.838	−97.865	−97.849	−98.220	−16.905	−112.442	−11.184	−17.305	−16.684	0.997	2.443	2.288
MPWB1K	−98.961	−92.408	−97.753	−97.175	−97.760	−16.933	−111.969	−11.265	−17.229	−16.518	0.996	2.604	2.445
MPWKCIS1K	−99.048	−91.670	−97.502	−97.956	−97.578	−16.613	−112.564	−10.915	−17.038	−16.457	0.996	2.653	2.508
MPW1PW91	−99.759	−93.092	−99.120	−98.607	−99.062	−16.319	−113.202	−11.270	−16.828	−17.013	0.998	1.911	1.815
PBE1KCIS	−98.657	−92.505	−97.877	−97.898	−97.852	−16.066	−112.775	−11.299	−17.462	−17.099	0.997	2.440	2.293
TPSS	−100.498	−93.012	−99.364	−97.872	−97.824	−17.041	−112.798	−11.291	−17.375	−16.749	0.998	2.008	1.860

[a] In conjugation with the Dunning's augmented basis set, aug-cc-pVTZ

[b] As the measures against the results derived from the MP2/ aug-cc-PVTZ level of theory

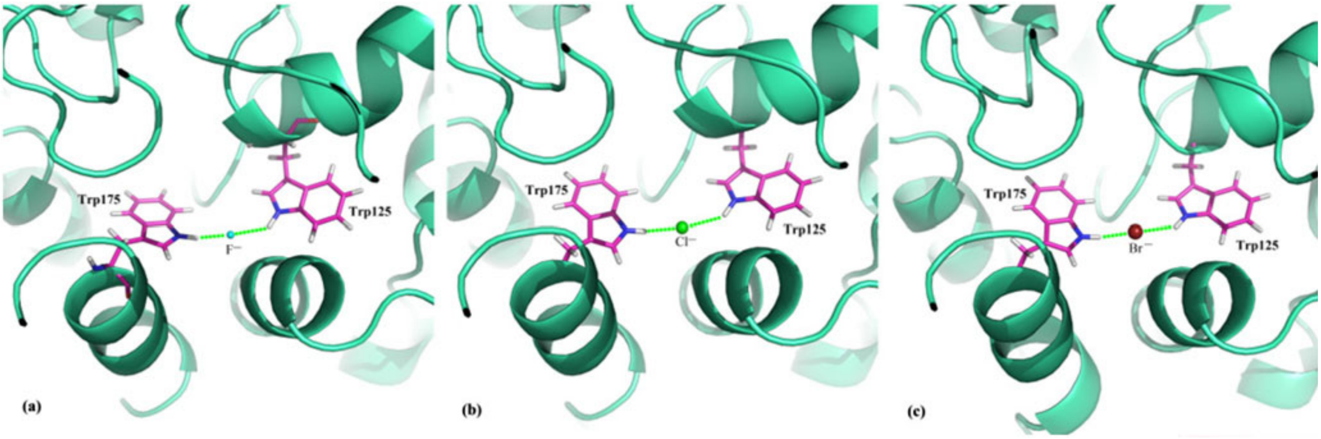


Fig. 2 Superposition of model QM layers, which were optimized with the presence of halide ions (PDB: 2eda). Binding by and fluoride (**a**), chloride (**b**), bromide (**c**). It is worth noting that the QM/MM optimization procedure for the iodide ion model structure did not reach the convergence

crystal structures of haloalkane dehalogenase (DhlA) in complex with F^-, Cl^-, Br^- and I^- were selected to perform ONIOM-based QM/MM calculations, in which the halide ions and the vicinal residues Trp125 and Trp175 that directly bound to the halide ions were included in the QM layer.

The complex structures of these studied systems were fully optimized using the two-layered QM/MM scheme, as described in the section of ONIOM-based QM/MM calculations on real systems. The electrostatic interactions between QM and MM layers were treated in terms of the mechanical embedding strategy to save computational cost. The optimized structures of QM layer with the presence of halide ions are depicted in Fig. 2, and corresponding geometrical and energetic parameters are assembled in

Table 11 Geometrical and energetic parameters for halide motifs depicted in Fig. 2

PDB Ion		1CIJ F^-		1CIJ Cl^-		1CIJ Br^-	
$d(N\cdots H)_{Trp125}$ (Å)	CS[a]	1.000		1.000		1.000	
	OS[b]	1.053		1.019		1.021	
$d(H_{Trp125}\cdots X^-)$ (Å)	CS	2.616		2.616		2.616	
	OS	1.695		2.360		2.524	
$d(N\cdots H)_{Trp175}$ (Å)	CS	1.000		1.000		1.000	
	OS	1.101		1.029		1.034	
$d(H_{Trp175}\cdots X^-)$ (Å)	CS	2.300		2.300		2.300	
	OS	1.437		2.135		2.271	
$\angle N_{Trp125}H_{Trp125}X^-$ (°)	CS	172.465		172.465		172.465	
	OS	159.640		145.920		137.113	
$\angle H_{Trp125}X^-H_{Trp175}$ (°)	CS	164.941		164.941		164.941	
	OS	162.916		166.603		174.556	
$\angle N_{Trp175}H_{Trp175}X^-$ (°)	CS	145.584		145.584		145.584	
	OS	163.320		169.408		169.821	
Energy (kcal/mol)		$Trp125\cdots F^-$	$Trp175\cdots F^-$	$Trp125\cdots Cl^-$	$Trp175\cdots Cl^-$	$Trp125\cdots Br^-$	$Trp175\cdots Br^-$
ΔE_{int} [MP2/6-311++G(df,pd)]		−36.016	−45.147	−20.148	−22.611	−17.806	−20.659
ΔE_{int} [MP2/aug-cc-pVTZ]		−37.577	−47.050	−22.068	−24.556	−19.583	−22.198
ΔE_{int} [DFT[d]/6-311++G(df,pd)]		−37.130	−46.086	−20.910	−22.701	−17.854	−20.346
ΔE_{int} [DFT[d]/aug-cc-pVTZ]		−37.062	−46.180	−21.410	−23.236	−17.987	−20.469

[a] Crystal structure

[b] Optimized structure

[c] Root mean-square derivation of the QM layer of the models relative to X-ray crystal structures

[d] From F^- to Br^-, the DFT functional is B98, M05-2X, and B97-1, respectively

Table 11. As seen, the calculated interatomic $X^{-} \cdots H{-}N$ distances are in the range 1.437~2.524 Å. These values are smaller than the sums of the van der Waals radii of the atoms involved. In fact, the QM/MM optimizations fail to converge for the model system with the presence of iodide ion, implying a morbid feature associated with the potential energy surface of this artifact, since the iodine ion is too large to be inserted in the halophilic site of DhlA. Furthermore, the halide-binding energies are calculated to be in the range −19.583 to −47.050 kcal mol^{-1} at the MP2/aug-cc-pVTZ level via single-point calculations. Note that these values are significantly smaller than the interaction energy of isolated fluoride-binding system and very close to those of isolated chloride- and bromide-binding counterparts, indicating that in this protein structure, the distances to the two tryptophans (residues Trp125 and Trp175) are optimized for chloride and bromide (around 2 Å). Unfortunately, fluoride binding requires a much smaller binding distance of about 1 Å, which cannot be satisfied at both sides in such short QM/MM simulation time. From Table 11 it is seen that, in comparison with the MP2 method, the DFT method gives binding energies very close to those obtained at the MP2 level with the same basis set. These results suggest that the B98, M05-2X, and B97-1 are good choices for accurately determining the fluoride-, chloride-, and bromide-binding energy of moderate systems, respectively.

Conclusions

In this article, we report a systematical comparison of 31 DFT methods in reproducing the energetic behaviors of halide series (F^{-}, Cl^{-}, Br^{-}, and I^{-}) binding to the polar and charged moieties of proteins. All model complex structures are fully optimized by using the MP2 method in conjunction with the aug-cc-pVDZ basis set. Two basis sets, 6-311++G(df,pd) and aug-cc-pVTZ, are used to calculate the interaction energy ΔE_{int} involved in halide complexes so as to check the stability of the results obtained using various DFT theories. Meanwhile, the performance of 31 DFT methods and other methods, including one noncorrelation *ab initio* theory (HF), two semi-empirical methods (AM1 and PM3) and one mechanical force field (UFF), has been assessed on the basis of the database obtained with MP2 calculations. The results are imparted with the following remarks: (1) most DFT methods perform well in determining ΔE_{int} of halide-binding complexes, among the tested DFT functionals, besides, the hybrid functionals generally yield deviations generally smaller than the corresponding pure ones; (2) the performance of the relatively small basis set, 6-311++G(df,pd), is an appropriate choice that could precisely describe the ΔE_{int} of fluoride and chloride interacting with model protein moieties; (3) the HF, AM1, and PM3 methods tested in this work have a strong tendency to underestimate binding energies of all halide adducts, especially for fluoride-binding complexes, while the UFF method can't be used to describe the interaction energy of halide-binding complexes; (4) the widely used function, B3LYP, seems not to be the best functional for describing the ΔE_{int} of halide-moiety interactions; (5) the B98, B97-1, and M05 give the lowest RMSE for fluoride-binding energies, the best performances of chloride-binding energies are obtained with M05-2X, MPW1B95, and MPW1PW91, the best results of bromide-binding energies are determined by B97-1, PBEKCIS, and PBE1KCIS, meanwhile, B97-1, MPW1PW91, and TPSS give rise to the lowest RMSE for iodide-binding energies. In addition, the PBE1KCIS functional provides accuracies close to the computationally expensive MP2 method for the calculation of the ΔE_{int} of halide adducts.

References

1. Takashima K, Riveros JM (1998) Mass Spectrom Rev 17:409, and reference cited therein
2. Zavitsas AA (2001) J Phys Chem B 105:7805–7817
3. Hosoda H, Mori H, Sogoshi N, Nagasawa A, Nakabayashi SJ (2004) Phys Chem A 108:1461–1464
4. Schleich T, von Hippel PH (1969) Biopolymers 7:861–877
5. Zhang Y, Cremer PS (2006) Curr Opin Chem Biol 10:658–663
6. Kunz W, Lo Nostro P, Ninham BW (2004) Curr Opin Colloid Int Sci 9:1–18
7. Ninham BW (1999) Adv Colloid Int Sci 83:1–17
8. Aroti A, Leontidis E, Dubois M, Zemb T, Brezesinski G (2007) Colloids Surf A Physicochem Eng Aspects 303:144–158
9. Curtis RA, Lue L (2006) Chem Eng Sci 61:907–923
10. Prausnitz J, Foose L (2007) Pure Appl Chem 79:1435–1444
11. Kunz W (2009) Specific ion effects in nature and technology, 1st edn. World Scientific Publishing
12. Boström M, Williams D, Ninham BW (2001) Langmuir 17:4475–4478
13. Boström M, Williams D, Ninham BW (2001) Phys Rev Lett 87:168103/1–168103/4
14. Böstrom M, Tavares FW, Bratko D, Ninham BW (2005) J Phys Chem B 109:24489–24494
15. Boström M, Kunz W, Ninham BW (2005) Langmuir 21:2619–2623
16. Jungwirth P, Tobias DJ (2001) J Phys Chem B 105:10468–10472
17. Lund M, Vácha R, Jungwirth P (2008) Langmuir 24:3387–3391
18. Hofmeister F (1888) Arch Exp Pathol Pharmakol 24:247–260
19. Ninham BW, Yaminsky V (1997) Langmuir 13:2097–2108
20. Boström M, Lonetti B, Fratini E, Baglioni P, Ninham BW (2006) J Phys Chem B 110:7563–7566
21. Laage D, Hynes JT (2006) Science 311:832–835
22. Zhou P, Ren Y, Tian F, Zou J, Shang Z (2010) J Chem Theor Comput 6:2225–2241
23. Zhou P, Tian F, Liu X, Ren Y, Shang Z (2010) J Phys Chem B 114:15673–15686
24. Lu Y, Shi T, Wang Y, Yang H, Yan X, Luo X, Jiang H, Zhu WJ (2009) J Med Chem 52:2854–2862
25. Lu Y, Wang Y, Xu Z, Yan X, Luo X, Jiang H, Zhu WJ (2009) J Phys Chem B 113:12615–12621

26. Alzate-Morales JH, Caballero J, Jague AV, Nilo FDG (2009) J Chem Inf Model 49:886–899
27. Zhao Y, Truhlar DG (2005) J Chem Theor Comput 1:415–432
28. Wesolowski TA, Parisel O, Ellinger Y, Weber J (1997) J Phys Chem A 101:7818–7825
29. Zhao Y, Schultz NE, Truhlar DG (2006) J Chem Theor Comput 2:364–382
30. Riley KE, Opt Holt BT, Merz KM Jr (2007) J Chem Theor Comput 3:407–433
31. Cybulski SM, Seversen CE (2005) J Chem Phys 122:14117
32. Dahlke EE, Truhlar DG (2005) J Phys Chem B 109:15677–15683
33. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian 03, Revision C.01. Gaussian Inc, Wallingford, CT
34. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA Jr, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam NJ, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas O, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ (2009) Gaussian 09. Gaussian Inc, Wallingford, CT
35. Perdew JP (1986) Phys Rev B 33:8822–8824
36. Becke AD (1988) Phys Rev A 38:3098–3100
37. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785–789
38. Perdew JP (1991) In: Ziesche P, Eschig H (eds) Electronic structure of solids. Akademie, Berlin, pp 11–20
39. Hamprecht FA, Cohen AJ, Tozer DJ, Handy NC (1998) J Chem Phys 109:6264–6271
40. Adamo C, Barone V (1998) J Chem Phys 108:664–675
41. Perdew JP, Burke K, Ernzerhof M (1996) Phys Rev Lett 77:3865–3868
42. Rey J, Savin A (1998) Int J Quantum Chem 69:581–590
43. Krieger JB, Chen J, Iafrate GJ, Savin A (1999) Electron Correl Mater Prop p463
44. Toulouse J, Savin A, Adamo C (2002) J Chem Phys 117:10465–10473
45. Staroverov VN, Scuseria GE, Tao J, Perdew JP (2003) J Chem Phys 119:12129–12137
46. Tao J, Perdew JP, Staroverov VN, Scuseria GE (2003) Phys Rev Lett 91:146401/1–146401/4
47. Becke AD (1993) J Chem Phys 98:5648–5652
48. Stephens PJ, Devlin FJ, Chabalowski CF, Frisch MJ (1994) J Phys Chem 98:11623–11627
49. Frisch MJ, Trucks GW, Schlegel HB, Gill PMW, Johnson BG, Wong MW, Foresman JB, Robb MA, Head-Gordon M, Replogle ES, Gomperts R, Andres JL, Raghavachari K, Binkley JS, Gonzalez C, Martin RL, Fox DJ, Defrees DJ, Baker J, Stewart JJP, Pople JA (1993) Gaussian 92/DFT Revision F.2; Gaussian Inc Pittsburgh PA
50. Perdew JP, Kurth S, Zupan A, Blaha P (1999) Phys Rev Lett 82:2544–2547
51. Lynch BJ, Fast PL, Harris M, Truhlar DG (2000) J Phys Chem A 104:4811–4815
52. Zhao Y, Truhlar DG (2004) J Phys Chem A 108:6908–6918
53. Handy NC, Cohen A (2001) J Mol Phys 99:403–412
54. Hoe WM, Cohen AJ, Handy NC (2001) Chem Phys Lett 341:319–328
55. Xu X, Goddard WA (2004) PNAS 101:2673–2677
56. Zhao Y, González-García N, Truhlar DG (2005) J Phys Chem A 109:2012–2018
57. Zhao Y, Lynch BJ, Truhlar DG (2005) Phys Chem Chem Phys 7:43–52
58. Zhao Y, Schultz NE, Truhlar DG (2005) J Chem Phys 123:161103/1–161103/4
59. Zhao Y, Truhlar DG (2008) Theor Chem Acc 120:215–241
60. Zhao Y, Truhlar DG (2006) J Phys Chem A 110:13126–13130
61. Vosko SH, Wilk L, Nusair M (1980) Can J Phys 58:1200–1211
62. Slater JC (1974) The self-consistent field for molecular and solids, vol 4. McGraw-Hill, New York
63. Lu Y, Wang Y, Zhu W (2010) Phys Chem Chem Phys 12:4543–4551
64. Zhao Y, Truhlar DG (2007) J Chem Theory Comput 3:289–300
65. Gapeev A, Dunbar RC (2002) J Am Soc Mass Spectrom 13:477–484
66. Zhao YX, Wang SG (2005) Chin Chem Lett 16:1555–1558
67. Goossen LJ, Koley D, Hermann HL (2005) Organometallics 24:2398–2410
68. Lu Y, Zou J, Wang H, Yu Q, Zhang H, Jiang Y (2005) J Phys Chem A 109:11956–11961
69. Lu Y, Zou J, Fan J, Zhao W, Jiang Y, Yu Q (2009) J Comput Chem 30:725–732
70. Lu SY, Jiang YJ, Zhou P, Zou JW, Wu TX (2010) Chem Phys Lett 485:348–353
71. Zhou P, Zou JW, Tian FF, Shang ZC (2009) J Chem Inf Model 49:2344–2355
72. Chao SD, Li AH (2007) J Phys Chem A 111:9586–9590
73. Benedek NA, Latham K, Snook IK, Yarovsky I (2006) J Phys Chem B 110:19605–19610
74. Park H, Yoon J, Seok C (2008) J Phys Chem B 112:1041–1048
75. Rubicelia V, Jorge G, David AD, Benjamin PH (2000) J Am Chem Soc 122:4750–4755
76. Tsuzuki S, Houjou H, Nagawa Y, Goto M, Hiratani K (2001) J Am Chem Soc 123:4255–4258
77. Dkhissi A, Blossey R (2007) Chem Phys Lett 439:35–39
78. Erin RJ, Robert AW, Gino AD (2004) Chem Phys Lett 394:334–338
79. Glukhovtsev MN, Pross A, Radom L (1995) J Am Chem Soc 117:2024–2032
80. Sanov A, Faeder J, Parson R, Lineberger WC (1999) Chem Phys Lett 313:812–819
81. Rappe AK, Bernstein ER (2000) J Phys Chem A 104:6117–6128
82. Chesnut DB, Moseley RW (1969) Theor Chim Acta 13:230–248
83. Boys SF, Bernardi F (1970) Mol Phys 19:553–566
84. Krivov GG, Shapovalov MV, Dunbrack RL Jr (2009) Proteins 77:778–795
85. Kabsch W, Sander C (1983) Biopolymers 22:2577–2637
86. Word JM, Lovell SC, Richardson JS, Richardson DC (1999) J Mol Biol 285:1735–1747
87. Zhou P, Tian F, Lv F, Shang Z (2009) Proteins 76:151–163
88. Bas DC, Rogers DM, Jensen JH (2008) Proteins 73:765–783

89. Zhao Y, Cheng T, Wang R (2007) J Chem Inf Model 47:1379–1385
90. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) J Am Chem Soc 117:5179–5197
91. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) J Comput Chem 25:1157–1174
92. Yerushalmi R, Brandis A, Rosenbach-Belkin V, Baldridge KK, Scherz A (2006) J Phys Chem A 110:412–421
93. Foster JP, Weinhold F (1980) J Am Chem Soc 102:7211–7218
94. Arshadi M, Yamdagni R, Kebarle P (1970) J Phys Chem 74:1475–1482
95. Hiraoka K, Mizuse S, Yamabe S (1988) J Phys Chem 92:3943–3957

J Mol Model (2012) 18:2099–2104
DOI 10.1007/s00894-011-1230-1

ORIGINAL PAPER

Molecular dynamics simulations on the aggregation behavior of indole type organic dye molecules in dye-sensitized solar cells

Ananda Rama Krishnan Selvaraj · Shuji Hayase

Received: 11 December 2010 / Accepted: 22 August 2011 / Published online: 9 September 2011

Abstract In $Ti0_2$ nanostructured dye-sensitized solar cells indole based organic dyes D149, D205 exhibits greater power conversion efficiency. Such organic dye molecules are easily undergone for aggregation. Aggregation in dye molecules leads to reduce electron transfer process in dye-sensitized solar cells. Therefore, anti-aggregating agents such as chenodeoxycholic acid are commonly added to organic dye solution in DSSCs. Studying aggregation of such dye molecules in the absence of semiconductors gives a detailed influence of anti-aggregating agents on dye molecules. Atomistic level of molecular dynamics (MD) simulations were performed on aggregation of indole type dye molecules D149, D205 and D205-F with anti-aggregating agent chenodeoxy cholic acid using AMBER program. The trajectories of the MD simulations were analyzed with order parameters such as radial atom pair distribution functions g(r), diffusion coefficients and root mean square deviations values. MD results suggest that addition of chenodeoxy cholic acid to dyes significantly reduces structural arrangement and increases conformational flexibility and mobility of dye molecules. The influence of semi-perfluorinated alkyl chains in indole dye molecules was analyzed. The parameters such as open-circuit voltage (V_{oc}) and power conversion efficiency (η) of dye-sensitized solar cells are corroborated with flexibility and diffusion values of dye molecules.

Keywords MD simulation · Diffusion coefficients (D) · Dye-sensitized solar cells · Flexibility · Indole dyes · Radial atom pair distribution function g(r) · Root mean square deviations (RMSD)

A. R. K. Selvaraj (✉) · S. Hayase
Graduate School of Life Science and Systems Engineering,
Kyushu Institute of Technology,
2-4 Hibikino, Wakamatsu, Kitakyushu,
Fukuoka, Japan 808-0196
e-mail: ananda.selvaraj@chemie.uni-halle.de

Introduction

Dye-sensitized solar cells (DSSCs) are current topic of research in the field of green chemistry and renewable energy resources. Regan et al. [1] have shown the significance of dye-sensitized solar cells. Greater performance of DSSCs above 11% power conversion efficiency under standard AM 1.5 solar illumination was obtained with electrolytes consisting of organic solvents [2, 3]. Organic dye molecules have considerable importance in comparison with inorganic dye molecules by their applicability to green chemistry in DSSCs. Several types of organic dyes have been reported including near IR dyes for DSSCs applications [4–6]. However, greater power conversion efficiency of DSSCs was observed in indole based organic dye molecules D149 and D205 [7–9]. The limitations of organic dyes are reduced in molar absorption coefficient and increased dye aggregation. Especially, dye aggregation suppresses electron transport from the excited dye molecule to TiO_2 semiconductor surface resulting in lower DSSCs performance. Therefore, an anti-

Table 1 Performance of DSSCs with D149 and D205 dyes in the absence and presence of CDCA [Ref.8]

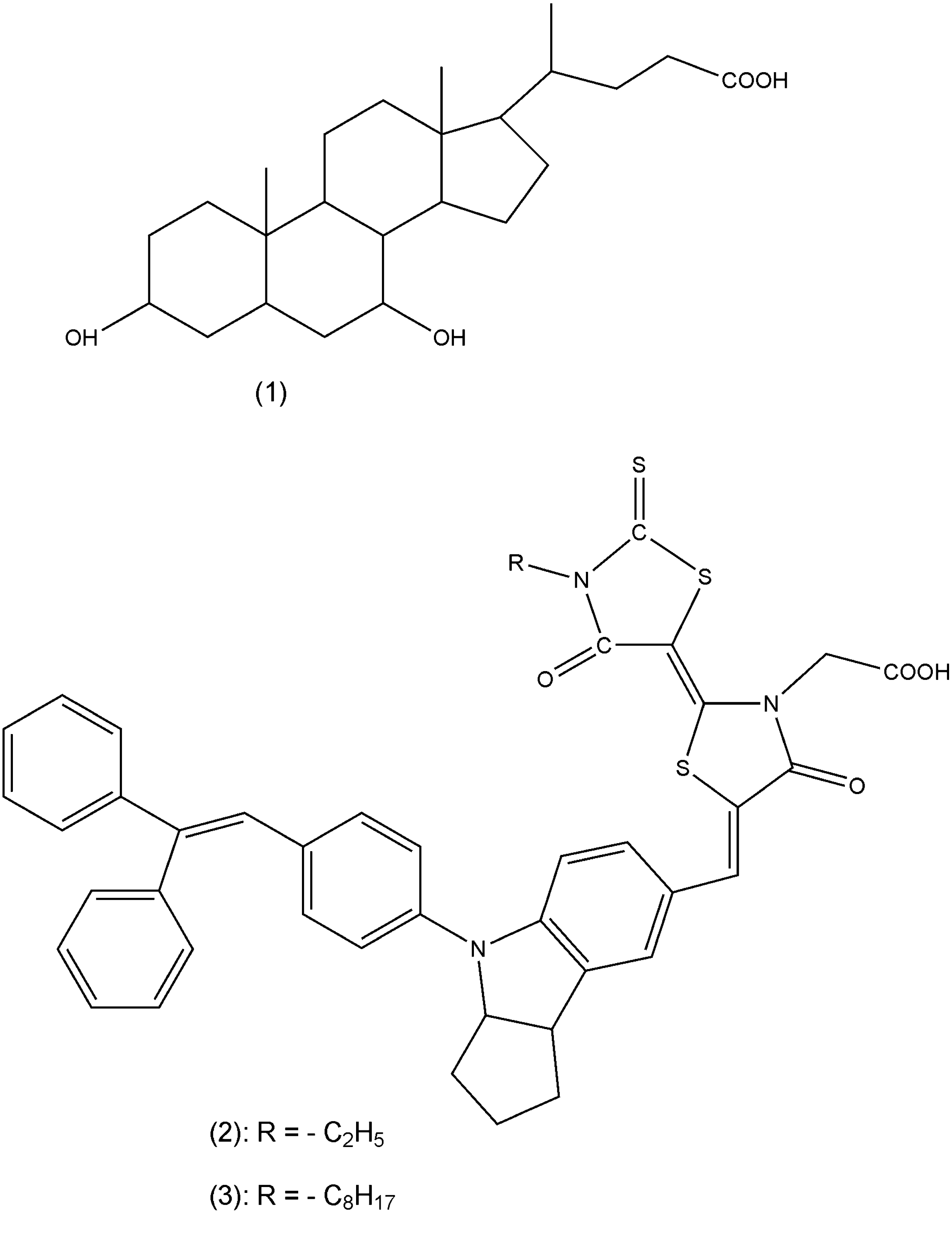
Photovoltaic characteristics	In the absence of CDCA		In the presence of CDCA	
	D149	D205	D149	D205
Short-circuit current (J_{sc}) (mA/cm^2)	19.08	18.99	19.86	18.68
Open-circuit voltage (V_{oc}) (V)	0.638	0.656	0.644	0.710
Fill factor (FF)	0.682	0.678	0.694	0.707
Efficiency (%)	8.26	8.43	8.85	9.40

aggregating agent such as chenodeoxy cholic acid (CDCA) is used with organic dyes to enhance the electron transport for larger power conversion efficiency of DSSCs [10–12]. Ito et al. show the effect of adding CDCA to indole dye molecules D149 (ethyl), D205 (octyl alkyl chain) in DSSCs performance [8]. The performance of DSSCs with D149 and D205 dyes and CDCA is presented in Table 1. D205 dye with octyl alkyl chain shows larger open circuit voltage V_{oc} and power conversion efficiency values than D149 dye with an ethyl alkyl chain in DSSCs. Moreover, addition of CDCA to D205 and D149 dye molecules results in increased DSSCs performance in both dyes. However, exact contribution of CDCA with organic dye molecules in DSSCs performance is not clearly reported. There are limited theoretical studies performed on the role of CDCA in aggregation of organic dyes with TiO_2 semiconductor surface. Recently, time dependent density functional theory studies on isolated D149, D102 and D131 dye molecules were carried out with respect to their conformational and absorption spectra [13]. Also, ab initio and second order Möller-Plesset perturbation level of theories were adopted to study the aggregation of D102 and D149 dyes in the presence of TiO_2 surface in which two dye molecules are attached on the $(TiO_2)_{82}$ surface to consider the interaction between dye molecules and TiO_2 surface [14].

Shlyk-Kerner et al. and Nojiri et al. showed the relation between conformational flexibility of the molecule and the electron transfer process. Mainly, in the photosynthesis process protein flexibility and their conformational changes are very important for conversion of solar energy into electrochemical potential as well as in biological systems where the electron transfer process completely depends on the functional protein's flexibility [15, 16]. Therefore in this work, we performed atomistic level of molecular dynamics (MD) simulations on the clusters (128 molecules) of D149, D205 and D205-F dye molecules with CDCA in order to get some qualitative information about aggregation, conformational flexibility and diffusion values of dye molecules in the presence of CDCA. In this simulation study, we have not considered TiO_2 semiconductor surface because we are interested in the effect of adding CDCA to dye molecules on flexibility and aggregation of dyes.

Results and discussion

The structures of the considered dye molecules and CDCA are presented in Fig. 1. D205-F model dye is different from D205 dye by the semi-perfluorinated alkyl chain (see Fig. 1). MD simulations were carried out on clusters of dyes and CDCA with implementing AMBER10 version using “GAFF” force field [17, 18]. The applicability of this force field on the organic molecules was shown with in the AMBER program [19–22]. Atomic net charges for the molecules were adapted from the fit to reproduce electrostatic potential within the HF/6-31G(d) level. The procedure is consistent with the defined atomic charges for amino-acid fragments in the AMBER program [23]. The starting structures of clusters were arranged by 128 monomers. MD simulations were performed in five different clusters. The clusters are **(1)**. CDCA/128cluster, **(2)**. D149/128cluster, **(3)**. D205/128cluster, **(4)**. D205-F/128cluster, **(5)**. mixture: (64CDCA+64D149)/128cluster. Clusters were arranged by (4x4x4x2) layer arrangement of monomers. In cluster **(5)**, CDCA and D149 molecules were arranged alternatively in layer arrangement. The MD simulations on clusters were performed with total simulation time of 6 ns with time step of 2 fs at temperature 300 K including heating phase. The systems were heated from 0 K to 300 K in the first 500 ps time period and equilibrated with NVT condition for 2.5 ns at 300 K. Finally systems were equilibrated for 3 ns with NPT condition at 300 K. The final 3 ns MD run have been considered for analysis of trajectories. The periodic boundary conditions were adopted during the MD run. The shake algorithm used for the hydrogen atoms and heavy atoms bonds during MD run [24]. The non-bonded interactions were calculated with a cut-off radius of 800 pm. The MD results were analyzed with AMBER10 standard analysis tool PTRAJ. The aggregation and structure formation in the clusters can be



(1)

(2): R = - C_2H_5

(3): R = - C_8H_{17}

(4): R = - C_2H_4 - C_6F_{13}

Fig. 1 Structures of dye molecules with anti-aggregating agent: (**1**). Chenodeoxycholic acid (CDCA); (**2**). Dye: D149; (**3**). Dye: D205; (**4**). Model dye: D205-F; The reference carbon atom "C^*" is in the carboxylic group (−C^*OOH) of CDCA and dye molecules (which is considered in g(r) calc.)

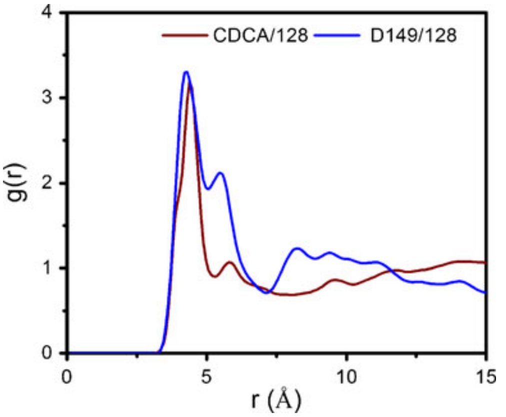


Fig. 2 Radial atom pair distribution function g(r) for CDCA in cluster (**1**) CDCA/128 and D149 in cluster (**2**) D149/128; [T=300 K, t=3 ns]

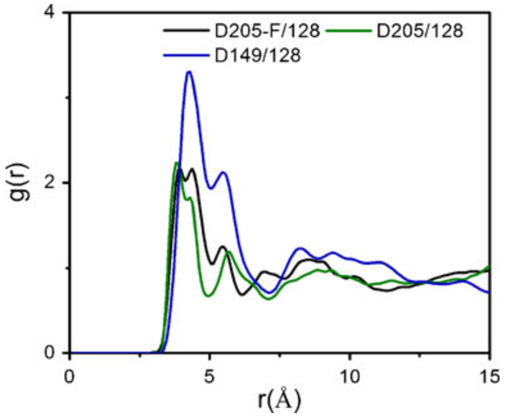


Fig. 4 Radial atom pair distribution function g(r) for the cluster D149 in (**2**): D149/128; D205 in cluster (**3**). D205/128 and D205-F in cluster (**4**). D205-F; [T=300 K, t=3 ns]

analyzed by the calculation of radial atom pair distribution function g(r) [25]. The g(r) values were calculated related to the reference carbon atom "C" in the carboxylic group of the dye and CDCA molecules (see Fig. 1). The conformational flexibility of dye molecules core part and terminal (alkyl /semi-perfluorinated alkyl) chain part estimated by root mean square deviation (RMSD) values of the corresponding atoms. The mobility of dye molecules in the clusters was investigated by diffusion coefficients which is calculated by Einstein model $<\Delta r^2> = 6Dt$ within the MD trajectories where $<\Delta r^2>$ is the mass weighted mean square displacements and D can be obtained by a linear fit [25].

The calculated radial atom pair distribution function g(r) results for the clusters are given in Figs. 2 and 3. In Fig. 2, g(r) curves of cluster of CDCA **(1)** and cluster of D149 **(2)** molecules are shown. In results both molecules show maxima of g(r) slightly different values but the pattern of both curves almost same. In Fig. 3, g(r) values of cluster **(5)** **(**mixture: D149/64+CDCA/64) is presented. In a mixture of CDCA and D149 molecules, lower g(r) value and broaden curve are observed for D149 molecules. It indicates comparatively reduced long range order in D149 dye molecules than CDCA molecules. Moreover, a distinct g (r) curve of CDCA indicates more structure formation in CDCA molecules than D149 molecules in a mixture of such molecules. The g(r) results for the clusters **(3)** D205/128 and cluster **(4)** D205-F/128 are given with cluster **(2)** D149/128 in Fig. 4. It results in influence of semi-perfluorinated, ethyl and octyl alkyl chains on structure formation. The maxima of g(r) value for D149 (ethyl) are larger than D205 (octyl) and D205-F (semi-perfluorinated octyl alkyl chain) which indicates reduced aggregation in D205, D205-F dyes in comparison to D149 dye. The g(r) values of D205-F dye molecules show two maxima at 3.9Å and 4.3Å

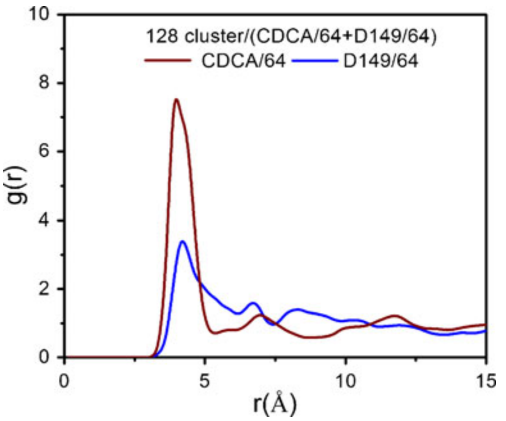


Fig. 3 Radial atom pair distribution function g(r) for CDCA and D149 in cluster (**5**) (CDCA/64+D149/64); [T=300 K, t=3 ns]

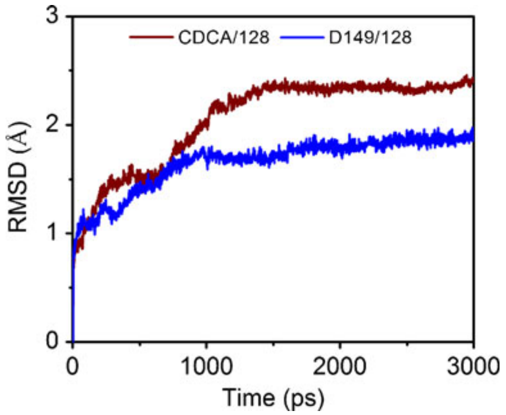


Fig. 5 Root mean square deviations values for CDCA and D149 in cluster (**1**). CDCA/128 and cluster (**2**). D149/128; [T=300 K, t=3 ns]

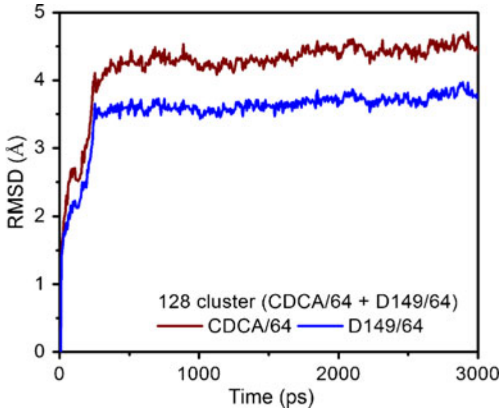


Fig. 6 Root mean square deviations values for CDCA and D149 in cluster (**5**). (CDCA/64+D149/64) ; [T=300 K, t=3 ns]

which results in the possibility of different arrangement of D205-F dye molecules.

The conformational flexibility of dye molecules are estimated by root mean square (RMSD) values. The RMSD values of clusters **(1)** CDCA/128, cluster **(2)** D149/128 and cluster **(5)** mixture; [(CDCA/64+D149/64)/128] are given in Figs. 5 and 6. In individual clusters CDCA molecules show larger conformational flexibility than D149 dye molecules (see Fig. 5). However, addition of CDCA to D149 dye molecules slightly increases conformational flexibility of both molecules in cluster **(5)** than in their individual clusters **(1)** and **(2)** (see Fig. 6). The conformational flexibility of the core and terminal chain parts of D205 and D205-F dye molecules was estimated in cluster **(3)** D205/128 and cluster **(4)** D205-F/128. The RMSD values for these two clusters **(3)** and **(4)** are given in Fig. 7. The core part of D205-F dye molecule shows larger conformational flexibility than D205 dye molecules. It indicates semi-perfluorinated alkyl chain have influence on conformational flexibility of core part of D205-F which is not observed by octyl alkyl chain (D205).

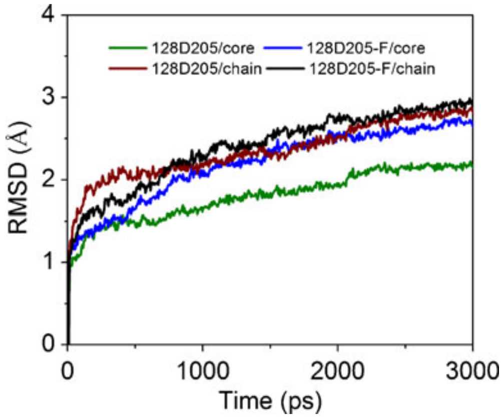


Fig. 7 Root mean square deviations values of core and terminal chain part of D205, D205-F dye molecules in cluster (**3**): D205/128 and cluster (**4**): D205-F/128; [T=300 K, t=3 ns]

Table 2 Diffusion coefficients of the clusters (T=300 K, t=3ns)

Cluster name	D ($10^{-12}m^2s^{-1}$)
(1) CDCA/128	47.80
(2) D149/128	24.80
(3) D205/128	46.88
(4) D205-F/128	73.64
(5) (CDCA+D149)/128; **CDCA/64**	34.98
(5) (CDCA+D149)/128; **D149/64**	30.46

The calculated diffusion coefficients values of five different clusters are presented (see Table 2). The D205-F (semi-perfluorinated alkyl chain) dye shows larger diffusion value than D149, D205 dyes and CDCA molecules. The addition of CDCA to D149 dye molecules leads to larger diffusion value for D149 and lower value for CDCA molecules. The D205 (octyl) dye shows larger diffusion value than D149 (ethyl) molecule. The calculated diffusion coefficients values are supporting conformational RMSD findings. Moreover, diffusion value of dye molecules show correlation with open-circuit voltage (V_{oc}) and power conversion efficiency (η) values of DSSCs (see Tables 1 and 2).

Conclusions

Atomistic MD simulation results adding CDCA to D149 dye molecules significantly increases conformational flexibility, diffusion values of dye molecules as well as reduce aggregation in dye molecules. Dye molecule D205 with longer alkyl chains shows larger mobility than D149 dye with shorter alkyl chain. Replacing alkyl chains (D205) by semi-perfluorinated alkyl chains (D205-F) in dye molecules leads to increase in conformational flexibility and significantly much larger diffusion coefficients values in perfluorinated dye molecules. Increased conformational flexibility, diffusion values of dye and reduced aggregation of dye are directly correlated with faster electron transfer from excited state of dye to TiO_2 semiconductor in DSSCs. Therefore designing highly flexible and conjugated dye molecules are necessary for complete harvesting of sun light in DSSCs. Thus open-circuit voltage and power conversion efficiency of DSSCs are directly correlated with larger conformational flexibility and diffusion values of organic dye molecules.

Acknowledgment Dr. Ananda Rama Krishnan Selvaraj is thankful to Kyushu Institute of Technology, Kitakyushu, Fukuoka, Japan for post-doctoral fellowship. He is also thankful to Dr. Takeomi Morita for his help on installation of AMBER program.

References

1. Regan BO, Grätzel M (1991) Nature 353:737–740
2. Chiba Y, Islam A, Watanabe Y, Komiya R, Koide K, Han L (2006) Jpn J Appl Phys 45:L638–L640
3. Gao F, Wang Y, Shi D, Zhang J, Wang M, Jing X, Humphry-Baker R, Wang P, Zakeeruddin SM, Grätzel M (2008) J Am Chem Soc 130:10720–10728
4. Grätzel M (2009) Acc Chem Res 42:1788–1798
5. Mishra A, Fischer MKR, Baurele P (2009) Angew Chem Int Edn 48:2474–2499
6. Burke A, Schmidt-Mende L, Ito S, Grätzel M (2007) Chem Commun 234–236
7. Horiuchi T, Miura H, Sumioka K, Uchida S (2004) J Am Chem Soc 126:12218–12219
8. Ito S, Miura H, Uchida S, Takata M, Sumioka K, Liska P, Comte P, Pechy P, Grätzel M (2005) Chem Commun 5194–5196
9. Kuang D, Uchida S, Humphry-Baker R, Zakeeruddin SM, Grätzel M (2008) Angew Chem Int Edn 47:1923–1927
10. Kay A, Grätzel M (1993) J Phys Chem 97:6272–6277
11. Neale NR, Kopidakis N, van de Lagemaat J, Grätzel M, Frank AJ (2005) J Phys Chem B 109:23183–23189
12. Sakaguchi S, Pandey SS, Okada K, Yamaguchi Y, Hayase S (2008) App Phys Express 1(105001):1–3
13. Bahers TL, Pauporte T, Scalmani G, Adamo C, Ciofini I (2009) Phys Chem Chem Phys 11:11276–11284
14. Pastore M, De Angelis F (2010) ACS Nano 4:556–562
15. Shlyk-Kerner O, Samish I, Kaftan D, Holland N, Sai PSM, Kless H, Scherz A (2006) Nature 442:827–830
16. Nojiri M, Koteishi H, Nakagami T, Kobayashi K, Inoue T, Yamaguchi K, Suzuki S (2009) Nature 462:117–120
17. Case DA, Pearlman DA, Caldwell JW, Cheatham TE, Wang J, Ross WS, Simmerling CL, Darden TA, Merz KM, Stanton RV, Cheng AL, Vincent JJ, Crowley M, Tsui V, Gohlke H, Radmer RJ, Duan Y, Pitera J, Massova I, Seibel GL, Singh UC, Weiner PK, Kollman PA (2002) AMBER 10. University of California, San Francisco
18. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) J Comput Chem 26:1668–1688
19. Friedemann R, Naumann S, Brickmann J (2001) Phys Chem Chem Phys 3:4195–4199
20. Monecke P, Friedemann R, Naumann S, Csuk R (1998) J Mol Model 4:395–404
21. Ananda Rama Krishnan S, Weissflog W, Pelzl G, Diele S, Kresse H, Vakhovskaya Z, Friedemann R (2006) Phys Chem Chem Phys 8:1170–1177
22. Ananda Rama Krishnan S, Weissflog W, Friedemann R (2007) J Mol Model 13:907–917
23. Jakalian A, Bush BL, Jack DB, Bayly CL (2000) J Comput Chem 21:132–146
24. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) J Comput Phys 23:327–341
25. Allen MP, Tildesley DJ Computer simulations of liquids. Clarendon, Oxford

J Mol Model (2012) 18:2105–2115
DOI 10.1007/s00894-011-1229-7

ORIGINAL PAPER

A B3LYP and MP2(full) theoretical investigation into explosive sensitivity upon the formation of the molecule-cation interaction between the nitro group of 3,4-dinitropyrazole and H^+, Li^+, Na^+, Be^{2+} or Mg^{2+}

Shan Du · Yong Wang · Li-zhen Chen · Wen-jing Shi · Fu-de Ren · Yong-xiang Li · Jian-long Wang · Duan-lin Cao

Received: 11 May 2011 / Accepted: 22 August 2011 / Published online: 9 September 2011

Abstract The explosive sensitivity upon the formation of molecule-cation interaction between the nitro group of 3,4-dinitropyrazole (DNP) and H^+, Li^+, Na^+, Be^{2+} or Mg^{2+} has been investigated using the B3LYP and MP2(full) methods with the 6-311++G** and 6-311++G(2df,2p) basis sets. The bond dissociation energy (*BDE*) of the C3–N7 trigger bond has also been discussed for the DNP monomer and the corresponding complex. The interaction between the oxygen atom of nitro group and H^+ in DNP…H^+ is partly covalent in nature. The molecule-cation interaction and bond dissociation energy of the C3–N7 trigger bond follow the order of DNP…Be^{2+} > DNP…Mg^{2+} > DNP…Li^+ > DNP…Na^+. Except for DNP…H^+, the increment of the trigger bond dissociation energy in comparison with the DNP monomer correlates well with the molecule-cation interaction energy, natural charge of the nitro group, electron density $\rho_{BCP(C3–N7)}$, delocalization energy $E^{(2)}$ and NBO charge transfer. The analyses of atoms in molecules (AIM), natural bond orbital (NBO) and electron density shifts have shown that the electron density of the nitro group shifts toward the C3–N7 trigger bond upon the formation of the molecule-cation interaction. Thus, the trigger bond is strengthened and the sensitivity of DNP is reduced.

Keywords 3,4-dinitropyrazole · Molecule-cation interaction · MP2 · Sensitivity · Trigger bond

Introduction

The search for new and thermally stable insensitive explosives has long been a primary goal in the field of energetic materials chemistry in order to avoid the catastrophic explosions in use and meet the requirements of military applications [1–4]. Therefore, recently much attention has been paid to investigate the relationship between the sensitivity and structure of the energetic compounds [5–21]. Introducing the desensitizing agents (such as aquadag, stearic acid and superpolymer) and certain functional groups into the structures of explosives have become the main methods to reduce explosive sensitivity [22].

The experimental measure of the sensitivity is dangerous and difficult. Thus, the theoretical prediction of the sensitivity and selection to the way of reducing the sensitivity become very urgent. Peter et al. have examined the effects of electric fields upon the trigger bonds using the B3PW91/6-31 G** method. It was found that the fields interact favorably energetically with the molecules and increase the stretching frequencies of the trigger bonds. The results show that the field-induced effects can have a direct bearing upon sensitivity to accidental detonation [23, 24]. Furthermore, many theoretical investigations have shown that the explosive sensitivity has a good linear relationship with the bond dissociation energy (*BDE*) of the trigger bond or the charge of nitro group [9, 10, 12–14, 16, 17, 20,

S. Du · Y. Wang · L.-z. Chen · F.-d. Ren (✉) · Y.-x. Li · J.-l. Wang · D.-l. Cao
College of Chemical Engineering and Environment,
North University of China,
Taiyuan 030051, China
e-mail: renfude@hotmail.com

W.-j. Shi
The Third Hospital of Shanxi Medical University,
Taiyuan 030051, China

21]. In particular, it has been confirmed that the intermolecular interaction can reduce the explosive sensitivity [22]. It is well known that molecule-cation interaction is one of the strongest interactions [25–31]. This suggests that the molecule-cation interaction may reduce available explosive sensitivity. However, to our knowledge, no investigation into the effect of the molecule-cation interaction on the explosive sensitivity or the *BDE* of the trigger bond has been presented.

For the explosive with nitro group, many researchers believe that the weakest bond linked nitro group, such as C–NO_2, N–NO_2 or O–NO_2, is the trigger spot [32, 33]. If the molecule-cation interaction can occur between cation and nitro group, the $\pi_3{}^4$ system of nitro group may be destroyed due to the π-electron rearrangement upon metal cation addition. Thus, the π-electron might transfer easily from nitro group to the C–N, N–N and O–N bonds. As a result, the trigger bond may be strengthened and the sensitivity of the explosive with nitro group might be reduced.

Recently, our group has been devoted to investigations on the synthesis and properties of 3,4-dinitropyrazole (DNP) [34], which is reported as a melt-cast explosive for replacing TNT [35]. To find out the way of reducing the sensitivity of DNP, in this paper, we will investigate theoretically the correlation of the sensitivity with the molecule-cation interaction between the nitro group of DNP and H^+, Li^+, Na^+, Be^{2+} or Mg^{2+}. On the other hand, lots of investigations have shown that the intensity of trigger bond correlates with the bond length and bond dissociation energy of trigger bond as well as the nitro group charges [16, 36]. Therefore, we will also analyze the changes of the bond length and bond dissociation energy of the trigger bond as well as nitro group charge upon the formation of the molecule-cation interaction. The analyses of atoms in molecules (AIM), natural bond orbital (NBO) and electron density shifts will be applied to explain the nature of these changes.

Computational details

As a cheap and effective approach, density functional theory (DFT) is feasible to optimize the geometry of the high energetic materials, while for the energetic stability, the value by the MP2 method is quite close to the experimental result [37–42]. In addition, the high quality basis set is a crucial factor for calculating the property of the complex [43, 44]. Taking the factors above in balance, we decide to use the DFT-B3LYP and MP2(full) methods with the 6-311++G** and 6-311++G(2df,2p) basis sets in this work.

All calculations have been performed with Gaussian 03 programs [45]. The title complexes have been fully optimized using the DFT-B3LYP and MP2(full) methods with the 6-311++G** and 6-311++G(2df,2p) basis sets, and the structures corresponding to the minimum energy points at the molecular energy hypersurface (NImag=0) have been obtained. Single point energy calculations have been carried out at the same levels. The NBO method [46] and the shifts of the electron density [47] that accompanies the formation of the complex have been analyzed at B3LYP/6-311++G(2df,2p) level, and the topological charge density has been displayed by the AIM method [48] using AIMPAC program [49] at the same level.

The *BDE* of the C–NO_2 bond has been calculated. It is defined as:

$$BDE = E_{(R\cdot)} + E_{(\cdot \mathrm{NO_2} \ldots \mathrm{M})} - E_{(\mathrm{RNO_2} \ldots \mathrm{M})}. \tag{1}$$

R· is 4-nitropyrazole radical. M is H^+, Li^+, Na^+, Be^{2+} or Mg^{2+}.

Molecule-cation interaction ($E_{\mathrm{int.}}$) has been investigated with the definition of the energy difference between the complex and isolated monomer.

$$E_{\mathrm{int.}} = E_{(\mathrm{DNP} \ldots \mathrm{M})} - E_{(\mathrm{DNP})} - E_{(\mathrm{M})} \tag{2}$$

$E_{\mathrm{int.}}$ is corrected with the basis set superposition error (BSSE) [50, 51] and zero-point energy (ZPE) corrections.

The nitro group charge Q_{NO2} is calculated as Eq. 3.

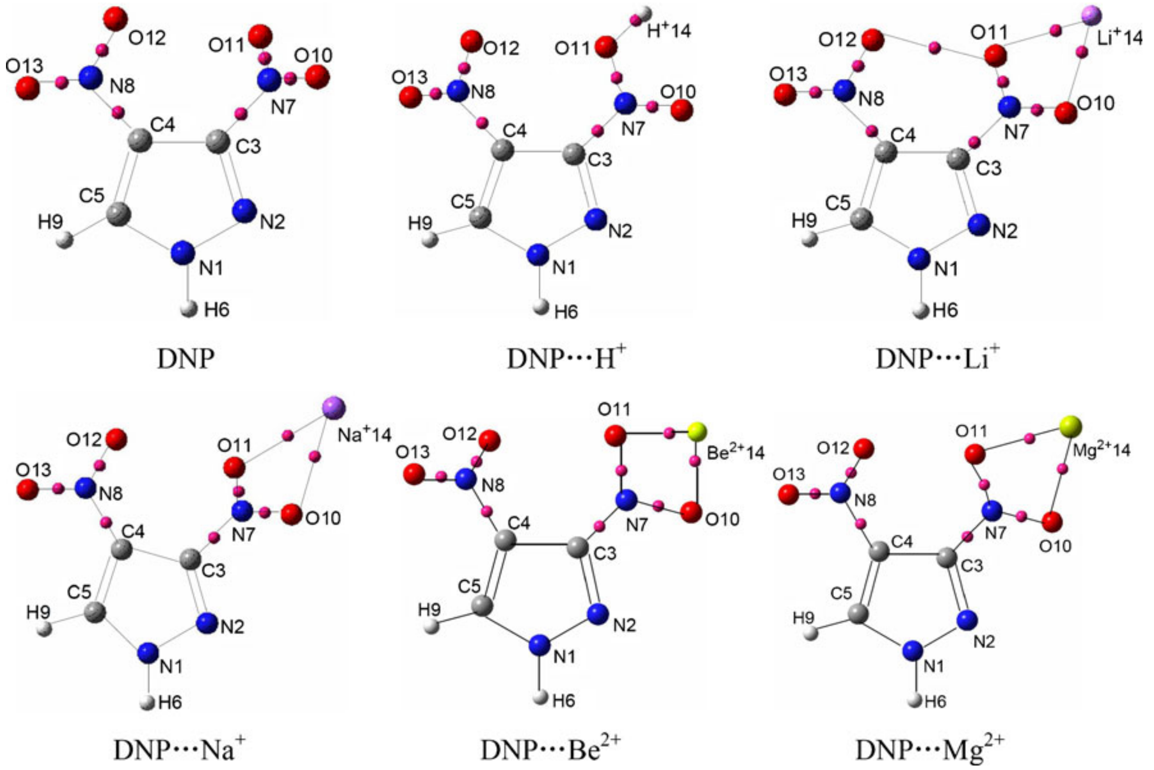
$$Q_{\mathrm{NO_2}} = Q_{\mathrm{N}} + Q_{\mathrm{O1}} + Q_{\mathrm{O2}} \tag{3}$$

The Q_{N}, Q_{O1} and Q_{O2} are the charges on the N and the O atoms of the nitro group, respectively.

Results and discussion

The structures and bond critical points (BCPs) of the complexes are shown in Fig. 1. The selected geometric parameters are listed in Table 1. The molecule-cation interaction energies of the complexes and the bond dissociation energies of trigger bonds are presented in Table 2. The analyses of AIM and NBO, Mulliken and natural charges are given in Tables 3, 4 and 5, respectively.

Our preliminary calculations show the C3–N7 bond is longer than the C4–N8 bond in the DNP monomer at both B3LYP/6-311++G(2df,2p) and MP2(full)/6-311++G** levels, indicating that the C3–N7 bond might be the trigger bond. So in this work, we will mainly pay attention to the molecule-cation interaction between C3–NO_2 and H^+, Li^+, Na^+, Be^{2+} or Mg^{2+}.



DNP DNP···H^+ DNP···Li^+

DNP···Na^+ DNP···Be^{2+} DNP···Mg^{2+}

Fig. 1 Molecular structures, bond critical points of the complexes at B3LYP/6-311++G(2df,2p) level. Small purple spheres (unlabeled) represent bond critical points

Table 1 Selected bond length (in Å) of the DNP monomer and the complexes

Parameters	DNP		DNP…H^+		DNP…Li^+		DNP…Na^+		DNP…Be^{2+}		DNP…Mg^{2+}	
O11…M14[a]			0.983[b]	0.982[c]	2.043	2.031	2.397	2.392	1.592	1.576	2.046	2.016
			0.983[d]		2.076		2.438		1.618		2.074	
C3–N7	1.469	1.466	–0.062[e]	–0.062	–0.045	–0.044	–0.032	–0.031	–0.144	–0.145	–0.107	–0.107
	1.451		–0.050		–0.025		–0.016		–0.103		–0.078	
C4–N8	1.441	1.436	0.023	0.024	0.016	0.017	0.011	0.013	0.030	0.031	0.027	0.028
	1.443		0.013		0.006		0.004		0.022		0.016	
N7–O10	1.217	1.215	1.184	1.182	1.237	1.236	1.231	1.229	1.317	1.316	1.277	1.277
	1.229		1.202		1.241		1.237		1.305		1.271	
N7–O11	1.218	1.216	1.335	1.333	1.238	1.236	1.233	1.230	1.311	1.309	1.275	1.273
	1.231		1.319		1.242		1.238		1.296		1.271	
N8–O12	1.222	1.220	1.219	1.217	1.221	1.219	1.223	1.221	1.219	1.217	1.220	1.218
	1.230		1.228		1.231		1.232		1.227		1.228	
N8–O13	1.227	1.224	1.215	1.213	1.217	1.216	1.218	1.216	1.205	1.203	1.209	1.207
	1.231		1.226		1.226		1.226		1.221		1.224	

[a] M14 is H^+, Li^+, Na^+, Be^{2+} or Mg^{2+} in the corresponding complexes

[b] At B3LYP/6-311++G** level

[c] At B3LYP/6-311++G(2df,2p) level

[d] At MP2(full)/6-311++G** level

[e] The difference of C–N bond in the complex in comparison with the monomer DNP

Table 2 Interaction energy $-E_{\text{int.}}$ (kJ mol^{-1}) and bond dissociation energy (*BDE* (kJ mol^{-1})) of the complexes

Parameters	DNP	DNP…H$^+$		DNP…Li$^+$		DNP…Na$^+$		DNP…Be^{2+}		DNP…Mg^{2+}	
$E_{\text{int.}}$		782.02 [a]	–[b]	171.37	168.85	127.88	125.60	937.31	933.04	531.79	528.07
		(750.28 [c])[d]		163.87		122.55		927.03		524.68	
		786.80 [e]	–	174.47	171.75	129.77	126.35	963.43	962.10	544.41	540.77
		(755.00)		166.73		123.36		955.45		537.20	
		772.08[f]	–	166.29	155.12	123.52	114.84	884.16	855.48	487.65	472.11
		(741.18)		149.89		111.62		850.58		468.38	
$BDE_{\text{C3–N7}}$[g]	275.58 276.98	520.44	519.04	449.66	449.19	407.93	407.35	1021.98	1022.47	760.26	754..92
	375.76	675.68		534.98		496.78		1069.67		823.30	
$BDE_{\text{C4–N8}}$	312.85 319.46	261.99	266.88	277.54	281.96	287.32	291.58	243.05	247.77	251.82	252.44
	400.8	323.75		365.43		378.29		–		–	

[a] The energies are uncorrected at B3LYP/6-311++G** level

[b] The interaction energies with BSSE-corrected

[c] The values in parenthesis are energies with ZPE corrections

[d] The interaction energies with ZPE and BSSE correction

[e] The energies are uncorrected at B3LYP/6-311++G(2df,2p) level

[f] The energies are uncorrected at MP2(full)/6-311++G** level

[g] The BDE is the difference between the complex and the 4-nitro group radical as well as the 3-nitropyrazole radical with the M cation

Structure of the complex

From Fig. 1, all the complexes are C_1 symmetry. As can be seen from Table 1, the O11…M14 distance is 2.076 and 2.438 Å in DNP…Li$^+$ and DNP…Na$^+$ at MP2(full)/6-311++G** level, respectively. In the molecule-cation complex Li$^+$(CH$_3$NO$_2$), the O…Li$^+$ distance is predicted to be 2.930 Å using CNDO method [52]. For the molecule-cation system of paranitroaniline (PNA) with Na$^+$, the distance of O…Na$^+$ is 2.302 Å at the B3LYP/6-31+ G* level [53]. For comparison, we have also obtained the corresponding O…Li$^+$ and O…Na$^+$ distances of 2.091 and 2.371 Å by employing the MP2(full)/6-311++G** method for Li$^+$…(CH$_3$NO$_2$) and PNA…Na$^+$, respectively. Comparing the results at the MP2(full)/6-311++G** level, it can be seen that the O…Li$^+$ and O…Na$^+$ distances in the title complexes are close to those in Li$^+$(CH$_3$NO$_2$) and PNANa$^+$, respectively. In BeO and MgO, the experimental values of O–Be and O–Mg bond lengths are 1.331 and 1.749 Å, respectively [54]. The

Table 3 The selected bond critical point properties (in a.u.) within the complexes and that of the monomer DNP at B3LYP/6-311++G(2df,2p) level

Parameters	DNP	DNP…H$^+$	DNP…Li$^+$	DNP…Na$^+$	DNP…Be^{2+}	DNP…Mg^{2+}
ρ_{BCP}(O11…M14)	–	0.3493	0.0246	0.0186	0.0981	0.0484
$\nabla^2\rho_{\text{BCP}}$(O11…M14)	–	−2.7301	0.1532	0.1041	0.6028	0.3190
ρ_{BCP}(O10…M14)	–	–	0.0240	0.0171	0.1010	0.0489
$\nabla^2\rho_{\text{BCP}}$(O10…M14)	–	–	0.1481	0.0930	0.6205	0.3227
ρ_{BCP}(C3–N7)	0.2775	0.2902	0.2872	0.2826	0.3354	0.3140
$\nabla^2\rho_{\text{BCP}}$(C3–N7)	−0.8156	−0.7338	−0.8567	−0.8683	−0.4923	−0.6164
ρ_{BCP}(N7–O10)	0.5170	0.5608	0.4923	0.4999	0.4033	0.4431
$\nabla^2\rho_{\text{BCP}}$(N7–O10)	−1.0752	−1.3329	−0.9403	−0.9849	−0.5016	−0.6872
ρ_{BCP}(N7–O11)	0.5160	0.3879	0.4927	0.4987	0.4109	0.4478
$\nabla^2\rho_{\text{BCP}}$(N7–O11)	−1.0692	−0.5281	0.1228	−0.9833	−0.5441	−0.7220
ρ_{BCP}(C4–N8)	0.2882	0.2805	0.2827	0.2839	0.2808	0.2796
$\nabla^2\rho_{\text{BCP}}$(C4–N8)	−0.8794	−0.7913	−0.8218	−0.8375	−0.7485	−0.7790
ρ_{BCP}(N8–O12)	0.5104	0.5149	0.5122	0.5099	0.5147	0.5133
$\nabla^2\rho_{\text{BCP}}$(N8–O12)	−1.0270	−1.0427	−1.0323	−1.0216	−1.0366	−1.0310
ρ_{BCP}(N8–O13)	0.5050	0.5196	0.5165	0.5156	0.5331	0.5286
$\nabla^2\rho_{\text{BCP}}$(N8–O13)	−0.9946	−1.0660	−1.0519	−1.0487	−1.1404	−1.1221

Table 4 NBO occupation numbers for the C3–N7 and M (H^+, Li^+, Na^+, Be^{2+}, Mg^{2+}) bonds , their respective orbital energies ε, the second-order perturbation energies $E^{(2)}$ and the NBO charge transfer of the monomers DNP in their complexes (Q) at B3LYP/6-311++G(2df,2p) level

Parameters	DNP…H^+	DNP…Li^+	DNP…Na^+	DNP…Be^{2+}	DNP…Mg^{2+}
Occ.(O10/O11)[a]	1.7723sp$^{99.99}$sp$^{99.99}$	1.9044sp$^{99.99}$	1.9015sp$^{99.99}$	1.8740sp$^{6.75}$	1.8830sp$^{23.01.}$
ε\{(O10/O11)\}[b]	−0.5986	−0.5352	−0.5049	−0.9085	−0.7582
Occ.(M14)*	**0.4724sp$^{99.99}$sp$^{99.99}$**	0.0360sp$^{0.11}$	0.0313sp$^{0.06}$	0.1468sp$^{0.08}$	0.1271sp$^{0.02}$
ε\{(M14)*\}	**−0.3710[d]**	0.0098	−0.0704	−0.2429	−0.3852
$E^{(2)}{}_{(O10/O11)\rightarrow(M14)*}$[c]	203.54	17.31	9.99	171.88	77.45
Occ.(N7–O11)	**1.9870sp$^{3.33}$s**	1.9955sp$^{2.26}$sp$^{3.21}$	1.9937sp$^{2.19}$sp$^{3.17}$	1.9933sp$^{2.69}$sp$^{4.41}$	1.9936sp$^{2.46}$sp$^{3.65}$
ε\{(N7–O11)\}	**−1.0146**	−1.2853	−1.2711	−1.4043	−1.4041
Occ.(C3–N7)*	0.0773sp$^{2.65}$sp$^{1.45}$	0.0885sp$^{2.65}$sp$^{1.62}$	0.0936sp$^{2.68}$sp$^{1.71}$	0.0533sp$^{2.31}$sp$^{1.18}$	0.0634sp$^{2.44}$sp$^{1.38}$
ε\{(C3–N7)*\}	0.0935	0.1162	0.1171	0.0429	0.0404
$E_{(N7-O11)\rightarrow(C3-N7)*}{}^{(2)}$	16.10	4.39	3.93	7.15	6.19
Q(DNP)[e]	0.00	46.50	36.04	230.71	145.73

[a] Occ.: occupation number. The Occ.(O10/O11) in DNP…H^+ is Occ.(O11) and in DNP…Li^+ or DNP…Na^+ DNP…Be^{2+} or DNP…Mg^{2+} is Occ.(O10)

[b] In a.u. The ε\{(O10/O12)\} in DNP…H^+ is ε (O11) and in DNP…Li^+ or DNP…Na^+ DNP…Be^{2+} or DNP…Mg^{2+} is ε(O10)

[c] In kJ mol^{-1}

[d] The bold Occ.(M14)* and ε\{(M14)*\} for DNP…H^+ is Occ.(N7–O10)*and ε\{(N7–O10)*\}. The bold Occ.(N7–O11) and ε\{(N7–O11)\} for DNP…H^+ is Occ.(O11–H14) and ε\{(O11–H14)\}

[e] In me

O11…M14 distance is 1.618 and 2.074 Å in DNP…Be^{2+} and DNP…Mg^{2+} at MP2(full)/6-311++G** level, respectively, which are only about 0.3 Å larger than those in the ionic compounds BeO and MgO, suggesting that the molecule-cation interactions in title complexes might be significant. For the complex DNP…H^+, the O11–H14 bond length is only 0.983, 0.982 and 0.983 Å at B3LYP/6-311++G**, B3LYP/6-311++G(2df,2p) and MP2(full)/6-311++G** levels, respectively. Here, we have also calculated the structure of H_3O^+ and the O–H bond length is found to be 0.9803, 0.9791 and 0.9778 Å at B3LYP/6-311++G**, B3LYP/6-311++G(2df,2p) and MP2 (full)/6-311++G** levels, respectively. The O–H bond lengths in DNP…H^+ are very close to those in H_3O^+ at three levels, indicating that the interaction between O11 and H14 in DNP…H^+ is partly covalent in nature.

As is shown in Table 1, the order of the O11…M14 distance is DNP…Be^{2+} < DNP…Mg^{2+} < DNP…Li^+ < DNP…Na^+ at three levels, suggesting that the order of the molecule-cation interaction might be DNP…Be^{2+} > DNP…Mg^{2+} > DNP…Li^+ > DNP…Na^+.

From Table 1, the C3–N7 bond length in molecule-cation complex decreases in comparison with that in the DNP monomer. For example, the decrease is 0.025, 0.016, 0.103 and 0.078 Å in DNP…Li^+, DNP…Na^+ , DNP…Be^{2+} and DNP…Mg^{2+} at MP2(full)/6-311++G** level, respectively, showing that the C3–N7 trigger bond is strengthened upon the formation of the molecule-cation interaction. The stronger the trigger bond of the explosive molecule, the greater the insensitivity [36], suggesting that the sensitivity might be reduced upon the formation of the molecule-cation interaction. The decrease of the C3–N7 bond length is the same order of DNP…Be^{2+} > DNP…Mg^{2+} > DNP…H^+ > DNP…Li^+ > DNP…Na^+ at three levels, suggesting that the sensitivity might be DNP…Be^{2+} < DNP…Mg^{2+} < DNP…H^+ < DNP…Li^+ < DNP…Na^+.

The N7–O10 and N7–O11 bond lengths are almost close to each other in the DNP monomer and the complexes DNP…Li^+, DNP…Na^+, DNP…Be^{2+} and DNP…Mg^{2+}. However, in the complex DNP…H^+, the N7–O10 bond length decreases while the N7–O11 bond length increases. The difference between them is up to 0.151, 0.151 and 0.117 Å at B3LYP/6-311++G**, B3LYP/6-311++G(2df,2p) and MP2(full)/6-311++G** levels, respectively, suggesting that the conjugated system of nitro group is destroyed and the π-electron rearrangement occurs upon the H^+ addition. It should be noted that, in the complex DNP…Li^+, DNP…Na^+, DNP…Be^{2+} or DNP…Mg^{2+}, the distance of metal cation with the two oxygen atoms of the nitro group is equal to each other, but in DNP…H^+, the distance of H^+…O11 is only 0.983 Å while that of H^+…O10 is up to 2.244 Å at MP2(full)/6-311++G** level.

Energies and stabilities

Table 2 gives the molecule-cation interaction energy and bond dissociation energy in the DNP monomer and complex. From Table 2, for the DNP…H^+, the interaction energy is up to 782.02, 786.80 and 772.08 kJ mol^{-1} at B3LYP/6-311++G**, B3LYP/6-311++G(2df,2p) and MP2 (full)/6-311++G** levels, respectively. These values are close to the σ-binding energy of our previous study on the system of the HB with H$^+$ (894.03 kJ mol^{-1} at B3LYP/6-311++G(2df,2p) level) [25], suggesting that the σ-binding interaction may have occurred between O11 and H$^+$, as is in agreement with the analysis of the structure of DNP…H$^+$.

As is shown in Table 2, the molecule-cation interaction energy is the same order of DNP…Be^{2+} > DNP…Mg^{2+} > DNP…Li$^+$ > DNP…Na$^+$ at three levels, as is in agreement with the analysis of structure. In DNP…Li$^+$ and DNP…Na$^+$, the molecule-cation interaction energy is only 166.29 and 123.52 kJ mol^{-1} at MP2(full)/6-311++G** level, respectively, while for DNP…Be^{2+} and DNP…Mg^{2+}, it is up to 884.16 and 487.65 kJ mol^{-1}, respectively. The molecule-cation interaction in the complex of the alkali metal cation with DNP is much weaker than that in the corresponding system of the alkaline-earth metal cation, as is very similar to the cation-π interaction [55]. For example, in our previous investigation, for the complex HB=BH…Li$^+$ or HB=BH…Na$^+$, the cation-π interaction is just equal to 112.88 or 75.84 kJ mol^{-1}, and it amounts to 708.08 and 405.45 kJ mol^{-1} for the Be^{2+} and Mg^{2+} complexes at MP2 (full)/aug-cc-pVTZ level, respectively [55].

The proportion of corrected molecule-cation interaction energies for the complexes to their total inteaction energies, defined as $[(-E_{\text{int.}})-(-E_{\text{int.(BSSE)}})]/(-E_{\text{int.}})$, is only 1.78%, 2.64% and 7.02% at B3LYP/6-311++G**, B3LYP/6-311++G(2df,2p) and MP2(full)/6-311++G** levels, respectively. Although it is not notable at three levels, the BSSE corrections for molecule-cation interaction energies are not negligible. In fact, there is a standard computational protocol, which requires BSSE corrections for molecule-cation interaction energies. Only in case of a complete basis set, the correction for BSSE is not needed. The ZPE correction is up to 4.06%, 4.04% and 4.00% at B3LYP/6-311++G**, B3LYP/6-311++G(2df,2p) and MP2(full)/6-311++G** levels, respectively. In our previous investigation on the cation-π bonded complexes of cations (Li$^+$, Na$^+$, Be^{2+} and Mg^{2+}) with the B=B double bond, the ZPE corrections amount to 2.52% and 2.77% at B3LYP/6-311++G(2df,2p) and MP2(full)/6-311++G(2df,2p) levels [55].

As can be seen from Table 2, the bond dissociation energy of the C3–N7 bond in the DNP monomer is 275.58, 276.98 and 375.76 kJ mol^{-1} at B3LYP/6-311++G**, B3LYP/6-311++G(2df,2p) and MP2(full)/6-311++G** levels, respectively. The results from the B3LYP methods are close to the experimental values of the bond dissociation energies of the C–N bonds linked the 1-nitro group of the 1,3-dinitrobenzene (278.10 kJ mol^{-1}) and 1,4-dinitrobenzene (280.19 kJ mol^{-1}) [56]. In surprise,

Table 5 Changes of Mulliken charges (in e) and natural charges (in e) of nitro group in the complexes in comparison with the DNP monomer

Parameters	DNP		DNP…H$^+$		DNP…Li$^+$		DNP…Na$^+$		DNP…Be^{2+}		DNP…Mg^{2+}	
N7O10O11	−0.1591[a]	−0.1682[b]	0.2681	0.2725	0.1839	0.2155	0.0050	0.0256	0.4194	0.2694	0.1923	0.2308
	−0.1678[c]	−0.1671[d]	0.2963	0.1347	0.1827	−0.1792	−0.0623	−0.1349	0.3247	−0.6228	0.0544	−0.4065
N8O12O13	−0.2306	−0.2949	0.1266	0.1298	0.0784	0.0890	0.0651	0.0706	0.2800	0.2803	0.2001	0.2056
	−0.2169	−0.2413	0.1043	0.0966	0.0586	0.0673	0.0222	0.0497	0.1217	0.2270	0.0733	0.1243
C3	0.0264	0.0833	−0.0095	0.0470	−0.0726	−0.0002	−0.1011	−0.0465	0.1853	0.2846	0.2671	0.2073
	−0.0362	0.2866	0.0438	−0.0561	−0.1252	−0.0491	−0.1723	−0.0422	0.2121	0.0277	0.2363	−0.0315
C4	−0.1065	0.1738	0.0964	−0.0713	−0.0078	−0.1398	0.0297	−0.0709	0.0199	−0.2128	−0.0222	−0.2193
	−0.0968	−0.0235	0.0591	0.0707	0.0102	0.0453	0.0791	0.0330	−0.0660	0.1246	0.1099	0.0909
M[e]			−0.6632	−0.6595	−0.3337	−0.4278	−0.1507	−0.2147	−1.3698	−1.2223	−0.8838	−0.9437
			−0.6626	−0.4735	−0.2935	−0.0465	−0.0972	−0.0360	−1.2176	−0.2307	−0.6726	−0.1457

[a] Mulliken charge at B3LYP/6-311++G** level

[b] Mulliken charge at B3LYP/6-311++(2df,2p) level

[c] Mulliken charge at MP2(full)/6-311++G** level

[d] Natural charge at B3LYP/6-311++G(2df,2p) level

[e] M are H$^+$, Li$^+$, Na$^+$, Be^{2+}, Mg^{2+} corresponding the complexes in the row

the dissociation energy obtained at MP2/6-311++G** level is about one hundred kJ mol^{-1} greater than the one obtained at B3LYP/6-311++G** level. Many theoretical investigations have shown that the B3LYP method correctly describes the *BDE* value. However, the MP2 method cannot be used to adequately describe the *BDE* value [57–59]. In 1995, the investigations by Jursic et al. confirmed that, for the O–O and O–C bond dissociation energies, the MP2 model gave unsatisfactory results, and calculation with the DFT-B3LYP method was required in order to obtain the satisfactory bond dissociation energy [60]. In 2002, Budyka et al. found that electron correlation correction at the MP2/6-31 G*//HF/6-31 G* level overestimated the C–N *BDE* value by about 67 kJ mol^{-1} compared to the experiment for $PhNHCH_3$, and B3LYP-calculated *BDE* value was in good agreement with experimental one [57]. Barckholtz et al. have also calculated the *BDE*s of the C-H and N-H bonds in monocyclic aromatic molecules [61]. They have also found that the B3LYP method yields *BDE*s that are on average lower than experiment by ~5-10 kJ mol^{-1}, and it provides the best agreement of the computed *BDE*s of the smaller aromatic hydrocarbons with experiment [61]. However, most of the MP2 *BDE*s calculated are about 72.0 kJ mol^{-1} higher than experiment. In addition, the MP2 calculations of the radical species suffer from significant spin contamination, with $< S^2 >$ as high as 1.4 for the radicals formed from benzene and the azabenzenes [61]. Indeed, in title compounds, for the MP2 method, the values of $< S^2 >$ are up to 1.15 for the nitropyrazole radical, while for the B3LYP method, they are about 0.75. Thus, those results from the MP2 method could not be close to experimental values due to the serious spin contamination. Therefore, in this paper, B3LYP is selected to elucidate the trends in the calculated bond dissociation energies of the complexes.

From Table 2, the bond dissociation energy of the C3–N7 bond in complex is greater than that in the DNP monomer. Especially for DNP…Be^{2+}, the bond dissociation energy of the C3–N7 bond is up to 1021.98, 1022.47 and 1069.67 kJ mol^{-1} at B3LYP/6-311++G**, B3LYP/6-311++G(2df,2p) and MP2(full)/6-311++G** levels, respectively. It is three times more than that in the DNP monomer (275.58, 276.98 and 375.76 kJ mol^{-1}) at three levels. This result shows that the strength of the C3–N7 trigger bond is enhanced and the explosive sensitivity is reduced upon the formation of molecule-cation interaction, as is in agreement with the analysis of structure. The order of the C3–N7 bond dissociation energy is DNP…Be^{2+} > DNP…Mg^{2+} > DNP…H^{+} > DNP…Li^{+} > DNP…Na^{+} > DNP at three levels, which is in accordance with the molecule-cation interaction energy.

The correlation between the bond dissociation energy of C3–N7 bond and the C3–N7 bond length is given in Fig. 2 at B3LYP/6-311++G(2df,2p) level. The correlation coefficient is up to 0.9940.

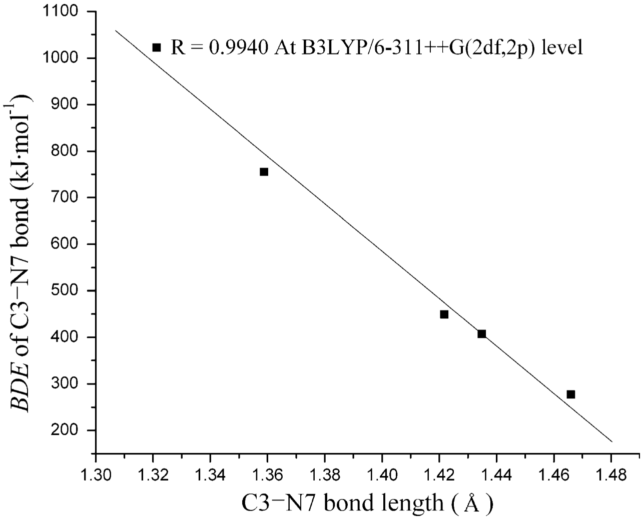


Fig. 2 The C3–N7 bond dissociation energy (*BDE*) of the DNP monomer and complex versus the C3–N7 bond length except for DNP…H^{+}

$$BDE = -5.087 \times 10^3 r + 7.707 \times 10^3 \tag{4}$$

BDE is in kJ·mol^{-1} and *r* is the C3–N7 bond length (in Å).

The relationship between the molecule-cation interaction energy ($E_{\mathrm{int.}}$) and increment of the C3–N7 bond dissociation energy in comparison with the DNP monomer (ΔBDE) is shown in Fig. 3 at B3LYP/6-311++G(2df,2p) level. The correlation coefficient is 0.9976 and they fit the Eq. 5:

$$\Delta BDE = -0.741 E_{\mathrm{int.}} + 45.895 \tag{5}$$

ΔBDE and E_{int} are in kJ·mol^{-1}.

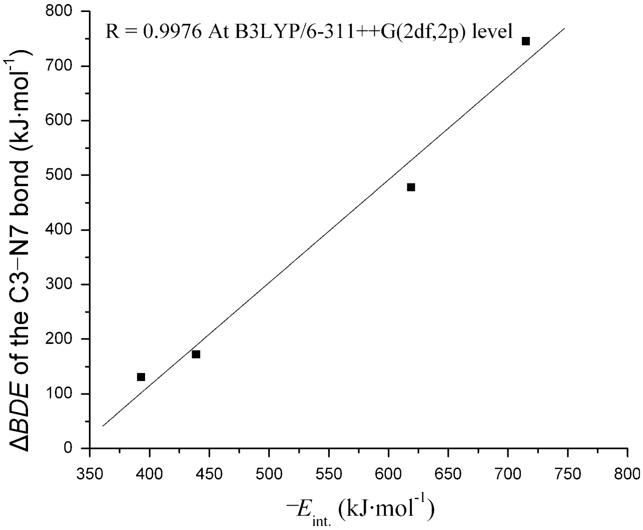


Fig. 3 The increment of C3-N7 bond dissociation energy (Δ*BDE*) in complex in comparison with the monomer DNP except for DNP…H^{+} versus interaction energy ($-E_{\mathrm{int.}}$)

AIM analysis

According to the AIM analysis at B3LYP/6-311++G(2df,2p) level, there is a bond path linking the cation M with the oxygen atom of the nitro group accompanied by a BCPs (see Fig. 1). From Table 4, except for DNP…H$^+$, the values of the electron densities $\rho_{BCP(O10\ldots M14)}$ are within the range of 0.0186−0.0484 a.u., which just falls into the common accepted values for intermolecular interactions (0.002–0.04 a.u.) [48]. Moreover, their Laplacians$\nabla^2\rho_{BCP}$ are all positive, suggesting the typical closed-shell kind of interactions in complexes. It is noted that, the $\rho_{BCP(O10\ldots H14)}$ value in DNP…H$^+$ is 0.3493, and its $\nabla^2\rho_{BCP}$ is negative, indicating that the interaction between O11 and H14 in DNP…H$^+$ is partly covalent, as is in agreement with the analysis of structure.

As can be seen from Table 3, the electron density ρ_{BCP} at the C3–N7 bond in the complex is larger than that in the DNP monomer. The increment in comparison with the DNP monomer is the order of DNP…Be^{2+} > DNP…Mg^{2+} > DNP…Li$^+$ > DNP…Na$^+$, as is in accordance with the bond length and bond dissociation energy of the C3–N7 bond as well as molecule-cation interaction energy of complex.

The correlation between the electron densities ($\rho_{BCP(C3-N7)}$) and the increment of the C3–N7 bond dissociation energy in comparison with the DNP monomer (ΔBDE) at B3LYP/6-311++G(2df,2p) level is also found (see Fig. 4). The correlation coefficient is up to 0.9995.

$$\Delta BDE = 1.167 \times 10^4 \rho_{BCP(C3-N7)} - 3.175 \times 10^3 \qquad (6)$$

In general, charge density at the BCP of a given bond can be used as an estimator of the bond strength. Therefore, the $\rho_{BCP(C3-N7)}$ values at the C3–N7 BCP should display a certain correlation with the *BDE* involved in the radical formation. The reason is the energy difference between the radicals and initial molecule mainly depends on the strength of the C3–N7 bond broken, which is similar to many of the previous investigations [62–64].

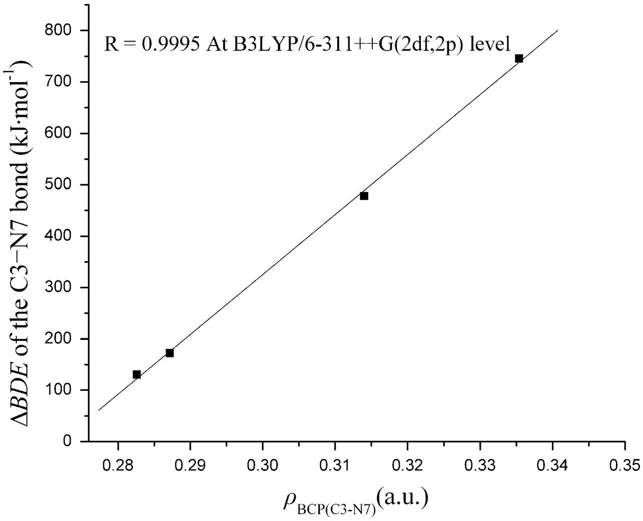


Fig. 4 The increment of C3-N7 bond dissociation energy (ΔBDE) in complex in comparison with the monomer DNP except for DNP…H$^+$ versus the bond critical point properties ($\rho_{BCP(C3-N7)}$)

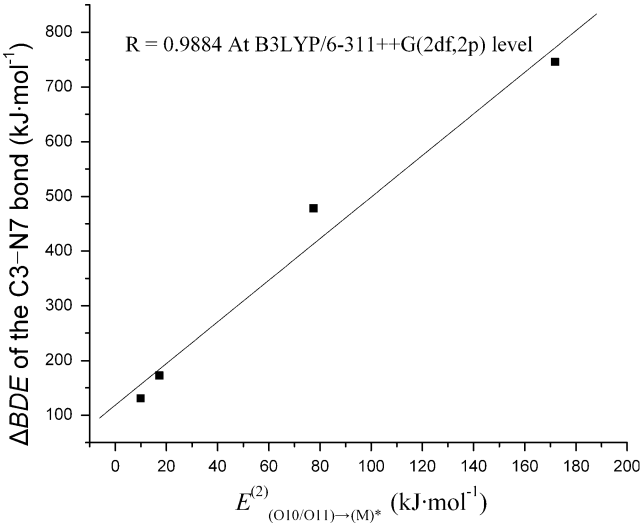


Fig. 5 The increment of C3–N7 bond dissociation energy (ΔBDE) in complex in comparison with the monomer DNP except for DNP…H$^+$ versus the delocalization energy $E_{(O10/O11)\rightarrow(M)^*}^{(2)}$

NBO analysis

The NBO results show all the complexes have two units except for DNP…H$^+$. The delocalization effects between

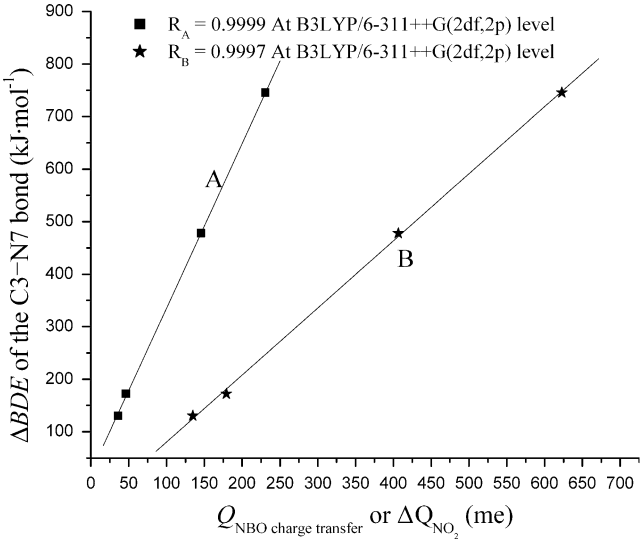


Fig. 6 The increment of the C3-N7 bond dissociation energy (ΔBDE) in complex in comparison with the monomer DNP except for DNP…H$^+$ versus NBO charge transfer ($Q_{NBO\ charge\ transfer}$ see A) and the change of the nature charge in the nitro group (ΔQ_{NO2} see B)

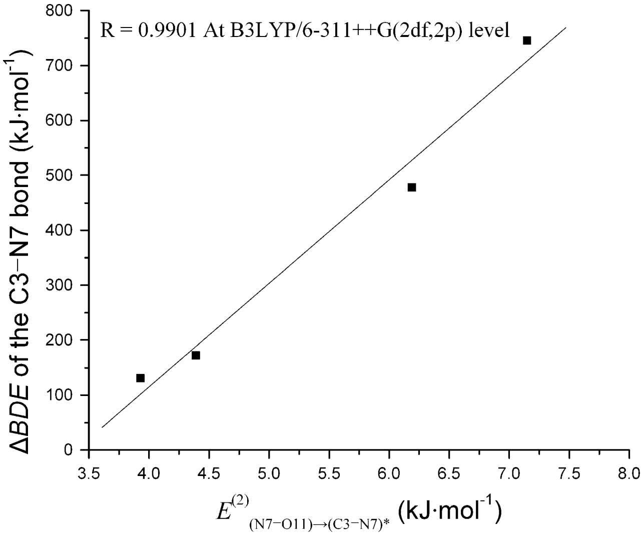


Fig. 7 The increment of C3-N7 bond dissociation energy (Δ*BDE*) in complex in comparison with the monomer DNP except for DNP…H^+ versus the delocalization energy ($E_{(N7-O11)\to(C3-N7)}{}^{(2)}$)

two units can be identified from the presence of off-diagonal elements of the Fock matrix in the NBO basis, and the strengths of these delocalization interactions, $E^{(2)}$ [46], can be estimated by second-order perturbation theory. The delocalization interactions $E_{(O10/O11)\to(M)^*}{}^{(2)}$ have stabilized the systems by 17.31, 9.99, 171.88 and 77.45 kJ mol^{-1} for DNP…Li^+, DNP…Na^+, DNP…Be^{2+} and DNP…Mg^{2+} (see Table 4). The correlation between the delocalization energy $E_{(O10/O11)\to(M)^*}{}^{(2)}$ and the increment of C3–N7 bond dissociation energy in comparison with the DNP monomer (Δ*BDE*) at B3LYP/6-311++G(2df,2p) level is given in Fig. 5. The correlation coefficient is up to 0.9884.

$$\Delta BDE = 3.801E^{(2)}{}_{(O10/O11)\to(M)^*} + 118.611 \tag{7}$$

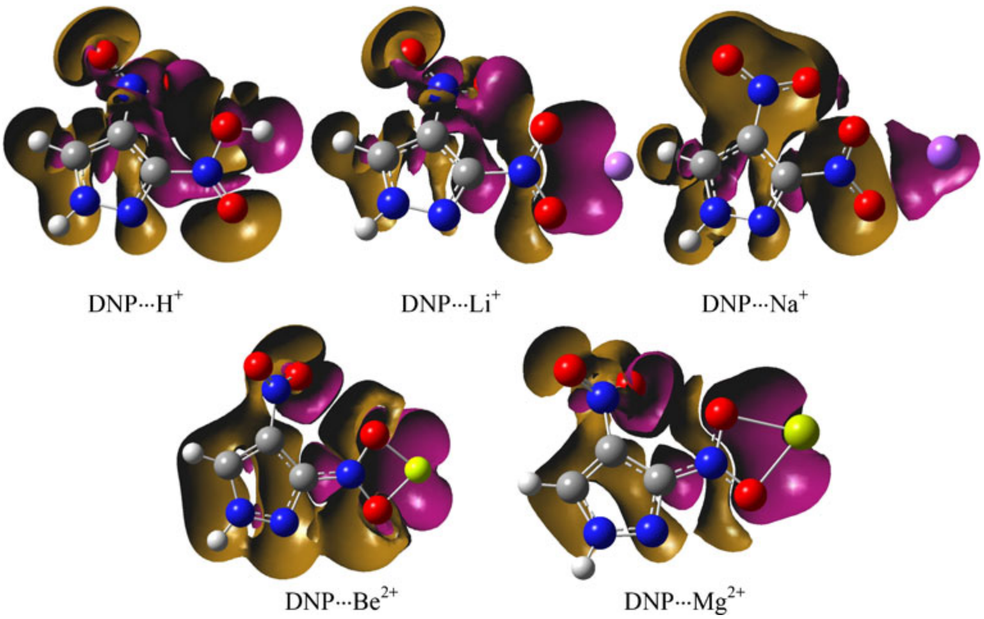
NBO gives the value of net charge transfer, which is evaluated to be from DNP to M by 46.50, 36.04, 230.71 and 145.73 me for DNP…Li^+, DNP…Na^+, DNP…Be^{2+} and DNP…Mg^{2+}, respectively. The charge transfer of DNP…Be^{2+} is the greatest, suggesting the molecule-cation interaction may be the strongest, as is consistent with the above. The correlation between the NBO charge transfer ($Q_{\text{NBO charge transfer}}$) and the increment of C3–N7 bond dissociation energy in comparison with the DNP monomer (Δ*BDE*) at B3LYP/6-311++G(2df,2p) level is given in Fig. 6 (see A), and the correlation coefficient is up to 0.9999.

$$\Delta BDE = 3.137Q_{\text{NBO charge transfer}} + 21.518 \tag{8}$$

To our interest, the delocalization interaction is also found between the N7–O11 bond and C3–N7 anti-bond orbitals. $E_{(N7-O11)\to(C3-N7)^*}{}^{(2)}$ is 16.10, 4.39, 3.93, 7.15 and 6.19 kJ mol^{-1} in DNP…H^+, DNP…Li^+, DNP…Na^+, DNP…Be^{2+} and DNP…Mg^{2+}, respectively. This result shows that the order of this delocalization interaction is DNP…Be^{2+} > DNP…Mg^{2+} > DNP…Li^+ > DNP…Na^+. The correlation between the delocalization energy ($E_{(N7-O11)\to(C3-N7)^*}{}^{(2)}$) and the increment of C3–N7 bond dissociation energy in comparison with the DNP monomer Δ*BDE* at B3LYP/6-311++G(2df,2p) level is shown in Fig. 7. The correlation coefficient is up to 0.9901.

$$\Delta BDE = 188.406\,E^{(2)}{}_{(N7-O11)\to(C3-N7)^*} - 638.713 \tag{9}$$

Fig. 8 Shifts of electron density as a result of formation of the complex between DNP and H^+, Li^+, Na^+, Be^{2+}, Mg^{2+}. Purple regions denote gain, and yellow regions represent loss



Recently, lots of investigations have indicated that the more negative charges the nitro groups carry, the more insensitive the explosives are [16, 17, 21]. From Table 5, the natural charge of nitro group O10–N7–O11 in the DNP monomer is −0.1671 e. The difference between the natural charge of nitro group in the complex and that in the DNP monomer is 0.1347, –0.1792, –0.1349, –0.6228 and −0.4065 e in DNP…H^+, DNP…Li^+, DNP…Na^+, DNP…Be^{2+} and DNP…Mg^{2+}, respectively. These results show that the natural charge of nitro group has reduced in complex in comparison with the DNP monomer except for DNP…H^+, indicating that much negative charge concentrates on the nitro group O10–N7–O11. Thus, the sensitivity is reduced upon the formation of the complex, as is in accordance with the above analysis. The order of the natural charge of nitro group is DNP…Be^{2+} > DNP…Mg^{2+} > DNP…Li^+ > DNP…Na^+, indicating that the order of the sensitivity follows DNP…Be^{2+} < DN…Mg^{2+} < DNP…Li^+ < DNP…Na^+, as is in agreement with the analyses of structure, energy, AIM and NBO. The relationship between the change of the nature charge of the nitro group (ΔQ_{NO2}) and the increment of C3–N7 bond dissociation energy in comparison with the DNP monomer (ΔBDE) at B3LYP/6-311++G(2df,2p) level is shown in Fig. 6 (see B). The correlation coefficient is up to 0.9997.

$$\Delta BDE = -1.278\,\Delta Q_{\text{Natural charge}} - 47.566 \qquad (10)$$

Analysis of the electron density shifts

It is known that changes in the electron density distribution in both donors and acceptors are the most important consequence of the formation of the non-bonded interaction [65, 66]. To display visually the nature of the molecule-cation interaction of DNP with H^+, Li^+, Na^+, Be^{2+} and Mg^{2+}, the shifts of electron density is calculated and illustrated in Fig. 8. Purple regions represent the accumulation of additional electron density; yellow regions indicate loss of density.

As is shown in Fig. 8, the M cation is filled with much purple area and the nitro group O10–N7–O11 is around yellow region, suggesting that the electron density of the nitro group has been lost toward cation and the molecule-cation interaction has formed between DNP and M. Moreover, the purple area around the M is the most significant in DNP…Be^{2+}, indicating that the molecule-cation interaction between Be^{2+} and DNP is the strongest. It is noted that, for DNP…H^+, the purple area concentrates on O–H bond, showing that the interaction between the oxygen atom of the nitro group and H^+ is partly covalent, as is in agreement with the analysis of geometry and energy.

It is interesting that much purple area is around the C3–N7 bond, showing that the electron density also shifts from the nitro group toward the C3–N7. It is well known that the more intensive an electron between two atoms, the more chances foroverlapping. As a result, the strength of the C3–N7 bond is improved. As can be seen in Fig. 8, the purple area around the C3–N7 bond of DNP…Be^{2+} is the most significant, suggesting that the strength of the C3–N7 bond in DNP…Be^{2+} is the most and the sensitivity is the lowest, as is in agreement with the bond length and bond dissociation energy of the C3–N7 bond.

Thus, we can draw a conclusion that the electron density shift from the nitro group O10–N7–O11 to the C3–N7 bond upon the formation of the molecule-cation interaction. The C3–N7 bond is enhanced and the sensitivity is reduced, as is in according with the analyses of structure, energy, AIM and NBO.

Conclusions

The explosive sensitivity upon the formation of molecule-cation interaction between the nitro group of DNP and H^+, Li^+, Na^+, Be^{2+} or Mg^{2+} has been investigated using the B3LYP and MP2(full) methods with the 6-311++G** and 6-311++G(2df,2p) basis sets. The bond dissociation energy of the C3–N7 trigger bond has also been discussed. The interaction between the oxygen atom of nitro group and H^+ in DNP…H^+ is partly covalent in nature. The molecule-cation interaction and bond dissociation energy of the C3–N7 trigger bond follow the order of DNP…Be^{2+} > DNP…Mg^{2+} > DNP…Li^+ > DNP…Na^+. The increment of the *BDE* of the trigger bond correlates well with the molecule-cation interaction energy, natural charge of the nitro group, electron density $\rho_{\mathrm{BCP(C3-N7)}}$, delocalization energy $E^{(2)}$ and NBO charge transfer except for DNP…H^+. The analyses of AIM, NBO and electron density shifts have shown that the electron density of the nitro group shifts toward the C3–N7 trigger bond upon the formation of the molecule-cation interaction. Thus, the trigger bond is strengthened and the sensitivity of DNP is reduced.

References

1. Hu TP, Ren FD, Ren J (2009) J Mol Struct Theochem 909:13–18
2. Richard RM, Ball DW (2008) J Mol Struct Theochem 851:284–293
3. Qiu L, Gong XD, Zheng J, Xiao HM (2009) J Hazard Mater 166:931–938
4. Wang G, Gong X, Liu Y, Xiao H (2009) Spectrochim Acta Part A 74:569–574
5. Liu Y, Gong XD, Wang LJ, Wang GX, Xiao HM (2011) J Phys Chem A 115:1754–1762

6. Keshavarz MH, Pouretedal HR (2010) Propell Explos Pyrot 35:175–181
7. Pospíšil M, Vávra P, Concha MC, Murray JS, Politzer P (2010) J Mol Model 16:895–901
8. Zhao J, Xu DH, Cheng XL (2010) Struct Chem 21:1235–1240
9. Li JS (2010) J Hazard Mater 174:728–733
10. Li JS (2010) J Hazard Mater 180:768–772
11. Atalar T, Jungová M, Zeman S (2009) J Energy Mater 27:200–216
12. Cao CZ, Gao S (2007) J Phys Chem B 111:12399–12402
13. Zhao J, Cheng XL, He B, Yang XD (2006) Struct Chem 17:501–507
14. Song XS, Cheng XL, Yang XD, He B (2006) Propell Explos Pyrot 31:306–310
15. Zhang CY (2006) J Phys Chem A 110:14029–14035
16. Zhang CY, Shu YJ, Huang YG, Zhao XD, Dong HS (2005) J Phys Chem B 109:8978–8982
17. Rice BM, Hare JJ (2002) J Phys Chem A 106:1770–1783
18. Vaullerin M, Espagnacq A, Morin-Allory L (1998) Propell Explos Pyrot 23:237–239
19. McNesby KL, Coffey CS (1997) J Phys Chem B 101:3097–3104
20. Sharma J, Beard BC, Chaykovsky M (1991) J Phys Chem 95:1209–1213
21. Tan BS, Long XP, Peng RF, Li HB, Jin B, Chu SJ, Dong HS (2010) J Hazard Mater 183:908–912
22. Ren WZ, Wang ZS (2004) Explosive theory and practice. China North Chemical Industries Corp Press, Nanjing, China
23. Peter P, Jane SM, Pat L (2009) Int J Quant Chem 109:3–7
24. Peter P, Jane SM, Pat L (2009) Int J Quant Chem 109:534–539
25. Xu WZ, Ren FD, Ren J, Liu SN, Yue Y, Wang WL, Chen SS (2010) J Mol Model 16:615–627
26. Cao DL, Ren FD, Feng YQ, Liu SN, Chen SS (2010) J Mol Model 16:589–598
27. Cao DL, Ren FD, Liu SN, Chen SS (2009) J Mol Struct Theochem 913:221–227
28. Rodrguez-Otero J, Cabaleiro-Lago EM, Peña-Gallego A (2008) Chem Phys Lett 452:49–53
29. Mohajeri A, Karimi E (2006) J Mol Struct Theochem 774:71–76
30. Cheng JG, Zhu WL, Tang Y, Li Z, Chen KX, Jiang HL (2006) Chem Phys Lett 422:455–460
31. Molina JM, Dobado JA, Melcho S (2002) J Mol Struct Theochem 589–590:337–347
32. Owens FJ (1996) J Mol Struct Theochem 370:11–16
33. Politzer P, Murray JS, Lane P, Sjoberg P (1991) Chem Phys Lett 181:78–82
34. Li YX, Du S, Wang JL (2011) Acta Crystallogr E67:o1369
35. Price dr D, Morris dr J (2009) Synthesis of New Energetic Melt-Pour Candidates, BAE SYSTEMS, HSAAP
36. Xiao HM (1993) Molecular orbital theory of nitro-compound. National Defence Industry Press, Peking, China
37. Richard RM, Ball DW (2006) J Mol Struct Theochem 776:89–96
38. Richard RM, Ball DW (2007) J Mol Struct Theochem 806:113–120
39. Macias AT, Norton JE, Evanseck JD (2003) J Am Chem Soc 125:2351–2360
40. Suwattanamala A, Magalhaes AL, Gomes JANF (2005) Chem Phys 310:109–122
41. Ruan C, Yang Z, Hallowita N, Rodgers MT (2005) J Phys Chem A 109:11539–11550
42. Amunugama R, Rodgers MT (2002) J Phys Chem A 106:9718–9728
43. Tanaka N, Tamezane T, Nishikiori H, Fujii T (2003) J Mol Struct Theochem 631:21–28
44. Novoa JJ, Mota F (2000) Chem Phys Lett 318:345–354
45. Frisch MJ, Trucks GA, Schlegel HB, Scuseria GE, Robb MA, Cheseman JR, Montgomery JA, Vreeven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochtersky JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi L, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2003) Gaussian 03, Revision B.03. Gaussian Inc, Pittsburgh, PA
46. Reed AE, Curtis LA, Weinhold F (1988) Chem Rev 88:899–926
47. Scheiner S, Kar T (2002) J Phys Chem A 106:1784–1789
48. Bader RFW (1990) Atoms in molecules-a quantum theory. Oxford University Press, Oxford, UK
49. König FWB, Bader RFW, Tang TH (1982) J Comput Chem 3:317–328
50. Duijineveldt FB, Duijineveldt-van de Rijdt JCM, Lenthe JHV (1994) Chem Rev 94:1873–1885
51. Boys SF, Bernardi F (1970) Mol Phys 19:553–566
52. Brakaspathy R, Singh S (1987) Proc Indian Acad Sci (Chem Sci) 99:253–260
53. Zhang J, Ha TK, Knochenmuss R, Zenobi R (2002) J Phys Chem A 106:6610–6617
54. Huber KP, Herzberg G (1979) Molecular spectra and molecular structure. IV. Constants of Diatomic Molecules. Reinhold, New York
55. Wu YJ, Ren FD, Li BC (2009) J Mol Struct Theochem 909:79–85
56. Luo YR (2003) Handbook of bond dissociation energies in organic compounds. CRC Press
57. Budyka MF, Zyubina TS, Zarkadis AK (2002) J Mol Struct Theochem 594:113–125
58. Jursic BS (1996) J Mol Struct Theochem 366:103–108
59. Brink T, Haeberlin M, Jonsson M (1997) J Am Chem Soc 119:4239–4244
60. Jursic BS, Martin RM (1996) Int J Quantum Chem 59:495–501
61. Barckholtz C, Barckholtz TA, Hadad CM (1999) J Am Chem Soc 121:491–500
62. Fuchs M, Niquet YM, Gonze X, Burke K (2005) J Chem Phys 122:094116–094128
63. Matta CF, Castillo N, Boyd RJ (2006) J Chem Phys 125:204103–204115
64. Mata I, Molins E, AlkortaI, Espinosa E (2009) J Chem Phys 130:044104–044119
65. Ebrahimi A, Roohi H, Habibi M, Hasannejad M (2006) Chem Phys 327:368–372
66. Scheiner S (2011) J Chem Phys 134:094315–094323

J Mol Model (2012) 18:2117–2133
DOI 10.1007/s00894-011-1228-8

ORIGINAL PAPER

Structural insights into human GPCR protein OA1: a computational perspective

**Anirban Ghosh · Uddhavesh Sonavane ·
Sai Krishna Andhirka · Gopala Krishna Aradhyam ·
Rajendra Joshi**

Received: 17 June 2011 / Accepted: 18 August 2011 / Published online: 22 September 2011

Abstract Human ocular albinism type 1 protein (OA1)—a member of the G-protein coupled receptor (GPCR) superfamily—is an integral membrane glycoprotein expressed exclusively by intracellular organelles known as melanocytes, and is responsible for the proper biogenesis of melanosomes. Mutations in the *Oa1* gene are responsible for the disease ocular albinism. Despite its clinical importance, there is a lack of in-depth understanding of its structure and mechanism of activation due to the absence of a crystal structure. In the present study, homology modeling was applied to predicting OA1 structure following thorough sequence analysis and secondary structure predictions. The predicted model had the signature residues and motifs expected of GPCRs, and was used for carrying out molecular docking studies with an endogenous ligand, L-DOPA and an antagonist, dopamine; the results agreed quite well with the available experimental data. Finally, three sets of explicit molecular dynamics simulations were carried out in lipid bilayer, the results of which not only confirmed the stability of the predicted model, but also helped witness some differences in structural features such as rotamer toggle switch, helical tilts and hydrogen bonding pattern that helped distinguish between the agonist- and antagonist-bound receptor forms. In place of the typical “D/ERY”-motif-mediated “ionic lock”, a hydrogen bond mediated by the “DAY” motif was observed that could be used to distinguish the agonist and antagonist bound forms of OA1. In the absence of a crystal structure, this study helped to shed some light on the structural features of OA1, and its behavior in the presence of an agonist and an antagonist, which might be helpful in the future drug discovery process for ocular albinism.

Keywords G-protein coupled receptor · Ocular albinism · L-DOPA · Dopamine · Homology modeling · Molecular docking · Molecular dynamics simulation

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1228-8) contains supplementary material, which is available to authorized users.

A. Ghosh · U. Sonavane · R. Joshi (✉)
Bioinformatics Group,
Centre for Development of Advanced Computing (C-DAC),
Pune University Campus,
Pune 411007, India
e-mail: rajendra@cdac.in

S. K. Andhirka · G. K. Aradhyam (✉)
Department of Biotechnology,
Indian Institute of Technology Madras,
Chennai 600036, India
e-mail: agk@iitm.ac.in

Introduction

G protein-coupled receptors (GPCRs) comprise the largest superfamily of membrane proteins in the human body, being coded by 3–4% of the entire genome [1]. These extensive and diverse membrane receptors are involved in several kinds of signal transduction pathways using their cognate G-proteins as mediators for transmitting signals. These receptors are responsible for the metabolic, physiological and neurological maintenance of most biological systems in all eukaryotes [2]. Various stimuli, such as light, ions, odorants and specific ligands—hormones, amino acids, nucleotides etc.—bind to the extracellular regions of these receptors, bringing about conformational changes leading to the activation of downstream G-proteins. The physiological importance of GPCRs is underlined by the fact that several diseases, for example retinitis pigmentosa, migraine, asthma, hypertension, congestive heart-failure, Parkinson’s, schizophrenia and glaucoma, are caused by their malfunctioning [3]. As a consequence, these receptors have become important drug targets (up to 50–60% of the drugs on the market target these receptors), largely for molecules that either enhance or inhibit signal amplitude by modulating their structure [2, 4].

Until recently, structure-based drugs designed to target GPCRs were generated mostly using an empirical structure of the receptor and its extracellular loop regions. The crystal structure of rhodopsin [5, 6] offered a real template and changed the mode of analysis of possible structures for new receptors. Almost a decade later, more structures have been published, including that of human β2-adrenergic receptor (β2-AR) [7] and its subsequent active-state structures [8, 9], human A2A-adenosine receptor [10], squid rhodopsin [11], human CXCR4 receptor [12] and human metarhodopsin II [13]. The advantage with these receptors is that people now have receptor structures with agonist ligands bound to them, thereby providing the active state conformations of the receptors. As the natural abundance of most of these receptors is very low, and large scale expression and purification of recombinant proteins has proved to be extremely difficult, structural studies on these membrane receptors remain an arduous task. In the absence of a structure for the GPCR of interest, structure-based drug discovery and other functional studies related to their modes of activation rely on reasonable molecular models of GPCRs generated through various computational approaches, such as homology modeling based on the present set of crystal structures and simulation studies [14–20]. These new structures have become useful for GPCR structure–function studies and the application of molecular dynamics (MD) simulations through the use of knowledge-based constraints to refine the generated homology models [21–24] will help provide novel insights into the mode of action of these newer receptors. Moreover, the activation mechanism—the process by which these receptors switch between their active and inactive conformations—of most GPCRs (e.g., rhodopsin) spans a millisecond time-scale, making them very difficult to study [25, 26].

Ocular albinism (OA) is a genetic disorder caused by a hypopigmentation of ocular tissues [27, 28]. OA can be linked to a gene on the X chromosome, *Oa1*, that produces a 404-amino-acid protein: a GPCR referred to as ocular albinism type I protein or OA1 (GenBank: GP143) [29, 30]. Evidence to this effect has come from work by Schiaffino et al. [31, 32] showing that OA1 binds to several G_α- as well as $G_{\beta\gamma}$-subunits, and Innamorati et al. [33], who illustrated its β-arrestin association even in the absence of any ligand. OA1 is a unique member of the GPCR superfamily, being a fully intracellular protein localized primarily to the endolysosomal compartment and melanosomes rather than to the cell surface [34, 35]. A more recent study by Lopez et al. [36] has shown that L-DOPA acts as an endogenous ligand for OA1, while the receptor shows a complicated trafficking behavior in the presence of agonist L-DOPA and antagonist dopamine.

The present communication reports a thorough structural investigation of OA1 from a computational perspective, with particular emphasis on the early events of its possible activation mechanism. The seven transmembrane (7TM) helices of OA1—a feature of all GPCRs—and their location with respect to its primary sequence, were established through analysis of primary and secondary structure. This led to the final construction of a model for OA1, wherein the stereochemical qualities were validated fully through extensive MD simulations. The mode of binding of OA1 with its endogenous ligand L-DOPA and antagonist dopamine was also investigated. The key structural differences of the receptor in the absence and presence of an agonist and an antagonist, as elucidated through MD simulation studies carried out for all the three systems—apo-OA1, OA1-L-DOPA and OA1-dopamine—in the presence of a lipid bilayer and explicit water, are illustrated. The MD simulations not only refined the predicted model of OA1 but also helped witness some of the signature “switches” that are seen during the normal activation of GPCRs like rhodopsin and β2-AR and help to distinguish between the different forms of the receptor in the presence of agonist/antagonist.

Materials and methods

Homology modeling

In the absence of a crystal structure, the 3D structure of OA1 was predicted by homology modeling using extensive primary and secondary structural knowledge. First, sequence alignments were performed to identify conserved residues and motifs that might have structural and functional implications. The sequence of OA1, consisting of 404 residues, was retrieved from the SWISSPROT database [37] and used as a query to search the PFam database [38] for homologous sequences. It was also used as a query for the PSI-BLAST search [39], in order to isolate distant homologs, since GPCR proteins share a very low percentage of sequence identity. From these two searches, a total of 45 related sequences were identified and subjected to multiple sequence alignment (MSA). MSA was performed using ClustalX (version 2.0.10) [40] to identify conserved regions. Secondary structure information was also applied from predictions derived using the TMHMM [41] and PredictProtein [42] transmembrane prediction servers to identify the exact location of 7TM helices, intracellular loops (ICLs), and extracellular loops (ECLs).

The crystal structure of human β2-AR (PDB ID: 2RH1) [43]—an amine GPCR—was used as a template for the construction of the OA1 model since it was found to be a more suitable template than bovine rhodopsin (PDB ID: 1U19) [44] in terms of sequence identity. Using knowledge of primary and secondary structure (as mentioned above), the final alignment between OA1 and β2-AR was edited manually using BioEdit (version 7.0.9.0) [45] to retain high equivalence of conserved regions. The 3D model was finally generated using the software MODELLER (version 9.6) [46] with 2RH1, using the 2.4 Å resolved crystal structure of β2-AR retrieved from the RCSB database as a template. As a result, 50 models were generated for OA1, and rated according to the GA341 and DOPE scoring functions available with MODELLER. The loop regions were refined using the in-built "loopmodel" class available with MODELLER. The stereochemical properties of the final selected model were validated using the PROCHECK [47] program and VERIFY3D server [48].

Molecular docking

Docking of OA1 was performed using the QM-polarized ligand docking (QPLD) workflow of the Schrödinger software suite (Maestro version 9.0.109) [49] with the endogenous ligand L-DOPA, as well as the possible antagonist dopamine, for a more accurate QM/MM docking [50]. The ligands L-DOPA and dopamine were built using the Maestro suite in the Schrödinger suite with only the polar aromatic hydrogen atoms, which is in compliance with the gromos96 force field. Internally, QPLD uses QSite and Glide to carry out the docking algorithm. Initially, QPLD performs a conventional docking to produce the prescribed number of initial poses for each ligand. With these initial docked poses, QM/MM single energy calculations are run on each of the poses with only the ligand as the QM region, producing new sets of atomic charges on the ligand by ESP (electrostatic potential) fitting. Finally, a re-docking is performed with these new atomic charges and the best scoring pose is selected. In the present study, QPLD (version 2.0) was used, which employs Glide (version 5.5) and QSite (version 5.5). The grid was generated to define the binding cavity as the entire transmembrane region, and a flexible ligand model was used throughout the docking protocol. The initial number of poses for the ligand was set to ten and, for the QM part, B3LYP function was used in density functional theory (DFT) calculations. All calculations were performed in extra precision (XP) mode.

Molecular dynamics simulations

Three systems were built for explicit MD simulation in lipid bilayer: (1) apo-OA1, (2) OA1-L-DOPA and (3) OA1-dopamine. MD simulations were performed using the GROMACS version 4.0.7 package [51] with gromos ffG43a2 force field, extended to improve the lipid components of the force field. The topology and other force field parameter for both ligands (L-DOPA and dopamine) were obtained from the PRODRG server [52] and were examined carefully for any discrepancies with previous results. The lipid bilayer consisted of a pre-equilibrated layer of 288 molecules of 1-palmitoyl-2oleoyl-*sn*-glycerophosphocholine (POPC), generously gifted by X. Periole. All the Berger lipid parameters including those for POPC were obtained from P. Tieleman's site at http://moose.bio.ucalgary.ca/Downloads. The protein was inserted carefully into the lipid bilayer using the InflateGro program [53]. The entire lipid bilayer was inflated and then slowly compressed around the protein until an area per lipid value of 69 $Å^2$ was reached, which is just above the experimental value of 65 $Å^2$ for pure POPC. Each compression step was followed by a round of steepest descent energy minimization to relax the lipid molecules, keeping the protein restrained. The entire system was then solvated with a single-point charge (SPC) water model and neutralized with Cl^- counter-ions. Each of the three systems contained a total of 135,218; 135,237; and 135,232 atoms respectively.

All three systems were then energy minimized using the steepest descent algorithm present in the GROMACS package. A 100 ps position-restraining simulation was then carried out to restrain the protein by a 1,000 kJ mol^{-1} harmonic constraint to relieve the close contacts with POPC and water under NVT ensemble conditions, with a V-rescale (modified Berendsen) temperature coupler [54]. This was followed by another 1 ns equilibration run under NPT ensemble conditions, before a final production run of 15 ns. The three systems were then run at 310 K, i.e., above the phase transition temperature of pure POPC, to ensure that the lipids maintained their proper density, and 1 bar pressure under isothermal-isobaric ensemble (NPT) for 15 ns each. Nosé-Hoover (which is used widely for membrane NPT simulations) temperature and Parrinello-Rahman pressure couplers were used to maintain the temperature and pressure values with the protein, ligands, lipids and water (plus ions) molecules coupled separately with a coupling constant of τ_t=0.1 ps. Semi-isotropic pressure coupling was set with τ_p=2 ps, allowing the bilayer to deform in the *x*–*y* plane independently of the *z*-axis. Since interfacial systems like membrane–water systems have a tendency to move laterally, the motion of the bilayer center-of-mass (COM) and solvent COM were reset separately so that the overall COM for the system is unchanged as the phases may drift in opposite directions. A time-step of 2 fs was used throughout with periodic boundary conditions. LINCS constraint algorithm [55] was used to maintain the geometry of the molecules.

Long-range electrostatic interactions were calculated using the particle-mesh Ewald (PME) method. Van der Waal's interactions and Coulomb interactions were cut off at 12 Å with updates every five steps. All simulations were performed on PARAM Yuva, using 64 Intel Xeon 2.93 GHz Quad Core processors. The results were analyzed using the in-built analysis package of GROMACS, LIGPLOT [56], XMGRACE [57] and in-house-developed scripts. The trajectories were visualized using VMD [58] and all the images were rendered using PyMol [59]. The overall health of the simulated systems was also checked with respect to temperature, pressure and potential energy of the systems. The data shown in Supplementary Fig. S1 show that all three simulated systems were in thermodynamic equilibrium during the production simulation runs, confirming the convergence of individual trajectories. Various properties of the lipid bilayer, like area per lipid and bilayer thickness (Supplementary Fig. S2) were also checked to ensure that the lipid bilayer did not enter a gel phase during the simulations.

Results and discussion

Homology modeling

OA1 or GPR143 (GenBank) is a 404-residue-long protein, expressed exclusively on the intracellular organelles known as melanosomes. The sequence of OA1 was retrieved from the NCBI database (NCBI Reference Sequence: NP_000264). To retrieve sequences homologous to, or belonging to the same superfamily as, OA1, the FASTA sequence of OA1 was submitted to the PFam database. The search returned a total of 23 sequences, all belonging to the GPCR superfamily. Next, a PSI-BLAST search was performed against the *nr* (non-redundant) database of NCBI using OA1 as the query sequence. PSI-BLAST mainly helps retrieve distantly related homologous sequences. Finally, with a total of 45 sequences retrieved from PFam database and PSI-BLAST search, MSA was performed to identify the conserved residues and motifs of OA1 that are present throughout the GPCR superfamily, e.g., the "DAY" ("D/ERY" in other GPCRs) motif at the end of intracellular helix TM3, and residues like C116, C184, D78, W162 and P210, mutations in which lead to OA disease in humans. Some specific proline residues that are present in the middle of the TM helices, e.g., P96 on TM2, P210 on TM5 and P300 on TM7, that confer distinctive "kinks" were also identified. Apart from primary sequence alignments, secondary structure predictions were also performed to identify the exact locations of the TM helices and the loop regions. Predictions from the TMHMM and PHD webservers were used to reach a consensus about the exact stretch of TM helices. Table 1 shows the residue-wise distribution of TM helices, ICLs, ECLs and the conserved residues contained within them, along with some of the mutations that have been reported to be causative agents of OA [60].

Table 1 Residue-wise distribution of transmembrane (TM) helices, intracellular (IC) and extracellular (EC) loops along with their conserved residues

Topology	Residue range	Conserved residues and mutations
Extracellular tail	1 – 27	R5C mutation
TM1	28 – 53	G35D mutation
ICL1	54 – 72	
TM2	73 - 101	D78N, G84R mutations
ECL1	102 – 112	
TM3	113 – 146	"DAY" motif; C116R, G118E, Q124R, W133R, A138V mutations
ICL2	147 – 151	
TM4	152 – 175	W162; S152N, A173D mutations
ECL2	176 – 190	C184
TM5	191 – 224	P210
ICL3	225 – 243	G229V, T232K, E235K, I244K mutations
TM6	244 – 269	W257 in the middle of TM3; I261N mutation
ECL3	270 – 289	E271G
TM7	290 – 314	W292G, P300
Eighth helix	315 – 327	
Cytoplasmic tail	328 – 404	

GPCRs are known to exhibit structural conservation, particularly in the TM domains, in the form of 7TM helices. This is often considered a signature property of GPCRs. Therefore, in spite of very low sequence similarity, homology modeling and other computational prediction methods are often used to predict the structures of several GPCRs based on their TM conservation [14–20]. The same principle has been followed here and, despite having an overall identity of only 17.8% (and 30.7% similarity) with β2-AR, 2RH1—the human β2-AR crystal structure—was used as the template for modeling OA1. The final alignment used for model building is shown in Fig. 1; the extents of the TM helices are shown, and the conserved residues are marked in solid colors. The final alignment was edited manually in order to arrive at this alignment, which has the maximum number of conserved residues aligned together.

This alignment was then provided as input to MODELLER to produce homology models for OA1. All the models produced were energy minimized using conjugate gradient algorithms and short MD simulations, as part of the MODELLER protocol in order to refine the side-chain orientations. The loop regions were refined using the in-

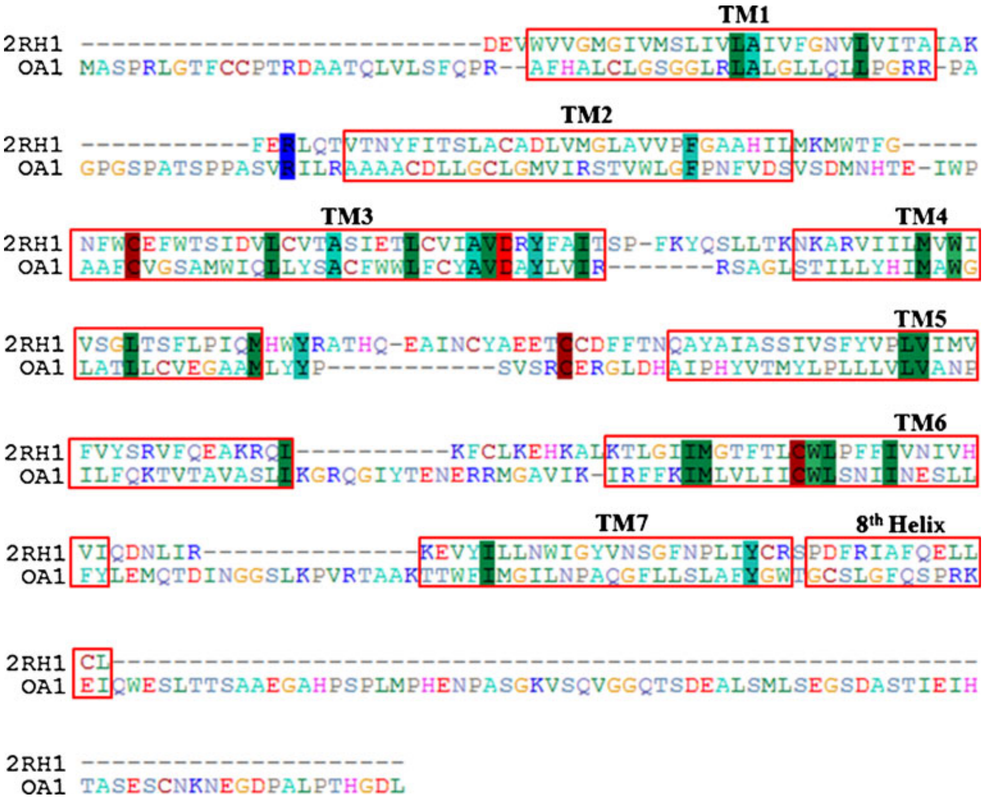

TM1
2RH1 --------------------------DEVWVVGMGIVMSLIVLAIVFGNVLVITAIAK
OA1 MASPRLGTFCCPTRDAATQLVLSFQPR--AFHALCLGSGGLRLALGLLQLLPGRR-PA
TM2
2RH1 -----------FERLQTVTNYFITSLACADLVMGLAVVPFGAAHIILMKMWTFG----
OA1 GPGSPATSPPASVRILRAAAACDLLGCLGMVIRSTVWLGFPNFVDSVSDMNHTE-IWP
TM3
TM4
2RH1 NFWCEFWTSIDVLCVTASIETLCVIAVDRYFAITSP-FKYQSLLTKNKARVIILMVWI
OA1 AAFCVGSAMWIQLLYSACFWWLFCYAVDAYLVIR-------RSAGLSTILLYHIMAWG
TM5
2RH1 VSGLTSFLPIQMHWYRATHQ-EAINCYAEETCCDFFTNQAYAIASSIVSFYVPLVIMV
OA1 LATLLCVEGAAMLYYP-----------SVSRCERGLDHAIPHYVTMYLPLLLVLVANP
TM6
2RH1 FVYSRVFQEAKRQI----------KFCLKEHKALKTLGIIMGTFTLCWLPFFIVNIVH
OA1 ILFQKTVTAVASLIKGRQGIYTENERRMGAVIK-IRFFKIMLVLIICWLSNIINESLL
TM7
8th Helix
2RH1 VIQDNLIR--------------KEVYILLNWIGYVNSGFNPLIYCRSPDFRIAFQELL
OA1 FYLEMQTDINGGSLKPVRTAAKTTWFIMGILNPAQGFLLSLAFYGWTGCSLGFQSPRK
2RH1 CL--
OA1 EIQWESLTTSAAEGAHPSPLMPHENPASGKVSQVGGQTSDEALSMLSEGSDASTIEIH
2RH1 ---------------------
OA1 TASESCNKNEGDPALPTHGDL

Fig. 1 Target template alignment. The final alignment of the ocular albinism type 1 (OA1) protein sequence (NCBI Reference Sequence: NP_000264) with 2RH1, the PDB entry for human β2-AR, as supplied to MODELLER. This final alignment was obtained using the BioEdit program. *Red outlined boxes* Transmembrane helices (TM). Amino acids are colored according to their chemical properties; *solid colored bars* conserved residues

built "loopmodel" class of MODELLER. Out of the 50 models generated by MODELLER, the best model was chosen based on the GA341 and DOPE scores reported by MODELLER. The stereochemical properties of the selected structure were then analyzed further.

The Ramachandran plot (Fig. 2) obtained from PROCHECK analysis confirms the stereochemical stability of the generated model, with 98.8% of the residues falling within the allowed regions (91.6% in the "most favored regions" and 7.2% in the "additional allowed region"). The only residue found in the disallowed region, D189, is located in region ECL2 of the model. The structure was also evaluated using the VERIFY3D program. The portions of the structure that were found to occur in the negative region of the VERIFY3D plot (not shown), were found to occur in the N-terminal (residues 1–28) and C-terminal (residues 325–404) regions. The residues found in the disallowed region of the Ramachandran plot and negative region of the VERIFY3D plot, correlated mainly to the loop regions of the structure, which showed very low sequence identity with the template sequence. Moreover, the occurrence of conserved domains and residues in the predicted model seem to tally well with other GPCR crystal structures, e.g., the presence of the "DAY" motif at the end of TM3, D78 on TM2, W162 on TM4, P210 on TM5, sequential C256 and W257 in the middle of TM6, C116 on TM3 and C184 on ECL3. Another signature feature of GPCRs is the presence of a Trp residue in the middle of TM6, which acts as a "toggle switch" upon activation [61–63]. In the predicted model, W257 was also found to occur in the middle of TM6 and acts as the toggle switch. N299 and P300—part of the "NPxxY" conserved motif—were found on TM7. Most of the Pro residues were found to occur in the loop regions, with a few exceptions, e.g., P96 on TM2, P201and P210 on TM5 and P300 on TM7, which confers characteristic kinks on the TM helices. The 7TM helices, ICLs and ECLs are shown schematically in Fig. 3. The figure also highlights the conserved GPCR residues, missense mutations that cause OA and missense mutations that abolish binding of the receptor to G proteins, as well as the signature features discussed above.

Molecular docking

OA1 is known to have a single saturable binding pocket for both L-DOPA and dopamine [36]. The structures of the ligands are shown in Supplementary Fig. S3. The docking studies performed here confirmed this, revealing that both L-DOPA and dopamine bind to OA1 in roughly the same pocket surrounded by TM3, TM5, TM6 and ECL2 extending from TM4. From among the ten poses for each of the ligands generated by the QPLD workflow, the final selection was based on the GScore provided by Glide. The GScore takes into account multiple factors, such as hydrogen bonds, hydrophobic, Van der Waal's, Coulomb and polar interactions in the binding pocket, as well as the

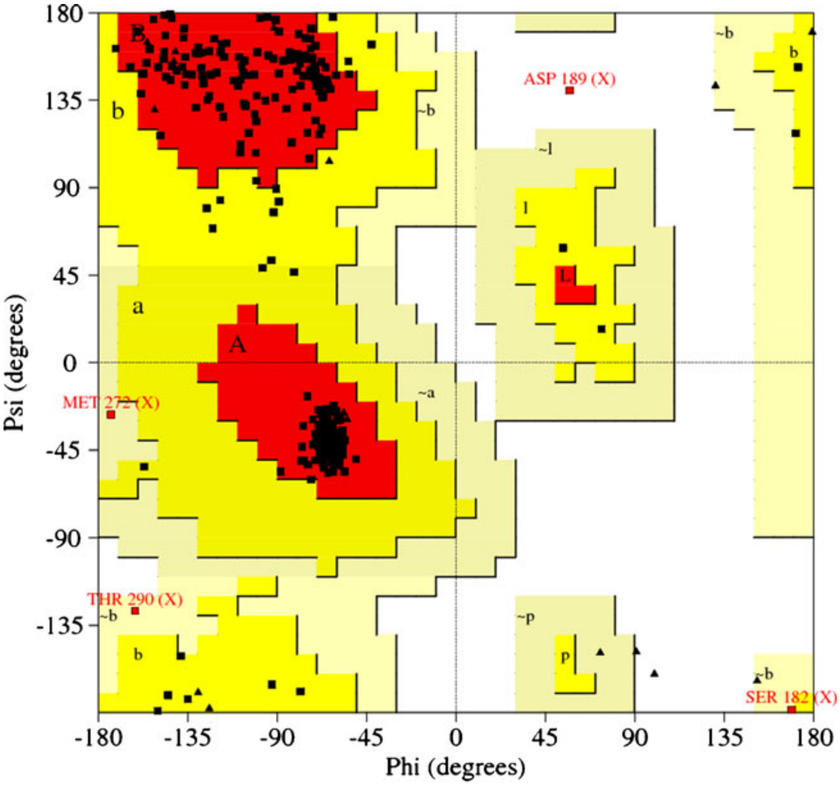


Fig. 2 Stereochemical property analysis of the predicted model of OA1. Ramachandran Plot obtained from PROCHECK analysis of the selected model from MODELLER. The plot confirms the stereochemical quality of the model, with 98.8% of the residues falling within the “allowed regions”. The only residue found in the disallowed region, D189, is located in the loop region of the model (ECL2)

Plot statistics

Residues in most favoured regions [A,B,L]	316	91.6%
Residues in additional allowed regions [a,b,l,p]	25	7.2%
Residues in generously allowed regions [~a,~b,~l,~p]	3	0.9%
Residues in disallowed regions	1	0.3%
	----	------
Number of non-glycine and non-proline residues	345	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	33	
Number of proline residues	24	

Total number of residues	404	

penalty for buried polar groups and freezing rotatable bonds. The conformation of L-DOPA finally selected had a GScore of −5.99 kcal mol^{-1} while that of dopamine was −4.97 kcal mol^{-1}. The subsequent 15 ns MD simulations further refined the binding modes of the ligands to the receptor. Figure 4a and b show the binding pockets within OA1 of L-DOPA and dopamine, respectively, at the end of 15 ns of MD simulation, as seen in the top view from the extracellular side. The selected docked poses of both ligands were found to be similar conformationally to that of the recently determined crystal structure of β2-AR in the active state with a bound ligand [8, 9]. Both ligands were found to be stabilized in their binding pockets through a network of hydrogen bonds, as discussed in detail below. Some residues engaged in hydrogen bonding with the docked ligands have been found to be associated with OA upon mutation, e.g., Q124, I261 and W292 [30, 34]. Indeed this provides further support for the selection of the particular docked poses of the ligands in binding pocket of OA1. Even after 15 ns of MD simulation, both ligands were found to remain in that binding pocket with ECL2 closing as a lid over them, as is also observed in the crystal structures of rhodopsin, β2-AR and other GPCRs [5–10, 61, 62].

Molecular dynamics simulations

A set of three simulations, each performed for a time-scale of 15 ns, helped to shed some light on the conformational changes of OA1 and its behavior in the presence of an agonist (L-DOPA) and a possible antagonist (dopamine). All three simulations, apo-OA1, OA1 with L-DOPA and OA1 with dopamine, were carried out for 15 ns each, after embedding the protein in a POPC bilayer and solvating the entire system with explicit water and counter-ions (Cl^{-}). However, comparisons between the three systems were

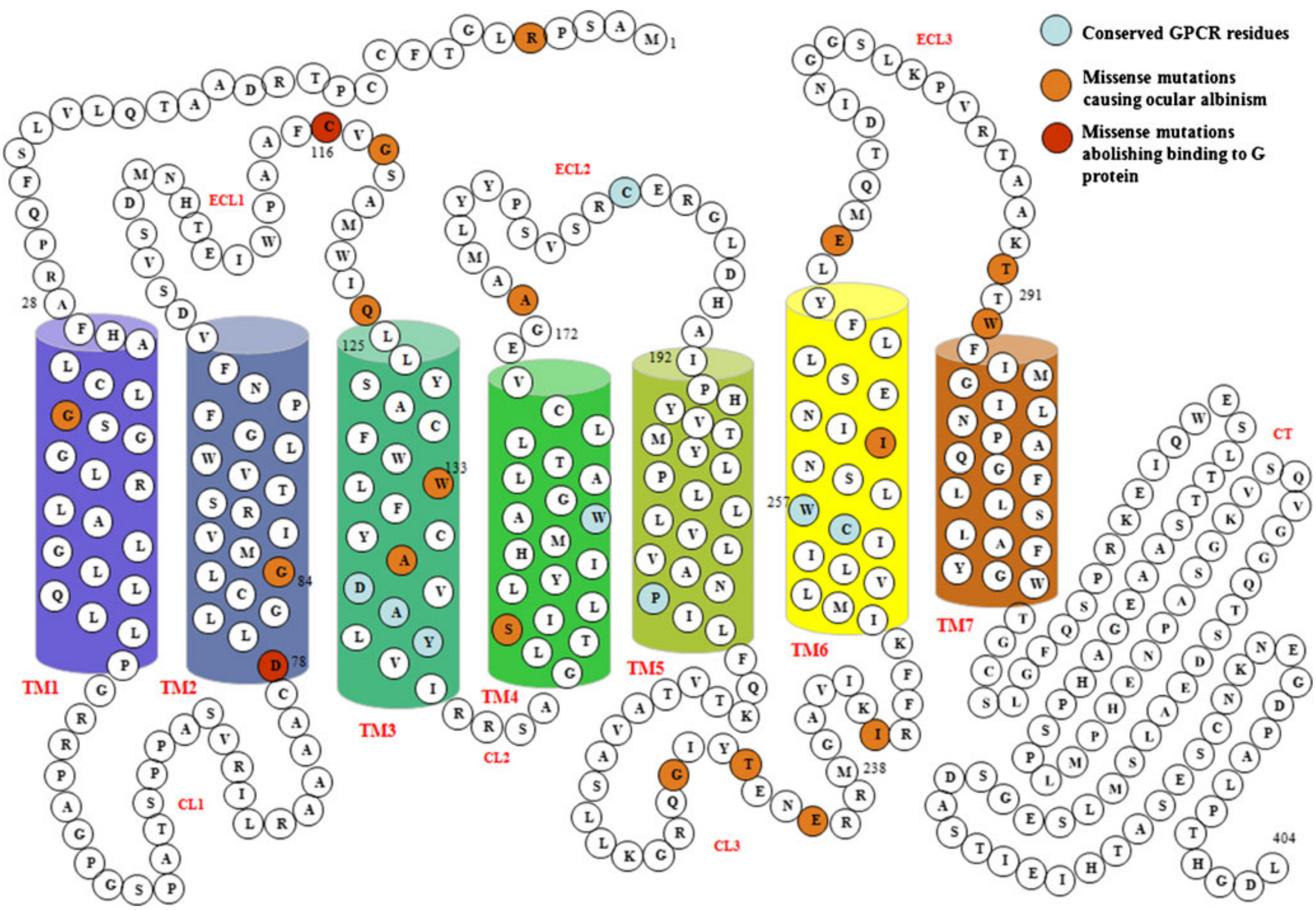


Fig. 3 Schematic representation of OA1 structure showing the 7TM helices, intracellular loops (ICLs) and extracellular loops (ECLs). The conserved GPCR residues, missense mutations that cause ocular albinism (OA) and missense mutations that abolish binding of the receptor to G proteins are also shown, as well as the "DAY" motif at the end of TM3 and "rotamer toggle switch" (W257 in the middle of TM6)

performed from 5 ns onwards, considering the first 5 ns as the equilibration phase required to allow the systems to become stabilized.

Apo-OA1 simulations

Upon comparing the initial structure of apo-OA1 obtained from homology modeling to the structure obtained after 15 ns of simulation, some differences were observed, mainly in the orientation of the loop regions. Plots of root mean square deviation (RMSD; Fig. 5) clearly show that the 7TM helices maintained their arrangements as compared to the loop regions. The low RMSD value of about 3 Å for the 7TM helices shown in Fig. 5a, which was maintained from 5 ns onwards, clearly indicates that the 7TM helices of OA1 maintained their structure and quaternary packing during the simulation. The RMSD values shown in Fig. 5a are quite stable, which in turn validates the quality of the generated model. The low RMSD values are indications of the fact that the protein maintains its 3D structure and arrangement in the POPC bilayer during the simulation. Hence, the MD simulation thus verifies the structural properties of the predicted model and further refines it [21, 24]. However, Fig. 5b shows a much higher deviation to about 10 Å, accounting for the overall movement of the entire protein. The higher value of RMSD is due mainly to the rearrangement of the long cytoplasmic tail (consisting of about 79 residues) and other loop regions, whose fluctuation is clearly shown in the root mean square fluctuation (RMSF) plot in Fig. 5c, which takes into account fluctuations of the residues between 5 ns and 15 ns.

A closer look at the RMSD values (Fig. 6a) of the individual TM helices (including the eighth helix, which remains parallel to the lipid bilayer and is a characteristic of some GPCRs) shows that, in the apo-OA1 system, TM5 shows a steady increase in RMSD value to about 2.5 Å. But TM3 and TM6, which are involved in the activation process of most GPCRs, including rhodopsin and β2-AR, do not show much deviation, with a steady RMSD value of about 2 Å and 1.5 Å, respectively. The radius of gyration plot (Fig. 7a) shows a decreasing trend in its values in the production phase of the last 10 ns, giving a clear indication that the protein attains a more compact

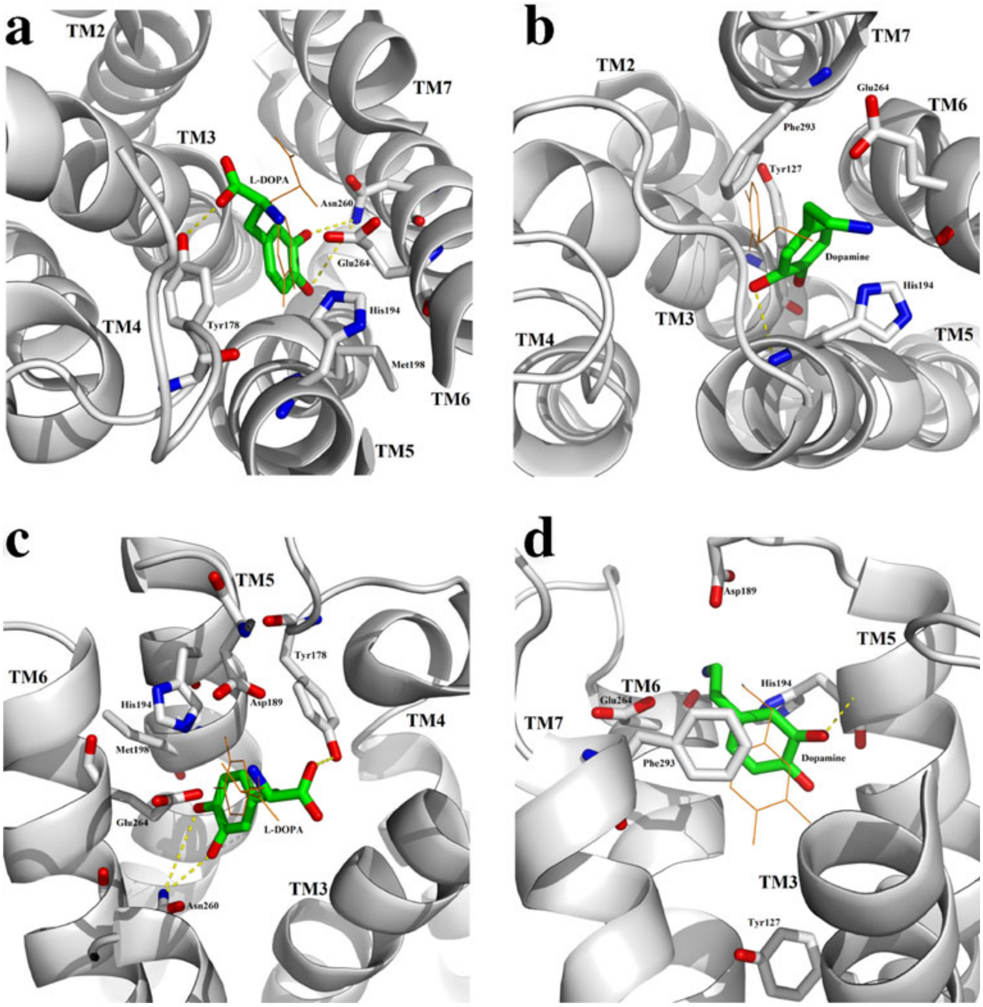


Fig. 4 Receptor-ligand interactions. Top (**a**, **b**) and side (**c**, **d**) views of the binding pockets of the two ligands L-DOPA (**a**, **c**) and dopamine (**b**, **d**) within the receptor. L-DOPA and dopamine bind to OA1 in roughly the same pocket surrounded by TM2, TM3, TM5, TM6 and ECL2. The initial (0 ns) docked pose of the ligands are shown in *orange* "line" representation, while "stick" representation in *green* (carbon atoms), *red* (oxygen atoms) and *blue* (nitrogen atoms) represent the average structures of the ligands forming hydrogen bonds with OA1. L-DOPA is seen to form hydrogen bonds with Y178 (on ECL2) and N260 (on TM6), and dopamine forms a hydrogen bond with H194 (on TM5). OA1 residues with which the ligands form stable hydrogen bonds (Y127, M198, E264 and F293) are also shown in "stick" representation. The hydrogen atoms are not shown in the figure. The figure was rendered using PyMOL

form as the long cytoplasmic tail and other loop regions fall back into their native conformations, which is in accordance with the RMSF plot. The fall in solvent accessible surface area (SASA) values (Fig. 7b) for the last 10 ns also reflects the same phenomenon of the loops attaining more compact forms.

OA1-L-DOPA simulation

L-DOPA has been shown to be an agonist for OA1 that activates the receptor upon binding [36]. When the OA1-L-DOPA system was simulated, a reduced motion of the loop regions was observed, which can be seen from the RMSD value ranging below 10 Å for the entire structure shown in Fig. 5b, indicating a more compact structure overall. Thus L-DOPA seems to stabilize the loop regions, ECL2 (residue 176–190) in particular, which will also be shown later through analysis of hydrogen bonding patterns. The RMSF plot shown in Fig. 5c also indicates this by clearly showing reduced fluctuation of ECL2 in the presence of L-DOPA as compared to the other two systems. However, there is a small increase in the RMSD value of the 7TM regions, which indicates movement of the 7TM helices relative to the ligand molecule during the simulation. A closer look at the RMSD values of the individual helices (Fig. 6b) shows that the RMSD values of TM5 and TM6 increase gradually with time. TM5 shows a particularly marked increase in RMSD value to about 3 Å. In the presence of L-DOPA, the intracellular end of TM5 from residue 213 to 224 shows a marked deviation from the normal of the *z*-axis, which is not observed in the other two systems. There is also an interesting hydrogen bond formation between TM5 and TM3 in this region, which is discussed below. The RMSF plot reveals a comparatively higher fluctuation of TM2 (residue 73–101) in the presence of L-DOPA than in the other two systems. The movements of TM2, TM5 and TM6, which also surround the ligand, clearly discriminate the conformations of the receptor in the presence of the agonist and antagonist, which is also observed during the activation of rhodopsin, β2-AR and other GPCRs [63–66]. From the RMSF plot (Fig. 5c), it can be seen that

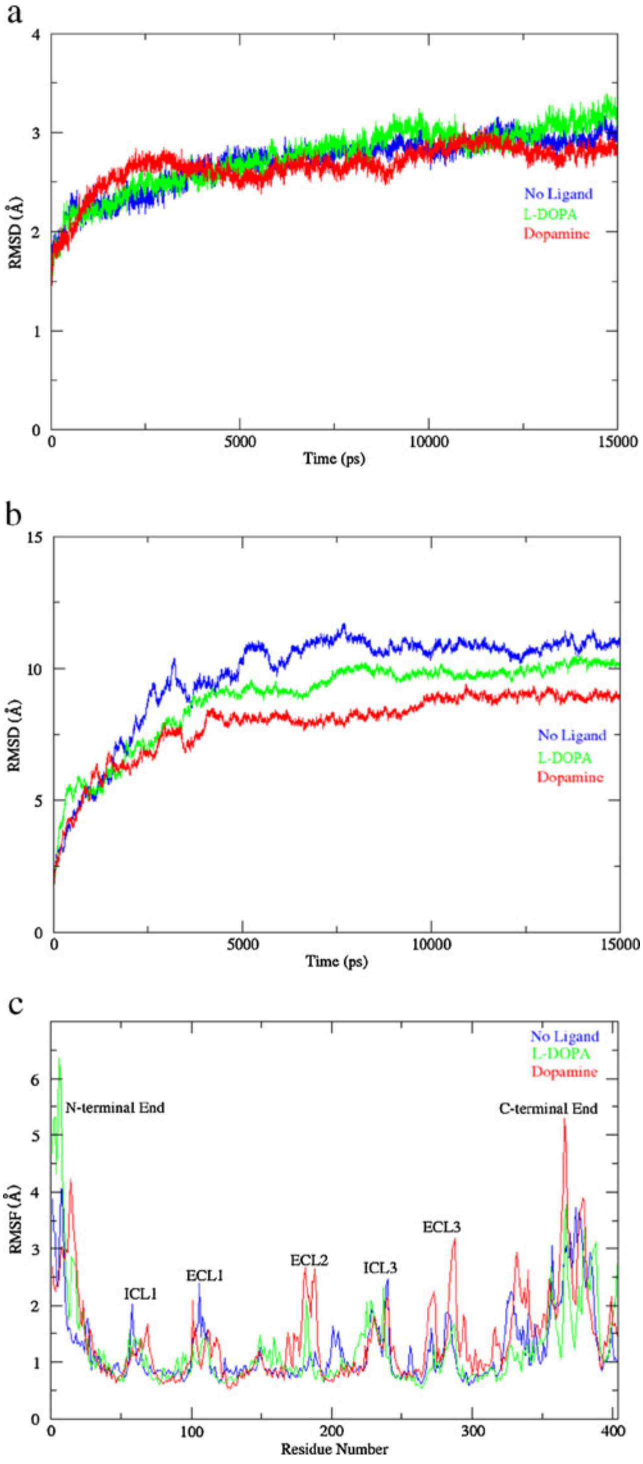


Fig. 5 Root mean square deviation (RMSD) and root mean square fluctuation (RMSF) plots. RMSD plots of the 7 TM helices (**a**) and the entire protein (**b**) are shown for all three simulated systems. (**c**) RMSF plot of the three systems. The colors indicate apo-OA1 (*blue*), OA1-L-DOPA (*green*) and OA1-dopamine (*red*) systems. The plots were prepared using XMGRACE

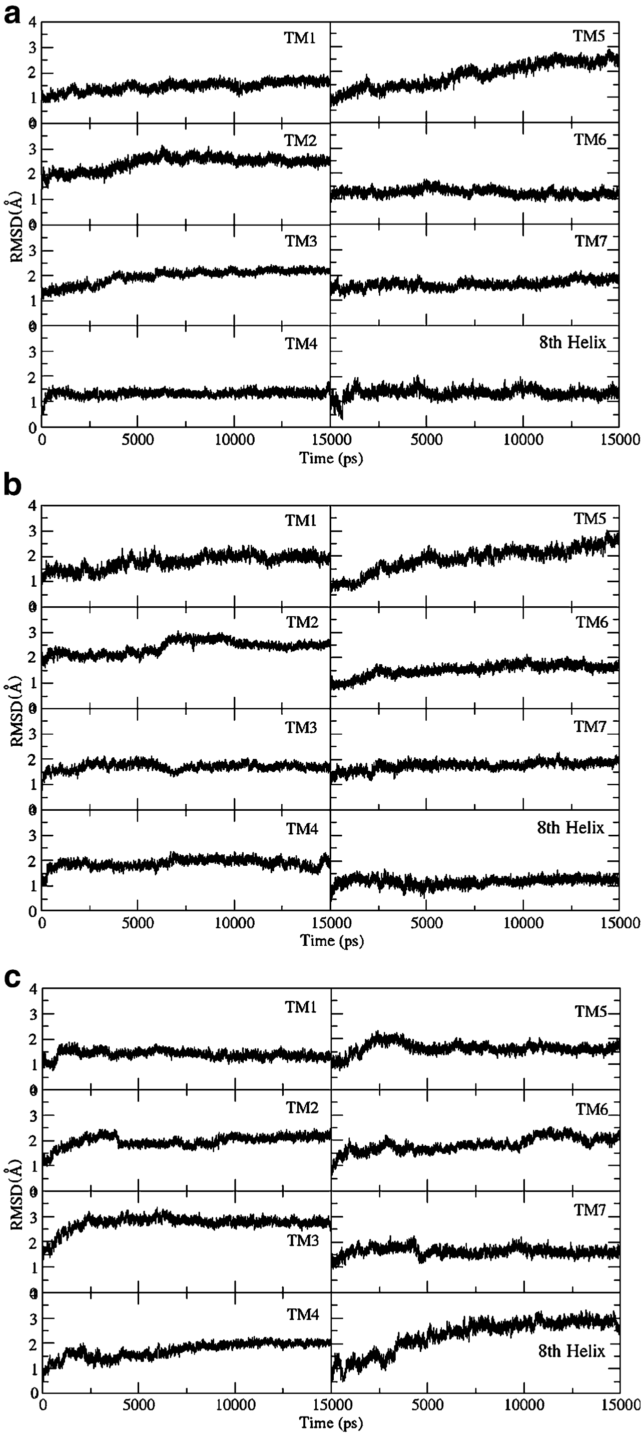


Fig. 6 RMSD plots of individual TM helices. **a**, **b** and **c** show the RMSD of the individual TM helices of apo-OA1, OA1-L-DOPA and OA1-dopamine systems, respectively. The plots also include the RMSD of the eighth helix, which remains parallel to the lipid bilayer. The plots were prepared using XMGRACE

ECL1 (residues 102–112) and ICL3 (residues 224–244) exhibit increased fluctuations. The fall in the radius of gyration values (Fig. 7a) reveals that the structure of OA1 becomes more compact in the presence of L-DOPA, whereas there is an increase in the SASA value (Fig. 7b) from 20,000 Å^2 to 21,500 Å^2 as compared to that of the apo-OA1 system from 5 ns onwards. This can be

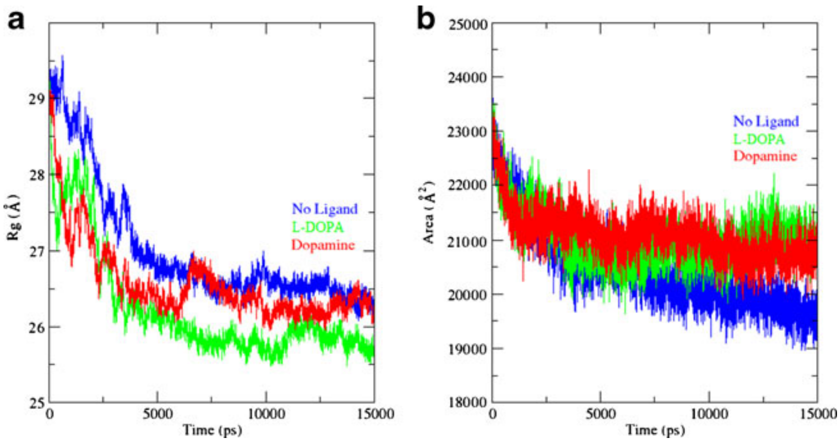


Fig. 7 Radius of gyration and solvent accessible surface area (SASA) plots. Plots show the change in **a** radius of gyration and **b** SASA for all three systems over time. The sharp fall in the SASA plot is due mainly to rearrangements of the loop regions and the cytoplasmic tail. The plots were prepared using XMGRACE

explained by the observation that the TM helices in the intracellular side, particularly TM5, open up, which is reflected in the SASA values, whereas the radius of gyration falls because the cytoplasmic tail and other loop regions become more compact. Visual inspection of the trajectory of the OA1-L-DOPA system also reveals lateral movements of TM1 and TM2 with respect to the *z*-axis, leading to increased accessibility to the binding pocket, whereas an opening in the intracellular end due to the movements of TM5 and TM6 might facilitate G-protein binding. This phenomenon is also reflected in Fig. 8b, which clearly shows that the tilt angles of TM1 and TM2 (with respect to the *z*-axis) for OA1-L-DOPA system fluctuate above 40°, whereas in the apo-OA1 (Fig. 8a) and OA1-dopamine (Fig. 8c) systems, the value remains below 40°. The other TM helices, however, maintain tilt angles of around 15–20°.

OA1-dopamine simulation

Dopamine is thought to act as an antagonist against OA1, which competes and binds in the same pocket as L-DOPA [36]. The results obtained in the present study indicate that, in comparison to L-DOPA, dopamine imparts more rigidity to OA1, making it more compact than even the apo-OA1 structure. The 7TM helices show an average RMSD of 2.5 Å ,while the RMSD value of the entire structure lies in the region of 8 Å (Fig. 5a,b) from 5 ns onwards, which is lower compared to the apo-OA1 and OA1- L-DOPA systems for which the values are in the range of 11 Å and 9 Å, respectively. Fluctuation in the radius of gyration value, which is relatively high compared to that in the presence of L-DOPA, shows that dopamine does not affect the loop regions so much. A look at the individual TM helix RMSD (Fig. 6c) shows enhanced movement of the eighth helix from its initial conformation, with a value of about 3 Å. The eighth helix shows minimal fluctuation in the other two simulations. TM5 maintains a relatively stable value below 2 Å from 5 ns onwards. The decreased movements of these regions are reflected in the RMSF plot (Fig. 5c), which also shows increased fluctuation for ECL2 (residue 176–190) and ECL3 (residue 270–291). Interestingly, the final structure obtained after 15 ns simulation shows the occurrence of an anti-parallel β-sheet on ECL2, which is not observed in the other two simulations. However, the template 2RH1 shows the presence of an α-helix on ECL2 [7]. The ICLs show no notable movement in the presence of dopamine. The SASA value for OA1-dopamine remains almost constant at about 21,000 Å^2 from 5 ns onwards (Fig. 7b), indicating a closed conformation of the GPCR as compared to OA1-L-DOPA system.

Agonist- and antagonist-induced structural changes of the receptor

Hydrogen bonding pattern

Hydrogen bonds play an important role in describing how a small molecule like L-DOPA or dopamine interacts with a protein, and in bringing about conformational changes switching it from one conformational state to another. OA1 is known to have a single saturable binding site for both L-DOPA and dopamine [36]. Docking studies performed here have also shown that both L-DOPA and dopamine dock into the same binding pocket surrounded by TM3, TM5, TM6 and ECL2, and stabilizes the entire complex by forming a dynamic network of hydrogen bonds with the receptor. All the different hydrogen bonds that have been formed between OA1 and the ligands throughout the entire simulation period of 15 ns are listed in Supplementary Table S1. Figure 4a and b show the top view from the extracellular side; Fig. 4c and d show the side view of the binding pocket, showing the average structures of L-DOPA forming hydrogen bonds with Y178 (on ECL2) and N260 (on TM6), and dopamine forming hydrogen bonds with H194 (on TM5), respectively. Figure 4

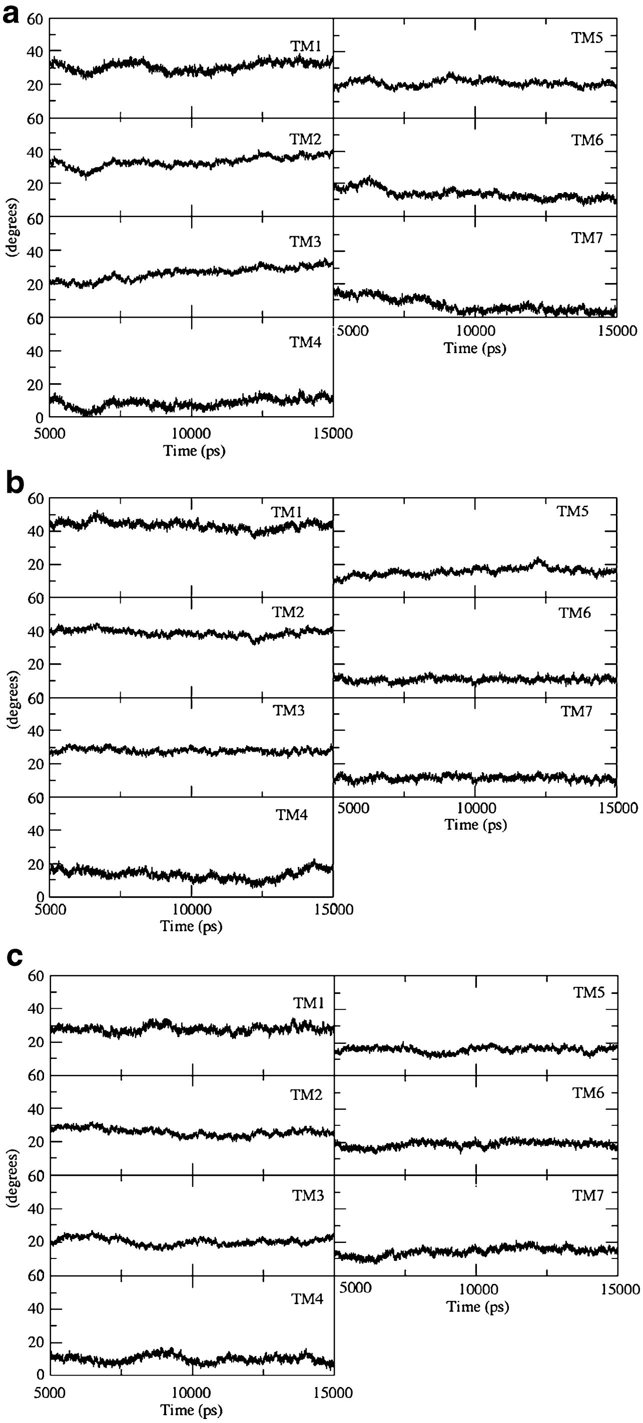


Fig. 8 Axis tilts of 7 TM helices. Tilts of the individual 7TM helices with respect to the *z*-axis were plotted for the entire simulation length for **a** apo-OA1, **b** OA1-L-DOPA and **c** OA1-dopamine systems. The tilt angle was calculated as the angle between a best-fit cylinder to the helix axis relative to the bilayer normal, using the GROMACS analysis package. All three systems show fluctuations in the tilt angles of all the 7 TM helices, with the difference between the maximum and minimum tilt angle shown by each helix ranging from 10° to 15°. The plots were prepared using XMGRACE

also shows, in “stick” representation, the other residues of the receptor with which the ligands persistently form stable hydrogen bonds. These hydrogen bonds are found consistently throughout the simulation as seen from Table 2. The two hydrogen bonds formed by Y178 on ECL2 with L-DOPA remain stable from 2 ns onwards with intermittent breakages, while the two formed by N260 remain from 1.8 ns onwards (Fig. 9a). Two more dominant hydrogen bonds were also observed throughout the entire 15 ns simulation between L-DOPA and D189 on ECL2 (Fig. 9a). The hydrogen bonds formed by dopamine with E264 and H194 were found to occur for a considerable time period (Fig. 9b) during the production phase. Dopamine also forms an intermittent hydrogen bond with D189 on ECL2 (Fig. 9b). As a result, ECL2, which acts as a closing lid, does not remain so firmly over the binding pocket as it does in the case of L-DOPA. Also, both L-DOPA and dopamine were found to form hydrogen bonds with Q124 on TM3, the mutation of which to Arg causes OA [34]. Dopamine also forms a hydrogen bond with I261 and W292, mutation of which to Asn and Gly, respectively, have also been reported in patients suffering from OA [34].

Figure 10 shows the distribution of clusters of OA1 conformations that have maintained zero, one, two, three, four, five, six or seven hydrogen bonds with either L-DOPA (black bars) or dopamine (grey bars). It is clear that L-DOPA forms a higher number (maximum of 7) of hydrogen bonds with OA1 than dopamine, throughout the entire simulation period. On average, L-DOPA maintains about two to three hydrogen bonds with the receptor, while dopamine maintains an average of one hydrogen bond throughout the simulation period or does not form hydrogen bonds with the receptor at all. These extra hydrogen bonds might help L-DOPA to stimulate OA1, thus acting as an agonist. It is also clear from Supplementary Table S1 that the total number of hydrogen bonds formed by dopamine is much higher than the number formed by L-DOPA with the receptor (50 vs 32), but most of these hydrogen bonds formed by dopamine are transient and maintained for only a few pico-seconds, while L-DOPA forms a more stable network of hydrogen bonds. Table 2 lists the hydrogen bond pairs between OA1 and the two ligands that are maintained at a distance below 3.5 Å, thus indicating hydrogen bonds, for at least 2 ns. From Table 2 it is clear that L-DOPA forms a strong network of hydrogen bonds with residues on TM5, TM6 and ECL2, which is also reflected in Fig. 9. As suggested for other GPCRs [5–10, 61, 62], in the case of OA1, ECL2 also plays an important role in keeping ligands in the binding pocket of the receptor.

Another interesting observation with respect to the hydrogen bonding pattern is that, in the OA1-L-DOPA system, a hydrogen atom of the aromatic ring of Y142 on TM3 was found to form a stable hydrogen bond with an oxygen atom of L212 on TM5 from 4 ns onwards (Fig. 11a,

Table 2 Hydrogen bonds formed between the receptor and L-DOPA and dopamine that are maintained for at least 2 ns continuously during molecular dynamics (MD) simulations

Atom pairs (donor–acceptor)	Receptor topology	Remarks
Hydrogen bonds formed between OA1 and L-DOPA		
ASN260HD22–LDP405OE2	TM6	Maintained at ≈2.5 Å from 1.8 ns onwards
MET198H–LDP405OE2	TM5	Maintained at ≈3 Å for first 2 ns, then increased over 7 Å
TYR178HH–LDP405O	ECL2	Initial at 10 Å, reduced to ≈1.7 Å from 2 ns onwards with intermittent increase to 3 Å
TYR178HH–LDP405OXT	ECL2	Initial at 10 Å, reduced to ≈1.5 Å from 2 ns onwards with intermittent increase to 3 Å
LDP405HZ–MET198N	TM5	Maintained at ≈3.2 Å with fluctuations throughout
LDP405HE2–HISB194NE2	TM5	Maintained at ≈3 Å for first 2.5 ns
LDP405HE2–HISB194O	TM5	Maintained at ≈3.5 Å for first 2.5 ns
LDP405HE2–ASN260OD1	TM6	Maintained at ≈3.4 Å from 5 ns onwards with fluctuations
LDP405H1–ASP189OD1	ECL2	Maintained at ≈3.4 Å throughout
LDP405H1–ASP189OD2	ECL2	Maintained at ≈3 Å throughout
LDP405H1–GLU264OE1	TM6	Maintained at ≈3 Å between 9 and 12 ns
LDP405H1–GLU264OE2	TM6	Maintained at ≈3 Å for first 5 ns and last 2.5 ns
Hydrogen bonds formed between OA1 and dopamine		
ILE297H–DOP405O1	TM7	Maintained at ≈3.5 Å between 1.5 and 3.4 ns and between 6.2 and 9.6 ns
ILE261H–DOP405O1	TM6	Maintained below 3.5 Å for the first 5.5 ns
TYR127HH–DOP405O2	TM3	Maintained at ≈3.5 Å for first 10 ns with fluctuations
DOP405H12–GLY187O	ECL2	Maintained at ≈3 Å for first 6 ns and then sudden jump above 10 Å
DOP405H12–ASP189OD1	ECL2	Maintained at ≈3.3 Å for first 12 ns
DOP405H12–HISB194ND1	TM5	Maintained at ≈3 Å for the last 9 ns
DOP405H12–GLU264O	TM6	Maintained at ≈3 Å between 6 and 13 ns
DOP405H11–PHE293O	TM7	Maintained at ≈3.3 Å between 6 and 10 ns

b), which was not seen in the other two systems. This hydrogen bond clearly differentiates the conformations of the receptor in the presence of the agonist and the antagonist.

Rotamer toggle switch on TM6

The transition from the active to inactive state of GPCRs involves the release of an important molecular constraint referred to as the "rotamer toggle switch", which is represented on TM6 by a change in rotameric conformation on a tryptophan residue [67–69]. The crystal structures of both rhodopsin and β2-AR show the presence of this toggle switch. In rhodopsin, W265 acts as the switch, which toggles upon activation, while in β2-AR this function is provided by W286. In OA1, there is also a Trp residue in the middle of TM6, whose indole ring side-chain faces the interior of the protein, as observed in the crystal structures of both rhodopsin and β2-AR. However, the recent elucidation of the active-state crystal structures of human β2-AR [8, 9] and human metarhodopsin II [13], show that the indole ring torsion angles of W286 and W265 are similar to those of their inactive-state crystal structures. Hence, there is a fair amount of debate regarding consideration of this rotamer toggle switch as a discriminating feature between the active and inactive states of GPCRs. In most GPCRs, this tryptophan residue occurs in a "WxP" motif, but in the case of OA1 this motif is absent from TM6. However, from the current observations it can be envisaged that W257 on TM6 acts as this toggle switch and distinguishes between the two distinct conformations of the receptor. The W257 torsion angle formed between CA–CB–CG–CD1 of the 0 ns structure of OA1 (shown in white color in Supplementary Fig. S4) has a value of 112°. Supplementary Fig. S4 depicts the two distinct conformations of the W257 rotamer in white (0 ns) and blue (after 15 ns simulation) colors. Interestingly, rotameric toggling is observed to occur quite clearly in both the apo-OA1 (Supplementary Fig. S4a) and OA1-dopamine systems (Supplementary Fig. S4c) with the CA–CB–CG–CD1 torsion angle values changing to 41° and −165°, respectively, from the initial value of 112°. However, the position

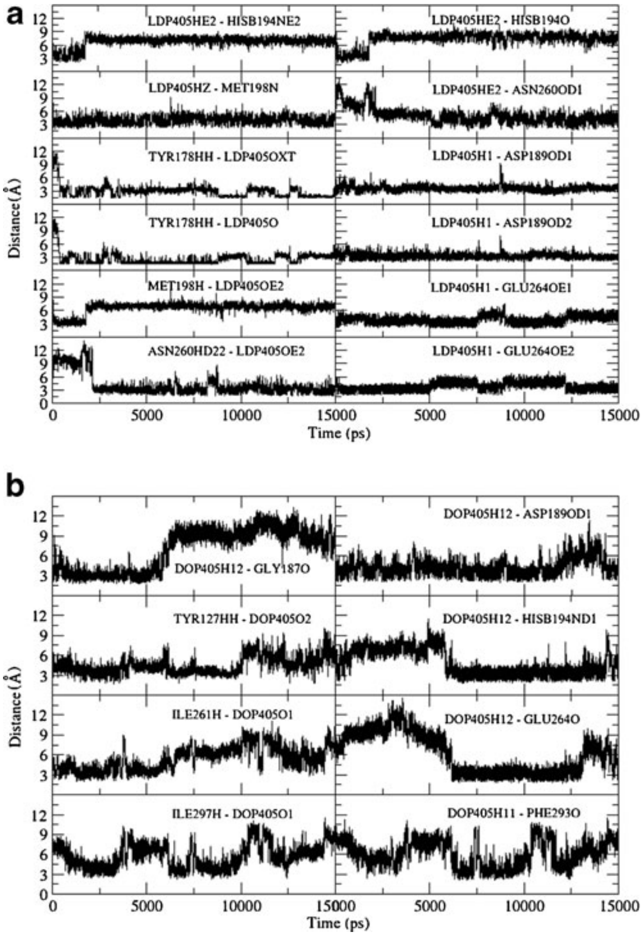


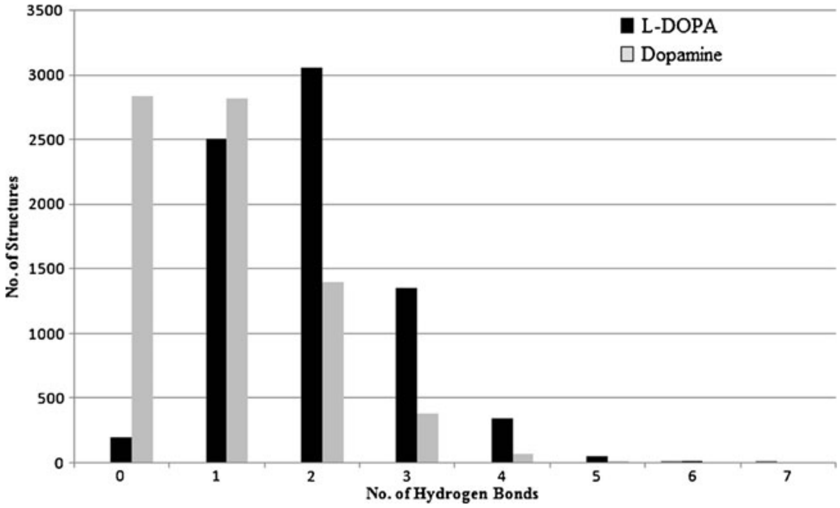
Fig. 9 Distance fluctuation between receptor–ligand hydrogen bond donor–acceptor pairs. The plot shows the distance fluctuation between those hydrogen bond donor–acceptor pairs that are maintained below 3.5 Å for at least 2 ns during the simulation, for **a** the OA1-L-DOPA system, and **b** the OA1-dopamine system, between the receptor and the ligands. It can be seen that L-DOPA forms a greater number of stable hydrogen bonds with OA1 than does dopamine. The plots were prepared using XMGRACE

of the W257 rotamer in the presence of L-DOPA remains same throughout the 15 ns simulation (Supplementary Fig. S4b). The aromatic rings of both L-DOPA and dopamine are found to be docked roughly in the same position, surrounded by TM3, TM5, TM6 and ECL2, and engage this toggle switch. The toggling of the W257 rotamer also depends on the relative movement of TM6, and both events occur in a concerted manner.

Helical tilts

Site-directed spin labeling (SDSL) studies have shown that rigid body motions, in terms of helical tilts of TM6 and TM3, play a vital role in the photo-activation of rhodopsin—the most well understood system in the GPCR superfamily [70–74]. Hence, we examined whether OA1 also shows any such rigid body motions represented in terms of helical tilts, in the presence of the two ligands. The helical tilts were calculated for all TM helices for all three systems during the 15 ns simulations. The tilt angle was calculated as the angle between a best-fit cylinder to the helix axis relative to the bilayer normal. Each helix was divided into top and bottom sections on the basis of the kinks and bends present in each helix. The difference between the maximum and minimum tilt angle shown by each helix ranged between 10° and 15°. For the apo-OA1 system, all the TM helices were observed to show some fluctuations at around roughly 5 ns (Fig. 8a). Among them, TM1, TM2, TM4, TM5 and TM6 show some abrupt changes in their tilt angles around 6 ns. In the OA1-L-DOPA system, it was observed that the tilt angle values for all the helices decrease initially and the value for TM5 then increases by about 10° from 5 ns onwards, due mainly to movement at the intracellular end of the helix. The relatively higher tilt values of TM1 and TM2 were correlated with visual observation of the trajectory, which revealed lateral

Fig. 10 Variation in the number of conformations of OA1 having a different number of hydrogen bonds with the ligand. The histogram shows the number of OA1 conformations having a particular number of hydrogen bonds with L-DOPA (*black bars*) or dopamine (*grey bars*). The conformations were analyzed every 2 ps. It can be seen that the number of conformations of OA1 having hydrogen bonds with L-DOPA is greater than that with dopamine



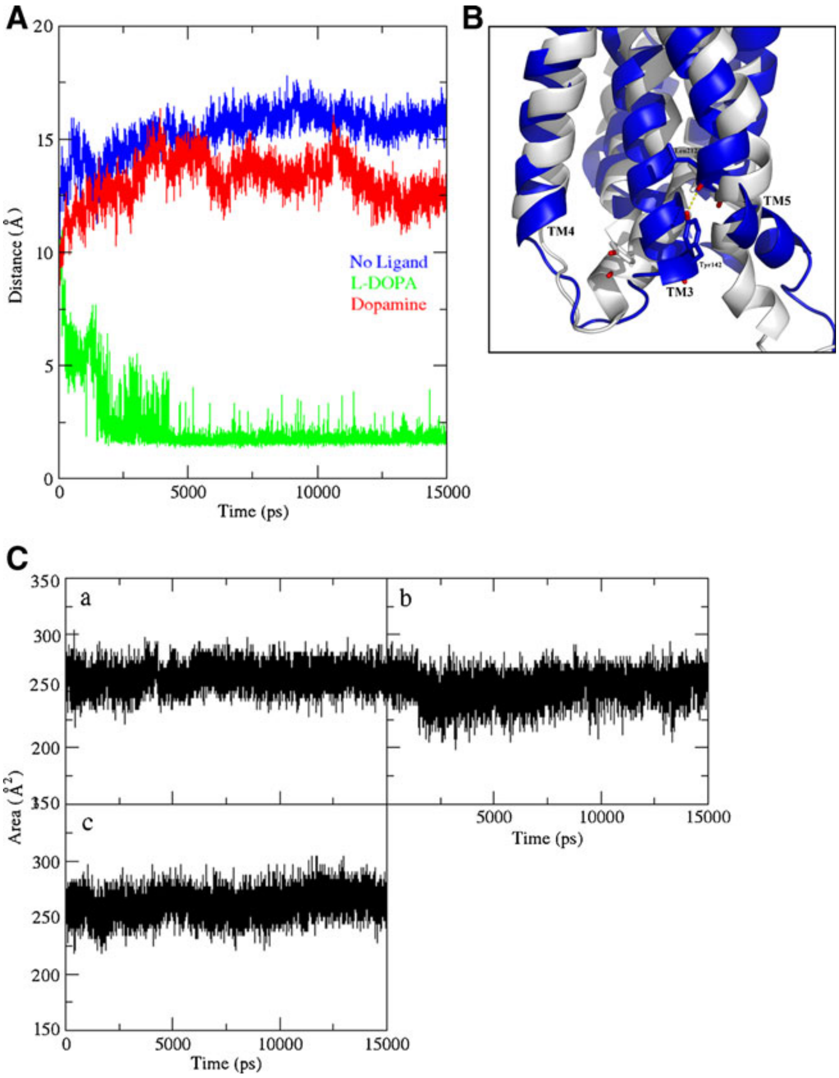


Fig. 11 "DAY" motif analysis. **A** Distance fluctuation between the hydrogen atom on the aromatic ring of Y142 on TM3 and the oxygen atom on L212 on TM5, which form a hydrogen bond in the presence of L-DOPA (*green*). This hydrogen bond is shown diagrammatically in **B**. The relative movement of TM5 away from the *z*-axis normal in the presence of L-DOPA (*blue*) can be seen, while in the presence of dopamine (*white*), TM5 does not show such deviation. **C** Variations in the SASA of the "DAY" motif for the systems (a) apo-OA1, (b) OA1-L-DOPA and (c) OA1-dopamine. The plots were prepared using XMGRACE and the figure was rendered using PyMOL

movement of TM1 and TM2 creating a slight opening of the binding pocket. The OA1-dopamine system showed some random fluctuations in tilt angle values, with the maximum being shown by TM3, TM4 and TM6. These abrupt changes in tilt angles, besides fast fluctuations, show that the TM helices of OA1, and GPCRs in general, are dynamic structures with various dynamical modes.

"DAY" motif on TM3

Extensive experimental and computational studies have shown that the activation process of both bovine rhodopsin and human β2-AR involves the breaking of an "ionic lock" [75–79] that is formed between the Arg residue on the "D/ERY" motif at the end of TM3—a signature of these GPCRs [69, 80]—and a conserved Glu residue on TM6. However, in the case of OA1 this Arg residue is replaced by an Ala residue, forming a "DAY" motif at the end of TM3. Being a non-polar amino acid, Ala is unable to form any salt bridge with E264 on TM6. However, some variations in the SASA value of the "DAY" motif for the OA1-L-DOPA system were observed. From Fig. 11c it can be seen that the apo-OA1 and OA1-dopamine systems (Fig. 11c) maintain almost a constant SASA value of about 270 Å^2 throughout the simulation for the "DAY" motif, while OA1-L-DOPA system (Fig. 11c) shows a slight decrease in the value at the beginning, followed by a gradual increase to about 270 Å^2. An interesting observation revealed that, in the OA1-L-DOPA system, the end residue of the "DAY" motif, Y142, on TM3 formed a stable hydrogen bond with L212 on the middle of TM5 from 5 ns onwards. The distance fluctuation plot (Fig. 11a) between the hydrogen atom on the aromatic ring of Y142 and the oxygen atom of L212 clearly shows that L-DOPA helps the receptor to form and maintain this hydrogen bond (shown diagrammatically in Fig. 11b). As a result, the lower end of TM3 holds the middle of TM5 together. However, the intracellular end of TM5 (residues 213–224) forms a bend at this point of the helix and deviates away from the *z*-axis normal, as can be clearly seen from Fig. 11b. This forms an opening in the intracellular end of the receptor surrounded by TM3, TM5 and TM6 which might act as the possible binding site of G-protein to OA1. The same can also be inferred from the gradual increase of receptor SASA in presence of L-DOPA as shown in Fig. 7b. However, this hydrogen bond was found to be absent in the case of both the apo-OA1 and OA1-dopamine systems where the intracellular end of TM5 did not deviate substantially to form this opening. In both these cases, the aromatic ring of Y142 was found to point away from the TM helices. This hydrogen bond mediated by the "DAY" motif is another feature that distinguishes between the agonist- and antagonist-bound forms of the receptor.

Conclusions

Being a GPCR that is expressed exclusively on intracellular organelles—the melanosomes—OA1 poses a unique challenge as a system for study. The present study attempted to address several issues, starting from model building to MD simulations in the presence and absence of ligands—steps that are vital to the proper understanding of the actual mechanism of the functioning of GPCRs and that are also useful in the drug discovery process. The predicted structure of OA1 was found to stand up to scrutiny, with about 92% of the residues falling within the most favored regions of the Ramachandran plot, while its quality was further refined and established through MD simulations. The simulations showed that the structure is stable in an explicit lipid bilayer environment over the period of the performed simulations, and that the helical conformations, including the eighth helix—a characteristic feature of rhodopsin family of GPCRs, which remains parallel to the lipid bilayer—were preserved.

The predicted structure of OA1 is attributed with some of the signature characteristics of GPCRs in conformationally correct positions, e.g., the "DAY" motif at the end of TM3, the rotameric toggle switch in the form of W257 in the middle of TM6, and some of the Pro residues in the middle of TM2, TM5 and TM7 that make the characteristic kinks in the TM helices. The binding pockets on OA1 for both L-DOPA and dopamine show that both ligands bind in a region similar to the ligand binding pocket of β2-AR. The predicted binding sites and energetics of these ligands agree quite well with the available experimental results, showing that both ligands bind to a single binding pocket. The selected binding poses of the ligands were found to form hydrogen bonds with some important residues such as Q124, I261 and W292, mutations in which are known to cause OA. It was also seen that ECL2 acts as a "closing lid" over the binding pocket in the case of both ligands and in a way stabilizes the conformation of that region, which is also observed for β2-AR. In β2-AR, ECL2 harbors a small α-helix, while the predicted model of OA1 did not show the presence of any such helical conformation. Interestingly, among the structures obtained after 15 ns of simulation, in the presence of dopamine OA1 showed the occurrence of an anti-parallel β-sheet on ECL2, whereas the apo-OA1 and OA1-L-DOPA structures retained only the loop conformations on ECL2.

Further investigations through MD simulations helped to shed more light on some of the structural features of the receptor in the presence of L-DOPA and dopamine. There is a clear difference in the rotameric conformations of W257 in the agonist- and antagonist-bound forms of the receptor. The hydrogen bonding pattern shows that L-DOPA imparts more stability to OA1 through a network of hydrogen bonds that is stronger than that in the presence of dopamine. Most of the hydrogen bonds formed with TM5 and TM6, suggesting that TM5 and TM6 might play an important role in the conformational switching of OA1. On average, L-DOPA maintains two to three hydrogen bonds with OA1 throughout the simulation. These bonds, which are mediated by the agonist, might stimulate the receptor in long time-scale cellular processes. The reliability of the MD simulations is also verified by the fact that the ligands were found to interact with residues like Q124, W292 and I261, mutations in which cause OA. The MD results also emphasize the fact that these ligands have bound to the correct pockets on the receptor and play a role in keeping the receptor in distinct conformations by engaging important residues. However, no "ionic lock" was observed between TM3 and TM6, which can be explained by the fact that OA1, unlike rhodopsin family GPCRs, does not have a "D/ERY" motif, but instead has a "DAY" motif. This "DAY" motif in turn was found to be involved in a hydrogen bond between TM3 and the middle of TM5 only in the presence of L-DOPA, which distinguishes the two conformations of OA1. As a consequence, the cavity formed due to movement of the intracellular end of TM5 away from the *z*-axis normal, and surrounded by TM3 and TM6, might act as the binding pocket for G-proteins. Besides these, the fluctuations harbored by the TM helices of OA1 reveal the dynamic nature of the receptor switching between several dynamical modes.

The present study on OA1 sheds light on the structural aspects of an important GPCR—a protein superfamily of paramount medical importance. This computational modeling study in conjunction with MD simulations has successfully predicted the structure and provided insight into the structural features of the receptor and its dynamically changing conformations in the presence of an agonist and an antagonist. This might help to design future experiments and further the understanding of this novel receptor.

Acknowledgments A.G., U.B.S. and R.R.J. gratefully acknowledge the Department of Information Technology (DIT), Government of India, New Delhi, for providing financial support. This work was performed using the "Bioinformatics Resources and Applications Facility (BRAF)" at C-DAC, Pune, funded by DIT, New Delhi. A.G. K. and A.S.K. gratefully acknowledge funding support from IIT Madras, Department of Science and Technology (DST) and Department of Biotechnology (DBT), Government of India.

References

1. Schöneberg T, Schulz A, Gudermann T (2002) The structural basis of G-protein-coupled receptor function and dysfunction in human diseases. Rev Physiol Biochem Pharmacol 144:143–227
2. Lundstrom K (2009) An overview on GPCRs and drug discovery: structure-based drug design and structural biology on GPCRs. Methods Mol Biol 552:51–66
3. Wilson S, Bergsma D (2000) Orphan G-protein coupled receptors: novel drug targets for the pharmaceutical industry. Drug Des Discov 17:105–114
4. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5:993–996
5. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H et al (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. Science 289:739–745
6. Ballesteros J, Palczewski K (2001) G protein-coupled receptor drug discovery: implications from the crystal structure of rhodopsin. Curr Opin Drug Discov Dev 4:561–574
7. Rasmussen SG, Choi HJ, Rosenbaum DM, Kobilka TS, Thian FS et al (2007) Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. Nature 450:383–387
8. Rasmussen SG, Choi HJ, Fung JJ, Pardon E, Casarosa P et al (2011) Structure of a nanobody-stabilized active state of the β(2) adrenoceptor. Nature 469:175–180
9. Rosenbaum DM, Zhang C, Lyons JA, Holl R, Aragao D et al (2011) Structure and function of an irreversible agonist-β(2) adrenoceptor complex. Nature 469:236–240
10. Jaakola VP, Griffith MT, Hanson MA, Cherezov V, Chien EY et al (2008) The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. Science 322:1211–1217
11. Murakami M, Kouyama T (2008) Crystal structure of squid rhodopsin. Nature 453:363–367
12. Wu B, Chien EY, Mol CD, Fenalti G, Liu W et al (2010) Structures of the CXCR4 chemokine GPCR with small molecule and cyclic peptide antagonists. Science 330:1066–1071
13. Choe HW, Kim YJ, Park JH, Morizumi T, Pai EF et al (2011) Crystal structure of metarhodopsin II. Nature 471:651–655
14. Patny A, Desai PV, Avery MA (2006) Homology modelling of G-protein coupled receptors and implications in drug design. Curr Med Chem 13:1667–1691
15. Kanagarajadurai K, Malini M, Bhattacharya A, Panicker M, Sowdhamini R (2009) Molecular modeling and docking studies of human 5-hydroxytryptamine 2A (5-HT$_{2A}$) receptor for the identification of hotspots for ligand binding. Mol BioSys 5:1877–1888
16. Miedlich SU, Gama L, Seuwen K, Wolf RM, Breitwieser GE (2004) Homology modeling of the transmembrane domain of the human calcium sensing receptor and localization of an allosteric binding site. J Biol Chem 279:7254–7263
17. Dastmalchi S, Church WB, Morris MB (2008) Modelling the structures of G protein-coupled receptors aided by three-dimensional validation. BMC Bioinforma 9:S14
18. Niv MY, Skrabanek L, Filizola M, Weinstein H (2006) Modeling activated states of GPCRs: the rhodopsin template. J Comput Aided Mol Des 20:437–448
19. Costanzi S (2008) On the applicability of GPCR homology models to computer-aided drug discovery: a comparison between in silico and crystal structures of the β2-adrenergic receptor. J Med Chem 51:2907–2914
20. Lavecchia A, Cosconati S, Novellino E (2005) Architecture of the human urotensin II receptor: comparison of the binding domains of peptide and non-peptide urotensin II agonists. J Med Chem 48:2480–2492
21. Periole X, Weinstein H (2002) Key issues in computational simulation of GPCR function. J Comput Aided Mol Des 16:841–853
22. Ivetac A, Sansom MS (2008) Molecular dynamics simulations and membrane protein structure quality. Eur Biophys J 37:403–409
23. Fan H, Mark AE (2004) Refinement of homology-based protein structures by molecular dynamics simulation techniques. Protein Sci 13:211–220
24. Kobilka B, Schertler GF (2008) New G-protein-coupled receptor crystal structures: insights and limitations. Trends Pharmacol Sci 29:79–83
25. Klein-Seetharaman J (2002) Dynamics in rhodopsin. ChemBioChem 3:981–986
26. Vilardaga JP, Bünemann M, Krasel C, Castro M, Lohse MJ (2003) Measurement of the millisecond activation switch of G protein-coupled receptors in living cells. Nat Biotechnol 21:807–812
27. Shen B, Samaraweera P, Rosenberg B, Orlow SJ (2001) Ocular albenism type I: more than meets the eye. Pigment Cell Res 14:243–248
28. Incerti B, Cortese K, Pizzigoni A, Surace EM, Varani S et al (2000) Oa1 knock-out: new insights on the pathogenesis of ocular albinism type I. Hum Mol Genet 9:2781–2788
29. Bassi MV, Schiaffino MV, Renieri A, De Nigris F, Galli L et al (1995) Cloning of the gene for ocular albinism type I from the distal short arm of the X chromosome. Nat Genet 10:13–19
30. Schiaffino MV, Bassi MV, Galli L, Renieri A, Bruttini M et al (1995) Analysis of the OA1 gene reveals mutations in only one-third of the patients with X linked ocular albinism. Hum Mol Genet 4:2319–2325
31. Schiaffino MV, d'Addio M, Alloni A, Baschirotto C, Valetti C et al (1999) Ocular albinism: evidence for a defect in an intracellular signal transduction system. Nat Genet 23:108–112
32. Schiaffino MV, Tacchetti C (2005) The Ocular Albinism type I (OA1) protein and the evidence for an intracellular signal transduction system involved in melanosome biogenesis. Pigment Cell Res 18:227–233
33. Innamorati G, Piccirillo R, Bagnato P, Palmisano I, Schiaffino MV (2006) The melanosome/lysosomal protein OA1 has properties of a G protein coupled receptor. Pigment Cell Res 19:125–135
34. d'Addio M, Pizzigoni A, Bassi MT, Baschirotto C, Valetti C et al (2000) Defective intracellular transport and processing of OA1 is a major cause of ocular albinism type 1. Human Mol Genet 9:3011–3018
35. Palmisano I, Bagnato P, Palmigiano A, Innamorati G, Rotondo G et al (2008) The ocular albinism type 1 protein, an intracellular G protein coupled receptor, regulates melanosome transport in pigment cells. Human Mol Genet 17:3487–3501
36. Lopez VM, Decatur CL, Stamer WD, Lynch RM, MacKay BS (2008) L-DOPA is an endogenous ligand for OA1. PLoS Biol 6:e236
37. Bairoch A, Apweiler R (1998) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucleic Acids Res 26:38–42
38. Bateman A, Coin L, Durbin R, Finn RD, Hollich V et al (2004) The Pfam protein families database. Nucleic Acids Res 32:138–141
39. Altschul SF, Madden TL, Schäffer AA (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

40. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948
41. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J Mol Biol 305:567–580
42. Rost B, Yachdav G, Liu J (2004) The PredictProtein Server. Nucleic Acids Res 32:W321–W326
43. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SGF, Thian FS et al (2007) High-resolution crystal structure of an engineered human β2-adrenergic G protein-coupled receptor. Science 318:1258–1265
44. Okada T, Sugihara M, Bondar AN, Elstner M, Entel P et al (2004) The retinal conformation and its environment in rhodopsin in light of a new 2.2 Å crystal structure. J Mol Biol 342:571–583
45. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Series 41:95–98
46. Šali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815
47. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26:283–291
48. Eisenberg D, Lüthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol 277:396–404
49. Schrödinger Suite (2009) QM-polarized ligand docking protocol; Glide version 5.5; Jaguar version 7.6; QSite version 5.5. Schrödinger, LLC, New York, NY
50. Chung JY, Hah JM, Cho AE (2009) Correlation between performance of QM/MM docking and simple classification of binding sites. J Chem Inf Model 49:2382–2387
51. Hess B, Kutzner C, van der Spoel D, Lindahl EJ (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. Chem Theor Comput 4:435–447
52. Schuettelkopf AW, van Aalten DMF (2004) PRODRG—a tool for high-throughput crystallography of protein–ligand complexes. Acta Crystallographica D60:1355–1363
53. Kandt C, Ash WL, Tieleman DP (2007) Setting up and running molecular dynamics simulations of membrane proteins. Methods 41:475–488
54. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. J Chem Phys 126:014101
55. Hess B (2008) P-LINCS: a parallel linear constraint solver for molecular simulations. J Chem Theor Comput 4:116–122
56. Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Prot Eng 8:127–134
57. XMGRACE: http://plasma-gate.weizmann.ac.il/Grace/
58. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14:33–38
59. DeLano WL (2003) The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA
60. Schnur RE, Gao M, Wick PA, Keller M, Benke PJ et al (1998) OA1 mutations and deletions in X-linked ocular albinism. Am J Hum Genet 62:800–809
61. Unal H, Jagannathan R, Bhat MB, Karnik SS (2010) Ligand-specific conformation of extracellular loop-2 in the angiotensin II type 1 receptor. J Biol Chem 285:16341–16350
62. Conner M, Hawtin SR, Simms J, Wootten D, Lawson Z et al (2007) Systematic analysis of the entire second extracellular loop of the V(1a) vasopressin receptor: key residues, conserved throughout a G-protein-coupled receptor family, identified. J Biol Chem 282:17405–17412
63. Dunham TD, Farrens DL (1999) Conformational changes in rhodopsin. Movement of helix f detected by site-specific chemical labeling and fluorescence spectroscopy. J Biol Chem 274:1683–1690
64. Knierim B, Hofmann KP, Ernst OP, Hubbell WL (2007) Sequence of late molecular events in the activation of rhodopsin. Proc Natl Acad Sci USA 104:20290–20295
65. Kobilka BK (2007) G protein coupled receptor structure and activation. Biochim Biophys Acta 1768:794–807
66. Kobilka BK (2002) Agonist-induced conformational changes in the beta2 adrenergic receptor. J Pept Res 60:317–321
67. Shi L, Liapakis G, Xu R, Guarnieri F, Ballesteros JA et al (2002) Beta2 adrenergic receptor activation. Modulation of the proline kink in transmembrane 6 by a rotamer toggle switch. J Biol Chem 277:40989–40996
68. Bhattacharya S, Hall SE, Li H, Vaidehi N (2008) Ligand-stabilized conformational states of human β2 adrenergic receptor: insight into G-protein-coupled receptor activation. Bio Phys J 94:2027–2042
69. Bhattacharya S, Hall SE, Vaidehi N (2008) Agonist-induced conformational changes in bovine rhodopsin: insight into activation of G-protein-coupled receptors. J Mol Biol 382:539–555
70. Farahbakhsh ZT, Hideg K, Hubbell WL (1993) Photoactivated conformational-changes in rhodopsin—a time-resolved spin-label study. Science 262:1416–1419
71. Hubbell WL, Cafiso D, Altenbach C (2000) Identifying conformational changes with site-directed spin labeling. Nature Struct Biol 7:735–739
72. Langen R, Cai K, Altenbach C, Khorana HG, Hubbell WL (1999) Structural features of the C-terminal domain of bovine rhodopsin: a site directed spin-labeling study. Biochemistry 38:7918–7924
73. Farrens D, Altenbach C, Yang K, Hubbell WL, Khorana HG (1996) Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin. Science 274:768–770
74. Crozier PS, Stevens MJ, Forrest LR, Woolf TB (2003) Molecular dynamics simulations of dark-adapted rhodopsin in an explicit membrane bilayer: coupling between local retinal and larger scale conformational change. J Mol Biol 333:493–514
75. Vogel R, Mahalingam M, Lüdeke S, Huber T, Siebert F et al (2008) Functional role of the “Ionic Lock”—an interhelical hydrogen-bond network in family a heptahelical receptors. J Mol Biol 380:648–655
76. Ballesteros JA, Jensen AD, Liapakis G, Rasmussen SGF, Shi L et al (2001) Activation of the β2-adrenergic receptor involves disruption of an ionic lock between cytoplasmic ends of transmembrane segments 3 and 6. J Biol Chem 276:29171–29177
77. Dror RO, Arlow DH, Borhani DW, Jensen MØ, Piana S et al (2009) Identification of two distinct inactive conformations of the beta2-adrenergic receptor reconciles structural and biochemical observations. Proc Natl Acad Sci USA 106:4689–4694
78. Romo TD, Grossfield A, Pitman MC (2010) Concerted interconversion between ionic lock substates of the beta(2) adrenergic receptor revealed by microsecond timescale molecular dynamics. Biophys J 98:76–84
79. Sgourakis NG, Garcia AE (2010) The membrane complex between transducin and dark-state rhodopsin exhibits large-amplitude interface dynamics on the sub-microsecond timescale: insights from all-atom MD simulations. J Mol Biol 398:161–173
80. Fanelli F, De Benedetti PG (2006) Inactive and active states and supramolecular organization of GPCRs: insights from computational modeling. J Comput Aided Mol Des 20:449–461

ORIGINAL PAPER

Crystal structure and conformational analysis of *s-cis*-(acetylacetonato)(ethylenediamine-N,N′-diacetato)-chromium(III)

Development of vibrationally optimized force field (VOFF)

Jong-Ha Choi · Svetozar R. Niketić · Ivana Djordjević · William Clegg · Ross W. Harrington

Received: 19 January 2011 / Accepted: 11 July 2011 / Published online: 25 September 2011

Abstract The crystal structure of [Cr(edda)(acac)] (edda= ethylediamine-N,N′-diacetate; acac=acetylacetonato) has been determined by a single crystal X-ray diffraction study at 150 K. The chromium ion is in a distorted octahedral environment coordinated by two N and two O atoms of chelating edda and two O atoms of acac, resulting in *s-cis* configuration. The complex crystallizes in the space group $P2_1/c$ of the monoclinic system in a cell of dimensions a= 10.2588(9), b=15.801(3), c=8.7015(11) Å, β =101.201(9)° and Z=4. The mean Cr-N(edda), Cr-O(edda) and Cr-O(acac) bond distances are 2.0829(14), 1.9678(11) and 1.9477(11) Å while the angles O-Cr-O of edda and O-Cr-O of acac are 171.47(5) and 92.72(5)°, respectively. The crystal structure is stabilized by N–H···O hydrogen bonds linking [Cr(edda)-(acac)] molecules in distinct linear strands. The visible electronic and IR spectroscopic properties are also discussed. An improved, physically more realistic force field, Vibrationally Optimized Force Field (VOFF), capable of reproducing structural and vibrational properties of [Cr(edda)(acac)] was developed and its transferability demonstrated on selected chromium(III) complexes with similar ligands.

Keywords acac · Chromium(III) · Complex · Crystal structure · edda · Molecular mechanics · *s-cis* geometry · Spectral properties

Introduction

Ethylenediamine-N,N′-diacetate or 2,2′-[ethane-1,2-diylbis-(azanediyl)]diacetato (edda) has two acetate and two amino groups, and it can act as a quadridentate ligand in the complexation of the chromium(III) ion. The coordination of an acyclic edda ligand gives three possible geometric isomers of a $[Cr(edda)L_2]$ complex (L=unidentate). These isomers are commonly referred to as *trans*, symmetrical-*cis* (*s-cis*) and unsymmetrical-*cis* (*u-cis*) [1, 2]. When L_2 is a bidentate ligand such as acetylacetonato (acac), only *s-cis* and *u-cis* geometric diastereoisomers are possible (Scheme 1) since the *trans* isomer can not be expected to form with a chelating acac ligand.

The complex (acetylacetonato)(ethylenediaminediacetato)chromium(III) has been prepared, its *s-cis* isomer isolated, and *u-cis* isomer identified in solution [3]. It was already known that the *s-cis* geometry of the chelating edda ligand is favoured in most Co(III) and Cr(III) complexes [2–5]. It was suggested that the observed strain of the diamine ring in edda may be a contributing factor in determining the configuration of the edda ligand [6].

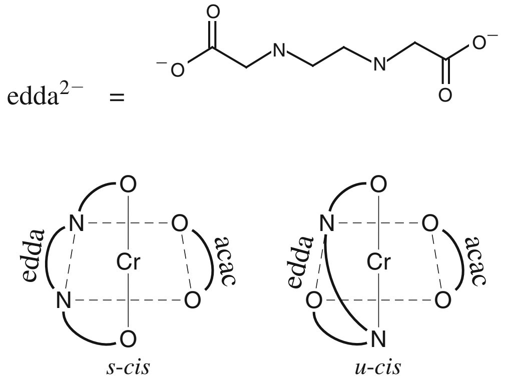
The assignment of geometrical configuration of some chromium(III) complexes with mixed ligands can be

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1185-2) contains supplementary material, which is available to authorized users.

J.-H. Choi (✉)
Department of Chemistry, Andong National University,
Andong 760-749, South Korea
e-mail: jhchoi@andong.ac.kr

S. R. Niketić (✉) · I. Djordjević
Chemistry Center, IHTM, University of Belgrade,
11158 Belgrade, Serbia
e-mail: niketic@gmx.com

W. Clegg · R. W. Harrington
School of Chemistry, Newcastle University,
Newcastle upon Tyne NE1 7RU, UK



Scheme 1 Diastereoisomers of [Cr(edda)(acac)]

performed by inspection of the d-d absorption and infrared spectra [7, 8]. It was also suggested that deuteration of the edda ligand at the acetate proton positions makes it possible to distinguish a chemically non-equivalent set of geminal deuterons on the chromium(III) complex by ^{2}H nuclear magnetic resonance spectroscopy [7].

The position of the spin-allowed transitions in the electronic spectra, the number of bands, and their extinction coefficients are usually reliable indicators for distinguishing the geometric isomers [9]. However, it should be noted that the assignments based on spectroscopic properties are not always conclusive [10]. Thus we should be very cautious in assigning the stereochemistry of a metal complex in the absence of single-crystal X-ray structural data. Furthermore, it is not easy to assign the geometry because the visible absorption maxima for the *s-cis* and *u-cis* isomers of [Cr(edda)(acac)] are very close to each other [3].

In this work, the crystal structure of [Cr(edda)(acac)] is determined to confirm the configuration of the edda ligand and the bidentate coordination of acac. In addition, due to the availability of sufficiently accurate structural and vibrational data (as shown below), the [Cr(edda)(acac)] molecule was used as a target structure for the optimization of molecular mechanics force field (FF) based on the requirement for satisfactory reproduction of conformations, energies, as well as the vibrational properties. This new force field we call Vibrationally Optimized Force Field (VOFF). A series of previous comprehensive conformational studies on edta–type Cr(III) compounds [1, 11–15] resulted in a force field capable of satisfactory reproducing the structures of several different families of diastereoisomeric complexes, and also their energetics [11] to the extent that in every case the most stable structure (in other words, the global minimum in molecular mechanics modeling) always corresponded to the one identified by an X-ray diffraction study. Even if such a force field has an advantage of being fully transferrable within a well defined range of structures, there is a need for a realistic force field able to provide a more accurate description of the interaction forces governing the motions of the atomic nuclei in the structures under investigation. In that respect VOFF represents a major improvement over the current force field [11] for edta-type Cr(III) complexes, as it ensures the consistent treatment of both the molecular structure and the vibrational frequencies. VOFF thus approaches the Consistent Force Field (CFF) philosophy in the way it was conceptualized by Lifson and his school [16–19], on which our CFF program for coordination compounds [20] was initially founded. To our knowledge the full incorporation of spectroscopic data in force field optimization in the realm of coordination chemistry has been rather limited.

Experimental

Synthesis and physical measurements

The ligand ethylenediamine-N,N′-diacetic acid was obtained from Aldrich Chemical Co. and used as supplied. All chemicals were reagent grade materials and used without further purification. The complex *s-cis*-[Cr(edda)-(acac)] was prepared according to a published procedure [4]. Recrystallization of the crude product from methanol–water (1:1) solution gave reddish violet crystals that were suitable for crystallographic analysis. Fujii *et al.* [4] reported the formula as [Cr(edda)(acac)]·2H_2O on the basis of elemental analysis and thermogravimerty. Subsequent authors [3, 5, 21], who prepared the complex by Fujii's method, did not present any experimental confirmation of its composition, but mainly carried over its original [4] designation as dihydrate. However, as will be shown below, we were not able to confirm the presence of any molecules of lattice water in our sample. The UV-visible absorption spectrum was recorded with an HP 8453 diode array spectrophotometer. The infrared spectra were obtained with a Mattson Infinities series FTIR spectrometer (KBr pellet in 1800–400 cm^{-1} region, and Nujol mull on PE film in 490–50 cm^{-1} region), and Thermo Scientific Nicolet 6700 and iS10 FTIR spectrometers (KBr pellet or ATR attachment, 4000–400 cm^{-1} range; 1 cm^{-1} resolution).

Crystal structure analysis

A reddish violet crystal (approximate dimensions of 0.52×0.42×0.40 mm) was mounted with inert oil on the top of a glass fibre. Single-crystal X-ray diffraction data were collected at 150 K on a Nonius KappaCCD diffractometer using graphite-monochromated Mo Kα radiation (λ = 0.71073 Å).

The diffraction data were measured using COLLECT [22]; unit cell parameters were refined and intensities obtained with EvalCCD [23]. Absorption corrections were applied by SADABS [24] based on symmetry-equivalent and repeated reflections. The structure was solved by direct methods and refined by full-matrix least-squares on F^2 using the SHELXTL [25]. Molecular graphics were produced using DIAMOND-3 [26] and ORTEP3 [27]. Non-hydrogen atoms were refined anisotropically; hydrogen atoms were first located in a difference map, then N–H hydrogen atoms were freely refined and C-H hydrogen atoms were constrained to ride on the parent carbon atom, with C-H=0.99 Å and $U_{iso}(H)=1.2U_{eq}(C)$. Table 1 contains a summary of crystal parameters, data collection and refinement.

Force field calculations

The calculations were performed with a locally modified version (rev. 2009) of the original Consistent Force Field (CFF) conformational program [20]. The components of the force field (potential functions for bond stretching, angle bending, torsional, non-bonded and electrostatic contributions, as well as the starting values of required parameters for coordinated polyamino-polycarboxylato structures) have been developed and particularized earlier [11]. Additional force field components needed for modelling of the acetylacetonato metal-chelate rings were introduced as required. In view of the commonly accepted [28] pseudo-aromatic nature of β-diketonato metal chelate rings the planarity on each sp^2 C atom of acac was maintained using ”out-of-plane” angle functions and parameters similar to those described by Morino and Shimanouchi [29].

The essentially diagonal force field is parameterized on the basis of three different types of carbon atom (tetrahedral, trigonal aromatic, and carbonyl carbon), three different types of oxygen atom (carboxylate oxygen ligator, acetylacetonato oxygen ligator, and carbonyl oxygen), and one type each of nitrogen, hydrogen, and metal atom. The same parameters were used for aliphatic carbon atoms from diamine and carboxylate chelate rings, as well as for the methyl groups on the acac ring. Point charge parameters were assigned on the basis of the results of Mulliken population analysis performed in quantum mechanics calculations with the ORCA program [30], and adjusted automatically so that their sum is zero.

Geometry optimizations were carried out using a combination of steepest-descent, Davidon-Fletcher-Powell and Newton-Raphson methods following the protocol described in detail elsewhere [11, 20]. The rms gradient of the total energy $<10^{-7}$ kJ/mol Å was taken as the convergence criterion in all energy minimizations. Refinement of force field parameters was first done batchwise, e.g. for suitable small (not necessarily non-overlapping) fragments, taken repeatedly in turn, in which interaction forces could be approximately taken as being decoupled from the remainder of the structure, the required FF parameters were least-squares fitted against the corresponding geometry (X-ray) data and experimental vibrational frequencies. Then the whole set of FF parameters was treated in a likewise manner to allow for further refinement due to small cross-term effects. In this way we hopefully avoided the risk of getting physically unrealistic FF parameter values out of a fully automated optimization.

Table 1 Crystallographic data, data collection and refinement for *s-cis*-[Cr(edda)(acac)]

Formula	$C_{11}H_{17}CrN_2O_6$
M_r	325.3
Crystal system	Monoclinic
Space group	$P2_1/c$
a(Å)	10.2588(9)
b(Å)	15.801(3)
c(Å)	8.7015(11)
β (°)	101.201(9)
V (Å^3)	1383.7(3)
Z	4
D_c (g cm^{-3})	1.561
μ(Mo Kα) (mm^{-1})	0.853
θ range (°)	4.6 to 27.5
Reflections collected	17923
Independent reflections	3163 (R_{int}=0.0355)
Reflections with $F^2>2\sigma$	2670
Min. and max. transmission	0.665 and 0.727
Refined parameters	191
R [$F^2>2\sigma$]	0.0272
R_w (F^2, all data)	0.0712
Goodness-of-fit on F^2	1.093
Largest diff. peak and hole	0.35 and −0.35 e Å^{-3}

Results and discussion

Crystallography

Selected bond lengths and angles are listed in Table 2. An ellipsoid plot of the complex together with the atomic labelling is illustrated in Fig. 1 (hydrogen atoms are shown as spheres of arbitrary radii). The dianionic edda ligand is tetradentate and it coordinates to the chromium(III) ion such that the two carboxylate oxygen atoms occupy *trans* positions while the two amine nitrogen atoms occupy *cis* positions, resulting in an *s-cis* configuration. The two oxygen atoms of acac coordinate to the remaining positions of the chromium(III) ion *trans* with respect to the ethylenediamine N donors. Thus the complex has a distorted octahedral geometry. The O4-Cr-O6 angle is 171.47(5) while the N1-Cr-O2 and N2-Cr-O1 angles are 173.41(5) and 173.85(5)°, respectively. The distortion is largely caused by the restricted bite angles of the chelating ligands.

Table 2 Selected bond distances (Å) and angles (◦) for *s-cis*-[Cr(edda)-(acac)]

Cr–O1	1.9509(11)	Cr–O2	1.9445(11)
Cr–O4	1.9623(11)	Cr–O6	1.9732(11)
Cr–N1	2.0802(13)	Cr–N2	2.0855(14)
O1–C2	1.282(2)	O2–C4	1.282(2)
O3–C6	1.213(2)	O4–C6	1.316(2)
O5–C11	1.223(2)	O6–C11	1.306(2)
N1–C7	1.485(2)	N1–C8	1.486(2)
N2–C9	1.485(2)	N2–C10	1.484(2)
C1–C2	1.507(2)	C2–C3	1.394(2)
C3–C4	1.395(2)	C4–C5	1.504(2)
C6–C7	1.526(2)	C8–C9	1.522(2)
C10–C11	1.525(2)		
O1–Cr–O2	92.71(5)	O1–Cr–O4	93.08(5)
O1–Cr–O6	92.54(5)	O1–Cr–N1	91.74(5)
O1–Cr–N2	173.85(5)	O2–Cr–O4	92.31(5)
O2–Cr–O6	93.82(5)	O2–Cr–N1	173.41(5)
O2–Cr–N2	91.00(5)	O4–Cr–O6	171.48(5)
O4–Cr–N1	82.60(5)	O4–Cr–N2	91.67(5)
O6–Cr–N1	90.83(5)	O6–Cr–N2	82.32(5)
N1–Cr–N2	84.99(5)	Cr–O1–C2	125.80(11)
Cr–O2–C4	125.34(10)	Cr–O4–C6	118.19(10)
Cr–O6–C11	117.69(10)	Cr–N1–C7	108.19(9)
Cr–N1–C8	106.80(9)	Cr–N2–C9	107.03(9)
Cr–N2–C10	106.96(10)	O1–C2–C1	115.42(15)
O1–C2–C3	124.73(15)	O2–C4–C3	124.99(15)
O2–C4–C5	115.37(15)	O3–C6–O4	124.14(15)
O3–C6–C7	120.84(15)	O4–C6–C7	115.00(13)
N1–C7–C6	112.24(12)	N1–C8–C9	109.51(13)
N2–C9–C8	109.80(13)	N2–C10–C11	111.54(13)
O5–C11–O6	124.66(15)	O5–C11–C10	120.40(16)
O6–C11–C10	114.93(14)		

The Cr-N(edda) bond distances [2.0802(13) and 2.0856 (14) Å] are within the expected range for chromium(III)–N (secondary amine) bonds and agree well with many literature values, e.g. for *cis*-$[Cr(cyclam)(ONO)_2]NO_2$ (cyclam= 1,4,8,11-tetraazacyclotetradecane), $[Cr_2(\mu\text{-}OH)_2(nta)_2]^{2-}$(nta= nitrilotriacetate), *cis*-β-$[Cr(2,2,3\text{-}tet)Cl_2]ClO_4$ (2,2,3-tet= 1,4,7,11-tetraazaundecane), *trans*-$[Cr(Me_2tn)_2Cl_2]Cl$ (Me_2tn= 2,2-dimethylpropane-1,3-diamine), *trans*-$[Cr(3,2,3\text{-}tet)F(H_2O)](ClO_4)_2$ (3,2,3-tet=1,5,8,12-tetraazatetradecane) and *trans*-$[Cr(15aneN_4)F_2]ClO_4$ ($15aneN_4$=1,4,8,12 tetraazacyclopentadecane) [31–36].

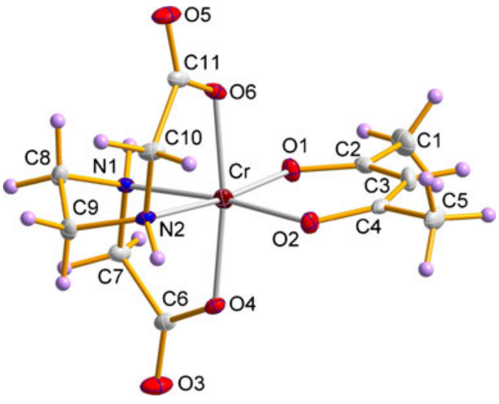


Fig. 1 Molecular structure of the Δ enantiomer of [Cr(edda)(acac)]

The Cr-O bond lengths [1.9623(11) and 1.9733(11) Å] for the carboxylate groups in edda can be compared to the Cr–O distances of 1.959(4), 1.956(4) and 1.9733(11) Å found in $[Cr(cyclam)(ox)]^+$, $[Cr(dpt)(glygly)]^+$ (dpt = di(3-aminopropyl)-amine; glygly=glycylglycinate) and $[Cr(edma)_2]^+$ (edma= ethylenediaminemonoacetate) complexes, respectively [37–39]. The slightly longer Cr-O4 and Cr-O6 bonds involve the atoms O4 and O6 linked to the secondary NH groups of neighbouring molecules by hydrogen bonds. The average length [1.9477(11) Å] for Cr-O(acac) is very close to the values of 1.951(7) and 1.952(3) Å found in $[Cr(acac)_3]$ [40] and [Cr(cyc-*b*)(acac)]$(ClO_4)_2$ [41]. Delocalized C-O and C-C bond lengths for acac are 1.282(2) and 1.395(2) Å, respectively. The internal bond lengths and angles of acac are in good agreement with those of $[Cr(acac)_3]$ and [Cr(cyc-*b*)(acac)]$(ClO_4)_2$ [40, 41]. The five-membered chelate ring of en in edda adopts a *gauche* conformation, with a bite angle at chromium(III) of 84.99(5)°, and the N-C-C-N torsion angle is ±53.12°.

Hydrogen bonding

The crystal structure is supported by hydrogen bonds between secondary NH groups of edda and the carboxylate groups of neighbouring edda molecules. Table 3 contains the distances and angles of hydrogen bonds. This strong intermolecular hydrogen-bonded network enhances the stabilization of the crystal structure and provides an interesting supramolecular organization of the solid [Cr(edda)(acac)], which may be described as follows.

Each complex is connected on one side to the neighbouring [Cr(edda)(acac)] molecule by a pair of equivalent, almost linear N–H···O hydrogen bonds (D···A=2.09 Å, ∠(N–H···O)=175.7°), and on the opposite side to another neighbouring [Cr(edda)(acac)] molecule by a similar pair of nearly

Table 3 Geometry of the hydrogen bonds for [Cr(edda)(acac)]

Hydrogen bond	N–H (Å)	H···O (Å)	N···O (Å)	N–H···O (°)
N(1)–H(1)···O(6)[a]	0.83(2)	2.09(2)	2.9242(18)	175.7(18)
N(2)–H(2)···O(4)[b]	0.85(2)	2.06(2)	2.9124(17)	177.6(17)

[a] [−*x*, 1−*y*, 1−*z*]

[b] [1−*x*, 1−*y*, 1−*z*]

linear, equivalent N–H···O hydrogen bonds (D···A=2.06 Å, ∠(N–H···O)=177.6°). In this way the complex molecules form distinct infinite one-dimensional strands in the direction of the crystallographic *a*-axis.

Since both the donor and acceptor atoms of the N–H···O hydrogen bonds are ligating atoms, each pair of the concomitant equivalent hydrogen bonds together with the corresponding metal atoms forms an eight-membered ring (HB-ring). Consecutive HB-rings thus yield a spiro-type structure with metal atoms as spiro junctions. Denoting the atoms from left and right neighbours with single and double primes, respectively, the HB-rings are: Cr-N1-H1-O6′-Cr′-N1′-H1′-O6 and Cr-N2-H2-O4″-Cr″-N2″-H2″-O4. A structural diagram (Supplementary Fig. S1) in which the acac rings as well as the edda atoms not directly involved in hydrogen bonding are partly hidden, clearly emphasizes the similarity between the two types of HB-rings and their almost regular chair-like conformation due to the linearity of hydrogen bonds. The angle between the least-squares planes of the adjacent HB-rings is about 55°. Its expected deviation from the ideal O_h value, $\cos^{-1}(1/2)$, is a consequence of a slight irregularity of the HB-ring and of the coordination octahedron.

The polymeric structure of a strand of HB-rings involves an alternating sequence of Δ and Λ [Cr(edda)(acac)] molecules (Fig. 2). Each diad in a strand is, therefore, racemic (i.e. there is an inversion centre at the mid-point of each 8-membered HB-ring) causing the strands to be syndiatactic — both with respect to the absolute configuration of [Cr(edda)(acac)] units, and to the chiralities of tertiary nitrogen atoms. The achiral (syndiatactic) structure of the strands is a requirement for the overall centrosymmetric space group ($P2_1/c$) of the [Cr(edda)(acac)] crystal structure.

In addition to these strong hydrogen bonds, which are the predominant packing and stabilizing factor, neighbouring strands interact with each other through a number of weaker van der Waals forces (Supplementary Table S1). Their orientation is transverse with respect to the direction of hydrogen-bonded strands (the crystallographic *a*-axis) as shown in a projection (Supplementary Fig. S2) of the crystal structure on the *bc*-plane. Among the shortest contacts of the latter type are those between the methyl groups of acac chelate rings and the CH or CH_2 groups of the edda chelates (and *vice versa*) from neighbouring molecules in adjacent strands.

All crystallographical evidence (low R factor, absence of voids or of significant residual electron density unaccounted for, and incompatibility of the present hydrogen bond network with any presence of lattice H_2O molecules) strongly indicates the anhydrous nature of [Cr(edda)(acac)] crystal sample studied in this work.

Spectroscopic properties

The UV-visible absorption spectrum exhibits two principal bands, one at 18725 cm^{-1} (ν_1), and the other at 25905 cm^{-1} (ν_2), corresponding to the ${}^4A_{2g}\rightarrow{}^4T_{2g}$ and ${}^4A_{2g}\rightarrow{}^4T_{1g}$ (O_h) transitions, respectively [42]. In order to have some point of reference for the splitting of the bands, we have fitted the band profiles to four Gaussian curves. A deconvolution procedure on the experimental band pattern yielded maxima at 18090, 19510, 25000 and 26590 cm^{-1} for the non-cubic split levels of ${}^4T_{2g}$ and ${}^4T_{1g}$, respectively.

A complete assignment of the molecular vibrations for [Cr(edda)(acac)] is done in connection with the optimization of force field parameters (see below). Therefore, we present here only the principal spectroscopic features that confirm the coordination mode of edda and point to the characteristic group frequencies of the complex molecule. The FT-infrared spectra show a strong band around 1677 cm^{-1} due to the asymmetric ν_a(COO) stretching mode of edda. The symmetric stretching mode of the carboxylate group occurs at 1285 cm^{-1}. The lack of absorptions between 1700–1750 cm^{-1} indicates that the carboxylate groups of edda are certainly coordinated to the central chromium(III) ion. The value of the frequency separation (Δν) between the antisymmetric and symmetric carboxylate stretching vibrations can be used for predicting the coordination mode of the carboxylate group [43]. In unidentate coordination a redistribution of electron density takes place, which shifts the

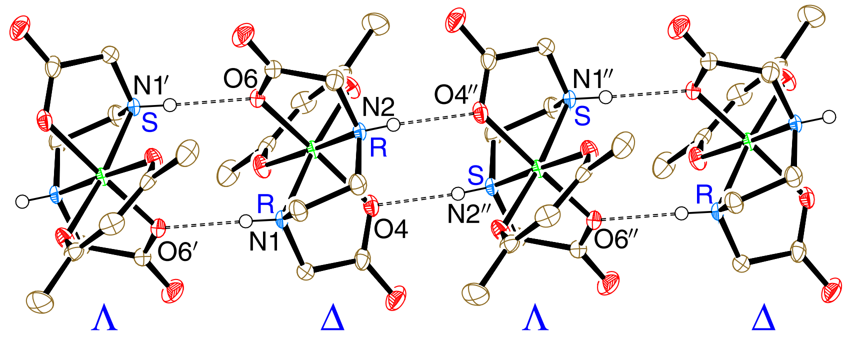


Fig. 2 1D strand of hydrogen-bonded ···Δ-[Cr(R,R-edda)(acac)]··· Λ-[Cr(S,S-edda)(acac)]··· in the crystal structure of [Cr(edda)(acac)]

Table 4 Relative energies (ΔE) and energy contributions (in kcal/mol) for bonds (E_b), angles (E_θ), torsions (E_φ), van der Waals (E_{vdw}), Coulomb (E_c) and out-of-plane (E_{oop}) terms for energy minimized structures of [Cr(edda)(acac)]

Isomer	Conf.	ΔE	E_b	E_θ	E_ϕ	E_{vdw}	E_c	E_{oop}
s-cis	Λ(*S*,*S*;λ)	0.00	7.15	11.25	0.53	15.73	−33.22	0.054
	Λ(*S*,*S*;δ)	3.55	5.69	17.03	0.53	15.29	−33.51	0.019
u-cis	Λ(*R*,R^{mer};λ)	2.05	6.78	14.74	0.64	14.02	−32.73	0.085
	Λ(*R*,S^{mer};δ)	2.47	6.82	14.62	0.53	14.33	−32.36	0.037

asymmetric carboxylate stretch to higher wavenumbers in comparison to ionic carboxylate. Consequently the Δν value for unidentate carboxylate coordination is higher. By contrast, bidentate coordination shifts the position of the antisymmetric carboxylate stretch to lower wavenumbers in comparison to the ionic group and thus lowers the value of Δν. The difference between $\nu_{as}(COO^-)$ and $\nu_s(COO^-)$ of about 392 cm^{-1} is consistent with the unidentate coordination mode for the O-bonded carboxylate group of edda. This coordination behaviour is in accordance with the crystal structure. The strong absorption bands at 1569 and 1528 cm^{-1} are due to ν(C=C) coupled with ν(C=O), and ν(C=O) coupled with ν(C=C), respectively. The absorption band around 750 cm^{-1} is assigned to the C-H out-of-plane bending mode of acac [43]. The two bands at 1461 and 1430 cm^{-1} can be assigned to CH_2 bending modes. The strong absorption at 1069 cm^{-1} may be assigned to a CH_2 twisting mode. The two absorptions at 848 and 830 cm^{-1} and one absorption at 797 cm^{-1} are assigned to CH_2 wagging and en ring deformation frequencies, respectively. The strong absorption at 1023 cm^{-1} may be assigned to a CH_2 twisting mode. The peaks at 502 and 479 cm^{-1} can be assigned to the Cr-N and Cr-O stretching modes, respectively. Finally, the conspicuous absence of stretching O-H and bending H-O-H vibrations from 3550–3200 cm^{-1} and 1630–1600 cm^{-1} region, respectively, is another proof that lattice H_2O molecules are absent in [Cr(edda)(acac)] crystal sample studied in this work. However, the UV-visible and infrared spectroscopic data are not able to give any definite evidence whether the complex [Cr(edda)(acac)] has *s-cis* or *u-cis* geometry of the chelated edda.

Table 5 Gradient norm values (kJ/mol/Å), free enthalpies (kJ/mol) and populations for CFF minimized structures of [Cr(edda)(acac)] at 298.16 K

Isomer	Conf.	$\nabla\times10^7$	G	ΔG	n
s-cis	*Λ*(S,S;λ)	4.5	1074.561	0.000	0.949
	Λ(S,S;δ)	0.46	1088.869	14.308	0.003
u-cis	*Λ*(R,R^{mer};λ)	10.3	1084.483	9.922	0.035
	Λ(R,S^{mer};δ)	0.39	1086.778	12.217	0.014

MM modeling

The stereochemisty of octahedral complexes containing quadridentate $edda^{2-}$ ligand has been described by Legg *et al.* [6], and reiterated by many subsequent authors, notably by Kaizaki *et al.* [44, 45]. However, CFF modelling requires (particularly in elucidating statistical thermodynamics) a consideration of all theoretically possible configurations and conformations, which represent local minima on the potential energy surface of [Cr(edda)(acac)]. To that end we performed a systematic search of the conformational space of [Cr(edda)(acac)] and found four pairs of enantiomeric structures (two for *s-cis* and two for *u-cis* configuration) that satisfy the loop closure constraints for all chelate rings of $edda^{2-}$. Their stereochemical designation independent of atom numbering is as follows. For the *s-cis* diastereoisomer (of C_2 symmetry) the same chirality of two equivalent N ligators is determined by the overall absolute configuration of [Cr(edda)(acac)], or *vice versa*, which turns out to be *S,S* for the Λ (or *R,R* for the Δ). In addition, the en ring may adopt one of two normal *gauche* conformations (δ or λ) without a change of chirality on N atoms—an interesting detail which has not been noted hitherto—leading to two energetically distinct forms, *e.g.*: Λ(*S*,*S*;λ) and Λ(*S*,*S*;δ). The foregoing labels define the overall absolute configuration and (in parantheses) the chirality of N atoms and (after a semicolon) the chirality of the en ring of $edda^{2-}$.

On the other hand, in the *u-cis* diastereoisomer (of C_1 symmetry) two N ligators are nonequivalent: one (N^{fac}) belongs to the facial junction of the coordinated edda, and the other (N^{mer}) to the meridional one. The chirality of the N^{fac} is linked to the overall absolute configuration of [Cr(edda)(acac)] as in the *s-cis* diastereoisomer, but as Λ(*R*) or Δ(*S*), whereas the chirality of N^{mer} is related to the conformation of the en ring, and can be either *S* or *R*. This choice generates two energetically distinct forms, *e.g.*: Λ(R^{fac},R^{mer};λ) and Λ(R^{fac},S^{mer};δ), the superscripts (or at least one of them) being necessary and self-explanatory.

It should be pointed out that the overall absolute configuration both for *s-cis* and for *u-cis* diastereoisomer of [Cr(edda)(acac)] is unambiguously assigned on the basis of the IUPAC rule for tris(bidentate) octahedral complexes

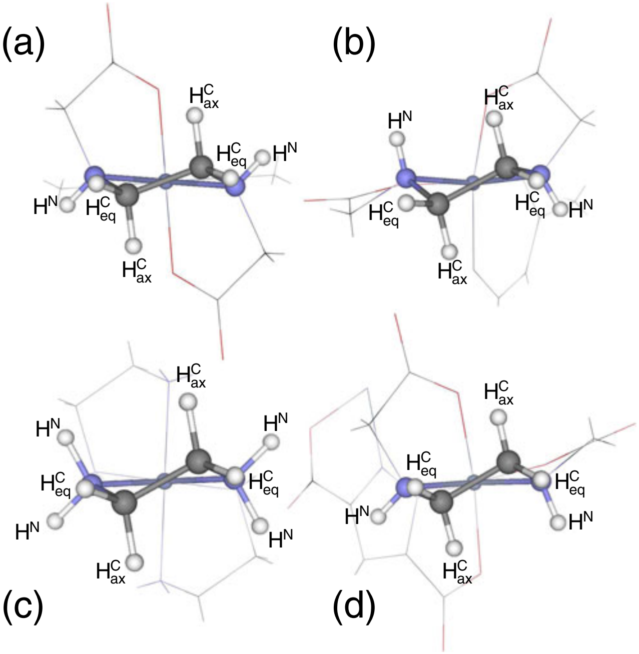


Fig. 3 *Gauche* conformations of the ethylenediamine fragment or chelate ring in: (**a**) *s-cis* and (**b**) *u-cis* isomer of [Cr(edda)(acac)], (**c**) X-ray structure of $[Cr(en)_3]^{3+}$ ion, and (**d**) a fragment of dinuclear $[(edda)Cr(\mu\text{-}pzdc)Cr(edda)]^-$ species. For clarity, only H atoms of en are labelled in a self-explanatory way

(IR-9.3.4.12) taking as reference the helicity of two carboxylato rings of edda together with the acac ring. With respect to the pseudo-C_3 axis defined in this way, the en ring in, e.g. Λ(*S,S;λ*) *s-cis* form is *lel* and in Λ (*S,S;δ*) it is *ob*. However, *lel* and *ob* labels do not apply to the *u-cis* configuration.

Energy minimization and geometry optimization yielded one stable conformation for each of the four aforementioned diastereoisomers (Table 4 and Fig. S5 of the Supporting material). The global minimum corresponds to the Λ(*S*,*S*;λ) [or Δ(*R*,*R*;δ)] configuration of the *s-cis*-[Cr(edda)(acac)] complex (with the *lel* conformation of en), which is the geometry found by the X-ray structure determination (Table S2 of the Supporting material shows a detailed comparison). Our former [11] force field (optimized only on the basis of structural data) yields qualitatively the same result. Therefore, as in all previous cases [11], MM succeeded in correlating the most stable structure of an edta-type complex with the one found experimentally by X-ray diffraction. For stereochemically relatively simple [Cr(edda)(acac)] species this result was hardly surprising: the previous MM calculations on systems containing an edda ligand [1], as well as all reported crystal structures [44, 46–49] (except one, discussed below), confirm the general tendency for the quadridentate edda ligand to adopt preferentially the *s-cis* geometry in coordination to Cr(III). Moreover, there is a remarkable similarity of the edda backbone conformation among all the crystal structures containing a Cr(edda) fragment [44, 46] and our [Cr(edda)(acac)] structure, as illustrated in Fig. S4 (of the Supporting material) showing a comparison of all endocyclic torsion angles involving non-hydrogen atoms of the three fused chelate rings of Cr(edda) fragment.

Structural details of geometry-optimized *s-cis* and *u-cis* isomers of the title complex follow the same pattern as in the previous MM investigations of edta-type complexes [11]. Thus, energy minimization produces an *s-cis* structure with exact C_2 symmetry. In both *s-cis* and *u-cis* isomers the en ring adopts the energetically preferred twist-boat (*gauche*) conformation, similar to the one found in isolated M(en) chelate rings in, *e.g.* $[M(en)_3]^{3+}$ structures (with M= trivalent metal of the first transition series). The central en ring is in fact the fragment of the edda backbone showing the least amount of conformational variation among different Cr(edda)-containing structures (cf. torsion angles in Fig. S4 of the Supporting material).

Isomer distribution

Free enthalphy was calculated using standard formulae of statistical thermodynamics [50] and previously described procedures [51, 52]. Averaging over all internal degrees of freedom was carried out at 298.16 K. External motion was quenched. Conformer population was obtained from Boltzmann distribution. The degeneracy factor for the *u-cis* form (point group C_1) was taken as twice that of the *s-cis* form (point group C_2). The results (Table 5) show an isomer distribution consisting of approximately 95% *s-cis* and 5% *u-cis* form of [Cr(edda)(acac)]. Quantitative experimental data are not available, apart from the ^{2}H NMR study of Bianchini and Legg [3] on the *u-cis* to *s-cis* isomerization of the closely related $[Cr(edda)(H_2O)_2]^+$ complex, in which an *s-cis*:*u-cis* ratio of 0.8:0.2 was reported. Table 5 also gives the final gradient norm values,

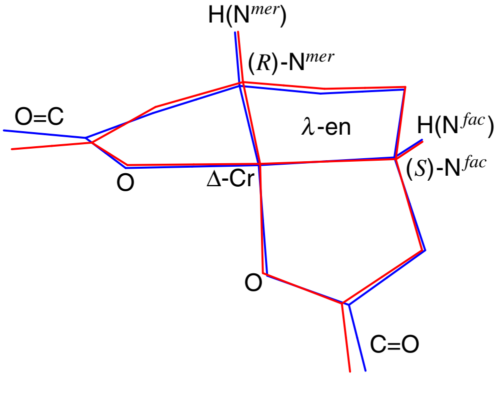


Fig. 4 Comparison between Cr(edda) fragment in VOFF optimized Δ(S^{fac},R^{mer},λ) structure of *u-cis*-[Cr(edda)(acac)] (blue) and the the same moiety in the crystal structure of $[(edda)Cr(\mu\text{-}pzdc)Cr(edda)]^-$ [44] (red)

which together with the positive definite Hessian and the absence of negative vibrational frequencies confirm the true minima in all cases.

The case of *u-cis* isomer and VOFF

The necessity for parameter optimization against experimental vibrational frequencies in the quest for more physically accurate MM force field is convincingly shown in the case of *u-cis* diastereoisomer of [Cr(edda)(acac)]. Our prevoius force field [11] predicted that *u-cis* would be less stable than the *s-cis* form (as in this work) but with a substantially greater energy difference than the one shown in Table 4. The strain was essentially localized on the central en ring, which adopted an unsymmetrical conformation (resembling envelope flattened along the Cr-N bond shared with the axial carboxylate ring, and with the apex on the opposite methylene carbon) due to the meridional coordination of the equatorial carboxylate ring. By contrast, the present VOFF produced a much less strained structure of the *u-cis*

Table 6 Calculated and observed vibrational frequencies for *s-cis*-[Cr(edda)(acac)]

#	Calc.	Obs.	Diff.	Dominant assignment [PED, %]
1	3139	3131	−8	ss(NH)
2	3139		−8	as(NH)
3	3104	2955	−149	s(CH)^{m}[90]
4	2932	2916	−16	as(CH)^{e}
5	2925	2914	−11	as(CH)^{g}[84],as(CH)^{e}[16]
6	2924		−10	as(CH)^{g}
7	2920	2911	−9	as(CH)^{e}[80],as(CH)^{g}[14]
8	2914		−3	as(CH)^{Me}[92]
9	2914		−3	as(CH)^{Me}[92]
10	2911		0	as(CH)^{Me}
11	2911		0	as(CH)^{Me}
12	2891	2850	−41	ss(CH)^{e}
13	2864	2837	−27	as(CH)^{g}[64],as(CH)^{e}[36]
14	2862		−25	ss(CH)^{g}
15	2858		−21	as(CH)^{e}[64],as(CH)^{g}[36]
16	2807	2798	−9	ss(CH)^{Me}
17	2807		−9	as(CH)^{Me}
18	1716	1687	−29	as(CO)^{g}[24],ad(CCO)^{g}[24]
19	1716	1677	−39	as(CO)^{g}[24],ad(CCO)^{g}[24]
20	1674	1640	−34	ad(CNH)^{g,e}[56],ad(NCH)^{g,e}[24]
21	1671		−31	ad(CNH)^{g,e}[52],ad(NCH)^{g,e}[22]
22	1579	1571	−8	as(CC)^{a}[22],as(CO)^{a}[12],ad(CCH)^{m}[22],ad(CCO)^{a}[8]
23	1453	1525	+68	ss(CO)^{a}[10],ss(CC)^{a}[8],ad(CCC)^{a}[5],t(CC)^{Me}[24]
24	1426	1461	+35	ab(HCH)^{e,g}[32],ad(NCH)^{e,g}[52]
25	1423	1429	+6	ab(HCH)^{e,g}[38],ad(NCH)^{e,g}[54],
26	1366	1381	+15	ad(HCH)^{Me}[44],t(CC)^{Me}[54]
27	1364		+17	sd(HCH)^{Me}[44],t(CC)^{Me}[54]
28	1344	1364	+20	ad(HCH)^{Me}[38],t(CC)^{Me}[60]
29	1341		+23	ad(HCH)^{Me}[40],t(CC)^{Me}[64]
30	1321	1320	−1	ad(CCH)^{g,e}[22],ad(HCH)^{g,e}[22],ad(NCH)^{g,e}[40]
31	1304	1303	−1	ad(NCH)^{g,e}[42],ad(HCH)^{g,e}[30]
32	1275		+28	sd(NCH)^{g,e}[40],sd(HCH)^{e}[20]
33	1258		+45	ad(NCH)^{g,e}[34],ad(HCH)^{e}[20]
34	1229	1283	+54	as(CO)^{g}[26],ad(CCO)^{g}[18],ad(OCO)^{g}[18],ad(HCH)^{g}[10]
35	1228		+55	ss(CO)^{g}[26],sd(CCO)^{g}[18],sd(OCO)^{g}[18],sd(HCH)^{g}[10]
36	1207	1224	+17	as(CCH)^{Me}[22],ad(CCH)^{m}[24],ad(HCH)^{Me}[28]

Table 6 (continued)

#	Calc.	Obs.	Diff.	Dominant assignment [PED, %]
37	1184	1199	+15	sd(CCH)^{Me}[48],sd(HCH)^{Me}[50]
38	1182		+17	ad(HCH)^{Me}[44],ad(CCH)^{Me}[40],ad(CCH)^{m}[14]
39	1164	1143	+21	ad(CrNH)^{e}[30],ad(CNH)^{e,g}[26],ad(NCH)^{e}[24]
40	1158		+15	sd(CrNH)^{e}[32],sd(CNH)^{e,g}[30],sd(NCH)^{e,g}[20]
41	1148	1120	−28	sd(CCH)^{e}[34],sd(HCH)^{e}[32],sd(NCH)^{g,e}[22],s(CC)^{e}[5]
42	1144		−24	sd(NCH)^{g,e}[36],sd(CNH)^{g,e}[26],sd(HCH)^{g}[14]
43	1130		−10	sd(CNH)^{e,g}[28],sd(NCH)^{g,e}[36],sd(HCH)^{g,e}[18]
44	1067	1067	0	tw(HCH)^{g}[30],ad(CCH)^{g}[42]
45	1066		+1	sd(HCH)^{g}[30],sd(CCH)^{g}[42]
46	1019	1021	+2	tw(HCH)^{Me},ad(CCH)^{m}[64]
47	1001	988	−2	as(CN)^{g,e}[16],ad(CCH)^{e}[20],ad(NCH)^{e,g}[34]
48	999			sd(acac)
49	996	988	+2	as(CN)^{g,e}[18],ad(CCH)^{e}[18],ad(NCH)^{e,g}[38],as(HCH)^{e}[16]
50	975	975	0	s(CC)^{e}[5],ss(CN)^{e,g}[20],ad(CCH)^{e}[10],ad(NCH)^{g,e}[10]
51	952			ad(CCH)^{e}[44],ad(HCH)^{e}[24],ad(NCH)^{e}[22]
52	936	934	−2	as(NC)^{e}[10],ad(CCH)^{e}[12],aad(NCH)^{e}[10],ad(CCH)^{g}[18]
53	885	909	−24	ad(CCH)^{Me}[58]
54	877		−32	ad(CCH)^{Me}[50],ad(CCH)^{m}[10],t(acac)
55	851	849	−2	ad(CCH)^{Me}[34],ad(CCH)^{m}[6],t(acac)
56	837	830	−7	ad(CCH)^{Me}[56],t(acac)
57	809	797	−12	ss(NC)^{e}[8],s(CC)^{e}[8],ad(CCH)^{g}[28],ad(CCH)^{e}[18]
58	775	759	−16	as(CrO)^{a}[8],ad(CrOC)^{a}[14],ad(OCC)^{a}[10],ad(OCC)^{Me}[12]
59	754	743	−11	oop(CCH)^{m},t(Me groups)
60	735	737	+2	sd(CrOC)^{a}[16],sd(OCC)^{a}[24],sd(CCC)^{a}[14],sd(CCC)^{Me}[16],sd(CCH)^{m}[14]
61	716	693	−23	ad(CCH)^{g}[42],ad(NCH)^{g}[14],ad(CCH)^{e}[6],t(CC)^{g}[14]
62	713		−20	t(Me groups)
63	704		−11	sd(CCH)^{g}[44],sd(NCH)^{g}[16]
64	669	672	+3	ad(CCH)^{Me}[8],t(acac)
65	663	662	−1	ad(CCH)^{e}[76],ad(NCH)^{e}[14],ad(CCH)^{g}[10]
66	628	629	+1	as(CrO)^{g}[18],as(CC)^{g}[12],ad(OCO)^{g}[26],as(CCO)^{g}[26]
67	618	612	−6	as(CrO)^{g}[6],as(CC)^{g}[8],ad(OCO)^{g}[16],as(CCH)^{e}[20],as(CCO)^{g}[14]
68	594	597	+3	as(CrO)^{g}[6],ad(CrOC)^{g}[10],ad(CCH)^{e}[12],ad(CCO)^{g}[18]
69	590	575	−15	ss(CrO)^{g}[10],sd(CrOC)^{g}[14],ad(CCH)^{e}[26],ad(CCO)^{g}[14]
70	549			t(Me groups)[54],t(acac)[44]
71	545	545	0	ad(CCH)^{e}[66],t(CC)^{e}[10]
72	530	531	+1	ss(CrO)^{a}[10],ss(CC)^{Me}[10],(CCH)^{e}[18],d(CCC)^{a}[5],t(CO)^{a}[10]
73	517	518	+1	as(CrN)^{e}[10],as(CC)^{g}[8],as(CC)^{Me}[6]
74	502	502	0	ss(CrN)^{e}[6]

isomer, with clearly discernible equatorial and axial hydrogens on both methylene carbons of the en ring having an almost normal *gauche* conformation. Such a conformation of the Cr(edda) moiety was indeed found in one of the crystallographic reports [44] for a dinuclear chromate(III) complex [(edda)Cr(μ-pzdc)Cr(edda)]$^{-}$ (where pzdc=pyrazole-3,5-dicarboxylate bridge) in which one of the two Cr(edda) units was found to be trapped in a less common (and considered as thermodynamically unstable [44]) *u-cis* configuration. It occurs actually in one out of nine crystallographically unique Cr(edda) fragments among all the structures retrieved from the CSD up to September 2009.

This result leads to two conclusions. First, the *u-cis* coordination of edda is more accessible (ΔE≈2 kcal/mol,

Table 6 (continued)

#	Calc.	Obs.	Diff.	Dominant assignment [PED, %]
75	485	479	−6	as(CrO)^{a}[14],as(CC)^{Me}[20]
76	474	467	−7	t(edda)[66],oop(COO)^{g}[10]
77	473		−6	t(edda)[74],oop(COO)^{g}[12]
78	431	439	+8	(all skeletal def.)
79	416		+23	(all skeletal def.)
80	414	417	+3	as(CC)^{g}[6],as(CC)^{Me}[8]
81	391	393	+2	as(CrO)^{g}[10],as(CC)^{g}[16],as(CC)^{Me}[10],as(CrN)^{e}[10],ad(NCC)^{e}[12],ad(CNCr)^{e}[10]
82	388			oop(CCC)^{m}[26],t(CC)^{Me}[34],t(acac)[42]
83	375	364	−11	ss(CrO)^{g}[6],ss(CC)^{g}[20],ss(CrN)^{e}[8]
84	326	343		ss(CrO)^{a}[6],ss(CC)^{Me}[10],ss(CC)^{g}[6]
85	316			as(CrO)^{g}[8],as(CC)^{g}[8]
86	306			ss(CC)^{Me}[14],d(CCC)^{a}[12]
87	276	279	+3	ss(CrO)^{a}[6],sd(OCrO)^{g}[6],sd(CCO)^{a}[14]
88	272		+7	as(CrO)^{a}[8],as(CC)^{Me}[6]
89	258			ss(CrO)^{g}[8],sd(OCO)^{g}[8],sd(OCC)^{g}[6]
90	256			ad(NCC)^{e}[22],t(NC)^{e}[14],t(NC)^{g}[6]
91	232	241	+9	t(NC)^{e}[14],t(CC)^{e}[10],t(CC)^{g}[22],t(CO)^{g}[20]
92	212			as(CrO)^{a}[4],as(CO)^{g}[18]
93	184	170	−14	ad(OCrO)^{a,g}[10],t(OC)^{g}[28]+(skeletal def.)
94	176		−6	ad(OCrO)^{a,g}[6],d(OCrO)^{g}[3],d(NCrN)^{e}[4],t(OC)^{g}[30]
95	170		0	ad(OCrO)^{a,g}[10],ad(OCrN)^{a,e}[6],t(CrO)^{a}[24],t(CO)^{a}[34]
96	166		+4	ss(CrO)^{g}[6],sd(OCrO)^{g}[7],t(CO)^{g}[30]
97	129	118	−11	(edda skeletal def.)
98	117	102	−15	ad(OCrO)^{a,g}[20],ad(OCrN)^{a,e}[8],t(CO)^{a}[48]
99	111		−9	ad(OCrO)^{a,g}[8],ad(OCrN)^{g,e}[10],ad(NCC)^{e}[8],t(NC)^{g}[20],t(CC)^{g}[26]
100	77	77	0	sd(OCrO)^{a,g}[18],d(OCrO)^{g}[13],sd(OCrN)^{g}[14]
101	64			ad(OCrO)^{a,g}[10],ad(OCrN)^{a,e}[90]
102	52			ad(OCrO)^{a,g}[10],ad(OCrN)^{g,e}[10],ad(OCrN)^{a,e}[6],t(CrN)^{e}[20],t(NC)^{e}[16]
103	35			t(acac)
104	30			sd(OCrO)^{a,g}[38]
105	22			t(CC)^{a}[36]

Abbreviations: s = stretch, b = bend, t = torsion, oop = out-of-plane bend, r = rock, w = wag, tw = twist, ss = symmetric stretch, as = antisymmwetric stretch, sd = symmetric deformation, ad = antisymmetric deformation. Superscripts: e = ethylenediamine fragment of edda, g = carboxylato fragments of edda, a = acac, m = acac ring CH, Me = acac methyl groups.

cf. Table 4) than what has been usually assumed, so that in the case of [(edda)Cr(μ-pzdc)Cr(edda)]$^{-}$ only one intramolecular hydrogen bond is presumably [44] sufficient to sustain it. VOFF is thus able to offer a more physically realistic description of this stereochemistry both in terms of energetics and in a more even distribution of steric strain in unfavourable meridionally fused five-membered metal-chelate rings. Second, the interaction forces that shape the ethylenediamine ring (either isolated or fused in multidentate structures) give rise dominantly to the *gauche* (twist-boat) conformations with distinct axial and equatorial positions on methylene carbons. The en ring conformation is thus essentially the same in *s-cis* and *u-cis* isomer of [Cr(edda)(acac)], as well as in the X-ray structure of $[Cr(en)_3]^{3+}$ ion [53], and in the above mentioned fragment of dinuclear $[(edda)Cr(\mu\text{-}pzdc)Cr(edda)]^-$ structure, as shown in Fig. 3, and also in other chromium(III) complexes for which X-ray data are available.

A final substantiation that the VOFF is physically realistic is provided again by the X-ray structure of $[(edda)Cr(\mu\text{-}pzdc)Cr(edda)]^-$ [44]. Having identified equilibrium geometries of all diastereoisomers of [Cr(edda)(acac)] (see Tables 4 and 5) it was straightforward to pinpoint the $\Delta(R,R,\delta)$ [or $\Lambda(S,S,\lambda)$] configuration for the *s-cis* Cr(edda) moiety, and the $\Delta(S^{fac},R^{mer},\lambda)$ [or $\Lambda(R^{fac},S^{mer},\delta)$] configuration for the *u-cis* Cr(edda) moiety of $[(edda)Cr(\mu\text{-}pzdc)Cr(edda)]^-$[44]. (In the original paper [44] they were designated as Cr(1) and Cr(2), respectively, but the chiralities of N atoms were incorrectly assigned.) Therefore, the observed *u-cis* structure of the Cr(edda) fragment in the binuclear species appears to be entrapped both in a less favourable ligand configuration and in a less favourable conformation of its en ring or, equivalently, the chirality of the *mer* nitrogen. The similarity between thermodynamically less stable $\Delta(S^{fac},R^{mer},\lambda)$ structure calculated with VOFF and the one observed in the crystal structure of $[(edda)Cr(\mu\text{-}pzdc)Cr(edda)]^-$ [44] (Fig. 4) is noteworthy, bearing in mind that CFF modelling did not include any of the conformationally responsive factors (intramolecular hydrogen bonding, nonbonded interactions) present in the binuclear species.

Vibrational frequencies and transferability of VOFF

The 37-atom *s-cis* isomer of [Cr(edda)(acac)] has C_2 symmetry and 105 fundamental vibrations, classified as $\Gamma^{vib} = 52A \oplus 53B$, with all the modes being IR active. Experimental vibrational frequencies were used in the parameter optimization of the VOFF. Final calculated frequencies using VOFF, with band assignments (confirmed by visualization using the program MOLEKEL [54, 55]) and PED's [56], are compared to the experimental values in Table 6. The agreement for 92 recorded peak positions is characterized by the sample correlation coefficient of 0.999658 for a linear function f(x)=x fitting (Fig. S6 of the Supporting material). Overall RMS deviation is 24.8 cm^{-1} (for N=92), which after removal of five outliers reduces to 15.5 cm^{-1} (for N=87). Most of the differences ($\Delta\nu$) are within an acceptable range and distributed as shown in a frequency histogram (Fig. 5) on which a hypothetical normal distribution curve is superimposed.

Band assignments using VOFF generally agree with those reported recently for $[Cr(acac)_3]$ [57–59] and related edta-type complexes of chromium(III) [60]. Finally, vibrational frequencies calculated for $[Cr(acac)_3]$ using the present VOFF are in accord with the published values [57], which offers a good prospect for the transferability of VOFF and a

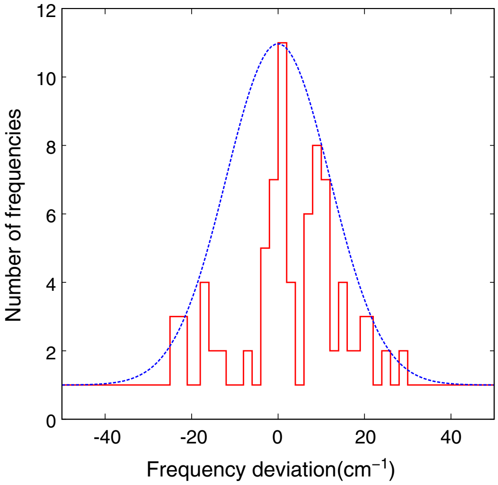


Fig. 5 Histogram showing the distribution of differences between the calculated and all experimental vibrational frequencies (using bin-width of 2.0 cm^{-1}) with a normal distribution curve (dashed line)

convenient tool to study normal modes of different conformations that may not be experimentally accessible.

Conclusions

The X-ray crystallography for [Cr(edda)(acac)] shows that the chromium(III) ion is in a distorted octahedral environment, coordinated by two nitrogen and two oxygen atoms of edda, and two oxygen atoms of acac in *cis* positions. The average Cr-N(edda), Cr-O(edda) and Cr-O(acac) bond distances are 2.0829(14), 1.9678(11) and 1.9477(11) Å, respectively. The complex has an *s-cis* configuration, and is stabilized in the solid state by significant intermolecular N–H···O hydrogen bonds forming infinite chains of [Cr(edda)(acac)] molecules. An improved force field, Vibrationally Optimized Force Field (VOFF), was developed by fitting the FF parameters to structural data and experimental vibrational frequencies of [Cr(edda)(acac)], which in addition to reproducing structures and energetics of edta-type complexes of chromium(III) is able to reproduce observed vibrational frequencies in a consistent way. The ability of VOFF to account more realistically for fine conformational details, as well as its probable transferability was demonstrated, notably in the case of a binuclear Cr(III) complex containing *s-cis* and *u-cis* configuration of coordinated edda, the stereochemistry of which was revisited and clarified.

Supplementary material

Full crystallographic data have been deposited with the Cambridge Crystallographic Data Centre, CCDC No. 719577. Copies of this information may be obtained free of charge from The Director, CCDC, 12 Union Road, Cambridge CB21EZ, UK (e-mail: deposit@ccdc.cam.ac.ukor fax: +44-1223-336-033 or url: http://www.ccdc.cam.ac.uk). Supporting material (pp. 11) contains information on hydrogen bonding, intermolecular contacts, comparison of calculated and experimental geometry of [Cr(edda)(acac)] and Cr(edda) fragment, torsional angles of the Cr(edda) fragment from CSD, stereoscopic representation of VOFF equilibrium structures of all four diastereoisomers of [Cr(edda)(acac)], and the correlation coefficient for the fitting of vibrational frequencies.

Acknowledgements IDj and SRN acknowledge financial support from the Serbian Ministry for Science through Grant No. 142017Г.

References

1. Grubišić S, Gruden-Pavlović M, Radanović DD, Perić M, Niketić SR (2009) J Mol Struct 919:54
2. Radanović DJ (1984) Coord Chem Rev 54:159, and references cited therein
3. Bianchini RJ, Legg JI (1986) Inorg Chem 25:3263
4. Fujii Y, Kyuno E, Tsuchina R (1969) Bull Chem Soc Jpn 42:1301
5. Guarddalabene J, Gulnac S, Keder N, Shepherd RE (1979) Inorg Chem 18:22
6. Halloran LJ, Caputo RE, Willett RD, Legg JI (1975) Inorg Chem 14:1762
7. Wheeler D, Kaizaki S, Legg JI (1982) Inorg Chem 21:3248
8. Choi JH, Hong YP, Park YC (2002) Spectrochim Acta 58A:1599
9. Hay PW, Tarafder MT (1991) J Chem Soc Dalton Trans, p 823
10. Stearns DM, Armstrong WH (1992) Inorg Chem 31:5178
11. Grubišić S, Niketić SR, Radanović DD, Rychlewska U, Warzajtis B (2005) Polyhedron 24:1701
12. Grubišić S, Radanović DD, Rychlewska U, Warzajtis B, Drašković NS, Djuran MI, Niketić SR (2007) Polyhedron 26:3437
13. Grubišić S, Gruden-Pavlović M, Niketić SR, Sakagami-Yoshida N, Kaizaki S (2007) J Coord Chem 60:851
14. Grubišić S, Gruden-Pavlović M, Niketić SR, Kaizaki S, Sakagami-Yoshida N (2003) Inorg Chem Commun 6:1180
15. Grubišić S, Gruden M, Niketić SR, Sakagami-Yoshida N, Kaizaki S (2002) J Mol Struct 609:1
16. Lifson S, Warshel A (1968) J Chem Phys 49:5116
17. Warshel A, Lifson S (1970) J Chem Phys 53:582
18. Warshel A, Levitt M, Lifson S (1970) J Mol Spectrosc 33:84
19. Ermer O, Lifson S (1973) J Am Chem Soc 95:4121
20. Niketić SR, Rasmussen K (1977) The consistent force field: a documentation. Lecture notes in chemistry, Vol. 3. Springer, Berlin
21. Hatfield WE, MacDougall JJ, Shepherd RE (1981) Inorg Chem 20:4216
22. *COLLECT.* Delft, The Netherlands (1998)
23. Duisenberg AJM, Kroon-Batenburg LMJ, Schreurs AMM (2003) J Appl Crystallogr 36:220
24. Sheldrick GM (2003) SADABS. University of Göttingen, Germany
25. Sheldrick GM (2008) Acta Crystallogr A64:112
26. Brandenburg K, Putz H (2004) DIAMOND (version 3). Crystal Impact GbR, Bonn, Germany
27. Farrugia LJ (1997) J Appl Crystallogr 30:565
28. See, e.g. a web page by H. S. Rzepa, *Is Co(acac)$_3$ aromatic?*, 2005, http://www.ch.ic.ac.uk/local/organic/tutorial/rzepa9/
29. Morino Y, Shimanouchi T (1978) Pure Appl Chem 50:1707
30. F. Neese, *ORCA – An ab initio, Density Functional, and Semiempirical Program Package, v2.6*. Institut für Physikalische und Theoretische Chemie, Universität Bonn, Germany (2008). Single-point UKS-DFT/B3LYP calculations on the geometries obtained by molecular mechanics (MM) geometry optimization in a self-consistent way
31. Choi JH, Oh IG, Lim WT, Park KM (2004) Acta Crystallogr C60: m238
32. Choi JH, Suzuki T, Kaizaki S (2003) Acta Crystallogr E59: m812
33. Choi JH, Choi SY, Hong YP, Ko SO, Ryoo KS, Lee SH, Park YC (2008) Spectrochim Acta 70A:619
34. Choi JH, Clegg W, Nichol GS, Lee SH, Park YC, Habibi MH (2007) Spectrochim Acta 68A:796
35. Choi JH, Lee U (2008) Acta Crystallogr E64:m1186
36. Choi JH, Oh IG, Ryoo KS, Lim WT, Park YC, Habibi MH (2006) Spectrochim Acta 65A:1138
37. Choi JH, Oh IG, Suzuki T, Kaizaki S (2004) J Mol Struct 694:39
38. Choi JH, Suh IH, Kwak SH (1995) Acta Crystallogr C51:1745
39. Choi JH, Suzuki T, Subhan MA, Kaizaki S, Park YC (2002) Acta Crystallogr C58:m409
40. Morosin B (1965) Acta Crystallogr 19:131
41. Byun JC, Han CH (2005) Bull Korean Chem Soc 26:1395
42. Lever ABP (1984) Inorganic electronic spectroscopy, 2nd edn. Elsevier, Amsterdam
43. Nakamoto K (1997) Infrared and Raman spectra of inorganic and coordination compounds, 5th edn. Wiley, New York
44. Sakagami N, Nakahanada M, Ino K, Hioki A, Kaizaki S (1996) Inorg Chem 35:683 [ZOTPIH]
45. Sakagami-Yoshida N, Teramoto M, Hioki A, Fuyuhiro A, Kaizaki S (2000) Inorg Chem 39(5717):10
46. Srdanov G, Herak R, Radanović DJ, Veselinović DS (1980) Inorg Chim Acta 38:37 [EDDACR]
47. Green CA, Place H, Willett RD, Legg JI (1986) Inorg Chem 25:4672 [FAWTOM]
48. T. Yonemura, R. Nakayama, N. Sakagami, K.I. Okamoto, T. Ama, H. Kawaguchi, T. Yasui, Chem. Lett. p. 215 (1998). [HIWSEL]
49. Zabel M, Poznyak AL, Pavlovsky VI (2007) Zh Strukt Khim 48:747 [PITJUY, PITKAF]
50. Janz GJ (1967) Thermodynamic properties of organic compounds. Academic, New York
51. Hald NCP, Rasmussen K (1978) Acta Chem Scand A32:879
52. Niketić SR, Rasmussen K (1981) Acta Chem Scand A35:623
53. Raymond KN, Ibers JA (1968) Inorg Chem 7:2333
54. Flükiger P, Lüthi HP, Portmann S, Weber J (2000) MOLEKEL *4.3*. Mano (2000–2002)
55. Portmann S, Lüthi HP (2000) Chimia 54:766
56. Califano S (1976) Vibrational states. Wiley, London, p 235
57. Diaz-Acosta I, Baker J, Cordes W, Pulay P (2001) J Phys Chem A 105:238
58. Slabzhennikov SN, Ryabchenko OB, Kuarton LA (2008) Russ J Coord Chem 34:551
59. Sato H, Taniguchi T, Nakahashi A, Monde K, Yamagishi A (2007) Inorg Chem 46:6755
60. Radanović DJ, Ristanović VM, Stojčeva-Radovanović B, Todorovska AD, Sakagami N, Iino A, Kaizaki S (1999) Transition Met Chem 24:403

J Mol Model (2012) 18:2147–2152
DOI 10.1007/s00894-011-1235-9

ORIGINAL PAPER

Investigating the electronic properties of silicon nanosheets by first-principles calculations

Ernesto Chigo Anota · Alejandro Bautista Hernández · Miguel Castro · Gregorio Hernández Cocoletzi

Received: 21 June 2011 / Accepted: 26 August 2011 / Published online: 25 September 2011

Abstract Using first-principles total energy calculations within the density functional theory (DFT), we investigated the electronic and structural properties of graphene-like silicon sheets. Our studies were performed using the LSDA (PWC) and GGS (PBE) approaches. Two configurations were explored: one corresponding to a defect-free layer (h-Si), and the other to a layer with defects (d-Si), both of which were in the armchair geometry. These sheets contained clusters of the C_nH_m type. We also investigated the effects of doping with group IV-A elements. Structural stability was studied by only considering positive vibration frequencies. Results showed that both h-Si and d-Si present a corrugated structure with concavity. h-Si sheets were found to be ionic (D.M.=0.33 Debye) with an energy gap (HOMO–LUMO) of 0.77 eV in the LSDA theory and 0.76 eV in the GGS approach, while d-Si sheets were observed to be covalent (D.M.=2.78 D), and exhibited semimetallic electronic behavior (HOMO–LUMO gap= 0.32 eV within the LSDA theory and 0.33 eV within the GGS approach). d-Si sheets doped with one carbon or one germanium preserved the polarity of the undoped d-Si sheets, as well as their semimetallic electronic behavior. However, when the sheets were doped with two C or two Ge atoms, or with one of each atom (to give $Si_{52}CGeH_{18}$), they retained the semimetallic behavior, but they changed from having ionic character to covalent character.

Keywords Silicon · Isocoronene · Cluster C_nH_m · DFT theory

Introduction

In recent years, the electronic and structural properties of graphene layers and graphene-like layers have attracted a great deal of interest from researchers. Hexagonal graphene sheets were isolated in 2004 [1], and silicon organic nanosheets ($Si_6H_4Ph_2$; Fig. 1) were recently synthesized [2]. On the other hand, silicon nanosheets with a thickness of 1 nm were produced by Harada et al. [3] using the exfoliation method, and $Si_6H_3(OH)_3$ nanosheets with a thickness of 0.7 nm have been produced by Nakano [4]. Recent theoretical studies by Ciraci et al. [5] have predicted the existence of hexagonal silicon sheets with a nonplanar (corrugated) configuration that exhibit semimetallic behavior. Noting the experimental and theoretical studies described above, we investigated the electronic and structural properties of hexagonal silicon nanosheets (h-Si) and defect-modified silicon nanosheets (d-Si). The results of that work are discussed in this paper. Our first-principles total energy calculations were performed using density functional theory.

E. Chigo Anota (✉)
Cuerpo Académico de Ingeniería en Materiales,
Facultad de Ingeniería Química, Benemérita Universidad
Autónoma de Puebla,
C.U. San Manuel,
C. P. 72570 Puebla, México
e-mail: ernesto.chigo@correo.buap.mx

A. Bautista Hernández
Facultad de Ingeniería, Benemérita Universidad Autónoma
de Puebla,
Apartado Postal J-39, Puebla 72570, México

M. Castro
Departamento de Física y Química Teórica, DEPg-Facultad
de Química, Universidad Nacional Autónoma de México,
México D.F., C.P. 04510, México

G. Hernández Cocoletzi
Benemérita Universidad Autónoma de Puebla,
Instituto de Física "Luis Rivera Terrazas",
Apartado Postal J-48, Puebla 72570, México

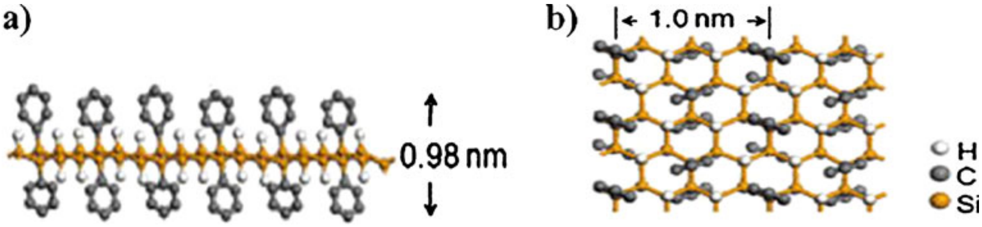


Fig. 1 The model used to study organic silicon nanosheets with the following chemical composition: $Si_6H_4Ph_2$

We used a C_nH_m-type cluster model to investigate the structure of h-Si, which has the chemical composition $Si_{54}H_{18}$ and contains 19 hexagons, and the isocoronene model [6] to study the d-Si configuration, which has the chemical composition $Si_{54}H_{18}$ and comprises three pentagons, three heptagons, and 13 hexagons. The armchair model was used to construct the structures of both h-Si and d-Si [7–9].

Our work was inspired by the experimental and theoretical studies of graphene with defects reported in [10–12]. In that work, samples were prepared by defect engineering using ion irradiation [13].

Methodology

To study the electronic and molecular structures of the hexagonal silicon sheets, we applied the circular C_nH_m model, which has been widely used to represent graphene sheets [14], group III nitrogen alloys [15–17] and group IV-A carbon alloys [18–20]. The isocoronene model [6, 21, 22] (consisting of one central hexagon, three pentagons, three heptagons, and 13 hexagons) was invoked for IV-A carbon alloys with defects. The calculation method we used has already been described in a previous work on first-principles total energy calculations employing density functional theory (DFT) [23, 24] performed using the DMOL^3 code [25]. The exchange-correlation energies were treated using the local spin density approximation (LSDA) with Perdew–Wang parameterization [26] and the generalized gradient spin approximation with Perdew–Burke–Ernzerhof [27] parameterization. A base set of atomic orbitals with double polarization, DNP, which included the *p* orbitals of hydrogen and the *d* orbitals of silicon was employed [28, 29]. We used a multiplicity of 1 for neutral systems and a multiplicity of 2 for doped systems. All-electron calculations with a tolerance of 10^{-6} Ha were performed to achieve convergence of the total energy. The circular structure free of defects had a diameter of 2.18 nm and the structure with defects had a base width of 2.18 nm and a height of 2.06 nm. The orbital (base function) had a cutoff radius of 0.4 nm with a tolerance of 10^{-6} Ha for total energy convergence. Structural stability was achieved using the non-negative vibration frequency criterion [30]; the optimized configurations are therefore the lowest local minima. On the other hand, the binding energies for the silicon nanosheets were obtained from the formula $E_b = (nE_{Si} + mE_H) - (E_{Si54H18} + E_{ZPE})$, where n is the number of silicon atoms and m is the number of hydrogen atoms.

To validate the model, we calculated the cohesive energies $[E_C = [nE(Si) + mE(H) - E(Si_{54}H_{18})]/(n+m)]$ of several structures with different sizes. Results showed an energy difference of 0.003 a.u. between the two structural configurations, so we believe that the nanosheets are well represented by the considered models.

Results and discussion

Geometry optimization, polarity, and gap: undoped case

We shall start our presentation of the results of our study by describing the optimization of the geometry and the determination of the polarity and the energy gap for the undoped geometry. First-principles calculations using LSDA (with PWC parameterization) and GGS (with PBE parameterization) show that it is possible to construct hexagonal silicon nanosheets (h-Si; Fig. 2a) and hexagonal silicon nanosheets with defects (d-Si; Fig. 2b), in agreement with the binding energies and the non-negative vibrational frequency criterion (Table 1). The geometry optimization results yield nonplanar (corrugated) geometries for both h-Si and d-Si, in agreement with the results of Ciraci et al. [5]. Also note that the d-Si system exhibits a curvature of 32.72° according to the LSDA theory and 35.76° according to the GGS approach[1] [31], with sp^2 and sp^3 hybridization observed for the Si–Si bonds. The bond lengths in the h-Si systems are quite similar for both the LSDA (2.23 Å) and GGS (2.24 Å) theories, and both compare well with the reported value of 2.25 Å [5]. In contrast, bond lengths for

[1] The curvatures of the graphene and boron nitride sheets are predicted by the model we employed in our studies. However, zinc oxide sheets do not show any bending, suggesting that the curvature is dependent on the structure rather than the model. We also used the same model to investigate the structures of hexagonal germanium (h-Ge) sheets and defect-modified germanium (d-Ge) sheets, and the results indicated corrugated surfaces that show semimetallic behavior (Chigo Anota E, Salazar Villanueva M, submitted to J Nanomat.) A calculation that utilizes periodical theory has recently been published that focuses on similar structures with larger unit cells, in order to demonstrate that amorphous graphene consisting of pentagons and hexagons displays only a small curvature.

the d-Si systems vary between 2.19 and 2.26 Å in the LSDA formalism and between 2.23 and 2.27 Å in the GGS approach. There is an energy difference between the two sheets: 0.74 eV in the LSDA approach and 0.94 eV in the GGS theory. The appearance of curvature in the d-Si structure suggests that it is possible to construct a C_{60}-like atomic configuration.

The dipole moment of the h-Si structure displays ionic character (0.33 D in both the LSDA and the GGS approaches), similar to the graphene sheet [32], but the dipole moment of the d-Si structure displays covalent character (2.78 D in the LSDA approximation and 2.74 D in the GGS approximation), similar to the graphene sheet in the presence of lattice defects (which has a dipolar moment of 3.07 D) (see Chigo Anota E, Salazar Villanueva M, submitted; and previous footnote). We argue that the lack of symmetry induced by the incomplete hexagonal structure (the structure consists of pentagons and heptagons) produces the change in polarity. According to the HOMO–LUMO energy difference used to determine the energy gap, both the h-Si and the d-Si structures behave as semimetals with energy gaps of 0.77 eV in the LSDA formalism and 0.76 eV in the GGS theory for h-Si, and 0.32 eV in the LSDA formalism and 0.33 eV in the GGS theory for d-Si. This behavior contrasts with those exhibited by graphene systems, provided that in the intrinsic graphene an energy gap of 1.94 eV is obtained (for the hexagonal structure) to yield a semiconductor and in the doped graphene an energy gap of 0.83 eV is obtained to resemble a semi-metallic structure (see Chigo Anota E, Salazar Villanueva M, submitted; and previous footnote), so the structural configuration determines the electronic structure of this bidimensional system.

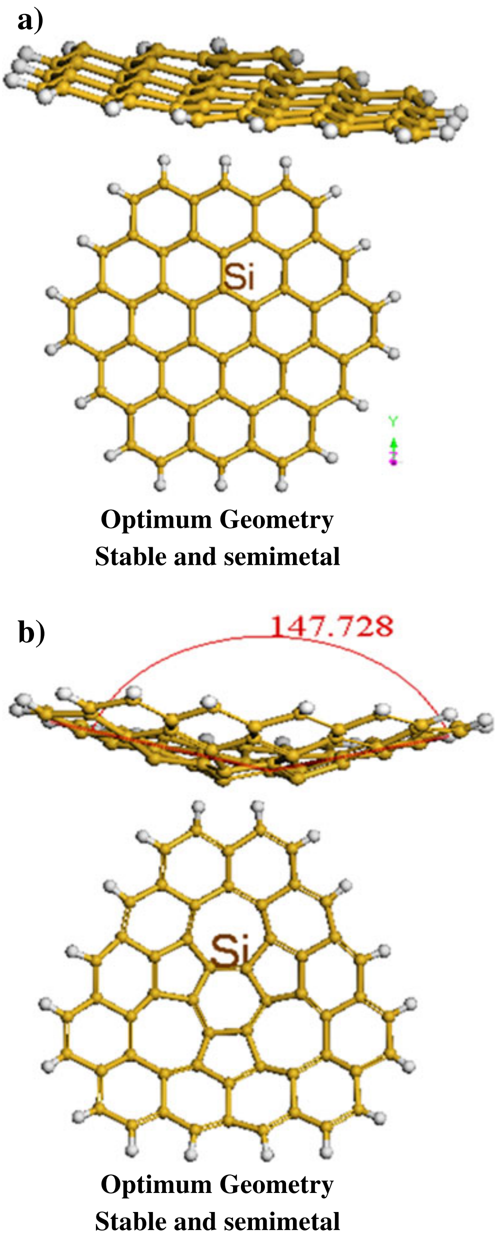


Fig. 2 **a** Optimized structure of the hexagonal silicon nanosheet (h-Si layer). **b** Corresponding structure with defects (d-Si geometry) oriented along the *xy* plane

Results reported by Guzmán et al. [33] for a hexagonal hydrogenated silicon structure of the graphane type indicate an energy gap of 2.2 eV, which differs from our values. On the other hand, studies of h-Si by Ciraci et al. [5] show semimetallic character, similar to our results.

Geometry optimization, polarity, and gap: doped case

In this section, we describe calculations aimed at optimizing the doped case, as well as determining its polarity and energy gap. The h-Si structure doped with carbon (Fig. 3a–f) produces stable configurations with corrugated shapes that have a central planar form in the vicinity of the doping atom, similar to graphene. When the h-Si is doped with germanium, the sheet exhibits a corrugated shape, as induced by the hexagonal germanium structure, which has a corrugated shape ([5]; also see Chigo Anota E, Salazar Villanueva M, submitted; and previous footnote). For the corrugated h-Si system, it was found that doping with carbon induces a twist (twist angle: 7.8°; Fig. 3a), while doping with germanium does not affect the structure. In these cases, we only used the LSDA approach when the results given by the GGS theory were the same as those obtained with LSDA. When the h-Si was doped with 1.85% C (Fig. 3a), which was inserted into the central hexagon, a $Si_{53}CH_{18}$ configuration was obtained for both structures, and a Si–C bond length of 1.82 Å, along with *sp* and sp^2 hybridization and a small variation in the Si–Si bond length of 10^{-2} Å (see Table 1). On the other hand, when this structure was doped with 1.85% Ge (Fig. 3b), which again was inserted into the central hexagon, a $Si_{53}GeH_{12}$ configuration was obtained for both structures), and a Si–Ge bond length of 2.36 Å, along with sp^2 and sp^3

Table 1 Bond lengths, dipole moments, energy gaps (HOMO–LUMO), binding energies, and thermodynamic properties of graphene-like silicon sheets

Cluster		Bond length (Å)			Dipolar moment (D)	Gap (HOMO–LUMO, in eV)	Binding energy (eV)	Thermodynamic properties		
		Si–Si	Si–C	Si–Ge				*S* (kcal/mol)	C_p	*H*
h-Si		2.25				Semimetal [5]+				
		2.248 [33]								
		2.247 [35]								
h-SiH						2.2 [34]++				
		2.319 [33]				2.0 (LDA) [33]				
h-Si*	LSDA	2.23			0.33	0.77	10.42	91.2	40.71	142.71
	GGS	2.24			0.33	0.76	9.09	95.22	43.56	140.62
	Doped*, LSDA									
	$Si_{53}CH_{18}$	2.19–2.26	1.82		0.48	0.75	10.55	92.26	41.10	144.24
	$Si_{23}GeH_{18}$	2.23–2.27		2.27	0.39	0.75	10.40	91.64	41.26	142.20
	Double-doped									
	$Si_{52}CGeH_{18}$	2.20–2.25	1.83	2.27	0.50	0.74	10.52	94.31	42.93	143.93
d-Si*	LSDA	2.24–2.28			2.78	0.32	10.29	94.69	44.43	141.19
	GGS	2.26–2.31			2.74	0.33	8.99	93.54	43.36	139.99
	Doped*, LSDA									
	$Si_{53}CH_{18}$	2.19–2.27	1.81		2.03	0.29	10.42	94.98	43.80	143.22
	$Si_{53}GeH_{18}$	2.24–2.27		2.36	2.85	0.26	10.25	95.82	45.39	140.53
	Double-doped									
	$Si_{52}CGeH_{18}$	2.24–2.27	1.82	2.30	1.56	0.32	10.48	94.71	43.29	142.91

+ Theoretical solid-state simulation

++ Tight binding calculation for a graphene-like sheet

* Present work

hybridization and a small variation in the Si–Si bond length of 10^{-2} Å (see Table 1). The semimetallic character was preserved with an energy gap of 0.75 eV (Table 1).

Let us now explore the d-Si structure when it is doped with a germanium atom (Fig. 3d). The resulting geometry is corrugated and the lattice preserves the sp^2 and sp^3 hybridization in the vicinity of the dopant, in a similar fashion to what was seen for the h-Si structure. The structure displays covalent character (2.85 D) and semimetallic behavior, and it has an energy gap of 0.26 eV. It bends 36.31° as a result of the doping, 4.04° more than the undoped d-Si and 13.75° more than the carbon-doped structure.

A third doping case was also explored, which corresponds to two silicon atoms being replaced by one carbon atom and one germanium atom (doping ratio: 3.70%), yielding the chemical composition $Si_{52}CGeH_{18}$ at the central hexagon for both h-Si and d-Si (Fig. 3e and f). After this kind of doping, the bond lengths in the h-Si structure are 2.20–2.25 Å for Si–Si, 1.83 Å for C–Si, and 2.27 Å for Si–Ge. The structure retains its polarity (0.48 D when doped with carbon and 0.39 D when doped with germanium) and semimetallic behavior (energy gap of 0.75 eV). On the other hand, the polarity of the d-Si structure decreases by almost 50% upon doping, but it retains its semimetallic behavior.

Thermodynamic properties

We now consider the thermodynamic properties of the structures: entropy, heat capacity, and enthalpy. The d-Si atomic structure is sensible to the transformation of configuration from a hexagonal one to another which contains pentagons, heptagons and hexagons. The geometric behavior is associated with the entropy variation produced by the change of structure from the nonplanar h-Si to the concave d-Si. The results for the heat capacities suggest that the structure has ceramic character. On the other hand, the enthalpy yields information about the thermal stability of each system.

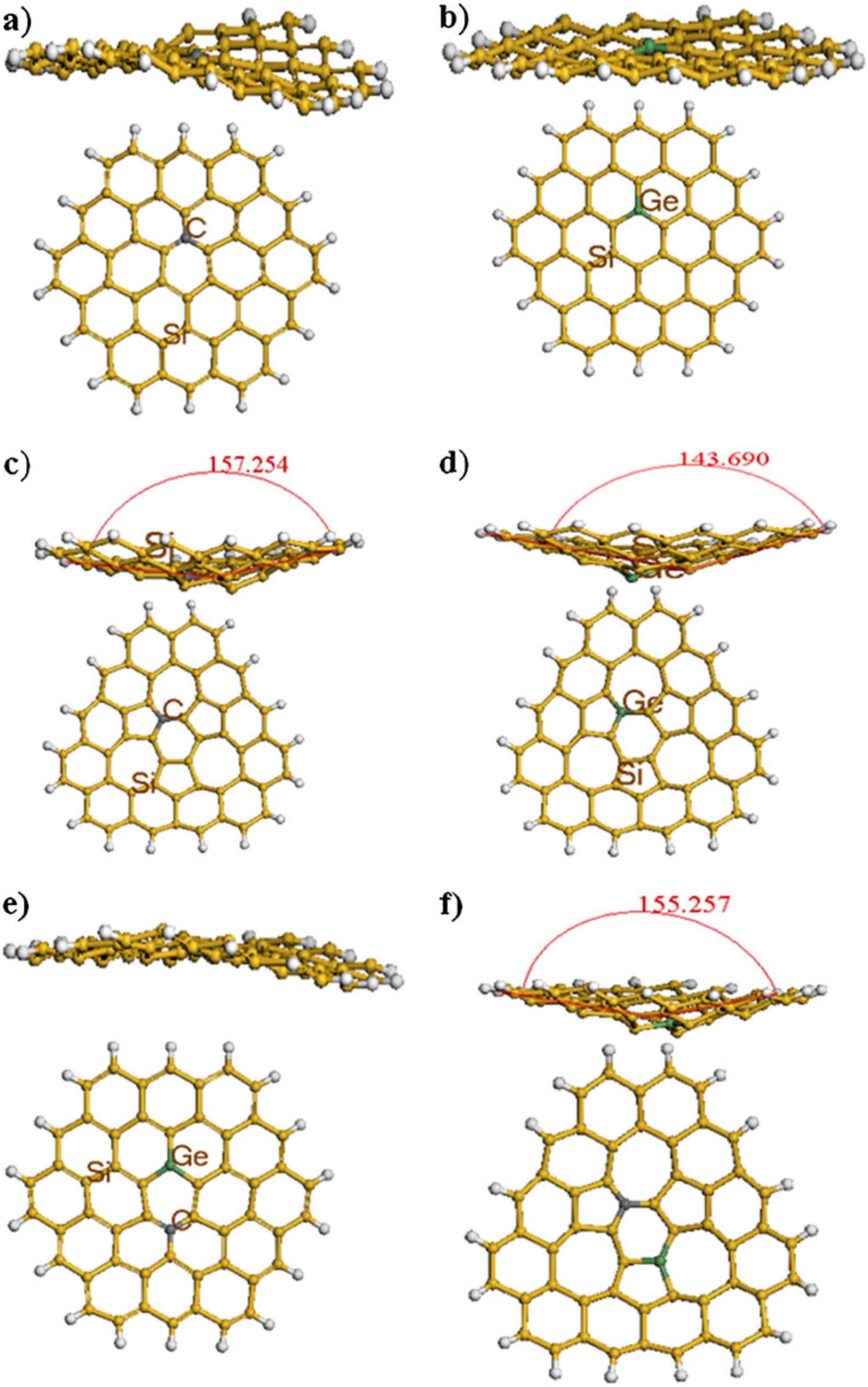


Fig. 3 Doped h-Si and d-Si structures. **a** h-Si structure doped with a carbon atom. **b** h-Si structure doped with a germanium atom. **c** d-Si structure doped with a carbon atom. **d** d-Si structure doped with a germanium atom. **e** h-Si structure doped with a carbon atom and a germanium atom. **f** d-Si structure doped with a carbon atom and a germanium atom

Conclusions

We have investigated the electronic and structural properties of hexagonal silicon nanosheets (h-Si) and defect-modified silicon nanosheets (d-Si). The total energy results show that the structure of h-Si has a lower total total energy than the d-Si structure (by 0.74 eV according to the LSDA formalism and 0.94 eV according to the GGS theory). Results for the binding energies indicate that it is plausible to fabricate d-Si sheets. The ground-state structure of the sheet displays concavity, which in turn suggests the possibility of obtaining a fullerene-type structure. It is also worth noting that the main characteristic displayed by the d-Si structure is that it retains its semimetallic behavior even after the structure has been doped with one or two atoms of impurities. The non-hexagonal low-symmetry structure induces strong variations in polarity, resulting in a covalent molecular structure. Based on these results, it appears possible to obtain h-Si and d-Si sheets, which are very promising materials for applications in the optoelectronics industry.

Acknowledgments This work has been supported by the projects Vicerrectoria de Investigación y Estudios de Posgrado-Benemérita Universidad Autónoma de Puebla (CHAE-ING11-I), Facultad de Ingeniería Química- Benemérita Universidad Autónoma de Puebla (2010–2011), Cuerpos Académicos de Ingeniería en Materiales (BUAP-CA-177), Física Computacional de la Materia Condensada (BUAP-CA-194), and Propiedades Mecánicas, Electrónicas y Estructurales de Materiales (BUAP-CA-236). Work done by GHC was partially supported by Consejo Nacional de Ciencia y Tecnología (83982).

References

1. Novoselov KS, Geim AK, Morozov SV, Jiang D, Zhang Y, Dubonos SV, Grigorieva IV, Firsov AA (2004) Science 306:666–669
2. Sugiyama Y, Okamoto H, Mitsuoka T, Morikawa T, Nakanishi K, Ohta T, Nakano H (2010) J Am Chem Soc 132:5946–5947
3. Harada M, Matsushita Y (2007) Activ Rep Neutron Scatt Res Exp Rep 14:252. http://quasi.issp.u-tokyo.ac.jp/actrep/actrep-14/pdf/ISSPNSL_report_252.pdf
4. Nakano H (2005) R&D Rev Toyota CRDL 40(3):51. http://www.tytlabs.co.jp/japanese/review/rev403pdf/403_051nakano.pdf
5. Şahin H, Cahangirov S, Topsakal M, Bekaroglu E, Akturk E, Senger RT, Ciraci S (2009) Phys Rev B 80:155453
6. Ciesielski A, Cyrański MK, Krygowski TM, Fowler PW, Lillington M (2006) J Org Chem 71:6840–6845
7. Martínez JI, Cabria I, López MJ, Alonso J (2009) J Phys Chem C 113:939–941
8. You YM, Ni Zh H, Yu T, Shen ZX (2008) Appl Phys Lett 93:163112
9. Zeng H, Zhi C, Zhang Z, Wei X, Wang X, Guo W, Bando Y, Golberg D (2010) Nano Lett 10:5049–5055
10. Akcöltekin S, Bukowska H, Peters T, Osmani O, Monnet I, Alzaher I, d'Etat BB, Lebius H, Schleberger M (2011) Appl Phys Lett 98:103103
11. Dinadayalane TC, Leszczynski J (2007) Chem Phys Lett 434:86–91
12. Dinadayalane TC, Leszczynski J (2010) Struct Chem 21:1155–1169
13. Lahiri J, Lin Y, Bozkurt P, Oleynik II, Batzill M (2010) Nature Nanotech 5:326–329
14. Chigo Anota E (2009) Superficies y Vacío 22:19–23
15. Chigo Anota E, Salazar Villanueva M, Hernández Cocoletzi H (2010) Phys Stat Solidi C 7:2252–2254
16. Chigo Anota E, Salazar Villanueva M, Hernández Cocoletzi H (2010) Phys Stat Solidi C 7:2559–2561
17. Chigo Anota E, Hernández Cocoletzi H (2011) J Mol Model. doi:10.1007/s00894-011-1043-2
18. Chigo Anota E, Hernández Cocoletzi H, Bautista Hernández A, Sánchez Ramírez JF (2011) J Theor Comput Nanosci 8:637–641
19. Chigo Anota E, Murrieta Hernández G (2011) Rev Mex Fis 57:30–34
20. Chigo Anota E (2011) Superficies y Vacío 24:9–13
21. Chigo Anota E, Ramírez Gutiérrez RE, Escobedo Morales A, Hernández Cocoletzi G (2011) J Mol Model. doi:10.1007/s00894-011-1233-y
22. Galicia Hernández JM, Hernández Cocoletzi G, Chigo Anota E (2011) J Mol Model. doi:10.1007/s00894-011-1046-z
23. Kohn W, Becke AD, Parr RG (1996) J Phys Chem 100:12974–12980
24. Parr RG, Yang W (1989) Density-functional theory of atoms and molecules. Oxford University Press, Oxford
25. Delley B (1990) J Chem Phys 92:508–517
26. Perdew JP, Wang Y (1992) Phys Rev B 45:13244–13249
27. Perdew JP, Burke K, Ernzerhof M (1996) Phys Rev Lett 77:3865–3868
28. Delley B (1996) J Phys Chem 100:6107–6110
29. Delley B (2000) J Chem Phys 113:7756–7764
30. Foresman JB, Frisch Æ (1996) Exploring chemistry with electronic structure methods, 2nd edn. Gaussian, Inc., Pittsburgh, p 70
31. Li Y, Inam F, Kumar A, Thorpe MF, Drabold DA (2011) Phys Stat Solidi (b) 248:2082–2086
32. Hernández Rosas JJ, Ramirez Gutierrez RE, Escobedo Morales A, Chigo Anota E (2011) J Mol Model 17:1133–1139
33. Guzmán-Verri GG, Lew Yan Voon LC (2011) J Phys Condens Matter 23:145502
34. Lew Yan Voon LC, Sandberg E, Aga RS, Farajian AA (2010) App Phys Lett 97:163114
35. Takeda K, Shiraishi K (1994) Phys Rev B 50:14916

DOI 10.1007/s00894-011-1197-y

ORIGINAL PAPER

Structure-based characterization of the binding of peptide to the human endophilin-1 Src homology 3 domain using position-dependent noncovalent potential analysis

Chunjiang Fu · Gang Wu · Fenglin Lv · Feifei Tian

Received: 28 May 2011 / Accepted: 20 July 2011 / Published online: 27 September 2011

Abstract Many protein–protein interactions are mediated by a peptide-recognizing domain, such as WW, PDZ, or SH3. In the present study, we describe a new method called position-dependent noncovalent potential analysis (PDNPA), which can accurately characterize the nonbonding profile between the human endophilin-1 Src homology 3 (hEndo1 SH3) domain and its peptide ligands and quantitatively predict the binding affinity of peptide to hEndo1 SH3. In this procedure, structure models of diverse peptides in complex with the hEndo1 SH3 domain are constructed by molecular dynamics simulation and a virtual mutagenesis protocol. Subsequently, three noncovalent interactions associated with each position of the peptide ligand in the complexed state are analyzed using empirical potential functions, and the resulting potential descriptors are then correlated with the experimentally measured affinity on the basis of 1997 hEndo1 SH3-binding peptides with known activities, using linear partial least squares regression (PLS) and the nonlinear support vector machine (SVM). The results suggest that: (i) the electrostatics appears to be more important than steric properties and hydrophobicity in the formation of the hEndo1 SH3–peptide complex; (ii) P_{-4} of the core decapeptide ligand with the sequence pattern $P_{-6}P_{-5}P_{-4}P_{-3}P_{-2}P_{-1}P_0P_1P_2P_3$ is the most important position in terms of determining both the stability and specificity of the architecture of the complex, and; (iii) nonlinear SVM appears to be more effective than linear PLS for accurately predicting the binding affinity of a peptide ligand to hEndo1 SH3, whereas PLS models are straightforward and easy to interpret as compared to those built by SVM.

Keywords Peptide2 · Src homology 3 domain · Noncovalent interaction · Statistical modeling

Introduction

Recently, Russell and co-workers estimated that 15%–40% of all interactions in the cell are mediated through protein–peptide interactions [1, 2], meaning that, at its most extreme, nearly every protein is affected either directly or indirectly by peptide-binding events [3]. Such interactions are commonly mediated by specialized protein domains, among which the Src homology 3 (SH3) domain is the most abundant in eukaryotic genomes and presents in a wide variety of proteins, such as kinases, lipases, GTPases, and adaptor proteins, to orchestrate diverse cellular processes [4]. The SH3 domain family are conserved protein modules consisting of 50–70 residues, and these modules can specifically bind to contiguous proline-rich ligands characterized by a core region of 7–9 amino acids [5]. These SH3-binding peptides can be divided into classes I and II [6]. Their variable binding characteristics mean that the SH3 domain family show a broad specificity in terms of recognizing their peptide ligands.

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1197-y) contains supplementary material, which is available to authorized users.

C. Fu · G. Wu (✉)
Department of Cardiology and Department of Hepatobiliary Surgery,
Daping Hospital, The Third Military Medical University,
Chongqing 400010, China
e-mail: wg1118@sina.com

F. Lv · F. Tian
College of Bioengineering and Key Lab of Biorheological Science
and Technology, Chongqing University,
Chongqing 400044, China

Because domain–peptide interactions are usually weak and transient, and often depend upon post-translational modification, they tend to be underrepresented in experimental and computational studies, thus highlighting the need to develop new strategies to identify these interactions [4]. A number of methods have already been developed to dissect the interaction profiles of SH3 domains with their ligands, and to qualitatively or quantitatively analyze the binding potency underlying these interactions. From an experimental perspective, phage display has often been applied to determine the specificity of amino acid types at different positions along the peptide sequence; the resulting information is then used to build frequency matrices representing the amino acid preference at each position [7, 8]. In addition, high-density arrays of relatively short peptide chains can be efficiently synthesized by the positionally addressable synthesis of peptides on cellulose membranes (SPOT synthesis), and this technique has been employed to screen a large-scale sequence pool in order to find out the binding activities of the SH3 family [9, 10]. However, these methods may introduce bias due to the incomplete sampling of all possible peptides, leading to the arbitrary weighting of contributions of peptide positions based on their binding strengths and/or the random assignment of peptides bound to the SH3 domain in different binding modes. Moreover, it is too time-consuming and costly to synthesize all potential peptides found in the complete genomes and to perform a further domain–peptide binding assay. Alternatively, computational approaches have been exploited as a promising way to predict domain–peptide interactions. Early on, machine learning strategies such as Gibbs sampling [11], hidden Markov [12], neural network [13], and support vector machine [14] were introduced to qualitatively identify SH3 partners. These methods were trained under known SH3-binding and -unbinding peptides, and then used to distinguish whether a peptide would be recognized by SH3. Recently, the quantitative structure–activity relationship (QSAR) methodology was applied to predict the affinities of peptide candidates and to explain the structural basis for the binding of peptide to SH3 [15]. For example, Hou et al. employed molecular dynamics simulation and CoMFA/CoMSIA to examine the binding mode and potency of peptide to hAmph SH3 [16]. Later, this work was further generalized to decipher the protein recognition codes of diverse SH3 domains [17]. Zhou et al. employed divided physicochemical property scores coupled with genetic algorithm–Gaussian processes to perform a comparative study of a panel of culled SH3-binding peptides, and concluded that diverse properties contribute remarkably to the interactions between the hAmph SH3 and its peptide ligands [18]. Very recently, He et al. used principal property descriptors derived from amino acid rotamers (PDAR) to statistically predict the binding affinities of over 13,000 peptides to ten types of SH3 domains, and found that the electrostatics, hydrophobicity, and hydrogen bonds at core residue positions contribute significantly to SH3–peptide binding [19].

In this study, we present a novel structure-based approach that can be used to characterize the position-dependent noncovalent profiles of diverse peptides binding to the SH3 domain involved in human endophilin-1 (hEndo1), a protein that localizes in brain presynaptic nerve termini and participates at multiple stages in clathrin-coated endocytosis, from early membrane invagination to synaptic vesicle uncoating [20]. This hybrid method combines various molecular modeling techniques, including molecular mechanics/molecular dynamics simulation, virtual mutagenesis, interaction energy decomposition, solvent effect analysis, and statistical fitting and validation. Specifically, instead of traditional amino acid descriptors that ignore structural information about the receptor protein [21], we herein describe a structure-based nonbonding potential analysis protocol that dissects the position-dependent noncovalent interaction profile of the hEndo1 SH3 domain with its peptide ligands. This method is then used to analyze the interaction mode and binding potency of 1997 hEndo1 SH3 domain-binding peptides with known affinities, and to investigate the relative contributions of different residues and different noncovalent terms in the peptide sequence to the binding. We also demonstrate that the position-dependent noncovalent interaction descriptors derived from the newly proposed protocol are more effective for modeling and predicting SH3–peptide binding affinities if a nonlinear machine learning method is employed to perform the statistical modeling.

Methods and materials

Construction of the hEndo1 SH3 domain–peptide complex model

Twenty NMR structures of the hEndo1 SH3 domain were retrieved from the Protein Data Bank (PDB entry: 2dbm). We used the first copies of these structures to construct the model of the complex structure of hEndo1 SH3 domain with the ten core residues (RSPPRPPRER) of the peptide WSRSPPRPPRERFE—which was reported by Landgraf et al. [22] to be an effective binder to hEndo1 SH3. As suggested by Cestra et al. [23], this decapeptide adopts class I orientation binding to hEndo1 SH3, with the position pattern $P_{-6}P_{-5}P_{-4}P_{-3}P_{-2}P_{-1}P_{0}P_{1}P_{2}P_{3}$ (this numbering corresponds to the nomenclature suggested by Lim et al. [24]). The structure of the hEndo1 SH3 domain complexed with RSPPRPPRER was modeled based on the crystal structure of the Abl SH3 domain in complex with a 3BP-1 synthetic peptide (APTMPPPLPP) (PDB entry: 1abo). The obtained crude model was then subjected to a molecular dynamics (MD) simulation to eliminate existing collisions and distortions in the structure and to relax the complex. Details about the MD can be found in the publications of Hou et al. [14, 16], who have recently employed a similar protocol to construct the model of the complex of the hAmph1 SH3 domain with the peptide PLPRRPPRAA. Briefly, the crude model was solvated in a rectangular box that extended 8 Å away from any solute atom, and then the hydrogen atoms, water molecules, and all systematic components were minimized in turn without constraints. The MD procedure consisted of a gradual temperature increase from 50 K to 300 K over 50 ps, and a 1000 ps simulation for equilibration. The SHAKE algorithm was employed to constrain all bonds in the system to speed up the simulation [25]. All MD calculations were carried out in the AMBER9.0 package with the AMBER03 force field [26]. The equilibrated structure model of the hEndo1 SH3–RSPPRPPRER complex will be used in the following study (Fig. 1).

Virtual mutagenesis of the template to target peptides

To obtain the hEndo1 SH3 domain–target peptide complex model, the residues of the template peptide RSPPRPPRER in complex with hEndo1 SH3 were mutated in turn to the corresponding residue types of the target peptide. The virtual mutation of a peptide residue was implemented by two steps: the side chain of the residue being mutated was manually deleted from the template, and then a new side chain was added automatically using the rotamer-based SCWRL4 program [27]. Before the virtual mutagenesis protocol was applied, all water molecules and cofactors were removed from the template structure, and then a hydrogen-adding procedure using the REDUCE strategy was employed [28]. SCWRL and REDUCE were adopted here because these two programs have been previously demonstrated to give good performance when reproducing the experimentally determined structure data of peptides and proteins [29, 30].

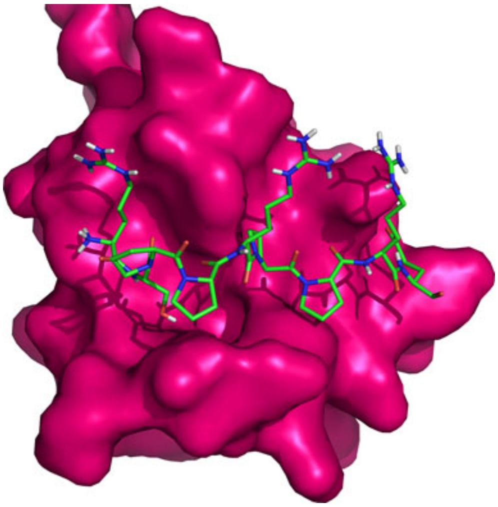


Fig. 1 Stereoview of the model of the complex of the hEndo1 SH3 domain with RSPPRPPRER. The hEndo1 SH3 domain was extracted from a 20-copy NMR structure (PDB entry: 2dbm), and the peptide RSPPRPPRER represents the ten core residues of the effective hEndo1 SH3-binder WSRSPPRPPRERFE, as reported by Landgraf et al. [22]

Position-dependent interaction energy analysis

The interaction energy of each position of a decapeptide ligand with the hEndo1 SH3 domain in the complexed state was decomposed into three components: an electrostatic term, a steric term, and a hydrophobic term. These can be readily calculated using a semi-empirical molecular mechanics approach and an empirical potential expression. The AMBER force field is widely used to describe the nonbonding interactions involved in biomacromolecular entities [31, 32], and was thus employed here to account for the electrostatic and steric terms, while a pairwise atomic hydrophobic potential developed in our lab [33] was utilized to characterize the hydrophobic interaction. In the AMBER force field, electrostatic and steric interactions between two protein (or peptide) atoms are quantified using Coulomb's law and the Lennard-Jones 6–12 equation, respectively. The formula of the additional hydrophobic potential is $U_{ij}^{hp} = -(S_i\rho_i + S_j\rho_j)e^{-d_{ij}}$[33], where d_{ij} is the distance between two atoms i and j, ρ_i represents the Eisenberg atomic solvation parameters [34], and S_i is the atomic solvent accessible surface area defined in the MSMS program [35]. Detailed descriptions of the procedure for calculating these nonbonding interactions can be found in our previous publications (with only slight modifications) [36, 37].

For a hEndo1 SH3–decapeptide complex, three interaction terms associated with each position of the decapeptide ligand can be calculated using the strategy described above. In this way, the position-dependent noncovalent interaction profile of the decapeptide with the hEndo1 SH3 domain is characterized by 3×10=30 nonbonding potential terms V_1–V_{30}, in which V_1, V_2, and V_3 represent, respectively, the electrostatic, steric, and hydrophobic interactions of position 1 of the decapeptide ligand with the hEndo1 SH3 domain; V_4, V_5, and V_6 represent those interactions for position 2; and so on.

Data set and statistical modeling

Since the contributions of different positions in a peptide and different nonbonding components at a particular position to the binding may not be identical, we further defined a linear weighting formula to correlate the 30 noncovalent terms with binding affinity:

$$Affinity_{(\log BLU)} = b_0 + b_1V_1 + b_2V_2 + \cdots b_{30}V_{30}. \qquad (1)$$

The values of the weights b_k were obtained by linearly fitting Eq. 1 to experimentally measured affinities on the basis of an elaborately selected, large-scale pool of hEndo1 SH3-binding peptides, using the sophisticated partial least squares regression (PLS) technique [37]. In addition, as a comparison, the nonlinear support vector machine (SVM) [38] was employed to perform correlation. The selected peptide panel contained 883 samples and was a subset of the 1997 peptides that were synthesized at high density on cellulose membranes using the SPOT synthesis technology. A chemoluminescence assay was then performed to determine the signal intensities of the peptides bound with the glutathione *S*-transferase (GST)-fused hEndo1 SH3 domain [22]. The SPOT signal intensities were measured in BLU, which is a quantitative indictor of the dissociation constant of the hEndo1 SH3–peptide complex. Each of these 883 selected peptides that were used to obtain the weight values b_k in Eq. 1 was assayed at least twice for its BLU value, and the remaining 1114 samples (with only one experimental result for each) were utilized as the test set to validate the reliabilities of the obtained models [see the "Electronic supplementary material" (ESM), Tables S1 and S2].

Results and discussion

Model development

In order to explore the relative contributions of different noncovalent components to the binding, we used the PLS method to separately correlate independent noncovalent terms and combinations of them with the binding affinities of the 883 training peptides. As shown in Table 1, model quality was improved obviously by increasing the number of noncovalent terms adopted, and the best model was obtained when all three noncovalent terms were fed into the modeling, suggesting that hEndo1 SH3–peptide binding is determined by diverse nonbonding properties. Among the three noncovalent components, the most important one seems to be the electrostatic term, which can give models with good stability (indicated by tenfold cross-validation q_{CV}^2) and strong fitting ability (measured by the fitted coefficient of determination r_{fit}^2) as compared to those without the electrostatic term included, although the secondary steric and hydrophobic components also significantly affect the binding behavior of the peptide to the hEndo1 SH3 domain. Here, we used the best model (i.e., the full-parameter one) to predict the 1114 test peptides, and we found a good correlation between the predicted and experimental affinities, as shown by the acceptable predictive coefficient r_{prd}^2=0.515 and the root-mean-square error of prediction (RMSP)=0.586.

Table 1 Statistics of PLS models using different combinations of noncovalent terms

Noncovalent combination[a]	r_{fit}^2	q_{CV}^2	RMSF
E	0.448	0.387	0.632
S	0.396	0.325	0.649
H	0.388	0.308	0.656
E + S	0.534	0.462	0.573
E + H	0.512	0.453	0.595
S + H	0.469	0.407	0.615
E + S + H	0.569	0.514	0.558

[a] *E* electrostatic term, *S* steric term, *H* hydrophobic term

The plots of fitted and predicted versus experimentally determined affinities for the training and test samples are shown in Figs. 2a and b, respectively. As can be seen, most samples are distributed around the fit lines, and only a few points (outliers) deviated greatly from the lines. We have examined these outliers and found that unusual structure and abnormal affinity appear to be the major factors that lead our model to significantly overestimate or underestimate the binding potency of these outliers. In addition, unavoidable errors associated with experimental assays are an important factor that undermines the accuracy of the statistical model. Furthermore, appreciable systematic error is present in the plots; that is, low-affinity peptides are commonly overestimated whereas high-affinity samples are underestimated by the model (indicated by the small slopes *k* of the fit lines). This phenomenon has also been observed in previous statistical investigations of hAmph1 SH3 domain-binding peptides [18, 19], and can be explained by the fact that some unknown factors, such as the interactions between separate residues in the peptide sequence and the conformational entropy loss during the binding, which were not considered in our method, may contribute marginally to the affinity.

In order to characterize the contributions of different residue positions to the peptide affinity, the linear relationships between independent positions and the binding affinities of training peptides were also calculated, and these are provided in Table 2. It can be seen that the contributions of independent positions to the peptide affinity are quite modest; only P_{-4}, P_{-3}, and P_1 seem to be relatively important in the binding, suggesting that none of these ten residue positions can independently dominate the binding behavior of the peptide to the hEndo1 SH3 domain, and the recognition and interaction of a peptide ligand with hEndo1 SH3 are co-determined by many

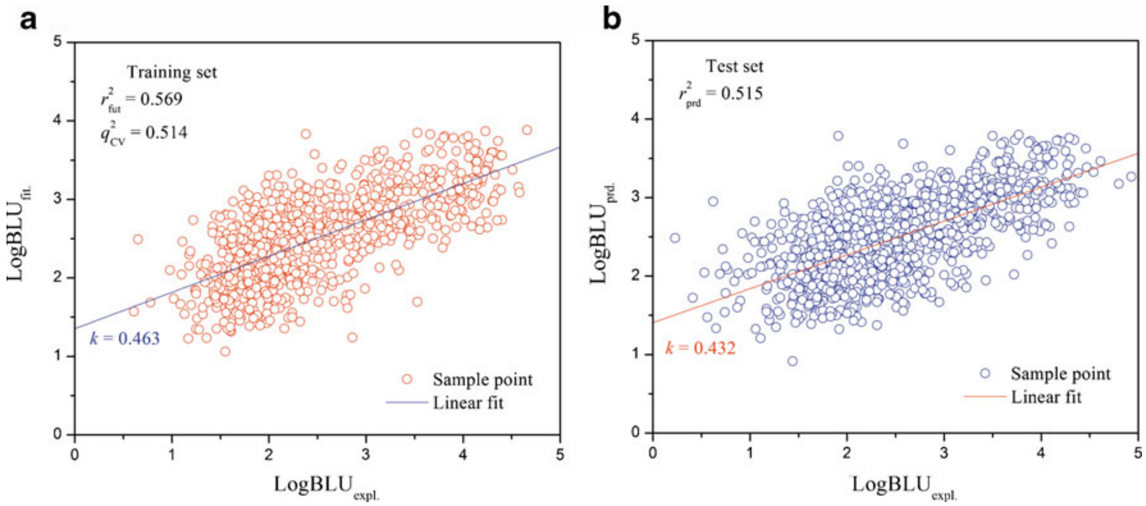


Fig. 2 **a** Plot of fitted versus experimental affinities for the 883 training samples. **b** Plot of predicted versus experimental affinities for the 1114 test samples

positions on the peptide. This finding can be further rationalized by analyzing the PLS coefficient plot (vide post).

Model analysis

The hidden information involved in the full-parameter PLS model was further analyzed in detail. The 883 training samples distributed in the top-two score space of PLS model are shown in Fig. 3. It is evident that the entire space can be divided into two regions: one at the bottom right of this plot, with only a few sample points (region 1), and another at the top left, which holds the majority of the samples (region 2). We investigated the structural difference between the peptide samples of these two regions, and found that the peptides present in region 1 are mainly those with abundant charge, such as DKPVKPPTKK, SRPTRPPEPR, and LGPAKPPAQQ, most of which have a relatively strong ability to bind with the hEndo1 SH3 domain, whereas the samples involved in region 2 are commonly hydrophobic and show moderate or weak affinity to hEndo1 SH3. Region 2 can be further partitioned into three parts, each of which represents a specific group of peptides with fine structural characteristics in common, and which differ in these characteristics from the peptides in the other two groups. Based on these considerations, the PLS score plot is thought to be capable of accurately characterizing the structures and properties of diverse peptides when they interact with the hEndo1 SH3 domain.

The PLS coefficient plot can also give information on the relative contributions to the peptide affinity of different positions in the peptide sequence and different noncovalent components at a given position. Even at first glance, it is clear from Fig. 4 that the contributions of the ten positions of the peptide are significantly different; the N-terminal residues seem to be more effective than the C-terminal ones at determining the binding behavior of the peptide ligand to hEndo1 SH domain. For example, the absolute values of the PLS coefficients at P_{-6}–P_{-2} are larger than those at P_{-1}–P_3; in particular, the position P_{-4} makes a substantial (negative) contribution to the binding potency of the peptide. In addition, different noncovalent components at different positions also contribute differently to the affinity, and the electrostatics appears to be more important than steric and hydrophobic factors. In the following section, based on the PLS coefficient plot, we will discuss the contributions of different positions in the decapeptide ligand to the binding affinity.

The most important position: P_{-4}

According to Fig. 4, the three noncovalent components at P_{-4} all make very significant negative contributions to the binding affinity of the peptide. In fact, P_{-4} is usually occupied by a conserved Pro residue, and any charged and/

Table 2 The linear relationship between independent residue positions and the binding affinities of training peptides

Statistics	P_{-6}	P_{-5}	P_{-4}	P_{-3}	P_{-2}	P_{-1}	P_0	P_1	P_2	P_3
r_{fit}^2	0.154	0.197	0.398	0.241	0.260	0.039	0.115	0.298	0.090	0.085
q_{CV}^2	0.041	0.093	0.215	0.136	0.162	-[a]	0.002	0.104	-[a]	-[a]

[a] Insignificant

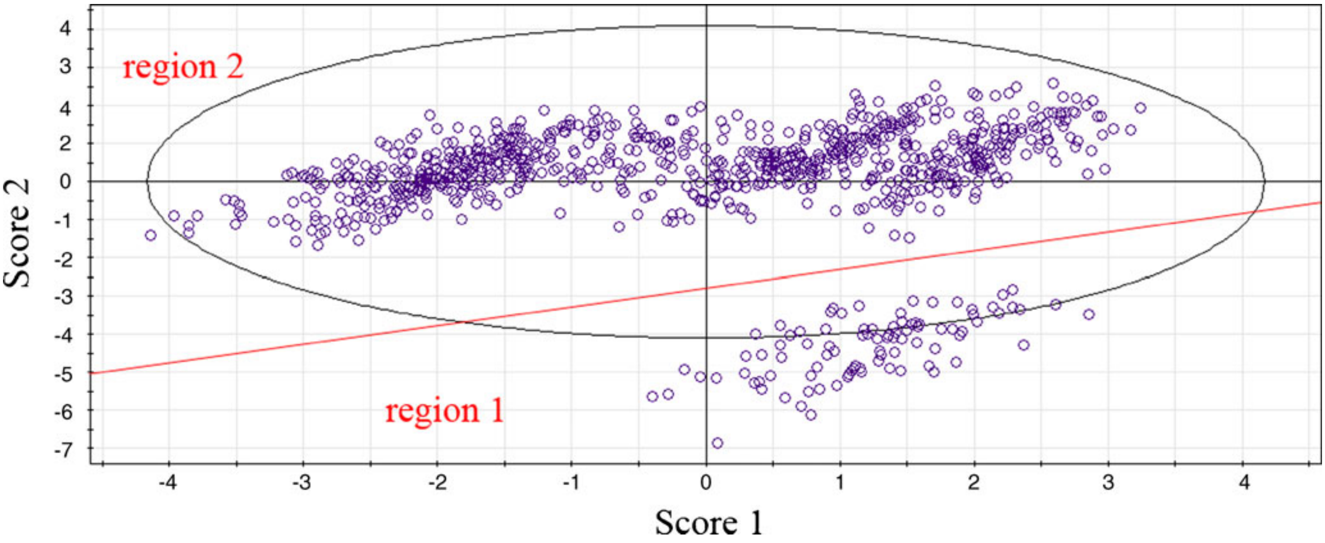


Fig. 3 The 884 peptide samples distributed in the top-two score space of the PLS model

or bulky substitution would fundamentally undermine the perfect match between the protein receptor and the peptide ligand. In the training samples, only four kinds of amino acids (i.e., Pro, Leu, Ile and Phe) are present at P_{-4}; aside from Pro, the other three types of amino acids are usually associated with low peptide affinity, revealing that strong hydrophobicity and a large volume at P_{-4} could impair the binding potency of the peptide. Nevertheless, the significant electrostatic effect also seems to disfavor peptide binding. This can be explained by noting that strong polarity is always associated with large volume; this would lead to a potential stereo-hindrance effect.

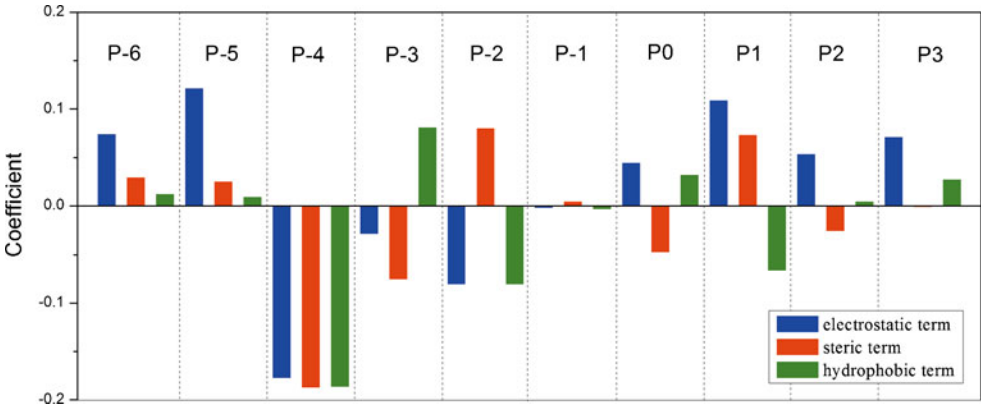
The secondary positions: P_{-6}, P_{-5}, P_{-3}, P_{-2}, and P_1

The electrostatics at P_{-6} and P_{-5} dominate over the steric and hydrophobic factors. By visually inspecting the structure model of the hEndo1 SH3–peptide complex, we found that the SH3 residues which make direct contact with these two positions are full of formal charge (e.g., Asp39 to P_{-6} and Glu19 and Glu23 to P_{-5}). Previously, Hou et al. used molecular mechanics/Poisson–Boltzmann solvent area method to perform a detailed investigation of the electrostatic interaction behavior of peptide with the Abl SH3 domain, which is highly homologous with the hEndo1 SH3 domain (both bind peptides in the class I orientation), and pointed out that the electrostatic free energy is particularly significant at the N-terminus of the peptide ligand, and thus it contributes substantial stability and specificity to the binding [4]. In addition, the fact that P_{-3}, P_{-2}, and P_1 are distributed in the middle of the peptide sequence can also exert an appreciable effect on the peptide affinity. It is suggested that nearly all of the noncovalent components play important roles at these positions, indicating that the hEndo1 SH3–peptide recognition and binding is affected by diverse physicochemical properties.

The insignificant positions: P_{-1}, P_0, P_2, and P_3

The PLS coefficients of the variable terms at P_{-1}, P_0, P_2, and P_3 are relatively small, suggesting limited contributions of these positions to peptide affinity. In fact, P_{-1} is

Fig. 4 The coefficients of the 30 variable terms in the PLS model



conservatively occupied by a Pro residue in all of the training samples, which is insignificant for statistical modeling. P_0 and P_2 are far from the body of the SH3 protein and hence can only form weak nonbonding interactions with SH3. P_3 is located at the C-terminus of the peptide ligand and would be better regarded as a marginal residue of the peptide.

Comparison of linear PLS to nonlinear SVM

Because of the complexity and polymorphism associated with the SH3 domain–peptide interaction, accurately predicting the binding affinity and recognition specificity of the peptide ligand to diverse SH3 domains is a great challenge in the bioinformatics community [15]. Previously, Wang and co-workers reviewed a lot of works on the statistical modeling and prediction of SH3 domain–peptide interaction behavior, and found that nonlinear machine learning methods appear to more effective than linear PLS in terms of predictive accuracy and reliability [14, 17]. Therefore, in this work we employed a sophisticated SVM technique to mine the hidden nonlinear relationship between the structural descriptors and the binding affinities of the peptide samples. For SVM regression, a coarse-grained grid-searching scheme using the root-mean-square error of cross-validation (RMSCV) as the objective function was carried out to determine the optimum combination of ε-insensitive loss function, penalty C, and kernel parameter σ^2. In this procedure, the ε, σ^2, and C values were tuned simultaneously in grids ranging from 0 to 1, 1 to 10, and 1 to 1000, with step sizes of 0.1-, 1-, and 10-fold, respectively. A detailed description of this procedure can be found elsewhere [39]. As a result, the optimal SVM model was constructed based on the 883 training peptides, with fitted coefficient of determination r_{fit}^2 and tenfold cross-validation q_{CV}^2 values of 0.614 and 0.547, respectively. This model was further used to predict the affinities of 1114 test samples, resulting in a predictive correlation coefficient r_{prd}^2 of 0.534. As might be expected, the stability and generalization ability of nonlinear SVM are significant improvements on those of linear PLS, and it reveals a strong nonlinear dependence in the hEndo1 SH3–peptide system. On the other hand, although the SVM shows good fitting and predictive power in the statistical modeling of hEndo1 SH3-binding peptides, the model built using this method is just a black box that is difficult to interpret in detail on a molecular basis, and it offers little insight into structural implications underlying the interaction of the peptide ligand with the hEndo1 SH3 receptor. Based on this consideration, we concluded that both PLS and SVM are useful in the modeling, prediction, and interpretation of peptide affinity to the SH3 domain—the use of diverse strategies is always more effective than a single technique when comprehensively investigating a specific problem.

Comparison of structure-based to sequence-based models

Previously, a number of sequence-based methods have been used to model and predict the binding affinities of peptides to diverse SH3 domains, such as Abl [4] and hAmph1 [18]. The sequence-based methods only consider structural information obtained from the primary sequence of the peptide [21]; they completely ignore the spatial properties and interaction behavior of the ligand in complex with the receptor. Here, for comparison purposes, two sophisticated amino acid descriptors (i.e., z-scale [40] and DPPS [41, 42]) were employed to develop sequence-based models for the hEndo1 SH3 domain-binding peptide data set. The statistics of QSAR models based on z-scale and DPPS as well as our PDNPA are tabulated in Table 3. As can be seen, both the fitting ability (r_{fit}^2) and predictive power (r_{pred}^2) of our method are obviously better than those of DPPS and, particularly, z-scale. This is expected if we consider that only indirect information about the SH3–peptide interaction was involved in the z-scale and DPPS models, whereas the nonbonding profile at the SH3–peptide interface was directly utilized in the PDNPA approach. In fact, evidence from several previous works already suggests that incorporating SH3–peptide complex structure properties into QSAR modeling can substantially improve both the statistical quality and the interpretability of the resulting models. For example, for a panel of hAmph1 SH3 domain-binding peptides [22], the published predictive powers (r_{pred}^2) of sequence-based (Liang et al. [43]) and structure-based (He et al. [32]) models were 0.530 and 0.705, respectively, suggesting that structure-based methods, at least for the SH3–peptide binding data set, should be more effective and reliable than sequence-based ones, although the former is more time-consuming and labor-intensive than the latter.

Table 3 Statistics of QSAR models based on different characterization methods

Method	Training set (883 samples)			Test set (1114 samples)	
	r_{fit}^{2} [a]	q_{CV}^{2} [b]	RMSF[c]	r_{pred}^{2} [d]	RMSP[e]
This work	0.569	0.514	0.558	0.515	0.586
z-Scale	0.395	0.370	0.662	0.324	0.698
DPPS	0.513	0.463	0.594	0.404	0.651

[a] r_{fit}^2 coefficient of determination for the fit to the training set. [b] q_{CV}^2 coefficient of determination for tenfold cross-validation of the training set. [c] *RMSF* root-mean-square error of fit to training set. [d] r_{pred}^2 coefficient of determination for predictions for the test set. [e] *RMSP* root-mean-square error of prediction for the test set

Conclusions

Accurate characterization of the noncovalent interaction profile of a peptide ligand with a protein receptor in the complex state is crucial for developing reliable statistical models that can quantitatively predict the binding affinities of diverse peptides, and to qualitatively explain the physicochemical properties and structural basis for these interactions. In this study, a new method, named position-dependent noncovalent potential analysis, was proposed for this purpose. In this procedure, structure models of the hEndo1 SH3 domain in complex with thousands of core decapeptide ligands were constructed using a combined strategy incorporating MD simulation and virtual mutagenesis based on high-resolution crystal structures. Subsequently, three noncovalent types that dominate interbiomolecular recognition and binding were calculated empirically for each position on the peptide ligand interacting with hEndo1 SH3. The resultant potential descriptors were used as structural variables to develop linear and nonlinear correlation models with experimental affinity values for 1997 peptides with known activities. After analyzing these built models, we can make the following conclusions. (i) Diverse physicochemical properties make significant contributions to the hEndo1 SH3–peptide binding. In particular, the electrostatics seems to be the dominant aspect in the binding. (ii) P_{-4} of the peptide ligand is the position that has the greatest influence on both the stability and specificity of the hEndo1 SH3–peptide complex, while P_{-6}, P_{-5}, P_{-3}, P_{-2}, and P_1 can confer moderate stabilization to the complex architecture, and P_{-1}, P_0, P_2, and P_3 have only a limited thermodynamic effect on the binding. (iii) Nonlinear SVM performs fairly well as compared to linear PLS when modeling the binding affinities of peptides to hEndo1 SH3, whereas the statistical models built by PLS are more interpretable and straightforward than those obtained by SVM.

Acknowledgments This work was supported by the Natural Science Foundation Project of Chongqing CSTC (grant no. 2009BA5068).

References

1. Neduva V, Linding R, Su Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 3:e405
2. Petsalaki E, Russell RB (2008) Peptide mediated interactions in biological systems: new discoveries and applications. Curr Opin Biotechnol 19:344–350
3. Vanhee P, Stricher F, Baeten L, Verschueren E, Lenaerts T, Serrano L, Rousseau F, Schymkowitz J (2009) Protein–peptide interactions adopt the same structural motifs as monomeric protein folds. Structure 17:1128–1136
4. Hou T, Chen K, McLaughlin WA, Lu B, Wang W (2006) Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain. PLoS Comput Biol 2:e1
5. Sparks AB, Rider JE, Hoffman NG, Fowlkes DM, Quillam LA, Kay BK (1996) Distinct ligand preferences of Src homology 3 domains from Src, Yes, Abl, Cortactin, p53bp2, PLCgamma, Crk, and Grb2. Proc Natl Acad Sci USA 93:1540–1544
6. Lim WA, Richards FM, Fox RO (1997) Structural determinants of peptide-binding orientation and of sequence specificity in SH3 domains. Nature 372:375–379
7. Rickles RJ, Botfield MC, Zhou XM, Henry PA, Brugge JS, Zoller MJ (1995) Phage display selection of ligand residues important for Src homology 3 domain binding specificity. Proc Natl Acad Sci USA 92:10909–10913
8. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295:321–324
9. Frank R (1992) Spot-synthesis: an easy technique for the positionally addressable, parallel chemical synthesis on a membrane support. Tetrahedron 48:9217–9232
10. Reineke U, Volkmer-Engert R, Schneider-Mergener J (2001) Applications of peptide arrays prepared by the SPOT-technology. Curr Opin Biotech 12:59–64
11. Reiss DJ, Schwikowski B (2004) Predicting protein–peptide interactions via a network-based motif sampler. Bioinformatics 20(suppl 1):i274–i282
12. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34:D247–D251
13. Ferraro E, Via A, Ausiello G, Helmer-Citterich M (2005) A neural strategy for the inference of SH3 domain–peptide interaction specificity. BMC Bioinformatics 6(Suppl 4):S13
14. Hou T, Zhang W, Case DA, Wang W (2008) Characterization of domain–peptide interaction interface: a case study on the amphiphysin-1 SH3 domain. J Mol Biol 376:1201–1214
15. Ivanciuc O (2009) Machine learning quantitative structure–activity relationships (QSAR) for peptides binding to the human amphiphysin-1 SH3 domain. Curr Proteomics 6:289–302
16. Hou T, McLaughlin W, Lu B, Chen K, Wang W (2006) Prediction of binding affinities between the human amphiphysin-1 SH3 domain and its peptide ligands using homology modeling, molecular dynamics and molecular field analysis. J Proteome Res 5:32–43
17. Hou T, Xu Z, Zhang W, McLaughlin WA, Case DA, Xu Y, Wang W (2009) Characterization of domain–peptide interaction interface. Mol Cell Proteomics 8:639–649
18. Zhou P, Tian F, Chen X, Shang Z (2008) Modeling and prediction of binding affinities between the human amphiphysin SH3 domain and its peptide ligands using genetic algorithm–Gaussian processes. Biopolymers (Pept Sci) 90:792–802
19. He P, Wu W, Yang K, Jing T, Liao K, Zhang W, Wang H, Hua X (2011) Exploring the activity space of peptides binding to diverse SH3 domains using principal property descriptors derived from amino acid rotamers. Biopolymers (Pept Sci) 96:288–301
20. Reutens AT, Begley CG (2002) Endophilin-1: a multifunctional protein. Int J Biochem Cell Biol 34:1173–1177
21. Zhou P, Tian F, Wu Y, Li Z, Shang Z (2008) Quantitative sequence–activity model (QSAM): applying QSAR strategy to model and predict bioactivity and function of peptides, proteins and nucleic acids. Curr Comput Aided Drug Des 4:311–321
22. Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, Volkmer-Engert R, Cesareni G (2004) Protein interaction networks by proteome peptide scanning. PLoS Biol 2:94–103

23. Cestra G, Castagnoli L, Dente L, Minenkova O, Petrelli A (1999) The SH3 domains of endophilin and amphiphysin bind to the proline-rich region of synaptojanin 1 at distinct sites that display an unconventional binding specificity. J Biol Chem 274:32001–32007
24. Lim WA, Richards FM, Fox RO (1994) Structural determinants of peptidebinding orientation and of sequence specificity in SH3 domains. Nature 372:375–379
25. Ryckaert J, Ciccotti G, Berendsen HJC (1977) Numerical integration of Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. J Comput Phys 23:327–341
26. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM (2005) The AMBER biomolecular simulation programs. J Comput Chem 26:1668–1688
27. Krivov GG, Shapovalov MV, Dunbrack RL Jr (2009) Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77:778–795
28. Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 285:1735–1747
29. Knapp B, Omasits U, Schreiner W (2008) Side chain substitution benchmark for peptide/MHC interaction. Protein Sci 17:977–982
30. Zhou P, Tian F, Lv F, Shang Z (2009) Geometric characteristics of hydrogen bonds involving sulfur atoms in proteins. Proteins 76:151–163
31. Hu L, Ai Z, Liu P, Xiong Q, Min M, Lan C, Wang J, Fan L, Chen D (2010) Predicting the binding affinity of epitope peptides with HLA-A*0201 by encoding atom-pair non-covalent interaction information between receptor and ligands. Chem Biol Drug Des 75:597–606
32. He P, Wu W, Wang H, Yang K, Liao K, Zhang W (2010) Toward quantitative characterization of the binding profile between the human amphiphysin-1 SH3 domain and its peptide ligands. Amino Acids 38:1209–1218
33. Zhou P, Tian F, Li Z (2007) A structure-based, quantitative structureactivity relationship approach for predicting HLA-A*0201-restricted cytotoxic T lymphocyte epitopes. Chem Biol Drug Des 69:56–67
34. Eisenberg D, McLachlan AD (1986) Solvation energy in protein folding and binding. Nature 319:199–203
35. Sanner MF, Olson AJ, Spehner JC (1996) Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 38:305–320
36. Tian F, Zhang C, Fan X, Yang X, Wang X, Liang H (2010) Predicting the flexibility profile of ribosomal RNAs. Mol Inf 29:707–715
37. Wold S, Sjöström M, Eriksson L (2001) PLS regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130
38. Cortes C, Vapnik V (1995) Support vector networks. Mach Learn 20:273–293
39. Zhou P, Tian F, Lv F, Shang Z (2009) Comprehensive comparison of eight statistical modelling methods used in quantitative structure–retention relationship studies for liquid chromatographic retention times of peptides generated by protease digestion of the *Escherichia coli* proteome. J Chromatogr A 1216:3107–3116
40. Hellberg S, Sjostrom M, Skagerberg B, Wold S (1987) Peptide quantitative structure–activity relationships, a multivariate approach. J Med Chem 30:1126–1135
41. Tian F, Yang L, Lv F, Yang Q, Zhou P (2009) In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure–activity relationship approach. Amino Acids 36:535–554
42. Tian F, Lv F, Zhou P, Yang Q, Jalbout AF (2008) Toward prediction of binding affinities between the MHC protein and its peptide ligands using quantitative structure–affinity relationship approach. Protein Pept Lett 15:1033–1043
43. Liang G, Chen G, Niu W, Li Z (2008) Factor analysis scales of generalized amino acid information as applied in predicting interactions between the human amphiphysin-1 SH3 domains and their peptide ligands. Chem Biol Drug Des 71:345–351

J Mol Model (2012) 18:2163–2174
DOI 10.1007/s00894-011-1234-x

ORIGINAL PAPER

Putative binding modes of Ku70-SAP domain with double strand DNA: a molecular modeling study

Shaowen Hu · Janice M. Pluth · Francis A. Cucinotta

Received: 24 June 2011 / Accepted: 26 August 2011 / Published online: 27 September 2011

Abstract The channel structure of the Ku protein elegantly reveals the mechanistic basis of sequence-independent DNA-end binding, which is essential to genome integrity after exposure to ionizing radiation or in V(D)J recombination. However, contradicting evidence indicates that this protein is also involved in the regulation of gene expression and in other regulatory processes with intact chromosomes. This computational study predicts that a putative DNA binding domain of this protein, the SAP domain, can form DNA-bound complexes with relatively high affinities (ΔG ≈ -20 kcal mol^{-1}). The binding modes are searched by low frequency vibration modes driven by the fully flexible docking method while binding affinities are calculated by the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) method. We find this well defined 5 kDa domain with a helix-extended loop-helix structure is suitable to form favorable electrostatic and hydrophobic interactions with either the major groove or the minor groove of DNA. The calculation also reveals the sequence specified binding preference which may relate to the observed pause sites when Ku translocates along DNA and the perplex binding of Ku with circular DNA.

Keywords Binding free energy · DNA repair · Ku70-SAP · Ku-DNA binding · Molecular dynamics simulation

Introduction

The Ku heterodimer (Ku70 and Ku80) is a multifunctional protein involved in DNA repair, V(D)J recombination, mobile-genetic-element biology, telomere maintenance, apoptosis and transcription [1]. It is a highly abundant nuclear protein (approximate 5×10^5 per nucleus) [2] and has also been shown to be present within the cytoplasm to varying degrees dependent upon culture confluencies [3]. In vertebrates, Ku initiates the non-homologous end joining (NHEJ) DNA repair pathway by specific recognition and tethering of the DNA ends at the site of the lesion [4]. Once bound at the DNA ends, Ku works as a scaffold protein to recruit other repair factors that are required in NHEJ, which, in mammalian cells, include DNA-PKcs (DNA-dependent protein kinase catalytic subunit), Artemis, polymerase μ and λ, and a complex of XLF (Cernunnos), XRCC4, and DNA ligase IV, etc. [5]. These proteins act together in a highly coordinated way to cleave the incompatible section, fill the gap, and ligate the strands of DNA [6]. It was reported recently that the recruitment of these enzymes is not necessarily in the exact order of nuclease-polymerase-ligation, but can have a wide range of flexibility disregarding the exact structure of the DNA broken ends [7, 8]. This observation underscores the key mediation role that Ku plays in NHEJ pathway.

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1234-x) contains supplementary material, which is available to authorized users.

S. Hu (✉)
Division of Space Life Sciences, Universities Space Research Association,
Houston, TX 77058, USA
e-mail: Shaowen.hu-1@nasa.gov

J. M. Pluth
Lawrence Berkeley National Laboratory, Life Sciences Division,
One Cyclotron Road, Building 74,
Berkeley, CA 94720, USA

F. A. Cucinotta
NASA, Lyndon B. Johnson Space Center,
2101 NASA Parkway,
Houston, TX 77058, USA

The atomic structures of the human Ku heterodimer and a complex with DNA have been determined using X-ray crystallography by Walker et al. in 2001 [9]. The two Ku subunits show sequential and topological similarity, each containing three well-organized regions: an N-terminal α/β domain, a central β-barrel domain, and a subunit-specific helical C-terminal arm. They intertwine to form a pseudo-symmetrical structure with a preformed ring structure extended from a broad base (Fig. 1a), allowing two turns of DNA to cradle inside, with Ku70 located proximal and Ku80 distal to the free end (Fig. 1b). Structural analysis indicates that the inner surface of the ring is lined with positively charged residues, complementing the negative charge of the DNA. No direct base contacts are observed in the Ku-DNA complex, providing an explanation for why Ku lacks significant sequence preference for DNA binding. The combination of the extended β-barrel cradle and a narrow ring allows Ku to interact with two turns of the DNA double helix while still exposing much of the DNA to the solvent, allowing other repair factors to access DNA and exert their functions.

Upon DNA binding, the α/β domains and the central β-barrel domains forming the ring structure show virtually no conformational change, while the Ku70 extreme carboxyl terminus attached to Ku80 α/β domain is displaced. This shift of Ku70 carboxyl-terminal end upon DNA binding has been confirmed recently by single-particle electron microscopy experiment [10], implying a role for this domain in mediating DNA binding. The human Ku70 C-terminal arm is composed of a highly random and flexible linker (residues 536–560) and a well-structured helix-extended loop-helix (HEH) fold (residues 561–609), also referred to as the SAP domain (named after three proteins containing this motif: SAF-A/B, Acinus and PIAS) [11]. Biochemical studies indicate that the Ku70 C-terminal arm is responsible for the high affinity binding of Ku heterodimer to DNA [12].

The structure of the SAP domain is similar to that of the DNA-binding domain of T4 endonuclease VII and the RNA-binding domain of bacterial transcriptional termination factor Rho, and has been defined as a putative DNA-binding motif [11, 13]. This region has also been designated as having a role in the observed pause sites when Ku translocates along the DNA molecule [9] and in the perplex binding of Ku with circular DNA [1, 14]. Chemical shift perturbation experiment has been conducted to locate the interfaces of the SAP region to interact with DNA, and a binding mode has been proposed [13]. As this domain is exposed to solvent in unbound structure and is distal to the DNA free end in the bound structure [9], it is most likely among the first functional groups of the Ku protein to interact with the broken DNA whenever it is introduced. A better understanding of how Ku recognizes DNA breaks and translocates along the DNA molecule will aid in more fully elucidating the DNA repair process following ionizing radiation, including differences between terresterial radiation such as X-rays and gammas-rays and space radiation comprised of heavy ions and protons [15, 16].

In recent years a wide range of theoretical methods have been developed for computational modeling of biological systems [17]. They give computer modeling enormous potential to make predictions for molecular systems and provide insights that can guide mechanistic understanding and molecular design [18]. To test whether the SAP domain defines a good DNA binding motif and how it plays a role on the recognition and binding unusual nucleic acids structures such as double strand breaks induced by ionizing radiation, we performed a series of molecular modeling studies, and identified several possible binding modes with DNA duplexes with relatively high binding affinities ($\Delta G \approx -20$ kcal mol^{-1}). Initial structures were constructed based on the information obtained by electrostatic potential map calculation and an NMR titration experiment [13]. They were examined by low vibrational modes driven fully-flexible docking protocol with implicit solvent, and 5 ns molecular dynamic simulation with explicit solvent. Structural and energetic analyses carried out on the last 3 ns stable trajectories indicated that this well-defined domain

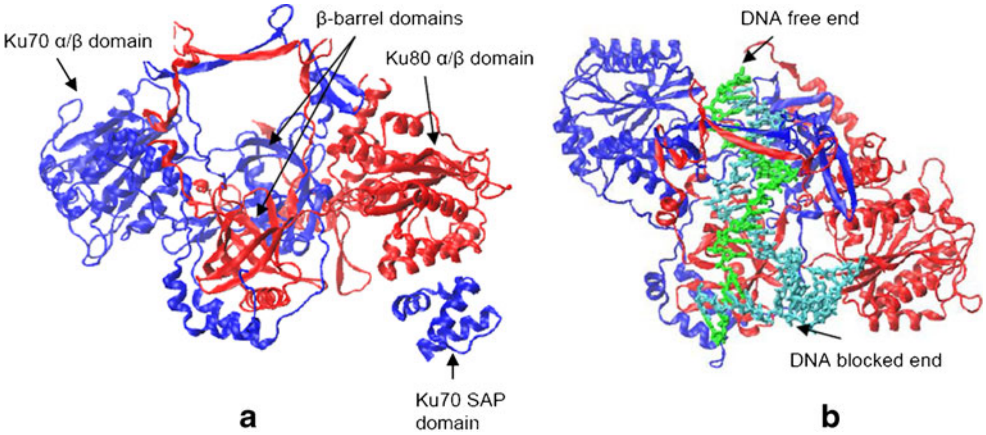


Fig. 1 Crystal structures of human Ku heterodimer (**a**) and its complex bound with DNA (**b**). Ku70 and Ku80 are colored blue and red, respectively. DNA is in licorice representation, with one end blocked with a three-way junction [9]. K70-SAP domain is attached to Ku80 α/β domain in apo-structure (**a**), but is dispatched in DNA bound structure (**b**)

with a helix-extended loop-helix structure can form favorable electrostatic and hydrophobic interactions with either the major groove or the minor groove of DNA.

Methods

Structural modeling

The atomic structure of the human Ku70-SAP domain has been determined by NMR [13] with code 1JJR in RCSB Protein Data Bank. It has three α-helices, encompassing residues 562–570, 578–587, and 596–606, with notations Ha, Hb, and Hc, respectively (left panel of Fig. 2). The inter-helical loops (La and Lb) are not well defined in experiment but contain several basic residues that could play important roles in DNA binding. It has a total charge of +2.0. Following the chemical shift perturbation experiment [13], we used two 10 base pair palindromic DNA duplexes as substrates, to ensure the binding of Ku70-SAP to each end of the DNA forms the same complex. One duplex was obtained from RCSB Protein Data Bank, with code 1CQO and upper strand sequence 5'-GCGTTAACGC-3' [19] (denoted as AT10 henceforth). Another duplex was constructed by the *nucgen* facility of AMBER software package [20], with upper strand sequence 5'-GCGCGCGCGC-3' (denoted as GC10). Both DNA duplexes have blunt ends and are in canonical B-form, and each has a charge of −18.0.

Chemical shift perturbation experiments identified several regions of SAP that interact with DNA [13]. These regions have the most conserved Arg or Lys residues. To aid the construction of starting structures of SAP-DNA complex, we used the adaptive Poisson-Boltzmann solver (APBS) method [21] to calculate the electrostatic potential surface of SAP. The mapped surface (Fig. 2) is consistent with the previous work [13], with two regions dominated by positive potential: the first covers helix Hb and loop Lb, the second covers the loop La. The two regions are nearly fused to each other on the surface of SAP (right panel of Fig. 2). Since DNA binding causes the largest chemical shift changes to occur in these positive electrostatic regions [13], our starting structures were constructed by manually placing the molecules AT10 and GC10 to these sites, by using the Weblab Viewer Lite software (Molecular Simulations Inc., San Diego, CA). The two molecules in each structure were docked with distances (between the closest atoms) of about 4–6 Å, with varying orientations and interfaces (Table 1). Two structures among them were built based on the proposed binding mode in [13].

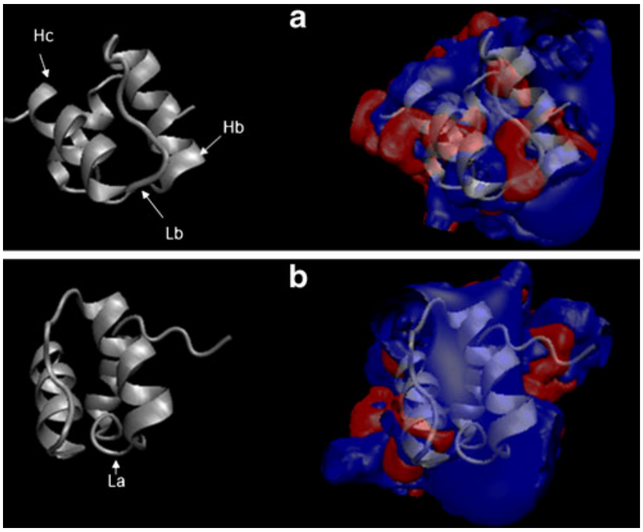


Fig. 2 Electrostatic properties of Ku70-SAP domain. Potential isocontours are shown at +3 kT/e (blue) and −3 kT/e (red) and obtained by APBS method at 150 mM ionic strength with a solute dielectric of 1 and a solvent dielectric of 78.5 [19]. (**a**) Side view of the patch along Helix Hb and Loop Lb. (**b**) Front view of the patch

Flexible LMOD docking

To test whether the hypothesized binding mode is actually the most energetically favored and to search for other possible binding modes, a full-flexible low frequency vibrational mode (LMOD) docking method [22–25] was first applied to these manually constructed starting structures. The LMOD method was designed to enable an exhaustive exploration of the potential energy hypersurface of molecules, based on eigenvector following (or mode following) methods [22], and has been found to be very efficient in some computational chemistry domains such as protein loop optimization, conformational analysis of complex systems and flexible docking [22–25]. The LMOD protocol used in this work was found by many trial-and-error calculations and could significantly build up the potential binding affinity between two macromolecules. Validation of this method was conducted with a system with known X-ray crystal structure (Supplementary material).

All docking calculations and the following simulations and analyses were performed with the AMBER 9 suite of programs [20] together with the Stony Brook modification ff99 force field [26]. The implicit solvent model of Onufriev-Bashford-Case [27] was used to represent the electrostatics of aqueous solution at the docking stage. GB/SA methodology (igb=5, cut=16.0 Å, rgbmax=12.0 Å, and surften=0.005 kcal mol^{-1}·Å^{-2}) was used and the salt concentration was set to 0.2 M, to represent the solvent and ionic effects, respectively. We found utilizing this model was essential to maintain the conformations of the molecules during the fully flexible docking process. During this process, the atoms of the molecules were subjected only to the forces imposed by the chosen force field and solvent model. Our preliminary calculation indicated that a simple charge screening function

Table 1 Interfaces and orientations of the starting structures of SAP-DNA complex. Interface 1 of SAP covers the region along Hb and Lb, and interface 2 covers regions along La (see Fig. 2). 'm' means the minor groove of DNA, while 'M' means the major groove. The duplexes are manually docked to SAP with their helical axis either parallel to the directions of the loops (||) or perpendicular to them (⊥). AT1c and GC1c were built based on the proposal of [13]

Name	SAP interface	DNA interface	DNA orientation	Name	SAP interface	DNA interface	DNA orientation
AT1a	1	m	||	GC1a	1	m	||
AT1b	1	M	⊥	GC1b	1	M	⊥
AT1c	1	M	||	GC1c	1	M	||
AT1d	1	m	⊥	GC1d	1	m	⊥
AT2a	2	m	||	GC2a	2	m	||
AT2b	2	M	⊥	GC2b	2	M	⊥
AT2c	2	M	||	GC2c	2	M	||
AT2d	2	m	⊥	GC2d	2	m	⊥

(ε=4r) could not reproduce reasonable binding geometries of our system, although several previous LMOD applications were based on this simple treatment [22–25].

In each of LMOD docking iterations, ten lowest vibrational modes were calculated, and three of them were randomly chosen to make ZIG-ZAG moves (in the range of 0.02-2.0) [20]. The system was carried out in this manner over a series of energy barriers until the lower energy endpoints were reached. They were then subjected to minimization with limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) quasi-Newton algorithm [28] to 0.1 kcal (mol·Å)$^{-1}$ of gradient RMS, and were subsequently collected in the pool with 50.0 kcal mol^{-1} energy window. The next iteration started with a structure chosen from this pool by metropolis Monte Carlo method. The Ku70-SAP, AT10 and GC10 were fully flexible, while the protein part was allowed ive5 times of explicit translation and rotation in each of the iterations. The 10 lowest vibrational modes were updated every 10 iterations till convergence was reached. Each starting structure was subjected to this same docking protocol for three times with three different random seeds (314159, 1000, and 2000).

Explicit solvent MD simulation

The docked structures with the most energy gains were placed into truncated octahedron periodic boxes of TIP3P water molecules, with counter ions to neutralize the total charge. The distances between the edges of the water box and the closest atom of the solute were at least 10 Å in all cases. The systems were minimized by 500 steps of minimization, with the solute constraint with 2.0 kcal mol^{-1}·Å^{-2} to all solute atoms. The particle mesh Ewald (PME) method was used to treat long-range electrostatic interactions, and bond lengths involving bonds to hydrogen atoms were constrained using SHAKE. The time-step for all MD simulations was 2.0 fs, with a direct-space, non-bonded cutoff of 9.0 Å. Translational center-of-mass motions were removed every 1000 steps. With the solute constraint with 2.0 kcal mol^{-1}·Å^{-2} to all solute atoms, canonical ensemble (NVT)-MD was carried out for 35 ps. This was followed by five rounds of 600 step energy minimizations on the entire system, by reducing the solute restraints gradually; 2.0 kcal mol^{-1}·Å^{-2} restraints on all solute atoms were again used during heating the entire system to 300 K. Then, with a time constant of 2.0 ps for heat-bath coupling, solute restraints were reduced gradually over 50 ps, while the systems underwent isothermal isobaric ensemble (NPT)-MD simulations to adjust the solvent density. After the equilibration phase, the production phase without any constraint was followed at 300 K and 1 atm for up to 5 ns. The last 3 ns trajectories were used to extract the snapshots for binding free energy calculation and structural analysis.

Binding free energy calculation

The binding affinity was approximated by MM-PBSA methodology [29] from each single trajectory, in which the protein and DNA structures were taken from the complex simulation. Snapshots taken every 20 ps from the last 3 ns of production phase simulation were evaluated for a total of 150 structures. The molecular mechanics energies (ΔE^{vdw}, ΔE^{elec}, and ΔE^{inter}) were evaluated in a single MD step in the *Sander* module using an infinite cutoff for nonbonded interactions. The electrostatic component of solvation free energy (ΔG^{PB}) was computed by the finite difference Poisson-Boltzmann method [30, 31], as implemented in the AMBER programs package. The reference system had a solvent dielectric of 1 and 0 M salt concentration. The solvated system had a solvent dielectric of 80 and 100 mM salt concentration. The non-polar contribution to the solvation free energy was approximated with the commonly used solvent-accessible surface area (SASA) model, $\Delta G^{SA}=\gamma(SASA)+\beta$, where $\gamma=0.00542$ kcal mol^{-1}·Å^{-2} and $\beta=0.92$ kcal mol^{-1} [32]. The SASA was estimated with a 1.4 Å solvent probe radius as implemented in *Sander*. Conformational entropies (ΔS) were calculated through normal-mode analysis with more sparse samples (15 structures of each trajectory), as suggested by AMBER manual.

To probe the key contributions and hot-spots of interfacial residues, MM-GBSA approach was applied to decompose the total binding affinity into contributions from each residue of Ku70-SAP and DNA. Focuses are on those that make the most contributions. We used GB=1 and the default parameters of the AMBER programs package.

Results and discussion

MD trajectory analysis of the unbounded monomers

For the MD simulations of unbound systems of SAP and AT10, the time-series of the root mean squared deviation (RMSD) of backbone heavy atoms are given in Fig. 3, with comparison with their NMR experimental structures. The structure of SAP are well maintained during the 5 ns simulation, with RMSD of 1.58 Å. The simulation of the AT10 duplex shows much larger fluctuations. The RMSD values of the backbone heavy atoms of AT10 vary between 1.0 Å and 3.3 Å, from the experimental starting structure, with more frequent oscillations. These results are comparable to those with similar simulations in literature. The canonical B-form, Watson-Crick hydrogen bonds, and the planarity of the bases are well maintained during the 5 ns MD simulation without constraints.

To test the quality of implicit solvent we used for LMOD docking, we carried out 1 ns GB/SA simulation for each monomer, starting directly from their 500 steps minimized structures, i.e., skipping the equilibration stage. Their RMSD values are shown in Fig. 4, indicating a reasonably accurate model alternative to the more expensive explicit model.

LMOD docking and binding modes prediction

With the LMOD protocol described above, most starting structures reach their low energy endpoints within 100 docking iterations, with few exceptions of converge till about 200 iterations. This represents an amenable task for this system. However, we find the ending structures are very sensitive to the starting geometries and searching parameters. Table 2 gives the minimized energies E and energy gains ΔE ($E^{complex}$ - $\Sigma E^{monomer}$) of the lowest energy structures of SAP-AT10 obtained by docking. For each of the eight starting structures, the ending structures are different with the same protocol but different random seeds. This implies the conformational space of this system is probably too rugged to allow an exhaustive exploration; especially, with all atoms fully flexible, the system is easily trapped in local minima on the potential energy hypersurface. Nevertheless, for most cases, this protocol can significantly build up the potential binding affinity between two macromolecules (Tables 2 and 3). From a further test of this method on a similar system with known bound structure (Supplemental material), we find that by carefully tuning the step size of ZIG-ZAG move, successful prediction of the energetic as well as conformational features of the published complex from unbound structure can be achieved.

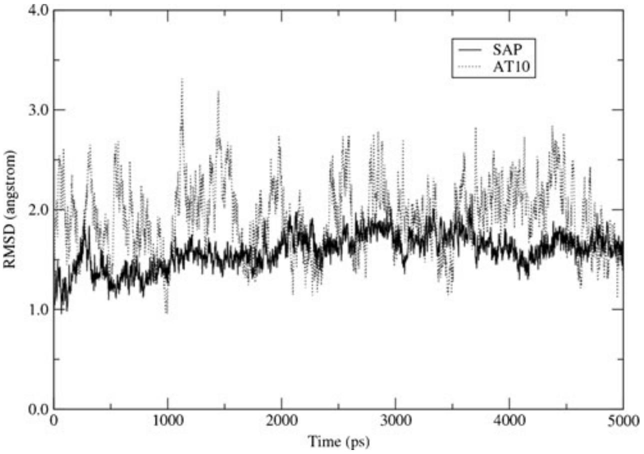


Fig. 3 Time-series of RMSD of backbone heavy atoms of the SAP and AT10 monomers over 5 ns of explicit solvent MD simulations, comparing to the NMR experimental results. The trajectory of SAP is more stable than that of DNA duplex

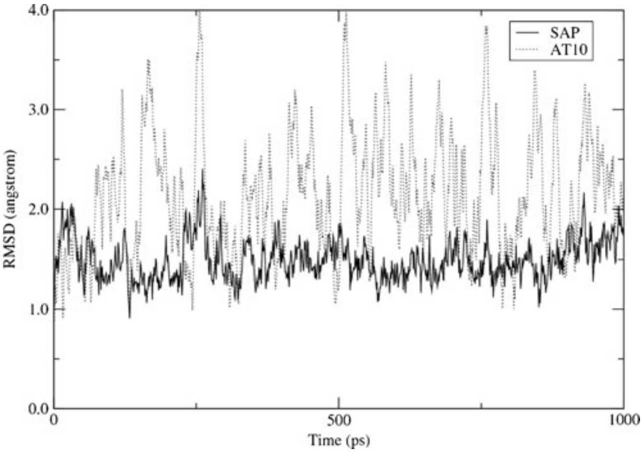


Fig. 4 Time-series of RMSD of backbone heavy atoms of the SAP and AT10 monomers over 1 ns of implicit solvent MD simulations, compared to the NMR experimental results

Table 2 Minimized energies E and energy gains ΔE (inside the brackets) of the results of docking of SAP-AT10. Dock 0, 1 and 2 denote the docking protocols with different random seeds (314159, 1000, 2000, respectively). Energy unit: kcal mol^{-1}

	AT1a	AT1b	AT1c	AT1d
Dock 0	−6211.2(−37.2)	−6189.3(−9.6)	−6253.4(−83.1)	−6221.6(−70.3)
Dock 1	−6206.2(−32.1)	−6197.4(−0.1)	−6219.5(−60.0)	−6217.1(−20.6)
Dock 2	−6210.0(−42.4)	−6203.9(−37.4)	−6209.7(−59.7)	−6201.8(−16.7)
	AT2a	AT2b	AT2c	AT2d
Dock 0	−6233.9(−47.3)	−6233.3(−81.3)	−6234.6(−103.7)	−6195.7(−2.6)
Dock 1	−6205.6(−20.3)	−6214.8(−25.1)	−6210.2(−49.2)	−6190.6(−15.4)
Dock 2	−6247.4(−75.5)	−6191.9(−6.4)	−6227.8(−57.6)	−6210.5(−36.8)

The lowest energy structure (AT1c0, obtained from AT1c by dock 0) among this search is obtained by a starting structure built upon the binding mode proposed by Zhang et al. [13]. While the detailed analysis of this binding mode is presented in the following section, it is worth noting that another complex AT2c0, starting from a totally different conformation, also converges into this mode. Figure 5 illustrates the energy gains of these two structures along the docking paths. After converged they acquire energy compensation of about 80 and 100 kcal mol^{-1}, respectively, indicating the potentency of the high binding affinity of this binding mode. Visualization of the docking paths also reveals the similarity of the protein-DNA interaction of these two complexes. Though the AT10 duplex is manually docked at interface 1 of SAP in AT1c0 (Fig. 2 and Table 1), its first close contact with SAP happens with one end at interface 2, after a series of low modes driven translations/rotations of SAP. Then the major groove and the other end of AT10 get involved in the interaction with SAP at interface 1. The converged structure of AT2c0 is obtained in the same way, but with much less movement of SAP due to its favorable starting orientation (Fig. 5).

An advantage of LMOD flexible docking is that no restraint is needed when the protein/DNA molecules are running along the docking path. Other popular protein-DNA docking programs such as HADDOCK [33] need extensive restraints to maintain the integrity of the DNA structure, which is highly fragile at abnormal conditions. In the future if experimental information becomes available, direct comparisons between the LMOD flexible docking approach and the HADDOCK method could be made. It should be noted that in LMOD docking the DNA duplex occasionally experiences structural damage, such as the opening of the base pairs at the terminals. However, we observed most of this damage can be remedied by the following energy minimization in each iteration. In addition, structures with damage generally have higher energies than the integral structures and are filtered out by the assigned energy window. Consequently the conformational features of both protein and DNA are well maintained in all 48 final structures in Tables 2 and 3.

With four different base pairs in GC10, the most favorable binding mode described above was visited again in the proposed starting structure (GC1c0). It gains the biggest binding energy (ΔE) among the 24 ending structures (Table 3). However, another structure, GC2a0, has a lower E than GC1c0.

Binding affinity and binding mode analysis with explicit solvent

Since the GBSA solvent model for LMOD docking includes all essential molecular mechanical forces and solvation effect, the entity ΔE in Tables 2 and 3 is quite similar to the binding energy $\Delta G^{MM\text{-}GBSA}$ in standard MM-GBSA free energy analysis [20]. This implies that in most of the ending structures the Ku70-SAP domain and DNA duplexes are favorably bound. Our explicit solvent calculations and binding affinity analyses were performed on the structures with ΔE<−45.0 kcal mol^{-1} in Tables 2 and 3. Totally 22 complexes were chosen from the docking results.

Table 3 Minimized energies E and energy gains ΔE (inside the brackets) of the results of docking of SAP-GC10. Energy unit: kcal mol^{-1}

	GC1a	GC1b	GC1c	GC1d
Dock 0	−6925.5(−77.1)	−6904.3(0.4)	−6957.3(−88.2)	−6914.3(−40.0)
Dock 1	−6930.1(−40.5)	−6943.0(−61.0)	−6894.9(0.9)	−6920.4(−45.7)
Dock 2	−6914.5(−46.2)	−6908.5(−25.5)	−6899.8(−44.2)	−6926.1(−50.1)
	GC2a	GC2b	GC2c	GC2d
Dock 0	−6963.8(−80.7)	−6950.3(−72.3)	−6890.1(0.0)	−6918.4(−46.9)
Dock 1	−6935.6(−54.6)	−6928.9(−45.5)	−6929.0(−36.4)	−6903.8(−17.5)
Dock 2	−6908.1(−0.1)	−6956.6(−46.9)	−6915.9(−43.0)	−6898.3(−31.3)

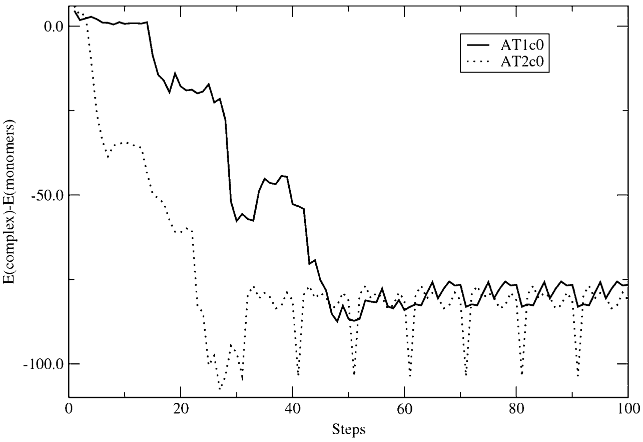


Fig. 5 Energy gains of complexes AT1c0 and AT2c0 along the paths of docking. AT1c0 converged at the 61st iteration and AT2c0 converged at the 31st iteration

In MD simulations, we find all bound complexes undergo various degrees of conformation and orientation adjustments. Figure 6 shows the RMSD of main chains of the three most stable complexes during 5 ns MD simulation, compared with the docked structures. These adjustments are comparable to those of the unbound monomers (Fig. 3). However, the magnitudes of the variation of RMSD are reduced significantly, both of the protein and DNA, demonstrating less flexibility of the molecules in complex. This phenomenon has been found in related simulations [34].

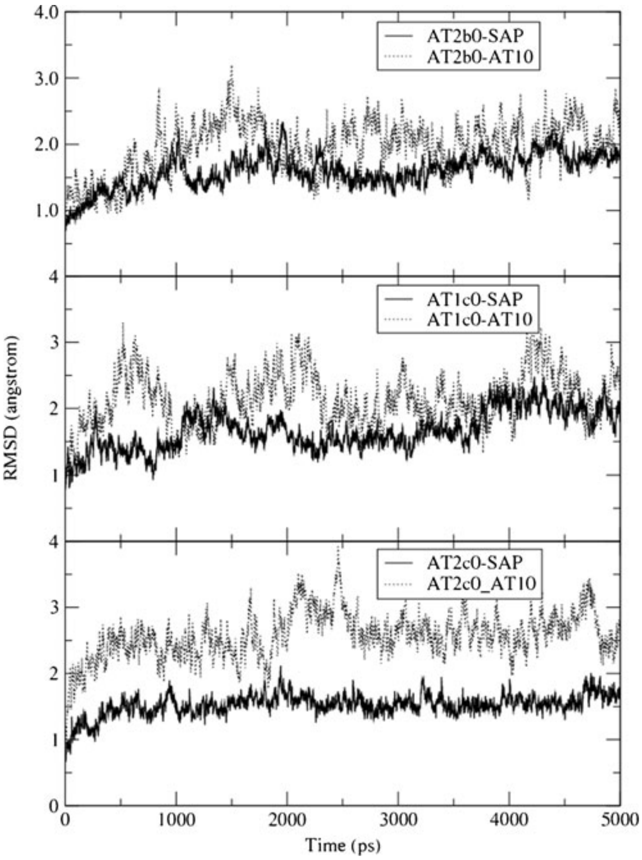


Fig. 6 Time-series of RMSD of backbone heavy atoms of three stable complexes over 5 ns of explicit solvent MD simulations, compared to their docked structures. The fluctuations of SAP and AT10 are smaller than those observed in monomer simulations (Fig. 3)

Compared to the energies obtained by LMOD docking procedure, the binding energies $\Delta G^{MM\text{-}PBSA}$ calculated by MM-PBSA method with trajectories from MD simulations are significantly scaled down. Particularly, when configuration entropy loses [18] are considered, most of the complexes with ΔE in the range (−60.0, -45.0) in Tables 2 and 3 could not be regarded as favorably associated, because the binding free energies of the complexes are positive. As all binding modes discussed below involve the direct contacts of the flexible loops of SAP with DNA, the molecules in the bound state can only access a narrower range of conformations than in the free state. It is common in protein-ligand system that the configuration entropy drops in complex and thus opposes binding [18]. Table 4 shows the results of MM-PBSA calculation for the complexes with negative binding free energies. This indicates the LMOD docking tends to overestimate binding affinities, like many other docking programs [18].

From these favorably bound structures four binding modes can be recognized (Fig. 7), among which modes I and II are major groove association, while III and IV are minor groove association. The assignment of these modes is done on the basis of visual similarity at the binding interfaces. In the following part of this section we discuss their structural and energetic features and particularly the differences caused by the sequence change of DNA.

a) Major groove interactions

Binding mode I has the most favorable binding free energy, which is adopted by AT2b0 with a few orientation changes from its starting conformation (Table 1). AT10 is located at interface 1, with its helix axis nearly parallel to La, major groove pointing to SAP, one end attached to the end of Lb, and the other end interacting with Ha (Fig. 7I). The overall electrostatic energy ($\Delta E^{elec}+\Delta G^{PB}$) is −12.44 kcal mol^{-1}, while the mean value of the van der Waals and hydrophobic interaction energies ($\Delta E^{vdw}+\Delta G^{SA}$) is −40.83 kcal mol^{-1}. This seems counter-intuitive since the electrostatic contribution should dominate the binding thermodynamics for this highly charged system. However, as has been observed in other protein-DNA and DNA-drug systems, there is an anti-correlation between the direct electrostatic component and the desolvation expense [35–37]. This desolvation expense is caused by the repulse of water molecules from the proximity of DNA by protein-DNA association, as water is essential to

Table 4 Energy contributions (kcal mol^{-1}) to the free energy of binding between SAP and DNA. Calculated by MM-PBSA and normal mode analysis methods. Binding modes are assigned based on visually overall conformational similarity

Complex	ΔE^{vdw}	ΔE^{elec}	ΔG^{PB}	ΔG^{SA}	$\Delta G^{MM\text{-}PBSA}$	$T\Delta S$	ΔG	Binding mode
AT2b0	−34.90	−1395.81	1383.37	−5.93	−53.27	33.55	−19.72	I
AT1c0	−28.41	−1382.78	1366.82	−5.53	−49.90	30.49	−19.41	II
AT2c0	−37.44	−1397.68	1387.54	−6.47	−54.05	35.65	−18.40	II
AT2a2	−58.11	−1261.26	1276.09	−7.07	−50.35	34.05	−16.30	III
AT2a0	−33.92	−1407.85	1401.50	−5.32	−45.59	33.34	−12.25	III
AT1d0	−48.68	−1159.62	1175.42	−5.87	−38.76	36.77	−1.99	IV
GC1c0	−26.88	−1264.64	1251.22	−4.95	−45.25	33.20	−12.05	II
GC2a0	−32.08	−1355.06	1353.27	−5.64	−39.51	31.43	−8.08	IV
GC2b0	−17.34	−1264.87	1244.18	−4.26	−42.29	34.25	−8.04	II
GC1c2	−17.10	−1166.48	1151.88	−3.05	−34.76	28.66	−6.10	II
GC1d2	−39.01	−1302.89	1316.41	−5.33	−30.81	28.01	−2.80	III

maintain the integrity of the highly charged DNA double helix. Thus, the low contribution of overall electrostatics to the binding energy is a general trend in DNA-ligand binding systems [37]. Five positively charged Lys residues 556, 570, 575, 595 and 596 of SAP [13] maintain close contacts with the backbone of DNA during the MD simulation, with energetic contributions −1.16, -4.40, -6.56, -7.89, and −7.95 kcal mol^{-1}, respectively. A hydrogen bond between Thr 572 and a phosphate oxygen atom is also well maintained. The residue Thr 572 contributes −2.52 kcal mol^{-1} to the overall binding affinity.

No close contact between the base pairs and amino acids is observed, and there is an apparent

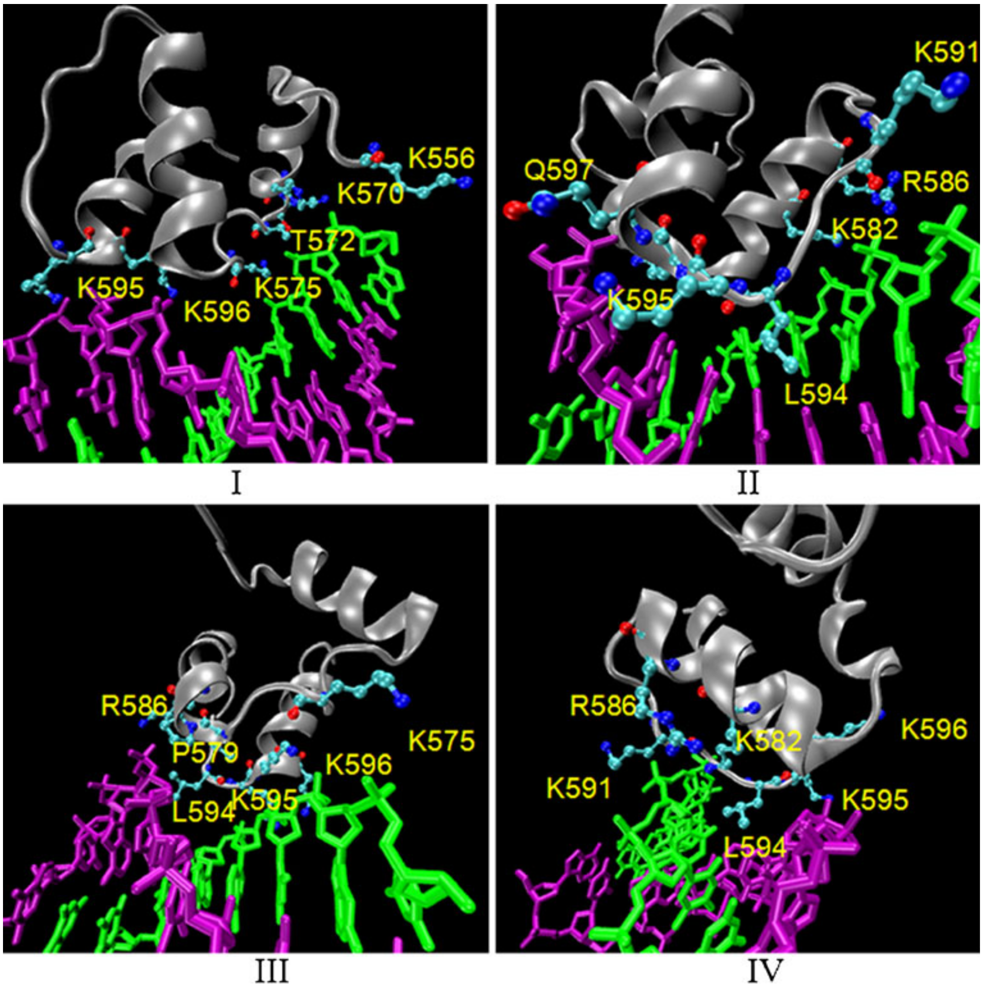


Fig. 7 Four binding modes of SAP-DNA complex (final snapshots of the 5 ns MD simulations of AT2b0, AT1c0, AT2a2, and AT1d0, respectively). The residues of SAP involved in electrostatic, hydrogen bonding, and hydrophobic interactions are given in CPK form

gap at the center of the SAP-AT10 interface. This means the major van der Waals and hydrophobic contributions come from the phosphate backbone interaction. Therefore the binding affinity of this mode should be non-sequence specific. It should be mentioned that this mode has not been visited by a stable SAP-GC10 complex. Constructing such a complex manually and then testing its stability is possible but has not been pursued in this study.

AT1c0 and AT2c0 share binding mode II as discussed in the previous section. Favorable electrostatic interactions are implied between the four Lys residues 582, 591, 595, 596, and one Arg residue 586 with the phosphate backbones from their persistent close contacts (Fig. 7II). The total contributions of these five residues to the binding affinities of these two systems are −31.10 and −35.01 kcal mol^{-1}, respectively. A hydrogen bond between Gln 597 and a phosphate oxygen atom is also well maintained during the MD simulation. This residue contributes −5.08 and −6.41 kcal mol^{-1}, respectively, to the binding affinities of the two systems. These interactions are similar to those in the binding mode Zhang et al. proposed [13]. However, in the modeled binding mode II, the residues 582 and 586 in helix Hb are not involved in groove interaction, but in phosphate backbone interaction. Furthermore, detailed analysis of the bound structures of these two complexes also unravels some subtle differences of the interactions. As AT2c0 starts docking from interface 2 of SAP, the AT10 duplex is a bit less stretched into interface 1 than in AT1c0, and there is one obvious interaction between Lys 575 and the end of AT10 backbone at interface 2, which is not present in AT1c0. The contributions of this residue are −1.07 and 0.02 kcal mol^{-1}, respectively, in the two systems. This interaction as well as the location of AT10 duplex in AT2c0 helps the AT10 duplex to secure more contact area with SAP than in AT1c0. However, although AT2c0 has a higher binding energy $\Delta G^{\text{MM-PBSA}}$, it also has a higher configuration entropy lose due to its tighter association [18], therefore, its binding free energy is a bit lower than AT1c0 (Table 4).

Interestingly, this mode is visited by three stable complexes of SAP-GC10 with less favorable binding affinities (GC1c0, GC2b0, and GC1c2 in Table 4). There are some subtle differences of interaction and relative orientation among them, like the two complexes discussed above. Comparing their structures with SAP-AT10 complexes, the GC10 duplex in these complexes is not as flexible as its counterpart AT10. This can be understood from the facts that, in the two SAP-AT10 complexes, AT10 is fully attached to the binding interfaces of SAP at both ends, but in three SAP-GC10 complexes, only one end of GC10 is closely attached while the other is not. Obviously the interactions between the backbones of GC10 and SAP basic residues are not as optimal as in SAP-AT10.

In the above major groove associated stable complexes, no strong interaction between amino acids and base pairs exist, which are common in other well-known protein-DNA binding motifs (reviewed in [38]). However, we observe persistent close contacts between some hydrophobic residues on the surface of SAP and some base pairs with hydrophobic sides (particularly, the four Thymine bases in AT10). The Leu 594 with an isobutyl side chain is among the few non-polar residues that are located on the surface of SAP, i.e., making no contribution to the integrity of the hydrophobic core of SAP [13]. Table 5 lists the average distances of the center of the side chain of Leu 594 to the methyl groups of four Thymine bases in AT1c0 and AT2c0 during their last 3 ns MD simulations. The distances are in the range of optimal van der Waals interaction, and the contributions of this residue are −2.82 and −2.79 kcal mol^{-1}, respectively, in these two systems. Due to the rigidity of GC10 and the lack of hydrophobic bases like Thymine groups in AT10, such interactions are absent in the three SAP-GC10 complexes discussed above. This should be the cause of the significantly lower contribution of van der Waals force and hydrophobic interaction in these complexes (Table 4).

b) Minor groove interactions

Mode III adopted by AT2a2 (Table 4) has a comparably high binding affinity as the above discussed modes, due to the favorable electrostatic association of residues Lys 575, 595, 596, and Arg 586 with phosphate backbone. They together contribute −23.18 kcal mol^{-1} to the binding energy. The partners in this mode adopt a similar orientation as in mode II, i.e., the DNA molecules cover the positive potential patches at interfaces 1 and 2. The non-polar contribution caused by solvent accessible surface area deduction of this complex is the largest among all stable structures, which

Table 5 The average distance (Å) between the center of isobutyl group of Leu 594 to the centers of methyl group of Thymine bases in two stable complexes during the last 3 ns MD simulations. All distances are in the range that favors van der Waals interaction

Complex	CH_3 of Thy 4	CH_3 of Thy 5	CH_3 of Thy 14	CH_3 of Thy 15
AT1c0	4.73	7.05	6.48	4.99
AT2c0	4.87	6.82	6.89	5.11

is −7.07 kcal mol^{-1}. This is mainly contributed by the non-polar residues that are located on the surface of SAP, including the above mentioned Leu 594. The other two aliphatic groups, Val 578 and Pro 579, located at the turn of La and Hb (Fig. 2), are deeply immersed into the minor groove and maintain close contacts with the hydrophobic bases at this side. These three non-polar residues together contribute −11.83 kcal mol^{-1} to the binding energy of AT2a2. The largest van der Waals contribution to the binding affinity among these structures, which is −58.11 kcal mol^{-1}, is also related to this close packing between SAP and AT10. However, we found the overall electrostatic interaction of this complex is +14.83 kcal mol^{-1}, e.g., unfavorable to association.

AT2a0 has the same binding mode of AT2a2, but the partners do not pack so closely as in AT2a2. Particularly, there is no contact between the surface hydrophobic groups and bases. Its electrostatic interaction is −6.35 kcal mol^{-1}, e.g., favorable to association, and its van der Waals and hydrophobic contributions are significantly reduced, compared to AT2a2 (Table 4). In GC1d2, a SAP-GC10 complex that bears this binding mode, the direct interaction between SAP and DNA mainly comes from the basic residues from interface 1, with only one such group (Lys 595) from interface 2. The Leu 594 is deeply immersed into the minor groove, which is responsible for its relatively high van der Waals and hydrophobic contributions (Table 4). This residue contributes −8.56 kcal mol^{-1} alone, based on the binding energy decomposition analysis. However, the overall electrostatic interaction of this complex is +13.52 kcal mol^{-1}, which is unfavorable to association.

Mode IV is similar to mode III in many ways. In both of the complexes AT1d0 and GC2a0 characterized as this mode, there are persistent contacts between the basic residues and the DNA backbones; Lys 591 at SAP interface 1 interacts with one phosphate chain while Lys 595 and 596 at interface 2 interact with the other chain, and the aliphatic group Leu 594 between them is deeply sucked into the minor groove. The groups of three charged residues contribute −13.65 and −13.02 kcal mol^{-1}, respectively, to the total binding energies of these two systems, while Leu 594 contributes −6.26 and −4.49 kcal mol^{-1}, respectively. The only difference between this mode and mode III lies on the orientation by which the DNA duplexes are packed on SAP; the helix axes of mode III complexes are parallel to the direction of loop Lb while in mode IV they are perpendicular to it. Like the complexes of mode III, the van der Waals interactions in these two complexes dominate their binding affinity. The combined electrostatic contribution in AT1d0 is +15.80 kcal mol^{-1} unfavorable to binding, and in GC2a0 is −1.19 kcal mol^{-1}, i.e., marginally favorable.

All complexes of major groove binding have significantly favorable electrostatic contributions to binding, while three out of five minor groove binding complexes have unfavorable electrostatic interactions, and the other two have very low favorable contributions (Table 4). This might be related to the different hydration forms of DNA in the major groove and minor groove. Ordered water shell along the DNA minor groove, termed as the spine of hydration, was reported for the d $(CGCGAATTCGCG)_2$ dodecamer [39]. This ordered form definitely needs more energy to compensate the desolvation, which has been discussed above. DNA hydration is a well-discussed subject [40, 41], and, interestingly, the molecular docking combining with MD simulation and binding affinity calculation of this work provides a consistent observation.

Biological implications

From this molecular modeling study on the stable complexes of SAP-DNA, the shape of human Ku70-SAP seems well designed to structurally associate with DNA in either broken or intact configurations. Two positively charged patches of residues, located at interface 1 and 2, respectively, are separated by a hydrophobic patch (Val 578, Pro 579 and Leu 594) on the surface of SAP. Other conserved hydrophobic residues of SAP are in the hydrophobic core and are responsible for the integrity of its structure [13]. The two positively charged patches can either form interactions with the two ends of DNA (AT10 and GC10), as in binding mode II and III, or bind the two sugar-phosphate backbones at the minor groove of DNA in binding mode IV. In either case, the surface non-polar residues, especially Leu 594, can form important van der Waals and hydrophobic interactions with the base pairs. These features can be observed in the SAP domains of Ku70 from several other species [13]. In the Ku70 proteins of hamster, mouse, Gallus, and Xenopus (Fig. 2A of Ref. [13]), sequence alignment indicate the residues of two positively charged patches on the surface of SAP domains, corresponding to Lys 582, 591, 595, 596 and Arg 586 of human Ku70 are well conserved, with some places of interchanging Lys and Arg residues. The intermediate hydrophobic patch on the surface can be observed in hamster and mouse only, with Val 578 and Pro 579 of human Ku70 conserved but Leu 594 replaced by Pro. In Gallus and Xenupos Ku70, Val 578 of human Ku70 is conserved in both species while Pro 579 is conserved only in Xenupos Ku70. Also at the position of Leu 594 of human Ku70, two small residues, Gly and Ser are present, respectively. It would be interesting to investigate how these changes may affect their binding with DNA.

The SAP domain is also found in several other nucleic acid-binding proteins and has been defined as a putative DNA-binding motif [11, 13]. Among the SAP domains of 31 different animal, plant, and fungal proteins (Fig. 1 of Ref. [11]), the structural pattern of two positively charged patches of residues separated by a hydrophobic patch can be observed in most cases. While the two positive charged patches similar to the positions of human Ku70 are clearly presented in call cases, 13 of them have a hydrophobic residue like Leu 594 in human Ku70 immediately before the second positively charged patch, and 28 of them have a hydrophobic residue like Val 578 in human Ku70 at the beginning of helix Hb (Fig. 1 of Ref. [11]). Based on our modeling, this pattern of structure may form a favorable association with DNA like modes II, III, and IV discussed above.

Concluding remarks

In this study the utilization of a low-frequency vibration mode driven molecular docking method combined with the implicit solvent model helped us find several favorable binding modes between the C-terminal DNA-binding domain of Ku70 and DNA. This method demonstrates the capability of building up the potential binding affinity between macromolecules and transforming several distinct starting conformations into one binding mode. In the system used to validate the docking method, the best docking solution achieved very good agreement with the published X-ray crystal structure (Supplemental material). Since the low-frequency vibrational modes represent large-amplitude, concerted atomic movement, this approach does not need to apply the extensive constraints on the molecules, but provides very effective structural perturbations to the systems to overcome energy barriers. It can be used as a fast and reliable way to predict how protein and DNA molecules may interact.

Chemical shift perturbation experiments previously conducted for this system proposed a binding mode of major groove interaction similar to mode II [13]. A limitation of such a method is that the involvement of the nucleotide could not be identified. In fact, all 11 complexes of the four binding modes in Table 4 are consistent with the published large chemical shift changes of SAP residues [13]. It should be noted that for binding mode I, although the residues Lys 570 and Thr 572 are not among the residues list of large chemical shift changes [13], they are close to the segment 573–577. The significant perturbation, found in this region along with the end segment of Lb involved positively charged residues 595–596 (Fig. 4. A in [13]), is in agreement with the interaction of mode I of this work. Compared to the protein-DNA interacting scenario depicted experimentally, the theoretical approach of this work provides more comprehensive information. It would be interesting to conduct experiments to obtain a full picture of such interactions, in which the functions of DNA can be identified.

This computational study strongly supports the DNA-binding capability of the SAP domain of Ku70. In addition, it has been observed that, along with the two positively charged patches on the surface of Ku70-SAP, the flexible linker section from residue 536 to 560 also contains at least two other positively charged patches of amino acids [13]. Along with the binding modes of SAP-DNA predicted in this study, it is also interesting to verify experimentally whether the SAP domain of Ku70 is the first functional group that recognizes the broken double strand of DNA. If so, this domain may function as an antenna to grab the free end of DNA, and then, with the assistance from the competitive bindings of different domains and modes, drag it into the cradle of the Ku heterodimer. Our further investigation will be conducted in this direction, in which the full length of Ku70, and eventually Ku80, will be considered, interacting with longer DNA duplexes.

Acknowledgments The authors thank Dr. István Kolossváry for kind help on LMOD calculations. Part of computations was performed on computers at TLC^2 of the University of Houston. Funding for this study was provided by the NASA Space Radiation Program, Risk Assessment Project.

References

1. Downs JA, Jacksons SP (2004) A means to a DNA end: the many roles of Ku. Nature Rev Mol Cell Biol 5:367–378
2. Anderson CW, Carter TH (1996) The DNA-activated protein kinase-DNA-PK. In: Jessberger R, Lieber MR (eds) Molecular Analysis of DNA Rearrangements in the Immune System. Springer, Heidelberg, pp 91–112
3. Fewell JW, Kuff EL (1996) Intracellular redistribution of Ku immunoreactivity in response to cell–cell contact and growth modulating components in the medium. J Cell Sci 109:1937–1946
4. Mahaney BL, Meek K, Lees-Miller SP (2009) Repair of ionizing radiation-induced DNA double-strand breaks by non-homologous end-joining. Biochem J 417:639–650
5. Weterings E, Chen DJ (2008) The endless tale of non-homologous end-joining. Cell Res 18:114–124
6. Polo SE (2011) Jackson SP (2011) Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. Genes Dev 25:409–433
7. Lieber MR (2008) The mechanism of human nonhomologous DNA end joining. J Biol Chem 283:1–5
8. Lieber MR, Lu H, Gu J, Schwarz K (2008) Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and

ligase of vertebrate non-homologous DNA end joining: relevance to cancer, aging, and the immune system. Cell Res 18:125–133

9. Walker JR, Corpina RA, Goldberg J (2001) Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. Nature 412:607–614
10. Rivera-Calzada A, Spagnolo L, Pearl LH, Llorca O (2007) Structural model of full-length human Ku70-Ku80 heterodimer and its recognition of DNA and DNA-PKcs 8:56–62
11. Aravind L, Koonin EV (2000) SAP — a putative DNA-binding motif involved in chromosomal organization. Trends Biochem Sci 25:112–114
12. Wang J, Dong X, Reeves WH (1998) A model for ku heterodimer assembly and interaction with DNA. J Biol Chem 273:31068–31074
13. Zhang Z, Zhu L, Lin D et al. (2001) The three-dimensional structure of the C-terminal DNA-binding domain of human Ku70. J Biol Chem 276:38231–38236
14. Dynan WS, Yoo S (1998) Interaction of Ku protein and DNA-dependent protein kinase catalytic subunit with nucleic acids. Nucleic Acids Res 26:1551–1559
15. Cucinotta FA, Pluth JM, Anderson JA et al. (2008) Biochemical kinetics model of DSB repair and induction of gamma-h2ax foci by non-homologous end joining. Radiat Res 169:214–222
16. Durante M, Cucinotta FA (2008) Heavy ion carcinogenesis and human space exploration. Nat Rev Cancer 8:465–472
17. Becker OM, MacKerell AD Jr, Roux B, Watanabe M (eds) (2001) Computational Biochemistry and Biophysics. Dekker, New York
18. Gilson MK, Zhou HX (2007) Calculation of protein-ligand binding affinities. Annu Rev Biophys Biomol Struct 36:21–42
19. Smith JA, Tsui VT, Chazin WJ, Case DA (to be published) NMR structure of the palindromic dna decamer d(GCGTTAACGC)$_2$
20. Case DA, Darden TA, Cheatham TE III et al. (2006) AMBER 9. University of California, San Francisco
21. Baker NA, Sept D, Joseph S et al. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA 98:10037–10041
22. Kolossváry I, Guida WC (1996) Low mode search. An efficient, automated computational method for conformational analysis: Application to cyclic and acyclic alkanes and cyclic peptides. Am Chem Soc 118:5011–5019
23. Kolossváry I, Guida WC (1999) Low-mode conformatinoal search elucidated: application to C39H80 and flexible docking of 9-deazaguanine inhibitors into PNP. J Comput Chem 20:1671–1684
24. Kolossváry I, Keserü GM (2001) Hessian-free low-mode conformational search for large-scale protein loop optimization: Application to c-jun Nterminal kinase JNK3. J Comput Chem 22:21–30
25. Keserü GM, Kolossváry I (2001) Fully flexible low-mode docking: Application to induced fit in HIV integrase. J Am Chem Soc 123:12708–12709
26. Hornak V, Abel R, Okur A et al. (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins 65:712–725
27. Onufriev A, Bashford D, Case DA (2004) Exploring native states and large-scale conformational changes with a modified Generalized Born model. Proteins 55:383–394
28. Liu DC, Nocedal J (1989) On the limited memory BFGS method for large scale optimization. Math Prog 45:503–528
29. Srinivasan J, Cheatham TE III, Cieplak P et al. (1998) Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices. J Am Chem Soc 120:9401–9409
30. Sharp KA, Honig B (1990) Electrostatic interactions in macromolecules - theory and applications. Annu Rev Biophys Biophys Chem 19:301–332
31. Sharp KA, Honig B (1990) Calculating total electrostatic energies with the nonlinear Poisson-Boltzmann equation. J Phys Chem 94:7684–7692
32. Cheatham TE III, Srinivasan J, Case DA, Kollman PA (1998) Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution. J Biomol Struct Dynam 16:265–280
33. van Dijk M, van Dijk ADJ, Hsu V et al. (2006) Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. Nucleic Acids Res 34:3317–3325
34. Gohlke H, Kiel C, Case DA (2003) Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. J Mol Biol 330:891–913
35. Misra VK, Honig B (1995) On the magnitude of the electrostatic contribution to ligand-DNA. Proc Nat Acad Sci USA 92:4691–4695
36. Singh SB, Kollman PA (1999) Calculating the absolute free energy of association of netropsin and DNA. J Am Chem Soc 121:3267–3271
37. Shaikh SA, Ahmed SR, Jayaram B (2004) A molecular thermodynamic view of DNA–drug interactions: a case study of 25 minor-groove binders. Arch Biochem Biophys 429:81–99
38. Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein-DNA complexes. Genome Biol 1:1–37
39. Kochoyan M, Leroy JL (1995) Hydration and solution structure of nucleic acids. Curr Opin Struct Biol 5:329–333
40. Berman HM (1991) Hydration of DNA. Curr Opin Struct Biol 1:423–427
41. Jayaram B, Jain T (2004) The role of water in protein-dna recognition. Annu Rev Biophys Biomol Struct 33343–361

J Mol Model (2012) 18:2175–2184
DOI 10.1007/s00894-011-1233-y

ORIGINAL PAPER

Influence of point defects on the electronic properties of boron nitride nanosheets

Ernesto Chigo Anota · Ramses E. Ramírez Gutiérrez · Alejandro Escobedo Morales · Gregorio Hernández Cocoletzi

Received: 4 June 2011 / Accepted: 24 August 2011 / Published online: 27 September 2011

Abstract Density functional theory was utilized to study the electronic properties of boron nitride (BN) sheets, taking into account the presence of defects. The structure considered consisted of a central hexagon surrounded by alternating pentagons (three) and heptagons (three). The isocoronene cluster model with an armchair edge was used with three different chemical compositions. In the first structure, three B–B bonds were formed where one B in the dimer was part of the central hexagon. In the second structure, three N–N–N bonds were formed at the periphery of the cluster, around the central hexagon. In the third structure, three N–N bonds were formed in a similar fashion to the first model. Our results indicated that the third structure was the most stable configuration; this exhibited planar geometry, semiconductor behavior, and ionic character. To explore the effects of doping, we replaced B and N atoms with C atoms, considering different atomic positions in the central hexagon. When an N atom was replaced with a C atom, the new structure was a semiconductor, but when a B atom was replaced with a C atom, the new structure was a semimetal. At the same time, the polarity increased, inducing covalent behavior. Replacing two N atoms with two C atoms also resulted in a semiconductor, while replacing two B atoms with two C atoms yielded a semimetal; in both cases the bonding was covalent. When three B (three N) atoms of the central hexagon were replaced with three C atoms, the new structure exhibited a transition to a conductor (remained a semiconductor) with low polarity. When monovacancies (N) and divacancies (B and N) were inserted into the lattice, the system was transformed into a covalent semiconductor. Finally, the electrostatic potential surface was calculated in order to explore intermolecular properties such as the charge distribution, which showed how the reactivity of the boron nitride sheets was affected by doping and orbital hybridization.

Keywords Boron nitride · DFT theory · Isocoronene · Electrostatic potential

Introduction

Graphene and graphene-like 2D layers have attracted the attention of scientists due to their suitability for applications in the technology industry. Boron nitride hexagonal (h-BN) sheets were prepared experimentally for the first time in 2005 by Geim and collaborators [1]. Since then, this material has led to the possibility of fabricating new optoelectronic devices, as recently reported [2]. Even though this material has been the subject of many investigations, there are still no reports of any inves-

E. C. Anota (✉) · A. E. Morales
Facultad de Ingeniería Química, Cuerpo Académico de Ingeniería en Materiales, Benemérita Universidad Autónoma de Puebla,
Ciudad Universitaria, San Manuel,
Puebla, Código Postal 72570, México
e-mail: echigoa@yahoo.es

R. E. R. Gutiérrez
Facultad de Ciencias Químicas,
Benemérita Universidad Autónoma de Puebla,
CP72570 Puebla, Pue, México

G. H. Cocoletzi
Instituto de Física 'Luís Rivera Terrazas',
Benemérita Universidad Autónoma de Puebla,
Apartado Postal J-48, Puebla 72570, México

tigations of lattice defects in h-BN sheets, performed in a similar fashion to those done for graphene sheets [3]. In a recent paper, Akcöltekin [4] performed theoretical and experimental investigations of lattice defects of graphene. These studies of defect engineering were performed using ion irradiation (Fig. 1) [5], and the isocoronene cluster model [6] ($C_{24}H_{12}$, Fig. 2a) was used to represent the d-BN armchair edge [7].

In a recent paper, Chigo [8] used the armchair edge circular model to investigate 2D carbon atomic structures, taking into account the presence of N atoms as doping atoms. Similar atomic configurations have been used to study the atomic structures of GaAlN and GaInN alloys [9], III-A nitrides [10], the adsorption of H_2O onto 2D h-BN [11], the doping of h-BN sheets with Li and F [12], the adsorption of O_3 onto h-BN sheets [13], the atomic structure of silicon carbide (either pure or with defects) [14], the effects of chemically modifying boron nitride oxide sheets, and vacancies and nitrogen dopants in boron nitride oxide.

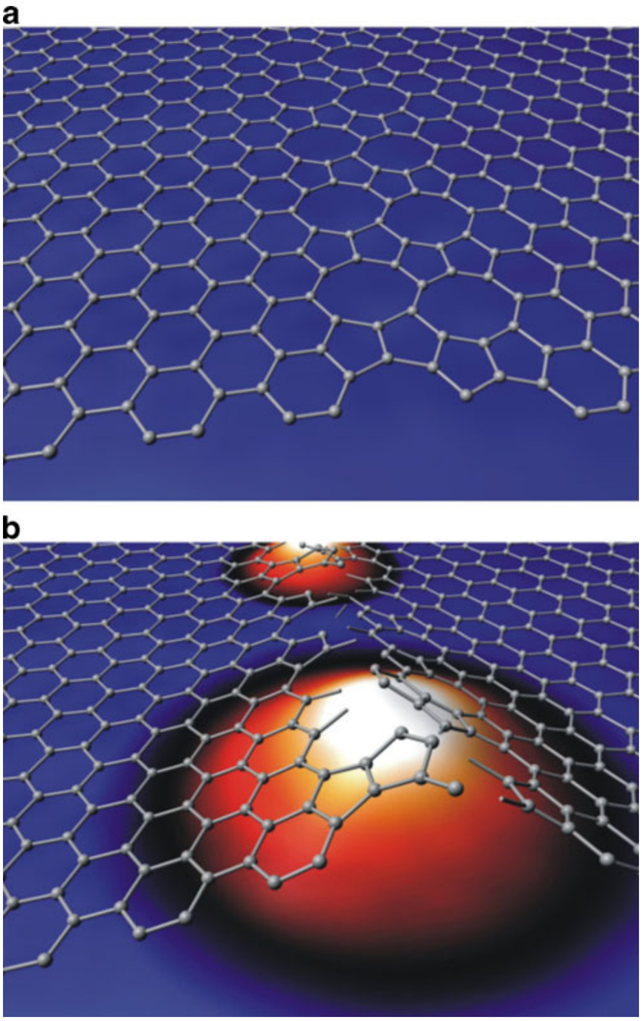


Fig. 1 Lattice model used to represent lattice defects in graphene. The graphene structure can be fabricated by ion irradiation [4]

Motivated by the work in [4], we investigated the atomic structure of h-BN sheets, taking into account lattice defects. The results of this investigation are summarized in this paper. We considered the atomic structure of the sheets to consist of a central hexagon surrounded by alternating pentagons (three) and heptagons (three) with three different chemical compositions, as depicted in Fig. 2b–d. Configuration 1 had three B–B bonds; configuration 2 had three N–N–N bonds at the periphery, around the central hexagon; while configuration 3 had three N–N bonds arranged in a similar fashion to configuration 1. We also explored the electronic properties of BN sheets as a function of the doping. To do this, we replaced one to three of the B or N atoms of the central hexagon with carbon (C) atoms. In addition, the effects of mono- and divacancies on the stable structures were studied. Finally, the chemical reactivities of the h-BN sheets were analyzed in terms of their electrostatic potential surfaces.

Computational methods

We performed first-principles total energy calculations to study boron nitride sheets, accounting for vacancies and doping effects, according to a procedure presented elsewhere [8–14]. Our calculations were performed using density functional theory (DFT) [15–19], as implemented in the DMOL^3 code [20]. The exchange and correlation energies were treated according to the local density approximation (LDA) with Perdew–Wang (PWC) [21] parametrization in the all-electron formalism, assuming that spin effects were unimportant for our system.

In the calculations, we used a double numeric polarized (DNP) atomic base (this includes a *p* orbital of hydrogen and *d* orbitals of carbon, boron, and nitrogen) for the core [20, 22, 23], with multiplicity equal to 1 (singlet) and zero charge (neutral) for the $B_{12}N_{12}H_{12}$ cluster, which has a base of 0.946 nm and a height of 0.98 nm (Fig. 2), in the absence of any doping. In the next step, a doped structure with a charge equal to zero and a multiplicity equal to 2 was considered, and the same conditions were applied to investigate structures with mono- and divacancies. Similar to the hexagonal boron nitride (h-BN) sheets, we apply the local density approximation to these d-BN sheets, assuming that spin does not affect their atomic and electronic structure.

The cutoff orbital radius used in the calculations was 0.40 nm, the self-consistency tolerance was 1.0×10^{-6} Ha, and the relaxation procedure took the positive vibration frequency criterion into account [24]. We

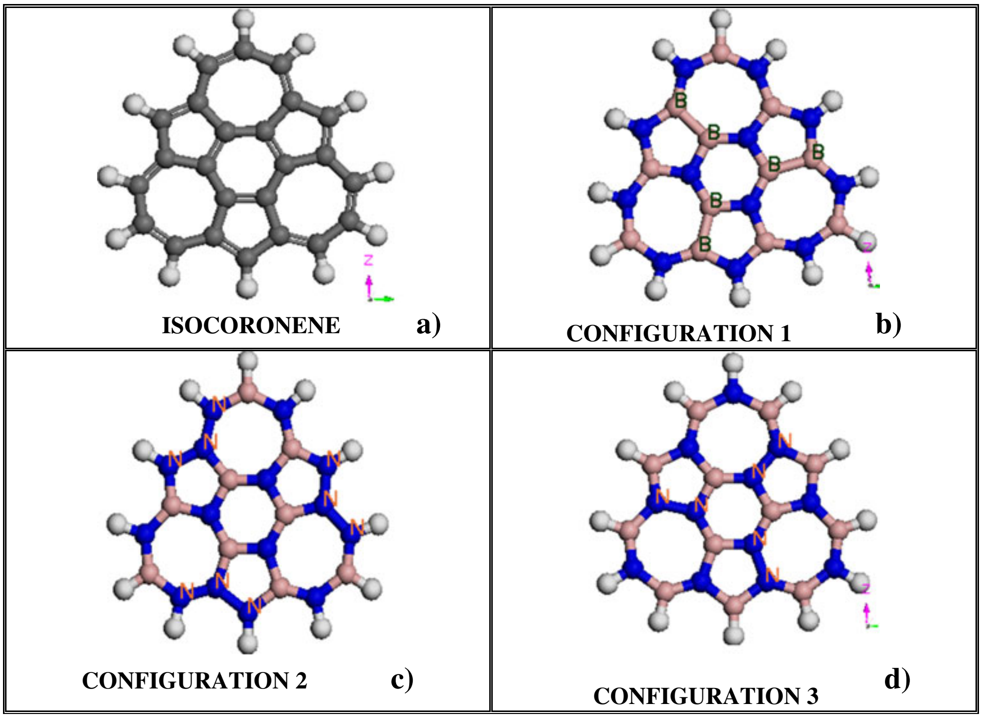


Fig. 2 **a** Coronene isomer ($C_{24}H_{12}$) model that was used to represent graphene. **b–d** Initial atomic geometries of the boron nitride nanosheets

determined the most stable configurations, polarities (dipole moments), binding energies, and energy gaps. Energy gaps were calculated as the energy difference between the HOMO and LUMO orbital energies. To validate our structural model, we used a procedure described in [25] for determining the cohesive energy for the following models: naphthalene, pyrene, coronene, and the cluster $B_{27}N_{27}H_{18}$. The value of the cohesive energy was 1.66 a.u./atom [26], assuming that size does not affect electronic properties.

To elucidate the static charge distribution, we calculated the electrostatic potential at van der Waals distances [27]. It is well known that any distribution of electrical charge, such as those of electrons or nuclei of molecules, creates an electrical potential in the surrounding space. This may be considered the potential of the molecule, which interacts with a point electrical charge [28]. A variety of methods to calculate the electrostatic potential are currently available in the literature, which display different levels of accuracy. In the present work, the molecules were investigated with the hybrid functional B3LYP [29] and the 6-31+G(d,p) [30, 31] basis using the GAUSSIAN03 [32] program. In the calculations, the electrostatic potential surfaces of the rings were generated by mapping the electrostatic potentials onto isosurfaces of the molecular electron density between 0.02 and 0.04 a.u. and by color coding using the program gOpenMol [33] to visualize the molecules.

Results and discussion

The isocoronene model of boron nitride sheets suggested the possibility of generating lattice defects in the structure with the formation of pentagons, hexagons, and heptagons.

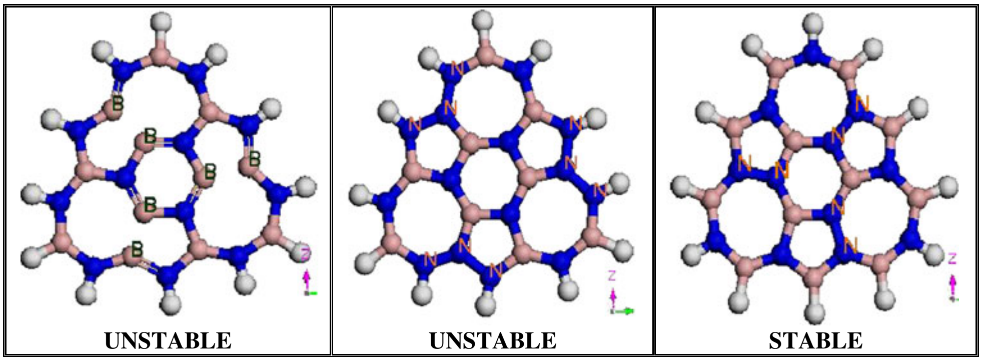


Fig. 3 The relaxed atomic geometries of boron nitride nanosheets

To carry out calculations, we applied the isocoronene cluster model with the armchair edge to three different chemical compositions. In the first model, three B–B bonds were formed, where the B atom of the dimer was part of the central hexagon. In the second model, three N–N–N bonds were formed at the periphery, around the central hexagon. The third model was similar to the first structure, but three N–N bonds were formed instead.

Studies of vacancy effects indicated that model 3 was the model with the lowest minimum energy (see Fig. 3), so this was considered the ground state of the system. In this stable configuration, the B–N bonds showed *sp* hybridization, which is very similar to what is reported in the literature [8, 13]. The system had no overall charge and the multiplicity was 1 (within the LDA formalism, with restricted spin). The structure exhibited a regular geometry, in contrast to the irregular geometry of graphene (as represented by isocoronene) at the C–C bond; see Table 1.

Doping effects were investigated, as shown in Fig. 4a and b. We replaced 4.16% of the B (N) by C, which led to stable, planar, 2D structures. The *sp* orbital hybridization of the C–N bonds alternated between *sp* and sp^2 throughout the entire lattice (Fig. 4a). Similarly, *sp* hybridization of the B–C bonds alternates with the sp^2 hybridization of the C–N bonds (Fig. 4b). These results agree with those obtained for CN structures with circular, rectangular, and triangular [34] shapes.

Replacing two B or N atoms with two C atoms also resulted in stable geometries, with sp^2 hybridization observed for the N–C bonds and *sp* hybridization for the B–C bonds; see Fig.4c.

Additional studies of the effects of doping were performed. We explored the structure obtained by replacing 25% of the central hexagon's atoms with C atoms, thus producing the following chemical composition: $B_9C_6N_9H_{12}$. This system had the geometry of model 3 (see Fig. 4c). After structural relaxation, the B–N bond length was 1.39 Å (in the central hexagon), which is somewhat different from the value of 1.44 Å obtained for the B–N bonds in the pentagons and 1.43 Å for those in the heptagons. These values are similar to those observed for hexagonal BN sheets with no defects [8]. The B–N bonds display *sp* hybridization and the C–C bonds show sp^2 hybridization.

When two B (N) atoms in the central hexagon are substituted by two C atoms, the structure behaves as semiconductor (semimetal) with an energy gap of 0.82 eV (0.45 eV) and a high polarity of 991.8×10^{-3} (1167.2×10^{-3}) D.

Doping with three C atoms is represented in Fig. 4d. Again, we replaced B and N atoms independently. After relaxation, in the stable atomic configurations, two of the three N–C bonds displayed sp^2 hybridization, while and the B–C bonds exhibited *sp* hybridization.

Monovacancies resulting from the absence of a B (N) atom induced sp^2 hybridization in B–N bonds inside pentagons, hexagons, and heptagons (Fig. 4f). However, divacancies (see Fig. 4e) led to a stable lattice structure with two hexagons, one pentagon, one tetragon, and one octagon, which presented bonding based on *sp* hybridization. This is similar to what was observed in a graphene oxide layer.

We determined the energy gap (3.42 eV) of the BN sheet with no defects based on the energy difference between the HOMO and LUMO orbitals. This energy gap indicates that the structure behaves as a semiconductor with ionic character, and has a dipole moment of 4.3×10^{-3} D. This result is in accord with the experimental value for the hexagonal BN sheet in the presence of vacancies [35]. In contrast, replacing the B site of the central hexagon with a C atom yielded a covalent atomic geometry of high polarity (1060.9×10^{-3} D), similar to what has been reported for the BN oxide sheet [36]. When the N atom was replaced instead, the structure transformed from a semimetal (energy gap of 0.33 eV) into a semiconductor (energy gap of 1.97 eV).

Replacing the three B (N) atoms of the central hexagon with three C atoms produced a stable conductor (semiconductor) with ionic character.

The electrostatic potential $V(r)$ of a molecule, based on the static charge distributions of the nuclei and electrons within it, can be analyzed to predict the reactivity of the molecule [37]. If a molecule has an electron density $\rho(r)$, then its electrostatic potential at any point r is given by

$$V(r) = \sum_{\mathrm{A}} \frac{Z_{\mathrm{A}}}{|R_{\mathrm{A}} - r|} - \int \frac{\rho(r')dr'}{|r' - r|} \tag{1}$$

Here, Z_{A} is the charge of nucleus A, located at $r = R_{\mathrm{A}}$. This potential has proven to be a particularly useful indicator of the sites or regions of a molecule to which an approaching electrophile will be attracted. It has been applied successfully to study interactions between reactants and to recognition in biological systems (e.g., in enzyme–substrate systems) [27, 37–39].

Using the approach described above, we characterized the electrostatic potential surface of the boron nitride sheet in terms of site-specific and global quantities. The calculated electrostatic potential surfaces for different BN models are presented in Figs. 5, 6, and 7. The electrostatic potential surface for model C1 is displayed in Fig. 5, where the positive (the red color represents boron atoms with an isosurface of 0.04 a.u.) and negative (the blue color represents nitrogen atoms with an isosurface of −0.02 a.u.) charge densities are clearly separated, indicating that there is no resonance effect. However, for model C2, as shown in Fig. 5c, it is possible to find sp^2 hybridization for the N–N–N bonds with high electron density (these are represented by

Table 1 Optimized atomic geometries, dipolar moments, energy gaps, and binding energies of the nanosheets

System	Bond distance (Å)					Dipolar moment (× 10^{-3} D)	Gap (HOMO–LUMO, eV)	Binding energy (eV)
	C–C	B–N	N–N	C–N	B–C			
[A]Isocoronene (graphene)	1.41(hexagon)					1.33	0.98	8.73
	1.39 (heptagon)							
	1.44 (pentagon)							
$C_{54}H_{18}$ (graphene)	1.41					2.9	1.94	19.18
$C_{24}H_{12}$ [5] (graphene)	1.41					0.2	4.06	
$B_{12}N_{12}H_{12}$ [8]		1.44				3.7	6.96	7.95
$B_{27}N_{27}H_{18}$ (doping with hexagon C)		1.44				13.4	4.84	17.16
BNO C1 [33]		1.43				1584	1.2	18.76
		1.47						
BNO C2 [33]		1.44				7200	2.4	18.88
		1.45						
BN configuration 1		1.43				7.3	3.76	Unstable structure
BN configuration 2		1.43				8.0	3.16	Unstable structure
			1.38					
			1.41					
BN		1.43				4.3	3.42	7.64
Configuration 3								
BN monovacancy								
B		1.43	1.38			2159.5	0.23	Unstable
N		1.43	1.38			1188.2	1.96	7.25
BN divacancy (B,N)		1.43	1.43			118.6	3.59	6.99
BN doped with C (replacing B)		1.43				1060.9	0.33	7.55
			1.41					
				1.36				
BN doped with C (replacing N)		1.43				349.5	1.97	7.66
		1.45						
		1.47	1.41					
				1.37				
BN hexagon doped with C	1.39	1.43	1.40			2.4	2.87	7.93
BN doped with 3C (replacing 3B)		1.44				2.2	0.0	7.73
		1.45			1.49			
BN doped with 3C (replacing 3N)		1.44	1.39	1.37		7.6	1.12	7.73
BN doped with 2C (replacing 2B)		1.42	1.39	1.39		1167.2	0.45	7.47
BN doped with 2C (replacing 2B)		1.43	1.40	1.38	1.50	991.8	0.82	7.697

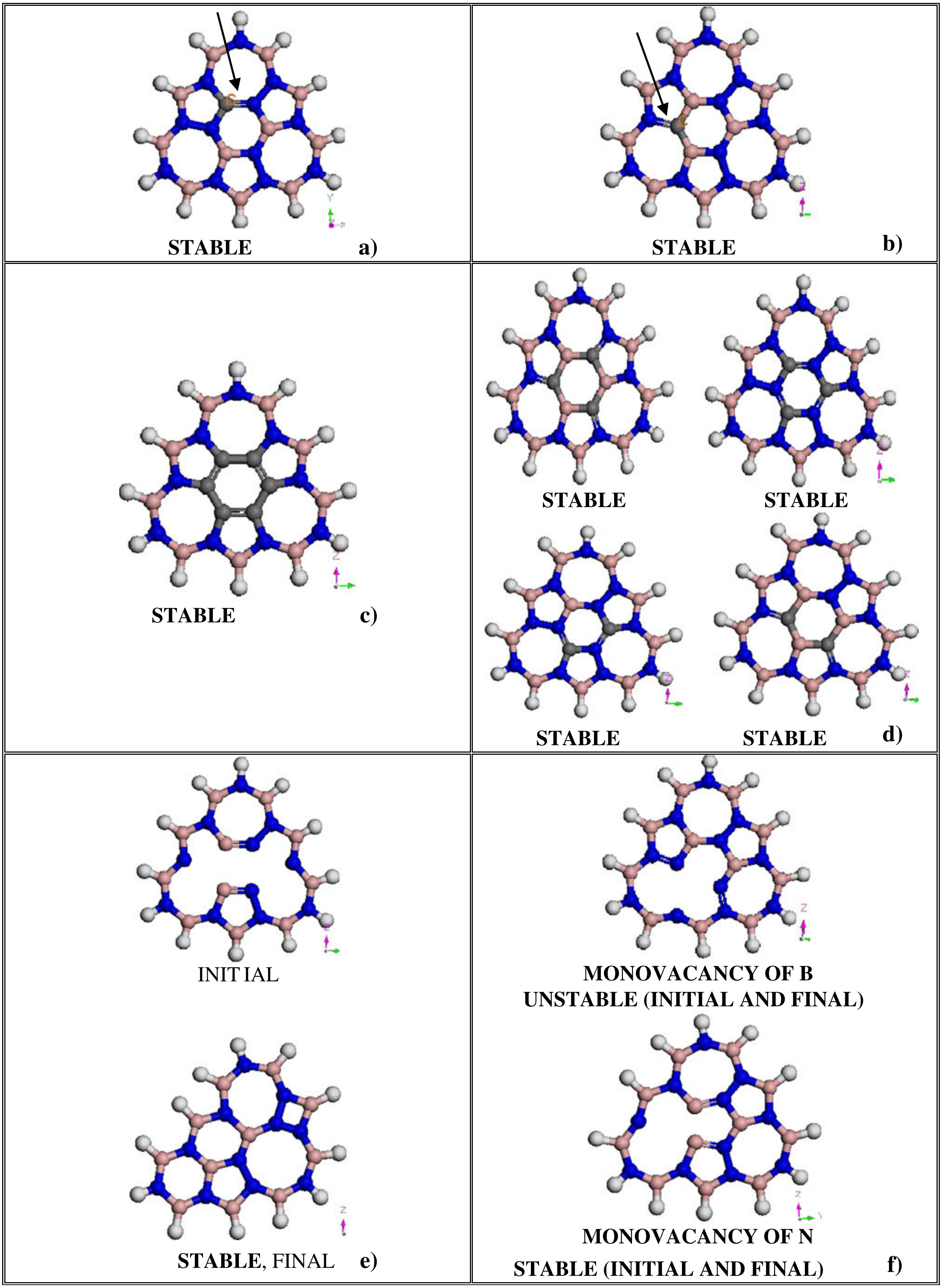

C
STABLE
a)
C
STABLE
b)
STABLE
c)
STABLE
STABLE
STABLE
STABLE
d)
INIT IAL
STABLE, FINAL
e)
MONOVACANCY OF B
UNSTABLE (INITIAL AND FINAL)
MONOVACANCY OF N
STABLE (INITIAL AND FINAL)
f)

Fig. 4a–f Atomic geometries of the boron nitride sheets after relaxation, taking into account carbon atom doping. Structures of the stable configurations are shown. **a** A carbon atom replaces a boron atom, **b** a carbon atom replaces a nitrogen atom, **c** doping of the central hexagon with carbon atoms, **d** replacing two or three B and N atoms with the equivalent number of C atoms, **e** initial and final geometries obtained with B and N divacancies, and **f** initial and final geometries of nanosheets with B and N monovacancies, respectively

the blue color, with an isosurface of −0.02 a.u.). Model C3 also displays sp^2 binding at the N–N bonds with the boron appearing as an isolated atom; similar effects are seen in carbon and boron/nitrogen [40, 41].

Similarly, we characterized the electrostatic potential surfaces of the different d-BN sheet models, as presented in Fig. 6. In the first model (see Fig. 6a), the atoms in the central ring have been replaced with carbon atoms, inducing distinctive charge distributions between the carbon atoms and between the carbons and the adjacent nitrogen atoms. The charge distribution due to the π electrons that move around the central benzene molecule is represented in a green color.

The model 1N by 1C (as presented in Fig. 6c), where one N atom in the central ring has been replaced with a carbon atom, shows similar electronegativities and hybrid-

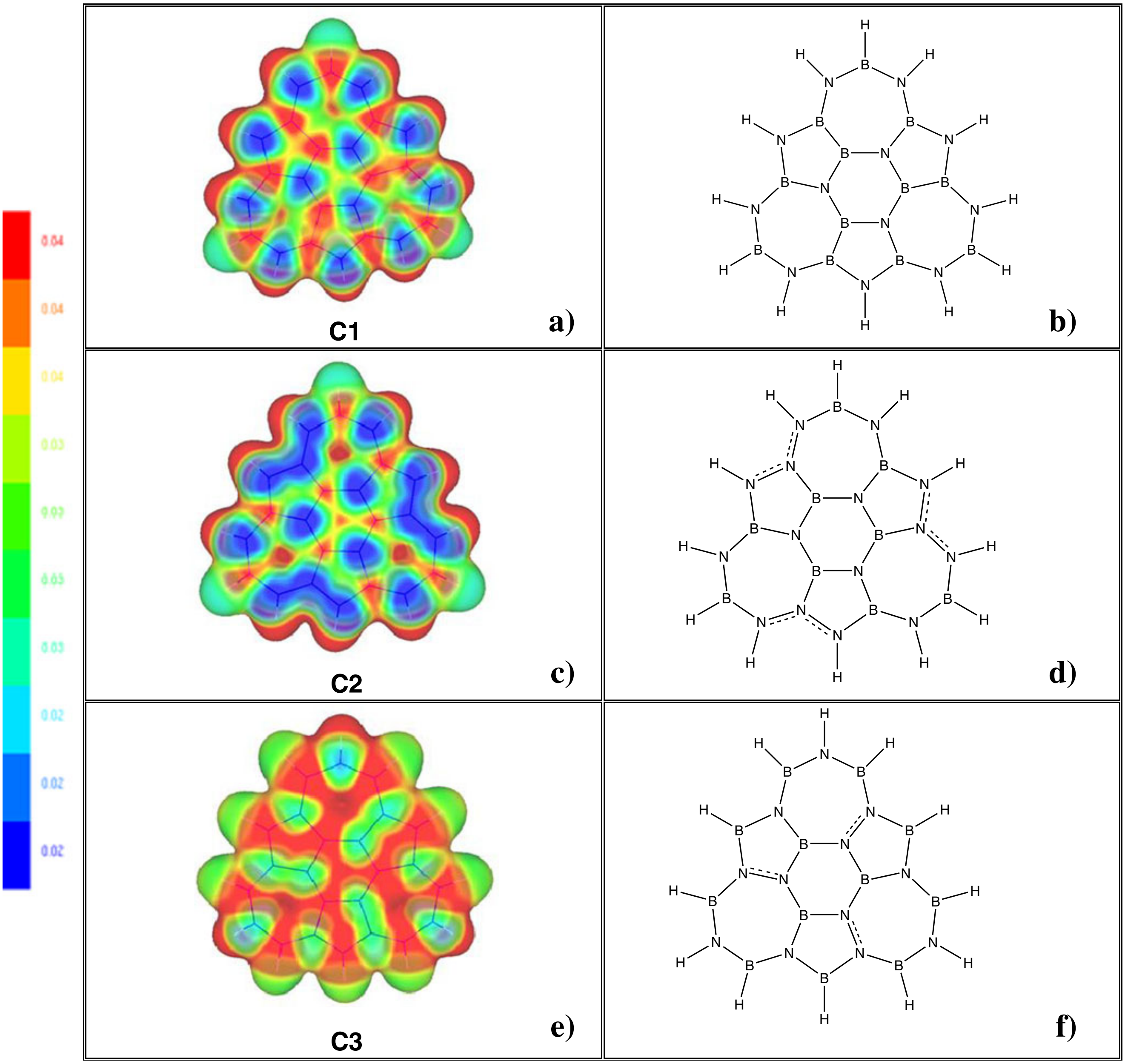


Fig. 5a–f Electrostatic potential surfaces for the three models C1, C2, and C3 (**a**, **c**, and **e**, respectively) used in the calculations. These represent boron nitride sheets with lattice defects. **b**, **d**, and **f** show the structural formulae of the models

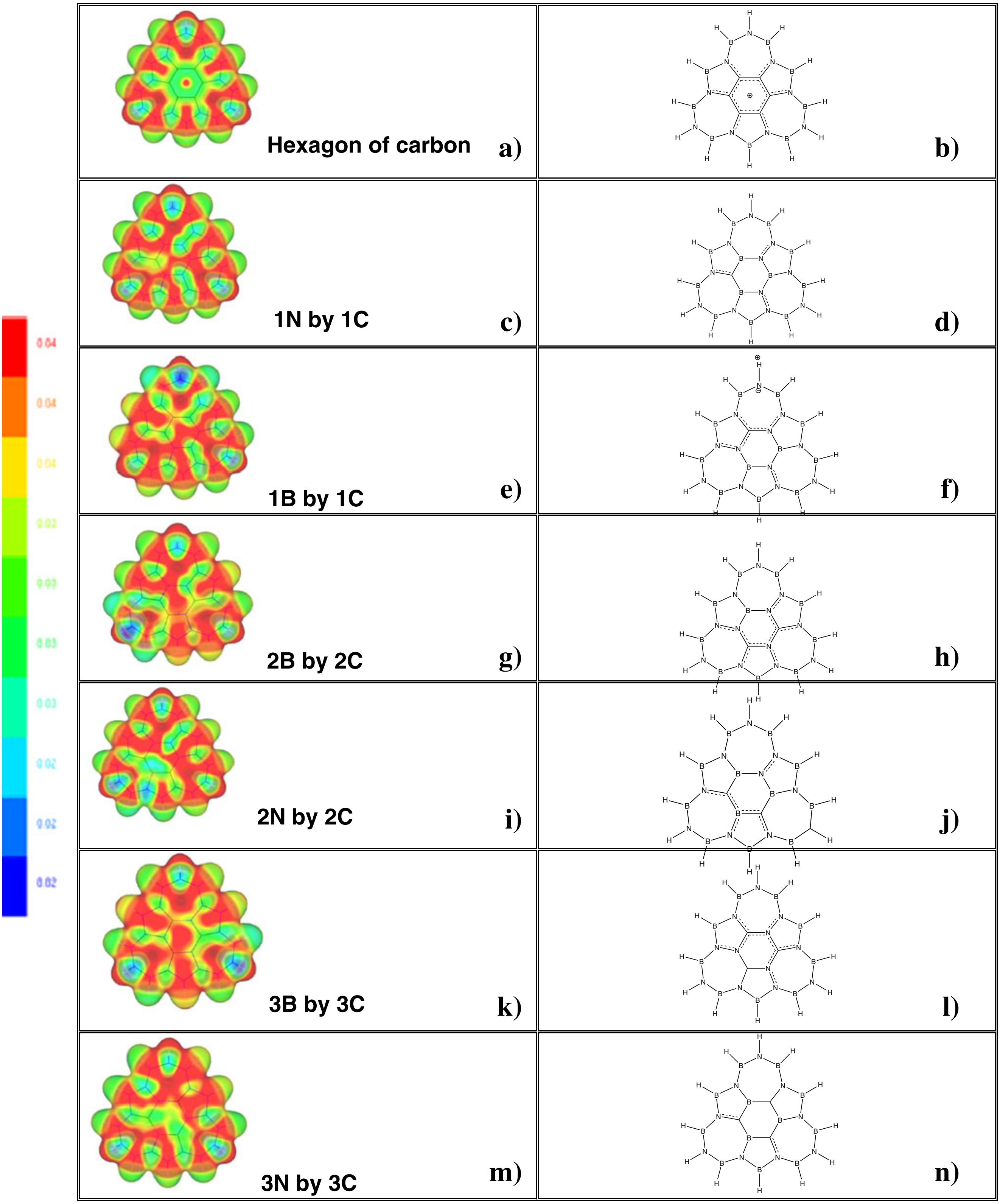


Fig. 6a–n Electrostatic potential surfaces of d-BN sheets that include doping. Each diagram represents a relaxed d-BN sheet that is doped with carbon atoms at a particular concentration (**a**, **c**, **e**, **g**, **i**, **k**, and **m**) and its corresponding structural formulae (**b**, **d**, **f**, **h**, **j**, **l**, and **n**)

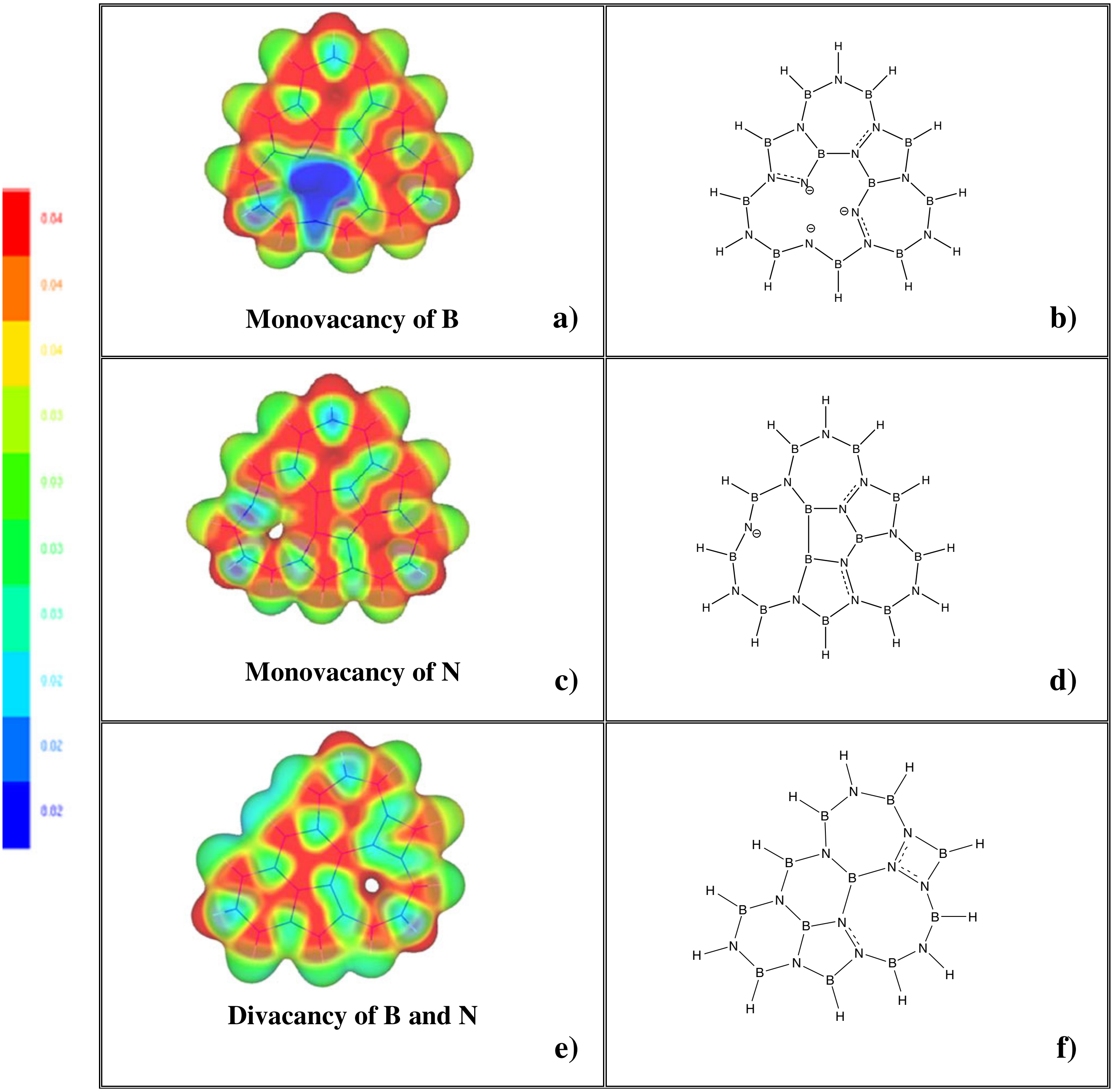


Fig. 7a–f Electrostatic potential surfaces of d-BN sheets with B (**a**) and N (**b**) monovacancies and with (B, N) divacancies (**c**) after atomic coordinate optimization, as well the corresponding structural formulae (**b**, **d**, and **f**, respectively)

izations to model C3. The hybridizations of the carbon and nitrogen atoms mean that C substitution barely changes the electrostatic potential surface of the sheet. The same behavior is observed when one B atom in the central ring is replaced with a C atom; the conjugation expands to include the nitrogen atoms adjacent to carbon, as depicted by the isosurface of 0.03 a.u. (shown as a green color in Fig. 6e).

Figure 6g represents the 2B by 2C model, and shows the charge distribution between the nitrogen and carbon atoms, and how the boron atoms are left isolated with positive charges. In contrast, in Fig. 6i, the charge distribution surrounding the boron atoms is a consequence of the environment. In the 2B by 2C and 2N by 2C models, the N–N bonds in the former model and the C–N bonds of the latter model support vibration modes that suggest resonant models. The models presented in Fig. 6a and c include boron and nitrogen monovacancies in their sheets, which cause large-scale positive or negative charge distributions around the absent atoms, as depicted in Fig. 7.

To complement our discussions, we have included the structural formula of all of the models, which indicate the positions of single and conjugated sp^2 orbitals (see Figs. 5b, d, and f, 6b, d, f, h, j, l, and n, and 7b, d, and f).

Conclusions

In this work, we have presented the results of defect engineering studies of boron nitride sheets. We modeled the systems as clusters of coronene isomers, consisting of a hexagon surrounded by three pentagons and three heptagons. Our molecular quantum mechanics studies showed the possible structures formed (with an armchair edge). We also demonstrated how the electronic (HOMO–LUMO gap) and atomic structures of BN sheets can be modified. Changing the polarity modifies mechanical properties such as the chemical hardness. The electric conductivity is modified by the presence of impurities at different concentrations (i.e., by replacing the boron or nitrogen atoms of the central hexagon in the structure with one, two, three, or six carbon atoms) as well as the presence of mono- and divacancies (boron–nitrogen) in the lattice. Studying the electrostatic potential surfaces of the different BN sheet models highlighted changes in the charge distribution caused by doping or lattice defects. Inspecting the isosurfaces elucidated sp^2 hybridization-based bond formation.

Acknowledgments This work was partially supported by VIEP-Benemérita Universidad Autónoma de Puebla (CHAE-ING1I-I), Facultad de Ingeniería Química-Benemérita Universidad Autónoma de Puebla (2010–2011), Cuerpo Académico Ingeniería en Materiales (BUAP-CA-177), Cuerpo Académico Física Computacional de la Materia Condensada (BUAP-CA-191), and VIEP-BUAP--EXC. Work performed by GHC was supported by Conacyt project # 83982. GHC's work was also partly performed while visiting the Department of Physical and Theoretical Chemistry, University of Saarland, Saarbruecken, Germany.

References

1. Novoselov KS, Jiang D, Schedin F, Booth TJ, Khotkevich VV, Morozov SV, Geim AK (2005) Proc Natl Acad Sci USA 102:10451–10453
2. Serrano J, Bosak A, Arenal R, Krisch M, Watanabe K, Taniguchi T, Kanda H, Rubio A, Wirtz L (2007) Phys Rev Lett 98:095503–095504
3. Appelhans DJ, Lin Z, Lusk MT (2010) Phys Rev B 82:073410–073414
4. Akcöltekin S, Bukowska H, Peters T, Osmani O, Monnet I, Alzaher I, d'Etat BB, Lebius H, Schleberger M (2011) Appl Phys Lett 98:103103–3
5. Lahiri J, Lin Y, Bozkurt P, Oleynik II, Batzill M (2010) Nature Nanotech 5:326–329
6. Ciesielski A, Cyranski MK, Krygowski TM, Fowler PW, Lillington M (2006) J Org Chem 71:6840–6845
7. Zeng H, Zhi C, Zhang Z, Wei X, Wang X, Guo W, Bando Y, Golberg D (2010) Nano Lett 10:5049–5055
8. Chigo Anota E (2009) Superficies y Vacío 22:19–23
9. Chigo Anota E, Hernández Cocoletzi H (2011) J Mol Model. doi:10.1007/s00894-011-1043-2
10. Chigo Anota E, Salazar Villanueva M, Hernández Cocoletzi H (2010) Phys Stat Solidi C 7:2252–2254
11. Chigo Anota E, Salazar Villanueva M (2009) Superficies y Vacío 22:23–28
12. Chigo Anota E, Salazar Villanueva M, Hernández Cocoletzi H (2010) Phys Stat Solidi C 7:2559–2561
13. Chigo Anota E, Hernández Cocoletzi H, Rubio Rosas E (2011) Eur Phys J D 63:271–273
14. Chigo Anota E, Hernández Cocoletzi H, Bautista Hernández A, Sánchez Ramírez JF (2011) J Comput Theor Nanosci 8:637–641
15. Kohn W, Becke AD, Parr RG (1996) J Phys Chem 10:12974–12980
16. Jones RO, Gunnarsson O (1989) Rev Mod Phys 61:689
17. Kohn W (1999) Rev Mod Phys 71:1253–1266
18. Parr R, Yang W (1989) Density functional theory of atoms and molecules, 1st edn. Oxford University Press, New York
19. Chigo Anota E, Rivas Silva JF (2005) Rev Col Fís 37:405–417
20. Delley B (1990) J Chem Phys 92:508–608
21. Perdew JP, Wang Y (1992) Phys Rev B 45:13244–13249
22. Delley B (1996) J Phys Chem 100:6107–6110
23. Delley B (2000) J Chem Phys 113:7756–7765
24. Foresman JB, Frisch Æ (1996) Exploring chemistry with electronic structure methods, 2nd edn. Gaussian, Inc., Pittsburgh, p 70
25. Hernández Rosas JJ, Ramírez Gutiérrez RE, Escobedo Morales A, Chigo Anota E (2010) J Mol Model 17:1133–1139
26. Galicia Hernández JM, Hernández Cocoletzi G, Chigo Anota E (2011) J Mol Model. doi:10.1007/s00894-011-1046-z
27. Weiner PK, Langridge R, Blaney JM, Schaefer R, Kollman P (1982) Proc Natl Acad Sci USA 79:3754–3758
28. Politzer P, Laurence PR, Jayasuriya K (1985) Environ Health Perspect 61:191–202
29. Becke AD (1993) J Chem Phys 98:5648–5652
30. Petersson GA, Al-Laham MA (1994) J Chem Phys 94:6081–6091
31. Petersson GA, Bennett A, Tensfeldt TG, Al-Laham MA, Shirley WA, Mantzaris J (1988) J Chem Phys 89:2193–2219
32. Frisch MJ et al (2004) Gaussian 03, revision C.02. Gaussian, Inc., Wallingford
33. CSC—IT Center for Science Ltd. (2011) g0penMol software download webpage.http://www.csc.fi/english/pages/g0penMol. Last accessed Jan 2011
34. Chigo Anota E, Hernández Cocoletzi H (2011) J Mol Model. doi:10.1007/s00894-011-1043-2
35. Alem N, Erni R, Kisielowski C, Rossell MD, Gannett W, Zettl A (2009) Phys Rev B 80:155425–155432
36. Chigo Anota E, Salazar Villanueva M, Hernández Cocoletzi H (2011) J Nanosci Nanotechnol 11:5515–5518
37. Peralta-Inga Z, Murray JS, Edward Grice M, Boyd S, O'Connor CJ, Politzer P (2001) J Mol Struc THEOCHEM 549:147–158
38. Scrocco E, Tomasi J (1973) Top Curr Chem 42:95–170
39. Naray-Szabo G, Ferenczy GG (1995) Chem Rev 95:829–847
40. Politzer P, Murray JS, Lane P, Concha MC, Jin P, Peralta-Inga Z (2005) J Mol Model 11:258–264
41. Peralta-Inga Z, Lane P, Murray JS, Boyd S, Grice ME, O'Connor CL, Politzer P (2003) Nano Lett 3:21–28

J Mol Model (2012) 18:2185–2198
DOI 10.1007/s00894-011-1236-8

ORIGINAL PAPER

Docking, molecular dynamics and quantitative structure-activity relationship studies for HEPTs and DABOs as HIV-1 reverse transcriptase inhibitors

Yating Mao · Yan Li · Ming Hao · Shuwei Zhang · Chunzhi Ai

Received: 21 June 2011 / Accepted: 26 August 2011 / Published online: 27 September 2011

Abstract As a key component in combination therapy for acquired immunodeficiency syndrome (AIDS), non-nucleoside reverse transcriptase inhibitors (NNRTIs) have been proven to be an essential way in stopping HIV-1 replication. In the present work, in silico studies were conducted on a series of 119 NNRTIs, including 1-(2-hydroxyethoxymethyl)-6-(phenylthio)thymine (HEPT) and dihydroalkoxybenzyloxopyrimidine (DABO) derivatives by using the comparative molecular field analysis (CoMFA), comparative molecular similarity indices analysis (CoMSIA), docking simulations and molecular dynamics (MD). The statistical results of the optimal model, the ligand-based CoMSIA one (Q^2=0.48, $R_{ncv}{}^2$ =0.847, $R_{pre}{}^2$=0.745) validates its satisfactory predictive capacity both internally and externally. The contour maps, docking and MD results correlate well with each other, drawing conclusions as follows: 1) Compounds with bulky substituents in position-6 of ring A, hydrophobic groups around position- 1, 2, 6 are preferable to the biological activities; 2) Two hydrogen bonds between RT inhibitor and the Tyr 318, Lys 101 residues, respectively, and a π-π bond between the inhibitor and Trp 188 are formed and crucial to the orientation of the active conformation of the molecules; 3) The binding pocket is essentially hydrophobic, which are determined by residues such as Trp 229, Tyr 318, Val 179, Tyr 188 and Val 108, and hydrophobic substituents may bring an improvement to the biological activity; 4) DABO and HEPT derivatives have different structures but take a similar mechanism to inhibit RT. The potency difference between two isomers in HEPTs can be explained by the distinct locations of the 6-naphthylmethyl substituent and the reasons are explained in details. All these results could be employed to alter the structural scaffold in order to develop new HIV-1 RT inhibitors that have an improved biological property. To the best of our knowledge, this is the first report on 3D-QSAR modeling of this series of HEPT and DABO NNRTs. The QSAR model and the information derived, we hope, will be of great help in presenting clear guidelines and accurate activity predictions for newly designed HIV-1 reverse transcriptase (RT) inhibitor.

Keywords 3D-QSAR · DABO · Docking · HEPT · HIV-1 RT inhibitor · MD

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1236-8) contains supplementary material, which is available to authorized users.

Y. Mao · Y. Li (✉) · M. Hao · S. Zhang
Department of Materials Science and Chemical Engineering,
Dalian University of Technology,
Dalian, 116023, Liaoning, China
e-mail: yanli@dlut.edu.cn

C. Ai
Lab of Pharmaceutical Resource Discovery,
Dalian Institute of Chemical Physics,
Graduate School of the Chinese Academy of Sciences,
Dalian, 116023, Liaoning, China
e-mail: icy@dicp.ac.cn

Introduction

Despite the big progress in pharmaceutical and surgical treatment, the acquired immunodeficiency syndrome (AIDS) caused by the pandemic form of sexually transmitted human immunodeficiency virus (HIV) is still one of the leading causes of death worldwide [1]. Reverse transcriptase (RT) plays a key role in the replication of HIV through changing single-stranded genomic RNA into double-stranded proviral DNA and thus becomes one of the main targets for the development of AIDS therapy and its inhibitors have accordingly attracted much research interest [2–4].

We all know that there are two main, proximal but different, active binding sites in HIV-1 RT, where one is the nucleoside binding site (NBP), and the other is the non-nucleoside binding pocket (NNBP) [5]. According to the requirements of these sites, some highly specific RT inhibitors are synthesized and used as nucleoside reverse transcriptase inhibitors (NRTIs) in AIDS therapy, such as zidovudine (AZT), disanoint (ddI) and zalcitabine (ddC) [6]. Several structurally different and potent compounds which are non-nucleoside reverse transcriptase inhibitors (NNRTIs) of RT have also been identified. They involve the tetrahydroimidazo[4,5,1-jk]-[1, 4]benzodiazpin-2(1H4)-one (TIBO), 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio) thymine (HEPT), nevirapine, pyridinone, bis(heteroaryl) piperazine (BHAP) and R-anilinophenylacetamide (R-APA), etc. [7]. Since the day of synthesis, these NNRTIs have attracted much research interest due to many advantages they owned, including lower toxicity, more stable chemical properties, slower metabolizing rate, as well as their slower emit rate from the human body than NRTIs. Actually, they interact with a specific allosteric site adjacent to the polymerase site of the HIV-1 RT rather than bind to cellular polymerases, which results in a non-competitive mechanism [8, 9].

In spite of the fact that NNRTIs are well tolerated with adverse effects occurring in the treatment of the first 6 weeks, the main limitation for all currently available NNRTIs is the low genetic barrier to resistance [10]. The popular anti-HIV drug policy is based on drug combination regimens. Considering the different classes of anti-HIV drugs currently available, i.e., NRTIs, nucleotide reverse transcriptase inhibitors (NtRTIs), NNRTIs, protease inhibitors (PIs), fusion inhibitors (FIs), coreceptor inhibitors (CRIs), and integrase inhibitors (INIs) [11, 12], the number of possible multi-drug combinations is high. However, the number of approved fixed-dose drug combinations is rather limited [13]. To catch up with these strategies, it will be of necessity to have in hand an arsenal of new compounds with enhanced activities or less vulnerability to viral drug resistance.

Computer simulation techniques potentially offer further means to design more effective drugs which exhibit potent activity toward drug-resistant strains of RT and explore the inhibition mechanisms. The quantitative structure-activity relationship (QSAR) [14] study has been one of the most effective computational approaches in drug design. Up till now, it has been successfully applied in many biological and medicinal studies like the FIXa inhibitors [15], BAZ-based DA D3 receptor antagonists [16], Aurora B inhibitors [17] and so on. A unique advantage of 3D-QSAR modeling is that it provides a direct way to explore and visualize the structure–activity relationship of the molecules, among which the comparative molecular field analysis (CoMFA) [18] and the comparative molecular similarity indices analysis (CoMSIA) [19] are the two most widely used approaches. Understanding the interactions between proteins and ligands is crucial for the rational design of novel drugs with potent pharmaceutical or functional effects. Thus when the experimental structure of an individual receptor or the receptor-ligand complex is available (usually obtained by X-ray crystallography or NMR), computational algorithms including especially the docking and molecular dynamics (MD) simulations are usually applied to identify the possible binding modes of the ligands at the active site.

Since the discovery of HEPT [20] as NNRTIs in 1989, more than 30 different classes of NNRTIs have been reported [21, 22]. Among them, dihydroalkoxybenzyloxopyrimidines (DABOs) [23] represented a significant class of NNRTIs which was developed in the last years. From the point of the view in chemistry, DABOs belong to the 4-pyrimidinone series like HEPTs. Due to the structural similarities between them (as illustrated in Fig. 1), HEPTs and DABOs, we assume, should obtain some similar structural requirements for anti-HIV activity, which has not yet been studied in detail up to now [24].

In the in silico research of HIV-1 inhibitors, the QSAR analyses of HIV-1 reverse transcriptase inhibitors [25, 26], HIV-1 protease inhibitors [27–29] and HIV-1 integrase inhibitors [30, 31] have been reported before. Nevertheless, the application of 3D-QSAR methodology for the design of the above HEPT and DABO types of RT inhibitors has received little attention. Thus, in the present work based on 119 HEPT and DABO derivatives, the largest dataset up-to-date to the best of our knowledge, we attempt to set up comprehensive 3D-QSAR studies with an aim to disclose their structural features impacting their HIV-1 RT inhibitory

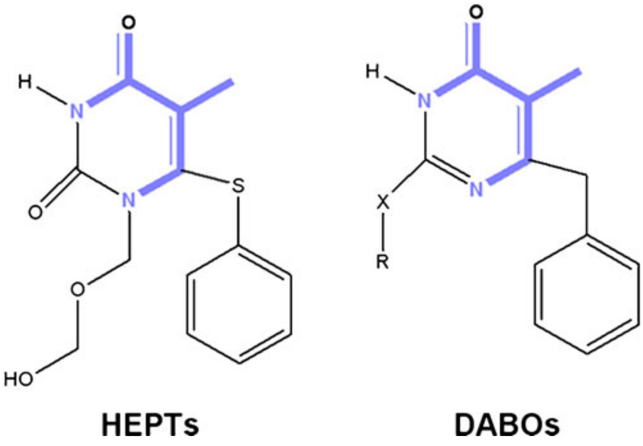


Fig. 1 Structures of HEPTs and DABOs (with X=S or N)

activities, by using integrated computational methods including 3D-QSAR, molecular docking simulations and molecular dynamics. To the best of our knowledge, this work is the first 3D-QSAR study for these two types of compounds, which will provide a platform for the design of novel HIV-1 RT inhibitors as important weapons in the fight against HIV.

Materials and methods

Dataset

Discarding those molecules with no confirmed anti-HIV activity, a total of 119 compounds (supporting Tables S1–S9) exhibiting HIV-1 RT inhibitory activities with IC_{50} values in the range of 0.017-237.740 μM were used to carry out the 3D-QSAR analysis (Table S10) [24–29]. Anti-HIV activity of all these compounds was determined in the same laboratory using the same procedures. All the anti-HIV activities used in the present study were expressed as $pIC_{50}=-\lg IC_{50}$, where IC_{50} is the concentration required to protect the cell against viral cytopathogenicity by 50% in MT-4 cells [24]. All molecular studies were performed using the molecular modeling package SYBYL 6.9 (Tripos Associates, St. Louis, MO). Partial atomic charges were calculated by the Gasteiger-Huckel method [30]. Energy minimization was performed using tripos force field and conjugate gradient method with convergence criterion set as 0.05 kcal mol^{-1} in this process.

3D-QSAR analysis

Training and test sets

All the compounds were grouped into a training set, for model generation and a test set, for model validation, containing 23 and 96 compounds (in approximately a ratio of 1:4), respectively. Both the training and the test sets were divided at random according to a representative range of structural variations and biological activities.

CoMFA and CoMSIA analyses

CoMFA method is a widely-used 3D-QSAR technique to correlate the biological activity of the compounds with their steric and electrostatic fields, which are calculated by putting the aligned molecules one by one, in a 3D regular lattice (2 Å spacing) extending at least 2 Å beyond the volumes of all investigated molecules on all axes. The van der Waals potential and Coulombic terms representing steric and electrostatic fields respectively, were calculated by the standard Tripos force field method. The column-filtering threshold value was set to 2.0 kcal mol^{-1} in order to increase the signal-noise ratio. A Csp^3 atom with a formal charge of +1 and a van der Waals radius of 1.52 Å served as a probe atom to generate steric (Lennard-Jones potential) and electrostatic (Coulombic potential) field energies, which were obtained by summing the individual interaction energies between each atom of the molecule and the probe atom at every grid point [31]. A 30 kcal mol^{-1} energy cut-off was used to avert infinity of energy values inside the molecule [18, 32].

CoMSIA method calculates five descriptors, including the steric, electrostatic and hydrophobic and the hydrogen bond donor and hydrogen bond acceptor properties. The similarity index descriptors were calculated by the same lattice box employed for the CoMFA calculations and a sp^3 carbon as a probe atom with +1 charge, +1 hydrophobicity and +1 HB donor and +1 HB acceptor properties [19, 33].

2D descriptor calculation and pre-processing

Presently, as a statistically satisfactory model could not be obtained only using the 3D descriptors, 2D descriptors are also employed to set up the model. The 2D descriptor was chosen according to the following procedures. To start with, the molecular structures of the inhibitors were built with the ISIS/Draw 2.3 program, and converted into the SMILES format to calculate the structural descriptors by DRAGON professional version 5.4 which were originally developed by the Milano Chemometrics and QSAR Research Group [34]. Next, Dragon calculated 929 molecular descriptors for each molecule where some descriptors were excluded by the following steps: 1) descriptors containing larger than 85% zero values were removed; 2) zero- and near zero- variance predictors were deleted; 3) the descriptors that have absolute correlations above 0.95 were omitted. After this, the number of original descriptors was reduced to 775, and then the correlations of these descriptors and the activity of the molecules were calculated by C program, with an attempt to find the most activity-relevant descriptor for further QSAR studies. As a result, the spectral moment 06 from the edge adjency matrix (ESpm06u), a kind of edge adjacency index, was employed as the independent variable in further 3D-QSAR analysis due to its highest correlation to the pIC_{50} values with an R^2 value of 0.166.

Partial least square analysis and model validation

The CoMFA and CoMSIA descriptors were taken as independent variables and pIC_{50} values were used as dependent variables to derive the 3D-QSAR in partial least-squares (PLS) method, an extension of the multiple regression analysis. This method can reduce a lot of original descriptors to a few principal components (PCs) which is linearly correlated to the original descriptors. Leave-one-out (LOO) cross-validation was also used to evaluate the predictive ability of the models. The cross-validated coefficient Q^2 was calculated by Eq. 1 as follows:

$$Q^2 = 1 - \frac{\sum_Y \left(Y_{predicted} - Y_{observed}\right)^2}{\sum_Y \left(Y_{observed} - Y_{mean}\right)^2} \qquad (1)$$

where $Y_{predicted}$, $Y_{observed}$, and Y_{mean} are predicted, observed and mean values of the target property (IC_{50}), respectively. $\sum \left(Y_{predicted} - Y_{observed}\right)^2$ is the predictive sum of squares (PRESS). The optimum number of components was used to derive the final regression models and it also corresponds to the lowest PRESS value [35]. Besides Q^2, the corresponding PRESS, the conventional correlation coefficient (R^2) and its standard error of estimate (SEE) were also calculated. Finally, the CoMFA/CoMSIA results were graphically represented by field contour maps, where the coefficients were generated by the field type "Stdev*Coeff".

Conformational sampling and alignment

The most crucial step for 3D-QSAR techniques is that the 3D structures of the molecules should be aligned based on a conformational template in an appropriate way. The template compound we chose was compound 19, owing to it having the most potent activity in the dataset (which is thus assumed to adopt a "bioactive conformation") [36]. All the compounds were fitted into the template by using the "Database Align" routine available in SYBYL. Figure 2a describes the common substructure for the alignment marked in bold. Figure 2b shows the resulting ligand-based alignment model.

Molecular docking studies

All compounds in the dataset were docked into the active site of HIV-1 reverse transcriptase (PDB accession code: 1RT1) by using the Surflex docking of SYBYL package. Our molecular docking operates as the steps below: 1) the protein structure was imported into Surflex and hydrogens were then added; 2) the protomol was generated in a ligand-based approach; 3) all the inhibitors were docked in the binding pocket and each of them got 50 possible active docking poses with different scores; 4) the docking conformations were saved for each compound, and were ranked on the basis of the scores; 5) the best ranking pose for every compound were extracted and aligned together for further QSAR analysis. During the above docking process, all the other parameters adopted default values [16].

Molecular dynamics simulations

The MD simulations were performed with GROMACS software package [37] using the GROMOS96 force field [38]. The molecular topology file for the ligand in protein was produced by the program PRODRG 2.5 [39, 40]. The simulation cell was a cubic periodic box whose size was 10.04Å *11.22Å* 12.55Å, and the minimum distance between box walls and the protein was set to be larger than 10Å. Eight chloride ions were placed randomly in the box in order to neutralize the total charge. In the simulation system, the number of all the atoms was 135470 including the protein complexes and waters. The remaining box volume was filled by the simple point charge (SPC) water [41]. Before the simulation, an energy minimization was applied to the full system with no constraints using the steepest descent integrator for 13000 steps, and then the system was equilibrated by a 200 ps MD simulation at 300 K. Finally, a 5 ns simulation was

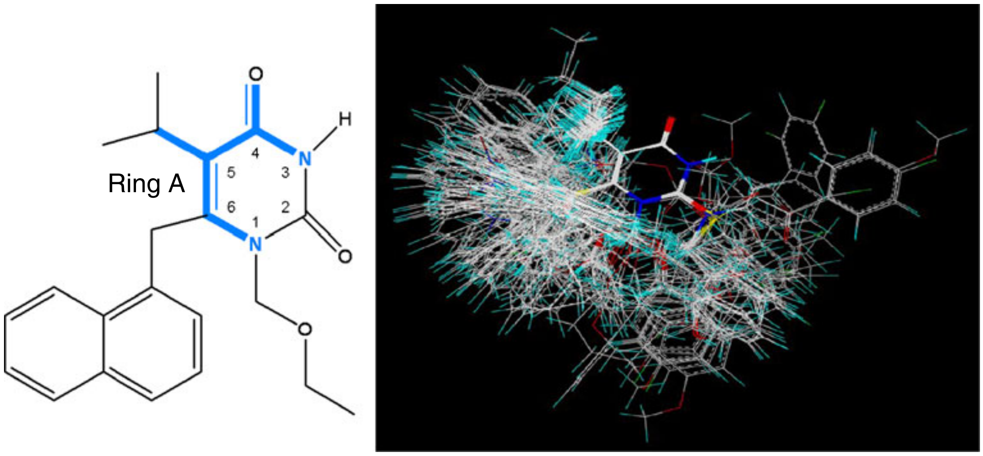


Fig. 2 Molecular alignment of compounds in the data set. (**a**) Common substructure of the molecules is shown in blue based on template compound **19**. (**b**) Ligand-based alignment of all the compounds

performed with a time step of 2 fs. During MD simulation process, main calculation methods and the standard parameters were set as listed: The model used normal pressure and temperature (NPT) ensemble at 300 K with periodic boundary conditions, the temperature remained constant by the Berendsen thermostat, the values of the isothermal compressibility were set to 4.5×10^{-5} bar^{-1} while the pressure was maintained at 1 bar using the Parrinello-Rahman scheme [42], electrostatic interactions were calculated using the particle mesh Ewald method [43], cut-off distances for the calculation of Coulomb and van der Waals interactions were 1.0 and 1.4 nm, respectively. All the MD simulations lasted for 5 ns in order to ensure that the whole systems were stable.

Results

CoMSIA analyses

In our research, both the ligand- and receptor- based 3D-QSAR studies were carried out, resulting in series of models. To evaluate the reliability of these models, all crucial statistical parameters were analyzed, including the Q^2 (leave-one-out), non cross-validated correlation coefficient ($R_{ncv}{}^2$), SEE, F-statistic values and predicted correlation coefficient ($R_{pre}{}^2$). As a result, the most optimal and robust one is the ligand-based CoMSIA model (Table 1) which is superior to all the receptor-based models as well as the ligand-based CoMFA ones, thus in the following parts only this model is further analyzed.

Table 1 Summary of CoMSIA and CoMFA results

PLS statistics	CoMSIA	CoMFA
Q^2	0.480	0.339
$R_{ncv}{}^2$	0.847	0.948
SEE	0.334	0.198
F	69.586	155.267
$R_{pre}{}^2$	0.745	0.737
SEP	0.397	0.403
OPN	7	10
Contribution:		
Steric	0.308	0.989
Hydrophobic	0.654	-
ESpm06u	0.038	0.011

Q^2, cross-validated correlation coefficient after the leave-one-out procedure; $R_{ncv}{}^2$, non-cross-validated correlation coefficient; SEE, standard error of estimate; F, ratio of $R_{ncv}{}^2$ explained to unexplained= $R_{ncv}{}^2/(1-R_{ncv}{}^2)$; $R_{pre}{}^2$, predicted correlation coefficient for the test set of compounds; SEP, standard error of prediction; OPN, optimal number of principal components

In addition, after trying all free possible combinations of the 3D field descriptors (i.e., the steric, electrostatic, hydrophobic, H-bond donor, and acceptor fields in CoMSIA model, and the steric and electrostatic fields in CoMFA) employed as the independent variables, the PLS still cannot obtain statistically satisfactory results. Table 2 shows the optimal results of this attempt, which obviously indicates the failure of using only 3D descriptors for the establishment of the QSAR models. Thus, the aid of two-dimensional descriptor is necessity. Presently, ESpm06u, the spectral moment 06 2D parameter calculated from the edge adjency matrix [44, 45] was used as an additional parameter to build the models, ending up with a CoMSIA model with satisfactory statistics as shown in Table 1. This evidently demonstrates that with the help of ESpm06u which has 3.8% relative contribution to the activity of the inhibitors, the model experiences a modest improvement from the model that lacks the 2D descriptor. The reason might be due to that ESpm06u is a parameter related to the molecular volume, which must have a close connection with the property of the HIV-1 RT inhibition.

As seen from Table 1, the optimal ligand-based CoMSIA model was built based on the employment of steric and hydrophobic field descriptors. Statistically, it used seven optimum numbers of components with an LOO cross-validated Q^2 of 0.480, SEE value of 0.334, and F value of 69.586 obtained, indicating its satisfactory internal predictive capacity. Besides, its high $R_{ncv}{}^2$ of 0.847 for the non-cross-validation presented the self-consistency of the model. While tested by the independent test set, this CoMSIA model exhibited good predictive ability with $R_{pre}{}^2$=0.745 and SEP=0.397. As to the relative contribution of descriptors, the contribution proportion of the hydrophobic feature to the model (65.4%) is 34.6% larger than that of the steric one (30.8%). Supporting Table S10 shows the experimental (pIC_{50Exp}), calculated (pIC_{50Calc})

Table 2 Summary of CoMSIA and CoMFA results without ESpm06u descriptor employed

PLS statistics	CoMSIA	CoMFA
Q^2	0.352	0.349
$R_{ncv}{}^2$	0.887	0.952
SEE	0.300	0.191
F	84.572	167.972
$R_{pre}{}^2$	0.952	0.716
SEP	0.1727	0.419
OPN	10	10
Contribution:		
Steric	0.341	1
Hydrophobic	0.659	-

and residual (pIC_{50Exp} – pIC_{50Calc}) potency values of the optimal CoMSIA model for all training and test set compounds. Figure 3 depicts the actual versus predicted pIC_{50} values plot for both the training (filled blue diamond) and test (filled black square) set molecules of the whole dataset based on ligand-based CoMSIA model.

It should be noted that during the modeling process, an initial inspection of the fitted/predicted activities identified four molecules (38, 54, 64, 99) which are regarded as outliers and then discarded in the model generation. The examination of outliers may, sometimes, provide additional information with the properties, and thus in the present study these chemicals are checked carefully: i) in structure, compound 64 belongs to skeleton type E (supporting Table S5). However, it exhibited the lowest potency with pIC_{50}=−1.41 in this skeleton type, a value much lower than the group's average activity of −0.32. Thus a different interaction mode from the other RT inhibitors, we speculate, might be the reason leading to its large, also the largest residual ($pIC_{50,cal}$ – $pIC_{50,exp}$) value among the whole dataset of −2.836. ii) Among all chemicals with skeleton type C in the data set (supporting Table S3), molecule 38 has the largest substituent, namely dibutylamine at position 2. While in activity, compound 38 has the lowest experimental activity as well as the largest experimental error value of 53.37 μM in the whole dataset. As a matter of fact, this is a much larger error than others since the error of other compounds in this skeleton only range from 0.00-14.06 μM. From the above analysis, we assume that it is either this unique structure or the lesser experimental precision which causes the molecule's large prediction residual. iii) As for compounds 54 and 99, they both have very high residuals between the experimental and predicted activity (with pIC_{50} residual of −0.965 and 1.359, respectively) and thus are treated as outliers. This discrepancy, we guess, on one side implies that these particular compounds may not be typical of the dataset that follows the general structure-activity rule, and on the other side, indicates the necessity to incorporate more accurate experimental data with more diversified molecular structures to the dataset with a purpose to improve the generalization ability of the 3D-QSAR models.

CoMSIA contour maps

One of the benefits about 3D-QSAR is that its results can be visualized through contour maps, which are calculated as the product of the field standard deviation (StDev) at each grid point and the coefficient from the PLS analysis (StDev*-Coeff), describing regions near the molecules where a substituent, with a particular peculiarity (in these models, steric and hydrophobic fields) is able to increase or to decrease the biological potency. Thus, presently, the hydrophobic and steric fields from the best CoMSIA model are also represented as 3D colored contour maps in Fig. 4a and b, respectively, using compound 19 shown as an example. The individual contributions from the hydrophobic and steric fields are 65.4% and 30.8%, respectively, indicating that hydrophobicity has a much greater impact on the peculiarity of the ligand than steric property.

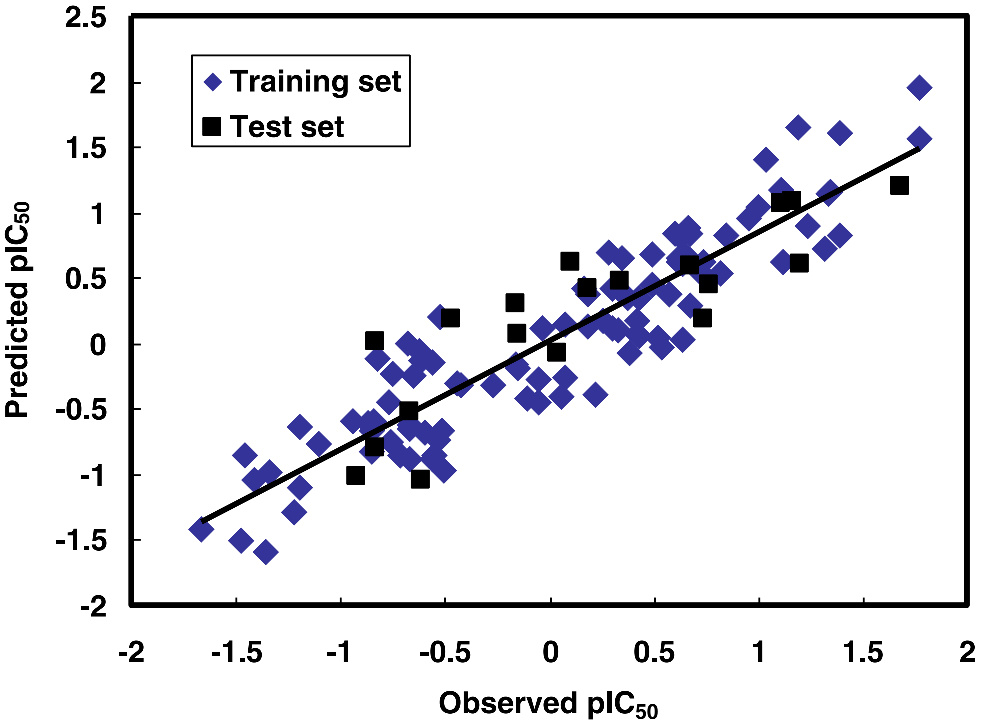


Fig. 3 The plots of the $pIC_{50\ Exp}$ versus the $pIC_{50\ Calc}$ values for the training and test sets of CoMSIA model

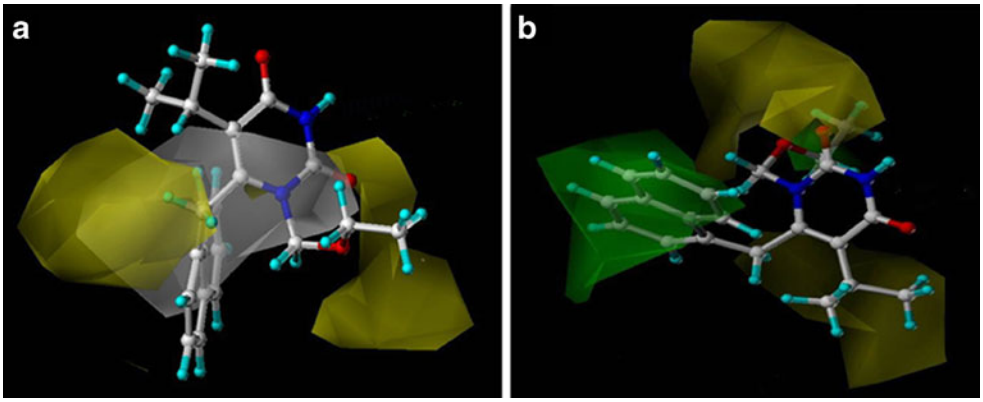


Fig. 4 CoMSIA StDev*Coeff contour plots. (**a**) Hydrophobic contour map (yellow/white) in combination with compound **19**. Yellow contours indicate regions where hydrophobic substituents enhance activity; and white contours indicate regions where hydrophilic substituents enhance activity; (**b**) Steric (green/yellow) contour map in combination with compound **19**. Green contours indicate regions where bulky groups increase activity; while yellow contours indicate regions where bulky groups decrease activity

Figure 4a depicts an overlay of the hydrophobic CoMSIA field on the compound. Yellow contours encompass areas where hydrophobic groups will enhance the biological activity, while a hydrophobic group located near the white regions will result in impaired biological activity. There are yellow isopleths on position- 1, 2 which indicates that hydrophobic groups (like -OMe, -OEt, -F, -Cl, -Br) are beneficial to the activity. This is illustrated by the example of compounds 21–23 with $PhCH_2$ at this position exhibiting a much higher activity than compounds 9–12 with the hydrophilic group $HOCH_2$. There is another yellow contour on position-6, so the conformation of the hydrophobic groups in these inhibitors in this position of the central core ring is favorable for interaction of the molecules with the RT. This can be verified by the larger potency of compounds 10 and 13 than molecules 1 and 2. A big white plot appears through the central core suggesting an increase in the activity with the presence of hydrophilic group (like hydroxy or amido) in this region.

In the CoMSIA steric field contour map in Fig. 4b, areas where steric bulk substituents enhance or decrease the potency are represented by green or yellow polyhedrons, respectively. A green contour near position-6 of the analogues implies that bulky substituents at this position strengthen the activity. Thus, molecules carrying a bulky substituent around the areas should be more active than those with a smaller or without substituent, which is illustrated by the fact that molecules 100 and 102 with a bulky substituent of 1-naphthyl exhibited higher potency (with pIC_{50} values of −0.5224 for compound 100 and 0.3468 for 102, respectively) than compounds 103 and 104 (with pIC_{50} values of −0.5478 and −0.8274, respectively) with a phenyl group in position-5. A yellow contour appearing above position- 1 and 2 suggests that bulky substituents at this position reduce the activity, which is proved by the fact that compounds 35–38 have the substituent structures of Pr, i-Pr, C_6H_{11} and Bu in this position, respectively, which are increasingly bulky in size, but the activities of them reduce gradually.

Docking results

Once the crystallography presents, docking is an attractive way to find the optimal orientation of the ligand in the binding pocket of the pharmaceutical target protein/enzyme, which cannot be completed only by QSAR studies. Thus, presently, molecular docking was also carried out using the crystal structure of HIV-1 RT (PDB ID: 1RT1) complexed with MKC- 442 (resolution value is 2.55 Å) obtained from the Protein Data Bank, due to the reason that the original ligand of MKC- 442 is very similar in structure to our dataset molecules. In the present work, all 119 compounds in the dataset were docked into the possible active site of HIV-1 RT crystal structures, and the optimal conformations of the molecules were determined, with the highest score of 7.65 obtained for template compound 19. All molecules in the series were set well in the binding pocket demonstrating the quality of the docking model. Diagram showing the interactions of RT with inhibitor 19 is provided in Fig. 5.

As seen from this figure, obviously the binding pocket is basically hydrophobic, which is made up of many acid residues such as Tyr 188, Tyr 183, Tyr 181, Trp 229, Phe 227, Val 179, Tyr 318 (Fig. 5a) and this observation correlates well with our previous contour map (Fig. 4a). For example, the two hydrophobic acid residues Trp 229 and Tyr 318 are in conformity with the yellow contour on position- 1 and 2 which indicate the preference of the molecules for hydrophobic environment. In addition, the presence of hydrophobic Val 179, Tyr 188 and Val 108 satisfies the yellow contour near position-6, and the existence of the hydrophilic Pro 95 is in line with the hydrophilic favored white contour above the ring, respectively.

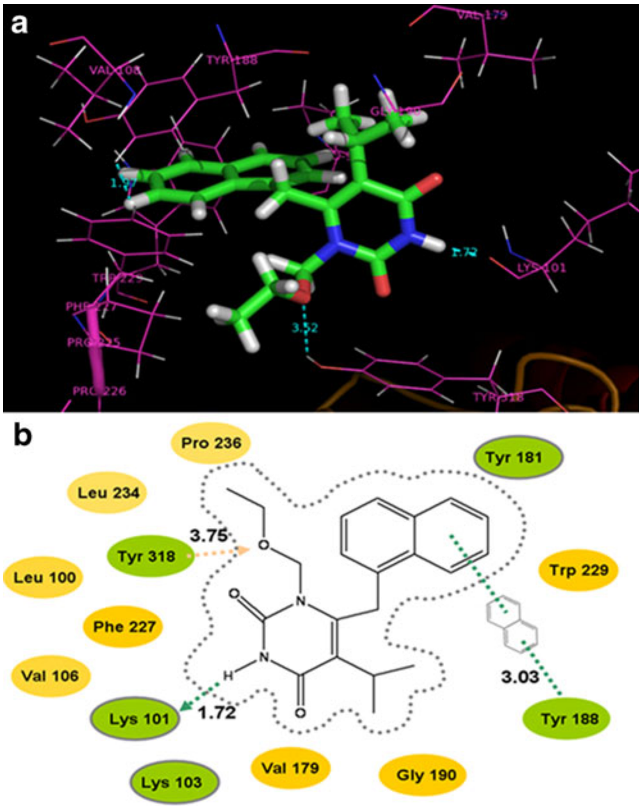


Fig. 5 The binding site formed around compound **19**. (**a**) Representative interactions with the amino acids. The dashed lines show the formation and distance of the hydrogen bonds. Active site amino acid residues are represented as lines, the inhibitor is shown as stick model, respectively. (**b**) The active site residues are represented as follows: polar residues in green, hydrophobic residues in yellow. Green and orange arrows indicate hydrogen bonding to side-chain and backbone atoms respectively. A naphthyl icon represents a π-π stacking interaction. The dotted contour reflects steric room for methyl substitution

Hydrogen bonding plays an important role in determining a molecule's physiological or biochemical role, and there is no exception in HIV-1 reverse transcriptase. We can see from Fig. 5b that two hydrogen bonds have been formed between the HIV-1 RT and ligand: 1) The backbone -CO- in Lys 101 (hydrogen bonding donor) forms a hydrogen bond with the -H atom in ring A with a distance of 1.72Å. 2) The substituent on position-1 contains an oxygen atom satisfying the requirement for a hydrogen bond acceptor with Tyr 318, and thus forms a H-bond with a distance of 3.52Å. The two H-bonds are vital, as they respectively pin the skeleton ring and one 'leg', i.e., the position-1 substituent in one side, of the molecules in the binding site. Besides these H-bonds, another phenomenon was also observed that the ring on the substituent on position-1 is also engaged in a π-π interaction with Tyr 188 (1.97Å), which fastens another 'leg', i.e., the position-6 substituent on the other side, of the inhibitors. Thus, a conclusion can be drawn that it is just the two hydrogen bonds and the π-π interaction that act a role as three anchors that fix the three-dimensional active orientation of the ligand in the binding pocket.

In addition, as revealed from previous 3D-QSAR analysis that steric interaction also contributes a lot to the model, the steric requirements of the binding pocket for the ligand is also analyzed and compared with the contour maps as follows: 1) Green contours are found unoccupied in Fig. 4b near the ligand site where any bulky substitution is favored. The reason may be that the presence of a phenyl ring will lead to a π-π interaction with Tyr 188, so a bulky substitution like a phenyl group will be beneficial to the increase in activity. 2) The presence of the yellow isopleths on position-5 of ring A can be explained by the fact that Gly 190 is only 2.10 Å away from the isopropyl group on position-5 of ring A. So when this position is substituted by any bulky molecule, it is obviously uneasy for the substituted new molecule to enter into the binding pocket. The yellow isopleths near position- 1 and 2 of the ring are taken up by Phe 227 and Pro 225, Pro 236, therefore any substituent bulkier than benzene is disfavored there. The above results, once again, demonstrate the reliability of the 3D-QSAR model by the good agreement between the docking results and the contour maps.

Molecular dynamics simulations

In order to take into account the protein flexibility (which cannot be fulfilled by the molecular docking process), the behavior of the docked complex is researched in a dynamic context, i.e., the MD simulation. Using the GROMOS96 force field, the MD simulations were conducted by the GROMACS package 4.0.7. The molecular topology files were created by the program PRODRG 2.5. We carried out 5 ns molecular dynamics simulations of HIV-1 RT with ligand 19 on the basis of the docked complex structure, so that a dynamical picture of the conformational changes in the HIV-1 RT binding site was taken. The RMSDs of the trajectory in regard to the initial structure ranging from 0.200 to 0.500 nm are presented in Fig. 6a. As a result, after 3000 ps the RMSD of the complex attains about 0.450 nm and almost remains this value for the whole process. This clearly indicates metastable conformation after 3000 ps of simulation for the docked complex structure. Figure 6b depicts a superposition of the average structure for the last 1 ns and the docked structure, where ligand 19 is shown in green stick for the initial complex and blue stick for the final average complex, separately. Obviously, there is no significant change between the docked model of the complex and the average structure obtained from MD simulations, which verifies the reasonability of the docking model.

The only difference is that naphthylmethyl substituent on position-6 of ring A in MD average structure has a torsion angle from ligand 19. This might result from the fact that the naphthylmethyl substituent on position-6 of ring A has a π-π interaction with a ring of Tyr 188. After the optimization in MD simulations, the naphthylmethyl substituent bends to the Tyr 188 in order to obtain a more stable conformation. This good superposition of the docking and MD conformations of the ligand and protein proves another time that our docked model is reliable.

Discussion

The reasonability of the docking pocket

To the best of our knowledge, up to date no research on docking studies of the HEPT, DABO and HIV-1 reverse transcriptase interactions has been reported, except for Hopkins AL's research in 1996 [8]. By docking three RT inhibitors, i.e., the MKC-442, TNK-651 and HEPT, into the HIV-1 NNBP, he observed the existence of many key hydrophobic residues in the binding pocket such as Tyr181, Tyr 188, Phe 227, and Trp 229, which is in good agreement with our observations that the pocket is basically hydrophobic and the ligand interaction with the hydrophobic acid residues is vital to their inhibitory activity. What is more, one of the two H-bonds formed in our docking pocket, the one between the Lys 101 and the ligand, was also observed in Hopkins' work, which verifies again the reliability of our docking model.

Apart from the above consistent results, subtle difference was also observed, i.e., the π-π interaction. Hopkins found a π-π bond formed between the 6-benzyl ring in HEPT and Tyr 181, while in our work it is between the 6-naphthylmethyl substituent of the inhibitor and Tyr 188. Two reasons may account for this difference: 1) In his work, only three HEPT molecules were docked while in ours all 119 molecules which belong to HEPT and DABO two types of RT inhibitors with IC_{50} activity range of 0.017-237.740 μM were docked to the binding site for analysis; 2) The compounds in our dataset all have a naphthyl sub-structure on position-2, while in his it is a benzyl ring on the same location.

Comparison between HEPT and DABO derivatives

The 119 compounds of our dataset include 73 HEPT and 46 DABO analogues. The reason we put these two derivatives together to analyze is that both DABOs and HEPTs belong to the 4-pyrimidinone series, and this structural similarity, we assume, might bring them some similar interaction mechanism in their anti-HIV activity. Just as expected, both the contour map and docking pocket analysis reveal their high similarity in the structural features impacting the RT inhibitory activities. However, the slight difference in the

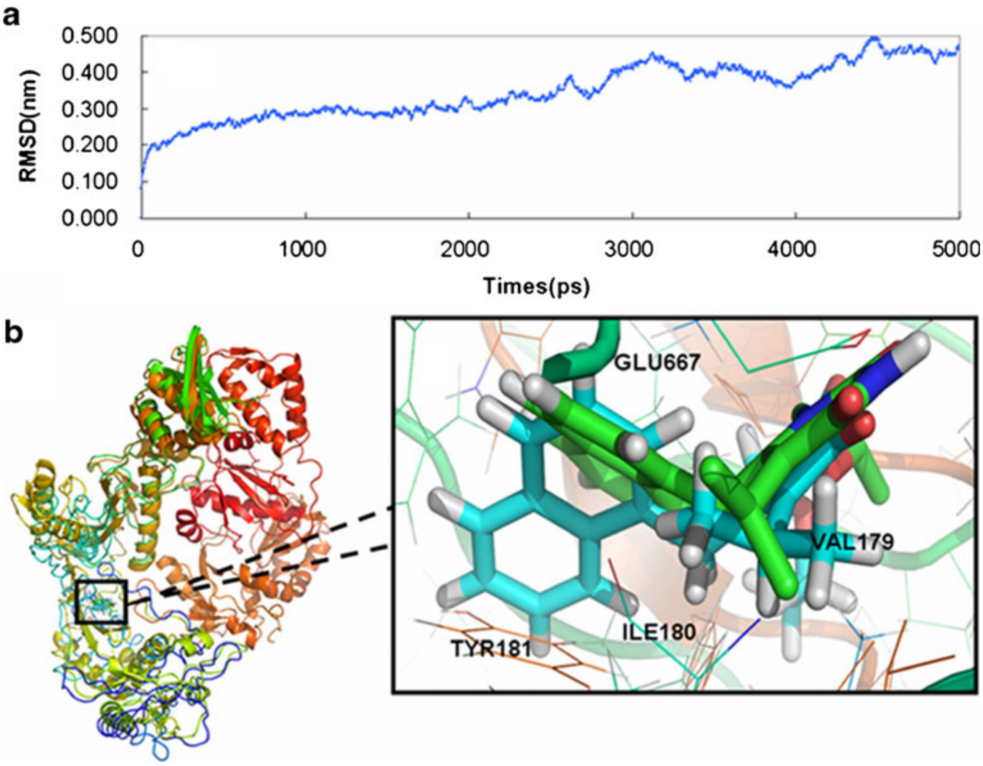


Fig. 6 MD simulation results. (**a**) Plot of the RMSD of docked complex versus the MD simulation time in the MD-simulated structures. (**b**) Structural superposition of the MD simulation and the initial structure for HIV-1 RT. The projection highlights the superimposed backbone atoms of the average structure of the last 1 ns of the MD simulation (blue) and the initial structure (green) for compound **19** and HIV-1 RT complex

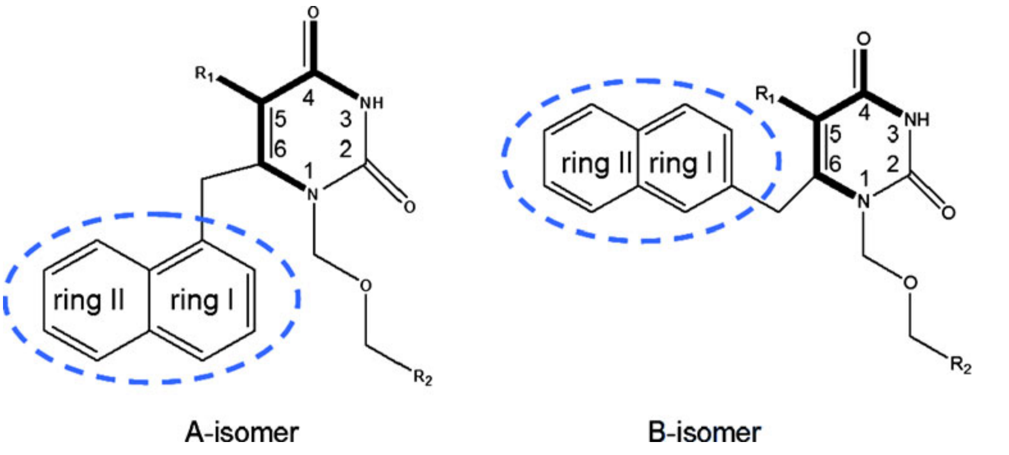
structure also leads to a big difference in biological potency. Firstly, the aldehyde group on position-2 in HEPTs brings more benefits to the compounds than other bulkier substituents on this position in DABOs on account of the yellow isopleths in that position in the contour map (Fig. 4b) that indicate that bulky substituents on position-2 will impair the potency. Secondly, an oxygen atom in the substituent on position-1 of HEPTs makes up a hydrogen bond with Tyr 318, which helps to improve the biological activity strongly by fixing the active conformation. Therefore, this substituent in HEPTs is important in enhancing the activity. In conclusion, these HEPT and DABO derivatives are not exactly the same in structure, but they interact with the HIV-1 RT in a similar mechanism.

Analysis of the two isomers of HEPTs

In structure, 16 6-naphthylmethyl substituted HEPT derivatives in our dataset can be divided into two types of location isomers, i.e., A- and B- isomers [46], which are structurally similar but greatly different in terms of potency. That is to say, all A-isomers exhibited higher pIC_{50} values than their corresponding B-isomers. For instance, compounds 10 and 1, a pair of isomers, are similar in structures but very different in potency (0.19 μM for A- and −1.66 μM for B-isomers, respectively). Table 3 shows all the structures and inhibitory activities of the HEPT isomers. This big difference, by docking analysis we assume, attributes directly to the distinct binding conformations of the 6-naphthylmethyl substituent. By analysis of the docking conformation of all pairs of A-/B-isomers, four possible reasons leading to the potency difference are found.

The first reason is due to one π-π bond formed between the 6-naphthylmethyl substituent of the inhibitor and one different amino acid residue of the protein (Tyr 188 for A-isomer but Trp 229 for B-isomer, respectively). Out of the eight pairs of isomers two pairs are such cases. Figure 7a and b show the docking results of molecules 1 and 10, 5 and 17, the two pairs of isomers, and the π-π bonds formed in the binding site, respectively. All A-isomers are shown in

Table 3 All HEPT isomers in the dataset[†]



Substitute		A-isomer		B-isomer	
R_1	R_2	No.	pIC_{50} (μM)	No.	pIC_{50} (μM)
Et	$HOCH_2$	10	0.1878	1	-1.6590
Me	H	13	-0.1617	2	-1.1933
Et	H	14	0.6716	3	-0.9174
Me	Me	16	0.6478	4	-0.5145
Et	Me	17	1.3872	5	-0.5043
Me	$PhCH_2$	21	1.3872	6	-0.6092
Et	$PhCH_2$	20	1.3872	7	-0.1611
n-Pr	$PhCH_2$	22	0.6308	8	-0.8253

[†] The 2-naphthyl ring (composed of rings I and II) highlighted in the dotted blue circles is the molecular region whose binding conformation determines the different potency of A- and B- isomers as we discovered

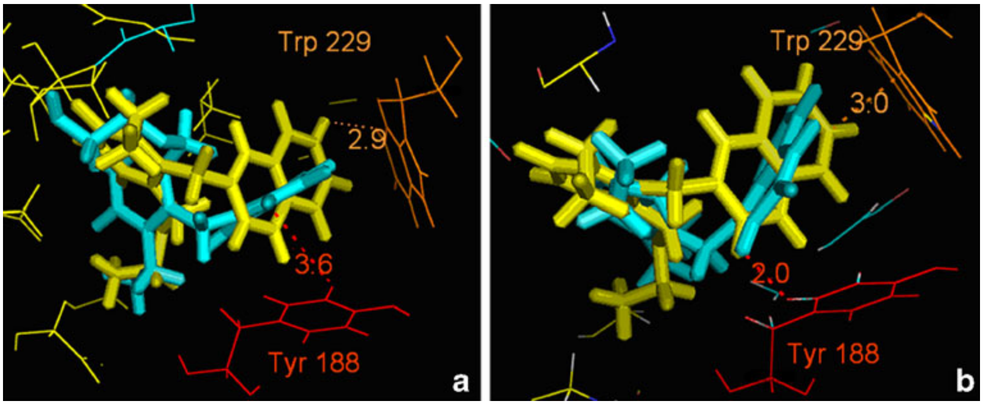


Fig. 7 The docking results of two pairs of isomers (compounds 1 and 10, 5 and 17) and the π-π bonds formed in the binding site. Tyr 188 and Trp 229 are shown in red and orange skeletons, respectively. All A-isomers are shown in blue and B-isomers in yellow, respectively. (**a**) A-isomer compound 10 in blue and corresponding B-isomer compound 1 in yellow are shown. (**b**) A-isomer compound 17 in blue and corresponding B-isomer compound 5 in yellow are presented

blue and B-isomers in yellow, respectively. If with no specification in other figures (Figs. 8, 9, 10), the colors of A- or B- isomers are the same meanings as depicted in Fig. 7. Two specific acid residues, Tyr 188 and Trp 229, are shown in red and orange skeletons, respectively. As shown in this figure, for both pairs a π-π bond is formed between the 6-naphthylmethyl substituent of the molecule and the protein, which to some extent fastens the binding of the ligand with the receptor. But the amino acid residue constructing this π-π bond is different in that for A-isomers, the bond is formed via Tyr 188 residue in the binding pocket, but for B-isomers the bond is built via another acid residue Trp 229. The different acid residues employed by A- and B- isomers to form the π-π bond may be the reason that leads to their distinct inhibitory activities.

The second reason may reside in the difference in the number of π-π bonds formed between the 6-naphthylmethyl substituent of the inhibitor and the acid residues of the protein, i.e., one bond for A-isomers but two bonds for B-ones. In fact, three pairs of A-/B-isomers (compounds 2 and 13, 4 and 16, 7 and 20) are such cases. Figure 8 shows the docking results of these pairs and the π-π bonds (in blue for A-isomer and in red dashed lines for B-isomer, respectively) formed in the binding pocket, where Tyr 188 is shown in red. As seen from this figure, clearly for all three pairs of isomers, the 6-naphthylmethyl substituent in A-isomers establishes one π-π bond with the Tyr 188 (with distance of 3.2, 3.2, 2.2 Å for A-isomer compounds 13, 16 and 20 respectively). While in corresponding B-isomers, this substituent builds two π-π bonds via the same residue (with distance of 5.1 and 4.2 Å for compound 2, 5.1 and 4.4 Å for compound 4, 2.7 and 2.5 Å for compound 7, respectively). Even though B-isomers form one more π-π bond within the binding pocket, the average bond distance of B-isomers are longer than those of A-isomers. Thus we assume the stronger π-π interaction between A-isomers and Tyr 188 residue in the binding pocket than corresponding B-isomers may result in the better inhibitory activity of A- than B- isomers.

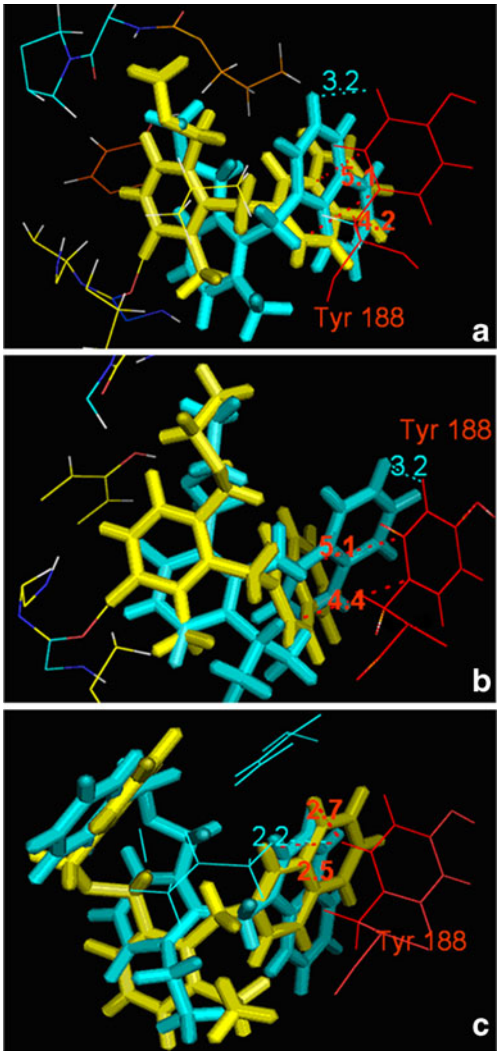


Fig. 8 The docking results of three pairs of isomers (compounds 2 and 13, 4 and 16, 7 and 20) and the π-π bonds formed in the binding site. Tyr 188 is shown in red skeleton. All A-isomers are shown in blue and B-isomers in yellow, respectively. The π-π bonds formed with A-isomers are in blue, and with B-isomers red. (**a**) A-isomer compound 13 in blue and corresponding B-isomer compound 2 in yellow are shown. (**b**) A-isomer compound 16 in blue and corresponding B-isomer compound 4 in yellow are shown. (**c**) A-isomer compound 20 in blue and corresponding B-isomer compound 7 in yellow are shown

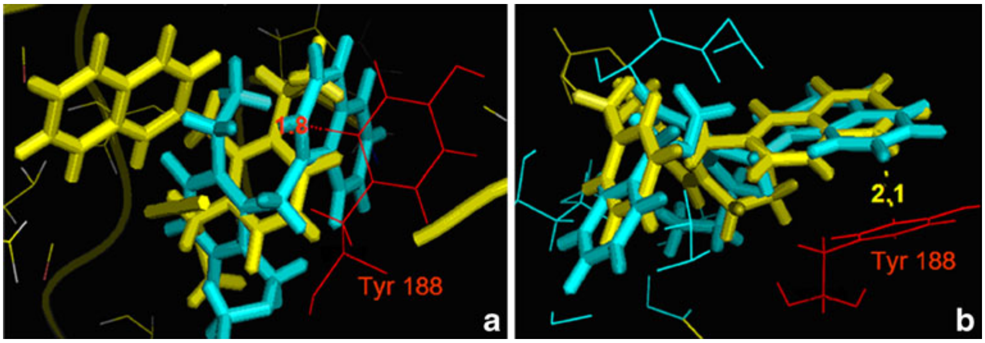


Fig. 9 The docking results of two pairs of isomers (compounds 3 and 14, 8 and 22) and the π-π bond (only formed in A-isomers) in the binding site. Tyr 188 is shown in red skeleton. All A-isomers are shown in blue and B-isomers in yellow, respectively. (**a**) A-isomer compound 14 in blue and corresponding B-isomer compound 3 in yellow are shown. (**b**) A-isomer compound 22 in blue and corresponding B-isomer compound 8 in yellow are presented

The greater potency of A- than B-isomers may also be ascribed to the third reason that all A-isomers form π-π bonds (between the 6-naphthylmethyl substituent and Tyr 188), but for some B-isomers no such interaction exists in the active site. Two pairs of isomers, compounds 3 and 14, 8 and 22, are such cases as shown in Fig. 9, where Tyr 188 is depicted in red skeleton. As seen from Fig. 9a, the 6-naphthylmethyl substituent in A-isomer compound 14 (in blue) establishes a π-π bond with a distance of 1.8Å with Tyr 188, which does not exist in corresponding B-isomer compound 3 (in yellow) at all. Figure 9b is the same situation where A-isomer compound 22 forms a 2.1Å π-π bond with Tyr 188, but B-isomer compound 8 does not. The intimate interactions caused by this π-π bond may account for the improvement of the potency of A-isomers compared with corresponding B-ones.

The last reason may lie in the different position that the π-π bonds are formed at between A- and B- isomers. Since many isomers in the dataset are under such conditions, only one pair (compounds 6 and 21) is shown as an illustration in Fig. 10, where the docking conformations and the π-π bonds formed in the binding pocket of the molecules are depicted. Tyr 188 is specifically shown in red skeleton. A-isomer compound 21 is shown in blue and corresponding B-isomer of compound 6 shown in yellow, respectively. The π-π bond formed between the A-isomer and Tyr 188 is shown in blue, and the bond formed between corresponding B-isomer and Tyr 188 is in red, respectively. It is true that all π-π bonds are created between Tyr 188 and the 6-naphthylmethyl substituent of the molecule as described above, but this substituent is still composed of two rings I and II (as seen in Table 3). It seems that when the π-π bond is formed between Tyr 188 and ring I, the potency of the inhibitor is stronger. While when the bond is constructed between the same acid residue and ring II, the inhibitory activity decreases to some extent. Actually among all eight pairs of A-/B-isomers we studied, A-isomers in six pairs (compounds 10, 13, 14, 17, 20, 21) have ring I connected to the Tyr 188 to form the π-π bond. However for B-isomers, things are complicated such that they either: 1) form two weak π-π bonds with Try 188 (with longer bond length) via both rings I and II as described previously, which cannot be compared with corresponding A-isomers, or 2) have no π-π interaction with the residue, or 3) form only one π-π bond with Try 188, in which case the bonds are found all built via the ring II in the 6-naphthylmethyl substituent of the molecules. B-isomers in three pairs (compounds 1, 5 and 6) are the last cases. For example, as shown in Fig. 10, Tyr 188 has a π-π bond with ring I in the 6-naphthylmethyl substituent of A-isomer compound 21, but with the ring II of B-isomer molecule 6. Thus we assume that the ring I connected π-π interaction may improve the potency of the inhibitor more than the ring II connected π-π interaction.

More conclusions would be made, if more pairs of A-B-isomers can be taken into the research. In a word, we believe it is the above reasons that cause the higher inhibitory

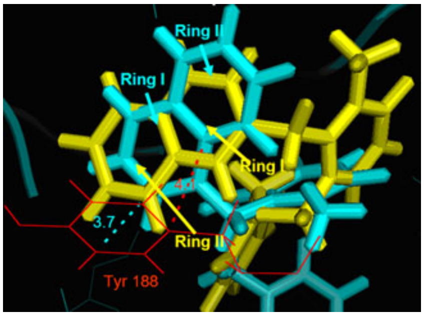


Fig. 10 The docking results of one pair of isomers (compounds 6 and 21) and the π-π bonds formed in the binding site. Tyr 188 is specifically shown in red skeleton. A-isomer of compound 21 is shown in blue and corresponding B-isomer of compound 6 in yellow, respectively. The π-π bond formed between the A-isomer and the Tyr 188 is shown in blue. The π-π bond formed between corresponding B-isomer and the Tyr 188 is in red

activities of A- than B- isomers, and the protruding direction of the naphthyl plane group is very crucial for the inhibitory activities of the RT ligands.

Conclusions

In this paper, predictive 3D-QSAR models were built on 119 HEPT and DABO HIV-1 reverse transcriptase inhibitors. The best prediction was developed by the ligand-based CoMSIA model with an LOO cross-validated Q^2 of 0.480, SEE value of 0.334, and F value of 69.586, $R_{ncv}{}^2$ of 0.847, $R_{pre}{}^2$ of 0.745 and SEP of 0.397, indicating its satisfactory predictive capability. Furthermore, the docking, MD results and the 3D-QSAR models correlated very well with each other, and the key acid residues of the binding pocket are identified. Our conclusions are: 1) The binding pocket of HEPT and DABO inhibitors are essentially hydrophobic, and hydrophobic substituents on position- 1, 2 and 6 are helpful for the potency. 2) Bulky groups in position-6 enhance, while in position- 1, 2 and 5 impair the activity. 3) Two hydrogen bonds are formed between Tyr 318 and the O atom in 1-substituent (3.52Å), Lys 101 and the H atom in ring A (1.72Å) respectively, and a π-π interaction is produced between Trp 188 and the ring on 1-substituent. It is just these three interactions that stabilize the ligand-RT complex, by acting as three anchors to fix the active conformation of the ligand in the binding pocket. 4) Despite the structural difference, DABO and HEPT derivatives employed a similar interaction mechanism to RT. 5) For HEPT derivatives, the activity difference between the two isomers (A and B) may be directly due to the distinct locations of the 6-naphthylmethyl substituent and the reasons are specified. All these results could be employed to alter the structural scaffold in order to develop new HIV-1 RT inhibitors that have an improved biological property.

Acknowledgments This work is financially supported by the National Natural Science Foundation of China (Grants No. 10801025). The authors are also very grateful to Prof. Ling Yang for access to SYBYL software.

Reference

1. John PB, François P, Patrick G (2011) Global trends in AIDS mortality. In: Richard GR, Eileen MC (eds) The international handbook of adult mortality. Springer, New York, pp 171–183
2. Barre-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, Dauguet C, Axler-Blin C, Vezinet-Brun F, Rouzioux C, Rozenbaum W, Montagnier L (1983) Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220:868–871
3. Gallo RC, Sarin PS, Gelmann EP, Robert-Guroff M, Richardson E, Kalyanaraman VS, Mann D, Sidhu GD, Stahl RE, Zolla-Pazner S, Leibowitch J, Popovic M (1983) Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). Science 220:865–867
4. Morris MC, Robert-Hebmann V, Chaloin L, Mery J, Heitz F, Devaux C, Goody RS, Divita G (1999) A new potent HIV-1 reverse transcriptase inhibitor. A synthetic peptide derived from the interface subunit domains. J Biol Chem 274:24941–24946
5. Gait MJ, Karn J (1995) Progress in anti-HIV structure-based drug design. Trends Biotechnol 13:430–438
6. Benbrik E, Chariot P, Bonavaud S, Ammi-SaId M, Frisdal E, Rey C, Gherardi R, Barlovatz-Meimon G (1997) Cellular and mitochondrial toxicity of zidovudine (AZT), didanosine (ddI) and zalcitabine (ddC) on cultured human muscle cells. J Neurol Sci 149:19–25
7. De Clercq E (1993) HIV-1-specific RT inhibitors: highly selective inhibitors of human immunodeficiency virus type 1 that are specifically targeted at the viral reverse transcriptase. Med Res Rev 13:229–258
8. Hopkins AL, Ren J, Esnouf RM, Willcox BE, Jones EY, Ross C, Miyasaka T, Walker RT, Tanaka H, Stammers DK, Stuart DI (1996) Complexes of HIV-1 reverse transcriptase with inhibitors of the HEPT series reveal conformational changes relevant to the design of potent non-nucleoside inhibitors. J Med Chem 39:1589–1600
9. Ren J, Esnouf R, Garman E, Somers D, Ross C, Kirby I, Keeling J, Darby G, Jones Y, Stuart D et al (1995) High resolution structures of HIV-1 RT from four RT-inhibitor complexes. Nat Struct Biol 2:293–302
10. Geretti AM (2006) Resistance to non-nucleoside reverse transcriptase inhibitors. Antiretroviral Resistance in Clinical Practice. Mediscript, London, Chapter 2
11. De Clercq E (2009) Anti-HIV drugs: 25 compounds approved within 25 years after the discovery of HIV. Int J Antimicrob Agents 33:307–320
12. De Clercq E (2009) The history of antiretrovirals: key discoveries over the past 25 years. Rev Med Virol 19:287
13. De Clercq E (2011) A 40-year journey in search of selective antiviral chemotherapy. Annu Rev Pharmacol Toxicol 51:1–24
14. Van de Waterbeemd H (1995) Introduction In Chemometric Methods in Drug Design. Chemometric Methods in Drug Design. VCH, Weinheim, chapter 1
15. Hao M, Li Y, Zhang S-W, Yang W (2011) Investigation on the binding mode of benzothiophene analogues as potent factor IXa (FIXa) inhibitors in thrombosis by CoMFA, docking and molecular dynamic studies. J Enzyme Inhib Med Chem. doi:10.3109/14756366.2011.554414
16. Liu J, Li Y, Zhang S, Xiao Z, Ai C (2011) Studies of new fused benzazepine as selective dopamine D3 receptor antagonists using 3D-QSAR, molecular docking and molecular dynamics. Int J Mol Sci 12:1196–1221
17. Zhang B, Li Y, Zhang H, Ai C (2010) 3D-QSAR and molecular docking studies on derivatives of MK-0457, GSK1070916 and SNS-314 as inhibitors against Aurora B kinase. Int J Mol Sci 11:4326–4347
18. Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967
19. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
20. Baba M, Tanaka H, De Clercq E, Pauwels R, Balzarini J, Schols D, Nakashima H, Perno CF, Walker RT, Miyasaka T (1989) Highly specific inhibition of human immunodeficiency virus type 1 by a novel 6-substituted acyclouridine derivative. Biochem Biophys Res Commun 165:1375–1381

21. De Clercq E (1999) Perspectives of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection. II. Farmaco 54:26–45
22. Campiani G, Ramunno A, Maga G, Nacci V, Fattorusso C, Catalanotti B, Morelli E, Novellino E (2002) Non-nucleoside HIV-1 reverse transcriptase (RT) inhibitors: Past, present, and future perspectives (2002). Curr Pharm Design 8:615–657
23. Artico M, Massa S, Mai A, Marongiu EM, Piras G, Tramontano E, Colla LP (1993) 3,4-Dihydro-2-alkoxy-6-benzyl-4-oxopyrimidines (DABOs): a new class of specific inhibitors of human immunodeficiency virus Type 1. Antiviral chemistry & chemotherapy, vol 4. International Medical Press, London, pp 361–368
24. He Y, Chen F, Yu X, Wang Y, De Clercq E, Balzarini J, Pannecouque C (2004) Nonnucleoside HIV-1 reverse transcriptase inhibitors; part 3. Synthesis and antiviral activity of 5-alkyl-2-[(aryl and alkyloxyl-carbonylmethyl)thio]-6-(1-naphthylmethyl) pyrimidin-4(3H)-ones. Bioorg Chem 32:536–548
25. Ji L, Chen F-E, Feng X-Q, De Clercq E, Balzarini J, Pannecouque C (2006) Non-nucleoside HIV-1 reverse Transcriptase inhibitors, part 7. Synthesis, antiviral activity, and 3D-QSAR investigations of novel 6-(1-naphthoyl) HEPT analogues. Chem Pharm Bull 54:1248–1253
26. Meng G, Chen FE, De Clercq E, Balzarini J, Pannecouque C (2003) Nonnucleoside HIV-1 reverse transcriptase inhibitors: Part I. Synthesis and structure-activity relationship of 1-alkoxymethyl-5-alkyl-6-naphthylmethyl uracils as HEPT analogues. Chem Pharm Bull (Tokyo) 51:779–789
27. Sun GF, Chen XX, Chen FE, Wang YP, De Clercq E, Balzarini J, Pannecouque C (2005) Nonnucleoside HIV-1 reverse-transcriptase inhibitors, part 5. Synthesis and anti-HIV-1 activity of novel 6-naphthylthio HEPT analogues. Chem Pharm Bull (Tokyo) 53:886–892
28. Sun GF, Kuang YY, Chen FE, De Clercq E, Balzarini J, Pannecouque C (2005) Non-nucleoside HIV reverse transcriptase inhibitors, part 6[1]: synthesis and anti-HIV activity of novel 2-[(arylcarbonylmethyl)thio]-6-arylthio DABO analogues. Arch Pharm (Weinheim) 338:457–461
29. Wang Y, Chen FE, Balzarini J, De Clercq E, Pannecouque C (2008) Non-nucleoside HIV-1 reverse-transcriptase inhibitors. Part 10. Synthesis and anti-HIV activity of 5-alkyl-6-(1-naphthylmethyl)pyrimidin-4(3H)-ones with a mono- or disubstituted 2-amino function as novel 'dihydro-alkoxy-benzyl-oxopyrimidine' (DABO) analogues. Chem Biodivers 5:168–176
30. Gasteiger J, Marsili M (1980) Iterative partial equalization of orbital electronegativity–a rapid access to atomic charges. Tetrahedron 36:3219–3228
31. Zhu Y-Q, Lei M, Lu A-J, Zhao X, Yin X-J, Gao Q-Z (2009) 3D-QSAR studies of boron-containing dipeptides as proteasome inhibitors with CoMFA and CoMSIA methods. Eur J Med Chem 44:1486–1499
32. Deshpande S, Jaiswal S, Katti SB, Prabhakar YS (2011) CoMFA and CoMSIA analysis of tetrahydroquinolines as potential antimalarial agents. SAR QSAR Environ Res 1:1–16
33. Gerhard K (2002) Comparative molecular similarity indices analysis: CoMSIA. In: Hugo K, Gerd F, Yvonne CM (eds) 3D QSAR in Drug Design. Springer, New York, pp 87–104
34. Todeschini R, Consonni V, Mauri A, Pavan M (2004) DRAGON - Software for the calculation of molecular descriptors. Rel. 5.2 for Windows. Talete srl, Milano, Italy
35. Klebe G (1994) The use of composite crystal-field environments in molecular recognition and the de Novo design of protein ligands. J Mol Biol 237:212–235
36. Caballero J, Quiliano M, Alzate-Morales JH, Zimic M, Deharo E (2011) Docking and quantitative structure-activity relationship studies for 3-fluoro-4-(pyrrolo[2,1-f][1,2,4]triazin-4-yloxy)aniline, 3-fluoro-4-(1H-pyrrolo[2,3-b]pyridin-4-yloxy)aniline, and 4-(4-amino-2-fluorophenoxy)-2-pyridinylamine derivatives as c-Met kinase inhibitors. J Comput Aided Mol Des 25:349–369
37. Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: A message-passing parallel molecular dynamics implementation. Comput Phys Commun 91:43–56
38. Lindahl E, Hess B, Van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 7:306–317
39. Aalten DMF, Bywater R, Findlay JBC, Hendlich M, Hooft RWW, Vriend G (1996) PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. J Comput Aided Mol Des 10:255–262
40. Schuttelkopf AW, van Aalten DMF (2004) PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. Acta Crystallogr Sect D Biol Crystallogr 60:1355–1363
41. Berendsen HFC, Postma JPM, Van Gunsteren WF, Hermans J (1981) Interaction models for water in relation to protein hydration. In: Pullman B (ed) Inermolecular forces. Reidel, Dordrecht, pp 331–342
42. Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. J Appl Phys 52:7182–7190
43. Lin JH, Perryman AL, Schames JR, McCammon JA (2002) Computational drug design accommodating receptor flexibility: The relaxed complex scheme. J Am Ceram Soc 124:5632–5633
44. Estrada E (1995) Edge adjacency relationships and a novel topological index related to molecular volume. J Chem Inf Comput Sci 35:31–33
45. Estrada E (1995) Edge adjacency relationships in molecular graphs containing heteroatoms: a new topological index related to molar volume. J Chem Inf Comput Sci 35:701–707
46. Meng G, Chen FE, De Clercq E (2003) Interactive study between two types of 1-[2-(Hydroxyethoxy)methyl]-6-naphthylmethylthymines and HIV-1 reverse transcriptase. Chin J Process Eng 3:24–28

J Mol Model (2012) 18:2199–2208
DOI 10.1007/s00894-011-1231-0

ORIGINAL PAPER

Automatic prediction of flexible regions improves the accuracy of protein-protein docking models

Xiaohu Luo · Qiang Lü · Hongjie Wu · Lingyun Yang ·
Xu Huang · Peide Qian · Gang Fu

Received: 10 May 2011 / Accepted: 22 August 2011 / Published online: 27 September 2011

Abstract Computational models of protein-protein docking that incorporate backbone flexibility can predict perturbations of the backbone and side chains during docking and produce protein interaction models with atomic accuracy. Most previous models usually predefine flexible regions by visually comparing the bound and unbound structures. In this paper, we propose a general method to automatically identify the flexible hinges for domain assembly and the flexible loops for loop refinement, in addition to predicting the corresponding movements of the identified active residues. We conduct experiments to evaluate performance of our approach on two test sets. Comparison of results on test set I between algorithms with and without prediction of flexible regions demonstrate the superior recovery of energy funnels in many target interactions using the new loop refinement model. In addition, our decoys are superior for each target. Indeed, the total number of satisfactory models is almost double that of other programs. The results on test set II docking tests produced by our domain assembly method also show encouraging results. Of the three targets examined, one exhibits energy funnel and the best models of the other two targets all meet the conditions of acceptable accuracy. Results demonstrate that the automatic prediction of flexible backbone regions can greatly improve the performance of protein-protein docking models.

Keywords Protein-protein docking · Backbone flexibility · Flexible hinge · Domain assembly

Introduction

Protein-protein interactions underlie intracellular signaling cascades, the dynamic regulation of cellular structure, and tissue organization. Complex protein-protein interaction networks have been mapped in several organisms using such methods as yeast two-hybrid [1] and mass spectrometry [2]. In Protein Data Bank(PDB), however, only a small fraction of these potential complexes has been characterized by experimental techniques such as X-ray crystallography, nuclear magnetic resonance(NMR) and electron microscopy [3]. This gap between the known and potential interacting proteins can be bridged by computational protein-protein docking models that generate one structural model or the best structural candidate models selected based on the structures of the individual proteins. The challenges faced when modeling protein-protein interactions include the accuracy of the binding site prediction, the accuracy of the flexible region prediction, the choice of effective sampling strategy, and the type of computational

Supported by National Natural Science Foundation of China under the grant number 60970055, 61170125. Qiang Lü

Q. Lü (✉) · P. Qian
School of Computer Science and Technology, Soochow
University, Jiangsu Provincial Key Lab for Information
Processing Technologies,
Shizi Street,
Suzhou, Jiangsu P.O.Box 158, 215006, People's Republic of China
e-mail: qiang@suda.edu.cn

X. Luo · H. Wu · L. Yang · X. Huang
School of Computer Science and Technology,
Soochow University,
Shizi Street,
Suzhou, Jiangsu P.O.Box 158, 215006, People's Republic of China

G. Fu
Google New York,
76 Ninth Avenue,
New York, NY 10011, USA

model. These choices constitute the greatest challenges for theoretical computation of protein- protein docking.

Based on the thermodynamic hypothesis of Anfinsen [4], native proteins always adopt lowest potential energies. Therefore, protein-protein docking can be modeled as a problem of minimizing the complex energy by sampling the degrees of freedom of the different parts of the receptor and ligand. Early methods treated the two docking partners as rigid bodys [5, 6], while later methods [7, 8] allowed flexibility only at the side chain. The performance of these methods has been extensively evaluated using blind structural predictions of more than 20 protein complexes in the critical assessment of predicted interactions(CAPRI) experiments [9–11]. For those test cases in which significant backbone conformational changes are observed upon formation of the complex, no current method is able to consistently generate models close to the correct docking conformation. For these hard CAPRI targets, the challenge of accounting for backbone conformational changes requires the incorporation of explicit backbone flexibility predictions in the protein-protein docking model. Bastard et al. [12] proposed the first docking method that incorporated an ensemble of possible loop conformations by a multicopy representation using a reduced model with up to three pseudo-atoms per amino acid. This model allowed for an extensive exploration of all possible orientations of the docking partners. The docking process starts from regularly distributed positions and orientations of the ligand around the whole receptor and each starting configuration is submitted to energy minimization during which the best fitting loop conformation is selected based on the mean-filed theory. The docking results showed that introducing loop flexibility on the isolated protein form during docking largely improves the accuracy of the model prediction of relative position of the partners in comparison with rigid body docking. Schneidman-Duhovny et al. [13] used software HingeProt [14] to automatically partition the docked partners into rigid parts and hinge regions(a strategy distinct from RosettaDock conducted by Wang et al. [15]) and proposed a method to assemble the flexible molecule into new conformations with good shape complementarity with the rigid molecule. Recent studies analyzing the perturbations of the backbone at the time of docking have employed RosettaDock, but the best positions to allow protein backbone flexibility are still unclear. In most cases, flexible regions are usually identified by visually comparing the native proteins and the bound complex [15].

The accuracy of these modeling results is a fundamental test of our understanding of the energetics of macromolecular interactions. Two classic types of backbone conformational changes [15] have been proposed, loop refinement for local variable regions (Fig. 1a) and hinge motions for domain assembly that allow domains to move relative to one another (Fig. 1b).

In a previous modeling study, backbone conformational changes at hinges and predefined flexible loops were described, but no methods were given to identify them. The flexible regions at the interface of the two docking partners are often determined empirically. Accounting only for these observed flexible regions limits the accuracy of the model. The hinge definition used for this paper, a segment of the polypeptide chain that can result in significant movement of the domains on either side, is the same as that in RosettaDock. A more accurate description is that the two hinge-linked domains manifest the maximal density of intradomain contacts and the minimal density of interdomain contacts [16, 17]. Due to such distribution of the residue-residue contacts, the forcefield of the polar and hydrophobic sidechains has made the local perturbation of the backbone movements (Fig. 1a) as well as the distinct domain rearrangements during protein-protein interactions (Fig. 1b). The target of this paper is to design an algorithm as well as the corresponding prediction mechanism to identify the flexible region that consists of the domain-linked hinges and the flexible loops automatically to serve for the protein-protein docking with backbone flexibility.

Results

In this section, we conduct experiments to evaluate the performance of our method on two test sets. Test set I consists of 25 test cases, which are taken from ref. [15]. We enforce the protocol of docking with loop minimization on the flexible loops automatically identified by our method with a backrub sampling protocol [18–20]. Our results are then compared to those from ref. [15].

Test set II consists of three difficult CAPRI targets selected by us: 1FAK, 1Y64 and 2I9B. The first two targets are selected from the benchmark set of Chen et al. (2003) [21] and target 2I9B is selected from the benchmark set of Hwang et al. (2010) [22]. These three targets possess distinct domain rearrangements during the formation of the complex, and the hinge motions are very different. Thus, only by identifying the flexible hinges accurately can we obtain good docking models. We present our docking results using the protocol of docking with domain assembly by fragment insertion in the identified flexible hinges at the low-resolution stage and using loop refinement with a backrub sampling strategy [18–20] in the identified flexible loops at the high-resolution stage.

As for metrics to evaluating performance of docking, three measurements are widely used as standard evaluation criteria; they are the fraction of native contacts (Fnat), ligand C_α rmsd (Lrmsd), and interface C_α rmsd (Irmsd).

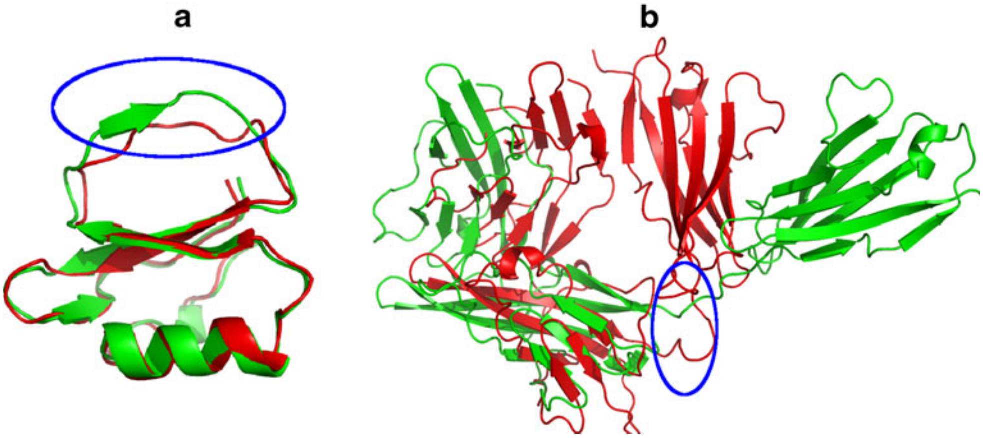


Fig. 1 Two classic types of backbone conformational changes in protein- protein docking. (a) Superimposition of the unbound ligand 1ACB and the native complex. The red segment in the blue ellipse has been marked for the flexible loop. (b) Superimposition of the unbound receptor 1IRA and the native complex. The red segment in the blue ellipse has been marked for the flexible hinge. The native complex is shown in green and the unbound partner in red

Furthermore, a CAPRI-style measurement is used to combine these three metrics to determine the prediction accuracy of docking models [10], namely, high accuracy for models with Fnat≥50% AND (Lrmsd≤1.0 Å OR Irmsd≤ 1.0 Å) [1], medium accuracy for models with Fnat≥30% AND (Lrmsd≤5.0 Å OR Irmsd≤2.0 Å), acceptable accuracy for models with Fnat≥10% AND (Lrmsd≤10.0 Å OR Irmsd≤5.0 Å) and incorrect for models with Fnat<10%. Also we count the numbers of models better than acceptable accuracy among the top three ranking models as another indicator for showing energy funnel of result decoys.

Test set I results

Table 1 compares the results attained using the protocol of docking with loop minimization on the identified flexible loops with the results from Wang et al. [15].

The first two blocks are directly from Wang et al. [15] for comparison. The third block is the quality of the best model in the decoy set given the known native structure. This third block is further expanded in a separate Table 2 in order to fully investigate the performance of the decoys generated by our method for each target.

The last two blocks in Table 1 list our results selected from decoys by interface energy and combined energy respectively (see Methods for detail).

As revealed in the third block of Table 1, the best solution is superior for each comparison target, and the total number of acceptable models is nearly double that of the competitor. Table 2 presents the detailed results of evaluation criteria from the decoys. In essence, these results demonstrate that the quality of our predictions is acceptable and superior to those of the competitor whether energy funnels are achieved or not. For 60% of the targets (Table 2, column Mc), more than 1% of the decoys meet medium accuracy, with 9.4% of decoys reaching medium accuracy for the best target. For 56% of the targets (Table 2, column BLc), more than 1% of the decoys are superior to the best value derived by the competitor. In one case, 52% of decoys are superior to the best value derived by the competitor.

It remains a challenge to choose the best conformation without knowing the native structure of the protein. For this reason, we also analyze the results based on the other two selection criteria we developed (see Methods for the detail descriptions about the two criteria). The results are listed in the 4th and 5th block of Table 1, respectively.

Test set I results by interface energy

As revealed in the 4th block of Table 1, we obtain about the same total number of energy funnels by interface energy as in ref. [15]. It is worth noting, however, that for targets 1CSE, 1FSS, 1MAH, 1MLC, and 2KAI, our method recovers energy funnels lost by ref. [15]. For each of these five targets, the near-native models selected are closer to the best solution than for the other targets, and the energy funnel surrounding the native structure is more evident with more models pushed into the funnel tip (Table 1). Similar conclusions can also be drawn for target 2PTC and 2SIC, possibly because the flexible loops identified reflect the true situation during protein-protein docking or because the sampling strategy employed can handle the degrees of freedom introduced by these flexible loops.

The selected results could be further improved if using better methods to select the best models from the decoys, but this is still a challenging problem. Targets like 1GLA, 1TGS, and 2PCC that lost energy funnels in this paper still have many good models that meet medium accuracy, demonstrating the accuracy of the identified flexible loops in protein-protein docking. Other results indicated that selection criteria based on the interface energy are too simple, however. For targets 1DQJ, 1WQ1, and 1ACB, there are no models that reach medium accuracy, mainly due to the inaccurate identifi-

[1] The parentheses operator is prior to AND operator

Table 1 Overall results comparison on test set I

	Unbound[a], RLX[b], BBmin[c], interface energy[d]				Unbound, RLX, BBmin, binding energy[d]				Unbound, LoopMin[e], the best model				Unbound, LoopMin, interface energy[f]				Unbound, LoopMin, combined energy[f]			
PDB	N3[j]	BC[g]	BL[h] (Å)	BI[i](Å)	N3	BC	BL	BI	decoy Size	BC	BL	BI	N3	BC	BL	BI	N3	BC	BL	BI
1CSE	0	0.000	14.33	7.24	0	0.116	5.40	3.07	1000	0.721	2.09	1.09	3	0.706	2.92	1.24	3	0.691	2.81	1.29
1FSS	0	0.044	12.02	4.49	0	0.178	7.61	2.87	1000	0.741	1.62	1.80	1	0.741	1.88	1.82	3	0.741	1.62	1.80
1MAH	0	0.308	8.61	4.14	0	0.231	4.68	2.16	1000	0.524	3.20	2.27	1	0.492	3.96	2.32	2	0.492	3.76	2.38
1MLC	0	0.152	7.81	3.20	0	0.061	29.25	8.36	1000	0.576	3.01	1.96	2	0.576	3.01	1.96	2	0.576	3.01	1.96
2KAI	0	0.000	22.37	10.77	0	0.000	19.93	8.27	1000	0.721	2.96	1.34	3	0.689	3.98	1.34	3	0.656	3.98	1.34
2PTC	3	0.513	3.96	0.98	3	0.538	3.75	0.98	1000	0.828	0.80	0.96	3	0.810	3.33	0.96	3	0.810	2.73	1.08
2SIC	2	0.409	6.41	1.34	1	0.386	4.61	1.29	1000	0.828	2.30	1.29	3	0.828	4.67	1.32	3	0.828	3.42	1.32
1BRC	2	0.667	3.75	1.26	3	0.667	3.83	1.20	1100	0.804	2.36	1.61	2	0.804	3.90	1.92	1	0.804	3.90	1.92
1AVW	3	0.660	4.87	0.99	2	0.447	4.80	1.29	1000	0.841	2.46	1.36	3	0.818	5.52	1.54	2	0.727	6.30	1.63
1BRS	3	0.750	3.00	1.30	3	0.750	3.00	1.30	1200	0.681	2.47	1.70	1	0.660	3.49	2.14	1	0.596	3.99	2.18
1CHO	3	0.625	1.88	0.74	3	0.375	2.25	1.03	2900	0.889	1.99	1.30	3	0.889	4.35	1.46	2	0.867	3.74	1.39
1UGH	3	0.459	3.50	1.64	0	0.189	14.76	7.19	1100	0.582	2.93	2.02	2	0.582	4.06	2.55	1	0.388	4.46	2.27
2SNI	3	0.738	4.26	1.26	2	0.738	5.05	1.43	1100	0.758	2.80	1.70	2	0.710	5.33	1.93	2	0.677	3.66	1.84
1GLA	2	0.619	2.94	1.04	0	0.190	16.23	4.97	1500	0.800	1.80	1.22	0	0.520	8.43	2.23	0	0.240	5.91	3.58
1TGS	3	0.563	2.94	1.47	2	0.417	3.18	1.64	3300	0.739	1.66	2.41	0	0.217	11.05	5.05	2	0.522	4.38	2.93
2PCC	0	0.222	5.44	3.28	0	0.222	5.47	3.18	2000	0.750	1.70	1.36	0	0.250	6.28	3.68	1	0.400	3.72	2.71
1WEJ	0	0.094	8.55	4.54	0	0.406	5.46	2.44	1100	0.741	2.39	1.55	0	0.407	5.46	3.29	0	0.407	12.98	4.03
1AVZ	0	0.048	14.95	6.29	0	0.238	12.85	6.25	1000	0.404	3.71	1.84	0	0.259	11.46	5.41	0	0.148	10.23	9.01
1MDA	0	0.000	9.72	4.21	0	0.000	17.16	7.97	1000	1.000	4.43	1.82	0	0.667	8.41	2.90	0	0.000	11.39	4.49
1DFJ	3	0.488	3.66	1.53	2	0.395	3.95	1.24	1100	0.574	2.27	1.63	0	0.298	6.30	3.46	1	0.574	3.35	1.63
1AHW	0	0.289	7.15	2.24	1	0.422	6.34	1.65	8000	0.519	2.80	1.87	0	0.423	9.90	2.89	0	0.423	9.90	2.89
1BVK	0	0.000	11.59	6.06	0	0.000	9.85	5.91	1000	0.636	4.51	2.04	0	0.636	6.43	2.34	0	0.212	12.46	3.68
1DQJ	0	0.000	21.49	9.54	0	0.000	18.63	11.78	1000	0.483	5.05	2.60	0	0.350	13.94	4.01	0	0.350	12.07	4.01
1WQ1	0	0.222	6.18	3.21	2	0.333	4.10	1.85	1000	0.320	3.50	3.57	0	0.200	6.24	5.25	0	0.200	6.24	5.25
1ACB	2	0.487	7.79	1.64	3	0.538	2.91	1.36	1700	0.246	3.71	3.72	0	0.193	10.62	4.85	0	0.140	6.54	4.19
Total[k]	12	13	10	12	12	13	11	12	-	24	24	18	13	19	11	10	16	19	15	11

[a] Unbound: for each docking, backbone conformations of the starting structures are taken from the independently solved structures. [b] RLX: the starting structures are prepared by the prerelaxing procedure. [c] BBmin: the models are generated using the flexible-backbone protocol. [d] interface energy and [d] binding energy: selection criterions in ref [15]. [e] LoopMin: the models were generated using the protocol of docking with loop minimization with the flexible loops identified automatically in this paper. [f] interface energy and [f] combined energy: selection criterions proposed in this paper. [g] BC is the best fraction of Fnat of the top three ranking models. [h] BL is the best Lrmsd(Å) of the top three models. [i] BI is the best Irmsd(Å) of the top three models. [j] N3 is the number of models among the top three ranking models with at least medium accuracy (see the 3rd paragraph of Results section), and N3>0 indicates that an energy funnel exists. [k] Total is the number of cases with N3>0, BC≥0.3, BL≤5.0 and BI≤2.0 respectively

Table 2 Detail performance of decoys on test set I

PDB	Rc[a]	Lc[b]	Mc(%)[c]	BLc(%)[d]	BCc[e]	BIc[f]
1CSE	1	1	52(5.2)	59(5.9)	167	83
1FSS	1	1	22(2.2)	58(5.8)	94	17
1MAH	1	1	8(0.8)	11(1.1)	9	0
1MLC	2	1	3(0.3)	13(1.3)	15	11
2KAI	2	1	47(4.7)	520(52)	362	437
2PTC	1	1	94(9.4)	45(4.5)	74	2
2SIC	1	1	51(5.1)	24(2.4)	66	0
1BRC	1	1	34(3.09)	6(0.55)	10	0
1AVW	1	1	45(4.5)	7(0.7)	23	0
1BRS	1	1	12(1)	2(0.17)	0	0
1CHO	1	1	126(4.34)	0(0)	55	0
1UGH	1	1	13(1.12)	2(0.18)	5	0
2SNI	1	1	32(2.9)	20(1.82)	3	0
1GLA	1	1	95(6.33)	14(0.93)	30	0
1TGS	1	1	99(3)	9(0.27)	52	0
2PCC	1	1	32(1.6)	48(2.4)	162	74
1WEJ	2	1	11(1)	19(1.73)	59	4
1AVZ	1	1	9(0.9)	453(45.3)	99	579
1MDA	1	1	6(0.6)	228(22.8)	303	380
1DFJ	1	1	5(0.46)	2(0.18)	2	0
1AHW	2	1	17(0.21)	62(0.78)	15	0
1BVK	2	1	2(0.2)	43(4.3)	436	229
1DQJ	1	1	0(0)	295(29.5)	421	595
1WQ1	1	1	0(0)	2(0.2)	0	0
1ACB	1	1	0(0)	0(0)	0	0

[a] Rc: the polypeptide chain number of the docking receptor. [b] Lc: the polypep- tide chain number of the docking ligand. [c] Mc: the number of the decoys that meets medium accuracy(shown energy funnel). [d] BLc, [e] BCc, and [f] BIc: the model numbers from the decoys which are superior to the best value selected by the interface energy and binding energy from the corresponding targets in ref. [15], respectively. For ease of comparison, [c] Mc and [d] BLc also give the percentage calculated by Mc and BLc over the decoy size (listed in the third block, Table 1) of the corresponding target in the brackets, respectively

cation of the flexible loops and inefficiency of the employed sampling strategy (see Discussion).

Test set I results by combined energy

Compared to either the results from ref. [15] or the results selected by interface energy using our model, there is a significant improvement using combined energy. In total, 16 targets show energy funnels, 12% higher than the ratio attained by interface energy(52%). Those targets that show energy funnels selected by interface energy also show energy funnels selected by combined energy. For targets 1CSE, 1FSS, 1MAH, 1MLC, and 2KAI, the values of the best Lrmsd and the best Irmsd of the top three models (the values of BL and BI in Table 1) are slightly better than those selected by interface energy. The three new targets 1DFJ, 1TGS, and 2PCC that show energy funnels demonstrate the superior performance of our generated decoys. Moreover, target 2PCC recovers the energy funnel lost in ref. [15], whether selected by the interface energy or by the binding energy.

Compared to these selected models, however, the values of the best Lrmsd of the top three models (the values of BL in Table 1) are very poor for targets 1BVK, 1DQJ, 1WEJ, 1AVZ, and 1MDA in ref. [15]. Using our program, the values of the best Fnat and the best Irmsd of the top three models (the values of BC and BI in Table 1) of the first three targets all meet acceptable accuracy and they are also comparably better than the results from ref. [15]. These five targets have the common feature that the ratio of Mc (Table 2) is comparably lower than the Mc ratio of the targets that show energy funnels (targets 1AVZ and 1WEJ for examples). Better models would be expected if a better selection scheme is used. These five targets highlight the fact that good model selection depends mainly on the quality of the decoy set generated.

Test set II results

Table 3 presents the detailed results for targets 1FAK, 1Y64, and 2I9B using the protocol of docking with domain assembly.

Results are divided into four blocks. The first block describes the flexible hinge and the segment identified by our approach in the corresponding chain of the docking partner. The second block gives the quality of the best model in the decoy set, and the last two blocks are our results selected with which using interface energy and combined energy respectively.

Results of target 1FAK

For target 1FAK, we capture the correct flexible area that causes the rearrangements of the two hinge-linked domains during protein-protein interactions. The best model in the decoy set has an Irmsd of 3.53 Å and an Lrmsd of 2.89 Å with respect to the native complex. Both the models selected by interface energy and combined energy show energy funnels and all the other performance criteria are acceptable (the 3rd and 4th blocks in Table 3).

Results of target 1Y64

For target 1Y64, we only capture the subsegment of the right flexible area that causes the domain motions during protein-protein interactions. Though the best model in the decoy set has an Irmsd of 4.15 Å with respect to the native

Table 3 Overall results on test set II

			Unbound, HingeMotion[a], LoopMin, the best model				Unbound, HingeMotion, LoopMin, interface energy				Unbound, HingeMotion, LoopMin, combined energy			
PDB	hinge region	hinge residues	decoy Size	BC	BL	BI	N3	BC	BL	BI	N3	BC	BL	BI
1FAK	[447..477:L]	CLPA…CEQYC	8×10^3	0.448	2.89	3.53	2	0.397	3.63	4.60	3	0.448	3.81	3.69
1Y64	[50..55:B]	FAAREI	8×10^3	0.466	12.68	4.15	0	0.397	20.07	5.31	0	0.345	14.08	6.38
2I9B	[23..32:A]	CPKKFGGQHC	8×10^3	0.441	6.69	6.08	0	0.271	18.15	7.54	0	0.034	25.69	8.64
Total	-	-	-	3	1	0	1	2	1	0	1	2	1	0
1Y64[1]	[50..61:B]	FAAREIKSLASK	8×10^3	0.655	4.45	1.84	0	0.466	25.40	4.44	0	0.379	14.64	3.32
2I9B[1]	[23..32:A]	CPKKFGGQHC	8×10^3	0.576	7.49	2.45	0	0.237	11.36	4.58	0	0.237	11.36	4.58
2I9B[2]	[23..32:A]	CPKKFGGQHC	8×10^3	0.746	6.16	2.20	0	0.542	8.30	2.57	0	0.610	8.85	2.20

[a] HingeMotion: our protocol of docking with domain assembly by fragment insertion in the identified flexible hinge at the low-resolution stage

complex (Fig. 2b), the quality of the results selected by either interface energy or combined energy (row 1Y64, Table 3) is below the medium accuracy, and therefore is unacceptable. Studying why this happens reveals that it is caused by the inaccurate identification of the complete flexible regions. We re-identify the flexible regions by manually specifying the right flexible hinge, and repeat 1Y64 experiment. We name this new experiment 1Y64[1] in Table 3. Although the models selected in 1Y64[1] do not show energy funnel either, the best model in the decoy set has an Irmsd of 1.84 Å as well as an Lrmsd of 4.45 Å (Fig. 2a, row 1Y64[1], Table 3). The quality of the generated decoys indicates the great importance of accurately identifying the flexible hinge in protein-protein docking with domain assembly.

Results of target 2I9B

The results for this test case demonstrate the superior performance of the scheme proposed to identify the flexible regions in protein-protein docking (Fig. 3).

The results also indicate that the flexible loops greatly impact the formation of the complex during domain assembly. Similar to the simulation of 1FAK, our model capture the correct flexible area that causes the domain motions during protein-protein interactions. The motions in the flexible loops at the high-resolution stage, however, do not simulate the true situation of the backbone flexibilities, especially for those targets with many long flexible loops (see Discussion).

We have provided highly accurate identification of the flexible loops. To further test the accuracy of our model, we have implemented two other experiments to test the performance of our identified flexible hinges.

First, we employ the bound receptor where flexible loops are densely distributed and the unbound ligand for experiment 2I9B[1]. The quality of the decoy set and the models selected (row 2I9B[1], Table 3) are better than those from experiment 2I9B. The number of good models is rather small (data not shown) because the three-dimensional structure of the backbone of the identified flexible loop [1, 7:CDCLNGG] in the chain of the unbound ligand biases the sampling of the conformational degrees of freedom of

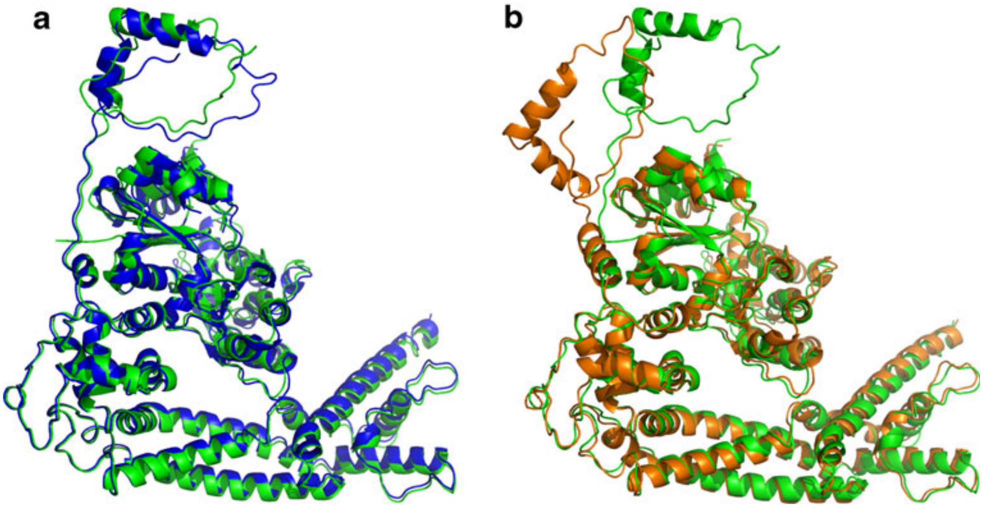


Fig. 2 Docking with domain assembly for target 1Y64. (a) Superimposition of the native complex (green) and the best model (blue) generated with the true flexible hinge based on residues at the protein-protein interface. (b) Superimposition of the native complex (green) and the best model (brown) generated with the identified flexible hinge based on residues at the protein-protein interface

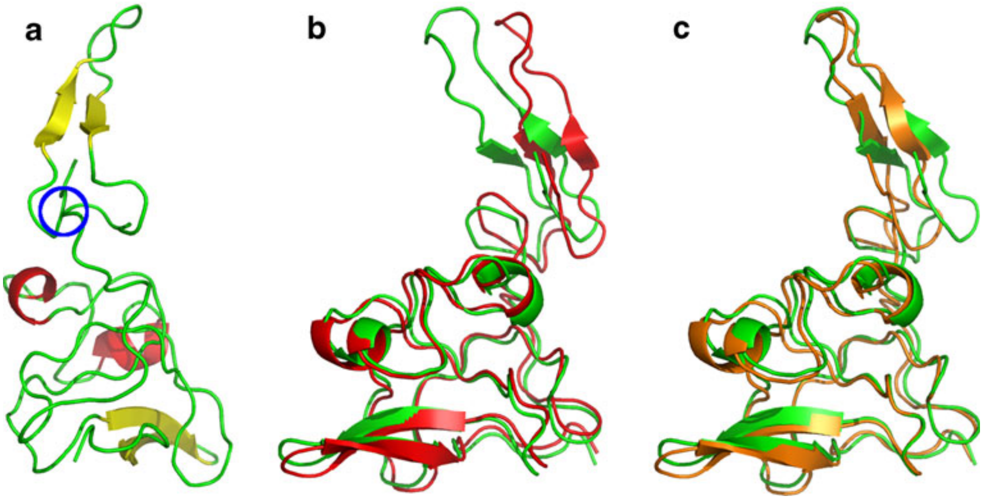


Fig. 3 Docking with domain assembly for target 2I9B. (a) The conformation of chain A of the best model in experiment 2I9B[1]. (b) Superimposition of the native complex (green) and the unbound ligand 2I9B(red). The receptor is not shown. (c) Superimposition of the native complex (green) and the conformation of chain A of the best model (brown) in experiment 2I9B[2]. The receptor is not shown. The docking model with domain assembly has impacted docking

the flexible hinge for the atomic clashes formed within the docking monomer (Fig. 3a, the blue ellipse has been marked for the tendency of atomic clash). For experiment 2I9B[2], we implement fragment insertion in the identified flexible hinge and flexible loop [1, 7] at the low-resolution stage and other identified flexible loops in the unbound ligand for loop refinement at the high-resolution stage. Although we do not attain energy funnel, the quality of the generated decoys are relatively good (Fig. 3c, Table 3), and the models selected are distinctly better than those from the two above experiments (experiment 2I9B and 2I9B[1]).

Discussion

Protein molecules are dynamic and protein-protein interactions are often accompanied by conformational changes in both the backbone and the side chains of the two docking partners. The flexible regions automatically identified by this new program can accurately predict the local regions on which the right perturbations are applied. Restricting movement to defined regions greatly reduces complexity of the search space. The comparison results on test set I demonstrate the enhanced performance of this method for identifying the flexible loops, while the docking results on test set II demonstrate that the quality of the docking models depended on the correct identification of the flexible hinges and flexible loops as well as the correct motions applied to corresponding flexible regions.

The backrub sampling protocol at the high-resolution stage is not as successful for those targets that are sensitive to the contacts at the interface area between the two docking partners because the residues inside the flexible loop rotate around the axis defined by the two boundary residues, even if the correct flexible loops are identified. For target 1ACB, better models would not be expected even if more decoys are generated because the best value of the fraction of the native contacts in the decoy set is smaller than the predefined threshold 30% of medium accuracy. This is due to the two flexible loops at the interface of the two docking monomers that would form stable parallel β−sheets in the complex (Fig. 1a). The backrub sampling could not effectively locate the right positions of the two flexible loops, causing poor repacking performance of the rotamers in the interface area. Similar results are found for targets 1DQJ and 1WQ1, but the mechanism is different for the very different situation of the surface flexible loops. For target 1WQ1, α−helies and β−strands are densely packed at the interface area with short flexible loops that are far from each other. Small perturbations in such flexible regions are sensitive to the Irmsd (Table 1, the best model) and the fraction of native contacts. For target 1DQJ, however, although small perturbations in the identified flexible loops are sensitive to the value of the Lrmsd (Table 1, the best model), better models would be expected if more models are generated.

In our current implementation, the identified flexible loops are tackled in random order. The best perturbations of the current flexible loops are only related to the current environment with no influence on the sampling of the other flexible loops. A new routine could be employed to further improve the performance of the sampling strategy by changing the current platform to alow parallel execution. Different sampling strategies could be combined with different object functions to search the best near-native conformation.

The flexible regions that are focused on the loops of the docking partners have inherent limitations and this has been well proved in this paper, as exemplified by targets 2I9b and 1Y64. For target 2I9B, the flexible loops include not only those defined by the program, but may also include the two parallel or anti-parallel β-sheets. For target 1Y64, the known true flexible hinge segment [50, 61] identified visually is composed of a segment with loop secondary structure [50, 55] and a segment with a stable α-helix secondary structure[56, 61]. The proposed scheme focuses on the identification of loops, so only part of the hinge [50, 55] was identified. The above experiments demonstrate that the flexible regions are of great importance during protein-protein docking. Precise detection of these flexible regions is necessary to accurately model protein-protein interactions during docking. The scheme proposed in this paper to identify the flexible regions may be too naive and simple because the flexible regions are a reflection of the combined effects of protein-protein interactions. Indeed, for those test cases where the flexible regions are not detected precisely, we only analyze the docking reaction roughly and completely neglect the detailed atomic contacts and the microenvironments of the group residue-residue interactions. The limitations mentioned above could be eliminated by machine learning algorithms that use the docking partners and their native protein complexes based on the published docking benchmarks to identify flexible regions by detecting the binding modes of the flexible residues and their environment. This technology has been successfully applied to many hard problems by technically eliminating noise interference and data sparseness in very high dimensional space.

Methods

Our software is implemented based on Rosetta3.1 source code (downloaded from the web site http://www.rosettacommons.org/software/ upon license agreement).

Key residues in flexible regions

Some specific residues play critical role in flexible regions. In our approach, four such residues are taken as the necessary component of flexible regions. ASN, ASP, GLU, and GLN are hydrophilic residues with similar polar side chains. Residues ASN and GLN have the same terminal side chain structure (H_2N-C_γ=O and H_2N- C_δ=O respectively), as do residues ASP and GLU (O^--C_γ=O and O^--C_δ=O respectively). Compared to other hydrophilic residues, the C=O bond of the side chain of these polar residues plays a vital role in backbone conformational changes during protein-protein docking. These polar hydrophilic residues are stabilized inside the core of the complex by hydrogen bonds or salt bridges that deflect the dihedral angles of these four active residues. When candidate flexible regions are identified (see next subsections), we will count the numbers of such key residues within the candidate regions. Only when the counter is greater than a threshold, can those regions be further considered as flexible loops or hinges.

Identification of flexible loops

First, we perform DSSP [23] to identify the secondary structures that partition the polypeptide chains into segments. The loop segments identified by DSSP are the first initial candidates for the selection procedure. We then apply biochemical constraints to reduce the number of initial candidates. If the number of residues between the two adjacent loops is smaller than four, the program merges the two loops to a single loop with the number of residues not larger than 30 [24]. If there are several such segments, it is possible that these short consecutive segments form several stable β-sheets and we take this into account to merge the corresponding segments based on the disulfide bond with one residue in the led segment and the other in the subsequent segment. A single β-sheet could not form a domain. To the contrary, a single α-helix can be regarded as a domain when the linked hinge meets some specific conditions. As for the derived candidate loops, the flexible loop is determined by the number of active residues proposed in Key residues in flexible regions, and the threshold is predefined to be no smaller than two. That is, a loop possessing no such active residue will have no flexibility during protein-protein docking and we ignore the flexibility of a loop that has only one such active residue.

Identification of the flexible hinges

Residue contact network

Our approach takes the positions of the C_α atoms as the nodes, and the connection between the two C_α atoms as the edge [25]. If the distance between two C_α atoms is not more than 5 Å, then the weight corresponding to the edge is 1. Otherwise it is 0. Thus, the adjacent matrix [A] is given by:

$$A_{ij} = \begin{cases} 1, & \text{distance}(i,j) \leq 5\text{Å}, \\ 0, & \text{otherwise}. \end{cases} \tag{1}$$

K_i is the degree of node-i of its neighbors,

$$K_i = \sum_{j=1, j\neq i}^{N} A_{ij}, \tag{2}$$

where N is the total number of the nodes. Let E_i be the edge set connected by node-*i*. Then C_i defines the clustering degree of node-*i*

$$C_i = \frac{\|E_i\|}{2} \cdot K_i, \tag{3}$$

where $||E_i||$ denotes the element number of the set. Obviously, a high value of K_i means that node-i has a large number of edges and is a degree center in the network [26]. The number of the element in set E_i describes the electron density around node-i, while C_i is a measure of the local properties to characterize the distribution of the atomic interactions with the larger value of the hydrophobic core.

Residue contact environment

Whether the chosen segment is the true hinge and the number of such segments are both uncertain. We must extract the features that can be consistently applied to cover the most likely situations [16, 27]. So the linchpin of the extracted features is to explore the similarity of the interdomain contacts to enlarge the ratio of formula(4). Our approach modified conventional modeling by applying the summation of the average clustering degree of the pair wise residues as the intradomain interactions while using the summation of the degree of the centroid of the pair wise residues as the interdomain interactions.

$$\rho(i,j) = \frac{\sum_{k=1}^{i-1}\sum_{s=k+1}^{i} A_{ks}\cdot\overline{C_{ks}} \cdot \sum_{m=j}^{N-1}\sum_{p=m+1}^{N} A_{mp}\cdot\overline{C_{mp}}}{\left(\sum_{q=1}^{i}\sum_{t=j}^{N} A_{qt}\cdot K_{qt}\right)^2}, \tag{4}$$

where $\overline{C_{ks}} = (C_k + C_s)/2$ relates to the average clustering degree of node-k and node-s while $\overline{C_{mp}}$ corresponding to the other, K_{qt} denotes the degree of the centroid of node-q and node-t. Here the centroid is the nearest residue to the center of node-q and node-t. These local properties, rather than the simple count of the atomic interactions, and the accumulation of such local properties could be interpreted as the global properties for the non-polar hydrophobic residues buried inside to form a comparatively steady structure.

Feature extraction

The candidate loops derived from the DSSP were parsed by applying biochemistry constraints. First, a quaternion [i, j, length, ρ] for each loop is constructed where *i* indicates the start position of the residue and j the end position. The length stands for the number of residues of the segment, and ρ is the eigenvalue of the corresponding loop calculated by formula(4). In cases where there are multiple polypeptide chains, the procedure handles them one by one. First, for each docking partner we sort the loop set based on the eigenvalue in descending order and exclude those segments with no interdomain interactions that would be potential flexible hinges.

$$\overline{len} = \frac{\sum_{k=1}^{M} length_k}{M}, \tag{5}$$

where $\overline{len}$ denotes the average length of the candidate loops and M is the total number of the items in the candidate loop set.

$$def = \frac{length_k^2 \cdot \rho_k^2}{\overline{len}^2 \cdot \rho k - 1 \cdot \rho k + 1} \geq \delta, \tag{6}$$

Formula(6) is used to find the break loop in the sorted loop set based on ρ in descending order such that the ratio *def* is no less than the predefined threshold δ=5. If there is only one such candidate, then it is regarded as a candidate hinge without ambiguity; otherwise, further iterative process is needed to identify the flexible hinges.

The iterative identification procedure can be described as follows since protein domains are the steady compact subunits in protein structures. In the first stage, as for the candidate hinges, the two candidates are merged if they are space adjacently with a number of linked residues smaller than five. In the second stage, we update the candidate loop set with the merged candidates and use the same method and parameters described above to calculate its properties with the same physical constraints described above. Finally, the flexible hinges can be determined. First, there must be at least one active residue(as proposed in Key residues in flexible regions) in the flexible hinge area. Similar to the strategy for candidate loops identification, the last step is to parse the disulfide bonds with one residue inside the flexible hinge area and the other out of this segment; if detected, the flexible hinge area is extended to the parsed residue to form the ultimate flexible hinge area.

Model selection

Model selection by interface energy is the same as that in Wang et al. [15] except that the 25% lowest energy models are ranked based on interface energy. For model selection by combined energy, 50% of the models are first selected based on the score of energy item interchain contact. Secondly, these selected models are sorted based on the interface energy of the lowest 15%. Thirdly, the derived models are ranked based on the score of the energy function that is used in the low-resolution stage to select the top ten models. Finally, the best models with both lower total energy and interface energy in the latter seven are selected to exchange those in the set top three.

Acknowledgements Great thanks to the developers of Rosetta3.1.

References

1. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403:623–627

2. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147
3. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, Westbrook J (200) The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol 7:957–959
4. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230
5. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem A, Aflalo C, Vakser I (1992) Molecular surface recognition: determination of geometric fit between protein and their ligands by correlation techniques. Proc Natl Acad Sci USA 89:2195–2199
6. Chen R Weng ZP (2002) Docking unbound proteins using shape complementarity, desolvation, and electrosatics. Proteins 47:281–294
7. Gray JJ, Moughon S, Wang C, Kuhlman OSB, Rohl CA, Baker D (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 331:281–299
8. Wang C, Schueler-Furman O, Baker D (2005) Improved side-chain modeling for proteinprotein docking. Protein Sci 14:1328–1339
9. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S (2003) CAPRI: a critical assessment of predicted interactions. Proteins 52:2–9
10. Mendez R, Leplae R, De Maria L, Wodak SJ (2003) Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins 52:51–67
11. Mendez R, Leplae R, Lensink MF, Wodak SJ (2005) Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. Proteins 60:150–169
12. Bastard K, Prevost C, Zacharias M (2006) Accounting for loop flexibility during proteinprotein docking. Proteins 62:956–969
13. Schneidman-Duhovny D, Nussinov R, Wolfson HJ (2007) Automatic prediction of protein interactions with large scale motion. Proteins 69:764–773
14. Emekli U, Schneidman-Duhovny D, Wolfson HJ, Nussinov R, Haliloglu T (2008) HingeProt: automated prediction of hinges in protein structures. Proteins 70:1219-1227
15. Wang C, Bradley P, Baker D (2007) Protein-protein docking with backbone flexibility. J Mol Biol 373:503–519
16. Rossmann MG, Liljas A (1974) Recognition of structural domains in globular proteins. J Mol Biol 85:177–181
17. Siddiqui AS, Barton GJ (1995) Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definition. Protein Sci 4:872-884
18. Davis IW, Arendall WB, Richardson DC, Richardson JS (2006) The backrub motion: how protein backbone shrugs when a sidechain dances. Structure 14:265–274
19. Smith AC, Kortemme T (2008) Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain rediction. J Mol Biol 380:742–756
20. Friedland GD, Linares AJ, Smith CA Kortemme T (2008) A simple model of backbone flexibility improves modeling of side-chain conformational variability. J Mol Biol 380:757–774 (2008)
21. Chen R, Mintseris J, Janin J, Weng Z (2003) A protein-protein docking benchmark. Proteins 52:88–91 (2003)
22. Hwang H, Vreven T, Janin J, Weng Z (2010) Protein-protein docking benchmark version 4.0. Proteins 78:3111–3114 (2010)
23. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637
24. Jiang HY, Blouin C (2006) Ab initio construction of all-atom loop conformations. J Mol Model 12:221–228 (2006)
25. Greene LH Higman VA (2003) Uncovering network systems within protein structures J Mol Biol 334:781–791
26. Amitai G, Shemesh A, Sitbon E, Shklar M, Venger DNI, Pietrokovski S (2004) Network analysis of protein structures identifies functional residues. J Mol Biol 344:1135–1146 (2004)
27. Maiorov VN, Abagyan RA (1997) A new method for modeling large-scale rearrangements of protein domains. Proteins 27:410–424

J Mol Model (2012) 18:2209–2218
DOI 10.1007/s00894-011-1238-6

ORIGINAL PAPER

Molecular dynamics modeling of the sub-THz vibrational absorption of thioredoxin from *E. coli*

Naser Alijabbari · Yikan Chen · Igor Sizov · Tatiana Globus · Boris Gelmont

Received: 15 July 2011 / Accepted: 8 September 2011 / Published online: 27 September 2011

Abstract Sub-terahertz (THz) vibrational modes of the protein thioredoxin in a water environment were simulated using molecular dynamics (MD) in order to find the conditions needed for simulation convergence, improve the correlation between experimental and simulated absorption frequencies, and ultimately enhance the predictive capabilities of computational modeling. Thioredoxin from *E. coli* was used as a model molecule for protocol development and to optimize the simulation parameters. The empirically parameterized software packages Amber 8 and 10 were used in this work. Using atomic trajectories from the constant energy and volume MD simulations, thioredoxin's sub-THz vibrational spectra and absorption coefficients were calculated in a quasi-harmonic approximation. An optimal production run length ~100 ps was found, in agreement with experimental data on thioredoxin relaxation dynamics. At the same time, a new procedure was developed for averaging correlation matrices of atomic coordinates in MD simulations. In particular, the open source package ptraj was edited to improve a matrix-analyzing function. Averaging only six matrices gave much more consistent results, with absorption peak intensities exceeding those from the individual spectra and a rather good correlation between simulated vibrational frequencies and experimental data.

Keywords THz absorption · Vibrational modes · Thioredoxin · Convergence · Molecular dynamics

N. Alijabbari · Y. Chen · I. Sizov · T. Globus (✉) · B. Gelmont
Department of Electrical & Computer Engineering,
University of Virginia,
351 McCormick Rd., PO Box 400743, Charlottesville,
VA 22904, USA
e-mail: tg9a@virginia.edu

Introduction

Terahertz (THz) vibrational spectroscopy is an emerging technique for characterizing biomolecules and species. Radiation in the THz range interacts with low-frequency internal molecular motions involving weak hydrogen bonds and nonbonded interactions between different functional groups by exciting these motions [1, 2]. The resonant frequencies of such motions usually occur below 300 cm^{-1} (or 9 THz). THz vibrational spectroscopy highlights these motions as resonance peaks in transmission (absorption) spectra at specific frequencies. The spectroscopic patterns of different biological molecules or bacterial cells are unique and can be used as their fingerprints. The ability to directly detect the low-frequency vibrations of the weakest bonds between groups of atoms is quite different from visible or IR spectroscopic characterization, which mainly probes the stronger bonds between neighboring atoms. At the same time, this technique does not damage living species [3].

The spectral range below 1 THz is the most attractive for practical applications mainly because of the relatively low-level disturbance from liquid water (2.5 orders of magnitude less absorption compared to the far IR) and water vapor absorption. Thus, sensors do not require evacuation or purging with nitrogen. The spectra in the sub-THz range are rich with resonance features with a spectral line width of ~0.5 cm^{-1} [4, 5], which is determined by the energy relaxation time for the low-frequency motions ~ 7×10^{-11} s [5]. This estimate, which is based on a comparison between experimental and modeling results, predicts that good spectral resolution is important for observing vibrational modes.

Due to the relatively low absorption of biomaterials in the sub-THz range, genetic material (DNA and RNA), proteins, and other molecules can all contribute to the THz signature of bacterial cells/spores [6, 7]. At the same time, the utilization of the sub-THz spectroscopic technique to identify biological macromolecules and cells requires a robust theoretical model to better understand the physical motions associated with the sub-THz absorption features. This understanding can be achieved by improving the predictive capabilities of computational modeling.

In this study, the protein thioredoxin from *Escherichia coli* is used as a model molecule to simulate sub-THz vibrational absorption using the software packages Amber 8 [8] and 10 [9]. Proteins contribute up to ~50% of the dry weight of *E. coli*, and thioredoxin comprises a quarter of that 50%, so thioredoxin is thought to be a significant contributor to *E. coli* THz absorption spectra. The molecular dynamics of proteins are widely studied, but we do not know of any works on the simulation of the low-frequency vibrational modes and the absorption spectra of thioredoxin resulting from these motions. We are interested most of all in the problem of simulation convergence. In our earlier study, we compared simulated and experimental sub-THz vibrational spectral features from thioredoxin [10]. The results of 100 ps constant temperature and pressure simulations were used to obtain simulated spectra of thioredoxin. The experimental procedure for the THz characterization of biological materials in water was described in [11]. Spectra were measured with a Fourier transform spectrometer and a cooled Si bolometer operating at 1.7 K. Some absorption features predicted by our earlier MD simulations [10] agreed reasonably well with experimental data when the default simulation parameters were used. However, the calculated spectra were highly sensitive to the parameter values, and reproducibility was poor. At that time, there was no systematic analysis of the problem.

In our current work, we use a crystal structure of thioredoxin that has been refined by the stereochemically restrained least-squares procedure at 1.68Å resolution [12]. The molecular structure (pdb ID: 2TRX) is optimized using molecular dynamic (MD) simulations at room temperature and atmospheric pressure. The covalent bond and angle energy, the proper and improper torsions, and the nonbonded interactions, including the electrostatic and van der Waals interactions, are taken into account using the AMBER 03 force field [13]. The effect of the liquid content inside the bacterial cell is emulated explicitly through the use of TIP3P water molecules [14]. The pre-equilibrated box of water provided in Amber is used to build an initial set of atomic coordinates for the system of water molecules and protein. MD simulations with periodic boundary conditions are performed to equilibrate the solvent and solute. Using atomic trajectories from room-temperature MD simulations, the oscillator strengths are calculated for each normal mode in a quasi-harmonic approximation. Finally, absorption coefficient spectra are calculated [5] for three different orientations of the molecule with respect to the electric field polarization and averaged. In these calculations, an oscillator dissipation factor (or damping coefficient) γ =0.5 cm^{-1} is used, as estimated from the spectral widths of resolved features in our experiments performed using a Bruker FTS-66 spectrometer with a spectral resolution of ~0.3 cm^{-1}. Due to possible contributions from several different modes occurring at frequencies that are close together, this spectral line width gives us an upper limit of γ, which is reciprocal to the vibration relaxation time or the timescale of internal molecular motions, τ. Thus, the lowest limit for the timescale of vibrational motions corresponding to the observed spectral features in the sub-THz range is estimated as τ=1/(γc)~ 70 ps, where c is the velocity of light. This estimated damping coefficient of 0.5 cm^{-1} is of the same order as γ= 1 cm^{-1}, as found in [15–17] at higher frequencies (~1 THz) in experiments based on photomixing technology for high-resolution spectroscopy in biosystems.

The length of the dissipation time is one important problem concerning THz vibrational modes in biological molecules. A number of studies of the problem have been performed using MD simulation and the Langevin equation, along with an analysis of inelastic neutron scattering [18, 19] and other experimental techniques [20]. Nevertheless, the entire mechanism that determines dissipation is still not completely understood. The dissipation time can be sensitive to various factors, including temperature [21], oscillation frequency [18], and specific interactions between the molecule and its water shell [22]. The estimates from inelastic neutron scattering lead to very large broadening of low-frequency motions. Possible reasons for the differences between experiments and simulations have been discussed in [23], in particular the much higher vibrational density of states in simulations compared to neutron scattering experiments. It is known from experiments that “proteins exist in an ensemble of structures, described by an energy landscape” [24], and that neutron scattering spectra result from averaging over different protein conformations or substrates. These motions, however, are quite different from quasi-harmonic vibrational modes in THz, and especially in the sub-THz spectral range, for both time and displacement scales [21]. Weak THz vibrations associate with displacements at distances on the order of only ~0.1–1.0 Angstrom. These oscillations can survive for a relatively long time, since many slow relaxation processes that are important for conformational changes are not involved.

The gamma factor can depend on frequency [18], and we expect a lower value in our sub-THz range compared to the higher frequency THz region. However, in our simulations, we used the frequency-independent gamma factor as a first try, and think that this is a reasonable approach since our frequency range is rather narrow, from 10 to 25 cm^{-1}.

MD simulations are usually used as the principal theoretical method for studying protein dynamics [25].

However, the accuracy of the simulation and the ability to reproduce the experimental results relies on many different factors during the simulation process. For nucleic acids, it was demonstrated that the choice of force field is important [26]. Force fields developed by Amber are widely used and refined [27], and they show good ability to reproduce many properties of nucleic acids [28] and proteins. Also, the choice of water environment has an effect on the accuracy of the simulated results [29]. Perryman et al. [30] developed an "automated reformatting protocol" and tested many sets of simulation parameters to obtain reproducible results. At the same time, the authors of [30, 31] hypothesize that only certain mini-proteins and protein fragments can currently be represented accurately by MD simulations.

The problem of poor simulation convergence has been discussed in the works of many authors. The authors of [32] state that "When repeated from slightly different but equally plausible initial conditions, MD simulations of protein equilibrium dynamics predict different values for the same dynamic property of interest" [33–35]. The same authors [32] state that the variations occur because of insufficient sampling of protein's conformational space, an effect known as the "sampling problem" [34–38]. Most pico-to-nanosecond simulations of proteins and nucleic acids in water are considered to be not well equilibrated in some way, and are thought to contain "rare events" [39–41]. There are also those who still believe that the simulation length is crucial to the convergence of results [42, 43]. Nevertheless, even several 40 ns MD simulations of HPr and T4 lysozyme failed to achieve convergence in sampling [38]. Not only did these simulations fail to provide a complete picture of the protein's conformational space, they also suggest that this goal will remain unattainable in the foreseeable future.

When investigating the convergence problem [37], it was noted that multiple picosecond simulations with different initial conditions (in their case the velocities assigned) can sample more of a protein's conformational distribution than a single nanosecond trajectory. The authors cite that a typical trajectory samples (i.e., it is trapped) in a single, localized region of conformational space with few possible transitions between conformational regions. Extending the simulation time makes it more probable that a transition to another region of conformational space will occur, yet a single trajectory is believed to be unlikely to be representative of the full range of conformations that are thermally accessible to the system. However, when using multiple trajectories, each trajectory can sample a different phase space compared to that sampled by any other trajectory. The authors base these conclusions on vacuum simulations, but believe that they are applicable to explicitly solvated samples, with conformational transitions occurring more slowly in solvated samples.

Thus, although a sufficiently long simulation length is required to guarantee the stability of the system, there is no general consensus on the factors that contribute to the convergence of the simulation. Our previous simulations were dependent on initial conditions, and this situation has also been observed by Elofsson et al. when simulating thioredoxin [33].

In this study, MD simulations of sub-terahertz (THz) vibrational modes of the protein thioredoxin was conducted in order to find the conditions needed for the simulation to converge, improve the correlation between experimental and simulated spectra, and ultimately enhance the predictive capabilities of computational modeling.

We checked the consistency and accuracy of MD simulations of the sub-THz vibrational modes by comparing the results of simulations with different initial conditions, protocols and parameters to the experimental results. It was demonstrated that using the constant energy simulation protocol (NVE) during the production run yields more accurate results than the constant temperature regime (NPT) for several reasons. Constant energy simulations that do not involve frequent exchange with the external bath for temperature regulation induce fewer disturbances into the trajectories of atoms, and they prevent the transitions of proteins into different conformations. At the same time, the NVE protocol requires that more attention is paid to the choice of starting energy in the production run. Better simulation convergence and improved consistency between simulated vibrational frequencies and experimental data were obtained by using a new procedure for averaging correlation matrices of atomic coordinates in MD simulations. We also found that the optimal time to use when dividing the production run into equal subintervals to calculate individual correlation matrices is ~100 ps. This result is in general agreement with relaxation dynamics timescales for the thioredoxin active center, coupled protein–water fluctuations [44, 45], and our experimental data on the spectral width of vibrational modes [10].

Methods

Our modeling work focuses on complexes of thioredoxin and water. Thioredoxin is a relatively small protein with 108 amino acid residues in a known sequence, which allows for relatively quick serial simulations. With an 8 Å water shell, this system of 12,154 atoms can still be effectively simulated using Amber. However, Amber is empirically parameterized to correctly represent the structural behavior of nucleic acids and proteins, which would be needed to predict non-bond-breaking conformational changes [46]. It was not specifically created to simulate low-frequency vibrational modes and THz absorption. Hence, it is necessary to perform a systematic study of the default simulation parameters and protocols provided in Amber, with the goal of achieving better simulation accuracy and convergence. We have not investigated many parameters, only several that are most relevant to the physics of low-frequency vibrations and which may be important for simulation convergence. In our study, we used the following very general guiding principles:

1) The simulation protocol has to ensure the consistency of spectra between different simulation subintervals and between different simulations within one localized region without conformational change
2) Within an optimal domain of a particular parameter, the resulting vibrational frequencies and absorption spectra should not be sensitive to the parameter's exact value

Building on our previous work, the empirically parameterized molecular mechanics force field FF03 [13] and the water model TIP3P [14] are used. Previously, we utilized several other water models to analyze liquid water properties at sub-THz frequencies [47]. While the specific choices of a force field and a water model may have an effect on the resulting absorption spectra, they are unlikely to influence the convergence of results.

The experimental procedure used to measure the sub-THz spectra of thioredoxin on a polycarbonate (PC) membrane substrate has already been described in our papers [10, 11]. The only difference is that the spectrum of dry material, which was included at that time in the averaged experimental results, is eliminated in this work, and only the spectrum of the material in a water solution is used. This is more relevant to our MD simulation. At the same time, solvated biomaterial presents much higher vibrational peak intensities than dry samples, resulting in more reproducible and reliable spectral features in experiments.

Preparation steps

The basic MD simulation procedure in Amber consists of preparation and production run stages. In the preparation stage, a PDB file containing information about thioredoxin's atomic coordinates and connectivity is used to generate a topology file. The missing hydrogen atoms are added and the molecule is solvated with an 8 Å shell of TIP3P water. At this stage, the water molecules have not felt the influence of the solute and there are gaps between the solvent and solute, and between the solvent and box edges. A 1000-point energy minimization step is conducted using steepest descent followed by conjugate gradient algorithms [46], while the protein is held fixed in place by a force constant of 5 kcal mol^{-1} Å^{-2}. Next, the solvent and solute are relaxed together for 2500 steps, which allows for the whole system to reach a local potential energy minimum.

There are three different statistical ensembles available in Amber: constant volume and temperature (NVT), constant temperature and pressure (NPT), and constant volume and energy (NVE). The NVT ensemble is used to raise the temperature to 293 K by scaling the velocities of atoms using a Langevin algorithm [48]. In this heating process, bonds involving hydrogen are fixed. It takes ~16 ps in our case, and the protein atoms are restrained using a 10 kcal mol^{-1} Å^{-2} force constant to ensure that the temperature is raised without causing any drastic disturbances to the solute structure. Constant pressure periodic boundary conditions with isotropic position scaling are then used to scale the system volume during 100 ps to reach a density of ~1 g/cm^3. The following system properties are checked to ascertain the quality of the equilibrium: total energy, temperature, and density. During these steps, a 10Å real space cutoff is used with a 2 fs integration time step. The parameters for particle mesh Ewald charge grid spacing, pressure relaxation, etc. are left at their default values.

Thus, the preparation procedure consists of several steps, including a solvation step, followed by energy minimization of the total system (thioredoxin and water molecules), heating to room temperature, and density adjustment. Once the system has attained experimental values of temperature and density, the MD random velocities from the Maxwellian distribution at room temperature are assigned to all atoms in the system based on the seed given by the pseudo-random number generator (ig), followed by another equilibration step (NPT ensemble) for further energy minimization. In the NPT step, the center of mass translation (COM) is periodically removed every 1 ps. This measure prevents the molecule from leaving the periodic box in the following long production run where the COM velocity check is turned off.

Production simulation

During the production run, the microcanonical ensemble or constant energy and volume protocol is used, since it produces more consistent results compared to the NPT ensemble, as will be shown in the "Results" section. The molecule is allowed to move freely during this step without temperature and pressure regulation, while the trajectories of atoms are recorded for further derivation of vibrational modes of the absorption spectrum.

It is our observation that the parameters used in the production run are critical to the resulting absorption spectra. The values of interest are the integration time between two consecutive MD steps, the direct sum tolerance, the charge grid spacing, the cutoff radius, and the width of the nonbonded "skin." Nevertheless, a search for the optimal values of these parameters can be completed only after improving the convergence. The values of the abovementioned parameters during the production run are as follows: 1 fs integration time step, direct sum tolerance of 10^{-6} [49, 50], SHAKE tolerance of 10^{-6} [51], charge grid spacing ~2 Å, cutoff radius range of 10 Å, and width of the nonbonded skin of 2 Å. The molecular trajectories are saved every 0.05 ps, and production simulation lengths of 10–600 ps were used. It was difficult to choose the values of certain parameters, so they are left at their default values. The saved trajectories from the NVE production simulations are then converted to the covariance matrix of atomic displacements using the quasiharmonic analysis. Utilizing the relation between the covariance matrix and the inverse of the force-constant matrix [52, 53], it is possible to find the latter matrix. The eigenvectors and eigenvalues (eigenfrequencies) of the normal modes are then determined by diagonalizing the force-constant matrix.

Absorption coefficients

The absorption coefficient α as the function of the frequency ν can be approximately calculated through the relationship between α and the imaginary part of the dielectric permittivity [5, 54]:

$$\alpha(\nu) = \gamma\nu^2 \sum_k \frac{S_k}{\left(\nu^2 - \nu_k^2\right)^2 + \gamma^2\nu^2},$$

where ν_k are the normal mode frequencies calculated by diagonalizing the force-constant matrix, and S_k are the oscillator strengths computed for all vibrational modes k. The oscillator strength calculated along each normal mode k is proportional to the squared dipole moment variation. The value of oscillator dissipation for all vibrational modes in the sub-THz range is taken from our experimental work [10] as γ=0.5 cm^{-1}, which corresponds to a relaxation dynamics timescale of~70 ps.

Results

Several findings are presented in this work. We found that using the NVE ensemble in a production run gives the most stable results, and that the starting total energy of the system (ETOT) at the beginning of the production run significantly affects the results. Additionally, the length of the production run and the method of averaging results from the production run considerably influence the absorption spectra obtained.

Total energy of the system in the NVE production simulation

As described in the "Methods" section, after random velocities are assigned to all atoms within the Maxwellian velocity distribution, the resulting system is further equilibrated using the NPT ensemble. In Fig. 1, the total energy during this NPT step is plotted against time after initial velocities have been assigned using three different random numbers. As can be seen, there is initially a large jump in the total energy of the system, followed by a slower decline. At the same time, we found that, depending on the assigned velocities, the equilibration times are variable. In Fig. 1, it takes about 5–35 ps before the system energy reaches a common value of around −28500 kcal mol^{-1}. The total energy of the system still continues to decline, but the fluctuations make it difficult to determine the equilibration minimum. A typical picture of total energy fluctuations (~0.3%) in the NPT ensemble is given in Fig. 2. The true value of ETOT corresponding to the equilibration minimum is hidden within these fluctuations. However, at this moment, we need to choose a value for the total energy that will be constant in the following NVE production run, as will be discussed below.

To start the production simulation, the first decision to be made is to choose between NPT and NVE ensembles. Production runs in the NPT ensemble using Berendsen temperature regulation seemed promising, since the correlation with the experimental spectra was not bad for some values of the simulation parameters. However, the repeated removal of molecular translation, the adjustment of volume, and the changes in system temperature caused by coupling to a surrounding bath seemed to lead to noise in the total energy of the system, as shown in Figs. 1 and 2. In particular, we found that the exchange of energy between the system and the external thermostat disturbs the system and leads to poor reproducibility of simulation results. For this reason, a constant total energy protocol (NVE) that does not require these corrections is more preferable. Hence, simulations with constant energy were conducted for all of the production runs presented in this work. Since a constant energy protocol does not permit significant energy changes, the last value of the energy during equilibration locks the energy of the system to almost a single value during the production run. We observed that simulation results in the NVE regime are very sensitive

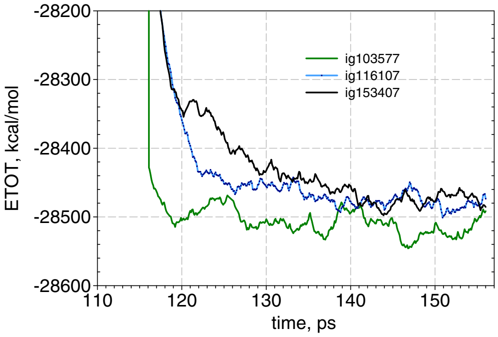


Fig. 1 Total energy in the NPT equilibration step as a function of time after Maxwellian velocities are assigned using three different random numbers

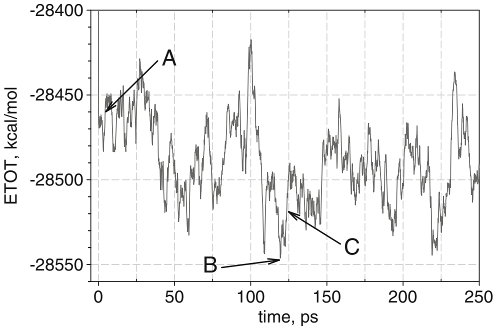


Fig. 2 Fluctuation of the total energy of the system (thioredoxin and water molecules) after initial velocity assignment at time zero. Point *A* represents the system at a high energy level, while point *B* represents the system at a lower energy level

to the value of the total energy of the system chosen at the end of the NPT equilibration step. Vibrational spectra calculated from production runs that started at points A, B, and C (Fig. 2) give rather different results. The consequence of starting the production run at a high energy level (point A) is shown in Fig. 3. As can be seen, after a sufficiently long simulation (> than 300 ps), some modes have become extremely intense. There is a large spike at 21 cm^{-1} that reaches a maximum absorption value of nearly 20 (arbitrary units) at 560 ps. We suggest that such a peak is a consequence of the energy level being far from the equilibrium value. The effect disappears if the production run is performed at lower energy levels (points C and B) that are closer to the minimum (Fig. 4). The fluctuations make it difficult to determine the value of ETOT that corresponds to the system's equilibration state. In our study, we tested different values of ETOT between points A and B. The best correspondence to our experimental data and better convergence are obtained for ETOT values within the interval of ~10 kcal mol^{-1} around level C, which is higher than the most minimal energy observed at point B. Our findings are consistent with other studies [55, 56], where authors have shown that the energy of the modeled protein is not always the minimum that corresponds to its native state when a standard force field such as AMBER FF03 or FF99 is used.

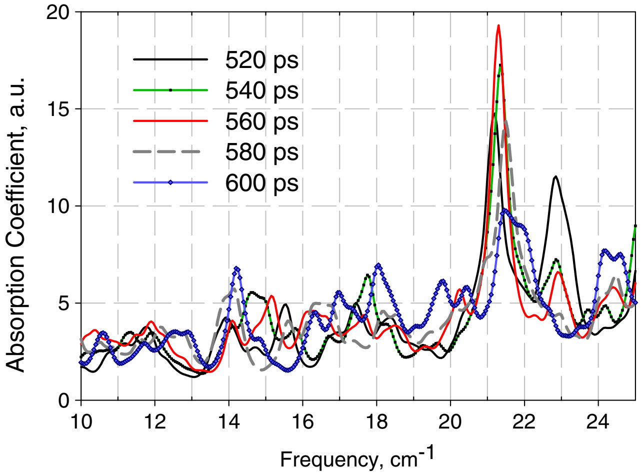


Fig. 3 Absorption spectra for 520–600 ps NVE production runs conducted at the highest energy level A in Fig. 2 (ETOT=−28460 kcal mol^{-1})

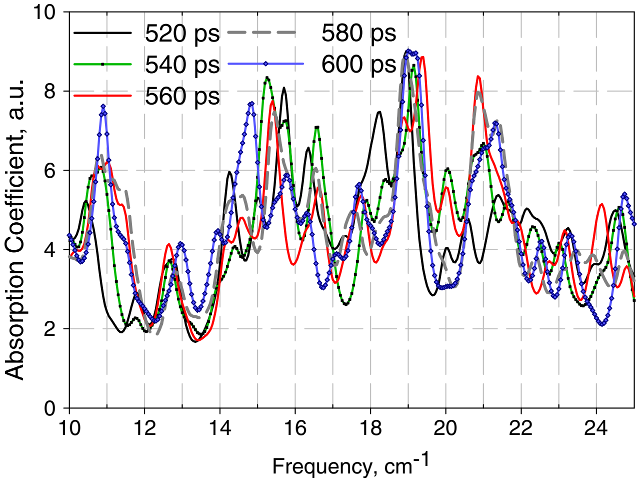


Fig. 4 Absorption spectra for 520–600 ps NVE production runs conducted at energy level B in Fig. 2 (ETOT=−28540 kcal mol^{-1})

Time of production simulation

Since the trajectories file from the production run contains superposed oscillations/vibrational modes, the shortest time period of the production run can be evaluated from the lowest frequency in our spectral range (10 cm^{-1}). An even better estimate would be to use 10–20 oscillating periods of the lowest frequency as the minimum time of the production run. One period of oscillations at 10 cm^{-1} is less than or equal to 3 ps, which gives us the estimate for the minimal production time as 30–60 ps or larger. Figure 5 confirms that simulation times of 5–15 ps are not long enough. Absorption amplitudes are damped and frequencies are occasional. The results of simulations with longer production times are shown in Fig. 6. It is clear that after 100 ps there is no increase in peak intensity.

We relate this result to the relaxation timescale for the protein's dynamics in water solution. It was shown that the robust quenching dynamics of thioredoxin's active center leads to a timescale of ~100 ps [44]. There is also evidence that the relaxation timescale for the dynamics of protein-water motions is around 90 ps [45]. In addition, our own experimental data demonstrate that the spectral width of vibrational modes is ~0.3–0.5 cm^{-1} [10], which yields a relaxation time of ~70–110 ps. These estimates give us the upper limit for the production time, since the vibrational phase is lost in longer simulations. Thus, a 90–100 ps production run is a justifiable estimate. This scale is much smaller than is needed to observe events like structural changes, which occur

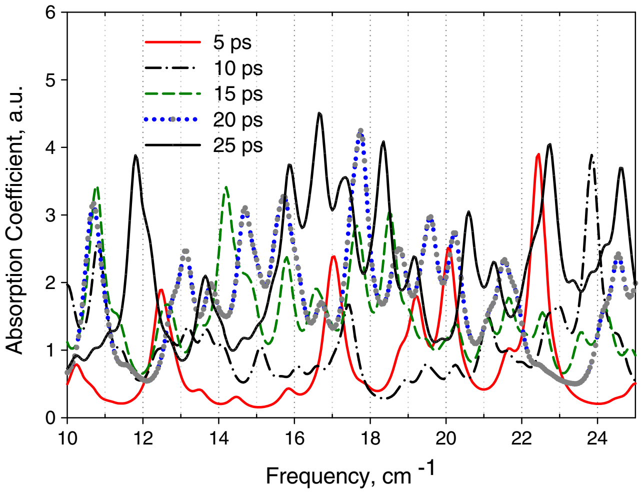


Fig. 5 Simulated absorption spectra with production runs of between 5 to 25 ps; integration step 0.25 fs. Shorter simulation times result in damped absorption amplitudes and reduced peak density and occasional frequency

in the millisecond domain. Another result discerned from Fig. 6 is that there is no consistency between individual production runs. Our computational experiments show that we are not able to get consistent or reproducible results even at much longer simulation times of up to 1200 ps.

From these results, we can conclude that an averaging procedure must be developed with subinterval durations of about 100 ps to achieve more reliable results.

Developing a new averaging procedure for vibrational absorption spectra simulation

As a first attempt at solving the problem, two relatively simple averaging procedures were tested: averaging individual absorption spectra and averaging MD trajectories from different trajectory sections of the same simulation or sections from simulations with different initial velocities. In both of these cases, averaging gives low vibrational peak intensities.

Figure 7 demonstrates individual 100 ps simulated spectra from six sections of the same 600 ps trajectory, the result of averaging the absorption coefficient spectra from these six sections, and the experimental spectrum. The lack of convergence is obvious, and since individual absorption spectra are not correlated, averaging leads to a damped spectrum. Thus, averaging the absorption coefficient does not improve the situation; neither does averaging individual absorption spectra calculated from fragments of trajectories with different initial velocities.

Figure 8 plots the results from averaging five and eight equally spaced sections of a trajectory from a 400 ps constant energy (NVE) simulation. As can be seen, the amplitudes of the resulting absorption peaks are greatly reduced compared to the case without averaging, indicating that individual spectra are poorly correlated or are not correlated at all. Thus, averaging sections of the trajectory does not seem to improve the simulated results either, and most importantly convergence is not observed.

The third procedure we applied was to average correlation (mass-weighted covariance) matrices, which can be computed in Amber for further quasi-harmonic analysis. In a typical simulation, the correlation matrix of atomic coordinates is calculated from the atomic trajectories recorded after the entire production run. In our new analytic procedure, we divide a single production run into individual sections with equal time intervals. For each of these sections, a correlation matrix is then obtained. The average correlation matrix can be found by summing all of the matrices and dividing by the number of matrices. Finally,

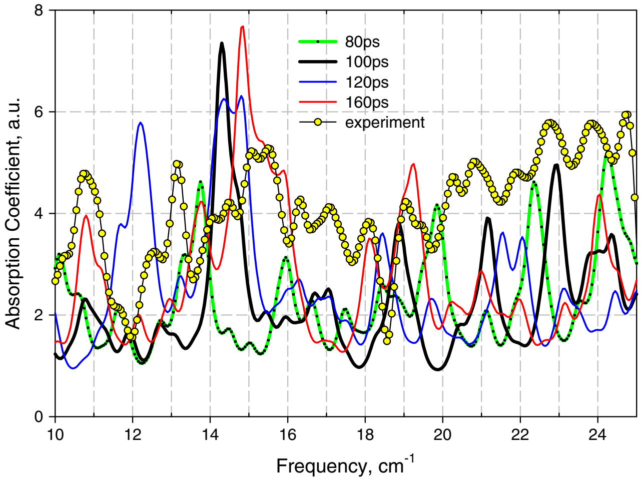


Fig. 6 Simulated absorption spectra with production run lengths of between 80 and 160 ps demonstrate that there is no significant increase in peak intensity after 100 ps

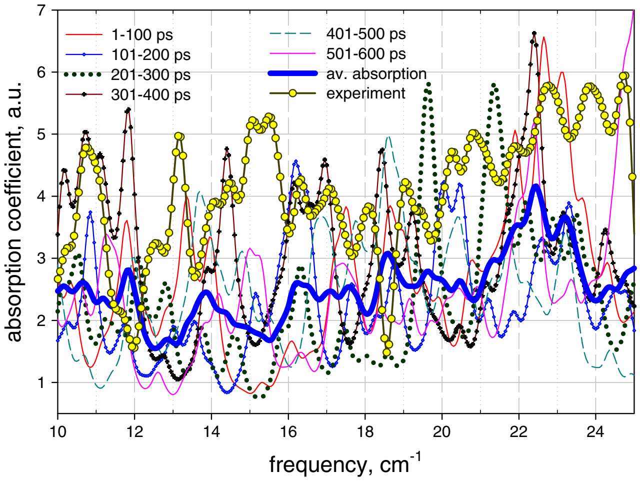


Fig. 7 Six individual simulated spectra, each obtained from a 100 ps simulation, the absorption coefficient averaging results, and the experimental spectrum. Individual absorption spectra are not correlated

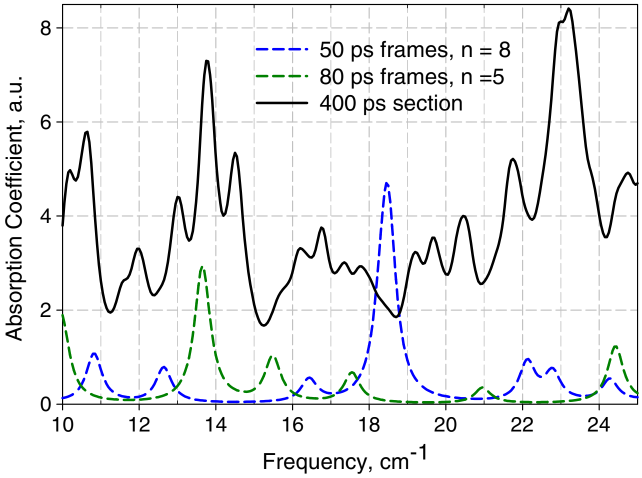


Fig. 8 Absorption spectrum from a 400 ps constant energy (NVE) simulation (*solid line*), and averaging results: eight and five equally spaced sections

the force constant matrix is obtained in the classical limit from the atomic displacement correlation matrix for an array of coupled harmonic oscillators. The creation of the force constant matrix and its diagonalization (to find frequencies of oscillations) are functions provided in Amber's ptraj package. However, the ptraj module only allows the correlation matrix to be created from the native atom trajectory file for a subsequent computation of the force constant matrix. We edited the open source package ptraj (Amber Tools 1.2, http://www.chpc.utah.edu/~cheatham/ptraj-9.9i.tar.gz) and recompiled it to accept an external source for the averaged correlation matrix with nine significant digits after the decimal point. Figure 9 plots the experimental absorption spectrum together with modeling results from two different procedures: averaging the six individual absorption spectra shown in Fig. 7, and averaging the correlation matrices for six 100 ps simulation intervals within a 600 ps simulation. Averaging the absorption spectra results in low peak intensities. Averaging correlation matrices produces larger peak intensities and better agreement between the entire spectrum and the experiment.

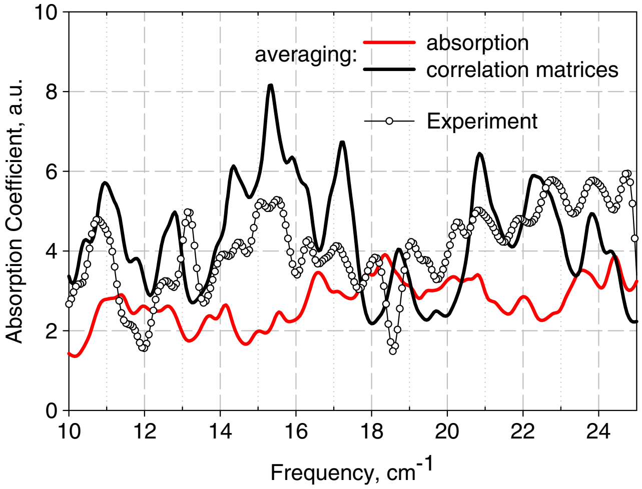


Fig. 9 Comparison of the experimental and simulated results obtained using two averaging procedures: averaging the absorption coefficient and averaging matrices (6 intervals, 100 ps each, the same simulation parameters). The initial 600 ps NVE production run was started at the energy level C in Fig. 2

Thus, we found that the third method is the best way of averaging results to reduce noise in the system. We also noted that NVE production runs that are started with similar values of ETOT within a range of ~10 kcal mol^{-1} produce reasonably similar absorption spectra when this procedure for averaging correlation matrices is used.

A large number of protein conformations might lead not only to spectroscopic peak broadening but also to different peak frequencies. Our computational experiments have shown that the optimal value of a simulation parameter should give the highest intensities of vibrational modes over the entire spectrum if the production run is carried out using a constant energy ensemble and the total energy of the system (ETOT) is close to the (local) equilibration minimum. Note that a constant energy simulation that does not involve frequent exchange with the external bath for temperature regulation induces less disturbance into the trajectories of atoms, and also prevents protein molecules from transitioning into different conformations. In order to optimize the simulation parameters, we must be able to compare the results from different simulations and those obtained with different atom velocities from the Maxwellian distribution. We found that this can be done if the difference in ETOT is not higher than ~10 kcal mol^{-1}.

Conclusions

We studied the accuracy and convergence of MD simulations of the sub-THz vibrational modes by comparing simulations with different initial conditions, protocols, and parameters to the experimental results. Our main findings are as follows:

- Using the NVE ensemble in a production run gives more stable results than the NPT regime.
- The choice of the starting total energy of the system (ETOT) at the beginning of the production run significantly affects the results, and this energy must be close to the equilibration minimum.
- The choice of the production run length considerably influences the obtained absorption spectra. The optimal production run length (~100 ps) can be derived from the vibrational relaxation time.
- A new averaging procedure for mass-weighted covariance matrices of atomic trajectories in MD simulation has been developed. This procedure significantly improved

modeling convergence, thus allowing the further optimization of simulation parameters to achieve a better correlation between the experimental and simulated spectra. We plan to conduct parameter optimization in a continuation of this work.

Acknowledgments This work is supported by the Defense Threat Reduction Agency, grant # HDTRA1-08-1-0038.

References

1. Van Zandt L, Saxena V (1992) Identifying and interpreting spectral features of DNA in the microwave-submillimeter range. In: Structure and function. Adenine, New York, p 237
2. Duong TH, Zakrzewska K (1997) Calculation and analysis of low frequency normal modes for DNA. J Comput Chem 18:796–811
3. Smye SW, Chamberlain JM, Fitzgerald AJ, Berry E (2001) The interaction between terahertz radiation and biological tissue. Phys Med Biol 46:R101–R112
4. Li X, Globus T, Gelmont B, Salay LC, Bykhovski A (2008) Terahertz absorption of DNA decamer duplex. J Phys Chem A 112:12090–12096
5. Globus T, Bykhovskaia M, Woolard D, Gelmont B (2003) Sub-millimetre wave absorption spectra of artificial RNA molecules. J Phys D 36:1314–1322
6. Bykhovski A, Globus T, Gelmont B, Woolard D (2006) An analysis of the THz frequency signatures in the cellular components of biological agents. In: Proc SPIE Defense Security Symp, Orlando, FL, USA, 17–21 April 2006, pp V 6212-8, pp 132–141
7. Bykhovski A, Xiaowei L, Globus T, Gelmont B, Woolard D, Samuels AC, Jensen JO (2005) THz absorption signature detection of genetic material of *E. coli* and *B. subtilis*. In: Proc SPIE Chem Biol Standoff Detection III, Boston, MA, USA, 24 Oct 2005, pp 59950N-59950N-10
8. Case DA, Pearlman DA, Caldwell JW, Cheatham TE III, Wang J, Ross WS, Simmerling CL, Darden TA, Merz KM, Stanton RV, Cheng AL, Vincent JJ, Crowley M, Tsui V, Gohlke H, Radmer RJ, Duan Y, Pitera J, Massova I, Seibel GL, Singh UC, Weiner PK, Kollman PA (2004) AMBER 8. University of California, San Francisco
9. Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Ross WS, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossváry I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, Sagui C, Babin V, Kollman PA (2008) AMBER 10. University of California, San Francisco
10. Bykhovski A, Globus T, Khromova T, Gelmont B, Woolard D (2008) Resonant terahertz spectroscopy of bacterial thioredoxin in water: simulation and experiment. In: Woolard D, Jensen J (eds) Spectral sensing research for water monitoring applications and frontier science and technology for chemical, biological and radiological defense (Selected Topics in Electronics and Systems 48). World Scientific, Singapore, pp 367–375
11. Globus T, Woolard D, Crowe TW, Khromova T, Gelmont B, Hesler J (2006) Terahertz Fourier transform characterization of biological materials in a liquid phase. J Phys D 39:3405–3413
12. Katti SK, LeMaster DM, Eklund H (1990) Crystal structure of thioredoxin from *Escherichia coli* at 1.68 A resolution. J Mol Biol 212:167–184
13. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24:1999–2012
14. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926
15. Plusquellic DF, Siegrist K, Heilweil EJ, Esenturk O (2007) Applications of terahertz spectroscopy in biosystems. Chem Phys Chem 8:2412–2431
16. Zhang H, Siegrist K, Douglas KO, Gregurick SK, Plusquellic DF (2008) THz investigations of condensed phase biomolecular systems. Methods Cell Biol 90:417–434
17. Korter TM, Plusquellic DF (2004) Continuous-wave terahertz spectroscopy of biotin: vibrational anharmonicity in the far-infrared. Chem Phys Lett 385:45–51
18. Moritsugu K, Smith JC (2005) Langevin model of the temperature and hydration dependence of protein vibrational dynamics. J Phys Chem B 109:12182–12194
19. Smith J, Becker T, Fischer S, Noé F, Tournier A, Ullmann G, Kurkal V (2006) Physical and functional aspects of protein dynamics. In: Poon WCK, Andelman D (eds) Soft condensed matter physics in molecular and cell biology. Taylor & Francis, New York, pp 225–241
20. Vinh NQ, Allen SJ, Plaxco KW (2011) Dielectric spectroscopy of proteins as a quantitative experimental test of computational models of their low-frequency harmonic motions. J Am Chem Soc 133:8942–8947
21. Cusack S, Doster W (1990) Temperature dependence of the low frequency dynamics of myoglobin. Measurement of the vibrational frequency distribution by inelastic neutron scattering. Biophys J 58:243–251
22. LeBard DN, Matyushov DV (2010) Ferroelectric hydration shells around proteins: electrostatics of the protein–water interface. J Phys Chem B 114:9246–9258
23. Balog E, Smith JC, Perahia D (2006) Conformational heterogeneity and low-frequency vibrational modes of proteins. Phys Chem Chem Phys 8:5543–5548
24. Frauenfelder H, McMahon BH, Austin RH, Chu K, Groves JT (2001) The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. Proc Natl Acad Sci USA 98:2370–2374
25. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. Nat Struct Biol 9:646–652
26. van der Spoel D, van Buuren AR, Tieleman DP, Berendsen HJ (1996) Molecular dynamics simulations of peptides from BPTI: a closer look at amide–aromatic interactions. J Biomol NMR 8:229–238
27. Pérez A, Marchán I, Svozil D, Sponer J, Cheatham TE III, Laughton CA, Orozco M (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophys J 92:3817–3829
28. Beveridge DL, McConnell KJ (2000) Nucleic acids: theory and computer simulation, Y2K. Curr Opin Struct Biol 10:182–196
29. Mahoney MW, Jorgensen WL (2000) A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. J Chem Phys 112:8910
30. Perryman AL, Lin JH, Andrew McCammon J (2006) Optimization and computational evaluation of a series of potential active site inhibitors of the V82F/I84V drug-resistant mutant of HIV-1 protease: an application of the relaxed complex method of structure-based drug design. Chem Biol Drug Des 67:336–345
31. Seibert MM, Patriksson A, Hess B, van der Spoel D (2005) Reproducible polypeptide folding and structure prediction using molecular dynamics simulations. J Mol Biol 354:173–183
32. Likić VA, Gooley PR, Speed TP, Strehler EE (2005) A statistical approach to the interpretation of molecular dynamics simulations of calmodulin equilibrium dynamics. Protein Sci 14:2955–2963

33. Elofsson A, Nilsson L (1993) How consistent are molecular dynamics simulations? Comparing structure and dynamics in reduced and oxidized *Escherichia coli* thioredoxin. J Mol Biol 233:766–780
34. Auffinger P, Louise-May S, Westhof E (1995) Multiple molecular dynamics simulations of the anticodon loop of tRNA[Asp] in aqueous solution with counterions. J Am Chem Soc 117:6720–6726
35. Likić VA, Prendergast FG (2001) Dynamics of internal water in fatty acid binding protein: computer simulations and comparison with experiments. Proteins 43:65–72
36. Straub JE, Rashkin AB, Thirumalai D (1994) Dynamics in rugged energy landscapes with applications to the S-peptide and ribonuclease A. J Am Chem Soc 116:2049–2063
37. Caves LS, Evanseck JD, Karplus M (1998) Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. Protein Sci 7:649–666
38. Hess B (2002) Convergence of sampling in protein simulations. Phys Rev E 65(3 Pt 1):031910
39. Chandrasekhar I, Clore GM, Szabo A, Gronenborn AM, Brooks BR (1992) A 500 ps molecular dynamics simulation study of interleukin-1[beta] in water: correlation with nuclear magnetic resonance spectroscopy and crystallography. J Mol Biol 226:239–250
40. Balsera MA, Wriggers W, Oono Y, Schulten K (1996) Principal component analysis and long time protein dynamics. J Phys Chem 100:2567–2572
41. Clarage JB, Romo T, Andrews BK, Pettitt BM, Phillips GN Jr (1995) A sampling problem in molecular dynamics simulations of macromolecules. Proc Natl Acad Sci USA 92:3288–3292
42. Amadei A, Ceruso MA, Di Nola A (1999) On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. Proteins 36:419–424
43. Daggett V (2000) Long timescale simulations. Curr Opin Struct Biol 10:160–164
44. Qiu W, Wang L, Lu W, Boechler A, Sanders DAR, Zhong D (2007) Dissection of complex protein dynamics in human thioredoxin. Proc Natl Acad Sci USA 104:5366–5371
45. Li T, Hassanali AA, Kao YT, Zhong D, Singer SJ (2007) Hydration dynamics and time scales of coupled water-protein fluctuations. J Am Chem Soc 129:3376–3382
46. Lewars EG (2003) Computational chemistry: introduction to the theory and applications of molecular and quantum mechanics, 1st edn. Springer, Berlin
47. Globus T, Bykhovski A, Khromova T, Gelmont B, Tamm LK, Salay LC (2007) Low-terahertz spectroscopy of liquid water. In: Proc SPIE Terahertz Phys Devices Syst II, Boston, MA, USA, 11 Sept 2007, 67720S
48. Leach A (2001) Molecular modelling: principles and applications, 2nd edn. Prentice Hall, Harlow
49. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. J Chem Phys 98:10089–10092
50. Essmann U, Perera L, Berkowitz M, Darden T, Lee H, Pedersen L (1995) A smooth particle mesh Ewald method. J Chem Phys 103:8577–8593
51. Ryckaert J, Ciccotti G, Berendsen H (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. J Comput Phys 23:327–341
52. Karplus M, Kushick JN (1981) Method for estimating the configurational entropy of macromolecules. Macromolecules 14:325–332
53. Levy RM, Karplus M, Kushick J, Perahia D (1984) Evaluation of the configurational entropy for proteins: application to molecular dynamics simulations of an α-helix. Macromolecules 17:1370–1374
54. Globus T, Woolard D, Bykhovskaia M, Gelmont B, Werbos L, Samuels AC (2003) THz-frequency spectroscopic sensing of DNA and related biological materials. IJHSES 13:903–936
55. Ji CG, Zhang JZH (2009) Electronic polarization is important in stabilizing the native structures of proteins. J Phys Chem B 113:16059–16064
56. Wroblewska L, Skolnick J (2007) Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? I. Large scale AMBER benchmarking. J Comput Chem 28:2059–2066

J Mol Model (2012) 18:2219–2226
DOI 10.1007/s00894-011-1241-y

ORIGINAL PAPER

A theoretical study on the reaction mechanism of O_2 with C_4H_9• radical

Hong-chen Du · Xue-dong Gong

Received: 16 May 2011 / Accepted: 12 September 2011 / Published online: 27 September 2011

Abstract Ab initio calculations have been performed using the complete basis set model (CBS-QB3) to study the reaction mechanism of butane radical (C_4H_9•) with oxygen (O_2). On the calculated potential energy surface, the addition of O_2 to C_4H_9• forms three intermediates barrierlessly, which can undergo subsequent isomerization or decomposition reaction leading to various products: HOO•+C_4H_8, C_2H_5•+CH_2CHOOH, OH•+C_3H_7CHO, OH•+cycle-C_4H_8O, CH_3•+$CH_3CHCHOOH$, CH_2OOH•+C_3H_6. Five pathways are supposed in this study. After taking into account the reaction barrier and enthalpy, the most possible reaction pathway is C_4H_9•+O_2→IM1→TS5→IM3→TS6→IM4→TS7→OH•+cycle-C_4H_8O.

Keywords C_4H_9 · O_2 · CBS-QB3 · Reaction mechanism · Theoretical study

Introduction

Hydrocarbon radicals are common intermediates in many chemical processes, and reactions of them with oxygen are important elementary steps in the atmospheric processes and combustion of hydrocarbons [1–5]. Deep understanding on the combustion reaction mechanism of hydrocarbon is an urgent scientific goal [1–17], and investigations on the reactions of reactive species such as hydrocarbon radicals can provide significant insights into combustion, hydrocarbon synthesis, interstellar space, and atmospheric chemistry. The reactivity of radicals which is very different from closed-shell molecules is extremely important in the reactions, e.g., between oxygen and hydrocarbon radicals. The generation of radicals and their participation in subsequent branching reaction steps affect the outcome of combustion processes. However, because of the difficulty in producing these transient species, often limited information is available under well-defined experimental conditions. Theoretical studies on the reaction of hydrocarbon radicals are also quite scarce except for some calculations on a few simple systems, among which the reactions of allyl radical (C_3H_5•) [18–24] and ethane radical (C_2H_5•) [25–36] have attracted more attentions than others. Estupiñán et al. studied the reactions of C_2H_5, n-C_3H_7, and i-C_3H_7 radicals with O_2 using the technique of laser photolysis/long-path frequency-modulation spectroscopy and ab initio method [37]. Wilke et al. studied the mechanism of the elimination of HO_2 from ethylperoxy (C_2H_5OO) using CCSD(T) method, and the calculated results agree well with the experiment [38]. Basevich et al. investigated the mechanisms of the oxidation and combustion of different alkanes [39–41]. Yet, many aspects of alkane combustion are still surrounded by controversy and confusion.

In this work, we examine some possible mechanisms of butane oxidation, i.e., the reaction of the butane radical (C_4H_9•) with oxygen molecule using the ab initio method.

Calculation methods

Calculations were carried out to obtain the lowest doublet potential energy surfaces for the reaction of C_4H_9• with O_2 using the complete basis set (CBS-QB3) [42] method

H.-c. Du · X.-d. Gong
Department of Chemistry,
Nanjing University of Science and Technology,
Nanjing 210094, China

X.-d. Gong (✉)
State Key Laboratory of Explosion Science and Technology,
Beijing Institute of Technology,
Beijing 100081, China
e-mail: gongxd325@mail.njust.edu.cn

implemented in the Gaussian 03 package [43]. This method provides optimized geometries and frequencies at the B3LYP/6-311 G(2 d,d,p) level. The frequencies were scaled by 0.99 prior to calculating thermodynamic functions. All transition states were verified by both the presence of single imaginary frequency and the connection between the designated reactants and products through intrinsic reaction coordinate (IRC) analysis [44].

Results and discussion

Five possible pathways have been investigated. The equilibrium geometries of the reactants, intermediates, transition states, and products are presented in Fig. 1. Figure 2 illustrates the schematic diagrams of the potential energy surface for the reactions. The energies of the various species relative to that of the reactants C_4H_9•+O_2 are listed in Table 1.

Pathway A

R→TS1→P1

This reaction pathway involves a single step of the direct hydrogen abstraction from the butane radical by oxygen molecule, C_4H_9•+O_2→HOO•+C_4H_8. In this reaction, no other intermediate forms. β-hydrogen abstraction by oxygen through the transition state TS1 leads to the formation of butylene, the reaction barrier is calculated to be 39.7 kJ mol^{-1}. Wagner et al. studied the reactions of C_2H_5•+O_2 experimentally and theoretically, and proposed that for the reaction of R•+O_2→HOO•+R_{-H}, the activation energy of direct abstraction of H atom is 20–40 kJ mol^{-1} [30], which is close to the value we obtained here. IRC calculations indicate that during the reaction process, the C3-H11 bond is elongated from 1.106 Å in C_4H_9• to 1.244 Å in TS1, then it ruptures to produce HOO• radical and butylene. The process is exothermic by −67.47 kJ mol^{-1}.

Pathway B

$$\begin{tikzcd}
& & & & \mathbf{TS3} \arrow[r] & \mathbf{P2} \\
\mathbf{R} \arrow[r] & \mathbf{IM1} \arrow[r] & \mathbf{TS2} \arrow[r] & \mathbf{IM2} \arrow[ur] \arrow[dr] & & \\
& & & & \mathbf{TS4} \arrow[r] & \mathbf{P3}
\end{tikzcd}$$

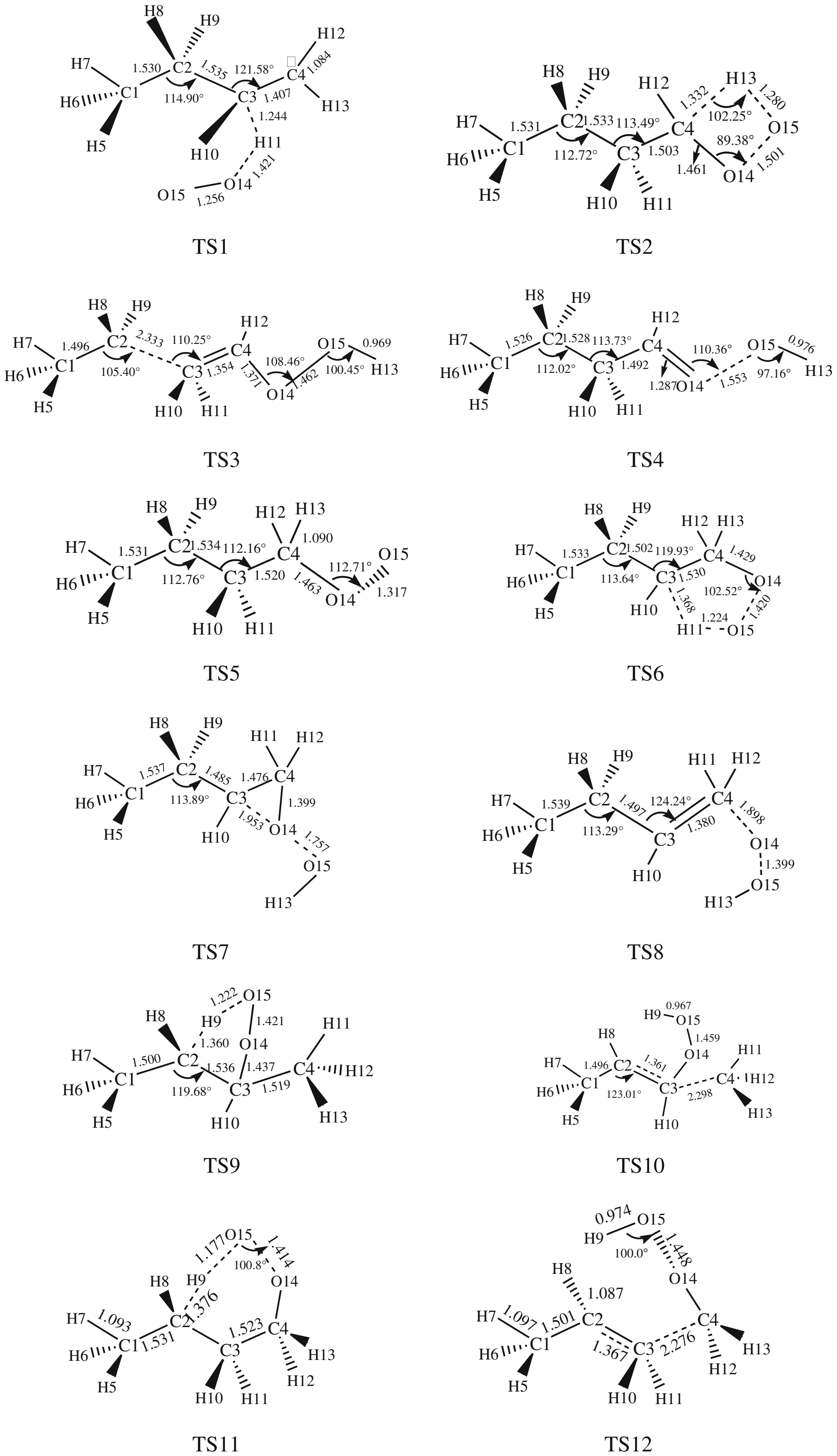
In this pathway, the O atom of O_2 attacks the C4 atom of C_4H_9• to form the intermediate C_4H_9OO• (denoted as IM1), this process is barrierless and exothermic by −148.46 kJ mol^{-1} with respect to the reactant C_4H_9•+O_2 at the CBS-QB3 level. Then IM1 undergoes a hydrogen (H13) migration from C4 to the neighboring O15 to form IM2 through a four-membered ring transition state TS2 with a barrier of 185.4 kJ mol^{-1}. During this process, the breaking bond C4-H13 is elongated from 1.093 Å in IM1 to 1.332 Å in TS2 and the new formed bond O15-H13 changes to 0.969 Å with a strained angle O14-O15-H13 (100.64°). As depicted in Fig. 2, IM2 is 45.44 kJ mol^{-1} more reactive than IM1, and can undergo two different bond dissociations: (1) C2-C3 ruptures through transition state TS3 with a high-barrier 119.38 kJ mol^{-1} to produce C_2H_5• and C_2H_3OOH (P2). The C2-C3 bond is elongated from 1.548 Å in IM2 to 2.333 Å in TS3, while the C3-C4 is shortened from 1.490 Å to 1.354 Å, indicating that the C3-C4 bond in TS3 is delocalized; (2) O14-O15 ruptures through TS4 with a barrier height 9.0 kJ mol^{-1} to form OH• and C_3H_7CHO (P3). In this process, the length of O14-O15 bond changes from 1.455 Å in IM2 to 1.553 Å in TS4 and the C4-O14 bond is shortened from 1.369 Å to 1.287 Å. Obviously, the channel to produce P3 with the lower barrier and larger exothermicity is more favorable. The rate-determing step of the pathway is IM1→TS2→IM2.

Pathway C

$$\begin{tikzcd}
& & & & & & \mathbf{TS7} \arrow[r] & \mathbf{P4} \\
\mathbf{R} \arrow[r] & \mathbf{IM1} \arrow[r] & \mathbf{TS5} \arrow[r] & \mathbf{IM3} \arrow[r] & \mathbf{TS6} \arrow[r] & \mathbf{IM4} \arrow[ur] \arrow[dr] & & \\
& & & & & & \mathbf{TS8} \arrow[r] & \mathbf{P1}
\end{tikzcd}$$

As shown in pathway B, IM1 can be formed from the reactants barrierlessly. IM1 may proceed an isomerization to IM3 by the migration of O15 atom through TS5. The barrier height is only 4.7 kJ mol^{-1}, that is to say this process is happens very easily, the exothermicity of this process is about null (−0.6 kJ mol^{-1}). IM3 can subsequently undergo a hydrogen migration of H11 from C3 to O15 to form IM4 through a five-membered ring transition state TS6 with a barrier of 143.6 kJ mol^{-1}, which is close to the activation energy for the similar reaction step of the combustion of ethane (143.5±10.0 kJ mol^{-1}) [45]. C3-H11 is 1.094 Å in IM3 and 1.368 Å in IM4; O14-O15 is 1.318 Å and 1.420 Å respectively in IM3 and IM4. The process from IM3 to IM4 is endothermic and the reaction enthalpy is 59.5 kJ mol^{-1}. Two competion reaction channels exist for IM4, the more favorable channel is the decomposition reaction by the bond cleavage of O14-O15 leading to OH•+cycle-C_4H_8O, in which the bond length of O14-O15 is elongated from 1.458 Å in IM4 to 1.757 Å in TS7. O14 gradually transfers to C3 and C3-O14 is shortened from 1.953 Å in TS7 to 1.434 Å in the product (P4). The activation energy for this process is 45.1 kJ mol^{-1}. Another channel involves the cleavage of the C4-O14 bond through TS8, which results in the same products (P1) as in pathway A. The corresponding energy barrier is 48.3 kJ mol^{-1}, higher than that (39.7 kJ mol^{-1}) of the pathway A. Considering the reaction endothermicity and barrier, the channel leading to P4 through TS7 is more favorable. The rate-determing step of the channel is IM3→TS6→IM4.

Fig. 1 Geometrical parameters of the species on the potential energy surface for the reaction of C_4H_9• with O_2 at the CBS-QB3 level. (bond lengths are in angstroms and angles are in degrees)



TS1

TS2

TS3

TS4

TS5

TS6

TS7

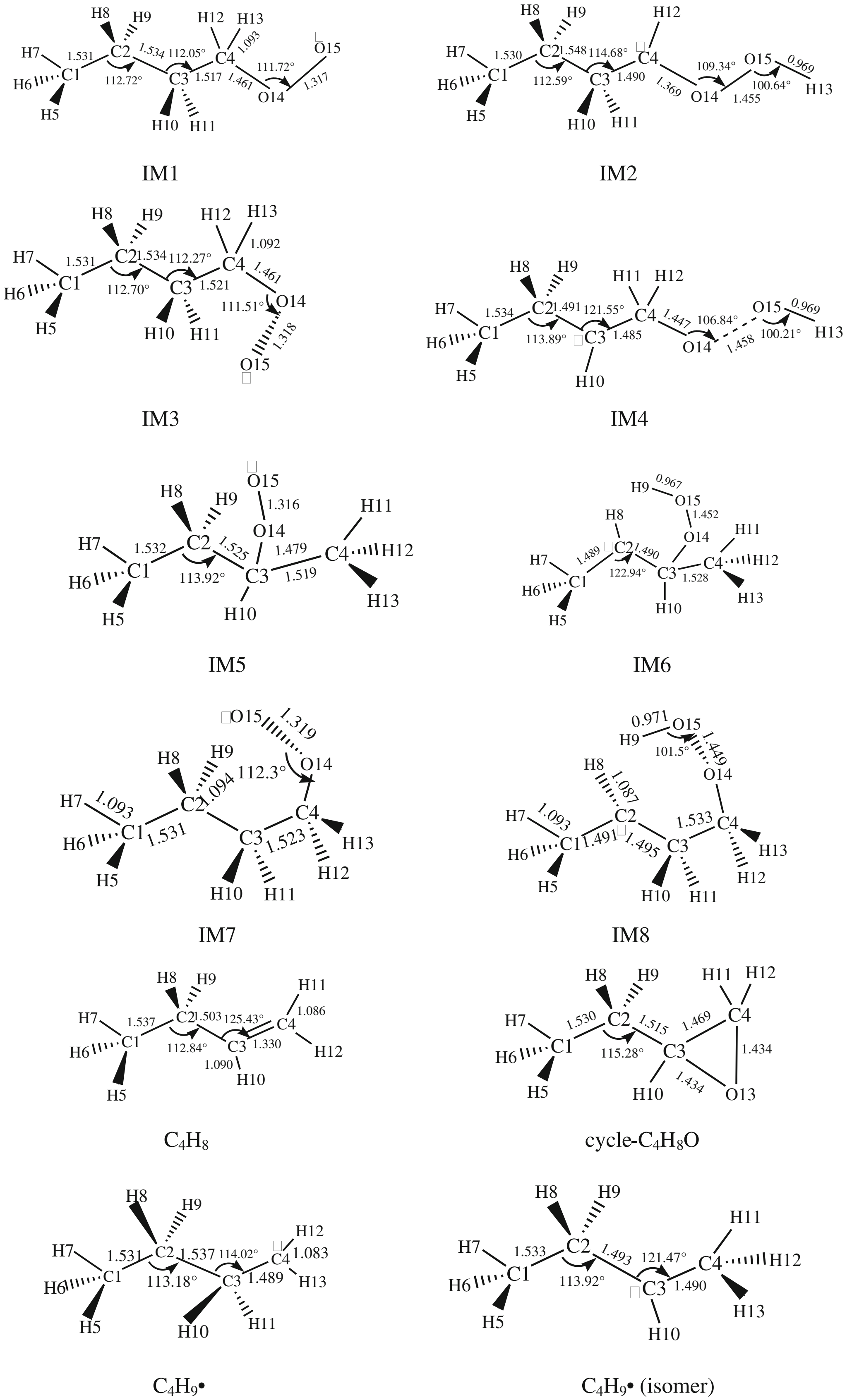
TS8

TS9

TS10

TS11

TS12


IM1
H8 H9 H12 H13 H7 H6 H5 C1 C2 C3 C4 O14 O15 H10 H11
1.531 1.534 112.05° 1.093 112.72° 1.517 1.461 111.72° 1.317
IM2
H8 H9 H12 H7 H6 H5 C1 C2 C3 C4 O14 O15 H13 H10 H11
1.530 1.548 114.68° 112.59° 1.490 1.369 109.34° 1.455 100.64° 0.969
IM3
H8 H9 H12 H13 H7 H6 H5 C1 C2 C3 C4 O14 O15 H10 H11
1.531 1.534 112.27° 1.092 1.461 112.70° 1.521 111.51° 1.318
IM4
H8 H9 H11 H12 H7 H6 H5 C1 C2 C3 C4 O14 O15 H13 H10
1.534 1.491 121.55° 1.447 106.84° 0.969 113.89° 1.485 1.458 100.21°
IM5
O15 1.316 H8 H9 O14 H11 H7 1.532 C2 1.525 1.479 C4 H12 H6 C1 113.92° C3 1.519 H5 H10 H13
IM6
H9 0.967 O15 1.452 H8 O14 H11 H7 1.489 C2 1.490 C4 H12 H6 C1 122.94° C3 1.528 H5 H10 H13
IM7
O15 1.319 H8 H9 112.3° O14 1.094 H7 1.093 C2 C4 H6 C1 1.531 C3 1.523 H13 H5 H10 H11 H12
IM8
0.971 O15 H9 101.5° 1.449 H8 1.087 O14 H7 1.093 C2 1.533 C4 H6 C1 1.491 1.495 C3 H13 H5 H10 H11 H12
C4H8
H8 H9 H11 H7 1.537 C2 1.503 125.43° C4 1.086 H6 C1 112.84° C3 1.330 H12 1.090 H5 H10
cycle-C4H8O
H8 H9 H11 H12 H7 1.530 C2 1.515 1.469 C4 H6 C1 115.28° C3 1.434 H5 H10 1.434 O13
C4H9•
H8 H9 H12 H7 1.531 C2 1.537 114.02° C4 1.083 H6 C1 113.18° C3 1.489 H13 H5 H10 H11
C4H9• (isomer)
H8 H9 H11 H7 1.533 C2 1.493 121.47° C4 H12 H6 C1 113.92° C3 1.490 H5 H10 H13

IM1 IM2

IM3 IM4

IM5 IM6

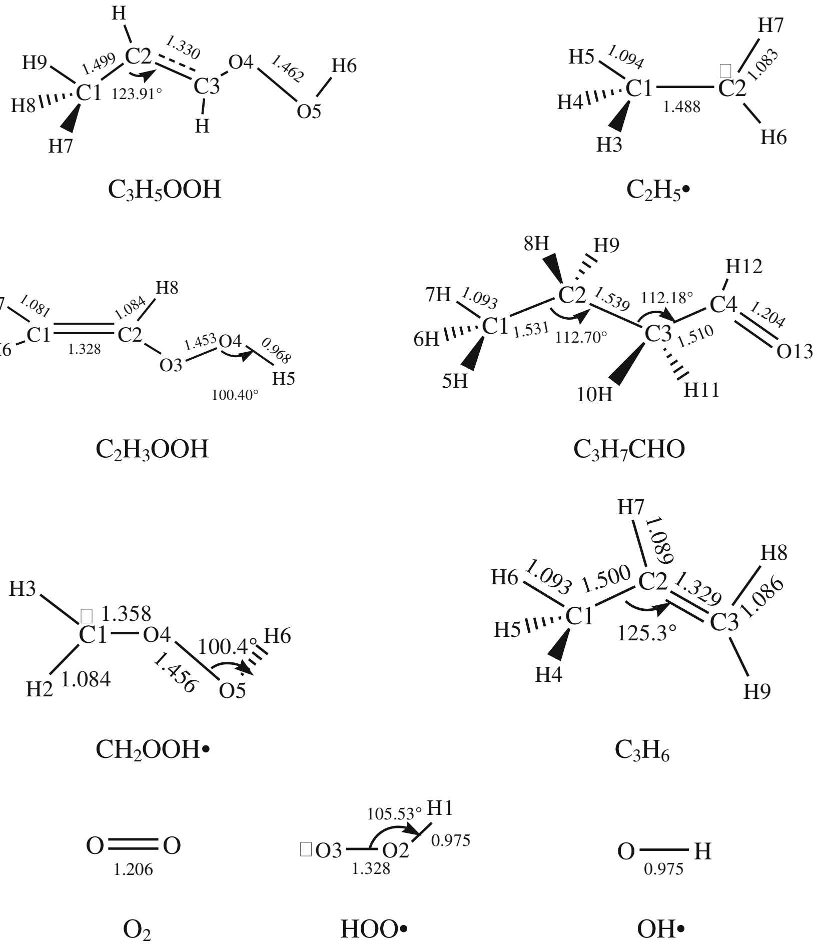
IM7 IM8

C_4H_8 cycle-C_4H_8O

C_4H_9• C_4H_9• (isomer)

Fig. 1 (continued)

Fig. 1 (continued)



C$_3$H$_5$OOH

C$_2$H$_5$•

C$_2$H$_3$OOH

C$_3$H$_7$CHO

CH$_2$OOH•

C$_3$H$_6$

O$_2$

HOO•

OH•

H9 H8 H7 C1 1.499 C2 H 1.330 123.91° C3 H O4 1.462 O5 H6
H5 1.094 H4 H3 C1 1.488 C2 H7 1.083 H6
H7 1.081 H6 C1 1.328 C2 1.084 H8 O3 1.453 O4 0.968 H5 100.40°
8H H9 7H 1.093 6H 5H C1 1.531 C2 1.539 112.70° C3 112.18° 1.510 10H H11 C4 H12 1.204 O13
H3 C1 1.358 O4 H2 1.084 1.456 O5 100.4° H6
H7 1.089 H6 1.093 H5 H4 C1 1.500 C2 1.329 125.3° C3 H8 1.086 H9
O O 1.206
O3 O2 1.328 105.53° H1 0.975
O H 0.975

Pathway D

R1→IM5→TS9→IM6→TS10→P5

In this pathway, the unpaired electron of C$_4$H$_9$• is located on the C3 atom, so when O$_2$ attacks C$_4$H$_9$•, the initial step is the addition of O$_2$ to C3 producing IM5, which is 19.0 kJ mol^{-1} lower than IM1, then H9 transfers to the neighboring O15 to form IM6 through a five-membered-ring transition state TS9 with a barrier of 135.1 kJ mol^{-1}, the length of C2-H9 changes from 1.093 Å in IM5 to 1.360 Å in TS9. H9-O15 in TS9 is 1.222 Å with a strained angle H9-O15-O14 of 93.9°, and in IM6, it is 0.967 Å and the unpaired electron is on the C2 atom. Afterward, C3-C4 of IM6 elongates from 1.528 Å to 2.298 Å resulting in CH$_3$+C$_3$H$_5$OOH (P5) through the transition state TS10 with a barrier of 129.6 kJ mol^{-1}. In TS10, C2-C3 displays a resonance feature with a length of 1.361 Å which is close to the bond length of C=C double bond. From IM5 to IM6 the reaction is overall endothermic, and the reaction enthalpy is 58.7 kJ mol^{-1}, and from IM6 to P5, it is 108.0 kJ mol^{-1}. The rate-determing step of the channel is IM5→TS9→IM6.

Pathway E

R→IM7→TS11→IM8→TS12→P6

The barrierless addition of O$_2$ to the terminal carbon of C$_4$H$_9$• can also produce the intermediate IM7, an isomer of IM1. The process has an exothermicity of −148.69 kJ mol^{-1}, which is basically equal to that of the process producing IM1 (−148.46 kJ mol^{-1}). Then H9 in IM7 migrates to O15 to give the intermediate IM8 by overcoming a barrier of 92.2 kJ mol^{-1} through the six-membered-ring transition state TS11. In this process, C2-H9 is elongated from 1.094 Å in IM7 to 1.376 Å in TS11, and H9-O15 is shortened from 1.177 Å in TS11 to 0.971 Å in IM8. The reaction process from IM7 to IM8 is endothermic (56.5 kJ mol^{-1}). IM8 can subsequently undergo a bond dissociation of C3-C4 via TS12 to form P6. In this process, C3-C4 changes from 1.533 Å in IM8 to 2.276 Å in TS12. The process from IM8 to P6 is also endothermic (92.9 kJ mol^{-1}). The rate-determing step of the channel is IM8→TS12→CH$_2$OOH•+C$_3$H$_6$ (P6).

In the five reaction pathways, six products (P1-P6) are produced and the corresponding reaction enthalpy are −67.47, 1.85, -250.81, -159.22, -3.44, and 0.78 kJ mol^{-1}

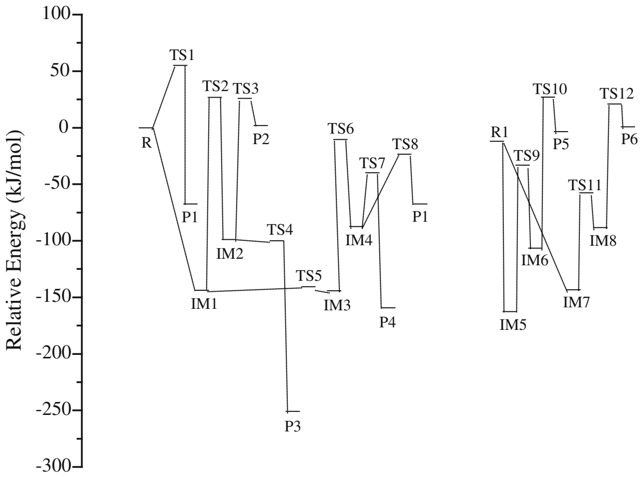


Fig. 2 A diagram of potential energy surface for various reaction pathways of C_4H_9•+O_2 at the CBS-QB3 level

respectively. In terms of the reaction exothermicity, the more favorable pathways are those resulting in the products P1, P3, and P4, which are all highly exothermic. However, the reactions leading to P1 and P3 have to pass through the transition states (TS1 and TS2 respectively) that are higher in energy than the reactants, that is to say, extra energy is needed to overcome the energy barrier to make these reactions happen, while for the reaction to give the product P4, all species involved in the pathway are lower in energy than the reactants, so no extra energy is demanded to make the reaction proceed. The reaction happens spontaneously. Therefore, the main pathway of the reaction is R→IM1→TS5→IM3→TS6→IM4→TS7→P4. The changes in free energy and enthalpy of the reaction are −149.62 and −159.22 kJ mol^{-1} respectively.

Table 1 Relative energies (in kJ mol^{-1}) of various species at the CBS-QB3 level

Species	Energy	Species	Energy
R (C_4H_9•+O_2)	0.00	IM2	−98.83
R1 (C_4H_9• isomer+O_2)	−11.98	IM3	−144.29
TS1	55.17	IM4	−87.43
TS2	26.74	IM5	−162.63
TS3	25.98	IM6	−106.61
TS4	−99.97	IM7	−143.27
TS5	−140.68	IM8	−88.26
TS6	−10.38	P1 (HOO•+C_4H_8)	−67.47
TS7	−39.74	P2 (C_2H_5•+C_2H_3OOH)	1.85
TS8	−23.50	P3 (OH•+C_3H_7CHO)	−250.81
TS9	−33.19	P4 (OH•+cycle-C_4H_8O)	−159.22
TS10	27.15	P5 (CH_3•+C_3H_5OOH)	−3.44
TS11	−57.51	P6 (CH_2OOH•+C_3H_6)	0.78
TS12	20.90	P6 (CH_2OOH•+C_3H_6)	0.78
IM1	−143.67		

Table 2 Relative energies (kJ mol^{-1}) of the species involved in the main reaction pathway at the CCSD(T)/cc-PVTZ and CBS-QB3 levels

Species	CCSD(T)/ cc-PVTZ	CBS-QB3
R	0	0
IM1	−151.01	−131.69
TS5	−148.04	−128.70
IM3	−152.87	−132.31
TS6	4.77	1.61
IM4	−83.08	−75.45
TS7	−17.83	−27.76
P4	−150.68	−147.24

To test the reliability of the calculation results obtained at the level of CBS-QB3, we have also calculated the energies at the higher theoretical level of CCSD(T)/cc-pVTZ for the main reaction pathway. A comparison of the results at the two levels is shown in Table 2 and Fig. 3. Obviously, there is a good linear correlation (R=0.994) between them and the differences between them are not that significant (standard deviation SD=7.16 kJ mol^{-1}), so the conclusions drawn at the two levels should be similar. Considering the computational expense and feasibility, we think the CBS-QB3 method is acceptable and the results obtained from it are reliable. In fact, this method has been used in many similar studies, e.g., in reference [34].

Conclusions

The ab initio CBS-QB3 method has been used to study the reaction mechanism of C_4H_9• radical with oxygen molecule. The activation energies and reaction enthalpies of five different reaction pathways are obtained and analyzed. Calculated results show that the main reaction pathway is C_4H_9•+O_2 (R)→IM1→TS5→IM3→TS6→IM4→TS7→OH•+cycle-C_4H_8O (P4).

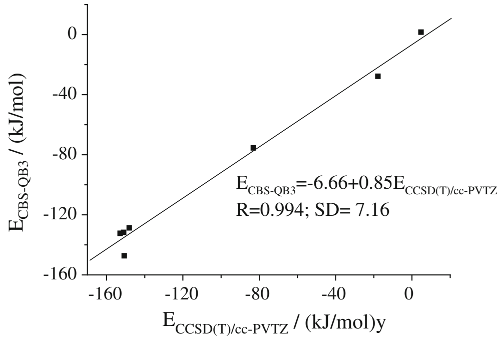


Fig. 3 Correlation between the relative energies of the species involved in the main reaction pathway at the CCSD(T)/cc-PVTZ and CBS-QB3 levels

Acknowledgments The project is supported by the Foundation of State Key Laboratory of Explosion Science and Technology of China (Grant No. KFJJ10-12 M) and NSAF (Grant No. 11076017).

References

1. Minkoff GJ, Tipper CFH (1962) Chemistry of combustion reactions. Butterworths, London
2. Gardiner WC (1984) Combustion chemistry. Springer, New York
3. Hucknall KJ (1985) Chemistry of Hydrocarbon Combustion. Chapman and Hall, New York
4. Pollard RT (1997) In: Bamford CH, Tipper CFH (eds) Comprehensive chemical kinetics. Elsevier, New York 17:249–367
5. McKay G (1977) Gas-phase oxidations of hydrocarbons. Prog Energy Combust Sci 3:105–126
6. Knox JH (1965) A new mechanism for the low temperature oxidation of hydrocarbons in the gas phase. Combust Flame 9:297–310
7. Benson SW (1976) Thermochemical Kinefics. John Wiley, New York
8. Fish A (1968) The Cool Flames of Hydrocarbons. Angew Chem Int Edn Engl 7:45–60
9. Baker RR, Baldwin RR, Walker RW (1975) Addition of n-butane to slowly reacting mixtures of hydrogen and oxygen at 480 °C. Part 2. Formation of oxygenated products. J Chem Soc Faraday Trans 1(71):756–779
10. Benson SW, Nangia PS (1979) Some unresolved problems in oxidation and combustion. Acc Chem Res 12:223–228
11. Baldwin RR, Bennett JP, Walker RW (1980) Addition of n-pentane to slowly reacting mixtures of hydrogen+oxygen at 480 ° C. J Chem Soc Faraday Trans 1(76):1075–1092
12. Lenhardt TM, McDade CE, Bayes KD (1980) Rates of reaction of butyl radicals with molecular oxygen. J Chem Phys 72:304–310
13. Ruiz RP, Bayes KD (1984) Rates of reaction of propyl radicals with molecular oxygen. J Phys Chem 88:2592–2595
14. Cox RA, Cole JA (1985) Chemical aspects of the autoignition of hydrocarbon—air mixtures. Combust Flame 60:109–123
15. Gulati SK, Mather S, Walker RW (1987) Arrhenius parameters for the addition of HO_2 radicals to pent-1-ene, hex-1-ene, and *cis*- and *trans*-hex-2-ene over the range 400–520 °C. J Chem Soc Faraday Trans 2(83):2171–2179
16. Gulati SK, Walker RW (1988) Arrhenius parameters for the reaction i-C_3H_7+ $O_2 \rightarrow C_3H_6$+ HO_2. J Chem Soc Faraday Trans 2 (84):401–407
17. Stothard ND, Walker RW (1990) Arrhenius parameters for the addition of HO_2 radicals to (*E*)-but-2-ene over the range 400–520 °C. J Chem Soc Faraday Trans 86:2115–2119
18. Kaiser RI, Sun W, Suits AG, Lee YT (1997) Crossed beam reaction of atomic carbon, C (^3P_j), with the propargyl radical, C_3H_3: Observation of diacetylene, C_4H_2. J Chem Phys 107:8713–8716
19. Yamaguchi M (1996) A CASSCF study of photochemical cyclization of the first excited Ã $^2B_1$ state of the allyl radical. J Mol Struct Theochem 365:143–149
20. Longuet-Higgins HC, Abrahamson EW (1965) The electronic mechanism of electrocyclic reactions. J Am Chem Soc 87:2045–2046
21. Schultz T, Fischer I (1997) The nonradiative decay of the allyl radical excited B 2A1 state studied by picosecond time-resolved photoelectron spectroscopy. J Chem Phys 107:8197–8200
22. Deyerl HJ, Fischer I, Chen P (1999) Photodissociation dynamics of the allyl radical. J Chem Phys 110:1450–1462
23. Stranges D, Stemmler M, Yang X, Chesko JD, Suits AG, Lee YT (1998) UV photodissociation dynamics of allyl radical by photofragment translational spectroscopy. J Chem Phys 109:5372–5382
24. Slagle IR, Bernhardt JR, Gutman D, Hanning-Lee MA, Pilling MJ (1990) Kinetics of the reaction between oxygen atoms and allyl radicals. J Phys Chem 94:3652–3656
25. Kaiser EW (1995) Temperature and pressure dependence of the C_2H_4 yield from the reaction C_2H_5+O_2. J Phys Chem 99:707–711
26. Bozzelli JW, Dean AM (1990) Chemical activation analysis of the reaction of ethyl radical with oxygen. J Phys Chem 94:3313–3317
27. Green WH (1994) Predictive chemical kinetics: Density functional and hartree–fock calculations on free-radial reaction transition states. Int J Quantum Chem 52:837–847
28. Ignatyev IS, Xie Y, Allen WD, Schaefer HF (1997) Mechanism of the C_2H_5+O_2 reaction. J Chem Phys 107:141–155
29. Quelch GE, Gallo MM, Schaefer HF III (1992) Aspects of the reaction mechanism of ethane combustion. Conformations of the ethylperoxy radical. J Am Chem Soc 114:8239–8247
30. Wagner AF, Slagle IR, Sarzynski D, Gutman D (1990) Experimental and theoretical studies of the ethyl+oxygen reaction kinetics. J Phys Chem 94:1853–1868
31. Atkinson DB, Hudgens JW (1997) Chemical kinetic studies using ultraviolet cavity ring-down spectroscopic detection: self-reaction of ethyl and ethylperoxy radicals and the reaction O_2+$C_2H_5 \rightarrow$ $C_2H_5O_2$. J Phys Chem A 101:3901–3909
32. Shen D, Moise A, Pritchard HO (1995) Theoretical calculation of intramolecular reactions in methylperoxyl and ethylperoxyl radicals. J Chem Soc, Faraday Trans 91:1425–1430
33. Skancke A, Skancke PN (1990) Ab initio studies of parts of the potential surface for the system C_2H_4+HO_2. J Mol Struct Theochem 207:201–215
34. Hans-Heinrich C, Chitralkumar VN, Anthony MD (2005) Detailed modeling of the reaction of C_2H_5+O_2. J Phys Chem A 109:2264–2281
35. Chad YS, Joseph WB, Anthony MD, Albert YC (2002) Detailed Kinetics and Thermochemistry of C_2H_5+O_2: Reaction kinetics of the chemically-activated and stabilized CH_3CH_2OO• Adduct. J Phys Chem A 106:7276–7293
36. Jonathan CR, Wesley DA, Henry FS (2000) The C_2H_5+O_2 reaction mechanism: high-level ab Initio characterizations. J Phys Chem A 104:9823–9840
37. Estupiñán EG, Klippenstein SJ, Taatjes CA (2005) Measurements and modeling of HO2 formation in the reactions of n-C_3H_7 and i-C_3H_7 radicals with O_2. J Phys Chem B 109:8374–8387
38. Wilke JJ, Allen WD, Schaefer HF III (2008) Establishment of the C_2H_5+O_2 reaction mechanism: A combustion archetype. J Chem Phys 128:074308(1–9)
39. Basevich VY, Belyaev AA, Frolov SM (2009) Mechanisms of the oxidation and combustion of normal alkanes: Passage from C_1-C_4 to C_2H_5. Russ Chem B 3:629–635
40. Basevich VY, Belyaev AA, Frolov SM (2010) Mechanisms of the oxidation and combustion of normal alkanes: Transition from C_1–C_5 to C_6H_{14}. Russ Chem B 4:634–640
41. Basevich VY, Belyaev AA, Posvyanskii VS, Frolov SM (2010) Mechanism of the oxidation and combustion of normal paraffin hydrocarbons: Transition from C_1–C_6 to C_7H_{16}. Russ Chem B 4:985–994
42. Montgomery JA, Frisch MJ, Ochterski JW, Petersson GA (1999) A complete basis set model chemistry VI. Use of density functional geometries and frequencies. J Chem Phys 110:2822–2827
43. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C,

Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniela AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Jonson B, Chen W, Wong MW, Gonzalez C, Pople JA (2003) Gaussian 03. Gaussian Inc, Pittsburgh, PA

44. Gonzalez C, Schlegel HB (1989) An improved algorithm for reaction path following. J Chem Phys 90:2154–2161
45. Baldwin RR, Pickering IA, Walker RW (1980) Reactions of ethyl radicals with oxygen over the temperature range 400–540 °C. J Chem Soc Faraday Trans 1(76):2374–2382

J Mol Model (2012) 18:2227–2240
DOI 10.1007/s00894-011-1227-9

ORIGINAL PAPER

Characteristic vibration patterns of odor compounds from bread-baking volatiles upon protein binding: density functional and ONIOM study and principal component analysis

Witcha Treesuwan · Hajime Hirao · Keiji Morokuma · Supa Hannongbua

Received: 3 May 2011 / Accepted: 16 August 2011 / Published online: 27 September 2011

Abstract As the mechanism underlying the sense of smell is unclear, different models have been used to rationalize structure–odor relationships. To gain insight into odorant molecules from bread baking, binding energies and vibration spectra in the gas phase and in the protein environment [7-transmembrane helices (7TMHs) of rhodopsin] were calculated using density functional theory [B3LYP/6-311++G(d,p)] and ONIOM [B3LYP/6-311++G(d,p):PM3] methods. It was found that acetaldehyde ("acid" category) binds strongly in the large cavity inside the receptor, whereas 2-ethyl-3-methylpyrazine ("roasted") binds weakly. Lys296, Tyr268, Thr118 and Ala117 were identified as key residues in the binding site. More emphasis was placed on how vibrational frequencies are shifted and intensities modified in the receptor protein environment. Principal component analysis (PCA) suggested that the frequency shifts of C–C stretching, CH_3 umbrella, C=O stretching and CH_3 stretching modes have a significant effect on odor quality. In fact, the frequency shifts of the C–C stretching and C=O stretching modes, as well as CH_3 umbrella and CH_3 symmetric stretching modes, exhibit different behaviors in the PCA loadings plot. A large frequency shift in the CH_3 symmetric stretching mode is associated with the sweet-roasted odor category and separates this from the acid odor category. A large frequency shift of the C–C stretching mode describes the roasted and oily-popcorn odor categories, and separates these from the buttery and acid odor categories.

Keywords Odorant category · Quantum chemical calculation · Molecular vibration · Binding energy · Principal component analysis

Introduction

Studies on the recognition and classification of odors have expanded markedly, driven by efforts to develop a new multidisciplinary discipline: molecular gastronomy [1]. Perception is an auto-bio-activation process acting through multi-stimulation G-protein coupled receptors (GPCRs). The GPCR mechanism is divided into four classes: A, B, C and F/S, with rhodopsin, rhodopsin-like and olfactory

W. Treesuwan · S. Hannongbua
Department of Chemistry, Faculty of Science,
Kasetsart University,
Bangkok, Thailand

W. Treesuwan · S. Hannongbua (✉)
Center of Nanotechnology, Kasetsart University,
Bangkok, Thailand
e-mail: fscisph@ku.ac.th

W. Treesuwan
Institute of Food Research and Product Development,
Kasetsart University,
Bangkok, Thailand

H. Hirao · K. Morokuma
Fukui Institute for Fundamental Chemistry, Kyoto University,
Kyoto 606-8103, Japan

H. Hirao
Division of Chemistry and Biological Chemistry,
School of Physical and Mathematical Sciences,
Nanyang Technological University,
21 Nanyang Link,
Singapore 637371, Singapore

receptors (OR) all being classified within class A [2]. Perception of the characteristic flavors of foods is mediated by small molecules binding noncovalently to the OR [3]. The process of smell perception is initiated when an odorant is captured in the upper nasal cavity where the nasal mucous membrane is located [4]. The aqueous olfactory mucus adsorbs hydrophobic odorants with the help of soluble proteins—the so-called odorant binding proteins (OBPs) [5]. The odorant binding site is assumed to be situated at the upper part of the molecule toward the N-terminal domain of the seven transmembrane helices (7TMHs) [6]. Upon binding of an odorant, the receptor is activated and this then causes a conformational change. In the inactive form, receptors are coupled with G-proteins composed of α, β and γ subunits, and guanine diphosphate (GDP) binds to the α subunit. The conformational change increases the affinity of the receptor for the α subunit and GDP is released. As a result, guanine triphosphate (GTP) binds to the α subunit, which subsequently dissociates from the $\beta\gamma$ subunits, activating downstream effectors for signal transduction [7]. A few experiments have indicated that an odorant compound is likely to give rise to multiple active conformations of the receptor, in either a sequential or a parallel manner [8, 9].

A better understanding of odor properties will help applications in biosensor design, perfumery and molecular gastronomy. At present, the detailed mechanisms of odorant–receptor interactions and subsequent activation are still unclear. It is important to understand the mechanisms underlying odor perception at the molecular level. Two basic features of odor recognition—shape and vibration—have so far been proposed to account for the olfaction mechanism and the process of signaling from odors based on molecular properties. Dyson proposed “vibration theory”, which links molecular vibrations and odor [10]. Vibration theory assumes that ORs recognize the unique vibrations of odorant molecules, which are subsequently translated into odors in the brain. Recently, vibration theory was revived by Turin, who carried out frequency calculations at the HF/3-21G* level of theory to distinguish several odor categories, namely, bitter almonds, musks, ambers, woods, sandalwoods and violets [11]. Whilst protein activation starts from ligand binding as the mechanism of basic biomolecular recognition, the origin of odorant triggering may be the correlated vibrations between the odorant and the amino residues of the receptor. The technical limitations of observing molecular vibration mean that no ligand–protein spectroscopy is available [12]. Brookes and co-workers [13] reported that only the specific conformation with the right vibrational frequency of odorant would satisfy OR recognition. Although important insights were derived from Turin’s study, there remains much to be done theoretically to fully understand odor chemistry. For example, higher-level calculations should be performed on models in the presence of receptors to identify more accurately the characteristic vibrations of odorants in their binding sites. Since molecular vibrations will change depending on the surrounding environment, such as a solvent or inside a protein, it is important to find out how vibrations are enhanced or attenuated in such situations. This type of information can be derived from hybrid quantum mechanical/molecular mechanical (QM/MM) calculations [14].

Theoretical calculations have been shown to provide useful insight into the physical basis of ligand binding systems. For example, the strengths of hydrogen bonds in transmembrane proteins were evaluated successfully at the MPWB1K/6-31+G(d,p) level [15]. Active conformations of Maillard products and odorants such as diacetylformoin were identified by calculations at the B3LYP/6-31++G (3df,2pd) level [16]. ONIOM (B3LYP/6-31G(d,p):PM3) calculations showed that the conformation and energy of TIBO (tetrahydro-imidazo[4,5,1-jk][1,4]-benzodiazepin-2-one) inside human immunodeficiency virus type-1 reverse transcriptase (HIV-1 RT) are different from those in vacuo [17, 18]. Such results indicate that the effect of the protein environment should be taken into account for accurate descriptions of biomolecular systems.

In this work, volatile compounds from bread baking were selected as the target odorant class because of their crucial importance in the food processing and bakery industries. We employed DFT and the ONIOM method [19] in an attempt to understand the properties of odorant molecules from bread baking products [20]. The vibrational frequencies of odorants were determined in the gas phase and in the protein complex. We identified a key vibration pattern characteristic of each odor category. Our observations provide useful structural and energetic information for the design of molecular vibrational experiments.

Computational methods

Calculations on aroma compounds in gas phase

Aroma compounds (I–VII) produced during bread baking were detected by an electronic nose [20]. A series of molecules in this class (Fig. 1) were chosen for our theoretical analyses. Conformational analysis was carried out using the AM1 method to find out the lowest energy conformations of each compound. For the lowest energy conformations obtained, full geometry optimization at the B3LYP/6-311++G(d,p) level was performed, followed by vibrational frequency and IR intensity analyses. All calculations were done using the Gaussian03 program [21].

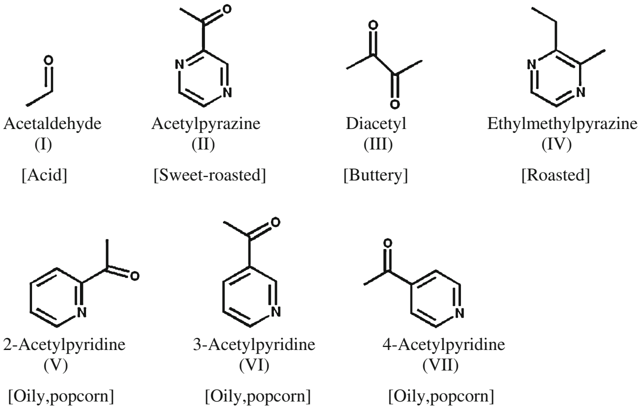


Fig. 1 Bread baking aroma compounds studied (*I–VII*)

Setup of ligand-receptor complexes

A major difficulty in modeling olfactory receptors is that no crystal structure has been solved to date. Crystallographic structures containing complete 7TMHs have been obtained only for bovine rhodopsin (pdb codes 1F88, 1HZX and 1L9H). Although most GPCR family A receptors share a similar fold of the 7TMH domain [3], the percent similarity between human olfactory receptor (hOR) from OR1D2 gene and the bovine rhodopsin is very low (~17%), and therefore we decided not to perform homology modeling for hOR but to directly use the geometry from 1L9H, which has the highest resolution of all available geometries.

Orientations and conformations of odorants in 7TMHs were obtained by a docking method using Gold 3.0 [22]. Protein preparation involved addition of hydrogen atoms using the Biopolymer module as implemented in the Sybyl6.9 program. In preparing ligand coordinates for docking, fully optimized gas-phase geometries described earlier were utilized. The orthosteric site located towards the N-terminus of 7TMHs was assumed to be the target for docking. Docking was performed using default settings. For each compound, the docked conditions yielding the three highest GoldScore values were selected for further analysis.

Vibration calculations

Gas phase calculations

Reasonable IR vibration values for organic molecule can be obtained at the B3LYP/6-31+G(d,p) [23], B3LYP/6-31++G (d,p) [24] and B3LYP/6-311++G(d,p) [25] levels of theory. The advantage of B3LYP/6-311++G(d,p) is that it accurately predicts not only IR properties but also NMR properties, as documented by Vailikhit [26]. Therefore, for isolated aroma compounds in gas phase, fully optimized structures as described above were subjected to vibration calculations at the B3LYP/6-311++G(d,p) level of theory.

Ligand/protein complex

All initial complex structures were at first relaxed to remove bad contacts. The whole complex model was then minimized in vacuo by molecular mechanics with the Duan force field [27] using the AMBER program [28]. Subsequently, the size of the model was reduced by selecting all residues within 7Å of odorants. To maintain the same environment for all odorants, all complex structures were aligned before reducing the size of model; thus, all complex models have the same residues. Note, the protein geometries are slightly different depending on the binding ligand, since the preparatory energy minimizations (see above) for the entire enzyme were carried out individually with the corresponding bound ligand. This smaller model contained 703 atoms from 7TMHs and was optimized by the PM3 method [29] under the condition of fixed protein backbone atoms (except peptide hydrogen and oxygen atoms).

A two-layered ONIOM2 method was applied to the protein cluster model to obtain the vibrational frequencies of the odorants in the protein. The complex models were then optimized at the ONIOM(B3LYP/6-311++G(d,p):PM3) level with backbone atoms (Cα, N, carbonyl carbon) fixed. Only the ligand was treated at the higher level of theory while surrounding residues were treated by low level calculations using the PM3 method. All protein residues were fixed in the frequency calculation using the ONIOM2 approach (B3LYP/6-311++G(d,p):PM3). The relative error as produced from constrained calculations was generally small and did not affect the results, which makes this a useful approach. The frequencies obtained were checked and it was verified that all frequencies were positive to ensure the presence of the actual configuration, not a transition configuration from the optimization. It should be mentioned here that frequency calculations on the entire system should be done ideally without any geometry constraints. However, performing the calculations with a flexible protein will give a large number of vibrational modes that may well be delocalized over the entire system, and this prevents a clear understanding of the effect of the protein on the vibration of a ligand. Therefore, the complex geometries were minimized at these theoretical levels with protein atoms fixed.

The calculated and experimental spectra and vibrational modes of isolated ligands were compared to evaluate the accuracy of the B3LYP/6-311++G(d,p) calculations. In addition, comparison of the calculated IR spectra in the gas phase with those in the binding site allowed us to investigate the effect of the receptor on molecular vibrations.

Other active volatiles

To understand the characteristic trends of a variety of molecules, the vibrational frequencies of other odorants were also examined. The odorant compounds examined are listed in Table 1. Two or three compounds from each category were selected as representative cases. Frequency calculations were performed for these compounds with the same protocol as described above to obtain frequencies in isolate and bound states.

Results and discussion

Binding of odorants in the receptor site

The features of compound binding with 7TMHs will now be discussed. The volumes of pocket cavities in the protein and the molecular surface area of the odorants were analyzed (Fig. 2). The volumes of pocket cavities in the upper part of 7TMHs were calculated using Swiss- PDB Viewer (SPDBV4.0) [30]. In the present study, the smallest

Table 1 Odorant molecules from different odor categories. Compounds used for frequency calculations are underlined

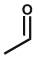
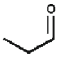
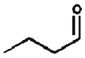
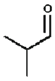
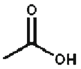
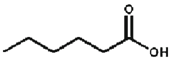
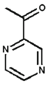
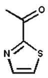
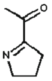
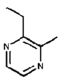
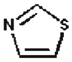
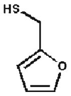
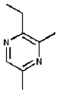
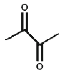
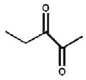
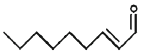
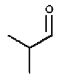
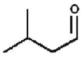
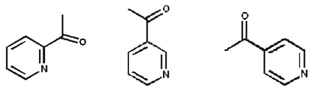
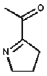
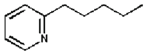
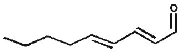
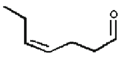
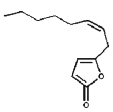
Compounds		Odor-quality	References
		[Acid]	
Acetaldehyde		Acid, Pungent	[20, 34]
Propanal		Pungent	[34]
Butanal		Green, pungent	[34]
2-Methylpropanal		Pungent, malty, green	[34]
Acetic acid		Sour, pungent, Rancid	[35, 36]
Hexanoic acid		Pungent, musty	[35]
		[Sweet roasted]	
Acetylpyrazine		Sweet roasted, Roasty	[20, 37]
		Hazelnut, praline, cake	[38]
2-Acetyl-2-thiazole		Roasty	[39]
2-Acetyl-1-pyrroline		Roasty, Roasty-popcorn-like	[35, 37]
		[Roasted]	
2-Ethyl-3-methylpyrazine		Roasted	[20]
Thiazole		Roast, cracker	[36]
2-Furanmethanethiol		Roast, meat	[36]

Table 1 (continued)

		Roasted, coffee-like	[37]
		Roasted	[39]
		Smoke, roasted	[39]
2-Ethyl-3,5-dimethylpyrazine		Potato-like, roasty	[37]
		Roasted	[39]
		[Buttery]	
Diacetyl		Buttery	[20, 34-37, 39]
2,3-Pentadione		Buttery	[34, 36, 39]
(E)-2-Nonenal		Buttery, oily	[39]
2-Methylbutanal		Buttery, oily	[39]
3-Methylbutanal		Buttery, oily	[39]
		[Oily, Popcorn]	
Acetylpyridine		Oily, Popcorn	[20]
			
2-Acetyl-1-pyrroline		Roasty, Roasty-popcorn-like	[35, 37]
2-Pentylpyridine		Fatty, tallowy	[37]
(E,E)-2,4-Nonadienal		Fatty, waxy	[37]
		Fatty, floral	[39]
(Z)-4-Heptenal		Fatty, oily, creamy	[39]
6-Dodecen-γ-lactone		Fatty	[39]

compound is acetaldehyde and the largest is 2-ethyl-3-methylpyrazine, and the molecular surface volumes ranged from 85.61 to 216.58 Å^3 as shown in Fig. 2b. By comparison, the binding pocket cavities in 7TMHs can have volumes as large as 602 Å^3. The comparison clearly shows that the molecular surface volume of odorants is much smaller than that of the binding pocket (cavity A in Fig. 2a). This confirms that the odorant molecules are not fit tightly inside the binding pocket, unlike enzyme–substrate binding, which uses "lock and key" and "induced fit" mechanisms. Our results also support the experimental findings of Triller and co-workers [31].

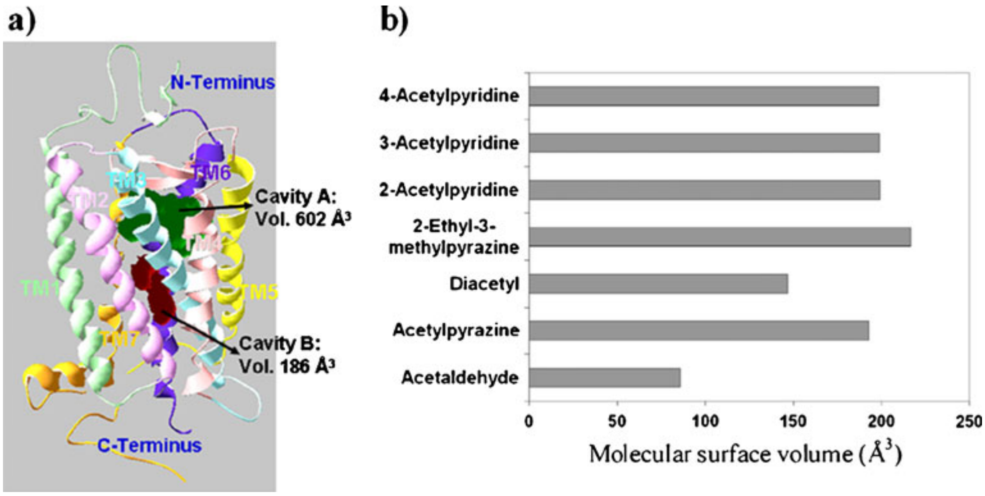


Fig. 2 **a** Binding pocket cavity (*green*) in the 7-transmembrane helices (7TMHs; Å), and **b** molecular surface volume of odorant molecules (Å)

Figure 3 compares the orientations of odorants obtained from our docking simulations. Acetaldehyde, acetylpyrazine and 2-ethyl-3-methylpyrazine bind rigidly inside the pocket, while diacetyl and acetylpyridine bind loosely with some degree of flexibility. This indicates that compounds with oily odor quality, such as diacetyl (buttery) and acetylpyridine (oily, popcorn), could possibly bind in a number of conformations, while odor compounds in other categories may have restricted conformations. Ala117, Thr118, Tyr268 and Lys296 have been found to play key roles in interactions inside the binding pocket. These residues form hydrogen bonding, electrostatic, dipole–dipole and/or hydrogen–π interactions with odorant compounds. The carbonyl groups of acetaldehyde and acetylpyrazine form a hydrogen bond with Lys296 with an O···H distance of 1.8Å ,while the carbonyl group of diacetyl and acetylpyridine do not. A flip of C=O group in acetylpyrazine results in a conformation (m2) devoid of a hydrogen bond interaction with Lys296, and this has significantly weaker binding energy (−4.99 kcal mol^{-1}).

The binding energy (BE) of odorants in the protein model was evaluated at the ONIOM[B3LYP/6-311++G(d, p):PM3] level from the energy of the optimized structures of the protein–odorant complex and the free protein, with the protein backbone fixed (see Computational methods section), and free odorant states, according to Eq. 1

$$\mathrm{BE}^{\mathrm{ONIOM}} = E^{ONIOM}_{Complex} - E^{PM3}_{Pocket} - E^{B3LYP}_{Ligand} \qquad (1)$$

Where $E^{ONIOM}_{Complex}$is the energy of complex structure, as calculated by the ONIOM method. E^{PM3}_{Pocket} and $E^{B3LYP}_{Ligand}$are the energies of the binding pocket and odor-active ligand, respectively. Their binding energies without entropy and solvation effect varied from −4.20 to −17.50 kcal mol^{-1}

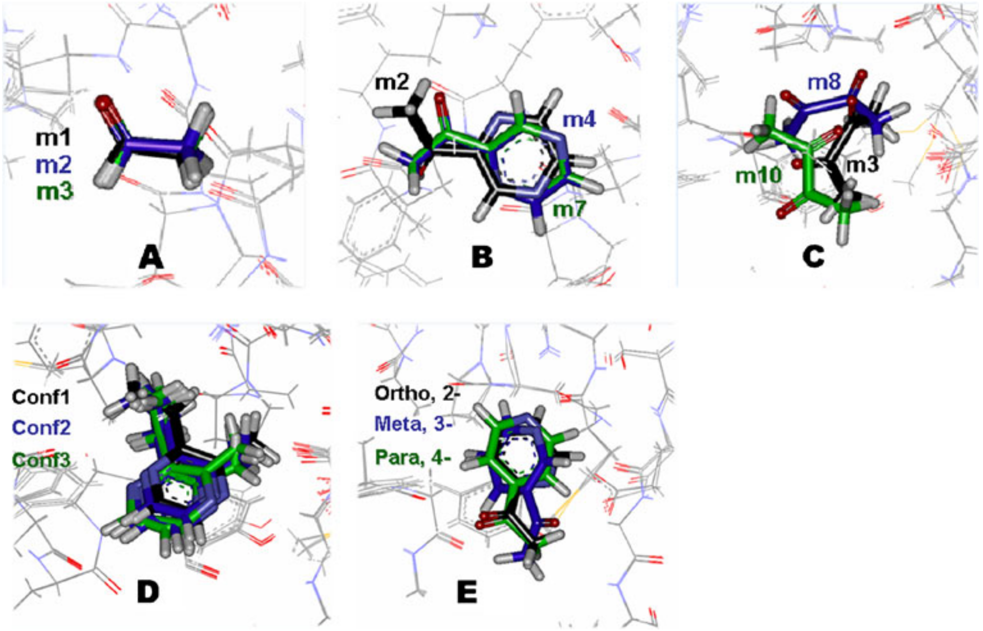


Fig. 3a–e Possible binding conformations of odorant molecules and their alignment inside the binding pocket of 7TMHs, after ONIOM[B3LYP/6-311++G(d,p):PM3] optimization with all protein atoms fixed. Distinct docking conformations are indicated by black, blue and green colors. **a** Acetaldehyde, **b** acetylpyrazine, **c** diacetyl, **d** 2-ethyl-3-methylpyrazine (including all conformations), and **e** acetylpyridine (including all isomers)

which are comparable to the values observed for several drug-enzyme complex systems [32, 33]. The conformation with the strongest binding energy was selected as a representative state for each odor category. As seen in Fig. 4, the binding strength was ranked in the following order: acetaldehyde~acetylpyrazine>diacetyl>acetylpyridine>2-ethyl-3-methylpyrazine. These are well accounted for hydrogen bond formation with Lys296; acetaldehyde and acetylpyrazine form strong hydrogen bonds, while 2-ethyl-3-methylyrazine does not. A clearer understanding of the binding process could be gained by calculation of the binding free energy, which is beyond the scope of this study.

Vibration spectra and characteristic spectrum patterns of odorants in the isolated state

The calculated IR vibration spectra (convolved with a Gaussian function of a width of 100 cm^{-1}) are compared in Fig. 5 for two odorants, with the experimental IR spectra taken from the National Institute of Standards and Technology (NIST) [40]. Although the peaks obtained theoretically are shifted slightly compared with experimental values, their spectra are similar. For example, in Fig. 5a and b, the experimental spectrum pattern of acetylpyrazine shows a moderate peak at ~3,100 cm^{-1}, a strong peak at ~1,700 cm^{-1}, and a group of moderate peaks at ~700–1,500 cm^{-1}. The calculated spectrum shows the same pattern of a moderate peak at 3,168 cm^{-1}, a strong peak at 1,762 cm^{-1}, and a group of moderate peaks at 864–1,462 cm^{-1}. This demonstrates that the vibrational spectra calculated at the B3LYP/6-311++G(d,p) level are quite reliable.

We calculated the gas phase vibrational spectra of acetaldehyde, 2-ethyl-3-methylpyrazine, acetylpyrazine, diacetyl, and acetylpyridine (including all isomers of 2-, 3- and 4-acetylpyridine) as representatives of their different classes; acid (pungent), roasted, sweet-roasted, buttery, and oily-popcorn categories, respectively. The overlayed spectra of odor qualities are shown in Fig. 6. A characteristic pattern for each odor quality emerged, as follows:

(1) Acid (pungent) odorants typically have a sharp and strong peak for a C=O stretching mode at ~1,700 cm^{-1}, a sharp peak of O–H stretching at 2,500–3,000 cm^{-1}, and a few smaller peaks in the 500–1,500 cm^{-1} region.
(2) Roasted odorants exhibit diverse peaks with weak and medium intensity from C–H bending modes in the 500–1,500 cm^{-1} region.
(3) Sweet-roasted odor-quality is associated with a spectrum pattern containing a group of C–H bending modes with medium to high intensity at 500–1,500 cm^{-1}, and a strong peak for a C=O stretching mode at ~1,750 cm^{-1}.
(4) Buttery odorants typically have a sharp and strong peak for a C=O stretching mode at ~1,700 cm^{-1}, small peaks at 500–1,500 cm^{-1}, and a weak and small peak of C–H stretching, or two at 3,000–3,200 cm^{-1}.
(5) Oily, popcorn odor quality is associated with the spectrum pattern of a group of dense peaks with medium to high intensity at 500–1,500 cm^{-1}, a prominently strong peak for a C–H bending mode at 1,270 cm^{-1}, a strong peak of C=O stretching at ~1,700 cm^{-1}.

These results indicate that odorant compounds in different odor categories are distinguishable based on their vibrational spectra. Although our finding supports the conclusion drawn by Turin [11], the effect of an olfactory receptor on vibrations should be investigated. Therefore, vibrational frequencies inside previously described 7TMHs protein models will be discussed in the next section.

Vibration spectra of odorants in the complex state

The spectra derived from odorants inside 7TMHs proteins obtained by the ONIOM(B3LYP/6-311++G(d,p):PM3)

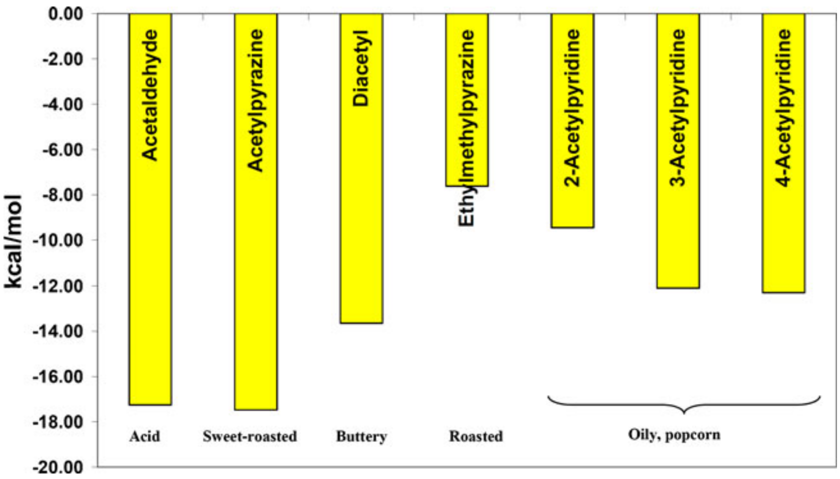


Fig. 4 Binding energies (kcal mol^{-1}) from the ONIOM [B3LYP/6-311++G(d,p):PM3] calculations

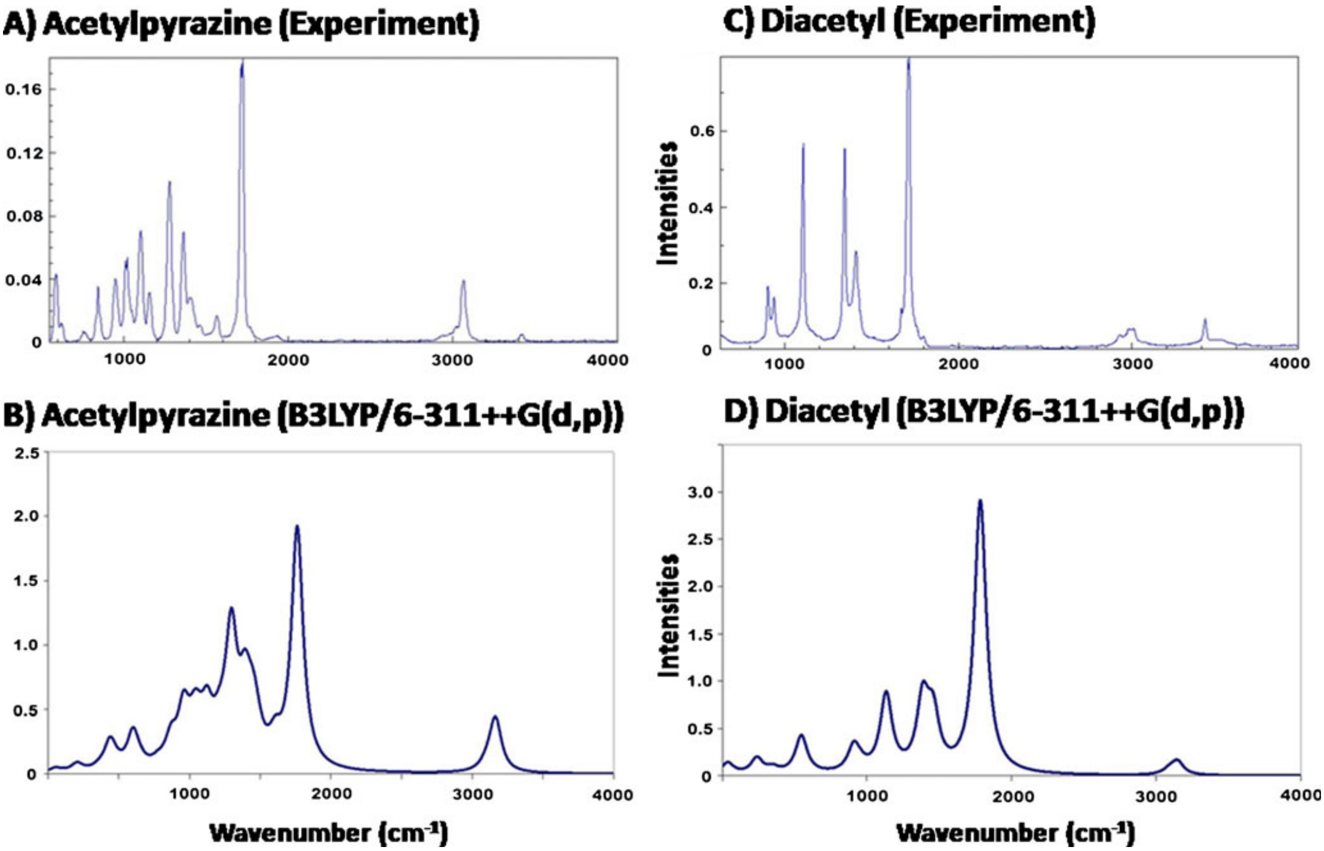


Fig. 5 Infrared (IR) spectra obtained from experiment and calculation at B3LYP/6-311++G(d,p) level for acetylpyrazine (**a**, **b**) and dicetyl (**c**, **d**)

method are shown in Fig. 7 along with patterns in the gas phase in black. The differences in the vibrational frequencies of odorants inside 7TMHs and in the isolated state were relatively small. However, changes in the peak intensities were sometimes large. New vibrations with low frequencies (100–300 cm^{-1}) emerged in the complex state for all odorants, but they are associated with translation and rotation modes of odorants trapped in the protein.

The key frequencies, magnitude and directions of shifts and changes in their intensities for each odor category are summarized in Table 2. In the odorant/protein complex state, the νC=O stretching mode at 1,700–1,800 cm^{-1} is always shifted to lower energy (red shift), while the intensities of many of the vibrational modes increase, especially in the range of 1,000–1,700 cm^{-1}. Characteristic changes in intensities are also found for each odor category. The intensity of the σC–H (H–C$_{ring}$–C$_{ring}$–H) bending mode is increased by 0.21 km mol^{-1} in the roasted odor category while it decreased by 0.24 km mol^{-1} in the oily-popcorn odor category. The spectrum differences observed for compounds in the roasted odor category, which are devoid of νC=O, were small compared to those of compounds in other odor categories. Higher intensity and shifted frequency of the vibration mode in the complex results from interactions of the ligands with residues of the membrane protein such as Lys296, Tyr268, Glu182, Thr119, Thr118 and Ala117. For example, acetic acid in the complex state shows significant changes in frequency and intensity of the O–H stretching mode at ~3,450 cm^{-1} (red shift by 320 cm^{-1} and intensity changes by 11.061 km mol^{-1}) because it stretches toward the carboxylate oxygen of Glu182, with which the ligand forms a hydrogen bond with an O···H distance of 1.56 Å. The same phenomenon is observed for 2-acetyl-1-pyrroline, where frequency shifts and higher intensities are found at ~1,700 and 2,900 cm^{-1}, as the carbonyl oxygen of 2-acetyl-1-pyrroline stretches toward the hydroxyl group of Thr119, which forms a hydrogen bond with an O···H distance of about 1.83 Å. This leads to the frequency shift of the C=O stretching mode at ~1,700 cm^{-1}, while the shift of the C–H stretching mode at ~2,900 cm^{-1} was induced by the hydrogen bond with the C=O group.

In the previous section we have seen that odorant molecules in different odor category could be distinguished by their spectrum patterns in the isolated state. Since the odor is actually sensed when the ligand is bound to the receptor, spectrum patterns in the complex state may serve as better fingerprints of odor categories. However, we did not observe major differences between the spectrum patterns in the complex state and those in the isolated state.

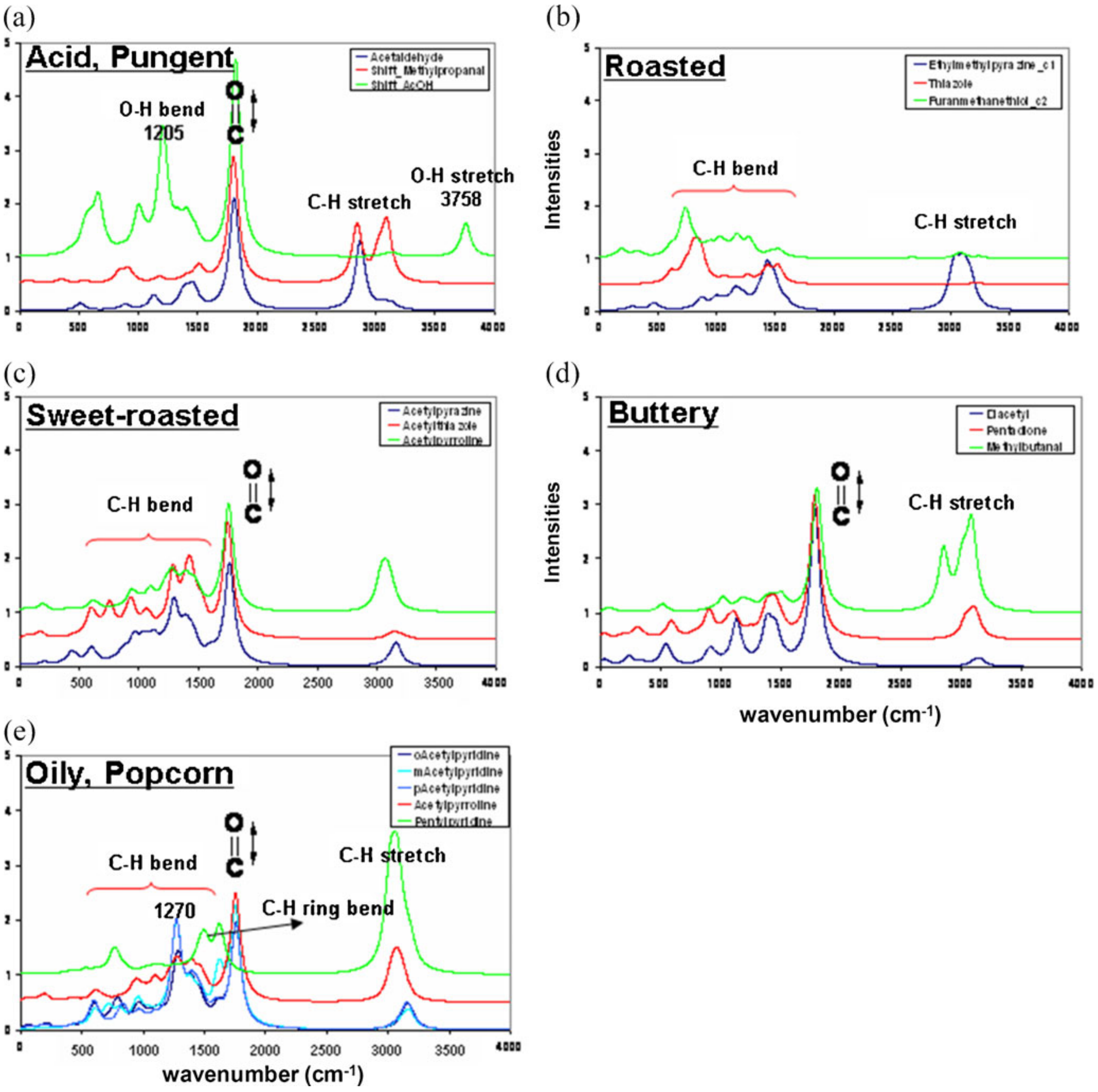
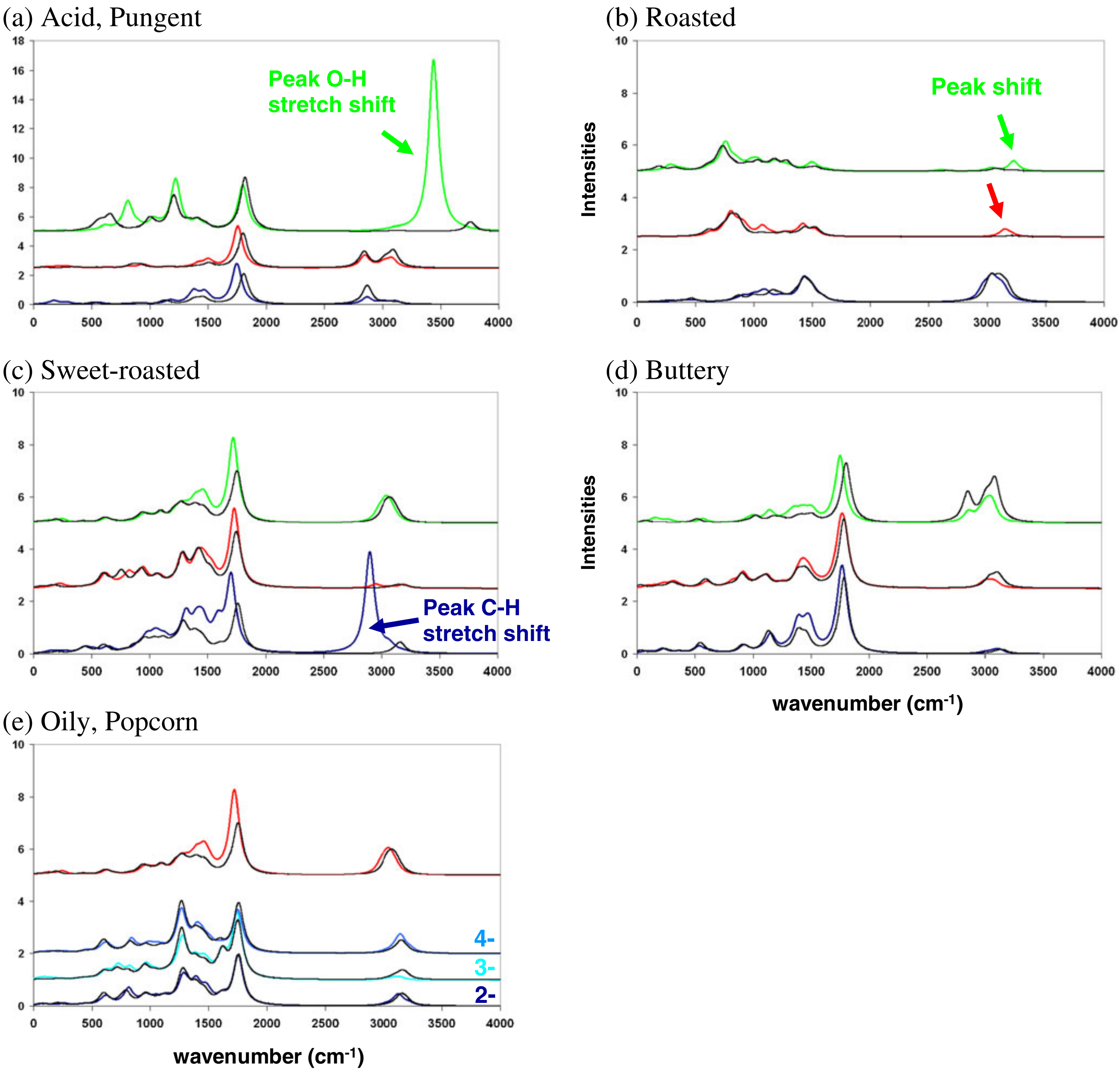


Fig. 6a–e Overlay of the calculated IR spectrum at B3LYP/6-311++G (d,p) from odorants in isolation state in the same odor-quality. **a** Acid, pungent (acetaldehyde, 2-methylpropanal and acetic acid in *blue*, *red* and *green*, respectively); **b** roasted (2-ethyl-3-methylpyrazine, thiazole and 2-furanmethanethiol in *blue*, *red* and *green*, respectively); **c** sweet-roasted (acetylpyrazine, 2-acetylthiazole and 2-acetyl-1-pyrroline in *blue*, *red* and *green*, respectively); **d** buttery (diacetyl, 2,3-pentadione and 3-methylbutanal in *blue*, *red* and *green*, respectively); and **e** oily, popcorn (acetylpyridine (all isomers), 2-acetyl-1-pyrroline and 2-pentylpyridine in *blue*, *red* and *green*, respectively)

Therefore, we conclude that the spectrum patterns in the isolated state can be used to distinguish between different odor categories. Further studies of the frequency shifts and intensity changes triggered by odorant/receptor interactions are needed to clarify the role played by the receptor in odor recognition.

Principal component analysis

The IR frequency calculations clearly demonstrate that vibration frequencies, generated from vibration modes of odorants, change when these compounds bind to the protein membrane. Low-frequency vibration modes such as the



(a) Acid, Pungent

(b) Roasted

(c) Sweet-roasted

(d) Buttery

(e) Oily, Popcorn

Fig. 7 Overlay of the calculated IR spectrum at ONIOM[B3LYP/6-311++G(d,p):PM3] from odorants in the complex state . See Fig. 6 for details. *Black dashed line* Calculated spectrum in the isolation state. Note that 2-pentylpyridine is omitted in the presentation for oily, popcorn odor-quality because it binds in a manner different from those of other compounds

Table 2 Frequencies (cm^{-1}) and values in parenthesis represent frequency shifts (cm^{-1}) and changes in intensity (km mol^{-1}) in the protein environment obtained in each odor category

Odor category	Frequencies (frequency shift, intensity change)
Acid (pungent)	1,757–1,800$^{\nu C=O}$ (−42, 0)
Roasted	754–1,137$^{\sigma C\text{-}H}$ (0, +0.21), 3,050–3,242$^{\nu C\text{-}H}$ (−43, 0)
Sweet-roasted	1,300–1,730 (0, +0.88), 1,709–1,738$^{\nu C=O}$ (−38, 0), 2,903–3,054$^{\nu C\text{-}H}$ (−161, 0)
Buttery	1,400–1,764 (0, +0.46), 1,754–1,764$^{\nu C=O}$ (−27, 0), 3,020–3,039 (−55, 0), 3,139–3,288 (0, −0.11)
Oily, popcorn	245–262 (+53, 0), 1,270–1,289 $^{\sigma C\text{-}H}$ (0, −0.24)

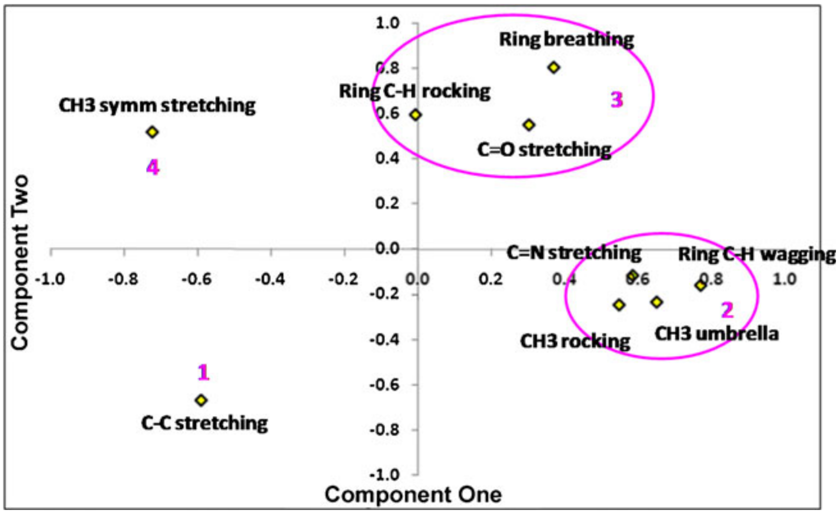


Fig. 8 Principal component analysis (PCA) loadings plot from components one and two

CH_3 rocking, ring C–H wagging and ring breathing modes are shifted consistently to higher frequencies (higher energy). In contrast, high-frequency vibration modes such as the CH_3 stretching and C=O stretching modes are shifted consistently to lower frequencies (lower energy). Vibration frequencies in the middle range ~750–1,600 cm^{-1} underwent either blue shift or red shift. To better understand the complex patterns of the vibration shifts, further analysis was carried out based on the statistical method of principal component analysis (PCA).

PCA allows extraction of key representations from a large data set containing several variables. On the vibration studied here, our PCA is concerned with the frequency shifts of vibrational modes as the input variables. In the calculations obtained, a vibration spectrum is composed of a multitude of vibration modes; however, only nine vibration-shift descriptors are selected as the common vibration modes in our compound series. The PCA model derived from the frequency shifts of the CH_3 rocking, ring C–H wagging, ring breathing, C–C stretching, C=N stretching, ring C–H rocking, CH_3 umbrella, C=O stretching and CH_3 symmetric stretching modes account for approximately 55% of the total variance in the descriptors. The relationship between the frequency shifts and odor categories was analyzed by loadings and scores plots of PCA.

The PCA loading plot indicates the correlations between descriptors, which are the vibration modes, and is shown in Fig. 8. Descriptors are interpreted as being highly correlated if they are located close to each other in the plot; therefore, nine frequency-shift variables can be reduced into four groups. The first group contains the C–C stretching mode. The second group contains the ring C–H wagging, CH_3 umbrella, C=N stretching and CH_3 rocking modes. The third group contains the ring breathing, ring C–H rocking and C=O stretching modes. The fourth group contains the CH_3 symmetric stretching mode. Descriptors in groups one and three are inversely intercorrelated, as are interactions between groups two and four. In addition, points at the extreme *x*- and *y*-axes indicate the descriptors have strong effects.

Separation among different odor categories can be observed from the PCA scores plot as shown in Fig. 9. Compounds with the acid, roasted and sweet-roasted odor qualities are clearly separated. Although compounds with the buttery and oily-popcorn odor quality are distributed close to the origin of the *x*- and *y*-axes, they are well classified. In fact, the first component describes the acid and sweet-roasted odor quality. The descriptors for some of the modes, such as the ring C-H wagging, CH_3

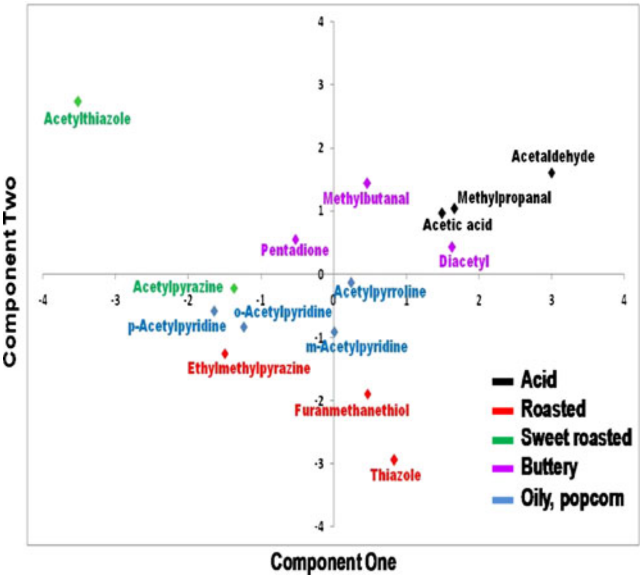


Fig. 9 PCA scores plot from components one and two. Odor-categories are colored as following: *black* acid odor; *red* roasted; *green* sweet roasted; *purple* buttery; *blue* oily, popcorn

umbrella, C=N stretching and CH_3 rocking modes, show strong positive loading, and therefore these values will increase in magnitude in going from the sweet-roasted to the acid odor quality. In contrast, the frequency shift of the CH_3 symmetric stretching mode was found to have strong negative loading on component one, indicating that this value increases in ligands going from the acid to the sweet-roasted odor quality. This negative loading shows opposite effect to odor quality, compared with the above mentioned vibrational modes, which have positive loading. The second component helps describe the acid, buttery, oily-popcorn and roasted odor quality. Frequency-shift descriptors such as ring breathing, ring C-H rocking and C=O exhibit marked positive loading on component two; therefore, these values will increase in magnitude in going from the roasted to the buttery odor quality. The opposite trend was observed for shifts of the C–C stretching mode. Thus, for this frequency mode, negative loading can be seen on both components one and two (Fig. 8). These results indicate that odor perception has to do with frequency shifts of odorants in the receptor. Thus, our results support the hypothesis that the receptor may sense the vibration shifts of the odorant compounds for odor identification. Our results also demonstrate that PCA allows us to determine the relationship between the frequency shift and odor quality.

Conclusions

To better understand the smell sensation and structure-odor relationship of molecules, a series of theoretical calculations were performed. Vibrational properties of key volatile compounds from bread baking such as acetaldehyde, acetylpyrazine, diacetyl, 2-ethyl-3-methylpyrazine and acetylpyridine, which induce acid (pungent), roasted, buttery, sweet-roasted, and oily-popcorn odors, respectively, were examined by using the B3LYP/6-311++G(d,p) method for the isolated state and the ONIOM(B3LYP/6-311++G(d,p): PM3) method for the complex state, employing a 703-atom model of bovine rhodopsin as the receptor. We found that these odorant molecules bind to the membrane protein less tightly than drug/enzyme systems that bind in a “lock and key” manner.

Odorants in each odor quality exhibit characteristic vibration spectrum patterns in the isolated state as well as in the complex state. Even though the vibrational frequencies of odorants in the gas phase and in the complex are quite similar, differences, especially in intensity, are apparent. Interactions, such as hydrogen bonding, between odorant molecules and the surrounding residues not only shift the vibration frequencies but also sometimes yield higher intensities for the corresponding vibration modes. If the vibration properties of odorant molecules are related to the odor as suggested by previous studies, vibrational spectrum patterns in the gas phase would provide “fingerprints” of odor categories for a series of compounds. To gain more detailed insight into the vibration of ligands in complex with protein membranes, we also employed PCA to explore the relationship between the frequency shifts and the odor categories. We found significant frequency shifts of the C–C stretching, CH_3 umbrella, C=O stretching and CH_3 symmetric stretching modes, which allow the oder category to be characterized. The loadings plot demonstrates that the frequency shifts of the C–C stretching mode exhibit positive loading, whereas the shifts of the C=O stretching mode exhibit negative loading. Thus, these vibrational modes show opposite behavior in terms of odor category identification. The same interpretation was obtained for the frequency shift of the CH_3 umbrella and CH_3 symmetric stretching modes. The four groups obtained by PCA clearly classify the five odor categories. The acid and sweet-roasted odor categories correlate strongly with the frequency shift of the CH_3 symmetric stretching and CH_3 umbrella modes. The acid, buttery, oily-popcorn and roasted odor categories are well described by the frequency shifts of the C–C stretching and CH_3 symmetric stretching modes. These results therefore support the hypothesis that 7TMH receptors detect the characteristic vibration shifts of molecules in order to precisely identify the odor. Because of the lack of a crystal structure for the human OR, we used a crystal structure of bovine rhodopsin for our analysis. However, we are aware that rhodopsins and ORs have several differences in structure as well as in physiological function. Hence, further studies using more reliable models are needed to firmly correlate vibrational frequency shifts and smelling sensation.

Acknowledgments This work was initiated when one of the authors (W.T.) was at FIFC under the JENESYS (Japan-East Asia Network of Exchange for Students and Youths) program, supported by the Japan Society for the Promotion of Science (JSPS). The Thailand Research Fund (RTA5380010) is gratefully acknowledged for research support. Research facilities were made available through the generosity of the Laboratory for Computational and Applied Chemistry (LCAC), the National Center of Excellence in Petroleum, Petrochemical Technology and Advanced Materials, the Center of Nanotechnology at Kasetsart University, and the National Research University Program. Support from the Kasetsart University Research and Development Institute (KURDI) is also acknowledged. Thanks are due to Dr. Matthew Paul Gleeson for a valuable discussion on the multivariate analysis and for reading of the manuscript.

References

1. This H (2002) Molecular gastronomy. Angew Chem Int Edn 41:83–88

2. Kristiansen K (2004) Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function. Pharmacol Ther 103:21–80
3. Arora G, Cormier F, Lee B (1995) Analysis of odor-active volatiles in Cheddar cheese headspace by multidimensional GC/MS/sniffing. J Agric Food Chem 43:748–752
4. Bignetti E, Damiani G, De Negri P, Ramoni R, Avanzini F, Ferrari G, Rossi GL (1987) Specificity of an immunoaffinity column for odorant-binding protein from bovine nasal mucosa. Chem Senses 12:601–608
5. Flower DR (1994) The lipocalin protein family: a role in cell regulation. FEBS Lett 354:7–11
6. Tota MR, Strader CD (1990) Characterization of the binding domain of the beta -adrenergic receptor with the fluorescent antagonist carazolol. Evidence for a buried ligand binding site. J Biol Chem 265:16891–16897
7. Cabrera-Vera TM, Vanhauwe J, Thomas TO, Medkova M, Preininger A, Mazzoni MR, Hamm HE (2003) Insights into G protein structure, function, and regulation. Endocr Rev 24:765–781
8. Seifert R, Wenzel-Seifert K (2002) Constitutive activity of G-protein-coupled receptors: cause of disease and common property of wild-type receptors. Naunyn-Schmiedebergs Arch Pharmacol 366:381–416
9. Schoneberg T, Schultz G, Gudermann T (1999) Structural basis of G-protein-coupled receptor function. Mol Cell Endocrinol 151:181–193
10. Dyson GM (1928) Some aspects of the vibration theory of odor. Perfum Essent Oil Rec 19:456–459
11. Turin L (2002) A method for the calculation of odor character from molecular structure. J Theor Biol 216:367–385
12. Zarzo M (2007) The sense of smell: molecular basis of odorant recognition. Biol Rev 82:455–479
13. Brookes JC, Horsfield AP, Stoneham AM (2009) Odour character differences for enantiomers correlate with molecular flexibility. J R Soc Interface 6:75–86
14. Valiev M, Kawai R, Adams JA, Weare JH (2003) The role of the putative catalytic base in the phosphoryl transfer reaction in a protein kinase: first-principles calculations. J Am Chem Soc 125:9926–9927
15. Park H, Yoon J, Seok C (2008) Strength of C-alpha -H···O:C hydrogen bonds in transmembrane proteins. J Phys Chem B 112:1041–1048
16. Hattotuwagama CK, Drew MGB, Nursten HE (2006) Quantum mechanics studies of the tautomers of diacetylformoin, an important Maillard product and odorant. THEOCHEM 775:67–76
17. Abraháo-Júnior O, Nascimento PGBD, Galembeck SE (2001) Conformational analysis of the HIV-1 virus reverse transcriptase nonnucleoside inhibitors: TIBO and nevirapine. J Comput Chem 22:1817–1829
18. Saen-Oon S, Kuno M, Hannongbua S (2005) Binding energy analysis for wild-type and Y181C mutant HIV-1 RT/8-Cl TIBO complex structures: quantum chemical calculations based on the ONIOM method. Proteins Struct Funct Genet 61:859–869
19. Svensson M, Humbel S, Froese RDJ, Matsubara T, Sieber S, Morokuma K (1996) ONIOM: A multilayered integrated MO+MM method for geometry optimizations and single point energy predictions. A test for Diels-Alder reactions and Pt(P(t-Bu)3)2+H2 oxidative addition. J Phys Chem 100:19357–19363
20. Ponzoni A, Depari A, Falasconi M, Comini E, Flammini A, Marioli D, Taroni A, Sberveglieri G (2008) Bread baking aromas detection by low-cost electronic nose. Sens Actuators B 130:100–104
21. Frisch MJT, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA Jr, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian 03. Gaussian Inc, Wallingford CT
22. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748
23. Singh D, Srivastava SK, Ojha AK, Asthana BP, Singh RK (2007) DFT study of hydrogen bond bridging mode of pyridine and diazenes in water environment. THEOCHEM 819:88–94
24. Ayala AP, Siesler HW, Wardell SMSV, Boechat N, Dabbene V, Cuffini SL (2007) Vibrational spectra and quantum mechanical calculations of antiretroviral drugs: Nevirapine. J Mol Struct 828:201–210
25. Irle S, Bowman JM (2000) Direct ab initio variational calculation of vibrational energies of the $H_2O\cdots Cl^-$ complex and resolution of experimental differences. J Chem Phys 113:8401–8403
26. Vailikhit V, Treesuwan W, Hannongbua S (2007) A combined MD-ONIOM2 approach for 1H NMR chemical shift calculations including a polar solvent. THEOCHEM 806:99–104
27. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman PA (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24:1999–2012
28. Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, Pearlman DA, Crowley M, Walker RC, Zhang W, Wang B, Hayik S, Roitberg A, Seabra G, Wong KF, Paesani F, Wu X, Brozell S, Tsui V, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Beroza P, Mathews DH, Schafmeister C, Ross WS, Kollman PA (2006) AMBER 9. University of California, San Francisco
29. Stewart JJP (1989) Optimization of parameters for semiempirical methods. I Method. J Comput Chem 10:209–220
30. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer. An environment for comparative protein modeling. Electrophoresis 18:2714–2723
31. Triller A, Boulden EA, Churchill A, Hatt H, England J, Spehr M, Sell CS (2008) Odorant–receptor interactions and odor percept: a chemical perspective. Chem Biodiversity 5:862–886
32. Kuno M, Hannongbua S, Morokuma K (2003) Theoretical investigation on nevirapine and HIV-1 reverse transcriptase binding site interaction, based on ONIOM method. Chem Phys Lett 380:456–463
33. Saen-oon S, Kuno M, Hannongbua S (2005) Binding energy analysis for wild-type and Y181C mutant HIV-1 RT/8-Cl TIBO complex structures: quantum chemical calculations based on the ONIOM method. Proteins 61:859–869
34. Qian M, Reineccius G (2003) Potent aroma compounds in Parmigiano Reggiano cheese studied using a dynamic headspace (purge-trap) method. Flavour Frag J 18:252–259

35. Gassenmeier K, Schieberle P (1994) Comparison of important odorants in puff-pastries prepared with butter or margarine. LWT Food Sci Technol 27:282–288
36. Song H, Xia L (2008) Aroma extract dilution analysis of a beef flavouring prepared from flavour precursors and enzymatically hydrolysed beef. Flavour Frag J 23:185–193
37. Schieberle P (1996) Odour-active compounds in moderately roasted sesame. Food Chem 55:145–152
38. Rega B, Guerard A, Delarue J, Maire M, Giampaoli P (2009) On-line dynamic HS-SPME for monitoring endogenous aroma compounds released during the baking of a model cake. Food Chem 112:9–17
39. d'Acampora Zellner B, Dugo P, Dugo G, Mondello L (2008) Gas chromatography-olfactometry in food flavour analysis. J Chromatogr A 1186:123–143
40. Afeefy HY, Liebman JF, Stein SE (2010) In: Linstrom PJ, Mallard WG (eds) Vibrational energy. NIST Chemistry WebBook, NIST Standard Reference Database Number 69, National Institute of Standards and Technology, Gaithersburg MD, 20899, http://webbook.nist.gov

J Mol Model (2012) 18:2241–2246
DOI 10.1007/s00894-011-1242-x

ORIGINAL PAPER

DFT studies of COOH tip-functionalized zigzag and armchair single wall carbon nanotubes

Elżbieta Chełmecka · Karol Pasterny · Teobald Kupka · Leszek Stobiński

Received: 10 July 2011 / Accepted: 12 September 2011 / Published online: 1 October 2011

Abstract Structure and energy calculations of pristine and COOH-modified model single wall carbon nanotubes (SWCNTs) of different length were performed at B3LYP/6-31G* level of theory. From 1 to 9 COOH groups were added at the end of the nanotube. The differences in structure and energetics of partially and fully functionalized SWCNTs at one end of the nanotube are observed. Up to nine COOH groups could be added at one end of (9,0) zigzag SWCNT in case of full functionalization. However, for (5,5) armchair SWCNT, the full functionalization was impossible due to steric crowding and rim deformation. The dependence of substituent attachment energy on the number of substituents at the carbon nanotube rim was observed.

E. Chełmecka
Division of Statistics, Department of Instrumental Analysis,
Medical University of Silesia,
30, Ostrogórska Street,
41-200 Sosnowiec, Poland

K. Pasterny
A. Chełkowski Institute of Physics, University of Silesia,
4, Uniwersytecka Street,
40-007 Katowice, Poland

T. Kupka (✉)
Faculty of Chemistry, University of Opole,
48, Oleska Street,
45-052 Opole, Poland
e-mail: teobaldk@yahoo.com

L. Stobiński
Institute of Physical Chemistry, Polish Academy of Sciences,
44/52, Kasprzaka 44/52,
01-224 Warsaw, Poland

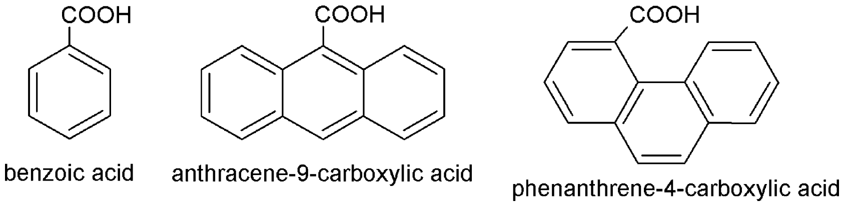
L. Stobiński
Faculty of Materials Science and Engineering,
Warsaw University of Technology,
Wołoska 141,
02-507 Warsaw, Poland

Keywords Carboxylation energy · COOH functionalization · DFT · End-substitution · Zigzag and armchair SWCNT

Introduction

Three different types of carbon nanotubes are experimentally observed: armchair, zigzag and chiral [1]. These carbon structures are finished with semispheres containing pentagons and hexagons, being formally parts of fullerenes. Due to their structure, the CNTs are hydrophobic, strongly interact with light and possess interesting electrical and physical properties [1–3].

Modification of hydrophobic carbon nanotubes by allowing stronger intermolecular interactions, leading to solubility is expected upon addition of selected small molecules covalently bonded to the (a) end, (b) surface, or (c) both the end and surface (mixed) of SWCNTs [2, 4–13]. The rim structure of SWCNT in case of zigzag or armchair open ended CNT shows a different pattern. Zigzag carbon nanotubes show metallic or semiconductor properties and their ends shows "saw-tooth" like shape. Functionalized carbon nanotubes are promising candidates in material sciences and nanomedicine [1–3]. For example, OH, COOH or NH_2 functionalized CNT are easily transformed and could bear longer chains connecting antibodies or drugs. End-substituted SWCNT are by-products of mild oxidation and carboxylic, carbonyl and hydroxyl groups are frequently formed. Most previous works [4–6] concentrated on single functionalized SWCNTs and the impact of substituent on physical properties of modified *versus*



Scheme 1 Small model molecules (benzoic, anthracene-9-carboxylic and phenathrene-4-carboxylic acids

pristine structure (for example, density of states, HOMO-LUMO gap). The pristine CNTs are insoluble in water and organic solvents and this is a serious hindrance in their industrial applications, for example, as efficient nanocomposites [1, 14, 15].

Earlier works demonstrated a possibility of transforming inert and hydrophobic CNTs, into soluble forms [14–17]. This was accomplished by subsequent chemical modification of water soluble CNTs, containing COOH groups [14–17]. Unfortunately, little is known about the systematic changes of energy of zigzag and armchair SWCNTs upon consecutive replacement of rim hydrogen atoms by COOH groups.

Density functional theory (DFT) and, in particular, the exchange-correlation B3LYP hybrid density functional is widely used in molecular modeling studies to predict structure, spectroscopic parameters and energy changes of small, middle and large size molecules [7, 8, 18–20]. Due to the large size of CNTs, the DFT calculations with relatively small basis sets (3-21G or 6-31G*), and also AM1 and PM3 semiempirical methods, have been used for theoretical description of molecular structure and other parameters of finite models of CNTs [8].

In this study, as an extension of our previous works on hydroxyl substituted SWCNTs [19, 21, 22], we would like to get a more detailed information at the atomistic level on the open-end CNT modification with COOH, up to full substitution with nine (zigzag) or ten (armchair) carboxylic groups.

Computational methods

All calculations were performed using Gaussian 09 program [23]. Reliable exchange-correlational B3LYP hybrid density functional and basis sets of relatively small size (3-21G and 6-31G*), enabling completing fairly large scale calculations were selected. Full structure optimization of unsubstituted open-ended (with dangling bonds on carbon saturated with hydrogen atoms), and COOH-modified SWCNTs were performed. Several models of SWCNT were selected, including (9,0) zigzag and (5,5) armchair structures with one and three layers (strings) of hexagon units. IR and Raman harmonic frequencies were calculated in case of one layer with one to 9/10 COOH substituents. All positive frequencies ensured ground state structure of the optimized system.

For comparison purposes only, the calculations with small model molecules including methane, benzene, anthracene and phenanthrene before, and after replacing one hydrogen atom with the carboxyl group were performed

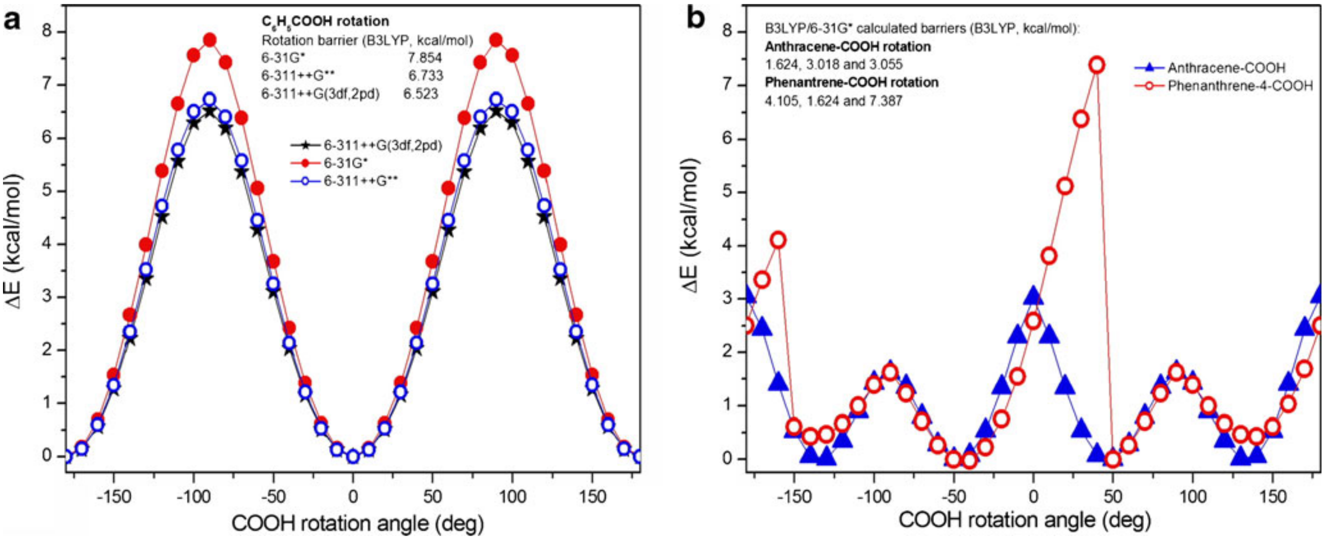


Fig. 1 Energy change upon rotation of COOH substituent relative to ring plane in (**a**) benzoic acid and (**b**) anthracene-9-carboxylic acid and phenathrene-4-carboxylic acid

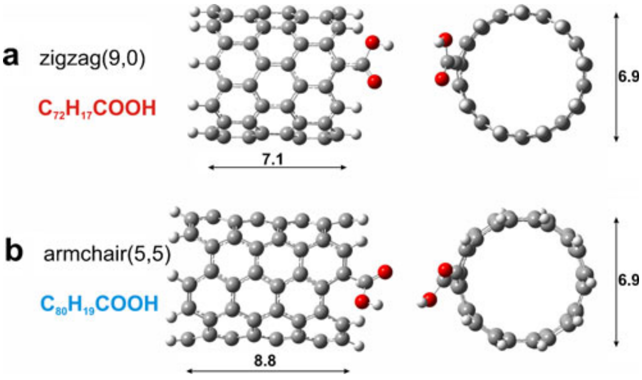


Fig. 2 Optimized structures of model (**a**) zigzag and (**b**) armchair SWCNTs with a single COOH substituent at the rim (dimensions in Å)

at the same level of theory. In addition, to verify the basis set quality impact, these calculations were performed with a large basis set (6-311++G(3df,2pd).

Energy of one COOH group formation at nanotube terminated initially with H atoms was calculated by considering a hypothetical reaction:

$$\mathrm{SWCNT-H+CH_3COOH\rightarrow SWCNT-COOH+CH_4}. \quad (1)$$

Energies of adding subsequent groups (ΔE in kcal mol^{-1}) were assumed as follows:

$$\Delta E_n = \left[E\left(\mathrm{SWCNT(COOH)}_n\right)+E(\mathrm{CH_4})\right] - \left[E\left(\mathrm{SWCNT(COOH)}_{n-1}\right)+E(\mathrm{CH_3COOH})\right], \quad (2)$$

where n=1,2,…9 (10).

Initially, the calculations were conducted at B3LYP/3-21G level of theory. Qualitatively, the changes in energies obtained with smaller basis set (3-21G) were similar to those, obtained at B3LYP/6-31G*. Thus, the final results, obtained with the larger basis set will be only discussed.

Results and discussion

The carboxylic group can be considered as an asymmetric substituent with two different ends (O atom *vs.* OH group) and their position in respect to the rim of the CNT be positioned in a way which minimizes the interactions with the neighboring H-atoms and/or forms H-bonds with other COOH substituents. This was tested on model systems (Scheme 1) by rotating the COOH substituent relative to the aromatic ring plane (changing the dihedral angle C_{ring}–C_{ring}–C=O).

The energy landscape of COOH rotation in case of monosubstituted benzene, phenanthrene and anthracene are shown in Fig. 1. In case of benzene, the energy minimum, corresponding to favorable carboxylic orientation, coplanar with the ring, is observed and the perpendicular position, e.g., at both sides of the ring, are about 8 kcal mol^{-1} higher. In case of anthracene, the energy minimum corresponds to about 45 degree deviation of COOH plane from rings plane and there are also two maxima of the same height (at about 3 kcal mol^{-1}) for the perpendicular orientation. The rotation of COOH in phenanthrene at position 4 (see Scheme 1) leads to an asymmetric shape of energy curve. The basis set effect on the position of energy maxima upon COOH rotation in C_6H_5COOH is also shown in Fig. 1a. Thus, upon improving the basis set quality from 6-31 G* to 6-311++G** and 6-311++G(3df,2pd) the barrier height slightly decreases (from 7.85 to 6.73 and 6.52 kcal mol^{-1}). It is apparent that

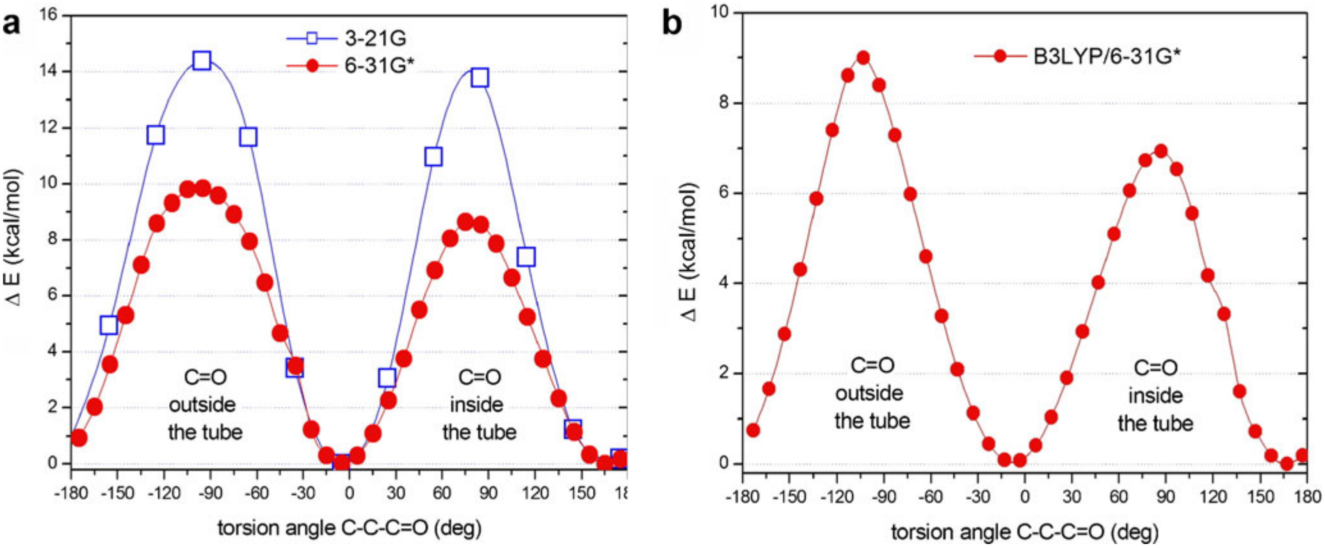


Fig. 3 Energy change upon rotation of COOH substituent at the rim of model (**a**) zigzag and (**b**) armchair SWCNTs

the barrier height decreases by about 1.3 kcal mol^{-1} upon significantly improving the basis set quality.

In Fig. 2a and b are shown optimized structures of zigzag and armchair CNT consisting of three ring layers with a single COOH substituent at the rim.

In Fig. 3 are shown energy landscapes of single COOH group rotation attached to zigzag and armchair SWCNTs. In this case, the preferred geometry is observed for both -OH and =O ends of carboxylic group outside the tube (on the circumference). Two energy maxima are observed for COOH group oriented along the tube radius and the slightly lower one corresponds to OH being outside the tube. In case of zigzag CNT, the energy minimum corresponds to C–C–C=O angle of about −5 degrees, (substituent on the circumference) and the highest maximum corresponds to about −90 degrees (C=O outside the tube). The other maximum (with C=O inside, or oriented toward the tube center) is slightly lower. Similarly to Fig. 1a, the improvement of basis set quality from 3-21G to 6-31G* leads to energy barrier lowering by about 4 kcal mol^{-1}. In the case of carboxylic group rotation at the armchair rim, the situation is similar and the corresponding barrier heights are 9 and 7 kcal mol^{-1} and the energy minimum is observed at about −10 degrees.

Up to nine carboxylic substituents were placed consecutively at the zigzag rim (see Fig. 4a), forming stable structures. In this case, a kind of threefold symmetry was observed. Nevertheless, some funnel shape deformation and increase of the tube-end diameter was observed.

The armchair model consisting of three hexagon layers with up to nine COOH groups at one rim was also stable. In addition, upon complete functionalization of one hexagon layer (the shortest armchair nanotube model) a stable system was also observed (Fig. 4b). However, all attempts to obtain fully functionalized one end of a longer tube, containing three layers of hexagons, failed. This was probably due to steric crowding at the relatively rigid tube skeleton end.

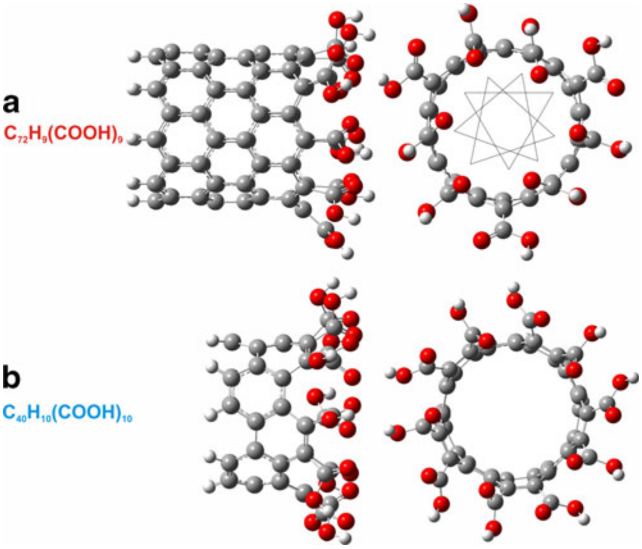


Fig. 4 Optimized structures of model (**a**) zigzag and (**b**) armchair SWCNTs fully functionalized with COOH substituents at the rim. Threefold symmetry is indicated for zigzag nanotube

Table 1 Comparison of carboxylation and hydroxylation energy (kcal mol^{-1}) calculated at B3LYP/6-311++G(3df,2pd) level for two model compounds according to Eq. 2

	CH_4	C_6H_6
-COOH		
ΔE		−3.8
Δ(E+ZPV)		−4.9
-OH[a] [19]		
ΔE	−29.4	−39.9
Δ(E+ZPV)	−26.5	−38.4

[a] in agreement with formula (2) an opposite sign to that in ref. [19] is given

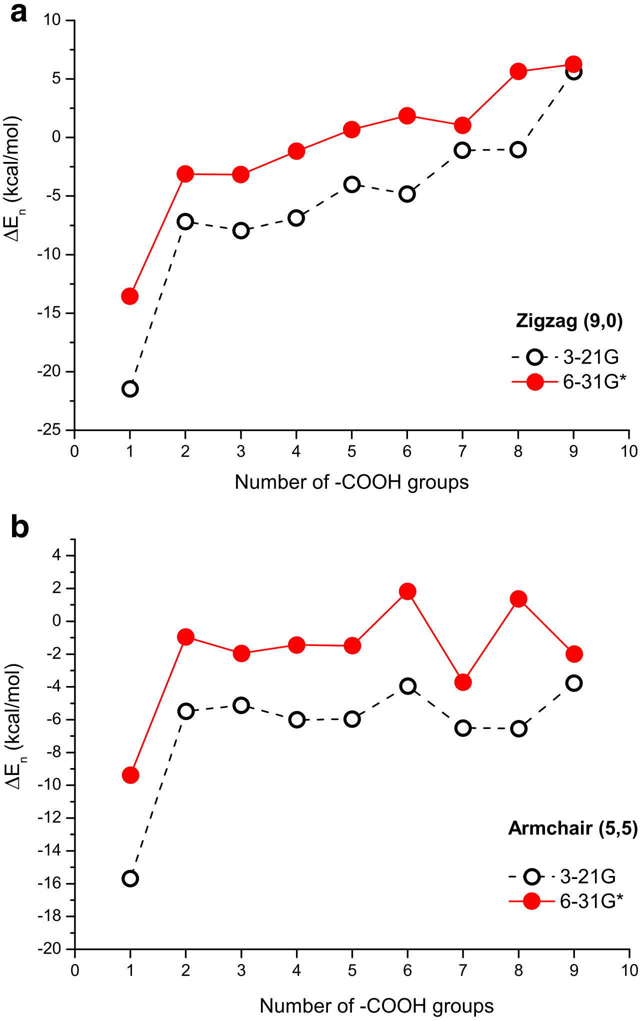


Fig. 5 Dependence of B3LYP predicted carboxylation energy according to Eq. 2 for model zigzag and (**b**) armchair SWCNTs functionalization at two basis set sizes. For better visualization the data points are connected

In the next step, starting from model systems of methane and benzene, we examined the energetics of substitution process calculated according to Eq. 2 (see Table 1). Addition of ZPV correction changes the substitution energy slightly while hydroxylation is more favorable in the case of benzene.

The relative carboxylation energy, calculated with Eq. 2 *vs*. the number of COOH substituents for zigzag and armchair are displaced in Fig. 5a and b.

It is evident from Fig. 5a that the carboxylation energy for the first hydrogen atom at the rim of zigzag nanotube formed from three layers is about −13 kcal mol^{-1}. This differs from the calculated previously [19] hydroxylation energy of about −35 kcal mol^{-1}. Addition of the second COOH group differs by about 10 kcal mol^{-1}. Significantly smaller energy increments are needed for adding three to nine carboxylic groups.

In Fig. 5b is shown a similar carboxylation energy dependence on replacement of consecutive hydrogen atoms at the rim of armchair CNT model. The first carboxylation energy is higher than for zigzag model (about –8.5 vs. –13 kcal mol^{-1}) whereas for the second group this energy is about −1 vs. –3 kcal mol^{-1}, for armchair and zigzag models, respectively. However, some oscillation of energy is observed, with minima located at odd numbers of COOH. This resembles the results for hydroxylation energy pattern observed previously for armchair CNT [21, 22]. The reason of this behavior was explained earlier as a result of different H-bond ring pattern formation at the rim. The results presented in Fig. 5 indicate higher reactivity of zigzag versus armchair SWCNT rim toward carboxylation. This is in agreement with earlier observations by Kim et al. [24].

Conclusions

The present density theory studies using B3LYP hybrid functional indicate a possibility of COOH-functionalization of one end of zigzag CNT with one to nine substituents. However, the end of armchair nanotube cannot be fully functionalized with COOH groups (one hydrogen atom remains unsubstituted).

On the basis of the performed B3LYP/6-31G* calculations it appears that the replacement of one hydrogen atom at the rim of the zigzag CNT model is a more exothermic process than for armchair model (−13 vs −8.5 kcal mol^{-1}). This indicates a higher reactivity of zigzag CNT toward carboxylation. A gradual and nearly linear increase of energy is observed for subsequent carboxylation, starting from two to nine groups on a zigzag nanotube end.

Acknowledgments This work was supported by the Polish Council for Science through the Development Grants for the years 2008–2011 (NR 15-0011-04/2008, KB/72/13447/IT1-B/U/08) and the calculations were performed using a hardware and software at WCSS Wrocław. We also acknowledge the kind support of Medical University of Silesia, University of Silesia and University of Opole and fruitful discussions with Prof. Małgorzata Broda.

References

1. Saito R, Dresselhaus MS, Dresselhaus G (1998) Physical properties of carbon nanotubes. Imperial College Press, London
2. Khabashesku VN, Margrave JL, Barrera EV (2005) Functionalized carbon nanotubes and nanodiamonds for engineering and biomedical applications. Diamond Rel Mat 14:859–866
3. Calvert P (1999) A recipe for strength. Nature 399:210–211
4. Kar T, Akdim B, Duan X, Pachter R (2006) Open-ended modified single-wall carbon nanotubes: A theoretical study of the effects of purification. Chem Phys Lett 423:126–130
5. Veloso MV, Filho AGS, Filho JM, Fagan SB, Mota R (2006) *Ab initio* study of covalently functionalyzed carbon nanotubes. Chem Phys Lett 430:71–74
6. Kuzmany H, Kukovecz A, Simon F, Holzweber M, Kramberger C, Pichler T (2004) Functionalization of carbon nanotubes. Synth Met 114:113–122
7. Kar T, Scheiner S, Roy AK (2008) The effect on acidity of size and shape of carboxylated single-wall carbon nanotubes. A DFT-SLDB study. Chem Phys Lett 460:225–229
8. Wongchoosuk C, Udomvech A, Kerdcharoen T (2009) The geometrical and electronic structures of open-end fully functionalized single-walled carbon nanotubes. Curr Appl Phys 9:352–358
9. Kar T, Scheiner S, Patnaik SS, Bettinger HF, Roy AK (2010) IR characterization of tip-functionalized single-wall carbon nanotubes. J Phys Chem C 114:20955–20961
10. Kar T, Adkim B, Duan X, Pachter R (2006) Open-ended modified single-wall carbon nanotubes: A theoretical study of the effects of purification. Chem Phys Lett 423:126–130
11. Salzmann CG, Llewellyn SA, Tobias G, Ward MAH, Huh Y, Green MLH (2007) The role of carboxylated carbonaceous fragments in the functionalization and spectroscopy of a single-walled carbon-nanotube material. Adv Mater 19:883–887
12. Zhao J, Lu JP, Han J, Yang CK (2003) Noncovalent functionalization of carbon nanotubes by aromatic organic molecules. Appl Phys Lett 82:3746–3751
13. Vigoloa B, Mamane V, Valsaque F, Le TNH, Thabit J, Ghanbaja J, Aranda L, Fort Y, McRae E (2009) Evidence of sidewall covalent functionalization of single-walled carbon nanotubes and its advantages for composite processing. Carbon 47:411–419
14. Tasis D, Tagmatarchis N, Georgakilas V, Prato M (2003) Soluble carbon nanotubes. Chem Eur J 9:4000–4008
15. Sun YP, Fu K, Lin Y, Huang W (2002) Functionalized carbon nanotubes: Properties and applications. Acc Chem Res 35:1096–1104
16. Chen J, Hamon MA, Hu H, Chen Y, Rao AM, Eklund PC, Haddon RC (1998) Solution properties of single-walled carbon nanotubes. Science 282:95–98
17. Liu J, Rinzler AG, Dai H, Hafner JH, Bradley RK, Boul PJ, Lu A, Iverson T, Shelimov K, Huffman CB, Rodriguez-Macias F, Shon YS, Lee TR, Colbert DT, Smalley RE (1998) Fullerene pipes. Science 280:1253–1256
18. Foresman JB, Frisch A (eds) (1996) Exploring chemistry with electronic structure methods. Gaussian Inc, Pittsburg, PA
19. Chełmecka E, Pasterny K, Kupka T, Stobiński L (2010) Density functional theory studies of OH-modified open-ended single-wall

zigzag carbon nanotubes (SWCNTs). J Mol Struct Theochem 948:93–98

20. Stobinski L, Tomasik P, Lii CY, Chan HH, Lin HM, Liu HL, Kao CT, Lu KS (2003) Single-walled carbon nanotube-amylopectin complexes. Carbohydr Polym 51:311–316
21. Chełmecka E, Pasterny K, Kupka T, Stobiński L (2011) DFT studies of OH-functionalized open-ended zigzag, armchair, and chiral single wall carbon nanotubes. Phys Status Solidi A 208:1774–1777
22. Chełmecka E, Pasterny K, Kupka T, Stobiński L (2011) OH-functionalized open-ended single wall armchair carbon nanotubes (SWCNT) studied by density functional theory. J Mol Model. doi:10.1007/s00894-011-1181-6
23. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin R. L, Fox DJ (2009) Gaussian 09, Revision A.02. Gaussian. Wallingford, CT
24. Kim C, Seo K, Kim B, Park N, Choi YS, Park KA, Lee YH (2003) Tip-functionalized carbon nanotubes under electric fields. Phys Rev B 68:115403(1–7)

J Mol Model (2012) 18:2247–2256
DOI 10.1007/s00894-011-1245-7

ORIGINAL PAPER

Why is the crystal shape of TATB is so similar to its molecular shape? Understanding by only its root molecule

Chaoyang Zhang · Bin Kang · Xia Cao · Bin Xiang

Received: 2 June 2011 / Accepted: 14 September 2011 / Published online: 1 October 2011

Abstract We present an understanding of the quasi-regular or regular hexagonal enlargement of 1,3,5-triamino-2,4,6 (TATB) from its root molecule to its bulk crystal, by only its root molecule. That is, the mechanism of regular hexagonal TATB molecules stacking to a quasi-regular or regular hexagonal TATB crystal was discussed using a combined method of a density functional theory BLYP and Dreiding forcefield, and a series of static scanning calculations. As a result, we found that there are two styles of forming the most energetically favored TATB dimers: a hydrogen bonding along the molecular plane and an offset π-stacking vertical to the plane, just leading to the outspread and the thickening of the regular hexagon during the crystal growth, respectively. At the same time, it was found that the rotation of one TATB layer in any parallel stacked double-layer should overcome a very high energy barrier. It suggests that the TATB molecules or layers are arranged on the crystal face always along the special orientation of a regular hexagon and other orientations are strongly thermodynamically forbidden, resulting in a hexagonal crystal bulk.

Keywords 1,3,5-triamino-2,4,6-trinitrobenzene (TATB) · Crystal shape · Hydrogen bonding

C. Zhang (✉) · B. Kang · X. Cao
Laboratory of Material Chemistry, Institute of Chemical Materials,
China Academy of Engineering Physics (CAEP),
P.O.Box 919-327, Mianyang, Sichuan,
People's Republic of China 621900
e-mail: zcy19710915@yahoo.com.cn

X. Cao · B. Xiang
College of Chemistry and Chemical Engineering,
Chongqing Universtiy,
Chongqing, People's Republic of China 400044

Introduction

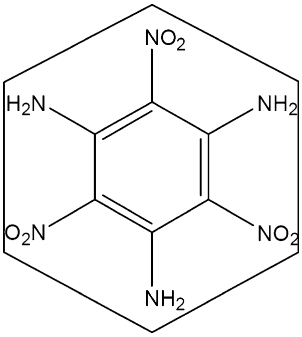
Crystal morphology including crystal shape and size is important to many industrial processes due to its remarkable effects on the qualities and the functionalities of intermediates and final products. Only for explosives, it has already been found that their crystal shapes and sizes can influence their sensitivities versus external stimuli. For example, for octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazocine (HMX) crystals with different shapes and sizes, its spherically-shaped crystal is less sensitive to mechanical stimuli than its needle shaped one or its sheet shaped one in the case of a same size; the smaller crystal is less sensitive to mechanical stimuli but more sensitive to heat than the bigger one in the case of a similar shape [1, 2]. According to this, the explosive crystals are prepared with different shapes and sizes to satisfy the various practical requirements. So to speak, it is also important to control the crystal shapes and sizes of synthesized explosives as to synthesize some new explosive compounds with comprehensively good properties. As a matter of fact, the bottleneck of development of the traditional organic explosive compounds containing C, H, O and N atoms appeared tens of years ago, due to the intrinsic energy-sensitivity contradiction in them: high energy usually goes with high sensitivity denoting low safety [3, 4]. Therefore, it is in a certain sense easier to realize or of more practical interest to enhance the crystal qualities of the existing explosives such as to increase the purity, to control the shape and size, to decrease the imperfection, and so forth.

In the current *era of material design*, people can to a certain extent make prior computer-and-program-aided predictions before they decide to begin a crystal shape engineering [5]. Even though there are still many difficulties in understanding and predicting crystal morphology, lots of endeavors have been done to develop and perfect them, such as describing a crystal energy landscape [6], improving the attachment energy model (AE) [7, 8] in view of the effects of solvent and some dynamic factors [9–11]. By choosing designed solvents or inhibitors, people can get some crystals with expected crystal habits [12]. This is the reason for our optimism, but we cannot ignore the current difficulty in predicting accurately the crystal structures and habits from their root molecules.

1,3,5-triamine-2,4,6-trinitrobenzene (TATB) is a representative insensitive explosive [13] and a special nonlinear optical material [14]. Currently, we found that the shape of the TATB crystal refined from its dimethyl sulfoxide (DMSO) solution [15] in Fig. 1 is very similar to that of its root molecule, a regular hexagon. The regular hexagonal shape of TATB molecule in Fig. 2 can be regarded as an augment of the regular hexagonal benzene ring by adding three amino groups and three nitro groups alternately onto the benzene ring. That is to say, the macro shape of the TATB crystal is very similar to the micro one of its molecule.

This enlargement from a root molecule to its bulk crystal without any obvious shape change is so rare and interesting that it attracts our attention and motivates us to understand the phenomenon. Obviously, it should be a typical case to explore the relationship between molecular and crystal shapes. We just intend to understand this enlargement by *only* a root TATB molecule without respect to its lattice structures and detailed crystallization conditions. TATB in fact has very low solubilities in almost all traditional solvents excluding the newly-found ionic liquids with good solubilities for TATB [16]. It implies the small and negligible solute-solvent interactions and the strong solute-solute interactions during the crystallization from the traditional solvents such as DMSO, which is helpful for us to simplify an assumption of thermodynamic and kinetic conditions for understanding. That is to say, the influence of solvent is not considered in the assumption.

Fig. 2 Regular hexagonal shape of TATB molecule



Methodologies

From the (quasi-) regular hexagonal shapes of the TATB crystals shown in Fig. 1, we can find that there are two sides involved in the crystal growth of TATB: the outspread of a regular hexagon and its thickening with a necessarily invariable orientation. That is, the above-mentioned enlargement from a root molecule to a crystal bulk includes the outspread and the thickening. Obviously, we will not

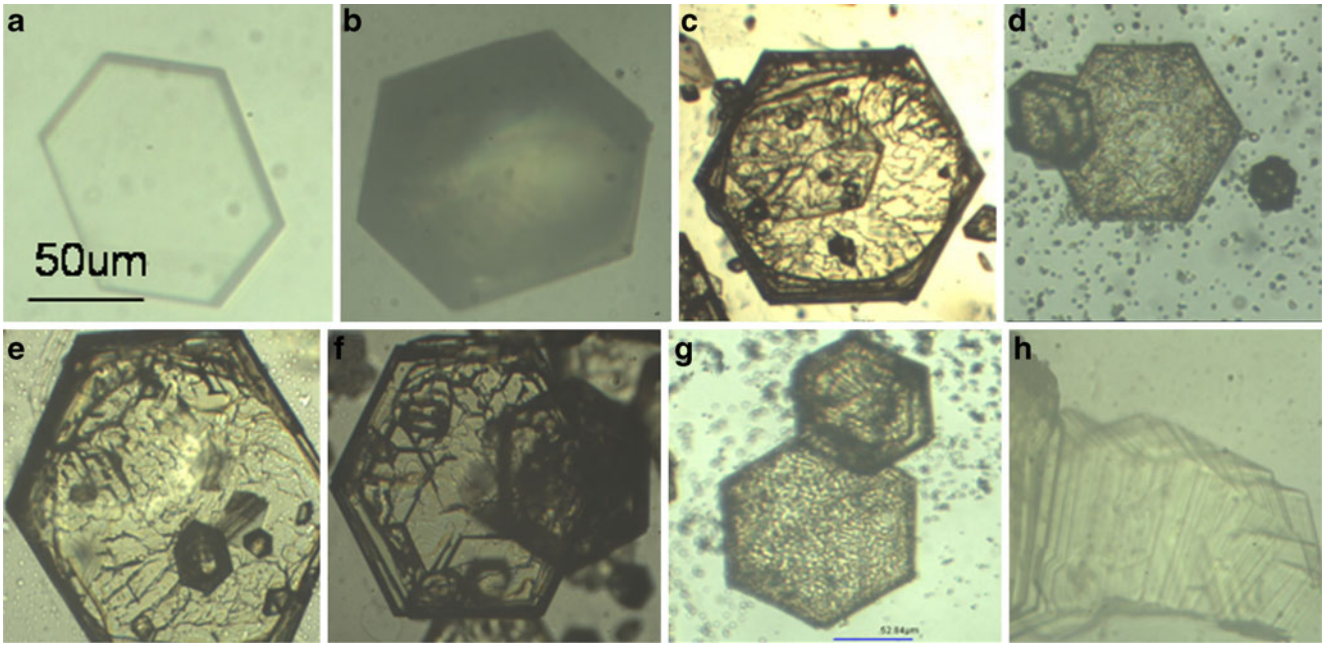


Fig. 1 Optical images of the crystallized TATB from DMSO. (**a**) and (**b**) show the only one hexagonal shell of crystallized TATB; (**c-h**) show the superimpositions without twist of many hexagonal layers with different sizes of crystallized TATB

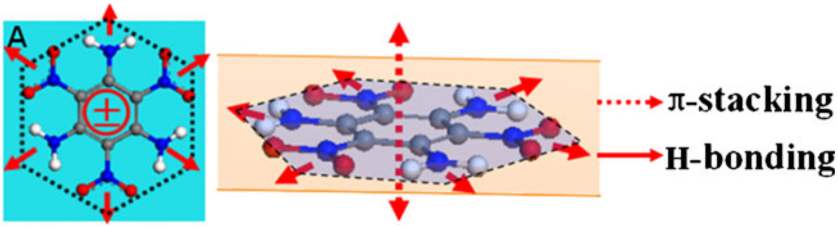


Fig. 3 Six red arrows and ⊕ (or a double-headed arrow) point to the crystal growth orientations determined by the TATB molecular structure, the intermolecular H-bonding in the molecular plane and the π-stacking perpendicular to the plane, respectively

see a thickened hexagon if these hexagons spread not in a consistent orientation.

This is actually a topic of crystal packing in which the dimer interactions are undoubtedly the basic ones. Some research on the molecular arrangement in crystal have been carried out based on the dimer interactions. For example, Dunitz and Gavezzotti offered a method for a quantitative description of crystal packing by molecular pairs and applied it to the hexamorphic crystal system of 5-methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile [17]. Also, the idea of dimer interactions was adopted here for understanding the enlargement. As illustrated in Fig. 3, there are eight main possible orientations of TATB crystal growth, i.e., in fact the orientations of molecular arrangement in crystal: six of the planar hydrogen bonded interactions leading to the outspread, and two of the π-stacking interactions leading to the thickening. We therefore arranged TATB dimers according to these orientations.

As to TATB dimers, two typical cases in terms of its crystal packing in Fig. 4, Dimers *a* and *c*, have been discussed using quantum chemical methods [18–20]. An all-atom molecular forcefield for TATB was established by Gee et al. [18] to simulate its isobaric thermal expansion and isothermal compression under hydrostatic pressures, resulting in good agreement with experiment. In this forcefield, the intermolecular interaction potential is determined by the single-point energies of TATB dimers calculated using the MP2/6-31 G(d, p) method. It confirms that it is feasible to make a forcefield reliable to a bigger system, based on *ab initio* calculation results of some much smaller systems involved in the bigger one. Also, a similar combined method was employed for our calculations. That is, the interaction energy (*ΔE*) of a TATB dimer can be obtained after three steps: (1) relax the TATB molecule and calculate the electrostatic potential (ESP) charges of all atoms using BLYP/DNP method [21, 22]. The results are shown in Fig. 5; (2) keep the relaxed TATB molecule rigid, assign each atom ESP charge, and arrange the TATB dimers in a static scanning way in view of some of the most possible styles; and (3) calculate the single-point energy of each dimer and a single TATB molecule using Dreiding forcefield [23]. *ΔE* is the energy difference after forming a dimer. It is obviously a combined method in that the molecular geometries and charges are from density functional theory (DFT) calculations, the interaction function is of Dreiding forcefield and the double TATB molecules are arranged in a static scanning manner.

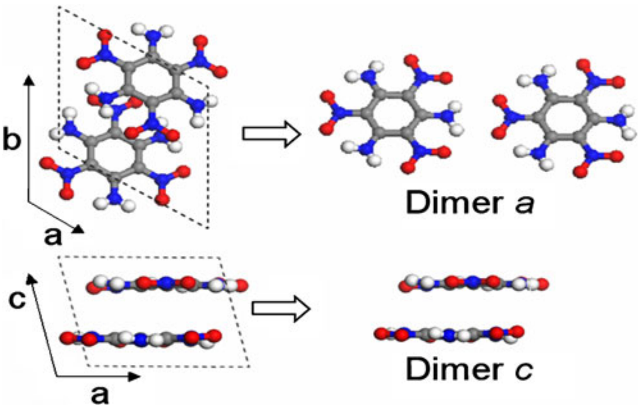


Fig. 4 Two kinds of dimers in terms of the crystal packing of TATB

To verify the reliability of the combined method to the TATB dimers, we compared *ΔE* calculated using it and the MP2 method, respectively. Gee et al.'s method [18] was referred to arrange the TATB dimers to compute *ΔE*: (1) make double TATB molecules in a plane and in a head-to-tail contact manner just as along the *a* crystallographic axis of TATB, and make double TATB molecules eclipsed

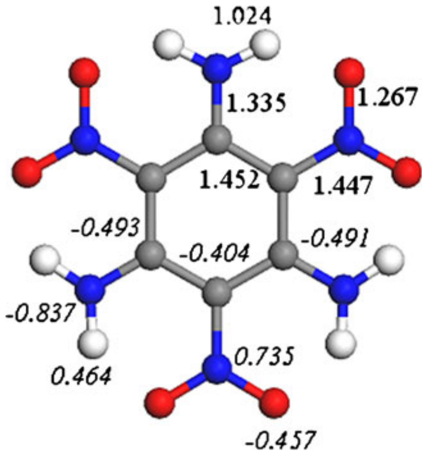


Fig. 5 Bond lengths (unit in Å) and ESP charges (unit in e) of TATB molecules derived from BLYP/DNP calculations

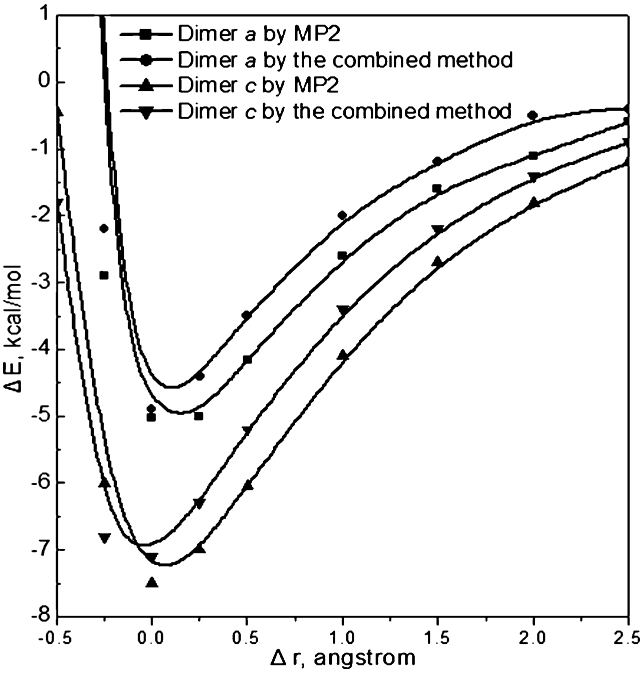


Fig. 6 Interaction energies (ΔE) of two kinds of TATB dimers derived from the MP2 method (cited from ref [18]) and the combined method. Δr is the displacing distance of one TATB molecule in the dimer along an relative orientation like the *a* or *c* crystallographic axis

stacked just like in the unit cell of TATB, like Dimers *a* and *c* in Fig. 4, respectively; and (2) fix one TATB molecule and displace another one in the dimer along the *a* (for Dimer *a*) or the *c* (for Dimer *c*) crystallographic axis a distance Δr and calculate the corresponding ΔE. As shown in Fig. 6, the ΔE derived from our combined method is qualitatively comparable to that of MP2/6-31 G(d, p) calculations [18]: they have the same changing trend and the same site corresponding to the most attraction. We think that our combined method is qualitatively adequate to understand the enlargement. By the way, we have also combined other forcefield functions, including COMPASS [24], CVFF [25] and UFF [26], with the above DFT, the BLYP/DNP method. However, we found that they could not give comparable values to MP2 results.

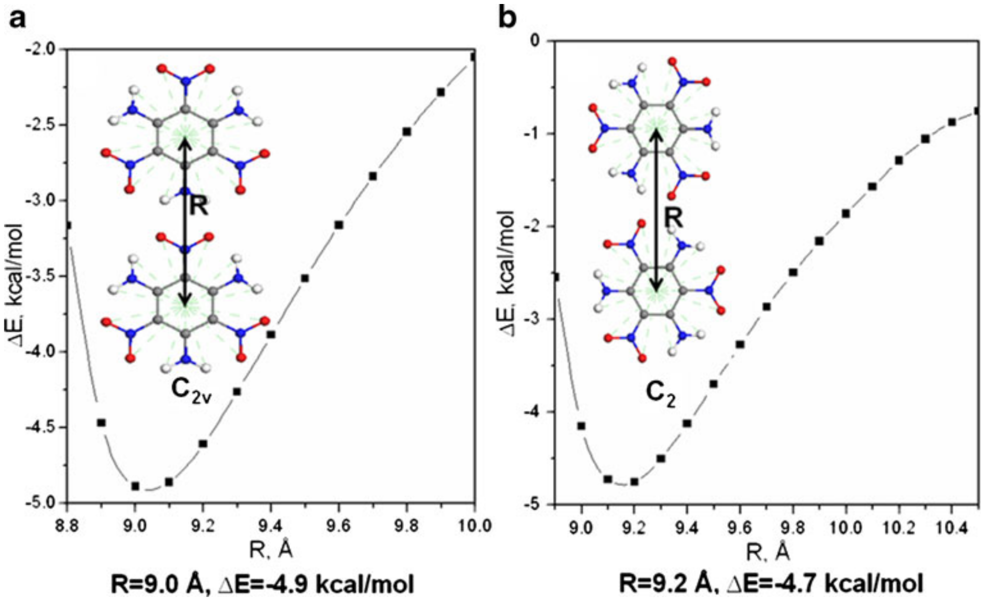
Furthermore, to be closer to the practical "birth and spread" mechanism model of crystal growth, we constructed some stacked double-layers and applied the combined method to discuss the thermodynamic resulted from the orientation variation of one layer overlaid on another layer in each double-layer. This is important to understand the thickening, as stressed later.

Results and discussion

Outspread of the regular hexagon

As illustrated in Fig. 3, the outspread of a regularly hexagonal TATB molecule to a hexagonal crystal bulk is mainly attributed to the planar intermolecular hydrogen bonding. Therefore the dimer interaction calculations were carried out for two styles of arrangement of double TATB molecules in one plane in terms of the preferable formation of hydrogen bonds. As a result, we can confirm that the arrangement in Fig. 7a (*style a*) is more reasonable than that in Fig. 7b (*style b*), by comparing ΔE and the possibility of the molecular arrangement for the planar outspread. As shown in Fig. 7, the biggest dimer attraction of *style a* is 4.9 kcal mol^{-1}, slightly more than that of *style b*, 4.7 kcal mol^{-1}. And *style a* has a little shorter intermolecular distance than *style b*, 0.2 Å, which is helpful for the impact molecular packing in crystal.

Fig. 7 Dimer interactions in the case of the preferable formation of hydrogen bonds. R is the distance between the centroids of double TATB molecules



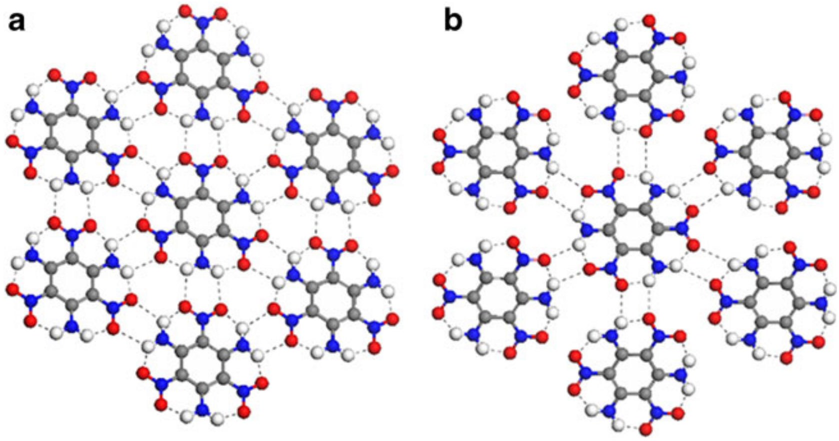


Fig. 8 Arranging TATB molecules for the planar outspread according to the two styles of the preferable formation of hydrogen bonds. The dash shows the hydrogen bonds

Also, we arranged six TATB molecules around a central TATB molecule in terms of the styles of preferable formation of hydrogen bonds as shown in Fig. 8a and b. As illustrated in Fig. 8, the planar outspread of TATB according to *style a* is much more energetically favored than that according to *style b*: for the former, all the TATB molecules are linked through hydrogen bonds not only between the central molecule and the molecules around it but also among these neighboring surrounding molecules; for the latter, the hydrogen bonds form only through the central TATB molecule and the surrounding molecules at the cost of the repulsion among the neighboring surrounding molecules. It is therefore confirmed that the planar hexagonal outspread according to *style a* is reasonable.

Now, we can discuss the outspread according to the dominant style of the formation of hydrogen bonds as mentioned above, with an assumption that this process is thermodynamically and kinetically controlled by the more intermolecular hydrogen bonds and the more free molecules around a nucleus or a growing cluster, respectively. Namely, more intermolecular hydrogen bonds can decrease more energy and increase stability, and more free molecules can increase the velocity of molecule packing. This assumption should be reasonable due to the negligible solute-solvent interactions and the strong solute-solute interactions in TATB crystallization from DMSO. For example, in Fig. 9c and e, ⊗ points to the next growth site because the free molecules lessen their nearby even though there may form the same quantity of hydrogen bonds. Figure 9 shows a stepwise outspread around a center TATB molecule, which is determined by the above mentioned thermodynamic and kinetic rule. These steps gradually form the bigger and bigger concentric hexagons to a final hexagonal shell with a "birth and spread" 2-Dimension growth mechanism. Obviously, this compact and perfect outspread to a regular hexagonal shell takes place only by *style a*.

Thickening of the regular hexagon

As indicated in Fig. 3, the thickening of a hexagonal layer of TATB to a hexagonal crystal bulk is mainly attributed to

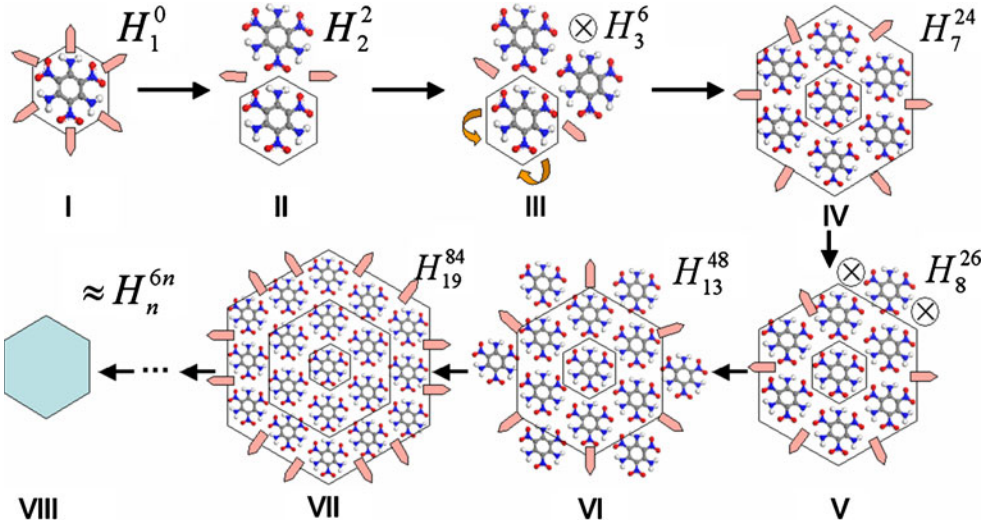


Fig. 9 Plots indicating the "birth and spread" mechanism of TATB to form a regular hexagon. Orange warheads point to the preferential growth orientations and ⊗ points to the next ones. The right superscript and subscript of letter H are the numbers of the intermolecular H-bonds and the molecules in a cluster, respectively. The process tends to form more intermolecular H-bonds relative to the finite number of molecules (6 1). Both the first molecule (**a**) and the last grown (**h**) have a regular hexagonal shape

the interlayer π-stacking interactions. Therefore, the cases of π-stacking interactions in TATB dimers were taken into account.

We firstly focus on the parallel π-stacking. As illustrated in Fig. 10a and b, the nitro-amino superposition π-stacking is much more energetically favored than the nitro-nitro one. The difference of the total energy between them is mainly resulted from the electrostatic interactions. Obviously, the nitro-nitro stacking causes more repulsion than the nitro-amino one. This can be examined in Fig. 10c by rotating the top TATB molecule along its molecular plane through its centroid: the electrostatic repulsion decreases when the rotation angle θ increases from 0 to 60°, just corresponding to the nitro-nitro and nitro-amino stacking, respectively. Considering that the strongest π-stacking interaction appears usually as an eclipsed one[27], we translated the top TATB molecule along an orientation of one of its C-Nitro (or C-Amino) bonds as indicated in Fig. 10d. As expected, the strongest π-stacking interaction takes place at D=−4.0 Å, showing an eclipsed π-stacking indeed, that is, double nitro-benzene stacking. Apparently, this eclipsed π-stacking interaction is the foundation of thickening a regular hexagonal root TATB molecule to a hexagonal bulk crystal. In the most energetically favored TATB dimer, the

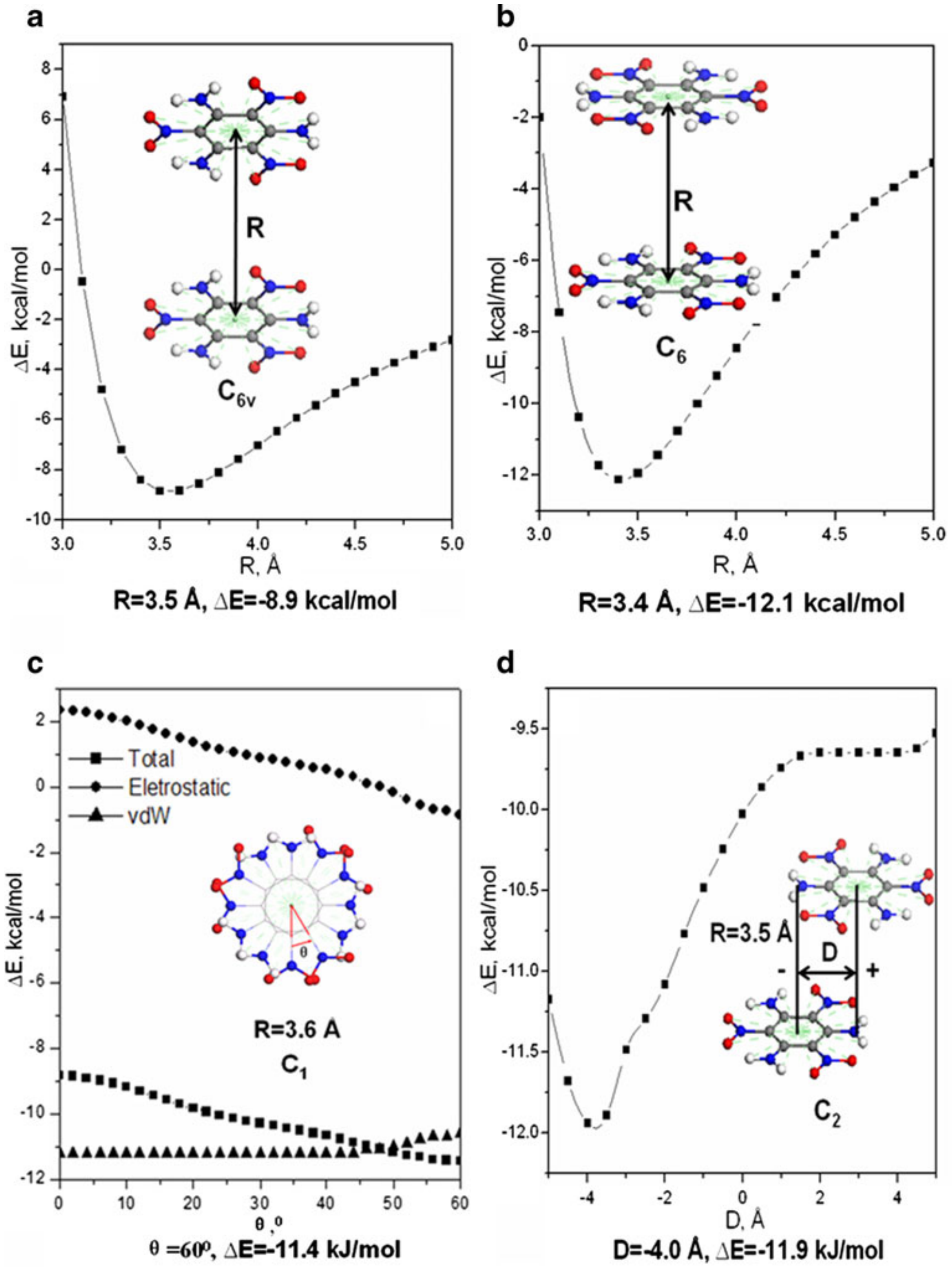


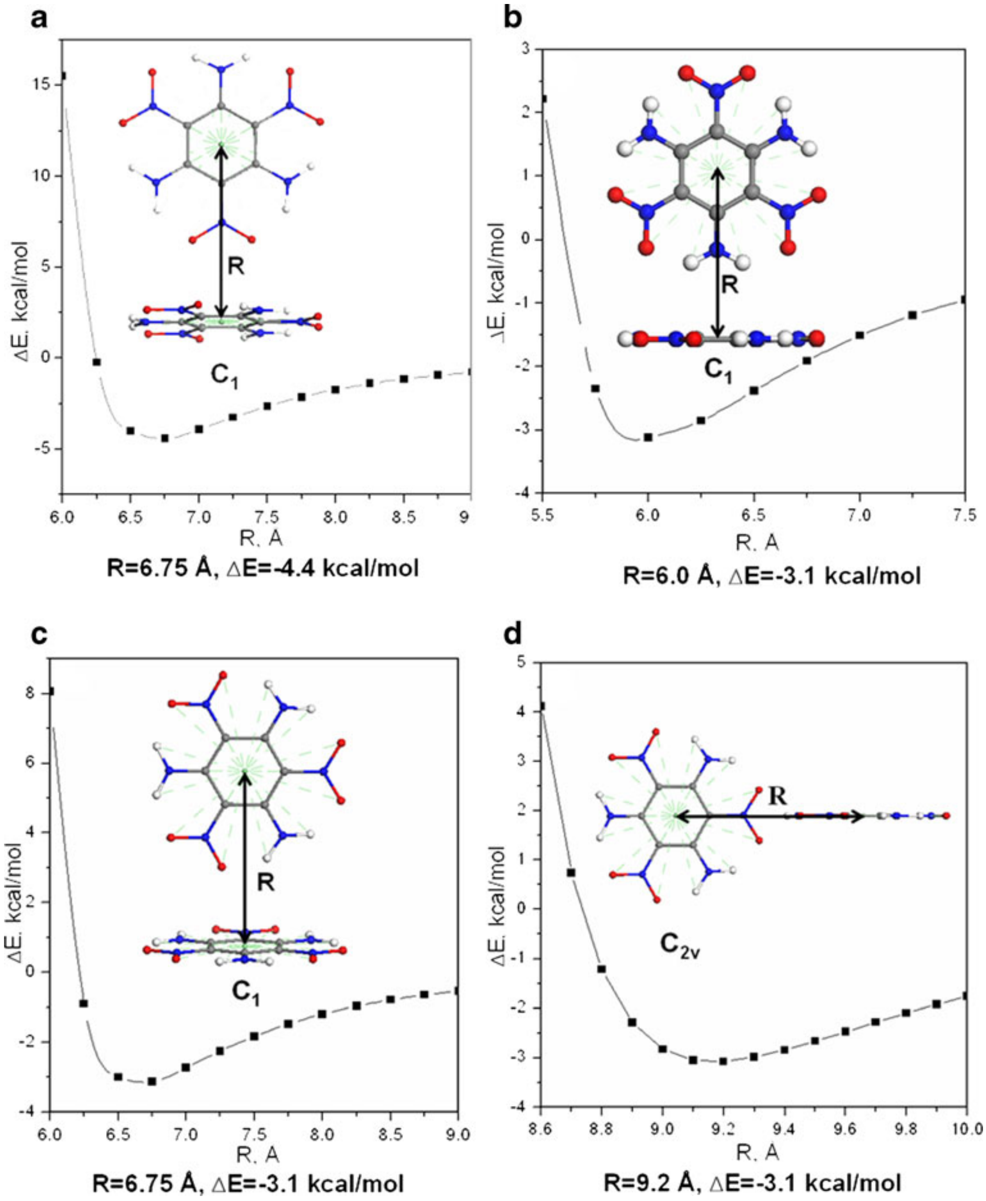
Fig. 10 Dimer interactions in the case of the parallel π-stacking. (**a**) and (**b**) are of the face-to-face π-stacking, in which only the intermolecular distances can be changed. Two TATB molecules are completely superposed in (**a**) (nitro-nitro stacking), and are superposed with 60^0 rotation of the above molecule along its molecular plane in (**b**) (nitro-amino stacking); (**c**) is of the face-to-face π-stacking too in which the intermolecular distance and one TATB molecule keeps fixed and another TATB molecule rotates around its centroid along the molecular plane; (**d**) is of eclipsed π-stacking, in which the intermolecular distance and the molecular orientations are fixed, and only the sliding distance D of the above molecule change. The negative D shows a left translation, and the positive D shows a right one

orientation of the regular hexagonal shape of a TATB molecule is not changed even though it rotates around its centroid 120° due to its symmetry of D_{3h}, equal to the thickening of a regular hexagon. That is, by translation, double hexagons can be completely superposed with each other.

Also, another case of the π-stacking, T-shaped stacking, was considered. As showed in Fig. 11a and b, both a nitro-π interaction (4.4 kcal mol^{-1}) and an amino-π attraction (3.1 kcal mol^{-1}) are less attractive than an eclipsed parallel π-stacking one (11.9 kcal mol^{-1}) in Fig. 10d. Considering the periodicity of molecular packing in crystal, we think that the two cases in Fig. 11a and b should appear simultaneously. Then, the average value of the attraction energy in the two cases, 3.8 kcal mol^{-1}, is much less than that of the parallel π-stacking, suggesting the molecular packing in crystal according to this style is impossible. Other two T-shaped stacking in Fig. 11c and d are also impossible due to the great disadvantage of thermodynamics. From above discussion of the π-stacking interactions, we can confirm that the strongest dimer attraction is of an eclipsed parallel π-stacking shown in Fig. 10d, which determines in principle the hexagonally thickening.

Otherwise, to be closer to the practice of crystal growth, we extended the hexagonal TATB molecule in π-stacking interactions to a hexagonal layer composed of many TATB molecules and examined whether the stacked layers have a consistent orientation. As illustrated in Fig. 12, if the stacked layers have different orientations, there will be no thickened hexagon, i.e., a hexagonal crystal bulk. As indicated in Fig. 13, we established a model to calculate the rotation barrier to examine the possibility of the stacked

Fig. 11 Dimer interactions in the case of T-shaped π-stacking. R is the distance between the centroids of double TATB molecules



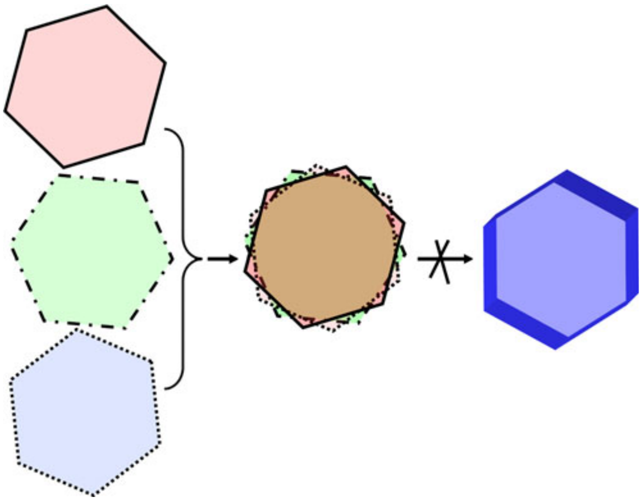


Fig. 12 A plot showing no hexagonal thickening if these hexagons are overlaid with different orientations

layers with different orientations: (1) construction of the layers. The hexagonal bottom layer is composed of 271 TATB molecules arranged like Fig. 9, that is, there are 10 TATB molecules on each edge of the layer; The top layer is similar to the bottom one, with no more TATB molecules than the bottom one. The distance between two neighboring TATB molecules in the layer is 9.0 Å (centroid-centroid distance) according to Fig. 7a; (2) arrangement of double layers. Make the double layers face-to-face stacked, and the centroids of double layers and two double-headed arrows in Fig. 13 overlaid, respectively. For interlayer distance, we selected two values: one is 3.5 Å in terms of the π-stacking in the crystal and its usual interplanar distance [27, 28]; and another is 3.008 Å considering the crystal density of benzene derivatives of CHON high energy materials usually less than 2 g/cm^{3} ^{4}, namely, 3.008 Å is an extreme value corresponding to the density of 2 g/cm^{3}; and (3) displacement and rotation. Fix the bottom layer and displace the top layer up in the figure plane 4 Å according to Fig. 10d, to simulate the top layer growing on the bottom layer. The rotation angle θ is zero this time as an initial case. Then rotate the top layer around its centroid clockwise. Calculate the total energy of every scanning step and obtain the rotation barrier, i.e., relative energy (RE) to the case of θ=0.

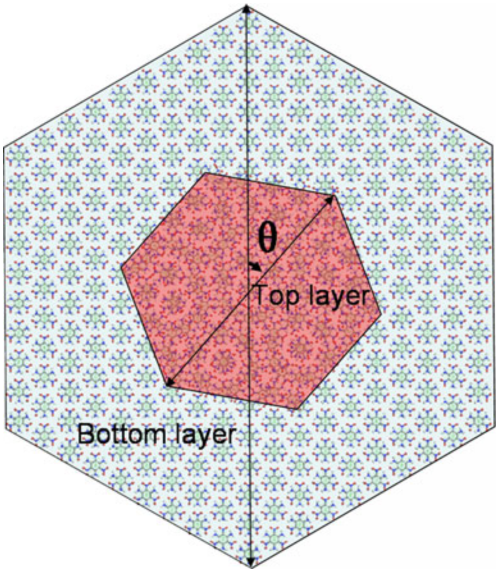


Fig. 13 Model for calculating the rotation barrier of the top layer relative to the bottom layer

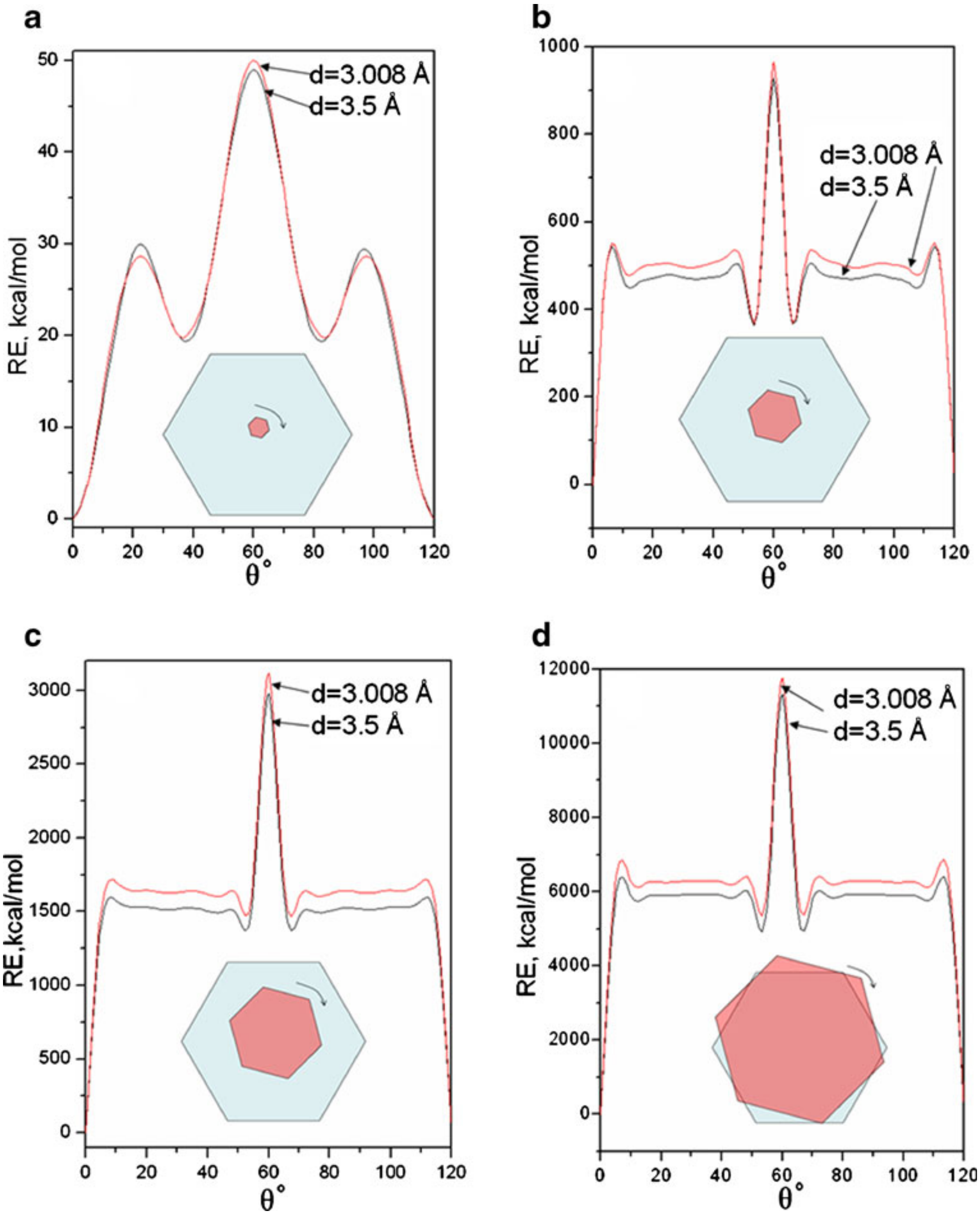
The calculated RE of four cases of each edge of the top layer containing 1, 3, 5 and 10 TATB molecules are shown in Figs. 14a–d, respectively. We can find from the figure similar results without exception: (1) there is a period of 120^{o} in the θ-RE curve, suggesting three possible orientations for the top layer precipitated on the bottom layer along; and (2) the smallest RE=0 occurs at θ=0 (or 120^{o}) and the largest occurs at θ=60^{o}, implying the orientation deviation of the top layer relative to the bottom layer is strongly thermodynamically forbidden, namely, the orientations of the two hexagons of double layers should be consistent with each other (θ=0 or 120^{o}). It should be attributed to the symmetry of the top layer, and in nature the symmetry of the TATB molecule.

Through above discussion, we can understand the regular or quasi-regular hexagonal bulk crystal shape from a regular hexagonal root molecule: (1) according to the assumed growth rule in Fig. 9, a regular hexagonal TATB cluster forms; (2) a TATB molecule stacks on the cluster plane as a birth, and step (1) repeats; (3) steps (1) and (2) repeat to form a thickened hexagon, which is actually a layer in final crystal packing as presented in Fig. 15c (similar to the visible crystal bulk in Fig. 1a and b) and 15 d (similar to the visible crystal bulk in Fig. 1c to h). And some non-uniformity in practical growth causes a quasi-regular hexagon. However, all the edge angles are always kept 120^{o} attributed to that the stacking is permissible only along a regular hexagonal orientation and other orientations are thermodynamically forbidden. In a word, in the mechanism of TATB crystallization, the birth and enlargement of the regular hexagon are thermodynamically and kinetically controlled, and the orientation of the hexagon is strongly thermodynamically controlled.

Conclusions

Taking a typical example, TATB, we have understood the relationship between molecular and crystal shapes by *only* the root TATB molecule, without its experimental crystal data or extensively applied models based on unit cell structures for predicting the crystal shape such as BFDH,

Fig. 14 Calculated barriers of the rotation scanning (RE). From a to d, the edges of the top hexagons (in red) are composed of 1, 3, 5 and 10 TATB molecules, respectively



equilibrium or attachment energy. These *withouts* imply the simplicity and legibility of our understanding, which is different from other conventional ones considering many influence factors. This may be attributed to the high rigidity

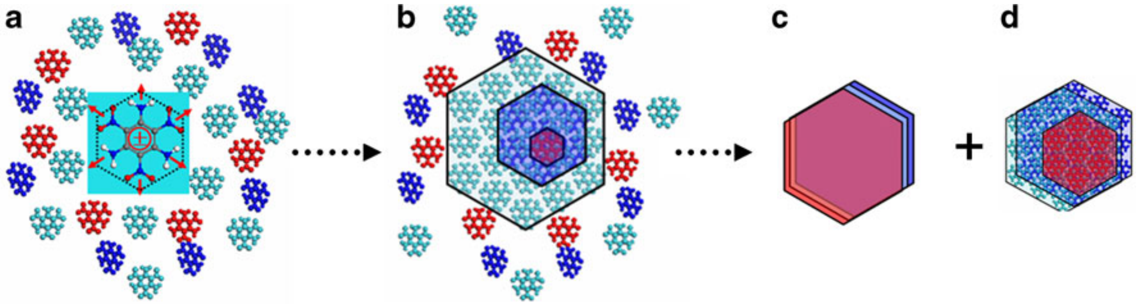


Fig. 15 Plots indicating the TATB growth process: (**a**) is the case before nucleation, (**b**) shows the nucleation and growth, (**c**) and (**d**) are a single hexagon layer and a multi hexagon layers stacked packing, corresponding to the practice illustrated in Fig. 1a and b, and c to h, respectively

and high symmetry of the TATB molecule, the strong and simple-patterned solute-solute interactions, and the weak and negligible solute-solvent interactions, that is, straightforward internal factors (effect of solute itself) but few external factors (effect of solvent). As a matter of fact, to understand well or predict a crystal shape is very challenging up to the present, due to the difficulties in accurately describing crystallization mechanism under different conditions. And the doubt of crystal predictions cannot be eliminated yet. Accurate predictions are based on good understandings. The understanding of TATB crystal shape is expected to be useful to other predictions.

Acknowledgments Financial support from the funds of China Academy of Engineering Physics (CAEP) (No 2011A0302014 and No 2010A0302012), the State Key Development Program for Basic Research of China (973–61383), and the National Natural Science Foundations of China (No 21173199 and No 10972025) is greatly appreciated.

References

1. Teipei U (2005) Energetic Materials. Wiley-VCH, Weinheim
2. Zhang C, Peng Q, Wang L, Wang X (2010) Thermal sensitivity of hmx crystals and hmx-based explosives treated under various conditions. Prop Explos Pyrotech 35:561–566, and references therein
3. Politzer P, Alper HE (1999) In: Computational chemistry: reviews of current trends. Leszczynski J (Ed) World Scientific, River Edge, NJ, pp 271–286 and references therein
4. Dong HS (2004) Strategy for developing energetic materials. Chin J Energ Mater 12:1–11, references therein
5. Lovette MA, Browning AR, Griffin DW, Sizemore JP, Snyder RC, Doherty MF (2008) Crystal shape engineering. Ind Eng Chem Res 47:9812–9833
6. Price SL (2008) From crystal structure prediction to polymorph prediction: interpreting the crystal energy landscape. Phys Chem Chem Phys 10:1996–2009
7. Berkovitch-Yellin Z (1985) Toward an ab initio derivation of crystal morphology. J Am Chem Soc 107:8239–8253
8. Docherty R, Clydesdale G, Roberts KJ, Bennema P (1991) Application of Bravais-Friedel-Donnay-Harker, attachment energy and Ising models to predicting and understanding the morphology of molecular crystals. J Phys D Appl Phys 24:89–99
9. Deij MA, Meekes H, Vlieg E (2007) The step energy as a habit controlling factor: application to the morphology prediction of aspartame, venlafaxine, and a yellow isoxazolone dye. Cryst Growth Des 7:1949–1957
10. Deij MA, Cuppen HM, Meekes H, Vlieg E (2007) Steps on surfaces in modeling crystal growth. Cryst Growth Des 7:1936–1942
11. Hammond RB, Pencheva K, Ramachandran V, Roberts KJ (2007) Application of grid-based molecular methods for modeling solvent-dependent crystal growth morphology: aspirin crystallized from aqueous ethanolic solution. Cryst Growth Des 7:1571–1574
12. Qiu SR, Wierzbicki A, Salter EA, Zepeda S, Orme CA, Hoyer JR, Nancollas GH, Cody AM, De Yoreo JJ (2005) Modulation of calcium oxalate monohydrate crystallization by citrate through selective binding to atomic steps. J Am Chem Soc 127:9036–9044
13. Zeman S (2003) In: Politzer P, Murray JS (eds) Energetic materials. Elsevier, Amsterdam, pp 25–52 and references therein
14. Voigt-Martin IG, Li G, Yakimanski A, Schulz G, Wolff JJ (1996) The Origin of nonlinear optical activity of 1,3,5-triamino-2,4,6-trinitrobenzene in the solid state: the crystal structure of a non-centrosymmetric polymorph as determined by electron diffraction. J Am Chem Soc 118:12830–12831
15. The TATB crystal was prepared as following: 4 g TATB offered by our institute was dissolved in 100 mL DMSO at 135 °C for 210 minutes; and then TATB was recrystallized during the temperature decreased naturally to the room temperature. The temperature controller of PROLINE-RP-845 of LAUDA, Germany, and the particle optical image workstation of VISION-218-D of Weison, were adopted for experiment and detection, respectively
16. See https://str.llnl.gov/June09/maxwell.html
17. Dunitz JD, Gavezzotti A (2005) Toward a quantitative description of crystal packing in terms of molecular pairs: application to the hexamorphic crystal system, 5-methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile. Cryst Growth Des 5:2180–2189
18. Gee RH, Roszak S, Balasubramanian K, Fried LE (2004) Ab initio based force field and molecular dynamics simulations of crystalline TATB. J Chem Phys 120:7059–7066
19. Roszak S, Gee RH, Balasubramanian K, Fried LE (2003) Molecular interactions of TATB clusters. Chem PhysLett 374:286–296
20. Song H, Xiao H, Dong H (2007) Intermoleuclar forces and gas geometries of tatb dimers. Acta Chimica Sinica 65:1101–1109
21. Delley B (1990) An all-electron numerical method for solving the local density functional for polyatomic molecules. J Chem Phys 92:508–517
22. Delley B (2000) From molecules to solids with the DMol3 approach. J Chem Phys 113:7756–7764
23. Mayo SL, Olafson BD, Goddard WA III (1990) Dreiding: a generic force field for molecular simulations. J Phys Chem 94:8897–8909
24. Sun H (1998) Compass: An ab initio force-field optimized for condensed-phase applicationsoverview with details on alkane and benzene compounds. J Phys Chem 102:7338–7364
25. Hagler AT, Lifson S, Dauber P (1979) Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 2. A benchmark for the objective comparison of alternative force fields. J Am Chem Soc 101:5122–5230
26. Rappe AK, Casewit CJ, Colwell KS, Goddard WA III, Skiff WM (1992) UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. J Am Chem Soc 114:10024–10035
27. Hunter CA, Sanders JKM (1990) The nature of π–π interactions. J Am Chem Soc 112:5525–5534
28. Zhang C (2011) Shape and size effects in π–π interactions: face-to-face dimers. J Comput Chem 32:152–161

J Mol Model (2012) 18:2257–2269
DOI 10.1007/s00894-011-1244-8

ORIGINAL PAPER

Molecular modeling study on the disassembly of dendrimers designed as potential antichagasic and antileishmanial prodrugs

Jeanine Giarolla · Kerly F. M. Pasqualoto · Daniela G. Rando · Márcio H. Zaim · Elizabeth I. Ferreira

Received: 8 June 2011 / Accepted: 13 September 2011 / Published online: 2 October 2011

Abstract A molecular modeling study was carried out to investigate the most likely enzymatic disassembly mechanism of dendrimers that were designed as potential antichagasic and antileishmanial prodrugs. The models contained *myo*-inositol (core), L-malic acid (spacer), and active agents such as 3-hydroxyflavone, quercetin, and hydroxymethylnitrofurazone (NFOH). A theoretical approach that considered one, two, or three branches has already been performed and reported by our research group; the work described herein focused on four (models A and B), five, or six branches, and considered their physicochemical properties, such as spatial hindrance, electrostatic potential mapping, and the lowest unoccupied molecular orbital energy (E_{LUMO}). The findings suggest that the carbonyl group next to the *myo*-inositol is the most promising ester breaking point.

Keywords Molecular modeling · Chagas disease · Leishmaniasis · Dendrimer disassembly · Dendrimeric prodrugs

J. Giarolla (✉) · K. F. M. Pasqualoto · M. H. Zaim · E. I. Ferreira
LAPEN, Department of Pharmacy, Faculty of Pharmaceutical Sciences, University of São Paulo (USP),
Av. Prof. Lineu Prestes, 580, Cidade Universitária,
São Paulo, SP, Brazil 05508-900
e-mail: jeaninegiarolla@yahoo.com.br

D. G. Rando
Department of Exact and Earth Sciences,
Federal University of São Paulo (UNIFESP),
Diadema, SP, Brazil

Introduction

The causative agent of the neglected tropical disease known as Chagas disease is the parasite *Trypanosoma cruzi*, which is transmitted to mammals by a bite from an insect vector [1]. This insect belongs to the family *Reduviidae* and subfamily *Triatominae*. The protozoan and the disease were described and discovered 101 years ago by the Brazilian scientist Carlos Chagas [2]. The infection occurs across the Americas, but especially in the region between the southern United States and southern Argentina and Chile. Almost 15 million people are currently infected and 90 million people are at risk of acquiring the disease [3]. Two drugs are available to treat Chagas disease, benznidazole and nifurtimox. However, the efficacies of both of these drugs are unclear in the chronic phase of the disease, they show high incidences of side effects, and they both also require long treatment times [4].

Almost 20 species of the protozoan genus *Leishmania* sp. can cause leishmaniasis, which is another neglected tropical disease, and is transmitted by the bite of a phlebotomine sandfly [5]. Clinical manifestations include cutaneous, mucocutaneous, and visceral leishmaniasis ("kala-azar;" fatal if not treated). Twelve million people are currently infected with this disease, and 60,000 die every year [6]. The disease is considered endemic to 88 countries, including countries in southern Europe, North Africa, the Middle East, Central and South America, and the Indian subcontinent. It is not, however, endemic to Southeast Asia and Australia [5, 7]. Unfortunately, the drugs that are available to treat it are toxic, expensive, and some of them require parenteral administration. These disadvantages can lead patients to abandon leishmaniasis treatment, resulting in the emergence of drug-resistant strains [5, 8–10]. Consequently, new chemotherapeutic agents against Chagas disease and leishmaniasis are urgently required.

Hydroxymethylnitrofurazone (NFOH), a nitrofurazone derivative synthesized in our laboratory, has proven to be a promising antichagasic compound; it presented enhanced in vitro activity [11] in trypomastigotes, especially amastigotes, in vivo activity in a murine model [12], and is four times less toxic than its prototype, benznidazole [13]. This derivative has stimulated several studies in our research group and in our co-workers' group. Flavonoids such as quercetin and 3-hydroxyflavone have also shown good in vitro activity against Chagas disease and leishmaniasis. However, in vivo, they were not active against leishmaniasis, or they presented a lower activity than miltefosine, which is the standard drug used in biological assays [14]. Therefore, it is necessary to design efficient transporters of these compounds in order to promote their use in therapeutic applications [15].

Prodrug design is a molecular modification process that is useful for improving drug characteristics; mainly their pharmaceutical, pharmacokinetic, and pharmacodynamic properties [16]. In this context, Chung and co-workers published a quite interesting review article in 2008 on twenty years of research in the field of the design of prodrugs for treating neglected and extremely neglected diseases [17].

A dendrimer is a unique class of synthetic molecules that provide significant control over the size, branching density, and surface functionality of molecules. These features imply that dendrimers are promising candidates for drug carriers [18]. Additionally, it has been observed that there is a better control over the release of a drug when it is covalently linked to a dendrimeric system, instead of forming a complex through encapsulation or electrostatic interactions [19, 20]. Dendrimer prodrug development may involve the following strategies: (1) the drug is linked directly to the dendrimer surface, establishing a covalent interaction for example [18, 21, 22]; (2) a linker molecule that is responsible for the interaction between branches is included; [18, 23] (3) each branch contains a drug molecule, leading to an exponential increase in the active agent with each subsequent generation [the patent on this approach was claimed by Giarolla and Ferreira in 2007 [24]); (4) drug molecules are bound to the dendritic structure through electrostatic, hydrophobic, and hydrogen-bonding interactions [21, 22]. There have already been studies that have explored the potential of dendrimers as prodrugs [25, 26].

The aim of this study was to apply molecular modeling methods as promising tools in a disassembly investigation of first-generation dendrimer prodrugs. The model prodrugs studied had four (models A and B), five, or six branches. They were composed of *myo*-inositol (the dendrimer core), L-malic acid (the spacer), and three potentially antichagasic and antileishmnaial bioactive agents: 3-hydroxyflavone, quercetin, and NFOH (see Fig. 1). Ester cleavage is probably performed enzymatically by nonspecific esterases. For this reason, the two carbonyl groups from L-malic acid were exploited. Similar preliminary studies have already been performed by our group for models containing one, two, or three branches [27].

Computational details

The computational procedure employed in this work was almost the same as that reported previously [27], but the dendrimers studied here had four (models A and B), five, or six branches. The three-dimensional (3D) structures of each dendrimer containing four, five, or six branches as well as the bioactive agents NFOH, 3-hydroxyflavone, and quercetin were constructed in their neutral forms using the HyperChem 7.51 software [28]. The crystallized structures retrieved from the Brookhaven Protein Data Bank (PDB) [29] and employed as standard geometries to build the 3-hydroxyflavone and quercetin models were 2g0l (NMR solution method) [30] and 1e8W (resolution 2.50 Å) [31], respectively. The NFOH 3D structure was constructed based on the crystallized structure of nitrofurazone (PDB entry code 1yki; resolution 1.70 Å) [32].

The energy of each model was minimized through the use of the MM+(molecular mechanics) force field (HyperChem 7.51 [28] and the MOLSIM 3.2 software package [33]), without any constrains. The MM+force field corresponds to the extended MM2 force field [34]. Partial atomic charges were calculated using the AM1 semiempirical method [35], also implemented in the HyperChem 7.51 program. The energy-minimization methods—steepest descent and conjugate gradient—were performed based on a set number of cycles or iterations for each procedure. The procedures were run sequentially, considering an energy convergence criterion. The energy-minimized models were used as initial structures to perform molecular dynamics (MD) simulations [2 ns; size step 0.001 ps at 300 K]. Trajectory files were recorded every 20 steps, resulting in 100,000 conformations for each model. A dielectric constant value of 3.5, which simulates the environment of the biological membranes, was used in the analysis of each model [36]. The hydration shell model proposed by Hopfinger [37] was employed to estimate the solvation energy contribution of the lowest energy conformation identified from the MD simulation, since the MOLSIM 3.2 software does not consider explicit water molecules during the MD simulations. Additionally, the hydrogen-bonding

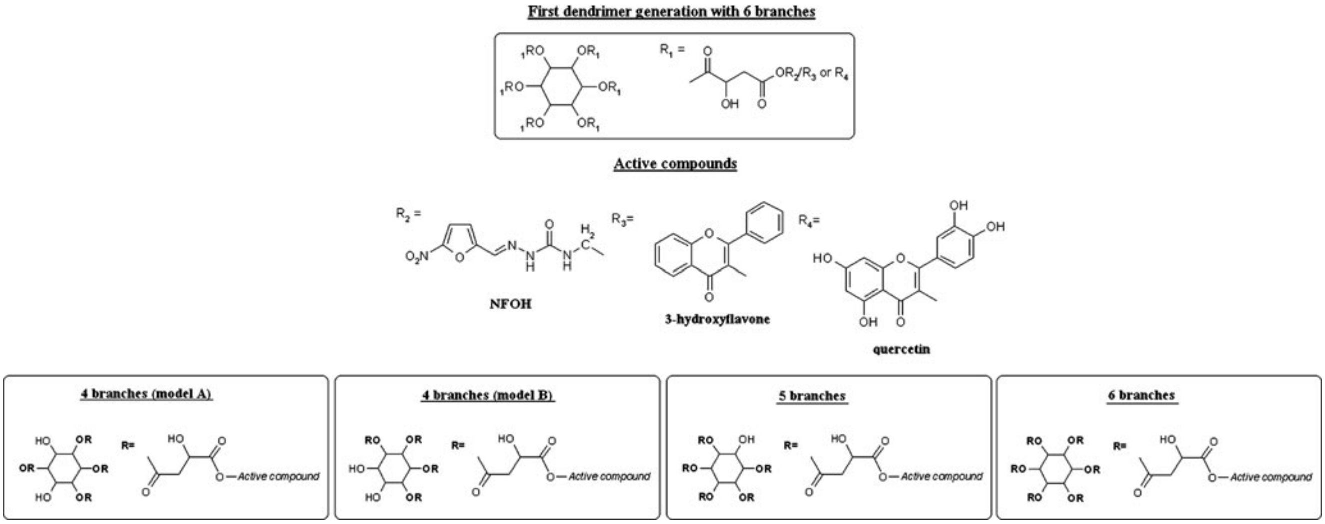


Fig. 1 Schematic representation of the dendrimer prodrugs studied in this work, which contained *myo*-inositol (core), L-malic acid (spacer), and the active agents (3-hydroxyflavone, quercetin, and NFOH). The skeletons of the systems containing four (models A and B), five, or six branches are also presented

energy contribution was computed for only the minimum energy conformers from MD simulations. The absence of explicit water molecules from the simulation analysis can lead to the generation and sampling of artifact states rather than the actual binding mode. On the other hand, the inclusion of explicit waters raises issues such as the assignment of water molecules and the degree of sampling needed to generate equilibrium/steady ensembles. These unknowns can also lead to artifact states. The use of an implicit solvation model as a hydration shell scheme appeared to be both a good compromise and a way to evaluate solvation effects [37].

The lowest energy conformation of each model was selected from MD simulation, its energy was minimized, and the electrostatic potential charges (ChelpG) were calculated using the ab initio method HF/3-21G* (Gaussian G03) [38]. The electrostatic potential (EP) of each model was mapped on a Connolly surface using a color ramp ranging from −8.5 to 8.5 e^{-2}. Negative values of EP (i.e., higher electronic densities) are depicted in red, while positive values (i.e., lower electronic densities) are shown in blue. The lowest unoccupied molecular orbital energy (E_{LUMO}) of each model was also computed and visualized using a range of −0.935 to 0.935 e^{-2}. Hydrogen bonds were displayed using the ViewerLite 4.2 program [39]. In this software, the bonds are shown between bond donors and acceptors throughout the entire molecule, and they are shown whether or not there are explicit hydrogen atoms. This can indicate intramolecular interactions that maintain the stability of the structure (branches), and can highlight intermolecular interactions with the aqueous medium solvent.

Results and discussion

As already mentioned, the two carbonyl groups from L-malic acid that were previously investigated [27] were also exploited in this study. These groups are assumed to be involved in dendrimer disassembly. The first carbonyl group is closer to the *myo*-inositol core, whereas the second is near the bioactive agent (3-hydroxyflavone, quercetin, or NFOH). The spatial hindrance, electronic density [map of electrostatic potential (MEP)], and LUMO distribution map were carefully evaluated. Those physicochemical properties may provide information indicating which of the moieties—the core or the bioactive agent—in the dendrimer system will be released first upon enzymatic action. A group presenting low spatial hindrance would be more likely to undergo enzymatic attack. The low electronic density and LUMO distribution on the carbonyl groups suggest that they are the most likely region to suffer an enzymatic nucleophilic attack.

The molecular electrostatic potential at a given point $p(x, y,z)$ in the vicinity of a molecule is the force acting on a positive test charge (a proton) located at p through the electrical charge cloud generated through the molecule's electrons and nuclei. Although the molecular charge distribution remains unperturbed by the presence of the external test charge (no polarization occurs), the electrostatic potential of the molecule is still a good guide to the molecule's reactivity towards positively or negatively charged reactants. This is typically visualized by mapping the electrostatic potential onto a surface that reflects the molecule's boundaries. This surface can be generated by overlapping the van der Waals radii of the molecule, through the use of algorithms that calculate the solvent-accessible surface of the molecule, or by employing a constant value for the electron density. The MEP can also be used to identify sites in the molecular system that act as proton donors or acceptors, or as nucleophiles or electrophiles, in terms of the drug–receptor molecular recognition process.

Table 1 presents the values of the total potential energy (E_{total}) obtained for the lowest energy models selected from MD simulations, the hydrogen-bonding energy contributions (E_{Hb}), the number of hydrogen bonds (Hb) in the selected models, and the E_{LUMO} values found for the models containing four (models A and B), five, or six branches plus the bioactive agent (3-hydroxyflavone, quercetin, or NFOH).

The E_{total} value corresponds to the sum of all energy contributions within each chosen model, such as those relating to stretching, bending, torsion, type 1–4 interactions (Lennard–Jones), van der Waals interactions, electrostatic interactions, hydrogen bonding, and solvation [36]. In general, the more negative the value of E_{total}, the more energetically favorable the system. However, more flexible molecular systems (i.e., those with more degrees of freedom) tend to present higher E_{total} values.

All dendrimer models containing NFOH as the bioactive agent were more energetically favorable (Table 1) than those with quercetin, except in the case of the model with four branches (model A). The six-branch dendrimers gave the most stable models for both quercetin and NFOH (E_{total}= −400.03 and −482.24 kcal mol^{-1}, respectively). These models also presented the most negative values for the E_{Hb} contribution (−571.63 and −512.31 kcal mol^{-1}, respectively). The number of intramolecular hydrogen bonds was 13 for both. The model with NFOH and four branches (model B) had a lower E_{total} value (−341.44 kcal mol^{-1}) and a higher E_{Hb} contribution (−361.79 kcal mol^{-1}) than the corresponding quercetin-based model. Additionally, it presented 11 intramolecular hydrogen bonds (see Table 1).

Otherwise, the set of models with 3-hydroxyflavone as the bioactive agent presented the highest E_{total} values (positive values), so this set was disregarded for further consideration.

The LUMO is a good indicator of electron-accepting ability. Therefore, E_{LUMO} indicates the ability of a molecule to accept electrons or act as an electrophile. The lower the E_{LUMO} value, the greater the ability of the system to act as an electron acceptor [40]. It is interesting to evaluate this property, for instance, when a nucleophilic enzymatic attack is needed to promote dendrimer disassembly. If we consider Table 1, all of the dendrimer models containing NFOH as the bioactive agent presented lower E_{LUMO} values than the respective models with 3-hydroxyflavone and quercetin.

The dendrimer models with four branches (model A) containing quercetin and NFOH as bioactive agents are presented in Fig. 2. The quercetin-based model showed the establishment of five intramolecular hydrogen bonds (bond lengths: 1.74, 2.16, 2.65, 2.85, and 2.86 Å). Moreover, in the CPK (Corey–Pauling–Koltun) or space-filling model, only minor spatial hindrance was noted at the carbonyl group near the core (*myo*-inositol). This feature renders this moiety more favorable to enzymatic attack. The same carbonyl group presented a region of low electronic density (green color), which can be visualized in the MEP. Therefore, the low density of electrons in this area could favor a nucleophilic enzymatic attack, which is needed for dendrimer disassembly. The LUMO distribution was

Table 1 E_{total}, E_{Hb}, and E_{LUMO} values, as well as the number of Hb found for the models containing dendrimers with four (models A and B), five, or six branches as well as the bioactive agent (3-hydroxyflavone, quercetin, or NFOH)

Dendrimer	E_{total} (kcal/mol)	E_{Hb} (kcal/mol)	Number of Hb	E_{LUMO} (kcal/mol)
Four branches (model A)				
3-Hydroxyflavone	82.10	−145.38	4	0.64
Quercetin	−257.65	−380.48	5	0.30
NFOH	−215.59	−245.53	11	−0.87
Four branches (model B)				
3-Hydroxyflavone	87.63	−118.14	2	0.97
Quercetin	−242.29	−365.17	8	0.74
NFOH	−341.44	−361.79	11	−0.90
Five branches				
3-Hydroxyflavone	116.10	−147.25	7	0.82
Quercetin	−390.95	−510.48	7	0.27
NFOH	−420.27	−450.54	13	−0.80
Six branches				
3-Hydroxyflavone	151.62	−138.31	5	0.49
Quercetin	−400.03	−571.63	13	0.36
NFOH	−482.24	−512.31	13	−0.83

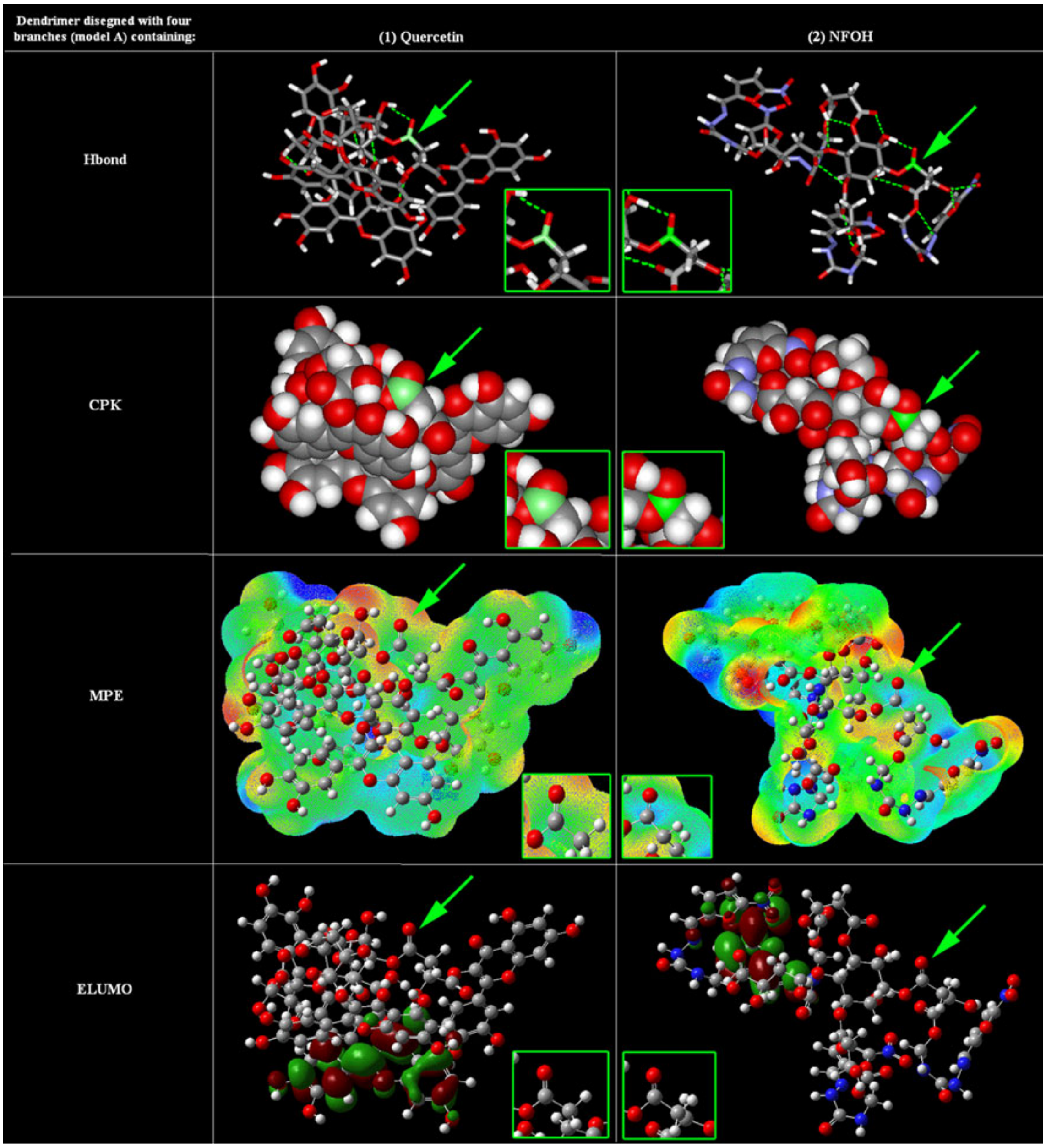


Fig. 2 The lowest energy conformations of the dendrimer with four branches (model A) containing (1) quercetin or (2) NFOH, as obtained from MD simulation. The intramolecular Hb (*green lines*) and the CPK or space-filling models are also presented (ViewerLite 4.2). The carbonyl group that shows the lowest spatial hindrance in the tube model and the CPK model is represented in *green* and also highlighted. MEPs are represented using a color ramp from −8.5 (*intense red*) to 8.5 (*intense blue*) e^{-2}, and the LUMO distribution uses a color ramp from −0.935 (*intense red*) to 0.935 (*intense blue*) e^{-2} (GaussView 3.0). The carbon atoms are shown in *gray*, oxygen in *red*, nitrogen in *blue*, and hydrogen atoms are depicted in *white*

investigated in the carbonyl group near quercetin. Although there is a favorable LUMO distribution in this region, it presents greater spatial hindrance than the carbon from the carbonyl group near the core. Thus, if we focus mainly on the effects of spatial hindrance and electronic distribution, it is clear that the *myo*-inositol will probably be released from the dendrimer system before the bioactive agent.

The NFOH model presented eleven intramolecular hydrogen bonds (1.82, 1.97, 2.20, 2.40, 2.44, 2.81, 2.89, 2.95, 2.99, 3.09, 3.19 Å) that appear to contribute to the conformational arrangement adopted by this system (see Fig. 2). The CPK or space-filling model indicated that the carbonyl group closest to the *myo*-inositol was the most likely to undergo enzymatic attack due to its low spatial hindrance. This finding corroborated the MEP analysis. The same carbonyl group presented a neutral/positive electronic density region (green/blue color), which would be an attractive area for nucleophilic enzymatic action and thus dendrimer disassembly. The LUMO distribution did not map to this carbonyl group.

Figure 3 presents the dendrimer models with four branches (model B) containing quercetin or NFOH. It is quite probable that for the model containing quercetin, dendrimer disassembly will occur in the carbonyl group near the core. The findings from the CPK and tube models and MEP indicate this. Regarding the CPK model, the carbonyl group near *myo*-inositol showed the lowest spatial hindrance, which is an important feature of any enzymatic

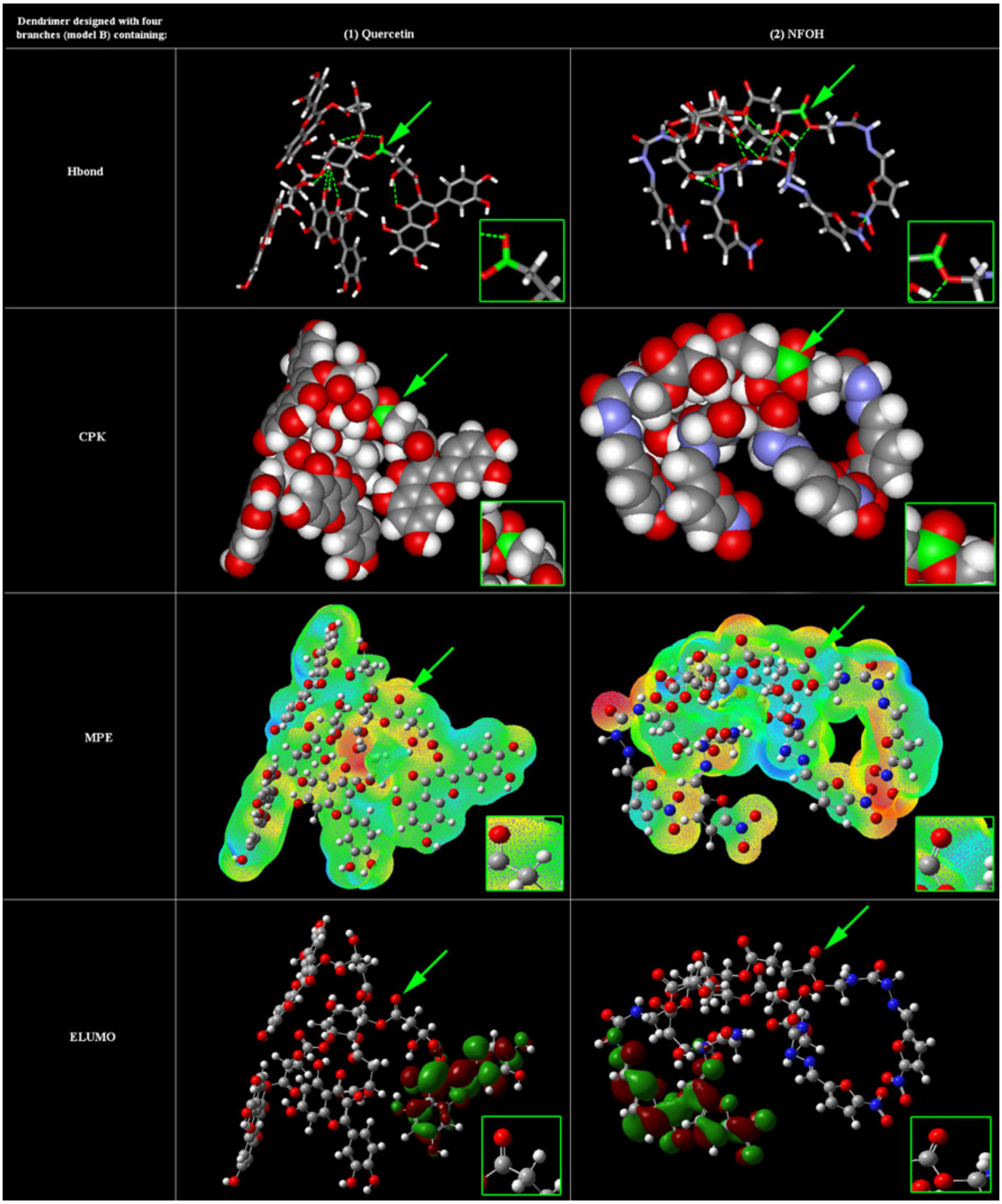


Fig. 3 The lowest energy conformations of the dendrimer with four branches (model B) containing (1) quercetin or (2) NFOH, as obtained from MD simulation. The intramolecular Hb (*green lines*) and the CPK or space-filling models are also presented (ViewerLite 4.2). The carbonyl group that shows the lowest spatial hindrance in the tube model and the CPK model is represented in *green* and also highlighted. MEPs are represented using a color ramp from −8.5 (*intense red*) to 8.5 (*intense blue*) e^{-2}, and the LUMO distribution uses a color ramp from −0.935 (*intense red*) to 0.935 (*intense blue*) e^{-2} (GaussView 3.0). The carbon atoms are shown in *gray*, oxygen in *red*, nitrogen in *blue*, and hydrogen atoms are depicted in *white*

attack needed for dendrimer disassembly. Moreover, in the tube model, the formation of eight intramolecular hydrogen bonds (2.09, 2.19, 2.20, 2.47, 2.64, 2.87, 2.95, 3.00 Å) that are relevant to the model's stability and conformational

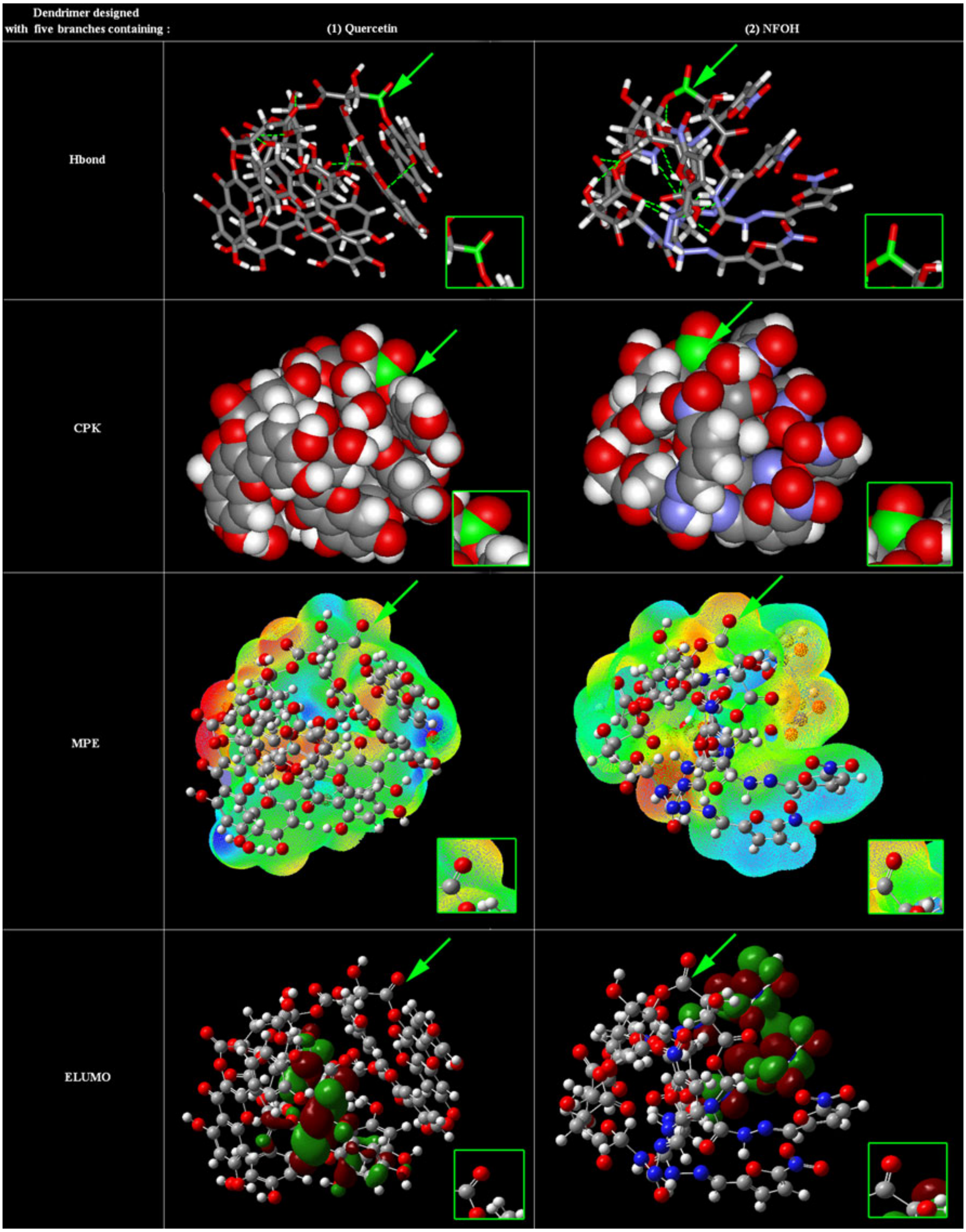


Fig. 4 The lowest energy conformations of the dendrimer with five branches containing (1) quercetin or (2) NFOH, as obtained from MD simulation. The intramolecular Hb (*green lines*) and the CPK models are also presented (ViewerLite 4.2). The carbonyl group that shows the lowest spatial hindrance in the tube model and the CPK or space-filling model is represented in *green* and also highlighted. MEPs are represented using a color ramp from −8.5 (*intense red*) to 8.5 (*intense blue*) e^{-2}, and the LUMO distribution uses a color ramp from −0.935 (*intense red*) to 0.935 (*intense blue*) e^{-2} (GaussView 3.0). The carbon atoms are shown in *gray*, oxygen in *red*, nitrogen in *blue*, and hydrogen atoms are depicted in *white*

arrangement was observed. The green color for the MEP on the this carbonyl carbon indicates a neutral area, which means that it is a region that could suffer an enzymatic nucleophilic attack. Although the LUMO was visualized in the other carbonyl group, all the other findings discussed here suggest that the ester break point will probably occur in a carbonyl near the core.

The corresponding dendrimer model designed with NFOH showed that the carbonyl group near the bioactive agent is the most likely of these groups to suffer an

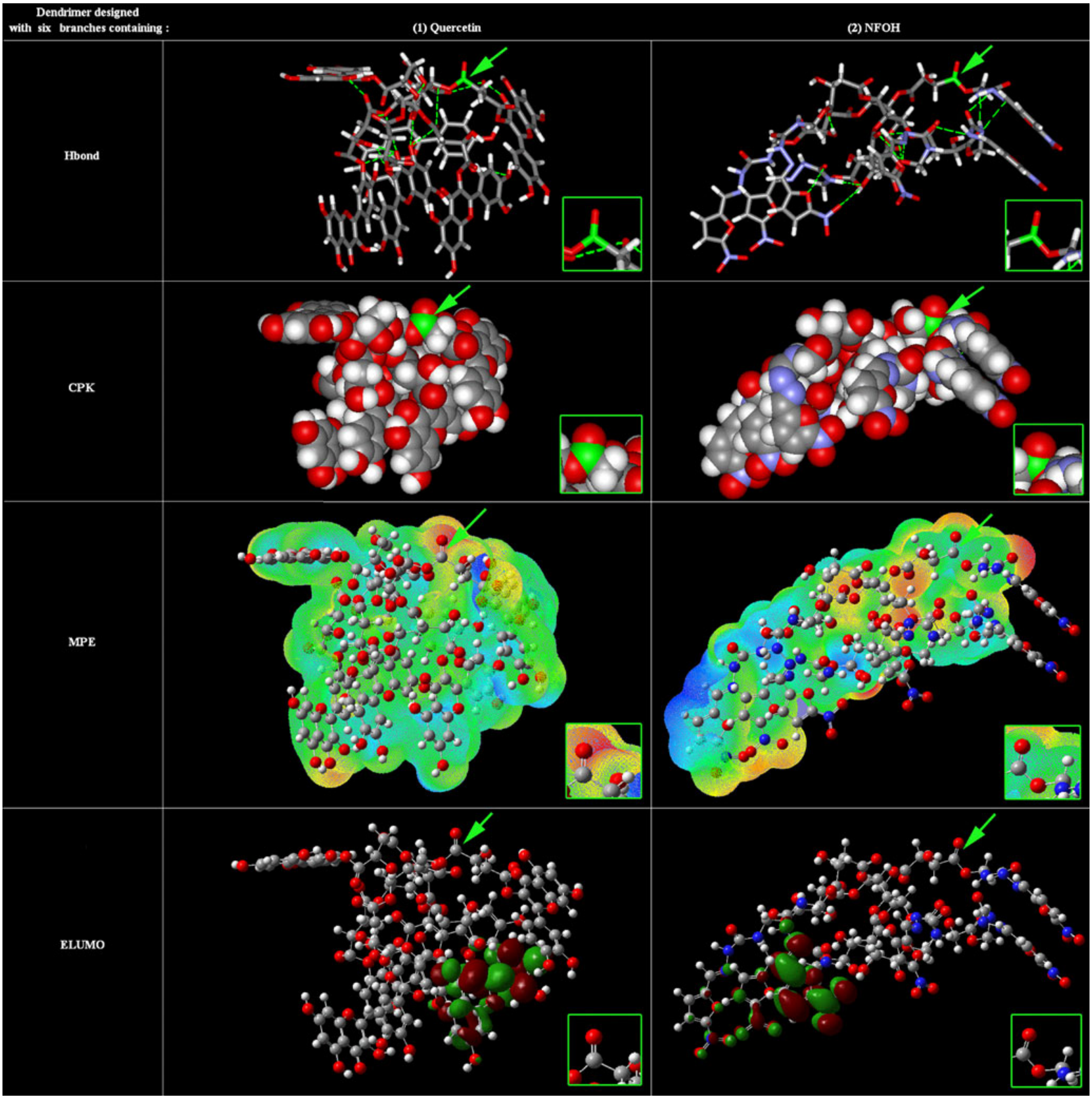
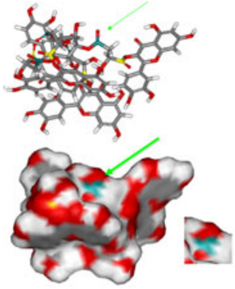
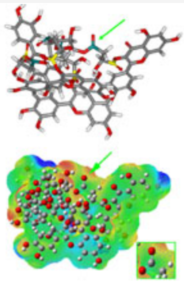
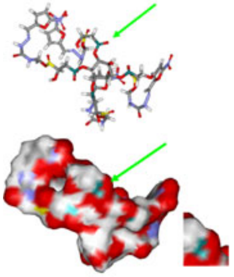
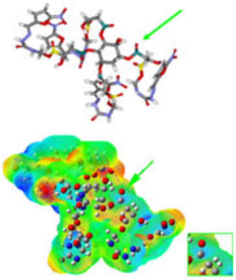
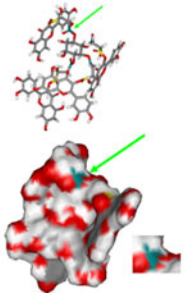
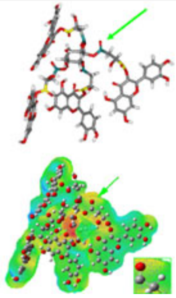


Fig. 5 The lowest energy conformations of the dendrimer with six branches containing (1) quercetin or (2) NFOH, as obtained from MD simulation. The intramolecular Hb (*green lines*) and the CPK models are also presented (ViewerLite 4.2). The carbonyl group that shows the lowest spatial hindrance in the tube model and the CPK or space-filling model is represented in *green* and also highlighted. MEPs are represented using a color ramp from −8.5 (*intense red*) to 8.5 (*intense blue*) e^{-2}, and the LUMO distribution uses a color ramp from −0.935 (*intense red*) to 0.935 (*intense blue*) e^{-2} (GaussView 3.0). The carbon atoms are shown in *gray*, oxygen in *red*, nitrogen in *blue*, and hydrogen atoms are depicted in *white*

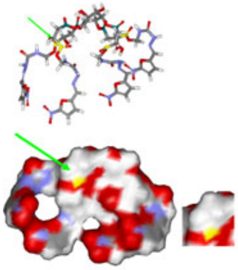
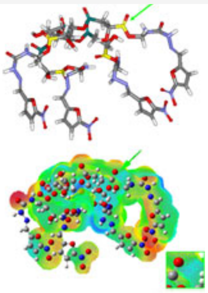
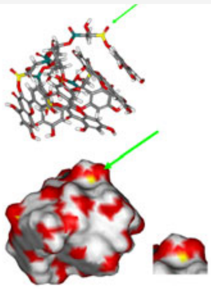
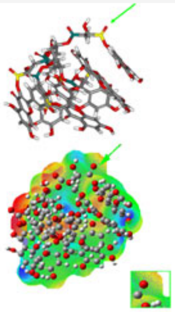
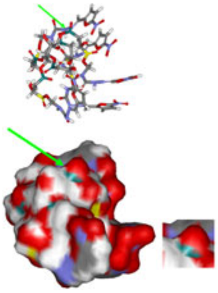
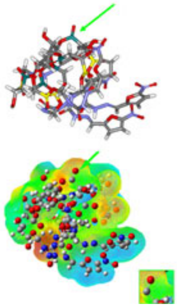
Table 2 Most elucidative physicochemical properties in relation to the region most likely to undergo an enzymatic nucleophilic attack, as evaluated for the dendrimers containing quercetin or NFOH with four (models A and B), five, or six branches

Dendrimer	Spatial hindrance (solvent-accessible molecular surface)	Electrostatic potential (MEP)	Carbonyl C more susceptible to the enzymatic attack
	Four branches (model A)		
Quercetin			near the core
NFOH			near the core
	Four branches (model B)		
Quercetin			near the core

enzymatic attack, considering the findings from the CPK and tube models as well as the MEP distribution. The lower spatial hindrance in the region of this carbonyl carbon can be visualized in the CPK model. Eleven intramolecular hydrogen bonds can be observed in the tube model (1.82, 2.10, 2.11, 2.14, 2.60, 2.66, 3.00, 3.05, 3.02, 3.08, 3.17 Å). Moreover, a green/blue (neutral/positive) color in this area of the MPE distribution indicates that enzymatic hydrolysis will probably occur in this region. However, controversially, the LUMO distribution did not occur in any of the L-malic acid carbonyl groups.

The five-branch dendrimer models containing quercetin or NFOH are presented in Fig. 4. In the model containing quercetin, the carbonyl near the bioactive agent appears to be a more likely group to suffer enzymatic action. The tube and CPK models as well as the MPE distribution explain this assumption. Seven intramolecular hydrogen bonds can be visualized in the tube model (1.98, 2.27, 2.41, 2.63, 2.98, 2.99, 3.09 Å). As discussed before, these intramolecular interactions contribute to the conformational arrangement of the system. Moreover, the carbon near quercetin presented the lowest steric hindrance, as seen in the CPK

Table 2 (continued)

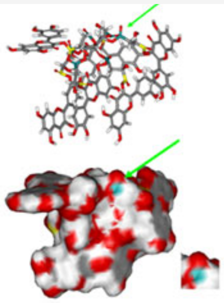
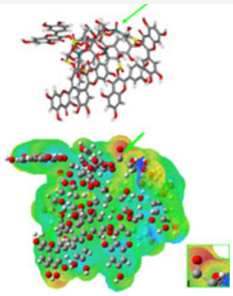
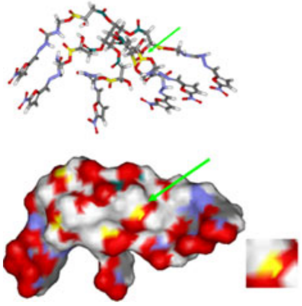
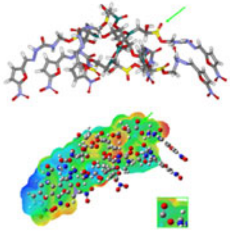
Dendrimer	Spatial hindrance (solvent-accessible molecular surface)	Electrostatic potential (MEP)	Carbonyl C more susceptible to the enzymatic attack
	Four branches (model B)		
NFOH			near bioactive agent
	Five branches		
Quercetin			near bioactive agent
NFOH			near the core

model. The same carbonyl group represented a neutral region (green color) in the MEP distribution, indicating that nucleophilic attack can occur in this portion of the system. The LUMO distribution was seen in another carbonyl group.

The dendrimer model containing NFOH indicated that the carbonyl carbon near *myo*-inositol is the most likely to suffer an enzymatic attack. The spatial disposition of this carbonyl group can be analyzed mainly using the tube and CPK models (see Fig. 4). Thirteen intramolecular hydrogen bonds were noted (1.99, 2.31, 2.67, 2.75, 2.77, 2.87, 2.90, 2.97, 3.03, 3.08, 3.15, 3.19, 3.20 Å). In addition, the same group represented a neutral region (green color) in the MEP distribution, indicating that it could suffer the nucleophilic attack needed for dendrimer disassembly. The LUMO distribution did not contribute to these findings, as it did not occur in the carbonyl group of interest.

In the dendrimer model with six branches containing quercetin, the carbonyl group near *myo*-inositol seemed to be the most likely to suffer an enzymatic approach and, consequently, a nucleophilic attack. Accordingly, in this particular model, dendrimer disassembly will probably start in the core, with the bioactive agents released afterwards. The CPK models as well as the MPE distributions can be

Table 2 (continued)

Dendrimer	Spatial hindrance (solvent-accessible molecular surface)	Electrostatic potential (MEP)	Carbonyl C more susceptible to the enzymatic attack
	Six branches		
Quercetin			near the core
NFOH			near bioactive agent

* Solvent-accessible surface area: carbonyl carbons near the core are in green; carbonyl carbons near the bioactive agent are in yellow (ViewerLite 4.2); MEPs: regions colored in green/blue indicate neutral/positive electronic density distribution (GaussView 3.0)

visualized in Fig. 5, and they confirm these assumptions. The models show that the lowest spatial hindrance is presented by the carbonyl carbon near *myo*-inositol; this is a fundamental feature of any enzymatic approach. There are also thirteen intramolecular hydrogen-bond interactions (2.12, 2.20, 2.31, 2.56, 2.65, 2.68, 2.72, 2.93, 3.05, 3.09, 3.12, 3.14, 3.16 Å). Moreover, in the MPE, the corresponding carbonyl group showed a green area, meaning that it is a neutral region, making it likely to suffer from enzymatic action. The LUMO distribution found for the dendrimer model with quercetin was in another carbonyl group, near the active agent. The corresponding model with NFOH presented thirteen intramolecular hydrogen bonds (1.89, 2.01, 2.79, 2.79, 2.83, 2.85, 2.86, 2.88, 2.87, 3.05, 3.06, 3.12, 3.15 Å). In the tube and CPK models, the lowest spatial hindrance was seen in the carbonyl group next to the active agent. Moreover, these carbonyl groups were a green/blue color (neutral/positive electron density) in the MEP distribution, indicating that they are the most likely to suffer a nucleophilic attack (see Fig. 5). The LUMO distribution of the NFOH-based dendrimer model was not seen in any of the groups investigated in this study.

The most important findings of this study for the investigated dendrimer systems containing quercetin or NFOH are summarized in Table 2. This table considers the most elucidative physicochemical properties in relation to defining the region most likely to suffer an enzymatic nucleophilic attack (carbonyl carbons near the core or near the bioactive agent). The spatial hindrance can be visualized through the solvent-accessible surface area (carbonyl carbons near the core are shown in green; carbonyl carbons near the bioactive agent are shown in yellow), whereas the electrostatic potential can be seen through the MEPs (regions colored in green/blue indicate neutral/positive electronic density distributions).

Conclusions

The molecular modeling study presented herein can be considered an important assessment, as it demonstrates the pre-disassembly behavior of dendrimers designed as potential antichagasic and antileishmanial prodrugs.

The molecular models were primarily analyzed in terms of physicochemical properties such as the spatial hindrance, the electrostatic potential map, and the lowest unoccupied molecular orbital energy. Based on the theoretical findings, the region most likely to suffer an enzymatic nucleophilic attack was determined. The carbonyl group next to the *myo*-inositol seems to be the most promising candidate for the point at which the ester breaks during dendrimer disassembly in the systems that contain quercetin as the bioactive agent, except for the five-branch dendrimer system. The dendrimers containing NFOH, on the other hand, show different release behavior for each size category.

The synthesis of these molecular models and release studies of them are currently being carried out, in order to hopefully validate the theoretical results obtained so far.

Acknowledgments The authors are grateful to the São Paulo Research Foundation (FAPESP) (01/01192-3, 06/00116-5, 07/59416-0, 08/54211-4, and 07/5461-0) and The National Council for Scientific and Technological Development (CNPq) for their financial support, and to the Chem21 Group, Inc., for the MOLSIM 3.2 academic license.

References

1. Coura JR, Dias JCP (2009) Epidemiology, control and surveillance of Chagas disease: 100 years after its discovery. Mem Inst Oswaldo Cruz 104:31–40
2. Coura JR, Viñas PA (2010) Chagas disease: a new worldwide challenge. Nature 465:S6–S7
3. Clayton J (2010) Chagas disease 101. Nature 465:S4–S5
4. Clayton J (2010) Chagas disease: pushing through the pipeline. Nature 465:S12–S15
5. Lindoso JAL, Lindoso AABP (2009) Neglected tropical disease in Brazil. Rev Inst Med Trop S Paulo 51:247–253
6. Neuber H (2008) Leishmaniasis. J Dtsch Dermatol Ges 6:754–765
7. Herwaldt BL (1991) Leishmaniasis. Lancet 354:1191–1199
8. Yardley V, Khan AA, Martin MB, Slifer TR, Araujo FG, Moreno SNJ, Docampo R, Croft SL, Oldfield E (2002) In vivo activities of farnesyl pyrophosphate synthase inhibitors against *Leishmania donovani* and *Toxoplama gondii*. Antimicrob Agents Chemother 46:929–931
9. Singh S, Sivakumar RJ (2004) Challenges and new discoveries in the treatment of leishmaniasis. Infect Chemother 10:307–315
10. Santos DO, Coutinho CER, Madeira MF, Bottino CG, Vieira RT, Nascimento SB, Bernardino A, Bourguignon SC, Corte-Real S, Pinho RT, Rodrigues CR, Castro HC (2008) Leishmaniasis treatment—a challenge that remains: a review. Parasitol Res 103:1–10
11. Chung MC, Güido RVC, Martinelli TF, Gonçalves MF, Polli MC, Botelho KCA, Varanda EA, Colli W, Miranda MTM, Ferreira EI (2003) Synthesis and in vitro evaluation of potential antichagasic hydroxymethylnitrofurazone (NFOH-121): a new nitrofurazone produg. Bioorg Med Chem 11:4779–4783
12. Davies C, Cardoso RM, Negrette OS, Mora MC, Chung MC, Basombrío MA (2010) Hydroxymethylnitrofurazone is active in a murine model of Chagas disease. Antimicrob Agents Chemother 54:3584–3589
13. Güido RVC, Ferreira EI, Nassute JC, Varanda EA, Chung MC (2001) Diminuição da atividade mutagênica do pró-fármaco NFOH-121 em relação ao NF (nitrofurazona). Rev Cienc Farm 22:319–333
14. Tasdemir D, Kaiser M, Brun R, Yardley V, Schmidt TJ, Tosun F, Ruedi P (2006) Antitrypanosomal and antileishmanial activities of flavonoids and their analogues: in vitro, in vivo, structure–activity relationship, and quantitative structure–activity relationship studies. Antimicrob Agents Chemother 50:1352–1364
15. Kayser O, Kiderlen EAF, Croft ESL (2003) Natural products as antiparasitic drugs. Parasitol Res 90:55–62
16. Chung MC, Silva ATA, Castro LF, Guido RVC, Nassute JC, Ferreira EI (2005) Latenciação e formas avançadas no transporte de fármacos. Rev Bras Cienc Farm 41:155–179
17. Chung MC, Ferreira EI, Santos JL, Giarolla J, Rando DG, Almeida AE, Bosquesi PL, Menegon RN, Blau L (2008) Prodrugs for the treatment of neglected diseases. Molecules 13:616–677
18. Svenson S (2009) Dendrimers as versatile platform in drug delivery applications. Eur J Pharm Biopharm 71:445–462
19. Aulenta F, Hayes W, Rannard S (2003) Dendrimers: a new class of nanoscopic containers and delivery devices. Eur Polym J 39:1741–1771
20. Uhrich K, Cannizzaro S, Langer R, Shakesheff K (1999) Polymeric systems for controlled drug release. Chem Rev 99:3181–3198
21. Liu M, Fréchet JMJ (1999) Designing dendrimers for drug delivery. Pharm Sci Technol Today 2:393–401
22. D'Emanuele A, Attwood D (2005) Dendrimer–drug interactions. Adv Drug Deliv Rev 57:2147–2162
23. Najlah M, Freeman S, Attwood D, D'Emanuele A (2007) In vitro evaluation of endrimer prodrugs for oral drug delivery. Int J Pharm 336:183–190
24. Giarolla J, Ferreira EI (2007) Patent 018070068714 (under analysis)
25. Lee CC, Gillies ER, Fox ME, Guillaudeu SJ, Fréchet JMJ, Dy EE, Szoka FC (2006) A single dose of doxorubicin-functionalized bow-tie dendrimer cures mice bearing C-26 colon carcinomas. Proc Natl Acad Sci USA 103:16649–16654
26. D'Emanuele A, Jevprasesphant R, Penny J, Attwood D (2004) The use of a dendrimer-propranolol prodrug to bypass efflux transporters and enhance oral bioavailability. J Control Rel 95:447–453
27. Giarolla J, Rando DG, Pasqualoto KFM, Zaim MH, Ferreira EI (2010) Molecular modeling as a promising tool to study dendrimer prodrugs delivery. J Mol Struct THEOCHEM 939:133–138
28. Hypercube Inc. (2002) HYPERCHEM, v.7.0 for Windows. Hypercube Inc., Gainesville
29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2011) The Protein Data Bank. http://www.pdb.org
30. Caddick S, Muskett FW, Stoneman RG, Woolfson DN (2006) Synthetic ligands for apo-neocarzinostatin. J Am Chem Soc 128:4204–4205
31. Walker EH, Pacold ME, Perisic O, Stephens L, Hawkins PT, Wvmann MP, Williams RL (2000) Structural determinants of phosphoinositide 3-kinase inhibition by wortmannin, LY294002, quercetin, myricetin, and staurosporine. Molecular Cell 6:909–919
32. Race PR, Lovering AL, Green RM, Ossor A, White SA, Searle PF, Wrighton CJ, Hyde EI (2005) Structural and mechanistic studies of *Escherichia coli* nitroreductase with the antibiotic nitrofurazone. Reversed binding orientations in different redox states of the enzyme. J Biol Chem 280:13256–13264
33. Doherty A (1997) MOLSIM: molecular mechanics and dynamics simulation software—user's guide, version 3.2. The Chem21 Group Inc., Lake Forest
34. Allinger NL (1977) Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. J Am Soc 99:8127–8134

35. Dewar MJS, Healy EF, Holder AJ, Yuan YC (1990) Comments on a comparison of AM1 with the recently developed PM3 method. J Comput Chem 11:541–542

36. Tokarski JS, Hopfinger AJ (1997) Prediction of ligand–receptor binding thermodynamics by free energy force field. J Chem Inf Comput Sci 37:792–811

37. Forsythe KH, Hopfinger AJ (1973) The Influence of solvent on the secondary structures of poly(L-alanine) and poly(L-proline). Macromolecules 6:423–437

38. Gaussian Inc. (1995–2003) Gaussian 03W for Windows, v.6. Gaussian Inc., Pittsburgh

39. Accelrys Inc. (2001) ViewerLite 4.2. Accelrys Inc., San Diego

40. McNaught AD, Wilkinson A, IUPAC (1997) Compendium of chemical terminology, 2nd edn (the “Gold Book”). Blackwell, Oxford (on-line corrected version: http://goldbook.iupac.org; created by Nic M, Jirat J, Kosata B, updates compiled by Jenkins A, ISBN 0-9678550-9-8; doi:10.1351/goldbook, 2006)

DOI 10.1007/s00894-011-1248-4

ORIGINAL PAPER

Combined DFT and BS study on the exchange coupling of dinuclear sandwich-type POM: comparison of different functionals and reliability of structure modeling

Bing Yin · GangLin Xue · JianLi Li · Lu Bai ·
YuanHe Huang · ZhenYi Wen · ZhenYi Jiang

Received: 3 February 2011 / Accepted: 20 September 2011 / Published online: 6 October 2011

Abstract The exchange coupling of a group of three dinuclear sandwich-type polyoxomolybdates $[MM'(AsMo_7O_{27})_2]^{12-}$ with MM'=CrCr, FeFe, FeCr are theoretically predicted from combined DFT and broken-symmetry (BS) approach. Eight different XC functionals are utilized to calculate the exchange-coupling constant J from both the full crystalline structures and model structures of smaller size. The comparison between theoretical values and accurate experimental results supports the applicability of DFT-BS method in this new type of sandwich-type dinuclear polyoxomolybdates. However, a careful choice of functionals is necessary to achieve the desired accuracy. The encouraging results obtained from calculations on model structures highlight the great potential of application of structure modeling in theoretical study of POM. Structural modeling may not only reduce the computational cost of large POM species but also be able to take into account the external field effect arising from solvent molecules in solution or counterions in crystal.

Keywords BS · DFT · Exchange coupling · POM · Structure modeling

B. Yin (✉) · G. Xue · J. Li · L. Bai · Z. Wen
Key Laboratory of Synthetic and Natural Functional Molecule Chemistry of Ministry of Education, College of Chemistry and Materials Science, Shaanxi Key Laboratory of Physico-Inorganic Chemistry, Northwest University,
Xi'an 710069, People's Republic of China
e-mail: rayinyin@gmail.com

B. Yin · Z. Wen · Z. Jiang
Institute of Modern Physics, Northwest University,
Xi'an 710069, People's Republic of China

Y. Huang
College of Chemistry, Beijing Normal University,
Beijing 100875, People's Republic of China

Introduction

Polyoxometalate (POM), also known as polyoxoanion, consists of plentiful polynuclear metal-oxygen clusters possessing high versality of structural and topological characters [1]. Due to its unique physical and chemical properties, POM exhibits great potential of application in many different areas of science and technology, e.g., medicine [2], catalysis [3], chemical analysis [4], etc. Therefore POM has caught extensive interests of scientists and researchers [1–6]. In the field of material science, the importance of POM as a suitable candidate of building block for the molecule-based materials with desirable multiple functions has been emphasized by the increasing number of recent research works [2, 5, 6].

Among the different types of POMs, nonclassical POMs of sandwich-like structures constitute a great subclass, which has received much attention in recent years [7–9]. Unlike sandwich-type polyoxotungstates, sandwich-type polyoxomolybdates are scarce because of the great difficulty in synthesis. Recently, our group reported the preparation and experimental characterization of a series of dinuclear sandwich-type heteropolymolybdates [10, 11]: $[MM'(AsMo_7O_{27})_2]^{12-}$ with MM'=CrCr, FeFe, FeCr. These POMs are abbreviated as Cr_2, Fe_2 and FeCr respectively here after. As shown in Fig. 1, the structural characters of these new sandwich-type POMs can be summarized $[MM'As_2O_{14}]^{16-}$, as the inclusion species, is sandwiched between two molybdenum oxide fragments $[Mo_7O_{20}]^{6+}$. Due to the existence of transition metal (TM) ions Cr(III) and Fe(III), the $[MM'As_2O_{14}]^{16-}$ unit may be paramagnetic. On the other side, the peripheral $[Mo_7O_{20}]^{2+}$ fragments, which contain only Mo(VI) and oxo ligands, are assumed to be diamagnetic.

Besides the novel structural features of these new sandwich-type POMs [10], remarkable magnetic property,

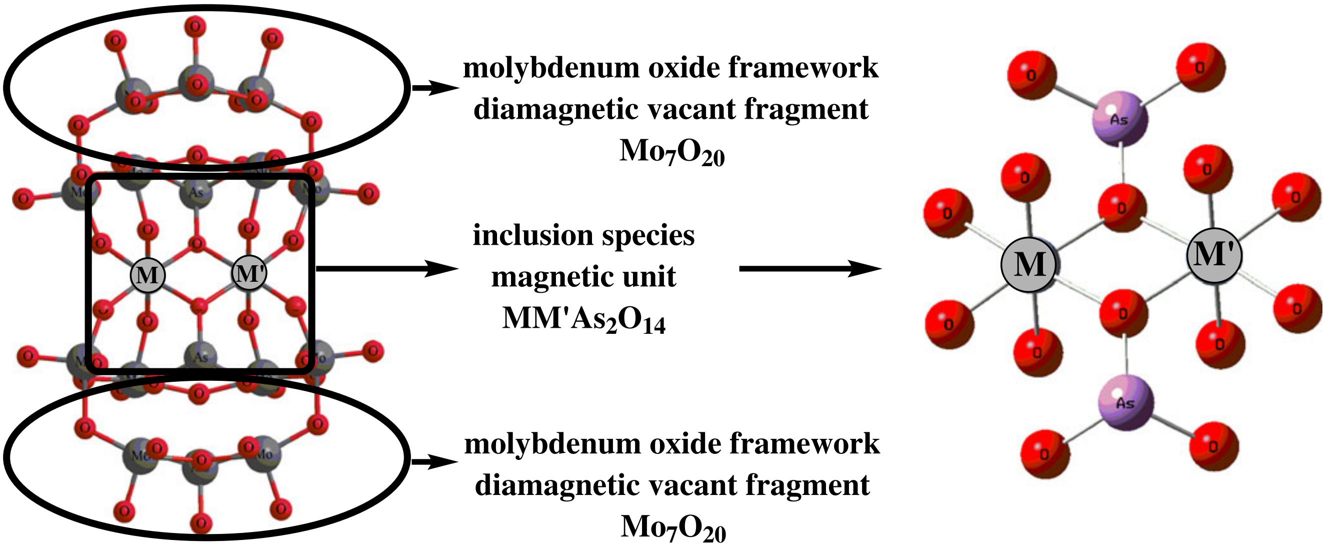

molybdenum oxide framework
diamagnetic vacant fragment
Mo7O20
inclusion species
magnetic unit
MM'As2O14
molybdenum oxide framework
diamagnetic vacant fragment
Mo7O20
M
M'
Mo
As
O

Fig. 1 Ball-and-stick representation of dinuclear sandwich-type polyoxomolybdates $[MM'(AsMo_7O_{27})_2]^{12-}$ and corresponding magnetic units

i.e., antiferromagnetic (AFM) exchange coupling, is demonstrated from accurate fitting of experimental data of variable-temperature magnetic susceptibility [11]. Therefore theoretical study of their properties, esp. the magnetism, is obviously necessary to achieve a better utilization of these new types of POMs. This necessity is the direct inspiration of this current computational work.

Compared with the extensive experimental researches on POM systems, the number of theoretical studies of POM is still limited mainly because of the intrinsic difficulties [12] of computation arising from their large sizes, the presence of multiple TM ions with unpaired electrons and high negative charges which can only be stabilized in the solution or crystal environments. However, as driven by the invaluable capability of computation in understanding the electronic and magnetic properties of POM, the interest in theoretical studies of POM is rapidly increasing with the help of modern computational methodology [12, 13], esp. density functional theory (DFT) [14, 15]. Therefore, DFT study on these new dinuclear sandwich-type POMs is reported here. Broken-symmetry (BS) method [16–20], originally developed for the theoretical study of magnetic exchange coupling [16], is used in conjunction with DFT method. One thing worth noting is that the magnetism of several POM systems has been well studied from combined DFT and BS (DFT-BS) approach [21–27].

In spite of its great success [17, 20], the accuracy of DFT-BS approach is still not a closed field, especially in TM systems where the diversity of DFT calculations with different functionals is frequently reported [28, 29]. Therefore, a careful choice of functionals is definitely not trivial as shown by the recent report on striking differences between DFT-BS results of pure and hybrid functionals [30]. According to our best knowledge, a systematic and comprehensive comparison among different functionals is still absent for POM systems. Thus eight different functionals, consisting of B3LYP [31, 32], O3LYP [32, 33], MPW1PW91 [34, 35], PBE0 [36], BP86 [31, 37], OPBE [33, 38], MPWPW91 [34, 35] and PBE [38], are examined in this work.

Theoretical methods and computational details

Due to its moderate computational effort and (partial) inclusion of electron correlation [12, 13, 39], DFT has become the most popular theoretical method for POM. In the field of molecular magnetism [40, 41], the most-commonly used phenomenological spin Hamiltonian (SH) is the Heisenberg-Dirac-Van Vleck (HDVV) one [42, 43] which implies an isotropic interaction between the magnetic centers [19, 20]. The formula of HDVV SH for dinuclear systems is written as:

$$H^{HDVV} = -JS_A \cdot S_B \qquad (1)$$

S_A and S_B are the total spin operators for magnetic centers A and B and J is the so-called exchange coupling constant. Positive sign of J indicates ferromagnetic (FM) coupling between magnetic centers, i.e., the ground state is spin parallel. Negative sign of J indicates AFM coupling, i.e., the ground state is spin antiparallel. The magnitude of J corresponds to the strength of exchange coupling.

$$E(S) = -\frac{1}{2}J \cdot S(S+1) \quad E(S) - E(S-1) = -J \cdot S \qquad (2)$$

According to HDVV SH, J could be calculated from the energy differences between spin eigenstates in principle [19] since the energies of various spin eigenstates of the system are determined only by J and the total spin S (Eq. 2) [40]. However, the direct application of Eq. 2 to real system is impractical because the calculations of intermediate or low spin state are usually problematic at DFT level [14].

BS technique, developed by Noodleman et al. [16], has been proven to be the most commonly used method [20] to solve this difficulty. It has been successfully applied in TM systems [44–49], organic diradicals [50–52] and hybrid systems consisting of both TM ions and organic radicals [53–55]. BS state is monodeterminantally constructed to have opposite spins essentially localized on different magnetic centers and thus it breaks the spatial symmetry from the view of spin distribution. In energetic aspect, BS state is a weighted average of spin eigenstates from a second-order perturbation approximation of configuration interaction up to double-excitation [16]. With the help of spin-projection technique, J could be calculated as:

$$J = \frac{2(E_{BS} - E_{HS})}{S^2_{\max}} \tag{3}$$

E_{HS} and E_{BS} are the energies of highest spin (HS) and BS state respectively, S_{max} is the value of maximum spin of the system. Recent studies have pointed out the possibility of approximating the energy of lowest spin (LS) state directly from that of BS state [18, 46, 48], thus J could also be calculated from non-projection equation as:

$$J = \frac{(E_{BS} - E_{HS})}{2S_A S_B + S_B} \tag{4}$$

Due to the high susceptibility of calculated J to geometry variation [18], experimental structures without optimization are mainly used for various functionals. Calculations are also performed on model structures consisting of non-optimized inclusion species $[MM'As_2O_{14}]^{16-}$ and 12 protons at optimized positions.

All the calculations are performed with Gaussian 03 program [56, 57] and the necessary initial guesses for the construction of BS state are generated with the help of NBO 5.0 code [58, 59]. For the calculations on full structures, all-electron basis TZVP [60] is used for Fe, Cr elements and pseudopotential basis LANL2DZ [61] is used for all the other elements. For the calculations on model structures, Fe and Cr elements are also described with TZVP, As and H elements are described with 6-31 G**, O element is described with 6-31+G*. The choice of these combinations of basis is based on recent benchmark calculations [47] concluding that high quality all-electron basis is necessary for magnetic centers to obtain high accuracy in the calculation of J. To ensure the reliability of calculation, the wavefunctions of BS states are converged with quadratic convergence (QC) [57] and tested with STABLE=OPT option [57] of Gaussian code.

Results and discussions

Reliability of obtained HS and BS state

The reliability of obtained HS and BS states is examined in the aspects of spin expectation value and spin density. The results of different functionals are quite close to each other and thus only the values of B3LYP and BP86, taken as the representatives of hybrid and GGA functionals respectively, are listed in Tables 1 and 2.

The expectation values of spin square operator of HS states, calculated with B3LYP on full structures, are 12.034, 30.018 and 20.029 for Cr_2, Fe_2 and FeCr respectively. The corresponding values of model structures are 12.030, 30.014 and 20.023 respectively. Similar results are also obtained from calculations with GGA functionals (Table 1). These values are nearly the same as the ideal values, which are 12, 30 and 20. Therefore, from the viewpoint of spin eigenvalue, all the functionals provide reliable descriptions on the HS states of different POM irrespective of the usage of full or model structures.

As shown in Table 1, the expectation values of spin square operator of BS states are~3, ~ 5 and~5 for Cr_2, Fe_2 and FeCr respectively. All these values are between the ideal values of HS and LS states of Cr_2 (12 and 0), Fe_2 (30 and 0) and FeCr (20 and 2). Therefore the obtained BS states are certainly weighted averages of all the possible spin eigenstates as required [16].

As shown in Table 2, both Mulliken population analysis (MPA) and natural population analysis (NPA) [59] demonstrate that the spin density of HS state is essentially localized on the magnetic centers of various POM. From B3LYP results, spin density on Cr(III) of Cr_2-full is 2.945

Table 1 Expectation values of spin square operator calculated with different functionals

	Cr2-full[a]		Fe2-full		FeCr-full	
	HS	BS	HS	BS	HS	BS
B3LYP	12.03	3.03	30.02	5.00	20.03	5.02
BP86	12.12	3.12	30.08	5.08	20.17	5.00
Eigen[b]	12	0	30	0	20	2
	Cr2-model[a]		Fe2-model		FeCr-model	
	HS	BS	HS	BS	HS	BS
B3LYP	12.03	3.02	30.01	5.00	20.02	5.02
BP86	12.03	3.01	30.01	5.08	20.02	5.00
Eigen	12	0	30	0	20	2

[a] Full indicates the full crystalline structures of certain POM. Model indicates the model structures of certain POM [b] the exact eigenvalues of spin square operator of HS and LS states of different POM deduced from the [Core]3 d^3 and [Core]3 d^5 electron configurations of free Cr (III) and Fe(III) ions. The values of LS states are shown at the columns of BS of this entry

Table 2 Atomic spin densities on the magnetic centers calculated with different functionals

	Cr2-full		Fe2-full		FeCr-full	
	HS	BS	HS	BS	HS	BS
B3LYP[a]	2.945/2.945[b]	2.935/-2.935	4.275/4.275	4.279/-4.279	4.261/2.298	4.254/-2.933
	2.759/2.759	2.742/-2.742	4.171/4.171	4.169/-4.169	4.130/2.749	4.119/-2.743
BP86	2.523/2.523	2.512/-2.512	4.294/4.294	3.870/-3.870	3.801/2.411	3.698/-2.371
	2.319/2.319	2.294/-2.294	4.153/4.153	3.708/-3.708	3.625/2.223	3.513/-2.173
	Cr2-model		Fe2-model		FeCr-model	
	HS	BS	HS	BS	HS	BS
B3LYP	3.054/3.054	3.045/-3.045	4.294/4.294	4.289/-4.289	4.276/3.072	4.273/-3.056
	2.792/2.792	2.778/-2.778	4.171/4.171	4.169/-4.169	4.105/2.806	4.098/-2.801
BP86	3.102/3.102	3.074/-3.074	4.294/4.294	4.083/-4.083	4.117/3.130	4.108/-3.083
	2.714/2.714	2.689/-2.689	4.153/4.153	3.867/-3.867	3.859/2.742	3.850/-2.716

[a] spin densities from MPA are on first entry and spin densities from NPA are on second entry [b] spin density on M is at the left side of " / " and spin density on M' is at the right side of "/ "

or 2.759 from MPA or NPA method. The corresponding values of Cr_2-model is 3.054 and 2.792 respectively. BP86 functional provide essentially the same results with the only deviation that its magnitude is moderately smaller than that of B3LYP. Therefore the localized spin distribution is not influenced by different structures, functionals or population methods used in calculations and this character confirms the use of HDVV SH in describing the title POMs [20].

As shown in Table 2, the requirement of localization of opposite spins on different magnetic centers is clearly obeyed by BS states obtained from NBO-generated initial guesses. In the cases of homobinuclear POM Cr_2 and Fe_2, the same magnitude of spin density is localized on the TM ions. With B3LYP functional, the spin densities of Cr(III) in Cr_2-full and Fe(III) in Fe_2-full are 2.935 and 4.279 from MPA or 2.742 and 4.169 from NPA. The corresponding values of Cr_2-model and Fe_2-model are 3.045 and 4.289 from MPA or 2.728 and 4.150 from NPA. Although with smaller magnitude of spin density, the same character is also demonstrated by BP86 functional as indicated in Table 2. Therefore obtained BS states meet the requirement in the aspect of spin density, no matter which different structures, functionals or population methods are used in calculations.

Description on the exchange coupling from calculations on full structures

Although different functionals provide the same results in the aspects of spin expectation value and spin density, their results in energetic aspect are of a high degree of diversity as shown in Table 3. All the GGA and O3LYP functionals fail in qualitative description on the exchange coupling of Cr_2-full and FeCr-full, as indicated by the large positive values of E_{BS}-E_{HS} (Table 3). That is to say, the calculations with these functionals on full structures lead to extremely strong FM exchange coupling in Cr_2 and FeCr, which is absolutely contrary to the accurate experimental results.

Although succeeding in qualitative description, the performance of hybrid functionals on full structures is still far away from being promising in the quantitative aspect. It is worth noting that the influence of the adoption of Eq. 3 or 4 on the calculated J values is not substantial.

MPW1PW91 and PBE0 functionals provide nearly the same numerical results which both underestimate the strengths of the exchange coupling of the title POM. For Cr_2-full, J values calculated with these two functionals are $-2\sim-3$ cm^{-1} whereas the experimental value is -12.52 cm^{-1}. The exchange coupling of Fe_2-full, predicted from MPW1PW91 and PBE0 functionals, is even negligible as shown by the calculated J values whose magnitudes are less than 1 cm^{-1}. However, the corresponding experimental value is -4.18 cm^{-1}. In the case of FeCr-full, the calculated J values are mainly between -5 and -6 cm^{-1}, still significantly less than the experimental value of -8.18 cm^{-1}.

An opposite tendency toward the overestimation of J of Fe_2-full is given by all the GGA and O3LYP functionals as shown in Table 3. Almost all the magnitudes of the calculated J values, varying from -7 to -60 cm^{-1}, are one-order larger than that of experimental J value.

For the calculations on full structures, the best numerical accuracy is given by the combination of B3LYP and Eq. 3 with the J values of -6.82, -3.58 and -8.25 cm^{-1} for Cr_2-full, Fe_2-full and FeCr-full respectively. However, compared with the experimental J values of -12.52, -4.18 and -8.18 cm^{-1} respectively, B3LYP is still incapable of reproducing the relative strengths of the exchange coupling among the title POMs.

Description on the exchange coupling from calculations on model structures

As shown in Table 4, theoretical description on the exchange coupling of the title POM is significantly improved by the calculations on model structures both

Table 3 Energy difference between HS and BS states and J calculated with different functionals on the full structures (in cm^{-1})

		B3LYP	O3LYP	MPW1	PBE0	BP86	OPBE	MPW	PBE
Cr_2-full	E_{BS}-E_{HS}[a]	−30.68	127.80	−14.42	−14.11	541.34	622.71	540.26	524.97
	J (Eq. 3)[b]	−6.82	28.40	−3.20	−3.32	120.30	138.38	120.06	116.66
	J (Eq. 4)[c]	−5.12	21.30	−2.40	−2.35	90.22	103.78	90.04	87.49
	J (exp)[d]	−12.52							
Fe_2-full	E_{BS}-E_{HS}	−44.77	−112.90	−9.524	−11.126	−731.48	−180.58	−753.14	−406.32
	J (Eq. 3)	−3.58	−9.03	−0.76	−0.89	−58.52	−14.45	−60.25	−32.50
	J (Eq. 4)	−2.98	−7.53	−0.63	−0.74	−48.76	−12.04	−50.21	−27.09
	J (exp)	−4.18							
FeCr-full	E_{BS}-E_{HS}	−66.01	90.06	−44.30	−45.51	120.78	852.88	582.90	512.98
	J (Eq. 3)	−8.25	11.26	−5.54	−5.69	15.10	106.61	72.86	64.12
	J (Eq. 4)	−7.33	10.01	−4.92	−5.06	13.42	94.76	64.77	57.00
	J (exp)	−8.18							

[a] energy difference calculated as E_{BS}-E_{HS} [b] J calculated from spin-projection Eq. 3 [c] J calculated from non-projection Eq. 4 [d] J from accurate fitting of experimental variable-temperature magnetic susceptibility in ref [11]

qualitatively and quantitatively. All the GGA functionals become successful in predicting AFM exchange coupling with the only exception of OPBE which fails in FeCr-model as shown by its positive value of E_{BS}-E_{HS} (Table 4). With the usage of model structures, O3LYP is capable of predicting AFM exchange coupling not only for Fe_2 but also for Cr_2. However, the use of model structures lead to qualitative failure of MPW1PW91 and PBE0 in the description on FeCr.

The improvement in quantitative aspect, arising from the use of model structures, is even more encouraging as shown in Table 4. The calculated J values of Cr_2-model, with PBE0 and MPW1PW91, are −9∼−12 cm^{-1}, which are quite close to the experimental value of −12.52 cm^{-1}. Similarly, the calculated J values of Fe_2-model are −3∼−5 cm^{-1} and thus they are also close to the experimental value of −4.18 cm^{-1}. Therefore the underestimate of the strengths of AFM coupling, arising from the calculations on full structures, is mainly remedied by the use of model structures for MPW1PW91 and PBE0. The most accurate values of J from O3LYP are −13.13 and −7.07 cm^{-1} for Cr_2-model and Fe_2-model respectively. These values are apparently close to the experiment values of −12.52 and −4.18 cm^{-1}.

The accuracy of B3LYP results is also remarkably increased by the use of model structures. The most accurate J values of Cr_2 are −6.82 cm^{-1} from the calculations on full structures and −13.98 cm^{-1} those on model structures, with the latter one more close to the experimental value of −12.52 cm^{-1}. In the case of Fe_2, although the most accurate J is provided by the calculations on full structure, the value from the calculations on model structures is −5.60 cm^{-1} with quite small deviation from the experimental value of −4.18 cm^{-1}.

Table 4 Energy difference between HS and BS states and J calculated with different functionals on the model structures (in cm^{-1})

		B3LYP	O3LYP	MPW1	PBE0	BP86	OPBE	MPW	PBE
Cr_2-model	E_{BS}-E_{HS}	−62.89	−78.76	−53.10	−54.51	−195.10	−143.89	−189.79	−206.56
	J (Eq. 3)	−13.98	−17.50	−11.80	−12.11	−43.36	−31.98	−42.18	−45.90
	J (Eq. 4)	−10.48	−13.13	−8.85	−9.08	−32.68	−23.98	−31.63	−34.43
	J (exp)	−12.52							
Fe_2-model	E_{BS}-E_{HS}	−84.07	−105.99	−51.81	−59.61	−578.73	−263.02	−506.38	−502.96
	J (Eq. 3)	−6.72	−8.48	−4.14	−4.77	−46.30	−21.04	−40.51	−40.24
	J (Eq. 4)	−5.60	−7.07	−3.45	−3.97	−38.58	−17.53	−33.76	−33.53
	J (exp)	−4.18							
FeCr-model	E_{BS}-E_{HS}	−1.27	10.51	14.26	26.24	−59.84	10.01	−51.66	−52.82
	J (Eq. 3)	−0.16	1.31	1.78	3.28	−7.48	1.25	−6.46	−6.60
	J (Eq. 4)	−0.14	1.17	1.58	2.92	−6.65	1.11	−5.74	−5.87
	J (exp)	−8.18							

For GGA functionals, although the overestimate of AFM magnetic coupling still exists in the calculations on model structures of Cr_2 and Fe_2, the magnitudes are usually smaller than those of full structures. For FeCr, with the help of model structures, GGA calculations are also able to provide results of high accuracy as shown by the obtained J values of $-6{\sim}-8$ cm^{-1}.

Discussions on the use of model structures

The improvement of numerical accuracy of DFT-BS calculations, arising from the use of model structures, is highly valuable in three aspects. First, the justification for using model structures, indicated here, implies the possibility of application of high-level multiconfigurational methods, e.g., CASSCF, CASPT2, DDCI, in the title POM and their analogues. In principle, the intermediate spin states of magnetic coupling systems can only be correctly described by multiconfigurational methods [19]. However, the exhaustive cost of these methods forbids the applications of them with the only exception of model structures, of which the sizes are largely reduced in comparison with full structures. Considering POM systems where accurate experimental data is unavailable, in the step of selection of suitable functionals from various candidates, the reliable references can only be obtained from calculations with multiconfigurational methods which are only practical in model structures. Therefore, the value of this point is especially important for POM system without reliable experimental results.

Second, the assumption of the title POM being constructed from magnetic inclusion species and diamagnetic peripheral fragments is supported by the accurate results of model structures, which are just combinations of inclusion species and protons. Therefore, the function exerted by peripheral fragments is implied to be mainly structural. That is to say, the characteristic magnetism of the title POM is essentially determined by the inclusion species. However, the structural features of these inclusion species, which are directly related to magnetic properties, are probably only available with the help of those peripheral fragments.

Third, the title POM studied here are all highly negative-charged anions, which can only be stable in solution or crystal environment. That is to say, the external field effect, arising from solvent molecule in solution or counterion in crystal, is quite important for POM systems [12, 13]. In this

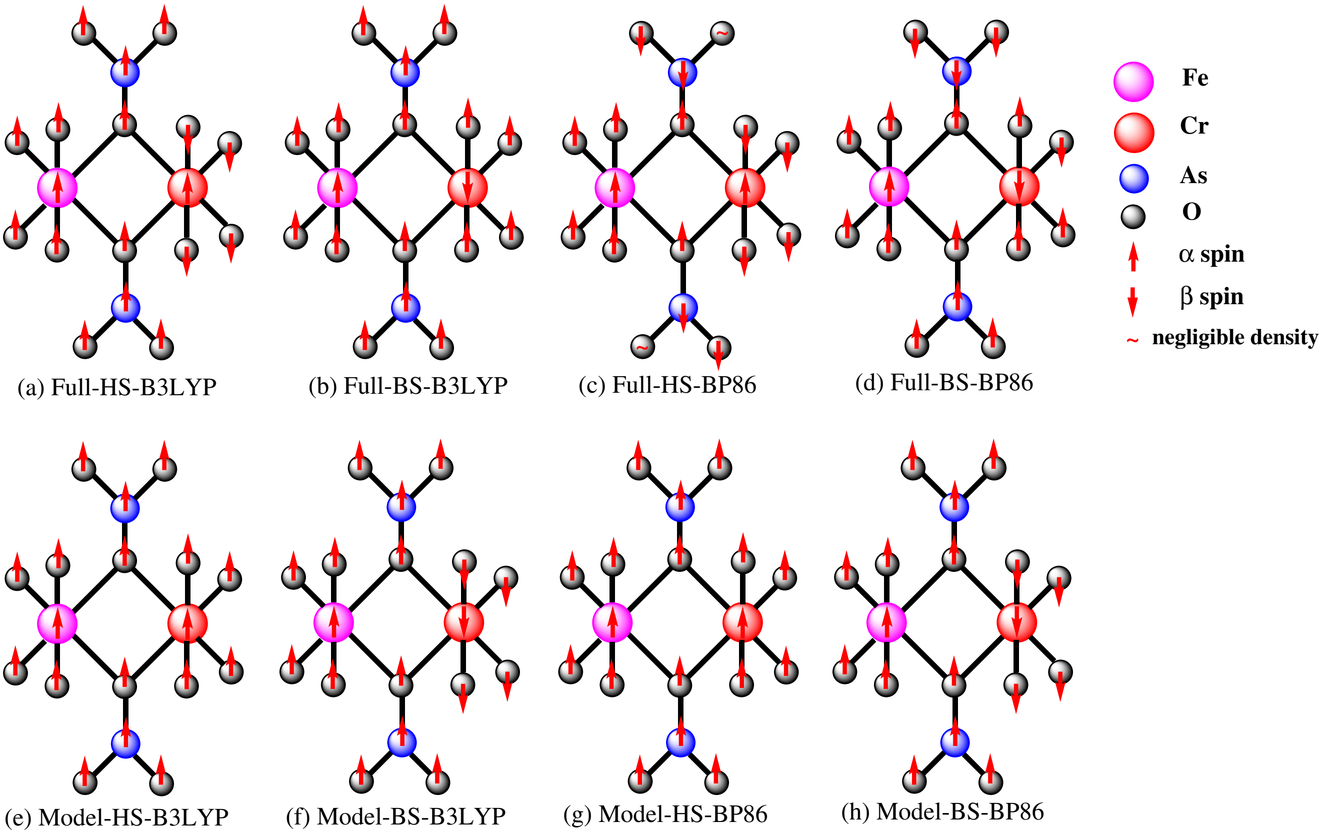


Fig. 2 Schematic representation of spin density distribution of heterobinuclear sandwich-type POM FeCr calculated from NPA with B3LYP and BP86 functionals (**a**) HS state of full structure with B3LYP (**b**) BS state of full structure with B3LYP (**c**) HS state of full structure with BP86 (**d**) BS state of full structure with BP86 (**e**) HS state of model structure with B3LYP (**f**) BS state of model structure with B3LYP (**c**) HS state of model structure with BP86 (**d**) BS state of model structure with BP86

work, solvent molecule or counterion is not included in the calculations. Therefore the increase in numerical accuracy of the calculations on model structures may arise from a better treatment of external field effect since the number of negative charge reduces from 12 to 4 with the use of model structures.

Based on these considerations, rational selection of model structures may be a suitable approach for POM systems due to the coexistence of lower computational cost and better treatment of external field effect, which may lead to higher numerical accuracy.

One thing worth noting is the qualitative difference between calculated J values of full and model structures for heterobinuclear sandwich-type FeCr as shown in Table 3 and 4. The reason for this difference is apparently important for future application of structure modeling for POM systems. Therefore spin density distribution of FeCr is schematically present in Fig. 2 in seeking a deep understanding.

From calculations with B3LYP, spin densities on terminal oxo ligands, coordinated to Cr(III) ion, are mainly determined by spin polarization mechanism [62, 63] in HS state, as shown by the opposite sign to central Cr(III) in Fig. 2a. However the dominating mechanism changes to spin delocalization [62, 63] in model structure, as shown by spin densities of the same sign to central Cr(III) in Fig. 2e. For BS state, B3LYP results also indicate the same change of mechanism in spin density, as shown in Fig. 2b and f.

Besides the change of spin densities on terminal oxo ligands coordinated to Cr(III), BP86 results also indicate a new change of spin densities on arsentic atoms as shown in the comparison between Fig. 2c and g or comparison between Fig. 2d and h. In HS state, spin densities on arsenous atoms change from −0.003 in full structure to 0.048 in model structure. The corresponding change of BS state is from −0.010 to 0.023. It is worth noting that spin densities on arsenous atoms from B3LYP calculation on full structure and from BP86 calculation on model structure are all α spin. Both of these two combinations of theoretical method and structure are capable of reproducing exactly experimental J values of FeCr.

Conclusions

For all the functionals examined here, BS states obtained from NBO-generated guesses meet the requirements in the aspects of spin expectation value and spin density. However the high diversity of calculated J values with different functionals emphasizes the necessity of a careful choice of functional, especially in DFT-BS study on the magnetism of POM system.

A great improvement in numerical accuracy of calculated J is achieved with the use of model structures. Only with a combination of model structures for homobinuclear POM and full structure for heterobinuclear POM, B3LYP, MPW1PW91 and PBE0 are capable of reproducing the relative strengths of AFM coupling among the title POM here. This fact highlights the great potential of structure modeling in theoretical study on POM systems, especially for those where accurate experimental results are not available.

Rational selection of model structures may not only reduce the computational cost but also be capable of providing better treatment of external field effect which is necessary for the existence of highly negative-charged POM anions.

The influence of the use of spin-projection or non-projection equation to calculating J is small for the title POM where there are several unpaired electrons in each magnetic center. The functionals based on OPTX [33] exchange functional are not suggested, especially in seeking for high accuracy of calculated J values.

The qualitative difference between results calculated on full and model structures for heterobinuclear POM may arise from the change of mechanism of spin distribution. Spin densities on arsentic atoms seem to be very important for the magnetism of heterobinuclear POM of this new type.

Due to the large sizes and the necessity of inclusion of electron correlation of POMs [12], DFT has become the first choice of theoretical method. However, current popular XC functionals are usually approximations of the unknown universal functional in original Hohenberg-Kohn formalism [14, 15] and thus DFT results may bear various problems, e. g., the fractional spin and fractional charge errors [64–67]. The usage of BS method may partially remedy this defect as shown in the results reported here as well as in previous studies [21–27]. However, the ultimate solution to this issue leans on the development of new functionals of high accuracy. Therefore the test of various functionals, especially those newly developed ones, is always valuable for theoretical studies of POM and it will be continued in our future works.

Acknowledgments This work is supported by the National Natural Science Foundation of China (No: 21103137, 20973133, 20972124), China Postdoctoral Science Foundation (No: 20110491677) and Northwest University (No: PR09050, 10NW04, BS10013, ZS11038). Y.B. thanks Prof Maxim V Borozov (Northwest University), Drs Tao Song (The University of Hong Kong), Jing Li (University of Illinoisat Urbana-Champaign) and Hongcun Bai (Beijing Normal University) for their kind help.

References

1. Pope MT (1983) Heteropoly and isopoly Oxometalates. Springer, Berlin
2. Pope MT, Muller A (eds) (1994) Polyoxometalate: from platonic solids to anti-retroviral activity. Kluwer, Dordrecht, The Netherlands

3. Kozhevnikov IV (2002) Catalysis by polyoxometalates. Wiley, Chichester, England
4. Hill CL (1998) eds. Chem Rev 98:1–390
5. Pope MT, Muller A (eds) (2001) Polyoxometalate chemistry from topology via self-assembly to application. Kluwer, Dordrecht, The Netherlands
6. Yamase T, Pope MT (eds) (2002) Polyoxometalate chemistry for nano-composite design. Kluwer, Dordrecht, The Netherlands
7. Coronado E, Gomez-Garcia CJ (1998) Chem Rev 98:273–296
8. Casan-Pastor N, Bas-Serra J, Coronado E, Pourroy G, Baker LCW (1992) J Am Chem Soc 114:10380–10383
9. Kikukawa Y, Yamaguchi S, Tsuchida K, Nakagawa Y, Uehara K, Yamaguchi K, Mizuno N (2008) J Am Chem Soc 130:5472–5478
10. Li L, Shen Q, Xue G, Xu H, Hu H, Fu F, Wang J (2008) Dalton Trans pp 5698–5700
11. Xu H, Li L, Liu B, Xue G, Hu H, Fu F, Wang J (2009) Inorg Chem 48:10275–10280
12. Poblet JM, Lopez X, Bo C (2003) Chem Soc Rev 32:297–308
13. Rohmer MM, Benard M, Blaudeau JP, Maestre JM, Poblet JM (1998) Coord Chem Rev 178:1019–1049
14. Koch W, Holthausen MC (2000) A chemist's guide to density functional theory. Wiley, Weinheim
15. Parr RG, Yang W (1989) Density functional theory of atoms and molecules. Oxford University Press, New York
16. Noodleman L, Davidson ER (1986) Chem Phys 109:131–143
17. Noodleman L, Peng CY, Case DA, Mouesca JM (1985) Coord Chem Rev 144:199–244
18. Ruiz E, Cano J, Alvarez S, Alemany P (1999) J Comput Chem 20:1391–1400
19. Illas F, Moreira PR, de Graaf C, Barone V (2000) Theor Chem Acc 104:265–272
20. Ciofini I, Daul CA (2003) Coord Chem Rev 239:187–209
21. Maestre JM, Poblet JM, Bo C, Casan-Pastor N, Gomez-Romero P (1998) Inorg Chem 37:3444–3446
22. Maestre JM, Lopez X, Bo C, Poblet JM, Casan-Pastor N (2001) J Am Chem Soc 123:3749–3758
23. Lopez X, de Graaf C, Maestre JM, Benard M, Rohmer MM, Bo C, Poblet JM (2005) J Chem Theor Comput 1:856–861
24. Rodrigues-Fortea A, de Graaf C, Poblet JM (2006) Chem Phys Lett 428:88–92
25. Duclusaud H, Borshch SA (2001) J Am Chem Soc 123:2825–2829
26. Zueva EM, Chermette H, Borshch SA (2004) Inorg Chem 43:2834–2844
27. Wang Y, Zheng G, Morokuma K, Geletii Y, Hill CL, Musaev DG (2006) J Phys Chem B 110:5230–5237
28. Swart M, Groenhof AR, Ehlers AW, Lammertsma K (2004) J Phys Chem A 108:5479–5483
29. Strassner T, Taige MA (2005) J Chem Theor Comput 1:848–855
30. Hopmann KH, Conradie J, Ghosh A (2009) J Phys Chem B 113:10540–10547
31. Becke AD (1993) J Chem Phys 98:5648–5652
32. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785–789
33. Handy NC, Cohen AJ (2001) Mol Phys 99:403–408
34. Adamo C, Barone V (1998) J Chem Phys 108:664–675
35. Perdew JP, Burke K, Wang Y (1996) Phys Rev B 54:16533–16539
36. Perdew JP, Burke K, Ernzerhof M (1997) Phys Rev Lett 78:1396
37. Becke AD (1988) Phys Rev A 38:3098–3100
38. Perdew JP, Burke K, Ernzerhof M (1996) Phys Rev Lett 77:3865–3868
39. Davidson ER (2000) eds. Chem Rev 100:351–818
40. Kahn O (1993) Molecular Magnetism. VCH Publisher, New York
41. Borras JJ, Coronado E, Tsukerblat BS, Georges R (1996) Molecular magnetism: from molecular assemblies to the devices. Kluwer, Dordrecht, The Netherlands
42. Heisenberg W (1928) Z Phys 49:619–636
43. VanVleck JH (1932) The theory of electric and magnetic susceptibilities. Oxford University Press, Oxford
44. Ruiz E, Alemany P, Alvarez S, Cano J (1997) J Am Chem Soc 119:1297–1303
45. Desplanches C, Ruiz E, Rodrigues-Fortea A, Alvarez S (2002) J Am Chem Soc 124:5197–5205
46. Ruiz E, Rodrigues-Fortea A, Alvarez S, Verdaguer M (2005) Chem Eur J 11:2135–2144
47. Ruiz E, Rodrigues-Fortea A, Tercero J, Cauchy T (2005) J Chem Phys 123:074102(1–10)
48. Manca G, Cano J, Ruiz E (2009) Inorg Chem 48:3139–3144
49. Zein S, Borshch SA (2005) J Am Chem Soc 127:16197–16201
50. Barone V, Bencini A, di Matteo A (1997) J Am Chem Soc 119:10831–10837
51. Barone V, Bencini A, Ciofini I, Daul CA (1999) J Phys Chem A 103:4275–4282
52. Barone V, di Matteo A, Mele F, Moreira IPR, Illas F (1999) Chem Phys Lett 302:240–248
53. Cauchy T, Ruiz E, Jeannin O, Nomura M, Fourmigue M (2007) Chem Eur J 13:8858–8866
54. Pardo E, Carrasco R, Ruiz-Garcia R, Julve M, Lloret F, Munoz MC, Journaux Y, Ruiz E, Cano J (2007) J Am Chem Soc 130:576–585
55. di Matteo A, Barone V (1999) J Phys Chem A 103:7676–7685
56. Frisch MJ, Trucks GW, Schlegel HB et al (2003) Gaussian03. Gaussian Inc, Pittsburgh PA
57. Foresman JB, Frisch Æ (1996) Exploring chemistry with electronic structure methods, 2nd edn. Gaussian Inc, Pittsburgh, PA
58. Glendening ED, Badenhoop JK, Reed AE, Carpenter JE, Bohmann JA, Morales CM, Weinhold F (2001) NBO 5.G. Theoretical Chemistry Institute, University of Wisconsin, Madison, WI
59. Weinhold F (2001) NBO 5.0 Program Manual: Natural Bond Orbita l Analysis Programs. Theoretical Chemistry Institute and Department of Chemistry, University of Wisconsin, Madison, WI 53706
60. Schafer A, Huber C, Ahlrichs R (1994) J Chem Phys 100:5289–5835
61. Hay PJ, Wadt WR (1985) J Chem Phys 82:299–310
62. Cano J, Ruiz E, Alvarez S, Verdaguer M (1998) Comment Inorg Chem 20:27–56
63. Ruiz E, Cirera J, Alvarez S (2005) Coord Chem Rev 249:2649–2660
64. Mori-Sanchez P, Cohen AJ, Yang WT (2009) Phys Rev Lett 102:06640
65. Mori-Sanchez P, Cohen AJ, Yang WT (2008) Phys Rev Lett 100:14640
66. Cohen AJ, Mori-Sanchez P, Yang WT (2008) Science 321:792
67. Cohen AJ, Mori-Sanchez P, Yang WT (2008) J Chem Phys 129:121104