

Computational characterization of the molecular structure and properties of Dye 7 for organic photovoltaics

Jesús Baldenebro-López · José Castorena-González ·
Norma Flores-Holguin · Joel Calderón-Guillén ·
Daniel Glossman-Mitnik

Received: 6 February 2011 / Accepted: 5 May 2011 / Published online: 20 May 2011
© Springer-Verlag 2011

Abstract Organic dyes have great potential for its use in solar cells. In this recent work, the molecular structure and properties of Dye 7 were obtained using density functional theory (DFT) and different levels of calculation. Upon comparing the molecular structure and the ultraviolet visible spectrum with experimental data reported in the literature, it was found that the M05-2X/6-31G(d) level of calculation gave the best approximation. Once the appropriate methodology had been obtained, the molecule was characterized by obtaining the infrared spectrum, dipole moment, total energy, isotropic polarizability, molecular orbital energies, free energy of solvation in different solvents, and the chemical reactivity sites using the condensed Fukui functions.

Keywords Molecular structure · Polarizability · $\Delta G(\text{solv})$ · Chemical reactivity

Abbreviations

DFT	Density functional theory
TD-DFT	Time-dependent density functional theory
IR	Infrared
UV-vis	Ultraviolet
Å	Angstrom

λ_{max}	Wavelength of maximum absorption
THF	Tetrahydrofuran
HOMO	Highest occupied molecular orbital
LUMO	Lowest unoccupied molecular orbital
$\Delta G(\text{solv})$	Free energy of solvation
IEF-PCM	Integral equation formalism of the polarized continuum model

Introduction

Photovoltaic devices have gained wide acceptance as a clean and renewable energy source [1]. These devices are based on the concept of charge separation at an interface between two materials with different conduction mechanisms [2, 3]. One important invention in this field is the photovoltaic dye-sensitized solar cell (DSSC) [4], which has been the subject of intense research due to its ability to convert solar energy into electrical energy [5, 6], as well as its low cost compared to solar cells that use polycrystalline silicon [7]. There are four main factors that affect the performance of a DSSC: the photosensitive dye, the anode, the cathode and the electrolyte. The dye plays a crucial role in enhancing the efficiency of the cell, which is why it is one of the most intensely studied factors [8]. In the present work, a theoretical study of the molecular structure and properties of a dye (Dye 7) was performed. This dye consists of a triphenylamine molecule that serves as electron donor group [9, 10], a thiophene to adjust the absorption spectra of the molecules [11], and a cyanoacrylic acid that acts as an acceptor group [12], as shown in Fig. 1. In the investigation described below, different levels of theory were used in order to establish the most appropriate methodology to study this dye. Besides optimizing the

J. Baldenebro-López · J. Castorena-González ·
J. Calderón-Guillén
Facultad de Ingeniería Mochis, Universidad Autónoma de
Sinaloa. Prol. Ángel Flores y Fuente de Poseidón, S.N,
C.P. 81223 Los Mochis, Sinaloa, México

N. Flores-Holguin · D. Glossman-Mitnik (✉)
Centro de Investigación en Materiales Avanzados, SC,
Complejo Industrial Chihuahua,
Miguel de Cervantes 120,
Chihuahua 31109, México
e-mail: daniel.glossman@cimav.edu.mx

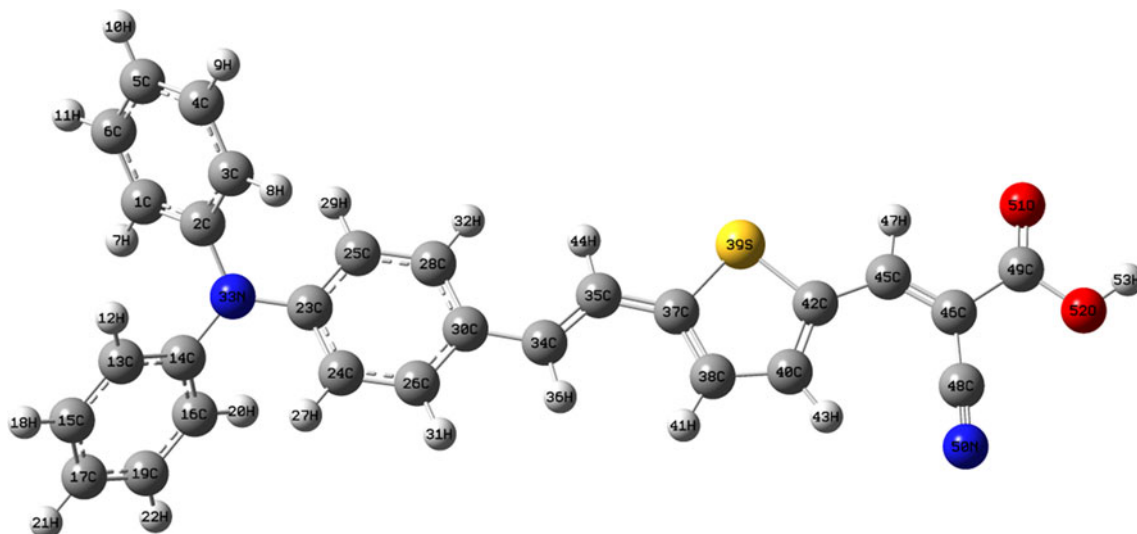


Fig. 1 Molecular structure of Dye 7: triphenylamine (donor), a thiophene (to adjust the absorption spectra of the molecules) and cyanoacrylic acid (acceptor)

geometry of Dye 7, the infrared and ultraviolet spectra were presented, the dipole moment calculated, as were the total energy, isotropic polarizability, molecular orbital energies, and the free energy of solvation in different solvents; Its chemically reactive sites were also discerned using the condensed Fukui functions.

Molecular modeling

Density functional theory (DFT) was used in this study [13]. DFT was developed by Walter Kohn in the 1960s, and was implemented in this study using the commercial software Gaussian 03W [14]. The geometry of the molecule in its fundamental state was obtained by the established technique in Gaussian 03W. The force constants and vibrational frequencies were determined by calculating analytical frequencies for stationary points obtained after optimizing the geometry. Both calculations were done at the same level of theory. The basis sets used in this study were 3-21G(d) and 6-31G(d) (for more details, see [15]). The density functionals used in this research were: BLYP [16, 18], B3LYP [16–19], PBE [20, 21], PBE1PBE [21], TPSS [22], TPSSh [23] and M05-2X [24]. A detailed description of these density functionals can be found in the updated bibliography of computational chemistry [25–28]. The calculation of the ultraviolet spectrum of the molecule Dye 7 was done via time-dependent DFT equations according to the method implemented in the Gaussian molecular package 03W [25, 29–32]. The equations were solved for 20 excited states. The infrared (IR) and ultraviolet-visible (UV-vis) spectra were analyzed and visualized using the program SWizard [33]. In all cases,

the displayed spectra show the calculated frequencies and absorption wavelengths.

The condensed Fukui functions were calculated using AOMIX molecular analysis software [34, 35], starting from single-point energy calculations.

Results and discussion

The molecular structure of Dye 7 was analyzed at different levels of theory, as mentioned above. The bond lengths obtained from our calculations as well as experimental data reported in the literature on the systems that comprise our dye are shown in Table 1 for the most representative bonds. It is clear that there is good agreement among the results obtained for different models for each bond. However, to check which methodology gives the most accurate results for the molecular structure of Dye 7, a statistical technique known as population standard deviation was applied (PSTD) to the results for the bond lengths. When applying this technique, the experimental result was used as a reference for the average value; in this way, the models that give the lowest deviation will give the best representations of our study system. It is important to note that we have not considered a tolerance level for the deviation, as our priority was to establish the model that best fits the experimental results. The results from the PSTD are shown in Table 2, which indicates that the most accurate methodology for the theoretical study is M05-2X/6-31G(d). The interatomic bond lengths (Å) and the angles (in degrees) calculated at this level of calculation are shown in Fig. 2.

A second validation of the models involved comparing the theoretical wavelengths of maximum absorption (λ_{\max})

Table 1 Bond lengths calculated at different levels of theory for Dye 7, and well as experimental data reported in the literature

Model	C1–C2	C3–C4	C4–C5	C1–C6	C1–H7	C13–H12	C13–C14	C16–C19	C15–C17	C23–C24	C24–C26
BLYP/3-21G*	1.415	1.404	1.408	1.404	1.089	1.089	1.415	1.404	1.408	1.421	1.395
B3LYP/3-21G*	1.403	1.394	1.397	1.394	1.082	1.082	1.403	1.394	1.397	1.408	1.387
PBE/3-21G*	1.412	1.401	1.406	1.401	1.092	1.092	1.412	1.402	1.406	1.418	1.393
PBE1PBE/3-21G*	1.400	1.391	1.395	1.391	1.084	1.084	1.400	1.391	1.395	1.405	1.385
TPSS/3-21G*	1.410	1.401	1.405	1.401	1.087	1.087	1.410	1.401	1.405	1.416	1.393
TPSSh/3-21G*	1.406	1.397	1.401	1.397	1.084	1.084	1.406	1.397	1.401	1.411	1.389
M05-2X/3-21G*	1.397	1.390	1.393	1.390	1.079	1.079	1.397	1.390	1.393	1.400	1.386
BLYP/6-31G(d)	1.414	1.404	1.407	1.403	1.092	1.092	1.414	1.403	1.407	1.420	1.395
B3LYP/6-31G(d)	1.403	1.394	1.397	1.394	1.085	1.085	1.403	1.394	1.396	1.408	1.387
PBE/6-31G(d)	1.410	1.400	1.403	1.399	1.094	1.094	1.410	1.400	1.403	1.416	1.392
PBE1PBE/6-31G(d)	1.399	1.390	1.393	1.390	1.086	1.086	1.399	1.390	1.393	1.403	1.384
TPSS/6-31G(d)	1.408	1.399	1.402	1.399	1.088	1.088	1.408	1.399	1.402	1.414	1.391
TPSSh/6-31G(d)	1.404	1.395	1.398	1.395	1.086	1.086	1.404	1.395	1.398	1.409	1.388
M05-2X/6-31G(d)	1.397	1.390	1.392	1.390	1.082	1.082	1.397	1.390	1.392	1.400	1.385
Experimental	1.397	1.397	1.397	1.397	1.084	1.084	1.397	1.397	1.397	1.397	1.397
Model	C26–C30	C24–H27	C2–N33	C30–C34	C34–C35	C34–H36	C35–C37	C37–C38	C37–S39	C38–C40	C38–H41
BLYP/3-21G*	1.425	1.089	1.448	1.459	1.371	1.095	1.444	1.412	1.764	1.409	1.088
B3LYP/3-21G*	1.411	1.082	1.434	1.456	1.355	1.088	1.442	1.397	1.741	1.404	1.080
PBE/3-21G*	1.422	1.091	1.436	1.454	1.369	1.098	1.440	1.412	1.748	1.405	1.090
PBE1PBE/3-21G*	1.407	1.083	1.424	1.452	1.352	1.089	1.440	1.395	1.726	1.401	1.081
TPSS/3-21G*	1.421	1.087	1.440	1.455	1.368	1.093	1.441	1.410	1.750	1.405	1.085
TPSSh/3-21G*	1.415	1.084	1.434	1.454	1.361	1.090	1.441	1.403	1.740	1.404	1.082
M05-2X/3-21G*	1.402	1.079	1.426	1.462	1.343	1.084	1.450	1.386	1.721	1.409	1.076
BLYP/6-31G(d)	1.423	1.092	1.439	1.455	1.374	1.097	1.443	1.410	1.766	1.405	1.091
B3LYP/6-31G(d)	1.410	1.085	1.427	1.453	1.358	1.089	1.442	1.396	1.743	1.400	1.084
PBE/6-31G(d)	1.419	1.093	1.428	1.449	1.370	1.098	1.438	1.408	1.748	1.400	1.092
PBE1PBE/6-31G(d)	1.405	1.085	1.417	1.449	1.353	1.090	1.439	1.392	1.727	1.397	1.084
TPSS/6-31G(d)	1.418	1.088	1.431	1.451	1.369	1.092	1.439	1.406	1.748	1.400	1.087
TPSSh/6-31G(d)	1.412	1.085	1.426	1.451	1.362	1.089	1.440	1.400	1.739	1.398	1.084
M05-2X/6-31G(d)	1.401	1.082	1.418	1.459	1.346	1.086	1.450	1.386	1.723	1.404	1.080
Experimental	1.397	1.084	1.418	1.475	1.334	1.099	1.475	1.370	1.714	1.423	1.079
Model	C42–S39	C40–C42	C35–H44	C42–C45	C45–C46	C46–C48	C46–C49	C48–N50	C49–O51	C49–O52	O52–S3H
BLYP/3-21G*	1.781	1.412	1.094	1.425	1.383	1.417	1.488	1.182	1.244	1.400	1.010
B3LYP/3-21G*	1.758	1.397	1.086	1.421	1.367	1.411	1.477	1.168	1.231	1.376	0.996
PBE/3-21G*	1.763	1.411	1.096	1.422	1.380	1.412	1.481	1.183	1.243	1.390	1.008
PBE1PBE/3-21G*	1.742	1.395	1.087	1.419	1.363	1.408	1.472	1.167	1.228	1.366	0.992
TPSS/3-21G*	1.765	1.409	1.091	1.422	1.379	1.413	1.480	1.181	1.243	1.394	1.005
TPSSh/3-21G*	1.756	1.403	1.088	1.421	1.372	1.411	1.476	1.175	1.237	1.383	0.999
M05-2X/3-21G*	1.735	1.387	1.082	1.428	1.353	1.413	1.473	1.159	1.224	1.366	0.991
BLYP/6-31G(d)	1.784	1.410	1.095	1.426	1.389	1.427	1.492	1.179	1.230	1.375	0.986
B3LYP/6-31G(d)	1.761	1.396	1.088	1.423	1.372	1.424	1.483	1.165	1.216	1.353	0.975
PBE/6-31G(d)	1.765	1.408	1.097	1.422	1.385	1.422	1.486	1.179	1.227	1.364	0.985
PBE1PBE/6-31G(d)	1.743	1.393	1.088	1.421	1.367	1.420	1.479	1.164	1.212	1.343	0.972
TPSS/6-31G(d)	1.765	1.406	1.090	1.423	1.384	1.423	1.484	1.176	1.226	1.366	0.983
TPSSh/6-31G(d)	1.756	1.400	1.088	1.422	1.376	1.422	1.481	1.170	1.220	1.356	0.978
M05-2X/6-31G(d)	1.739	1.387	1.084	1.430	1.358	1.428	1.482	1.157	1.209	1.342	0.971
Experimental	1.714	1.370	1.099	1.475	1.334	1.475	1.475	1.172	1.200	1.334	0.970

Table 2 Results of the population standard deviation for the bond lengths obtained with different models

Model	Population standard deviation
BLYP/3-21G*	0.0258
B3LYP/3-21G*	0.0189
PBE/3-21G*	0.0236
PBE1PBE/3-21G*	0.0175
TPSS/3-21G*	0.0232
TPSSh/3-21G*	0.0203
M05-2X/3-21G*	0.0150
BLYP/6-31G(d)	0.0238
B3LYP/6-31G(d)	0.0171
PBE/6-31G(d)	0.0214
PBE1PBE/6-31G(d)	0.0159
TPSS/6-31G(d)	0.0204
TPSSh/6-31G(d)	0.0179
M05-2X/6-31G(d)	0.0130

for Dye 7 obtained using the models with the value obtained experimentally in research conducted by Zhang et al. [4]. In their research, they found that λ_{\max} occurred at 432 nm in the solvent tetrahydrofuran (THF). Knowing this fact, it was possible to re-validate which of the models best represented our molecular system. Thus, the UV-vis spectrum of Dye 7 in THF solvent was calculated at the B3LYP/6-31G(d), BLYP/6-31G(d), M05-2X/6-31G(d), PBE1PBE/6-31G(d), PBE/6-31G(d), TPSS/6-31G(d) and TPSSh/6-31G(d) levels of theory using time-dependent DFT (TD-DFT). The results obtained are shown in Fig. 3.

The calculated UV-vis spectra are provided in Table 3, and it is quite apparent that the closest theoretical value of λ_{\max} to the experimental one is 444.44 nm, which was afforded by M05-2X/6-31G(d).

The calculated value of λ_{\max} is an important parameter which indicates that this molecular system should be considered for use as a functional material (a dye in this case) in a DSSC, as the value of this parameter for Dye 7 falls within the range of the solar spectrum of visible light [36].

At all of the levels of theory tested, the observed signal corresponded to the HOMO (highest occupied molecular orbital) to LUMO (lowest unoccupied molecular orbital) transition. Table 4 shows the results of TD-DFT calculations performed using the functional M05-2X and the basis set 6-31G(d), including the electronic state transitions, their corresponding wavelengths (in nm) and energies (in eV), as well as their assignments in terms of the orbitals involved in the transitions.

Taking into account the results for the molecular structure and UV-vis spectrum calculations, it is clear that the functional M05-2X and the basis set 6-31G(d) is the

most appropriate level of calculation to perform the rest of the characterization of Dye 7.

The infrared spectrum (IR) for Dye 7 calculated with M05-2X/6-31G(d) is shown in Fig. 4. The vibrational bands were assigned using the molecular visualization software for Windows ChemCraft. C–S stretching is observed as a peak at 584 cm^{-1} . At 796 cm^{-1} , a peak due to vibrations of carbon-chain hydrogens out of the plane of the aromatic rings can be seen, while another peak due to the corresponding vibrations for the bending of C–H in thiophene is present at 1091 cm^{-1} . Vibrations due to the bending of the bond O–H produce a peak at 1236 cm^{-1} . Other intense peaks include those due to the stretching of the C–N bond in the amine and the C–C in the thiophene occurring at 1392 cm^{-1} and 1521 cm^{-1} respectively; meanwhile, at 1677 cm^{-1} , a peak due to the double-bond stretching of C(45)=C(46) is noted. The peak at 1872 cm^{-1} represents the stretching of the double bond C=O, while the vibration at 2417 cm^{-1} corresponds to the stretching of the triple bond C≡N. The C–H vibrations for the aromatic rings occur at 3244 cm^{-1} , and stretching of the O–H bond is observed at 3784 cm^{-1} .

The molecular dipole moment is an experimental measure of the charge distribution in a molecule. The precision of the global distribution of electrons in a molecule is difficult to quantify, since it involves all multipoles. In this calculation, the values of the total energy of the system, the total dipole moment and the isotropic polarizability in the fundamental state obtained at the M05-2X/6-31G(d) level of calculation are –1736.99 a.u., 6.7374 debye and 433.06 bohr³. Moreover, the calculated energies of the HOMO and LUMO are 6.29 eV and 1.85 eV, respectively. These results are of great importance, since they can be used during synthesis to determine the solubility and chemical reactivity of the molecule, and they can also be employed in organic electronics and photovoltaics, as reported in different works [37–39].

The free energy of solvation $\Delta G(\text{solv})$ of the molecule was calculated for Dye 7 using M05-2X/6-31G(d) coupled with the integral equation formalism of the polarized continuum model (IEF-PCM) for different solvents. The solubility of a molecule depends on several kinetic and thermodynamic factors. However, the magnitude and sign of $\Delta G(\text{solv})$ can be used as an approximate index of solubility. In this sense, a negative sign and a large magnitude indicates increased solubility. The results of this calculation for the studied molecule can be summarized as follows: cyclohexane=–1.85 kcal mol^{–1}, chloroform=–4.68 kcal mol^{–1}, water=–5.02 kcal mol^{–1}, THF=–5.30 kcal mol^{–1}, acetone=–12.86 kcal mol^{–1}, ethanol=–14.43 kcal mol^{–1} and methanol=–15.03 kcal mol^{–1}. Based on these results, it appears that the molecule under investigation will be most soluble in methanol and ethanol.

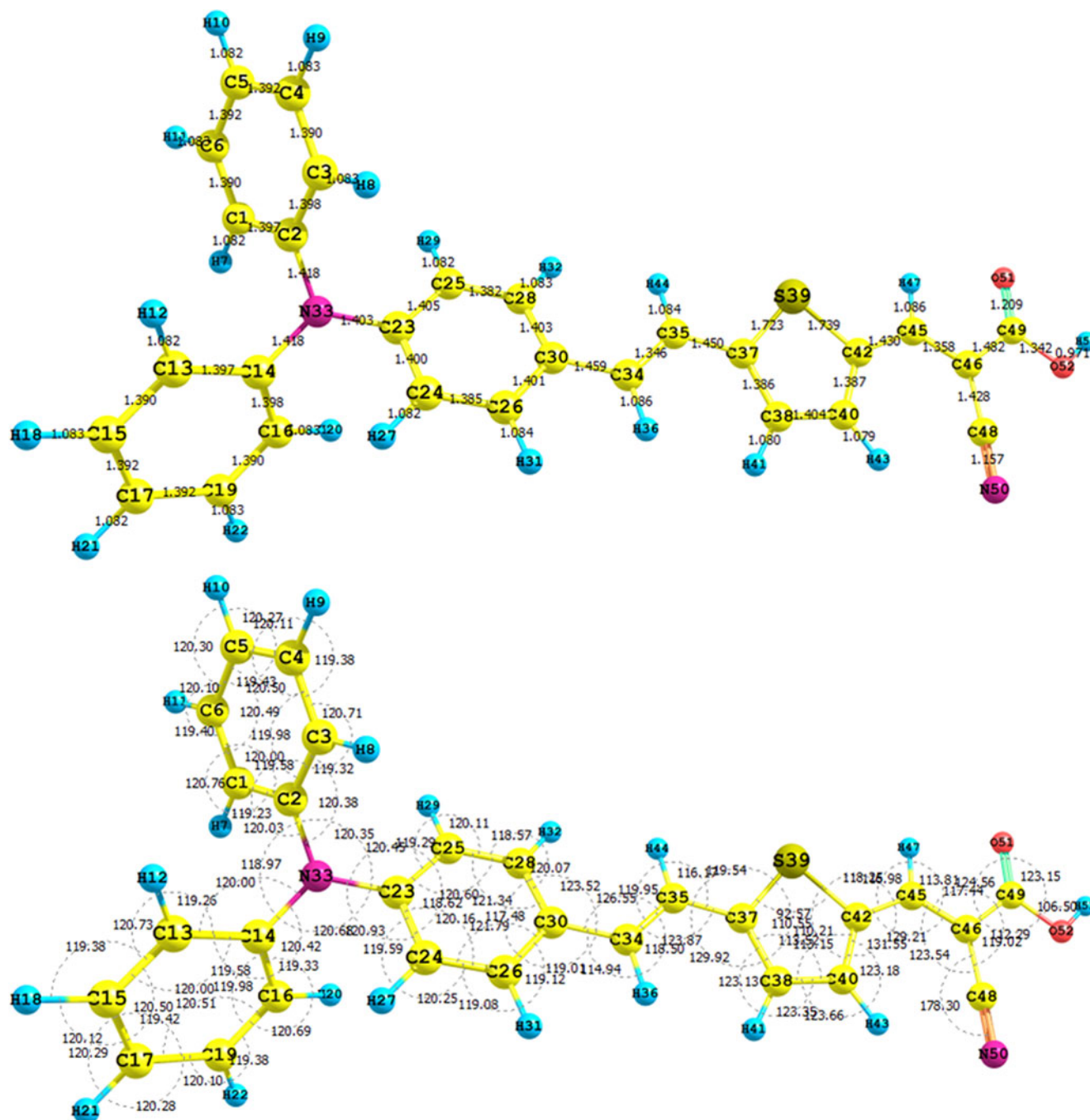


Fig. 2 Interatomic bond distances (Å) and bond angles (in degrees) for Dye 7 obtained at the M05-2X/6-31G(d) level of calculation

The HOMO and LUMO orbitals of Dye 7 calculated at the M05-2X/6-31G(d) level of theory are shown in Fig. 5. The HOMO orbital density is located over the double bonds of the carbon chain and the nitrogen (N33); meanwhile, the density of the LUMO orbital is concentrated over the C–C single bonds. This provides a good idea of the reactivity of the molecule.

The reactive sites can be identified through these orbital densities. The calculated HOMO and LUMO

densities shown in Fig. 5 indicate that electrophilic attack may occur preferentially at the C=C double bonds or at N33, while nucleophilic attack occurs at C–C single bonds.

The condensed Fukui functions can also be used to determine the reactivity of each atom in the molecule. The corresponding condensed Fukui functions are $f_k^+ = q_k(N+1) - q_k(N)$ (for nucleophilic attack), $f_k^- = q_k(N) - q_k(N-1)$ (for electrophilic attack) and $f_k^0 =$

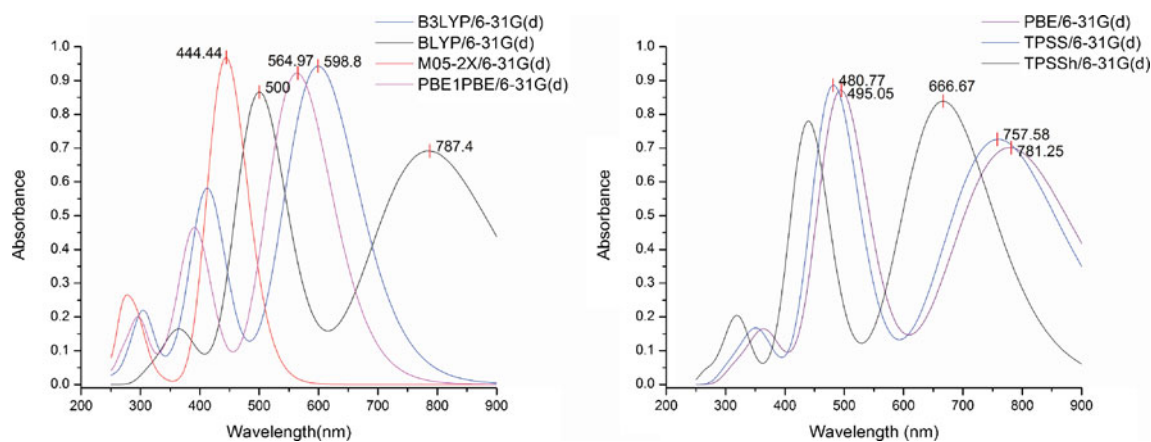


Fig. 3 Ultraviolet-visible (UV-vis) spectrum of Dye 7 calculated using time-dependent DFT (TD-DFT) with the basis set 6-31G(d) and the functionals used in this research

Table 3 Values of the wavelength of maximum absorption by Dye 7 calculated using the various models tested

Model	λ_{\max} (nm)
BLYP/6-31G(d)	500.00
B3LYP/6-31G(d)	598.80
PBE/6-31G(d)	495.05
PBE1PBE/6-31G(d)	564.97
TPSS/6-31G(d)	480.77
TPSSh/6-31G(d)	666.67
M05-2X/6-31G(d)	444.44
Experimental	432.00

$[q_k(N+1) - q_k(N-1)]/2$ (for radical attack), where q_k is the effective Mulliken charge of atom k in the molecule.

The calculations of the condensed Fukui functions for nucleophilic and electrophilic attacks were performed using AOMIX (a molecular analysis program), which gave the following results: $f_k^+ = 0.1876$ and $f_k^- = 0.1906$.

Electrophilic attack will occur at atoms that produce a negative charge, and where the Fukui function f_k^- is a maximum. This value confirms that the most probable site of electrophilic attack is N33. Nucleophilic attacks, on the other hand, will occur at atoms that produce a positive charge and where the Fukui function f_k^+ is a maximum.

Table 4 Electronic transition states of Dye 7 (calculated with TD-DFT at the M05-2X/6-31G(d) level of theory)

State	Wavelength (nm)	Energy (eV)	f	Assignment (H = HOMO, L = LUMO)
1	444.1	2.79	1.7821	S H-0→L+0(+73%) H-1→L+0(11%)
2	325.2	3.81	0.0338	S H-1→L+0(+61%) H-0→L+1(+12%) H-0→L+0(+8%)
3	295.5	4.2	0.2897	S H-0→L+1(+64%) H-0→L+0(12%) H-1→L+0(9%)
4	280.2	4.42	0.0564	S H-0→L+2(+79%)
5	271.3	4.57	0.2703	S H-0→L+3(+81%) H-1→L+3(+8%)
6	268.1	4.62	0.0851	S H-7→L+0(+71%) H-4→L+0(+12%)
7	252.1	4.92	0.0215	S H-0→L+5(+33%) H-0→L+4(9%) H-6→L+0(+8%) H-1→L+1(+8%) H-3→L+0(+6%)
8	250.8	4.94	0.0052	S H-1→L+1(+18%) H-4→L+0(12%) H-0→L+5(12%) H-7→L+0(+9%) H-0→L+4(6%)
9	243.4	5.09	0.0001	S H-9→L+0(+67%) H-9→L+8(12%) H-9→L+1(9%)
10	241	5.15	0.0261	S H-0→L+6(+57%) H-4→L+0(+10%) H-1→L+6(+8%) H-3→L+3(5%)
11	239.2	5.18	0.0028	S H-1→L+1(+24%) H-4→L+0(+23%) H-2→L+0(14%) H-0→L+4(11%) H-0→L+6(8%)
12	234.1	5.3	0.0824	S H-3→L+0(+40%) H-0→L+5(24%) H-6→L+0(+14%)
13	227.8	5.44	0.001	S H-0→L+7(+47%) H-1→L+7(39%)
14	223.2	5.55	0.0095	S H-0→L+4(+49%) H-1→L+1(+26%) H-0→L+1(+7%)
15	221	5.61	0.0054	S H-2→L+0(+68%) H-4→L+0(+17%)
16	215.4	5.76	0.0469	S H-6→L+0(+51%) H-3→L+0(33%)
17	209.6	5.92	0.0093	S H-5→L+0(+88%) H-5→L+1(+6%)
18	206.6	6	0.0012	S H-1→L+10(+17%) H-15→L+0(16%) H-0→L+10(12%) H-9→L+8(6%)
19	206	6.02	0.0678	S H-8→L+0(+42%) H-4→L+1(+15%) H-10→L+0(13%) H-4→L+0(+6%)
20	203.1	6.1	0.0158	S H-1→L+2(+35%) H-2→L+3(10%) H-2→L+2(+6%) H-0→L+5(+6%)

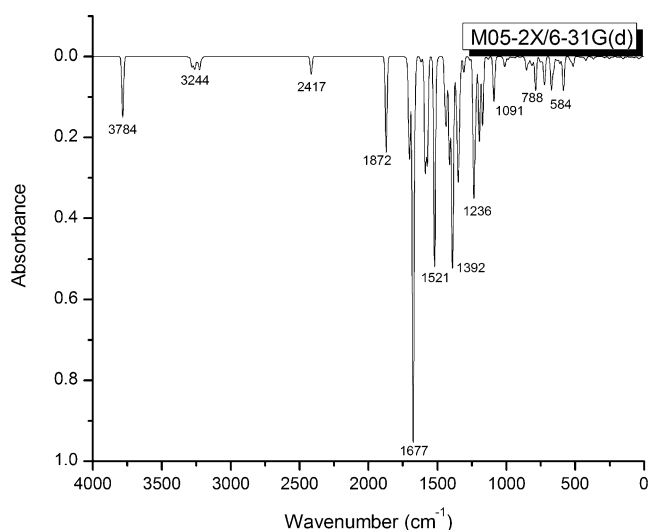


Fig. 4 Infrared spectrum of Dye 7 calculated at the M05-2X/6-31G(d) level of theory

Thus, the atom most likely to suffer a nucleophilic attack is C45.

Conclusions

In this work, a general comparison of the optimizations of the molecular structure and the ultraviolet spectrum in THF solvent achieved with different density functionals and basis sets was performed. This comparison indicated that the functional that gave results that were closest to the experimental results was M05-2X, along with the basis set 6-31G(d), so this level of theory was then used to study Dye 7 molecule, which is intended for use in photovoltaic devices. The total energy of this system, its dipole moment,

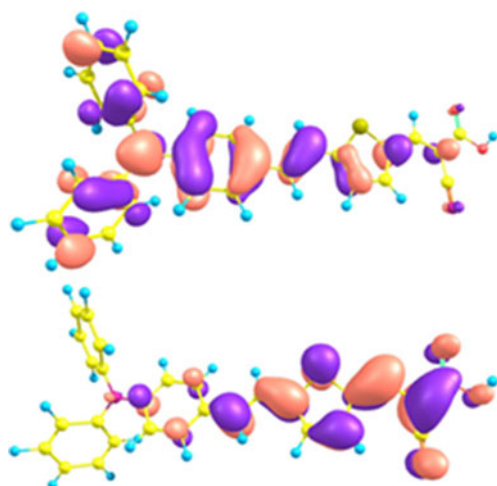


Fig. 5 HOMO and LUMO orbitals of Dye 7 calculated at the M05-2X/6-31G(d) level of theory

its isotropic polarizability, its molecular orbitals and its infrared spectrum were calculated using M05-2X/6-31G(d).

The free energy of solvation $\Delta G(\text{solv})$ of the molecule, calculated using the same level of theory along with the integral equation formalism of the polarized continuum model (IEF-PCM), indicates that the molecule is potentially soluble in methanol and ethanol.

The M05-2X/6-31G(d) methodology can be used as an useful tool for studying the molecular structure and electronic properties of Dye 7, as well as other structures derived from it.

Acknowledgments This work was made possible by the support of Universidad Autónoma de Sinaloa through the Facultad de Ingeniería Mochis, by the PROFAPI2010/033 project, Centro de Investigación en Materiales Avanzados, S.C. (CIMAV), and Consejo Nacional de Ciencia y Tecnología.

References

- Grätzel M (2006) Photovoltaic performance and long-term stability of dye-sensitized mesoscopic solar cells. *C R Chimie* 9:578–583. doi:10.1016/j.crci.2005.06.037
- Grätzel M (2003) Dye-sensitized solar cells. *J Photochem Photobiol C* 4:145–153. doi:10.1016/S1389-5567(03)00026-1
- Grätzel M (2004) Conversion of sunlight to electric power by nanocrystalline dye-sensitized solar cells. *J Photochem Photobiol A* 164:3–14. doi:10.1016/j.jphotochem.2004.02.023
- Fan Z, Yan H, Jin S, Xiao Z, Wei L, Chun M, Yong H, Mao F, Zhishan B, Qing M (2009) Triphenylamine-based dyes for dye-sensitized solar cells. *Dyes Pigments* 81:224–230. doi:10.1016/j.dyepig.2008.10.012
- Hwang S, Lee J, Park C, Lee H, Kim C, Park C, Lee M, Lee W, Park J, Kim K, Park N, Kim C (2007) A highly efficient organic sensitizer for dye-sensitized solar cells. *Chem Commun* 2007:4887–4889. doi:10.1039/b709859f
- Buscaino R, Baiocchi C, Barolo C, Medana C, Grätzel M, Nazeeruddin M, Viscardi G (2008) A mass spectrometric analysis of sensitizer solution used for dye-sensitized solar cell. *Inorg Chim Acta* 361:798–805. doi:10.1016/j.ica.2007.07.016
- Tachan Z, Rühle S, Zaban A (2010) Dye-sensitized solar tubes: a new solar cell design for efficient current collection and improved cell sealing. *Sol Energ Mater Sol Cells* 94:317–322
- Shen P, Liu Y, Huang X, Zhao B, Xiang N, Fei J, Liu L, Wang X, Huang H, Tan S (2009) Efficient triphenylamine dyes for solar cells: effects of alkyl-substituents and π -conjugated thiophene unit. *Dyes Pigments* 83:187–197. doi:10.1016/j.dyepig.2009.04.005
- Chang Y, Chow T (2009) Dye-sensitized solar cell utilizing organic dyads containing triarylene conjugates. *Tetrahedron* 65:4726–4734
- Casanova D, Rotzinger F, Grätzel M (2010) Computational study of promising organic dyes for high-performance sensitized solar cells. *J Chem Theor Comput* 6:1219–1227. doi:10.1021/ct100069q
- El-Shishtawy R (2009) Functional dyes, and some hi-tech applications. *Int J Photoenergy* 2009:1–21. doi:10.1155/2009/434897
- Hagberg D, Edvinsson T, Sun L (2006) A novel organic chromophore for dye-sensitized nanostructured solar cells. *Chem Commun* 2245–2247. doi:10.1039/b603002e
- Parr R, Yang W (1989) Density-functional theory of atoms and molecules. Oxford University Press, Oxford

14. Frisch MJ et al (2004) Gaussian 03W. Gaussian Inc., Wallingford
15. Foresman J, Frisch A (1996) Exploring chemistry with electronic structure methods. Gaussian Inc., Pittsburgh
16. Becke A (1993) Density-functional thermochemistry. III. The role of exact exchange. *J Chem Phys* 98:5648–5652. doi:10.1063/1.464913
17. Becke A (1988) Density functional exchange energy approximation with correct asymptotic behavior. *Phys Rev A* 38:3098–3100. doi:10.1103/PhysRevA.38.3098
18. Lee C, Yang W, Parr R (1988) Development of the Colle–Salvati correlation-energy formula into a functional of the electron density. *Phys Rev B* 37:785–789. doi:10.1103/PhysRevB.37.785
19. Stephens P, Devlin F, Chabalowski C, Frisch M (1994) Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional Force fields. *J Phys Chem* 98:11623–11627. doi:10.1021/j100096a001
20. Ernzerhof M, Scuseria G (1999) Assessment of the Perdew–Burke–Ernzerhof exchange–correlation functional. *J Chem Phys* 110:5029–5036. doi:10.1063/1.478401
21. 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
22. Tao J, Perdew J, Staroverov V, Scuseria G (2003) Climbing the density functional ladder: non-empirical meta-generalized gradient approximation designed for molecules and solids. *Phys Rev Lett* 91:1–4. doi:10.1103/PhysRevLett.91.146401
23. Staroverov V, Scuseria G, Tao J, Perdew J (2003) Comparative assessment of a new nonempirical density functional: molecules and hydrogen-bonded complexes. *J Chem Phys* 119:12129–12137. doi:10.1063/1.1626543
24. Zhao Y, Schultz N, Truhlar D (2006) Design of density functionals by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions. *J Chem Theor Comput* 2:364–382. doi:10.1021/ct0502763
25. Lewars E (2003) Computational chemistry: introduction to the theory and applications of molecular and quantum mechanics. Kluwer, Norwell
26. Young D (2001) Computational chemistry: a practical guide for applying techniques to real-world problems. Wiley, New York
27. Jensen F (2007) Introduction to computational chemistry. Wiley, Chichester
28. Cramer C (2002) Essentials of computational chemistry: theories and models. Wiley, Chichester
29. Burke K, Werschnik J, Gross E (2005) Time-dependent density functional theory: past, present, and future. *J Chem Phys* 123:1–9. doi:10.1063/1.1904586
30. Stratmann R, Scuseria G, Frisch M (1998) An efficient implementation of time-dependent density-functional theory for the calculation of excitation energies of large molecules. *J Chem Phys* 109:8218–8224. doi:10.1063/1.477483
31. Bauernschmitt R, Ahlrichs R (1996) Treatment of electronic excitations within the adiabatic approximation of time dependent density functional theory. *Chem Phys Lett* 256:454–464. doi:10.1016/0009-2614(96)00440-X
32. Casida M, Jamorski C, Casida K, Salahub D (1998) Molecular excitation energies to high-lying bound states from time-dependent density-functional response theory: characterization and correction of the time-dependent local density approximation ionization threshold. *J Chem Phys* 108:4439–4449. doi:10.1063/1.475855
33. Gorelsky S (2010) SWizard program. <http://www.sg-chem.net/>, accessed 29 Sept 2010
34. Gorelsky S (2010) AOMix program. <http://www.sg-chem.net/>, accessed 10 Sept 2010
35. Gorelsky S, Lever A (2001) Electronic structure and spectra of ruthenium diimine complexes by density functional theory and INDO/S. Comparison of the two methods. *J Organomet Chem* 635:187–196. doi:10.1016/S0022-328X(01)01079-8
36. Green M (1982) Solar cells: operating principles, technology, and systems applications. Prentice-Hall, Upper Saddle River
37. De Angelis F, Fantacci S, Sgamelloti A (2007) An integrated computational tool for the study of the optical properties of nanoscale devices: application to solar cells and molecular wires. *Theor Chem Acc* 117:1093–1104. doi:10.1007/s00214-006-0224-z
38. Weng Y, Wang Y, Asbury J, Ghosh H, Lian T (2000) Back electron transfer from TiO₂ nanoparticles to FeIII(CN)₆³⁻: origin of non-single-exponential and particle size independent dynamics. *J Phys Chem B* 104:93–104. doi:10.1021/jp992522a
39. Sharma S, Inamdar A, Im H, Kim B, Patil P (2011) Morphology dependent dye-sensitized solar cell properties of nanocrystalline zinc oxide thin films. *J Alloys Compd* 509:2127–2131. doi:10.1016/j.jallcom.2010.10.163

Problems with molecular mechanics implementations on the example of 4-benzoyl-1-(4-methyl-imidazol-5-yl)-carbonylthiosemicarbazide

Agata Siwek · Katarzyna Świderek · Stefan Jankowski

Received: 12 March 2011 / Accepted: 3 May 2011 / Published online: 28 May 2011
© Springer-Verlag 2011

Abstract Results from force fields implemented in HyperChem, a program frequently used in studies of bioactive compounds, have been compared using the example of the conformational analysis of a 1-carbonylthiosemicarbazide that exhibits strong antibacterial activity. By comparing these results with the original force fields and the experimental NMR ROESY spectrum, it was shown that these implementations can lead to erroneous results.

Keywords Molecular mechanics · Amber · CHARMM · OPLS · Conformational search · 1-Carbonylthiosemicarbazide · HyperChem

Introduction

In the past 20 years, quests for pharmaceutically active compounds have evolved from the “spray-and-pray” approach to computer-guided studies with high throughput analyses of chemical libraries. While not always successful,

this enhanced approach has resulted in recent years in the near-mandatory augmentation of bioactivity screening results with some form of molecular modeling of the studied compounds. Out of necessity (in the case of QM/MM [1] calculations for large receptors and enzymatic systems), or in order to save time, these calculations are usually carried out at low levels of theory, such as molecular mechanics. Most frequently, these calculations aim to establish descriptors that can be used in structure–activity relationships (SAR, QSAR, etc.), and during the initial phase they usually attempt to find the most stable conformation of the ligand. Even at such an early stage, this aim is not very easy to define, as a ligand bound to a receptor or in the active site of an enzyme may adopt a conformation that is quite different from the most stable one. Furthermore, the dielectric properties of the receptor or the active site are usually different (lower) than those of a bulk aqueous solution. It is thus not clear which conformation of a molecule should be sought. Different conformations result in different values of the descriptors used in the SAR analysis, so the results obtained from calculations may or may not have any relevance to the desired structure–activity relationship. It appears that the safest mode of action is to proceed from the most stable conformation, which can be obtained from the conformational search [2]. Alternatively, results from molecular dynamics or simulated annealing can be used.

Yet another problem arises from a technical standpoint: the number of parameters employed in most force fields is insufficient to describe ligands, and the missing ones should be evaluated prior to further studies. Since this is time-consuming, several approximate solutions are implemented, including the use of similar parameters, the on-the-fly evaluation of these missing parameters from the existing parameters, or the introduction of default generic parameters

A. Siwek (✉)
Department of Organic Chemistry, Faculty of Pharmacy,
Medical University,
Chodzki 4a,
20–093 Lublin, Poland
e-mail: agata.siwek@am.lublin.pl

K. Świderek
Institute of Applied Radiation Chemistry,
Technical University of Lodz,
Zeromskiego 116,
90–924 Lodz, Poland

S. Jankowski
Institute of Organic Chemistry, Technical University of Lodz,
Zeromskiego 116,
90–924 Lodz, Poland

in place of the missing ones. In studies of compounds with potential pharmaceutical activity, calculations are usually carried out using widely available software that does not necessarily include full implementations of molecular mechanical force fields. Furthermore, the starting conformation is frequently the one that corresponds to the local energy minimum closest to the structure introduced via the graphical interface of the program. This protocol can lead to erroneous results: a fact not fully appreciated in the literature. Herein, we illustrate these problems using the example of the conformational analysis of a 1-carbonylthiosemicarbazide using three protein-oriented force fields implemented in HyperChem, a program that is frequently employed in such calculations [3–66].

Experimental methods

Chemistry

4-Benzoyl-1-(4-methyl-imidazol-5-yl)-carbonylthiosemicarbazide was synthesized by the routine protocol [67–70]. In short, 4-methyl-imidazole-5-carboxylic acid hydrazide was reacted with benzoyl isothiocyanate as described in [71]. Its NMR spectrum was recorded on a Bruker Avance II Plus spectrometer at 25°C in DMSO- d_6 using the solvent methyl group signal as the internal standard (δ_H 2.50 or δ_C 40.0, respectively). The ROESY spectrum was recorded with a spin-lock time of 1 s.

Computational details

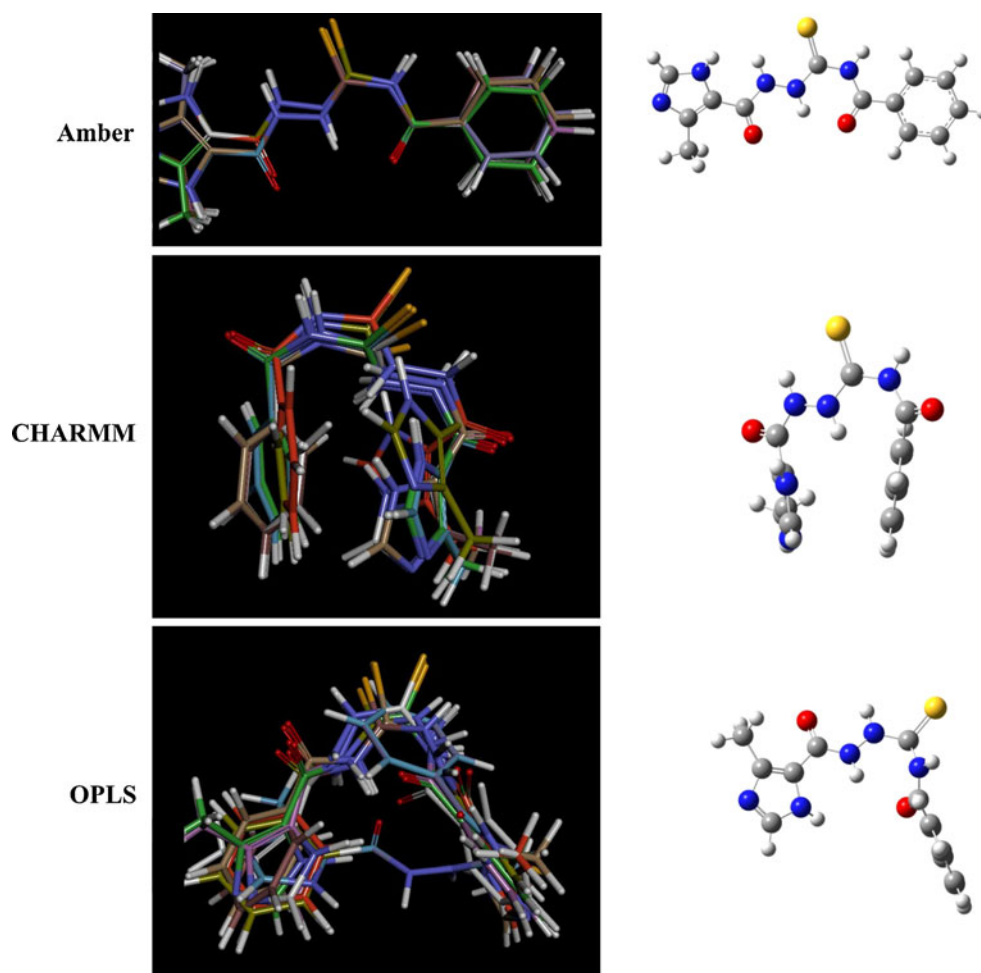
The conformational search engine of the HyperChem program [72] was employed at the MM level of theory, using the Amber99 [73], CHARMM [74], and OPLS [75] force fields as implemented in version 8. All dihedral angles along the longest chains between and including substituents were considered. Optimization of the structures obtained from the conformational searches was performed in Gaussian09 [76] using the M05-2X DFT functional [77, 78] with the 6-31+G(d,p) basis set [79–81] and the SMD continuum solvent model [82]. Theoretical NMR shifts were obtained using the B3PW91 DFT functional [83, 84] with the 6-31+G(d) basis set for optimization and 6-311++G(2df,p) [85] for energy calculations. Solvent was modeled using the continuum IEF-PCM model [86–88]. The default threshold was used in all geometry optimization calculations, while it was increased tenfold to 0.01 kcal mol⁻¹Å⁻¹ in MM calculations. All structures obtained at QM levels were confirmed to be stationary points corresponding to energy minima by vibrational analysis ($3n - 6$ normal modes of vibration, where n stands for the number of atoms in the molecule).

Results and discussion

When used in lieu of the crystal structure of the pharmaceutically active form of a drug bound into the active (or allosteric) site of an enzyme, computations can provide only speculative suggestions regarding the source of its bioactivity. Nevertheless, useful information can be inferred in favorable cases. One of the most important questions that must be taken into account in order to calculate the energetics of binding is the selection of the appropriate conformational states. 1-Carbonylthiosemicarbazide skeletons have six rotatable bonds, leading to a huge conformational space, which should be sampled in order to find the most stable conformations. Such a large space cannot be searched systematically at either the ab initio or DFT level. The problem is further complicated by the fact that there is no guarantee that the most stable conformations are relevant to bioactivity. In fact, in terms of the chemical reactivities of these compounds, we have shown that dehydrocyclizations occur from the energy-rich conformation due to the geometrical requirements of these reactions [89]. In the case of enzymatic reactions, conformational flexibility is not restricted to aqueous solution but frequently also plays an important role in protein-bound ligands [90]. The problem of conformational preference is also interesting from another point of view; flexible binding and docking studies frequently rely on geometries that have been optimized using empirical force fields, so it is interesting to see how these perform when applied to molecules that they were not optimized for. We therefore carried out extensive conformational searches for a 1-carbonylthiosemicarbazide at the molecular mechanics level in order to compare structures obtained from different force fields with the experimental results.

We used the conformational search engine implemented in HyperChem with three force fields used in the modeling of proteins: Amber, CHARMM, and OPLS. All seven dihedral angles (six from the carbonylthiosemicarbazide skeleton and one from the benzoyl moiety) along the longest chain between the aromatic rings were considered. The ten most stable structures obtained with each force field are overlaid in Fig. 1. All of these structures were further optimized at the DFT level, and the results of these calculations are compared graphically in Fig. 2. The most stable structure obtained from each force field is illustrated by the structure in the last column of Fig. 1. Confusingly, the most stable conformations obtained with these three force fields are very different. Table 1 compares the torsional angles of the conformers of lowest energy. The atom numbering used to define angles is given in Fig. 3. Amber calculations yielded a stretched structure that is practically planar. In the structure obtained with CHARMM, aromatic rings are stacked, while the planes of these rings are nearly perpendicular and the chain between them is looped due to the presence of the internal

Fig. 1 Most stable conformations obtained with Amber, CHARMM, and OPLS force fields



hydrogen bond in the structure obtained with OPLS. These results highlight the fact that protocols that employ molecular mechanics only should be used (e.g., those used in most molecular dynamics and docking studies) with caution.

Because of these significant differences in the optimal conformations obtained with different force fields, we have carried out ROESY $^1\text{H-NMR}$ experiments to deduce which

of the conformations presented in Fig. 1 represents the structure in solution. For this purpose, proton NMR spectra and ROESY experiments were carried out in DMSO. Complete assignment of resonances was based on the set of COSY, HSQC and HMBC spectra [67–70]. An examination of through-space proton–proton interactions by means of ROESY experiments highlighted correlations within the benzamide residue only; the ROESY spectrum (see Fig. 4) shows only a weak correlation between protons at 11.88 (C(S)NH $\underline{\text{C}}$ (O)) and 7.99 ppm (Phe $_{2,6}$) and a strong correlation between ortho (7.99 ppm) and meta (7.55)

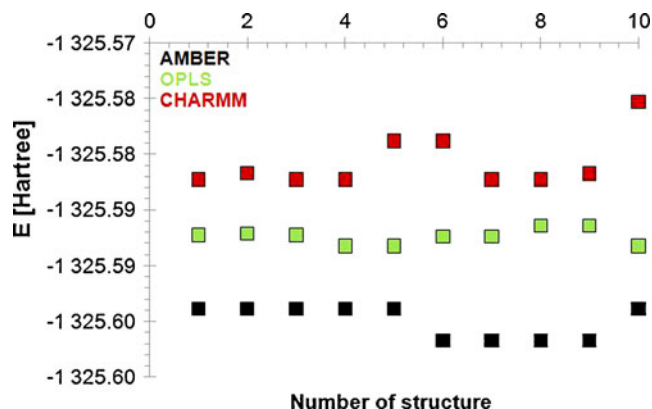


Fig. 2 Comparison of the energies of the top ten structures obtained from conformational searches using Amber, CHARMM and OPLS

Table 1 Torsional angles of the conformers with the lowest energies

Torsional angle	Amber	CHARMM	OPLS
5-4-6-8	-17.8	40.7	-23.1
4-6-8-9	178.8	21.8	174.6
6-8-9-10	180.0	164.4	179.1
8-9-10-12	178.8	-165.4	-7.4
9-10-12-13	2.7	-6.9	80.1
10-12-13-15	180.0	39.6	175.6
12-13-15-20	0.1	-136.7	28.4

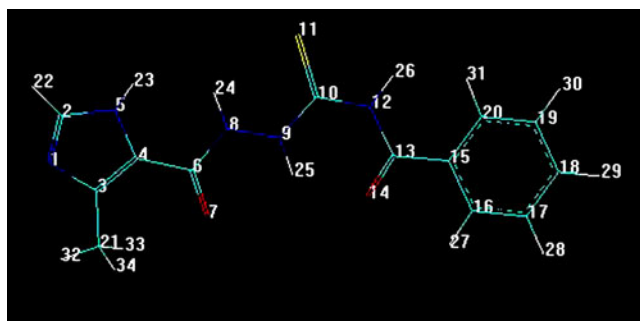


Fig. 3 Atom numbering used to define torsional angles

phenyl protons. From these results, we were able to conclude that hydrogen atoms of terminal aromatic rings are separated by at least 5 Å. This excludes the stacked structure obtained from CHARMM calculations from being the dominant conformation, because in this structure phenyl ring rotation is hindered geometrically, and the distances between hydrogen atoms from different rings are in the range of 3.2 to 4.8 Å. Also, the structure obtained from OPLS calculations is questionable because some contacts between hydrogen atoms of terminal rings are at a distance of about 4.7–4.8 Å, although this cannot be rigorously excluded from considerations based on the NMR results. The assignment of NMR signals was confirmed by theoretical calculations that were carried at the B3PW91/6-311++G(2df,p)//B3PW91/6-31+G(d) level using a continuum solvent model of DMSO solution. The DFT-optimized geometries agree with those obtained with the Amber force field, and are practically identical in the gas phase, in DMSO, and in the aqueous solution.

Since the OPLS results could not be rigorously rejected, we have compared them with the analogous calculations carried out using the OPLS-2005 implementation in the Impact program of the Schrodinger package [91]. This implementation contains all types of atoms present in the studied compound, so we assume that the results obtained

Fig. 4 ROESY spectrum (*left*) and the enlargement of the central part (*right*)

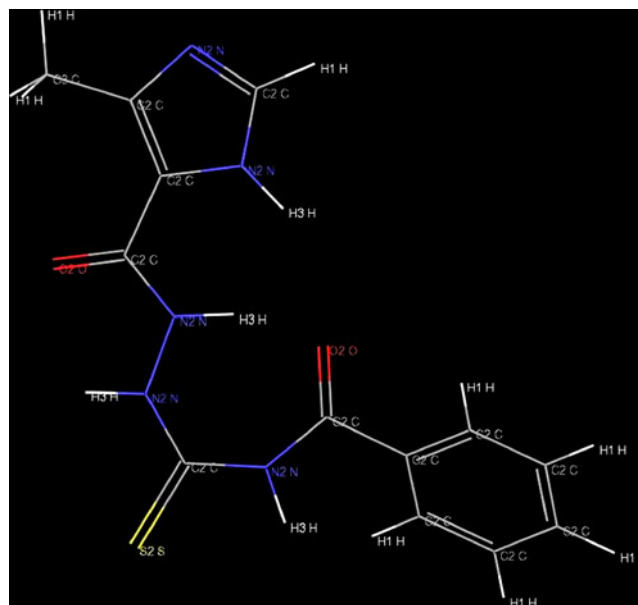
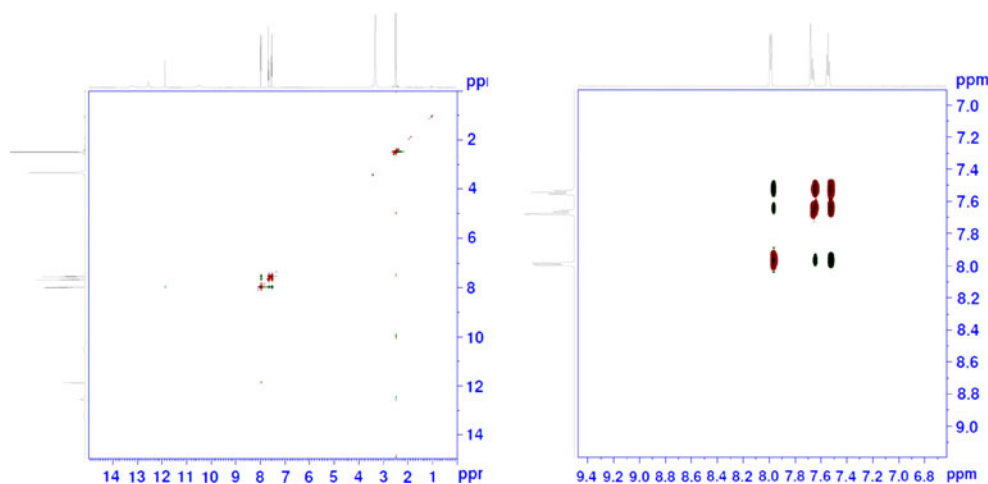


Fig. 5 The most stable conformation obtained using the OPLS-2005 implementation in the Schrodinger package

using it are the correct results for the OPLS force field. In particular, the structures that are most stable in the HyperChem implementation of Amber, CHARMM, and OPLS (see Fig. 1) were reoptimized using the Impact implementation of OPLS. The Amber and CHARMM structures did not change significantly. The conformation obtained from the best HyperChem OPLS structure underwent the most changes (see Fig. 5), and in fact approached the one obtained using Amber, with the whole molecule being almost planar. The only major difference is one dihedral angle (4-6-8-9), which is nearly 180° in Amber but nearly 0° in OPLS-2005 (the Impact implementation of OPLS), making the molecule less elongated. Most importantly, however, in OPLS-2005, the structure obtained from Amber as implemented in HyperChem is the most stable; it is 1.6 kcal mol⁻¹ more stable than the one obtained from

the HyperChem implementation of OPLS and 16.9 kcal mol⁻¹ more stable than the one obtained from the HyperChem CHARMM implementation.

Conclusions

We have compared the performances of three popular force fields used to model enzymatic reactions and in docking studies when they were applied to find the conformation of a 1-carbonylthiosemicarbazide that exhibits strong antibacterial activity [67–70]. By comparing the theoretical results with experimental NMR data, we concluded that the Amber99 and OPLS-2005 force fields yielded the correct structure. Our finding parallels a recent report on a similar force field comparison for peptides [92]. Our results indicate that incomplete implementations of the force fields (like those used in HyperChem) can lead to erroneous results.

References

- Warshel A, Levitt M (1976) Theoretical studies of enzymatic reactions: dielectric electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J Mol Biol* 103:227–249
- Chen IJ, Follope N (2011) Is conformational sampling of drug-like molecules a solved problem? *Drug Develop Res* 72:85–94
- Guével RL, Oger F, Lecorgne A, Dudasova Z, Chevance S, Bondon A, Barath P, Simonneaux G, Salbert G (2009) Identification of small molecule regulators of the nuclear receptor HNF4 α based on naphthofuran scaffolds. *Bioorg Med Chem* 17:7021–7030
- Paoletta S, Steventon GB, Wildeboer D, Ehrman TM, Hylands PJ, Barlow DJ (2008) Screening of herbal constituents for aromatase inhibitory activity. *Bioorg Med Chem* 16:8466–8470
- Narayanamy S, Thirumamagal BTS, Johnsamuel J, Byun Y, Al-Madhoun AS, Usova E, Cosquer GY, Yan J, Bandyopadhyaya AK, Tiwari R, Eriksson S, Tjarks W (2006) Hydrophilically enhanced 3-carboranyl thymidine analogues (3CTAs) for boron neutron capture therapy (BNCT) of cancer. *Bioorg Med Chem* 14:6886–6899
- Farkas V, Vass E, Hanssens I, Majer Z, Hollósi M (2005) Cyclic peptide models of the Ca²⁺-binding loop of α -lactalbumin. *Bioorg Med Chem* 13:5310–5320
- Khlebnikov AI, Schepetkin IA, Quinn MT (2008) Structure–activity relationship analysis of N-benzoylpyrazoles for elastase inhibitory activity: a simplified approach using atom pair descriptors. *Bioorg Med Chem* 16:2791–2802
- Dheyongera JP, Geldenhuys WJ, Dekker TG, Matsabisa MG, Van der Schyf CJ (2005) Antimalarial activity of thioacridone compounds related to the acronycine alkaloid. *Bioorg Med Chem* 13:1653–1659
- Plonska-Ocypa K, Sicinski RR, Plum LA, Grzywacz P, Frelek J, Clagett-Dame M, DeLuca HF (2009) 13-Methyl-substituted des-C,D analogs of (20S)-1 α ,25-dihydroxy-2-methylene-19-norvitamin D3 (2MD): synthesis and biological evaluation. *Bioorg Med Chem* 17:1747–1763
- Abdel-Aziz AA-M, El-Subbagh HI, Kunieda T (2005) Lewis acid-promoted transformation of 2-alkoxy-pyridines into 2-aminopyridines and their antibacterial activity. Part 2: remarkably facile C–N bond formation. *Bioorg Med Chem* 13:4929–4935
- Katritzky AR, Dobchev DA, Tulp I, Karelson M, Carlson DA (2006) QSAR study of mosquito repellents using Codessa Pro. *Bioorg Med Chem Lett* 16:2306–2311
- Fiorentino A, D'Abrosca B, Pacifico S, Iacovino R, Mastellone C, Di Blasio B, Monaco P (2006) Distachyasins: a new antioxidant metabolite from the leaves of *Carex distachya*. *Bioorg Med Chem Lett* 16:6096–6101
- Stout EP, Prudhomme J, Le Roch K, Fairchild CR, Franzblau SG, Aalbersberg W, Hay ME, Kubanek J (2010) Unusual antimalarial meroditerpenes from tropical red macroalgae. *Bioorg Med Chem Lett* 20:5662–5665
- El-Ayaan U, Abdel-Aziz AA-M, Al-Shihry S (2007) Solvatochromism, DNA binding, antitumor activity and molecular modeling study of mixed-ligand copper(II) complexes containing the bulky ligand: bis[N-(p-tolyl)imino]acenaphthene. *Eur J Med Chem* 42:1325–1333
- Thakur A, Thakur M, Bharadwaj A, Thakur S (2008) SAR and QSAR studies: modelling of new DAPY derivatives. *Eur J Med Chem* 43:471–477
- Gupta AK, Gupta RA, Soni LK, Kaskhedikar SG (2008) Exploration of physicochemical properties and molecular modeling studies of 2-sulfonyl-phenyl-3-phenyl-indole analogs as cyclooxygenase-2 inhibitors. *Eur J Med Chem* 43:1297–1303
- Abdel-Aziz AA-M (2007) Novel and versatile methodology for synthesis of cyclic imides and evaluation of their cytotoxic, DNA binding, apoptotic inducing activities and molecular modeling study. *Eur J Med Chem* 42:614–626
- El-Kerdawy MM, El-Bendary ER, Abdel-Aziz AA-M, El-wassef DR, Abd El-Aziz NI (2010) Synthesis and pharmacological evaluation of novel fused thiophene derivatives as 5-HT_{2A} receptor antagonists: molecular modeling study. *Eur J Med Chem* 45:1805–1820
- Rescifina A, Chiacchio U, Corsaro A, Piperno A, Romeo R (2011) Isoxazolidinyl polycyclic aromatic hydrocarbons as DNA-intercalating antitumor agents. *Eur J Med Chem* 46:129–136
- da Silva SL, Calgarotto AK, Maso V, Damico DCS, Baldasso P, Veber CL, Villar JAFP, Oliveira ARM, Comar M Jr, Oliveira KMT, Marangoni S (2009) Molecular modeling and inhibition of phospholipase A₂ by polyhydroxy phenolic compounds. *Eur J Med Chem* 44:312–321
- Alieva IN, Mustafayeva NN, Gojayev NM (2006) Conformational analysis of the N-terminal sequence Met1–Val60 of the tyrosine hydroxylase. *J Mol Struct* 785:76–84
- de Sousa AS, Fernandes MA, Nxumalo W, Balderson JL, Jeftić T, Cukrowski I, Marques HM (2008) Sc(III) porphyrins. The molecular structure of two Sc(III) porphyrins and a re-evaluation of the parameters for the molecular mechanics modelling of Sc(III) porphyrins. *J Mol Struct* 872:47–55
- Gaber M, El-Daly SA, El-Sayed YSY (2009) Synthesis, spectral, thermal and theoretical studies of Cu(II) complexes with 3-[4'-dimethylaminophenyl]-1-(2-pyridyl)prop-2-en-1-one (DMAPP). *J Mol Struct* 922:51–57
- Mautner FA, Taylor ER, Rozas DM, Massoud SS (2009) Synthesis and structural characterization of dicopper(II) and dipalladium(II) complexes of 1,1,2,2-tetrakis(carboxamido-2-methylpyridyl)ethane. *J Mol Struct* 936:250–256
- Bairaga HR, Mukhopadhyay BP, Bhattacharya S (2009) Role of the conserved water molecules in the binding of inhibitor to IMPDH-II (human): a study on the water mimic inhibitor design. *J Mol Struct THEOCHEM* 908:31–39
- Pichierri F (2010) Macrodipoles of potassium and chloride ion channels as revealed by electronic structure calculations. *J Mol Struct THEOCHEM* 950:79–82
- Jankowski CK, Martel J-L, Femandjian S, Maroun RG (2005) Study of potential HIV-1 inhibition glutaric dialdehyde adducts. *J Mol Struct THEOCHEM* 731:83–87

28. Gikas E, Bazoti FN, Tsarbopoulos A (2007) Conformation of oleuropein, the major bioactive compound of *Olea europaea*. *J Mol Struct THEOCHEM* 821:125–132
29. Fendri A, Frikha F, Miled N, Bacha AB, Gargouri Y (2007) Modulating the activity of avian pancreatic lipases by an alkyl chain reacting with an accessible sulfhydryl group. *Biochem Biophys Res Commun* 360:765–771
30. Lorin A, Lins L, Stroobant V, Brasseur R, Charlotheaux B (2007) Determination of the minimal fusion peptide of bovine leukemia virus gp30. *Biochem Biophys Res Commun* 355:649–653
31. Shi YH, Song YL, Lin DH, Tan J, Roller PP, Li Q, Long YQ, Song GQ (2005) Binding affinity difference induced by the stereochemistry of the sulfoxide bridge of the cyclic peptide inhibitors of Grb2-SH2 domain: NMR studies for the structural origin. *Biochem Biophys Res Commun* 330:1254–1261
32. Dupiereux I, Zorzi W, Lins L, Brasseur R, Colson P, Heinen E, Elmoualij B (2005) Interaction of the 106–126 prion peptide with lipid membranes and potential implication for neurotoxicity. *Biochem Biophys Res Commun* 331:894–901
33. Katsara M, Yuriev E, Ramsland PA, Deraos G, Tselios T, Matsoukas J, Apostolopoulos V (2008) Mannosylation of mutated MBP83–99 peptides diverts immune responses from Th1 to Th2. *Mol Immunol* 45:3661–3670
34. Solórzano-Vargas RS, Vasilevko V, Acero G, Ugen KE, Martinez R, Govezensky T, Vazquez-Ramirez R, Kubli-Garfias C, Cribbs DH, Manoutcharian K, Gevorkian G (2008) Epitope mapping and neuroprotective properties of a human single chain FV antibody that binds an internal epitope of amyloid-beta 1–42. *Mol Immunol* 45:881–886
35. Zoroddu MA, Medici S, Peana M (2009) Copper and nickel binding in multi-histidinic peptide fragments. *J Inorg Biochem* 103:1214–1220
36. Kulon K, Valensin D, Kamysz W, Valensin G, Nadolski P, Porciatti E, Gaggelli E, Kozłowski H (2008) The His–His sequence of the antimicrobial peptide demegen P-113 makes it very attractive ligand for Cu^{2+} . *J Inorg Biochem* 102:960–972
37. Christofis P, Katsarou M, Papakyriakou A, Sanakis Y, Katsaros N, Psomas G (2005) Mononuclear metal complexes with Piroxicam: synthesis, structure and biological activity. *J Inorg Biochem* 99:2197–2210
38. Mucha A, Bal W, Jeżowska-Bojczuk M (2008) Comparative studies of coordination properties of puromycin and puromycin aminonucleoside towards copper(II) ions. *J Inorg Biochem* 102:46–52
39. Deconinck E, Xu QS, Put R, Coomans D, Massart DL, Vander Heyden Y (2005) Prediction of gastro-intestinal absorption using multivariate adaptive regression splines. *J Pharm Biomed Anal* 39:1021–1030
40. Al Omari AA, Al Omari MM, Badwan AA, Al-Sou'od KA (2011) Effect of cyclodextrins on the solubility and stability of candesartan cilexetil in solution and solid state. *J Pharm Biomed Anal* 54:503–509
41. Deconinck E, Hancock T, Coomans D, Massart DL, Heyden YV (2005) Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. *J Pharm Biomed Anal* 39:91–103
42. Pescitelli G, Bilia AR, Bergonzi MC, Vincieri FF, Di Bari L (2010) Cyclodextrins as carriers for kavalactones in aqueous media: Spectroscopic characterization of (*S*)-7,8-dihydrokavain and β -cyclodextrin inclusion complex. *J Pharm Biomed Anal* 52:479–483
43. Kumar RR, Perumal S (2007) A facile synthesis and highly atom economic 1,3-dipolar cycloaddition of hexahydropyrido[3,4-*c*][1,5]benzothiazepines with nitrile oxide: stereoselective formation of hexahydro-[1,2,4]oxadiazolo[5,4-*d*]pyrido[3,4-*c*][1,5]benzothiazepines. *Tetrahedron* 63:7850–7857
44. Campayo L, Calzado F, Cano MC, Yunta MJR, Pardo M, Navarro P, Jimeno ML, Gómez-Contreras F, Sanz AM (2005) New acyclic receptors containing pyridazine units. The influence of π -stacking on the selective transport of lipophilic phenethylamines. *Tetrahedron* 61:11965–11975
45. Saghiyan AS, Dadayan SA, Petrosyan SG, Manasyan LL, Geolchanyan AV, Djamgaryan SM, Andreyan SA, Maleev VI, Khrustalev VN (2006) New chiral NiII complexes of Schiff's bases of glycine and alanine for efficient asymmetric synthesis of α -amino acids. *Tetrahedron Asymmetr* 17:455–467
46. Alajarín M, López-Leonardo C, Berná J, Sánchez-Andrada P (2007) Center-to-propeller and propeller-to-propeller stereocontrol in a series of macrobicyclic tri- λ -5-phosphazenes. *Tetrahedron Lett* 48:3583–3586
47. Garmy N, Taïeb N, Yahi N, Fantini J (2005) Apical uptake and transepithelial transport of sphingosine monomers through intact human intestinal epithelial cells: physicochemical and molecular modeling studies. *Arch Biochem Biophys* 440:91–100
48. You Z, Omura S, Ikeda H, Cane DE (2007) Pentalenolactone biosynthesis: molecular cloning and assignment of biochemical function to PtlF, a short-chain dehydrogenase from *Streptomyces avermitilis*, and identification of a new biosynthetic intermediate. *Arch Biochem Biophys* 459:233–240
49. Bouffieux O, Berquand A, Eeman M, Paquot M, Dufrière YF, Brasseur R, Deleu M (2007) Molecular organization of surfactin-phospholipid monolayers: effect of phospholipid chain length and polar head. *Biochim Biophys Acta* 1768:1758–1768
50. Concu R, Dea-Ayuela MA, Perez-Montoto LG, Prado-Prado FJ, Uriarte E, Bolás-Fernández F, Podda G, Pazos A, Munteanu CR, Ubeira FM, González-Díaz H (2009) 3D Entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in *Leishmania* parasites. *Biochim Biophys Acta* 1794:1784–1794
51. Petraccone L, Martino L, Duro I, Oliviero G, Borbone N, Piccialli G, Giancola C (2007) Physico-chemical analysis of G-quadruplex containing bunch-oligonucleotides. *Int J Biol Macromol* 40:242–247
52. Noble JE, Wang L, Cole KD, Gaigalas AK (2005) The effect of overhanging nucleotides on fluorescence properties of hybridising oligonucleotides labelled with Alexa-488 and FAM fluorophores. *Biophys Chem* 113:255–263
53. Taraszewska J, Koźbiał M (2005) Complexation of Ketoconazole by native and modified cyclodextrins. *J Incl Phenom Macro* 53:155–161
54. Baker ES, Dupuis NF, Bowers MT (2009) Aminoglycoside antibiotics: A-site specific binding to 16S. *Int J Mass Spectrom* 283:105–111
55. Matsingou Ch, Dimas K, Demetzos C (2006) Design and development of liposomes incorporating a bioactive labdane-type diterpene. In vitro growth inhibiting and cytotoxic activity against human cancer cell lines. *Biomed Pharmacother* 60:191–199
56. Liu H, Du Y-M, Kennedy JF (2007) Hydration energy of the 1,4-bonds of chitosan and their breakdown by ultrasonic treatment. *Carbohydr Polym* 68:598–600
57. Banerjee A, Sengupta PK (2006) Encapsulation of 3-hydroxyflavone and fisetin in β -cyclodextrins: excited state proton transfer fluorescence and molecular mechanics studies. *Chem Phys Lett* 424:379–386
58. Singh BK, Jetley UK, Sharma RK, Garg BS (2007) Synthesis, characterization and biological activity of complexes of 2-hydroxy-3,5-dimethylacetophenoneoxime (HDMAOX) with copper(II), cobalt(II), nickel(II) and palladium(II). *Spectrochim Acta A* 68:63–73
59. Melnig V, Apostu MO, Tura V, Ciobanu C (2005) Optimization of polyurethane membranes. Morphology and structure studies. *J Membrane Sci* 267:58–67
60. TarushiA CP, Psomas G (2007) Synthesis, characterization and interaction with DNA of mononuclear metal complexes with oxolinic acid. *Polyhedron* 26:3963–3972

61. Setaki D, Tataridis D, Stamatou G, Kolocouris A, Foscolos GB, Fytas G, Kolocouris N, Padalko E, Neyts J, De Clercq E (2006) Synthesis, conformational characteristics and anti-influenza virus A activity of some 2-adamantylsubstituted azacycles. *Bioorg Chem* 34:248–273
62. Parks WM, Bottrill AR, Pierrat OA, Durrant MC, Maxwell A (2007) The action of the bacterial toxin, microcin B17, on DNA gyrase. *Biochimie* 89:500–507
63. Fatiha M, Khatmi DE, Largate L (2010) Theoretical approach in the study of the inclusion processes of sulconazole with β -cyclodextrin. *J Mol Liq* 154:1–5
64. Hutter MC (2006) Stability of the guanine-cytosine radical cation in DNA base pairs triplets. *Chem Phys* 326:240–245
65. Deconinck E, Ates H, Callebaut N, Van Gysegem E, Heyden YV (2007) Evaluation of chromatographic descriptors for the prediction of gastro-intestinal absorption of drugs. *J Chromatogr A* 1138:190–202
66. Furuşjö E, Svenson A, Rahmberg M, Andersson M (2006) The importance of outlier detection and training set selection for reliable environmental QSAR predictions. *Chemosphere* 63:99–108
67. Ezabadi IR, Camoutsis C, Zoumpoulakis P, Geronikaki A, Soković M, Glamocilija J, Ćirić A (2008) Sulfonamide-1,2,4-triazole derivatives as antifungal and antibacterial agents: synthesis, biological evaluation, lipophilicity, and conformational studies. *Bioorg Med Chem* 16:1150–1161
68. Kuş C, Ayhan-Kılıçgil G, Özbey S, Kaynak FB, Kaya M, Çoban T, Can-Eke B (2008) Synthesis and antioxidant properties of novel *N*-methyl-1,3,4-thiadiazol-2-amine and 4-methyl-2 H-1,2,4-triazole-3(4H)-thione derivatives of benzimidazole class. *Bioorg Med Chem* 16:4294–4303
69. Önkol T, Doğruer DS, Uzun L, Adak S, Özkan S, Şahin MF (2008) Synthesis and antimicrobial activity of new 1,2,4-triazole and 1,3,4-thiadiazole derivatives. *J Enzym Inhib Med Chem* 23:277–284
70. Shams HZ, Mohareb RM, Helal MH, Mahmoud AE (2007) Synthesis, structure elucidation, and biological evaluation of some fused and/or pendant thiophene, pyrazole, imidazole, thiazole, triazole, triazine, and coumarin systems based on cyanoacetic 2-[(benzoylamino)thioxomethyl]hydrazide. *Phosphorus Sulfur Silicon Relat Elem* 182:237–263
71. Siwek A, Stączek P, Wujec M, Stefańska J, Kosikowska U, Malm A, Jankowski S, Paneth P (2011) Biological and docking studies of topoisomerase IV inhibition by thiosemicarbazides. *J Mol Model* doi:10.1007/s00894-010-0889-z
72. HyperCube Inc. (2007) HyperChem 8.0.3. HyperCube Inc., Gainsville
73. Cornell WD, Cieplak P, Bayly ChI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PAJ (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* 117:5179–5197
74. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, 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
75. Jorgensen WL, Tirado-Rives J (1988) The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc* 110:1657–1666
76. Frisch MJ et al. (2009) Gaussian 09, revision A.02. Gaussian Inc., Wallingford
77. Zhao Y, Schultz NE, Truhlar DG (2006) Design of density functionals by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions. *J Chem Theor Comput* 2:364–382
78. Zhao Y, Truhlar DG (2008) Density functionals with broad applicability in chemistry. *Acc Chem Res* 41:157–167
79. Ditchfield R, Hehre WJ, Pople J (1971) Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules. *J Chem Phys* 54:724–728
80. Clark T, Chandrasekhar J, Spitznagel GW, Pvr S (1983) Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li–F. *J Comput Chem* 4:294–301
81. Frisch MJ, Pople JA, Binkley JS (1984) Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets. *J Chem Phys* 80:3265–3269
82. Marenich AV, Cramer CJ, Truhlar DG (2009) Universal solvation model based on solute electron density and a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J Phys Chem B* 113:6378–6396
83. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. *J Chem Phys* 98:5648–5652
84. Burke K, Perdew JP, Wang Y (1998) In: Dobson JF, Vignale G, Das MP (eds) *Electronic density functional theory: recent progress and new directions*. Plenum, New York (eds)
85. McLean AD, Chandler GS (1980) Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11–18. *J Chem Phys* 72:5639–5648
86. Cancès MT, Mennucci B, Tomasi J (1997) A new integral equation formalism for the polarizable continuum model: theoretical background and applications to isotropic and anisotropic dielectrics. *J Chem Phys* 107:3032–3041
87. Cossi M, Barone V, Mennucci B, Tomasi J (1998) Ab initio study of ionic solutions by a polarizable continuum dielectric model. *Chem Phys Lett* 286:253–260
88. Mennucci B, Tomasi J (1997) Continuum solvation models: a new approach to the problem of solute's charge distribution and cavity boundaries. *J Chem Phys* 106:5151–5158
89. Siwek A, Paneth P (2007) Computational studies of the cyclization of thiosemicarbazides. *J Phys Org Chem* 20:463–468
90. Lodola A, Sirirak J, Fey N, Rivara S, Mor M, Mulholland AJ (2010) Structural fluctuations in enzyme-catalyzed reactions: determinants of reactivity in fatty acid amide hydrolase from multivariate statistical analysis of quantum mechanics/molecular mechanics paths. *J Chem Theor Comput* 6:2948–2960
91. Schrödinger, LLC (2009) Impact, version 56107. Schrödinger, LLC, New York
92. Aliev AE, Courtier-Murias D (2010) Experimental verification of force fields for molecular dynamics simulations using Gly-Pro-Gly-Gly. *J Phys Chem B* 114:12358–12375

Density functional theory studies on the inclusion complexes of cyclic decapeptide with 1-phenyl-1-propanol enantiomers

Hongge Zhao · Yanyan Zhu · Mingqiong Tong ·
Juan He · Chunmei Liu · Mingsheng Tang

Received: 8 October 2010 / Accepted: 5 May 2011 / Published online: 28 May 2011
© Springer-Verlag 2011

Abstract Cyclic peptides are exciting novel hosts for chiral and molecular recognition. In this work, the inclusion complexes of cyclic decapeptide (**CDP**) with the 1-phenyl-1-propanol enantiomers (**E-PP**) are firstly studied using the density functional theory (DFT) B3LYP method. Our calculated results indicated that *S*(-)-1-phenyl-1-propanol (**S-PP**) could form a more stable inclusion complex with **CDP** than that of *R*(+)-1-phenyl-1-propanol (**R-PP**). The obvious differences in binding energy and thermodynamics data suggest that the cyclic decapeptide could differentiate the two enantiomers. Furthermore, molecular dynamics simulation results have supported the conclusions obtained by DFT. The current investigation shows that cyclic peptide is a desirable host molecule for chiral and molecular recognition.

Keywords Chiral recognition · Cyclic peptide · 1-phenyl-1-propanol · Inclusion complex

Introduction

Inclusion complex is the focus of current host-guest chemistry and supramolecular chemistry [1–5]. Experimental [6–8] and theoretical [9–13] investigations on this topic have been actively pursued for decades. Particularly,

studies on searching the desired host molecules dominate the scene. Many examples of host molecules, such as cyclodextrins (CDs) [14–16], macrocyclic antibiotics [17, 18], proteins [19] and chiral micelles [20] are available now. The representative host molecule cyclodextrins (CDs) have received much attention because they can separate many enantiomers by forming inclusion complexes with specific guest molecules [21, 22], this characteristic has been successfully applied to many fields including solubility enhancement, drug delivery, chemical protection, separation technology, and supramolecular chemistry [23, 24]. Another reason of the popularity of CDs is that the high symmetry and rigidity of their structures facilitate the study of inclusion complexes by NMR techniques [25]. However, this lack of conformational flexibility is a limitation regarding efficiency of inclusion complex. It's difficult for the CDs molecules to adjust their geometries to fit the guest molecules in an optimal interaction mode. Notably, these conformational disadvantages of CDs are just good qualities for cyclic peptides which are polypeptide linked by amino acid residues. In recent years, cyclic peptides have been synthesized and used as anticancer, antimalarial, antibacterial drug carriers and enzyme inhibitors, where they act as host molecules to form inclusion complexes with biological molecules [26–29].

Understanding the structural details of the inclusion complexes of cyclic peptides with guest molecules may help us delineate the features that are responsible for the remarkable potency of cyclic peptides. However, knowledge of the precise interaction mechanism of cyclic peptides with enantiomers of a chiral molecule at the molecular level is still very limited [30]. Especially, conformations and structures of cyclic peptides are not yet clear experimentally. Some theoretical studies on the

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1119-z) contains supplementary material, which is available to authorized users.

H. Zhao · Y. Zhu (✉) · M. Tong · J. He · C. Liu · M. Tang (✉)
Department of Chemistry, Zhengzhou University,
Zhengzhou, Henan Province 450001, People's Republic of China
e-mail: zhuyan@zzu.edu.cn

M. Tang
e-mail: mstang@zzu.edu.cn

structures of cyclic peptides have been performed over the past several years [31–34]. Guangju Chen and co-workers have reported the structural characteristics of an important type of cyclopeptides formed by cyclo[(- β^3 -HGly) $_4$] based on density function theory (B3LYP) [35]. Their results provide us with new insights into the formation of polypeptide. Inspired by the study of Guangju Chen and co-workers on cyclopeptide, we have performed a density functional theory (DFT) study of the interactions between cyclic peptides and enantiomers, which may have much theoretical and practical importance. In the present work, we focused on the structure of model cyclic peptide derived from glycine. The glycine, as the simplest amino acid, was used to create a cyclopeptide template. Cyclic decapeptide (assigned as **CDP**) (shown in Chart 1a), constructed with ten identical glycines, is used as a receptor that is capable of including trapping the guest molecules inside the peptide cavity possibly caused by the conformational flexibility and noncovalent interactions.

1-phenyl-1-propanol, a chiral molecule existing in a couple of enantiomers forms (assigned as **E-PP**, **E=R** or **S**, shown in Chart 1b), is a good candidate for constructing a simple model to study chiral discrimination. The separation of the enantiomers of 1-phenyl-1-propanol has already been carried out in the experiment [36]. However, the separation result is not desirable. In this work, **CDP** and **E-PP** are selected as host molecule and guest molecules, respectively, to investigate the conformational and structural features of **CDP/E-PP** and the interaction of **CDP** with **E-PP**.

Computational methods

The search of the energy minimum

In this paper, the selected initial structure of **CDP** is E-type backbone due to the E-type backbone of cyclopeptide is

more stable than that of B-type when the number of amino acid residues is equal to or bigger than 10 [37].

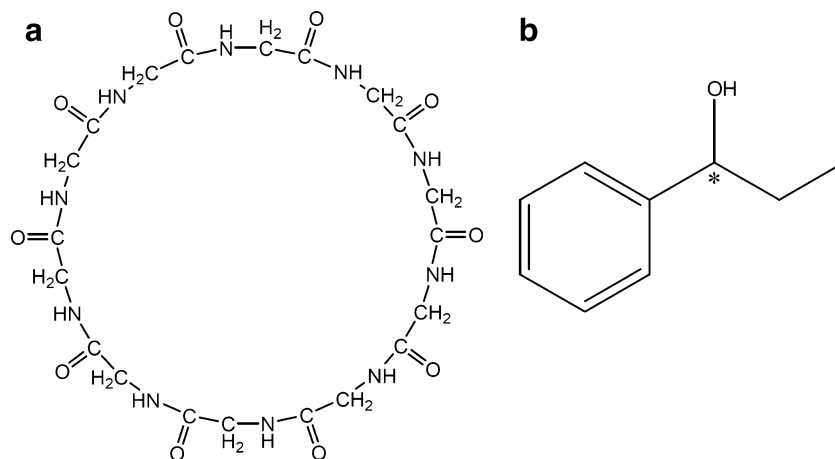
The coordinate system used to define the inclusion process of **CDP** with **E-PP** is shown in Chart 2, which was adopted from a previous work [38]. Briefly, the **CDP** ring was positioned symmetrically around the Z-axis, such that all oxygen and nitrogen atoms in the glycine are in the XY plane. The **E-PP** molecule was docked into the cavity of **CDP** along with the Z-axis. Multiple initial positions were generated by movement of **E-PP** along the Z-axis. The relative position between **CDP** and **E-PP** was measured by the Z-coordinate of the labeled carbon atom of **E-PP** (shown in Chart 2). In order to find a more stable structure of **CDP/E-PP**, we calculated all of the structures of each E-PP molecule by scanning θ , circling around the Z-axis, at 20° intervals from -180° to 180° and scanning the Z-coordinate at 0.3 Å intervals with semi-empirical calculations (PM3), which can be currently applied in biochemical systems with its improved description of the interactions between non-bonded atoms, e.g., hydrogen bond and steric effects [39]. All of local energy minimum structures from potential energy surface (PES) by scan calculations were fully optimized at the B3LYP/3-21G level of theory [40, 41]. Subsequently, **CDP/E-PP** with the lowest energy obtained by B3LYP/3-21G calculations were fully optimized using the basis set of 6-31+G(d,p). Additionally, the frequency calculations for **CDP/E-PP** were also carried out to verify the optimized structures to be energy minima without any imaginary frequency.

Definition of the binding energy (BE)

In order to investigate the driving forces leading to **CDP/E-PP** between **CDP** and **E-PP**, the binding energy (BE) upon **CDP/E-PP** for the minimum energy structure is evaluated from the following equation.

$$BE = E[\text{CDP/E-PP}] - E[\text{E-PP}] - E[\text{CDP}] \quad (1)$$

Chart 1 Schematic representation for the conformations of (a) cyclic decapeptide (**CDP**) and (b) 1-phenyl-1-propanol enantiomers (**E-PP**)



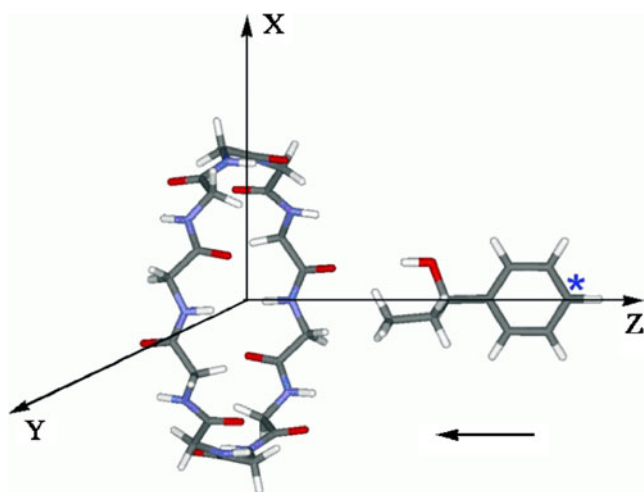


Chart 2 Coordinate systems used to define the inclusion process (H atoms included in **CDP** are omitted)

Where, $E[\text{CDP/E-PP}]$, $E[\text{E-PP}]$ and $E[\text{CDP}]$ represent the energies of **CDP/E-PP**, **E-PP** and **CDP**, respectively. The magnitude of BE would be a sign of the driving force toward **CDP/E-PP**. A negative value of BE means that the corresponding **CDP/E-PP** is energetically stable; the more negative the BE is, the more stable the complex is.

The deformation energies of **CDP** and **E-PP** were calculated by Eq. 2 and 3 [42].

$$\text{DE}[\text{E-PP}] = E[\text{E-PP}]_{\text{sp}}^{\text{opt}} - E[\text{E-PP}]_{\text{opt}} \quad (2)$$

$$\text{DE}[\text{CDP}] = E[\text{CDP}]_{\text{sp}}^{\text{opt}} - E[\text{CDP}]_{\text{opt}}, \quad (3)$$

where $\text{DE}[\text{E-PP}]$ and $\text{DE}[\text{CDP}]$ are the deformation energies of **E-PP** and **CDP** respectively; $E[\text{E-PP}]_{\text{sp}}^{\text{opt}}$ and $E[\text{E-PP}]_{\text{opt}}$ are the single point energy of **E-PP** on the configuration taken from the optimized **CDP/E-PP** and the energy of the optimized geometry of **E-PP** respectively; $E[\text{CDP}]_{\text{sp}}^{\text{opt}}$ and $E[\text{CDP}]_{\text{opt}}$ are the single point energy of **CDP** on the configuration taken from the optimized **CDP/E-PP** and the energy of the optimized geometry of **CDP** respectively.

Thermodynamic analysis for the inclusion process of **CDP** with **E-PP**

The geometries of the two inclusion complexes were fully optimized without any geometrical or symmetry constraints using the B3LYP/6-31+G(d,p) method. The frequencies were performed for the evaluation of the enthalpy changes (ΔH) and Gibbs free energy changes (ΔG) of the inclusion process between **CDP** and **E-PP**.

Moreover, the electronic properties of **CDP/E-PP** were studied using the natural bond orbital (NBO) analysis at

the B3LYP/6-31+G(d,p) level of theory [43]. NBO calculations quantify the H-bond interactions between host and guest molecules via the determination of the stabilization energy $E^{(2)}$. The stabilization energy $E^{(2)}$ related to the delocalization trend of electrons from donor to acceptor orbital is calculated via perturbation theory. A large stabilization energy $E^{(2)}$ between a lone pair LP(Y) of an atom Y and an antibonding σ^* (X—H) orbital is generally indicative of a strong X—H...Y hydrogen bond [44]. Basis set superposition error (BSSE) of binding energies is calculated by using the counterpoise corrections method [45]. All calculations were carried out using the GAUSSIAN 03 program package [46].

Results and discussion

Most stable conformation and binding energy

Two obtained PESs are shown in Fig. 1. It can be seen that the inclusion processes of **CDP** with **E-PP** are energetically favorable. Interestingly, most of energy minima structures locate at approximately $Z=0$ Å for **E-PP** approaches. Based on the related scanned energy minima at the level of PM3, B3LYP/3-21G calculations were performed to optimize **CDP/E-PP** as presented in Fig. 2. Other possible locations and angles of **E-PP** were examined using the B3LYP method, which were shown to be energetically less favorable and therefore not listed. Based on the B3LYP/3-21G optimized equilibrium geometries of the **CDP/E-PP**, calculations at the B3LYP/6-31+G(d,p) level were then performed.

The BE values including BSSE corrections for most stable inclusion configurations are listed in Table 1. The BE values for **CDP/S-PP** and **CDP/R-PP** are -19.94 and -10.54 kJ mol^{-1} , respectively, which demonstrate that **CDP** can form stable complexes with **E-PP**. The **CDP/S-PP** was more favorable than **CDP/R-PP** by an energy difference of 9.40 kJ mol^{-1} , suggesting that **S-PP** is bound more firmly by **CDP**.

Energies of the inclusion complexes

To investigate the thermodynamics of the inclusion process, the statistical thermodynamic calculations were performed at the B3LYP/6-31+G(d,p) level of theory. The calculated results are listed in Table 1. It is obvious that the inclusion process of **CDP** with **E-PP** are exothermic judged from the negative enthalpy changes. The negative enthalpy changes also suggest that both the inclusion processes are enthalpically favorable. On the other hand, the enthalpy change of **CDP/S-PP** (-20.93 kJ mol^{-1}) is about 8.30 kJ mol^{-1} lower than that of

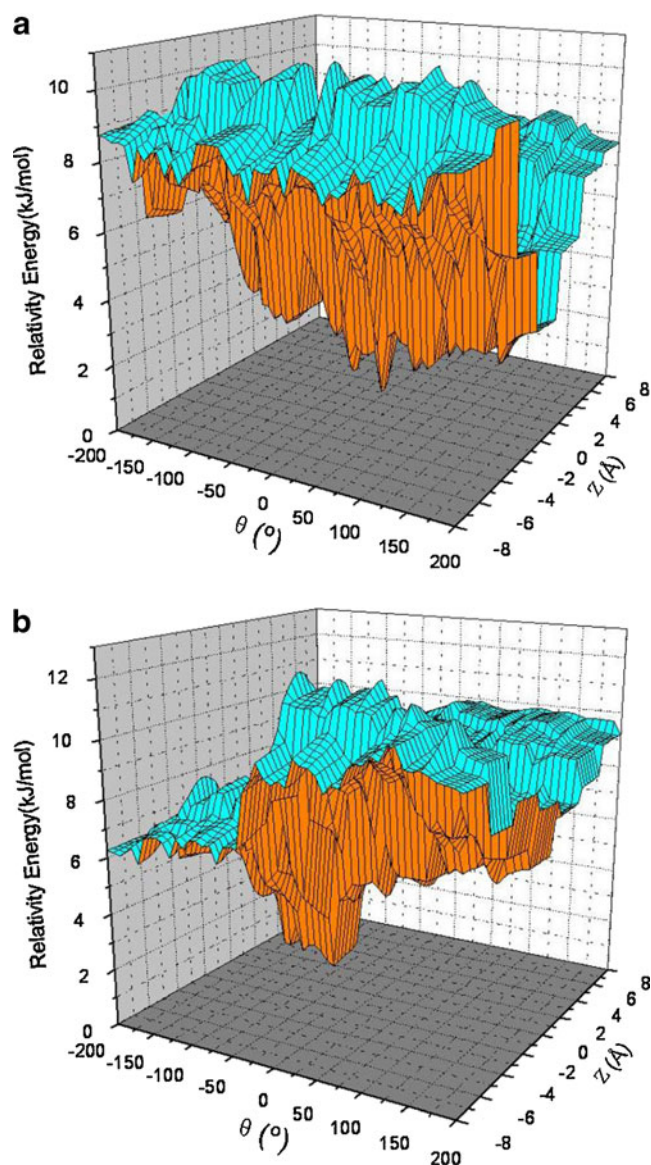


Fig. 1 Scan of total energy of the inclusion complex of the **E-PP** enantiomers into **CDP** at different positions (z) and orientations (θ): (a) **S(-)-1-phenyl-1-propanol (S-PP)** and **CDP**; (b) **R(+)-1-phenyl-1-propanol (R-PP)** and **CDP**. The position of the **E-PP** molecule was determined by the Z -coordinate of the labeled carbon atom (*) in the phenyl group. θ refers to the angle of each guest molecule circling around the Z -axis of the system

CDP/R-PP ($-12.63 \text{ kJ mol}^{-1}$). The thermodynamic results indicate that the **S-PP** structure is preferred to form inclusion complex with **CDP** based on enthalpy grounds.

One interesting feature of the guest is its conformational flexibility. A better guest conformational flexibility is favorable to the host–guest interactions, it makes it possible for the guest molecule to modify its conformation to ensure a better penetration [47]. Investigation of the deformation energy of the chosen guest **E-PP** at the B3LYP/6-31+G(d,p) level of theory (as shown in Table 1)

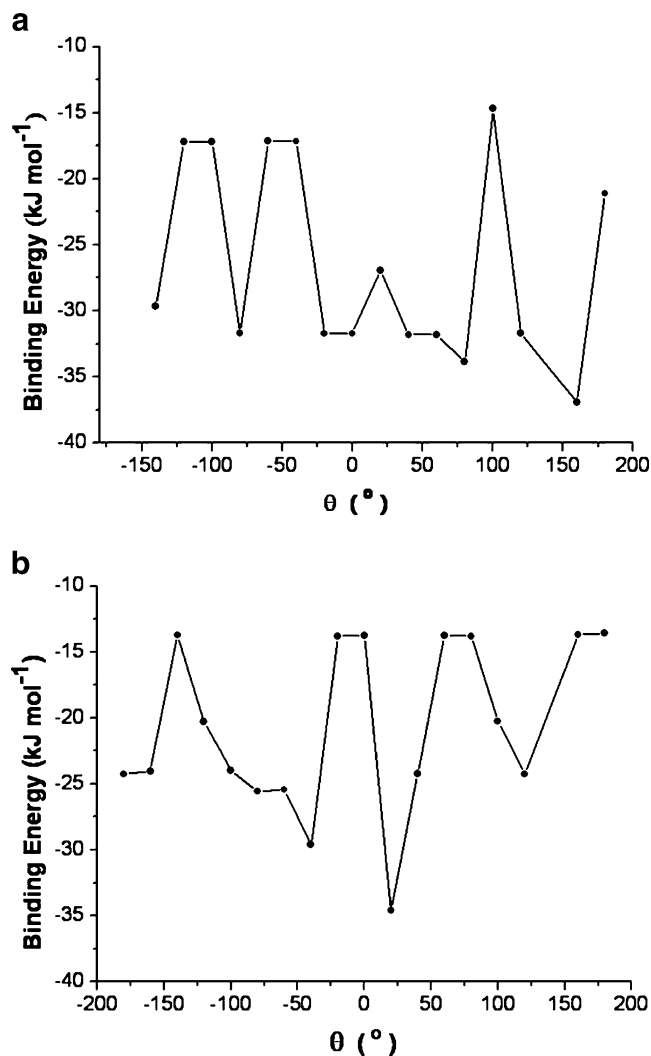


Fig. 2 B3LYP/3-21G stabilization energy including BSSE correction of the **CDP/E-PP**: (a) **S-PP** and **CDP**; (b) **R-PP** and **CDP**. θ refers the start angle of each guest molecule into **CDP** circling around the Z -axis of the system

demonstrated that the deformation of **S-PP** requires slightly more energy to adapt conformation to fit the cavity of **CDP** than that of **R-PP** as indicated by the **DE[E-PP]** data of about 2.29 and 0.93 kJ mol^{-1} respectively. On the other hand, there are some distortion of **CDP** in the inclusion process as well. **CDP** needs 3.68 kJ mol^{-1} to adapt conformational adaptation for **CDP/S-PP** and 1.48 kJ mol^{-1} for **CDP/R-PP**, indicating that the deformation of **CDP** is advantageous for the inclusion complex formation.

Conformational characteristics of **CDP/E-PP**

The favorable structures of **CDP/E-PP** optimized at the B3LYP/6-31+G(d,p) level are graphically presented in Fig. 3. Figure 3a shows that for **CDP/S-PP**, the phenyl of

Table 1 The binding energies and thermodynamic parameters upon the inclusion complexes of **CDP/S-PP** and **CDP/R-PP** at the B3LYP/6-31+G(d,p) level of theory

Parameter	CDP/S-PP	CDP/R-PP
BE ^a (kJ mol ⁻¹)	-27.67	-18.79
BSSE(kJ mol ⁻¹)	7.73	8.25
BE ^b (kJ mol ⁻¹)	-19.94	-10.54
DE[E-PP] ^c (kJ mol ⁻¹)	2.29	0.93
DE[CDP] ^d (kJ mol ⁻¹)	3.68	1.48
ΔH° (kJ mol ⁻¹)	-20.93	-12.63
ΔG° (kJ mol ⁻¹)	28.47	37.14
ΔS° (J mol ⁻¹ K ⁻¹)	-165.67	-166.95
ΔH_{pcm}^e (kJ mol ⁻¹)	-6.52	5.37
ΔG_{pcm}^f (kJ mol ⁻¹)	-38.34	-29.42
ΔS_{pcm}^g (J mol ⁻¹ K ⁻¹)	106.72	116.69

^a BE is the binding energy upon complex^b BE is the binding energy including the basis set superposition error (BSSE) correction^c **DE[E-PP]** is the deformation energy of **E-PP**^d **DE[CDP]** is the deformation energy of **CDP**^e ΔH_{pcm} is the enthalpy change obtained by PCM model^f ΔG_{pcm} is the Gibbs free energy change obtained by PCM model^g ΔS_{pcm} is the entropy change obtained by PCM model

S-PP is almost totally encapsulated in the cyclic decapeptide cavity. While the OH group remains on the rim of the **CDP**, which is in favor of formation of H-bond with some groups of **CDP**. The optimized geometries reveal that there are two hydrogen bond interactions between **CDP** and **S-PP**. Figure 3b shows clearly that for the **CDP/R-PP**, the phenyl of **R-PP** are partially included in **CDP** and the orientation of OH group directed toward the inside of the **CDP** cavity, allowing the lone pair of the oxygen atom to act as an H-bond acceptor thus stabilizing the **CDP/R-PP**.

Hydrogen bond analysis and NBO analysis

To investigate the reason why geometries of **CDP/E-PP** are different, the hydrogen bond and NBO analyses are further performed at the B3LYP/6-31+G(d,p) level of theory. The detailed information of intermolecular hydrogen bond interactions for **CDP/E-PP** are listed in Table 2 and shown in Fig. 4. It can be seen from Table 2 that the distinct differences for hydrogen bond interactions occur in the different inclusion complexes. In the **CDP/S-PP** structure, there are two hydrogen bond interactions. One occurs between O1 of **CDP** and O11 of **S-PP**, a strong hydrogen bond O11 – H11...O1 ($d_{\text{H}\dots\text{O}} = 1.93\text{\AA}$). The other occurs between the O11 of **S-PP** and C10 of **CDP**, a weak

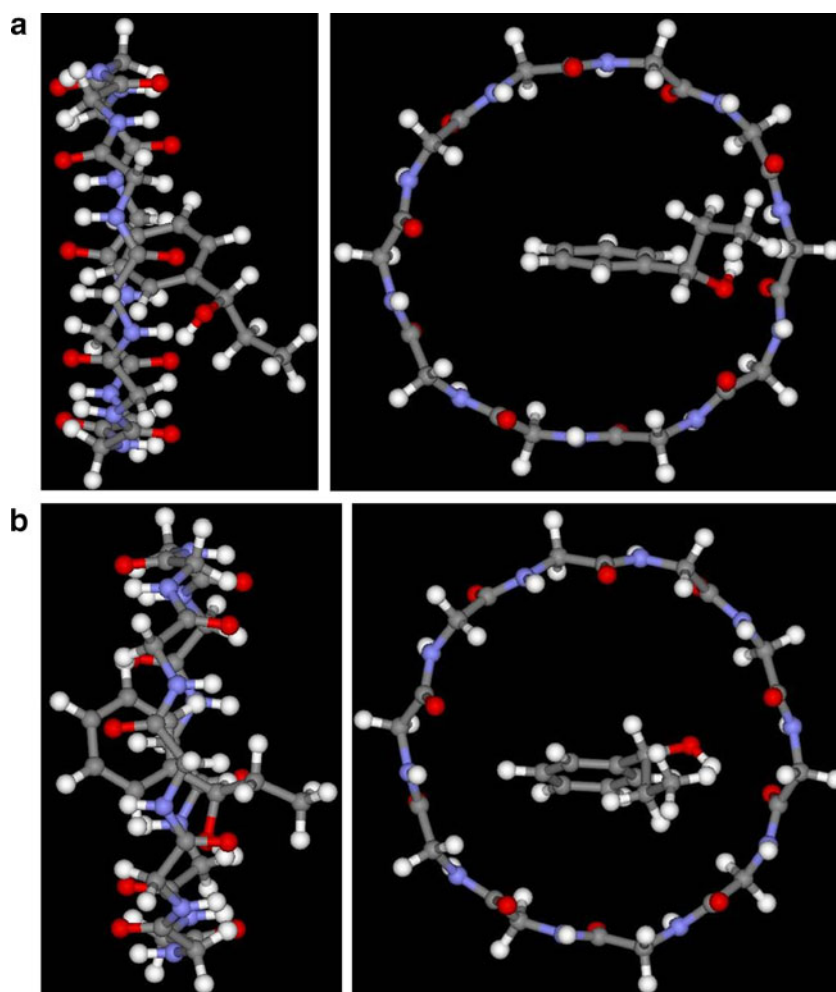
hydrogen bond C10 – H10...O11 ($d_{\text{H}\dots\text{O}} = 2.48\text{\AA}$). In the **CDP/R-PP** structure, only one weak hydrogen bond is formed. Namely, the O5 atom of **CDP** donates a hydrogen bond ($d_{\text{H}\dots\text{O}} = 2.22\text{\AA}$) to O11 of **R-PP**. Distinctly, the intermolecular hydrogen bonds play a crucial role in the stability of inclusion complexes conformational change. It was suggested that the contribution of the O – H...O hydrogen bond interactions to the structural stability in **CDP/S-PP** is greater than those in **CDP/R-PP**. This explains why the BE for the **CDP/S-PP** is 9.40 kJ mol⁻¹ lower than that of **CDP/R-PP**.

The following NBO analyses confirm the occurrence of these intermolecular hydrogen bonds. The stabilization energies $E^{(2)}$ calculated at the B3LYP/6-31+G(d,p) level of the established H-bond in the **CDP/E-PP** are listed in Table 2. Significant interaction energies are obtained for the expected hydrogen bonds, especially for the O – H...O one. The interaction energy of the O – H...O hydrogen bond of **CDP/S-PP** is 23.24 kJ mol⁻¹, which is a conventional hydrogen bond (16–25 kJ mol⁻¹ for O – H...O hydrogen bonds in carbohydrates) [48]. The interaction energy of the O – H...O hydrogen bond of the **CDP/R-PP** is 3.61 kJ mol⁻¹, which belongs to a typical weak hydrogen bond for which energies vary between 2.1 and 8.4 kJ mol⁻¹ [49]. Noteworthy, one extra C – H...O hydrogen bond was observed for **CDP/S-PP**. Quantum mechanical calculations have been performed to determine the energetic of the C – H...O bonds in the complexes, which are far below values of conventional hydrogen bonding [50, 51], but appreciably above energies of van der Waals contacts. Briefly, these hydrogen bond interactions play important roles in the inclusion processes of **CDP** with **E-PP**.

Molecular dynamics simulations

Regarding the identification of the preferred inclusion modes, it would be more realistic to select a set of inclusion complex structures, besides a single optimized configuration. Inclusion phenomena are dynamic in nature; therefore the establishment of host-guest intermolecular interactions cannot be analyzed from a single structure [52]. Perhaps, other unexplored inclusion complexes can lead to different hydrogen bonding patterns. Molecular dynamics (MD) simulations could provide such a view [53]. To obtain the possible inclusion modes between **CDP** and **E-PP**, the two guests, **R-PP** and **S-PP**, were firstly docked into **CDP** by using AutoDock 4.0 program [54]. The grid map of 32×32×32 points and a grid-point spacing of 0.375 Å have been employed during the dock processes. One better-scoring representative from 1000 predication inclusion models for **CDP** with **E-PP** has been selected as an initial structure for MD simulations.

Fig. 3 Energy-minimized structure obtained by B3LYP/6-31+G(d,p) calculations for **CDP/E-PP**: (a) side view (left) and top view (right) for **CDP/S-PP** and (b) side view (left) and top view (right) for **CDP/R-PP**



All MD simulations were carried out using the AMBER9 [55] package with the AMBER force fields of parm99 [56, 57] and gaff [58]. The systems were explicitly solvated by using the TIP3P water potential inside a box large enough to ensure the solvent shell extended to 10 Å in all directions of each system studied. For the equilibration of the investigated systems, the following procedures were carried out. First, 22500 steps energy minimization were carried out to remove unfavorable contacts. Then the systems were heated over 100 ps from 0 to 300 K with a little restrains of 10 kcal mol⁻¹ Å⁻². The equilibration time for each simulation was 500 ps (NPT) followed by 10 ns of data collection for trajectory analysis, that is, 5000

structures for each simulation were saved for further data analysis by uniformly sampling the trajectory.

With the help of a 10 ns long molecular dynamics, it is shown that the **CDP/E-PP** are stable in water environment. The detailed information of intermolecular hydrogen bonds interactions for the **CDP/E-PP** in the course of the simulations are listed in Table 3. The quantities and lifetimes of H-bonds reflect the ability of **CDP** to bind **E-PP**, respectively. These observations clearly show that the lifetimes and number of H-bonds for **CDP/S-PP** are longer and larger than those of **CDP/R-PP**. Specifically, the longest lifetime of H-bond for **CDP/S-PP** is up to 99.58% of the simulation times, while the longest

Table 2 The electron donors, electron acceptors and the corresponding $E^{(2)}$ energies, distances and angles obtained at the B3LYP/6-31+G(d,p) level of theory

	Donor	Acceptor	H...A (Å)	D...A (Å)	D-H...A (°)	$E^{(2)}$ (kJ mol ⁻¹)
CDP/S-PP	LP O1	BD*O11-H11	1.93	2.87	160.36	23.24
	LP O11	BD*C10-H10	2.48	3.40	140.94	8.41
CDP/R-PP	LP O5	BD*O11-H11	2.22	2.93	129.74	3.61

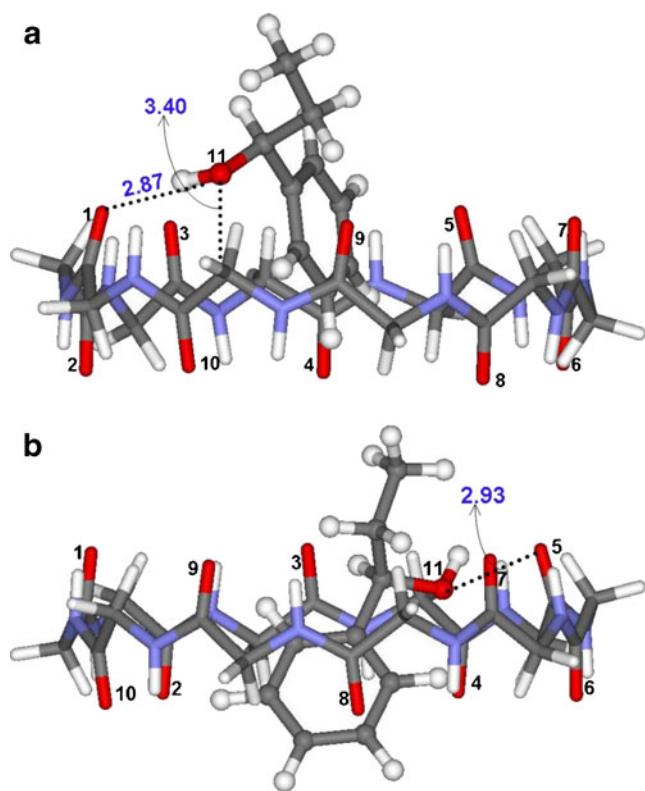


Fig. 4 Host–guest hydrogen bonds are presented by dotted lines: (a) **CDP/S-PP** and (b) **CDP/R-PP**. H in white, C in gray, N in blue, O in red

lifetime of H-bond for **CDP/R-PP** is only 44.06% of the simulation times. Compared to **R-PP**, **S-PP** exhibits the anticipated binding propensity to associate with **CDP**. The MD results indicate that the **CDP/S-PP** system is appreciably more stable than **CDP/R-PP**. Briefly, the MD results support the conclusions obtained by B3LYP/6-31+G(d,p).

Solvent effects

Complex phenomena take place in condensed phase. Thus, solvent plays a critical role in the intermolecular interactions that lead to the formation of inclusion complexes, especially in the case of polar compounds with hydrogen donor/acceptor groups. To consider the role of the solvent, the polarized continuum model (PCM) [59–61] has been

Table 3 The intermolecular hydrogen bonds of **CDP/S-PP** and **CDP/R-PP** during molecular dynamics (MD) simulations

Complex	Donor	Acceptor	Lifetime (%)	Distance (Å)
CDP/S-PP	O1	O11-H11	99.58	2.82
	O11	C10-H10	95.82	3.65
CDP/R-PP	O5	O11-H11	44.06	2.99

employed to simulate the solvent effects as implemented within the solvent reaction field based on the optimized structures as listed in Table 1. As shown in Table 1, positive entropy changes (ΔS_{pcm}) in the two inclusion processes are 106.72 and 116.69 J mol⁻¹ K⁻¹ respectively, which are attributed to the releasing of water molecules in the cavity of **CDP**. Based on the discussions above, it can be concluded that entropy effects on the stability of the **CDP/E-PP** are favorable factors, that is, the formations of **CDP/E-PP** are entropy driven processes in aqueous solution.

Conclusions

In this work, **CDP/E-PP** have been investigated theoretically using the density functional theory (DFT) B3LYP method. Almost all possible locations of **E-PP** with **CDP** were taken into account to obtain the most stable conformation of **CDP/E-PP**. The optimized structures and the binding energy (BE) indicate that **CDP/S-PP** is more stable than **CDP/R-PP**. The conformational characteristics of **CDP/E-PP** show that the distinct differences for hydrogen bond interactions occur in the different **CDP/E-PP**. For **CDP/S-PP**, the better stabilization may be attributed to the formation of two hydrogen bonds between **CDP** and **E-PP**. For **CDP/R-PP**, only one hydrogen bond has been formed between **CDP** and **E-PP**, which might account for the stabilization of **CDP/E-PP**. The NBO analyses confirm the occurrence of these intermolecular hydrogen bonds: the NBO results show that there is one conventional hydrogen bond and one weak hydrogen bond in the **CDP/S-PP** inclusion complex while there is only one weak hydrogen bond in the **CDP/R-PP** inclusion complex. Briefly, these hydrogen bond interactions will contribute to the overall stability and structure of the inclusion complexes of **CDP** with **E-PP**. Furthermore, the MD simulation results are in agreement with the conclusions obtained by the B3LYP/6-31+G(d,p) method.

Additionally, the thermodynamic calculated results demonstrated that enthalpy changes (ΔH) are prominent in the inclusion processes. The enthalpy changes suggest that the formation of **CDP/E-PP** is an enthalpy driven process. Their obvious differences in binding energy and enthalpy change suggest that **CDP** could well distinguish **E-PP**. Take the solution effects into account, the entropy is still a favorable driving force for the formation of **CDP/E-PP**. The current studies provide a revealing insight into conformational characteristics and thermodynamics properties for **CDP/E-PP** at the molecular level. The observations in this work indicate that **CDP** is a desirable host molecule for chiral and molecular recognition.

Acknowledgments The work described in this paper was supported by the National Natural Science Foundation of China (No. 21001095) and China Postdoctoral Science Foundation (No. 20100480858).

References

- Gellman SH (1997) *Chem Rev* 97:1231–1232
- Breslow R, Dong SD (1998) *Chem Rev* 98:1997–2012
- Lee WY, Park CH, Kim S (1993) *J Am Chem Soc* 115:1184–1185
- Song LX, Wang HM, Yang Y (2007) *Acta Chim Sinica* 65:1593–1599
- De Sousa FB, Denadai AML, Lula IS, Lopes JF, Dos Santos HF, De Almeida WB, Sinisterra RD (2008) *Int J Pharm* 353:160–169
- Khedkar JK, Gobre W, Pinjari RV, Gejji SP (2010) *J Phys Chem A* 114:7725–7732
- Maheshwari A, Sharma D (2010) *J Incl Phenom Macro* 68:453–459
- Jug M, Mennini N, Melani F, Maestrelli F, Mura P (2010) *Chem Phys Lett* 500:347–354
- Wen XH, Liu ZY, Zhu TQ (2005) *Chem Phys Lett* 405:114–117
- Zoppi A, Quevedo MA, Delrivo A, Longhi MR (2010) *J Pharm Sci* 99:3166–3176
- Dos Santos HF, Duarte HA, Sinisterra RD, De Melo Mattos SV, De Oliveira LFC, De Almeida WB (2000) *Chem Phys Lett* 319:569–575
- Snor W, Liedl E, Weiss Greiler P, Virmstein H, Wolschann P (2009) *Int J Pharm* 381:146–152
- Barbiric DJ, Castro EA, de Rossi RH (2000) *J Mol Struct THEOCHEM* 532:171–181
- Seridi L, Boufelfel A (2011) *J Mol Liq* 158:151–158
- Chankvetadze B (1997) *J Chromatogr A* 792:269–295
- Fanali S (2000) *J Chromatogr A* 875:89–122
- Armstrong DW, Nair UB (1997) *Electrophoresis* 18:2331–2342
- Ward TJ, Oswald TM (1997) *J Chromatogr A* 792:309–325
- Haginaka J (2000) *J Chromatogr A* 875:235–254
- Otsuka K, Terabe S (2000) *J Chromatogr A* 875:163–178
- Castillo N, Boyd RJ (2005) *Chem Phys Lett* 416:70–74
- Kim H, Jeong K, Lee S, Jung S (2002) *J Comput Aided Mol Des* 16:601–610
- Stella VJ, Rao VM, Zannou EA, Zia V (1999) *Adv Drug Deliv Rev* 36:3–16
- Schneiderman E, Stalcup AM (2000) *J Chromatogr B* 745:83–102
- Coleman AW (1998) Kluwer Academic Publishers, p 103
- Kobayashi J, Tsuda M, Nakamura T, Mikami Y, Shigemori H (1993) *Tetrahedron* 49:2391–2402
- Gulavita NK, Gunasekera SP, Pomponi SA, Robinson EV (1992) *J Org Chem* 57:1767–1772
- Ferrante F, La Manna G (2007) *J Comput Chem* 28:2085–2090
- Lewis JP, Pawley NH, Sankey OF (1997) *J Phys Chem B* 101:10576–10583
- Maier NM, Schefzick S, Lombardo GM, Feliz M, Rissanen K, Lindner W, Lipkowitz KB (2002) *J Am Chem Soc* 124:8611–8629
- Kim KS, Cui C, Cho SJ (1998) *J Phys Chem B* 102:461–463
- Zhu YY, Tang MS, Shi XY, Zhao YF (2007) *Int J Quantum Chem* 107:745–753
- Teranishi M, Okamoto H, Takeda K, Nomura K, Nakano A, Kalia RK, Vashishta P, Shimojo F (2009) *J Phys Chem B* 113:1473–1484
- Chen GJ, Su S, Liu RZ (2002) *J Phys Chem B* 106:1570–1575
- Tan HW, Qu WW, Chen GJ, Liu RZ (2003) *Chem Phys Lett* 369:556–562
- Khattabi S, Cherrak DE, Mihlbachler K, Guiochon G (2000) *J Chromatogr A* 893:307–319
- Okamoto H, Nakanishi T, Nagai Y, Kasahara M, Takeda K (2003) *J Am Chem Soc* 125:2756–2769
- Yan CL, Xiu ZL, Li XH, Hao C (2007) *J Mol Graph Model* 26:420–428
- Liu L, Guo QX (2004) *J Incl Phenom Macrocycl Chem* 50:95–103
- Becke AD (1993) *J Chem Phys* 98:5648–5652
- Lee C, Yang W, Parr RG (1988) *Phys Rev B* 37:785–789
- Ohashi M, Kasatani K, Shinohara H, Sato H (1990) *J Am Chem Soc* 112:5824–5830
- Glendening ED, Reed AE, Carpenter JE, Weinhold F, NBO Version 03.01, included in the GAUSSIAN 03 package of programs
- Zhu YY, Chen ZF, Guo ZJ, Wang Y, Chen GG (2009) *J Mol Model* 15:469–479
- van Duijneveldt FB, van Duijneveldt-van de Rijdt JGCM, van Lenthe JH (1994) *Chem Rev* 94:1873–1885
- Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven JT, 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, 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, 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
- Rekharsky MV, Inoue YI (1998) *Chem Rev* 98:1875–1917
- Starikov EB, Saenger W, Steiner Th (1998) *Carbohydr Res* 307:343–346
- Uccello Barretta G, Balzano F, Sicoli G, Paolino D, Guccione S (2004) *Bioorg Med Chem* 12:447–458
- Desiraju GR (1996) *Chem Res* 29:441–449
- Steiner T (1997) *Chem Commun* 727–734
- Yu YM, Christophe C, Cai WS, Shao XG (2006) *J Phys Chem B* 110:6372–6378
- Cai WS, Sun TT, Liu P, Christophe C, Shao XG (2009) *J Phys Chem B* 113:7836–7843
- Morris GM, Goodse DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) *J Comput Chem* 19:1639–1662
- Case DA, Darden TA, Cheatham TE, 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
- Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T (2003) *J Comput Chem* 24:1999–2012
- Lee MC, Duan Y (2004) *Proteins* 55:620–634
- Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) *J Comput Chem* 25:1157–1174
- Rehbein J, Hiersemann M (2009) *J Org Chem* 74:4336–4342
- Peles DN, Thoburn JD (2008) *J Org Chem* 73:3135–3144
- Takano Y, Houk KN (2005) *J Chem Theor Comput* 1:70–77

Ab initio simulation of the effect of the potential of water on the electronic structure of arginine

Xingrong Wang · Haoping Zheng

Received: 26 February 2011 / Accepted: 11 May 2011 / Published online: 28 May 2011
© Springer-Verlag 2011

Abstract As a part of the task of constructing the equivalent potential of water in order to obtain a reliable electronic structure for a protein, the equivalent potential of water for an arginine molecule was constructed by performing first-principles, all-electron, ab initio calculations. The process consisted of three steps. First, the electronic structure of arginine was calculated using a free cluster calculation. Then, the minimum-energy geometric structure of the system $\text{Arg}^+ + 9\text{H}_2\text{O}$ was found using free cluster calculations. Then, based on the optimized geometric structure of the $\text{Arg}^+ + 9\text{H}_2\text{O}$ system, the electronic structure of Arg^+ in the potential of water was calculated using the SCCE method. Finally, by performing SCCE calculations, the effect of water on the electronic structure of Arg^+ was simulated with dipoles. The results show that the effect of water on the electronic structure of Arg^+ is to broaden the energy gap tenfold, and to increase the eight eigenvalues below the HOMO by about 0.0546 Ry on average. The water potential can be accurately simulated using dipoles.

Keywords Arginine · Electronic structure · Water · Free cluster calculation · Self-consistent cluster-embedding calculation

Introduction

Researchers at the Human Genome Project have now finished mapping the body's 25,000 genes, but scientists

are still hard at work on an even greater task: researching the geometric structures and biological functions of proteins. Elucidating the electronic structures of proteins allows us to better understand some of the actions of proteins. It is, however, difficult to calculate the electronic structure of a protein, as it requires an incredible amount of computational effort. In the last two decades, the fields of computational condensed matter physics and quantum chemistry have both focused on developing so-called $O(N)$ methods, for which the computational effort scales linearly with the number (N) of atoms [1–17]. Self-consistent cluster-embedding calculation (SCCE) [16, 17] is an $O(N)$, first-principles, all-electron, ab initio calculation method that is based on density functional theory. Unlike traditional calculation methods, the one-electron wavefunctions obtained with the SCCE method are localized; i. e., each one-electron wavefunction localizes only in a part of the region occupied by the system. The advantage of the SCCE method is that the localized valence electrons of the material can be described well, and the computational effort can be greatly reduced while maintaining the calculation precision. SCCE calculations have been successfully applied to several insulators, semiconductors, metals, crystals with defects and impurities, and surfaces [17–23]. SCCE calculations were first applied to calculate the electronic structure of a protein in 2000 [24]. So far, the electronic structures of three proteins with four three-dimensional structures have been obtained with SCCE [25–27]. Another protein was calculated by both Sato et al. [28] and Yoshihiro et al. [29] using their own methods.

Our previous protein calculations did not take the influence of solvent into account due to limited computational capacity. In general, a protein in water as a solvent has a biological function, and its geometric structure is different from that in the isolated state. In other words, almost no biological processes can occur without the presence of solvent. Thus, it is necessary to take the

X. Wang · H. Zheng (✉)
Physics Department, Tongji University,
Shanghai 200092, China
e-mail: zhenghp@tongji.edu.cn

X. Wang
e-mail: enjoymaself@126.com

influence of solvent into account in electronic structure calculations of proteins.

Several continuous medium models for water have been developed [30–37]. Concerning the macro properties of a macroscopic system, the effect of a large number of water molecules can be reasonably considered to represent the effect of a continuous medium. However, for an active site of a protein (which is usually located on the tip of the lateral chain of a residue, and is comparable in size to a water molecule) and its localized molecular orbitals, a conductor-like polarizable continuous medium is clearly not an acceptable model for water. We really need to consider the effect of the nearest individual water molecules. However, it is impossible to add a large amount of water molecules to the electronic structure calculation of protein in solution due to the intense computational effort required, though this computational effort can be greatly reduced using the SCCE method. Thus, in order to calculate the electronic structure of a protein in solution, it is necessary to construct an equivalent potential of water that is simple, easy-to-use and requires little additional computational effort.

It is reasonable to construct equivalent potentials of water for the electronic structures of the 20 kinds of amino acid (the building blocks of proteins). These equivalent potentials can then be used to represent the impact of the potential of water on a protein. Equivalent potentials of water for 13 amino acids have already been constructed: those for cysteine (Cys), lysine (Lys⁺), histidine (His), glutamic acid (Glu⁻), alanine (Ala), asparagine (Asp⁻), serine (Ser), threonine (Thr), asparagine (Asn), glycine (Gly), leucine (Leu), proline (Pro) and isoleucine (Ile) [38–48]. In this paper, we report the equivalent potential of water for arginine (Arg⁺).

This work is based on two considerations. (1) Our purpose is not to mimic the situation found in water at room temperature, but to mimic the potential that acts on the electronic structure of Arg⁺ due to the presence of water. From the viewpoint of the electronic structure of Arg⁺, it is reasonable to consider only the nearest water molecules that form hydrogen bonds with Arg⁺, and minimize the total energy. The reasons for this are as follows. (i) In our calculation, the Arg⁺ was fixed within its solvated structure. (ii) There are only a limited number of the nearest water molecules that can form hydrogen bonds with an Arg⁺. (iii) Although they fluctuate significantly, the water molecules surrounding the Arg⁺ are those that have the highest probability of being at the positions that minimize the total energy of the Arg⁺+9H₂O system. (iv) The potential of a removed dipole (such as a polar water molecule located far from the Arg⁺) attenuates as $1/r^2$. So the nearest water molecules that form hydrogen bonds with the Arg⁺ and minimize the total energy, at least to a first-order approx-

imation, contribute most of the effect of the potential of the solvent on the Arg⁺ electronic structure, no matter where the other water molecules are distributed. (2) Dipoles made up of point charges can easily be added to the SCCE calculation with almost no additional computational effort, and do not increase the CPU time. On the other hand, the average potentials of polar water molecules can be reasonably simulated by dipoles. Thus, in this work, we chose to use dipoles to simulate the potential of water.

Basic theory

The “free cluster calculation” and the “self-consistent cluster-embedding (SCCE) calculation” methods, which are based on density functional theory (DFT) [49, 50], have been described in detail elsewhere (see references [16, 17, 47] and the website <http://www.esprotein.org.cn>). Especially, we refer readers to [47]—which describes such calculations for glycine (Gly)—for further details concerning DFT, the free cluster calculation, the SCCE calculation, and the computational procedure.

The Arg⁺+9H₂O system, the calculation process and results

Before the calculation, we determined that the number of “nearest” water molecules around Arg⁺ was nine, according to hydrogen bonding and our experience. We then obtained both the optimized geometric structure and the electronic structure of the Arg⁺+9H₂O system with the free-cluster calculation method. Evidently, just nine water molecules cannot adequately describe the effect of water on the geometric structure of Arg⁺ considering the degrees of freedom of the geometric structure of Arg⁺ in solution. However, it is appropriate to use nine water molecules to describe the effect of water on the electronic structure of Arg⁺ in solution:

- (1) Both the properties and the biological functions of an amino acid depend mainly on the molecular orbitals near the HOMO (highest occupied molecular orbital), which are easily affected by the solvent environment. The molecular orbitals that are much lower than the HOMO are barely affected by the solvent. Therefore, we used nine water molecules to simulate the effect of water on the molecular orbitals near the HOMO. The degrees of freedom is small if we choose ten molecular orbitals near the HOMO, so the use of nine water molecules should be appropriate.
- (2) The valence electrons in the amino acid Arg⁺ are all localized. The molecular orbitals near the HOMO are

mainly localized around the amidogen (H_3N^+), carboxyl (COO^-) or lateral chain. Each H atom of Arg^+ may form a hydrogen bond with an O atom of a water molecule; likewise, each O atom of Arg^+ may form a hydrogen bond with an H atom of a water molecule, and one O atom can form at most two hydrogen bonds with two H atoms. Considering the fact above, and that a water molecule has one O atom and two H atoms, nine water molecules can form hydrogen bonds with neither more nor less than 15 H atoms and 2 O atoms of Arg^+ , so long as they stay at the right positions.

- (3) There is no doubt that the greater the number of molecules used the better the result will be. However, the following two facts indicate that nine water molecules may be the best choice. First, when a water molecule is near the amino acid, it will impede the approaches of other water molecules. The potential of a removed dipole (such as a polar water molecule) attenuates as $1/r^2$. Therefore, a certain number of water molecules that form hydrogen bonds with Arg^+ , at least to a first-order approximation, contribute most of the effect of the solvent on the electronic structure of Arg^+ . Second, suppose that eleven water molecules are positioned around Arg^+ . In this case, 30 geometric structures would be obtained, all of which have almost the same total energy. This degeneracy could make the calculations impossible and nonsensical. On the other hand, the complex interactions among water molecules are of no interest here. Thus, nine water molecules were applied in our calculations.

Distribution of the water molecules

The coordinates of the 27 atoms of Arg^+ listed in Table 1 were found from a PDB structure file provided by the Laboratory of Mass Spectrometry and Gaseous Ion Chemistry at the Rockefeller University (<http://prowl.rockefeller.edu/aainfo/struct.htm>).

The polar water molecules mainly affect the charged H_3N^+ and COO^- of Arg^+ , as well as its lateral chain. At the start, the nine water molecules were distributed in positions surrounding the Arg^+ according to our experience. Three were placed in the neighborhood of the H_3N^+ , with each water molecule's oxygen end oriented toward one of the hydrogen atoms of H_3N^+ . One water was near the COO^- , with its hydrogen end oriented toward the two oxygen atoms of COO^- , while five others were placed near the H atoms of the lateral chain. Each water molecule, depending on its initial position and orientation, was able to form hydrogen bonds with H or O atoms of Arg^+ and thus to lower the total energy of the $\text{Arg}^+ + 9\text{H}_2\text{O}$ system.

Table 1 Atomic coordinates of Arg^+

No.	Atom	X (Å)	Y (Å)	Z (Å)
1	C _α	2.807	0.022	1.011
2	C	1.442	-0.625	0.785
3	C	0.374	0.463	0.679
4	C	-0.992	-0.183	0.454
5	N	-2.019	0.863	0.352
6	C	-3.234	0.288	0.151
7	N	-4.307	1.041	0.030
8	N	-3.342	-1.066	0.076
9	N	3.128	0.924	-0.137
10	C	3.862	-1.054	1.116
11	O	4.528	-1.158	2.132
12	O	4.056	-1.850	0.145
13	H	2.787	0.600	1.932
14	H	1.209	-1.282	1.621
15	H	1.462	-1.204	-0.136
16	H	4.053	1.362	0.015
17	H	2.405	1.661	-0.209
18	H	3.147	0.373	-1.014
19	H	0.607	1.120	-0.156
20	H	0.354	1.042	1.601
21	H	-1.225	-0.840	1.290
22	H	-0.971	-0.762	-0.467
23	H	-1.802	1.473	-0.424
24	H	-5.116	0.454	-0.113
25	H	-4.432	1.587	0.870
26	H	-4.308	-1.319	-0.075
27	H	-2.781	-1.406	-0.692

The von Barth and Hedin [51] form of the exchange-correlation potential, as parameterized by Rajagopal and coworkers [52] was used in the calculations. An optimized linear combination of the Gaussian basis sets of C, N, O, and H atoms was also used [53–57]; parts of the original bases were uncontracted, several diffuse bases were inserted, and two polarization functions were added. They were the same as those used in the electronic structure calculations of proteins [24–27]; i.e., C, 8s6p, 26 Gaussian bases; N, 8s7p, 29 Gaussian bases; O, 8s7p, 29 Gaussian bases; H, 8s1p, 11 Gaussian bases. The total number of Gaussian bases was 954. The space occupied by the $\text{Arg}^+ + 9\text{H}_2\text{O}$ system was divided into 594,002 grid points so as to calculate the exchange-correlation energy.

The software Free Cluster Calculation was developed by the group of Prof. Callaway in the Department of Physics and Astronomy, Louisiana State University (USA) [58]. The electronic structures of many molecules and clusters have been calculated using this software [59–65]. By solving the Kohn–Sham equation self-consistently, we

obtained the electronic structure, the total energy, and the force exerted on each atomic nucleus.

Adjusting the nine water molecules

For the $\text{Arg}^+ + 9\text{H}_2\text{O}$ system, only the positions of water molecules relative to Arg^+ were adjusted. In other words, during the optimization, all of the nuclei in Arg^+ were fixed, and the nuclei in the water molecules were moved according to the forces while the geometric structure of each water molecule was fixed. Using a first-principles, all-electron, ab initio method, the electronic structure of the $\text{Arg}^+ + 9\text{H}_2\text{O}$ system was calculated self-consistently. Each time one of the nine water molecules was moved along the direction of applied force while the other water molecules remained fixed. Twelve step lengths, ranging from small to very large, were used for adjustments in order to avoid the adjustment from falling into a local minimum. Therefore, in all probability, our final optimized structure does not depend on the initial geometry. The position of the water molecule that gave the lowest total energy among the twelve step lengths was then saved. Then we moved the position of another water molecule that was subjected to the strongest applied force. Once the positions of all nine water molecules had all been adjusted in this manner, one round was finished and the second round was begun. After hundreds of rounds of such adjustments, the total energy of the system hovered near a particular value, which meant that the optimized geometric structure of the $\text{Arg}^+ + 9\text{H}_2\text{O}$ system had been obtained. The total energy was -2577.5250 Ry, 1.1193 Ry lower than that of the initial configuration. The atomic coordinates of the nine water molecules are given in Table 2. Figure 1 shows the optimized configuration of the whole $\text{Arg}^+ + 9\text{H}_2\text{O}$ system, the atoms of which are numbered according to Tables 1 and 2. In order to more clearly depict the positions of water molecules relative to Arg^+ , three parts of the $\text{Arg}^+ + 9\text{H}_2\text{O}$ system are shown individually in Figs. 2, 3, and 4.

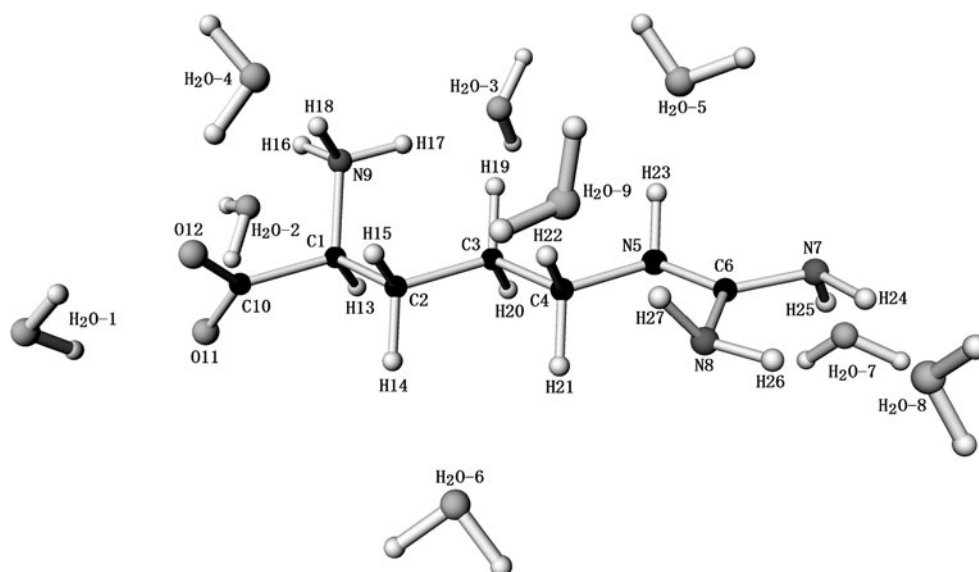
In principle, we can not obtain the global minimum. However, we are sure that the total energy of the geometric structure is very close to the minimum, as the room for further adjustment is very small. Therefore, the calculated electronic structure of Arg^+ under the influence of water should be a good approximation to the real electronic structure of Arg^+ in aqueous solution for the following reasons. First, we did not intend to explore the positions of water molecules relative to Arg^+ . Second, there are no fixed hydrogen bonds between the water molecules and the protein in solution, so there are no fixed relative positions between them. The water molecules are most probably at the positions that minimize the total energy of the system. Third, in order to reduce the total

Table 2 Final atomic coordinates of the nine water molecules

Water molecule	Atom	X (Å)	Y (Å)	Z (Å)
1	O28	6.2340	-3.1642	1.2762
	H29	5.9581	-2.5443	1.9514
	H30	5.6411	-2.9999	0.5429
2	O31	5.2578	1.4501	1.5159
	H32	5.1485	0.6634	2.0502
	H33	5.9858	1.9184	1.9245
3	O34	1.5349	3.4552	-0.1350
	H35	1.4722	3.8896	0.7157
	H36	1.3979	4.1557	-0.7728
4	O37	3.4152	-0.7761	-2.3060
	H38	3.7272	-1.4388	-1.6897
	H39	4.1544	-0.6307	-2.8965
5	O40	-2.0128	1.5450	-2.2426
	H41	-2.8409	1.6984	-2.6975
	H42	-1.3405	1.7722	-2.8849
6	O43	-0.5368	-2.9396	2.4153
	H44	0.1590	-3.3980	2.8864
	H45	-1.3331	-3.1429	2.9060
7	O46	-4.6120	2.7440	2.1828
	H47	-4.0549	2.8875	2.9478
	H48	-5.5039	2.8594	2.5104
8	O49	-6.2333	-1.2269	-0.3169
	H50	-6.8282	-1.4392	0.4022
	H51	-6.7092	-1.4850	-1.1063
9	O52	-1.6697	-2.0717	-2.2289
	H53	-1.2055	-2.9083	-2.1994
	H54	-1.7061	-1.8500	-3.1593

computational effort, charge-density fitting was used in both the free-cluster calculation and the band structure calculation. A pseudo charge density that differs from the real one but can give a total energy which is very close to that calculated using the real charge density was used to calculate the electronic structure. It is believed that the electronic structure obtained using this pseudo charge density is a good approximation.

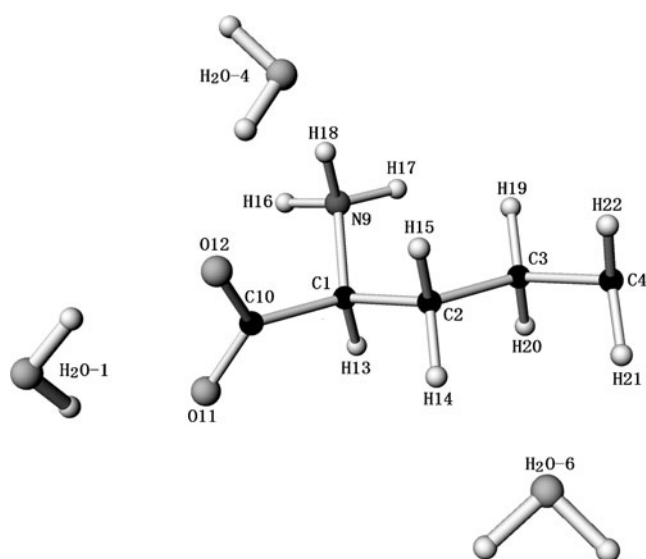
We used the free cluster calculation to optimize the geometric structure of the $\text{Arg}^+ + 9\text{H}_2\text{O}$ system instead of the molecular dynamics method (MD) due to the following reasons. (1) Our ultimate goal was to obtain the electronic structure of Arg^+ in water accurately, which means no pseudopotential, no adjustable parameters in the calculation, and the use of an adequate set of Gaussian bases which was the same as that used to calculate the electronic structure of the protein. In fact, for the $\text{Arg}^+ + 9\text{H}_2\text{O}$ system containing 54 atoms (184 electrons), we used 954 Gaussian bases, which is much larger than the number that can be used in MD. It is also well known that pseudopotential and Gaussian bases will

Fig. 1 Geometric structure of the $\text{Arg}^+ + 9\text{H}_2\text{O}$ system

affect the electronic structure as well as the geometric structure. (2) We did not want to change the structure of Arg^+ in solution and the structures of water molecules at all during the search for the lowest-energy conformation of $\text{Arg}^+ + 9\text{H}_2\text{O}$ complex. It may be difficult for MD to guarantee this. By the way, our free cluster calculation can partially perform the relaxation automatically, but the computational effort associated with this is large due to the large Gaussian bases.

Electronic structure of Arg^+ in the potential of water

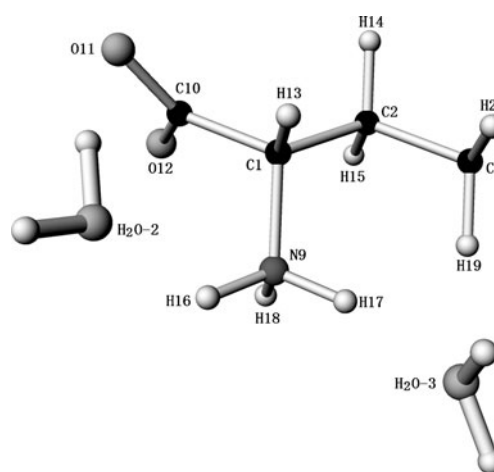
Based on the optimized geometric structure, the system was then divided into ten embedded clusters for the self-consistent

**Fig. 2** Positions of water molecules 1, 4, and 6 relative to the lateral part of arginine. The oxygen atom of water 4 forms a hydrogen bond with H18, and one of its hydrogen atoms connects with O12. Each oxygen atom of the carboxyl group connects to one hydrogen atom of water 1. The oxygen atom of water 1 is shared by H14 and H21

cluster-embedding calculation (SCCE) [66]. Arg^+ was set as the first cluster, and the nine water molecules were set as the other nine clusters. In the SCCE calculation, the total potential was the same as that in the free cluster calculation, but the localized one-electron wavefunctions were substituted for extended one-electron wavefunctions: each one-electron wavefunction was then localized in the region of a cluster. Thus, the electronic structure of Arg^+ was isolated from that of the water molecules.

The SCCE calculation involves two kinds of iterations: intra-cluster iteration and inter-cluster iteration:

- (i) *Intra-cluster iteration.* For each embedded cluster, the Kohn–Sham equation [50] was calculated self-consistently: the $\rho_1(\vec{r})$ of the embedded cluster was self-consistently changed during the iterations, while the rest of the system served as its fixed environment $\rho_2(\vec{r})$.

**Fig. 3** Positions of waters 2 and 3 relative to the backbone of arginine. One hydrogen atom of water 2 points to O11. Its oxygen atom points to H16. The hydrogen atoms H17 and H19 share the oxygen atom of water 3

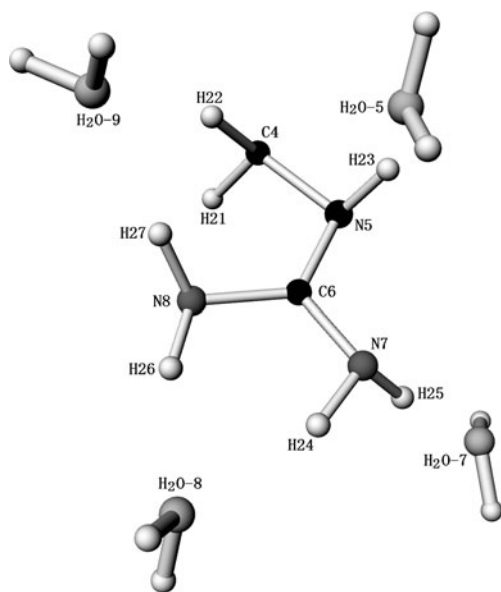


Fig. 4 The positions of the water molecules 5, 7, 8, and 9 relative to arginine. The oxygen atom of water 8 is shared by H26 and H24. The oxygen atom of water 9 is shared by H22 and H27. The oxygen end of water 5 points to H23. The oxygen atom of water 7 is closest to H25

(ii) *Inter-cluster iteration.* The ten embedded clusters were synchronously calculated by ten CPUs. After the intra-cluster iterations of all ten embedded clusters had converged, the results were used to construct new environments $\rho_2(\vec{r})$ for each embedded cluster, and a new inter-cluster iteration was started. After ten inter-cluster iterations, converged results were obtained.

Listed in Table 3 are the eigenvalues and Mulliken populations of ten orbitals near the HOMO of Arg^+ under the influence of nine water molecules.

Table 3 Some of the eigenvalues and Mulliken populations of Arg under the potential of the nine water molecules

State	Energy (Ry)	Mulliken population							
		C		N		O		H	
		<i>s</i>	<i>p</i>	<i>s</i>	<i>p</i>	<i>s</i>	<i>p</i>	<i>s</i>	<i>p</i>
49	-0.3481	-2.3765	0.0030	5.2051	1.8548	0.0004	-0.0004	-3.6936	0.0072
48	-0.4155	-0.0684	0.4853	0.1707	0.5371	0.0000	0.0001	-0.1314	0.0066
The above states are unoccupied									
47	-0.5212	0.0059	-0.0141	0.0715	-0.0058	0.0781	0.9236	-0.0595	0.0003
46	-0.5539	0.0404	0.0827	0.0315	0.0123	-0.0269	0.9071	-0.0473	0.0002
45	-0.5700	0.2849	-0.1268	0.0280	0.0046	-0.0288	0.8138	0.0238	0.0004
44	-0.7704	0.0613	0.1555	-0.0053	0.6350	-0.0004	0.0277	0.1174	0.0088
43	-0.7929	0.0110	0.5269	-0.0168	0.1214	-0.0109	0.3000	0.0638	0.0045
42	-0.8201	-0.0061	0.4960	0.0393	0.1067	0.0046	0.0855	0.2638	0.0102
41	-0.8319	0.0247	0.1362	-0.0202	0.5882	0.0021	0.0183	0.2369	0.0138
40	-0.8536	0.0538	0.3869	-0.0102	0.1325	-0.0011	0.2928	0.1393	0.0060

Because of the tiny populations associated with the *d* electrons, they are not given in this table and the tables below

Simulating the potential of water by dipoles

After the electronic structure of Arg^+ in the potential of water had been obtained, nine dipoles were substituted for the nine water molecules: the O atom of each water molecule was replaced with a negative point charge, while the two H atoms were replaced with a positive charge located at the middle of the line connecting the two H atoms. The electronic structure of Arg^+ was recalculated using the SCCE calculation by adjusting the point charges according to the difference in the electronic structure of Arg^+ obtained when the water potential and the dipole potential were considered.

It is important to note that the electronic structure of $\text{Arg}^+ + 9\text{H}_2\text{O}$ obtained with the free cluster calculation cannot be used as fitting criteria to adjust the dipoles in the SCCE calculation, as the two systems have different numbers of electrons and different distribution regions of the electrons. It is the electronic structure of Arg^+ in the $\text{Arg}^+ + 9\text{H}_2\text{O}$ system obtained using the SCCE calculation that can be used to approximate the electronic structure of Arg^+ in water, and is suitable for use as fitting criteria. To aid in the evaluation of the difference in electronic structure between $\text{Arg}^+ + 9\text{H}_2\text{O}$ and $\text{Arg}^+ + 9$ dipoles, two quantitative criteria were established:

(1) The mean square deviation of eigenvalues

$$\overline{\Delta E^\sigma} = \frac{1}{N\sigma} \left[\sum_{n=1}^{N\sigma} (\varepsilon_n^\sigma - \varepsilon_{n0}^\sigma)^2 \right]^{1/2}, \quad (\text{a})$$

where ε_n^σ and ε_{n0}^σ are the eigenvalues of the n^{th} molecular orbital with spin σ calculated in this section

and in the previous section, respectively. N^σ is the number of electrons with spin σ .

- (2) The equivalent mean square deviation of charge density

$$\overline{\Delta C^\sigma} = \frac{1}{N} \left[\sum_{i=1}^N \sum_{j=1}^N \left(\sum_{n=1}^{N^\sigma} C_{ni}^{\sigma*} C_{nj}^\sigma - \sum_{n=1}^{N^\sigma} C_{ni0}^{\sigma*} C_{nj0}^\sigma \right)^2 \right]^{1/2}, \quad (b)$$

where C_{ni}^σ and C_{ni0}^σ are the expansion coefficients of the eigenfunctions of the n^{th} molecular orbitals with spin σ calculated in this section and in the previous section, respectively. N is the number of Gaussian bases used to expand the one-electron wavefunction. (Note the charge density:

$$\begin{aligned} \rho^\sigma(\vec{r}) &= \sum_{n=1}^{N^\sigma} |\psi_n^\sigma(\vec{r})|^2 = \sum_{n=1}^{N^\sigma} \left[\sum_{i=1}^N C_{ni}^{\sigma*} U_i^*(\vec{r}) \right] \left[\sum_{j=1}^N C_{nj}^\sigma U_j(\vec{r}) \right] \\ &= \sum_{i=1}^N \sum_{j=1}^N \left[\sum_{n=1}^{N^\sigma} C_{ni}^{\sigma*} C_{nj}^\sigma \right] U_i^*(\vec{r}) U_j(\vec{r}) \end{aligned}$$

where the Gaussian bases $U_i(\vec{r})$ are identical in the two calculations.)

The electronic structure of Arg^+ in the dipole potential is considered to be proximate to that in the water potential when criteria (a) and (b) are minimized. In practice, however, we mainly use criterion (a), because the reliability of criterion (b) can be impacted by the charge density fitting procedure [58, 67] used in the calculations.

The values of the two criteria after a new combination of dipole charge values were calculated with SCCE were then compared with the previous values. After hundreds of adjustments, the value of criterion (a) decreased from 5.4700×10^{-3} to 1.9246×10^{-3} , which indicated that the electronic structure of Arg^+ of the system $\text{Arg}^+ + 9$ dipoles was mostly proximate to the system $\text{Arg}^+ + 9\text{H}_2\text{O}$.

There are a number of studies in the literature that address the problem of achieving a simplified discrete solvent representation using point charges; each has its own advantages. On the website <http://www.lsbu.ac.uk/water/models.html>, 23 water models and more than 1200 papers can be found on this subject. The review written by Guillot [68] listed 46 distinct water models, which indirectly indicates their lack of success in quantitatively reproducing the properties of real water. In our previous studies [40, 42], two popular three-charge water models, TIP4P-FQ [69] and SPC [70], were tried as well as the dipole model. However, when calculating the electronic structure of an amino acid, the more complicated three-charge water models did not give a better fit than the simple dipole model (this will be discussed in another paper). Two facts should be emphasized here. First, we are not attempting to construct a general water model, since this is a complex problem with a

lot of associated issues, but rather to construct an equivalent potential of water especially for electronic structure calculations of proteins in solution. For example, the dipoles constructed for Arg^+ will be applied to Arg^+ peptides located on the surface of a protein. Second, we are not attempting to construct an exact special equivalent potential of water. Actually, this does not exist: it is impossible to make the electronic structure of Arg^+ in the dipole potential absolutely identical to that in the water potential. Our aim is rather to construct a simple and easy-to-use potential which, at least to a first-order approximation, contributes most of the effect of the solvent on the electronic structure of the protein.

The distance between the positive charge and negative charge of a dipole was fixed at 0.5858 Å during the calculations. The initial charges of the nine dipoles were all set to be 0.5e. The charges of the nine dipoles were then adjusted in turns until the two criteria reached their minima. The final charges and coordinates of the nine dipoles are given in Table 4. An ichnography of the geometric structure of $\text{Arg}^+ + 9$ dipoles is shown in Fig. 5. Table 5 lists the eigenvalues and Mulliken populations of ten orbitals near the HOMO of Arg^+ in the $\text{Arg}^+ + 9$ dipoles system.

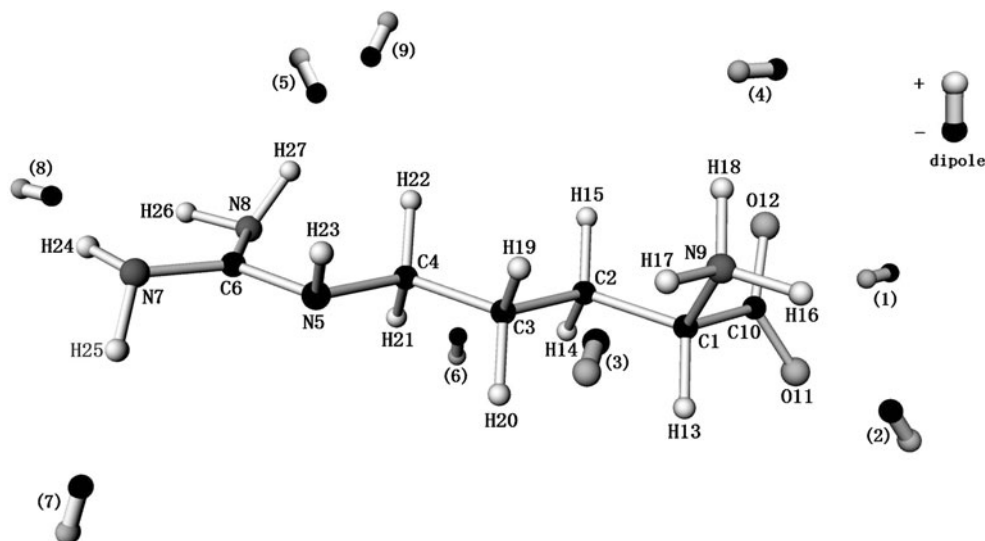
Results analysis

The electronic structure of Arg^+ was calculated to study the influence of water on the electronic structure of Arg^+ . The

Table 4 Point charges and coordinates of the nine dipoles

Dipole	Charge (e)	X (Å)	Y (Å)	Z (Å)
1	-0.69	6.2340	-3.1642	1.2762
	0.69	5.7996	-2.7721	1.2471
2	-1.11	5.2578	1.4501	1.5159
	1.11	5.5672	1.2909	1.5159
3	-0.83	1.5349	3.4552	-0.1350
	0.83	1.4350	4.0226	-0.0286
4	-2.15	3.4152	-0.7761	-2.3060
	2.15	3.9408	-1.0347	-2.2931
5	-0.77	-2.0128	1.5450	-2.2426
	0.77	-2.0907	1.7353	-2.7912
6	-0.37	-0.5368	-2.9396	2.4153
	0.37	-0.5870	-3.2705	2.8962
7	-0.70	-4.6120	2.7440	2.1828
	0.70	-4.7794	2.8735	2.7291
8	-0.46	-6.2333	-1.2269	-0.3169
	0.46	-6.7687	-1.4621	-0.3520
9	-0.60	-1.6697	-2.0717	-2.2289
	0.60	-1.4558	-2.3792	-2.6794

Fig. 5 Geometric structure of the Arg⁺ + 9 dipoles system



total energy of an isolated Arg⁺ molecule is -1207.4910 Ry. The eigenvalues and Mulliken populations of ten orbitals near the HOMO of Arg⁺ are listed in Table 6.

The eigenvalues of states 40–49 of Arg⁺ in the three cases (in the potential of dipoles, water molecules and no potential) are shown in Table 7. The last row gives the energy gap between states 48 and 47. A sketch map is shown in Fig. 6 that compares the three sets of eigenvalues.

The properties of Arg⁺ are determined mainly by the molecular orbitals near the HOMO. Tables 3, 5 and 6 show that the molecular orbitals below the HOMO are similar in the potential of water and the dipole potential. We now provide a detailed description of eight orbitals below the

LUMOs of the two cases. Orbitals 46 and 47 are mainly occupied by the $2p$ electrons of O in COO⁻. Orbital 45 is a hybridized state contributed to mainly by the $2p$ electrons of O and the $2s$ electrons of C in COO⁻, as well as the $2s$ and $2p$ electrons of C α . Orbital 44 is mainly occupied by the $2p$ electron of N5. Orbital 43 is a hybridized state that is mainly occupied by $2p$ electrons from the C and O atoms of the COO⁻. Orbital 42 is occupied mainly by $2p$ electrons from two carbon atoms (C2, C3) in the backbone. Orbital 41, a hybridized state, is occupied by $2p$ electrons from the N atom of the lateral chain. Orbital 40 is mainly occupied by $2p$ electrons from the COO⁻ and a carbon atom (C2).

Table 5 Some of the eigenvalues and Mulliken populations of Arg⁺ in the potential of 9 dipoles

State	Energy (Ry)	Mulliken population							
		C		N		O		H	
		<i>s</i>	<i>p</i>	<i>s</i>	<i>p</i>	<i>s</i>	<i>p</i>	<i>s</i>	<i>p</i>
49	-0.2912	-0.7409	0.1198	1.7448	0.6819	0.0001	-0.0001	-0.8154	0.0098
48	-0.4140	-0.0069	0.4920	0.0157	0.4051	0.0000	0.0001	0.0878	0.0064
The above states are unoccupied									
47	-0.5281	0.0117	-0.0358	0.0313	0.0077	0.0618	0.9496	-0.0268	0.0005
46	-0.5441	0.0150	-0.0071	0.0162	0.0053	0.0008	0.9800	-0.0103	0.0001
45	-0.5811	0.2094	-0.0338	0.0221	0.0150	-0.0194	0.8207	-0.0142	0.0002
44	-0.7707	0.0197	0.3342	0.0246	0.4127	0.0021	0.1451	0.0540	0.0076
43	-0.7887	0.0054	0.3208	-0.0091	0.3693	0.0038	0.2365	0.0675	0.0059
42	-0.8200	0.0101	0.4360	0.0060	0.1902	0.0350	0.1045	0.2072	0.0110
41	-0.8305	0.0486	0.1619	-0.0198	0.5116	0.0324	0.0533	0.1981	0.0140
40	-0.8527	0.0364	0.3525	0.0052	0.0336	0.1043	0.2934	0.1684	0.0062

Table 6 Some of the eigenvalues and Mulliken populations of an isolated Arg⁺

State	Energy (Ry)	Mulliken Population							
		C		N		O		H	
		<i>s</i>	<i>p</i>	<i>s</i>	<i>p</i>	<i>s</i>	<i>p</i>	<i>s</i>	<i>p</i>
49	−0.4178	−1.3125	−0.0126	0.7820	0.0546	−0.0006	0.0005	1.4756	0.0131
48	−0.5335	0.0063	0.4915	0.0008	0.3822	0.0006	0.0156	0.0968	0.0061
The above states are unoccupied									
47	−0.5430	−0.0073	0.1494	−0.0627	0.0065	0.0008	0.8892	0.0232	0.0009
46	−0.5561	0.0409	−0.0081	0.0018	0.0017	0.0005	0.9707	−0.0076	0.0001
45	−0.6032	0.1940	0.0939	−0.0549	−0.0067	−0.0206	0.7913	0.0029	0.0002
44	−0.8227	0.0108	0.4833	−0.0023	0.0185	0.0150	0.4676	0.0052	0.0018
43	−0.8563	0.0127	0.2101	−0.0052	0.0109	0.2388	0.4997	0.0319	0.0012
42	−0.8960	0.0318	0.1385	−0.0492	0.7592	0.0005	0.0123	0.0969	0.0101
41	−0.9304	0.0096	0.3972	0.0089	0.0469	0.1358	0.2931	0.1027	0.0058
40	−0.9431	0.0587	0.1847	−0.0139	0.5961	0.0277	0.0676	0.0662	0.0129

The orbitals in the case of free Arg⁺ are different from the former two cases, except for orbitals 45, 46, and 47. Orbital 44 is occupied by the 2*p* electrons of the C and O atoms of COO[−]. Besides those 2*p* electrons, orbital 43 is also occupied by the 2*p* electron of N5. Orbital 42 is mainly occupied by the 2*p* electrons of the N3 atom of the lateral chain. Orbital 41 is highly hybridized and is mainly occupied by the 2*p* electrons of C and O atoms of COO[−] and the 2*p* electron of C2C3. Orbital 40 is occupied primarily by the 2*p* electrons of N7 and N8.

By comparing the second column with the third one in Table 7, and Fig. 3a with Fig. 3b, it was found that the

influence of water does not change the electronic structure of Arg⁺. Its main effect is to broaden the energy gap tenfold, and to increase the eigenvalues of all orbitals by about 0.0546 Ry on average.

By comparing columns 2 and 3 of Table 7 and Fig. 3b with Fig. 3c, we found that below the HOMO, the eigenvalues of Arg⁺ in the dipole potential are very close to those in the potential of water, except for orbitals 45 and 46. Compared to the water case, orbital 45 is lowered by 0.0111 Ry while orbital 46 increases by 0.0098 Ry in the dipole case. Moreover, the energy gap in the case of water is very close to that obtained in the dipole potential. Because unoccupied orbitals make no contribution to the charge density, it safe to conclude that the potential of dipoles gives a good simulation of the effect of water on the electronic structure of Arg⁺.

Although the equivalent potential of water was especially constructed for our calculation of the electronic structure of the protein, it is transferable. We have been using dipoles to construct the equivalent potentials of water for the electronic structures of 20 amino acids. It is possible to apply the obtained equivalent potentials to SCCE calculations as well as any first-principles, all-electron, ab initio calculation method that is used to calculate the electronic structure of the protein in solution. Amino acids lose their water molecules when they turn into amino acid residues, and then combine into several interlaced polypeptide chains, thereby forming a protein. Each chain has only one N-terminal H₃N⁺, one C-terminal COO[−], and many lateral chains. For the electronic structure of a protein with a known geometric structure in solution, water solvent does not affect the molecular

Table 7 Three sets of eigenvalues of Arg⁺

State	Eigenvalues (Ry)		
	No potential	Potential of water molecules	Potential of dipoles
49 (unoccupied)	−0.4178	−0.3481	−0.2912
48 (unoccupied)	−0.5335	−0.4155	−0.4140
47 (<i>E_F</i>)	−0.5430	−0.5212	−0.5281
46	−0.5561	−0.5539	−0.5441
45	−0.6032	−0.5700	−0.5811
44	−0.8227	−0.7704	−0.7707
43	−0.8563	−0.7929	−0.7887
42	−0.8960	−0.8201	−0.8200
41	−0.9304	−0.8319	−0.8305
40	−0.9431	−0.8536	−0.8527
<i>E_g</i> = <i>E</i> ₄₈ − <i>E</i> ₄₇	0.0095	0.1057	0.1141

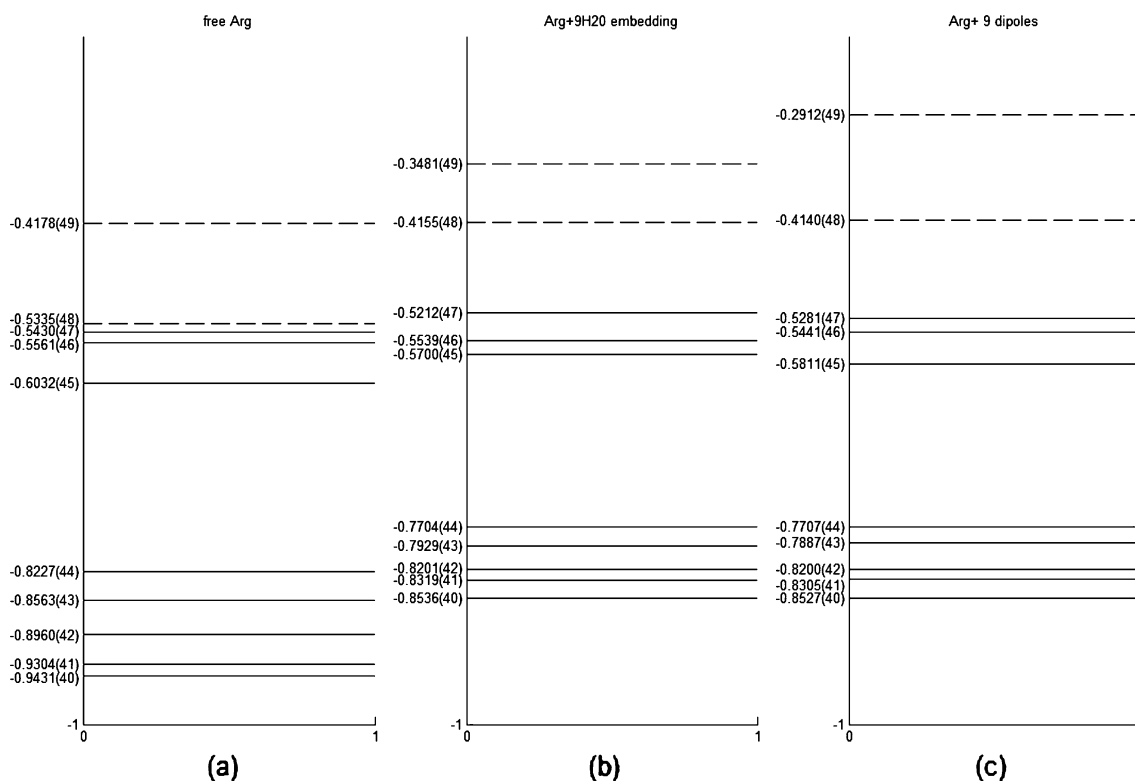


Fig. 6 Ten eigenvalues (47 is the eigenvalue of the HOMO orbital) of Arg^+ in three cases: the isolated state (a), in the potential of water (b), and in the potential of dipoles (c)

orbitals of amino acid residues in the interior of the protein due to screening effects. Thus, we only need to put the dipoles (when using our dipole potentials) near the N-terminus and C-terminus of a chain, and near the tips of lateral chains of the amino acid residues that are on the surface of the protein. A program will soon be developed that yields both the charges and the locations of the dipoles around an amino acid residue located on the surface of a protein so long as the atomic coordinates of the protein are provided. This should help us to calculate the electronic structure of a protein in solution (to determine its active sites and reactive sites) more reliably.

Conclusions

The equivalent potential of water for the electronic structure of Arg^+ in solution was successfully simulated with dipoles by performing first-principles, all-electron, ab initio calculations. Three steps were performed in the simulation process. First, the geometric structure of the $\text{Arg}^+ + 9\text{H}_2\text{O}$ system was optimized by free cluster calculation. Second, based on the optimized structure, the electronic structure of Arg^+ in the potential of water molecules was obtained with the SCCE calculation.

Finally, using dipoles, the electronic structure of Arg^+ in the potential of water was simulated.

Comparing the electronic structure of Arg^+ among the three cases (no potential, in the potential of water, and in the dipole potential), the main effects of water on the electronic structure of Arg^+ were a broadening of the energy gap tenfold, and an increase in the eight eigenvalues below the HOMO by about 0.0546 Ry on average. The effect of water on the electronic structure of Arg^+ can be simulated well by dipoles: for the molecular orbitals under the HOMO, the eigenvalues in the latter two cases are very close to each other. Dipoles are simple and easy to use, and thus are suitable for simulating the effect of water on the electronic structure of Arg^+ . Employing the equivalent potential represented by dipoles requires almost no additional computational effort, and adds no more CPU time.

The work required to construct equivalent potentials for other amino acids will soon be finished. They will then be directly applied to calculations of the electronic structures of proteins in solution.

Acknowledgments This work was supported by the National Natural Science Foundation of China (grant no. 30970694). The work was also supported by the Shanghai Supercomputer Center. The calculations were performed on the supercomputer DAWN 5000A of the Shanghai Supercomputer Center of China.

References

- Yang WT (1991) Direct calculation of electron density in density-functional theory. *Phys Rev Lett* 66:1438–1441
- Cortona P (1991) Self-consistently determined properties of solids without band structure calculations. *Phys Rev B* 44:8454–8458
- Galli G, Parrinello M (1992) Large scale electronic structure calculations. *Phys Rev Lett* 69:3547–3550
- Mauri F, Galli G, Car R (1993) Orbital formulation for electronic-structure calculations with linear system-size scaling. *Phys Rev B* 47:9973–9976
- Li XP, Nunes RW, Vanderbilt D (1993) Density-matrix electronic-structure method with linear system-size scaling. *Phys Rev B* 47:10891–10894
- Ordejon P, Drabold DA, Martin RM, Grumbach MP (1995) Linear system-size scaling methods for electronic-structure calculations. *Phys Rev B* 51:1456–1476
- Yang WT, Lee TS (1995) A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules. *Chem Phys* 103:5674–5678
- Kohn W (1996) Density functional and density matrix method scaling linearly with the number of atoms. *Phys Rev Lett* 76:3168–3171
- Ordejón P, Artacho PE, Soler JM (1996) Self-consistent order-N density-functional calculations for very large systems. *Phys Rev B* 53:10441–10444
- Baer R, Gordon MH (1997) Sparsity of the density matrix in Kohn–Sham density functional theory and an assessment of linear system-size scaling methods. *Phys Rev Lett* 79:3962–3965
- Klessinger M, Mcweeny R (1965) Self-consistent group calculations on polyatomic molecules. *J Chem Phys* 42:3343–3354
- Li JB, McWeeny R (2002) VB2000: Pushing valence bond theory to new limits. *Int J Quantum Chem* 89:208–216
- Wesolowski TA, Warshel A (1993) Frozen density functional approach for *ab initio* calculations of solvated molecules. *J Phys Chem* 97:8050–8053
- Wesolowski TA (2006) One-electron equations for embedded electron density: challenge for theory and practical payoffs in multi-level modeling of soft condensed matter. In: Leszczynski J (ed) *Computational chemistry: reviews of current trends*, vol X. World Scientific, Singapore, pp 1–82
- Govind N, Wang YA, da Silva AJR, Carter EA (1998) Accurate *ab initio* energetics of extended systems via explicit correlation embedded in a density functional environment. *Chem Phys Lett* 295:129–134
- Zheng H (1997) One-electron approach and the theory of the self-consistent cluster-embedding calculation method. *Phys Lett A* 226:223–230
- Zheng H (1993) Self-consistent cluster-embedding calculation method and the calculated electronic structure of NiO. *Phys Rev B* 48:14868–14883
- Zheng H (1995) Electronic structure of CoO. *Phys B* 212:125–138
- Zheng H, Rao BK, Khanna SN, Jena P (1997) Electronic structure and binding energies of hydrogen-decorated vacancies in Ni. *Phys Rev B* 55:4174–4181
- Zheng H, Wang Y, Ma G (2002) Electronic structure of LaNi₅ and its hydride LaNi₅H₇. *Eur Phys J B* 29:61–69
- He J, Zheng H (2002) The electronic structure of GaN and a single Ga-vacancy in GaN crystal. *Acta Phys Sin* 51:2580–2588
- Lin S, Zheng H (2005) Electronic structure of the surface of LaNi₅ crystal. *Acta Phys Sin* 54:4680–4687
- Zhen H, Lin S (2006) First-principles calculation of LaNi₅ surface. *J Phys Conf Ser* 29:129–140
- Zheng H (2000) Electronic structure of trypsin inhibitor from squash seeds in aqueous solution. *Phys Rev E* 62:5500–5508
- Zheng H (2000) First principle *ab initio* calculation of the electronic structure of protein molecule. *Prog Phys* 20:291–300
- Zheng H (2002) *Ab initio* calculations of the electronic structures and biological functions of protein molecules. *Mod Phys Lett B* 16:1151–1162
- Zheng H (2003) Electronic structures of *Ascaris* trypsin inhibitor in solution. *Phys Rev E* 68:051908-1–051908-8
- Sato F, Yoshihiro T, Era M, Kashiwagi H (2001) Calculation of all-electron wavefunction of hemoprotein cytochrome c by density functional theory. *Chem Phys Lett* 341:645–651
- Yoshihiro T, Sato F, Kashiwagi H (2001) Distributed parallel processing by using the object-oriented technology in ProteinDF program for all-electron calculations on proteins. *Chem Phys Lett* 346:313–321
- Lazaridis T, Karplus M (1999) Effective energy function for proteins in solution. *Proteins Struct Funct Genet* 35:133–152
- Lazaridis T, Karplus M (1998) Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol* 288:477–487
- Onsager L (1936) Electric moment of molecules in liquids. *J Am Chem Soc* 58:1486–1493
- Klamt A, Schuurmann G (1993) COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J Chem Soc Perkin Trans* 2:799–803
- Guo H, Karplus M (1994) Solvent influence on the stability of the peptide hydrogen bond: a supramolecular cooperative effect. *J Phys Chem* 98:7104–7105
- Schaefer M, Karplus MA (1996) Comprehensive analytical treatment of continuum electrostatics. *J Phys Chem* 100:1578–1599
- Eckert F, Klamt A (2002) Fast solvent screening via quantum chemistry: COSMO-RS approach. *AIChE J* 48:369–385
- Foresman JB, Keith TA, Wiberg KB (1996) Solvent effects. 5. Influence of cavity shape, truncation of electrostatics, and electron correlation on *ab initio* reaction field calculations. *J Phys Chem* 100:16098–16104
- Wang X, Zheng H, Li C (2006) The equivalent potential of water molecules for electronic structure of cysteine. *Eur Phys J B* 52:255–263
- Li C, Zheng H, Wang X (2007) The equivalent potential of water molecules for electronic structure of lysine. *Sci China Ser G* 50:15–30
- Li C, Zheng H, Wang X (2007) The equivalent potential of water molecules for the electronic structure of histidine. *J Phys Condens Matter* 19:16102-1–16102-15
- Zhang T, Zheng H, Yan S (2007) Equivalent potential of water molecules for electronic structure of glutamic acid. *J Comput Chem* 28:1848–1857
- Yan S, Zheng H, Zhang T (2008) The equivalent potential of water molecules for electronic structure of alanine. *Mol Phys* 106:1427–1439
- Zhang T, Zheng H, Yan S (2008) Equivalent potential of water molecules for electronic structure of aspartic acid. *J Comput Chem* 29:1780–1787
- Wang X, Zheng H (2009) Simulation of water potential for the electronic structure of serine. *Chin Phys B* 18:1968–1178
- Shen X, Gao Y, Zheng H (2009) The equivalent dipole potential of water for the electronic structure of threonine. *Mol Phys* 107:1393–1405
- Gao Y, Shen X, Zheng H (2010) Equivalent potential of water for electronic structure of asparagines. *Int J Quantum Chem* 110:925–938

47. Min P, Zheng H (2010) Equivalent potential of water for electronic structure of glycine. *J Mol Model* 17:111–124
48. Wang X, Zheng H (2011) A computer simulation of the electronic structure of leucine in solution. *J Solution Chem* (accepted)
49. Hohenberg P, Kohn W (1964) Inhomogeneous electron gas. *Phys Rev B* 136:864–871
50. Kohn W, Sham L (1965) Self-consistent equations including exchange and correlation effects. *J Phys Rev A* 140:1133–1138
51. von Barth U, Hedin L (1972) A local exchange-correlation potential for the spin polarized case: I. *J Phys C* 5:1629–1637
52. Rajagopal AK, Singhal S, Kimball J (1979) As quoted by Rajagopal AK (unpublished). In: Prigogine GI, Rice SA (eds) *Advance in chemical physics*, vol 41. Wiley, New York, p 59
53. van Duijneveldt FB (1971) *IBM J Res Dev* 945:16437
54. Lie GL, Clementi E (1974) *J Chem Phys* 60:1275–1287
55. Poirier RA, Daudel R, Mezey PG, Csizmadia IG (1982) *Int J Quantum Chem* 21:799–811
56. Huzinaga S (1965) *J Chem Phys* 42:1293–1302
57. Poirier R, Kari R, Csizmadia IG (1985) *Handbook of Gaussian basis sets*. Elsevier, New York
58. Chen H (1988) *Electronic structure of clusters: applications to high-T_c superconductors* (Ph.D. dissertation). Louisiana State University, Baton Rouge
59. Chen H, Callaway J, Misra PK (1988) Electronic structure of Cu-O chains in the high-T_c superconductor YBa₂Cu₃O₇. *Phys Rev B* 38:195–203
60. Chen H, Callaway J (1991) Local electronic structure and magnetism of 3d transition-metal impurities (Cr, Mn, Fe, Co, and Ni) in La_{2-x}Sr_xCuO₄. *Phys Rev B* 44:2289–2296
61. Zheng H, He J (2001) Limitations of conventional one electron approximation methods. *J Tongji Univ* 29:593–597
62. Xu W, Zheng H (2003) Theoretic calculations of Co and Ni clusters with different sizes. *J Tongji Univ* 31:374–378
63. Lin S, Zheng H (2004) Electronic structure of new oxygen molecule O₄. *J Tongji Univ* 32:551–555
64. Hao J, Zheng H (2004) Theoretical calculation of structures and properties of Ga₆N₆ cluster. *Acta Phys Sin* 53:1044–1049
65. Zheng H, Hao J (2005) Ab initio study of the electronic properties of the planar Ga₅N₅ cluster. *Chin Phys* 14:529–532
66. Zheng H (1993) *Self-consistent cluster-embedding calculation method and the electronic structure of NiO and CoO* (Ph.D. dissertation). Louisiana State University, Baton Rouge
67. Guillot B (2002) A reappraisal of what we have learnt during three decades of computer simulations on water. *J Mol Liq* 101:219–260
68. Rick SW (2001) Simulations of ice and liquid water over a range of temperatures using the fluctuating charge model. *J Chem Phys* 114:2276–2283
69. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1981) *Intermolecular forces*. Reidel, Dordrecht, p 331
70. Robinson GW, Zhu SB, Singh S, Evans MW (1996) *Water in biology, chemistry and physics: experimental overviews and computational methodologies*. World Scientific, Singapore

Influence of stereochemistry on proton transfer in protonated tripeptide models

Namat Ali Soliman · Petr Kulhánek · Jaroslav Koča

Received: 16 February 2011 / Accepted: 2 May 2011 / Published online: 28 May 2011
© Springer-Verlag 2011

Abstract Vectorial proton transfer among carbonyl oxygen atoms was studied in two models of tripeptide via quantum chemical calculations using the hybrid B3LYP functional and the 6-31++G** basis set. Two principal proton transfer pathways were found: a first path involving isomerization of the proton around the double bond of the carbonyl group, and a second based on the large conformational flexibility of the tripeptide model where all carbonyl oxygen atoms cooperate. The latter pathway has a rate-determining step energy barrier that is only around half of that for the first pathway. As conformational flexibility plays a crucial role in second pathway, the effect of attaching methyl groups to the alpha carbon atoms was studied. The results obtained are presented for all four possible stereochemical configurations.

Keywords Conformational rearrangement · Density functional theory · Protonated peptides · Proton transfer

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1116-2) contains supplementary material, which is available to authorized users.

N. A. Soliman · P. Kulhánek · J. Koča
Faculty of Science - National Centre for Biomolecular Research,
Masaryk University,
Kamenice 5,
CZ-625 00 Brno, Czech Republic

P. Kulhánek · J. Koča (✉)
Central European Institute of Technology (CEITEC),
Masaryk University,
Kamenice 5,
CZ-625 00 Brno, Czech Republic
e-mail: jkoca@chemi.muni.cz

Introduction

Recently, interactions of peptides or proteins with protons have been very extensively studied both experimentally and theoretically [1–13]. There is no doubt that such interactions occurring in a water environment play crucial roles in all biological systems, such as in the catalysis involved in many enzymatic reactions [14, 15], especially amide bond hydrolysis [1], as well as bioenergetic proton transport [16, 17]. From a structural point of view, the formation of hydrogen-bonded bridges between water and protonated peptides can change molecular geometries and conformational equilibria, resulting in different biological activities. When a longer peptide chain is available, the proton can interact with more groups, and proton transfer may occur among them.

Oligopeptides and proteins contain several positions to which the proton can attach. These are most notably the terminal amino group, carbonyl oxygen atoms, amide nitrogen atoms of the peptide backbone, and basic side chains (lysine, arginine, histidine). The best position on the peptide backbone to attach the proton is the terminal amino group, because it has the highest basicity. However, peptides contain other groups that are less basic. These are mainly oxygen and nitrogen atoms of peptide bonds. Proton interactions with such atoms are not particularly strong but they may play an important role under certain circumstances. The hydrolysis of a peptide bond in an acid solution is a good example. In the gas phase, the importance of this type of interaction is increasing, as it is expected to play a key role in peptide fragmentation processes when peptides are analyzed by mass spectrometry [18, 19]. Such methods are becoming more and more important with the development of soft ionization methods

such as a fast atom bombardment (FAB) [20], matrix-assisted laser desorption/ionization (MALDI) [21], and electrospray ionization (ESI) [22].

Because oxygen and nitrogen atoms are regularly distributed along the peptide chain, they can support vectorial proton transfer along it. This idea has been the subject of several computational studies [2–5]. These studies concluded that proton transfer between oxygen and nitrogen atoms within a single amide group is accompanied by a high energy barrier [2] of about $39.1 \text{ kcal mol}^{-1}$. The speed of this process can be increased in the presence of some protic compounds. However, in this case, the mechanism changes from direct proton transfer to transfer accompanied by proton exchange, and the energy barriers are still higher than those in processes involving only carbonyl oxygen atoms as interaction positions. The proton can interact with these by two ways. In both of them, the proton lies in the plane of the amide bond and is attached to the oxygen from either the C_α or the nitrogen side (e.g., two such structures are *E/Z* isomers). Two principal proton transfer steps have been recognized [5, 23]. In the first, the proton jumps between adjacent carbonyl oxygen atoms; this transfer has almost no barrier. The second step is proton rotation (isomerization) around the double bond of a single carbonyl group. The energy barriers of the latter process were found to be within the range of $16\text{--}23 \text{ kcal mol}^{-1}$. This proton isomerization is strongly influenced by the presence of a single water molecule [24]. In this case, the rate-determining step, which is still isomerization, has almost half the energy barrier (8 kcal mol^{-1}) of the previous pathway. It was recently found that the isomerization step can be bypassed if there is stronger cooperation between three carbonyl oxygen atoms in longer oligopeptides [23]. In the work presented here, we will focus on this possibility in more detail. Because this cooperation between oxygen atoms requires higher flexibility of the peptide chain, we will also examine how side chains influence the proton transfer mechanism.

In theory, a particularly long peptide chain should be used as a model to appropriately describe the proton transfer (Fig. 1). However, the flexibility of such a chain would lead to a very complicated potential energy surface and, of course, highly involved calculations. Therefore, we decided to use *N*-acetylglycyl-*N*¹-methylglycinamide (AGA) as the reference model for the peptide and all

possible stereoisomers of *N*-acetylalanyl-*N*¹-methylalanyl-amide (ALA) (Fig. 2). We assume that the chemical and structural properties of the middle amide group and the internal parts of the terminal amide groups are very similar to the properties of the peptide groups in the corresponding polypeptide. Therefore, we assume that our results are transferable to longer peptides.

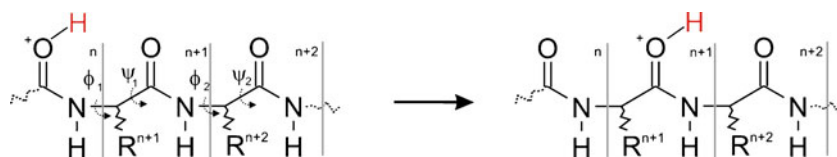
The results we have obtained show two different proton pathways along the peptide chain. The basic ideas of these two mechanisms were reported in a short communication [23]. In this paper, we will discuss them in more detail and, in particular, we will show how different configurations of C_α can influence them. As we were interested in the fastest proton transfer pathway, all possible stereoisomers were considered, not only the natural one (which is the (*S,S*)-configuration).

Computational details

All stationary points on the potential energy surface (PES) presented here were localized at the density functional theory (DFT) level using the hybrid B3LYP functional (with Becke's three-parameter exchange functional [25] and the correlation functional [26, 27] from Lee, Yang, and Par). The 6-31++G** basis set [28–30] with polarization and diffuse functions on both heavy and hydrogen atoms was used with this method. DFT calculations were performed with the Gaussian 98 (G98) molecular modeling package [31]. Minima and first-order transition states were found using the standard optimization technique implemented in G98. The optimization of the first-order transition states was initiated by explicitly calculating the Hessian at the HF/6-31++G** level of theory. The nature of each stationary point was determined by vibrational analysis using the same method and basis set. All of the minima presented in this study have all-real vibrational frequencies, and the first-order transition states have only one imaginary vibrational frequency (the values of the imaginary frequencies of the transition structures are summarized in Table S1 in the “Electronic supplementary information,” ESM).

Thermodynamic functions such as the enthalpy and Gibbs energy were calculated for 298.15 K and 101325 Pa. Zero point vibrational energies that were used in the calculations were not scaled. Although the studied system

Fig. 1 Proton transfer in a peptide chain



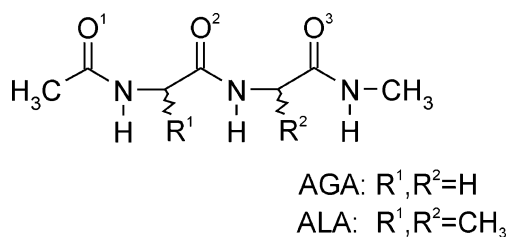


Fig. 2 Reference models for calculations

contains methyl groups, the calculated thermochemical energies were not corrected to these possible low energy barrier internal rotors [32], as we do not expect them to make a key contribution to the final results. The relative energies are differences between the energies of individual structures and a reference structure that has all energy quantities equal to zero.

Structures of the reference model AGA that were used for final calculations were found with the B3LYP functional using a smaller basis set: starting from 6-31G* to 6-31+G**. Employing a smaller basis set enabled the use of thinner integration grids for integral evaluation. This led to a significant increase in the speed of potential energy scan calculations, as described below. Trial structures for this level of calculation were found in all cases by proton coordinate driving. The driving was performed by a relaxed potential energy surface scan, as implemented in G98. This technique is based on incremental changes in the proton's internal coordinate along with the optimization of the remaining degrees of freedom. The distance between the proton and the oxygen atom, or the dihedral angle—which defines position around a carbonyl double bond—were usually chosen as the driven internal coordinate, with the step size ranging from 0.1 to 0.15 Å for distance and from 10 to 15° for dihedral angle. The structure with maximum energy was then considered to be the trial transition state and submitted to the full transition-state optimization, as described above. The structures with the minimum energy were considered to be trial minima and were fully optimized.

Structures of the model ALA that were used for final calculations were constructed from structures of the reference AGA model. Two hydrogen atoms on two C_α

atoms were replaced with methyl groups. At first, only the methyl groups were optimized when creating structures, but then the relaxation of the whole structure was performed at the finest level of theory described above.

From here on, **S1** and **S2** are the initial and final structures, respectively. **I_x** and **T_x** refer to the intermediate and transition-state structures, respectively, where **x** is an integer that distinguishes between the different structures. Common numbering is used for all mechanisms. Therefore, it is easy to identify the same intermediates on the various pathways found.

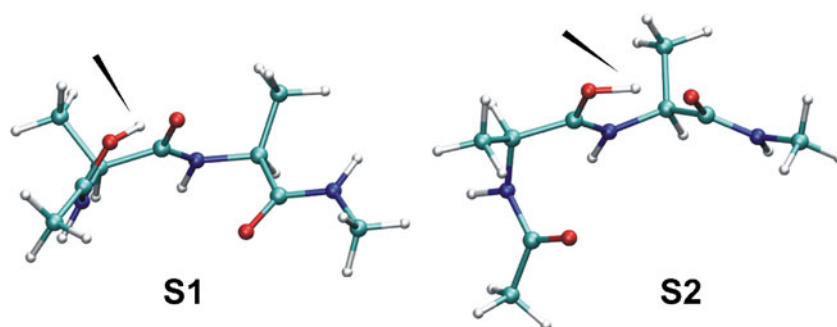
Results and discussion

In our previous study, we reported two possible mechanisms of proton transfer in the AGA peptide model. In the first one, the rate-determining step is proton isomerization around a carbonyl C=O double bond. This mechanism will henceforth be denoted as mechanism A. The second mechanism avoids the isomerization step and will henceforth be denoted mechanism B. Based the stationary points found on the potential energy surface, we will show the nature of each mechanism. We will also discuss how they are influenced by the presence of methyl groups attached to the C_α atoms of the peptide backbone (the ALA model).

Boundary structures

Intramolecular proton transfer between the two structures **S1** and **S2** has been studied for the AGA and ALA models. These structures are shown in Fig. 3 for the (*S,S*)-ALA model. In the initial state **S1**, the proton is bound to the first carbonyl oxygen atom by a regular covalent bond with an average bond length of 1.092 Å (calculated from all of the models studied here). For comparison, the O–H bond length in water and methanol molecules calculated to the same level of theory is 0.965 Å. This increase in bond length by about 0.13 Å is due to proton stabilization by the second carbonyl oxygen. The associated interaction can be

Fig. 3 Initial (**S1**) and final (**S2**) structures for proton transfer in the (*S,S*)-ALA model



considered to be a very strong hydrogen bond because of the very short distance between the interacting partners and the good orientation with the second carbonyl oxygen lone pair. On average, it is only about 1.349 Å long, whereas a regular hydrogen bond [33] is usually at least 2.2 Å in length. A similar interaction pattern was also observed for the final structure **S2**, where the proton is bound to the second carbonyl oxygen and stabilized by the third one. Various energetic and geometric parameters describing the structures **S1** and **S2** are summarized in Table 1. It is worth noting that **S1** and **S2** are not the most stable structures; there is probably a more energetically stable structure than **S1**. In this structure, the oxygen atom O3 would interact with the hydrogen atoms of the second amide. A similar situation applies to **S2**. In this case, the O1 oxygen would interact with the second amide's hydrogen atom. These other (probably more stable) structures were not considered in our study because the oxygen atoms are not pre-organized in an orientation that is suitable for vectorial proton transfer.

Before we proceed any further with our description of the proton transfer pathways, let us first discuss the thermodynamics of the whole process. In a long peptide chain, the initial and final states should have almost the same energy, as the environment around the unit that binds the proton is similar in both states. However, this is not the case in the models that we are studying.

The calculated reaction free energies are not even close to zero (Table 1); they vary from -3.81 to 1.87 kcal mol $^{-1}$. The main reason for this is the stabilizing effect of the remaining carbonyl oxygen. In structure **S1**, this carbonyl oxygen is located on the third carbonyl group. It interacts with the carbon of the second carbonyl group, whose oxygen only stabilizes the proton. The opposite situation exists in structure **S2**. The first carbonyl oxygen interacts with the carbon of the second carbonyl group, whose oxygen is directly bound to the transferred proton. This observed imperfection in our models could be improved by

using longer models with extended and more uniform stabilization of the unit bearing the proton. However, the structures presented in our work can serve as the basis for such prospective studies.

Due to symmetry constraints, the reaction energies for proton transfer in the (*R,R*)-ALA and (*S,S*)-ALA models have to be the same. The same condition applies to the (*R,S*)-ALA and (*S,R*)-ALA models. However, the calculated reaction energies (Table 1) contradict this constraint, which means that the initial and final structures for the (*R,R*)/(*S,S*) pair and the (*R,S*)/(*S,R*) pair presented here are terminal structures for different proton transfer pathways.

Mechanism A

In this section, proton transfer will be described in the **S1** → **S2** direction. At the beginning, the proton is situated between the O1 (for numbering, see Fig. 2) and O2 oxygen atoms (structure **S1**). The proton transfer starts with the proton jumping between these two oxygen atoms (**S1** → **T1** → **I1**). The situation in structure **I1** is very similar to that in **S1**, but here the proton is bonded to the second oxygen and stabilized by the first oxygen. The difference in geometry between **S1** and **S2** is small, and the energy barrier for the proton jump is smaller than 0.2 kcal mol $^{-1}$ (Table 2). We only found this step for the reference AGA model and the (*R,R*) and (*R,S*) configurations of the ALA model. For the (*S,R*) and (*S,S*) configurations, our attempts to localize either **I1** or **T1** were not successful. Therefore, it appears that the proton can occupy a broad space in this region, which consists of either two shallow minima separated by very small barrier or one very flat minimum. Thus, minima **S1** and **I1** should be considered the initial states of proton transfer, because the error associated with the computational method used is far larger than the barriers found. This is also reflected in the calculated free energies, which show that the free energy minimum actually corresponds to the transition structure **T1**.

Table 1 Comparison of the reaction electronic (ΔE_r) and free (ΔG_r) energies of proton transfer and selected geometrical parameters of the initial (**S1**) and final (**S2**) proton transfer states

	ΔE_r	ΔG_r	S1			S2		
			d_1	d_2	d_3	d_4	d_5	d_6
AGA	0.63	0.94	1.092	1.352	3.035	1.080	1.360	2.850
(<i>R,R</i>)-ALA	0.01	0.36	1.102	1.335	3.033	1.103	1.330	3.008
(<i>R,S</i>)-ALA	-2.91	-3.81	1.089	1.359	2.870	1.086	1.355	3.030
(<i>S,R</i>)-ALA	1.72	1.87	1.092	1.345	3.055	1.135	1.282	2.789
(<i>S,S</i>)-ALA	-1.11	-1.08	1.086	1.356	2.884	1.099	1.331	2.813

All energies are in kcal mol $^{-1}$; distances are in Å.

d_1 distance $\text{O}^1\text{-H}$; d_2 distance between $\text{O}^1\text{-H}$ and O^2 ; d_3 distance between $\text{C}(=\text{O}^2)$ and O^3 ; d_4 distance $\text{O}^2\text{-H}$; d_5 distance between $\text{O}^2\text{-H}$ and O^3 ; d_6 distance between O^1 and $\text{C}(=\text{O}^2)$

Table 2 Comparison of the relative electronic (ΔE) and free (ΔG) energies of mechanism A

	ΔE					ΔG				
	AGA	(<i>R,R</i>)-ALA	(<i>R,S</i>)-ALA	(<i>S,R</i>)-ALA	(<i>S,S</i>)-ALA	AGA	(<i>R,R</i>)-ALA	(<i>R,S</i>)-ALA	(<i>S,R</i>)-ALA	(<i>S,S</i>)-ALA
S1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T1	0.20	0.07	0.13	^a	^a	-0.93	-1.29	-1.29		
I1	0.19	-0.04	0.05	0.00	0.00	-0.38	-0.74	-0.49	0.00	0.00
T2	17.92	18.87	17.76	18.28	17.08	17.73	19.03	17.20	18.26	16.49
S2	0.63	0.01	-2.91	1.72	-1.11	0.94	0.36	-3.81	1.87	-1.08

All energies are in kcal mol⁻¹

^a Transition state was not found

The proton transfer proceeds with the rotation of the proton around the second carbonyl double bond, yielding structure **S2** (**I1**→**T2**→**S2**). In transition state **T2**, the proton is located almost perpendicular to the plane of the corresponding amide group (Fig. 4). Thus, the proton is transferred along one amide unit of the triamide chain in two steps (**S1**→**T1**→**I1** and **I1**→**T2**→**S2**). Proton transfer via this mechanism in a similar system has already been reported in the literature [5].

The rate-determining step of mechanism A is clearly the second step, which includes proton isomerization, and has an activation energy of 17–19 kcal mol⁻¹ (Table 2). In the transition state **T2**, the transferred proton is stabilized by two hydrogen bonds, whereas it is stabilized by only one in the initial structure **I2**. From this point of view, we would expect the barrier to be lower than that for a simpler system without such extra stabilization. However, the energy barrier to proton isomerization in protonated formamide [24] is only 10.5 kcal mol⁻¹ (calculated at the B3LYP/6-31+G** level). The explanation for this conflicting observation is likely the different stabilizations of the proton in both states. The stabilizing hydrogen bond in the initial state **I2** is very short (about 1.3 Å; see Table 1 for the similar structure **S1**). In the transition state, hydrogen bonds are much longer (Table 3). The shift towards the initial state is about 0.7 Å or even 1.2 Å for the second hydrogen bond. The angle between the donor,

proton, and acceptor in these hydrogen bonds is also sharper (about 120 and 100°, respectively).

The barrier height is also influenced by the presence of methyl groups and their different stereoconfigurations. The greatest increase in activation energy was observed for the (*R,R*) configuration, which has a barrier that is about 0.95 kcal mol⁻¹ higher than that for the AGA model. In contrast, the greatest decrease in activation energy towards the AGA model (about -0.84 kcal mol⁻¹) was found for the (*S,S*) configuration. In summary, increased energy barriers were observed for the configurations (*R,R*) and (*S,R*), whereas decreased energy barriers were observed for the (*R,S*) and (*S,S*) configurations.

However, the pathways presented here represent only a subset of all possible proton transfers that might occur in the systems studied. Since the (*R,R*)/(*S,S*) configurations are enantiomers, there must be two pathways—one for each configuration—that have the same energy profile but different pathway geometries; indeed, these pathway geometries must be mirror images of each other. Thus, the pathway found for (*R,R*) configuration might occur for the (*S,S*) configuration with the same energy profile, but their pathway geometries would be mirror images. This is why we can conclude that the presence of methyl groups always leads to a lower barrier to the isomerization step for any stereoisomer than the barrier in the AGA model, even though

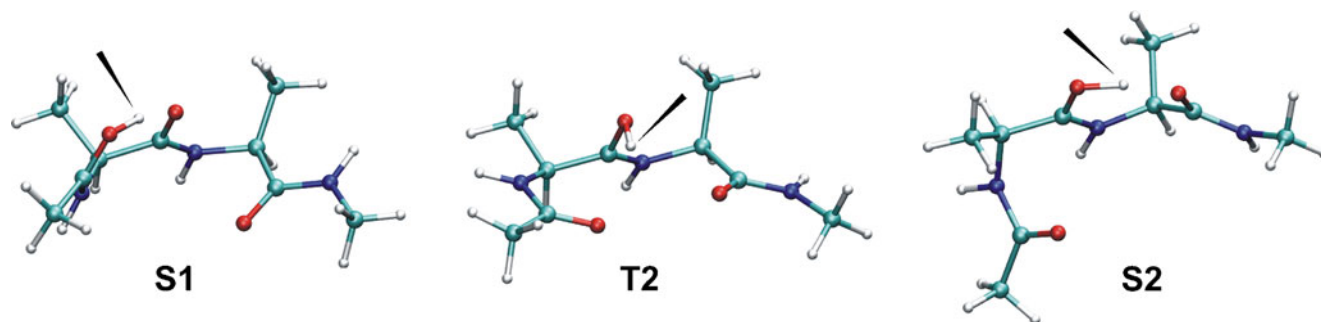


Fig. 4 Proton transfer in the protonated (*S,S*)-ALA triamide (mechanism A). *Black arrow* indicates the transferred proton

Table 3 Selected geometrical parameters of the transition structure **T2**

	AGA	(<i>R,R</i>)-ALA	(<i>R,S</i>)-ALA	(<i>S,R</i>)-ALA	(<i>S,S</i>)-ALA
d_1	0.982	0.984	0.984	0.983	0.983
d_2	2.126	2.063	2.038	2.073	2.052
d_3	2.544	2.546	2.463	2.614	2.542
τ	88.4	89.3	85.3	87.0	82.8

d_1 distance $\text{O}^2\text{-H}$; d_2 distance between $\text{O}^2\text{-H}$ and O^1 ; d_3 distance between $\text{O}^2\text{-H}$ and O^3 ; τ dihedral angle $\text{C-C}=\text{O}^2\text{-H}$. All distances are in Å. All angles are in degrees

the pathways for the (*S,R*) and (*R,R*) stereoisomers shown in Table 2 exhibit higher barriers.

Mechanism B

In the previous section, it was shown that the rate-determining step of mechanism A is proton isomerization around the carbonyl double bond. This indicates that every step in which the proton must shift from the plane of amide group will be disadvantaged. In the following section, an alternative scenario for proton transfer is presented. The pathway found benefits from cooperation among all of the oxygen atoms, which keeps the transferred proton closer to the planes of the amide groups.

The proton transfer is initiated from structure **S1** with the rotation of the first amide group bearing the proton through dihedral angle ϕ (Fig. 5, Table 4). As result of this step, the proton is stabilized by the third oxygen atom instead of the second (**S1**→**T3**→**I2**). In the next step, the proton jumps from the first to the third oxygen (**I2**→**T4**→**I3**). The proton transfer then proceeds with a similar reorganization to that in the first step of this mechanism, but this time the change in conformation is achieved with rotation through the dihedral angle ψ_2 . During this step, the stabilization of the proton with the first oxygen is replaced with its stabilization with the second oxygen (**I3**→**T5**→**I4**). After this step, the proton is situated on the opposite side of the second amide group. In the last step, the proton jumps between the third and second oxygen atoms (**I4**→**T6**→**S2**).

The second conformational change (**I3**→**T5**→**I4**) is the rate-determining step of mechanism B. It has a barrier of 8–10 kcal mol⁻¹, which nearly 50% lower than that for mechanism A (17–19 kcal mol⁻¹). A similar barrier was found for the first conformational change (**S1**→**T3**→**I2**), which is in the range 6–8 kcal mol⁻¹. During these two conformational changes, the proton stays very close to the amide plane; the out-of-plane deviation does not exceed 20° (Table 5). Thus, it is likely that the observed barrier is due to weaker stabilization of the proton with the adjacent oxygen in the transition states **T5** and **T3** than in the corresponding minima. Indeed, the average elongations of this stabilizing hydrogen bond are 0.51 and 0.45 Å for the steps **I3**→**T5** and

S1→**T3**, respectively, which correlate with the observed energy barriers.

Exceptional deviation of the proton from the amide plane was only observed for the transition state **T4** (Table 5). The average deviation is about 34°, which is half of the deviation required by mechanism A (90°). The barrier to this proton jump step is in the range 0.4–3.0 kcal mol⁻¹, depending on the direction. This is significantly lower than that for the isomerization step from mechanism A (17–19 kcal mol⁻¹) and those for the previously discussed conformational changes **I3**→**T5**→**I4** and **S1**→**T3**→**I2**. On the other hand, this barrier is higher than the barriers to proton jumps between adjacent oxygen atoms such as those that occur in mechanism A (**S1**→**T1**→**I1**) and mechanism B (**I4**→**T6**→**S2**). The final structure of the step including the transition state **T4** is the intermediate **I3**, which is the most stable structure along the proton transfer pathway. However, if the free energy is taken into account, the structure corresponding to the global minimum along the proton path shifts to **T4**. The proton pathway then consists of the two main steps **I3**→**T5**→**T4** and **T4**→**T5**→**S2**. However, it is important to note that such a conclusion is dependent on the reliability of the method used for the free-energy calculation. This may not be precise enough to provide very accurate results. According to a deeper analysis, the most significant differences between the electronic and free energies are due to zero point energy (ZPE) corrections. In minima **I2** and **I3**, the vibrations of the proton–oxygen bond and the corresponding stabilizing hydrogen bond (e.g., d_1 and d_2 in Table 6) are included in the ZPE, whereas these vibrations become part of the proton jump in transition state **T4** (corresponding to a single imaginary vibration). Thus, they are excluded from the ZPE for structure **T4**. This causes a dramatic drop in the ZPE for structure **T4** towards the minima **I2** and **I3**. Since the harmonic approximation is used in the evaluation of vibration modes, this effect may be overestimated. A similar conclusion applies to the proton jump **S1**→**T1**→**I1** found in mechanism A.

As discussed before, the proton transfer is mostly influenced by two conformational changes, so introducing methyl substituents onto the C_α atoms in ALA model may have a greater impact on the proton transfer than observed in mechanism A. This hypothesis is actually confirmed by the root-mean-square deviations of the rate-determining barriers of all possible stereoisomers of ALA towards the AGA model, which are 0.73 and 0.98 kcal mol⁻¹ for mechanisms A and B, respectively. Moreover, methyl groups always increase the barrier in mechanism B.

Conclusions

We have presented a theoretical study of proton transfer among carbonyl oxygen atoms in tripeptide models. Two

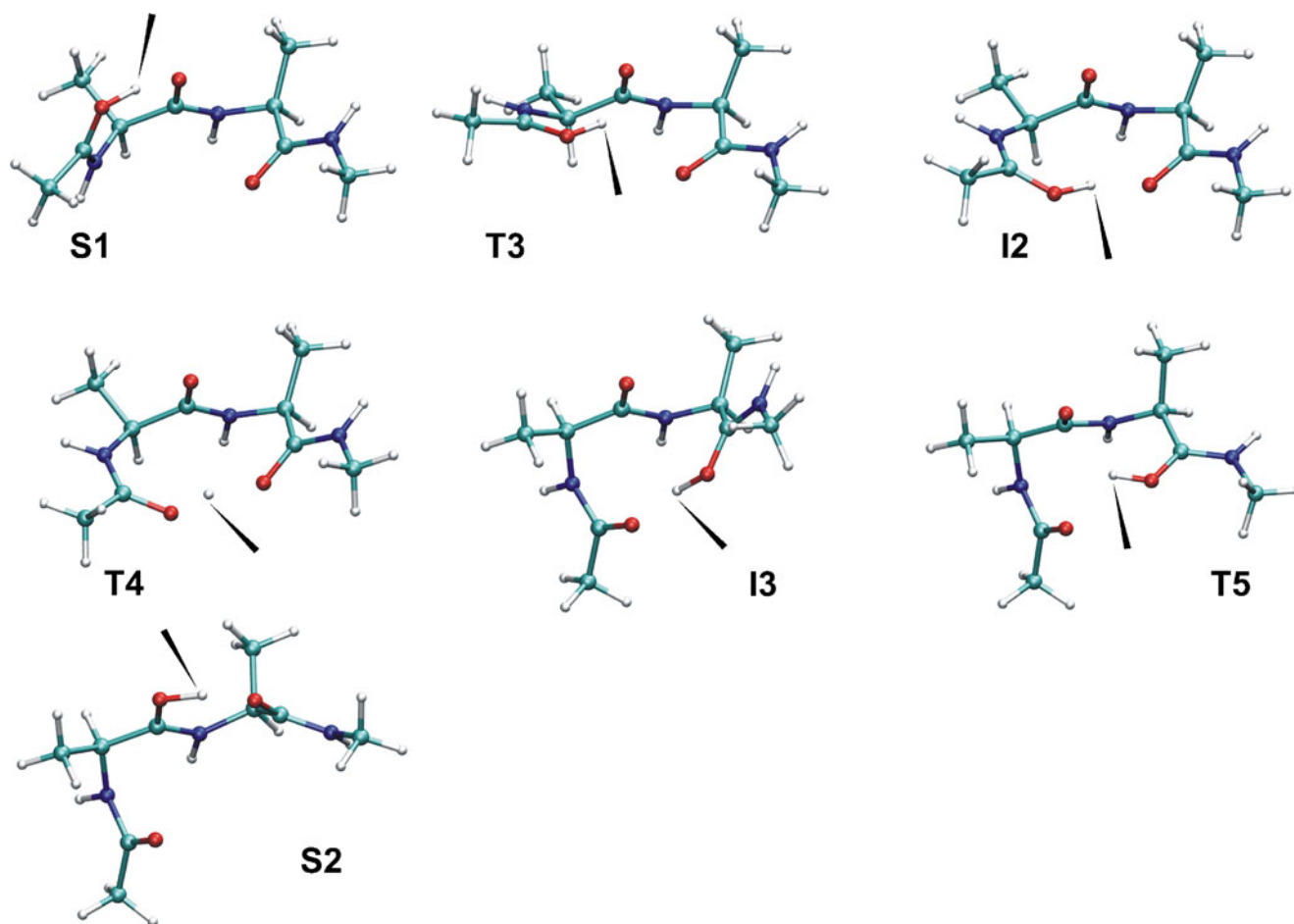


Fig. 5 Proton transfer in protonated (*S,S*)-ALA triamide (mechanism B). *Black arrow* indicates the transferred proton. (Structures **I4** and **T6** are not shown because they were not found for the (*S,S*)-ALA model)

principal proton transfer pathways were found. The first involves isomerization of the proton around the double

bond of the carbonyl group. This is the rate-determining step, with a barrier ranging from 17 to 19 kcal mol⁻¹.

Table 4 Comparison of the relative electronic (ΔE) and free (ΔG) energies for mechanism B

	ΔE					ΔG				
	AGA	(<i>R,R</i>)-ALA	(<i>R,S</i>)-ALA	(<i>S,R</i>)-ALA	(<i>S,S</i>)-ALA	AGA	(<i>R,R</i>)-ALA	(<i>R,S</i>)-ALA	(<i>S,R</i>)-ALA	(<i>S,S</i>)-ALA
S1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T3	6.08	7.69	6.99	6.64	5.95	8.34	9.65	7.92	8.47	6.79
I2	-0.21	1.72	1.16	-0.32	-0.87	1.59	3.42	2.17	1.33	0.00
T4	0.19	1.98	1.28	-0.11	-0.71	-0.28	3.58	2.36	1.46	0.39
I3	-2.44	-3.23	-4.39	-2.46	-3.62	-0.01	-1.22	-2.31	0.39	-1.47
T5	5.88	6.06	5.19	6.53	5.63	8.46	8.79	6.69	9.59	7.38
I4	0.34	0.00	^a	^a	^a	-0.65	0.47	^a	^a	^a
T6	0.39	0.09	^b	^b	^b	-0.69	-0.39	^b	^b	^b
S2	0.63	0.01	-2.91	1.72	-1.11	0.94	0.36	-3.81	1.87	-1.08

All energies are in kcal mol⁻¹

^a Same as **S2**

^b Does not exist

Table 5 Deviation of the proton from the amide plane during mechanism B

	ω					δ				
	AGA	(<i>R,R</i>)-ALA	(<i>R,S</i>)-ALA	(<i>S,R</i>)-ALA	(<i>S,S</i>)-ALA	AGA	(<i>R,R</i>)-ALA	(<i>R,S</i>)-ALA	(<i>S,R</i>)-ALA	(<i>S,S</i>)-ALA
S1	160.6	158.9	159.0	163.1	163.2	19.4	21.1	21.0	16.9	16.8
T3	-157.7	-159.7	-159.2	-158.3	-158.4	22.4	20.3	20.9	21.7	21.6
I2	-164.8	-165.4	-163.5	-164.1	-163.1	15.2	14.7	16.5	16.0	16.9
T4	-134.8	-148.8	-150.7	-146.7	-150.4	45.2	31.2	29.3	33.3	29.6
I3	-11.8	-14.7	-15.6	-16.1	-17.1	11.8	14.7	15.6	16.1	17.1
T5	0.9	4.1	-1.1	5.0	0.3	0.9	4.1	1.1	5.0	0.3
I4	7.8	10.1	^a	^a	^a	7.8	10.1	^a	^a	^a
T6	158.3	157.5	^b	^b	^b	21.7	22.5	^b	^b	^b
S2	164.7	160.8	165.4	162.5	167.0	15.3	19.2	14.6	17.5	13.0

ω Dihedral angle $\underline{C-C=O^x-H}$; δ deviation of the proton out of the amide group plane. All angles are in degrees

^a Same as **S2**

^b Does not exist

This barrier is about 8 kcal mol⁻¹ higher than that in protonated formamide, which is likely due to more efficient proton stabilization in the local minima than that in the transition state. The barrier is also affected by the presence of methyl groups attached to the C_α atoms. These decrease the barrier by about 0.16 and 0.84 kcal mol⁻¹ for the enantiomeric couples (*R,S*)/(*S,R*) and (*S,S*)/(*R,R*), respectively.

An alternative pathway to the mechanism including an isomerization step was also found. This pathway eliminates the isomerization step using a series of conformational changes where all three oxygen atoms cooperate. The barrier to the rate-determining step of this second mecha-

nism is in the range of only 8–10 kcal mol⁻¹, which is nearly half of the corresponding barrier in the previous case. The rate-determining step and another step with a similar barrier mostly depend on the flexibility of the peptide chain. It was shown that introducing methyl substituents into the ALA model, regardless of the stereo-configuration, increases the barrier of the rate-determining step. The increases are about 0.67 and 0.93 kcal mol⁻¹ for the enantiomeric pairs (*R,S*)/(*S,R*) and (*S,S*)/(*R,R*), respectively. Thus, we can conclude that, even with the introduction of methyl groups, the second mechanism is still the most favorable, and that differences between the stereoisomers are small.

Table 6 Selected geometric parameters of the protonated models in mechanism B

	d_1					d_2				
	AGA	(<i>R,R</i>)-ALA	(<i>R,S</i>)-ALA	(<i>S,R</i>)-ALA	(<i>S,S</i>)-ALA	AGA	(<i>R,R</i>)-ALA	(<i>R,S</i>)-ALA	(<i>S,R</i>)-ALA	(<i>S,S</i>)-ALA
S1	1.092	1.102	1.089	1.092	1.086	1.352	1.335	1.359	1.345	1.356
T3	0.991	0.994	0.995	0.995	0.998	1.871	1.798	1.786	1.790	1.750
I2	1.035	1.038	1.040	1.035	1.036	1.501	1.483	1.475	1.494	1.488
T4	1.150	1.092	1.089	1.087	1.077	1.262	1.352	1.355	1.364	1.382
I3	1.523	1.522	1.529	1.513	1.522	1.026	1.026	1.024	1.026	1.024
T5	2.049	2.081	2.004	2.060	1.986	0.985	0.984	0.986	0.985	0.987
I4	1.340	1.346	^a	^a	^a	1.100	1.096	^a	^a	^a
T6	1.214	1.206	^b	^b	^b	1.191	1.199	^b	^b	^b
S2	1.080	1.103	1.086	1.130	1.099	1.360	1.330	1.355	1.282	1.331

d_1 distance of the proton from the closest carbonyl oxygen; d_2 length of the stabilizing hydrogen bond. All distances are in Å

^a Same as **S2**

^b Does not exist

Acknowledgments The access to the MetaCentrum supercomputing facilities provided under the research intent MSM6383917201 is appreciated. This work was supported by the Ministry of Education of the Czech Republic, under contracts MSM0021622413 and LC06030 (J.K.). The research leading to these results also received funding from the European Community's Seventh Framework Programme under grant agreement no. 205872 (P.K.).

References

- Pelmenschikov V, Blomberg M, Siegbahn P (2002) A theoretical study of the mechanism for peptide hydrolysis by thermolysin. *J Biol Inorg Chem* 7:284–298
- Rodriguez C, Cunje A, Shoeib T, Chu I, Hopkinson A, Siu K (2000) Solvent-assisted rearrangements between tautomers of protonated peptides. *J Phys Chem A* 104:5023–5028
- Rodriguez C, Cunje A, Shoeib T, Chu I, Hopkinson A, Siu K (2001) Proton migration and tautomerism in protonated triglycine. *J Am Chem Soc* 123:3006–3012
- Paizs B, Suhai S (2001) Theoretical study of the main fragmentation pathways for protonated glycyglycine. *Rapid Commun Mass Spectrom* 15:651–663
- Paizs B, Csonka I, Lendvay G, Suhai S (2001) Proton mobility in protonated glycyglycine and N-formylglycyglycinamide: a combined quantum chemical and RRKM study. *Rapid Commun Mass Spectrom* 15:637–650
- Smith R, Loo J, Barinaga C, Edmonds C, Udseth H (1990) Collisional activation and collision-activated dissociation of large multiply charged polypeptides and proteins produced by electrospray ionization. *J Am Soc Mass Spectrom* 1:53–65
- Campbell S, Rodgers M, Marzluff E, Beauchamp J (1995) Deuterium exchange reactions as a probe of biomolecule structure. Fundamental studies of gas phase H/D exchange reactions of protonated glycine oligomers with D₂O, CD₃OD, CD₃CO₂D, and ND₃. *J Am Chem Soc* 117:12840–12854
- Cassady C, Carr S, Zhang K, Chungphillips A (1995) Experimental and ab-initio studies on protonations of alanine and small peptides of alanine and glycine. *J Org Chem* 60:1704–1712
- Chaudhuri C, Jiang J, Wu C, Wang X, Chang H (2001) Characterization of protonated formamide-containing clusters by infrared spectroscopy and ab initio calculations. II. Hydration of formamide in the gas phase. *J Phys Chem A* 105:8906–8915
- Csonka I, Paizs B, Lendvay G, Suhai S (2000) Proton mobility in protonated peptides: a joint molecular orbital and RRKM study. *Rapid Commun Mass Spectrom* 14:417–431
- Martin R (2001) In: Sigel A (ed) Probing of proteins by metal ions and their low-molecular-weight complexes (Metal Ions in Biological Systems, vol 38). CRC Press, Boca Raton, pp 1–23
- MacDonald B, Thachuk M (2008) Gas-phase proton-transfer pathways in protonated histidylglycine. *Rapid Commun Mass Spectrom* 22:2946–2954
- Fu H, Fu A (2007) Theoretical study on the reaction mechanism of proton transfer in alaninamide. *J Mol Struct THEOCHEM* 818:163–170
- Richard J, Amyes T (2001) Proton transfer at carbon. *Curr Opin Chem Biol* 5:626–633
- Hur O, Niks D, Casino P, Dunn M (2002) Proton transfers in the beta-reaction catalyzed by tryptophan synthase. *Biochemistry* 41:9991–10001
- Tomashek J, Brusilow W (2000) Stoichiometry of energy coupling by proton-translocating ATPases: a history of variability. *J Bioenerg Biomembr* 32:493–500
- Senior A (1990) The proton-translocating atpase of *Escherichia coli*. *Annu Rev Biophys Bioeng* 19:7–41
- Green MK, Lebrilla CB (1997) Ion-molecule reactions as probes of gas-phase structures of peptides and proteins. *Mass Spectrom Rev* 16:53–71
- Papayannopoulos I (1995) The interpretation of collision-induced dissociation tandem mass-spectra of peptides. *Mass Spectrom Rev* 14:49–73
- Barber M, Bordoli R, Sedgwick R, Tyler A (1981) Fast atom bombardment of solids (fab): a new ion-source for mass-spectrometry. *J Chem Soc Chem Commun* 325–327
- Hillenkamp F, Karas M, Beavis R, Chait B (1991) Matrix-assisted laser desorption ionization mass-spectrometry of biopolymers. *Anal Chem* 63:A1193–A1202
- Fenn J, Mann M, Meng C, Wong S, Whitehouse C (1990) Electrospray ionization: principles and practice. *Mass Spectrom Rev* 9:37–70
- Kulhanek P, Schlag E, Koca J (2003) A novel mechanism of proton transfer in protonated peptides. *J Am Chem Soc* 125:13678–13679
- Kulhanek P, Schlag E, Koca J (2003) Mechanism of proton transfer in short protonated oligopeptides 1N-methylacetamide and N-2-acetyl-N-1-methylglycinamide. *J Phys Chem A* 107:5789–5797
- Becke A (1993) Density-functional thermochemistry. 3. The role of exact exchange. *J Chem Phys* 98:5648–5652
- Lee C, Yang W, Parr R (1988) Development of the Colle–Salvetti correlation-energy formula into a functional of the electron-density. *Phys Rev B* 37:785–789
- Miehlich B, Savin A, Stoll H, Preuss H (1989) Results obtained with the correlation-energy density functionals of Becke and Lee, Yang and Parr. *Chem Phys Lett* 157:200–206
- Hehre W, Ditchfie R, Pople J (1972) Self-consistent molecular-orbital methods. 12. Further extensions of Gaussian-type basis sets for use in molecular-orbital studies of organic molecules. *J Chem Phys* 56:2257–2261
- Harihara P, Pople J (1973) Influence of polarization functions on molecular-orbital hydrogenation energies. *Theor Chim Acta* 28:213–222
- Clark T, Chandrasekhar J, Spitznagel G, Pvr S (1983) Efficient diffuse function-augmented basis-sets for anion calculations. 3. The 3-21+G basis set for 1st-row elements, Li-F. *J Comput Chem* 4:294–301
- Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA Jr, Stratmann RE, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN, Strain MC, Farkas O, Tomasi J, Barone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Ochterski J, Petersson GA, Ayala PY, Cui Q, Morokuma K, Salvador P, Dannenberg JJ, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Baboul AG, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Andres JL, Gonzalez C, Head-Gordon M, Replogle ES, Pople JA (1998) Gaussian 98, revision A.9. Gaussian Inc., Pittsburgh
- Ayala P, Schlegel H (1998) Identification and treatment of internal rotation in normal mode vibrational analysis. *J Chem Phys* 108:2314–2325
- Desiraju G, Steiner T (1999) The weak hydrogen bond in structural chemistry and biology. Oxford University Press, Oxford

NMR and NQR parameters of the SiC-doped on the (4,4) *armchair* single-walled BPNT: a computational study

Mohammad T. Baei · S. Zahra Sayyad-Alangi ·
Ali Varasteh Moradi · Parviz Torabi

Received: 17 March 2011 / Accepted: 15 May 2011 / Published online: 28 May 2011
© Springer-Verlag 2011

Abstract The structural properties, NMR and NQR parameters in the pristine and silicon carbide (SiC) doped boron phosphide nanotubes (BPNTs) were calculated using DFT methods (BLYP, B3LYP/6-31G*) in order to evaluate the influence of SiC-doped on the (4,4) *armchair* BPNTs. Nuclear magnetic resonance (NMR) parameters including isotropic (CS^I) and anisotropic (CS^A) chemical shielding parameters for the sites of various ^{13}C , ^{29}Si , ^{11}B , and ^{31}P atoms and quadrupole coupling constant (C_Q), and asymmetry parameter (η_Q) at the sites of various ^{11}B nuclei were calculated in pristine and SiC-doped (4,4) *armchair* boron phosphide nanotubes models. The calculations indicated that doping of ^{11}B and ^{31}P atoms by C and Si atoms had a more significant influence on the calculated NMR and NQR parameters than did doping of the B and P atoms by Si and C atoms. In comparison with the pristine model, the SiC-doping in Si_pC_B model of the (4,4) *armchair* BPNTs reduces the energy gaps of the nanotubes and increases their electrical conductance. The NMR results showed that the B and P atoms which are directly bonded to the C atoms in the SiC-doped BPNTs have significant changes in the

NMR parameters with respect to the B and P atoms which are directly bonded to the Si atoms in the SiC-doped BPNTs. The NQR results showed that in BPNTs, the B atoms at the edges of nanotubes play dominant roles in determining the electronic behaviors of BPNTs. Also, the NMR and NQR results detect that the Fig. 1b (Si_pC_B) model is a more reactive material than the pristine and the Fig. 1a (Si_BC_p) models of the (4,4) *armchair* BPNTs.

Keywords Boron phosphide nanotubes · NMR · NQR · Silicon carbide

Introduction

Since the synthesis of carbon nanotubes (CNTs) by Ijima in 1991 [1], single-walled carbon nanotubes (SWCNTs) have attracted great interest owing to their physical and chemical properties [1–3] and applications as novel materials [4, 5]. The electronic properties of CNTs depend on their tubular diameter and chirality. Many investigations have been undertaken to investigate non-carbon based nanotubes, which exhibit electronic properties independent of these features. Among these, boron nitride nanotubes (BNNTs) and boron phosphide nanotubes (BPNTs), which are made from the group III and V elements neighboring C in the Periodic Table, are an interesting subject of many studies [6–10]. Boron phosphide nanotubes (BPNTs) are inorganic proportion of carbon nanotubes (CNTs) and have good physical properties for a broad variety of applications [11]. However, the properties of BNNTs have been studied more often than those of BPNTs [12, 13], further study of the electronic properties of BPNTs remains interesting.

Nuclear magnetic resonance (NMR) [13, 14] and nuclear quadrupole resonance (NQR) [15] spectroscopy are the best

M. T. Baei (✉) · S. Z. Sayyad-Alangi
Department of Chemistry, Azadshahr Branch,
Islamic Azad University,
Azadshahr, Golestan, Iran
e-mail: baei52@yahoo.com

A. V. Moradi
Department of Chemistry, Gorgan Branch,
Islamic Azad University,
Gorgan, Iran

P. Torabi
Department of Chemistry, Mahshahr Branch,
Islamic Azad University,
Mahshahr, Iran

techniques to study the electronic structure properties of matters. There is the known similarity between the properties of the electronic structures of BP and silicon carbide (SiC) nanotubes [16]. Moreover, doping of BPNTs by Si and C atoms may be able to yield changes in the interactions between the nanotube and foreign atoms or molecules. Therefore, the objective of the present work is to study the properties of the electronic structure of SiC-doped BPNTs by performing density functional theory (DFT) calculations of the NMR and NQR parameters of representative (4,4) *armchair* BPNT models (Fig. 1). The electronic structure properties, including bond lengths, bond angles, tip diameters, dipole moments (μ), energies, band gaps, NMR, and NQR parameters in both pristine and the SiC-doped BPNT structures, are investigated by calculations of the chemical shielding (CS) tensors including isotropic and anisotropic chemical shielding parameters at the sites of various ^{13}C , ^{29}Si , ^{11}B , and ^{31}P atoms and NQR calculations in sites of ^{11}B atom.

Computational methods

In the present work, the electronic structure properties of BPNTs were studied by using representative models of (4,4) *armchair* BPNTs in which the ends of nanotubes were saturated by hydrogen atoms. Each of the representative models has three forms (Fig. 1), namely the pristine model (Fig. 1c) and models where B and P atoms are doped by Si and C atoms, respectively ($\text{Si}_\text{B}\text{C}_\text{P}$, Fig. 1a), or B and P atoms are doped by C and Si atoms, respectively ($\text{Si}_\text{P}\text{C}_\text{B}$, Fig. 1b). We investigated the influence of the SiC-doping on the properties of the (4,4) *armchair* single-walled BPNT. The hydrogenated models of (4,4) *armchair* single-walled BPNTs and the SiC-doped of BPNT have 72 atom with formulas of $\text{B}_{28}\text{P}_{28}\text{H}_{16}$ and $\text{SiCB}_{27}\text{P}_{27}\text{H}_{16}$, respectively. In the first step, the structures were allowed to relax by all atomic geometrical parameters in the optimization at the DFT levels of B3LYP and BLYP exchange-functional and 6-31G* standard basis set. Then, the CS tensors were calculated in the optimized structures by using B3LYP and BLYP/6-31G* for the sites of various ^{13}C , ^{29}Si , ^{11}B , and ^{31}P atoms and NQR parameters of ^{11}B . It is noted that, in DFT methods, B3LYP is more popular due to its more reliable results in comparison with experiments [17, 18] and in a previous study, it has been found that the NMR parameters calculated by B3LYP and B3PW91 levels are in good agreement [17]. Also, in DFT methods, the computations based on the BLYP functional could yield reliable results for the properties of the electronic structure of nanotubes [6]. Therefore, all of the calculations were studied in both of B3LYP and BLYP levels. The calculated CS tensors in the principal axis system (PAS) with the order

of $\sigma_{33} > \sigma_{22} > \sigma_{11}$ [19] were converted into measurable NMR parameters [isotropic chemical shielding CS (CS^I) and anisotropic chemical shielding CS (CS^A) parameters] using Eqs. 1 and 2 [20] and the NMR parameters of ^{13}C , ^{29}Si , ^{11}B , and ^{31}P atoms for the investigated models of the (4,4) *armchair* single-walled BPNTs are summarized in Table 2.

$$\text{CS}^I(\text{ppm}) = 1/3(\sigma_{11} + \sigma_{22} + \sigma_{33}) \quad (1)$$

$$\text{CS}^A(\text{ppm}) = \sigma_{33} - 1/2(\sigma_{11} + \sigma_{22}) \quad (2)$$

For NQR parameters, Computational calculations do not directly detect experimentally measurable NQR parameters, nuclear quadrupole coupling constant (C_Q), and asymmetry parameter (η_Q). Therefore, Eqs. 3 and 4 are used to calculate EFG tensors to their proportional experimental measurable parameters; C_Q is the interaction energy of nuclear electric quadrupole moment (eQ) with the electric field gradient (EFG) tensors at the sites of quadrupole nuclei but asymmetry parameter (η_Q) is a quantity of the EFG tensors, deviation from tubular symmetry at the sites of quadrupole nuclei. The nuclei that have $I > 1/2$ (I =nuclear spin angular momentum) are active in NQR spectroscopy. The calculated EFG tensor eigenvalues in the principal axis system (PAS) with the order of $|q_{zz}| > |q_{yy}| > |q_{xx}|$ were converted into measurable NQR parameters (nuclear quadrupole coupling constant (C_Q), and asymmetry parameter (η_Q)) using Eqs. 3 and 4. The standard Q values ($Q(^{11}\text{B}) = 40.59$ mb) reported by Pyykkö [21] are used in Eq. 3. The NQR parameters of ^{11}B nuclei for the investigated models of the (4,4) *armchair* single-walled BPNTs are summarized in Table 3. All calculations were carried out by using the Gaussian 03 suite of programs [22].

$$C_Q(\text{MHz}) = e^2Qq_{zz}h^{-1} \quad (3)$$

$$\eta_Q = \left| \frac{(q_{xx} - q_{yy})}{q_{zz}} \right| \quad 0 < \eta_Q < 1 \quad (4)$$

Results and discussion

Structures of the (4,4) *armchair* BPNTs

The structural properties consisting of the B–P bond lengths, bond angles, tip diameters, dipole moments (μ), energies and band gaps for the investigated models of the (4,4) *armchair* BPNTs are summarized in Table 1. There are two forms of triple SiC-doped BPNTs for the (4,4) *armchair* model, where the B and P atoms are doped by Si and C atoms (Fig. 1a, $\text{Si}_\text{B}\text{C}_\text{P}$) or the B and P atoms are doped by C and Si atoms (Fig. 1b, $\text{Si}_\text{P}\text{C}_\text{B}$). There are B–P,

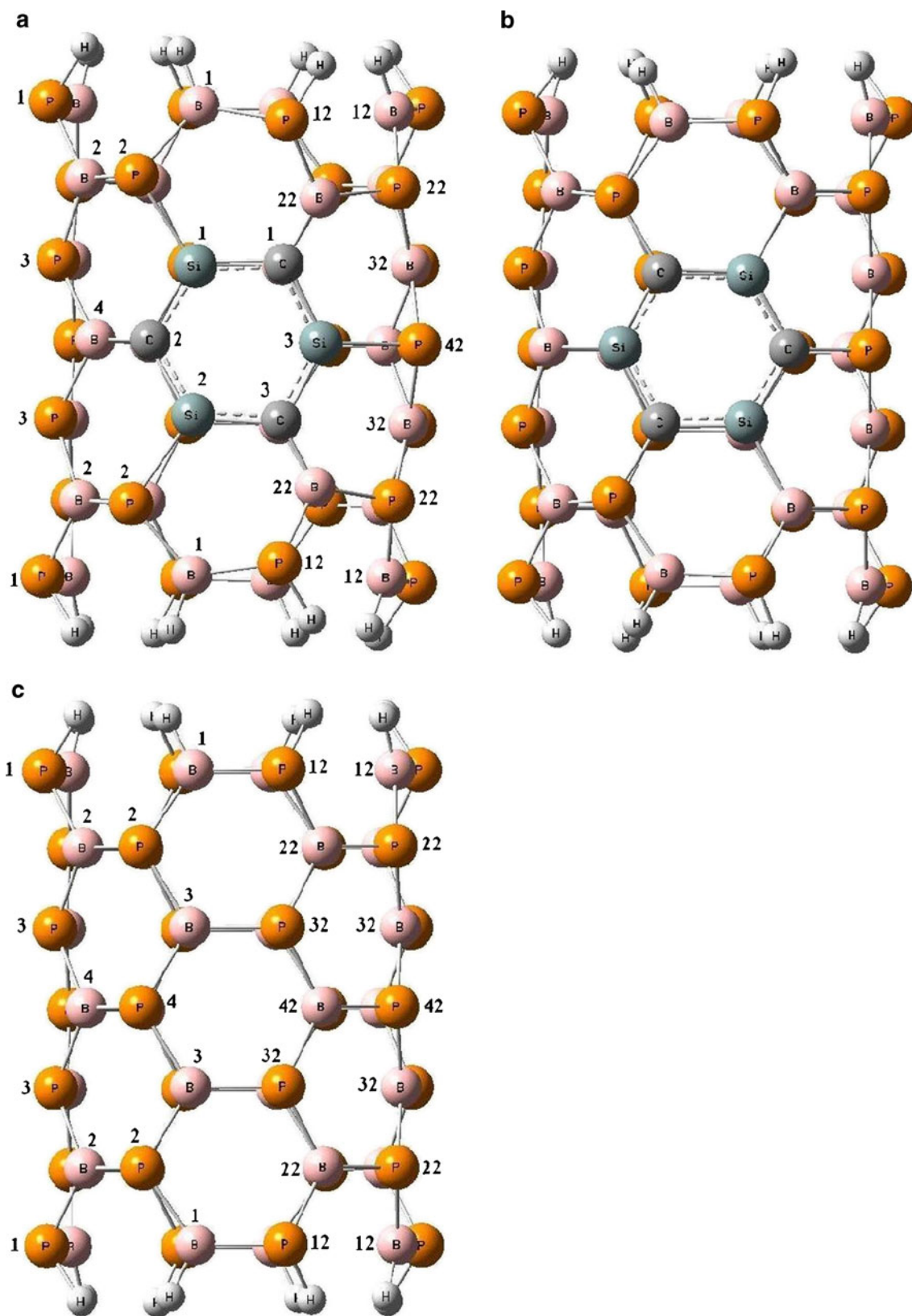


Fig. 1 (a), (b) Two-dimensional (2D) views of triple Si_BC_P and Si_PC_B doped (4,4) *armchair* BPNTs, (c) 2D views of pristine (4,4) *armchair* single-walled BPNTs

Table 1 Structural properties of representative (4,4) *armchair* BPNT models with BLYP and B3LYP/6-31G* methods

Property	Si _B C _P (Fig. 1a)			Si _P C _B (Fig. 1b)			Pristine (4,4) BPNT		
	BLYP	B3LYP	BLYP-B3LYP	BLYP	B3LYP	BLYP-B3LYP	BLYP	B3LYP	BLYP-B3LYP
Bond length (Å)									
B-H	1.198	1.191	0.007	1.197	1.191	0.006	1.197	1.190	0.007
P-H	1.425	1.413	0.012	1.425	1.414	0.011	1.425	1.412	0.013
B ₁ -P ₁₂	1.884	1.874	0.010	1.892	1.885	0.007	1.892	1.881	0.011
B ₁ -P ₂	1.910	1.900	0.010	1.876	1.866	0.010	1.887	1.879	0.008
B ₁₂ -P ₂₂	1.886	1.876	0.010	1.888	1.876	0.012	1.887	1.879	0.008
B ₂ -P ₁	1.906	1.896	0.010	1.915	1.905	0.010	1.911	1.889	0.022
B ₂ -P ₂	1.928	1.915	0.013	1.900	1.888	0.012	1.901	1.891	0.010
B ₂ -P ₃	1.907	1.896	0.011	1.912	1.904	0.008	1.908	1.897	0.011
B ₂₂ -P ₁₂	1.937	1.927	0.010	1.887	1.876	0.011	1.911	1.889	0.022
B ₂₂ -P ₂₂	1.948	1.936	0.012	1.883	1.869	0.014	1.901	1.891	0.010
B ₂₂ -P ₃₂	–	–	–	–	–	–	1.908	1.897	0.011
B ₃ -P ₂	–	–	–	–	–	–	1.911	1.901	0.010
B ₃ -P ₄	–	–	–	–	–	–	1.905	1.894	0.011
B ₃ -P ₃₂	–	–	–	–	–	–	1.895	1.882	0.013
B ₃₂ -P ₂₂	1.910	1.899	0.011	1.912	1.903	0.009	1.911	1.901	0.010
B ₃₂ -P ₄₂	1.926	1.915	0.011	1.906	1.896	0.010	1.905	1.894	0.011
B ₄ -P ₃	1.956	1.945	0.011	1.887	1.865	0.022	1.902	1.892	0.010
B ₄ -P ₄	–	–	–	–	–	–	1.897	1.885	0.012
B ₄₂ -P ₄₂	–	–	–	–	–	–	1.897	1.885	0.012
B ₄₂ -P ₃₂	–	–	–	–	–	–	1.902	1.892	0.010
P ₂ -Si ₁	2.217	2.199	0.018	–	–	–	–	–	–
P ₄₂ -Si ₂	2.225	2.206	0.019	–	–	–	–	–	–
B ₂₂ -C ₁	1.539	1.532	0.007	–	–	–	–	–	–
B ₄ -C ₂	1.516	1.508	0.008	–	–	–	–	–	–
Si ₁ -C ₁	1.801	1.784	0.017	1.810	1.793	0.017	–	–	–
Si ₁ -C ₂	1.801	1.785	0.016	1.814	1.795	0.019	–	–	–
Si ₂ -C ₁	1.811	1.794	0.017	1.815	1.794	0.021	–	–	–
P ₂ -C ₁	–	–	–	1.821	1.809	0.012	–	–	–
P ₄₂ -C ₂	–	–	–	1.832	1.825	0.007	–	–	–
B ₂₂ -Si ₁	–	–	–	1.979	1.977	0.002	–	–	–
B ₄ -Si ₂	–	–	–	1.964	1.961	0.003	–	–	–
Average B-P	1.918	1.907	0.011	1.896	1.885	0.011	1.902	1.890	0.012
Average Si-C	1.804	1.788	0.016	1.813	1.794	0.019	–	–	–
Bond angles (°)									
P ₁ -B ₂ -P ₂	122.301	122.157	0.144	126.064	126.034	0.030	124.511	124.264	0.247
P ₂ -B ₁ -P ₁₂	121.149	121.188	0.039	124.866	125.230	0.364	123.069	123.356	0.287
P ₁₂ -B ₂₂ -P ₂₂	116.872	116.687	0.185	129.236	129.233	0.003	124.511	124.264	0.247
P ₃ -B ₂ -P ₂	120.925	120.875	0.050	120.389	120.349	0.040	121.134	121.261	0.127
P ₂ -Si ₁ -C ₁	121.206	121.314	0.108	–	–	–	–	–	–
P ₂ -C ₁ -Si ₁	–	–	–	119.808	120.098	0.290	–	–	–
Si ₁ -C ₁ -B ₂₂	120.328	120.300	0.028	–	–	–	–	–	–
C ₁ -Si ₁ -B ₂₂	–	–	–	119.191	118.748	0.443	–	–	–
B ₄ -C ₂ -Si ₁	120.167	119.980	0.187	–	–	–	–	–	–
B ₄ -Si ₂ -C ₁	–	–	–	114.504	114.857	0.353	–	–	–
C ₂ -Si ₁ -C ₁	121.926	121.932	0.006	–	–	–	–	–	–
Si ₁ -C ₁ -Si ₂	113.305	113.406	0.101	119.756	118.649	1.102	–	–	–
C ₁ -Si ₂ -C ₁	124.200	124.124	0.076	118.556	119.345	0.789	–	–	–

Table 1 (continued)

Property	Si _B C _P (Fig.1a)			Si _P C _B (Fig.1b)			Pristine (4,4) BPNT		
	BLYP	B3LYP	BLYP-B3LYP	BLYP	B3LYP	BLYP-B3LYP	BLYP	B3LYP	BLYP-B3LYP
Si ₁ -C ₂ -Si ₁	115.449	115.584	0.135	119.645	118.800	0.845	–	–	–
Diameter tip (Å)	7.817	7.755	0.062	7.974	7.916	0.058	7.905	7.833	0.072
μ (Debye)	1.341	1.465	0.124	0.542	0.611	0.069	0.000	0.000	0.000
Energy (keV)	-276.118	-276.146	0.028	-276.116	-276.143	0.027	-279.272	-279.300	0.028
Band gaps (eV)	1.83	2.99	1.160	1.29	2.31	1.02	1.75	2.95	1.200

B–C, Si–P, and Si–C bonds in the Fig. 1a (Si_BC_P) and there are B–P, B–Si, C–P, and Si–C bonds in the Fig. 1b (Si_PC_B). We have optimized investigated models of the (4,4) *armchair* BPNTs employing the BLYP and B3LYP/6-31G* computational levels. We have also compared the BLYP results with the results of B3LYP study on the BPNT models. The average B–P bond length in pristine the (4,4) BPNT (Fig. 1c) were 1.902 and 1.890 Å in BLYP and B3LYP levels, but this value was changed in the triple SiC-doped BPNTs models in Fig. 1a and b, due to the influence of the triple SiC-doped on the BPNTs. In the models, the average B–P bond length for the Fig. 1a (Si_BC_P) were 1.918 and 1.907 Å and for the Fig. 1b (Si_PC_B) were 1.896 and 1.885 Å in BLYP and B3LYP levels. In Fig. 1, the atoms of the BPNTs are numbered in order to describe the relevant structural parameters. In Table 1, bond lengths distances and bond angles and properties of the electronic structure are listed for the BPNTs. The calculated results showed that values of the B–P bond lengths have slightly different in the investigated models of the (4,4) *armchair* BPNTs. In the Fig. 1a, the value of bond length of the P₄₂-Si₂ is the largest whereas that of the B₂₂-C₂ one is the smallest among different types of the bonds in the investigated models. The value of the Si–C bond lengths are almost same in the investigated models of the (4,4) *armchair* BPNTs. The bond angles showed slightly difference in comparison to the pristine model, The bond angle of P₁₂-B₂₂-P₂₂ undergo changes from 124° for the pristine model to 116° and 129° for the Fig. 1a (Si_BC_P) and the Fig. 1b (Si_PC_B) yielding some structural deformations.

Furthermore, in the Fig. 1a (Si_BC_P) model, it should be noted that B and P atoms slightly relax inwardly while in the Fig. 1b (Si_PC_B) model, B and P atoms relax outwardly of the nanotube surface yielding different diameters of 7.817 and 7.755 Å for the Fig. 1a (Si_BC_P) mouth and 7.974 and 7.916 Å for the Fig. 1b (Si_PC_B) mouth, whereas in the pristine model, the diameters are 7.905 and 7.833 Å in BLYP and B3LYP levels, respectively. It must be noted that the significant changes of geometries are just for those

atoms placed in the nearest neighborhood of the triple SiC-doped BPNTs and those of other atoms remained almost unchanged. The calculated energies and the values are almost the same for the two forms a and b of the SiC-doped BPNTs. However, the band gaps showed differences between the two forms (Fig. 1a and b). In comparison with the pristine model, the band gap of the Fig. 1a is closer to the pristine model, whereas Fig. 1b is significantly reduced in the (4,4) *armchair* BPNTs and increases their electrical conductance. These results showed that the doping of B and P atoms by C and Si atoms (Fig. 1b, Si_PC_B) has more influence on the band gap of the BPNTs than does doping of the B and P atoms by Si and C atoms (Fig. 1a, Si_BC_P). The values of dipole moments (μ) of the SiC-doped BPNTs structures (Fig. 1a and b) detect notable changes with respect to the pristine model. Also, the values of dipole moments (μ), for Fig. 1a are more than Fig. 1b. It is important to note that the point charges are balanced in the pristine (4,4) *armchair* BPNTs but these conditions are corrupted in the SiC-doped BPNTs structures. We have also compared the BLYP results with the results of B3LYP on the pristine and the SiC-doped models of BPNTs (see Table 1). The calculated results showed that the changes of the B–P bond lengths and tip diameters were almost negligible, whereas there are slight changes in bond angles and dipole moments (μ) in the two levels. The comparison of the optimized energies showed that the calculated energies values with B3LYP method are more than the BLYP method. Also there are the most significant changes between the B3LYP and BLYP levels in band gaps. The values of band gap energies for (Fig. 1a, Si_BC_P) and (Fig. 1b, Si_PC_B) in B3LYP were increased about 1.16 eV and 1.02 eV with respect to BLYP level. Mirzaei investigated the electronic structure properties of the (4,4) *armchair* BPNT just in BLYP level [16]. Our calculations results are very close to theirs. To our knowledge, B3LYP study on the electronic structure properties of BPNT surfaces has not been reported. Therefore, all of the calculations were studied in both of B3LYP and BLYP levels. An interesting conclusion that can be drawn from

these pathways is that the obtained calculated results of the B3LYP in BPNTs have significantly different band gap energies with respect to the BLYP level, unlike BNNTs, the results of the BLYP are very similar to those of the B3LYP [13, 23].

NMR parameters of the (4,4) *armchair* BPNTs

The NMR parameters for the investigated models of the (4,4) *armchair* BPNTs are summarized in Table 2. In the pristine model of the (4,4) *armchair* BPNTs, there are 28 B and 28 P atoms in the considered model and the NMR parameters are separated into four layers (Table 2 and Fig. 1c). In the model, the values of NMR parameters in each of the groups were the same, however, results of the Table 2 shows that the calculated NMR parameters are not similar for different groups which means that the CS parameters for the atoms of each layer have equivalent chemical environment and electrostatic properties. In the first layer, B1 to B14 have almost the smallest values of the CS^I parameters but the largest values of the CS^A parameters among the B atoms in the pristine model of the (4,4) *armchair* BPNTs. In second layer, B2 to B24, values of the CS^I parameters of the layer almost are equal to the first layer but values of the CS^A parameters of the layer are smaller than the first layer. In third and fourth layers, B3 to B34 and B4 to B44, values of the CS^I parameters of the layers are larger than two previous B groups but values of the CS^A parameters of the layers are decreased.

The P atom has a lone pair of electrons in the valance shell, therefore there are differences between the properties of the electronic structures of B and P atoms. In first layer, P1 to P14 have largest values of the CS^I parameters but the smallest values of the CS^A parameters among the P atoms in the pristine model of the (4,4) *armchair* BPNTs, unlike the NMR parameters of B atoms. In second layer, P2 to P24, values of the CS^I parameters of the layer are smaller than the first layer but values of the CS^A parameters of the layer are significantly increased. In third and fourth layers, P3 to P34 and P4 to P44, values of the CS^I and CS^A parameters of the layers are smaller than two previous P groups. The changes of the values of CS^I and CS^A parameters for P atoms are important just from the first group to the second one. In Fig. 1a ($Si_B C_P$), the B3 and B42 atoms are doped by Si atoms and P32 and P4 are doped by C atoms, which results in B–C, Si–P, and Si–C bonds. The calculated results in Table 2 show that among the B atoms of Fig. 1a ($Si_B C_P$), B4 and B22 are directly bonded to C atoms; hence, both CS^I and CS^A parameters show important changes due to the SiC-doping. However CS^I parameters for B1, B2, and B32 atoms that indirectly bonded to C atoms and for other B atoms show some changes due to the SiC-

doping, but changes of the CS^A values of the B atoms are almost negligible except for B2 that CS^A value for the atom show some changes due to the SiC-doping. Among the P atoms of Fig. 1a ($Si_B C_P$) in comparison with the pristine model, P2 and P42 are directly bonded to Si atoms, the greatest changes in the NMR parameters are observed for P2 and P42 atoms, and both the CS^I and CS^A parameters show significant changes because of the contribution to the chemical bonding with the Si atoms. Except for the change in the CS^I parameters for the P22 and P3 atoms, the changes in the CS^I and CS^A parameters are not very important for the other P atoms, which are indirectly bonded to the SiC-doped (4,4) *armchair* BPNTs in Fig. 1a ($Si_B C_P$).

In Fig. 1b ($Si_P C_B$), the P4 and P32 atoms are doped by Si atoms and the B3 and B42 atoms are doped by C atoms on the (4,4) *armchair* BPNTs, which yield B–Si, C–P, and Si–C bonds. Among the B atoms, the most important changes in both NMR parameters (CS^I and CS^A) are observed for the B4 and B22 atoms, which are directly bonded to Si atoms and the atoms show some changes due to the SiC-doping. The CS^I and CS^A parameters of other B atoms which are indirectly bonded to the SiC-doped (4,4) *armchair* BPNTs in Fig. 1b ($Si_P C_B$) do not exhibit any significant changes due to the SiC-doping. In Fig. 1b ($Si_P C_B$), the P2 and P42 atoms are directly bonded to C atoms; hence, both CS^I and CS^A parameters of the atoms show significant changes due to the SiC-doping. However, NMR parameters for P12, P22, and P3 atoms, which indirectly bonded to Si and C atoms, exhibit some significant changes due to the SiC-doping.

The values of the NMR parameters (CS^I and CS^A) of the ^{13}C and ^{29}Si atoms in the SiC-doped (4,4) *armchair* BPNTs are summarized in Table 2. The results in Table 2 show that the values of the CS^I and CS^A parameters of the ^{13}C and ^{29}Si atoms in Fig. 1b ($Si_P C_B$) are larger than those in Fig. 1a ($Si_B C_P$) in the (4,4) *armchair* BPNTs except that the CS^I parameter of atom C1 and the CS^A parameter of atom Si1 in Fig. 1a ($Si_B C_P$) are larger than in Fig. 1b ($Si_P C_B$). Comparison of the calculated NMR parameters in Fig. 1a and b shows that the properties of the electronic structure of the Fig. 1b ($Si_P C_B$) of the SiC-doped (4,4) *armchair* BPNT, where the B atoms are doped by C atoms and the P atoms are doped by Si atoms ($Si_P C_B$) are more influenced than those of Fig. 1a ($Si_B C_P$), where the B atoms are doped by Si atoms and the P atoms are doped by C atoms ($Si_B C_P$). This trend is in agreement with the change in the band gap of Fig. 1b ($Si_P C_B$) in comparison with the pristine model of the (4,4) *armchair*. The band gaps of Fig. 1a ($Si_B C_P$) and the pristine model are almost the same, but the band gaps of Fig. 1b ($Si_P C_B$) are smaller than those of the pristine model (Table 1). Also, the results of Table 2 show that there are some significant differences in NMR parameters between

Table 2 NMR parameters (ppm) of representative (4,4) *armchair* BPNT models in the sites of various ^{11}B , ^{31}P , ^{13}C , and ^{29}Si

nucleus	$\text{Si}_\text{B}\text{C}_\text{P}$ (Fig. 1a)				$\text{Si}_\text{P}\text{C}_\text{B}$ (Fig. 1b)				Pristine (4,4) BPNT			
	BLYP		B3LYP		BLYP		B3LYP		BLYP		B3LYP	
	CS^I	CS^A	CS^I	CS^A	CS^I	CS^A	CS^I	CS^A	CS^I	CS^A	CS^I	CS^A
B ₁	33.6	128.2	33.8	138.0	40.9	128.2	41.5	132.3	36.5	128.1	36.3	136.1
B ₁₂	35.2	129.7	35.2	137.5	35.7	127.8	36.7	131.1	36.5	128.1	36.3	136.1
B ₁₃	36.0	129.1	35.8	137.3	38.3	129.6	37.1	136.4	36.5	128.1	36.3	136.1
B ₁₄	36.6	128.2	36.7	135.9	35.0	129.8	35.2	136.6	36.5	128.1	36.3	136.1
B ₂	30.0	118.6	30.0	130.0	36.1	111.1	36.1	121.0	35.5	111.5	35.6	121.5
B ₂₂	45.9	90.3	47.1	97.7	38.4	120.8	39.2	138.5	35.5	111.5	35.6	121.5
B ₂₃	35.9	113.8	35.5	123.7	35.5	112.6	34.3	121.2	35.5	111.5	35.6	121.5
B ₂₄	35.6	112.4	35.6	122.2	36.8	112.3	36.2	121.3	35.5	111.5	35.6	121.5
B ₃	–	–	–	–	–	–	–	–	39.7	98.1	39.6	109.3
B ₃₂	31.2	100.4	31.8	112.7	36.7	100.3	37.8	101.6	39.7	98.1	39.6	109.3
B ₃₃	40.0	97.6	40.0	108.2	40.3	105.1	40.5	110.0	39.7	98.1	39.6	109.3
B ₃₄	37.0	97.8	37.2	108.5	40.2	105.2	40.5	110.6	39.7	98.1	39.6	109.3
B ₄	54.7	67.9	56.5	76.5	46.5	122.9	47.7	123.1	43.6	100.6	43.8	111.4
B ₄₂	–	–	–	–	–	–	–	–	43.6	100.6	43.8	111.4
B ₄₃	42.7	94.7	42.6	106.0	42.5	108.6	42.3	112.2	43.6	100.6	43.8	111.4
B ₄₄	43.8	101.5	43.9	112.6	44.9	106.0	45.3	110.7	43.6	100.6	43.8	111.4
P ₁	383.0	156.4	410.7	146.7	387.0	146.4	411.4	138.4	387.2	150.0	414.8	141.7
P ₁₂	389.0	159.4	416.4	154.6	384.5	155.6	407.0	164.6	387.2	150.0	414.8	141.7
P ₁₃	385.3	158.8	413.6	148.8	389.1	150.0	415.3	143.0	387.2	150.0	414.8	141.7
P ₁₄	389.1	146.4	416.3	138.4	390.9	146.5	416.0	138.4	387.2	150.0	414.8	141.7
P ₂	398.3	215.1	436.4	204.3	320.8	140.6	339.7	142.3	329.6	260.0	362.5	247.1
P ₂₂	338.1	262.8	369.7	246.4	310.0	280.2	328.8	286.2	329.6	260.0	362.5	247.1
P ₂₃	333.5	256.6	364.5	244.3	345.3	248.2	360.3	242.6	329.6	260.0	362.5	247.1
P ₂₄	327.1	262.3	358.6	248.1	346.8	250.8	362.5	245.1	329.6	260.0	362.5	247.1
P ₃	334.8	249.6	366.3	239.8	320.8	363.8	339.2	367.5	324.4	256.3	355.4	243.1
P ₃₂	–	–	–	–	–	–	–	–	324.4	256.3	355.4	243.1
P ₃₃	319.7	258.4	351.7	246.4	335.8	245.1	354.6	230.4	324.4	256.3	355.4	243.1
P ₃₄	322.8	260.7	354.9	245.3	332.8	255.1	352.7	246.3	324.4	256.3	355.4	243.1
P ₄	–	–	–	–	–	–	–	–	325.7	257.1	358.9	245.0
P ₄₂	400.8	174.1	436.8	166.0	324.2	191.2	342.6	202.9	325.7	257.1	358.9	245.0
P ₄₃	328.1	252.9	360.2	241.3	325.2	235.4	356.2	245.9	325.7	257.1	358.9	245.0
P ₄₄	326.0	249.5	359.6	238.1	327.8	246.8	357.9	253.3	325.7	257.1	358.9	245.0
Si ₁	226.1	224.0	231.8	217.7	290.4	169.6	296.5	171.2	–	–	–	–
Si ₂	248.0	133.4	249.4	140.5	285.4	170.7	290.7	169.6	–	–	–	–
C ₁	77.9	134.6	88.7	140.5	75.2	168.6	72.1	163.9	–	–	–	–
C ₂	72.8	119.4	84.7	124.4	80.5	162.0	91.3	160.9	–	–	–	–

the BLYP and B3LYP levels. An interesting conclusion that can be drawn from these pathways is that the obtained calculated results for B and P atoms which are directly bonded to C atoms in the SiC-doped BPNTs have significant changes in the NMR parameters with respect to B and P atoms which are directly bonded to Si atoms in the SiC-doped BPNTs.

^{11}B electric field gradient tensors of the (4,4) *armchair* models

The NQR parameters at the sites of various ^{11}B nuclei for the optimized investigated models of the (4,4) *armchair* BPNT are summarized in Table 3. There are 28 B atom in the considered models of the (4,4) *armchair* and the NQR

Table 3 NQR parameters of representative (4,4) *armchair* BPNT models in the sites of various ^{11}B

nucleus	$\text{Si}_\text{B}\text{C}_\text{P}$ (Fig. 1a)				$\text{Si}_\text{P}\text{C}_\text{B}$ (Fig. 1b)				Pristine (4,4) BPNT			
	BLYP		B3LYP		BLYP		B3LYP		BLYP		B3LYP	
	C_Q (MHz)	η_Q	C_Q (MHz)	η_Q	C_Q (MHz)	η_Q	C_Q (MHz)	η_Q	C_Q (MHz)	η_Q	C_Q (MHz)	η_Q
B_{12}	3.56	0.38	3.71	0.38	3.55	0.36	3.67	0.37	3.56	0.36	3.71	0.37
B_2	3.19	0.41	3.41	0.36	3.15	0.11	3.41	0.14	3.21	0.33	3.41	0.28
B_{32}	2.87	0.12	3.13	0.08	2.84	0.24	3.08	0.22	3.07	0.14	3.29	0.11
B_4	2.92	0.02	3.14	0.05	3.43	0.19	3.79	0.23	3.24	0.13	3.45	0.10

parameters are separated into four layers based on the likenesses of the calculated EFG tensors in each layer. The results of Table 3 show that the calculated NQR parameters are not similar for various nuclei, therefore, the electrostatic environment of BPNT is not equivalent in length in both the nanotube models. In Fig. 1, B_{12} atom shows the position of the first layer, B_2 shows the position t of the second layer, B_{32} shows that of the third layer, and B_4 shows the position t of the fourth layer in the considered armchair models. The B_{12} -layer is placed at the end of the tubes and includes both B and P atoms. In the (4,4) *armchair* models, values of $C_Q(^{11}\text{B}_{12})$ is the largest among other ^{11}B nuclei (see Table 3) that shows more orientation of EFG tensor eigenvalues along the z-axis of electronic distribution at the sites of $^{11}\text{B}_{12}$ nuclei and electrostatic environment of B_{12} is stronger rather than the other layers in the length of the tube. Other searches showed that the nanotubes grow from their ends; therefore the properties of the end nuclei in nanotubes are important in their growth and synthesis [6, 24]. Therefore in the BPNTs the B atoms placed at the edge of the (4,4) *armchair* nanotubes play important roles in determining the electronic behavior of the (4,4) *armchair* BPNTs, because the geometrical properties of this layer are different from the other layers. The B_2 -layer and B_{32} -layer are placed at the second and third layers in the considered models of the (4,4) *armchair* BPNTs. The values of $C_Q(^{11}\text{B})$ significantly reduced (see Table 3) in the models. In the first layer of atoms in the nanotubes, the B-P bond distances are almost 1.88 Å, but in the second and third layers the B-P bond distances are larger than the first layer. Therefore, the significant different between NQR parameters in the first layer and the other layers due to the change of the geometrical parameters. The B_4 -layer is placed at the fourth layer in the considered models of the (4,4) *armchair* BPNTs and the values of $C_Q(^{11}\text{B})$ significantly increased in the models. Comparison of the calculated $C_Q(^{11}\text{B})$ and η_Q parameters in the considered models of the (4,4) *armchair* BPNTs shows that the values of $C_Q(^{11}\text{B})$ and η_Q of the first layer are almost the same and have similar effects for the two SiC doping processes.

Also, in the second and third layers, B_2 and B_{32} atoms, the electronic sites of the B atoms of the layers ($C_Q(^{11}\text{B})$), for both of SiC-doped models shows similar effects, but η_Q of the $\text{Si}_\text{P}\text{C}_\text{B}$ -doped model shows stronger effects than for the $\text{Si}_\text{B}\text{C}_\text{P}$ -doped model. In the fourth layer, B_4 , the values of $C_Q(^{11}\text{B})$ and η_Q of the B atoms of the $\text{Si}_\text{P}\text{C}_\text{B}$ -doped model shows greater changes than for the $\text{Si}_\text{B}\text{C}_\text{P}$ model. Therefore, the electronic sites of the B atoms in the $\text{Si}_\text{P}\text{C}_\text{B}$ -doped model of the (4,4) *armchair* BPNT shows more changes than for the $\text{Si}_\text{B}\text{C}_\text{P}$ -doped model. In agreement with the results of the structural properties of the $\text{Si}_\text{P}\text{C}_\text{B}$ -doped model. We have also compared the BLYP results with the results of B3LYP, the results of Table 3 show that there are some significantly different in NQR parameters between the levels, values of NQR parameters of B3LYP level are greater than BLYP level.

Conclusions

We studied the electronic structure properties including bond lengths, bond angles, tip diameters, dipole moments (μ), energies, band gaps, the NMR and NQR parameters of the pristine and the silicon-carbide (SiC) doped boron phosphide nanotubes (BPNTs) by mean of DFT calculations. On the basis of our calculations, the values of the B-P bond lengths and bond angles were changed in the triple SiC-doped models due to the influence of the triple SiC-doped on the BPNTs with respect to the pristine model. The values of the Si-C bond lengths are almost the same in the SiC-doped models. The values of dipole moments (μ) of the SiC-doped BPNTs structures (Fig. 1a and b) detect notable changes with respect to the pristine model. Also, the values of dipole moments (μ), for Fig. 1a are more than Fig. 1b. In comparison with the pristine model, the band gap of Fig. 1a is closer to the pristine model, whereas the band gap Fig. 1b is significantly reduced with respect to pristine in the (4,4) *armchair* BPNTs and increases its electrical conductance. This results showed that the doping of the B and P atoms by C and Si atoms (Fig. 1b, $\text{Si}_\text{P}\text{C}_\text{B}$) had more significant influence on the band gap of the BPNT than did doping of the B and P atoms by Si and C atoms (Fig. 1a,

Si_BC_P). Also there are the most significant changes between the B3LYP and BLYP levels in band gaps. The NMR parameters for the pristine model are separated into four layers and the NMR values for the ¹¹B and ³¹P atoms, which are directly bonded to Si and C atoms in the SiC-doped models, are significantly changed. Comparison of the calculated NMR parameters in Fig. 1a and b shows that the properties of the electronic structure of Fig. 1b (Si_PC_B) of the SiC-doped (4,4) *armchair* BPNT are more influenced than those of Fig. 1a (Si_BC_P). The values of NQR of the first layers belonging to those B atoms placed at the edges of the BPNT nanotubes were stronger rather than the other layers in the length of the tube, shows the dominant role of B atoms in determining the electronic behavior of BPNT. The electronic sites of the B atoms in the Si_PC_B-doped model of the (4,4) *armchair* BPNT shows more changes than for the Si_BC_P-doped model. Finally, the NMR and NQR results detect that the Fig. 1b (Si_PC_B) model is a more reactive material than the pristine and the Fig. 1a (Si_BC_P) models of the (4,4) *armchair* BPNTs. Also, values of NQR parameters of B3LYP level are greater than BLYP level.

References

- Ijima S (1991) Nature 354:56–58
- Derycke V, Martel R, Appenzeller J, Avouris PH (2002) Appl Phys Lett 80:2773–2775
- Liu C, Fan YY, Liu M, Cong HT, Cheng HM, Dresselhaus MS (1999) Science 286:1127–1129
- Zurek B, Autschbach J (2004) J Am Chem Soc 126:13079–13088
- Nojeh A, Lakatos GW, Peng S, Cho K, Pease RFW (2003) Nano Lett 3:1187–1190
- Hou S, Shen Z, Zhang J, Zhao X, Xue Z (2004) Chem Phys Lett 393:179–183
- Zhang M, Su ZM, Yan LK, Qiu YQ, Chen GH, Wang RS (2005) Chem Phys Lett 408:145–149
- Erkoc S (2001) J Mol Struct THEOCHEM 542:89–93
- Ferreira VA, Leite Alves HW (2008) J Cryst Growth 310:3973
- Mirzaei M, Gihai M (2010) Physica E 42:1667
- Golberg D, Bando Y, Tang CC, Zhi CY (2007) Adv Mater 19:2413–2432
- Mirzaei M (2009) Z Phys Chem 223:815
- Mirzaei M (2009) Physica E 41:883–885
- Bovey FA (1988) Nuclear magnetic resonance spectroscopy. Academic, SanDiego
- Das TP, Han EL (1958) Nuclear quadrupole resonance spectroscopy. Academic, New York
- Mirzaei M (2011) J Mol Model 17:89–96
- Mirzaei M, Hadipour NL (2006) J Phys Chem A 110:4833
- Mothana B, Ban F, Boyd RJ (2005) Chem Phys Lett 401:7
- Drago RS (1992) Physical methods for chemists, 2nd edn. Saunders College Publishing, Florida
- Mirzaei M, Seif A, Hadipour NL (2008) Chem Phys Lett 461:246–248
- Pykkö P (2001) Mol Phys 99:1617–1629
- Frisch MJ et al. (2003) Gaussian 03, Revision B03. Gaussian Inc, Pittsburgh
- Mirzaei M, Nouri A (2010) J Mol Struct THEOCHEM 942:83–87
- Bengu E, Marks LD (2001) Phys Rev Lett 86:2385–2387

Unique example of amyloid aggregates stabilized by main chain H-bond instead of the steric zipper: molecular dynamics study of the amyloidogenic segment of amylin wild-type and mutants

Workalemahu Mikre Berhanu · Artëm E. Masunov

Received: 16 August 2010 / Accepted: 6 March 2011 / Published online: 28 May 2011
© Springer-Verlag 2011

Abstract Most proteins do not aggregate while in their native functional states. However, they may be disturbed from their native conformation by certain change in the environment, and form unwanted oligomeric or polymeric aggregates. Recent experimental data demonstrate that soluble oligomers of amyloidogenic proteins are responsible for amyloidosis and its cytotoxicity. Human islet amyloid polypeptide (IAPP or amylin) is a 37-residue hormone found as fibrillar deposits in pancreatic extracts of nearly all type II diabetics. In this study we performed *in silico* mutation analysis to examine the stability of the double layer five strand aggregates formed by heptapeptide NNFGAIL segment from amyline peptide. This segment is one of the shortest fragments that can form amyloid fibrils similar to those formed by the full length peptide. The mutants obtained by single glycine replacement were also studied to investigate the specificity of the dry self-complementary interface between the neighboring β -sheet layers. The molecular dynamics simulations of the aggregates run for 20 ns at 330 K, the degree of the aggregate disassembly was investigated using several geometry analysis tools: the root mean square deviations of the C_{α} atoms, root mean square fluctuations per residue, twist angles, interstrand distances, fraction of the secondary structure elements, and number of H-bonds. The analysis

shows that most mutations make the aggregates unstable, and their stabilities were dependent to a large extent on the position of replaced residues. Our mutational simulations are in agreement with the previous experimental observations. We also used free binding energy calculations to determine the role of different components: nonpolar effects, electrostatics and entropy in binding. Nonpolar effects remained consistently more favorable in wild type and mutants reinforcing the importance of hydrophobic effects in protein-protein binding. While entropy systematically opposed binding in all cases, there was no clear trend in the entropy difference between wildtype and glycine mutants. Free energy decomposition shows residues situated at the interface were found to make favorable contributions to the peptide-peptide association. The study of the wild type and mutants in an explicit solvent could provide valuable insight into the future computer guided design efforts for the amyloid aggregation inhibitor.

Keywords Aggregation · Amylin · Amyloid fibril · β sheet · Binding free energy · Cross- β structure · Hydrogen bond · MM-PBSA · Molecular dynamic simulation · NNFGAIL · Oligomer · Secondary structure

Introduction

A number of human diseases known as amyloidoses are associated with the presence of amyloid plaques in organs and tissues [1]. The main constituents of these plaques are fibrillar aggregates arising from the pathological self-assembly of normally soluble proteins. The etiology of amyloidoses is poorly understood, and the causative agents in cellular toxicity have been associated with soluble

W. M. Berhanu
NanoScience Technology Center and Department of Chemistry,
University of Central Florida,
Orlando, FL 32826, USA

A. E. Masunov (✉)
NanoScience Technology Center, Department of Chemistry,
and Department of Physics, University of Central Florida,
Orlando, FL 32826, USA
e-mail: amasunov@knights.ucf.edu

oligomers fibrils [2]. The fibrillar products of aggregation share common structural features: they are enriched in β -sheet structure and possess a common cross- β sheet motif, in which the β -strands lay perpendicular to the main axis of the fibril [12–16]. In most cases, the atomic structure of the fibrils is not known, however, recent studies employing solid-state NMR have provided details on inter- and intra-molecular interactions within several specific fibrils, and have shed light on mechanisms of aggregation [3].

Amylin, also called islet amyloid polypeptide or IAPP, is a peptide that is co-secreted with insulin by pancreatic β -cells. Amylin is the major peptide or protein component of the islet amyloid found in the pancreas of approximately 90% of type 2 (or non-insulin-dependent) diabetes patients [4, 5]. The amylin sequence contains 37 residues: KCNTATCATQRLANFLVHSSNNFGAILSSTNVGSNTY. Of these residues two underlined ones are positively charged at neutral or higher pH, and 14 are hydrophobic. Thus, the peptide has hydrogen bonding capacity throughout the backbone and at 23 side chains. Although the sequence of IAPP is strongly conserved over a number of animal species, IAPP-derived amyloid is only formed by humans, cats, and few non-human primates [6]. Rodents do not form pancreatic amyloid, although rat IAPP sequence differs from its human analog by only six amino acid residues. Considering that five of these six amino acid residues are located between residues 20 and 29, amyloidogenicity of amylin has been attributed to these residues. It has been shown that the synthetic decapeptide amylin (20–29) is able to form fibrils with morphology that is similar to the fibrils formed by the complete amylin sequence [7]. The aggregation-prone region of the peptide (residues 20–29) has been identified through comparison of amylin variants from different species with variable amyloidogenic propensity. Similar to the full length amylin, several fragments from human amylin form amyloid fibrils in vitro [8–12]. These fragments include residues 8–20, 14–20, 20–29, 22–27, 23–27, and 30–37. Among these peptides, only residues 20–29 had been suggested to be the cause of amyloidogenic propensity of full length human amylin.

Recently Wiltzius et al. [13] have published the structure of the heptapeptide NNFGAIL (residues 22–27 of the human islet amyloid polypeptide) from X-ray diffraction data. The structure of the NNFGAIL segment aggregate is not a typical steric zipper (i.e., pairs of tightly packed, highly complementary β -sheets with interdigitated side-chains [9]). It contains a pronounced bend in the backbone facilitated by a Gly in the fourth position. This bend allows the side chain of Asn in the second position to turn inward and form hydrogen bond to the backbone carbonyl of the Gly residue (Fig. 1). The structure also lacks the

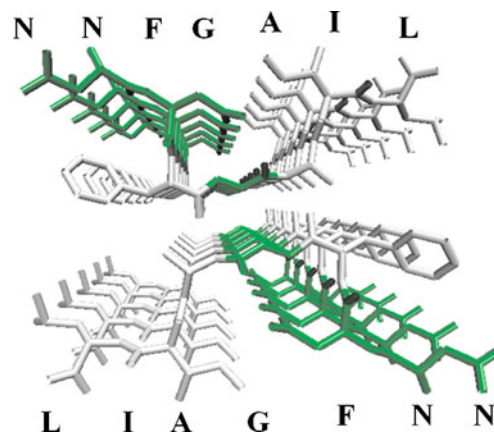


Fig. 1 The structure of NNFGAIL (21–27): it lacks of any side-chain interlocking making this peptide an exception among other steric zipper amyloid structures reported by Eisenberg group. Instead, the association is between the main chain–main chain interface formed by Phe, Gly, and Ala residues from opposing sheets. Polar residues are colored in green, and hydrophobic ones are colored in gray

interdigitized side chains of the steric zipper motif. This peptide instead has a tight main chain–main chain interface formed by Phe, Gly, and Ala residues from opposing sheets. This interface has a significant shape complementarity and large interface area, as established in the Ref. [13]. The main-chain carbonyl of the Phe is tilted within the backbone and is hydrogen bonding with a neighboring main-chain amide across this dry interface [13]. The Phe side chain adopts a rotamer that favors this mainchain packing. The short NNFGAIL fragment is a good model system because it is one of the shortest fragments that can form amyloid fibrils similar to those formed by the full length peptide and the fibrils are also toxic to the pancreatic cell line [11, 14].

While atomistic characterization of the fibril form of this amyloid peptide has been emerging [13], the structures of the early aggregation species, including monomer and small oligomers, remain poorly understood. To date, atomic information for the aggregation of the amylin peptide in explicit water is still limited [15]. Investigating the structural fluctuations of the amylin fragment NNFGAIL and computational mutation studies will provide knowledge of the relative important of different regions of NNFGAIL in stabilizing this short segment oligomer. The mutational studies may help in identifying the residues that stabilize the NNFGAIL oligomer and can be used for designing drugs to inhibit amylin aggregation targeted at the less flexible portion. Understanding the dynamic behavior of the amyloidogenic peptides is expected to provide insights into the possible mechanism of amyloid formation. Amyloidogenic sequences tend to lack Pro and Gly, presumably as they are destabilizing in β structure [16]. Conservation of glycine and proline residues at

structurally strategic positions appears to serve the purpose of aggregation prevention [17]. Experiments with de novo peptides and proteins as well as with mutated forms of naturally occurring proteins, have elucidated features of polypeptide sequence which inhibit aggregation and fibril formation [17].

A number of inhibitors of aggregation of the amyloid- β (A β) peptide have been examined for their potential application to the treatment of Alzheimer's disease [18]. By contrast, only a limited number of investigations have been made into the similar design of compounds to inhibit aggregation of amylin [19]. Mutagenesis has also been used to probe secondary structure and inter-sheet side-chain packing. Single proline substitutions within the 20–29 fragment of amylin revealed that substitution of residues 22, 24 and 26–28 destabilizes fibrils and alters the kinetics of fibril formation [20]. Abedini et al. [21] designed a variant of the amyloidogenic 8–37 region of human amylin with proline substitutions at positions 17, 19 and 30. Compared to the wild-type, the mutant had dramatically greater solubility. Clearly, more detailed structural descriptions of the human amylin is very important for the molecular understanding of the oligomerization process and for the development of innovative therapeutic and diagnostic approaches in type 2 diabetics.

Computational studies have complemented experiments to provide insights into amyloid formation. Molecular dynamic simulation has given considerable insights into the mechanisms of formation of amyloid fibrils [4]. The advantage of molecular simulation is that most if not all relevant structural, kinetic, and thermodynamic observables of a chemical system can be calculated at one time, in the context of a molecular model [5]. MD simulations have provided insights into (a) the intrinsic propensities of peptide fragments to associate in amyloid-like states; (b) the energetic factors stabilizing these aggregates; (c) the possible structural states of either oligomeric precursors or larger assemblies [22–25]; (d) the molecular level mechanisms of interactions of inhibitor with amyloid polypeptide [26–28]; (e) amyloid aggregation pathway and kinetics of amyloid association [29, 30]; (f) mechanism of membrane disruption effect of amyloid [31].

Some previous molecular dynamic simulation studies have focused on the fibrillogenic properties of short human amylin peptides of seven to ten residues, aiming to identify regions likely to be responsible for the amyloidogenic properties of full length amylin [25, 32, 33]. Wu et al. [25] performed a series of molecular dynamic studies on the formation of ordered aggregates of hexapeptide NFGAIL. They observed that the main growth mode was elongation along the β -sheet hydrogen bonds through primarily a two-

stage process. They found that the peptides initially attached to the surface of the ordered oligomer, by hydrophobic forces, then moved quickly to the β -sheet edges, and formed stable β -sheet hydrogen bonds. Addition of peptides to the existing oligomer notably improves the order of the peptide aggregate in which labile outer layer β -sheets were stabilized, and provides good templates for further elongation.

In the present study we performed MD simulations to explore the stability of smallest aggregation segment of amylin (NFGAIL) in solution and examine its dependence on the position of residue mutation using the tiny β -breaker amino acid glycine. We also estimated the binding free energy for dimerization of the β -sheet into a double layer. In addition we investigated the most promising, potential structural target for further drug design based on the structure-stability information of the wild type and mutants.

Computational details

System setup

The microcrystal structure and coordinate of the NFGAIL assembly with a pair of β -sheets of five strands (SH2-ST5 model) has been determined by Wiltzius et al. [13] and was kindly provided by Dr. M. Sawaya. The model does not have the typical cross- β spine and lacks any side-chain interlocking. The structure instead shows association between the main chains interface formed by Phe, Gly, and Ala residues from neighboring sheets with parallel β -strands (Fig. 1).

Molecular dynamics simulation

The molecular dynamic (MD) simulation was performed using AMBER11 package [34] with an all atom amber99SB force field and explicit TIP3P water models. Each of the NFGAIL segment of amylin models and the corresponding mutants were solvated by explicit water molecules in octahedral box that extends 10 Å from the protein atoms (Table 1). Counterions were added to the box by randomly replacing water molecules to neutralize the system. Each system was energy minimized to remove bad contact by using conjugate gradient method with the peptide constrained and then to relax the atoms without position constraints. The system was then subjected to 50 ps of heating procedure while constraining the backbone atoms of the protein to allow relaxation of water and ions, followed by 500 ps equilibration run without any constraints. The production times were 20 ns for different simulations for the NFGAIL heptapeptide amylin models

Table 1 Summary of the NNFGAIL oligomeric models and simulation system

Model	Systems	Sheet/strand organization	Simulation box size (Å)	Simulation time (ns)	T(K)
Wilde type, WT	Two sheet, five strands (NNFGAIL)	Parallel /Antiparallel	61.28×61.28×61.28	20	330
Single point mutants					
N1G	Two sheet, five strands (GNFGAIL)	Parallel /Antiparallel	59.78×59.78×59.78	20	330
N2G	Two sheet, five strands (NGFGAIL)	Parallel /Antiparallel	61.11×61.11×61.11	20	330
F3G	Two sheet, five strands (NNGGAIL)	Parallel /Antiparallel	62.02×62.02×62.02	20	330
A5G	Two sheet, five strands (NNFGAIL)	Parallel /Antiparallel	61.43×61.43×61.43	20	330
I6G	Two sheet, five strands (NNFGAGL)	Parallel /Antiparallel	61.43×61.43×61.43	20	330
F3Y	Two sheet, five strands (NNPGAIL)	Parallel /Antiparallel	61.45×61.45×61.45	20	330
I6P	Two sheet, five strands (NNFGAPL)	Parallel /Antiparallel	61.54×61.54×61.54	20	330
L7G	Two sheet, five strands (NNFGAIG)	Parallel /Antiparallel	59.98×59.98×59.98	20	330
Double point mutants					
N2GF3G	Two sheet, five strands (NNGGAIL)	Parallel /Antiparallel	62.67×62.67×62.67	20	330
F3GI6G	Two sheet, five strands (NNGGAGL)	Parallel /Antiparallel	61.42×61.42×61.42	20	330

and the corresponding mutants. Constant pressure (1 atm) and temperature (330 K) on the system was maintained by isotropic Langevin barostat and a Langevin thermostat. This somewhat elevated temperature was chosen to accelerate convergence to the equilibrium without destroying the fibrils (according to experimental data, the fibrils fully dissociate above ~373 K, but are stable below ~330 K) [35, 36]. Electrostatic interactions were calculated by using the particle mesh Ewald (PME) method [37]. Most PME parameters, including the cutoff distance of 12 Å, were kept at the values default in AMBER11 package [38]. The SHAKE algorithm [39] was used for bond constraints and the time step was 2 fs for all simulations. Each system was simulated for 20 ns and the trajectories were saved at 4.0 ps intervals for further analysis. VMD (visual molecular dynamics) [40] program was used for the visualization of trajectories. Hydrogen bond occupancies and structure RMSDs was calculated using PTRAJ module available within AMBER. A hydrogen bond was assigned if the distance between donor D and acceptor A is ≤ 3.5 Å and the angle D-H ...A $\geq 120^\circ$ [41]. The MM-PBSA single trajectory approach implemented as script (MMPBSA.py) in AMBER11 [34], was used to calculate the binding energy for non-covalent association between the β -sheets within the double layer. In this approach an assumption is made that no significant conformational changes occur upon binding, i.e., structural adaptation is negligible and the snapshots for all three species can be obtained from the single trajectory of the complex by separating it into two components. To calculate binding free energies in MM-PBSA method, the explicit water simulations were used to generate the trajectory followed by the implicit Poisson-Boltzmann/surface area method to calculate solvation energy terms.

The gas phase and the solvation free energies were calculated over the course of the 20 ns of the MD production trajectories. This approach was previously used to study the thermodynamics of amyloidogenic peptides by Wu et al. [25].

Mutant studies

The NNFGAIL heptapeptide amylin model was mutated to produce the corresponding glycine, proline and tyrosine mutants (see Fig. 1 and Table 1) to examine the stability of the NNFGAIL aggregate, understand the driving force for aggregation and find ways to prevent the fibril formation. The wildtype 5-stranded double layer aggregate was used as the starting structure to generate the single point glycine mutants of NNFGAIL. The coordinates of all starting structures of the mutants were built from the wild-type NNFGAIL by substituting the side-chains of the targeted residues [42] using Sirius visualization program [43]. The wild types were mutated to examine the effect of the side-chain interactions of the amino acids involved in stabilizing the sheet to sheet and strand-to-strand association of the different amyloid peptide fragments. The structure of the designed mutant was first minimized for 500 steps using the steepest decent algorithm with the backbone of the protein restrained before being subjected to the simulation. While most of the in silico mutations were done with glycine, two of the mutants (NNYGAIL, F3Y and NNFGAPL, I6P) were derived from the wild type sequence by replacement of the phenylalanine side with tyrosine side chain, and the isoleucine with proline. The resulting structures were minimized for 500 steps using the steepest decent algorithm before the simulation.

Results and discussion

Relative structure stability of amylin oligomers

The conformational change and the conservation of the oligomers were monitored by the time evolution of the backbone root mean square (RMSD) and root mean square fluctuation (RMSF) through the simulation relative to their initial energy minimized structure as shown in Fig. 2a and b. The RMSDs provide useful information on relative stability of the oligomers, and were previously used in stability analyses of amyloid oligomers with β -sheet structure [44–46].

RMSD

The conformation change and oligomers stability of the NNFGAIL and its mutants was monitored by the time evolution of the RMSD. The RMSD of the wildtype (WT) and its mutants are shown in Fig. 2. The results indicate that the mutants F3G and I6G have RMSD of about 5.0 Å, more than double the value of the wildtype. This reflects

significant instability of the aggregates assembled from these mutants. Two other mutants (N1G and N2G) have RMSD value of about 4 Å that also corresponds to the reduced stability of their aggregates. The other single point mutants A5G, I6P and L7G have an RMSD of about 2.5 to 3.50 Å. These results indicate the important role of the Asn2, Phen3 and Ile6 in stabilizing the oligomers. While replacement of Phen3 with Gly has dramatic effect on aggregate stability, its replacement with Tyr has virtually no effect, and the F3Y mutant is as stable as the WT with similar RMSD ~2 Å. These results reveal stabilizing role of the π - π interaction between the aromatic side chains on the aggregate stability, in agreement with previous experimental observations [47].

The result of double point glycine mutation shows the RMSD for non-adjacent amino acid mutant F3GI6G is significantly higher (RMSD ~12 Å) than for the mutant N2GF3G with adjacent amino acid (RMSD ~6 Å). One can rationalize this result as follows. When a replacement is done on the non-adjacent amino acids at positions 3 and 6 this disrupts the main chain-main chain interactions to a larger degree, than does the replacement of the adjacent amino acids at positions 2 and 3. Thus one can expect the effect on the overall structural organization to be affected more by

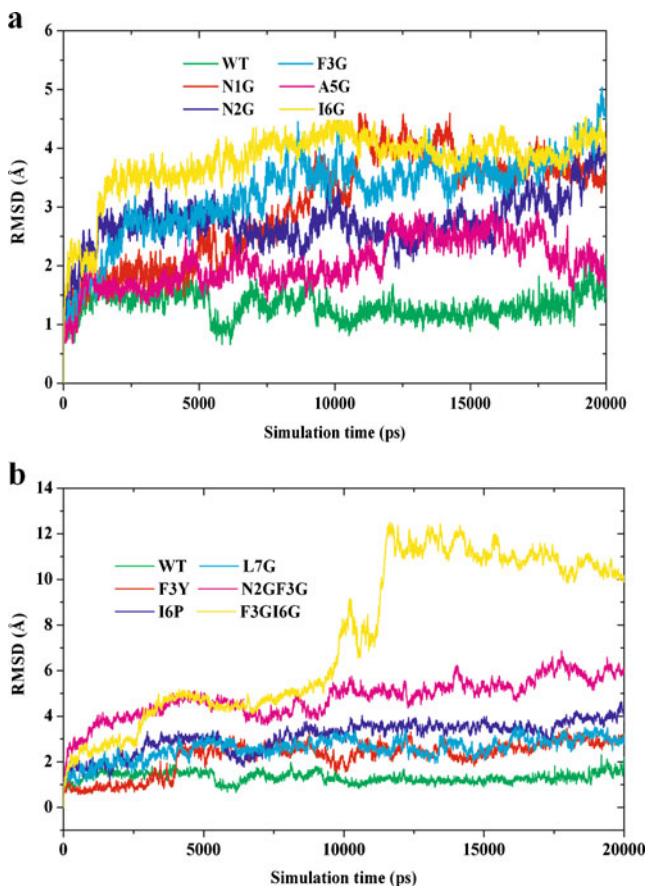


Fig. 2 Time evolution of the backbone RMSD of the 5 β -strands double-sheet wildtype amylin NNFGAIL sequence and its mutants

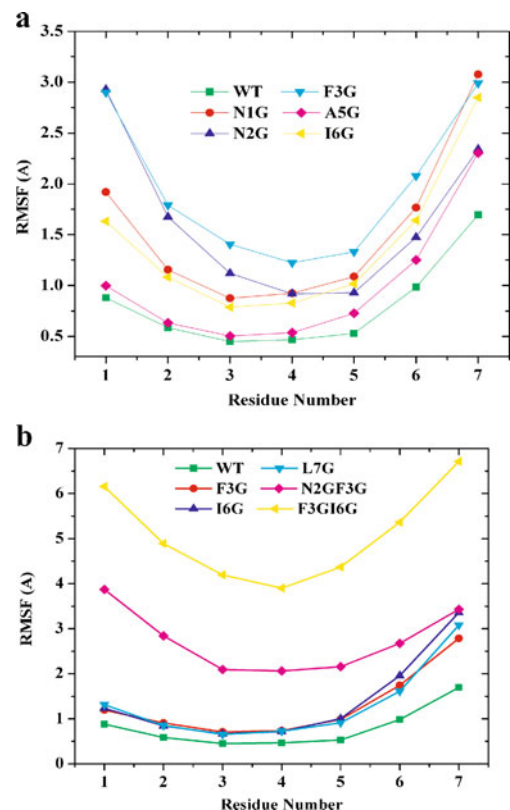


Fig. 3 Average RMSF values for 5 β -strands double-sheet wildtype (NNFGAIL) and its mutants

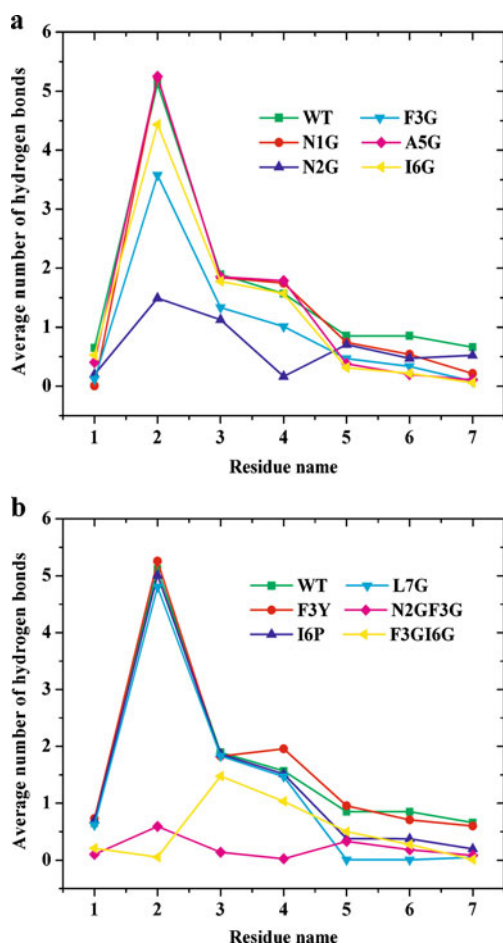


Fig. 4 Average numbers of hydrogen bonds for individual residues from the simulations of the wild-type models of amylin (NNFGAIL) and its mutants

substitution of amino acids at non-adjacent positions. It is clear from the RMSD results that the mutations are indeed drastically reducing the aggregate stability in most cases, except for the mutation of Phe with Tyr.

RMSF

The residue-based root mean square fluctuation (RMSF) of the backbones was used to assess the local dynamics and flexibility of each residue using PTRAJ tool in AMBER. Figure 3 shows the RMSF values of atomic positions by each residue, computed throughout the simulation for wildtype and its mutants. Among the single point mutants the RMSF values for F3G, N2G and N1G are much larger and this followed by A5G, I6P and L7G. The wildtype (WT) and mutated sequences (Phe \rightarrow Tyr) have the smallest RMSF values. In the case of double glycine mutants N3GI6G with non-adjacent amino acid shows an enhanced flexibility in both the terminal and central region compared to the other double point mutant N2GF3G, with adjacent amino acid mutants.

The RMSF results for the wildtype and the mutants indicates that all chains have common characteristics of small variation for the three central residues whereas large variations for the two terminal residues, suggesting that the central residues are more rigid than the residues in the termini regions. This is in agreement with the trend reported by Zheng et al. [48]. The lowest fluctuations in all cases were observed by residue 3 suggesting a low inter-chain mobility and a great compactness in this portion. This is a promising target for further drug design based on the structure stability information. One can suggest new “amyloid inhibitors” capable of interacting specifically with this portion of the aggregates. Single point alanine (Phe23 \rightarrow Ala), and proline substitution (Asn22 \rightarrow Pro), (Ile26 \rightarrow Pro), (Lys27 \rightarrow Pro) were found to inhibit the aggregation of amylin [20, 49, 50]. Porat et al. [47, 51] showed that no amyloid formation could be observed under the experimental conditions when the phenylalanine was replaced with alanine. On the other hand modification of the phenylalanine to tryptophan (that

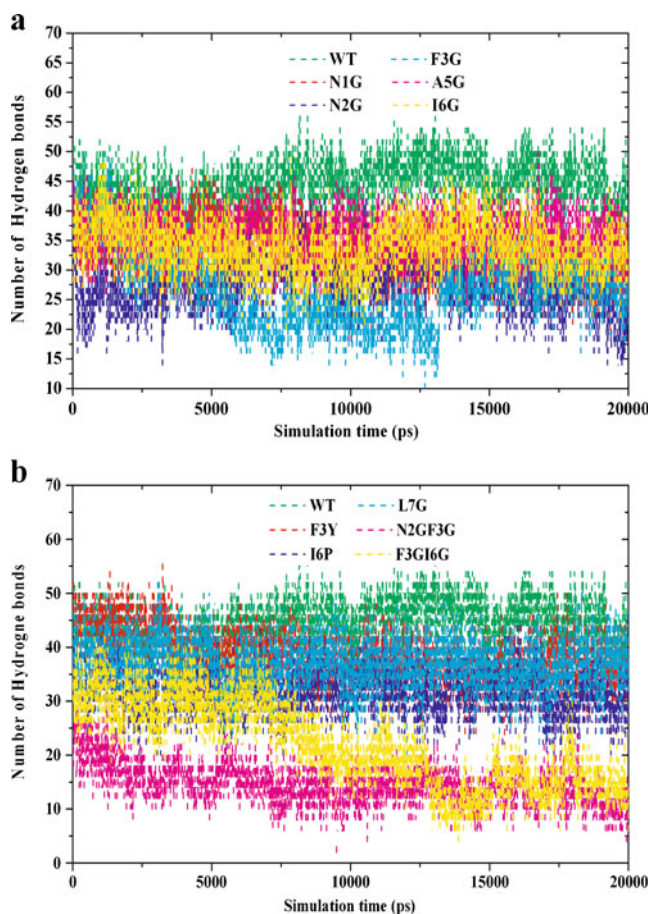


Fig. 5 Time evolution of the total number of hydrogen bonds between peptides (backbone and side chain) obtained from the wild type and mutants. The wildtype preserved the number of hydrogen bond while the mutants loss hydrogen bonds

is less hydrophobic than phenylalanine) allows the formation of amyloid-like structure. Our simulation is in agreement with the above experimental observation in that the single point glycine substitution of Asn22, Phe23, Ile26 and Lys27 increase the structural disorder of the mutant while (Phe23→Tyr) substitution retains the structural stability of the wild type. The result of the simulation indicates that *in silico* mutation in combination with MD simulation is useful in identifying the critical amino acid that stabilize the amyloid peptide and may be useful in designing peptidomimetics amyloid aggregation inhibitors [52, 53].

Hydrogen bond analysis

The analysis of the number of hydrogen bonds of individual residues, averaged from 20 ns simulations for the wild type and mutants are shown in Fig. 4. One can see that the average number of hydrogen bonds for the central residues is larger than those for the two terminal residues for all cases, consistent with the RMSF results.

The larger flexibility of these residues is due their exposure to the water and formation hydrogen bonds with the water molecules rather than peptides. The hydrogen bonds between the two terminal residues and water molecules are weak and easily break for both wild type and mutant aggregates. Because glycine and proline cannot make hydrogen bonds, total counts of hydrogen bonds in all the mutants are smaller than the wild-type (Fig. 4).

In general, the average number of hydrogen bond per residue for the wild type and the F3Y mutant is the largest. Mutants with a larger RMSD and RMSF have smaller hydrogen bonds per residue (Fig. 4). The smallest average hydrogen bond per residue for a single mutant were found for N2G, I6G, and I6G, suggesting that N2 and I6 are key residues for NNFGAIL aggregation. The replacement of either N2 or I6 with the beta breaker amino acid Gly or Pro resulted in a significant reduction of hydrogen bond in the central residues of the peptide making these mutants structurally unstable. The Asn in the second position in the wildtype

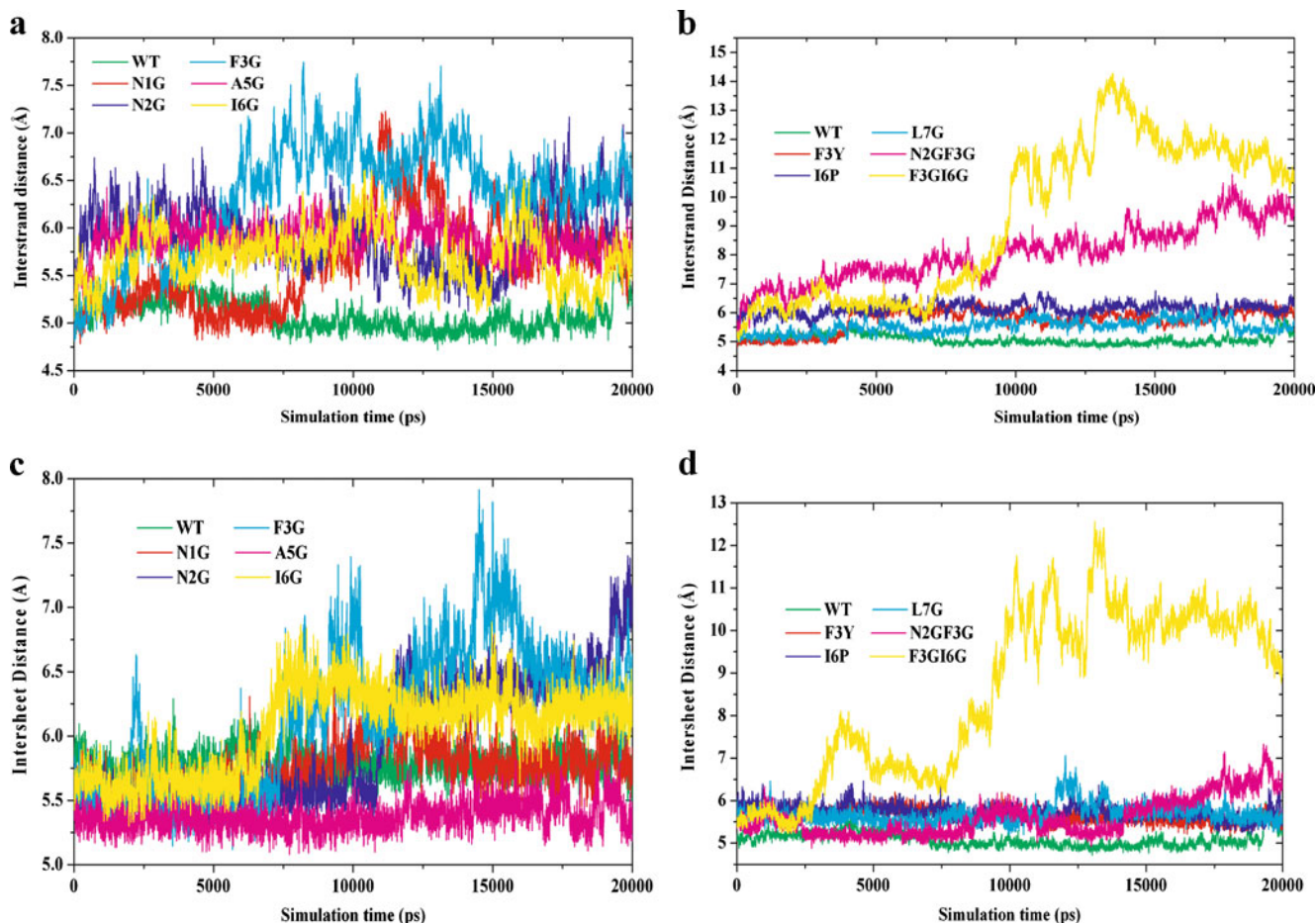


Fig. 6 Time evolution of inter-strand distances of the wild-type and its mutant models of NNFGAIL (**a–b**) and sheet-to-sheet distances for the wild-type of NNFGAIL and its mutants models (**c–d**) between central residues (3–4–5) Phe-Gly-Ala of the wildtype and its mutants

turn inward and forms a hydrogen bond to the backbone carbonyl of the Gly residue at position four. The replacement of Asn2 with Gly significantly decreases the hydrogen bonds in the central region making this mutant the most unstable. This is also evident in the RMSD, RMSF (see above), inter-sheet and inter-strand distances (see below). The two main forces stabilizing proteins are the hydrophobic effect and hydrogen bonding [54]. The intra-sheet hydrogen bonds were previously found to be necessary to stabilize the main conformational pattern of amyloid fibrils, β -sheets [33].

To monitor the stability of the oligomers in this study, we also analyzed the change in the number of the total (backbone and side chain) peptide-peptide hydrogen bonds. The hydrogen bonding was found to be stable during the simulation for the wildtype and F3Y. The mutants N2GF3G and F3GI6G are the least stable and the hydrogen bonding interaction in their aggregates disappear rapidly (Fig. 4). The result of the analysis of the total hydrogen bond indicates that the wild type and N1G, F3Y, A5G, L7G mutants preserved about 80% of the original hydrogen bonding with respect to the wildtype. The mutants I6G and I6P, F3G and N2G preserved ~70%, 60% and 40% of the original hydrogen bonding respectively (Fig. 5). In the case of the double mutants, N2GF3G lost more than 60% of the original hydrogen bonds while F3GI6G lost only about 30% of the hydrogen bonding with respect to the minimized structure of the wildtype (Fig. 5). The effect of mutation on the preservation of the hydrogen bonding compared to the wildtype is due to the fact that the mutation reduces the side chain hydrogen bonds especially in N2G (Asn in the second position to the backbone carbonyl of the Gly residue at position four).

Inter-strand (d_{strand}) and inter-sheet (d_{sheet}) distances

To examine the structural stability of the wildtype and the corresponding mutant oligomers we also analyzed the inter-strand and inter-sheet distances. The d_{strand} is calculated by averaging the distance between each residue in one strand and its corresponding residue in adjacent strand in the same sheet, whereas d_{sheet} is calculated by averaging the mass center distance between each strand center of mass in one sheet and its corresponding strand in the adjacent sheet [48]. The inter-sheet and inter-strand distances for wild type and mutants are shown in Fig. 6.

The inter-sheet distance for both the wildtype and the mutants N1G, A5G, F3Y, I6P and L7G were found to be within the 5.5 to 6.0 Å which is very close to initial the inter-sheet distance of ~5.0 Å for the central region consisting the residues 3 to 5 in the double layer oligomer.

This suggests the structure remains stable during the 20 ns simulation for the above mutant. The result of the inter-sheet distance for the mutants was found to be large for the mutants N2G, F3G, I6G and N2GF3G with a value of 6–7.5 Å. In the case of the mutant F3GI6G the inter-sheet distance were within the range of 5.5–11.0 Å indicating the tendency of the sheets to come apart making these particular mutants structurally unstable. The result suggests the wild type and the F3Y form the most stable aggregates (Fig. 6). The N2GF3G and F3GI6G have larger variation of about 6.0 Å in its inter-strand distance (between 5.0 to 13.0 Å). The N2GF3G and F3GI6G are unstable and this structural modification will disaggregate the mutant oligomers. The inter-strand distances for remaining mutants were found within the range of 5.0 to 7.5 Å.

Secondary structure contents

We analyzed the secondary structure of the oligomers using the DSSP algorithm written by Kabsch and Sander [55]. This algorithm is based on identification of hydrogen-bonding (H-bonding) patterns and recognizes seven types

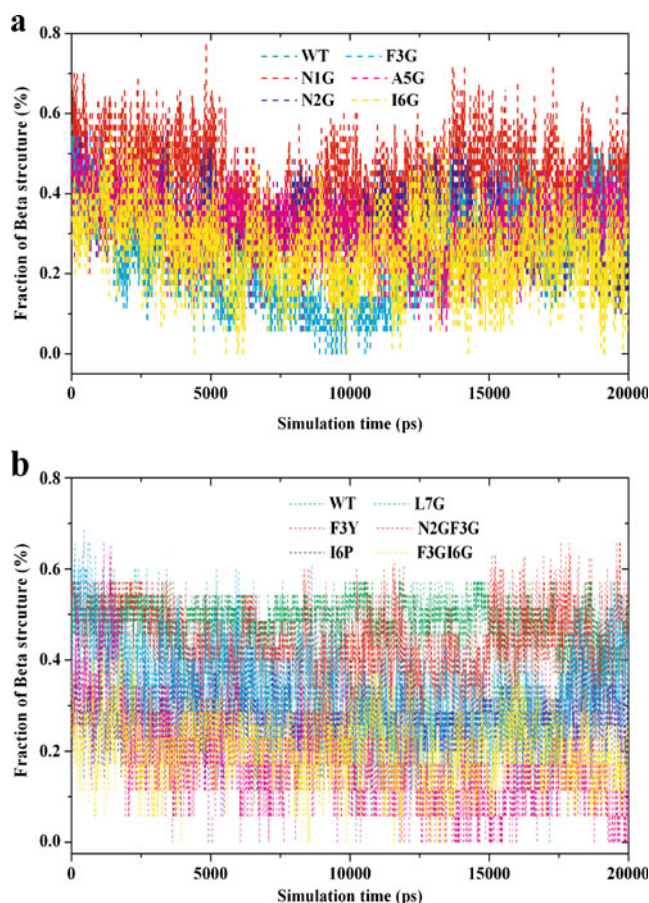


Fig. 7 Time evolution of β -strand contents of the wildtype amylin NNFGAIL sequence and mutants

of secondary structures which can be grouped into three classes: helix (α -helix, 3_{10} -helix, π -helix), β -strand (isolated β -bridge, extended β -sheet) and loop (turn, bend). The secondary structure of the wildtype NNFGAIL segment of amylin oligomer and its mutants as calculated by the Kabsch and Sander algorithm of AMBER during the 20 ns simulations are shown in Fig. 7.

The result of the secondary structure analysis of the wildtype oligomer and the F3Y mutant shows greatest stability at 330 K for the 20 ns simulation, which is confirmed by the conservation of the β -sheet content throughout the whole simulation time (as shown in Fig. 7). The glycine and proline mutants were found to be unstable with lack of preserving the β -sheet content during the simulation. The single point mutants (N2G, F3G, I6G and I6P) and both double mutants were found to preserve the β -sheet content to lesser extent indicating the reduced stability of these mutants. The results of β sheet content are in agreement with the RMSD and RMSF result in that those structures with highest RMSD and RMSF fluctuation were also found to be lacking the ability to preserve secondary structure and conformational stability.

Sheet-to-sheet binding energy

To further quantify the driving force underlying the β -sheet association of the studied wildtype amylin models and mutants, we calculated the interaction energy between β -sheets. The trajectories were first extracted from explicit MD trajectories by excluding water molecules. The solvation energies of double-layer sheets and each single-layered sheet were calculated using the MM-PBSA (Molecular Mechanics-Poisson-Boltzmann/Surface Area) module [56, 57] in the AMBER package. The determination of the binding free

energy following the MM-PBSA approach has been described in the past and has been shown to be a good method for comparing binding energies between similar peptides [58]. In the MM-PBSA calculation, the dielectric constant of water is set to 80 and no distance cutoff is used. The binding energy between two β -sheets was calculated by

$$\langle \Delta G_{\text{binding}} \rangle = \langle \Delta G_C \rangle - \langle \Delta G_A \rangle - \langle \Delta G_B \rangle \tag{1}$$

Where C, A and B stands for complex (the double-layer sheet), sheet1 and sheet2. The free energy of each system X = A, B, or C was computed as a sum of the three terms [57, 59]:

$$\langle \Delta G_X \rangle = \langle E_{\text{MM}} \rangle + \langle \Delta G_{\text{solv}} \rangle - T \langle S \rangle \tag{2}$$

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}):

$$E_{\text{MM}} = E_{\text{int}} + E_{\text{ele}} + E_{\text{vdw}} \tag{3}$$

ΔG_{solv} accounts for the solvation energy which can be divided into the polar and nonpolar part:

$$\Delta G_{\text{solv}} = \Delta G_{\text{PB}} + \Delta G_{\text{SA}} \tag{4}$$

The polar part ΔG_{PB} accounts for the electrostatic contribution to solvation and is obtained by solving the linear Poisson-Boltzmann equation in a continuum model of the solvent.

The second term ΔG_{SA} is nonpolar contribution to solvation free energy that is linearly dependent on the solvent accessible surface area (SASA):

$$\Delta G_{\text{SA}} = \gamma \text{SASA} + b \tag{5}$$

Table 2 Binding free energy components (kcal mol⁻¹) and standard deviations calculated with MM-PBSA for wild type and mutants of the amylin (NNFGAIL) oligomer double-layers (SH2-ST5 models): ΔE^{ele} , nonsolvent electrostatic potential energy; ΔG_{PB} , electrostatic contributions to the solvation free energy calculated with Poisson-

Boltzmann equation; G_{SA} , nonpolar contributions to solvation free energy; ΔE^{vdw} , van der Waals potential energy; $T\Delta S$, the entropic contribution calculated to the free energy of binding; $\Delta G_{\text{binding}}$, calculated binding free energy

Type	$\langle \Delta E^{\text{ele}} \rangle$	$\langle \Delta E^{\text{vdw}} \rangle$	$\langle \Delta G_{\text{PB}} \rangle$	$\langle \Delta G_{\text{SA}} \rangle$	$\langle \Delta G_{\text{subtotal}} \rangle$	$\langle T\Delta S \rangle$	$\langle \Delta G_{\text{binding}} \rangle$
WT	-299.5±24.4	-81.9±5.6	305.8±26.4	-6.7±0.5	-82.3±5.3	-31.6±3.9	-50.7(-10.1) ^a
N1G	-170.8±36.1	-82.1±7.4	190.2±33.6	-6.6±0.9	-69.4±6.0	-32.9±3.0	-36.5(-7.3)
N2G	-179.5±27.7	-81.3±6.0	188.7±26.8	-7.1±0.5	-79.2±7.2	-30.3±5.3	-48.9(-9.8)
F3G	-457.8±82.8	-70.5±7.9	466.2±86.8	-6.8±1.1	-68.8±9.3	-29.5±3.8	-39.3(-7.9)
A5G	-287.7±28.9	-88.6±6.3	305.9±30.3	-7.5±0.6	-78.0±5.7	-36.6±4.2	-44.4(-8.9)
I6G	-183.0±32.2	-74.2±7.5	197.5±32.9	-6.4±0.8	-66.2±7.9	-23.9±5.5	-42.3(-8.5)
F3Y	-286.5±35.4	-82.9±7.5	300.1±36.2	-6.8±0.7	-76.3±6.5	-28.6±3.9	-47.7(-9.5)
I6P	-208.2±46.2	-81.7±7.2	232.2±45.6	-6.9±0.9	-64.6±7.4	-34.2±3.9	-30.4(-6.1)
L7G	-229.9±47.9	-71.4±6.5	244.2±47.6	-5.62±0.6	-62.7±7.3	-27.2±3.1	-35.5(-7.1)

^a The averaged binding free energy per strand between two beta sheets is in brackets.

The ΔG_{SA} were calculated using AMBER11 default parameter for γ and b (5). Finally, the entropic term in Eq. 2 was calculated with the normal mode analysis [56]. Since this calculation is computationally expensive, $-\Delta S$ was averaged over 100 frames of the MD trajectory (1 frame taken at an interval of 50 frames from the total of 5000 frames). We have used a single molecular dynamics trajectory protocol, which can qualitatively estimate the free energy consequences of many mutations [56, 60].

Detailed characterization of individual energy terms of the calculated binding free energy are shown in Table 2. An inspection of the free energy components for the wild types and mutants investigated in this study reveals that the electrostatic component of the solvation free energy ΔG_{PB} is destabilizing (positive), while the nonpolar component G_{SA} is stabilizing (negative). This is expected, since the complex formation desolvates the monomers, and reduces solvent-accessible surface area. Entropy component was found to contribute unfavorably to binding, since complexation reduces freedom of motion for the monomers. The electrostatic interaction between sheets is stabilizing. These observations are consistent with previous calculations of the components of the free energy of solvation [57, 61]. However, the less favorable electrostatics in each case is compensated by highly favorable nonpolar component of the free energy. In each case, favorable nature of the nonpolar interaction mostly originates from the van der Waals interaction energy ΔG^{vdW} , as opposed to the nonpolar component of solvation ΔG_{SA} . There did not appear to be a clear trend for the entropy change upon binding ($T\Delta S$).

The values of the total binding free energy in all cases are negative. They are reported in the Table 2. These results indicate that the structurally stable models have the lowest binding free energy, while the models which are structurally unstable were found to have higher binding free energy. Despite the fact that N2G mutation does not lower the binding free energy significantly compared to the WT ($1.8 \text{ kcal mol}^{-1}$), the mutation of N2 to G (as in N2G) leads to the loss of the conformation in mutant N2G indicating the important role of N2 in maintaining the stability of the oligomer aggregate. Even though this residue is not found at the protein-protein interface (it is located near to the N terminal region) and thus do not contribute much to the association energy, it forms side chain hydrogen bonds to the backbone carbonyl of the G4 residues at position four. The side chain hydrogen bonds appear to be important in stabilizing the initial conformation. The comparison of the geometry analysis for the wildtype and the N2G mutant as shown in Figs. 2a, 3a, 4a, 5a, 6a and 7a indicates the role of this specific residue in the stabilizing the system. The trend in the calculated binding free energy is in agreement with the observed instability based on RMSD, RMSF, interstrand,

and inter-sheet distances. Those aggregate oligomer models which show structural instability were found to have unfavorable binding energy compare to the stable ones.

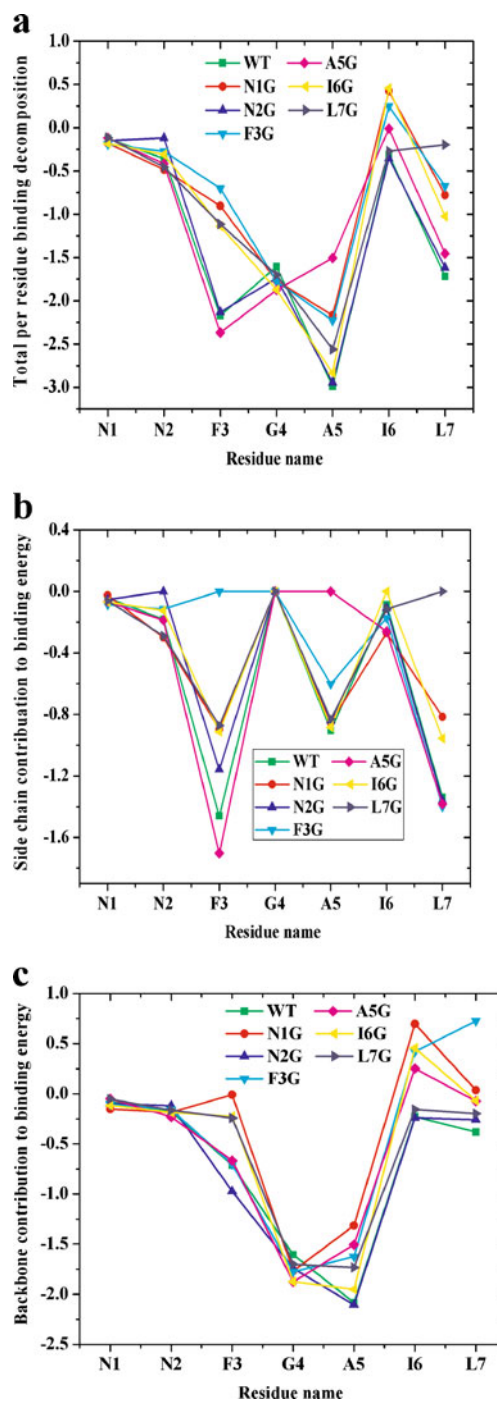


Fig. 8 MM-PBSA per residue decomposition of the binding free energy of the wildtype NNFGAIL and mutants. The energy term corresponds to the sum of backbone and side chain contribution (a), side chain (b) and backbone (c) for the wildtype and single point glycine mutants. A negative value indicates residue makes favorable contribution

In order to identify the residues that contribute the most to the calculated overall binding energy, we used a residue-by-residue decomposition protocol. Binding free energy decomposition at the atomic level allows to evaluate the contribution of each residue to the total binding free energy, as well as the contributions of its side-chain and backbone. In the past, MM-PBSA approach has been applied to estimate the binding energy for protein–protein [61, 62] and protein–ligand systems [60, 63, 64]. It has also been used to predict the effect of residue mutations on the binding energy of protein–protein systems with the “computational alanine scanning” [57, 65]. An alternative approach is MM-GBSA, where the electrostatic contribution to the solvation energy is determined using a generalized Born (GB) model [56]. Despite its generally lower accuracy [66], the GB model has two advantages. First, the GB method is much faster than the PB method. Second, GB allows one to decompose easily and rapidly the electrostatic solvation energy, and thus the binding free energy, into atomic contributions from only one calculation [67].

A decomposition of the binding free energy in the context of PB calculations is also possible [66, 68–70] but requires separate, time-consuming calculations. The MMPBSA.py script in AMBER11 implements per-residue decomposition with both PB and GB implicit solvent models [34]. The PB non-polar solvation component is currently not decomposable. However, the non-polar solvation remains constant in both the wild type and mutants (see Table 2) and is much smaller than the other energy terms. Thus, we used the MM-PBSA decomposition to plot Fig. 8a–c. As one can see, the residues making the most favorable contributions to the binding free energy between the two sheets are residues Phe3, Gly4, and Ala5 (Fig. 8a–c). Their contribution to the binding free energy ranges from -0.5 to -3.0 kcal mol⁻¹. These residues are

situated at the interface between the two sheets and form stable hydrogen bonds between their backbone atoms and van der Waals interactions between their side-chains. The contribution of the side-chains to the association of the five stranded double layer oligomers is larger than that of the backbone atoms, underlining their importance. Mutation of the side chains at the interface to the smallest amino acid glycine resulted in the reduced side chain continuation of the targeted amino acid and this leads to reduced total binding free energy of the sheet to sheet association.

These results are in agreement with the RMSD, RMSF and secondary structure analysis in that the wildtype and the Phe → Tyr mutants are more stable whereas the Phe → Gly and Ile → Pro mutants are less stable. The analysis of the MD simulation indicates that the asparagine, phenylalanine and isoleucine residues are important in the formation and stabilization of the oligomers. The result thus can be used in the rational design of peptiomimetic aggregation inhibitors.

Twisting

Amyloid fibrils typically exhibit twisted β -sheets, as observed by electron microscopy and solid state NMR. Since twisted β -sheets optimize the hydrogen bonds, side-chain stacking, and electrostatic interactions, it is commonly accepted that twisted sheets are more stable than flat ones. While twisting, the β -sheets pairs remain to be complimentary via the steric zippers [44, 71, 72]. The twisting in SH2-ST5 aggregate of NNFGAIL heptapeptide was evaluated by considering pairs of dihedral angles, one per each sheet of the pair. Each dihedral angle is calculated from the coordinates of the C ^{α} (Asn2) and the C ^{α} (Ile6) atom of the second and the fourth strand of the

Table 3 Comparison of the average twist angles in the SH2-ST5 aggregate of NNFGAIL and its mutants

Model	Simulation systems	Average overall twist angle (°)	Average interstrand twist angle (°)
Wild type, WT	Two sheet, five strands (NNFGAIL)	13.3±4.7	4.4±1.6
Single point mutants			
N1G	Two sheet, five strands (GNFGAIL)	7.1±7.7	2.4±2.5
N2G	Two sheet, five strands (NGFGAIL)	37.3±9.7	12.4±3.2
F3G	Two sheet, five strands (NNGGAIL)	28.7±15.9	9.5±5.3
A5G	Two sheet, five strands (NNFGGIL)	8.9±5.7	3.0±1.9
I6G	Two sheet, five strands (NNFGAGL)	9.0±8.1	5.4±2.7
F3Y	Two sheet, five strands (NNPGAIL)	1.9±10.0	0.6±3.3
I6P	Two sheet, five strands (NNFGAPL)	0.1±8.0	0.05±2.7
L7G	Two sheet, five strands (NNFGAIG)	9.7±8.2	3.2±2.7
Double point mutants			
N2GF3G	Two sheet, five strands (NNGGAIL)	68.3±44.7	22.8±14.9
F3GI6G	Two sheet, five strands (NNGGAGL)	6.6±49.3	2.2±16.4

sheet. Twisting angles have been computed by using the three inner strands [72]. As shown in Table 3, for the wildtype NNFGAIL model, the average twist of $\sim 13.3^\circ$ is observed with an estimated twist of 4.4° ($13.3^\circ \div 3$) between consecutive strands. This value is much smaller than the previous analyses that estimated a twist of $10\text{--}11^\circ$ between consecutive strands of the GNNQQNY model [23]. The smaller twist angle observed for the NNFGAIL and its mutant compared to GNNQQNY model (SH2-ST5) could be due to the lack of the steric zipper in the NNFGAIL and its mutant.

Several groups have shown that peptides or peptidomimetics can inhibit A β aggregation [33]. Our simulation indicates the peptides N2G, N3F, N2GF3G, I6G and I6P could be a starting point for designing peptidomimetic inhibitor of amylin. The synthetic peptides suffer from a disadvantage of being able to undergo self-amyloidosis, which limits their application in therapeutics development. The strategy of N-methylation of peptide amide bonds has been a well-known protein-design approach to suppress H-bonding ability of an NH group and to restrict the conformation of the backbone. The identification of amino acid important in stabilizing the amyloid aggregate using MD simulation based on a single point mutation with beta breaker amino acid and combining this with N-methylation of the peptide amide could be a variable option. The designed molecule can prevent fibrils from forming and break down already formed fibrils [73, 74].

Conclusions

In this work we report the effect of single point mutations on amyloidogenic propensity of the NNFGAIL peptide (the shortest aggregation prone region of amylin) with all-atom explicit solvent molecular dynamics simulation. The results suggest that the aggregates formed by the wildtype and F3Y mutant are more stable than by other mutant peptides. The free energy calculations indicate that hydrophobic interactions play key role in the stability of amyloid oligomers. In silico mutations confirmed that Asn2, Phe3 and Ile6 are key residues that stabilize the aggregation for the NNFGAIL segment. Single and double mutation of these specific amino acids with beta breaker amino acids indicate the dramatic stability loss for the double-layer oligomer. These results are helpful to understand the factors important for the early peptides aggregation into amyloid fibril-like assemblies. Results from this work indicate that the most important forces that are responsible for the stability of the peptide-peptide complexes are hydrophobic and Van der Waals ones, while electrostatic component of the solvation energy destabilizes the complexes. Based on per residue decomposition of the binding free energy, mutation of the residues from the interface region decreases their contributions, while for the terminal residues the contributions remain the same. The study of the wild type

and mutants in explicit solvent may provide valuable insight to guided future amyloid aggregation inhibitor design efforts.

Acknowledgments This work was supported in part by the National Science Foundation (CHE0832622), and used the resources of the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The authors are thankful to Dr. Michael Sawaya for providing the initial aggregate models, and for his helpful discussions. WMB also thanks Dr. Zhengji Zhao of NERSC for her help with the software installation.

References

1. Chiti F, Dobson CM (2006) *Annu Rev Biochem* 75:333–366. doi:10.1146/annurev.biochem.75.101304.123901
2. Kitamura A, Kubota H (2010) *FEBS J* 277:1369–1379. doi:10.1111/j.1742-4658.2010.07570.x
3. Antzutkin ON, Leapman RD, Balbach JJ, Tycko R (2002) *Biochemistry* 41:15436–15450. doi:10.1021/bi0204185
4. Cooper GJS, Willis AC, Clark A, Turner RC, Sim RB, Reid KBM (1987) *Proc Natl Acad Sci USA* 84:8628–8632
5. Westermark P, Wernstedt C, Wilander E, Hayden DW, Obrien TD, Johnson KH (1987) *Proc Natl Acad Sci USA* 84:3881–3885
6. Hoppener JWM, Oosterwijk C, Nieuwenhuis MG, Posthuma G, Thijssen JHH, Vroom TM, Ahren B, Lips CJM (1999) *Diabetologia* 42:427–434
7. Glenner GG, Eanes ED, Wiley CA (1988) *Biochem Biophys Res Commun* 155:608–614
8. Nilsson MR, Raleigh DP (1999) *J Mol Biol* 294:1375–1385
9. Sawaya MR, Sambashivan S, Nelson R, Ivanova MI, Sievers SA, Apostol MI, Thompson MJ, Balbirnie M, Wiltzius JJW, McFarlane HT, Madsen AO, Riekel C, Eisenberg D (2007) *Nature* 447:453–457. doi:10.1038/nature05695
10. Griffiths JM, Ashburn TT, Auger M, Costa PR, Griffin RG, Lansbury PT (1995) *J Am Chem Soc* 117:3539–3546
11. Tenidis K, Waldner M, Bernhagen J, Fischle W, Bergmann M, Weber M, Merkle ML, Voelter W, Brunner H, Kapurniotu A (2000) *J Mol Biol* 295:1055–1071
12. Jaikaran E, Higham CE, Serpell LC, Zurdo J, Gross M, Clark A, Fraser PE (2001) *J Mol Biol* 308:515–525. doi:10.1006/jmbi.2001.4593
13. Wiltzius JJW, Sievers SA, Sawaya MR, Cascio D, Popov D, Riekel C, Eisenberg D (2008) *Protein Sci* 17:1467–1474. doi:10.1110/ps.036509.108
14. Andreatto E, Yan LM, Tatarek-Nossol M, Velkova A, Frank R, Kapurniotu A (2010) *Angew Chem Int Ed* 49:3081–3085. doi:10.1002/anie.200904902
15. De Simone A, Pedone C, Vitagliano L (2008) *Biochem Biophys Res Commun* 366:800–806. doi:10.1016/j.bbrc.2007.12.047
16. Williams AD, Portelius E, Kheterpal I, Guo JT, Cook KD, Xu Y, Wetzel R (2004) *J Mol Biol* 335:833–842. doi:10.1016/j.jmb.2003.11.008
17. Tzotzos S, Doig A (2010) *Protein Sci* 19:327–348. doi:10.1002/pro.314
18. Hawkes CA, Ng V, McLaurin J (2009) *Drug Develop Res* 70:111–124. doi:10.1002/ddr.20290
19. Potter KJ, Scrocchi LA, Warnock GL, Ao ZL, Younker MA, Rosenberg L, Lipsett M, Verchere CB, Fraser PE (2009) *Biochim Biophys Acta Gen Subj* 1790:566–574. doi:10.1016/j.bbagen.2009.02.013

20. Moriarty DF, Raleigh DP (1999) *Biochemistry* 38:1811–1818
21. Abedini A, Raleigh DP (2006) *J Mol Biol* 355:274–281. doi:10.1016/j.jmb.2005.10.052
22. Vitagliano L, Stanzione F, De Simone A, Esposito L (2009) *Biopolymers* 91:1161–1171. doi:10.1002/bip.21182
23. Esposito L, Pedone C, Vitagliano L (2006) *Proc Natl Acad Sci USA* 103:11533–11538. doi:10.1073/pnas.0602345103
24. Zheng J (2008) MB, Chang Y, Nussinov R. *Front Biosci* 13:3919–3930
25. Wu C, Lei HX, Duan Y (2005) *J Am Chem Soc* 127:13530–13537. doi:10.1021/ja050767x
26. Wu C, Lei HX, Wang ZX, Zhang W, Duan Y (2006) *Biophys J* 91:3664–3672. doi:10.1529/biophysj.106.081877
27. Raman EP, Takeda T, Klimov DK (2009) *Biophys J* 97:2070–2079. doi:10.1016/j.bpj.2009.07.032
28. Berhanu WM, Masunov AE (2010) *Biophys Chem* 149:12–21. doi:10.1016/j.bpc.2010.03.003
29. Wang J, Tan CH, Chen HF, Luo R (2008) *Biophys J* 95:5037–5047. doi:10.1529/biophysj.108.131672
30. Xu WX, Ping J, Li WF, Mu YG (2009) *J Chem Phys* 130:164709. doi:10.1063/1.3123532
31. Xu YC, Shen JJ, Luo XM, Zhu WL, Chen KX, Ma JP, Jiang HL (2005) *Proc Natl Acad Sci USA* 102:5403–5407. doi:10.1073/pnas.0501218102
32. Zanuy D, Nussinov R (2003) *J Mol Biol* 329:565–584. doi:10.1016/s0022-2836(03)00491-1
33. Zanuy D, Porat Y, Gazit E, Nussinov R (2004) *Structure* 12:439–455. doi:10.1016/j.str.2004.02.002
34. Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, Duke RE, Luo R, Walker RC, Zhang W, Merz KM, Roberts B, Wang B, Hayik S, Roitberg A, Seabra G, Kolossváry I, Wong KF, Paesani F, Vanicek J, Liu J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Cai Q, Ye X, Wang J, Hsieh MJ, Cui G, Roe DR, Mathews DH, Seetin MG, Sagui C, Babin V, Luchko T, Gusarov S, Kovalenko A, Kollman PA (2010) AMBER 11 University of California, San Francisco
35. Sasahara K, Naiki H, Goto Y (2005) *J Mol Biol* 352:700–711. doi:10.1016/j.jmb.2005.07.033
36. Meersman F, Dobson CM (2006) *BBA-Proteins Proteom* 1764:452–460. doi:10.1016/j.bbapap.2005.10.021
37. Darden T, York D, Pedersen L (1993) *J Chem Phys* 98:10089–10092
38. 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. doi:10.1002/jcc.20290
39. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) *J Comput Chem* 23:327–341
40. Humphrey W, Dalke A, Schulten K (1996) *J Mol Graph* 14:33–38
41. Fabiola F, Bertram R, Korostelev A, Chapman MS (2002) *Protein Sci* 11:1415–1423. doi:10.1110/ps.4890102
42. Paparcone R, Pires MA, Buehler MJ (2010) *Biochemistry* 49:8967–8977. doi:10.1021/bi100953t
43. Center SDSUc (2009)
44. Zheng J, Jang H, Ma B, Tsai CJ, Nussinov R (2007) *Biophys J* 93:3046–3057. doi:10.1529/biophysj.107.110700
45. Buchete NV, Hummer G (2007) *Biophys J* 92:3032–3039. doi:10.1529/biophysj.106.100404
46. Huet A, Derreumaux P (2006) *Biophys J* 91:3829–3840. doi:10.1526/biophysj.106.090993
47. Porat Y, Mazor Y, Efrat S, Gazit E (2004) *Biochemistry* 43:14454–14462. doi:10.1021/bi048582a
48. Zheng J, Ma BY, Tsai CJ, Nussinov R (2006) *Biophys J* 91:824–833. doi:10.1529/biophysj.106.083246
49. Azriel R, Gazit E (2001) *J Biol Chem* 276:34156–34161
50. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM (2003) *Nature* 424:805–808. doi:10.1038/nature01891
51. Porat Y, Stepensky A, Ding FX, Naider F, Gazit E (2003) *Biopolymers* 69:161–164. doi:10.1002/bip.10386
52. Bartolini M, Andrisano V (2010) *ChemBioChem* 11:1018–1035. doi:10.1002/cbic.200900666
53. Dasilva KA, Shaw JE, McLaurin J (2009) *Exp Neurol* 223:311–321
54. Pace CN (2009) *Nat Struct Mol Biol* 16:681–682. doi:10.1038/nsmb0709-681
55. Kabsch W, Sander C (1983) *Biopolymers* 22:2577–2637
56. 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) *Acc Chem Res* 33:889–897. doi:10.1021/ar000033j
57. Massova I, Kollman PA (1999) *J Am Chem Soc* 121:8133–8143
58. Campanera JM, Pouplana R (2010) *Molecules* 15:2730–2748. doi:10.3390/molecules15042730
59. Chong LT, Duan Y, Wang L, Massova I, Kollman PA (1999) *Proc Natl Acad Sci USA* 96:14330–14335
60. Wang JM, Morin P, Wang W, Kollman PA (2001) *J Am Chem Soc* 123:5221–5230. doi:10.1021/ja003834q
61. Gohlke H, Kiel C, Case DA (2003) *J Mol Biol* 330:891–913. doi:10.1016/s0022-2836(03)00610-7
62. Wang W, Kollman PA (2000) *J Mol Biol* 303:567–582. doi:10.1006/jmbi.2000.4057
63. Kuhn B, Kollman PA (2000) *J Am Chem Soc* 122:3909–3916
64. Lee TS, Kollman PA (2000) *J Am Chem Soc* 122:4385–4393
65. Huo S, Massova I, Kollman PA (2002) *J Comput Chem* 23:15–27
66. Lafont V, Schaefer M, Stote RH, Altschuh D, Dejaegere A (2007) *Proteins* 67:418–434. doi:10.1002/prot.21259
67. Zoete V, Meuwly M, Karplus M (2005) *Proteins* 61:79–93. doi:10.1002/prot.20528
68. Archontis G, Simonson T, Karplus M (2001) *J Mol Biol* 306:307–327. doi:10.1006/jmbi.2000.4285
69. Carrascal N, Green DF (2010) *J Phys Chem B* 114:5096–5116. doi:10.1021/jp910540z
70. Hensch ZS, Tidor B (1999) *Protein Sci* 8:1381–1392
71. Periole X, Rampioni A, Vendruscolo M, Mark AE (2009) *J Phys Chem B* 113:1728–1737. doi:10.1021/jp8078259
72. De Simone A, Esposito L, Pedone C, Vitagliano L (2008) *Biophys J* 95:1965–1973. doi:10.1529/biophysj.108.129213
73. Amijee H, Madine J, Middleton DA, Doig AJ (2009) *Biochem Soc Trans* 37:692–696. doi:10.1042/bst0370692
74. Sellin D, Yan LM, Kapurniotu A, Winter R (2010) *Biophys Chem* 150:73–79. doi:10.1016/j.bpc.2010.01.006

Theoretical study on the structural, vibrational, and thermodynamic properties of the $(\text{Br}_2\text{GaN}_3)_n$ ($n=1-4$) clusters

Qi-Ying Xia · Qing-Fu Lin · Wen-Wei Zhao

Received: 23 February 2011 / Accepted: 11 May 2011 / Published online: 28 May 2011
© Springer-Verlag 2011

Abstract The molecular geometries, vibrational properties, and thermodynamic properties of the clusters $(\text{Br}_2\text{GaN}_3)_n$ ($n=1-4$) were studied at the B3LYP/6-311+G* level. The optimized clusters $(\text{Br}_2\text{GaN}_3)_n$ ($n=2-4$) were all found to possess a cyclic structure consisting of Ga atoms bridged by the α -nitrogen of the azide groups. A discussion of the relationships between the geometrical parameters and the degree of oligomerization n is provided. Features in the IR spectra were assigned by vibrational analysis. Trends in thermodynamic properties with temperature and degree of oligomerization n are discussed. Thermodynamic analysis of the gas-phase reaction showed that the formation of the clusters $(\text{Br}_2\text{GaN}_3)_n$ ($n=2-4$) is thermodynamically favorable considering the enthalpies at 298.2 K. The calculated results for the Gibbs free energies were negative, which indicates that the oligomerizations can occur spontaneously at 298.2 K.

Keywords $(\text{Br}_2\text{GaN}_3)_n$ ($n=1-4$) clusters · Density functional theory (DFT) · Structural feature · IR spectra · Thermodynamic properties

Introduction

The chemical vapor deposition (CVD) of gallium nitride (GaN) has attracted extensive interest because of its applications in high-power, high-efficiency optoelectronic devices [1]. The most successful method of producing GaN is to react trimethylgallium or triethylgallium with ammonia at a temperature in excess of 900 °C [2]. Such elevated growth temperatures, however, result in thermal stresses in the cooling films as well as a loss of stoichiometry due to nitrogen deficiency. Alternative synthetic routes to stoichiometric gallium nitride materials involve the use of single-source precursors containing strong Ga–N bonds. Particularly promising are precursors that contain the azide (N_3) ligand as the nitrogen source. Organometallic gallium azides such as $(\text{R}_2\text{GaN}_3)_3$ ($\text{R} = \text{CH}_3, \text{C}_2\text{H}_5$), $[(\text{CH}_3)\text{ClGaN}_3]_4$ and $[(\text{CH}_3)\text{BrGaN}_3]_3$ have been used to deposit stoichiometric GaN of reasonable crystal quality and chemical purity [3–5]. To eliminate the possibility of carbon inclusion, McMurrin et al. investigated several related routes for GaN synthesis that utilize a new class of inorganic azide compounds which incorporate hydrogen and halide ligands [6–11] instead of organic groups. For example, GaN films were prepared at temperatures as low as 200 °C using $(\text{H}_2\text{GaN}_3)_n$ ($n=2-3$) as the precursor in chemical vapor deposition processes [7, 8]. Moreover, they also used ab initio methods and normal-mode analysis to calculate the vibrational properties of the trimeric $(\text{H}_2\text{GaN}_3)_3$ C_{3v} and dimeric $(\text{H}_2\text{GaN}_3)_2$ D_{2h} forms of the compound. The vapor IR and mass spectra were consistent with the trimeric model of C_{3v} symmetry [8].

Q.-Y. Xia (✉) · Q.-F. Lin · W.-W. Zhao
School of Chemistry and Resources Environment,
Linyi University,
Linyi,
Shandong 276005, China
e-mail: xiaqiyang@163.com

Despite these extensive experimental investigations, the reliable structures, IR spectra and thermodynamic properties of many gas-phase precursors are unknown. One of the reasons for this is the difficulty associated with experimental detection. A good alternative route to investigating a wide variety of potential precursors is offered by theoretical computation. Thus, motivated by and based on our previous studies on the clusters $(\text{H}_2\text{Ga}\text{aN}_3)_n$ ($n=1-4$), $[(\text{CH}_3)_2\text{Ga}\text{aN}_3]_n$, and $[(\text{CH}_3\text{CH}_2)_2\text{Ga}\text{aN}_3]_n$ ($n=1-3$) [12–14], we performed density functional theory (DFT) investigations on $(\text{Br}_2\text{Ga}\text{aN}_3)_n$ ($n=1-4$) clusters. We hope that the calculated vibrational spectra prove to be useful reference data for experimentalists. The thermodynamic properties of $(\text{Br}_2\text{Ga}\text{aN}_3)_n$ ($n=1-4$) clusters are expected to provide useful information for the molecular design of novel gallium azides. In addition, the results shed some light on the emergence of bulk-like behavior with increasing cluster size.

Computational methods

All of the clusters in Fig. 1 were generated using the ChemBats3D software and fully optimized by the Beryny method at the DFT-B3LYP level with the 6-311+G* basis set [15, 16]. To characterize the nature of the stationary points and to determine the zero-point vibrational energy corrections, harmonic vibrational analyses were performed subsequently on each optimized structure at the same level with the Gaussian 03 program [17]. Since the DFT-calculated harmonic vibrational frequencies are usually larger than those observed experimentally, they were scaled by a factor of 0.96 [18]. On the basis of the principle of statistical thermodynamics [19], the standard molar heat capacity ($C_{p,m}^0$), standard molar entropy (S_m^0), and standard molar enthalpy (H_m^0) from 200 to 800 K were derived from the scaled frequencies using a program written in-house.

Results and discussion

Geometric structure

All of the optimized structures were characterized by the harmonic vibrational analyses as true local energy minima on the potential energy surfaces without any imaginary frequency. The monomer $\text{Br}_2\text{Ga}\text{aN}_3$ is a planar molecule with C_s symmetry (Fig. 1, **1A**), which is studied only as a starting point for the oligomerizations. The dimer $(\text{Br}_2\text{Ga}\text{aN}_3)_2$ with D_{2h} symmetry is produced by the head-to-tail dimerization of the $\text{Br}_2\text{Ga}\text{aN}_3$ monomers (Fig. 1, **2A**). The dimer $(\text{Br}_2\text{Ga}\text{aN}_3)_2$ with C_{2v} symmetry has been

reported experimentally previously [6]. Two types of trimer $(\text{Br}_2\text{Ga}\text{aN}_3)_3$ were obtained in the present study: a boat-like conformation **3A** (symmetric C_s) and a chair-like conformation **3B** (symmetric C_{3v}); the N_2 portion of the N_3 group is omitted from these figures to improve clarity. The trimer $(\text{Br}_2\text{Ga}\text{aN}_3)_3$ has not been reported theoretically or experimentally previously. However, trimeric structures of other group 13 azides have been observed, such as $(\text{H}_2\text{Ga}\text{aN}_3)_3$ [7, 8, 12] $(\text{Cl}_2\text{Ga}\text{aN}_3)_3$ [11], $[(\text{CH}_3)\text{Br}\text{Ga}\text{aN}_3]_3$ [5], $(\text{H}_2\text{Al}\text{N}_3)_3$ [20], and $[(\text{CH}_3)_2\text{Al}\text{N}_3]_3$ [21]. Four optimized tetramers $(\text{Br}_2\text{Ga}\text{aN}_3)_4$ possessing C_s , S_4 , C_i and C_2 symmetry with Ga_4N_4 core structures are presented in Fig. 1 (**4A–4D**; the N_2 portion of the N_3 group is again omitted to improve clarity). Among them, a structure with S_4 symmetry that has the N_3 alternatively up and down has

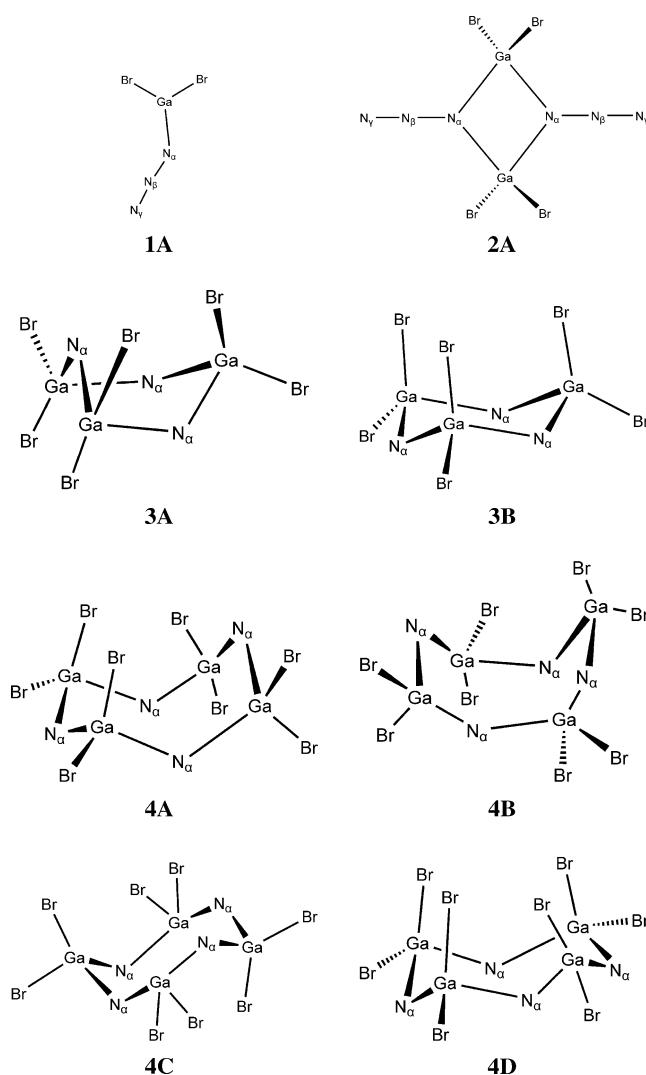


Fig. 1 Structures of the clusters. $\text{Br}_2\text{Ga}\text{NaN}_3$ monomer of C_s symmetry (**1A**); $(\text{Br}_2\text{Ga}\text{NaN}_3)_2$ dimer of D_{2h} symmetry (**2A**); $(\text{Br}_2\text{Ga}\text{NaN}_3)_3$ trimers of C_s (**3A**) and C_{3v} (**3B**) symmetry; $(\text{Br}_2\text{Ga}\text{NaN}_3)_4$ tetramers of C_s (**4A**), S_4 (**4B**), C_i (**4C**), and C_2 (**4D**) symmetry. The N_2 portion of the N_3 groups of the trimers and tetramers is omitted to improve clarity

been suggested for the tetramers $[(CH_3)ClGaN_3]_4$ [5] and $[HClGaN_3]_4$ [9] among the gallium azides, and as far as we are aware, the other three structures have not been reported previously [12].

The ranges of the optimized bond lengths and angles are presented in Table 1. Obviously, the $N_\alpha-N_\beta$ lengths are longer than the $N_\beta-N_\gamma$ distances for the investigated clusters $(Br_2GaN_3)_n$ ($n=1-4$). This can be interpreted as a higher bond order for the terminal N–N bond, showing pre-formation of the N_2 molecule. Moreover, we also find that the degree of oligomerization n is an important influence on the geometry. As n increases, the average $N_\beta-N_\gamma$ bond length decreases, while the average $N_\alpha-N_\beta$, Ga–Br and Ga– N_α bond lengths increase. These trends in average $N_\alpha-N_\beta$, Ga– N_α , and $N_\beta-N_\gamma$ bond lengths with n appear to be similar to those reported for the clusters $(H_2GaN_3)_n$ ($n=1-4$), $[(CH_3)_2GaN_3]_n$, and $[(CH_3CH_2)_2GaN_3]_n$ ($n=1-3$) [12–14]. The fact that the $N_\alpha-N_\beta$ and Ga–Br bond lengths increase shows it would be relatively easy to eliminate N_2 ($N_\beta-N_\gamma$) and Br^- groups to yield GaN material. For the monomer, the azide group is slightly bent, with $N_\alpha-N_\beta-N_\gamma$ angles of 174.4° . For $(Br_2GaN_3)_n$ ($n=2-4$), the azide groups are nearly linear, with $N_\alpha-N_\beta-N_\gamma$ angles in the range of $179.4-180.0^\circ$. The bond angles for $N_\alpha-Ga-N_\alpha$ and Ga– N_α –Ga increase as the cyclic clusters enlarge, while the $N_\beta-N_\alpha$ –Ga bond angle decreases. The Ga– N_α –Ga bond angles in the cyclic clusters are consistently larger than the $N_\alpha-Ga-N_\alpha$ bond angles, with the difference increasing with the size of the cluster.

Table 1 also reports the total energies (E) and zero point energies (ZPE) obtained at the B3LYP/6-311+G* level. The energies show that the order of stability is as follows: **3A**>**3B** and **4B**>**4C**>**4A**>**4D**. The most attractive trimer, which possesses C_s symmetry, is about $13.56 \text{ kJ mol}^{-1}$ lower in energy than that of the trimer possessing C_{3v} symmetry. For the tetramer, the difference in energy is about $8.30\sim 34.61 \text{ kJ mol}^{-1}$.

IR spectrum

As is well known, IR spectra are not only basic features of compounds, but they also provide an effective way to analyze or identify substances. The IR spectrum of a compound is also directly related to its thermodynamic properties. However, to the best of our knowledge, there are no experimental IR data for the title compounds. Therefore, it is important to predict IR spectra for both theoretical and practical reasons. Figure 2 provides the calculated IR spectra of $(Br_2GaN_3)_n$ ($n=1-4$) clusters at the B3LYP/6-311+G* level. Due to the complexity of vibrational modes, it is difficult to assign all of the bands in each spectrum.

Table 1 Ranges of the bond lengths (Å), bond angles ($^\circ$) and energies (kJ mol^{-1}) for the title clusters optimized at the B3LYP/6-311+G* level

	1A	2A	3A	3B	4A	4B	4C	4D
$N_\beta-N_\gamma$	1.133	1.127	1.121–1.124	1.124	1.122–1.123	1.122	1.123–1.124	1.123–1.124
$N_\alpha-N_\beta$	1.228	1.231	1.243–1.249	1.245	1.247–1.250	1.249	1.247	1.247–1.248
Ga– N_α	1.858	2.019	2.020–2.025	2.029	2.028–2.044	2.009–2.036	2.022–2.033	2.034–2.047
Ga–Br	2.273–2.289	2.301	2.302–2.308	2.293–2.313	2.291–2.315	2.305–2.311	2.296–2.310	2.283–2.320
$N_\alpha-N_\beta-N_\gamma$	174.4	180.0	179.6–179.7	179.8	179.9–180.0	179.5	179.8–179.8	179.4–179.9
Ga– N_α –Ga		101.7	124.5–128.4	127.8	129.1–129.7	127.7	127.9–130.1	129.5–134.2
N_α –Ga– N_α		78.3	96.1–97.8	98.4	99.4–100.7	101.0	100.2–100.9	97.5–100.5
$N_\beta-N_\alpha$ –Ga	123.7	129.1	115.8–118.2	116.1	114.0–115.7	114.7–116.6	114.3–116.6	111.4–115.5
E	–19002153.87	–38004455.71	–57006705.84	–57006692.28	–76008919.68	–76008939.81	–76008930.98	–76008905.20
ZPE	38.54	82.50	124.77	124.38	166.02	167.03	166.49	165.54

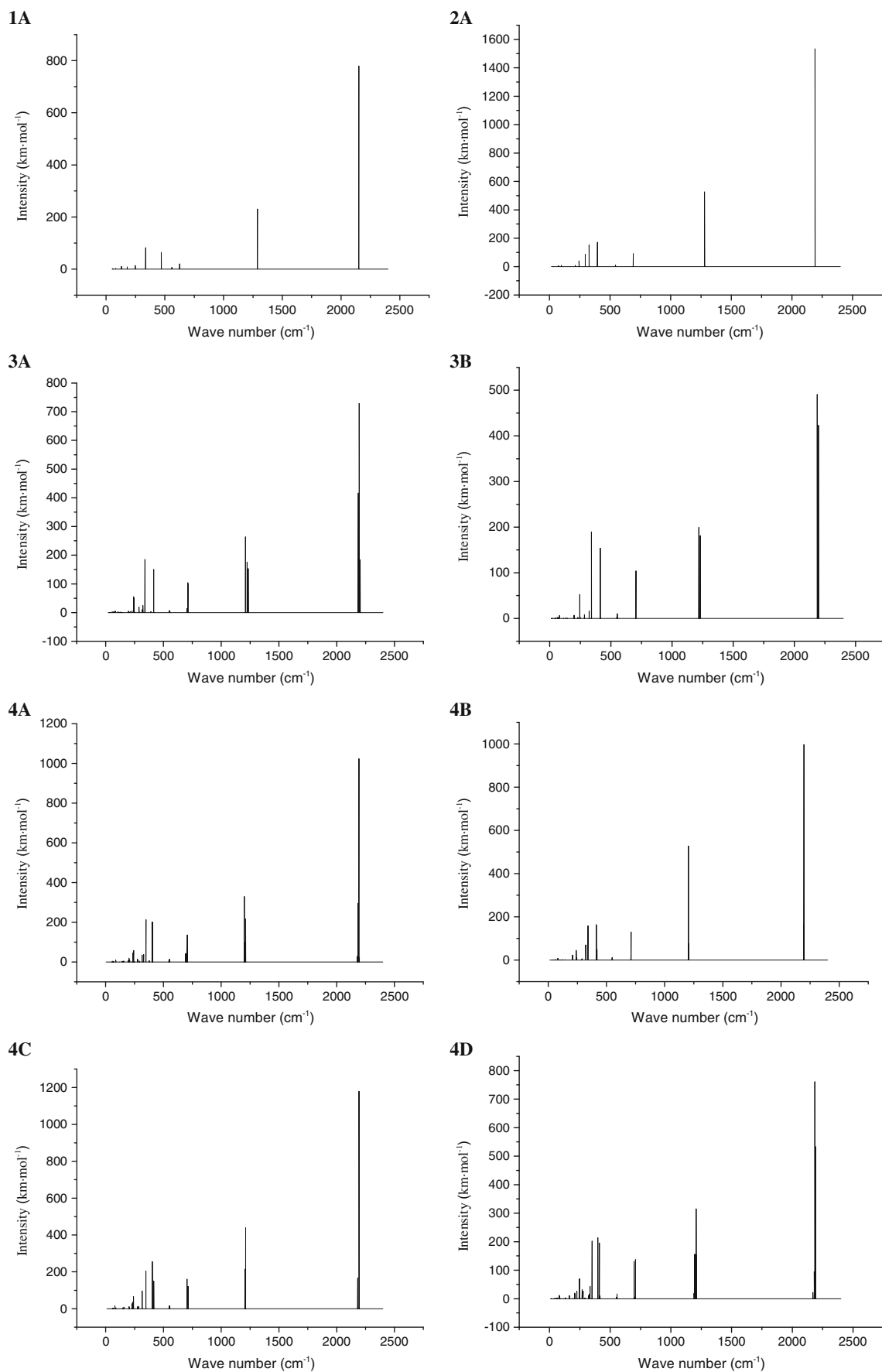


Fig. 2 The calculated IR spectra of $(\text{Br}_2\text{GaN}_3)_n$ ($n=1-4$) clusters at the B3LYP/6-311+G* level

Therefore, only some of the typical vibrational modes were analyzed and are discussed below.

It is clear from Fig. 2 that there are three main characteristic regions for the title compounds: those associated with N_3 asymmetric and symmetric stretching and the fingerprint region. The modes at $2152.3\sim 2202.5\text{ cm}^{-1}$ with the strongest absorption intensities are associated with characteristic N_3 asymmetric stretching vibrations, and in this region, the number of vibrations equals the number of azido groups. For example, **3A** has three bands at 2186.3, 2194.7, and 2202.5 cm^{-1} . The characteristic N_3 symmetric stretching modes are located in the frequency range of $1190.8\sim 1306.5\text{ cm}^{-1}$ with strong intensities. In this region, the number of vibrations again equals the number of azido groups. For example, **4B** has four bands at 1204.4, 1206.7, 1206.7, and 1214.0 cm^{-1} . The weak peak below 1150.0 cm^{-1} is the fingerprint region, which is associated with ring stretching, the asymmetric and symmetric stretching of Ga–Br, the wagging and scissoring of Br–Ga–Br, and N_3 deformation vibrations. This region can be used to identify isomers.

The clusters $(\text{Br}_2\text{GaN}_3)_n$ ($n=1-4$) have similar vibrational modes associated with the characteristic $\nu(\text{N}_3)$, $\nu_{\text{as}}(\text{N}_3)$ and fingerprint regions, but the degree of oligomerization n has an effect on the vibrational frequencies and intensities. The vibrations due to N_3 asymmetric stretching move to higher frequencies (the hypsochromic phenomenon) as the cluster becomes larger, but the vibrations due to N_3 symmetric stretching move to lower frequencies (bathochromic phenomenon).

Thermodynamic properties

Using the calculated IR spectra, the thermodynamic properties ($C_{p,m}^0$, S_m^0 and H_m^0) ranging from 200 to 800 K were also obtained based on the principle of statistic thermodynamics (see Table 2). Since there are no corresponding experimental values, no comparison can be made with them. The calculated thermodynamic functions and the established dependence of each on the temperature and the degree of oligomerization n would be helpful for further studies on other physical, chemical, and energetic properties of the gallium azides.

From these data, it was found that all of the thermodynamic functions clearly increase as the temper-

Table 2 Thermodynamic properties of the title clusters at different temperatures^a

	T	200	298.2	400	500	600	700	800
1A	$C_{p,m}^0$	92.30	102.90	109.55	114.02	117.37	119.97	122.04
	S_m^0	365.69	404.71	435.95	460.90	482.00	500.29	516.45
	H_m^0	13.78	23.41	34.25	45.44	57.02	68.89	81.00
2A	$C_{p,m}^0$	196.14	219.82	234.06	243.45	250.38	255.74	259.99
	S_m^0	553.02	636.23	702.98	756.27	801.30	840.31	874.75
	H_m^0	27.54	48.07	71.24	95.14	119.85	145.16	170.96
3A	$C_{p,m}^0$	292.72	329.02	351.05	365.47	376.00	384.08	390.43
	S_m^0	683.40	807.77	907.78	987.76	1055.36	1113.96	1165.67
	H_m^0	39.76	70.45	105.17	141.04	178.13	216.16	254.89
3B	$C_{p,m}^0$	301.55	337.77	359.70	374.04	384.52	392.56	398.88
	S_m^0	716.33	844.20	946.78	1028.67	1097.84	1157.74	1210.59
	H_m^0	41.46	73.01	108.62	145.35	183.30	222.17	261.75
4A	$C_{p,m}^0$	407.86	456.25	485.55	504.67	518.62	529.29	537.68
	S_m^0	886.47	1059.31	1197.81	1308.34	1401.64	1482.42	1553.67
	H_m^0	55.66	98.31	146.38	195.95	247.15	299.57	352.93
4B	$C_{p,m}^0$	406.83	455.46	484.95	504.19	518.22	528.95	537.38
	S_m^0	868.79	1041.26	1179.56	1289.97	1383.19	1463.91	1535.12
	H_m^0	55.39	97.95	145.96	195.47	246.62	299.01	352.34
4C	$C_{p,m}^0$	407.16	455.65	485.07	504.31	518.33	529.06	537.50
	S_m^0	880.19	1052.76	1191.10	1301.53	1394.78	1475.52	1546.74
	H_m^0	55.62	98.20	146.22	195.75	246.91	299.30	352.65
4D	$C_{p,m}^0$	407.96	456.28	485.60	504.75	518.70	529.39	537.78
	S_m^0	891.17	1064.03	1202.55	1313.08	1406.40	1487.20	1558.45
	H_m^0	55.92	98.57	146.65	196.22	247.43	299.86	353.23

^a Units: T (K); $C_{p,m}^0$ ($\text{J mol}^{-1}\text{ K}^{-1}$); S_m^0 ($\text{J mol}^{-1}\text{ K}^{-1}$); H_m^0 (kJ mol^{-1})

ature increases. This is because the main contributions to the thermodynamic functions come from the translations and rotations of molecules at lower temperatures, whereas the vibrational motion is intensified at higher temperatures and contributes more to the thermodynamic functions. Using monomer **1A** as an example, the relationships between the thermodynamic functions and the temperature (T) in the range 200–800 K can be expressed as follows:

$$\begin{aligned} C_{p,m}^0 &= 71.5500 + 0.1242T - 7.7444 \times 10^{-5}T^2 \\ S_m^0 &= 82.8199 + 0.4679T - 2.2214 \times 10^{-4}T^2 \\ H_m^0 &= -5.3920 + 0.0907T + 2.1806 \times 10^{-5}T^2 \end{aligned}$$

and the correlation coefficients R^2 are 0.9933, 0.9991, and 1.0000, respectively.

Meanwhile,

$$\begin{aligned} dC_{p,m}^0/dT &= 0.1242 - 1.5489 \times 10^{-4}T \\ dS_m^0/dT &= 0.4679 - 4.4428 \times 10^{-4}T \\ dH_m^0/dT &= 0.0907 + 4.3612 \times 10^{-5}T \end{aligned}$$

It is obvious that the gradients of $C_{p,m}^0$ and S_m^0 decrease as the temperature increases, while that of H_m^0 constantly increases. Isomers have similar thermodynamic function values at the same temperature due to the fact that they have similar geometric and electronic structures.

In addition, all of the thermodynamic functions increase as the degree of oligomerization n increases. The n -dependent relations for $C_{p,m}^0$, S_m^0 , and H_m^0 at 298.2 K can be expressed as follows (where the correlation coefficients are all more than 0.99):

$$\begin{aligned} C_{p,m}^0 &= -14.92 + 116.69n \\ S_m^0 &= 202.20 + 208.12n \\ H_m^0 &= -1.53 + 24.60n \end{aligned}$$

On average, $C_{p,m}^0$, S_m^0 and H_m^0 increase by 116.69 J mol⁻¹ K⁻¹, 208.12 J mol⁻¹ K⁻¹, and 24.60 kJ mol⁻¹, respectively, when another Br₂GaN₃ is added.

Based on the calculated thermodynamic functions, the theoretical entropies (ΔS^0), enthalpies (ΔH^0), and Gibbs free energies (ΔG^0) of various oligomerizations in the

Br₂GaN₃+ Br₂GaN₃ system at 298.2 K were evaluated and are compiled in Table 3. The values show that the order of stability is **3A**>**3B** and **4B**>**4C**>**4A**>**4D**, which is consistent with the results for the energies. All oligomerizations are disfavored by the entropy at 298.2 K, as shown in Table 3. The oligomerization enthalpies are negative at 298.2 K, which reveals that the oligomerizations are thermodynamically favorable. The Gibbs free energy (ΔG^0) at a given temperature was evaluated using the standard values for the enthalpy and entropy according to the equation $\Delta G^0 = \Delta H^0 - T\Delta S^0$. The values of ΔG^0 are negative at 298.2 K, which indicates that all of the oligomerizations can occur spontaneously.

Conclusions

Based on our theoretical studies on (Br₂GaN₃)_{*n*} ($n=1-4$) clusters, the following conclusions can be drawn:

- (1) The DFT/B3LYP method with the 6-311+G* basis set that was used to calculate Br₂GaN₃ clusters consisting of up to four molecules predicts that Br₂GaN₃ oligomerizes via the α -N atoms to form cyclic-like clusters (Br₂GaN₃)_{*n*} ($n=2-4$). The N _{α} -N _{β} , Ga-Br, and Ga-N _{α} bond lengths and the Ga-N _{α} -Ga and N _{α} -Ga-N _{α} bond angles all increase with the size of the cluster, while the N _{β} -N _{γ} bond length and the N _{α} -N _{β} -Ga bond angle decrease.
- (2) The calculated IR spectra have three main characteristic regions: the N₃ asymmetric stretching, the N₃ symmetric stretching, and the complicated fingerprint regions.
- (3) Thermodynamic properties all increase quantitatively with increasing temperature and degree of oligomerization n . The gradients of $C_{p,m}^0$ and S_m^0 with temperature decrease, but that for H_m^0 increases.
- (4) The oligomerizations are thermodynamically favorable in the gas phase as judged by the enthalpies at 298.2 K, and all of the oligomerizations can occur spontaneously.

Table 3 Oligomerization entropies, enthalpies and Gibbs free energies at 298.2 K

	$\Delta S_{298.2}^0$ (J mol ⁻¹ K ⁻¹)	$\Delta H_{298.2}^0$ (kJ mol ⁻¹)	$\Delta G_{298.2}^0$ (kJ mol ⁻¹)
1A →(1/2) 2A	-86.60	-70.76	-44.93
1A →(1/3) 3A	-135.45	-78.41	-38.01
1A →(1/3) 3B	-123.31	-73.16	-36.39
1A →(1/4) 4A	-139.88	-72.03	-30.32
1A →(1/4) 4B	-144.40	-76.91	-33.85
1A →(1/4) 4C	-141.52	-74.77	-32.57
1A →(1/4) 4D	-138.70	-68.47	-27.11

References

1. Stringfellow GB (1999) *Organometallic vapor-phase epitaxy: theory and practice*, 2nd edn. Academic, New York
2. Amano H, Sawaki N, Akasaki I et al (1986) *Appl Phys Lett* 48:353–355
3. Kouvetakis J, Beach DB (1989) *Chem Mater* 1:476–478
4. Atwood DA, Jones RA, Cowley AH et al (1990) *J Organomet Chem* 394:C6–C8
5. Kouvetakis J, McMurran J, Steffek C et al (2000) *Inorg Chem* 39:3805–3809
6. Dehnicke K, Krueger N (1978) *Z Anorg Allg Chem* 444:71–76
7. McMurran J, Kouvetakis J (1999) *Appl Phys Lett* 74:883–885
8. McMurran J, Dai D, Balasubramanian K et al (1998) *Inorg Chem* 37:6638–6644
9. McMurran J, Kouvetakis J, Nesting DC et al (1998) *J Am Chem Soc* 120:5233–5237
10. Crozier PA, Tolle J, Kouvetakis J et al (2004) *Appl Phys Lett* 84:3441–3443
11. McMurran J, Todd M, Kouvetakis J (1996) *Appl Phys Lett* 69:203–205
12. Xia QY, Xiao HM, Ju XH et al (2004) *Int J Quantum Chem* 100:301–308
13. Xia QY, Xiao HM, Ju XH et al (2005) *Chem J Chin Univ* 26:922–926
14. Xia QY, Ma DX, Yang JM (2009) *Chin J Energ Mater* 17:260–264
15. Lee C, Yang W, Parr RG (1988) *Phys Rev B* 37:785–789
16. Becke AD (1993) *J Chem Phys* 98:5648–5652
17. Frisch MJ, Trucks GW, Schlegel HB et al (2003) *Gaussian 03*, revision B.03. Gaussian, Inc., Pittsburgh
18. Pople JA, Schlegel HB, Krishnan R et al (1981) *Int J Quant Chem Quant Chem Symp* 15:269–278
19. Hill TL (1960) *Introduction to statistic thermodynamics*. Addison-Wesley, New York
20. Xia QY, Xiao HM, Ju XH et al (2004) *J Phys Chem A* 108:2780–2786
21. Xia QY, Xiao HM, Ju XH et al (2004) *Chin J Chem* 22:1245–1249

Topological properties of some PhSeX compounds

Nora Beatriz Okulik · Alicia H. Jubert ·
Eduardo A. Castro

Received: 31 October 2010 / Accepted: 12 May 2011 / Published online: 28 May 2011
© Springer-Verlag 2011

Abstract A theoretical study on the series of compounds “PhSeX”, where Ph=phenyl, Se=selenium and X=Cl, Br, I, CN or SCN, is reported and compared with previously reported experimental data. The molecular geometry for these PhSeX compounds was studied at the DFT/B3LYP level of calculation by means of the 6-311G(d,p) basis set. The equilibrium structures of the molecules were dependent on the method employed to compare the known solid structures. A topological study of the calculated PhSeX species, based on the AIM theory, was carried out to gain a deeper insight into the bonding nature and to find an explanation for the structural diversity exhibited by these PhSeX compounds. The results reported herein illustrate the subtle differences in the solid-state structures of PhSeX compounds.

Keywords Topological study · DFT · Pseudohalogen · Selenium · PhSeX

N. B. Okulik (✉)
Universidad Nacional del Chaco Austral,
Cte. Fernández 755,
3700 Pcia. R. Sáenz Peña, Chaco, Argentina
e-mail: nora@unca.edu.ar

A. H. Jubert
CEQUINOR, Dpto. de Química, Facultad de Ciencias Exactas
47 y 115 y Facultad de Ingeniería 1 y 47, Universidad Nacional
de La Plata,
1900 Buenos Aires, Argentina

E. A. Castro
INIFTA, Dpto. de Química, Facultad de Ciencias Exactas,
Universidad Nacional de La Plata,
1900 Buenos Aires, Argentina

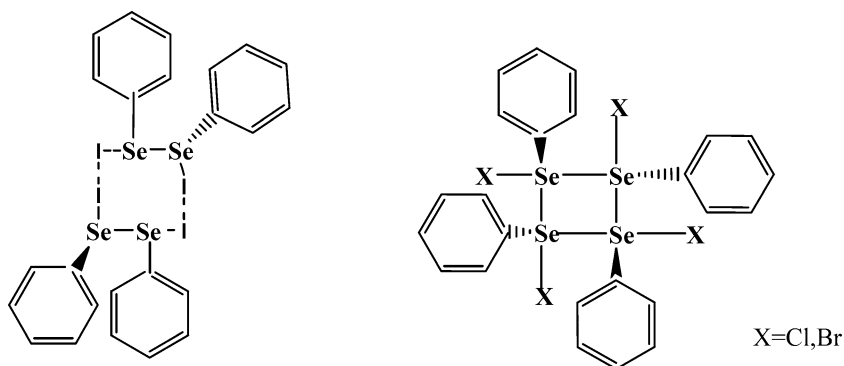
Introduction

Pseudohalogens are compounds that resemble the halogen elements, X_2 , in their chemistry, e.g., $(CN)_2$, cyanogen. Certain ions that have sufficient resemblance to halide ions are sometimes referred to as pseudohalide ions, e.g., N_3^- , SCN^- , CN^- , $SeCN^-$ [1]. The pseudohalide concept [2] has been used extensively in nonmetal chemistry in both experimental and theoretical research [3].

Phenylselenenyl halides and pseudohalides (such as PhSeCN) are versatile electrophilic reagents used in a variety of organic transformations [4, 5]. Due to this behavior, a considerable body of recent work has documented the structural delineation of the nature of various phenylselenenyl compounds.

Crystallographic studies on phenylselenenyl halides, PhSeX (X=Cl, Br, I) have revealed a number of structural motifs. For example, PhSeI exists as a centrosymmetric dimer, $(Ph_2Se_2I_2)_2$, in the solid state [6]. Two diphenyldiselenane molecules are coupled with two I_2 molecules in such a way that a slightly puckered eight-membered ring containing two Se-groups, and two I_2 -groups are formed. Since the angles in this ring are alternately approximately 90° and 180° , an almost square geometry results. Diphenyldiselenane coordinates a diiodine molecule, with one selenium atom acting as donor towards an iodine atom. One selenium atom acts as a donor towards iodine (Se–I: 2.992 Å), whilst the other behaves as a weak acceptor (Se–I: 3.588 Å). This charge transfer system differs from the analogous PhSeCl and PhSeBr (see Scheme 1). Both these latter compounds consist of a tetrameric “square” structure, $Ph_4Se_4Cl_4$ and $Ph_4Se_4Br_4$. The solid-state structures of PhSeCl [7] and PhSeBr [8] adopt a “square” motif where four PhSeX units are held together by weak selenium–selenium bonds to form Se_4 . The Se–Se–Se angles are,

Scheme 1 Different structural motifs observed in phenylselenenyl halides (PhSeX) compounds



however, close to the anticipated 90° for a square (motif) structure, forming an essentially planar ring. The conformation of the Se_4 square is such that the Se–X bonds lie in the Se_4 plane with two phenyl rings lying above (the plane), and two below the plane. The selenium atom therefore lies in a pseudo-trigonal bipyramidal “see-saw”. The structure is further linked through long X–X contacts to form planar sheets of selenium and bromine or chlorine atoms.

In contrast, the pseudohalogen derivatives PhSeCN and PhSeSCN consist of essentially monomeric units. In the former, the two independent molecules in the unit cell are loosely linked by $\text{Se}\cdots\text{N}$ contacts (see Fig. 2a) with $\text{Se}\cdots\text{N}\cdots\text{C}$ contact angles close to linearity. Both $\text{Se}\cdots\text{N}$ interactions are slightly shorter than the sum of the van der Waals radii for selenium and nitrogen. The Se–C bonds to the cyanide groups are considerably shorter than the Se–C bonds to the phenyl rings, reflecting the increased double bond character in the Se–C bond to the cyanide groups. In PhSeSCN, the thiocyanato group is coordinated to the selenium through the sulfur atom, and it would appear that the soft selenium center prefers to bind with sulfur, rather than with the harder nitrogen atom. The thiocyanate coordination mode was confirmed in solution by $^{13}\text{C}\{^1\text{H}\}$ NMR studies [9]. The Se–S bond length (2.221 Å) is close to the reported mean for covalent Se–S bonds (2.193 Å) [10] and the geometry of the SCN group is near linear, with S–C and $\text{C}\equiv\text{N}$ bond lengths similar to those observed for the free SCN^- ion in KSCN [11]. The bent geometry at both the selenium and sulfur atoms reflects the presence of lone pairs on both atoms, with angles S–Se–C close to 100° (see Fig. 1e). In the extended structure of PhSeSCN, individual molecules stack such that a weak interaction, $\text{Se}\cdots\text{N}$ of 3.348 Å is set up, between a Se atom of one molecule interacting with the N atom of another.

The shortest $\text{Se}\cdots\text{N}$ interaction between adjacent stacks (3.567 Å) is slightly longer than the sum of the van der Waals radii for selenium and nitrogen (3.45 Å). The $\text{Se}\cdots\text{N}$ contacts appear to be the dominant packing force in PhSeCN and PhSeSCN compounds, and are strong enough to preclude any formation of the weak lattice of Se_4 squares favored by $\text{Ph}_4\text{Se}_4\text{Cl}_4$ and $\text{Ph}_4\text{Se}_4\text{Br}_4$.

We previously reported a detailed theoretical and topological study of some pseudohalogen compounds [12, 13]. As part of a more general study on pseudohalide compounds, we report here a topological analysis of the phenyl series of compounds, PhSeX (X=halides and pseudohalides) in an attempt to find an explanation for the structural diversity exhibited by these PhSeX compounds.

Due to the fact that spectroscopic studies suggest that all these compounds exist as monomers in solution, in this paper we presented theoretical calculations in the gas phase in order to compare the results with experimental data. Different conformers of some of these compounds can be analyzed, although herein we report results from the study performed only on the lowest energy conformers.

A notable point in the present study is the use of density charge analysis based on the atoms-in-molecules ((AIM) theory to better explain chemical bonding character, since this procedure has proven extremely useful for this purpose.

Methods of calculation and computational details

Molecular geometries were optimized within the density functional theory (DFT) approach [14–16] at the B3LYP/6-311G(d,p) level. The B3LYP is a hybrid functional method based on the Becke’s three-parameter nonlocal exchange functional [17], with nonlocal correlation according to Lee et al. [18]. X-ray geometry was used as the starting input file. X-ray crystallographic data, with files in CIF format, for structures 1 and 2 were retrieved of the Cambridge Crystallographic Data Centre (CCDC 268776 and 268777) [19]. Densities used for topological analysis were obtained through single-point calculations on the above optimized geometries at the B3LYP/6-311++G(d,p) level. For iodine, we used the 6-311G(d) [20, 21] basis set, due to the fact that neither the 6-311G(d,p) nor 6-311++G(d,p) basis set are available for this atom. Frequency calculations were performed with the aim of assessing the nature of the stationary points. All calculations were carried out with the Gaussian 2003 package [22]. The analysis of the charge

electron density was performed using the PROAIM package [23].

Finally, in order to get the best possible agreement between calculated and observed structures, the root mean square deviation (RMSD) between the coordinates were calculated using the Qmol program [24]. With the results obtained, we think that some of the patterns followed in the analysis and interpretation of charge density of PhSeX compounds could be useful for the theoretical study of other derivatives.

AIM analysis

AIM theory [25], which is based on the critical points (CP) of the electronic density, $\rho(\mathbf{r})$, reveals insights into the nature of bonds. CPs are points where the gradient of the electronic density, $\nabla\rho(\mathbf{r})$, vanishes and are characterized by the three eigenvalues ($\lambda_1, \lambda_2, \lambda_3$) of the Hessian matrix of $\rho(\mathbf{r})$. The CPs are labeled as (r,s) according to their rank, r (number of nonzero eigenvalues), and signature, s (the algebraic sum of the signs of the eigenvalues).

Four types of CP are of interest in molecules: $(3, -3)$, $(3, -1)$, $(3, +1)$, and $(3, +3)$. A $(3, -3)$ point corresponds to a maximum in $\rho(\mathbf{r})$ and appears generally at nuclear positions. A $(3, +3)$ point indicates electronic charge depletion and is known as a cage CP. $(3, +1)$ points, or ring CPs, are merely saddle points. Finally, a $(3, -1)$ point, or bond critical point (BCP), is generally found between two neighboring nuclei indicating the existence of a bond between them.

Several properties that can be evaluated at the BCP constitute very powerful tools to classify the interactions between two fragments.

The two negative eigenvalues of the Hessian matrix (λ_1 and λ_2) at the BCP measure the degree of contraction of $\rho(\mathbf{r})$ perpendicular to the bond towards the CP, while the positive eigenvalue (λ_3) measures the degree of contraction parallel to the bond and from the BCP towards each of the neighboring nuclei. Different values of λ_1 and λ_2 at $(3,-1)$ BCPs denote an anisotropic spread of electrons quantified through the concept of ellipticity: $\varepsilon = \lambda_1/\lambda_2 - 1$, (with $\lambda_1 > \lambda_2$) where values of $\varepsilon \gg 1$ can be indicative of π bonding. Calculated properties of electronic density at the BCP are labeled with the subscript 'b' throughout this work.

In AIM theory, atomic interactions are classified according to two limiting behaviors, namely, shared interactions and closed-shell interactions. Shared interactions are characteristic of covalent and polarized bonds and their main features are large values of ρ_b , $\nabla^2\rho_b < 0$ and $E_b < 0$, E_b being the local electronic energy density of the system calculated at the BCP and defined as the sum of the local kinetic energy density and the local potential energy density, both computed at the BCP. In contrast, closed-shell interactions,

useful to describe ionic bonds, hydrogen bonds, and van der Waals interactions, are characterized by small values of ρ_b , $\nabla^2\rho_b > 0$ and $E_b > 0$.

Results and discussion

Geometric analysis

The optimized structures of single molecules of PhSeX (X = Cl, Br, I, CN and SCN) compounds in gas phase are shown in Fig. 1. Figure 2a shows the calculated structure of the two independent molecules in the unit cell of PhSeCN. The optimized structure of the square motif adopted by $\text{Ph}_4\text{Se}_4\text{Cl}_4$ in the solid-state structure is displayed in Fig. 2b. Selected bond lengths and bond angles are shown in both figures.

Although the lowest energy conformers of the calculated structures are highly symmetric, small differences between the crystal data and theoretical values are observed. These differences can be assigned to the fact that the X-ray structures were measured in a compacted crystalline form, whereas the calculations were performed for free isolated molecules.

A common measure of conformational similarity in structural bioinformatics is the minimum RMSD between the coordinates of two macromolecules. Using this idea, we think that the close structural relationship between the calculated and crystallographic observed structures is best illustrated with the RMSD overlay error than by comparing the paired lengths bonds, bond angles and torsion angles.

In this paper, we shall consider a general framework for feature comparison based on the following:

1. In the isolated molecules, alignment is considered good if (a) the molecules have a similar shape, and (b) their aromatic atoms and Se atoms overlap.
2. In the dimeric structure observed for PhSeCN, and in the tetrameric structure of $\text{Ph}_4\text{Se}_4\text{Cl}_4$, an alignment is good if (a) the molecules have a similar shape, and (b) their Se atoms and the groups attached to Se atoms overlap.

These statements can be justified as follows: in the first case the solid state interactions are rather different because of the possibility of rotation of the side chain. So, benzene ring superposition becomes more important. On the contrary, in dimeric and tetrameric structures the interaction between Se atoms and the groups attached to Se atoms are more important than the benzene rings.

The calculated halogenated structures agree satisfactorily with the corresponding experimental structures. For example, RMS is 0.05 in PhSeCl while it is only 0.025 for PhSeBr and 0.020 for PhSeI. A good alignment is also observed for other calculated isolated structures. RMS

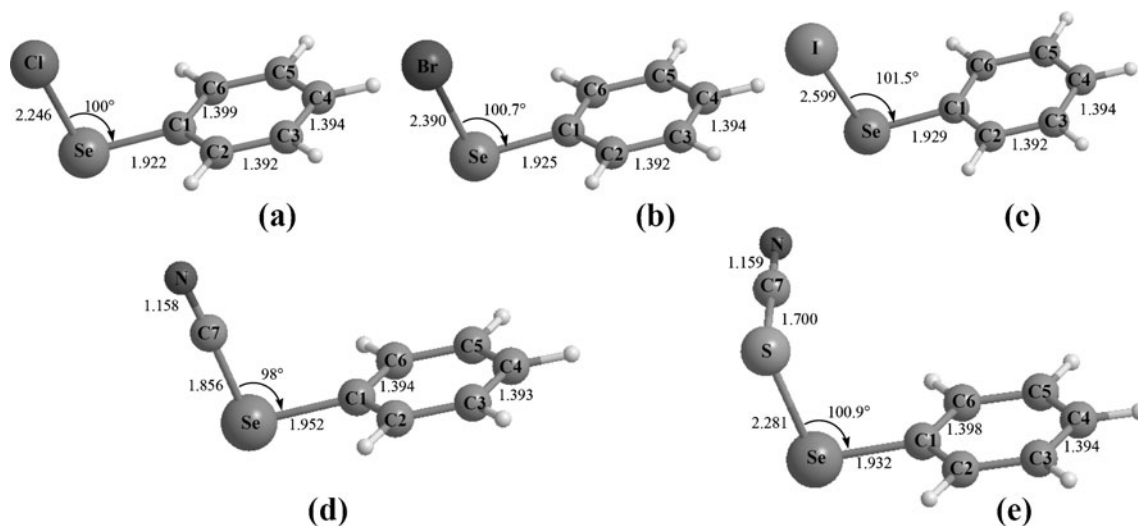


Fig. 1 Optimized structures of **a** PhSeCl, **b** PhSeBr, **c** PhSeI, **d** PhSeCN and **e** PhSeSCN calculated at the B3LYP/6-311++G(d,p) level. For iodine, we used the 6-311G(d) the B3LYP/6-311++G* level. The atomic labeling scheme and selected geometric parameters are indicated

overlay error is 0.025 for PhSeCN and only 0.017 for PhSeSCN. The optimized structures of the isolated molecules are superimposed on the crystallographic structures in

Fig. 2 Optimized structures of **a** the two independent molecules in the unit cell of PhSeCN, and **b** the square motif adopted by Ph₄Se₄Cl₄ in the solid-state structure, calculated at B3LYP/6-311++G(d,p). The atomic labeling scheme and bond lengths are indicated

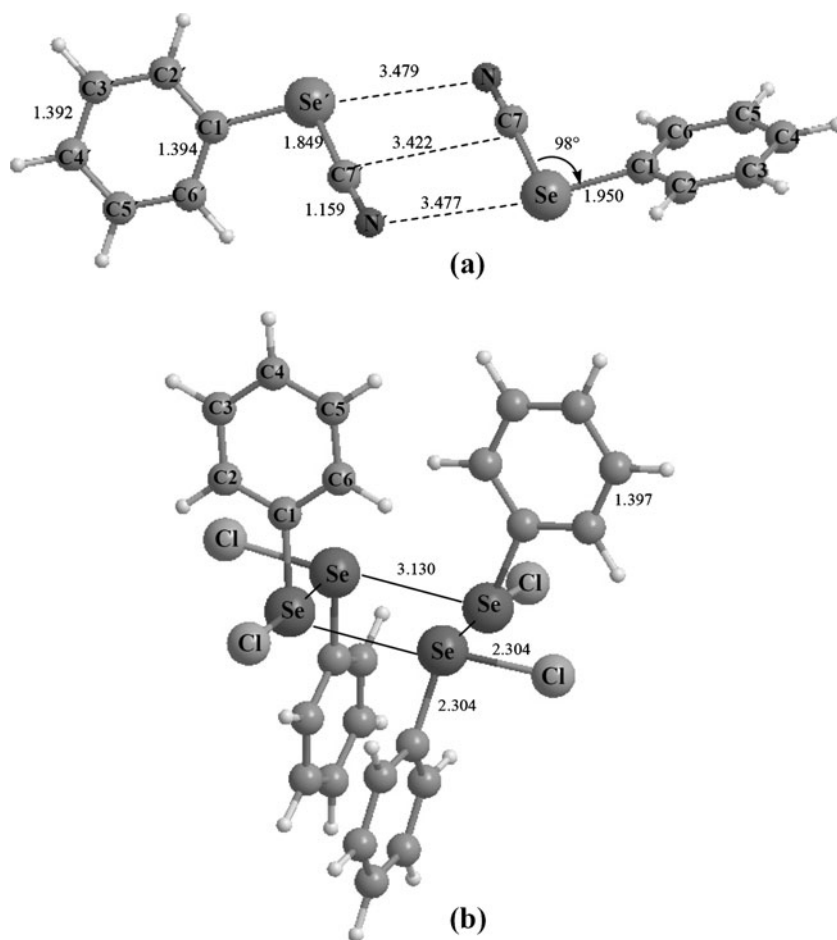
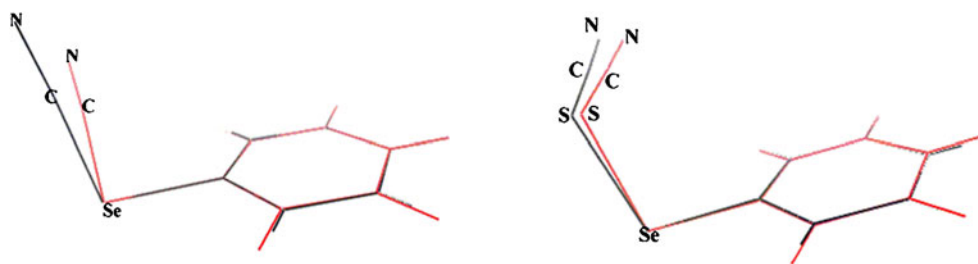


Fig. 3. It can be seen that the molecules have a similar shape and the benzene rings and Se atoms overlap completely.

Fig. 3 Overlaid structures of the calculated (*grey*) and solid state (*red*) structures of PhSeCN (*left*) and PhSeSCN (*right*)



The $\text{Se}\cdots\text{N}$ bonds between the two molecules of PhSeCN optimized as a dimer are somewhat longer (3.477 Å and 3.479 Å) than that in the crystal (3.329 or 3.444 Å). A RMSD value of 0.119 is obtained in the alignment of the dimeric structures. This low RMSD value reveals a good alignment due to the similar shape of the molecules and the fact that the Se, C and N atoms overlap (Fig. 4, left).

The increased RMS overlay error of the tetrameric structure of $\text{Ph}_4\text{Se}_4\text{Cl}_4$ (0.414) is most probably due to the increased number of atoms that must overlap; many of them belong to non-rigid structures but have four benzene rings with side chains with free rotation. According to the conventions adopted for comparison, in this case, Se atoms and the groups attached to Se atoms overlap each other but do not overlap benzene rings because preference was given to overlap of the side chains rather than rings, as illustrated in Fig. 4 (right). However, the RMSD value is acceptable for the alignment of both structures.

In summary, the results of these quantum chemical calculations overall correctly describe, to a good approximation, the experimentally observed peculiarities in molecular structures of the different species studied.

Topological analysis of electron density

Table 1 presents characteristics of the BCPs obtained from topological analysis of the electron density distributions of

the PhSeX (X=Cl, Br, I, CN and SCN) species studied. As mentioned in the section above on [Methods of calculation and computational details](#), the basis set employed for iodine [6-311G(d)] is different from that used for the rest of the atoms [6-311G(d,p)]. In order to check the sensibility of AIM results to the different basis set employed for iodine, for bromine compound, we performed AIM calculations with both basis sets, with similar results (Table 1). We consider, therefore, that we can safely analyze the trend of AIM results along the series, even when using a slightly different basis set for the iodine atom.

Topological analysis of BCPs in ρ of PhSeX (X=Cl, Br, I, CN, SCN) reveals that all bonds forming the phenyl ring correspond to covalent interactions, namely, a relatively large value for ρ_b and a negative value for $\nabla^2\rho_b$. The ellipticities of bonds forming the ring have relatively large numerical values, revealing their partial double bond character due electronic charge delocalization over the ring surface. The E_b values are negative as expected for covalent bonds. The topologic properties computed on C–C BCPs of the benzene ring are only slightly affected by the halogen atom, CN group or SCN group attached to selenium atom (at the same level of calculation in a C–C bond of benzene, they are: $\rho_b=0.3092$ a.u., $\nabla^2\rho_b=-0.8640$ a.u., $\epsilon=0.1999$ and $E_b=-0.3162$ a.u.).

Details of the electron density topology at the CP can provide more insight into the nature of a particular bond.

Fig. 4 Overlaid structures of the calculated (*grey*) and solid state (*red*) structures of dimer $(\text{PhSeCN})_2$ (*left*) and tetramer $\text{Ph}_4\text{Se}_4\text{Cl}_4$ (*right*)

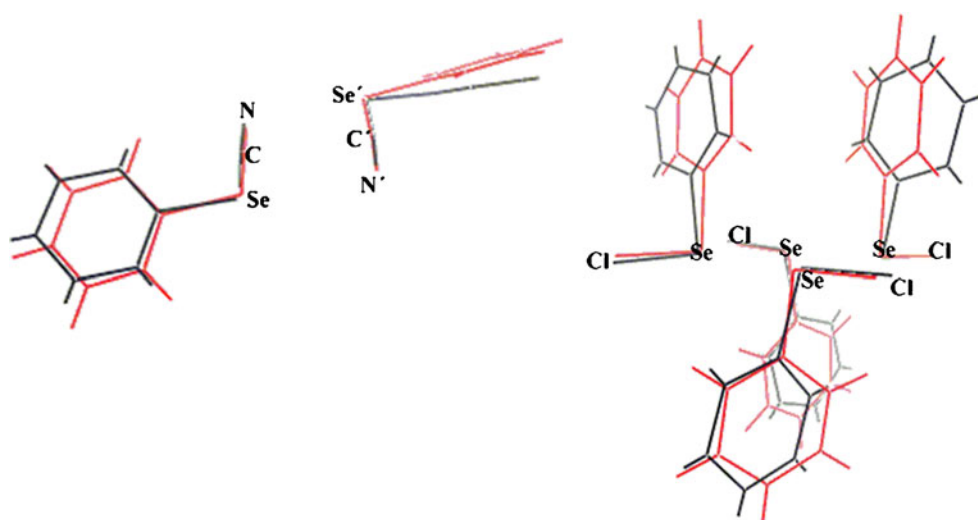


Table 1 Topological properties of charge density^{a,b} calculated at some bond critical points (BCPs) of phenylselenenyl halides (PhSeX; X=Cl, Br, I, CN, SCN)

Compound	Bond ^c	ρ_b	$\nabla^2\rho_b$	ϵ	E_b
PhSeCl	C1-C2	0.3061	-0.8422	0.1943	-0.3094
	C1-Se	0.1527	-0.1061	0.0696	-0.0871
	Se-Cl	0.0981	0.0490	0.1408	-0.0350
PhSeBr	C1-C2	0.3061 (0.3063)	-0.8415 (-0.8427)	0.1952 (0.1950)	-0.3092 (-0.3096)
	C1-Se	0.1511 (0.1513)	-0.1020 (-0.1023)	0.0620 (0.0618)	-0.0854 (-0.0857)
	Se-Br	0.0887 (0.0876)	0.0119 (0.0120)	0.1326 (0.1328)	-0.0315 (-0.0314)
PhSeI	C1-C2	0.3062	-0.8419	0.1970	-0.3096
	C1-Se	0.1495	-0.0976	0.0501	-0.0837
	Se-I	0.0744	-0.0040	0.1143	-0.0240
PhSeCN	C1-C2	0.3082	-0.8511	0.2097	-0.3145
	C1-Se	0.1448	-0.1071	0.0572	-0.0782
	Se-C7	0.1595	-0.0265	0.1919	-0.1035
	C7-N	0.4724	-0.2263	0.0116	-0.8536
PhSeSCN	C1-C2	0.3063	-0.8419	0.1988	-0.3101
	C1-Se	0.1499	-0.1032	0.0456	-0.0840
	Se-S	0.0999	0.0090	0.0649	-0.0369
	S-C7	0.2090	-0.3845	0.3608	-0.2396
	C7-N	0.4715	-0.2848	0.0085	-0.8492

^a ρ_b , $\nabla^2\rho_b$ y E_b in au^b Calculated at the B3LYP/6-311G(d,p) level. Values in parentheses were obtained using the 6-311G(d) basis sets for the halogen atom^c For atom labels, see Fig. 1

Large positive values of ρ_b and large negative values of the Laplacian are indicators of strong covalent bonds (see, for example, values for the C–C and C–N single bonds in Table 1). In case of a pure ionic bond (close shell interaction) one would expect a small value of ρ_b , indicating depletion of the electron density and positive values of the Laplacian. In our case we have considerable ρ_b values and positive Laplacian values for the Se–Cl (0.0981 a.u. and 0.0490 a.u.), Se–Br (0.0887 a.u. and 0.0119 a.u.) and Se–S (0.099 a.u. and 0.0090 a.u.), but (is) $E_b < 0$ in all cases. Clearly, here we have a superposition of two extreme cases that allows us to conclude that these bonds should be classified as strong, highly polarized covalent bonds.

When the C–X bonds are compared, a decrease in the electron density (0.0981 a.u., 0.0887 a.u. and 0.0744 a.u.) and a corresponding decrease of the electronic energy density (0.0350 a.u., 0.0310 a.u. and 0.0240 a.u.) can be seen. These findings can be interpreted as a decrease in covalent character of the bonds when going from Se–Cl to Se–I. Moreover, the features of the C–N bond are not affected by inclusion of the S atom in the group attached to selenium atom, those being strong covalent bonds (see Table 1). The electron density of the Se–C bond to the cyanide group (0.1591 a.u.) is slightly higher than the

electron density of the Se–C bonds to the phenyl rings (0.1445 a.u.). Accordingly, the ellipticity is 0.1919 against 0.0572, respectively, reflecting the double bond character in the Se–C bond to the sp hybridized cyanide carbon, compared with the sp² hybridized ring carbon.

Table 2 Topological properties of charge density^{a,b} calculated at selected BCPs of the PhSeCN dimer and Ph₄Se₄CN₄

Compound	Bond ^c	ρ_b	$\nabla^2\rho_b$	ϵ	E_b
(PhSeCN) ₂	C1-C2	0.3070	-0.8410	0.2116	-0.3130
	C1-Se	0.1454	-0.1077	0.1715	-0.0790
	Se-C7	0.1618	-0.0263	0.1964	-0.1066
	C7-N	0.4725	-0.2550	0.0105	-0.8532
	Se'-N	0.0051	0.0169	0.1654	0.0008
	C7'-C7	0.0051	0.0157	1.6579	0.0008
	Se-N'	0.0051	0.0169	0.1663	0.0008
Ph ₄ Se ₄ Cl ₄	C1-C2	0.3065	-0.8416	0.2030	-0.3107
	Cl-Se	0.1500	-0.1145	0.0402	-0.0835
	Se-Cl	0.0888	0.0574	0.1102	-0.0285
	Se-Se	0.0244	0.0401	0.1279	-0.0010

^a ρ_b , $\nabla^2\rho_b$ y E_b in au^b Calculated at the B3LYP/6-311G(d,p) level^c For atom labels, see Fig. 2

The atom bonding network that connects both PhSeCN units is an important feature of the dimeric structure of this compound, which seems to be aiding the crystallization. The dimeric structure shows three intermolecular interactions. The Se \cdots N distances are 3.48 Å and the C \cdots C distance is 3.42 Å, which are almost equal to the sum of the van der Waals radii of the selenium and nitrogen atoms (3.45 Å) and of the carbon atoms (3.40 Å). However, the experimental distances are shorter than the sum of the van der Waals radii, suggesting that there is appreciable interaction at the long Se \cdots N contact distance (3.33 Å and 3.44 Å, respectively).

Within the tetramer, the solid-state structures of Ph₄Se₄Cl₄, four PhSeCl units are held together by weak Se \cdots Se bonds forming an essentially planar ring of Se₄ and the Se–Cl bonds lie in the Se₄ plane, with two phenyl rings lying above the plane, and two below it.

The calculated Se \cdots Se distance (3.13 Å) is significantly smaller than the sum of van der Waals radii of the selenium atoms (3.80 Å) but longer than the covalent radii (3.80 Å), revealing the existence of the non-covalent interactions.

No significant changes in the topological properties of the bonds are observed when the dimeric structure of (PhSeCN)₂ is formed (Table 2). Indeed, when the corresponding bonds are compared, similar characteristics appear. For example, in the Se–C7 bond, the electronic density is 0.1595 a.u. and 0.1618 a.u. in the monomer and dimer, respectively. In the same direction, the Laplacian of the density is –0.0265 a.u. and –0.0263 a.u., respectively, in this bond. It is interesting to note that the three bonds that hold the two molecules together have similar topological properties. The low value of the electron density (0.0051 a.u.) and positive and low values of the Laplacian of the density (between 0.0157 a.u. and 0.0169 a.u.) at the CPs of the Se \cdots N and C \cdots C bonds indicates a weak interaction between the two molecules. The high value of the ellipticity at the BCP in the C–C bond linking the two molecules can be explained by the strong π character of this bond.

The topological properties of the bonds in the monomer were similar in the corresponding tetrameric structure. A topological analysis of BCPs in ρ of Ph₄Se₄CN₄ revealed that all bonds forming the phenyl ring correspond to covalent interactions ($\rho_b=0.3070$ a.u., $\nabla^2\rho_b=-0.8410$ a.u. and $E_b=-0.3130$ a.u.), and the ellipticities values ($\epsilon=0.2116$) reveal the partial double bond character, as in the PhSeCN monomeric structure.

As seen previously, weak interactions are observed taking into account the calculated and experimental Se \cdots Se distances. Accordingly, topological properties at the BCPs in ρ of Se–Se bonds correspond to weak interactions: low values of electron densities ($\rho_b=0.0244$ a.u.), positive and low values of the Laplacian of the density ($\nabla^2\rho_b=0.0401$ a.u.) and E_b values near to zero (–0.010 a.u.).

Conclusions

This paper reports a theoretical study of a “PhSeX” series of compounds, where Ph=phenyl, Se=selenium and X=Cl, Br, I, CN or SCN. The molecular geometry was calculated at DFT/B3LYP level of calculation by means of the 6-311 G(d,p) basis sets. The equilibrium structures of the molecules were found to depend on the method employed for comparison with previously reported experimental data. A topological study of the calculated PhSeX species, based on the AIM theory, illustrates the subtle differences in the solid-state structures of PhSeX compounds.

A decrease in the electron density and a corresponding decrease in the electronic energy density was observed when going from PhSeCl to PhSeI, which can be interpreted as a decrease in the covalent character of the C–X bonds. In PhSeCN, the ellipticity of the Se–C attached to the cyanide group is slightly higher than the ellipticity of the Se–C bonds to the phenyl rings, reflecting the double bond character of the former bond in the cyanide group. In PhSeSCN, the features of the C–N bond are not affected by the inclusion of the S atom in the group attached to the selenium atom.

The atom bonding network connecting two units seems to aid the crystallization process in the dimeric structure of (PhSeCN)₂, showing three intermolecular interactions.

In the structure of Ph₄Se₄Cl₄, four PhSeCl units are held together by weak Se \cdots Se bonds forming an essentially planar ring of Se₄ with non-covalent Se \cdots Se interactions.

Acknowledgments A.H.J. is a member of the Carrera del Investigador Científico, CIC, Buenos Aires and E.A.C. and N.B.O. are members of the career researcher of Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, Argentina.

References

- IUPAC (1995) Compendium of Chemical Terminology 67:1361
- Birkenbach L, Kellermann K (1925) Über Pseudohalogene (I). Ber Dtsch Chem Ges 58:786–794
- Holleman F, Wiberg N, Wiberg E (1995) Lehrbuch der Anorganischen Chemie, 101st edn. de Gruyter, Berlin
- Tiecco M (2000) Electrophilic Selenium, Selenocyclizations. Top Curr Chem 208:7–54
- Wirth T (2000) Organoselenium chemistry: modern developments in organic synthesis. Springer, Berlin
- Kubiniok S, du Mont WW, Pohl S, Saak W (1988) The reagent diphenyldiselenane/Iodine: no phenylselenenyl iodide but a charge transfer complex with cyclic moieties. Angew Chem Int Edn 27:431–433
- Barnes NA, Godfrey SM, Halton RTA, Mushtaq I, Parsons S, Pritchard RG, Sadler M (2007) A comparison of the solid-state structures of a series of phenylseleno-halogen and pseudohalogen compounds, PhSeX (X=Cl, CN, SCN). Polyhedron 26:1053–1060
- Barnes NA, Godfrey SM, Halton RTA, Mushtaq I, Pritchard RG, Sarwar S (2006) Reactions of Ph₄Se₄Br₄ with tertiary phosphines.

- Structural isomerism within a series of $R_3PSe(Ph)Br$ compounds. *Dalton Trans* 12:1517–1523
- Klapötke TM, Krumm B, Mayer P (2004) Chemistry of C_6F_5SeLi and C_6F_5SeCl : precursors to new pentafluorophenylselenium(II) compounds. *Z Naturforsch* 59b:547–553
 - Allen FH, Kennard O, Watson DG, Brammer L, Orpen AG, Taylor R (1987) Tables of bond lengths determined by X-ray and neutron diffraction. Part 1. Bond lengths in organic compounds. *J Chem Soc Perkin Trans 2*:S1–S19
 - Akers C, Peterson SW, Willett RD (1968) A refinement of the crystal structure of KSCN. *Acta Crystallogr B* 24:1125–1126
 - Okulik N, Jubert AH, Castro EA (2002) Theoretical study of new pseudohalogen CS_2N_3 and some related compounds. *J Mol Struct THEOCHEM* 589–590:79–87
 - Okulik NB, Jubert A, Castro E (2006) Bonding in some covalent derivatives of the 1,2,3,4-thiazole-5-thiolate anion. A topological study. *J Mol Struct THEOCHEM* 770:13–22
 - Hohenberg P, Kohn W (1964) Inhomogeneous electron gas. *Phys Rev* 136:864–B871
 - Kohn W, Sham LJ (1965) Self-consistent equations including exchange and correlation effects. *Phys Rev* 140:A1133–A1138
 - Parr RG, Yang W (1989) Density functional theory of atoms and molecules. Oxford University Press, Oxford
 - Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. *J Chem Phys* 98:5648–5652
 - 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
 - Cambridge Crystallographic Data Center, 12, Union Road, Cambridge CB2 1EZ, UK
 - Clark T, Chandrasekhar J, Spitznagel GW, PvR S (1983) Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li–F. *J Comput Chem* 4:294–301
 - Curtiss LA, McGrath MP, Blandeau JP, Davis NE, Binning RC Jr, Radom L (1995) Extension of Gaussian-2 theory to molecules containing third-row atoms Ga–Kr. *J Chem Phys* 103:6104–6113
 - Frisch MJ et al (2003) Gaussian 03, Revision C.02. Gaussian Inc, Wallingford, CT
 - Biegler-König FW, Bader RFW, Tang TH (1982) Calculation of the average properties of atoms in molecules II. *J Comput Chem* 3:317–328
 - Gans J, Shalloway D (2001) Qmol: A program for molecular visualization on Windows based PCs. *J Mol Graph Model* 19:557–559
 - Bader RFW (1990) Atoms in molecules. A quantum theory. Oxford University Press, Oxford

The role of CS₂ in CS₂/NMP mixed solvent in weakening the hydrogen bond of OH⋯N in coal: a DFT investigation

Baojun Wang · Liping Wang · Riguang Zhang · Lixia Ling

Received: 28 February 2011 / Accepted: 12 May 2011 / Published online: 31 May 2011
© Springer-Verlag 2011

Abstract The interaction processes of trace amounts of N-methyl-2-pyrrolidinone (NMP), CS₂/NMP (1:1 by volume) and pure NMP solvent with the hydrogen bond of OH⋯N in coal were constructed and simulated by density functional theory methods. The distances and bond orders between the main related atoms, and the hydrogen bond energy of OH⋯N were calculated. The calculated results show that pure NMP solvent does not weaken the hydrogen bond of OH⋯N in coal. However, trace amounts of NMP and CS₂/NMP (1:1 by volume) have a strong capacity to weaken the hydrogen bond of OH⋯N in coal. The H2–N3 distances are elongated from 1.87 Å to 3.80 Å and 3.44 Å, the bond orders of H2–N3 all disappear, and the corresponding hydrogen bond energies of OH⋯N in coal decrease from 45.72 kJ mol⁻¹ to 7.06 and 11.24 kJ mol⁻¹, respectively. These results show that CS₂ added to pure NMP solvent plays an important role in releasing the original capacity of NMP to weaken the hydrogen bond of OH⋯N in coal, in agreement with experimental observations.

Keywords N-methyl-2-pyrrolidinone · CS₂ · Coal · Hydrogen bond · Density functional theory

Introduction

Hydrogen bonds [1, 2] play a key role in the chemical properties and structures existing in coal [3, 4], with one of the main types of hydrogen bonds found in coal being OH⋯N [5, 6]. Solvent extraction is an effective method with which to investigate the composition and structure of coal. By weakening the hydrogen bonds in coal, solvents increase its solubility, allowing extraction to be achieved [7, 8]. It is well known that CS₂/N-methyl-2-pyrrolidinone (NMP) (1:1 v:v) is an excellent mixed solvent for extraction of many coals at room temperature [9], as NMP (see Fig. 1) has a strong association with CS₂ [10], and the complex of NMP and CS₂ may be the main reason why CS₂/NMP (1:1 by volume) has a high extraction yield for many coals.

Zong et al. [11] found that N-methylpyrrolidine-2-thione (NMPT) and CSO were produced by the reaction of NMP with CS₂ in their experiments. Subsequently, Wang et al. [12] and Fu et al. [13] studied the reaction mechanism of NMP with CS₂, obtaining possible reaction pathways, transition state (TS) and intermediates (IM) using quantum chemistry calculation methods. Comparing the structure parameters, activation energies and other related data, the work by Fu et al. proved to be more reasonable. Further, we consider that the IM obtained by Fu et al. may be a complex of CS₂ and NMP, which we think represents CS₂/NMP (1:1, v:v). Liu et al. [14] extracted three types of coal by CS₂/NMP(1:1, v:v), and showed that when all CS₂ and most of the NMP were removed, trace amounts of NMP still remained in strong interaction with coal, which means that NMP itself has extracting capacity. However, as NMP

B. Wang (✉) · L. Wang · R. Zhang · L. Ling
Key Laboratory of Coal Science
and Technology of Ministry of Education and Shanxi Province,
Taiyuan University of Technology,
Taiyuan 030024 Shanxi, China
e-mail: quantumtyut@126.com
e-mail: wangbaojun@tyut.edu.cn

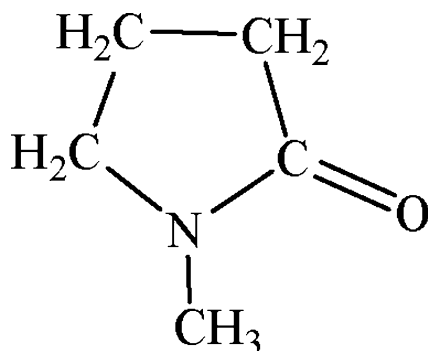


Fig. 1 The structure of N-methyl-2-pyrrolidinone (NMP)

molecules self-aggregate [15, 16], pure NMP is not an effective solvent for extraction of coal. In other words, aggregation between NMP molecules can inhibit the extraction capacity of pure NMP for coal. However, for a detailed understanding of the extraction process of different solvents, experimental information is not always sufficient and accompanying theoretical calculations can be helpful to clarify some essential questions.

Recently, density functional theory (DFT) has provided qualitative and quantitative insights into hydrogen bonds [17–19]. For example, Ireta et al. [17] used DFT method with Perdew-Burke-Ernzerh (PBE) functional to investigate a set of representative hydrogen bonded dimers in diverse geometric environments; the calculated results show that

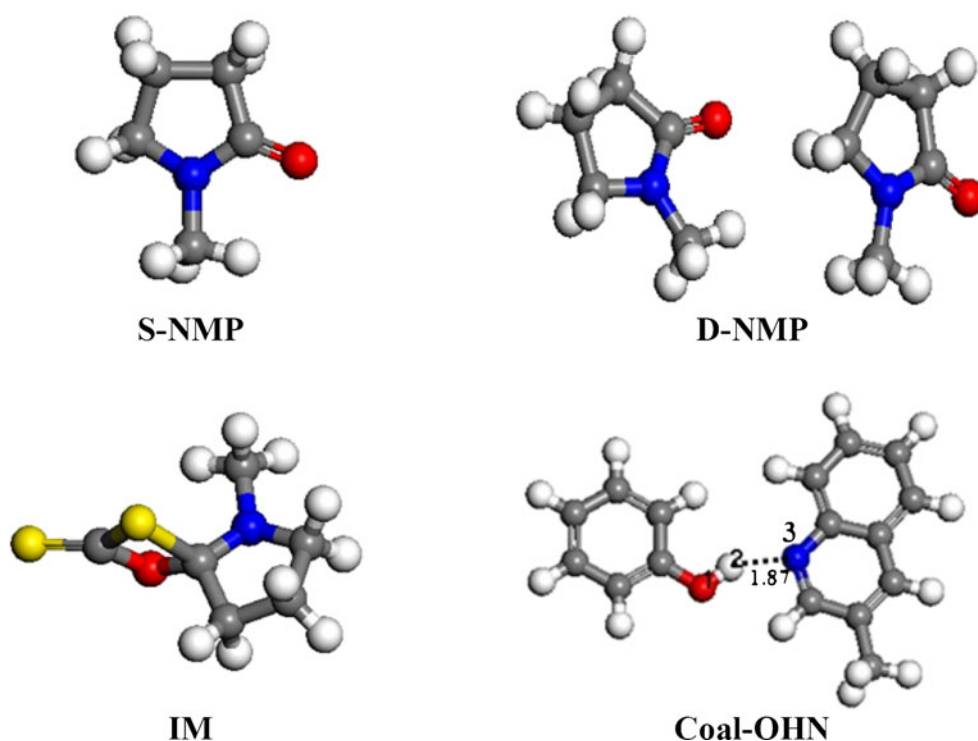
DFT-PBE is reliable for the description of hydrogen bond strengths and geometry parameters. Meanwhile, Korth et al. [18] studied the intra-molecular hydrogen bonding in 2-substituted phenols by DFT method; the results, such as conformations and enthalpies, are in agreement with experimental findings.

In this study, three solvent models of trace amounts of NMP solvent, CS₂/NMP (1:1, v:v) mixed solvent and pure NMP solvent, as well as the hydrogen bond of OH···N in coal were constructed to simulate the extraction processes of three different solvents for the hydrogen bond of OH···N in coal using DFT method. Based on changes in the hydrogen bond of OH···N, we examined the role of CS₂ in the CS₂/NMP (1:1, v:v) mixed solvent for the extraction process.

Model construction and computational method

In this study, single NMP (S-NMP) and double NMP (D-NMP) molecules were chosen to simulate trace amounts of NMP solvent and pure NMP solvent, respectively. IM obtained in previous studies by Fu et al. [13] were used to simulate CS₂/NMP (1:1, v:v) mixed solvent. The simplified model of the OH···N hydrogen bond in coal (Coal-OHN) reported in studies by Miura et al. [20], was applied to simulate the real bond in coal. Coal-OHN is sufficient to represent the functional group of OH···N in real coal,

Fig. 2 The constructed models of single NMP (S-NMP), double NMP (D-NMP), intermediates (IM) and the simplified model of the OH···N hydrogen bond in coal (Coal-OHN)



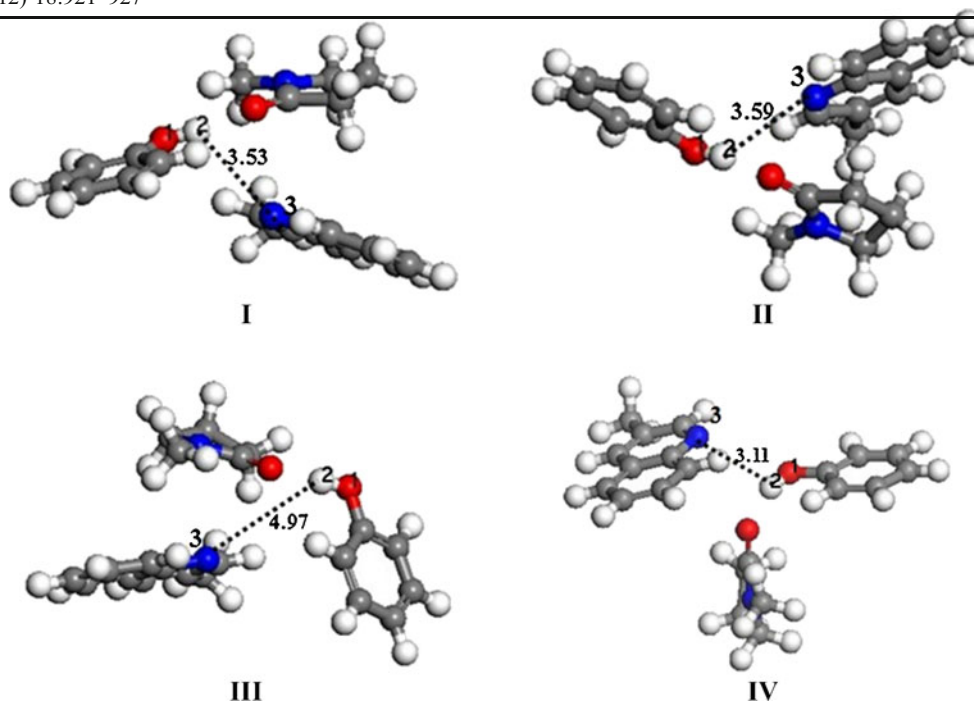


Fig. 3 Stable structures of S-NMP and Coal-OHN at different spatial locations

although it does not recreate the integrated structure of real coal. Models of S-NMP, D-NMP, IM and Coal-OHN are shown in Fig. 2.

A DFT method was adopted and calculations were performed using the Dmol³ program mounted on the Materials studio 4.4 package (<http://accelrys.com>). All models were optimized at the level of generalized gradient approximation (GGA) [21] using the PBE functional [22] together with the DND basis set [23]. Unrestricted spin was chosen. Total self-consistent field (SCF) tolerance criteria, integration accuracy criteria and orbital cutoff quality criteria were set at medium. The converge criterion judged by the energy, force and displacement are 2×10^{-5} Ha, 4×10^{-3} Ha/Å and 5×10^{-3} Å, respectively. Multipolar expansion is set at octupole.

Considering that solvent molecules may interact with the hydrogen bonds in coal from different directions as

a result of the complexity of the actual extraction processes, it is impossible to calculate all possible situations. Thus, we adopted the method of “Multi-point calculation, Overall average” [24], which means that the solvent models (S-NMP, D-NMP and IM) were placed in different spatial locations relative to the hydrogen bond in Coal-OHN, and the new composite models were then optimized to obtain their corresponding stable structures; thereafter, we averaged these related data. Using this method, the average data obtained were more convincing and reasonable.

Results and discussion

After many attempts, four typical and effective spatial locations were found for the extraction processes of three different solvents in Coal-OHN. With the purpose

Table 1 O1–H2 and H2–N3 distances in the simplified model of the OH \cdots N hydrogen bond in coal (Coal-OHN). S-NMP Single N-methyl-2-pyrrolidinone

	Coal-OHN with S-NMP					Coal-OHN without S-NMP
	I	II	III	IV	Avg ^a	
O1–H2 (Å)	1.00	1.00	1.00	1.00	1.00	1.00
H2–N3 (Å)	3.53	3.59	4.97	3.11	3.80	1.87

^a Average of I, II, III and IV

Table 2 Mayer bond orders of O1–H2 and H2–N3 in Coal-OHN

	Coal-OHN with S-NMP					Coal-OHN without S-NMP
	I	II	III	IV	Avg ^a	
O1–H2	0.62	0.60	0.60	0.62	0.61	0.64
H2–N3	– ^b	–	–	–	–	0.17

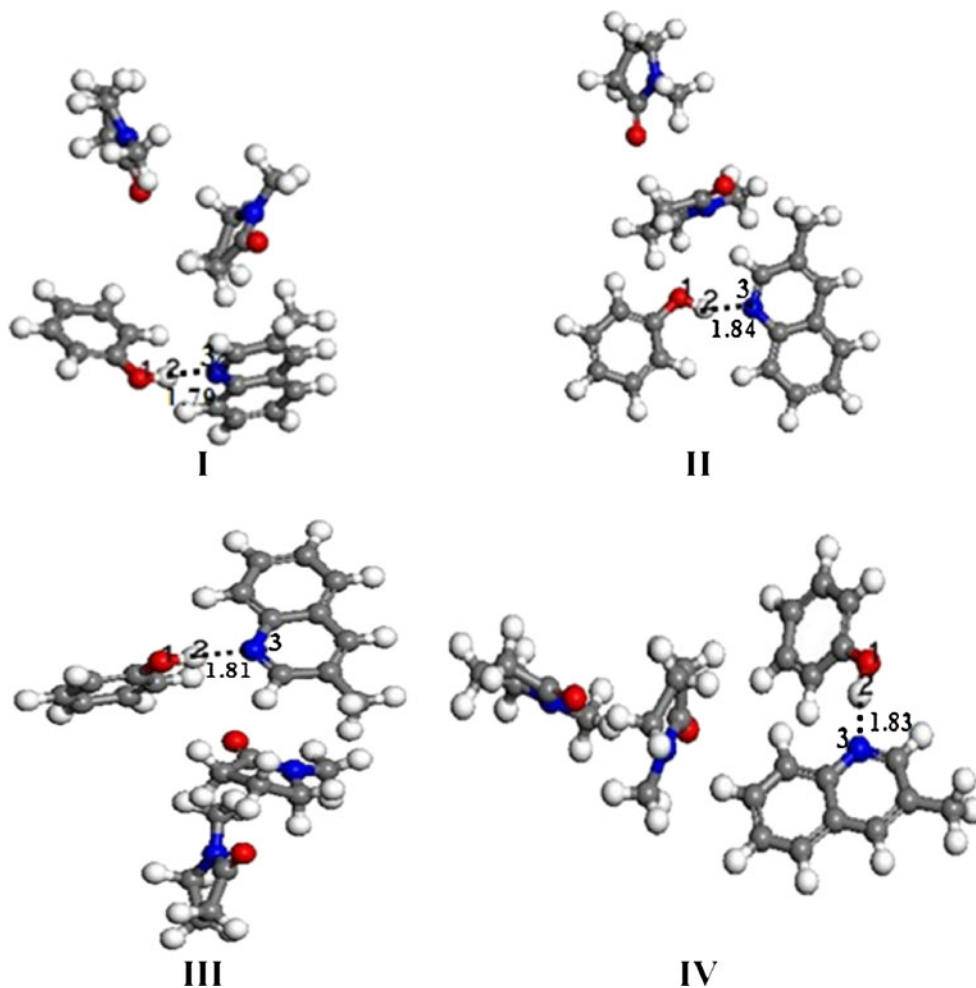
^a Average of I, II, III and IV

^b Bond order does not exist

Table 3 Energies of the relative parts of Coal-OHN

	Coal-OHN with S-NMP					Coal-OHN without S-NMP
	I	II	III	IV	Avg ^a	
E_{coal} (Ha)	-747.9988	-747.9993	-747.9985	-747.9999	-747.9991	-747.9828
E_{OH} (Ha)	-307.1858	-307.1854	-307.1854	-307.1858	-307.1856	-307.1730
E_{N} (Ha)	-440.8107	-440.8107	-440.8110	-440.8109	-440.8108	-440.7923
E_{HBE} (kJ mol ⁻¹)	6.01	8.36	5.49	8.36	7.06	45.72

^a Average of I, II, III and IV

Fig. 4 Stable structures of D-NMP with Coal-OHN at different spatial locations

of examining the capacities of different solvent models (S-NMP, D-NMP and IM) to weaken the hydrogen bond of $\text{OH}\cdots\text{N}$ in Coal-OHN, the atoms of O1, H2 and N3, which compose the $\text{OH}\cdots\text{N}$ hydrogen bond in Coal-OHN, were identified as the main research focus.

Composite structures of S-NMP and Coal-OHN

The four composite structures of S-NMP and Coal-OHN at different spatial locations were optimized, and the

Table 4 O1–H2 and H2–N3 distances in Coal-OHN. *D-NMP* Double NMP

	Coal-OHN with D-NMP					Coal-OHN without D-NMP
	I	II	III	IV	Avg ^a	
O1–H2 (Å)	1.00	1.00	1.00	1.00	1.00	1.00
H2–N3 (Å)	1.79	1.84	1.81	1.83	1.82	1.87

^a Average of I, II, III and IV

Table 5 Mayer bond orders of O1–H2 and H2–N3 in Coal-OHN

	Coal-OHN with D-NMP					Coal-OHN without D-NMP
	I	II	III	IV	Avg ^a	
O1–H2 (Å)	0.59	0.61	0.60	0.61	0.60	0.64
H2–N3 (Å)	0.18	0.17	0.17	0.16	0.17	0.17

^a Average of I, II, III and IV

corresponding stable structures (with no imaginary frequency) are shown in Fig. 3.

In order to investigate the capacity of S-NMP to weaken the hydrogen bond of OH···N in Coal-OHN, we analyzed the distances and Mayer bond orders of O1–H2 and H2–N3, as listed in Tables 1 and 2.

As shown in Tables 1 and 2, the distance and bond order between O1 and H2 were almost the same; however, we observed a large change between H2 and N3. The H2–N3 distance is stretched from 1.87 Å initially to 3.80 Å in Coal-OHN, which suggests that the hydrogen bond of OH···N in Coal-OHN with S-NMP is very weak. Moreover, the bond order also confirms that no bond order exists between H2 and N3 after S-NMP is added. Furthermore, we investigated the hydrogen bond energy (E_{HBE}) of OH···N in Coal-OHN. E_{HBE} is defined as:

$$E_{\text{HBE}} = E_{\text{OH}} + E_{\text{N}} - E_{\text{coal}} \quad (1)$$

Where E_{HBE} is the hydrogen bond energy, E_{OH} is the energy of the part containing OH in Coal-OHN, E_{N} is the energy of the part containing N in Coal-OHN, E_{coal} is the total energy of Coal-OHN. The energies of the related parts are shown in Table 3; we can see that E_{HBE} obviously decreases from 45.72 kJ mol⁻¹ to 7.06 kJ mol⁻¹ after adding S-NMP into Coal-OHN.

From Tables 1–3, we can see that the distances between H2 and N3 are stretched; the corresponding bond orders do not exist, and E_{HBE} decreases dramatically due to the addition of S-NMP into Coal-OHN, which

means that S-NMP can seriously damage the OH···N hydrogen bond in Coal-OHN.

Composite structures of D-NMP and Coal-OHN

The composite structures of D-NMP (which replaces the location of the S-NMP) and Coal-OHN were optimized, and the corresponding stable structures (with no imaginary frequency) are shown in Fig. 4.

Similarly, we analyzed the distances, Mayer bond orders of O1–H2 and H2–N3, and the energies of the related parts of Coal-OHN, as shown in Tables 4, 5, and 6.

The results reveal that there are few changes to the distances and bond orders of O1–H2 and H2–N3 and the E_{HBE} . In other words, D-NMP has no ability to weaken the hydrogen bond of OH···N in Coal-OHN.

Composite structures of IM and Coal-OHN

The composite structures of the IM (which replaces the location of S-NMP) and Coal-OHN were optimized and the corresponding stable structures (with no imaginary frequency) are shown in Fig. 5. The distances, Mayer bond orders of O1–H2 and H2–N3, and the energies of the related parts of Coal-OHN were calculated, as shown in Tables 7, 8, 9. The calculated results show that the addition of IM leads to the distances between H2 and N3 increasing significantly from 1.87 Å to 3.44 Å, see Table 7. The corresponding bond orders do not exist (see Table 8). Moreover, E_{HBE} decreases to 11.24 kJ mol⁻¹ from 45.72 kJ mol⁻¹, as shown in Table 9. The above results also suggest that the OH···N hydrogen bond in Coal-OHN is seriously damaged. Therefore, IM has a strong capacity to weaken the hydrogen bond of OH···N in coal-OHN.

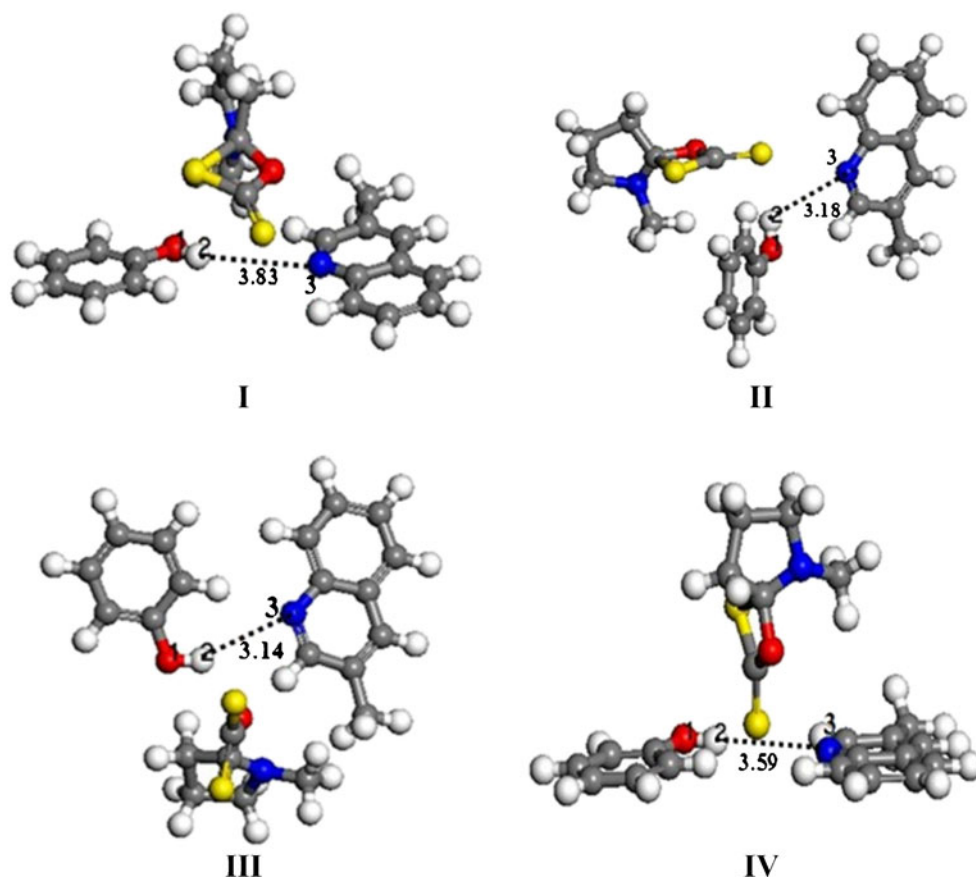
Finally, all the data were summarized and compared as presented in Table 10. It can be seen that the values obtained for Coal-OHN are very similar to those obtained with Coal-OHN with D-NMP, which suggests that the hydrogen bond of OH···N in Coal-OHN cannot be weakened even if D-NMP is added. Interestingly, in Coal-OHN with S-NMP and Coal-OHN with IM, the data are

Table 6 Energies of relative parts of Coal-OHN

	Coal-OHN with D-NMP					Coal-OHN without D-NMP
	I	II	III	IV	Avg ^a	
E_{coal} (Ha)	-748.0143	-748.0125	-748.0148	-748.0137	-748.0138	-747.9828
E_{OH} (Ha)	-307.1851	-307.1849	-307.1854	-307.1852	-307.1852	-307.1730
E_{N} (Ha)	-440.8104	-440.8105	-440.8106	-440.8107	-440.8106	-440.7923
E_{HBE} (kJ mol ⁻¹)	49.11	44.67	49.11	46.50	47.35	45.72

^a Average of I, II, III and IV

Fig. 5 Stable structures of IM and Coal-OHN at different spatial locations



also comparable. The distances between H2 and N3 are elongated from 1.87 Å to 3.80 Å and 3.44 Å, respectively. The bond orders all disappear; the corresponding E_{HBE} decreases from 45.72 kJ mol⁻¹ to 7.06 and 11.24 kJ mol⁻¹, respectively, which means that addition of both S-NMP and IM can seriously damage the hydrogen bond of OH···N in Coal-OHN.

Recently, Politzer et al. [25–27] found that a σ -hole (a region of positive charge) resulting from a deficiency of electron density exists on the outermost portions of some covalently bonded halogen atoms. Meanwhile, Wang et al. [28] thought that the lone electron pairs of halogen atom produce a region of negative electrostatic potential around the central part of halogen atom, which leaves the

possibility for the atom to act as an electron donor. Later, Wang et al. [29] studied the corresponding properties of chalcogen atoms (O, S, Se, Te, Po), suggesting that chalcogen atoms share similar characteristics with halogen atoms. On the basis of previous experimental facts and theoretical calculations, we think that the O in S-NMP and the S of C=S in IM also share these corresponding characters. Considering that it is sensitive to direction, occupies a small specific gravity, and is surrounded by negative electrostatic potential, the σ -hole cannot interact easily with negative sites (such as ring N) in our research system, although the σ -hole has a positive electrical property. The main reason why S-NMP and IM could weaken the OH···N in Coal-OHN may be that the lone

Table 7 O1–H2 and H2–N3 distances in Coal-OHN. IM Intermediates

	Coal-OHN with IM					Coal-OHN without IM
	I	II	III	IV	Avg ^a	
O1–H2 (Å)	1.00	1.00	1.00	1.00	1.00	1.00
H2–N3 (Å)	3.83	3.18	3.14	3.59	3.44	1.87

^a Average of I, II, III and IV

Table 8 Mayer bond orders of O1–H2 and H2–N3 in Coal-OHN

	Coal-OHN with IM					Coal-OHN without IM
	I	II	III	IV	Avg ^a	
O1–H2	0.63	0.61	0.62	0.61	0.62	0.64
H2–N3	– ^b	–	–	–	–	0.17

^a Average of I, II, III and IV

^b Bond order does not exist

Table 9 Energies of the relative parts of Coal-OHN

	Coal-OHN with IM					Coal-OHN without IM
	I	II	III	IV	Avg ^a	
E_{coal} (Ha)	-747.9672	-747.9713	-747.9565	-747.9700	-747.9663	-747.9828
E_{OH} (Ha)	-307.1730	-307.1735	-307.1720	-307.1734	-307.1730	-307.1730
E_{N} (Ha)	-440.7921	-440.7922	-440.7791	-440.7925	-440.7890	-440.7923
E_{HBE} (kJ mol ⁻¹)	5.49	14.63	14.11	10.71	11.24	45.72

^a Average of I, II, III and IV

electron pairs of O in S-NMP and S of C=S in IM lead to negative charge of the two atoms as electron donors, which could interact with positive H of OH \cdots N and compete with the negative N. As a result, the model hydrogen bond of OH \cdots N in Coal-OHN is weakened.

Considering that IM in our study comes from the addition of CS₂ into NMP, we hypothesize that CS₂ can destroy the aggregation in D-NMP and release the capacity of D-NMP to weaken the hydrogen bond of OH \cdots N in Coal-OHN, almost reaching the level of S-NMP.

Conclusions

Three different extraction processes were simulated successfully by performing DFT methods. The calculated results show that although D-NMP barely weakens the hydrogen bond of OH \cdots N in coal, S-NMP and IM do so, seriously and similarly. Thus, we conclude that NMP has an intrinsically high capacity to weaken the hydrogen bond of OH \cdots N in coal, but that this capacity is masked due to aggregation between NMP molecules. The addition of CS₂ to NMP destroys the aggregation of NMP and released the capacity of weakening the hydrogen bond of OH \cdots N. That is, CS₂ added into pure NMP solvent plays the role of releasing the original capacity of NMP to weaken the OH \cdots N hydrogen bond in coal.

Table 10 Related data of the different mixing processes

Models	Distance	Bond order	Hydrogen bond energy
	H2-N3 (Å)	H2-N3	E_{HBE} (kJ·mol ⁻¹)
Coal-OHN	1.87	0.17	45.72
Coal-OHN with S-NMP	3.80	– ^b	7.06
Coal-OHN with D-NMP	1.82	0.17	47.35
Coal-OHN with IM	3.44	–	11.24

^b Bond order does not exist

Acknowledgments The authors thank the anonymous reviewers for their helpful suggestions to improve the quality of our present paper. This work was supported financially by the National Natural Science Foundation of China (No. 20976115 and 20776093) and the Younger Foundation of Shanxi Province (No. 2009021015).

References

- Pimentel GC, McClellan AL (1960) The hydrogen bond. Freeman, San Francisco
- Margaret CE (1990) Acc Chem Res 23:120–126
- Painter PC, Sobkowiak M, Youtcheff J (1987) Fuel 66:973–978
- Nishioka M (1992) Fuel 71:941–948
- Li DT, Li W, Li BQ (2001) Chem Online 64:411–415
- Chen C, Gao JS, Yan YJ (1998) Energ Fuels 12:446–449
- Painter PC, Sobkowiak M, Valerie G (1998) Prepr Pap-Am Chem Soc Div Fuel Chem 43:913–915
- Lino M (2000) Fuel Process Technol 62:89–101
- Lino M, Takanoashi T, Ohsuga H et al (1988) Fuel 67:1639–1647
- Chen C, Kurose H, Lino M (1999) Energ Fuels 13:1180–1183
- Zong ZM, Peng YL, Qin ZH et al (2000) Energ Fuels 14:734–735
- Wang BJ, Wei XY, Xie KC (2004) Chin J Chem Eng 55:569–574
- Fu XB, Zhang C, Zhang DJ et al (2006) Chem Phys Lett 420:162–165
- Liu CM, Zong ZM, Jia JX et al (2008) Chin Sci Bull 53:183–190
- Shui HF, Wang ZC, Gao JS (2006) Fuel Process Technol 87:185–190
- Aparicio S, Davila MJ, Alcalde R (2009) Energ Fuels 23:1591–1602
- Ireta J, Neugebauer J, Scheffler M (2004) J Phys Chem A 108:5692–5698
- Korth HG, Heer MI, Mulder P (2002) J Phys Chem A 106:8779–8789
- Pan YP, McAllister MA (1997) J Am Chem Soc 119:7561–7566
- Miura K, Mae K, Li W et al (2001) Energ Fuels 15:599–610
- Perdew JP, Wang Y (1986) Phys Rev B 33:8800–8802
- Perdew JP, Burke K, Ernzerhof M (1996) Phys Rev Lett 77:3865–3868
- Delley B (2000) J Chem Phys 113:7756–7764
- Wang BJ (2006) Study on quantum chemistry of coal structure and reactivity. Dissertation, Taiyuan University of Technology
- Politzer P, Lane P, Concha MC et al (2007) J Mol Model 13:305–311
- Murray JS, Riley KE, Politzer P et al (2010) Aust J Chem 63:1598–1607
- Clark T, Hennemann M, Murray JS et al (2007) J Mol Model 13:291–296
- Wang WZ, Tian AM, Wong NB (2005) J Phys Chem A 109:8035–8040
- Wang WZ, Ji BM, Zhang Y (2009) J Phys Chem A 113:8132–8135

Ab initio DFT study of bisphosphonate derivatives as a drug for inhibition of cancer: NMR and NQR parameters

Hussein Aghabozorg · Beheshteh Sohrabi ·
Sara Mashkouri · Hamid Reza Aghabozorg

Received: 6 November 2010 / Accepted: 28 April 2011 / Published online: 2 June 2011
© Springer-Verlag 2011

Abstract DFT computations were carried out to characterize the ^{17}O and ^2H electric field gradient, EFG, in various bisphosphonate derivatives. The computations were performed at the B3LYP level with 6-311++G (d,p) standard basis set. Calculated EFG tensors were used to determine the ^{17}O and ^2H nuclear quadrupole coupling constant, χ and asymmetry parameter, η . For better understanding of the bonding and electronic structure of bisphosphonates, isotropic and anisotropic NMR chemical shieldings were calculated for the ^{13}C , ^{17}O and ^{31}P nuclei using GIAO method for the optimized structure of intermediate bisphosphonates at B3LYP level of theory using 6-311++G (d, p) basis set. The results showed that various substituents have a strong effect on the nuclear quadrupole resonance (NQR) parameters (χ , η) of ^{17}O in contrast with ^2H NQR parameters. The NMR and NQR parameters were studied in order to find the correlation between electronic structure and the activity of the desired bisphosphonates. In addition, the effect of substitutions on

the bisphosphonates polarity was investigated. Molecular polarity was determined via the DFT calculated dipole moment vectors and the results showed that substitution of bromine atom on the ring would increase the activity of bisphosphonates.

Keywords Bisphosphonate · DFT calculations · Electrical field gradient · NMR · NQR

Introduction

Derivatives of bisphosphonates are a novel class of drugs that have been registered for various clinical applications worldwide [1]. Clinical data confirm the role of bisphosphonate in treatment of bone metastatic cancer and multiple myeloma, breast, prostate and lung cancer. Recent reports suggest bisphosphonates treatment may have a direct effect on the tumor cells [2–6]. These compounds are also potent activators of human $\gamma\delta$ T cell [7, 8]. $\gamma\delta$ T cells have an important role in defense against many infectious organisms and are also involved in killing of tumor cell. $\gamma\delta$ T cells expressing the V γ 2V δ 2 (also known as V γ 9V δ 2) T cell receptor (TCR) play an important role in immune system surveillances [9–12].

Recently, various derivatives of these bisphosphonates have been investigated by Hartree-Fock theory with a 6-31G (p) basis set, e.g., atomic charges are calculated using the Merz-Sinng-Kollman (MSK) method [9, 10] in the Gaussian 98 program. It is reported that the activity of bisphosphonate compound increases with polarity for example, activity of 3, 4-Br $_2$ Ph is greater than 3, 4-Cl $_2$ Ph [13]. In addition, the nuclear magnetic resonance chemical shift and nuclear quadrupole resonance parameters are the most powerful properties available for structure determi-

H. Aghabozorg
Faculty of chemistry, North Tehran branch,
Islamic Azad University,
Tehran, Iran

B. Sohrabi (✉) · S. Mashkouri
Department of chemistry, Surface Chemistry Research Laboratory,
Iran University of Science and Technology,
P.O. Box 16765–163, Tehran, Iran
e-mail: sohrabi_b@yahoo.com

B. Sohrabi
e-mail: sohrabi_b@iust.ac.ir

H. R. Aghabozorg
Research Institute of Petroleum Industry (RIPI),
Tehran, Iran

nation at the molecular level [14–16]. Zhang and Oldfield have carried out an experimental and theoretical investigation of the ^{31}P shielding tensors in phosphonates and bisphosphonates [17]. In their work, calculations and experiment both indicate a large change in tensor orientation between neutral and negatively charged phosphonates. In addition, the isotropic and anisotropic shielding tensors calculations in phosphonates and bisphosphonates have opened a way to the determination of their protonation states when bound to proteins. These information which is not accessible from crystallographic studies, are used to facilitate the drug design [17]. In this research, NMR, NQR and polarity of some derivatives of bisphosphonates as anticancer drug are investigated by using the density functional theory (DFT) calculations with the Gaussian 98 suite of programs in B3LYP/6-31G(d, p) level.

Computational details

The calculations of NMR shieldings and chemical shifts are widely used. These calculations are important tool for determining of new structures at the molecular level [18, 19]. It is important that the gauge is determined for calculating of magnetic properties [20]. This problem is solved in principle by the gauge including atomic orbital (GIAO) method. This method is used both in Hartree-Fock and DFT methods [21]. The GIAO method is a good method because it is less sensitive to basis set quality. The calculated chemical shifts at the Hartree-Fock level are often reasonably accurate, particularly for organic molecules. However, for higher accuracy or in strongly correlated systems, it is necessary to take electron correlation into account [22, 23]. Therefore, DFT method must be used for determining of chemical shifts.

The density functional theory (DFT) studies are carried out using the Gaussian 98 suite of programs [24] and the geometry optimization is performed at the B3LYP/6-31G(d, p) level. To evaluate and ensure the optimized structures of the molecules, frequency calculations were carried out using analytical second derivatives. In all cases, only real frequencies were obtained for the optimized structures.

To calculate the ^2H and ^{17}O EFG tensors in the principal axis system, DFT method including B3LYP [25, 26] with the basis set of 6-311++G(d, p) is employed. To investigate the influence of the substitution on the EFG tensors, all calculations are performed for derivatives of non-nitrogen bisphosphonates.

In this work, gauge-included atomic orbital (GIAO) approach in DFT/B3LYP method is used in the chemical shielding tensor calculations the principal eigenvalues of

chemical shielding tensors σ_{11} , σ_{22} and σ_{33} are found to have the following relationship [27]:

$$\sigma_{33} > \sigma_{22} > \sigma_{11} \quad (1)$$

Chemical shielding anisotropy ($\Delta\sigma$) is obtained by $\Delta\sigma = \sigma_{33} - (\sigma_{22} + \sigma_{11})/2$, and chemical shielding isotropy (σ_{iso}) is obtained by $\sigma_{iso} = (\sigma_{11} + \sigma_{22} + \sigma_{33})/3$ [28].

For EFG tensors q_{xx} , q_{yy} and q_{zz} have the following relationship:

$$|q_{zz}| \geq |q_{yy}| \geq |q_{xx}| \quad (2)$$

The nuclear quadrupole coupling constant (χ) was obtained by

$$\chi(\text{MHz}) = e^2 Q q_{zz} / h \quad (3)$$

where “e” is the charge of electron, Q is the nuclear electric quadrupole moment, and “h” is the Planck’s constant [29]. Q value for ^{17}O and ^2H nuclei used in the calculation of χ values has been reported to be 25.78 and 2.86 mb (1 mb = $1 \times 10^{-31} \text{ m}^2$), respectively [30].

Another important parameter which refers to the deviation of charge distribution from cylindrical symmetry is the asymmetry parameter (η) obtained by [31].

$$\eta = \left| \frac{q_{yy} - q_{xx}}{q_{zz}} \right| \quad (4)$$

The DFT calculated dipole moment vectors show ρ_1 , ρ_3 and ρ_6 components.

The molecular polarity was obtained by $\rho = (\rho_1 + \rho_3 + \rho_6)/3$ [32].

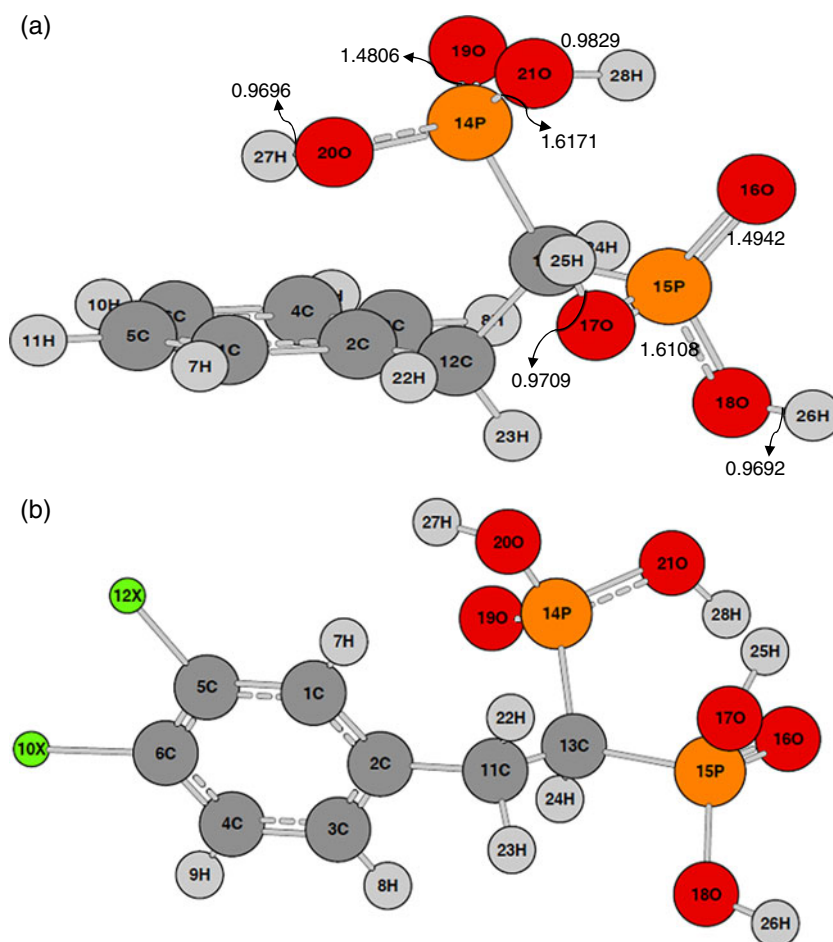
Results

^2H and ^{17}O EFG tensors, the nuclear quadrupole coupling constants (χ), asymmetry parameters (η) and ^{13}C , ^{17}O and ^{31}P chemical shielding are investigated for desired bisphosphonates and their derivatives (substituted dihalo, *ortho*, *meta* and *para*). The B3LYP/6-311++G(d, p) optimized geometries for desired bisphosphonate and its derivatives are shown in Fig. 1.

Electric field gradients

In this part, the DFT calculations at the B3LYP level of theory with the 6-311++G(d,p) basis set are carried out to study the substitution effect on the ^2H and ^{17}O EFG tensors of bisphosphonates. In addition, NQR parameters are used for investigating of the substitution effect on acidic properties of bisphosphonates. The calculated

Fig. 1 The optimized geometry using B3LYP/6-31G (d,p) for bisphosphonate (a) and its derivatives (X = Cl, Br, F and OMe) (b)



nuclear quadrupole coupling constants, χ , and asymmetry parameters, η , by EFG tensor principal components, q_{ii} , for these atoms are summarized in Tables 1 and 2.

Tables 1 and 2 indicate the influences of different substitutions on the calculated ^2H and ^{17}O EFG tensors. In addition, the effect of the substitutions on χ and η is studied. The results show that χ relate to charge density of the atom and its symmetry. χ increases with increase of the charge density and decreases with decrease of the atom symmetry. According to the results in Table 1, the decrease in the EFG tensor elements of oxygen atoms in $\text{P}=\text{O}$ bond can be a result of delocalized electrons. Two factors control the value of q_{zz} for a quadrupolar nucleus: the charge density at the nucleus and the symmetry of the EFG around the nucleus. The double bond in $\text{P}=\text{O}$ increases the charge density at both oxygen atoms (Fig. 1). Since the contribution of nonbonding electrons (lone pairs of p and d electrons) to the nonspherical charge distribution is greater than that of bonding electrons, the EFG is more asymmetric in atoms with nonbonding electron pairs due to the increased charge density. On the other hand, if the asymmetry of EFG increases, then q_{zz} and consequently χ would decrease. As a result, the competing

effects of charge density and EFG asymmetry on χ offset each other, leading to only a small increase in the χ values of the acidic ^{17}O at phosphonate groups and a decrease in the χ values of ^{17}O in $\text{P}=\text{O}$ bonds. Thus, we conclude that the χ values of ^{17}O are a good marker for distinguishing between the acidic and nonacidic forms of oxygen atoms in phosphonate groups. In addition, Table 1 indicates that the χ values of ^{17}O can be used as a good marker for investigating the effect of substitutions on the acidic properties of bisphosphonates.

Tables 1 and 2 indicate that the NQR parameters of the acidic hydrogen atom (28H) on the 14P in bisphosphonate change considerably with respect to other acidic hydrogens on the bisphosphonate. χ (28H) decreases by 55 KHz and η (28H) increases by 0.03 through position of this hydrogen atom with respect to benzene ring. Furthermore, the change of substitution affects the NQR parameters of the acidic hydrogen atom (28H) on the 14P in bisphosphonate and their derivative. The decrease of χ (28H) and increase of the O–H length in bisphosphonate and their derivatives lead to increasing acidity of the hydrogen atom on the 21O–28H bond (Fig. 1a).

Table 1 Nuclear quadrupole coupling constants (NQCC), χ , and asymmetry parameters calculated for ^2H and ^{17}O nuclei for bisphosphonate and its derivatives using DFT-B3LYP/6-311++G(d, p) level

Compounds Number	Acid		Di Bromo acid		Di Chloro acid		Di Fluoro acid		Chloro and Bromo		Meta-Br		Meta-Cl		Meta-F	
	χ	η	χ	η	χ	η	χ	η	χ	η	χ	η	χ	η	χ	η
H(13C)	24	187.80	0.02	186.45	0.03	186.39	0.03	190.21	0.02	186.26	0.03	185.90	0.03	186.46	0.03	187.34
O(14P)	19	3.36	1.01	3.11	2.09	3.08	2.12	3.39	1.91	3.07	2.13	3.24	1.95	3.24	1.95	3.32
	20	7.82	1.11	7.84	1.15	7.85	1.15	7.24	1.09	7.82	1.15	7.82	1.13	7.82	1.13	7.82
	21	7.12	1.11	6.96	1.15	6.96	1.16	7.71	1.13	6.96	1.17	7.03	1.14	7.03	1.14	7.08
O(15P)	16	3.16	2.20	3.09	2.29	3.08	2.30	3.31	2.06	3.07	2.32	3.13	2.25	3.13	2.25	3.15
	18	7.76	1.13	7.69	1.18	7.68	1.19	7.91	1.17	7.67	1.19	7.74	1.16	7.74	1.16	7.76
	17	8.13	1.13	8.11	1.13	8.10	1.13	7.12	1.10	8.09	1.13	8.11	1.13	8.11	1.13	8.12
H	25	290.76	0.11	295.54	0.11	295.52	0.11	295.07	0.11	296.67	0.11	293.76	0.11	294.93	0.11	291.97
	26	298.21	0.12	298.41	0.11	298.03	0.11	287.70	0.12	298.61	0.11	298.60	0.11	298.11	0.11	298.06
	27	292.72	0.12	296.97	0.11	297.16	0.11	297.22	0.11	296.44	0.11	296.85	0.11	296.11	0.11	293.495
	28	243.77	0.14	235.36	0.14	234.69	0.14	224.84	0.16	232.71	0.14	237.81	0.14	235.74	0.14	241.06
Compounds Number		Orto-Br		Orto-Cl		Orto-F		P-Br		P-Cl		P-F		P-OMe		
H(13C)	24	182.31	0.02	183.42	0.02	184.93	0.02	187.76	0.03	187.08	0.03	187.50	0.03	187.67	0.02	
O(14P)	19	3.01	2.21	2.93	2.31	3.19	2.00	3.34	1.84	3.26	1.93	3.30	1.88	3.16	2.20	
	20	7.87	1.14	7.88	1.14	7.82	1.13	7.83	1.12	7.82	1.13	7.83	1.12	7.77	1.13	
	21	6.91	1.18	6.90	1.19	6.97	1.15	7.09	1.12	7.02	1.14	7.06	1.13	8.13	1.13	
O(15P)	16	3.17	2.21	3.15	2.22	3.15	2.22	3.16	2.21	3.12	2.26	3.16	2.21	3.36	1.83	
	18	7.75	1.17	7.73	1.17	7.75	1.16	7.75	1.14	7.72	1.17	7.75	1.15	7.82	1.10	
	17	8.12	1.14	8.12	1.14	8.12	1.13	8.12	1.13	8.11	1.13	8.12	1.13	7.11	1.11	
H	25	294.94	0.11	296.00	0.11	294.73	0.11	291.85	0.11	294.58	0.11	292.47	0.11	292.15	0.12	
	26	297.90	0.11	298.45	0.11	298.40	0.11	297.92	0.11	298.13	0.11	298.21	0.11	242.71	0.11	
	27	297.23	0.11	297.88	0.11	295.69	0.11	294.61	0.11	295.45	0.11	294.74	0.11	291.57	0.11	
	28	231.91	0.14	228.82	0.15	234.12	0.14	242.27	0.14	237.14	0.14	239.96	0.14	298.29	0.14	

 χ of ^{17}O is in MHz and for ^2H in KHz

Table 2 Calculated the largest component of the EFG tensor, q_{zz} , for ^2H and ^{17}O nuclei in bisphosphonate and its derivatives using DFT-B3LYP/6-311++G (d, p) level

Compounds	Acid	Di Bromo acid	Di Chloro acid	Di Fluoro acid	Chloro and Bromo	Meta -Br	Meta-Cl	Meta-F
Number	q_{zz}	q_{zz}	q_{zz}	q_{zz}	q_{zz}	q_{zz}	q_{zz}	q_{zz}
H(13C)	24	0.28	-0.28	0.28	-0.28	-0.28	-0.28	-0.27
O(14P)	19	-0.38	-0.79	-0.79	-0.79	-0.79	-0.79	-0.78
	20	1.22	-1.40	-1.39	-1.39	-1.37	-1.38	-1.36
	21	-1.24	-1.24	-1.24	-1.25	-1.24	-1.23	-1.24
O(15P)	16	-0.83	-0.84	-0.84	-0.84	-0.84	-0.84	-0.84
	18	-1.36	-1.38	-1.38	-1.42	-1.38	-1.38	-1.37
	17	-1.43	-1.42	-1.42	-1.23	-1.42	-1.43	-1.43
H	25	-0.43	-0.44	-0.44	-0.33	-0.44	-0.44	-0.43
	26	-0.44	-0.44	-0.44	-0.43	-0.44	-0.44	-0.44
	27	-0.43	-0.44	-0.44	-0.44	-0.44	-0.44	-0.44
	28	-0.36	-0.35	-0.35	-0.33	-0.35	-0.35	-0.36
Compounds	Orto-Br	Orto-Cl	Orto-F	P-Br	P-Cl	P-F	P-OMe	
Number	q_{zz}	q_{zz}	q_{zz}	q_{zz}	q_{zz}	q_{zz}	q_{zz}	
H(13C)	24	-0.27	-0.27	-0.27	-0.28	-0.28	-0.28	
O(14P)	19	-0.80	-0.80	-0.79	-0.78	-0.78	-0.78	
	20	-1.39	-1.39	-1.37	-1.37	-1.37	-1.35	
	21	-1.24	-1.25	-1.24	-1.24	-1.24	-1.24	
O(15P)	16	-0.84	-0.84	-0.84	-0.84	-0.84	-0.84	
	18	-1.38	-1.39	-1.38	-1.37	-1.37	-1.37	
H	17	-1.43	-1.39	-1.43	-1.42	-1.43	-1.43	
	25	-0.44	-0.44	-0.44	-0.43	-0.43	-0.43	
	26	-0.44	-0.44	-0.44	-0.44	-0.44	-0.43	
	27	-0.44	-0.44	-0.44	-0.44	-0.44	-0.36	
	28	-0.34	-0.34	-0.35	-0.36	-0.36	-0.43	
							-0.44	

q_{zz} values in atomic units, 1 au=9.717365×10²¹ Vm⁻²

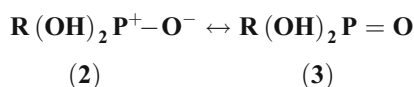
Table 3 The calculated chemical shieldings of ^{17}O and ^{31}P (Fig. 1) atoms B3LYP/6-311++G(d, p) in bisphosphonate and its derivatives

Compounds	Acid	Di Bromo acid	Di Chloro acid	Di Fluoro acid	Chloro and Bromo	Orto-Br	Orto-Cl	Orto-F					
Number of atoms	σ_{iso}	$\Delta\sigma$	σ_{iso}	$\Delta\sigma$	σ_{iso}	$\Delta\sigma$	σ_{iso}	$\Delta\sigma$					
P	14	363.87	229.99	222.19	365.21	222.05	343.52	157.84	365.33	222.51	364.37	224.99	
	15	338.00	172.60	337.83	169.71	337.98	168.92	348.46	167.91	336.51	163.06	336.76	167.57
O	19	137.62	68.80	143.96	60.73	144.17	60.18	169.91	54.65	143.69	59.35	142.32	62.97
	20	205.03	77.18	202.48	79.06	202.46	79.55	222.25	68.70	203.94	80.25	204.61	78.77
	21	225.76	95.75	228.59	106.07	228.29	107.10	240.17	107.20	227.36	105.84	226.29	101.51
	16	173.50	50.43	175.51	47.77	175.91	47.34	164.27	48.24	176.51	46.02	175.22	48.32
	17	239.77	97.24	237.78	93.95	237.51	94.08	218.21	88.08	236.69	94.04	238.40	95.56
	18	229.24	95.02	229.54	100.93	229.74	101.26	223.44	70.06	229.89	102.43	229.86	97.87
C	13	142.70	20.87	143.75	19.05	144.11	19.59	143.35	18.59	144.43	19.84	146.64	25.77
Compounds	Meta -Br	Meta-Cl	Meta-F	P-Br	P-Cl	P-F	P-OMe						
Number of atoms	σ_{iso}	$\Delta\sigma$	σ_{iso}	$\Delta\sigma$	σ_{iso}	$\Delta\sigma$	σ_{iso}	$\Delta\sigma$	σ_{iso}	$\Delta\sigma$	σ_{iso}	$\Delta\sigma$	
P	14	364.56	226.60	364.86	225.36	364.21	229.31	364.10	228.47	364.71	225.92	363.97	230.36
	15	338.17	170.39	337.90	169.15	338.20	171.81	338.70	172.87	337.95	170.16	337.59	171.72
O	19	140.65	63.95	140.94	63.77	138.49	67.27	137.85	67.87	139.18	65.62	197.92	68.94
	20	204.17	77.95	204.12	77.95	204.73	77.02	204.95	77.55	205.02	77.87	205.55	77.37
	21	226.67	100.31	226.35	100.21	225.76	95.95	226.44	97.72	226.08	100.12	225.94	95.62
	16	175.03	48.57	175.22	47.79	174.05	49.981	173.74	49.92	174.64	48.05	173.57	50.33
	17	238.75	95.03	238.23	94.94	239.69	96.84	239.57	96.57	238.26	94.99	239.48	97.42
	18	229.40	98.51	229.59	99.43	229.23	95.73	229.64	95.60	230.04	98.59	229.32	95.08
C	13	143.34	19.57	143.15	19.062	142.75	19.98	143.28	21.36	143.91	21.18	142.70	21.22

Investigation of NMR parameters (chemical shielding tensors) in bisphosphonates and their derivatives

In this part, the focus is on the effect of the structure of molecule on the ^{13}C , ^{31}P and ^{17}O NMR chemical shielding. To achieve this aim, DFT calculation is carried out at the B3LYP level of theory with 6-311++G (d, p) basis set for desired bisphosphonate acids and their derivatives (substituted dihalo, *ortho*, *meta* and *para*). The calculated chemical shielding tensors are reported as chemical shielding principal components (σ_{ij}), chemical shielding isotropy (σ_{iso}), and chemical shielding anisotropy ($\Delta\sigma$) in Table 3.

Phosphorous atom is in the second row of the periodic table and has empty *d* orbital which overlaps with filled *p* orbital of oxygen atom. Thus, the following resonance structures may be considered



The P = O double bond in (2) is expected to be quite different from a normal *pπ-pπ* double bond of, for example, a carbonyl group. The 3*d* orbitals are diffuse and more directional than 3*p* orbitals and most of the electron density in a *pπ-dπ* bond is expected to lie in the vicinity of the oxygen atom [33]. The obtained chemical shielding for ^{31}P and ^{17}O in Table 3 confirm the above discussion. The ^{31}P (15) chemical shielding of bisphosphonate is deshielded by 25.88 ppm from ^{31}P (14). If the ^{31}P NMR shielding tensors are considered in more detail, NMR spectra of phosphorus

may be interpreted by two types of phosphorus site in the studied phosphonates here: one is neutral with two hydroxyl groups, PO (OH)₂, while the other has one formal negative charge with one of the two hydroxyl groups which is deprotonated, PO (OH) O⁻. The results show that in bisphosphonates, the O = P-O⁻ group does not contain two almost equal P-O bond lengths (Table 4).

The difference in chemical shielding of two types of phosphorus site is consistent with an expected smaller contribution of resonance structure (2) for the ^{31}P (15). In addition, the bond length and bond order of 15P-22O and 14P-19O were given in Table 4. These results show that as the chemical shielding decreases, the bond length P-O decreases and its bond order increases. According to the Ramsey theory of nuclear magnetic shielding, the shielding of a nucleus can be separated into two main contributions, the diamagnetic shielding (σ^d) and the paramagnetic shielding (σ^p) [31, 32]. The diamagnetic shielding contribution describes the shielding of the nucleus from the external magnetic field by the surrounding electrons that induce a magnetic field opposite to the external one. The paramagnetic shielding contribution is a perturbation of the electron density currents that generally causes a decrease in the absolute shielding. In other words, the paramagnetic contribution is typically responsible for observed changes in chemical shifts for a given nucleus. Therefore, the ^{31}P NMR chemical shielding is dominated by paramagnetic shielding term, while the changes in the diamagnetic term are comparatively small. The results show that chemical shielding of ^{31}P (14) is more than that of ^{31}P (15), since ^{31}P

Table 4 The calculated bond lengths of between ^{17}O and ^{31}P (Fig. 1) atoms and polarity by using B3LYP/6-311++G (d, p) in bisphosphonate and its derivatives

Compounds	Acid	Di Bromo acid	Di Chloro acid	Di Fluoro acid	Chloro and Bromo acid	Orto-Br acid	Orto-Cl acid	Orto-F acid
14P-nO	19 1.4806	1.4811	1.4813	1.4981	1.4806	1.4813	1.4813	1.4816
	20 1.6171	1.6123	1.6082	1.5996	1.6148	1.6086	1.6064	1.6061
	21 1.6185	1.6228	1.6215	1.6117	1.6192	1.6252	1.6261	1.6232
15P-nO	16 1.4942	1.4944	1.4949	1.4980	1.4806	1.4958	1.4962	1.4962
	17 1.6108	1.6124	1.6104	1.6014	1.6148	1.6088	1.6088	1.6100
	18 1.6210	1.6229	1.6209	1.6137	1.6192	1.6222	1.6219	1.6226
Polarity	130.4	169.7	155.0	131.6	148.8	147.5	140.7	130.5
Compounds	Meta -Br acid	Meta-Cl acid	Meta-F acid	P-Br acid	P-Cl acid	P-F acid	P-OMe acid	
14P-nO	19 1.4813	1.4813	1.4816	1.4809	1.4812	1.4808	1.4807	
	20 1.6115	1.6099	1.6148	1.6079	1.6099	1.6138	1.6164	
	21 1.6208	1.6132	1.6194	1.6241	1.6205	1.6197	1.6190	
15P-nO	16 1.4946	1.4949	1.4943	1.4948	1.4948	1.4946	1.4947	
	17 1.6102	1.6101	1.6103	1.5996	1.6105	1.6105	1.6111	
	18 1.6213	1.6215	1.6207	1.6113	1.6220	1.6210	1.6216	
Polarity	149.6	142.3	131.0	148.8	143.5	131.0	149.8	

¹ n is the number of oxygen atoms in Fig. 1

(15) is vertical on benzene ring. Therefore, phosphorus atom should result in an increased polarization of the charge distribution of ^{31}P leading to an increased paramagnetic shielding. Table 3 shows that the magnitude of chemical shift is directly related to the energy gap between the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO). Furthermore, the magnitude of σ^p is inversely related to the energy gap between the relevant occupied and virtual orbitals, with a smaller energy gap leading to a larger chemical shielding. The required mixing may be visualized as rotations of an occupied orbital about one of the three Cartesian axes to produce constructive overlap with an appropriate virtual molecular orbital.

The polarity of the bisphosphonates

The polarity of the bisphosphonate and its derivatives is presented in Table 4. Zhang and et al. showed that if bisphosphonates polarity increases then their IC_{50} decreases but their activity increases [34]. The results in Table 4 show the type and position of substitution affect on polarity of bisphosphonates. The bromine substitution and its position on the ring affect more than other substitutions. In addition, the polarity of bisphosphonate is increased using O-Me substitution. On the other hand, electron donating groups such as O-Me increase the polarity of bisphosphonates. Since bromine atom electronegativity is less than chlorine and fluorine atoms, the bromine atom increases the polarity of bisphosphonates.

Summary and conclusions

Based on DFT calculations, it is concluded that the EFG tensors of oxygen atoms are good indicators to characterize the acidic property of hydrogen atoms in phosphate groups. The NQR parameters of oxygen atom in $\text{P}=\text{O}$ bond change significantly through delocalized electron. The results show that the electronic environments of oxygen atoms are affected by benzene ring and its substitutions. The position of 19O and 28H causes notable changes in the EFG tensors of these atoms. Furthermore, the change of substitution affects the NQR parameters of the acidic hydrogen atom on the 14P (28H) in bisphosphonates and their derivative. The decrease of χ (28H) and increase of the O–H bond length in bisphosphonates and their derivatives the increased acidity of the hydrogen atom in the 21 O–28H bond. The results show that chemical shielding of ^{31}P (14) is more than that of ^{31}P (15). Therefore, the phosphorus atom should result in an increased polarization of the charge distribution of ^{31}P leading to an increased paramagnetic shielding. The polarity of the bisphosphonates shows that the type and position of substitution affects this quantity. This effect for

bromine substitution is more than the others. Since bromine atom electronegativity is less than chlorine and fluorine atoms, the bromine atom increases the polarity of bisphosphonates.

References

- Hewitt RE, Lissina A, Green AE, Slay ES, Price DA, Sewell AK (2005) *Clin Exp Immunol* 139:101–111
- Neville-Webbe HL, Holen I, Coleman RE (2002) *Cancer Treat Rev* 28:305–319
- Coleman RE (2001) *Cancer Treat Rev* 27:165–176
- Coleman RE (2002) *Semin Oncol* 29:43–49
- Boissier S, Magnetto S, Frappart L, Cuzin B, Ebetino FH, Delma PD, Clezardin P (1997) *Cancer Res* 57:3890–3894
- Rosen LS, Gordon D, Tchekmedyan NS (2004) *Cancer* 100:2613–2621
- Wilhelm M, Kunzmann V, Eckstein S, Reimer P, Weissinger F, Ruediger T, Tony HP (2003) *Blood* 102:200–206
- Sanders JM, Ghosh S, Chan JM, Meints G, Wang H, Raker AM, Song Y, Colantino A, Burzynska A, Kafarski P, Morita CT, Oldfield E (2004) *J Med Chem* 47:375–384
- Hayday AC (2000) *Annu Rev Immunol* 18:975–1026
- Bendelac A (2001) *Nat Rev Immunol* 1:177–185
- Girardi M, Oppenheim DE, Steele CR, Lewis JM, Glusac E, Filler R, Hobby P, Sutton B, Tigelaar ER, Hayday AC (2001) *Science* 294:605–609
- Morita CT, Mariuzza RA, Brenner MB (2000) *Springer Semin Immunopathol* 22:191–217
- Zhang Y, Leon A, Song Y, Studer D, Haase Ch, Koscielski LA, Oldfield E (2006) *J Med Chem* 49:5804–5814
- Fritscher J (2004) *Phys Chem Chem Phys* 6:4950–4956
- Nakamura N, Masui H, Ueda T (2000) *Z Naturforsch* 55a:315–322
- Latosińska JN, Seliger J, Nogaj B (1999) *Magn Reson Chem* 37:878–880
- Zhang Y, Oldfield E (2004) *J Phys Chem B* 108:19533–19540
- Helgaker T, Jaszunski M, Ruud K (1999) *Chem Rev* 99:293–352
- Perczel A, Csaszar AG (2000) *J Comput Chem* 21:882–900
- Hameka HF (1963) *Advanced quantum chemistry*. Addison-Wesley, New York
- Hui-ding X, Yu-peng L, Kai-xiong Q, Bo L, Ya-ping C (2010) *Chem Res Chin Univ* 26:1016–1019
- Gauss J, Stanton JF (2002) In: Prigogine I, Rice SA (eds) *Advances in chemical physics*. Wiley, Chichester, pp 123–355
- Casabianca LB, Dios AC (2008) *J Chem Phys* 128:052201–10
- Frisch MJ et al (1998) *Gaussian 98, Revision A7*. Gaussian Inc, Pittsburgh
- Becke AD (1993) *J Chem Phys* 98:5648–5652
- Lee C, Yang W, Parr RG (1988) *Phys Rev B* 37:785–789
- Wolinski K, Hinton JF, Pulay P (1990) *J Am Chem Soc* 112: P8251–8260
- Daniel M, Jordan K, Maria M, Ioan A, Akash B, Kim Z, Erik RP (2007) *Chem Phys Chem* 8:1375–1385
- Tossell JA, Paolo L (1986) *J Phys B at Mol Phys* 19:3217–3226
- Pykkö P (2001) *Mol Phys* 99:1617–1629
- Garcia ME, Bennemann KH (1989) *Phys Rev B* 40:8809–8813
- Karelson M, Lobanov VS, Katritzky AR (1996) *Chem Rev* 96:1027–1043
- Mitchell KAR (1969) *Chem Rev* 89:157–178
- Zhang Y, Hudock MP, Krysiak K, Cao R, Bergan K, Yin F, Leon A, Oldfield E (2007) *J Med Chem* 50:6067–6079

Time-dependent density functional theory study on the electronic excited-state hydrogen bonding of the chromophore coumarin 153 in a room-temperature ionic liquid

Dandan Wang · Ce Hao · Se Wang · Hong Dong · Jieshan Qiu

Received: 7 February 2011 / Accepted: 15 May 2011 / Published online: 3 June 2011
© Springer-Verlag 2011

Abstract In the present work, in order to investigate the electronic excited-state intermolecular hydrogen bonding between the chromophore coumarin 153 (C153) and the room-temperature ionic liquid *N,N*-dimethylethanolammonium formate (DAF), both the geometric structures and the infrared spectra of the hydrogen-bonded complex C153–DAF⁺ in the excited state were studied by a time-dependent density functional theory (TDDFT) method. We theoretically demonstrated that the intermolecular hydrogen bond C₁=O₁⋯H₁–O₃ in the hydrogen-bonded C153–DAF⁺ complex is significantly strengthened in the S₁ state by monitoring the spectral shifts of the C=O group and O–H group involved in the hydrogen bond C₁=O₁⋯H₁–O₃. Moreover, the length of the hydrogen bond C₁=O₁⋯H₁–O₃ between the oxygen atom and hydrogen atom decreased from 1.693 Å to 1.633 Å upon photoexcitation. This was also confirmed by the increase in the hydrogen-bond binding energy from 69.92 kJ mol⁻¹ in the ground state to 90.17 kJ mol⁻¹ in the excited state. Thus, the excited-state hydrogen-bond strengthening of the coumarin chromophore in an ionic liquid has been demonstrated theoretically for the first time.

Keywords Hydrogen-bonding dynamics · Excited state · Hydrogen bond strengthening · Spectral shift

D. Wang · C. Hao (✉) · S. Wang · H. Dong · J. Qiu
State Key Laboratory of Fine Chemicals,
School of Chemical Engineering,
Dalian University of Technology,
Dalian,
116024 Liaoning, China
e-mail: haoce_dlut@126.com

Introduction

Numerous experimental and theoretical methods have been developed to investigate the nature of a hydrogen bond linking a solute with a polarizable functional group and a protic solvent. Since intermolecular hydrogen bonds are site-specific solute–solvent interactions, they play a fundamental role in the molecular photochemistry of organic and biological chromophores in solution [1–24]. Upon photoexcitation, the intermolecular hydrogen bonds formed between solute and solvent molecules will reorganize themselves as the result of differences in the charge distribution of the different electronic states; this process is termed *hydrogen-bonding dynamics*, and it is linked to photochemical and photophysical processes [25–31]. A strengthening of the hydrogen bond between C102 and phenol early during photoexcitation to the electronic excited state was first demonstrated by Zhao and Han theoretically [4], and since then a great deal of work has focused on hydrogen-bonding dynamics in the excited state [4–10]. Their work has already yielded much information on the structural and relaxation dynamics of hydrogen bonds after photoexcitation, which has aided our understanding of fluorescence-quenching phenomena in the excited state [32–35]. However, previous studies focused on the intermolecular hydrogen bonds that form between chromophores and traditional polar protic solvents, and while great progress has been made in this field [36–44], fewer studies have been conducted on the solute–solvent interactions between chromophores and room-temperature ionic liquids [45–47].

We have shifted the attention in our research away from traditional solvents and towards solvents containing ions—

room-temperature ionic liquids (RTILs). The properties of RTILs—a novel class of molten salts that mainly comprise organic cations and inorganic anions with melting points below room temperature—are currently receiving a great deal of attention at present [48–69]. Due to their unique ingredients, their properties differ drastically from those of conventional organic solvents [51, 52]. Recent work in this field has focused on the dynamics of the solvation of a solvatochromic probe in an RTIL, which has been explored through experimental studies and spectroscopic measurements [53–66]. For example, Mroncelli and coworkers [65] have monitored the steady-state spectra, rotation times, and time-resolved emission spectra of the probe 4-aminophthalimide (4-AP) in the ionic liquid 1-*n*-butyl-3-methylimidazolium hexafluorophosphate ([bmim⁺][PF₆⁻]). They found that the solvation energy of 4-AP in [bmim⁺][PF₆⁻] is comparable to those of 4-AP in highly polar but aprotic solvents, and they demonstrated that [bmim⁺][PF₆⁻] possesses essentially no hydrogen bond donating ability, so no hydrogen bonds form in the 4-AP [bmim⁺][PF₆⁻] system. The solvation and rotational dynamics of coumarin 153 (C153) in a series of phosphonium ionic liquids has also been reported by Mroncelli and coworkers [66]. To investigate the influence of specific hydrogen-bonding interactions on solvation and rotational dynamics in RTILs, Paul and Samanta [46] performed spectroscopic measurements to study the behavior of C153 in an alcohol-functionalized room-temperature ionic liquid, 1-(hydroxyethyl)-3-methylimidazolium bis(trifluoromethanesulfonyl)imide, abbreviated to [OH-emim][Tf₂N]. The presence of the OH group in [OH-emim][Tf₂N] makes it a good hydrogen bond donor, and the occurrence of hydrogen-bonding interactions between the probe molecule C153 and the hydroxylated cation was confirmed by experimental measurements; furthermore, the hydrogen bonding exerts a significant influence on the overall dynamics in RTILs [46]. Cation–anion hydrogen-bonding associations in the first solvation shell can also help to reduce the ultrafast component of the dynamics [46]. Besides investigations into the dynamics of RTILs, efforts have also been directed into the study of hydrogen bonds between ionic pairs [69]. For instance, Dhumal et al. [69] provided very useful insights into the intermolecular interactions between the 1-methyl-3-imidazolium cation and acetate anion by combining theoretical analysis with experimental methods. In other words, hydrogen bonding in RTILs is the focus of much important research.

C153 is widely utilized as a solvation probe to monitor the nature of a solvent, owing to its rigid structure and the large change in dipole moment that is caused by photoexcitation [67, 68]. C153 was also employed in the investigation reported by Seth et al., who found that a nonbonding

interaction formed between the cation and the anion from the optimized structure of *N,N*-dimethylethanolammonium formate (DAF) [45]. Moreover, they concluded that the rotational dynamics of C153 were hindered in DAF compared to the viscous flow of DAF, which is possibly due to hydrogen-bond formation for C153 in DAF [45]. Their work has played an important role in showing that hydrogen-bond formation affects the rotational dynamics of a solute, and thus affects investigations of the nature of RTILs. Significantly, their work aroused our interest in studying the hydrogen bonding that forms between solute and solvent in RTILs. To our knowledge, little theoretical work has been performed on the hydrogen bonding between solvatochromic probe molecules and solvent molecules in novel RTIL systems. To determine the precise nature of the hydrogen bonding in this novel system, further theoretical methods need to be adopted for excited-state geometry optimization and electronic transition calculations. TDDFT is accepted as a reliable method for excited-state computation, and it can also be used to calculate the IR spectrum in the electronically excited state [70–74]. Infrared spectra also reflect hydrogen-bonding dynamics, since such dynamics occur on an ultrafast timescale that is primarily dictated by the vibrational modes of the atoms engaged in the formation of the hydrogen bond [75–78]. Therefore, in this study, we were motivated to research the hydrogen-bonded dimer C153–DAF⁺ that forms between isolated C153 and the DAF⁺ cation in ionic liquid DAF, and the TDDFT method was performed to study the hydrogen dynamics of this hydrogen-bonded C153–DAF⁺ complex in the electronically excited state. The calculated absorption peak of C153 in DAF is 407 nm, which is in good agreement with experiment results [45]. At the same time, the basis set superposition error (BSSE) for the intermolecular hydrogen bond calculated using the MP2 method only accounts for a small proportion of the intermolecular hydrogen-bond binding energy. Furthermore, the calculated results for the proposed hydrogen-bonded complex C153–DAF⁺ are also consistent with the mechanism of hydrogen bond strengthening in the electronically excited state that was first demonstrated by Zhao and Han [4].

Computational details

All of the electronic structure calculations were carried out using the TURBOMOLE program suite. The ground-state geometric optimization was performed using the density function theory (DFT) method with Becke's three-parameter hybrid exchange function and the Lee–Yang–Parr gradient-correlation functional (B3LYP functional) [79]. The excited state electronic structures were calculated

using the time-dependent density functional theory (TDDFT) with the B3LYP functional. In both the ground-state and excited-state geometric optimizations, triple- ζ valence quality basis sets with one set of polarization functions (TZVP) were chosen [80]. Fine quadrature grids 4 were also employed [81]. Harmonic vibrational frequencies in the ground state and excited state were determined by diagonalizing the Hessian [82]. The excited-state Hessian was obtained by the numerical differentiation of analytical gradients using central differences and a default displacement of 0.02 bohr. The infrared intensities were determined from the gradients of the dipole moment [83]. The BSSEs were calculated at the MP2/TZVP level.

Results and discussion

Geometric structures in the ground state

The optimized geometry in the ground state of the hydrogen-bonded complex C153–DAF⁺, where the oxygen atom of the carbonyl group in C153 is linked to the hydrogen atom of the hydroxyl group in the cation DAF⁺ in the RTIL DAF is shown in Fig. 1. We chose the hydrogen-bonded complex C153–DAF⁺ here to study the ultrafast hydrogen-bonding dynamics for the following reason. As we know, in traditional polar protic solvents, the solvent molecules reorient themselves around the photoexcited solute molecules to form many solvation shells, but only the solvent molecules in the inner solvation shell can be taken into consideration early on in hydrogen-bonding dynamics that occur on the ultrafast timescale [4]. The model built by Zhao and Han for traditional polar protic solvents can be adopted for the hydrogen bonding in RTILs too, as the solvation time in most conventional solvents is extremely short (≤ 10 ps), but, the solvation time in neat RTILs is rather long (in the range of 0.1–10 ns) [65, 66]. In addition, only the hydrogen bond C₁=O₁⋯H₁–O₃ joining the C₁=O₁ group of isolated C153 and the O₃–H₁ group of the isolated cation in DAF is studied here.

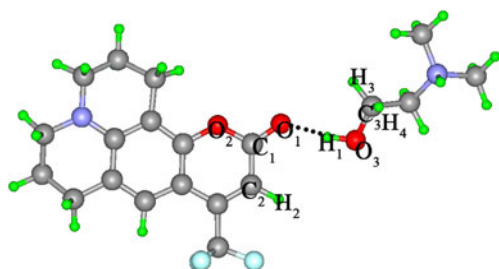


Fig. 1 Optimized geometric configuration of the hydrogen-bonded complex C153–DAF⁺; dotted line denotes the intermolecular hydrogen bond

Some calculated bond lengths, angles and dihedral angles in the hydrogen-bonded complex C153–DAF⁺ as well as the related monomers in the ground state are listed in Table 1. In the optimized geometric structure of the hydrogen-bonded complex C153–DAF⁺, the bond angles C₁=O₁⋯H₁ and O₁⋯H₁–O₃ are 29.89° and 169.4°, respectively. In addition, the calculated dihedral angle C₁=O₁⋯H₁–O₃ is 10.38°. The calculated length of the hydrogen bond C₁=O₁⋯H₁–O₃ between the oxygen atom and hydrogen atom is 1.693 Å, which, generally speaking, is shorter than the hydrogen bonds formed in traditional solvents [4–10]. In Table 1, the calculated bond length for C₁=O₁ in isolated C153 is 1.204 Å, which increases to 1.227 Å upon the formation of the intermolecular hydrogen bond C₁=O₁⋯H₁–O₃. At the same time, the O₃–H₁ bond in isolated cation DAF⁺ increases slightly in length from 0.965 Å to 0.987 Å upon the formation of the intermolecular hydrogen bond C₁=O₁⋯H₁–O₃. These changes are similar to the conditions in other traditional solvents [4–10]. The length of the C₁–C₂ bond in isolated C153 and that of the C₃–O₃ bond in the isolated cation DAF⁺ were calculated to be 1.450 Å and 1.410 Å, respectively, and they decrease to 1.433 Å and 1.396 Å upon the formation of the hydrogen-bonded complex C153–DAF⁺. The C₁–O₂ bond in isolated C153 shortens from 1.391 Å to 1.366 Å upon the formation of the hydrogen-bonded complex C153–DAF⁺. However, the lengths of the C₂–H₂, C₃–H₃ and C₃–H₄ bonds remain almost unchanged upon forming the hydrogen-bonded complex C153–DAF⁺.

Electronic spectra

To understand the nature of the excited states of C153 and its hydrogen-bonded dimer C153–DAF⁺, we need to investigate the properties of the low-lying electronically excited states in detail. The electronic excitation energies and corresponding oscillator strengths for the singlet excited states of the hydrogen-bonded dimer C153–DAF⁺ as well as the involved monomers were calculated using the TDDFT method, and the results are shown in Table 2. Both the isolated C153 and the hydrogen-bonded complex C153–DAF⁺ can be initially photoexcited to the S₁ state, since the S₁ states of both species have larger oscillator strengths than the other states. The absorption peak of the hydrogen-bonded complex C153–DAF⁺ was calculated to occur at 407 nm; in experiments, the absorption maximum of C153 in DAF is found at about 425 nm [45]. Thus, our theoretical calculation is very close to the experimental value. Interestingly, it should be noted that all of the excitation energies of the hydrogen-bonded complex C153–DAF⁺ are slightly redshifted compared with those of the isolated C153 in different electronically excited states, which indicates that the intermolecular hydrogen-bonding interactions can reduce the excitation energies of the

Table 1 Calculated bond lengths (Å), angles (°) and dihedral angles (°) of the isolated monomers and the hydrogen-bonded complex C153–DAF⁺ in the ground state

Parameter		C153	DAF ⁺	C153–DAF ⁺
Bond length (Å)	C ₁ =O ₁	1.204		1.227
	O ₁ ⋯H ₁			1.693
	O ₃ –H ₁		0.965	0.987
	C ₁ –O ₂	1.391		1.366
	C ₁ –C ₂	1.450		1.433
	C ₂ –H ₂	1.079		1.079
	C ₃ –O ₃		1.410	1.396
	C ₃ –H ₃		1.095	1.095
	C ₃ –H ₄		1.101	1.106
	Bond or dihedral angle (°)	C ₁ =O ₁ ⋯H ₁		
O ₁ ⋯H ₁ –O ₃				169.4
C ₁ =O ₁ ⋯H ₁ –O ₃				10.38

hydrogen-bonded dimer C153–DAF⁺. This may be useful information for us when investigating the changes in the hydrogen bond with different electronically excited states. Furthermore, the excitation energies of the isolated cation DAF⁺ are much larger than those of the isolated C153 and the hydrogen-bonded dimer C153–DAF⁺. The data mentioned above reveal that only the C153 moiety is electronically excited when the hydrogen-bonded complex C153–DAF⁺ is photoexcited to the S₁ state; the DAF⁺ moiety remains in its electronic ground state. Thus, the S₁ state of the hydrogen-bonded complex C153–DAF⁺ is defined as a locally excited (LE) state [83, 84]. From Table 2, we can also obtain information on the orbital transition that contributes to the S₁ state of the isolated C153 and the hydrogen-bonded complex C153–DAF⁺: both of the S₁ states correspond to the molecular orbital transition from the highest occupied orbital (HOMO) to the lowest unoccupied orbital (LUMO) according to our TDDFT calculations.

Frontier molecular orbitals

The two main frontier molecular orbitals involved in the first photoexcited state of the hydrogen-bonded complex C153–

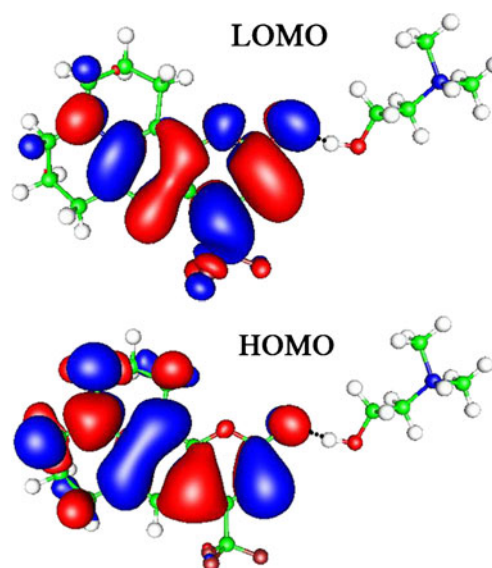
Table 2 Calculated electronic excitation energies (nm) and the corresponding oscillator strengths of the isolated monomers as well as the hydrogen-bonded complex C153–DAF⁺

	C153	DAF ⁺	C153–DAF ⁺
S ₁	380 (0.336)	191 (0.007)	407 (0.374)
	H→L 96.3%	H→L 98.8%	H→L 95.5%
S ₂	323 (0.024)	168 (0.002)	371 (0.000)
S ₃	282 (0.000)	156 (0.015)	332 (0.047)
S ₄	279 (0.040)	150 (0.014)	281 (0.064)
S ₅	258 (0.067)	148 (0.013)	277 (0.000)
S ₆	241 (0.045)	145 (0.002)	266 (0.000)

DAF⁺ are depicted in Fig. 2. It is clear that the electron densities of the HOMO and LUMO orbitals are localized on the C153 moiety. Thus, the S₁ state of the hydrogen-bonded complex C153–DAF⁺ has the characteristics of the LE state. After further observation, it is apparent that the electron density distribution of the HOMO is comparatively uniform, but it is deformed for the LUMO because the electron density reaches to the side of the C₁=O₁ group moiety. Consequently, the electron density of the C₁=O₁ group is strengthened in the first photoexcited state. This indicates that the electronic excitation may have a significant influence on the intensity of the hydrogen bond C₁=O₁⋯H₁–O₃.

Vibrational absorption spectra

Based on the optimized excited-state geometry, all of the IR spectra of the ground state and the S₁ state for the isolated

**Fig. 2** Frontier molecular orbitals (MOs) of the hydrogen-bonded complex C153–DAF⁺

C153 as well as the hydrogen-bonded complex C153–DAF⁺ were calculated using the DFT and the TDDFT methods, respectively, and the ground-state IR spectrum of DAF⁺ was also calculated for comparison.

The calculated IR spectra for both the isolated C153 and the hydrogen-bonded complex C153–DAF⁺ in different electronic states over the spectra range 1000–2000 cm⁻¹ are presented in Fig. 3, and the spectral regions of the C=O stretching band are indicated with red arrows. Electronic excitation from the ground state to the S₁ state of the isolated C153 induces a large redshift (of 251 cm⁻¹) in the stretching vibrational mode of the C₁=O₁ group from 1797 cm⁻¹ to 1546 cm⁻¹, while the stretching vibrational mode of the C₁=O₁ group in the ground state is redshifted by only 80 cm⁻¹ from 1797 cm⁻¹ to 1717 cm⁻¹ because of the hydrogen-bonding interactions. So, we can conclude that both the formation of the hydrogen bond C₁=O₁⋯H₁–O₃ and electronic excitation can cause the stretching vibrational mode of the C₁=O₁ group to redshift, whereas electronic excitation can produce a relatively large redshift in the stretching vibrational mode of the C₁=O₁ group.

The calculated IR spectra for the hydrogen-bonded complex C153–DAF⁺ in different electronic states over the spectral range 2000–4000 cm⁻¹ are shown in Fig. 4. Additionally, the O–H stretching band of the cation DAF⁺ is presented. The stretching vibrational frequencies of the O₃–H₁ group are depicted in Fig. 4. Upon observing the stretching vibrational mode of the O₃–H₁ group, it is clear that the stretching vibrational mode of the O₃–H₁ group in the ground state is significantly redshifted (by 472 cm⁻¹) from 3812 cm⁻¹ to 3340 cm⁻¹ owing to the formation of the hydrogen bond C₁=O₁⋯H₁–O₃. This suggests that the stretching vibrational mode of the O₃–H₁ group undergoes a larger shift than that of the C₁=O₁ group upon the formation of the hydrogen bond C₁=O₁⋯H₁–O₃ in the ground state. As a result, the O₃–H₁ group is more sensitive

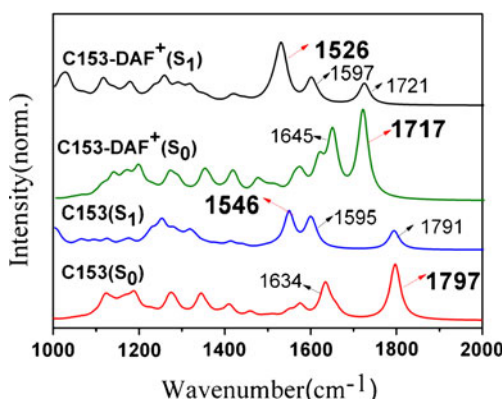


Fig. 3 Calculated IR spectra of the isolated C153 and the hydrogen-bonded complex C153–DAF⁺ in different electronic states across the spectral range of the C=O stretching absorption band

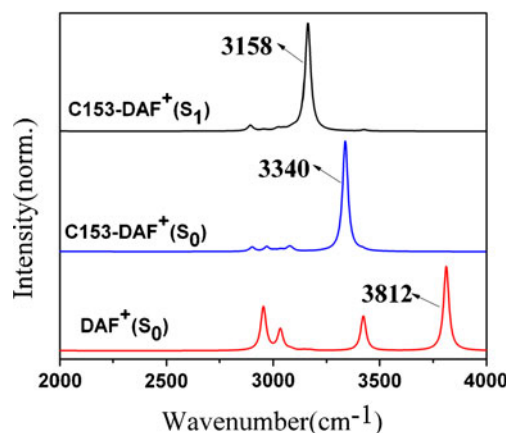


Fig. 4 Calculated O–H stretching bands of the isolated cation DAF⁺ and the hydrogen-bonded complex C153–DAF⁺ in different electronic states

to hydrogen-bonding interactions. As we discussed earlier, the DAF⁺ moiety remains in its ground state when the hydrogen-bonded complex C153–DAF⁺ is photoexcited to the S₁ state. These analyses indicate that the stretching vibrational mode of the O₃–H₁ group is an excellent mode to monitor the hydrogen-bonding dynamics of the hydrogen-bonded complex C153–DAF⁺.

Comparing Fig. 3 and Fig. 4, when the hydrogen-bonded complex C153–DAF⁺ is photoexcited to the S₁ state from the electronic ground state, a clear redshift in the stretching vibrational mode of the C₁=O₁ group occurs, as it changes from 1717 cm⁻¹ to 1526 cm⁻¹. This suggests that the electron density of the C₁=O₁ group of the hydrogen-bonded complex C153–DAF⁺ changes significantly upon electronic excitation, which is in accordance with the results we obtained in the molecular orbital (MO) analysis. At the same time, the stretching vibrational mode of the O₃–H₁ group also changes significantly, from 3340 cm⁻¹ to 3158 cm⁻¹.

Excited-state hydrogen-bond strengthening

As mentioned before, when the hydrogen-bonded complex C153–DAF⁺ is electronically excited to the S₁ state, the moiety of the isolated cation DAF⁺ in DAF remains in the ground state. Therefore, we calculated the binding energy of the hydrogen bond in the excited state by subtracting the energy of the isolated C153 in the S₁ state and the energy of the isolated cation in the ground state from the energy of the hydrogen-bonded complex C153–DAF⁺ in the S₁ state. The calculated binding energies of the hydrogen bonds as well as the corresponding hydrogen bond lengths in the ground and S₁ electronic states are shown in Table 3. Moreover, the hydrogen bond energies include BSSE corrections performed using the counterpoise method,

Table 3 Calculated bond lengths (Å), angles (°) and dihedral angles (°) for the hydrogen bond and the bonds very near to the hydrogen bond in different electronic states for isolated monomers and the hydrogen-bonded complex C153–DAF⁺. Calculated hydrogen-bond binding energies E_b (kJ mol⁻¹), BSSEs (kJ mol⁻¹), and BSSE-corrected hydrogen bond binding energies E_b^{BSSE} (kJ mol⁻¹) in different electronic states are also listed

Parameter		C153–DAF ⁺	
		S ₀	S ₁
Bond length (Å)	C ₁ =O ₁	1.227	1.242
	O ₁ ⋯H ₁	1.693	1.633
	O ₃ –H ₁	0.987	0.997
	C ₁ –O ₂	1.366	1.405
	C ₁ –C ₂	1.433	1.417
	C ₂ –H ₂	1.079	1.081
	C ₃ –O ₃	1.396	1.393
	C ₃ –H ₃	1.095	1.096
	C ₃ –H ₄	1.106	1.107
Bond or dihedral angle (°)	C ₁ =O ₁ ⋯H ₁	128.9	123.2
	O ₁ ⋯H ₁ –O ₃	169.4	172.4
	C ₁ =O ₁ ⋯H ₁ –O ₃	10.38	29.28
E_b		77.19	98.46
BSSE		-7.268	-8.292
E_b^{BSSE}		69.92	90.17

$$E_b^{\text{BSSE}} = E_b + \text{BSSE}$$

although the BSSE corrections did not affect the binding energies notably. It is clear that the hydrogen bond C₁=O₁⋯H₁–O₃ is strengthened in the excited state, since the BSSE-corrected binding energy for this hydrogen bond is greatly increased by 20.25 kJ mol⁻¹ from 69.92 kJ mol⁻¹ in the ground state to 90.17 kJ mol⁻¹ in the S₁ state. It is worth noting that the binding energies of the hydrogen bond in this novel system are quite large compared to those obtained with conventional solvents; binding energies for a hydrogen bond between a fluorescent probe and conventional solvent generally fall within the range 8–50 kJ mol⁻¹ [4–6, 36–39]. Lee et al. [84] have carried out some theoretical studies on a similar system. They found that cationic and anionic dimers with short and strong hydrogen bonds (SSHBs) have larger hydrogen-bond binding energies (by almost 100 kJ mol⁻¹) than neutral hydrogen-bonded dimers, and this leads to the polarization of proton-donating H atoms on proton-accepting O/N atoms [84].

When the dimer is photoexcited, the calculated length of the C₁=O₁ bond increases from 1.227 Å to 1.242 Å, while the O₃–H₁ bond length increases to 0.997 Å from 0.987 Å. Moreover, the length of the hydrogen bond O₁⋯H₁ is correspondingly shortened by 0.06 Å from 1.693 Å to 1.633 Å. Thus, the conclusion that the hydrogen bond is strengthened in the photoexcited state, as seen for conven-

tional solvents [4–10], can also be drawn for C153 hydrogen bonded to the cation DAF⁺.

To show the influence of electronic excitation on the whole hydrogen-bonded complex C153–DAF⁺, the changes in the lengths of the bonds adjacent to the hydrogen bond C₁=O₁⋯H₁–O₃ induced by electronic excitation are listed in Table 3. The C₁–O₂ bond is stretched, the C₁–C₂ bond is shortened, but the lengths of the C₂–H₂, O₃–C₃, C₃–H₃, and C₃–H₄ bonds are almost unchanged. The results basically correlate with the discussion of the LE characteristics of the S₁ state. The bond angles of C₁=O₁⋯H₁ and O₁⋯H₁–O₃ in the electronically excited state are practically the same as those in the ground state (they change by only -5.7° and 3.0° from 128.9° and 169.6°, respectively). On the other hand, the dihedral angle C₁=O₁⋯H₁–O₃ is significantly increased by 18.9° from 10.38° in the ground state to 29.28° in the S₁ state.

Conclusions

In this work, the electronic excited-state hydrogen-bonding dynamics of the chromophore C153 in the room-temperature ionic liquid DAF was studied using time-dependent density functional theory (TDDFT). We investigated the hydrogen bond C₁=O₁⋯H₁–O₃ that forms between isolated C153 and the cation DAF⁺ in DAF. Based on the geometric structures and the energies of the hydrogen-bonded complex C153–DAF⁺ for both the ground state and electronic excited state, it is clear that a short and strong hydrogen bond forms between C153 and the cation DAF⁺ in DAF. Also, the hydrogen bond C₁=O₁⋯H₁–O₃ decreases from 1.693 Å to 1.633 Å, and the corresponding binding energy corrected for BSSE increases from 69.92 kJ mol⁻¹ to 90.17 kJ mol⁻¹, upon photoexcitation. Using a molecular orbital analysis, we demonstrated that the S₁ state of the hydrogen-bonded complex C153–DAF⁺ has the characteristics of a locally excited (LE) state, and the electron density is concentrated on the C153 moiety. The calculated maximum absorption peak of the hydrogen-bonded complex C153–DAF⁺ coincides with the experimental results. In addition, we calculated the IR spectra of the ground state and the S₁ state of the hydrogen-bonded complex C153–DAF⁺ as well as C153, and the ground state of the cation DAF⁺ was also calculated. The vibrational absorption frequencies of the C=O group and the O–H group associated with the formation of the C₁=O₁⋯H₁–O₃ hydrogen bond both redshift due to photoexcitation; in other words, the hydrogen bond is strengthened when moving from the ground state to the S₁ state. Our calculations also revealed that the hydrogen bond C₁=O₁⋯H₁–O₃ linking the oxygen atom of the C=O group in C153 and the hydrogen atom of

the O–H group in cation DAF⁺ is stronger than those formed in conventional solvents; the polarization of the cation in DAF is responsible for the strengthening of the hydrogen bond C₁=O₁⋯H₁–O₃. More research effort should be directed into studying the hydrogen bonds formed in ionic liquids.

Acknowledgments This work was supported by the National Natural Science Foundation of China (grant nos. 21036006) and the Key Laboratory of Industrial Ecology and Environmental Engineering, China Ministry of Education.

References

- Han KL, Zhao GJ (2010) Hydrogen bonding and transfer in the excited state. Wiley, Chichester. doi:10.1002/9780470669143. ISBN 978-0-470-66677-7
- Banno M, Ohta K, Tominaga K (2008) Ultrafast vibrational dynamics and solvation complexes of methyl acetate in methanol studied by sub-picosecond infrared spectroscopy. *J Raman Spectrosc* 39:1531–1537
- Liu S, Kokot S, Will G (2009) Photochemistry and chemometrics—an overview. *J Photochem Photobiol C* 10:159–172
- Zhao GJ, Han KL (2007) Early time hydrogen-bonding dynamics of photoexcited coumarin 102 in hydrogen-donating solvents: theoretical study. *J Phys Chem A* 111:2469–2474
- Zhao GJ, Han KL (2007) Novel infrared spectra for intermolecular dihydrogen bonding of the phenol-borane-trimethylamine complex in electronically excited state. *J Chem Phys* 127:024306
- Zhao GJ, Han KL (2008) Time-dependent density functional theory study on hydrogen-bonded intramolecular charge-transfer excited state of 4-dimethylamino-benzonitrile in methanol. *J Comput Chem* 29:2010–2017
- Zhao GJ, Han KL (2008) Effects of hydrogen bonding on tuning photochemistry: concerted hydrogen-bond strengthening and weakening. *Chem Phys Chem* 9:1842–1846
- Zhao GJ, Han KL (2008) Site-specific solvation of the photoexcited protochlorophyllide a in methanol: formation of the hydrogen-bonded intermediate state induced by hydrogen-bond strengthening. *Biophys J* 94:38–46
- Zhao GJ, Chen RK, Sun MT et al (2008) Photoinduced intramolecular charge transfer and S₂ fluorescence in thiophene- π -conjugated donor-acceptor systems: experimental and TDDFT studies. *Chem Eur J* 14:6935–6947
- Zhao GJ, Han KL, Stang PJ (2009) Theoretical insights into hydrogen bonding and its influence on the structural and spectral properties of aquo palladium(II) complexes: cis-[(dppp)Pd(H₂O)₂]₂⁺, cis-[(dppp)Pd(H₂O)(OSO₂CF₃)]₂⁺(OSO₂CF₃)⁻, and cis-[(dppp)Pd(H₂O)₂](OSO₂CF₃)₂⁻. *J Chem Theory Comput* 5:1955–1958
- Duan LL, Fischer A, Xu YH, Sun LC (2009) Isolated seven-coordinate Ru(IV) dimer complex with [HOHOH]⁻ bridging ligand as an intermediate for catalytic water oxidation. *J Am Chem Soc* 131:10397–10399
- Nyhlén J, Duan LL, Åkermark B, Sun LC, Privalov T (2010) Evolution of O₂ in a seven-coordinate Ru(IV) dimer complex with a [HOHOH]⁻ bridge: a computational study. *Angew Chem Int Ed* 49:1773–1777
- Zhao GJ, Northrop BH, Stang PJ, Han KL (2010) Photophysical properties of coordination-driven self-assembled metallosupramolecular rhomboids: experimental and theoretical investigations. *J Phys Chem A* 114:3418–3422
- Nagasawa Y, Yartsev AP, Tominaga K, Johnson AE, Yoshihara K (1994) Temperature dependence of ultrafast intermolecular electron transfer faster than solvation process. *J Chem Phys* 101:5717–5724
- Woutersen S, Emmerichs U, Bakker HJ (1997) Femtosecond mid-IR pump-probe spectroscopy of liquid water: evidence for a two-component structure. *Science* 278:658–660
- Hamm P, Lim M, Hochstrasser RM (1998) Non-Markovian dynamics of the vibrations of ions in water from femtosecond infrared three-pulse photon echoes. *Phys Rev Lett* 81:5326–5329
- Sessler JL, Sathianathan M, Brown CT et al (2001) Hydrogen-bond-mediated photoinduced electron-transfer: novel dimethylaniline—anthracene ensembles formed via Watson–Crick base-pairing. *J Am Chem Soc* 123:3655–3660
- Zhao GJ, Han KL (2010) pH-controlled twisted intramolecular charge transfer (TICT) excited state via changing the charge transfer direction. *Phys Chem Chem Phys* 12:8914–8918
- Chan WS, Ma CS, Kwok WM, Phillips DL (2005) Time-resolved resonance Raman and density functional theory study of hydrogen-bonding effects on the triplet state of *p*-methoxyacetophenone. *J Phys Chem A* 109:3454–3469
- Benniston AC, Harriman A (2006) Charge on the move: how electron-transfer dynamics depend on molecular conformation. *Chem Soc Rev* 35:169–179
- Zhao GJ, Han KL (2009) Excited-state electronic structures and photochemistry of heterocyclic annulated perylenes (HAPs) materials tuned by heteroatoms: S, Se, N, O, C, Si, and B. *J Phys Chem A* 113:4788–4794
- Tsumura K, Furuya K, Sakamoto A, Tasumi M (2008) Vibrational analysis of *trans*-stilbene in the excited singlet state by time-dependent density functional theory: calculations of the Raman, infrared, and fluorescence excitation spectra. *J Raman Spectrosc* 39:1584–1591
- Sandanyaka ASD, Sasabe H, Takata T, Ito O (2010) Photoinduced electron transfer processes of fullerene rotaxanes containing various electron-donors. *J Photochem Photobiol C* 11:73–92
- Mallick A, Das P, Chattopadhyay N (2010) Photophysics of norharmane in solution phase: from homogeneous to micro-heterogeneous environments. *J Photochem Photobiol C* 11:62–72
- Kearley GJ, Fillaux F, Baron MH, Benington S, Tomkinson J (1994) A new look at proton transfer dynamics along the hydrogen bonds in amides and peptides. *Science* 264:1285–1289
- Zhao GJ, Northrop BH, Han KL, Stang PJ (2010) The effect of intermolecular hydrogen bonding on the fluorescence of a bimetallic platinum complex. *J Phys Chem A* 114:9007–9013
- Shynkar VV, Klymchenko AS, Piémont E, Demchenko AP, Mély Y (2004) Dynamics of intermolecular hydrogen bonds in the excited states of 4'-dialkylamino-3-hydroxyflavones. On the pathway to an ideal fluorescent hydrogen bonding sensor. *J Phys Chem A* 108:8151–8159
- Shirota H, Ushiyama H (2008) Hydrogen-bonding dynamics in aqueous solutions of amides and acids: monomer, dimer, trimer, and polymer. *J Phys Chem B* 112:13542–13551
- Yun C, You J, Kim J, Huh J, Kim E (2009) Photochromic fluorescence switching from diarylethenes and its applications. *J Photochem Photobiol C* 10:111–129
- Priyadarsini KI (2009) Photophysics, photochemistry and photobiology of curcumin: studies from organic solutions, bio-mimetics and living cells. *J Photochem Photobiol C* 10:81–95
- Mazur K, Heisler IA, Meech SR (2010) Ultrafast dynamics and hydrogen-bond structure in aqueous solutions of model peptides. *J Phys Chem B* 114:10684–10691
- Zhao GJ, Liu JY, Zhou LC, Han KL (2007) Site-selective photoinduced electron transfer from alcoholic solvents to the chromophore facilitated by hydrogen bonding: a new fluorescence quenching mechanism. *J Phys Chem B* 111:8940–8945

33. Karmakar R, Samanta A (2002) Steady-state and time-resolved fluorescence behavior of C153 and PRODAN in room-temperature ionic liquids. *J Phys Chem A* 106:6670–6675
34. Zhao GJ, Han KL (2009) Role of intramolecular and intermolecular hydrogen bonding in both singlet and triplet excited states of aminofluorenones on internal conversion, intersystem crossing, and twisted intramolecular charge transfer. *J Phys Chem A* 113:14329–14335
35. Zhao GJ, Han KL (2007) Ultrafast hydrogen bond strengthening of the photoexcited fluorenone in alcohols for facilitating the fluorescence quenching. *J Phys Chem A* 111:9218–9223
36. Liu YF, Ding JX, Shi DH, Sun JF (2008) Time-dependent density functional theory study on electronically excited states of coumarin 102 chromophore in aniline solvent: reconsideration of the electronic excited-state hydrogen-bonding dynamics. *J Phys Chem A* 112:6244–6248
37. Liu YF, Ding JX, Liu RQ, Shi DH, Sun JF (2009) Revisiting the electronic excited-state hydrogen bonding dynamics of coumarin chromophore in alcohols: undoubtedly strengthened not cleaved. *J Photochem Photobiol A* 201:203–207
38. Wei NN, Li P, Hao C, Wang R, Xiu ZL, Chen JW, Song P (2010) Time-dependent density functional theory study of the excited-state dihydrogen bond O–H···H–Si. *J Photochem Photobiol A* 210:77–81
39. Han KL, He GZ, Lou NQ (1996) Effect of location of energy barrier on the product alignment of reaction A+BC. *J Chem Phys* 105:8699–8704
40. Chu TS, Zhang Y, Han KL (2006) The time-dependent quantum wave packet approach to the electronically nonadiabatic processes in chemical reactions. *Int Rev Phys Chem* 25:201–205
41. Zhou LC, Zhao GJ, Liu JF, Han KL, Wu YK, Peng XJ, Sun MT (2007) The charge transfer mechanism and spectral properties of a near-infrared heptamethine cyanine dye in alcoholic and aprotic solvents. *J Photochem Photobiol A* 187:305–310
42. Chen RK, Zhao GJ, Yang XC et al (2008) Photoinduced intramolecular charge-transfer state in thiophene- π -conjugated donor-acceptor molecules. *J Mol Struct* 876:102–109
43. Chen TY, Zhang WP, Wang XQ, Zhao GJ (2009) Theoretical insight into stereodynamics of the O(¹D)+H₂ ($v=0-3, j=0$) \rightarrow OH+H reaction: a quasiclassical trajectory (QCT) study. *Chem Phys* 365:158–163
44. Liu YH, Zhao GJ, Li GY, Han KL (2010) Fluorescence quenching phenomena facilitated by excited-state hydrogen bond strengthening for fluorenone derivatives in alcohols. *J Photochem Photobiol A* 209:181–185
45. Seth D, Sarkar S, Sarkar N (2008) Solvent and rotational relaxation of coumarin 153 in a protic ionic liquid dimethylethanolammonium formate. *J Phys Chem B* 112:2629–2636
46. Paul A, Samanta A (2007) Solute rotation and solvation dynamics in an alcohol-functionalized room temperature ionic liquid. *J Phys Chem B* 111:4724–4731
47. Samanta A (2006) Dynamic Stokes shift and excitation wavelength dependent fluorescence of dipolar molecules in room temperature ionic liquids. *J Phys Chem B* 110:13704–13716
48. Endres F, Abedin SZE (2006) Air and water stable ionic liquids in physical chemistry. *Phys Chem Chem Phys* 8:2101–2116
49. Chiappe C, Pieraccini D (2005) Ionic liquids: solvent properties and organic reactivity. *J Phys Org Chem* 18:275–297
50. Welton T (1999) Room-temperature ionic liquids. solvents for synthesis and catalysis. *Chem Rev* 99:2071–2083
51. Jin H, Baker GA, Arzhantev S, Dong J, Maroncelli M (2007) Solvation and rotational dynamics of coumarin 153 in ionic liquids: comparisons to conventional solvents. *J Phys Chem B* 111:7291–7302
52. Aki SNVK, Brennecke JF, Samanta A (2001) How polar are room-temperature ionic liquids? *Chem Commun* 5:413–414
53. Karmakar R, Samanta A (2002) Solvation dynamics of coumarin-153 in a room-temperature ionic liquid. *J Phys Chem A* 106:4447–4452
54. Karmakar R, Samanta A (2003) Intramolecular excimer formation kinetics in room temperature ionic liquids. *Chem Phys Lett* 376:638–645
55. Karmakar R, Samanta A (2003) Dynamics of solvation of the fluorescent state of some electron donor-acceptor molecules in room-temperature ionic liquids, [BMIM][CF₃SO₂]₂N] and [EMIM][CF₃SO₂]₂N]. *J Phys Chem A* 107:7340–7346
56. Saha S, Mandal PK, Samanta A (2004) Solvation dynamics of Nile Red in a room temperature ionic liquid using streak camera. *Phys Chem Chem Phys* 6:3106–3110
57. Arzhantsev S, Ito N, Heitz M, Maroncelli M (2003) Solvation dynamics of coumarin 153 in several classes of ionic liquids: cation dependence of the ultrafast component. *Chem Phys Lett* 381:278–286
58. Li GY, Zhao GJ, Han KL, He GZ (2011) A TD-DFT study on the cyanide-chemosensing mechanism of 8-formyl-7-hydroxycoumarin. *J Comput Chem* 32:668–674
59. Lang B, Angulo G, Vauthey E (2006) Ultrafast solvation dynamics of coumarin 153 in imidazolium-based ionic liquids. *J Phys Chem A* 110:7028–7034
60. Chowdhury PK, Halder M, Sanders L et al (2004) Dynamic solvation in room-temperature ionic liquids. *J Phys Chem B* 108:10245–10255
61. Shim Y, Duan J, Choi MY, Kim HJ (2003) Solvation in molecular ionic liquids. *J Chem Phys* 119:6411–6414
62. Kobrak MN, Znamenskiy V (2004) Solvation dynamics of room-temperature ionic liquids: evidence for collective solvent motion on sub-picosecond timescales. *Chem Phys Lett* 395:127–132
63. Shim Y, Choi MY, Kim HJ (2005) A molecular dynamics computer simulation study of room-temperature ionic liquids. II. Equilibrium and nonequilibrium solvation dynamics. *J Chem Phys* 122:044511
64. Ingram JA, Moog RS, Ito N, Biswas R, Maroncelli M (2003) Solute rotation and solvation dynamics in a room-temperature ionic liquid. *J Phys Chem B* 107:5926–5932
65. Ito N, Arzhantsev S, Heitz M, Maroncelli M (2004) Solvation dynamics and rotation of coumarin 153 in alkylphosphonium ionic liquids. *J Phys Chem B* 108:5771–5777
66. Zhao GJ, Liu YH, Han KL, Dou YS (2008) Dynamic simulation study on ultrafast excited-state torsional dynamics of 9,9'-bianthryl (BA) in gas phase: real-time observation of novel oscillation behavior with the torsional coordinate. *Chem Phys Lett* 453:29–34
67. Ravi M, Soujanya T, Samanta A, Radhakrishnan TP (1995) Excited-state dipole moments of some coumarin dyes from a solvatochromic method using the solvent polarity parameter, E_T^N . *J Chem Soc Faraday Trans* 91:2739–2742
68. Dhupal NR, Kim HJ, Kiefer J (2009) Molecular interactions in 1-ethyl-3-methylimidazolium acetate ion pair: a density functional study. *J Phys Chem A* 113:10397–10404
69. Zhou LC, Liu JY, Zhao GJ, Shi Y, Peng XJ, Han KL (2007) The ultrafast dynamics of near-infrared heptamethine cyanine dye in alcoholic and aprotic solvents. *Chem Phys* 333:179–185
70. Han KL, He GZ (2007) Photochemistry of aryl halides: photodissociation dynamics. *J Photochem Photobiol C* 8:55–66
71. Chu TS, Han KL (2008) Effect of Coriolis coupling in chemical reaction dynamics. *Phys Chem Chem Phys* 10:2431–2441
72. Chai S, Zhao GJ, Song P, Yang SQ, Liu JY, Han KL (2009) Reconsideration of the excited-state double proton transfer (ESDPT) in 2-aminopyridine/acid systems: role of the intermolecular hydrogen bonding in excited states. *Phys Chem Chem Phys* 11:4385–4390

73. Li GY, Zhao GJ, Liu YH, Han KL, He GZ (2010) TD-DFT study on the sensing mechanism of a fluorescent chemosensor for fluoride: excited-state proton transfer. *J Comput Chem* 31:1759–1796
74. Deloncle R, Caminade AM (2010) Stimuli-responsive dendritic structures: the case of light-driven azobenzene-containing dendrimers and dendrons. *J Photochem Photobiol C* 11:25–45
75. Ramaswamy S, Rajaram RK, Ramakrishnan V (2003) Vibrational spectroscopic studies of L-argininium dinitrate. *J Raman Spectrosc* 34:50–56
76. Max JJ, Chapados C (2004) Infrared spectroscopy of acetone-water liquid mixtures. II. Molecular model. *J Chem Phys* 120:6625–6641
77. Molotsky T, Huppert D (2003) Site specific solvation statics and dynamics of coumarin dyes in hexane—methanol mixture. *J Phys Chem A* 107:2769–2780
78. Becke AD (1993) Density-functional thermochemistry. The role of exact exchange. *J Chem Phys* 98:5648–5652
79. Schäfer A, Huber C, Ahlrichs R (1994) Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J Chem Phys* 100:5829–5835
80. Zhao GJ, Han KL, Lei YB, Dou YS (2007) Ultrafast excited-state dynamics of tetraphenyl-ethylene studied by semiclassical simulation. *J Chem Phys* 127:094307
81. Ahlrichs R, Bär M, Haser M, Horn H, Kölmel C (1989) Electronic structure calculations on workstation computers: the program system turbomole. *Chem Phys Lett* 162:165–169
82. Halls JJM, Cornil J, Dos Santos DA et al (1999) Charge- and energy-transfer processes at polymer/polymer interfaces: a joint experimental and theoretical study. *Phys Rev B* 60:5721–5727
83. Lemaire V, Steel M, Beljonne D, Brédas JL, Cornil J (2005) Photoinduced charge generation and recombination dynamics in model donor/acceptor pairs for organic solar cell applications: a full quantum-chemical treatment. *J Am Chem Soc* 127:6077–6086
84. Lee HM, Kumar A, Kolaski M et al (2010) Comparison of cationic, anionic and neutral hydrogen bonded dimers. *Phys Chem Chem Phys* 12:6278–6287

Insights into the permeability of drugs and drug-like molecules from MI-QSAR and HQSAR studies

Ranajit N. Shinde · K. Srikanth · M. Elizabeth Sobhia

Received: 31 August 2010 / Accepted: 6 May 2011 / Published online: 3 June 2011
© Springer-Verlag 2011

Abstract Membrane-interaction QSAR (MI-QSAR) and Holographic QSAR (HQSAR) analyses have been performed on a diverse set of drugs and drug-like molecules. MI-QSAR combines a set of membrane-solute interaction properties calculated during molecular dynamics simulation with the set of classical solute descriptors to predict the biological behavior of drugs and drug-like molecules. HQSAR is a technique which employs fragment fingerprints or molecular holograms as predictive variables of biological activity. A data set of 60 structurally diverse molecules with permeability coefficients were used to construct significant MI-QSAR and HQSAR models of Caco-2 cell permeation. A statistically meaningful MI-QSAR model was obtained with $r^2=0.805$ and $q^2=0.696$. Subsequently, HQSAR models were developed on the same data set. The best HQSAR model ($r^2=0.915$, $q^2=0.539$) was obtained with fragment distinctions atom, bond, donor and acceptor with atom count 4 to 7. The predictions for training and test set molecules are in good agreement with experimental results and show the potential of models for untested compounds. This displays the importance of MI-QSAR and HQSAR analysis in estimating ADME properties characterized by the transport of solutes through biological membranes.

Keywords Descriptors · Fragment · HQSAR · Membrane · MI-QSAR · Molecular Dynamics Simulation · Solubility

Introduction

The discovery and development of a new drug product consists of multiple steps including discovering an active pharmaceutical ingredient (API), preclinical ADME/T testing, designing an optimum formulation for the API, clinical trials, etc. Discovery of APIs has been facilitated by the high-throughput methods, thereby leading to a large number of new chemical entities (NCE). The traditional trial and error approach of drug design is too costly and time consuming to meet the increasing demand for new drugs. It has been estimated that roughly 10% of the compounds that enter development eventually become marketed drugs and 40% of compounds fail due to poor pharmacokinetic properties. The ability to predict the parameters such as solubility, permeability and partition co-efficient certainly plays an important role in optimizing the drug-like molecules. Properties derived from molecular structure would have an impact on the drug discovery both in cost and time needed to bring a new compound to market. Informatics tools can be helpful to achieve the goal of reducing time and money needed in drug discovery. These tools help to streamline the drug discovery by rejecting the dead end leads in the earlier stages of development.

One of the bottlenecks of modern drug discovery and development is to characterize absorption, distribution, metabolism and excretion (ADME) properties of hits coming from combinatorial and natural product libraries or from rational synthesis in the early stage of drug discovery process [1]. Current in vitro and in vivo practices have concentrated on characterization of ADME properties of late stage molecules and the protocols are not always suitable for early stage molecules. The methods used for this require a large number of samples and high-

R. N. Shinde · K. Srikanth · M. E. Sobhia (✉)
Centre for Pharmacoinformatics, National Institute
of Pharmaceutical Education and Research (NIPER),
Sector 67,
S.A.S. Nagar, Punjab 160062, India
e-mail: mesophia@niper.ac.in

level of validation and quantification leading to high cost for testing. Also, *in vivo* activity screening and toxicological studies are not integrated to support ADME characterization to save time, resources and laboratory animals. Clearly, modified and tailored approaches are needed at different stages of drug discovery and development. The use of computational models in the prediction of ADME properties of compounds is growing rapidly in drug discovery as the benefits they provide in high throughput screening and early application in drug design are being realized [2].

Intestinal absorption is one of the most important ADME parameters for a molecule that is designed for the oral therapy. It is defined as the process of transfer of the molecule from apical side to basolateral side of enterocyte. Cellular membrane, described as phospholipid bilayer, plays a significant role in the absorption [3]. It makes hydrophilic and hydrophobic interactions with molecules and thus governs their transport across membrane. Most drugs are absorbed via transcellular route through passive diffusion; such are more lipophilic. In contrast, small, hydrophilic molecules and peptides pass through water filled pores that are formed by fusion of adjacent cells (paracellular route) [4]. Caco-2 cell lines are one of the most studied cells for their ability to predict the intestinal absorption. These cells are derived from human colorectal carcinoma and possess structural and functional similarities with enterocytes [5]. Many research groups have efficiently used Caco-2 cell monolayers to predict drug transport by different pathways across the intestinal epithelium. In the majority of studies the best correlation with the *in vivo* absorbed fraction is obtained for passively transported drugs. In case of other transport mechanisms variable results are obtained and in such cases it is advised to use Caco-2 cell permeability cautiously [6]. High predictivity of the intestinal absorption by Caco-2 cell lines prompted us to use Caco-2 permeability in the development of the QSAR models. Additionally computational Caco-2 permeability prediction models are another source that provides an inexpensive and fast way to assess the potential for intestinal permeability of a molecule which enables prioritization of molecules for *in vitro* and *in vivo* studies before their synthesis.

Computational Caco-2 permeability prediction models are another source that provides an inexpensive and fast way to assess the potential for intestinal permeability of a molecule. The majority of the computational studies carried out correlate Caco-2 permeability with the physicochemical properties of drug molecules. *In silico* prediction of oral absorption started taking shape with Lipsinki's 'rule-of-five' where permeability shown to be dependent on hydrogen bond acceptors, hydrogen bond donors, molecular weight and logP [7]. Afterward different simple

descriptors like molecular size, hydrophobicity, rotatable bonds, dynamic polar surface area, charge etc. [4, 8–11] and complex descriptors like quadratic indices of the molecular pseudograph's atom adjacency matrix [12], interaction descriptors derived from simulation of molecule transport across cell membrane [13], molecular orbital calculation [14], MolSurf-derived descriptors [15], Volsurf derived descriptors [16], etc. were used to predict the permeability. Here we used two approaches. MI-QSAR simulates the transfer of molecules through the membrane while HQSAR provides information about the importance of different structural fragments in the permeability. Kulkarni et al. has carried out membrane interaction QSAR on a set of 38 molecules [17]. It was observed that permeability is governed by factors viz. aqueous solubility, size and shape of molecule, intramolecular hydrogen bonding energy, membrane-solute interactions energy and conformational flexibility of the solute in the membrane.

Here we tried to find out important properties that govern the permeability of molecules through MI-QSAR and HQSAR methods with more molecules (60) than considered by Kulkarni et al. [17]. The dataset used cover a relatively wide range of chemical space and permeability.

Materials and methods

Data set

The apparent Caco-2 permeability coefficients for 60 structurally diverse compounds that are absorbed via transcellular and paracellular were selected from the different sources [18–22]. The majority of the data (41 compounds) were taken from a single source where experiments were performed under uniform experimental condition [23]. Briefly, Caco-2 cells were obtained from American Tissue Culture Collection, Rockville, MD and cells were seeded at a density of 80,000 cells cm^{-2} and allowed to grow and differentiate for up to 25 days. Cells of passage numbers 23 to 50 were used throughout. Prior to permeability experiments, the culture medium was replaced with the transport medium of pH 7.4 and equilibrated for 30 minutes at 37 °C. Drug solutions were prepared in HBSS at a final concentration of 0.01 to 0.1 mM. Initially apical side of the monolayers provided 1.5 ml of drug solution. The amount of solute permeated was determined by either moving the inserts to new wells containing fresh medium or taking a sample from the basolateral side and replacing it with fresh medium at discrete time intervals. Transport rates were then determined by plotting the cumulative amount permeated as a function of time.

Permeability experiments were performed in an incubator at 37 °C and an atmosphere of 5% CO₂ over the duration of two hours. Experiments were performed under sink condition where the concentration of the solute in the receiver side was less than 10% of the dose applied at all intervals of time. The permeability coefficient was then determined according to the following equation:

$$P_{\text{Caco-2}} = J(\text{ACO}) - 1 \quad (1)$$

In Eq. 1, J is the rate of appearance of solute in the receiver chamber, C_0 is the initial concentration of the solute in the donor chamber, and A is the surface area of the filter. Caco-2 cell permeability (P) of all compounds was represented in negative logarithm, i.e., $-\log P(P_{\text{Caco-2}})$. All the QSAR analyses were performed considering $P_{\text{Caco-2}}$ as dependent variable. In relation to the descriptors this implies that the positively correlated descriptors to $P_{\text{Caco-2}}$ will decrease the Caco-2 cell permeability. Table 1 contains the Caco-2 cell permeability values for 60 structurally diverse drugs and drug-like molecules used for the study. Representative molecules from the data set are shown in Chart 1.

MI-QSAR

Building molecules and phospholipid layer

The 3-dimensional structures of the solute molecules of the training set were built using the Sybyl7.1 package [24]. The phospholipid Dimyristoyl phosphatidylcholine (DMPC) was selected as the model phospholipid. It was built using available crystal structure data in HyperChem [25, 26]. The structure of a DMPC molecule is shown in Fig. 1. Building of the DMPC membrane monolayer was carried out in the MI-QSAR package installed on a Silicon Graphics Fuel Workstation [27]. Construction of the model monolayer was performed on the basis of information available in the literature [28]. An assembly of 25 DMPC molecules (5*5*1) in x , y , z directions, respectively, was used as the model membrane monolayer. The size of the monolayer simulation system was selected based on the work done by van der Ploeg and Berendsen [29].

A central DMPC molecules was removed from the equilibrated monolayer and a test solute molecule was inserted in the space created by the missing DMPC molecule. Each of the test solute molecules of the data set was inserted at three different positions in the DMPC monolayer with the most polar group of the solute molecule “facing” toward the head group region of the monolayer. Three molecular dynamics simulation (MDS) models were generated for each solute molecule for the trial positions of the solute molecule in the monolayer.

Table 1 The data set used in the QSAR analysis

S. No.	M. No.	Name	Caco-2 cell permeability (P) *10 ⁶	-Log P (P _{caco-2})
1	1	Acebutalol	0.51	6.29
2	2	Acyclovir	0.25	6.60
3	3	Alprenolol	25.30	4.60
4	4	Aminopyrine	36.50	4.44
5	5	Amoxicilline	0.80	6.10
6	6	Antipyrine	28.20	4.55
7	10	Caffeine	30.80	4.51
8	11	Cephalexin	0.50	6.30
9	12	Chloromphenicol	20.60	4.69
10	13	Chlorpromazine	19.90	4.70
11	14	Chlorthiazide	0.15	6.82
12	15	Cimetidine	0.74	6.13
13	16	Clodine	21.80	4.66
14	18	Decipramine	24.20	4.62
15	19	Desoxycorticosterone	21.20	4.67
16	20	Dexamethazone	12.20	4.91
17	21	Diazepam	33.40	4.48
18	22	Diltiazem	29.80	4.53
19	23	Dopamine	9.33	5.03
20	24	Doxorubicine	0.16	6.80
21	25	Enalapril	2.13	5.67
22	26	Erythromycine	3.73	5.43
23	27	Estradiol	16.90	4.77
24	28	Furesomide	0.12	6.92
25	29	Gancyclovir	0.38	6.42
26	30	Griseofulvin	36.60	4.44
27	31	Guanabenz	20.90	4.68
28	32	Hydrocartisone	14.00	4.85
29	33	Hydrochlorthiazide	0.51	6.29
30	34	Ibuprofen	52.50	4.28
31	35	Imipramine	14.10	4.85
32	36	Indomethacin	20.40	4.69
33	37	Labetalol	9.31	5.03
34	38	Mannitol	0.38	6.42
35	39	Meloxicam	19.50	4.71
36	40	Metaprolal	23.70	4.63
37	41	Methotrexate	1.20	5.92
38	42	Nadalol	3.88	5.41
39	43	Naproxen	39.50	4.40
40	44	Nicotine	19.40	4.71
41	45	Phenatoin	26.70	4.57
42	46	Phencyclidine	24.70	4.61
43	47	Pindolol	16.70	4.78
44	48	Piroxicam	35.60	4.45
45	49	Prazocine	43.60	4.36
46	50	Progesterone	23.70	4.63
47	51	Propronalol	21.80	4.66

Table 1 (continued)

S. No.	M. No.	Name	Caco-2 cell permeability (P) *10 ⁶	-Log P (P _{caco-2})
48	52	Quinidine	20.40	4.69
49	53	Ranitidine	0.49	6.31
50	54	Salicylic acid	22.00	4.66
51	55	Saquinavir	0.80	6.10
52	56	Scopolamine	11.80	4.93
53	57	Sucrose	1.71	5.77
54	58	Sulfa salazine	0.30	6.52
55	59	Telmisartan	15.10	4.82
56	60	Terbutaline	0.47	6.33
57	61	Testosterone	24.90	4.60
58	62	Timolol	12.80	4.89
59	63	Urea	4.56	5.34
60	64	Valproic acid	48.00	4.32

The three trial positions were,

- (1) Solute molecule in the head group region
- (2) Solute molecule between the head-group region and the aliphatic chains

- (3) Solute molecule in the tail region of the aliphatic chains.

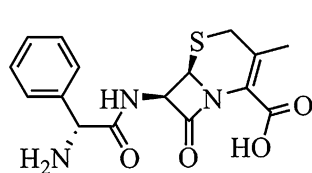
The lowest energy geometry of the solute molecule in the monolayer was sought using each of the three trial solute positions [30, 31]. Figure 2 shows the top view of acebutolol molecule docked into DMPC monolayer at the head position. The acebutolol molecule is shown at the center.

Molecular dynamic simulation (MDS)

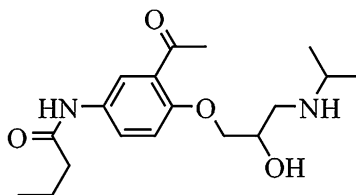
MI-QSAR uses molecular dynamics to find out the lowest energy conformation of the solute molecules. Molecular dynamics is a process which reproduces the time dependent motional behavior of a molecule. It assumes that the atoms in the molecule interact with one another according to the force field used. At regular time intervals the classical equation of motion represented by Newton's second law is solved:

$$F_i(t) = m_i a_i(t),$$

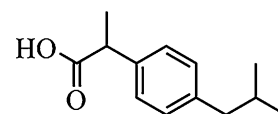
where F_i is the force on an atom i at time t , m_i is the mass of an atom i , and a_i is the acceleration of atom i at time t . The gradient of potential energy function is used to



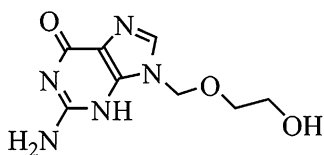
Cephalixin



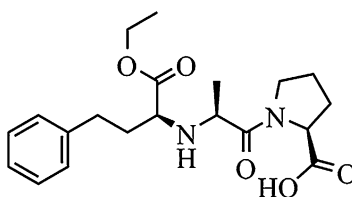
Acebutolol



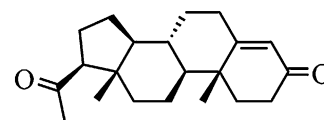
Ibuprofen



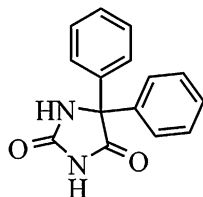
Acyclovir



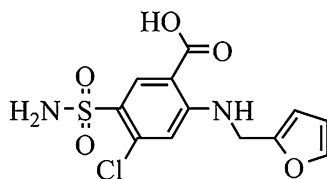
Enalapril



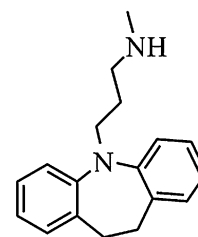
Progesterone



Phenytoin



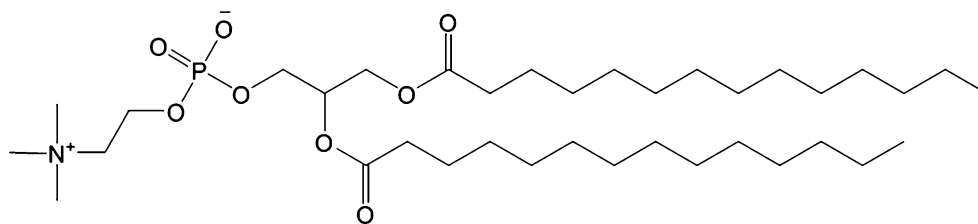
Furosemide



Desipramine

Chart 1 Representative molecules from the data set

Fig. 1 The chemical structure of the DMPC molecule



calculate the forces on the atoms while the initial velocities on the atoms are generated randomly at the beginning of the dynamics run. Based on the initial atom coordinates of the system, new positions and velocities on the atoms can be calculated at time t and atom will be moved to those new positions. As a result, a new conformation is generated. The cycle will then be repeated for a specified period of time steps. MDS calculations are used to estimate membrane-solute interaction properties and to "add" the resultant membrane-solute properties to the intramolecular physico-chemical property descriptors to provide an extended set of trial descriptors for building QSAR models. MDS was carried out using the Molsim package with an extended MM2 force field [32]. The temperature was set at 311K as the simulation temperature and held constant in the MDS by coupling the system to an external fixed-temperature bath [33]. All the remaining parameters were kept default in the study. The trajectory step size was 0.001 ps over a total simulation time of 10 ps for each test compound. Periodic boundary conditions (PBC) were applied for the DMPC monolayer model, but not for the test solute molecule. By using periodic boundary conditions it is possible to simulate an infinite system. Also by using PBC, simulations can be performed on relatively small systems in such a way that the system experiences forces in bulk fluid. The solute molecule was placed in one of the three positions and MDS was carried out. For each MDS, only one solute molecule was considered [34]. A trajectory plot of the total energy versus simulation time for acebutolol embedded in the model DMPC monolayer is shown in Fig. 3.

Calculation of descriptors, construction and testing of models

The descriptors were calculated using the MI-QSAR package. The descriptors calculated are classified into three groups.

- 1) Intramolecular solute descriptors
- 2) Intermolecular membrane solute descriptors
- 3) Dissolution and solvation solute descriptors

We calculated both the intra and intermolecular descriptors. Intermolecular solute membrane descriptors were derived from the MDS [34–37]. MI-QSAR models were constructed using the genetic function approximation

(GFA). The GFA algorithm uses a genetic algorithm to perform a search over the space of possible QSAR/QSPR models using the LOF score to estimate the fitness of each model. All intramolecular and intermolecular descriptors in the MI-QSAR trial descriptor pool were used as linear terms during the evolution of genetic function approximation to generate MI-QSAR models.

HQSAR

Molecular modeling studies were performed using the molecular modeling package SYBYL7.1 installed on a Silicon Graphics Fuel Workstation running on IRIX 6.5. The structures were sketched and minimized individually by using Powell's conjugate gradient method [38].

HQSAR is a technique which employs fragment fingerprints or molecular holograms as predictive variables of biological activity or other structurally related data [39]. Fragments with different atom counts were generated and fragments with 4 to 7 atoms were hashed into bins 1 to 85 of the fingerprint. These molecular fingerprints are broken into strings at fixed intervals as specified by a hologram length (HL) parameter. The HL determines the number of

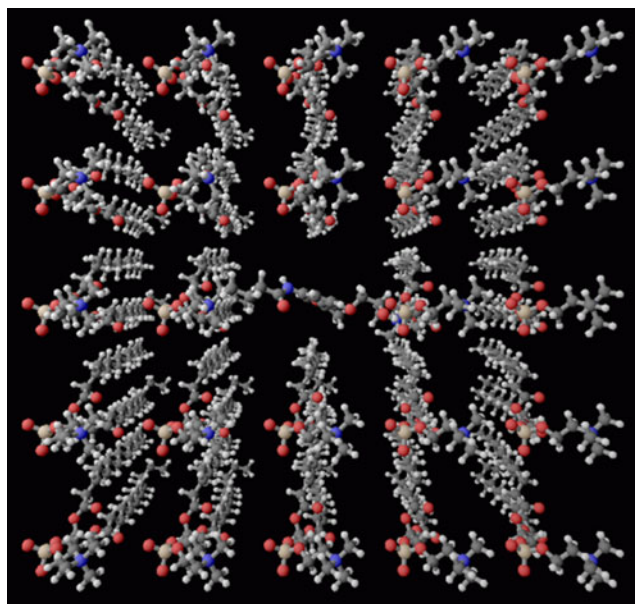
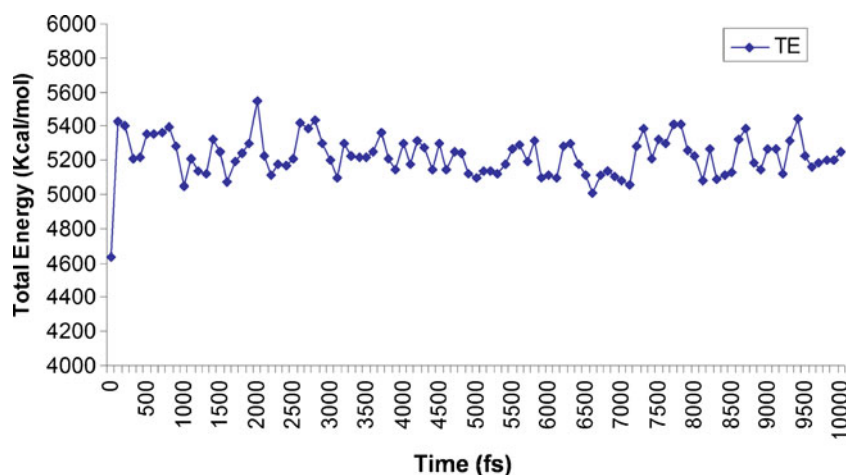


Fig. 2 The top view of acebutolol molecule docked into DMPC monolayer at head position. Acebutolol is placed at the center

Fig. 3 The molecular dynamics simulation (MDS) trajectory plot of the total energy versus step size for acebutalol embedded in the model DMPC monolayer



bins in the hologram into which the fragments are hashed. Each corresponding fragment SLN is then mapped to a pseudo-random integer in the range 0 to 231 using the CRC (cyclic redundancy check) algorithm. The integer generated by the CRC algorithm is unique and reproducible for each and every unique SLN string. The hashing then occurs by folding the pseudorandom integer for a particular SLN string into the bin range defined. In HQSAR, bins contain information about the number of fragments hashed into each bin. The optimal HQSAR model was derived from screening through the 12 default HL values, which were a set of 12 prime numbers ranging from 53 to 401. A schematic representation of the generation of molecular hologram is shown in Fig. 4.

Results and discussion

MI-QSAR

Various MI-QSAR Models for Caco-2 cell permeability were developed based on the genetic function approximation (GFA) optimization. As the correlation coefficient (r^2) changes with the number of terms in the QSAR equation, we took the cross validation correlation coefficient (q^2) as the limiting factor for a number of descriptors to be used in the model. The plot of (q^2) for 1–6 term MI-QSAR models versus the numbers of terms in the corresponding models is shown in Fig. 5. The (q^2) value increased till the number of descriptors in the equation reached up to six. When the number of descriptors in the equation increased above six, there was a decrease in (q^2) value of model. So the number of descriptors was restricted to six.

Model-1

In order to arrive at the first MI-QSAR model many models were tried. Various models were developed and the model

with $r^2=0.807$; $q^2=0.742$ (Model-1) was found to be the best and the results were statistically compatible with the literature. The equation for Model-1 is shown below and the correlation matrix for the Model-1 is shown in Table 2. Training set and test set predictions of Model-1 are shown in Tables 3 and 4 respectively.

$$P_{Caco-2} = 4.28549 - 0.048541 \times LE14 - 0.01117 \times EHBD + 0.000642 \times I_c - 0.035672 \times Ecoh + 0.017574 \times LETOR - 0.004383 \times LUMO \quad (1)$$

$$r^2=0.807; q^2=0.742; n=45$$

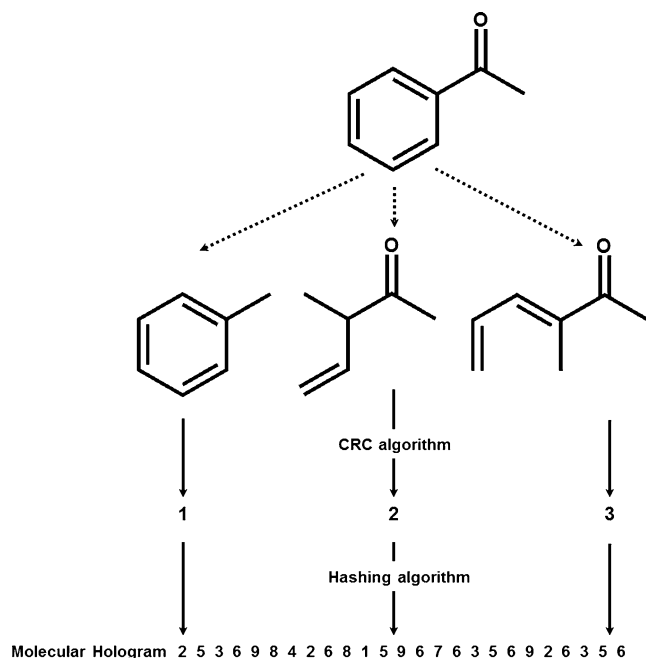
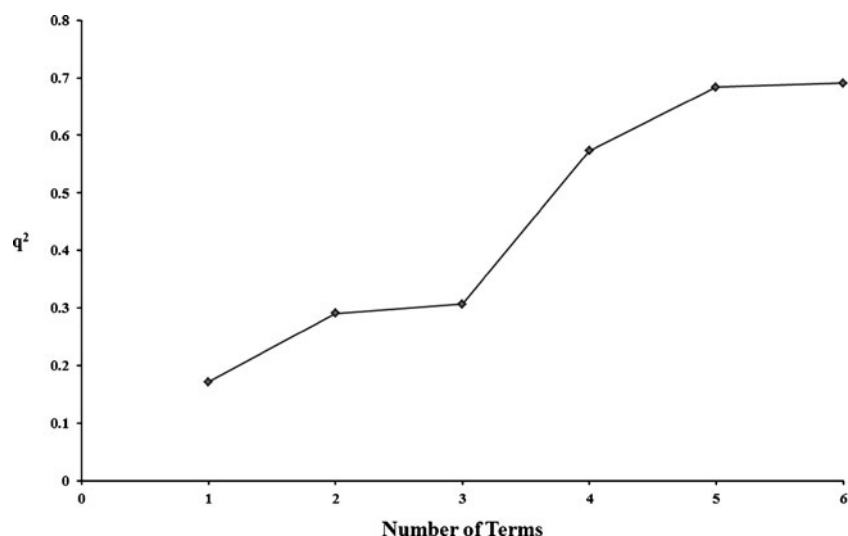


Fig. 4 Schematic representation of generation of molecular hologram

Fig. 5 Plot of cross-validation coefficient (q^2) versus number of terms in the corresponding MI-QSAR model



Where, n is the no. of molecules in training set r^2 is the correlation coefficient and q^2 is the cross validated correlation coefficient.

Descriptors I_c and *LETOR* of Model-1 showed negative correlation. This implies that the higher the solute torsional energy, *LETOR*, lower the permeability will be and vice versa. *LUMO*, *LE14* and *EHBD* are the other descriptors correlated positively to the activity. *LUMO* is the lowest occupied molecular orbital energy level. *LE14* denotes the solute 1–4 interaction energy. *EHBD* is the total complex hydrogen bonding energy. *Ecoh* and I_c are the cohesive energy descriptors, *Ecoh* is negatively correlated and I_c positively correlated. Model-1 is statistically robust but this model was not selected for further analyses due to the high inter correlation (0.70) between the descriptors *Ecoh* and *LE14* (Table 2).

Model-2

Model-1 was further refined to predict a meaningful QSAR equation. In this process different term models were obtained by considering the combination of

intramolecular, intermolecular and membrane interaction descriptors as a function of number of terms, i.e., descriptors included in a given MI-QSAR model. After many iterations of this exercise model-2 was obtained with a set of descriptors that are not inter-correlated. The results obtained for 1–6 term models are presented below. The 6-term model, i.e., model-2 was selected for final analysis. The training and test set prediction of this model are shown in Tables 3 and 4, respectively. The correlation matrix for this model is shown in Table 5.

Comparing the descriptors obtained for model 1 and 2, *LE14* is found common in both the models and other descriptors namely *Dipole*, *Chi-8*, *LEHBD*, *DEHBD* and I_c which are the new descriptors in Model-2. The calculated values for these descriptors are given in Table 6. Equation for this model is shown below.

The plot for actual P_{Caco-2} versus predicted P_{Caco-2} values of the training and test set molecules for model-2 is shown in Fig. 6. Predictivity of the test set was determined with the use of root mean squared error (RMSE) and mean error (ME). For model 1 the statistic was 0.88 (RMSE) and

Table 2 Correlation between Model-1 descriptors

	I_c	<i>Ecoh</i>	<i>LUMO</i>	<i>LETOR</i>	<i>LE14</i>	<i>EHBD</i>	P_{Caco-2}
I_c	1.00						
<i>Ecoh</i>	0.04	1.00					
<i>LUMO</i>	-0.25	-0.27	1.00				
<i>LETOR</i>	0.07	0.22	-0.02	1.00			
<i>LE14</i>	-0.12	0.70	-0.13	0.54	1.00		
<i>EHBD</i>	0.20	-0.66	-0.23	-0.11	-0.44	1.00	
P_{Caco-2}	0.21	0.54	-0.13	-0.03	-0.03	-0.63	1.00

Table 3 Actual, model-1 and model-2 predicted, $P_{\text{caco-2}}$ values for training set molecule

S. No.	M. No.	Actual	Model-1	Model-2
1	2	6.60	6.35	6.41
2	3	4.60	4.33	4.50
3	4	4.44	4.28	4.30
4	5	6.10	5.96	5.94
5	6	4.55	4.76	4.69
6	10	4.51	4.53	4.97
7	11	6.30	5.87	5.76
8	13	4.70	4.54	4.38
9	18	4.62	4.49	4.33
10	19	4.67	4.76	4.88
11	20	4.91	5.22	5.36
12	21	4.48	4.64	4.55
13	22	4.53	4.74	4.33
14	23	5.03	4.61	4.80
15	24	6.80	6.88	6.29
16	27	4.77	4.58	4.57
17	28	6.92	6.09	6.46
18	29	6.42	6.81	6.63
19	30	4.44	4.92	4.55
20	33	6.29	5.67	5.67
21	34	4.28	4.69	4.98
22	35	4.85	4.17	4.18
23	36	4.69	5.20	5.14
24	38	6.42	6.44	6.38
25	39	4.71	4.91	5.09
26	40	4.63	4.67	4.79
27	41	5.92	6.31	6.04
28	42	5.41	5.49	5.78
29	43	4.40	4.52	4.60
30	45	4.57	5.27	5.04
31	47	4.78	4.85	4.93
32	48	4.45	4.80	4.78
33	49	4.36	4.62	4.35
34	50	4.63	4.93	5.03
35	51	4.66	4.36	4.30
36	53	6.31	6.23	5.93
37	54	4.66	5.19	5.01
38	55	6.10	6.15	6.45
39	56	4.93	4.77	4.71
40	57	5.77	5.62	6.00
41	59	4.82	4.59	4.59
42	60	6.33	5.41	5.75
43	61	4.60	4.83	5.01
44	62	4.89	4.76	4.36
45	64	4.32	4.64	4.57

Table 4 Actual, model-1 and model-2 predicted, $P_{\text{caco-2}}$ values for test set molecules

S. No	Molecule	Actual $P_{\text{caco-2}}$	$P_{\text{caco-2}}$ (Model-1)	$P_{\text{caco-2}}$ (Model-2)
1	1	6.29	5.38	5.12
2	12	4.69	6.19	5.61
3	14	6.82	5.61	5.79
4	15	6.13	5.91	5.64
5	16	4.66	5.17	4.86
6	25	5.64	4.59	4.79
7	26	5.43	5.75	5.60
8	31	4.68	5.57	5.74
9	35	4.85	5.41	5.46
10	37	5.03	5.85	5.43
11	39	4.71	4.17	4.36
12	46	4.61	4.03	4.18
13	52	4.69	4.42	4.16
14	58	6.52	5.47	4.95
15	63	5.34	6.83	7.54

0.79 (ME) while for model 2 it was 0.96 (RMSE) and 0.80 (ME).

$$P_{\text{Caco-2}} = 4.17571 - 0.041073 \times LE14 - 1.6024 \times Dipole + 0.388243 \times Chi8 + 0.001484 \times I_c - 0.023404 \times LEHBD - 0.009813 \times DEHBD \quad (2)$$

$$r^2=0.805; q^2=0.696; n=45$$

The best MI-QSAR models for Caco-2 cell permeability, with different numbers of descriptor terms are as follows:

1-term Model

$$P_{\text{Caco-2}} = 4.83609 + 0.019949 \times LEHBD \quad (3)$$

$$r^2=0.353; q^2=0.163; n=45$$

2-term Model

$$P_{\text{Caco-2}} = 5.24901 - 0.02785 \times LEHBD - 0.025547 \times LE14 \quad (4)$$

$$r^2=0.502; q^2=0.288; n=45$$

3-term Model

$$P_{\text{Caco-2}} = 5.12795 - 0.008272 \times DEHBD - 0.022744 \times LEHBD - 0.025729 \times LE14 \quad (5)$$

$$r^2=0.571; q^2=0.322; n=45$$

Table 5 Correlation between model-2 descriptors

	LE14	LEHBD	DEHBD	Chi8	Dipole	Ic	P _{Caco-2}
LE14	1						
LEHBD	-0.52	1					
DEHBD	-0.27	0.50	1				
Chi8	0.57	-0.41	-0.19	1			
Dipole	-0.21	-0.13	0.04	0.04	1		
Ic	-0.12	0.14	0.20	0.01	0.35	1	
P _{Caco-2}	-0.02	-0.59	-0.52	0.37	0.15	0.20	1

4-term Model

$$P_{Caco-2} = 4.6808 - 0.020225 \times LEHBD - 0.032855 \times LE14 + 0.39337 \times Chi8 - 0.008614 \times DEHBD \quad (6)$$

$$r^2=0.701; q^2=0.580; n=45$$

5-term Model

$$P_{Caco-2} = 4.03869 + 0.001108 \times I_c - 0.035904 \times LE14 + 0.361843 \times Chi8 - 0.020611 \times LEHBD - 0.009813 \times DEHBD \quad (7)$$

$$r^2=0.770; q^2=0.688; n=45$$

6-term Model

$$P_{Caco-2} = 4.17571 - 0.041073 \times LE14 - 1.6024 \times Dipole + 0.388243 \times Chi8 + 0.001484 \times I_c - 0.023404 \times LEHBD - 0.009813 \times DEHBD \quad (8)$$

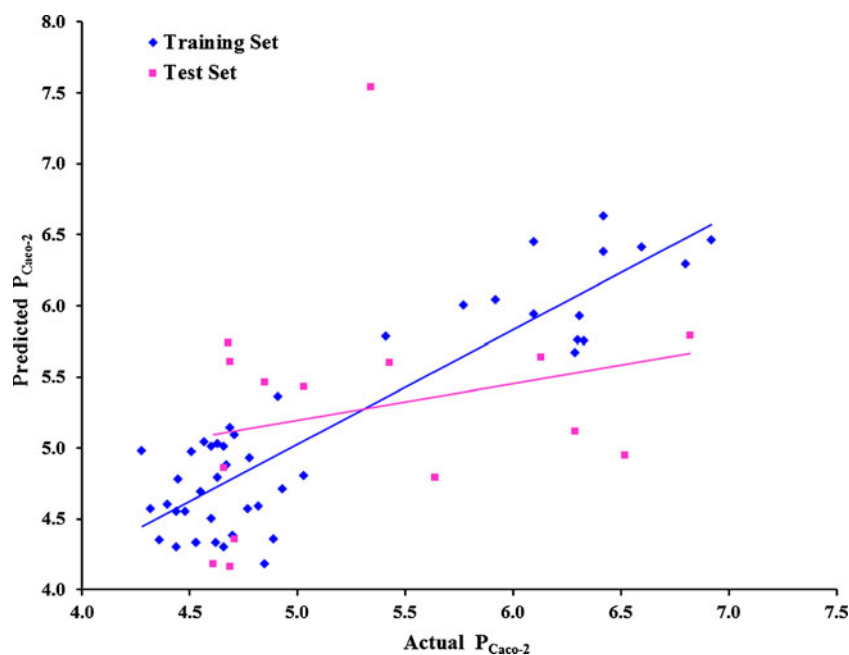
$$r^2=0.805; q^2=0.696; n=45.$$

The best MI-QSAR models for Caco-2 cell permeability were realized by considering the combination of general intramolecular solute, intermolecular dissolution/solvation-solute, and intermolecular membrane-solute descriptors. It appears from the analysis of Eqs. 3–8, that there is no specific descriptor which accounts for the variance of P_{Caco-2} across the training set. The first term model is a

Table 6 Descriptors values for test set molecules of model-2

S. No.	M. No.	P _{Caco-2}	LE14	LEHBD	DEHBD	Chi8	Dipole	Ic
1	1	6.29	16.08	-15.44	-12.09	1.83	0.14	438.14
2	12	4.69	19.61	-17.79	-92.98	1.59	0.13	372.30
3	14	6.82	15.43	-5.39	-46.38	3.05	0.23	594.15
4	15	6.13	-4.96	-5.64	-25.53	0.61	0.35	825.05
5	16	4.66	3.09	0.00	-12.54	0.81	0.23	499.37
6	25	5.64	31.36	-21.85	-12.50	1.64	0.08	510.76
7	26	5.43	70.92	-51.63	-82.43	6.36	0.25	195.53
8	31	4.68	-2.65	-3.11	-13.39	1.02	0.18	778.07
9	35	4.85	25.85	0.00	0.00	1.15	0.07	493.22
10	37	5.03	19.98	-22.81	-52.25	1.60	0.25	557.24
11	39	4.71	42.60	-16.99	-9.13	2.36	0.09	946.80
12	46	4.61	22.76	0.00	0.00	0.81	0.15	580.99
13	52	4.69	35.71	-3.82	-31.04	1.47	0.19	548.07
14	58	6.52	24.33	-11.13	-30.52	2.51	0.29	484.70
15	63	5.34	-20.70	0.00	-88.33	0.58	0.16	1158.80

Fig. 6 Plot for actual versus predicted P_{Caco-2} values of the training and test set molecules of model-2. The training set and test set molecules are shown in blue (diamond) and pink (square) spots, respectively



specific membrane interaction descriptor showing the influence of membrane system on solute molecules. A composite analysis of all the MI-QSAR models using Eqs. 2–8, suggests that the 4-term MI-QSAR model captures the essential features of the postulated mechanism responsible for solute membrane permeability as represented by P_{Caco-2} values. The 4-term model showed statistical improvement over the 1-term model. After successive refinement of the 4-term model, 5 & 6 term models were obtained with two additional descriptors. The 5 & 6 term models fit well with the training set and are also supported by the improved statistical values after the addition of each new descriptor. It is noted from the 1–6 term models that the regression coefficients of the descriptors are found to be remarkably similar to each other indicating the robustness of the models and their respective roles in predicting the permeability. Out of all the equations, we have selected the 6-term model for further analysis. One of the descriptor selected for the 6-term model is *LEHBD*. *LEHBD* is the intramolecular hydrogen bonding energy of the solute molecule when it is in the lowest membrane solute interaction state within the membrane. As this is correlated negatively, increasing the value will decrease the Caco-2 cell permeability. This is the descriptor which has the high correlation. The next descriptor is *LE14* which represents the Van der Waals and electrostatic energies associated with each set of atoms separated by one torsion angle in the solute molecule and all the DMPC molecules of the model membrane. This contribution to the total conformational energy measures the composite rigidity of an average torsion rotation of the entire solute-membrane system. As *LE14* increases, the

molecules of the membrane solute system, on the average are moving away from minimum energy conformer states and exploring more conformational states, thus expressing greater flexibility.

This greater flexibility results in a higher permeation coefficient of the solute molecule based on the negative regression coefficients for *LE14* in Eqs. 4–8. Presumably, an increase in conformational flexibility of the membrane solute system makes it easier for the solute to navigate through the membrane. *DEHBD* is the change in the hydrogen bonding energy of the entire membrane-solute system. Solute is relocated from free-space to the position corresponding to the lowest solute membrane interaction energy state of the model system. This is the change (complex - solute alone - membrane alone) of hydrogen bonding energy. *DEHBD* is the difference in the total hydrogen bond energy of the solute in the membrane minus the solute being present in free space and the membrane by itself. No hydrogen bonding can occur within, or between, DMPC molecules. Thus, the hydrogen bond energy of the membrane by itself is zero. The regression coefficients of this descriptor term are negative. It shows that if intramolecular hydrogen bonding of the solute decreases upon uptake into the membrane, a decrease in intramolecular solute hydrogen bonding should correspond to an increase in the conformational flexibility of the solute. Solute conformational flexibility within the membrane is very important for high permeability as other MI-QSAR model descriptors. *Chi-8* is the topological descriptor measuring the size and shape of a molecule. *Chi-8* which is one of the topological indices developed to encode both molecular size and shape information within a common measure.

Table 7 Statistics for different HQSAR models

Model	Fragment distinction	q ² (LOO)	r ² _{ncv}	S.E.	B.H.L.	Component
a	A/B/C	0.165	0.711	0.449	307	4
b	A/B/H	NO MODEL				
c	A/B/Ch	0.280	0.793	0.390	151	6
d	A/B/DA	0.539	0.915	0.249	151	6
e	A/B/C/H	0.030	0.560	0.554	83	4
f	A/B/C/Ch	0.248	0.868	0.312	257	6
g	A/B/C/DA	0.311	0.890	0.284	53	6
h	A/B/H/Ch	0.164	0.837	0.346	59	6
i	A/B/H/DA	0.277	0.894	0.279	401	6
j	A/B/Ch/DA	0.384	0.908	0.260	151	6
k	A/B/C/H/Ch	0.054	0.665	0.483	151	4
l	A/B/C/Ch/DA	0.380	0.933	0.211	199	6
m	A/B/C/H/DA	0.375	0.787	0.391	59	5
n	A/B/H/DA/Ch	0.243	0.876	0.302	199	6
o	A/B/C/H/Ch/DA	0.288	0.863	0.313	151	5

Caco-2 cell permeability is positively correlated to *Chi-8* in Eq. 8. Thus, the form of *Chi-8* in Eqs. 6–8 suggests that the more bulky/large is a solute molecule, the less will be its permeability through a Caco-2 cell membrane which makes intuitive sense. Dipole moment of the molecule increases as the permeability increases. As the polarizability increases the hydrogen bonding energy decreases and the molecule will be readily permeable.

Consistent with results of Kulkarni et al., model-2 consisted of intramolecular hydrogen bonding (LEHBD and LE14), size and shape of molecule (*Chi8*) and change of hydrogen bonding energy membrane-solute system as descriptors governing the permeability of the compounds [17]. In addition to this, the model provided a novel intermolecular descriptor, *Ic*, hypothetical crystal-melt transition temperature of a solute, as a negative contributor toward the permeability. It is a measure of crystal packing strength of a solute and describes the dissolution behavior of a solute. The dissolution of a drug in water is controlled by two types of interaction, intra-molecular and inter-molecular. The first describes how strongly the molecule

associates with the solvent. Compounds having a large number of polar groups such as sugar alcohols show more favorable interactions with water than themselves. This is often translated into greater solubility. The second defines the crystal packing strength, i.e., the affinity of the solute for itself, or how tightly bound the compound is to its own crystal lattice. The stronger the inter-molecular interactions, the higher the energy required to disintegrate molecules out of it, leading to lower solubility. Thus an increase in the crystal-melt transition temperature lowers the solubility which in turn lowers the permeability. This relationship suggests that as dissolution of a solute becomes difficult, the Caco-2 permeability gets lowered.

Model-2 showed high residual (actual=5.34, predicted=7.54) for the molecule 63 (urea). Urea possesses no intramolecular hydrogen bonding. Because of smaller size it makes hydrogen bonds with the water molecule that provide it high solubility in water. The training set is not sufficiently populated with molecules that show similar molecular weight (60.06 Da, lowest among all molecules considered), intramolecular binding and permeability as

Table 8 Statistical parameters obtained for HQSAR model-d with different atom count

Model	Atom count	q ² (LOO)	r ² _{ncv}	S.E.	B.H.L.	Component
d-1	1-4	0.565	0.846	0.277	257	6
d-2	2-5	0.538	0.894	0.279	151	6
d-3	3-6	0.473	0.890	0.285	307	6
d-4	4-7	0.539	0.915	0.249	151	6
d-5	5-8	0.119	0.892	0.282	199	6
d-6	6-9	0.286	0.725	0.438	307	4
d-7	7-10	0.382	0.781	0.396	97	5

that of urea. This can be attributed to the high residual (2.2) in the prediction. Thus the model can be improved by increasing the chemical space.

HQSAR

Model development and validation

HQSAR uses different parameters for developing the model. Parameters like atom count, hologram length and fragment distinctions are important while (when) generating the hologram fingerprints [40–42]. The atom parameters enable fragments determination based on elemental atom types while the bonds and connections consider the bond orders and hybridization states within fragments respectively [43]. Initially, for developing the model, we used the default parameters namely atoms (A), bonds (B) and connections (C) which gave poor r^2 (0.680) and q^2 (0.221) values. After trying various combinations of the default parameters we included other parameters like hydrogen atom (H) chirality (Ch), donor acceptor (DA) to develop robust models.

We developed numerous models with the combination of parameters such as A/B/C, A/B/H, A/B/Ch, A/B/DA, A/B/C/H, A/B/C/Ch, A/B/C/DA, A/B/H/Ch, A/B/H/DA, A/B/Ch/DA, A/B/C/H/Ch, A/B/C/Ch/DA, A/B/C/H/DA, A/B/H/DA/Ch, A/B/C/H/Ch/DA. Model-d built with four parameters atom, bond, donor and acceptor was found to best predictive over training set. It gave statistically significant r^2 and q^2 of 0.915 and 0.539 respectively as shown in Table 7.

As the second criterion for improving the statistical values we used the atom count which refers to the minimum and maximum length of the fragment in hologram finger print. The statistical parameters obtained for a variety of models developed using A/B/DA parameters with different atom count is shown in Table 8. There was a significant difference noticed in statistical values of models with reference to the atom counts. Models with increase or decrease in the atom count with respect to atom count of 4–7 showed reduced prediction power. So, the Model-d-4 developed with atom count of 4–7 with essential parameters namely atoms, bonds, donor and acceptor was used for further analysis. For this model, actual and predicted P_{Caco-2} values of the training and test set molecules are shown in Tables 9 and 10 respectively, while their plot is shown in Fig. 7.

This model showed less predictivity on the test set of 15 molecules ($r_{pred}^2=0.47$). Among the test set, three molecules [molecule no. 53 (ranitidine), 58 (sulfasalazine) and 60 (terbutaline)] showed very large residuals. This model was found to be sufficiently trained with the compounds containing guanidine group as it showed good predictivity for the compounds in training set [molecule 2 (residual:

Table 9 Actual and predicted P_{Caco-2} values for training set molecules of HQSAR model

S. No.	Molecule	Actual P_{Caco-2}	Predicted P_{Caco-2}	Residual
1	1	6.29	5.59	0.70
2	10	4.51	4.38	0.13
3	12	4.69	4.75	-0.06
4	13	4.70	4.85	-0.15
5	14	6.82	6.80	0.02
6	15	6.13	5.92	0.21
7	16	4.66	4.77	-0.11
8	2	6.60	6.45	0.15
9	20	4.91	4.87	0.04
10	21	4.48	4.34	0.14
11	22	4.53	4.55	-0.02
12	23	5.03	5.04	-0.01
13	24	6.80	6.84	-0.04
14	25	5.64	5.48	0.16
15	26	5.43	5.29	0.14
16	28	6.92	6.38	0.54
17	29	6.42	6.59	-0.17
18	30	4.44	4.36	0.08
19	32	4.85	4.82	0.03
20	33	6.29	6.73	-0.44
21	34	4.28	4.80	-0.52
22	35	4.85	4.84	0.01
23	37	5.03	4.97	0.06
24	38	6.42	6.60	-0.18
25	39	4.71	4.73	-0.02
26	40	4.63	4.94	-0.31
27	42	5.41	5.33	0.08
28	43	4.40	4.30	0.10
29	44	4.71	4.72	-0.01
30	45	4.57	4.74	-0.17
31	46	4.61	4.41	0.20
32	48	4.45	4.44	0.01
33	49	4.36	4.52	-0.16
34	50	4.63	4.60	0.03
35	51	4.66	4.54	0.12
36	52	4.69	4.66	0.03
37	55	6.10	6.16	-0.06
38	56	4.93	4.87	0.06
39	57	5.77	5.95	-0.18
40	59	4.82	4.71	0.11
41	61	4.60	4.77	-0.17
42	62	4.89	4.83	0.06
43	63	5.34	4.99	0.35
44	64	4.32	4.95	-0.63
45	6	4.55	4.71	-0.16

Table 10 Actual and predicted $P_{\text{caco-2}}$ values for test set molecules of HQSAR model

S. No.	Molecule	Actual $P_{\text{caco-2}}$	Predicted $P_{\text{caco-2}}$	Residual
1	3	4.60	5.07	-0.47
2	4	4.44	4.41	0.03
3	5	6.10	6.59	-0.49
4	11	6.30	5.85	0.45
5	18	4.62	4.80	-0.18
6	19	4.67	4.52	0.15
7	27	4.77	4.78	-0.01
8	31	4.68	5.16	-0.48
9	36	4.69	4.82	-0.13
10	41	5.92	5.34	0.58
11	47	4.78	4.93	-0.15
12	53	6.31	5.14	1.17
13	54	4.66	5.12	-0.46
14	58	6.52	5.51	1.01
15	60	6.33	5.17	1.16

0.15); molecule 15 (residual: 0.21); molecule 16 (residual: -0.11); molecule 29 (residual: -0.17)] and test set molecule 31 (residual: -0.13). Molecule 53, ranitidine, contains the guanidine group where the imine nitrogen is replaced with the carbon atom. This arrangement is absent in the training set molecules. This can be attributed to the weak prediction of permeability of this compound. Molecule 58 (sulfasalazine) contains the functional group diazene (N=N) that is absent in training set chemical space. This can be the cause of residual 1.01 in the prediction space. Tebutaline is a recemate and its bioavailability is stereoselective. Based on the plasma data, Borgstrom et al. have found that the oral bioavailability of (+)tebutaline is 7.5% and that of (-)

tebutaline is 14.8%. The bioavailability of (\pm)tebutaline is similar to that of (-)tebutaline [44]. Thus the permeability of the turbutaline is dependent on the isoforms and their equilibrium concentrations. This has lead to the weak prediction of terbutaline permeability. If these three outliers are removed, the model shows significantly better r_{pred}^2 of 0.70. Thus, two out of three outliers are because of insufficient molecular space covered in the training set. Because of this the model may not be applicable for the prediction of molecules having imine and diazine functional groups.

Color coding

The HQSAR module in SYBYL uses the color coding to show the atomic contributions to the activity. While the color codes red, red orange and orange show the favorable or positive contribution to the activity, the color codes yellow, green blue, green denote unfavorable or negative contribution to the activity. The white color code shows the intermediate contribution to the activity. Contribution map obtained for a few molecules by HQSAR analysis is shown in Fig. 8. Molecule 49 with three nitrogen atoms has red color code on the ring system which increases the solubility in water, thus enhances the absorption of the molecule. Molecule 43 has red-orange color on the naphthalene ring which indicates that the ring plays an important role in enhancing permeability by increasing the lipid solubility. The carboxylic acid group is coded with orange color which also further indicates toward enhanced solubility of the molecule.

The aromatic ether is a hydrogen bond acceptor and thus increases the water solubility. Molecule 48 has N-H group which is hydrogen bond donor and it makes the molecule

Fig. 7 Plot for actual versus predicted $P_{\text{Caco-2}}$ values of the training and test set molecules of HQSAR-d-4 model. The training set and test set molecules are shown in blue (diamond) and pink (square) spots, respectively

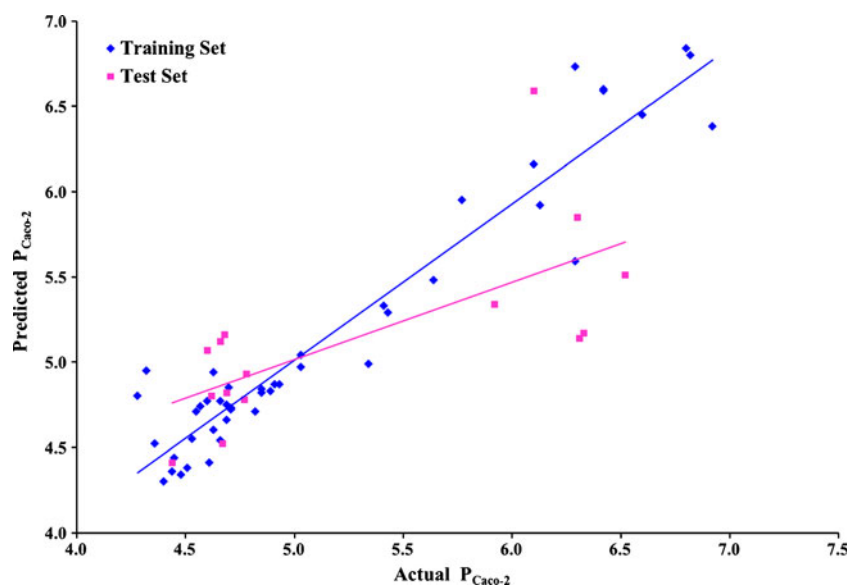
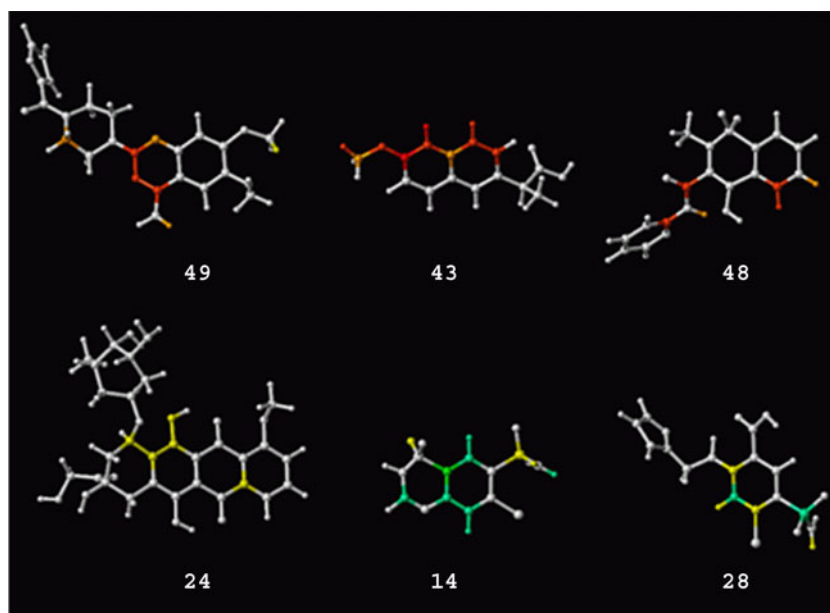


Fig. 8 Contribution map obtained for a few molecules by HQSAR analysis



highly polar thereby increasing its lipid solubility. Moreover, amide bond which is also highly polar increases the lipid solubility. In molecule 14 and 28, the sulfone group has the green color code which shows that the group may be responsible for a decrease in the absorption.

Fragment analysis

The final HQSAR model produced hundreds of fragments. The fragments produced are useful in predicting the

variability in biological activity. Although a direct correlation may not be established between the activity and all the fragments produced, these fragments provide useful hints toward improving the activity. Analysis of the fragments and their contributions shows that the fragments possessing positive values contribute favorably to the activity while the fragments possessing negative values contribute unfavorably to the activity. A few fragments showing positive and negative contributions toward the absorption of the various drugs are shown in Fig. 9.

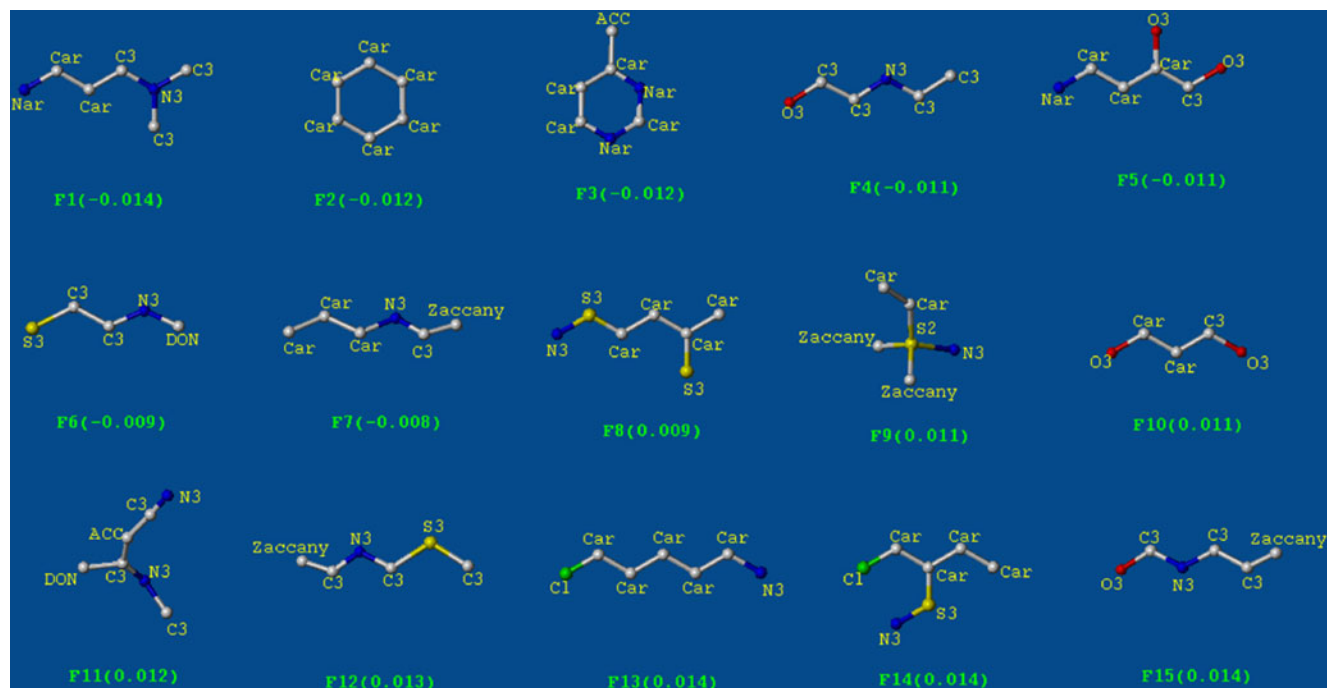


Fig. 9 A few fragments showing positive and negative contributions toward the permeability of the various drugs

Fragment 1 has a negative coefficient value of -0.014 . This fragment has the methyl group which may increase electron density on nitrogen and increases the polarity of the molecule thus making the molecule water soluble. As the number of hydrocarbon groups like methyl group increases, the lipid partitioning of the molecule is also nicely enhanced as seen in molecule 10. Fragment 2 has coefficient value of -0.012 and is derived from an aromatic ring system. The aromatic ring is a non-polar group and it will increase the lipid partitioning of the molecules. Fragment 3 is derived from molecule 49 with the coefficient value of -0.012 . This fragment 3 has hydrogen bond acceptors which may increase the water solubility. Fragment 4 has the coefficient value of -0.011 and the two hydrogen bond acceptor groups. It might positively contribute in their absorption which shows that it may contribute certainly for the absorption of the molecule. Fragment 5 has the coefficient value of -0.011 . This fragment is also derived from molecule 49 which has two oxygen with the two methyl groups substituted on them. These methyl groups are non-polar groups and increases the lipid partitioning of the molecule. Molecule 49 has both the fragment that are involved in lipid and water solubility which is important in absorption. Fragment 6 and 7 have negative coefficient values and positively contribute in absorption as seen in molecule 22.

Fragment 8 has the coefficient value of 0.009 ; it is observed that sulfur increases the electron density on the nitrogen while the hydrogen bonding property of the nitrogen decreases.

Fragment 10 has positive coefficient value of 0.011 . In this fragment the electron delocalization from one oxygen to another oxygen decrease the hydrogen-bonding-acceptor property of the oxygen and negatively affect the absorption as seen in molecule number 58.

Fragment 11 is cyanoguanidine with the coefficient value of 0.012 . This property may decrease the absorption of the molecule as seen in molecule 15 which is highly polar in nature and affects absorption. Fragment 12 has coefficient value of 0.013 and is from the beta lactum ring which is also more hydrophilic in nature.

Fragment 13 has the coefficient value of 0.014 . The halogen substitution in the molecules may play an important role in increasing the polarization and thus decrease the water solubility and absorption of the molecule as seen in molecule 28. Fragment 14 has the coefficient value of 0.014 . This fragment also has the chlorine substitution, thus may negatively contribute in their absorption. Fragment 15 has the coefficient value of 0.014 and is taken from the amide bond of the molecules which increases the hydrophilicity of the molecule. In spite of having acceptor groups, the hydrogen-bonding-acceptor properties decreases and negatively contributes in their absorption.

Conclusions

An important strength of the MI-QSAR approach is to build simple and statistically significant relationships such as Eqs. 1–3. Solute conformational flexibility within the membrane is important for high permeability. Decrease in the hydrogen bonding energy of the solute increases the absorption of the molecule. *Chi-8* points out that the smaller molecule will permeate more freely than the larger one. Dipole moment increases as the permeability increases.

Through HQSAR analysis we noted important features for improving the absorption of drugs and drug-like molecules. Maintaining the HLB is important for the absorption of the molecule. The presence of one or more chlorines shows negative contribution for absorption. The methyl group increases the lipophilicity thus affecting the absorption. The presence of a polar group will have a negative effect on absorption.

Acknowledgments The authors thank Department of Science and Technology (DST), Delhi, and Council of Scientific and Industrial Research (CSIR), New Delhi for financial assistance.

References

1. Van de Waterbeemd H, Gifford E (2003) ADMET in silico modelling: towards prediction paradise? *Nat Rev Drug Discov* 2:192–204
2. Wishart DS (2007) Improving Early Drug Discovery through ADME: An Overview. *Drugs R&D* 8:349–362
3. Singer SJ, Nicolson GL (1972) The fluid mosaic model of the structure of cell membranes. *Science* 175:720–731
4. Conradi RA, Burton PS, Borchardt RT (1996) Physico-chemical and biological factors that influence a drug's cellular permeability by passive diffusion. *Methods Princ Med Chem* 4:233–252
5. Delie F, Rubas WA (1997) A human colonic cell line sharing similarities with enterocytes as a model to examine oral absorption: advantages and limitations of the Caco-2 model. *Crit Rev Ther Drug Carrier Syst* 14:221–286
6. Artursson P, Palm K, Luthman K (2001) Caco-2 monolayers in experimental and theoretical predictions of drug transport. *Adv Drug Deliv Rev* 46:27–43
7. Lipsinki CA, Lomabardo 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
8. Camenisch G, Folkers G, van de Waterbeemd H (1996) Review of theoretical passive drug absorption models: historical background, recent developments and limitations. *Pharm Acta Helv* 71:309–327
9. Chan O, Stewart BH (1996) Physicochemical and drug-delivery consideration for oral drug bioavailability. *Drug Discov Today* 1:461–473
10. Palm K, Stenberg P, Luthman K, Artursson P (1997) Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm Res* 14:568–571
11. Camenisch G, Alsenz J, van de Waterbeemd H, Folkers G (1998) Estimation of permeability by passive diffusion through Caco-2

- cell monolayers using drugs' lipophilicity and molecular weight. *Eur J Pharm Sci* 6:313–319
12. Ponce YM, Cabrera Pérez MA, Zaldivar VR, Ofori E, Montero LA (2003) Total and local quadratic indices of the “molecular pseudograph's atom adjacency matrix”. application to prediction of caco-2 permeability of drugs. *Int J Mol Sci* 4:512–536
 13. Iyer M, Tseng YJ, Senese CL, Liu J, Hopfinger AJ (2007) Prediction and mechanistic interpretation of human oral drug absorption using MI-QSAR analysis. *Mol Pharm* 4:218–231
 14. Fujiwara SI, Yamashita F, Hashida M (2002) Prediction of caco-2 cell permeability using a combination of molecular calculation and neural networks. *Int J Pharm* 237:95–105
 15. Norinder U, Osterber T, Artursson P (1997) Theoretical calculation and prediction of caco-2 cell permeability using molsurf parametrization and pls statistics. *Pharm Res* 14:1786–1791
 16. Cruciani G, Crivori P, Carrupt PA, Testa B (2000) Molecular fields in quantitative structure-permeation relationships: the volsurf approach. *J Mol Struct THEOCHEM* 503:17–30
 17. Kulkarni A, Han Y, Hopfinger AJ (2002) Predicting Caco-2 cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis. *J Chem Inf Comput Sci* 42:331–342
 18. Liang E, Chessic K, Yazdani M (2000) Evaluation of an accelerated Caco-2 cell permeability model. *J Pharm Sci* 89:336–345
 19. Ponce YM, Perez MAC, Zaldivar VR, Diaz HG, Torrens F (2004) A new topological descriptors based model for predicting intestinal epithelial transport of drugs in Caco-2 cell culture. *J Pharm Sci* 7:186–199
 20. Zhu C, Jiang L, Chen TM, Hwang KK (2002) A comparative study of artificial membrane permeability assay for high throughput profiling of drug absorption potential. *Eur J Med Chem* 37:399–407
 21. Yee S (1997) In vitro permeability across Caco-2 cells (colonic) can predict in vivo (small intestinal) absorption in man—fact or myth. *Pharm Res* 14:763–766
 22. Pade V, Stavchansky S (1998) Link between drug absorption solubility and permeability measurements in Caco-2 cells. *J Pharm Sci* 87:1604–1607
 23. Yazdani M, Glynn SL, Wright JL, Hawi A (1998) Correlating partitioning and caco-2 cell permeability of structurally diverse small molecular weight compounds. *Pharm Res* 15:1490–1494
 24. SYBYL Molecular Modeling System, version 7.1, Tripos Inc., St. Louis, MO, 63144–2913, USA
 25. HYPERCHEM, version 4.5, Hypercube Inc, Waterloo, Ontario Canada N2L 3X2
 26. Van der Ploeg P, Berendsen HJC (1982) Molecular dynamics simulation of a bilayer membrane. *J Chem Phys* 76:3271–3276
 27. MI-QSAR, Version 1.0, The Chem21 Group Inc, 1780 Wilson Drive, IL, 60045, USA
 28. Kulkarni AS, Hopfinger AJ (1999) Membrane-interaction QSAR analysis: Application to the estimation of eye irritation by organic compounds. *Pharm Res* 16:1245–1253
 29. Hauser H, Pascher I, Pearson RH, Sundell S (1981) Preferred conformation and molecular packing of phosphatidylethanolamine and phosphatidylcholine. *Biochim Biophys Acta* 650:21–51
 30. Kulkarni A, Hopfinger AJ, Osborne R, Bruner LH, Thompson ED (2001) Prediction of eye irritation from organic chemicals using membrane-interaction QSAR analysis. *Toxicol Sci* 59:335–345
 31. Iyer M, Mishra R, Han Y, Hopfinger AJ (2002) Predicting blood–brain barrier partitioning of organic molecules using membrane-interaction QSAR analysis. *Pharm Res* 19:1611–1621
 32. Doherty DC, MOLSIM, Version 3.0, The Chem21 Group Inc, 1780 Wilson Drive, IL, 60045, USA
 33. 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
 34. Liu J, Li Y, Pan D, Hopfinger AJ (2005) Predicting permeability coefficient in ADMET evaluation by using different membranes-interaction QSAR. *Int J Pharm* 304:115–123
 35. Kodithala K, Hopfinger AJ, Thompson ED, Robinson MK (2002) Prediction of skin irritation from organic chemicals using membrane-interaction QSAR analysis. *Toxicol Sci* 66:336–346
 36. Iyer M, Tseng YJ, Senese CL, Liu J, Hopfinger AJ (2006) Prediction and Mechanistic Interpretation of Human Oral Drug Absorption Using MI-QSAR Analysis. *Mol Pharm* 4:218–231
 37. Li Y, Liu J, Pan D, Hopfinger AJ (2005) A study of the relationship between cornea permeability and eye irritation using membrane-interaction QSAR analysis. *Toxicol Sci* 88:434–446
 38. Powell MJD (1977) Restart procedures for the conjugate gradient method. *Math Program* 12:241–254
 39. Lewis DR (1997) HQSAR a new, highly predictive QSAR technique. *Tripos Tech Notes* 1:1–7
 40. Rodrigues CR, Flaherty TM, Springer C, McKerrow JH, Cohen FE (2002) CoMFA and HQSAR of acylhydrazide cruzain inhibitors. *Bioorg Med Chem Lett* 12:1537–1541
 41. Moda TL, Montanari CA, Andricopulo AD (2007) Hologram QSAR model for the prediction of human oral bioavailability. *Bioorg Med Chem* 15:7738–7745
 42. Honorio KM, Garratt RC, Andricopulo AD (2005) Hologram quantitative structure–activity relationships for a series of farnesoid X receptor activators. *Bioorg Med Chem Lett* 15:3119–3125
 43. Castilho MS, Guido C, Andricopulo AD (2007) Classical and hologram QSAR studies on a series of tacrine derivatives as butyrylcholinesterase inhibitors. *Lett Drug Des Discov* 4:106–113
 44. Borgström L, Nyberg L, Jönsson S, Lindberg C, Paulson J (1989) Pharmacokinetic evaluation in man of terbutaline given as separate enantiomers and as the racemate. *Br J Clin Pharmacol* 27:49–56

Theoretical study of crown ethers with incorporated azobenzene moiety

Yuan Miao · Xueye Wang · Dan Ouyang

Received: 17 March 2011 / Accepted: 16 May 2011 / Published online: 4 June 2011
© Springer-Verlag 2011

Abstract A series of crown ethers containing the azobenzene moiety incorporated into crowns of various sizes [Cr(O₆), Cr(O₇) and Cr(O₈)] and their corresponding alkali metal cation (Li⁺, Na⁺, K⁺, Rb⁺) complexes have been studied theoretically. The density functional theory (DFT) method was employed to elucidate the stereochemical structural natures and thermodynamic properties of all of the target molecules at the B3LYP/6-31 G(d) and LANL2DZ level for the cation Rb⁺. The fully optimized geometries had real frequencies, thus indicating their minimum-energy status. In addition, the bond lengths between the metal cation and oxygen atoms, atomic torsion angles and thermodynamic energies for complexes were studied. Natural bond orbital (NBO) analysis was used to explore the origin of the internal forces and the intermolecular interactions for the metal complexes. The calculated results show that the most significant interaction is that between the lone pair electrons of electron-donating oxygens in the *cis*-forms of azobenzene crown ethers (*cis*-ACEs) and the LP* (1-center valence antibond lone pair) orbitals of the alkali-metal cations (Li⁺, Na⁺, K⁺ and Rb⁺). The electronic spectra for the *cis*-ACEs [*cis*-Cr(O₆), *cis*-Cr(O₇) and *cis*-Cr(O₈)] are obtained by the time-dependent density functional theory (TDDFT) at the B3LYP/6-31 G(d) level. The spectra of the *cis*-isomers show broad $\pi \rightarrow \pi^*$ (S₀ → S₂) absorption bands at 310–340 nm but weaker $n \rightarrow \pi^*$ (S₀ → S₁) bands at 480–490 nm. The calculated results are in good agreement with the experimental results.

Keywords Azobenzene crown ethers (ACEs) · Photoisomerization · Preorganization · Switchable molecules · Time-dependent density functional theory (TDDFT)

Introduction

Supramolecular chemistry is a highly interdisciplinary field covering the chemical, physical, and biological features of complex chemical species that are held together and organized by means of intermolecular (noncovalent) bonding interactions. How well things fit together depends on their predisposition to do so, a matter frequently referred to as “preorganization.” Reliably predicting host–guest interactions is an important goal of supramolecular chemistry [1]. A molecular system with a preorganized and effectively functionalized recognition unit for guest molecules is ideal for host–guest interactions.

Nowadays, most molecular builders are very interested in constructing switchable molecular systems that can selectively bind different metal cations. The key to designing a successful system of this type involves the use of binding interactions that have well-defined, predictable geometric consequences. These are important aspects in the development of functional molecular devices of increasing complexity [2]. Ever since the first synthesis of crown ethers was reported by Pedersen [3], these molecules have been the focus of extensive study due to their ability to complex metal cations [4]. A large number of studies have shown that the binding properties of crown ethers are sensitive to change in conformation or effective size [5].

Azobenzenes comprise an interesting class of compounds that exhibit photoresponsive properties. Their photoisomerization properties have led to them becoming

Y. Miao · X. Wang (✉) · D. Ouyang
Key Laboratory of Environmentally Friendly Chemistry
and Applications of Ministry of Education, College of Chemistry,
Xiangtan University,
Xiangtan, Hunan 411105, People's Republic of China
e-mail: wxueye@xtu.edu.cn

among the most common used photoresponsive molecular switches [6]. They have been incorporated into a number of supramolecular frameworks to produce ionophores for transports and photoswitchable receptors [7]. Azobenzene has the ability to undergo isomerization between the straight *trans*-isomer and the bent *cis*-isomer with light irradiation ($trans \leftrightarrow cis$) and thermal induction ($cis \rightarrow trans$) (see Fig. 1), respectively. Because of their facile interconversion at appropriate wavelengths, azobenzenes have the potential to be used in optical switching and image storage devices [8–11] as well as molecular scissors [12] and as targets for coherent control in molecular electronics [13].

The basic requirement of a successful molecular switch is the presence of two distinct forms of the molecule that can be interconverted reversibly by means of an external stimulus, such as light, heat, pressure, magnetic or electric fields, a pH change or a chemical reaction [14]. Irradiating or heating azobenzene-containing materials induces reversible isomerization between the two isomers, making azobenzenes switchable molecules. The isomerization of azobenzenes is accompanied by significant changes in the absorption spectra and structures of the molecules. These changes can alter properties of their surrounding environment by switching them “on” or “off.”

The azobenzene moiety incorporated into the crowns (see Fig. 2) is used to change the size of the crowns and hence to modify the complexing properties of the molecules [15, 16]. The combination of a crown ether with an azobenzene moiety enables us to control ionic conductivity by light irradiation or thermal induction.

The azobenzene-type crown ethers (hereafter referred to as “ACEs”) Cr(O₇) and Cr(O₈), in which the 4 and 4' positions of azobenzene are linked by a polyoxyethylene chain, were synthesized and studied by Seiji Shinkai and co-workers [17]. Cr(O₇) and Cr(O₈) have azobenzene as an antenna and the crown ether as a functional group, and change their chemical and physical functions in response to photoirradiation or changes in temperature. Similarly, azobenzene derivatives have been utilized as light-driven or temperature-driven triggers to control the functions of metal ligands.

Computational methods are a promising way to calculate the structures and properties of complexes, such as their binding energies and absorbance spectra. In the work presented here, a family of ACEs [Cr(O₆), Cr(O₇) and Cr(O₈)] with rings of different sizes containing the azoben-

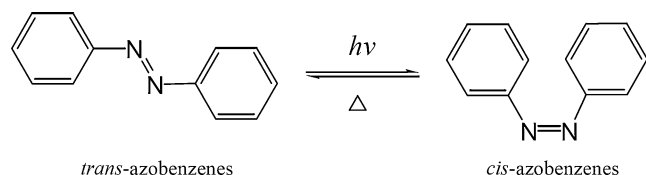


Fig. 1 Schematic diagram of the *trans* ↔ *cis* isomerization of azobenzenes

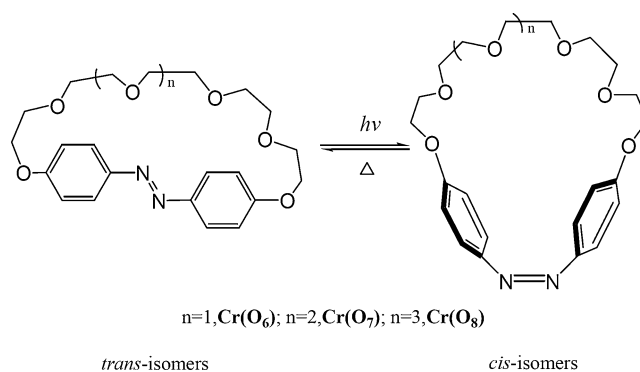


Fig. 2 Schematic diagram of the *trans* ↔ *cis* isomerization of crown ethers with an incorporated azobenzene moiety

zene moiety incorporated into the crown were studied theoretically.

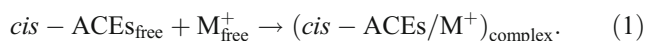
Theory and methods of calculation

In the framework of the density functional theory (DFT) approach, the B3LYP hybrid functional [18, 19] is one of the most preferred methods, as it has proven its ability to reproduce various molecular properties, including structural parameters and vibrational spectra. The combined use of the B3LYP functional and the standard split valence basis set 6-31 G(d) has been previously shown to provide an excellent compromise between the accuracy of the results and computational efficiency for large and medium-sized molecules [20–26]. Ground-state electronic structure calculations of all complexes were performed using density functional theory (DFT) methods as implemented using the Gaussian 03 software package [27]. The functional that was used throughout this study is B3LYP, consisting of a hybrid exchange functional, as defined by Becke’s three-parameter equation, and the Lee–Yang–Par correlation functional [18, 19]. The ground-state geometries were obtained in the gas phase by full geometry optimization, and the optimum structures, located as stationary points on the potential energy surfaces, were verified by the absence of imaginary frequencies. The standard 6–31 G(d) and LANL2DZ basis sets were found to be suitable for most ligands.

Time-dependent density functional theory (TDDFT) can model highly complex molecules like azobenzenes accurately, efficiently, and cost-effectively. In this study, TDDFT was used to model the absorption spectra of several azobenzene derivatives. The results show a reasonably good association between the theoretical and experimental values for the absorbance spectra of the azobenzenes. A natural bond orbital (NBO) population analysis was performed with the NBO 3.1 program as implemented in Gaussian [28–31]. NBO analysis represents a unique and powerful approach to evaluating the

origins of intermolecular interactions from a computational standpoint.

The binding energies, binding enthalpies, and Gibbs free energies in the gas phase for the complexes were calculated for the reaction



For this system, the binding energy ΔE can be expressed as follows:

$$\Delta E = E(\text{cis-ACEs}/\text{M}^+)_{\text{complex}} - [E(\text{cis-ACEs}_{\text{free}}) + E(\text{M}_{\text{free}}^+)] \quad (2)$$

Results and discussion

Optimized ground-state geometry

The structures of molecules play an especially important role in determining their chemical properties. The optimized stability structures for both the *trans* and *cis* forms of ACEs [Cr(O₆), Cr(O₇) and Cr(O₈)] were obtained at the B3LYP/6-31 G(d) level in the gas phase at 298 K, while unsubstituted *trans*- and *cis*-azobenzene were studied as reference compounds at the same level. The results of the analysis of all of the target molecules described above are depicted in Table 1, and their ground-state structures are presented in Fig. 3.

The *trans* isomer of azobenzene is about 15.1 kcal mol⁻¹ or 0.65 eV lower in energy than that of the *cis* isomer. This is only slightly higher than the experimental value of 0.6 eV [32]. The DFT results are very similar to some of the previous theoretical predictions [33–39]. The calculated results indicate that the phenyl rings of *trans*-azobenzene are 50.2° out of plane compared to those of the *cis* isomer,

and the distance between the 4 and 4' positions decreases from 9.079 Å to 6.562 Å for *trans*- and *cis*-azobenzene, respectively.

The energies of the *cis* ACEs are 16.3 kcal mol⁻¹ (*cis*-Cr(O₆)), 20.1 kcal mol⁻¹ (*cis*-Cr(O₇)), and 15.7 (*cis*-Cr(O₈)) kcal mol⁻¹ higher than those of their corresponding *trans* isomers, respectively. The optimized structures of the *trans* isomers of Cr(O₆), Cr(O₇), and Cr(O₈) are shown in Fig. 3, and the calculated parameters for them are listed in Table 1. The polyoxyethylene (–CH₂–O–CH₂–)_n (n = 1, 2, 3) chains between the two aromatic rings are almost linearly extended. The distances between the 4 and 4' positions of azobenzene of the three *trans* isomers are 8.945, 9.072 and 9.034 Å, respectively, which are all smaller than those of the unsubstituted *trans*-azobenzene (9.079 Å). The angle ∠NNCC for *trans*-Cr(O₆), *trans*-Cr(O₇), and *trans*-Cr(O₈) are 3.2°, 1.4°, and 8.4°, respectively. These results indicate that the phenyl rings of the *trans*-azobenzene unit in the ACEs are out of plane compared to those of the unsubstituted azobenzene, and there must be some steric restriction in play during the *trans* ↔ *cis* isomerization. The methylene chain of *trans*-Cr(O₈) undergoes a small amount of folding, and *trans*-Cr(O₆) shows the most restricted structure (see from Fig. 3). The *trans* isomers of the ACEs show poor preorganization because of the long loops in their structures; the isomers lack any affinity for metal cations according to the rules of supermolecule preorganization [40]. With respect to the *cis* ACEs, there is a crown loop in each of the target molecules. The preorganization of the *cis* forms of ACEs is enhanced, allowing them to coordinate with metal cations; they thus present an “on” state, while the *trans* forms of the ACEs are in an “off” state in terms of coordinating with metal cations. *Cis* forms of these crown ethers show affinity for metal cations. The sizes of the loops in the *cis*-type ACEs follows the order: *cis*-Cr(O₆) < *cis*-Cr(O₇) < *cis*-Cr(O₈). The arrangement of atoms in

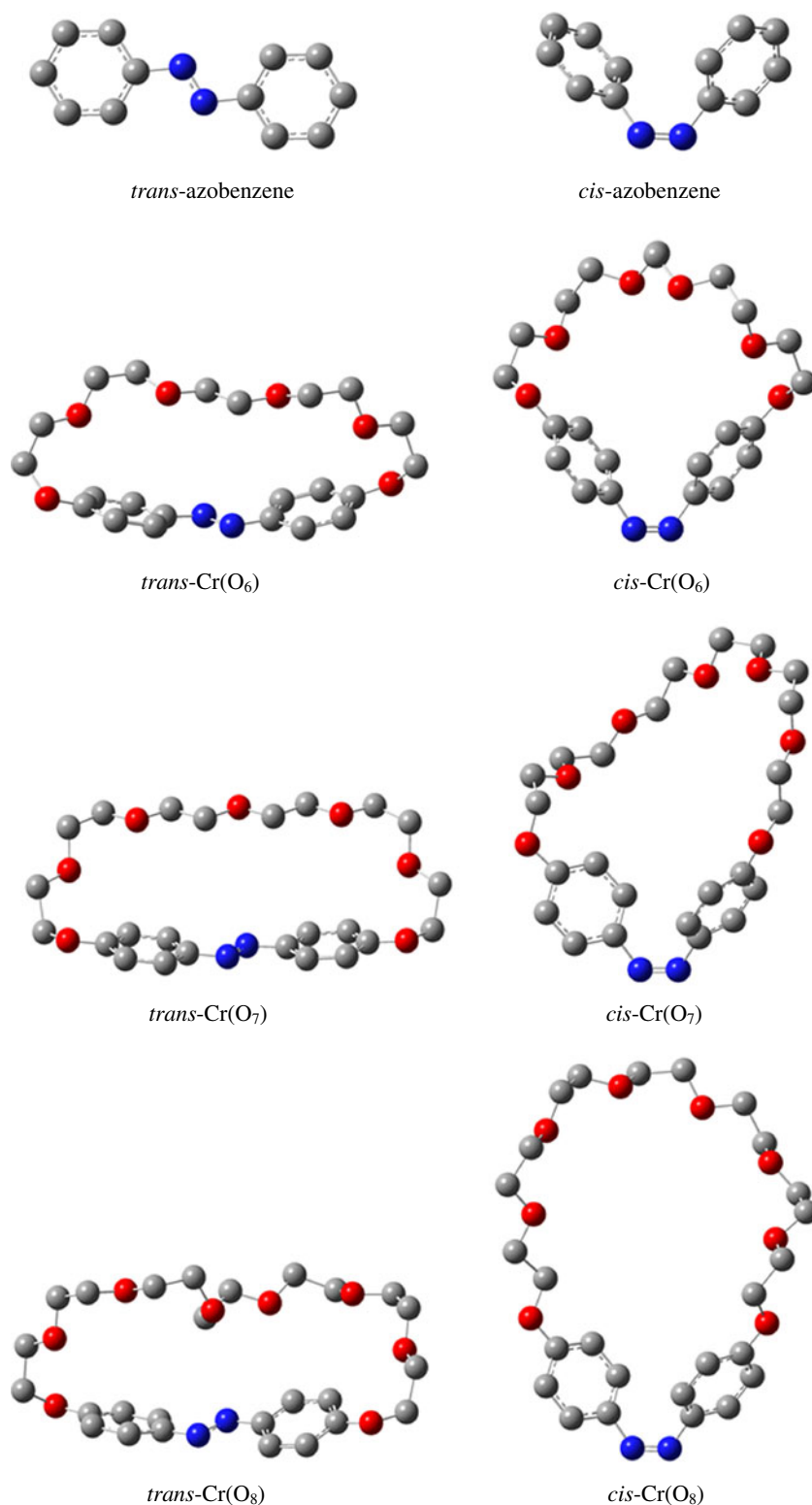
Table 1 Comparison of the calculated parameters of the *trans* and *cis* isomers of ligands optimized at the B3LYP/6-31 G(d) level

Ligand	Angle (°)		Distances (Å)			Energy ^a (kcal/mol)	
	∠CNC	∠NNCC	∠NNC	d _{NN}	d _{CN}		d _{C-C}
<i>Trans</i> -azobenzene	180.0	0	114.8	1.261	1.419	9.079	0
<i>Cis</i> -azobenzene	9.8	50.2	124.1	1.250	1.436	6.562	15.1
<i>Trans</i> -Cr(O ₆)	171.0	3.2	114.8	1.265	1.410	8.945	0
<i>Cis</i> -Cr(O ₆)	11.1	49.6	123.6	1.253	1.434	6.507	16.3
<i>Trans</i> -Cr(O ₇)	176.8	1.4	115.3	1.265	1.410	9.072	0
<i>Cis</i> -Cr(O ₇)	9.0	57.2	122.4	1.254	1.434	6.233	20.1
<i>Trans</i> -Cr(O ₈)	178.0	8.4	113.5	1.264	1.408	9.034	0
<i>Cis</i> -Cr(O ₈)	8.3	48.6	123.5	1.255	1.434	6.284	15.7

d_{C-C}: the distances (Å) between the 4 and 4' positions of azobenzene and azobenzene crown ethers (ACEs)

^a Energies are relative to the corresponding *trans* isomer of the ligand

Fig. 3 The optimized structures of the *trans* and *cis* isomers of azobenzene and azobenzene crown ethers (ACEs) at the B3LYP/6-31 G(d) level



the *cis* isomers is more relaxed than that in the *trans* isomers. In addition, the flexibility of the crown-like loop increases as the number of $-\text{CH}_2-\text{O}-\text{CH}_2-$ units between the 4 and 4' positions of azobenzene increases. These results show that in ACEs with a polyoxyethylene chain, the crown-like loop

appears in *cis* ACEs and disappears in *trans* ACEs, causing an “all-or-nothing” change in the ion-binding ability. The molecules of ACEs in their *cis* and *trans* isomer forms act as “switched-on” and “switched-off” crown ethers, respectively.

Optimized geometries of the complexes

The optimized structures of the *cis* ACE/ M^+ complexes [*cis*-Cr(O_6), *cis*-Cr(O_7) and *cis*-Cr(O_8)]/ Li^+ , Na^+ , K^+ and Rb^+] are given in Fig. 4, whereas the most important

parameters for these complexes, which were optimized by performing DFT at the B3LYP/6-31 G(d) and LANL2DZ level, are given in Table 2.

Upon inspecting Figs. 3 and 4 and Tables 1 and 2, it is clear that the distances between the 4 and 4' positions of the

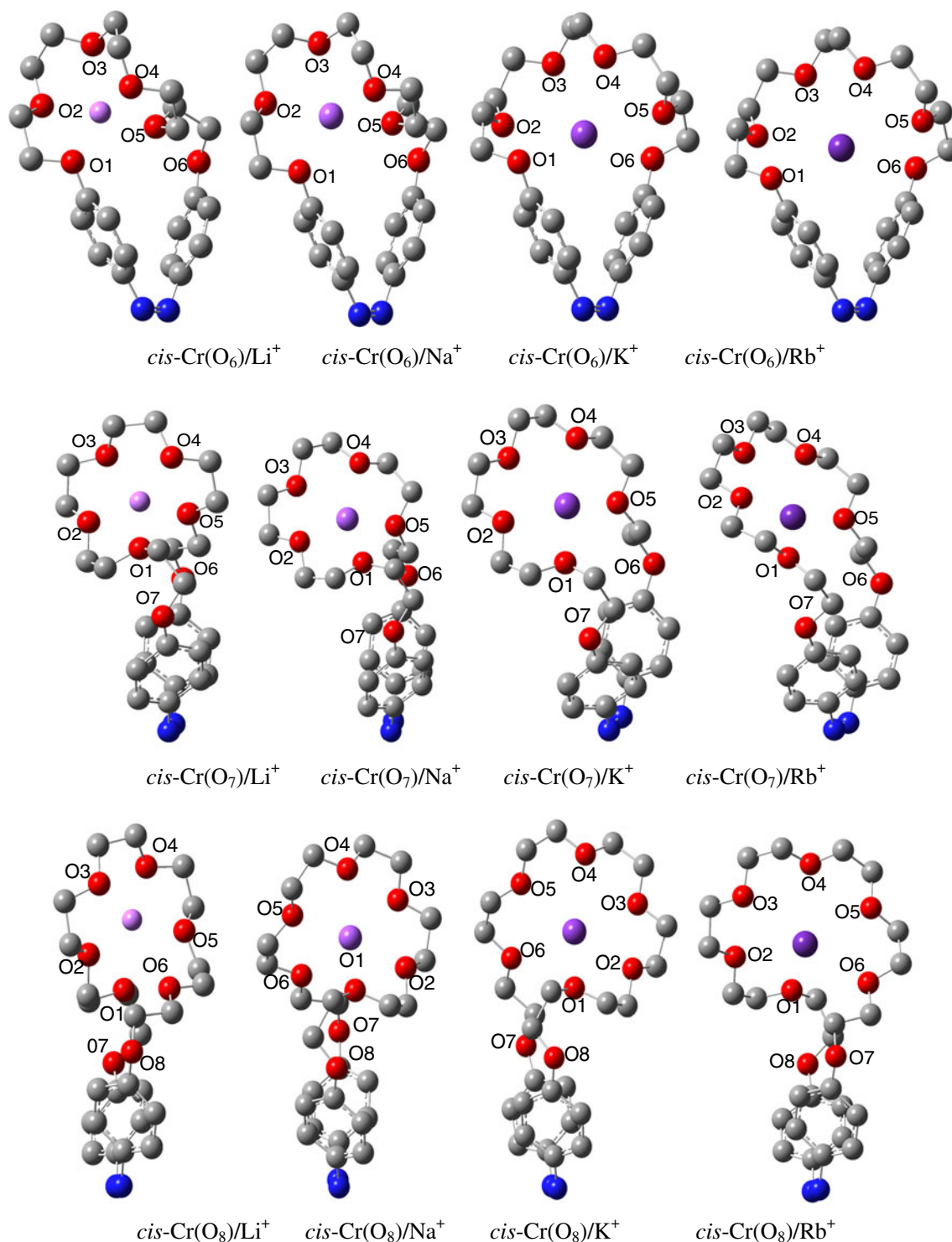


Fig. 4 The optimized structures of the *cis* isomers of azobenzene crown ethers (*cis* ACEs) complexed with Li^+ , Na^+ , K^+ and Rb^+ metal cations, obtained at the B3LYP/6-31 G(d) and LANL2DZ level of theory

Table 2 Selected parameters for the ACE/M⁺ (Li⁺, Na⁺, K⁺ and Rb⁺) complexes optimized at the B3LYP/6-31 G(d) and LANL2DZ level (distances in Å, dihedral angles in degrees °)

Parameter	<i>cis</i> -Cr(O ₆)				<i>cis</i> -Cr(O ₇)				<i>cis</i> -Cr(O ₈)			
	Li ⁺	Na ⁺	K ⁺	Rb ⁺	Li ⁺	Na ⁺	K ⁺	Rb ⁺	Li ⁺	Na ⁺	K ⁺	Rb ⁺
<i>r</i> ₁	2.155	2.405	3.370	3.709	2.105	2.413	2.796	3.137	3.290	2.675	2.934	3.065
<i>r</i> ₂	2.096	2.431	2.909	3.125	2.044	2.424	2.777	2.983	2.009	2.380	2.752	2.969
<i>r</i> ₃	2.343	2.486	2.864	3.027	2.193	2.470	2.776	3.015	2.055	2.511	2.843	3.012
<i>r</i> ₄	2.078	2.387	2.864	3.027	1.989	2.372	2.753	3.020	2.066	2.518	2.841	3.006
<i>r</i> ₅	2.094	2.457	2.907	3.125	2.226	2.439	2.813	3.097	2.016	2.363	2.781	2.983
<i>r</i> ₆	4.965	5.115	3.357	3.709	–	–	–	–	3.192	2.585	2.935	3.131
<i>d</i> _{C–C}	5.027	5.307	5.365	5.814	5.663	5.530	5.870	5.915	6.391	5.606	5.391	5.689
∠CNNC	7.0	7.7	7.5	8.6	8.3	7.3	8.4	8.5	8.5	8.1	6.7	8.4
∠NNCC	64.3	62.9	65.6	64.5	59.5	57.7	54.8	52.7	55.1	61.0	57.7	58.4

r: bond lengths (Å) between O atoms and alkali-metal cations

*d*_{C–C}: distances (Å) between the 4 and 4' positions of azobenzene

azobenzene crown ether loops all change greatly when the free ligand *cis* ACE [*cis*-Cr(O₆), *cis*-Cr(O₇) and *cis*-Cr(O₈)] coordinates with the alkali cations (Li⁺, Na⁺, K⁺ and Rb⁺). It can be assumed that the decreased distances *d*_{C–C} are due to the inductive effect arising from the O...M⁺ interactions. The smaller the number of methylene groups between the 4 and 4' positions of azobenzene, the stronger the restrictions on the crown-like ring. In addition, not all of the oxygen atoms can contribute to the formation of a crown-like ring in the *cis* isomers. The oxygen atoms in the crown loops do not all interact with M⁺ (Li⁺, Na⁺, K⁺ and Rb⁺) because they are too far away.

When the crown ether loop of *cis*-Cr(O₆) coordinates with an alkali-metal cation, the structural features of the dihedral angle ∠CNNC of the complex change significantly and show different properties to those of the metal-free *cis*-Cr(O₆) (11.1°). Upon inspecting Fig. 4 and Table 2, it is clear that in the complexes *cis*-Cr(O₆)/Li⁺ and *cis*-Cr(O₆)/Na⁺, the interatomic distances between Li⁺ and (O₁–O₆) are 2.155, 2.096, 2.343, 2.078, 2.094, and 4.965 Å; those between Na⁺ and (O₁–O₆) are 2.405, 2.431, 2.486, 2.387, 2.457, and 5.115 Å, respectively. It is clear that *r*₆ is 4.965 Å for *cis*-Cr(O₆)/Li⁺ and 5.115 Å for *cis*-Cr(O₆)/Na⁺. The bond length *r*₆ is much larger than the others in each of the complexes. The oxygen O₆ in the crown loop of *cis*-Cr(O₆) shows only weak interactions with Li⁺ and Na⁺ because it is too far away from them. The optimized structure of *cis*-Cr(O₆) with and without the cations K⁺ and Rb⁺ presents only small changes. This result can be attributed to the small size of the crown-like cavity but the big cation diameters of K⁺ and Rb⁺.

Turning our attention to the structures of the complexes *cis*-Cr(O₇)/M⁺ (Li⁺, Na⁺, K⁺ and Rb⁺), a polyoxyethylene loop is formed that is analogous to 15-crown-5 [40–42].

Not all of the donor oxygen atoms in *cis*-Cr(O₇) interact with metal cations. In the complexes, the average coordination bond lengths of the metal cations are 2.111, 2.424, 2.783, and 3.050 Å, respectively. The bond lengths for *cis*-Cr(O₇)/Li⁺ and Na⁺ are smaller than those for *cis*-Cr(O₇)/K⁺ and Rb⁺; in other words, there are stronger metal–oxygen interactions between the ligand *cis*-Cr(O₇) and Li⁺ and Na⁺ than K⁺ and Rb⁺. Li⁺ gives a shorter bond length with the donor O than Na⁺ does in these complexes. However, the structures of *cis*-Cr(O₇)/M⁺ (Li⁺ and Na⁺) from Fig. 4 indicate that Na⁺ improves the planarity of the oxygens compared to Li⁺. Therefore, Na⁺ fits with the crown-like ring better than Li⁺. *Cis*-Cr(O₇) cannot bind well with the large alkali metal cation Rb⁺, as can be seen from the structures in Fig. 4.

Based on the optimized structures of the complexes formed by the alkali cations Li⁺, Na⁺, K⁺ and Rb⁺ and the ligand *cis*-Cr(O₈), a 18-crown-6 crown-like ring is produced when *cis*-Cr(O₈) coordinates with metal cations. There are six donor oxygens that are mainly involved in the O...M⁺ interactions. *Cis*-Cr(O₈) can bind with both small and large alkali-metal cations, as can be seen from the structures shown in Fig. 4. Li⁺ is too small to coordinate with all six oxygen atoms in the crown-like ring of *cis*-Cr(O₈). The coordination bond lengths shown in Table 2 for the complex *cis*-Cr(O₈)/Li⁺ are 3.290, 2.009, 2.055, 2.066, 2.016 and 3.192 Å, respectively. It is clear that *r*₁ and *r*₆ are all much larger than the other bond lengths. The optimized structure shown in Fig. 4 indicates that Li⁺ is drawn to one side of the crown-like ring. The bond lengths indicate that the best match for the crown-like loop in the ligand *cis*-Cr(O₈) is Na⁺ according to the lock-and-key complementarity rule [43]. However, crown-6 ethers are known to prefer K⁺ according to experimental results [44–47]. The most

plausible reason for this difference between experiment and theory is that the calculations do not include the effect of the solvent, and Na⁺ is even more strongly solvated than K⁺ [48]; the calculations were performed for isolated molecules in the gas phase, but the experiments were done in aqueous solution.

Natural bond orbital analysis

For each donor NBO (*i*) and acceptor NBO (*j*), the stabilization energy (*E*₂) associated with *i*→*j* delocalization is explicitly estimated using the following equation [49–52]:

$$E_2 = \Delta E_{ij} = q_i \frac{F^2(i,j)}{\varepsilon_j - \varepsilon_i}, \tag{3}$$

where *q_i* is the *i*th donor orbital occupancy, ε_i and ε_j are the diagonal elements (orbital energies), and *F* (*i*, *j*) are the off-diagonal elements, respectively, associated with the NBO Fock matrix.

The results of second-order perturbation theory analysis of the Fock matrix for *cis*-ACE/M⁺ [*cis*-Cr(O₆), *cis*-Cr(O₇) and *cis*-Cr(O₈)/Li⁺, Na⁺, K⁺ and Rb⁺], obtained by NBO analysis, are summarized in Table 3. The interaction energies *E*₂ of the host–guest molecules *cis*-ACE/M⁺ are mainly dependent on the lone-pair electrons of O atoms of the crown ether and the LP* orbitals of the alkali-metal cation (Li⁺, Na⁺, K⁺ and Rb⁺); the N atoms in the azobenzene part do not appear to be as important.

For the complexes *cis*-Cr(O₆)/M⁺, the strong donor–acceptor interactions for *cis*-Cr(O₆)/Li⁺ between the lone-pair electrons of the electron-donating oxygens O₁, O₂, O₃, O₄, and O₅ and the LP* orbital of Li⁺ have stabilization energies of 3.05, 4.76, 4.47, 3.81, and 4.41 kcal mol^{−1}, respectively, which are much bigger than the corresponding energies *E*₂ for the complex *cis*-Cr(O₆)/Na⁺ (1.28, 3.72, 3.14, 2.92, and 3.32 kcal mol^{−1}). Obviously, one of the electron-donating oxygens, O₆, in the ligand *cis*-Cr(O₆) is not considered to interact with Li⁺ and Na⁺, and the *E*₂ data show a poor distribution. However, the interaction stabi-

Table 3 Results of second-order perturbation theory analysis of the Fock matrix for *cis*-ACE/M⁺ [*cis*-Cr(O₆), *cis*-Cr(O₇) and *cis*-Cr(O₈)/Li⁺, Na⁺, K⁺ and Rb⁺] within the NBO basis

<i>cis</i> -Cr(O ₆)		<i>cis</i> -Cr(O ₇)		<i>cis</i> -Cr(O ₈)	
Donor NBO(<i>i</i>) →Acceptor NBO (<i>j</i>)	<i>E</i> ₂ (kcal/mol)	Donor NBO(<i>i</i>) →Acceptor NBO (<i>j</i>)	<i>E</i> ₂ (kcal/mol)	Donor NBO(<i>i</i>) →Acceptor NBO (<i>j</i>)	<i>E</i> ₂ (kcal/mol)
LP O ₁ →LP* Li	3.05	LP O ₁ →LP* Li	4.91	LP O ₁ →LP* Li	3.10
LP O ₂ →LP* Li	4.47	LP O ₂ →LP* Li	4.52	LP O ₂ →LP* Li	3.54
LP O ₃ →LP* Li	4.41	LP O ₃ →LP* Li	4.48	LP O ₃ →LP* Li	3.29
LP O ₄ →LP* Li	3.81	LP O ₄ →LP* Li	4.01	LP O ₄ →LP* Li	3.30
LP O ₅ →LP* Li	4.76	LP O ₅ →LP* Li	5.01	LP O ₅ →LP* Li	3.51
				LP O ₆ →LP* Li	2.33
LP O ₁ →LP* Na	1.28	LP O ₁ →LP* Na	3.32	LP O ₁ →LP* Na	4.43
LP O ₂ →LP* Na	3.14	LP O ₂ →LP* Na	2.77	LP O ₂ →LP* Na	2.83
LP O ₃ →LP* Na	3.32	LP O ₃ →LP* Na	2.81	LP O ₃ →LP* Na	3.53
LP O ₄ →LP* Na	2.92	LP O ₄ →LP* Na	2.65	LP O ₄ →LP* Na	3.62
LP O ₅ →LP* Na	3.72	LP O ₅ →LP* Na	3.35	LP O ₅ →LP* Na	2.82
				LP O ₆ →LP* Na	4.18
LP O ₁ →LP* K	1.12	LP O ₁ →LP* K	2.45	LP O ₁ →LP* K	2.95
LP O ₂ →LP* K	2.36	LP O ₂ →LP* K	2.16	LP O ₂ →LP* K	2.38
LP O ₃ →LP* K	2.99	LP O ₃ →LP* K	2.01	LP O ₃ →LP* K	2.69
LP O ₄ →LP* K	2.99	LP O ₄ →LP* K	1.88	LP O ₄ →LP* K	2.72
LP O ₅ →LP* K	2.36	LP O ₅ →LP* K	2.55	LP O ₅ →LP* K	2.51
LP O ₆ →LP* K	1.11			LP O ₆ →LP* K	2.99
LP O ₁ →LP* Rb	0.57	LP O ₁ →LP* Rb	1.15	LP O ₁ →LP* Rb	1.62
LP O ₂ →LP* Rb	1.21	LP O ₂ →LP* Rb	1.03	LP O ₂ →LP* Rb	1.32
LP O ₃ →LP* Rb	1.64	LP O ₃ →LP* Rb	1.04	LP O ₃ →LP* Rb	1.51
LP O ₄ →LP* Rb	1.64	LP O ₄ →LP* Rb	1.02	LP O ₄ →LP* Rb	1.54
LP O ₅ →LP* Rb	1.21	LP O ₅ →LP* Rb	1.05	LP O ₅ →LP* Rb	1.34
LP O ₆ →LP* Rb	0.57			LP O ₆ →LP* Rb	1.76

zation energy E_2 between an electron-donating oxygen and K^+ or Rb^+ is weaker than that for $cis-Cr(O_6)/M^+$ (Li^+ and Na^+). Also, six electron-donating oxygens in $cis-Cr(O_6)$ all interact with the cations K^+ and Rb^+ . This result can be attributed to the small size of the crown-like cavity and the big cation diameters of K^+ and Rb^+ .

In $cis-Cr(O_7)/M^+$ (Li^+ , Na^+ , K^+ and Rb^+) complexes, the stronger donor–acceptor interactions mainly derive from the lone-pair electrons of the five electron-donating oxygens (O_1 , O_2 , O_3 , O_4 , O_5) and the LP* orbital of the alkali-metal cation M^+ (Li^+ , Na^+ , K^+ or Rb^+), and the interaction phenomenon is analogous to 15-crown-5/ M^+ . The stabilization energies E_2 for the complexes $cis-Cr(O_7)/Li^+$ and Na^+ are larger than those of the complexes $cis-Cr(O_7)/K^+$ and Rb^+ . For the complex $cis-Cr(O_7)/Li^+$, the data distribution for the stabilization energy E_2 of $cis-Cr(O_7)/Na^+$ is better than that of $cis-Cr(O_7)/Li^+$.

In the $cis-Cr(O_8)/M^+$ complexes, the strongest donor–acceptor interactions mainly come from the lone-pair electrons of the six electron-donating oxygens (O_1 , O_2 , O_3 , O_4 , O_5 , O_6) and the LP* orbital of the alkali-metal cation M^+ (Li^+ , Na^+ , K^+ or Rb^+), and the interaction phenomenon is analogous to 18-crown-6/ M^+ . The stabilization energy E_2 for $O_6 \dots Li^+$ in the $cis-Cr(O_8)/Li^+$ complex is $2.33 \text{ kcal mol}^{-1}$, which is much smaller than those of the other five oxygens (O_1 : 3.10 , O_2 : 3.54 , O_3 : 3.29 , O_4 : 3.30 , O_5 : $3.51 \text{ kcal mol}^{-1}$). The data distributions of the stabilization energies E_2 for the complexes $cis-Cr(O_8)/K^+$ and $cis-Cr(O_8)/Rb^+$ are better than those of the complexes $cis-Cr(O_8)/Li^+$ and $cis-Cr(O_8)/Na^+$. In the complex $cis-Cr(O_8)/K^+$, the strong donor–acceptor interactions between the lone-pair electrons of the electron-donating oxygens O_1 – O_6 and the LP* orbital of Li^+ have stabilization energies of 2.95 , 2.38 , 2.69 , 2.72 , 2.51 and $2.99 \text{ kcal mol}^{-1}$, respectively, which are much bigger than the corresponding stabilization energies E_2 of the complex

$cis-Cr(O_8)/Rb^+$ (1.62 , 1.32 , 1.51 , 1.54 , 1.34 and $1.76 \text{ kcal mol}^{-1}$).

Binding energies and stabilities

The calculated binding energies (ΔE^b), enthalpies (ΔH^b) and Gibbs free energies (ΔG^b) (298 K) of the ACE/ M^+ complexes [$cis-Cr(O_6)$, $cis-Cr(O_7)$, and $cis-Cr(O_8)/Li^+$, Na^+ , K^+ and Rb^+], based on reaction (1) at the B3LYP/6-31 G(d) and LANL2DZ level in the gas phase are listed in Table 4. When performing such a study, it is important to consider the large basis set superposition error (BSSE), which in most cases leads to overestimated interaction energies [53, 54]. One of the most commonly used methods of correcting for the BSSE is the counterpoise (CP) method [55]. Thus, the binding energies were corrected for the undesirable effects of the BSSE using the CP method at the B3LYP/6-31 G (d) level with relaxed fragment geometries.

Table 4 shows that the gas-phase binding energies (ΔE^b), binding enthalpies (ΔH^b) and Gibbs free energies (ΔG^b) at 298 K decrease for the three different free ligands $cis-Cr(O_6)$, $cis-Cr(O_7)$ and $cis-Cr(O_8)$ as the size of the alkali cation increases, in other words: $\Delta E_{ACEs/Li^+} > \Delta E_{ACEs/Na^+} > \Delta E_{ACEs/K^+} > \Delta E_{ACEs/Rb^+}$.

For $cis-Cr(O_6)/M^+$ (Li^+ , Na^+ , K^+ and Rb^+), the crown-like cavity ring of $cis-Cr(O_6)$ must undergo considerable folding/twisting to bring the binding sites in close proximity to the small cations Li^+ and Na^+ . These distortions enhance the host–guest intramolecular interactions. Although the backbone of the complex suffers much distortion and displays poor structural symmetry, the calculations are performed for isolated molecules in the gas phase (i.e., they do not include the intramolecular interactions of the studied complexes); therefore, the thermal energies of $cis-Cr(O_6)/Li^+$ and $cis-Cr(O_6)/Na^+$ are larger than those of $cis-Cr(O_6)/K^+$ and $cis-Cr(O_6)/Rb^+$. However, because the

Table 4 Calculated binding energies ΔE^b (kcal mol^{-1}), binding enthalpies ΔH^b (kcal mol^{-1}), and Gibbs free energies ΔG^b (kcal mol^{-1}) in the gas phase for the complexes at 298 K

Ligands	Metal cations	ΔE^b	E_{BSSE}	ΔE^b_{BSSE}	ΔH^b	ΔG^b
<i>Cis-Cr(O₆)</i>	Li^+	−92.9	28.9	−64.0	−93.5	−77.8
	Na^+	−72.2	22.0	−50.2	−72.8	−60.9
	K^+	−51.5	17.6	−33.9	−52.1	−41.4
	Rb^+	−23.2	13.8	−9.4	−33.9	−13.2
<i>Cis-Cr(O₇)</i>	Li^+	−105.4	18.2	−87.2	−106.0	−94.8
	Na^+	−83.5	18.2	−65.3	−84.1	−73.4
	K^+	−60.9	6.9	−54.0	−61.5	−51.5
	Rb^+	−30.7	8.2	−22.5	−31.4	−21.3
<i>Cis-Cr(O₈)</i>	Li^+	−109.8	15.1	−94.7	−110.4	−97.9
	Na^+	−90.4	14.4	−66.0	−91.0	−79.1
	K^+	−67.1	8.2	−58.9	−67.8	−58.45
	Rb^+	−33.3	6.9	−26.4	−33.9	−23.2

metal cations K^+ and Rb^+ are large but the crown-like cavity is small, the thermal energies of $cis\text{-Cr}(\text{O}_6)/K^+$ and $cis\text{-Cr}(\text{O}_6)/Rb^+$ are also small. For $cis\text{-Cr}(\text{O}_7)/M^+$ and $cis\text{-Cr}(\text{O}_8)/M^+$ (Li^+ , Na^+ , K^+ and Rb^+), the complexes suffer much distortion, display poor structural symmetry and so exhibit the biggest thermal energies for the cation Li^+ . Thus, the relationship between the cavity size of the crown ether and the cation diameter plays an important role in determining the thermal energies of complexes during coordination.

For $cis\text{-Cr}(\text{O}_6)/M^+$, $cis\text{-Cr}(\text{O}_7)/M^+$ and $cis\text{-Cr}(\text{O}_8)/M^+$ (Li^+ , Na^+ , K^+ and Rb^+), the different alkali cations show different trends. If we consider Na^+ , the thermal energy shows the relation $cis\text{-Cr}(\text{O}_6)/Na^+ < cis\text{-Cr}(\text{O}_7)/Na^+ < cis\text{-Cr}(\text{O}_8)/Na^+$, while the binding energies of the *cis* ACEs with Na^+ [$cis\text{-Cr}(\text{O}_6) < cis\text{-Cr}(\text{O}_7) < cis\text{-Cr}(\text{O}_8)$] indicate that the affinities of the *cis* ACEs for Na^+ increase as the number of $-\text{CH}_2-\text{O}-\text{CH}_2-$ units increases, enlarging the crown-like loop. As the loop enlarges, the rigidity of the crown ether is reduced, so it becomes easier for the ACE to bind with metal cations.

Absorption spectra

The absorption spectra of the *cis* isomers of azobenzene and ACEs [$cis\text{-Cr}(\text{O}_6)$, $cis\text{-Cr}(\text{O}_7)$ and $cis\text{-Cr}(\text{O}_8)$] were investigated by time-dependent density functional theory (TDDFT) with the 6-31 G(d) basis set. The calculated excitation energies (E_g), wavelengths of peak absorption (λ_{abs}) and the oscillator strengths (f) of all compounds in their optimized ground-state geometries are summarized in Table 5.

The absorption spectrum of *cis*-azobenzene shows two distinct bands: a strong $\pi \rightarrow \pi^*$ ($S_0 \rightarrow S_2$) absorption band peaking at about 270 nm and a much weaker $n \rightarrow \pi^*$ ($S_0 \rightarrow S_1$) band with a peak at around 470 nm. The results are in a good agreement with some of the previous studies [56, 57].

Obviously, the values of the parameters of the *cis* isomers of the ACEs are all different from those of *cis*-azobenzene. The spectra of the *cis* ACEs [$cis\text{-Cr}(\text{O}_6)$, $cis\text{-Cr}$

(O_7) and $cis\text{-Cr}(\text{O}_8)$] contain broad $\pi \rightarrow \pi^*$ ($S_0 \rightarrow S_2$) absorption bands with a characteristic peak at 310–340 nm in the near-UV region, and their oscillator strengths are much more intense. Weaker bands in the visible region (peak wavelengths: 480–490 nm) and lower oscillator strengths due to the $n \rightarrow \pi^*$ ($S_0 \rightarrow S_1$) transitions are also observed. The spectra of the *cis* ACEs present significant redshifts in comparison to the spectrum of unsubstituted *cis*-azobenzene. This result indicates that the size of the crown has a distinct influence on the absorption spectra of the *cis* ACEs.

In ref. [15], the experimental results indicated that both $\text{Cr}(\text{O}_7)$ and $\text{Cr}(\text{O}_8)$ give high yields of the *cis* isomer at about 360 nm, which is a little different from the peak wavelengths obtained in our calculations. This difference between the theoretical calculations and the experimental results arises because the calculations performed in this paper relate to the gas phase at 303 K, while the experiments were performed in liquids at 298 K. However, this theoretical study is still useful for predicting reactions and gauging trends.

Conclusions and perspectives

The ground-state electronic structures of azobenzene crown ethers [ACEs: $\text{Cr}(\text{O}_6)$, $\text{Cr}(\text{O}_7)$ and $\text{Cr}(\text{O}_8)$] and complexes of their *cis* isomers with the alkali-metal cations Li^+ , Na^+ , K^+ and Rb^+ were obtained by DFT methods at the B3LYP/6-31 G(d) level and LANL2DZ. The significant structural differences between the optimized *trans* and *cis* isomers of the ACEs indicate that the preorganization of the *trans* ACEs is poor and in an “off” state, while it is enhanced for the *cis* isomers and in an “on” state in relation to coordinating with alkali metal cations. These “molecular machines” can therefore be used as “on/off” switches as they can switch between different molecular structures and parameters. The *cis* isomers showed spherical recognition patterns in the binding of alkali-metal cations. In NBO

Table 5 Electronic transition data obtained by TDDFT for a family of azobenzene-type crown ethers [$cis\text{-Cr}(\text{O}_6)$, $cis\text{-Cr}(\text{O}_7)$ and $cis\text{-Cr}(\text{O}_8)$]

Ligands	Electronic transitions	TD-B3LYP/6-31 G(d) // LANL2DZ		
		E_g (eV)	Wavelength (nm)	f
<i>Cis</i> -azobenzene	$S_0 \rightarrow S_1$	2.60	465.30	0.0105
	$S_0 \rightarrow S_2$	4.65	266.79	0.5858
<i>Cis</i> -Cr(O_6)	$S_0 \rightarrow S_1$	2.49	497.71	0.0708
	$S_0 \rightarrow S_2$	3.74	331.24	0.1951
<i>Cis</i> -Cr(O_7)	$S_0 \rightarrow S_1$	2.57	482.33	0.0542
	$S_0 \rightarrow S_2$	3.97	312.69	0.1289
<i>Cis</i> -Cr(O_8)	$S_0 \rightarrow S_1$	2.53	490.81	0.0582
	$S_0 \rightarrow S_2$	3.78	328.25	0.1370

analysis, the main intermolecular charge-transfer interactions were between the LP* orbitals of the metal cations and the lone-pair electrons of the electron-donating O atoms of the *cis* ACEs, but not all of the donor oxygen atoms in the *cis* ACEs interact with metal cations. The interaction pattern of *cis*-CrO₇ with metal cations (M⁺) is analogous to 15-crown-5/M⁺, while that for *cis*-Cr(O₈)/M⁺ is analogous to 18-crown-6/M⁺. A time-dependent density functional theory (TDDFT) study of the *cis* ACEs afforded their absorption spectral parameters. The results of the TDDFT study indicate that the *cis* isomers have broad $\pi \rightarrow \pi^*$ (S₀ → S₂) absorption bands but weaker $n \rightarrow \pi^*$ (S₀ → S₁) bands, and good agreement between the theoretical and experimental values was seen.

Acknowledgments The author wish to acknowledge the financial support from the Scientific Research Fund of Hunan Provincial Education Department (no. 09A091).

References

- Kyba EP, Helgeson RC, Madan K, Gokel GW, Tarnowski TL, Moore SS, Cram DJ (1977) *J Am Chem Soc* 99:2564–2571
- Kovbasyuk L, Krämer R (2004) *Chem Rev* 104:3161–3187
- Pedersen CJ (1967) *J Am Chem Soc* 89:7017–7036
- Gokel GW (1991) *Crown ethers and cryptands*. Royal Society of Chemistry, Cambridge
- More MB, Glendening ED, Ray D, Feller D, Armentrout PB (1996) *J Phys Chem* 100:1605–1614
- Feringa BL (2001) *Molecular switches*. Wiley-VCH, Weinheim, p 454
- Balzani V, Scandola F (1991) *Supramolecular photochemistry*. Ellis Horwood, New York, pp 199–215
- Liu ZF, Hashimoto K, Fujishima A (1990) *Nature* 347:658–660
- Ikeda T, Tsutsumi O (1995) *Science* 268:1873–1875
- Sekkat Z, Dumont M (1992) *Appl Phys B* 54:486–489
- Hugel T, Holland NB, Cattani A, Moroder L, Seitz M, Gaub HE (2002) *Science* 296:1103–1106
- Muraoka T, Kinbara K, Kobayashi Y, Aida T (2003) *J Am Chem Soc* 125:5612–5613
- Zhang C, Du MH, Cheng HP, Zhang XG, Roitberg AE, Krause JL (2004) *Phys Rev Lett* 92:158301(1–4)
- Halabieh HE, Mermut O, Barrett CJ (2004) *Pure Appl Chem* 76:1445–1465
- Shinkai S, Nakaji T, Nishida Y, Ogawa T, Manabe O (1980) *J Am Chem Soc* 102:5860–5865
- Tahara R, Morozumi T, Nakamura H, Shimomura M (1997) *J Phys Chem B* 101:7736–7743
- Shinkai S, Minami T, Kusano Y, Manabe O (1983) *J Am Chem Soc* 105:1851–1856
- Becke AD (1993) *J Chem Phys* 98:5648–5652
- Lee C, Yang W, Parr RG (1988) *Phys Rev B* 37:785–789
- Korth HG, De Heer MI, Mulder P (2002) *J Phys Chem A* 106:8779–8789
- Johnson BG, Gill PMW, Pople JA (1993) *J Chem Phys* 98:5612–5626
- Chowdhury PK (2003) *J Phys Chem A* 107:5692–5696
- Chis V (2004) *Chem Phys* 300:1–11
- Asensio A, Kobko N, Dannenberg JJ (2003) *J Phys Chem A* 107:6441–6443
- Müller A, Losada M, Leutwyler S (2004) *J Phys Chem A* 108:157–165
- Goncalves NS, Cristiano R, Pizzolatti MG, da Silva Miranda F (2005) *J Mol Struct* 733:53–61
- Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JAJr, 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 TM, Al-Laham A, Peng CY, Nanayakkara A, Challacombe MP, Gill MW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2003) *Gaussian 2003W*, revision B.05. Gaussian Inc., Pittsburgh
- Reed AE, Curtiss LA, Weinhold F (1988) *Chem Rev* 88:899–926
- Reed AE, Weinhold F (1983) *J Chem Phys* 78:4066–4073
- Foster JP, Weinhold F (1980) *J Am Chem Soc* 102:7211–7218
- Reed AE, Weinstock RB, Weinhold F (1985) *J Chem Phys* 83:735–746
- Schulze FW, Petrick HJ, Cammenga HK, Klinge H (1977) *Z Phys Chem Neue Fol* 107:4743
- Cattaneo P, Persico M (1999) *Phys Chem Chem Phys* 1:4739–4743
- Ishikawa T, Noro T, Shoda TJ (2001) *Chem Phys* 115:7503–7512
- Tiago ML, Ismail-Beigi S, Louie SG (2005) *J Chem Phys* 122:094311(1–7)
- Cembran A, Bernardi F, Garavelli L, Gagliardi L, Orlandi G (2004) *J Am Chem Soc* 126:3234–3243
- Biswas N, Umpathy S (1997) *J Phys Chem* 107:7849–7858
- Mostad A, Romming C (1971) *Acta Chem Scand* 25:3561–3568
- Fliegl H, Kohn A, Hattig C, Ahlrichs R (2003) *J Am Chem Soc* 125:9821–9827
- Hopkins HP Jr, Norman AB (1980) *J Phys Chem* 84:309–314
- Smetana AJ, Popov AI (1980) *J Solution Chem* 9:183–196
- Lamb JD, Izatt RM, Swain CS, Christensen JJ (1980) *J Am Chem Soc* 102:475–479
- Ouchi M, Inoue Y, Kanzaki T, Hakushi T (1984) *J Org Chem* 49:1408–1412
- Pedersen C (1970) *J Am Chem Soc* 92:391–394
- Liu Y, Lu TB, Tan MY, Hakushi T, Inoue Y (1993) *J Phys Chem* 97:4548–4551
- Ouchi M, Inoue Y (1985) *Bull Chem Soc Jpn* 58:525–530
- Ouchi M, Inoue Y, Kanzaki T (1984) *Bull Chem Soc Jpn* 57:887–888
- Hill SE, Feller D (2000) *Int J Mass Spectrom* 201:41–58
- Adamovic I, Gordon MS (2005) *J Phys Chem A* 109:1629–1636
- Mo Y, Wu W, Song L, Lin M, Zhang Q, Gao J (2004) *Angew Chem Int Ed* 43:1986–1990
- Mo Y, Jiao H, Schleyer PvR (2004) *J Org Chem* 69:3493–3499
- Mo Y, Schleyer PvR, Wu W, Lin M, Zhang Q, Gao J (2003) *J Phys Chem A* 107:10011–10018
- Cramer CJ (2002) *Essentials of computational chemistry: theories and models*, 2nd edn. Wiley, New York
- Kim KS, Tarakeshwar P, Lee JY (2000) *Chem Rev* 100:4145–4186
- Boys SF, Bernardi F (1970) *Mol Phys* 19:553–566
- Crecca CR, Roitberg AE (2006) *J Phys Chem A* 110:8188–8203
- Nägele T, Hoche R, Zinth W, Wachtveitl J (1997) *Chem Phys Lett* 272:489–495

Molecular modeling studies on phosphonic acid-containing thiazole derivatives: design for fructose-1,6-bisphosphatase inhibitors

Ping Lan · Zhi-Wei Wu · Wan-Na Chen ·
Ping-Hua Sun · Wei-Min Chen

Received: 18 February 2011 / Accepted: 18 May 2011 / Published online: 5 June 2011
© Springer-Verlag 2011

Abstract Presently, an in silico modeling was carried out on a series of 63 phosphonic acid-containing thiazole derivatives as fructose-1,6-bisphosphatase (FBPase) inhibitors using CoMFA/CoMSIA and molecular docking methods. The CoMFA and CoMSIA models using 51 molecules in the training set gave r_{cv}^2 values of 0.675 and 0.619, r^2 values of 0.985 and 0.979, respectively. The systemic external validation indicated that our CoMFA and CoMSIA models possessed high predictive powers with r_o^2 values of 0.995 and 0.994, $r_{m(test)}^2$ values of 0.887 and 0.860, respectively. The 3D contour maps of the CoMFA and CoMSIA provided smooth and interpretable explanation of the structure-activity relationship for the inhibitors. Molecular docking studies revealed that a phosphonic group was essential for binding to the AMP binding site, and some key features were also identified. The analyses of the 3D contour plots and molecular docking results permitted interesting conclusions about the effects of different substituent groups at different positions of the common scaffold, which might guide the design of novel FBPase inhibitors with higher activity and bioavailability. A set of 60 new analogues were designed by utilizing the results revealed in the present study, and were predicted with significantly improved potencies in the developed models. The findings can be quite useful to aid the designing of new fructose-1,6-bisphosphatase inhibitors with improved biological response.

Keywords CoMFA · CoMSIA · Docking · Fructose-1,6-bisphosphatase

Introduction

In order to develop new therapeutics for the treatment of type-2 diabetes, many investigations have been carried out especially through the use of small-molecule compounds binding to various enzyme targets [1]. Significant effort has targeted fructose-1,6-bisphosphatase (FBPase, EC 3.1.3.11), a key regulatory enzyme of hepatic gluconeogenesis (GNG) pathway which catalyzes the irreversible reaction of hydrolysis of fructose-1,6-bisphosphate to fructose-6-phosphate [2]. A large number of radioisotope studies and several experiments have demonstrated that the GNG can account for up to 100% of the glucose produced by the liver in the non-absorptive state, moreover, the GNG flux is excessive in type-2 diabetes patients [3]. As a rate-limiting and highly regulated enzyme in the GNG pathway, FBPase is an attractive approach in the development of new anti-diabetic pharmaceuticals [4–6].

As a tetramer of four identical polypeptide chains, FBPase exists as a dimer of dimers. Regulation of FBPase enzymatic activity involves changes in quaternary structure between the active (R) and inactive (T) conformational states [7]. This enzyme is subject to competitive substrate inhibition by fructose-2,6-bisphosphate and to allosteric inhibition by AMP. Without effectors the FBPase exists in the active R-quaternary structure, AMP binds to the allosteric site and inhibits the FBPase by shifting the enzyme from R to T conformation or stabilizing the T state [8, 9]. Therefore, efforts over the past two decades have focused on designing AMP mimics capable of retaining the key binding interactions of AMP with the allosteric binding

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1134-0) contains supplementary material, which is available to authorized users.

P. Lan · Z.-W. Wu · W.-N. Chen · P.-H. Sun · W.-M. Chen (✉)
Guangdong Province Key Laboratory of Pharmacodynamic
Constituents of TCM and New Drugs Research,
College of Pharmacy, Jinan University,
Guangzhou 510632, People's Republic of China
e-mail: twmchen@jnu.edu.cn

site of FBPase, moreover, forming additional interactions designed to enhance inhibitory potency and specificity [10, 11]. Despite numerous discovery programs over the past decades targeting the AMP binding site of FBPase, few AMP mimics have emerged with suitable potency and specificity [12]. Unlike most drug binding sites which are hydrophilic and highly dependent on hydrogen bond interactions, the FBPase relies strictly on the structural similarity of AMP. Many compounds were proved to be unsuitable AMP mimics due to the strict structural requirements and other reasons [13]. Efforts to develop nucleoside analogs that generate nucleoside monophosphates (NMPs) inside cells capable of functioning as potent and specific AMP mimics were also unsuccessful. Most of these nucleoside analogs were unable to generate high intracellular levels of the NMP because of poor initial phosphorylation; furthermore, the generated NMPs can be rapidly phosphorylated to the corresponding nucleoside triphosphate [14]. Thus, many attentions have been paid to the development of non-nucleoside AMP mimics.

By using structure-guided drug design strategies, a series of phosphonic acids that bind to the allosteric AMP binding site of FBPase and serve as non-nucleoside AMP mimics with high inhibitory potencies and specificities for FBPase were reported recently [10–13, 15, 16]. These AMP mimics including purine phosphonic acid [11, 12], benzimidazole phosphonic acid [10, 13] and thiazole phosphonic acid [16]. A set of potent lead compounds have also been identified (Fig. 1), these AMP mimics were more potent than AMP and exhibited high specificity for the AMP binding site on FBPase, especially the thiazole phosphonic acid, which exhibited excellent inhibitory potency as well as good oral bioavailability [16].

In our previous study [9], we carried out a systemic molecular modeling and docking research on [5-(4-amino-1*H*-benzoimidazol-2-yl)-furan-2-yl]-phosphonic acid derivatives that function as AMP mimetics with FBPase inhibitory activities. Nevertheless, these compound series showed poor oral bioavailability due to their high molecular weight

[16]. Replacement of the benzoimidazole ring system with benzoimidazole smaller 5-membered thiazole of lower molecular weight resulted in significantly improved FBPase inhibitory activity and oral bioavailability [16]. To investigate the structure-activity relationship and the interaction between these newly synthesized inhibitors and the FBPase, herein we reported 3D-QSAR (CoMFA/CoMSIA), molecular docking and molecular design studies on a total set of 63 thiazole phosphonic acid compounds. The main aims of the present study are: (i) establish reliable and valuable drug design computational methods to predict the activity of new designed molecules; (ii) explore the regions in space where interactive fields may influence the activity and identify the accurate structure-activity relationship of these inhibitors; (iii) investigate the interaction of these compounds and the AMP binding region; (iv) design novel compounds based on the results taken from current work and predict their potencies.

During the past decades, quantitative structure-activity relationship (QSAR) methods especially three-dimensional quantitative structure-activity relationship (3D-QSAR) approaches, have been successfully employed to assist the design of new drug candidates, ranging from enzyme inhibitors to receptor ligands [17]. Furthermore, they have been extensively applied in connection to medicinal chemistry research as well as proteomics, metabolomics, and bioinformatics [18–20]. 3D-QSAR methods including comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) are powerful and versatile tools to build and design an activity model for a given set of compounds in rational drug design and related applications. In CoMFA, the biological activity of molecules is correlated with their steric and electrostatic interaction energies. In CoMSIA, similarity indices are calculated at regularly placed grid points for these molecules. CoMSIA includes five molecular descriptors named steric, electrostatic, hydrophobic, hydrogen bond donor and acceptor fields [21]. We have also employed systemic internal and external validations to evaluate the true predictive power of the CoMFA and CoMSIA models. Molecular docking was applied to investigate the FBPase-inhibitor interactions.

Based on the good performances of the 3D-QSAR (CoMFA/CoMSIA) and docking experiments, the developed models can not only help in understanding the structure-activity relationship of these compounds but also serve as a useful guide for the design of new inhibitors with better activities. We have designed a number of novel phosphonic acid containing thiazole derivatives by utilizing the structure analysis results obtained from previous and present work, which exhibited excellent predicted activities as well as binding affinities in the established 3D-QSAR and docking models. Meanwhile, based on the excellent performance of

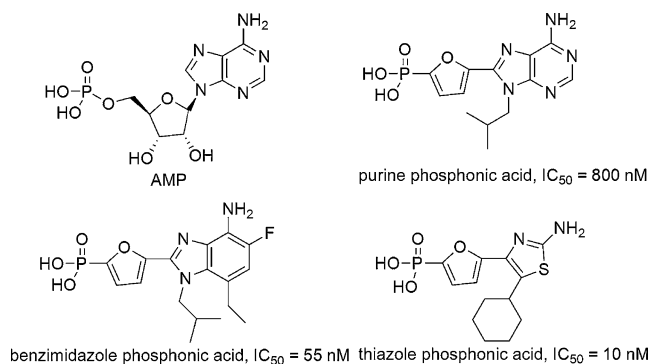


Fig. 1 Structures of AMP and novel AMP mimics

the external validation, the predicted potencies of these newly designed derivatives may be reliable.

Results and discussion

PLS analysis for CoMFA and CoMSIA models

A total set of 63 compounds were randomly segregated into training and test sets comprising 51 and 12 molecules, respectively. Structures and associated inhibitory activities were listed in Table 1 and Table 2. All of the structures were aligned into a lattice box by fitting with thiazole skeleton as a common structure. After careful selection, compound **23** displayed the highest pIC₅₀ value and was selected as a template. The aligned molecules were shown in Fig. 2.

The statistical parameters associated in CoMFA and CoMSIA models were listed in Table 3. In both CoMFA and CoMSIA methods, a cross-validated PLS analysis was performed using the six principal components, which have given the higher non-cross-validated correlation coefficient after the LOO procedure on the training set of the compounds in order to generate the corresponding CoMFA and CoMSIA contour maps.

The CoMFA model gave a good cross-validated correlation coefficient (r^2_{cv}) of 0.675 (> 0.6) which suggested that the model should be a useful tool for predicting the IC₅₀ values. A high non-cross-validated correlation coefficient (r^2) of 0.985 with a low standard error estimate (SEE) of 0.115 and excellent F value of 487.178 were obtained. Contributions of steric and electrostatic fields were 0.529 and 0.471, respectively. The excellent $r^2_{bootstrapping}$ (0.990 ± 0.004) and low $SEE_{bootstrapping}$ (0.091 ± 0.056) values indicated the robustness and statistical confidence of the generated CoMFA model. The actual and predicted pIC₅₀ values of the training set and test set by the CoMFA model were given in Table 2, and the graph of actual activity versus predicted pIC₅₀ of the training set and test set was illustrated in Fig. 3a.

The CoMSIA model incorporated steric (S), electrostatic (E), hydrophobic (H), hydrogen bond donor (D) and hydrogen bond acceptor (A) fields also gave a good cross-validated correlation coefficient (r^2_{cv}) of 0.619 (> 0.6) which indicated that the model should be a reliable tool for predicting the IC₅₀ values. A high non-cross-validated correlation coefficient (r^2) of 0.979 with a low standard error estimate (SEE) of 0.136 and excellent F value of 345.303 were obtained. Contributions of each field were 0.138 (S), 0.260 (E), 0.196 (H), 0.253 (D) and 0.153 (A), accordingly. The high $r^2_{bootstrapping}$ (0.987 ± 0.005) and low $SEE_{bootstrapping}$ (0.106 ± 0.066) values demonstrated the robustness and statistical confidence of the established CoMSIA model. The actual and predicted pIC₅₀ values of the training set and test set by the CoMSIA model were given

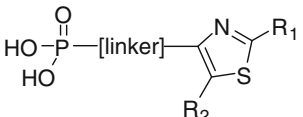
in Table 2, and the graph of actual activity versus predicted pIC₅₀ of the training set and test set was illustrated in Fig. 3b.

External validation analysis for CoMFA and CoMSIA models

The most important part of QSAR model development is the model validation. Most of the QSAR modeling methods employ the leave-one-out (LOO) cross-validation procedure which leads to the LOO correlation coefficient (r^2_{cv}). Normally, the r^2_{cv} is used as a criterion of both robustness and predictive ability of the model. In many cases, a high r^2_{cv} value (usually > 0.5) is considered as an indicator or even as the ultimate proof that the model is accurate or reliable [22]. Nevertheless, previous research revealed that validating a QSAR model only by a r^2_{cv} value is not enough and unacceptable. In fact, the high r^2_{cv} value does not imply automatically a high predictive power of the model. Even though the low value of r^2_{cv} can serve as an indicator of a low predictive ability, the opposite is not necessarily true [22]. In many cases, a model with high r^2_{cv} and r^2 values can be proved to be inaccurate. Moreover, although a model may exhibit a good predictive ability based on the statistics for the test set, it is not always sure that the model will perform well on a new set of data [23]. The only way to estimate the true predictive power of a model is to test it on an external validation. To evaluate the true predictive abilities of the established models, both the CoMFA and CoMSIA models were subjected to systemic external validation process, several statistics such as r^2_{pred} , $r^2_{m(test)}$, r_0^2 , R , a , b and k were employed. For the ideal model, the slope a is equal to 1, intercept b is equal to 0, and correlation coefficient R is equal to 1. 3D-QSAR models were considered acceptable if they satisfy all of the following conditions [22–24]:

$$r^2_{cv} > 0.5, r^2 > 0.6, [(r^2 - r_0^2)/r^2] < 0.1, 0.85 \leq k \leq 1.15 \text{ and } r^2_{m(test)} > 0.5.$$

The results of the external validation for both the CoMFA and CoMSIA models were shown in Table 4. The established CoMFA model using 12 molecules in the test set, gave a predictive correlation coefficient (r^2_{pred}) of 0.805, slope a value of 1.136 (close to 1), intercept b value of -0.995 (close to 0), an excellent $r^2_{m(test)}$ value of 0.887 (> 0.5) as well as high slope of regression lines through the origin (k) value of 0.995 ($0.85 \leq k \leq 1.15$), and the correlation coefficient (R) values of 0.940 (close to 1), the calculated $[(r^2 - r_0^2)/r^2]$ values of -0.010 (< 0.1) were obtained. These excellent external validation statistics indicated that the CoMFA model possessed a high accommodating capacity, and it may be reliable for being used to predict the activities of new derivatives.

Table 1 The structures and experimental IC₅₀ values of the training and test set molecules [16]


Compound No.	Substituent			IC ₅₀ (μM)
	linker	R ₁	R ₂	
1	2,5-furanyl	Me	<i>i</i> -Bu	0.1
2	2,5-furanyl	Et	<i>i</i> -Bu	0.4
3	2,5-furanyl	vinyl	<i>i</i> -Bu	1.2
4	2,5-furanyl	CH ₂ OH	<i>i</i> -Bu	0.22
5	2,5-furanyl	Cl	<i>i</i> -Bu	0.18
6	2,5-furanyl	SMe	<i>i</i> -Bu	0.89
7	2,5-furanyl	CN	<i>i</i> -Bu	2
8	2,5-furanyl	NHMe	<i>i</i> -Bu	1
9	2,5-furanyl	NHAc	<i>i</i> -Bu	10
10	2,5-furanyl	CONH ₂	<i>i</i> -Bu	2.75
11	2,5-furanyl	CSNH ₂	<i>i</i> -Bu	0.5
12	2,5-furanyl	Ph	<i>i</i> -Bu	13.5
13	2,5-furanyl	2-thienyl	<i>i</i> -Bu	8
14	2,5-furanyl	3-pyridyl	<i>i</i> -Bu	5
15	2,5-furanyl	NH ₂	H	0.45
16	2,5-furanyl	NH ₂	Me	0.12
17	2,5-furanyl	NH ₂	<i>n</i> -Pr	0.03
18	2,5-furanyl	NH ₂	<i>i</i> -Pr	0.028
19	2,5-furanyl	NH ₂	CF ₃ CH ₂	0.057
20	2,5-furanyl	NH ₂	neopentyl	0.012
21	2,5-furanyl	NH ₂	cyclobutyl	0.019
22	2,5-furanyl	NH ₂	cyclopentyl	0.021
23	2,5-furanyl	NH ₂	cyclohexyl	0.01
24	2,5-furanyl	NH ₂	cyclopropyl-CH ₂	0.02
25	2,5-furanyl	NH ₂	cyclopentyl-CH ₂	0.018
26	2,5-furanyl	NH ₂	cyclohexyl-CH ₂	0.059
27	2,5-furanyl	NH ₂	PhCH ₂	0.15
28	2,5-furanyl	NH ₂	morpholinyl-CH ₂	0.56
29	2,5-furanyl	NH ₂	Cl	0.07
30	2,5-furanyl	NH ₂	Br	0.05
31	2,5-furanyl	NH ₂	I	0.1
32	2,5-furanyl	NH ₂	1-morpholinyl	0.016
33	2,5-furanyl	NH ₂	EtS	0.033
34	2,5-furanyl	NH ₂	<i>n</i> -PrS	0.016
35	2,5-furanyl	NH ₂	<i>i</i> -PrS	0.024
36	2,5-furanyl	NH ₂	<i>t</i> -BuS	0.024
37	2,5-furanyl	NH ₂	PhS	0.3
38	2,5-furanyl	NH ₂	CO ₂ Et	0.014
39	2,5-furanyl	NH ₂	CO ₂ Bn	0.015
40	2,5-furanyl	NH ₂	<i>n</i> -PrSO	0.858
41	2,5-furanyl	NH ₂	Ph	0.014

Table 1 (continued)

Compound No.	Substituent			IC ₅₀ (μM)
	linker	R ₁	R ₂	
42	2,5-furanyl	NH ₂	2-MeO-Ph	0.043
43	2,5-furanyl	NH ₂	3-MeO-Ph	0.021
44	2,5-furanyl	NH ₂	4-MeO-Ph	0.022
45	2,5-furanyl	NH ₂	4-MeS-Ph	0.021
46	2,5-furanyl	NH ₂	4- <i>t</i> -Bu-Ph	0.088
47	2,5-furanyl	NH ₂	4-MeO ₂ C-Ph	0.014
48	2,5-furanyl	NH ₂	4-F-Ph	0.016
49	2,5-furanyl	NH ₂	4-Cl-Ph	0.013
50	2,5-furanyl	NH ₂	4-Ac-Ph	0.032
51	2,5-furanyl	NH ₂	4-MeSO ₂ -Ph	0.041
52	2,5-furanyl	NH ₂	4-Ph-Ph	0.034
53	2,5-furanyl	NH ₂	2-nathphyl	0.012
54	2,5-furanyl	NH ₂	2-furanyl	0.04
55	2,5-furanyl	NH ₂	2-thienyl	0.044
56	-CH ₂ OCO-	NH ₂	<i>n</i> -Pr	0.05
57	-CH ₂ NHCO-	NH ₂	2-thienyl	0.95
58	2,6-pyridyl	NH ₂	H	2
59	1,3-phenyl	NH ₂	H	1.3
60	1,3-phenyl	NH ₂	<i>n</i> -Pr	0.25
61	1,3-phenyl-(6-Me)	NH ₂	<i>n</i> -Pr	0.135
62	1,3-phenyl-(6-OMe)	NH ₂	<i>i</i> -Pr	0.21
63	1,3-phenyl-(6-F)	NH ₂	Ph	0.08

The CoMSIA model also using 12 molecules in the test set, gave a predictive correlation coefficient (r^2_{pred}) of 0.710, slope a value of 1.042 (close to 1), intercept b value of -0.365 (close to 0), an excellent $r^2_{\text{m(test)}}$ value of 0.860 (> 0.5) as well as high slope of regression lines through the origin (k) value of 0.994 ($0.85 \leq k \leq 1.15$), and the correlation coefficient (R) values of 0.895 (close to 1), the calculated $[(r^2 - r_0^2) / r^2]$ values of -0.015 (< 0.1) were obtained. It was indicated in this external validation process that the CoMSIA model exhibited slightly better predictive power than CoMFA model, and both the two models may be reliable for being used to predict the potencies of novel derivatives.

CoMFA versus CoMSIA

The conclusions derived from the PLS and external validation parts of present study demonstrated that both the CoMFA and CoMSIA models could be used reliably

to predict the FBPase inhibitory activities of these phosphonic acid-containing thiazole derivatives; moreover, they may be regarded as valuable tools to design new inhibitors with improved potencies against FBPase. Compared with CoMSIA, the CoMFA model displayed slightly better PLS statistics, which indicated that the CoMFA model possessed higher predictive power than CoMSIA. In the external validation analysis, the CoMFA model was also found to be slightly more reliable, it displayed better r^2_{pred} , $r^2_{\text{m(test)}}$, r_0^2 , R , and slope k values than the CoMSIA model.

Graphical interpretation of CoMFA model

To visualize the information content of the derived 3D-QSAR models, both the CoMFA and CoMSIA contour maps were generated by interpolating the products between the 3D-QSAR coefficients and their associated standard

Table 2 The actual pIC₅₀s, predicted pIC₅₀s (Pred.), their residuals (Res.) and Surflex-Dock total score (docking score) values of the training and test set molecules

Compound	pIC ₅₀	CoMFA		CoMSIA		Docking
No.	Actual	Pred.	Res.	Pred.	Res.	score
1	7.000	7.157	-0.157	7.055	-0.055	6.43
2	6.398	6.625	-0.227	6.836	-0.438	6.34
3*	5.921	6.580	-0.659	6.698	-0.777	6.07
4	6.658	6.730	-0.072	6.608	0.050	5.20
5	6.745	6.456	0.289	6.817	-0.072	6.14
6	6.051	5.947	0.104	6.109	-0.058	6.38
7*	5.699	5.782	-0.083	5.811	-0.112	5.48
8	6.000	6.017	-0.017	5.993	0.007	5.77
9	5.000	4.844	0.156	5.025	-0.025	5.44
10	5.561	5.674	-0.113	5.595	-0.034	5.20
11	6.301	6.234	0.067	6.089	0.212	4.58
12	4.870	4.940	-0.070	4.935	-0.065	6.34
13	5.097	5.169	-0.072	5.129	-0.032	6.24
14	5.301	5.175	0.126	5.088	0.213	6.32
15*	6.347	6.242	0.105	6.621	-0.274	4.31
16*	6.921	6.960	-0.039	7.061	-0.140	6.30
17	7.523	7.632	-0.109	7.511	0.012	7.38
18	7.553	7.466	0.087	7.558	-0.005	6.88
19*	7.244	7.412	-0.168	6.905	0.339	6.34
20	7.921	7.742	0.179	7.971	-0.050	7.23
21	7.721	7.665	0.056	7.626	0.095	4.97
22	7.678	7.642	0.036	7.602	0.076	6.84
23	8.000	8.062	-0.062	7.956	0.044	6.98
24	7.699	7.690	0.009	7.612	0.087	6.96
25*	7.745	7.542	0.203	7.590	0.155	6.67
26	7.229	7.359	-0.130	7.368	-0.139	7.20
27	6.824	6.798	0.026	6.916	-0.092	7.35
28	6.252	6.269	-0.017	6.186	0.066	5.91
29*	7.155	7.002	0.153	7.182	-0.027	3.50
30	7.301	7.076	0.225	7.284	0.017	4.42
31	7.000	7.213	-0.213	7.386	-0.386	4.95
32	7.796	7.914	-0.118	7.826	-0.030	5.93
33	7.482	7.493	-0.012	7.527	-0.046	7.33
34	7.796	7.775	0.021	7.604	0.192	5.57
35	7.620	7.657	-0.037	7.717	-0.097	7.30
36	7.620	7.610	0.010	7.447	0.173	7.31
37	6.523	6.548	-0.025	6.568	-0.045	5.56
38	7.854	7.757	0.097	7.652	0.202	6.08
39	7.824	7.794	0.030	7.930	-0.106	2.72
40	6.067	6.016	0.050	6.328	-0.262	6.47
41	7.854	7.683	0.171	7.695	0.159	6.49
42	7.367	7.562	-0.196	7.448	-0.082	4.98
43	7.678	7.654	0.024	7.568	0.110	6.47
44	7.658	7.528	0.130	7.443	0.215	5.49
45	7.678	7.709	-0.031	7.761	-0.083	6.92
46	7.056	7.122	-0.066	7.100	-0.044	7.03
47	7.854	7.844	0.010	7.864	-0.010	6.24

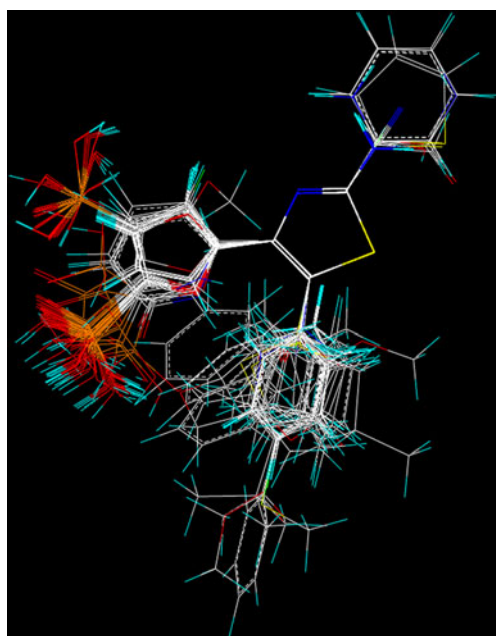
Table 2 (continued)

Compound	pIC ₅₀	CoMFA		CoMSIA		Docking
48	7.796	7.873	-0.077	7.687	0.109	6.16
49	7.886	7.935	-0.049	7.762	0.124	7.34
50*	7.495	7.353	0.142	7.621	-0.126	6.97
51	7.387	7.436	-0.049	7.419	-0.032	6.62
52*	7.469	7.448	0.020	7.477	-0.009	6.74
53	7.921	7.884	0.037	7.976	-0.055	6.46
54	7.398	7.302	0.096	7.342	0.056	6.60
55*	7.357	7.311	0.045	7.584	-0.228	6.40
56	7.301	7.199	0.102	7.351	-0.050	7.35
57	6.022	6.093	-0.071	6.056	-0.034	6.43
58	5.699	5.955	-0.256	5.833	-0.134	5.24
59	5.886	5.818	0.068	5.636	0.250	6.05
60*	6.602	6.788	-0.186	6.477	0.125	6.14
61*	6.870	6.974	-0.104	6.658	0.212	5.89
62	6.678	6.638	0.040	6.664	0.014	5.78
63	7.097	7.123	-0.026	7.065	0.032	5.17

* Test set molecules

deviations. They could rationalize the regions in 3D space around the molecules where changes in the steric, electrostatic, hydrophobic, hydrogen bond donor and acceptor fields were predicted to increase or decrease the activity.

Since the compound **23** was the most suitable molecule to illustrate the information taken from CoMFA and CoMSIA contour maps, it was selected as a reference structure. The CoMFA steric and electrostatic contour maps were shown in Fig. 4. The steric field is represented by

**Fig. 2** Alignment of the compounds used in the training set

green and yellow contours, in which green contours (80% contribution) indicate regions where bulky group would be favorable, while the yellow contours (20% contribution)

Table 3 PLS statistics of CoMFA and CoMSIA models

PLS statistics	CoMFA	CoMSIA
r_{cv}^2 ^a	0.675	0.619
r^{2b}	0.985	0.979
ONC ^c	6	6
SEE ^d	0.115	0.136
F value ^e	487.178	345.303
$r_{bootstrapping}^2$ ^f	0.990±0.004	0.987±0.005
SEE _{bootstrapping} ^g	0.091±0.056	0.106±0.066
Field contribution		
Steric	0.529	0.138
Electrostatic	0.471	0.260
Hydrophobic	-	0.196
H-bond donor	-	0.253
H-bond acceptor	-	0.153

^a the cross-validated correlation coefficient after the LOO procedure on the training set of compounds

^b the non-cross-validated correlation coefficient of the training set

^c the optimal number of principal components in the PLS model

^d the standard error of estimate

^e the value of Fisher test

^f the average of correlation coefficient for 100 samplings using bootstrapping procedure

^g the average standard error of estimate for 100 samplings using bootstrapping procedure

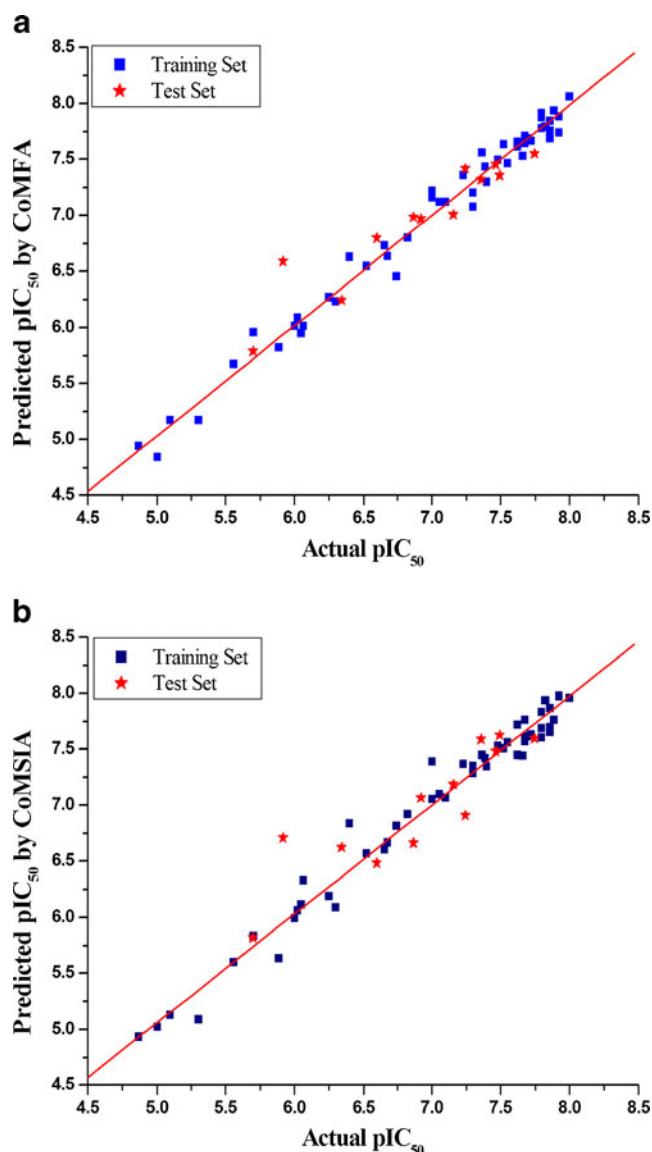


Fig. 3 Graph of actual versus predicted pIC_{50} of the training set and the test set using CoMFA (a) and CoMSIA (b)

represent regions where bulky group would decrease the activity. The electrostatic field is indicated by blue (80% contribution) and red (20% contribution) contours, which demonstrate the regions where electron-donating group and electron-withdrawing group would be favorable respectively.

In Fig. 4a, the huge green contour around the R_2 position suggested that a bulky group at this site would benefit the inhibitory activity. This consisted of the fact that compounds **20–26**, **32**, **34–36**, **38**, **39**, **41**, **43–45**, **47–49** and **53** which possessed a relative bulky substituent (e.g., cyclobutyl, cyclopentyl, cyclohexyl, 1-morpholinyl, phenyl, 3-MeO-phenyl, 4-MeO-phenyl, 4-MeS-phenyl, 4-MeO₂C-phenyl, 4-F-phenyl, 4-Cl-phenyl and 2-nathphyl) at R_2 exhibited excellent inhibitory potencies. On the other hand, compounds **15**, **16**, **58** and **59** bearing a minor group (e.g.,

H, methyl) at R_2 showed significantly decreased activities. For instance, compounds **15** ($pIC_{50}=6.347$), **16** ($pIC_{50}=6.921$), **17** ($pIC_{50}=7.523$), **18** ($pIC_{50}=7.553$) had an order for the activity of **15**<**16**<**17**<**18**, with the corresponding R_2 substituent -H, methyl, *n*-propyl, *i*-propyl, respectively. By comparing compounds **54** ($R_2=2$ -furanyl, $pIC_{50}=7.398$) and **55** ($R_2=2$ -thienyl, $pIC_{50}=7.357$) with compound **41** ($R_2=$ phenyl, $pIC_{50}=7.854$), it can be concluded that their activity discrepancies can be also explained by this green contour. Two green contours along with three yellow contours around the furanyl-2-phosphonic acid indicated that the steric field at this site exerted no significant effect on the inhibitory activity.

In Fig. 4b, the huge blue contour near the R_1 position revealed that an electron-donating substituent at this site would be favorable. In general, compared to compounds **15–63** with a strong electron-donating amino group at R_1 , the compounds **1–14** bearing electron-withdrawing groups (i.e., vinyl, -Cl, -CN, -NHAc, -CONH₂, -CSNH₂, phenyl, 2-thienyl and 3-pyridyl) or less electron-donating substituent (i.e., methyl, ethyl and -CH₂OH) at R_1 displayed significantly decreased potencies. Moreover, compounds **3** ($R_1=$ vinyl, $pIC_{50}=5.921$), **7** ($R_1=-$ CN, $pIC_{50}=5.699$), **12** ($R_1=$ phenyl, $pIC_{50}=4.870$), **13** ($R_1=2$ -thienyl, $pIC_{50}=5.097$), **14** ($R_1=3$ -pyridyl, $pIC_{50}=5.301$) had a strong electron-withdrawing group at R_1 were the most inactive inhibitors. Four red contours along with one blue contours around the R_2 site demonstrated that the electrostatic field at this position exerted no significant influence on the inhibitory potency. This was validated by the fact that both compounds **20–25** and **32** with electron-donating groups (i.e., neopentyl, cyclobutyl, cyclopentyl, cyclohexyl, cyclopropyl-CH₂, cyclopentyl-CH₂, 1-morpholinyl) and compounds **38**, **39**, **41**, **47–49** and **53** bearing electron-withdrawing substituent (i.e., -CO₂Et, -CO₂Bn, phenyl, 4-MeO₂C-phenyl, 4-F-phenyl, 4-Cl-phenyl, 2-nathphyl) at R_2 site showed excellent inhibitory activities with the pIC_{50} values ranging from 7.678 to 8.000.

Table 4 Results of the external validation for CoMFA and CoMSIA models

Parameters	CoMFA	CoMSIA
r^2_{pred}	0.805	0.710
Slope a	1.136	1.042
Intercept b	-0.995	-0.365
Correlation coefficient R	0.940	0.895
Slope k	0.995	0.994
r_0^2	0.995	0.994
$r^2_{m(test)}$	0.887	0.860
$[(r^2 - r_0^2) / r^2]$	-0.010	-0.015

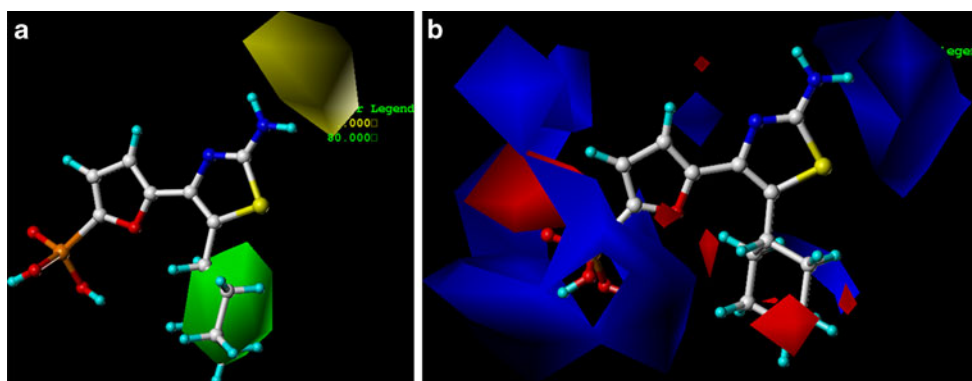


Fig. 4 Std* coeff contour maps of CoMFA analysis with 2 Å grid spacing in combination with compound **23**. (a) Steric fields: green contours indicate regions where bulky groups increase activity, while yellow contours indicate regions where bulky groups decrease activity,

and (b) Electrostatic fields: blue contours represent regions where electron-donating groups improve activity, while red contours represent regions where electron-withdrawing groups benefit activity

Graphical interpretation of CoMSIA model

Figure 5a–e provided the steric, electrostatic, hydrophobic, hydrogen bond donor and acceptor contours plots for compound **23** of the CoMSIA model. The CoMSIA electrostatic contour map was almost the same AS the corresponding CoMFA electrostatic contour map. In hydrophobic field, white (20% contribution) and yellow (80% contribution) contours highlighted areas where hydrophilic and hydrophobic properties were favored. In hydrogen bond donor field, the cyan (80% contribution) and purple (20% contribution) contours indicated favorable and unfavorable hydrogen bond donor groups. In hydrogen bond acceptor field, the magenta (80% contribution) and red (20% contribution) contours identified favorable and unfavorable positions for hydrogen bond acceptors.

Unlike the CoMFA steric contour map, in Fig. 5a, a huge yellow contour near the R₁ position indicated that a bulky substituent would decrease the activity. Most of the compounds possessed a relative minor amino group at R₁, meanwhile, compounds **8–14** with bulky substituent (phenyl, 2-thienyl, 3-pyridyl, -CONH₂, -CSNH₂, -NHAc, -NHMe) at this site were the most inactive inhibitors with their pIC₅₀ values ranging from 4.870 to 6.301. The green contour around the R₂ position was in agreement with the corresponding CoMFA steric contour map.

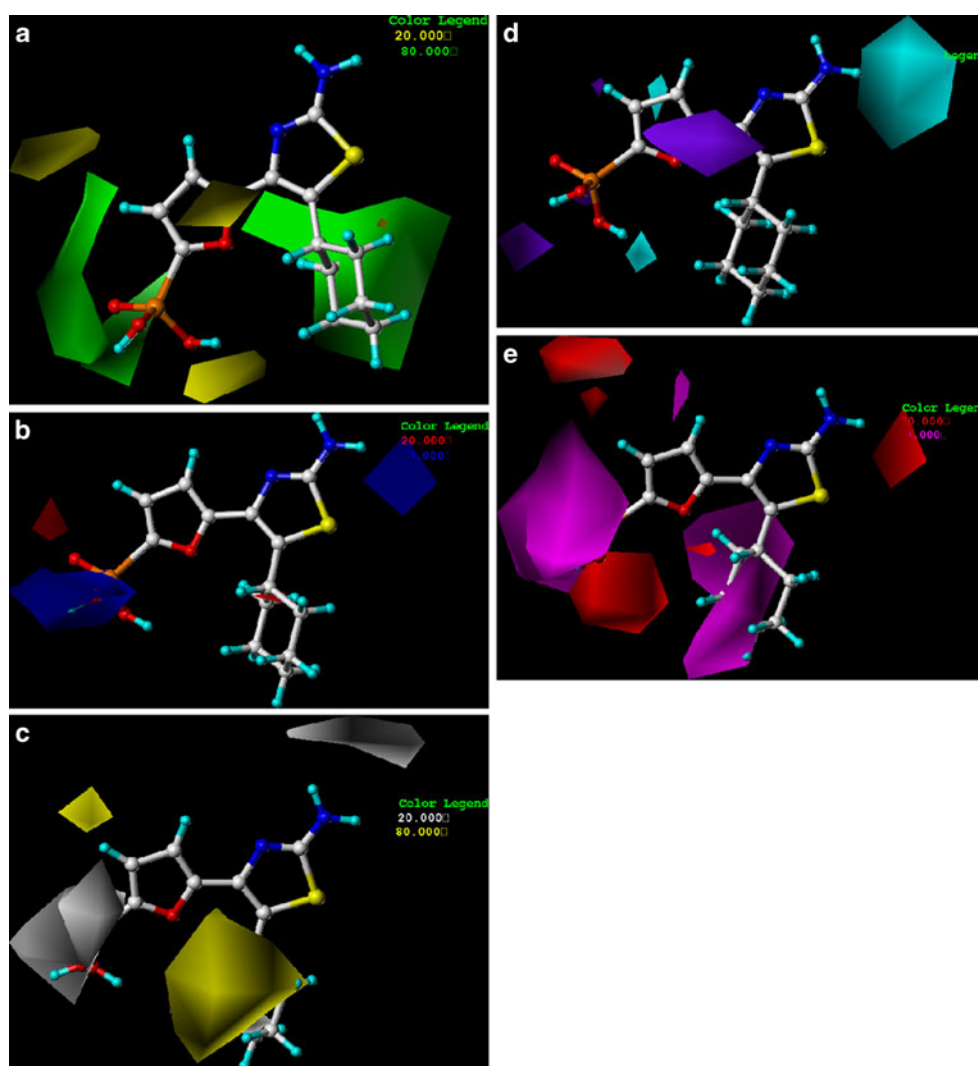
In Fig. 5c, a white contour near the R₁ position revealed that a hydrophilic group at this site may benefit the activity. This may be one of the reasons why compounds **15–63** with a hydrophilic amino group at R₁ showed better inhibitory potencies than compounds **1–15** bearing hydrophobic groups (i.e., methyl, ethyl, vinyl, -Cl, -CN, -NHAc, -CONH₂, -CSNH₂, phenyl, 2-thienyl and 3-pyridyl). Another huge yellow contour around the R₂ site suggested that a hydrophobic substituent may have the positive effect on the activity. This consisted of the fact that most of the potential

inhibitors such as **20–26**, **32**, **34–36**, **38**, **39**, **41**, **43–45**, **47–49** and **53** all possessed a hydrophobic substituent (e.g. cyclobutyl, cyclopentyl, cyclohexyl, 1-morpholinyl, phenyl, 3-MeO-phenyl, 4-MeO-phenyl, 4-MeS-phenyl, 4-MeO₂C-phenyl, 4-F-phenyl, 4-Cl-phenyl and 2-naphthyl) at R₂. Moreover, the hydrophobic favored yellow contour near the furanyl moiety indicated the extreme importance of the hydrophobic linker. In fact, almost every compound had a hydrophobic linker at this position. Meanwhile, two huge white contours around the phosphate group also demonstrated the extreme importance of the hydrophilic phosphate group. Each compound of the entire dataset possessed a phosphate group which was necessary for the inhibitory activity.

In Fig. 5d, the cyan contour near the R₁ position suggested that a hydrogen bond donor group at this site would increase the potency. This was validated by the fact that compounds **15–63** with a hydrogen bond donor amino group at R₁ exhibited much more excellent inhibitory potencies than compounds **1–3**, **5–7** and **12–14** without hydrogen bond donor groups. In fact, the amino group at R₁ acted as hydrogen bond donor and formed H-bond with the residues of the AMP binding region of FBPase, removing of it may result in decreased activity. A purple contour around the furanyl group indicated that a hydrogen bond acceptor linker was essential for the potency. In fact, almost every inhibitor possessed such a hydrogen bond acceptor linker. On the other hand, two purple contours along with one cyan contour near the phosphate group revealed that it may serve as hydrogen bond donor and acceptor at the same time and indicated the extreme importance of the phosphate group.

In Fig. 5e, the red contour near the R₁ site indicated that a hydrogen bond acceptor substituent would be unfavorable. This was in agreement with the observation taken from Fig. 5d. The oxygen atom of furanyl group was oriented towards a huge magenta contour, suggesting a

Fig. 5 Std* coeff contour maps of CoMSIA analysis with 2 Å grid spacing in combination with compound **23**. **(a)** Steric contour map. Green and yellow contours refer to sterically favored and unfavored regions. **(b)** Electrostatic contour map. Blue and red contours refer to regions where electron-donating and electron withdrawing groups are favored. **(c)** Hydrophobic contour map. White and yellow contours refer to regions where hydrophilic and hydrophobic substituent are favored. **(d)** Hydrogen bond donor contour map. The cyan and purple contours indicate favorable and unfavorable hydrogen bond donor groups. **(e)** Hydrogen bond acceptor contour map. The magenta and red contours demonstrated favorable and unfavorable hydrogen bond acceptor groups



hydrogen bond acceptor moiety at this site would be favored. Meanwhile, the phosphate group was surrounded by two huge purple and one red contours, indicating it may serve as hydrogen bond donor and acceptor at the same time. The observations taken from this hydrogen bond acceptor contour map satisfactorily matched the hydrogen bond donor contour map (Fig. 5d).

Molecular docking analysis

Molecular docking was employed to explore the binding mode between these phosphonic acid-containing thiazole derivatives and the receptor, furthermore, to examine the stability of 3D-QSAR models previous established. Since the crystal structure of FBPase was known, we replaced AMP with these derivatives to examine their bound to FBPase. The calculated Surflex-Dock total score of the entire database ranging from 2.32 to 7.38 were listed in Table 2. In general, the most active compounds **20–25**, **32**,

34–36, **38**, **39**, **41**, **43–45**, **47–49** and **53** with their pIC_{50} values ranging from 7.620 to 8.000 had a mean docking score value of 6.31; the active compounds **17–19**, **26**, **29–31**, **33**, **42**, **46**, **50–52**, **54–56** and **63** with their pIC_{50} values ranging from 7.000 to 7.553 had a mean docking score value of 6.18; the less inactive compounds **1**, **2**, **4–6**, **15**, **16**, **27**, **28**, **37**, **40** and **60–62** with the pIC_{50} values ranging from 6.051 to 7.000 had a mean docking score value of 6.08; the most inactive compounds **3**, **7–14** and **57–59** with the pIC_{50} values ranging from 4.870 to 6.301 had a mean docking score value of 5.76. Since the docking score represented the binding affinities and energies between these inhibitors and the pocket, it could be inferred that the enhanced affinities (higher docking scores) may result in improved inhibitory potencies. Meanwhile, their binding affinities were determined by their molecular structures. In order to identify the structural features responsible for the binding affinity, the most potential inhibitors **23** was selected for more detailed study.

Figure 6a showed the exact hydrogen bond binding mode between the selected compound **23** and the residues of AMP binding site of FBPase, the key residues and hydrogen bonds were labeled. Not surprisingly, the $-NH_2$ at R_1 position served as hydrogen bond donor by forming three H-bonds with the carbonyl group of Val17 and the hydroxyl group of Thr31, respectively. Furthermore, the carbonyl of the phosphate group acted as hydrogen bond acceptor and formed four H-bonds with the imino groups of Leu30, Gly28, Glu29 and the hydroxyl group of Thr27; one hydroxyl of the phosphate group served as hydrogen bond donor and acceptor at the same time and formed five H-bonds with the imino and hydroxyl groups of Thr27 as well as the water molecule; the other hydroxyl only acted as hydrogen bond acceptor by forming two H-bonds with the $-NH_3$ group of Lys112 and the hydroxyl group of Tyr113, respectively. The water molecule was important for the binding since it contributed three H-bonds. The observations obtained from this picture were in agreement with the corresponding CoMSIA hydrogen bond donor and acceptor contour maps, which indicated the extreme importance of the amino and phosphate groups for the inhibitory potency. This may be the reason for the poor activities of compounds **1–3**, **5–7** and **12–14** without hydrogen bond donor groups at R_1 . To validate this assumption, the most inactive compound **12** was selected for more detailed docking research. Figure 6b illustrated the hydrogen bond binding mode between compound **12** and the residues of AMP binding site for comparison. It can be found that the hydrogen bond interaction in Fig. 6b was less than Fig. 6a. In Fig. 6b, the carbonyl of the phosphate group acted as hydrogen bond acceptor and formed four H-bonds with the imino and hydroxyl groups of Thr27, Gly28 and the water molecule; one hydroxyl of the phosphate group served as hydrogen bond acceptor and formed three H-bonds with the imino and hydroxyl groups of Leu30, Glu29 and Tyr113; the other hydroxyl only acted as hydrogen bond acceptor by forming three H-bonds with the $-NH_3$ group of Lys112 and the hydroxyl group of Tyr113

as well as the water molecule. Compound **12** was unable to form hydrogen bond with Val17 and Thr31 due to lacking of the amino group at R_1 position.

The MOLCAD surface of AMP binding site was also developed and displayed with cavity depth (CD), electrostatic potential (EP), lipophilic potential (LP) and hydrogen bond site (HB) to further explore the interaction between these inhibitors and the receptor. These potentials on a protein surface can be used to find the sites that act attractively on ligands by matching opposite colors.

Figure 7a and b depicted the MOLCAD ribbon and multi-channel cavity depth potential (CD) surfaces structure of the binding site within the compound **23** in ball & stick (a) and space fill (b) formats, respectively. The cavity depth color ramp ranges from blue (low depth values=outside of the pocket) to light red (high depth values=cavities deep inside the pocket). As shown in Fig. 7a and b, the R_1 site of the thiazole scaffold and the phosphate group (along with the linker part) were oriented to yellow areas, which revealed that these parts of the compound **23** were anchored deep inside the binding region. It can be inferred that since the R_1 position was found to be very close to the binding surface, the space between the R_1 and the surface was narrow, thus a bulky substituent at this position may exert negative effects on the binding affinity, which will result in decreased activity. This was validated by the fact that compounds **8–14** with bulky substituent at R_1 were the most inactive inhibitors. Meanwhile, the cyclohexyl group at R_2 was located in a blue region, suggesting it was oriented to the entrance of the binding pocket. Since the space accommodating the R_2 substituent was large, it was capable for holding a bulky group. This may be the reason why compounds **20–26**, **32**, **34–36**, **38**, **39**, **41**, **43–45**, **47–49** and **53** bearing a bulky substituent at R_2 exhibited excellent inhibitory potencies. The observations and conclusions taken from Fig. 7 satisfactorily matched the corresponding CoMFA and CoMSIA steric contour maps.

The MOLCAD electrostatic potential (EP) surface of the binding region was shown in Fig. 8a. The color ramp for

Fig. 6 The binding modes between selected compound **23** (a) and compound **12** (b) with the AMP binding site of FBPase (PDB code 1FTA). Key residues and hydrogen bonds were labeled

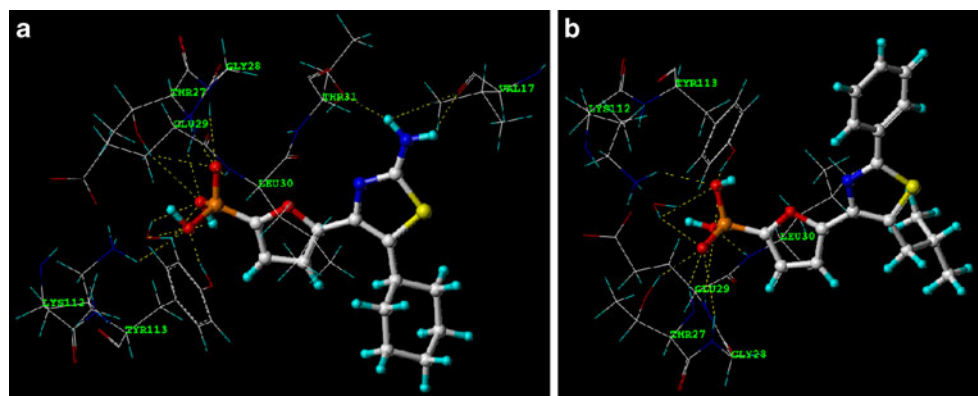
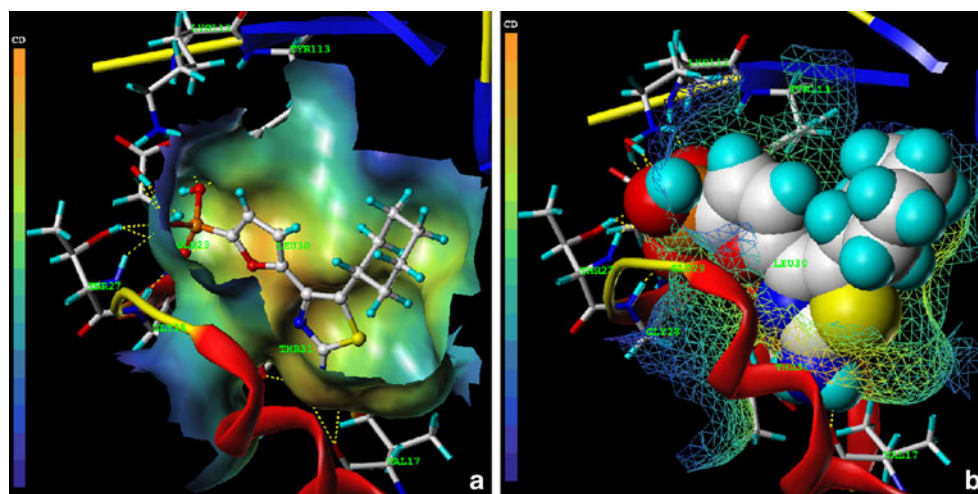


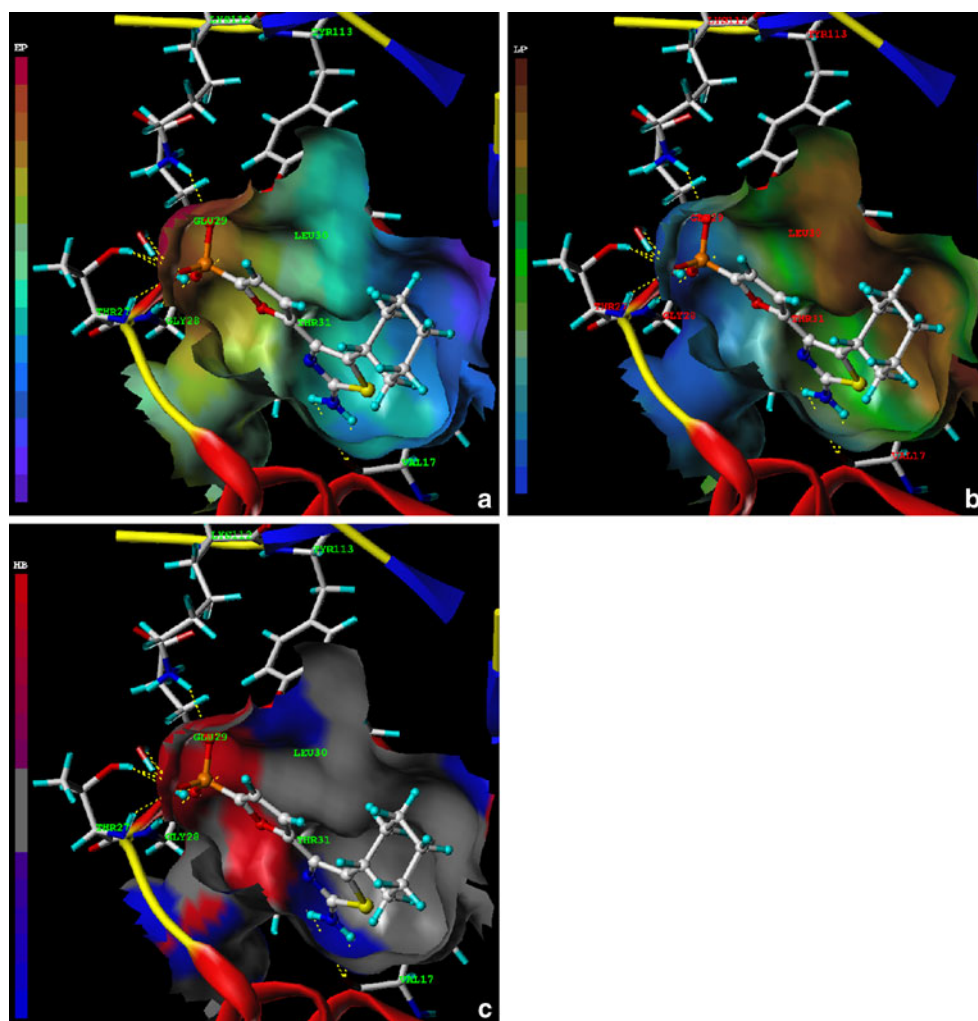
Fig. 7 The MOLCAD ribbon and multi-channel surfaces structure displayed with cavity depth potential of the AMP binding site within the compound **23**. Key residues and hydrogen bonds were labeled. The cavity depth color ramp ranges from blue (low depth values=outside of the pocket) to light red (high depth values=cavities deep inside the pocket)



EP ranges from red (most positive) to purple (most negative). The R_1 position was found in a blue area, indicating that electron-donating properties at this site were beneficial for the binding affinity. This consisted of the fact that compounds **3–7** and **10–14** without electron-donating groups at

R_1 displayed significantly decreased inhibitory potencies as well as binding affinities. The phosphate group and the furanyl linker were anchored in red and yellow areas, revealing that electron-withdrawing properties may be favored, which indicated the extreme importance of the

Fig. 8 The MOLCAD electrostatic potential (a), lipophilic potential (b) and hydrogen binding (c) surfaces of the AMP binding site of FBPase within the compound **23**. The color ramp for EP ranges from red (most positive) to purple (most negative); the color ramp for LP ranges from brown (highest lipophilic area of the surface) to blue (highest hydrophilic area); the color ramp for HP ranges from red (hydrogen bond donors) to blue (hydrogen bond acceptors)



electron-withdrawing phosphate group and the aromatic linkers. It can be inferred that a molecule without an aromatic linker may be unable to participate in aromatic (π - π) stacking interactions with the residues of the binding region. The observations and conclusions were in agreement with the corresponding CoMFA and CoMSIA electrostatic contour maps.

Figure 8b showed the MOLCAD lipophilic potential surface of the binding area, the color ramp for LP ranges from brown (highest lipophilic area of the surface) to blue (highest hydrophilic area). The R_1 position was oriented to a white region, which indicated that a hydrophilic substituent may benefit the binding affinity. This again revealed the importance of the amino groups of compounds 15–63. On the other hand, the R_2 site was located in a brown area, suggesting a hydrophobic group would increase the binding affinity. This consisted of the fact that most of the potential inhibitors such as 20–26, 32, 34–36, 38, 39, 41, 43–45, 47–49 and 53 all possessed a hydrophobic substituent at R_2 . Moreover, the phosphate group was found in a blue region, which demonstrated that the hydrophilic phosphate group was essential for binding to the residues of the FBPase. The observations and conclusions satisfactorily matched the corresponding CoMSIA hydrophobic contour map.

Figure 8c illustrated the MOLCAD hydrogen bonding sites of the binding surfaces, ligands can be docked to proteins by matching the patterns displayed on the surface, the color ramp for HB ranges from red (hydrogen donors) to blue (hydrogen acceptors). As shown in Fig. 8c, the amino group at R_1 was oriented to a wide blue surface, which indicated that the surface of this site were hydrogen bond acceptors, thus a hydrogen bond donor substituent would be favorable. Meanwhile, the R_2 position was located in a huge gray region indicating the surface of this region were not hydrogen bond acceptors nor donors, therefore, the hydrogen bond donor or acceptor field had no significant effect on R_2 . The oxygen atom of the furanyl linker was oriented to a red area, which indicated that the surfaces of this site were hydrogen bond donors, and a

hydrogen bond acceptor property may be favorable. This may be the reason for the inactivity of compounds 59–62 (linker=1,3-phenyl, 1,3-phenyl-(6-Me), 1,3-phenyl-(6-OMe)) without hydrogen bond acceptor linkers. The phosphate group was also anchored in a huge red surface, demonstrating a hydrogen bond acceptor group would enhance the binding affinity. Removal of the hydrogen bond acceptor of this position may result in poor binding affinity. The observations and conclusions taken from this hydrogen bonding sites were in agreement with the corresponding CoMSIA hydrogen bond contour maps.

Design of new inhibitors

The structure-activity relationship and binding features obtained by present 3D-QSAR and molecular docking analysis are summarized in Fig. 9. In detail, the minor, electron-donating, hydrophilic and hydrogen bond donor groups at R_1 position would be favored; the bulky, hydrophobic substituent at R_2 site would benefit the inhibitory potency; an electron-withdrawing, hydrophobic and hydrogen bond acceptor linker may be desirable; the electron-donating, hydrophilic, hydrogen bond donor and acceptor phosphate group was essential for binding to the AMP pocket. According to literature [16], in order to achieve acceptable OBAV, the molecular weight of these phosphonic acid-containing thiazoles should be limited to below 600. We have employed this combined useful information of the structural requirements as well as the synthetic availability of these derivatives to design a set of 40 new analogues showing excellent inhibitory activities in the 3D-QSAR models previously established. Furthermore, these molecules also exhibited good Surflex-Dock total score in the molecular docking experiments.

These molecules were designed by introducing minor, electron-donating, hydrophilic and hydrogen bond donor groups (i.e., -OH, -NH₂) at R_1 site; bulky or hydrophobic groups (i.e., substituted aromatic rings or substituted pyrrolyl) at R_2 position; electron-withdrawing, hydrophobic and hydrogen bond acceptor groups (e.g., 2,5-furanyl, 2,5-

Fig. 9 Structure-activity relationship and binding features obtained by present study

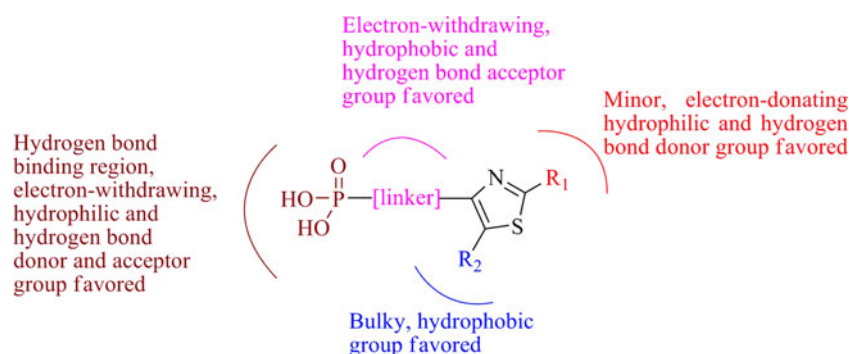


Table 5 Structure, predicted pIC₅₀ values, Surflex-Dock total score (score) and molecular weight (M. W.) of newly designed derivatives

No.	Substituent			Predicted pIC ₅₀		Score	M.W.
	linker	R ¹	R ²	CoMFA	CoMSIA		
D1	2,5-furanyl	-NH ₂		7.926	8.140	6.75	406.35
D2	2,5-furanyl	-NH ₂		9.228	7.733	7.38	396.31
D3	2,5-furanyl	-NH ₂		8.353	7.422	6.80	410.38
D4	2,5-furanyl	-NH ₂		8.066	7.823	7.46	365.34
D5	2,5-furanyl	-NH ₂		8.890	7.091	6.43	380.31
D6	2,5-furanyl	-NH ₂		8.285	7.274	6.25	440.36
D7	2,5-furanyl	-NH ₂		8.840	7.776	7.09	366.33
D8	2,5-furanyl	-NH ₂		8.523	7.755	7.84	394.38
D9	2,5-furanyl	-NH ₂		8.900	7.048	5.76	410.34

Table 5 (continued)

No.	linker	Substituent		Predicted pIC ₅₀		Score	M.W.
		R ¹	R ²	CoMFA	CoMSIA		
D10	2,5-pyrimidinyl	-OH		8.023	7.732	5.13	392.33
D11	2,5-pyridinyl	-OH		8.068	7.415	5.12	360.32
D12	2,5-pyridinyl	-OH		8.010	7.656	4.78	364.31
D13	2,5-furanyl	-NH ₂		7.521	8.054	5.71	315.29
D14	2,5-furanyl	-NH ₂		7.208	8.246	5.64	387.39
D15	2,5-pyridinyl	-NH ₂		7.690	8.053	5.70	370.36
D16	2,5-furanyl	-NH ₂		7.112	8.234	5.73	359.34
D17	2,5-furanyl	-OH		7.104	8.005	6.16	355.31
D18	2,5-pyridinyl	-OH		7.460	7.940	5.68	366.33
D19	2,5-furanyl	-NH ₂		7.038	8.312	6.16	327.25
D20	2,5-furanyl	-OH		7.204	8.066	5.42	328.24
D21	2,5-furanyl	-NH ₂		7.545	8.382	6.36	358.35

Table 5 (continued)

No.	Substituent			Predicted pIC ₅₀		Score	M.W.
	linker	R ¹	R ²	CoMFA	CoMSIA		
D22	2,5-furanyl	-NH ₂		7.744	8.098	5.47	329.31
D23	2,5-furanyl	-NH ₂		7.451	8.080	6.32	343.34
D24	2,5-furanyl	-NH ₂		8.284	7.447	5.67	351.32
D25	2,5-pyridinyl	-OH		8.093	7.416	5.06	380.38
D26	2,5-pyridinyl	-NH ₂		7.950	7.597	5.90	379.39
D27	2,5-furanyl	-NH ₂		7.327	8.045	6.69	368.37
D28	2,5-furanyl	-NH ₂		7.319	7.908	6.04	394.38
D29	2,5-furanyl	-OH		8.021	7.351	5.55	481.85
D30	2,5-pyridinyl	-OH		7.892	7.747	4.96	371.35
D31	2,5-pyrimidinyl	-NH ₂		7.877	7.909	4.90	391.34
D32	2,5-pyridinyl	-OH		7.832	7.862	4.66	370.36
D33	2,5-furanyl	-NH ₂		7.824	7.914	5.98	386.34
D34	2,5-pyridinyl	-OH		8.086	7.950	5.78	426.40
D35	2,5-pyridinyl	-OH		8.172	8.017	5.58	426.40
D36	2,5-pyridinyl	-OH		8.280	7.911	5.94	454.46

Table 5 (continued)

No.	Substituent		Predicted pIC ₅₀		Score	M.W.	
	linker	R ¹	R ²	CoMFA			CoMSIA
D37	2,5-furanyl	-OH		7.983	7.918	4.46	373.32
D38	2,5-furanyl	-OH		8.010	8.011	6.74	344.32
D39	2,5-furanyl	-NH ₂		8.066	8.142	3.16	355.26
D40	2,5-furanyl	-OH		8.359	8.364	3.71	364.53

pyridinyl and 2,5-pyrimidinyl) at the linker site; the phosphonic group remained. The structure and molecular weight as well as predicted activities and docking score of these designed molecules were shown in Table 5.

As shown in Table 5, the hydroxyl at R₁ position could also result in excellent predicted potencies (e.g., **D10-D12**, **D17**, **D18**, **D20**, **D25**, **D29**, **D30**, **D32**, **D34-D38** and **D40**). In the case of **D2-D12**, there were some discrepancies between their predicted pIC₅₀ values of CoMFA and CoMSIA. The predicted activities of CoMFA were found to be better than CoMSIA, it can be inferred that since the **D2-D12** possessed substituted phenyl groups at R₂ site, the CoMSIA hydrogen bond fields have no significant effect on these substituents, resulting in decreased predicted pIC₅₀ values. In the case of **D13-D23**, their predicted pIC₅₀ values of CoMSIA were better than CoMFA. It can be also be inferred that the CoMSIA hydrogen bond fields possessed effect on these substituted pyrrolyl groups, resulting in increased predicted activities. Compounds **D34-40** displayed excellent predicted pIC₅₀ values in both the CoMFA and CoMSIA models. Although some of the designed molecules showing excellent predicted potencies possessed 2,5-pyridinyl and 2,5-pyrimidinyl linker (e.g., **D10-D12**, **D15**, **D18**, **D25**, **D26**, **D30-D32** and **D34-D36**), most of the designed molecules had a 2,5-furanyl acted as linker, indicating that the best linker for these derivatives was the 2,5-furanyl.

Conclusions

A combination of 3D-QSAR (CoMFA/CoMSIA) and molecular docking studies was performed on a set of 63 FBPase inhibitors for designing new compounds with improved inhibitory potency. The 3D-QSAR study yielded stable and statistically significant predictive models as indicated by high cross-correlation coefficients. The established models were validated by a systemic external validation. Furthermore, the combination of the 3D-QSAR studies and the molecular docking calculations offered enough information to understand the structure-activity relationship and identified several important structural features influencing the inhibitory activity as well as binding affinity. The robust and predictive CoMFA and CoMSIA models were then utilized to design new molecules presenting improved inhibitory potency. The selected designed molecule was subsequently docked in the AMP binding region to check how it interacted with the AMP binding site. This model can be used to guide the rational design of new inhibitors presenting improved inhibitory activity against the FBPase. Moreover, these designed molecules can be synthesized to generate a greater number of phosphonic acid-containing thiazole derivatives with required pharmacokinetics for further clinical evaluations.

Computational details

The computational details were depicted in the supporting information of this paper according to references [25–30].

Acknowledgments We gratefully acknowledge support for this research from the National Natural Science Foundation of China (Grant No. 81072554).

References

- Heng S, Harris KM, Kantrowitz ER (2010) Designing inhibitors against fructose 1,6-bisphosphatase: Exploring natural products for novel inhibitor scaffolds. *Eur J Med Chem* 45:1478–1484
- Kitas E, Mohr P, Kuhn B, Hebeisen P, Wessel HP, Haap W, Ruf A, Benz J, Joseph C, Huber W, Sanchez RA, Paehler A, Benardeau A, Gubler M, Schott B, Tozzo E (2010) Sulfonylureido thiazoles as fructose-1,6-bisphosphatase inhibitors for the treatment of Type-2 diabetes. *Bioorg Med Chem Lett* 20:594–599
- Magnusson I, Rothman DL, Katz LD, Shulman RG, Shulman GI (1992) Increased rate of gluconeogenesis in type- II diabetes-mellitus-A C-13 nuclear-magnetic-resonance study. *J Clin Invest* 90:1323–1327
- Tsukada T, Tamaki K, Tanaka J, Takagi T, Yoshida T, Okuno A, Shiiki T, Takahashi M, Nishi T (2010) A prodrug approach towards the development of tricyclic-based FBPase inhibitors. *Bioorg Med Chem Lett* 20:2938–2941
- Hebeisen P, Kuhn B, Kohler P, Gubler M, Huber W, Kitas E, Schott B, Benz J, Joseph C, Ruf A (2008) Allosteric FBPase inhibitors gain 10^5 times in potency when simultaneously binding two neighboring AMP sites. *Bioorg Med Chem Lett* 18:4708–4712
- Kebede M, Favaloro J, Gunton JE, Laybutt R, Shaw M, Wong N, Fam BC, Aston-Mourney K, Rantzaou C, Zulli A, Proietto J, Andrikopoulos S (2008) Fructose-1,6-Bisphosphatase overexpression in pancreatic β -cells results in reduced insulin secretion. *Diabetes* 57:1887–1895
- Mendicino J, Kratowich N, Oliver RM (1972) Role of enzyme-enzyme interactions in regulation of gluconeogenesis-properties and subunit structure of fructose 1,6-diphosphatase from swine kidney. *J Biol Chem* 247:6643–6650
- Heng S, Gryncel KR, Kantrowitz ER (2009) A library of novel allosteric inhibitors against fructose 1,6-bisphosphatase. *Bioorg Med Chem* 17:3916–3922
- Lan P, Xie MQ, Yao YM, Chen WN, Chen WM (2010) 3D-QSAR studies and molecular docking on [5-(4-amino-1 H-benzimidazol-2-yl)-furan-2-yl]-phosphonic acid derivatives as fructose-1,6-biphosphatase inhibitors. *J Comput Aided Mol Des* 24:993–1008
- Dang Q, Kasibhatla SR, Xiao W, Liu Y, Dare J, Taplin F, Reddy KR, Scarlato GR, Gibson R, van Poelje PD, Potter SC, Erion MD (2010) Fructose-1,6-bisphosphatase inhibitors. 2. Design, synthesis, and structure-activity relationship of a series of phosphonic acid containing benzimidazoles that function as 5'-adenosinemonophosphate (AMP) mimics. *J Med Chem* 53:441–451
- Dang Q, Brwon BS, Liu Y, Rydzewski RM, Robinson ED, van Poelje PD, Reddy RM, Erion MD (2009) Fructose-1,6-bisphosphatase inhibitors. 1. Purine phosphonic acids as novel AMP mimics. *J Med Chem* 52:2880–2898
- Erion MD, Dang Q, Reddy MR, Kasibhatla SR, Huang J, Lipscomb WN, van Poelje PD (2007) Structure-guided design of AMP mimics that inhibit fructose-1,6-bisphosphatase with high affinity and specificity. *J Am Chem Soc* 129:15480–15490
- Dang Q, Kasibhatla SR, Reddy KR, Jiang T, Reddy MR, Potter SC, Fujitaki JM, van Poelje PD, Huang J, Lipscomb WN, Erion MD (2007) Discovery of potent and specific fructose-1,6-bisphosphatase inhibitors and a series of orally-bioavailable phosphoramidase-sensitive prodrugs for the treatment of type 2 diabetes. *J Am Chem Soc* 129:15491–15502
- Yamanaka G, Wilson T, Innaimo S, Bisacchi GS, Egli P, Rinehart JK, Zahler R, Colonno RJ (1999) Metabolic studies on BMS-200475, a new antiviral compound active against hepatitis B virus. *Antimicrob Agents Chemother* 43:190–193
- Dang Q, Kasibhatla SR, Jiang T, Fan K, Liu Y, Taplin F, Schulz W, Cashion DK, Reddy KR, van Poelje PD, Fujitaki JM, Potter SC, Erion MD (2008) Discovery of phosphonic diamide prodrugs and their use for the oral delivery of a series of fructose 1,6-bisphosphatase inhibitors. *J Med Chem* 51:4331–4339
- Dang Q, Liu Y, Cashion DK, Kasibhatla SR, Jiang T, Taplin F, Jacintho JD, Li H, Sun Z, Fan Y, DaRe J, Tian F, Li W, Gibson T, Lemus R, van Poelje PD, Potter SC, Erion MD (2011) Discovery of a series of phosphonic acid-containing thiazoles and orally bioavailable diamide prodrugs that lower glucose in diabetic animals through inhibition of fructose-1,6-bisphosphatase. *J Med Chem* 54:153–165
- Cichero E, Cesarini S, Mosti L, Fossa P (2010) CoMFA and CoMSIA analyses on 4-oxo-1,4-dihydroquinoline and 4-oxo-1,4-dihydro-1,5-, -1,6- and -1,8-naphthyridine derivatives as selective CB2 receptor agonists. *J Mol Model* 16:677–691
- Prado-Prado FJ, Uriarte E, Borges F, Gonzalez-Diaz H (2009) Multi-target spectral moments for QSAR and Complex Networks study of antibacterial drugs. *Eur J Med Chem* 44:4516–4521
- Prado-Prado FJ, Ubeira FM, Borges F, Gonzalez-Diaz H (2009) Multiple distance and triadic census analysis of antiparasitic drugs complex networks. *J Comput Chem* 31:164–173
- Trossini GHG, Guido RVC, Oliva G, Ferreira EI, Andricopulo AD (2009) Quantitative structure-activity relationships for a series of inhibitors of cruzain from *Trypanosoma cruzi*: Molecular modeling, CoMFA and CoMSIA studies. *J Mol Graph Model* 28:3–11
- Lan P, Huang ZJ, Sun JR, Chen WM (2010) 3D-QSAR and molecular docking studies on fused pyrazoles as p38 α mitogen-activated protein kinase inhibitors. *Int J Mol Sci* 11:3357–3374
- Golbraikh A, Tropsha A (2002) Beware of q^2 . *J Mol Graph Model* 20:269–276
- Roy PP, Roy K (2008) On some aspects of variable selection for partial least squares regression models. *QSAR Comb Sci* 27:302–313
- Roy PP, Pau S, Mitra I, Roy K (2009) On two novel parameters for validation of predictive QSAR models. *Molecules* 14:1660–1701
- SYBYL 8.1, Tripos Inc., 1699 South Hanley Rd., St. Louis, MO, 63144, USA
- Lan P, Chen WN, Xiao GK, Sun PH, Chen WM (2010) 3D-QSAR and docking studies on pyrazolo[4,3-h]quinazoline-3-carboxamides as cyclin-dependent kinase 2 (CDK2) inhibitors. *Bioorg Med Chem Lett* 20:6764–6772
- Khanfar MA, Youssef DTA, Sayed KAE (2010) 3D-QSAR studies of latrunculin-based actin polymerization inhibitors using CoMFA and CoMSIA approaches. *Eur J Med Chem* 45:3662–3668
- Mouchlis VD, Mavromoustakos TM, Kokotos G (2010) Molecular docking and 3D-QSAR CoMFA studies on indole inhibitors of GIIA secreted phospholipase A₂. *J Chem Inf Model* 50:1589–1601
- Lan P, Chen WN, Chen WM (2011) Molecular modeling studies on imidazo[4,5-b]pyridine derivatives as Aurora A kinase inhibitors using 3D-QSAR and docking approaches. *Eur J Med Chem* 46:77–94
- Pirhadi S, Ghasemi JB (2010) 3D-QSAR analysis of human immunodeficiency virus entry-1 inhibitors by CoMFA and CoMSIA. *Eur J Med Chem* 45:4897–4903

Studies of H4R antagonists using 3D-QSAR, molecular docking and molecular dynamics

Jing Liu · Yan Li · Hui-Xiao Zhang · Shu-Wei Zhang ·
Ling Yang

Received: 21 April 2011 / Accepted: 23 May 2011 / Published online: 7 June 2011
© Springer-Verlag 2011

Abstract Three-dimensional quantitative structure–activity relationship studies were performed on a series of 88 histamine receptor 4 (H4R) antagonists in an attempt to elucidate the 3D structural features required for activity. Several *in silico* modeling approaches, including comparative molecular field analysis (CoMFA), comparative similarity indices analysis (CoMSIA), molecular docking, and molecular dynamics (MD), were carried out. The results show that both the ligand-based CoMFA model ($Q^2=0.548$, $R_{\text{ncv}}^2=0.870$, $R_{\text{pre}}^2=0.879$, $\text{SEE}=0.410$, $\text{SEP}=0.386$) and the CoMSIA model ($Q^2=0.526$, $R_{\text{ncv}}^2=0.866$, $R_{\text{pre}}^2=0.848$, $\text{SEE}=0.416$, $\text{SEP}=0.413$) are acceptable, as they show good predictive capabilities. Furthermore, a combined analysis incorporating CoMFA, CoMSIA contour maps and MD results shows that (1) compounds with bulky or hydrophobic substituents at positions 4–6 in ring A (R2 substituent), positively charged or hydrogen-bonding (HB) donor groups in the R1 substituent, and hydrophilic or HB acceptor groups in ring C show enhanced biological activities, and (2) the key amino acids in the binding pocket are TRP67, LEU71, ASP94, TYR95, PHE263 and GLN266. To our best knowledge, this work is the first to

report the 3D-QSAR modeling of these H4R antagonists. The conclusions of this work may lead to a better understanding of the mechanism of antagonism and aid in the design of new, more potent H4R antagonists.

Keywords 3D-QSAR · H4R antagonist · CoMFA · CoMSIA · MD · Docking

Abbreviations

QSAR	Quantitative structure–activity relationship
3D-QSAR	Three-dimensional quantitative structure–activity relationship
HR	Histamine receptors
H4R	Histamine receptor 4
GPCR	G-protein-coupled receptor
H4R	Histamine receptor 4
CADD	Computer-aided drug design
CoMFA	Comparative molecular field analysis
CoMSIA	Comparative similarity index analysis
MD	Molecular dynamics
HB	Hydrogen bond
RMSD	Root mean square deviation
SEE	Standard error of estimates
SEP	Standard error of prediction
Q^2	Cross-validated correlation coefficient after the leave-one-out procedure
R_{ncv}^2	Non-cross-validated correlation coefficient
F ratio of R_{ncv}^2	Explained to unexplained R_{ncv}^2 ratio = $R_{\text{ncv}}^2 / (1 - R_{\text{ncv}}^2)$
R_{pre}^2	Predicted correlation coefficient for the test set of compounds
OPN	Optimal number of principal components
PLS	Partial least squares
PCs	Principal components
LOO	Leave-one-out

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1137-x) contains supplementary material, which is available to authorized users.

J. Liu · Y. Li (✉) · H.-X. Zhang · S.-W. Zhang
Department of Materials Science and Chemical Engineering,
Dalian University of Technology,
Dalian 116012 Liaoning, China
e-mail: yanli@dlut.edu.cn

L. Yang
Lab of Pharmaceutical Resource Discovery,
Dalian Institute of Chemical Physics,
Graduate School of the Chinese Academy of Sciences,
Dalian 116023 Liaoning, China

Introduction

Histamine is a biogenic amine that affects a variety of functions in the human body by playing a role in inflammation, gastric acid secretion, and neurotransmission [1]. Thus, the pharmaceutical industry has a long-standing interest in exploring histamine receptors (HR) as therapeutic targets for the treatment of various diseases [2]. To date, four histamine receptor subtypes (H1R, H2R, H3R, H4R) have been identified, and all of them are members of the G-protein-coupled receptor (GPCR) family [3]. H1 and H2 antagonists have long been used in the treatment of inflammatory and gastric hyperacidity diseases, respectively [4]. H3R stimulates the release of histamine and other neurotransmitters from neurons as a presynaptic autoreceptor, and H3 antagonists are useful in cognitive and memory disorders and obesity [4]. Histamine receptor 4 (H4R), identified in 2000, mediates its effects by coupling to G α i/o G-proteins, and has low homology with other histamine receptors: only 35% amino acid identity with H3R (58% homology in its transmembrane regions), and a much lower identities with H1R and H2R [5]. Modulation of H4 receptor activity provides an opportunity to treat inflammatory and allergic conditions [6].

Since its relatively recent discovery, H4R has been the focus of much attention [7]. In contrast to other histamine receptors, H4R has a distinct expression profile on immune cells, including mast cells, eosinophils, dendritic cells and T cells, exerting modulatory effects on cell function. Moreover, H4R appears to play a significant role in multiple functions of these cells, such as activation, migration, cytokine and chemokine production [8]. This suggests that the receptor plays a crucial role in immunological and inflammatory processes [9]. H4R is involved in immune or inflammatory responses because histamine signaling induces changes in cell shape and chemotaxis of mast cells as well as eosinophils, mast cell migration, and upregulation of adhesion molecules on monocytes. All these effects can be blocked by H4R antagonists [10].

From the above physiological reactions, one can deduce several potential clinical uses for H4R inverse agonists/antagonists in the broad field of anti-inflammatory therapy, such as in treatments for allergy and asthma, pruritus associated with allergy or autoimmune skin conditions, inflammatory bowel disease, rheumatoid arthritis, and pain [10]. JNJ777120, which is the first non-imidazole H4R antagonist reported by Jablonowski et al. in 2003 [11], shows good selectivity over other histamine receptors that also have interesting anti-inflammatory activities in vivo [11]. Furthermore, it has become a standard reference agent for evaluating H4 receptor activity in many laboratories [12]. Recently, based on JNJ777120, a series of new H4R antagonists were synthesized by Altenbach et al. [12]. Due to the facts that they currently represent the largest data set

(88 compounds in total) on H4R antagonists, and that they have quite different molecular structures from other groups of H4R antagonists that have been developed, they attracted our particular interest for further quantitative structure–H4R antagonist potency relationship studies.

The addition of computer-aided drug design (CADD) technologies to the drug discovery and development process could lead to a reduction of at least 50% in the cost of drug design [13]. QSARs, especially 3D-QSAR, is a CADD technology that has been applied widely throughout the world to prioritize untested chemicals for more intensive and costly experimental evaluations [14]. Molecular docking and molecular dynamics are therefore being utilized more and more in current drug design processes. The aim of the present study was to use the 88 new fused compounds mentioned above as a data set to identify the structural features that lead to H4R antagonist effects, through a combination of several *in silico* approaches, including CoMFA, CoMSIA, molecular docking and molecular dynamics. As far as we know, this study is the first 3D-QSAR study of this new series of H4R antagonists.

Materials and methods

Dataset and biological activity

A series of 88 2-aminopyrimidine-containing H4R antagonists were synthesized by Robert J. Altenbach and his colleagues as the data set for the QSAR studies described in this paper [12]. Their pK_b values ($pK_b = -\lg K_b$) were employed as their biological activities (Tables S1–S3, “Electronic supplementary material”). Based on their skeleton structures, these compounds were divided into skeleton types A–C. There were 71 type A compounds (Table S1), 6 type B (Table S2) and 11 type C compounds (Table S3). To generate 3D-QSAR models, the molecules were divided into a training (66 compounds) and a test set (22 molecules) for validating the quality of the models, in the ratio 3:1. The test molecules were selected randomly such that the data set showed high structural diversity and a wide range of activities. All molecular studies were performed using the molecular modeling package SYBYL 6.9 (Tripos Associates, St. Louis, MO, USA). Energy minimization was performed using the Tripos force field and the conjugate gradient method, with the convergence criterion set to 0.05 kcal mol⁻¹ in this process. Partial atomic charges were calculated by the Gasteiger–Huckel method [15].

Conformational sampling and alignment

It is known that the appropriate superimposition of the molecules being studied within a three-dimensional fixed

lattice is the key procedure in further CoMFA and CoMSIA studies [16]. In our present work, based on an atom-by-atom superimposition principle, molecular alignment was carried out using the ALIGN DATABASE command in SYBYL. Both ligand-based and receptor-based alignment rules were adopted here. In both types of alignment, compound **56**, which had the highest pK_b value (of 8.86), was chosen as the template molecule. Figure 1a shows the common substructure depicted in red, and Fig. 1b shows the resulted ligand-based alignment model. The receptor-based alignment we used is shown in Fig. 1c. For this alignment, after the docking process, none of the conformations of the compounds that showed optimal scores with the H4R protein presented a statistically significant result. Therefore, the optimal conformation of each molecule was selected through leave-one-out (LOO) cross-validation [17], which ensures that the correlation coefficient $R^2=0.41$. Then, the partial atomic charges of the molecules were calculated by the Gasteiger–Hückel method [15].

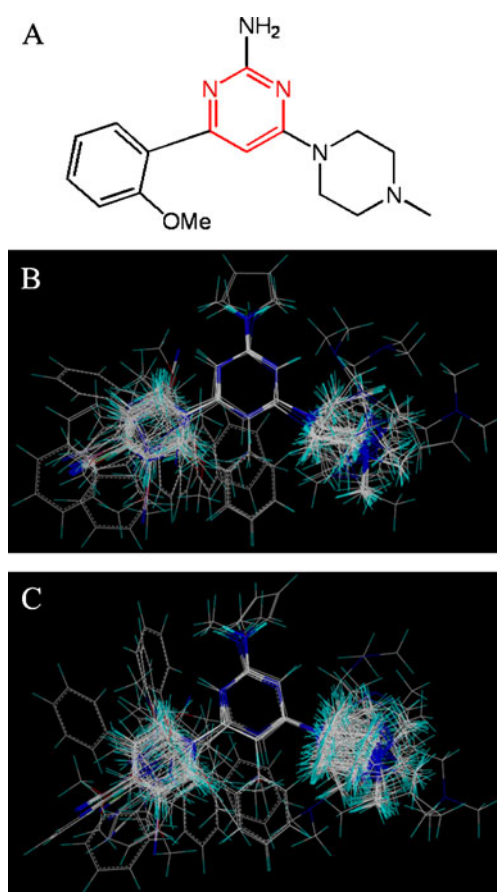


Fig. 1 Molecular alignments of all of the compounds in the data set. **a** The common substructure of the molecules is shown in red, based on template compound **56**. **b** Ligand-based alignments of all of the compounds. **c** Receptor-based alignments of all of the compounds

CoMFA and CoMSIA calculation

In CoMFA [18] analysis, all of the superimposed molecules were placed in a regular 3D lattice (2 Å spacing) extending at least 2 Å beyond the volumes of all investigated molecules on all axes. The van der Waals potentials and the Coulombic term representing the steric and electrostatic fields were calculated using the standard Tripos force field for CoMFA. A C_{sp3} atom with a formal charge of +1 and a van der Waals radius of 1.52 Å served as the 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 [19]. The cutoff value for both the steric and electrostatic fields was set to 30.0 kcal mol⁻¹, with a distance-dependent dielectric constant.

The CoMFA method only calculates the steric and electrostatic interactions, whereas CoMSIA [20] also includes the hydrophobic, HB donor and HB acceptor interactions. The basic assumption of CoMSIA is that appropriate sampling of the steric, electrostatic, hydrophobic and HB acceptor interactions generated around a set of aligned molecules with a probe atom can highlight all of the features relevant to their biological activities, and that changes in the binding affinities of the ligands are related to changes in molecular properties [21]. The aligned molecules were placed in a 3D lattice with regular grid points separated by 2 Å, similar to the lattice used in CoMFA studies. CoMSIA similarity index descriptors were also derived within a lattice box with a grid spacing of 2 Å and a C_{sp3} atom with a charge of +1 was used as the probe atom. A Gaussian function was used to evaluate the distance between the probe atom and each atom in the molecule. CoMSIA similarity indices (A_F) for a molecule j with an atom i at a grid point q were calculated as follows:

$$A_{F,k}^q(j) = - \sum \omega_{probe,k} \omega_{ik} e^{-\alpha r_{iq}^2}, \quad (1)$$

where $\omega_{probe,k}$ is the probe atom with a radius of 1 Å, a charge of +1, a hydrophobicity of +1, hydrogen bond donation of +1, and hydrogen bond acceptance of +1. ω_{ik} is the actual value of the physicochemical property k of atom i . r_{iq} is the distance between the probe atom at grid point q and item i of the test molecule [22].

Calculation and validation of the 3D-QSAR models

CoMFA and CoMSIA descriptors were used as the independent variables, and the corresponding pK_b values as the dependent variables, in the partial least squares (PLS) regression analyses performed during 3D-QSAR model development. The advantage of this method is that it can

reduce an initially large number of descriptors to a few principal components (PCs) that are linear combinations of the original descriptors [23]. In our PLS analysis, the cross-validation was performed using the leave-one-out (LOO) method, wherein one compound is removed from the dataset and its activity is predicted using the model derived from the rest of the dataset. The cross-validated Q^2 obtained from the PCs was considered [24]. Using this, a non-cross-validation analysis was then carried out, and the Pearson coefficient (R_{ncv}^2) and SEE were calculated [23].

During the PLS process, several statistical parameters, including Q^2 and R_{ncv}^2 , are needed to evaluate the reliability of the model generated. The cross-validated coefficient Q^2 is used as a statistical index of the predictive power of the model, and is calculated by the following equation, where the parameters $Y_{\text{predicted}}$, Y_{observed} and Y_{mean} are the predicted, actual and mean values of the target property, respectively [21]:

$$q^2 = 1 - \frac{\sum_Y (Y_{\text{predicted}} - Y_{\text{observed}})^2}{\sum_Y (Y_{\text{observed}} - Y_{\text{mean}})^2}. \quad (2)$$

When assessing the predictive power of the QSAR model derived from the training set, an independent test set was used, and their biological activities were predicted using this QSAR model. The predicted R^2 (R_{pre}^2) value was calculated using

$$R_{\text{pre}}^2 = (\text{SS} - \text{PRESS})/\text{SS}, \quad (3)$$

where SS is the sum of the squared deviations between the biological activity of the test set and the mean activity of the training set molecules, and PRESS is the sum of the squared deviations between the actual and the predicted activities of the test set molecules [25]. Finally, the CoMFA/CoMSIA results were presented graphically on field contour maps, where the coefficients were generated using the field type “Stdev*Coeff.”

Molecular docking

Molecular docking is the method most commonly used to calculate protein–ligand interactions, and it is efficient at predicting the potential ligand binding site(s) on the whole protein target [26]. In order to find the probable binding conformations and offer more insight into the interactions between H4R and its antagonist, molecular docking analysis was carried out using Surflex-Dock in the SYBYL package in our present work. In the docking process, an accurate 3D structure of the receptor is important. Due to the unavailability of the X-ray structure of H4R, in the present work we used a homology model built by Armin Buschauer et al. [9]. In their work, a satisfactory H4R

homology model was built based on a template of human $\beta 2$ -adrenoceptor, and optimized by an MD process [9]. Currently, most standard docking protocols incorporate ligand flexibility into the docking process while considering the protein to be a rigid structure [27]. Our molecular docking involves the following steps. (1) The protein structure was imported into Surflex and then hydrogens were added. (2) The protomol was generated using a ligand-based approach. During the protomol generation process, two particular parameters must be specified to form the appropriate binding pocket. One parameter, called *protomol_bloat*, determines how far the site should extend from a potential ligand; the other is the *protomol_threshold*, which determines how deep the atomic probes that are used to define the protomol can penetrate into the protein. In the present work, *protomol_bloat* was set to 0, and *protomol_threshold* was set to 0.50 when a reasonable binding pocket was obtained. (3) All of the antagonists were docked into the binding pocket, and 20 possible active docking conformations with different scores were obtained for each antagonist. During the docking process, all of the other parameters were assigned their default values.

Molecular dynamics simulations

The MD simulations were performed with the GROMACS software package [28] using the GROMOS96 force field [29]. The molecular topology file for the ligand in protein was generated by the program PRODRG 2.5 [30–33]. The simulation cell was a cubic periodic box with a side length of 95.99 Å, and the minimum distance between the protein and the walls of the box was set to 10 Å. In order to neutralize the total charge, seven chloride ions were placed randomly in the box. The total number of atoms in the simulated system was 87,326, including the protein complexes and waters. The remaining box volume was filled with simple point charge (SPC) waters [34].

Prior to the simulation, energy minimization was performed for the full system without constraints using the steepest descent integrator for 7500 steps, and then the system was equilibrated via 200 ps MD simulations at 300 K. Finally, a 5 ns simulation was performed with a time step of 2 fs. During the MD simulations, the standard parameters and main calculation methods were configured as follows. The model used an NPT ensemble at 300 K with periodic boundary conditions. The temperature was kept constant by the Berendsen thermostat, and the values of isothermal compressibility were set to $4.5 \times 10^{-5} \text{ bar}^{-1}$, while the pressure was maintained at 1 bar using the Parrinello–Rahman scheme [35]. Electrostatic interactions were calculated using the particle mesh Ewald method [36], and the cut-off distances for calculating Coulomb and van der Waals interactions were 1.0 and 1.4 nm, respectively.

All of the MD simulations lasted 5 ns to ensure that the whole systems were stable.

Results and discussion

CoMFA and CoMSIA statistical results

In our present work, all CoMFA and CoMSIA models were derived using the same training (66 molecules) and test (22 molecules) sets. To validate the reliability of these models, all of the vital statistical parameters were analyzed, including the leave-one-out Q^2 , the non-cross-validated correlation coefficient (R_{ncv}^2), SEE, F -statistic values, and the predicted correlation coefficient (R_{pre}^2). The statistical results obtained from standard CoMFA models constructed with steric and electrostatic fields are summarized in Table 1. For the ligand-based model, the optimum number of components (five) was determined by SAMPLS analysis implemented in SYBYL with a LOO cross-validated Q^2 of 0.548, which indicated that the model had good predictive capability. A high R_{ncv}^2 of 0.870 for the non-cross-validated final model showed the self-consistency of the model. The SEE was 0.410, and the F value was 79.992. When tested with the independent test set, the ligand-based CoMFA model exhibited satisfactory predictive ability, with $R_{pre}^2=0.879$ and $SEP=0.386$. In this ligand-based CoMFA model, the electrostatic features was found to contribute more to the activity (~53%) than the steric feature. For the receptor-

based model, a cross-validated Q^2 of 0.392 was obtained using five optimum components, demonstrating the poor predictive ability of the model. This failure is also seen when the external predictive capability of the receptor-based CoMFA model was evaluated with the independent test set, ending up with statistical results of $R_{pre}^2=0.248$ and $SEP=0.902$. Thus, the receptor-based model failed completely, and will not be discussed further in the present work.

As mentioned above, CoMSIA not only offered the same steric and electrostatic field information as CoMFA, but it also provided hydrophobic and hydrogen-bond (HB) donor and acceptor interaction information, which are all always relevant to the binding affinities [19]. Thus, different combinations of these additional three interaction fields with the steric and electrostatic ones may result in other useful or even better QSAR models. So, in our CoMSIA modeling process, all five field descriptors were calculated using the same data sets as used in CoMFA analysis, and they were then fitted together in every possible form to build appropriate CoMSIA models. Finally, the best ligand- and receptor-based models were obtained with the highest Q^2 values using the steric, electrostatic, hydrophobic and HB acceptor parameters (Table 1). The ligand-based CoMSIA model has a Q^2 value of 0.526 with five optimum components, an R_{ncv}^2 value of 0.866, a SEE value of 0.416, and an F value of 77.575. However, in this model, unlike in the ligand-based CoMFA model, the electrostatic feature provided the major contribution to the antagonist activity (~32.7%). Again, poor predictive results were obtained for the receptor-based CoMSIA model, with a Q^2 value of 0.269 using three optimum components, an R_{ncv}^2 value of 0.682, a SEE value of 0.630, and an F value of 44.238.

Normally, 3D-QSAR studies with a cross-validated Q^2 value >0.5 are considered to be statistically significant [37]. In addition, higher R_{ncv}^2 and F values as well as lower SEE values should also be considered the foundation of a reliable 3D-QSAR model. However, using the widely accepted LOO cross-validated Q^2 alone is insufficient to assess the predictive power of a QSAR model [38]. Thus, to validate the above four models (especially the two optimal ones—the ligand-based CoMFA and CoMSIA models), an external test set (22 molecules) that was independent of and represented 33.3% of the training set was used to predict the activities (pK_b values) of the compounds in it.

Before this validation, an initial inspection of the fitted/predicted activities revealed poor predictions for two compounds (**33** and **35**) which were considered outliers in our work for both the ligand-based and receptor-based models. Careful examination of outliers may provide additional information on their peculiarities; therefore, in this study, the two outliers were checked carefully. Compounds **33** and **35** had comparatively high residuals

Table 1 Summary of the CoMFA and CoMSIA results

PLS statistics	Ligand-based model		Receptor-based model	
	CoMFA	CoMSIA	CoMFA	CoMSIA
Q^2	0.548	0.526	0.392	0.269
R_{ncv}^2	0.870	0.866	0.909	0.682
SEE	0.410	0.416	0.343	0.630
F	79.992	77.575	119.738	44.238
R_{pre}^2	0.879	0.848	0.248	0.200
SEP	0.386	0.413	0.902	0.956
OPN	5	5	5	3
Contribution:				
Steric	0.47	0.095	0.512	0.080
Electrostatic	0.53	0.327	0.488	0.298
Hydrophobic		0.295		0.252
HB acceptor		0.282		0.370

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 explained to unexplained $R_{ncv}^2 = R_{ncv}^2 / (1 - R_{ncv}^2)$; R_{pre}^2 predicted correlation coefficient for the test set of compounds; SEP standard error of prediction; OPN optimal number of principal components

between the experimental and predicted activities—absolute pK_b residual values of 2.525 and 2.811 for the optimal CoMFA model, and 2.206 and 2.553 for the optimal CoMSIA model, respectively (both larger than 2)—and were thus treated as outliers in the model. This discrepancy suggests the need to recruit more (and more accurate) experimental data with more diverse molecular structures. After eliminating the outliers, both the CoMFA and CoMSIA models obtained from the ligand-based alignment exhibited good predictive abilities (with $Q^2 > 0.5$, $R_{pre}^2 > 0.84$, Table 1), indicating that the ligand-based alignment rule is more appropriate for the QSAR study of this dataset. However, for receptor-based alignment, both the CoMFA and the CoMSIA models yielded unsatisfactory predictions (CoMFA: $Q^2 = 0.269$, $R_{pre}^2 = 0.200$; CoMSIA: $Q^2 = 0.392$, $R_{pre}^2 = 0.200$).

The observed and predicted H4R inhibitory activities for both the ligand- and receptor-based CoMFA and CoMSIA models are shown in Table S4. Figure 2 shows the actual and predicted pK_b values plotted against each other for both the training (filled black squares) and test (filled blue diamonds) set molecules of the whole dataset, based on the ligand-based CoMFA and CoMSIA models. It is clear that all of the points are rather uniformly distributed around the regression line in the two figures, and the predicted activities are almost as accurate as the experimental data, indicating good correlation between the predicted and experimental activities of the dataset and the reliability of the obtained models.

3D-QSAR contour maps

Based on the ligand-based optimal CoMFA and CoMSIA models, various contour maps were constructed here to show the important features of the ligands. The results of the 3D-QSAR models were mapped using the “StDev*Coeff” mapping option “contoured by contribution.” The default levels of contour by contribution (80 for favored regions and 20 for disfavored regions) were employed during the contour analysis. In this study, compound **56** (Fig. 3), which was one of the most active compounds in the whole dataset (pK_b value of 8.86), is used as an example molecule in all subsequent CoMFA and CoMSIA contour maps (Figs. 4 and 5).

The steric and electrostatic fields from the best CoMFA model are shown in Fig. 4. Areas where steric bulk substituents increase the potency are represented by green polyhedrons, while areas where steric bulk substituents decrease the potency are represented by yellow polyhedrons (Fig. 4a). A large green contour was found near positions 4–6 in ring A (R2 substituent). Thus, molecules carrying a bulky substituent at these positions should be more active than those with no or a smaller

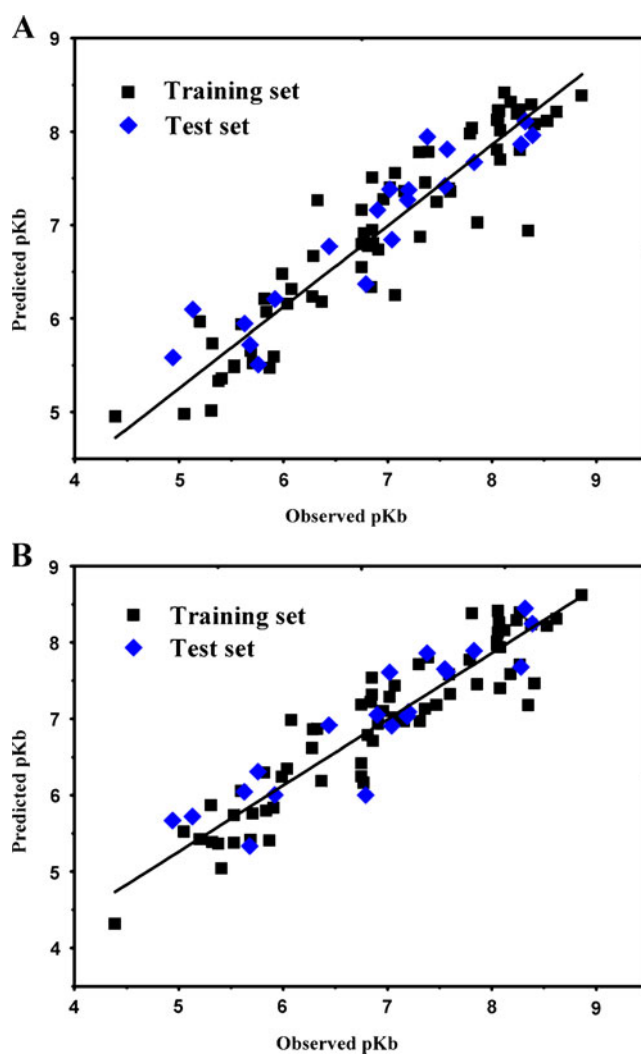


Fig. 2 Ligand-based correlation plots of the predicted versus the actual pK_b values for the training (filled black squares) and the test (filled blue diamonds) set compounds, based on **a** the CoMFA model and **b** the CoMSIA model, respectively

substituent, as illustrated by the higher potencies of compounds **3** ($pK_b = 8.53$) and **58** ($pK_b = 8.32$), which have a bulky substituent (4-CN-Ph), than the potencies of molecules **20–25** (with pK_b values of < 7.07), which have

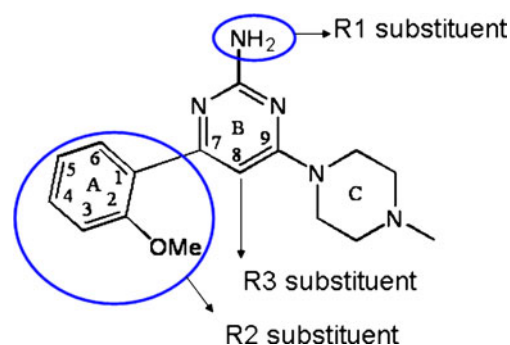


Fig. 3 The structure of compound **56** [12]

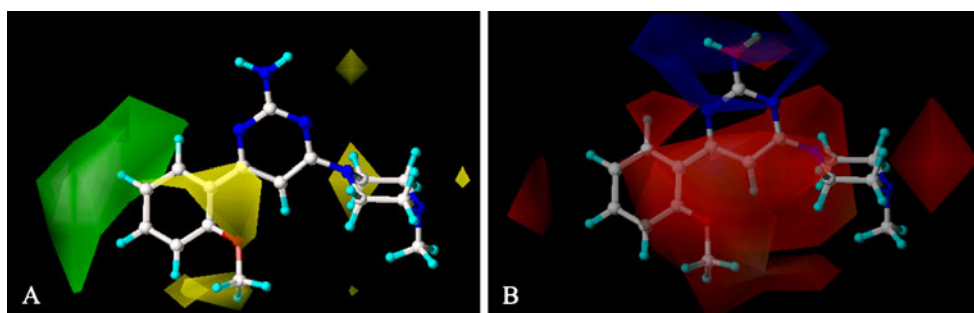


Fig. 4 CoMFA “StDev*Coeff” contour plots. **a** Steric (*green/yellow*) contour map in combination with compound **56**. *Green contours* indicate regions where bulky groups increase activity; *yellow contours* indicate regions where bulky groups decrease activity. **b** Electrostatic

contour map (*red/blue*) in combination with compound **56**. *Red contours* indicate regions where negatively charged groups increase activity; *blue contours* indicate regions where positively charged groups increase activity

the group *t*-Bu at this position. In contrast, three negative steric (*yellow*) regions appear mainly above positions 2 and 3 of ring A and positions 7 and 9 of ring B. This suggests that a bulky substituent at this position degrades the biological activities of the molecules, as illustrated by the fact that compounds **68** ($pK_b=6.85$) and **69** ($pK_b=6.79$) at positions 2 and 3 of ring A exhibited lower activities than those of **38** ($pK_b=8.28$) and **40** ($pK_b=7.55$), which have the substituents $-NHMe$ and $-NEt_2$ at the same position, respectively.

The CoMFA electrostatic contour plots for the highly active compound **56** are displayed in Fig. 4b. A blue contour indicates that the substituents should be electron deficient for high binding affinity with the protein, while a red color indicates that they should be electron rich to achieve this binding affinity. A large red contour appears around the R1 substituent, indicating active site favorable regions of high electron density, which explains the decreased potencies of compounds **23** ($pK_b=5.31$) and **24** ($pK_b=4.39$) considering the electronegative groups ($-Cl$

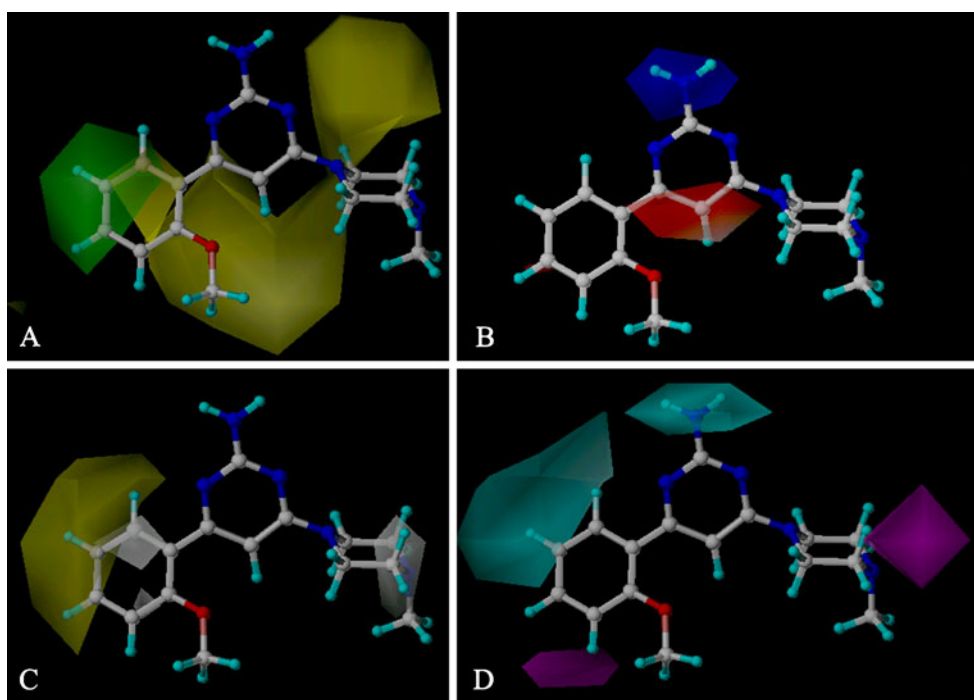


Fig. 5 CoMSIA “StDev*Coeff” contour maps. **a** Steric (*green/yellow*) contour map in combination with compound **56**. *Green contours* indicate regions where bulky groups increase activity; *yellow contours* indicate regions where bulky groups decrease activity. **b** Electrostatic contour map (*red/blue*) in combination with compound **56**. *Red contours* indicate regions where negative charges increase activity; *blue contours* indicate regions where positive charges increase activity. **c** Hydrophobic contour map (*yellow/white*) in combination with

compound **56**. *Yellow contours* indicate regions where hydrophobic substituents enhance activity; *white contours* indicate regions where hydrophilic substituents enhance activity. **d** HB acceptor contour map (*magenta/cyan*) in combination with compound **56**. *Magenta contours* indicate regions where HB acceptors on the receptor promote the affinity; *cyan contours* indicate regions where HB acceptors on the receptor degrade the affinity

and –OH) that they have in these areas. A large red isopleth around the R2 substituent, ring B and ring C shows that this area prefers negatively charged substituents. Due to the strong electronegativities of nitrogen and oxygen atoms, the activities of compounds **3** ($pK_b=8.53$), **12** ($pK_b=8.08$) and **56** ($pK_b=8.86$) are greater than those of compounds **15–19** with a *t*-Bu substituent in R2 substituents (their pK_b values are smaller than 7.02).

In our study, the CoMSIA model not only calculates the steric and electrostatic fields, but also uses the hydrophobic and HB acceptor fields to correlate with the antagonist activity. All contour maps of the four CoMSIA fields are shown in Fig. 5. The color scheme used in the CoMSIA steric and electrostatic field contour maps (Fig. 5a and b) is the same as that described for the CoMFA contour maps. The steric contour map of the CoMSIA model (Fig. 5a) shows similar results to that of the CoMFA one, with only a subtle difference in that all of the negative steric (yellow) regions located above ring A (R2 substituent) in the CoMSIA model are much larger in size than those in the CoMFA one. These results lead to the conclusion that compounds with bulky substituents in ring A and ring C may possess enhanced activity. Figure 6b shows the electrostatic contour map of the optimal CoMSIA model, where the red polyhedron is clearly smaller than in the CoMFA one in size. One small red contour is also observed near position 8 (R1 substituent), indicating that occupancy by an electronegative substituent in this region would promote the binding affinity to H4R.

Figure 5c depicts the hydrophobic contour maps of the CoMSIA models. Yellow contours encompass regions where a hydrophobic group will improve biological activity, while a hydrophobic group located near the white regions will result in impaired biological activity. A large yellow contour was found near the positions 4–6 in ring A (R2 substituent). Therefore, molecules carrying hydrophobic groups (like –OMe, –OEt,

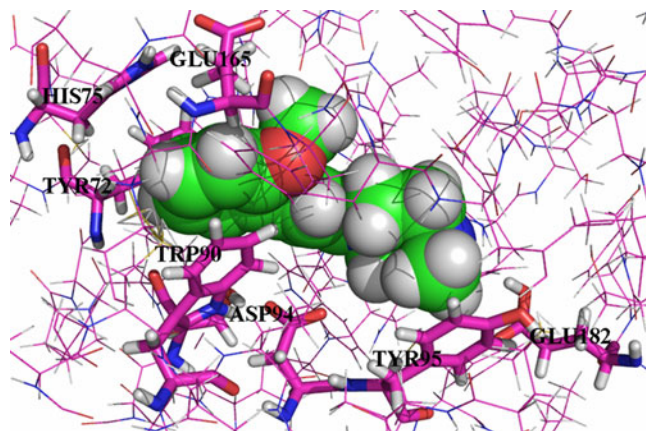


Fig. 6 The binding pocket in H4R. Molecule **56** is shown in *green*, and all key amino acids are shown in *pink*, respectively

–F, –Cl, –Br) tend to show enhanced activity, as exemplified by the higher activity of molecule **58** ($pK_b=8.32$), which has a Cl at position 4 of ring A. Two small white polyhedra appear below the plane of ring A, and one small white contour appears around ring C, indicating that hydrophilic (like hydroxy or amido) groups here are correlated with good antagonist activity.

The HB acceptor fields based on PLS analysis of the CoMSIA models are shown in Fig. 5d, where the magenta and cyan contours highlight areas in which HB acceptors and donors are preferred, respectively. One large cyan contour was found near the positions 4–6 in ring A, indicating that an HB donor at these positions may improve the activity. The other cyan contour that appears around R1 also has the same meaning; for instance, compounds with –NH₂ at the R1 position may have higher activities than those without an –NH₂ group at this position (compounds **20–25**). One small magenta contour appearing near position 3 in ring A, and another magenta contour appearing around ring C indicate that HB acceptors are preferred here. For example, compounds **52** ($pK_b=8.62$) with –OMe and **65** ($pK_b=8.06$) with –CN at position 4 of ring A possess higher activities than most of the other molecules in the data set. Also, the presence of an N atom in ring C, which acts as an HB acceptor, can improve the activity.

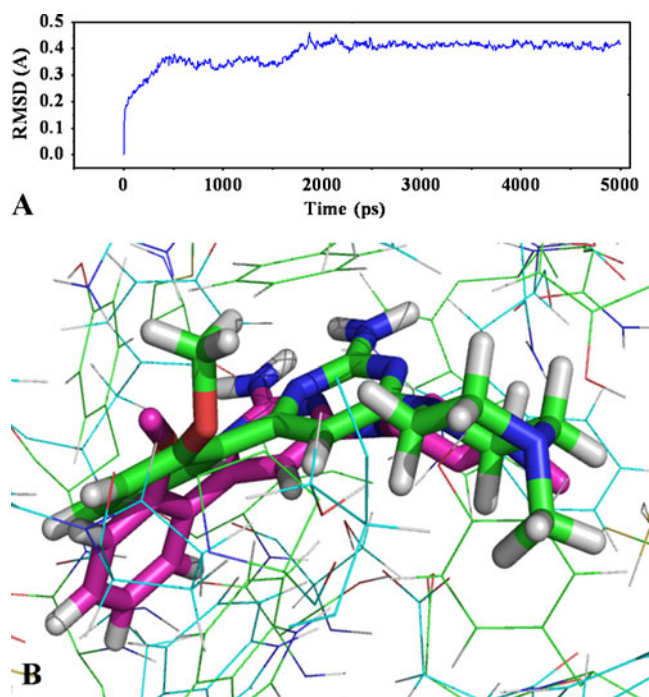
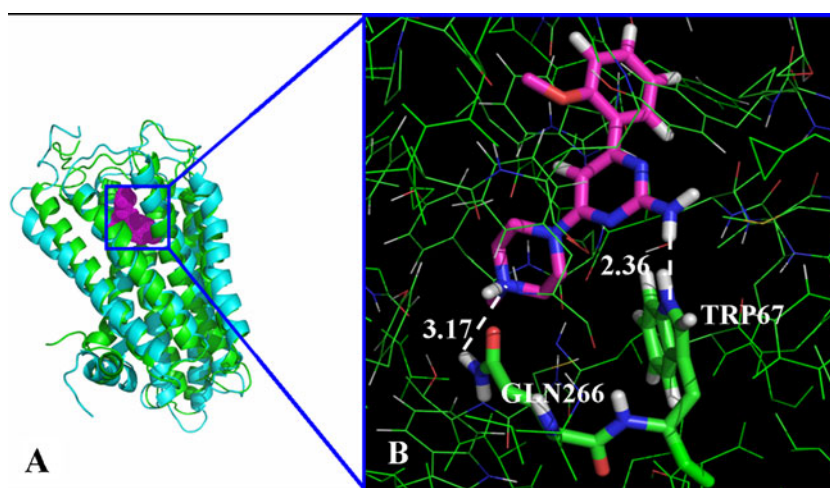


Fig. 7 MD simulation results. **a** Plot of the RMSD of docked complex versus MD simulation time in the MD-simulated structures. **b** View of superimposed backbone atoms of the average structure of the last 1000 ps of the MD simulation (*pink*) and the initial structure (*green*) for compound **56** and the H4R complex

Fig. 8 The binding pocket formed around molecule **56**. **a** Superposition of the MD simulation (*green*) and the initial structure (*cyan*) for H4R. *Pink dotted region* is the binding pocket of compound **56**. **b** HB interactions around the compound in the active docking pocket. The *black dotted line* shows the HBs formed and their bond lengths



Docking results

Docking protocols are widely used to predict the binding affinities of a number of ligands [39]. In this study, in order to explore the binding environment in which the ligand interacts within the H4R, docking studies were performed on these compounds. After all 88 compounds had been docked into the possible active site, we found that the highest CScore was 5.22. Figure 6 shows the binding pocket we generated (with molecule **56**, used as a template, shown in green). It is clear that this binding pocket is the same as that constructed by Buschauer et al. [9]. It is easily shown that many of the key amino acids (such as TYR72, HIS75, TRP90, ASP94, TYR95, GLU165 and GLU182) that interact with the H4R antagonist at the binding site are the same as those observed in the work of Armin Buschauer and his colleagues. All of the above results indicate that the binding pocket we found was appropriate for the study of H4R receptor antagonists.

Molecular dynamics simulations

In this study, a 5000 ps molecular dynamics simulation was carried out using H4R with ligand **56**, based on the docked complex structure. The main purpose of the simulation was to optimize the binding pocket and the correlation between H4R and molecule **56**. The root mean square deviation (RMSD) of the trajectory with respect to their initial structure ranged from 0.3 to 0.45 Å, as depicted in Fig. 7a. After 2000 ps, the RMSD of the complex was about 0.4 Å, and it almost remained at this level for the whole simulation process. This indicates that the docked complex structure is stable after 2000 ps of simulation [40]. A superposition of the average structure of the ensemble for the last 1 ns and the docked structure are shown in Fig. 7b, where the blue ribbon represents the initial structure of the docked complex, and the green ribbon represents the MD-

simulated structure, respectively. Compound **56** is represented in green for the initial complex and pink for the final average complex, respectively. In Fig. 7b, it is clear that the docked complex and the MD average structure occur at the same binding site, and their ring C regions are very similar.

As it was the most potent antagonist, compound **56** was chosen to illustrate the analysis of the MD results. In Fig. 8, the backbone $\text{-NH}_2\text{-}$ in GLN266 may form a hydrogen bond with the -N atom (HB acceptor) in ring C, at a distance of 3.19 Å. This binding mode also shows a strong

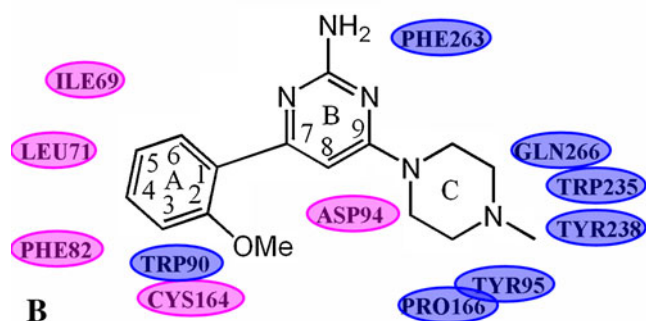
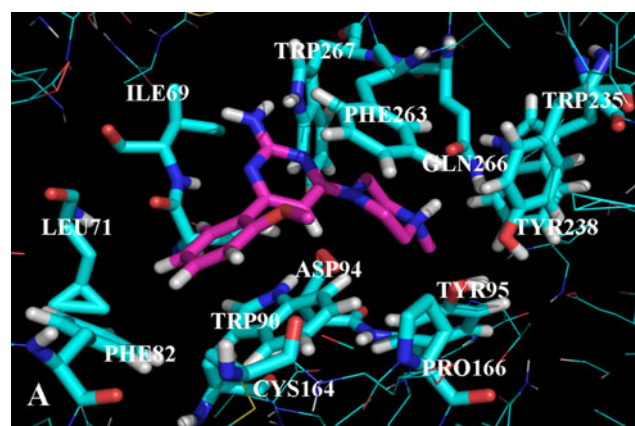


Fig. 9 The binding site formed around compound **56**. **a** Steric amino acid residues around the compound. **b** Positions of hydrophobic (*pink*) and hydrophilic (*blue*) amino acid residues

HB interaction between the N atom (HB donor) in the R1 substituent and the backbone –NH– in TRP67, at a distance of 2.36 Å. These MD results are very consistent with the results of our CoMSIA HB acceptor field contour map analysis, which shows that a compound with an HB acceptor group in ring C and an HB donor in the R1 substituent should show enhanced biological activity.

The steric amino acid residues around the compound at the binding site are shown in Fig. 9a. Clearly, no steric amino acid residues appear around positions 4–6 in ring A (R2 substituent). However, several crucial amino acid residues are found near some specific positions in the molecules. For example, TRP90, ASP94, TYR95 and CYS164 lie near position 2 (ring A) and positions 7 and 9 in ring B. These results further confirm our results from the CoMFA model (Fig. 4a), where bulky substituents at positions 4–6 improved the activity but bulky groups at position 2 of ring A and positions 7 and 9 of ring B impaired the activity. Furthermore, it is clear that the aromatic residue in PHE263 can participate in a π -stacking interaction with ring C in molecule 56. Thus, we can conclude that ring C plays an important role in this binding pocket, due to its ability to form hydrogen-bond and π -stacking interactions with the residues.

In Fig. 9b, hydrophobic amino acid residues ILE69, LEU71, PHE82 and CYS164 appear above the R2 substituent, especially at positions 4–6 of ring A, indicating that molecules with hydrophobic groups in these areas may possess higher binding affinities to H4R. The hydrophilic amino acid residues TYR95, PRO166, TRP235, TYR238 and GLN266 around ring C suggest that compounds with hydrophobic groups in this region may show reduced activity. These MD results correspond well to those from our previous CoMSIA analysis: in Fig. 5c (the hydrophobic field contour map), the yellow isopleth around positions 4–6 in ring A implies that the presence of hydrophobic groups in this region favor activity, and the white polyhedra that appear near ring C indicate a preference for hydrophilic groups in these areas.

These conclusions are highly consistent with the findings obtained from the CoMFA and CoMSIA contour maps, in that bulky or hydrophobic substituents at positions 4–6 of ring A can interact with the receptor, because they may well fit into the binding pocket. Hydrophilic groups on ring C tend to enhance the activity. In addition, an R1 substituent with an HB donor and an HB acceptor on ring C promote high activity.

Conclusions

In this paper, two 3D-QSAR models were built using the CoMFA and CoMSIA methods for the first time, utilizing a

total of 88 H4R antagonists. The optimal ligand-based models obtained exhibited good predictive abilities according to their Q^2 , R_{ncv}^2 , and R_{pre}^2 values. Furthermore, our MD results correlated well with those from the 3D-QSAR models. By analyzing both of the models and the derived contour maps, significant regions that influence the potency of H4R antagonists were identified: (1) bulky or hydrophobic substituents at positions 4–6 of ring A (R2 substituent) can enhance the biological activities of these compounds; (2) positively charged groups or HB donor groups on the R1 substituent may improve the binding affinity; (3) the presence of hydrophilic substituents or HB acceptor groups on ring C increases activity, and; (4) the key amino acids are TRP67, LEU71, ASP94, TYR95, PHE263 and GLN266. The correlations among the results obtained from QSAR, docking and MD studies should lead to a better understanding of the structural features needed for enhanced activity, and aid in the design of new, more potent H4R antagonists.

Acknowledgments We gratefully acknowledge Armin Buschauer and his colleagues for providing us with the homology model of the H4R protein structure. This work is supported by the National Natural Science Foundation of China (grant no. 10801025).

References

1. Hsieh GC, Chandran P, Salyers AK, Pai M, Zhu CZ, Wensink EJ, Witte DG, Miller TR, Mikusa JP, Baker SJ, Wetter JM, Marsh KC, Hancock AA, Cowart MD, Esbenshade TA, Brioni JD, Honore P (2010) H4 receptor antagonism exhibits anti-nociceptive effects in inflammatory and neuropathic pain models in rats. *Pharmacol Biochem Behav* 95:41–50
2. Crane K, Shih DT (2004) Development of a homogeneous binding assay for histamine receptors. *Anal Biochem* 335:42–49
3. Smits RA, Leurs R, de Esch IJ (2009) Major advances in the development of histamine H4 receptor ligands. *Drug Discov Today* 14(15–16):745–753
4. Kiss R, Noszal B, Racz A, Falus A, Eros D, Keseru GM (2008) Binding mode analysis and enrichment studies on homology models of the human histamine H4 receptor. *Eur J Med Chem* 43:1059–1070
5. Hsieh GC, Chandran P, Salyers AK, Pai M, Zhu CZ, Wensink EJ, Witte DG, Miller TR, Mikusa JP, Baker SJ, Wetter JM, Marsh KC, Hancock AA, Cowart MD, Esbenshade TA, Brioni JD, Honore P (2009) H4 receptor antagonism exhibits anti-nociceptive effects in inflammatory and neuropathic pain models in rats. *Pharmacol Biochem Behav* 95:41–50
6. Lee-Dutra A, Arienti KL, Buzard DJ, Hack MD, Khatuya H, Desai PJ, Nguyen S, Thurmond RL, Karlsson L, Edwards JP, Breitenbucher JG (2006) Identification of 2-arylbenzimidazoles as potent human histamine H4 receptor ligands. *Bioorg Med Chem Lett* 16:6043–6048
7. Cramp S, Dyke HJ, Higgs C, Clark DE, Gill M, Savy P, Jennings N, Price S, Lockey PM, Norman D, Porres S, Wilson F, Jones A, Ramsden N, Mangano R, Leggate D, Andersson M, Hale R (2010) Identification and hit-to-lead exploration of a novel series of histamine H4 receptor inverse agonists. *Bioorg Med Chem Lett* 20:2516–2519

8. Zhang M, Thurmond RL, Dunford PJ (2007) The histamine H4 receptor: A novel modulator of inflammatory and immune disorders. *Pharmacol Therapeut* 113:594–606
9. Igel P, Geyer R, Strasser A, Dove S, Seifert R, Buschauer A (2009) Synthesis and structure–activity relationships of cyanoguanidine-type and structurally related histamine H4 receptor agonists. *J Med Chem* 52:6297–6313
10. Sander K, Kottke T, Tanrikulu Y, Proschak E, Weizel L, Schneider EH, Seifert R, Schneider G, Stark H (2009) 2,4-Diaminopyrimidines as histamine H4 receptor ligands—scaffold optimization and pharmacological characterization. *Bioorg Med Chem* 17:7186–7196
11. Jablonowski JA, Grice CA, Chai W, Dvorak CA, Venable JD, Kwok AK, Ly KS, Wei J, Baker SM, Desai PJ, Jiang W, Wilson SJ, Thurmond RL, Karlsson L, Edwards JP, Lovenberg TW, Carruthers NI (2003) The first potent and selective non-imidazole human histamine H4 receptor antagonists. *J Med Chem* 46:3957–3960
12. Altenbach RJ, Adair RM, Bettencourt BM, Black LA, Fix-Stenzel SR, Gopalakrishnan SM, Hsieh GC, Liu H, Marsh KC, McPherson MJ, Milicic I, Miller TR, Vortherms TA, Warrior U, Wetter JM, Wishart N, Witte DG, Honore P, Esbenshade TA, Hancock AA, Brioni JD, Cowart MD (2008) Structure–activity studies on a series of a 2-aminopyrimidine-containing histamine H4 receptor ligands. *J Med Chem* 51:6571–6580
13. Taft CA, Da Silva VB, Da Silva CH (2008) Current topics in computer-aided drug design. *J Pharm Sci* 97:1089–1098
14. Papa E, Battaini F, Gramatica P (2005) Ranking of aquatic toxicity of esters modelled by QSAR. *Chemosphere* 58:559–570
15. Gasteiger J, Marsili M (1980) Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* 36:3219–3228
16. Artico M, Botta M, Corelli F, Mai A, Massa S, Ragno R (1996) Investigation on QSAR and binding mode of a new class of human rhinovirus-14 inhibitors by CoMFA and docking experiments. *Bioorg Med Chem* 4:1715–1724
17. Rao CR, Wu Y (2005) Linear model selection by cross-validation. *J Stat Plan Infer* 128:231–240
18. Wang X, Yang W, Xu X, Zhang H, Li Y, Wang Y (2010) Studies of benzothiadiazine derivatives as hepatitis C virus NS5B polymerase inhibitors using 3D-QSAR, molecular docking and molecular dynamics. *Curr Med Chem* 17:2788–2803
19. Zhu YQ, Lei M, Lu AJ, Zhao X, Yin XJ, Gao QZ (2009) 3D-QSAR studies of boron-containing dipeptides as proteasome inhibitors with CoMFA and CoMSIA methods. *Eur J Med Chem* 44:1486–1499
20. Wang Z, Li Y, Ai C, Wang Y (2010) In silico prediction of estrogen receptor subtype binding affinity and selectivity using statistical methods and molecular docking with 2-arylnaphthalenes and 2-arylquinolines. *Int J Mol Sci* 11:3434–3458
21. Li Y, Wang Y-H, Yang L, Zhang S-W, Liu C-H, Yang S-L (2005) Comparison of steroid substrates and inhibitors of P-glycoprotein by 3D-QSAR analysis. *J Mol Struct* 733:111–118
22. Xu M, Zhang A, Han S, Wang L (2002) Studies of 3D-quantitative structure–activity relationships on a set of nitro-aromatic compounds: CoMFA, advanced CoMFA and CoMSIA. *Chemosphere* 48:707–715
23. Kovalishyn VV, Kholodovych V, Tetko IV, Welsh WJ (2007) Volume learning algorithm significantly improved PLS model for predicting the estrogenic activity of xenoestrogens. *J Mol Graph Model* 26:591–594
24. Nayana MR, Sekhar YN, Nandyala H, Muttineni R, Bairy SK, Singh K, Mahmood SK (2008) Insight into the structural requirements of proton pump inhibitors based on CoMFA and CoMSIA studies. *J Mol Graph Model* 27:233–243
25. Thaimattam R, Daga P, Rajjak SA, Banerjee R, Iqbal J (2004) 3D-QSAR CoMFA, CoMSIA studies on substituted ureas as Raf-1 kinase inhibitors and its confirmation with structure-based studies. *Bioorg Med Chem* 12:6415–6425
26. Shahlaei M, Madadkar-Sobhani A, Mahnam K, Fassihi A, Saghale L, Mansourian M (2011) Homology modeling of human CCR5 and analysis of its binding properties through molecular docking and molecular dynamics simulation. *Biochem Biophys Acta* 1808:802–817
27. Sander T, Liljefors T, Balle T (2008) Prediction of the receptor conformation for iGluR2 agonist binding: QM/MM docking to an extensive conformational ensemble generated using normal mode analysis. *J Mol Graph Model* 26:1259–1268
28. Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: a message-passing parallel molecular dynamics implementation. *Comput Phys Commun* 91:43–56
29. 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
30. van Aalten DM, Bywater R, Findlay JB, Hendlich M, Hooft RW, 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
31. Barreca ML, Ortuso F, Iraci N, De Luca L, Alcaro S, Chimirri A (2007) Tn5 transposase as a useful platform to simulate HIV-1 integrase inhibitor binding mode. *Biochem Biophys Res Commun* 363:554–560
32. Liu R, Li X, Li Y, Jin P, Qin W, Qi J (2009) Effective removal of rhodamine B from contaminated water using non-covalent imprinted microspheres designed by computational approach. *Biosens Bioelectron* 25:629–634
33. Niu C, Xu Y, Luo X, Duan W, Silman I, Sussman JL, Zhu W, Chen K, Shen J, Jiang H (2005) Dynamic mechanism of E2020 binding to acetylcholinesterase: a steered molecular dynamics simulation. *J Phys Chem B* 109:23730–23738
34. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1981) In: B Pullman (ed) *Intermolecular forces*. Reidel, Dordrecht, pp 331–342
35. Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: a new molecular dynamics method. *J Appl Phys* 52:7182–7190
36. Lin J-H, Perryman AL, Schames JR, McCammon JA (2002) Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *J Am Chem Soc* 124:5632–5633
37. Dixit A, Kashaw SK, Gaur S, Saxena AK (2004) Development of CoMFA, advanced CoMFA and CoMSIA models in pyrroloquinazolinones as thrombin receptor antagonist. *Bioorg Med Chem* 12:3591–3598
38. Ashek A, Cho SJ (2006) A combined approach of docking and 3D QSAR study of [beta]-ketoacyl-acyl carrier protein synthase III (FabH) inhibitors. *Bioorg Med Chem* 14:1474–1482
39. Politi A, Durdagi S, Moutevelis-Minakakis P, Kokotos G, Mavromoustakos T (2010) Development of accurate binding affinity predictions of novel renin inhibitors through molecular docking studies. *J Mol Graph Model* 29:425–435
40. Wang Y, Li Y, Ma Z, Yang W, Ai C (2010) Mechanism of microRNA–target interaction: molecular dynamics simulations and thermodynamics analysis. *PLoS Comput Biol* 6:e1000866

Ab initio study of $M\text{Kr}_n^{2+}$ ($M = \text{Cu, Ag, and Au, } n = 1-6$) clusters

Xinying Li

Received: 1 April 2011 / Accepted: 24 May 2011 / Published online: 8 June 2011
© Springer-Verlag 2011

Abstract Quantum chemical calculations of the structures and stabilities of the title series at the CCSD(T) theoretical level are performed. Laplacian, electron density deformation, electron localization function and reduced density gradient analysis are investigated to explore the nature of the interaction. The results show that a covalent contribution occurs in the Kr-M^{2+} bonding.

Keywords Electron density property · Interaction Structure and stability

Introduction

Interactions of atomic ions with rare gas are of considerable importance and interest, since it provides a convenient inert environment for many experiments and technological processes. About five decades ago, Bartlett reported the first rare gas compound, xenon hexafluoroplatinate [$\text{Xe}^+(\text{PtF}_6)$] [1]. Gold is in fact generally regarded as the element whose chemistry is most affected by relativistic effects [2, 3]. This metal is nowadays used in several high-technology fields, like microelectronics and nanostructured materials science [4]. Seidel and Seppelt demonstrated the existence of the [AuXe_4] $^{2+}$ cation [5], it is very important because it supports the concept that the rare gas atoms can be directly bonded to the gold atom. Many investigations have been focused on the *M-rare gas* bonding in the recently examined rare gas-coinage metal halides, rare gas-MX (rare gas = Ar, Kr, Xe;

$M = \text{Cu, Ag, Au}$; and $X = \text{F, Cl, Br}$) [6–27]. In these species, the presence of chemical bonds between M and rare gas atoms is in sharp contrast to the conventional behavior, which is considered to be inert from the existing chemical intuitions. Pyykkö suggested that most of the bonding interaction is covalent in character [18, 19]; while it was questioned by saying that “covalency within the rare gas- Au^+ species appears to be unproven” [23]. In 2001, Walker and co-workers reported the unexpected experimental and theoretical determination of stable MAr_n^{2+} clusters (at the MP4/LANL2DZ level) [15]. Our previous results show that the interactions of singly charged M^+ -Kr series are stronger than those of M^+ -Ar series [28, 29]. However, the theoretical investigations including geometrical structures, electronic structures, especially the roles of interactions of small MKr_n^{2+} clusters are less reported. These motivated us to perform further calculation to investigate the M^{2+} -Kr and explore the nature of the interaction.

Reported here are the results of the calculations undertaken on the MKr_n^{2+} ($M = \text{Cu, Ag, and Au, } n=1-6$) series at the CCSD(T) theoretical level, with the aim of not only understanding the behavior of the systems under consideration, but also to give an insight into the nature of the interaction between Kr atom(s) and coinage metal atoms. It would be meaningful and interesting to give a description of structures and properties of this new class of compounds.

Computational details

The SDD relativistic pseudopotentials and the basis set SDB-cc-pVQZ, (14s10p3d2f1g)/[4s4p3d2f1g] are used to describe the Kr atom [30, 31]. The 19-valence electron relativistic pseudopotentials and the match basis set

X. Li (✉)
Institute for Computational Materials Science,
School of Physics and Electronics, Henan University,
Kaifeng 475004, People's Republic of China
e-mail: lxying@henu.edu.cn

((8s7p6d)/(6s5p3d)) of Dolg are employed for Au, Ag and Cu atoms [32, 33]. Pyykkö found that two *f*-type polarization functions are desirable for the correct description of the interaction and the inclusion of an additional *g* function on Au has a sizable effect on the computed bond energy and also leads to a significant shrinking of equilibrium distance, therefore, two *f* functions (0.20 and 1.19 for Au, 0.22 and 1.72 for Ag, 0.24 and 3.7 for Cu) [34], and one *g* function (1.1077) for Au are augmented to the basis sets [19]. The interactions in rare gas atoms containing compounds often require the inclusion of very high angular momentum functions for accurate description [35], the present *g*-function including high momentum basis set is proved to be sufficiently accurate and necessary to describe the interaction in our previous investigation [28].

The basis set superposition error (BSSE) is corrected by using counterpoise procedure of Boys and Bernardi [36]. The calculations were performed with the Gaussian 03 W program package [37].

Results and discussion

The equilibrium structures, binding energies (E_b), and natural population analyses (NPA) of M atom of MKr_n^{2+} ($n=1-6$) system are given in Table 1.

Structures

For $n = 1$ system, the CCSD(T) method obtained the equilibrium M^{2+} -Kr distances of 229.2, 242.5, and 240.4 pm, and binding energies of 2.4945, 2.5326 and 2.9173 eV, for $M = Cu, Ag$ and Au , respectively. The compact structure and enhanced stability compared to the singly charged M^+ -Kr calculated at the same theoretical level with the same basis sets (237.7, 268.3 and 255.8 pm; 0.7500, 0.5470 and 0.8010 eV; for $M = Cu, Ag$ and Au , respectively [28]) were found. A comparison along the M series shows that the order of the R_{M-Kr} distance is $R_{Cu-Kr} < R_{Au-Kr} < R_{Ag-Kr}$, similar to the trend of M^+ -Kr series. The binding energy order is $Cu < Ag < Au$; the Au-containing species is more stable and compact compared to Ag and Cu, and this stabilization is brought about by relativistic effects [18, 28].

The CCSD(T) global minimum energy structures of MKr_n^{2+} ($n=2-6$) are shown in Fig. 1. The most stable structures are of $D_{\infty h}$ (linear), C_{2v} (planar), D_{4h} (planar), C_{4v} (Pentahedron) and D_{4h} (octahedron) for $n=2-6$, respectively. Our results accords with Walker's structures of MAR_n^{2+} . One can find in Fig. 1 that the MKr_{n+1}^{2+} system can be formed by adding one Kr atom to the stable MKr_n^{2+} structures without obvious changes of structural parameters. All the MKr_2^{2+} ($M = Cu, Ag$ and Au) systems are found to

have the linear symmetry, with the M ions in the center connected to the two Kr atoms. For $n=3$ systems, the angles A_{314} are about 170 degree for Au and Ag, while that of Cu is only 105.8 degree. One can see from Table 1, the bond length of M^{2+} -Kr in larger clusters ($n=2-6$) are longer than that of MKr_2^{2+} . The M^{2+} -Kr "bond" is stronger than that of Kr-Kr "bond"; therefore, in the global minimum energy structure for larger clusters, all the Kr atoms are in direct contact with the central M^{2+} , allowing the maximum M^{2+} -Kr "bond" to be formed. Next, the Kr atoms are grouped in such a way that the number of Kr-Kr "bond" is maximized.

Stabilities

In order to understand the relative stability and size-dependent behavior, we have investigated the binding energy (E_b), average binding energy (E_{b-ave}), fragmentation energy (F_e) and the second-order difference of energies (Δ_2E). The energies are referenced to the separated-atom limit consisting of the ground state Kr $4s^24p^6$ and $M(II) s^0d^9$ state. The results are shown in Fig. 2. Here the binding energy, average binding energy, fragmentation energy and second-order difference of energies are defined as:

$$E_b(n) = E(M^{+2}) + nE(Kr) - E(MKr_n^{2+}) \quad (1)$$

$$E_{b-ave}(n) = [E(M^{+2}) + nE(Kr) - E(MKr_n^{2+})]/n + 1 \quad (2)$$

$$F_e(n, n-1) = E(MKr_{n-1}^{2+}) + E(Kr) - E(MKr_n^{2+}) \quad (3)$$

$$\Delta_2E(n) = E(MKr_{n-1}^{2+}) + E(MKr_{n+1}^{2+}) - 2E(MKr_n^{2+}), \quad (4)$$

where $E(\dots)$ is the total energy of the corresponding system.

The results collected in Table 1 clearly shows that the binding energies increase monotonically as the size of n increase, which means that these clusters can continuously gain energy during the growth process. One can see from Fig. 1 that the M^{2+} was found to be located between the Kr atoms, thus the M^{2+} -Kr "bonds" increase monotonically as the size of n increase while the Kr-Kr "bond" does not have the same behavior; thus it results in the monotonically increase of dissociation energies and irregular variable trend of the average binding energies, fragmentation energies and second-order energies.

As shown in Fig. 2, the global maximum E_{b-ave} , F_e and Δ_2E are found at $n=2$ for all the systems; it clearly shows the enhanced stabilities of the $n=2$ systems. The local maximum E_{b-ave} , F_e and Δ_2E are found at $n=4$; it indicates

Table 1 Global minimum energy structures, binding energies and NPAs calculated at CCSD(T) theoretical level

n		R_e/pm	A/degree	E_b/eV	NPA(M)
1	Cu	$R_{12}=229.2$		2.4945	1.83445
	Ag	$R_{12}=242.5$		2.5326	1.74609
	Au	$R_{12}=240.4$		2.9173	1.70576
2	Cu	$R_{12}=R_{13}=232.1$	$A_{314}=180.0$	4.5493	1.64701
	Ag	$R_{12}=R_{13}=246.4$	$A_{314}=180.0$	4.2672	1.60985
	Au	$R_{12}=R_{13}=243.0$	$A_{314}=180.0$	5.1124	1.49270
3	Cu	$R_{13}=R_{14}=235.8$ $R_{12}=237.5$	$A_{213}=105.8$	5.5631	1.54265
	Ag	$R_{13}=R_{14}=249.1$ $R_{12}=251.2$	$A_{213}=170.2$	5.4053	1.47417
	Au	$R_{13}=R_{14}=246.1$ $R_{12}=251.2$	$A_{213}=168.1$	6.3948	1.33102
4	Cu	$R_{12}=R_{13}=244.2$	$A_{213}=90.0$	6.5950	1.48538
	Ag	$R_{12}=R_{13}=259.4$	$A_{213}=90.0$	6.4392	1.38430
	Au	$R_{12}=R_{13}=257.3$	$A_{213}=90.0$	7.4888	1.26031
5	Cu	$R_{13}=R_{14}=247.9$ $R_{16}=258.8$	$A_{214}=165.4$	7.2178	1.45075
	Ag	$R_{13}=R_{14}=255.9$ $R_{16}=281.1$	$A_{214}=170.0$	6.8757	1.35192
	Au	$R_{13}=R_{14}=256.4$ $R_{16}=289.1$	$A_{214}=170.8$	7.8791	1.23419
6	Cu	$R_{12}=R_{16}=246.9$ $R_{13}=262.7$	$A_{213}=90.0$	7.6251	1.42949
	Ag	$R_{12}=R_{16}=257.8$ $R_{13}=287.3$	$A_{213}=90.0$	7.3080	1.33112
	Au	$R_{12}=R_{16}=258.0$ $R_{13}=296.3$	$A_{213}=90.0$	8.2417	1.22264

the enhanced stabilities of the $n=4$ systems compared to its $n=3$ and $n=5$ neighbors. While local minimum F_e and Δ_2E are found at $n=3$ and $n=5$ systems; it suggests the weaker stabilities. It is well known that coinage metal in the valence state $+I$, M^+ , very much prefer linear coordination [38], and the present results indicate that for $M(\text{II})$ valence state, they also prefer linear coordination. For $n=2$ and $n=4$ systems, the M^{2+} and Kr atoms are arranged in linear coordination and the corresponding systems have stronger stabilities. The plots of Δ_2E clearly show odd-even oscillation character (only for $n=2-5$). The linear coordinations are also found for $n=6$ systems and it is expected it also has stronger stability. Since the larger systems ($n>6$)

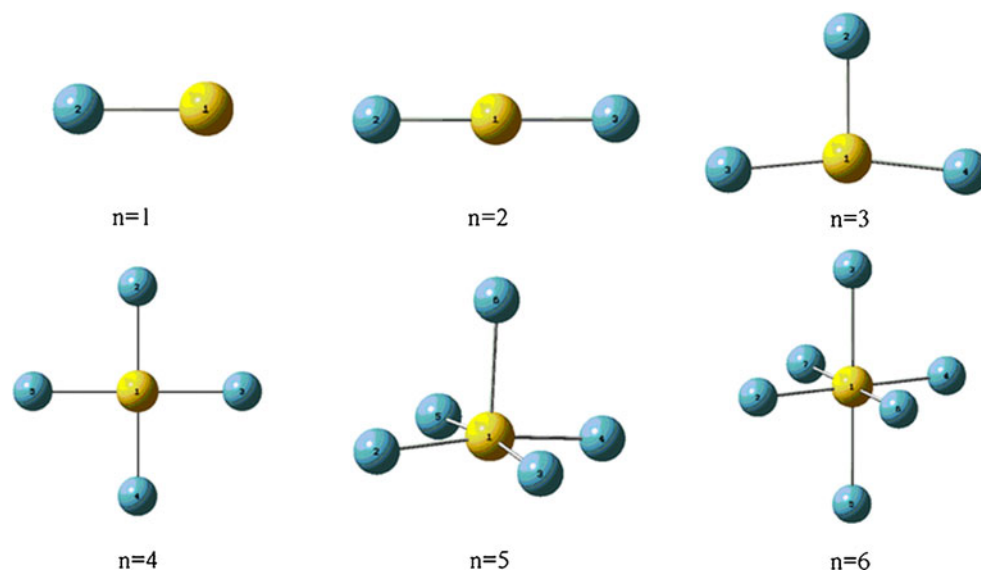
are not investigated in the present research, only the F_e of AuKr_6^{2+} and AgKr_6^{2+} shows the enhanced stabilities compared to their $n=5$ neighbors.

Electron density properties

Atoms in molecules (AIM)

According to the AIM theory from Bader [39] based on topological analysis of electron density, the chemical bonding can be characterized by the existence of a (3, -1) type of critical point (bond critical point, BCP; i.e., a point where the gradient vector field $\nabla\rho(r)$ is zero and $\rho(r)$

Fig. 1 CCSD(T) structures for AuKr_n^{2+} clusters with n in the range 1–6. Structures involving the other metal cations are very similar to these. Details of structures are given in Table 1



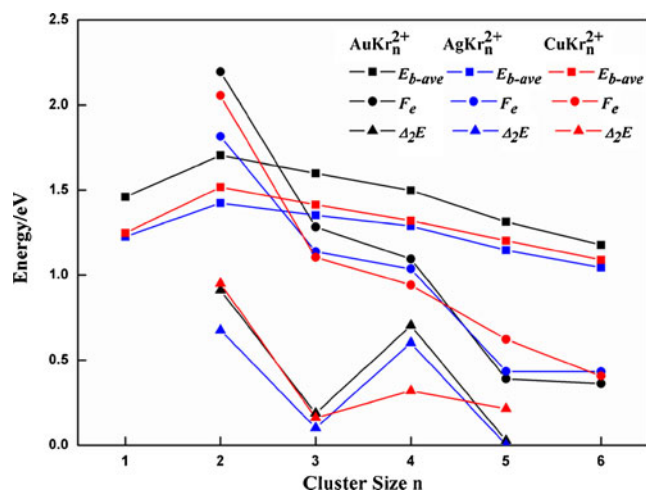


Fig. 2 CCSD(T) average binding energies, fragmentation energies, and the second-order difference of energies plotted as the function of cluster size n

possesses one positive and two negative curvatures. At a bond critical point, the electron density, $\rho(r)$, is maximum in the plane and minimum along the bond path between atoms

A and B.) and the corresponding bond path. Its nature is revealed by descriptors at the BCP such as the electron density $\rho(r)$ and Laplacian $\nabla^2\rho(r)$, which is composed of the sum of three curvatures, λ_1 , λ_2 , and λ_3 . For all the cases, λ_3 dominates λ_1 and λ_2 . Therefore, all Laplacians are positive and all $|\lambda_1/\lambda_3|$ ratios are very small. The chemical bondings are divided into two groups: shared interactions and closed shell interactions, which correspond to the negative and positive Laplacian value $\nabla^2\rho$. A negative value means that the electronic charge is concentrated in the inter-nuclear region, and is therefore shared by the two nuclei. A positive value indicates the depletion in the region and this is the case in all closed shell interactions. It is also suggested that positive local energy density, $E(r)$, the sum of the kinetic energy density $G(r)$ and the potential energy density $V(r)$, is necessary for the classification as closed shell interactions [40]. It can be seen from the equation, $E(r) = (\hbar/4m)\nabla^2\rho - G(r)$, that $E(r)$ may still be negative if $\nabla^2\rho(r)$ is positive due to the always positive kinetic energy density $G(r)$. The chemical bonding characterized by positive $\nabla^2\rho(r)$ and negative $E(r)$ is referred to as “intermediate type” [41]. The interactions listed in Table 2

Table 2 BCP properties calculated at CCSD(T) theoretical level

	BCP (A-B)	ρ (10^{-2})	λ_2 (10^{-2})	Lap(10^{-1})	$G(r)$ (10^{-2})	$V(r)$ (10^{-2})	$E(r)$ (10^{-2})
1	(2-1)Cu	7.5102	-6.9005	2.7123	8.431430	-10.08197	-1.650540
	(2-1)Ag	8.2377	-8.6654	1.5494	6.312401	-8.751228	-2.438826
	(2-1)Au	9.6983	-9.6135	1.6160	7.660748	-11.28142	-3.620672
2	(2,3-1)Cu	6.9252	-5.9810	2.6228	8.026862	-9.496497	-1.469634
	(2,3-1)Ag	7.1904	-6.6184	1.9129	6.550102	-8.317730	-1.767628
	(2,3-1)Au	8.8141	-7.9231	1.9684	7.868766	-10.81640	-2.947639
3	(2-1)Cu	6.3832	-6.5649	2.2596	6.691663	-7.734286	-1.042622
	(3,4-1)Cu	6.3353	-5.8059	2.4647	7.184110	-8.206350	-1.022240
	(2-1)Ag	6.4588	-6.0934	1.7422	5.794138	-7.232590	-1.438452
	(3,4-1)Ag	6.6934	-6.1891	1.8777	6.216861	-7.739356	-1.522494
4	(2-1)Au	7.3690	-6.7642	1.6700	6.209714	-8.244354	-2.034639
	(3,4-1)Au	8.1919	-7.5226	1.9537	7.420892	-9.957456	-2.536563
	Cu	5.3195	-4.3177	1.9256	5.641165	-6.468223	-0.827059
5	Ag	5.9695	-5.3841	1.7032	5.475847	-6.693673	-1.217826
	Au	6.8630	-6.0815	1.7078	6.023811	-7.778108	-1.754296
	(2,3,4,5-1)Cu	4.8634	-3.8025	1.7699	5.094683	-5.764562	-0.669878
6	(6-1)Cu	4.0499	-3.5108	1.2946	3.790749	-4.344937	-0.554187
	(2,3,4,5-1)Ag	5.7446	-5.0992	1.6603	5.278574	-6.406367	-1.127792
	(6-1)Ag	3.3809	-2.7586	1.1009	3.102502	-3.452578	-0.350075
	(2,3,4,5-1)Au	6.5858	-5.7445	1.6737	5.787358	-7.390374	-1.603016
	(6-1)Au	3.3754	-2.6335	1.0938	2.972034	-3.209449	-2.374146
6	(2,4,6,7-1)Cu	4.8228	-3.5907	1.8612	5.311095	-5.969106	-0.658011
	(3,5-1)Cu	3.7332	-2.9575	1.1147	3.339341	-3.891932	-0.552591
	(2,4,6,7-1)Ag	5.5371	-4.8237	1.6141	5.080071	-6.124664	-1.044592
	(3,5-1)Ag	5.4880	-4.7729	1.6003	5.027639	-6.054436	-1.026796
	(2,4,6,7-1)Au	6.3438	-5.4448	1.6366	5.567970	-7.044432	-1.476461
	(3,5-1)Au	2.9681	-2.2646	0.9394	2.494223	-2.639876	-0.145652

(very small BCP electron density ρ_{BCP} , negative $E(r)$ and positive Laplacian value) all fall into the intermediate type.

Density difference function (DDF)

We performed the changes in the electron density upon the formation of the interactions between the fragments to understand the nature of the M^{2+} -Kr interaction. The density difference function (DDF), $\Delta\rho(r)$, is defined as the difference between the total electron density and the promolecule density; this function is regarded as a deformation density associated with the redistribution of the electron charge when the system forms from the constituent atoms [42]. In Fig. 3 the contours of the DDF between the complexes and the non-interacting fragments (in the same positions) for the most stable structure, $AuKr_2^{2+}$, are plotted. Regions where the electron charges accumulate with respect to isolated atoms have positive $\Delta\rho(r)$ whereas regions of electron depletion have negative $\Delta\rho(r)$. For $AuKr_2^{2+}$, obvious changes of the charge distribution can be seen in Fig. 3, there is electron accumulation in the Kr-Au interaction region, and it enhances its stability. NBO analysis clearly shows that the $4p$ orbitals of Kr atom(s) play an important role in the strong Au-Kr interaction. Figure 3 also shows the electron depletion of the corresponding p orbital wherein the charge transfers from the Kr atom(s) to the Au atom.

Electron localization function (ELF)

The ELF describes how much the Pauli repulsion is efficient at a given point of the molecular space [43]. $ELF = 1$ correspond to a completely localized situation, 0 to a delocalized system, and 0.5 is the value one should obtain for the homogenous electron gas. It provides a rigorous basis for the analysis of the interaction. The local maxima of ELF define localization domains and they correspond to chemically interesting regions. Two main types of domains are obtained from the ELF partition of the real space: (a) core basins are located around nuclei, and always occur

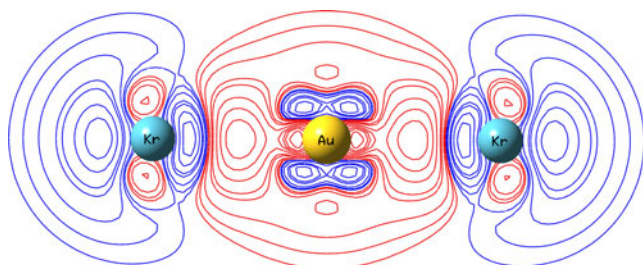


Fig. 3 Contours map of the density difference function at the plane of $AuKr_2^{2+}$. Red and blue lines are in regions of charge concentration ($\Delta\rho > 0$) and charge depletion ($\Delta\rho < 0$), respectively

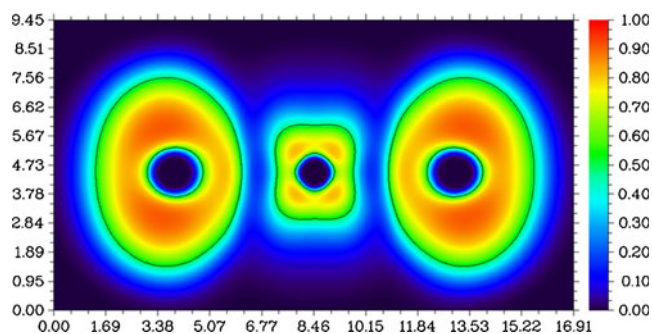


Fig. 4 2D plot of the electron localization function (ELF) in the plane of $AuKr_2^{2+}$ (ELF = 0.5 contour lines are plotted)

when the atomic number is larger than 2, and (b) valence basins are characterized by their synaptic orders. Mono-synaptic basins represent the lone pairs, whereas disynaptic basins belong to the covalent interaction. ELF analysis of $AuKr_2^{2+}$ clearly shows that there are only core basins and valence basins around the Au and Kr atoms in the system (Fig. 4), while valence basins located between the Au and Kr atom(s) could not be found, and in this region the ELF values are very small (about 0.1–0.25). However, we note that there are considerable deformations in the valence basin (it expands to the Au-Kr interaction region direction) around the Kr atom(s). Taking the binding energy and DDF analysis into account, it suggests that there is a covalent contribution component in the interaction.

Reduced density gradient (RDG)

Johnson and co-workers developed an approach to investigate the weak interactions in real space based on the electron density and its derivatives [44]. The RDG is a fundamental dimensionless quantity coming from the density and its first derivative ($RDG = 1 / (2(3\pi^2)^{1/3}) |\nabla\rho| / \rho^{4/3}$).

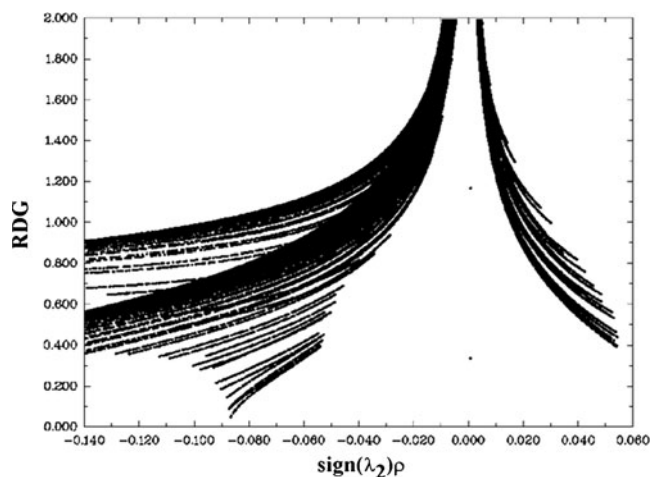


Fig. 5 Plots of the RDG versus the electron density ρ multiplied by the sign of λ_2 for $AuKr_2^{2+}$

The weak interactions can be isolated as regions with low electron density and low RDG value. The density values of the low-gradient spikes (the plot of *RDG* versus ρ) appear to be an indicator of the interaction strength. The sign of λ_2 is utilized to distinguish the bonded ($\lambda_2 < 0$) from nonbonded ($\lambda_2 > 0$) interactions. The plot of the RDG versus the electron density ρ multiplied by the sign of λ_2 can allow analysis and visualization of a wide range of interactions types. The results were calculated by Multiwfn and plotted by VMD program [45, 46].

One or more spikes are found in the low-density, low-gradient region (Fig. 5), indicative of weak interactions in the system and the electron density value at the RDG versus $\text{sign}(\lambda_2)\rho$ peaks itself provides the information about the strength of interaction. Large, negative values of $\text{sign}(\lambda_2)\rho$ are indicative of stronger attractive interactions (spikes in the left part in Fig. 5), while if it is large and positive, the interaction is repulsion (spikes in the right part in Fig. 5). Values near zero indicate very weak, van der Waals interactions [44]. We find that the RDG = 0.3 line crosses only the attractive interaction spikes while the RDG > 0.4 lines cross not only the attractive but also the repulsion spikes. The strength of the repulsion is smaller than that of the attraction and thus the attractions dominate in the system. RDG discriminates between different types of interactions. Very low density values (i.e., $\rho < 0.005\text{au}$) generally map to weaker dispersion interactions. Slightly higher density values (i.e., $0.005 < \rho < 0.05\text{au}$) map to stronger noncovalent interactions. The present electron density value, about 0.09au (negative λ_2), clearly shows the covalent character in the Kr-Au interaction.

Conclusions

Investigations of the MKr_n^+ series at the CCSD(T) theoretical level with extended basis sets provide reliable structures and stabilities as well as insights into its nature of the interaction and the electron properties. Analysis on the electron properties and the dissociation energies clearly shows the covalent component in character for the systems studied.

Acknowledgments Financial supports from the National Science Foundation of China (No. 10947141 and 10804027), National Science Foundation of Henan Province Education Department (No. 2010B140003) are gratefully acknowledged.

References

- Bartlett N (1962) Proc Chem Soc 1962:218
- Pyykkö P (1988) Chem Rev 88:563–594
- Schwarz H (2003) Angew Chem Int Ed 42:4442–4454
- Leonard RM, Bhuvanesh NSP, Schaak RE (2005) J Am Chem Soc 127:7326–7327
- Seidel S, Seppelt K (2000) Science 290:117–118
- Evans CJ, Gerry MCL (2000) J Chem Phys 112:1321–1329
- Evans CJ, Gerry MCL (2000) J Chem Phys 112:9363–9374
- Evans CJ, Lesarri A, Gerry MCL (2000) J Am Chem Soc 122:6100–6105
- Evans CJ, Rubino DS, Gerry MCL (2000) Phys Chem Chem Phys 2:3943–3948
- Cooke SA, Gerry MCL (2004) J Am Chem Soc 126:17000–17008
- Cooke SA, Gerry MCL (2004) Phys Chem Chem Phys 6:3248–3256
- Thomas JM, Walker NR, Cooke SA, Gerry MCL (2004) J Am Chem Soc 126:1235–1246
- Michaud JM, Cooke SA, Gerry MCL (2004) Inorg Chem 43:3871–3881
- Michaud JM, Gerry MCL (2006) J Am Chem Soc 128:7613–7621
- Walker NR, Wright RR, Barran PE, Cox H, Stace AJ (2001) J Chem Phys 114:5562–5567
- Ghanty TK (2005) J Chem Phys 123:074323
- Ghanty TK (2006) J Chem Phys 124:124304
- Pyykkö P (1995) J Am Chem Soc 117:2067–2070
- Schröder D, Schwarz H, Hrusak J, Pyykkö P (1998) Inorg Chem 37:624–632
- Yousef A, Shretha S, Breckenridge WH (2007) J Chem Phys 127:154309
- Bauschlicher CW, Partridge JH, Langhoff SR (1989) J Chem Phys 91:4733–4737
- Shen Y, BelBruno JJ (2005) J Phys Chem A 109:10077–10083
- Read JP, Buckingham AD (1997) J Am Chem Soc 119:9010–9013
- Bellert D, Breckenridge WH (2002) Chem Rev 102:1595–1622
- Belpassi L, Infante I, Tarantelli F, Visscher L (2008) J Am Chem Soc 130:1048–1060
- Breckenridge WH, Ayles VL, Wright TG (2008) J Phys Chem A 112:4209–4214
- Zeng T, Klobukowski M (2008) J Phys Chem A 112:5236–5242
- Xinying L, Xue C, Yongfang Z (2009) Aust J Chem 62:121–125
- Xin-Ying L, Xue C, Yongfang Z (2009) J Phys B-At Mol Opt Phys 42:065102
- Nicklass A, Dolg M, Stoll H, Preuss H (1995) J Chem Phys 102:8942–8952
- Martin JML, Sundermann A (2001) J Chem Phys 114:3408–3420
- Andrae D, Haeussermann U, Dolg M, Stoll H, Preuss H (1990) Theor Chim Acta 77:123–141
- Dolg M, Wedig U, Stoll H, Preuss H (1987) J Chem Phys 86:866–872
- Pyykkö P, Runeberg N, Mendizabal F (1997) Chem Eur J 3:1451–1457
- Chalasiński G, Szczeniński MM (1994) Chem Rev 94:1723–1765
- Boys SF, Bernardi F (1970) Mol Phys 19:553–566
- Frisch MJ, Trucks GW et al (2003) Gaussian 03 W. Gaussian Inc, Pittsburgh
- Roithová J, Schröder D (2009) Coord Chem Rev 253:666–677
- Bader RFW (1990) Atoms in Molecules. A Quantum theory. Clarendon, Oxford
- Cremer D, Kraka E (1984) Angew Chem Int Edit 23:627–628
- Nakanishi W, Hayashi S, Narahara K (2008) J Phys Chem A 112:13593–13599
- Pacios LF, Fernandez A (2009) J Mol Graph Model 28:102–112
- Becke AD, Edgecombe KE (1990) J Chem Phys 92:5397–5403
- Johnson ER, Keinan S, Mori-Sánchez P, Contreras-García J, Cohen AJ, Yang W (2010) J Am Chem Soc 132:6498–6506
- Lu T, “Multiwfn: Multifunctional wavefunction analyzer”, Version 1.5, <http://Multiwfn.codeplex.com>
- Humphrey W, Dalke A, Schulten K (1996) Visual Molecular Dynamics. J Mol Graph 14:33–38

Interactions of uranyl ion with cytochrome *b*₅ and its His39Ser variant as revealed by molecular simulation in combination with experimental methods

Dun Wan · Li-Fu Liao · Min-Min Zhao ·
Min-Long Wu · Yi-Mou Wu · Ying-Wu Lin

Received: 12 December 2010 / Accepted: 22 March 2011 / Published online: 9 June 2011
© Springer-Verlag 2011

Abstract The biological toxicity of uranyl ion (UO_2^{2+}) lies in interacting with proteins and disrupting their native functions. The structural and functional consequences of UO_2^{2+} interacting with cytochrome *b*₅ (cyt *b*₅), a small membrane heme protein, and its heme axial ligand His39Ser variant, cyt *b*₅ H39S, were investigated both experimentally and theoretically. In experiments, although cyt *b*₅ was only slightly affected, UO_2^{2+} binding to cyt *b*₅ H39S with a K_D of 2.5 μM resulted in obvious alteration of the heme active site, and led to a decrease in peroxidase activity. Theoretically, molecular simulation proposed a uranyl ion binding site for cyt *b*₅ at surface residues of Glu37 and Glu43, revealing both coordination and hydrogen bonding interactions. The information gained in this study provides insights into the mechanism of uranyl toxicity toward membrane protein at an atomic level.

Keywords Heme proteins · Metal-binding site · Peroxidase · Toxicity · Uranium

Introduction

Uranium is one of the heaviest naturally occurring elements on Earth, and is harmful to human health due to its long-

lived radioactivity and high toxicity. The biological toxicity results from the ability of the uranyl ion (UO_2^{2+}), the most stable form of uranium under physiological conditions [1], to interact with proteins [2–4] such as transferrin, ferritin and albumin, and to disrupt the native function of these biomolecules. Thus, for biological remediation purposes, it is necessary to understand the mechanism underlying these interactions at an atomic level. Recent studies show that UO_2^{2+} binds to proteins mainly through carboxylic acid groups such as those of aspartate (Asp) and glutamate (Glu) [5, 6], and histidine residues as in mutated NikR [7]. The second sphere hydrogen bonds involving uranyl oxo groups may enhance the interaction between UO_2^{2+} and proteins. On the other hand, due to the positive charge, UO_2^{2+} has a strong tendency to be absorbed at the negatively charged surfaces of membranes [8], where it has more chance to interact with the membrane proteins; to date, limited attention has been directed towards this aspect.

Cytochrome *b*₅ (cyt *b*₅) is a small heme protein with heme coordinated by two axial histidine ligands (His39 and His63). Cyt *b*₅ often binds to the microsome membrane via a short hydrophobic domain and functions as an electron transport in biological systems [9]. By replacing heme axial His39 with a non-coordinated residue such as serine (Ser), cyt *b*₅ has been converted into a peroxidase-like enzyme, cyt *b*₅ H39S, as a result of creating an open binding site for substrates [10]. A distinguishing characteristic of cyt *b*₅ is that the hydrophilic heme-binding domain is highly negatively charged due to the presence of a series of acidic residues surrounding the heme group, such as Glu37, Glu43, Glu44, Asp60 and Asp66; this is also known as the “acidic” cluster of cyt *b*₅ [11]. With this in mind, one might expect that UO_2^{2+} would tend to interact with cyt *b*₅ involving this region, thus offering us a suitable example with which to study the interactions between UO_2^{2+} and membrane proteins.

D. Wan · L.-F. Liao · M.-M. Zhao · M.-L. Wu · Y.-W. Lin (✉)
School of Chemistry and Chemical Engineering,
University of South China,
Hengyang 421001, People’s Republic of China
e-mail: linlinyong@hotmail.com

L.-F. Liao
e-mail: lf_liao@yahoo.com.cn

Y.-M. Wu
Institute of Pathogenic Biology, University of South China,
Hengyang 421001, People’s Republic of China

As shown herein, we investigated the interactions between UO_2^{2+} and *cyt b₅*, and its His39Ser variant *cyt b₅ H39S*, by UV-vis titration, and proposed a uranyl binding site in *cyt b₅* at Glu37 and Glu43 by molecular simulation. The impact of the uranyl ion on protein function was further investigated by evaluating peroxidase activity in presence of UO_2^{2+} ions. The structural and functional consequences of UO_2^{2+} binding revealed in this study provide valuable information for understanding the mechanism of biological toxicity of the uranyl ion at an atomic level.

Materials and methods

Materials

The lipase-solubilized bovine liver microsomal *cyt b₅* and the *cyt b₅ H39S* variant (kindly provided by Prof. Z.-X. Huang, Fudan University, Shanghai, China) were expressed and purified as described in a previous study [12]. Uranyl nitrate, guaiacol, hydrogen peroxide (30%), Bis-Tris, and other chemicals were commercial products and of analytical grade. Double distilled water was used throughout the experiment.

UV-vis studies

UV-vis spectra of uranyl titration were collected on a PerkinElmer Lambda 35 spectrometer at room temperature (25 °C). *Cyt b₅ H39S* was dissolved in 50 mM Bis-Tris buffer (pH 7.0) at a concentration of 10 μM , as calculated by an extinction coefficient of $\epsilon_{405}=80 \text{ mM}^{-1}\cdot\text{cm}^{-1}$ in its ferric state [12]. Up to five equivalents of uranyl ions (UO_2^{2+}) (2 mM uranyl nitrate solution) were titrated into the above protein solution, and spectra were recorded at every 0.5 equivalents with an interval of 30 min. Data were plotted by the double reciprocal plot method as used previously for determining the binding affinity of Cu(II) in a copper-binding site created in myoglobin [13]. *Cyt b₅*, with an extinction coefficient of $\epsilon_{413}=117 \text{ mM}^{-1}\cdot\text{cm}^{-1}$ in the ferric state [12], was titrated under the same conditions, and the spectra were recorded at each one equivalent of UO_2^{2+} .

Molecular simulation studies

The structure of *cyt b₅ H39S* was modeled based on the crystal structure of bovine liver microsomal *cyt b₅* (PDB entry 1CYO) [14], by using a procedure similar to previous modeling *cyt c* axial variants [15]. The heme axial ligand His39 was first mutated to a Ser using program VMD 1.8.7 (Visual Molecular Dynamics) [16]. The variant was solvated in a cubic box of TIP3 water, which extended 10 Å away from any given protein atom. The resultant system

was minimized for 1,000 ps using program NAMD 2.7 (Nanoscale Molecular Dynamics) [17] with the conjugate gradient method, subsequently equilibrated for 10 ps with a time step of 1 fs, then further minimized for 30,000 steps for analysis with the VMD program.

To model uranyl binding to *cyt b₅* and *cyt b₅ H39S*, namely, U-*cyt b₅* and U-*cyt b₅ H39S*, a water molecule (HOH 578) in crystal structure of *cyt b₅* forming hydrogen bonds with both Glu37 and Glu43 was changed to a UO_2^{2+} ion. The UO_2^{2+} -containing protein, after solvating in a cubic box of TIP3 water, was minimized with NAMD using 5000 minimization steps at 0 K, then 10,000 molecular dynamics steps (1 fs per step) via an VNT ensemble (where the number of particles *N*, the volume *V*, and the temperature *T* of the system were kept constant) at 310 K, by using a procedure described recently for modeling Zn(II) binding to a rationally designed Fe(II)-binding site above the heme group in myoglobin [18]. The system was further minimized for 30,000 steps before analysis with the VMD program. The parameters used for uranyl ion were explored in a previous study [8], with a bond length of U-O 1.77 Å and an angle of O-U-O 180°.

Peroxidase activity studies

By using a procedure similar to a previous study [10], the peroxidase activity of *cyt b₅ H39S* (5 μM) in the absence or presence of one to five equivalents of UO_2^{2+} was estimated in Bis-Tris buffer (2 mL, 50 mM, pH 7.0) at 25 °C, with guaiacol (10 mM) as a substrate. The reaction was initiated by the addition of hydrogen peroxide (final concentration 10 mM) to the starting mixtures and followed by monitoring the change in absorbance (for 30 min with 6 s intervals) of the product, tetraguaiacol, at 470 nm using an extinction coefficient of $\epsilon_{470}=26.6 \text{ mM}^{-1}\cdot\text{cm}^{-1}$ [19]. Control experiments were carried out for *cyt b₅* by following the same procedure, in the absence and presence of five equivalents of UO_2^{2+} in solution.

Results and discussion

Uranyl ion binding from UV-vis studies

Heme proteins are characterized by a Soret band around 400 nm, and the α -, β -bands around 500–600 nm in electronic absorption spectroscopy in the ultraviolet and visible region (UV-vis), which are closely linked to the heme coordination environments and heme iron oxidation states [20]. To determine if UO_2^{2+} can bind to *cyt b₅*, we titrated UO_2^{2+} into bovine liver microsomal *cyt b₅* or its His39Ser variant, *cyt b₅ H39S*, both in the oxidized state, and monitored the resulting UV-vis spectra. As shown in Fig. 1,

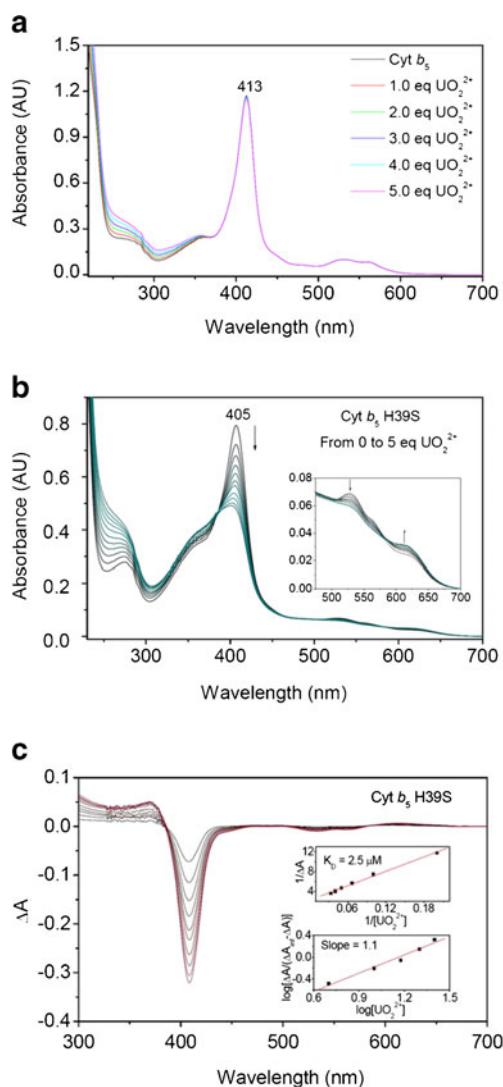


Fig. 1 UV-vis titration of oxidized *cyt b₅* (a) and *cyt b₅* H39S (b) (10 μM) with uranyl ion (UO_2^{2+}) in 50 mM Bis-Tris buffer (pH 7.0) at 25 °C. Difference spectra of uranyl ion binding to *cyt b₅* H39S (c). (Inset, top) Double reciprocal plot of the change in UO_2^{2+} concentration versus the change in absorbance. (Inset, bottom) Hill plot of the data

although only a slight decrease of the Soret band (413 nm) was observed for *cyt b₅* upon addition of UO_2^{2+} up to five equivalents (Fig. 1a), a dramatic decrease in the Soret band (405 nm) was observed under the identical conditions for the *cyt b₅* H39S variant where the heme group is five-coordinated by mutating an axial ligand His39 to a Ser39 (Fig. 1b). Concurrently, the 525 nm band decreased in intensity, and the 625 nm band increased in intensity, suggesting that the heme coordination state remains the same, whereas a disturbance of micro-environment occurs as a result of UO_2^{2+} binding.

Difference spectra upon UO_2^{2+} titration of *cyt b₅* H39S are shown in Fig. 1c. The changes in the Soret region were fitted to a double reciprocal plot (inset, top) and a Hill plot (inset,

bottom) with a slope of 1.1, which indicates that there is a single uranyl-binding site with a K_D of 2.5 μM in *cyt b₅* H39S. Through the loss of one heme axial ligand, His39, *cyt b₅* H39S exhibits lower stability compared to native *cyt b₅* [12]. On the other hand, with the same surface residues as in *cyt b₅*, the low stability of *cyt b₅* H39S facilitates probing of the interactions between UO_2^{2+} and *cyt b₅* by producing more obvious observations in UV-vis spectra upon uranyl ion titration.

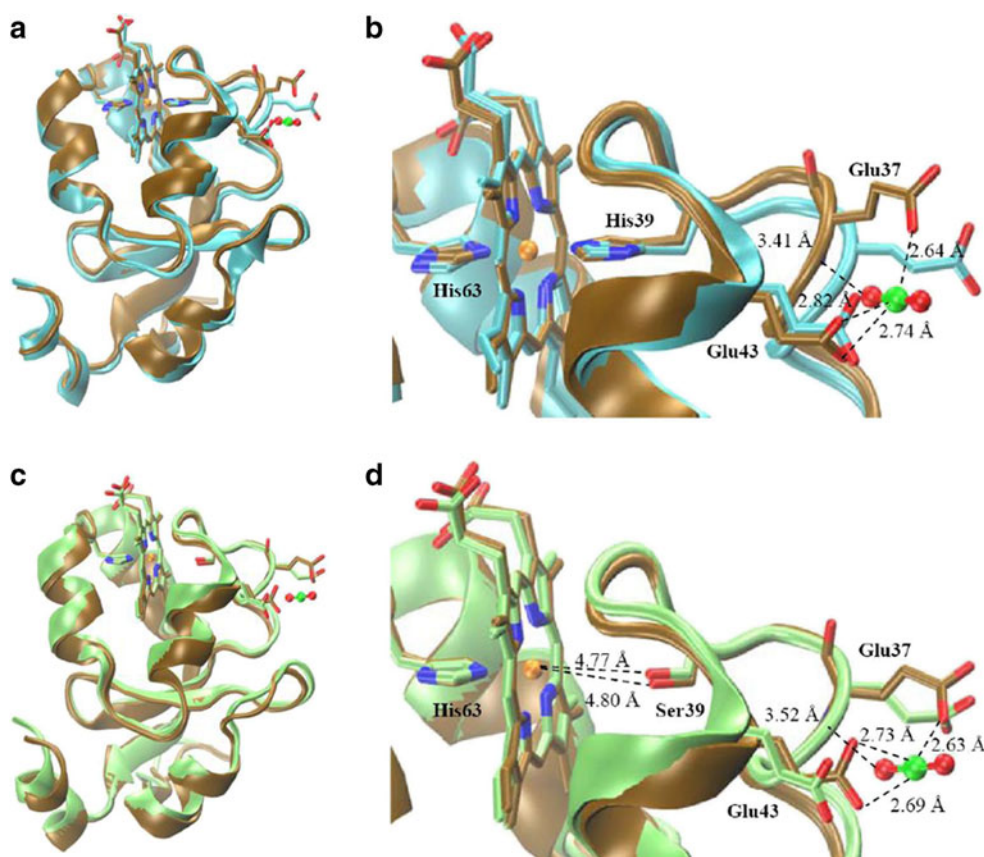
Uranyl ion binding from modeling studies

In order to reveal the uranyl ion binding site in *cyt b₅* and present a detailed view of the interactions between UO_2^{2+} and *cyt b₅* at an atomic level, we carried out molecular simulation studies for UO_2^{2+} binding to *cyt b₅* and the *cyt b₅* H39S variant. The *in silico* approach has been shown to be capable of providing structural insights into protein-uranyl interactions that might otherwise be difficult to obtain experimentally [5]. Based on the crystal structure of bovine liver microsomal *cyt b₅* (PDB entry 1CYO) [14], we first modeled a structure of *cyt b₅* H39S variant for comparison with the uranyl-bound form. To search for the UO_2^{2+} binding site in *cyt b₅*, we performed a close inspection of the crystal structure and found that water molecule HOH578 forms hydrogen bonds with Glu37 and Glu43 simultaneously. From our experiences in the rational design of metal-binding sites above the heme group in another heme protein, myoglobin [18, 21, 22], this water site might be a potential metal binding site for UO_2^{2+} . We thus changed the water to a UO_2^{2+} ion and carried out computer modeling for UO_2^{2+} bound forms of *cyt b₅*, namely, U-*cyt b₅* and U-*cyt b₅* H39S, respectively.

Figure 2a shows the spatial alignment of the crystal structure of *cyt b₅* and the simulated structure of U-*cyt b₅*. A detailed view of the heme group and the UO_2^{2+} binding site is presented in Fig. 2b. It can be seen that, in the presence of a UO_2^{2+} ion at the water-578 site, Glu37 and Glu43 coordinate to the U atom via one and two O atoms with a distance of 2.64 Å, 2.74 Å and 2.82 Å, respectively. The distances are close to the maximum values reported for carboxylate monodentate (2.61 Å) and bidentate (2.84 Å) ligand of uranyl ions in protein crystal structures, respectively [5]. A hydrogen bond is also formed between one uranyl oxo group and the backbone amide group of Glu37 (3.41 Å). As a consequence, the conformation was changed slightly in the region of Glu37. Meanwhile, no obvious alteration occurred to the heme axial ligand His39, in agreement with UV-vis studies (Fig. 1a).

By contrast, in the case of the *cyt b₅* H39S variant (Fig. 2c, d), Ser39 shifts slightly apart from the heme group in U-*cyt b₅* H39S with respect to *cyt b₅* H39S, where Glu37 and Glu43 coordinate to the U atom with shorter distances (2.63 Å, 2.69 Å

Fig. 2 Overlay of *cyt b₅* (cyan) with simulated U-*cyt b₅* (ochre) (a full view, **a**, and a binding site view, **b**), and simulated *cyt b₅* H39S (lime) with U-*cyt b₅* H39S (ochre) (a full view, **c**, and a binding site view, **d**). Residues Glu37 and Glu43, and two heme axial ligands, His39 and His63, as well as mutated Ser39, are highlighted



and 2.73 Å) compared to U-*cyt b₅*. This suggests that UO_2^{2+} binds tightly to the variant, despite the weakened hydrogen bonding interaction (3.52 Å). This observation can be attributed to the elimination of one heme axial ligand, His39, resulting in a conformation suitable for UO_2^{2+} coordination by both Glu37 and Glu43 with slight conformational changes. Note that beyond obtaining the basic structural information of uranyl binding, future simulations should be directed to perform quantitative free energy calculations to study the binding of UO_2^{2+} to *cyt b₅* as well as its H39S variant by developing parameters for uranyl ion.

Functional consequences of uranyl ion binding

To further probe the consequences of UO_2^{2+} binding to *cyt b₅* in terms of protein function, we evaluated the peroxidase activity of *cyt b₅* H39S [10], as affected by UO_2^{2+} ion binding. Peroxidase activity has been shown to be an efficient tool for studying conformational changes in the heme active site during folding and unfolding of heme proteins [23, 24]. As shown in Fig. 3, the initial rate of guaiacol oxidation decreases from $6.24 \mu\text{M}\cdot\text{min}^{-1}$ of *cyt b₅* H39S to $2.61 \mu\text{M}\cdot\text{min}^{-1}$ with five equivalents of UO_2^{2+} , suggesting a disruption of peroxidase activity by partial dissociation of the heme group from the heme binding domain,

as indicated in UV-vis spectra (Fig. 1b). Meanwhile, in the presence of one equivalent of UO_2^{2+} , U-*cyt b₅* H39S exhibits peroxidase activity ($5.98 \mu\text{M}\cdot\text{min}^{-1}$) similar to that of *cyt b₅* H39S, which agrees with the modeling result indicating that the heme active site was slightly altered (Fig. 2d), as well as UV-vis observations at one equivalent of UO_2^{2+} (Fig. 1b). In control experiments, it was interesting to observe that U-*cyt b₅* shows

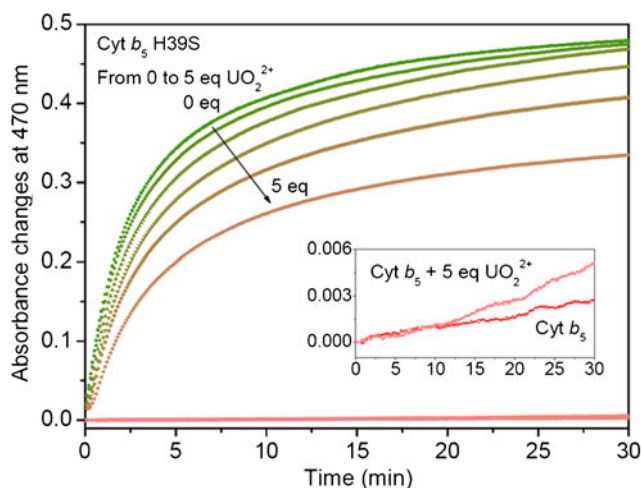


Fig. 3 Time-dependent guaiacol oxidation with H_2O_2 catalyzed by *cyt b₅* H39S in the presence of 0–5 equivalents of UO_2^{2+} (arrow) and by *cyt b₅* in the absence or presence of 5 equivalents of UO_2^{2+} (inset)

a slightly increased peroxidase activity compared to native *cyt b₅* (0.065 vs. 0.048 $\mu\text{M}\cdot\text{min}^{-1}$), due to the alteration of the heme active site by excess UO_2^{2+} ions.

Conclusions

In this study, we investigated the interactions of UO_2^{2+} and *cyt b₅* as well as the *cyt b₅* H39S variant using both experimental and theoretical methods. Based on experimental observations, a uranyl ion binding site was proposed in *cyt b₅* at the surface residues Glu37 and Glu43 by molecular simulation. These insights revealed at the atomic level shed light on the mechanism of uranyl toxicity in general. The impact of UO_2^{2+} on the structure and function of *cyt b₅*, in turn, may further interfere with the interactions of *cyt b₅* and its partners in biological systems, such as *cyt c* [25] and *cyt P450* [26]. These interactions are currently under investigation.

Acknowledgments We gratefully thank Prof. Zhong-Xian Huang at Fudan University, Shanghai, China, for providing the cytochrome *b₅* gene and proteins, and Dr. Tianlei Ying at the National Institutes of Health (NIH), USA, for discussions regarding simulation of the uranyl ion. NAMD and VMD were developed by the Theoretical Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign, USA. This work is supported by the National Natural Science Foundation of China (NSFC Nos. 20877038, 10975069).

References

- Gorden AE, Xu J, Raymond KN, Durbin P (2003) *Chem Rev* 103:4207–4282
- Vidaud C, Gourion-Arsiquaud S, Rollin-Genetet F, Torme-Celer C, Plantevin S, Pible O, Berthomieu C, Quéménéur E (2007) *Biochemistry* 46:2215–2226
- Montavon G, Apostolidis C, Bruchertseifer F, Repinc U, Morgenstern A (2009) *J Inorg Biochem* 103:1609–1616
- Michon J, Frelon S, Garnier C, Coppin F (2010) *J Fluoresc* 20:581–590
- Pible O, Guilbaud P, Pellequer JL, Vidaud C, Quéménéur E (2006) *Biochimie* 88:1631–1638
- Van Horn JD, Huang H (2006) *Coord Chem Rev* 250:765–775
- Wegmer SV, Boyaci H, Chen H, Jensen MP, He C (2009) *Angew Chem Int Edn* 48:2339–2341
- Lins RD, Vorpapel ER, Guglielmi M, Straatsma TP (2008) *Biomacromolecules* 9:29–35
- Vergères G, Waskell L (1995) *Biochimie* 7:604–620
- Wang WH, Wang YH et al (2002) *Chem Lett* 674–675
- Chudaev MV, Gilep AA, Usanov SA (2001) *Biochemistry (Moscow)* 66:667–681
- Wang WH, Lu JX, Yao P, Xie Y, Huang ZX (2003) *Protein Eng* 6:1047–1054
- Sigman JA, Kwok BC, Lu Y (2000) *J Am Chem Soc* 122:8192–8196
- Durley RC, Mathews FS (1996) *Acta Crystallogr D* 52:65–76
- Wang ZH, Lin YW, Rosell FI, Ni FY, Lu HJ, Yang PY, Tan XS, Li XY, Huang ZX, Mauk AG (2007) *ChemBiochem* 8:607–609
- Humphrey W, Dalke A, Schulten K (1996) *J Mol Graph* 14:33–38
- Kalé L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, Schulten K (1999) *J Comput Phys* 151:283–312
- Yeung N, Lin YW, Gao YG, Zhao X, Russell BS, Lei L, Miner KD, Robinson H, Lu Y (2009) *Nature* 462:1079–1082
- Baldwin DA, Marques HM, Pratt JM (1987) *J Inorg Biochem* 30:203–217
- Moore GR, Pettigrew GW (1990) *Cytochrome c: evolutionary, structural and physicochemical aspects*. Springer, Berlin
- Lin YW, Yeung N, Gao YG, Miner KD, Tian S, Robinson H, Lu Y (2010) *Proc Natl Acad Sci USA* 107:8581–8586
- Lin YW, Yeung N, Gao YG, Miner KD, Lei L, Robinson H, Lu Y (2010) *J Am Chem Soc* 132:9970–9972
- Lin YW, Wang WH, Zhang Q, Lu HJ, Yang PY, Xie Y, Huang ZX, Wu HM (2005) *ChemBiochem* 6:1356–1359
- Diederix RE, Ubbink M, Canters GW (2002) *Biochemistry* 41:13067–13077
- Ren Y, Wang WH, Case M, Qian W, McLendon G, Huang ZX (2004) *Biochemistry* 43:3527–3536
- Im SC, Waskell L (2011) *Arch Biochem Biophys* 507:144–153

The insertion reactions of the p-complex silylenoid H_2SiLiF with Si-X (X=F, Cl, Br, O, N) bonds

Yuhua Qi · Bing Geng · Zhonghe Chen

Received: 17 March 2011 / Accepted: 15 May 2011 / Published online: 10 June 2011
© Springer-Verlag 2011

Abstract The insertion reactions of the silylenoid H_2SiLiF with $\text{SiH}_3\text{XH}_{n-1}$ (X=F, Cl, Br, O, N; n=1, 1, 1, 2, 3) have been studied by DFT calculations. The results indicate that the insertions proceed in a concerted manner, forming $\text{H}_3\text{SiSiH}_2\text{XH}_{n-1}$ and LiF. The essence of H_2SiLiF insertion into Si-X bonds reactions are the donations of the electrons of X into the p orbital on the Si atom in H_2SiLiF and the σ electrons on the Si atom in H_2SiLiF to the positive SiH_3 group. The order of reactivity by H_2SiLiF insertion in vacuum indicates the reaction barriers decrease for the same-row element X from right to left and the same-family element X from up down in the periodic table. The insertion reactions in ether are similar to those in vacuum. The energy barriers in vacuum are higher than those in ether. The silylenoid insertions are thermodynamically exothermic both in vacuum and in ether.

Keywords DFT · Insertion reactions · Silylenoids · Theoretical study

Introduction

Silylenoids, R_2SiMX (X=halogen, M=alkali metal), are important intermediates in silicon hybrid and organosilicon

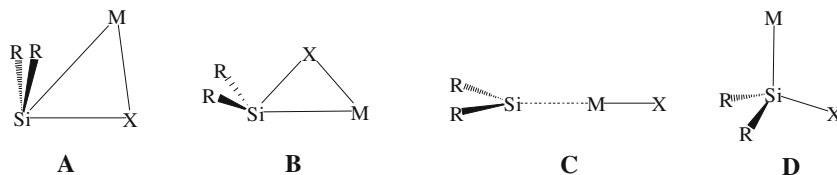
chemistry [1, 2]. As a kind of very reactive species, the preparation of silylenoids is very difficult. In 1995, Tamao et al. [3] reported the first experimental study of silylenoid chemistry and detected the existence of [(tert-butoxy) diphenylsilyl]lithium, $\text{Ph}_2\text{SiLi}(\text{O}i\text{Bu}-t)$. Recently, a great breakthrough has been made in the research of silylenoids. Lee et al. [4] reported the syntheses of stable halosilylenoids $(\text{Tsi})\text{X}_2\text{SiLi}$ ($\text{Tsi}=\text{C}-(\text{SiMe}_3)_3$; X=Br, Cl) at room temperature. In 1980, Clark et al. [5] theoretically studied the isomers of lithoflurosilylenoid H_2SiLiF by *ab initio* calculations for the first time. Since 1990s, we have studied some silylenoids such as $\text{R}_1\text{R}_2\text{SiMX}$ ($\text{R}_1, \text{R}_2=\text{H}, \text{F}, \text{OH}, \text{NH}_2, \text{Me}, \text{Et}$; X=F, Cl, Br; M=Li, Na, K, etc.) by quantum chemistry methods. Specifically, we have investigated their structures, their stability, isomerization, insertion reactions, and addition reactions [6–10]. Both experimental [3, 4] and theoretical [5–10] results show that silylenoids have ambiphilic character, nucleophilicity and electrophilicity, and can take part in many reactions. Such reactions were recognized as important and effective methods for preparation of the new silicon-bonded and heterocyclic silicon compounds.

For insertion reactions, experimental and theoretical studies have been concluded on the insertions of silylenoids into the bonds of H_2 [11], C-X (X=F, Cl, Br, O, N) [12] and Y-H (Y=C, Si, N, P, O, S, and F) [13]. However, the insertion reactions of silylenoids into Si-X (X=F, Cl, Br, O, N) have not been systematically reported to our knowledge. The elucidation of the mechanism of these insertion reactions can provide a new reaction mode of silicon-silicon bond formation. Previous calculations have shown that each silylenoid R_2SiMX , which can be regarded as the complex of the silylene R_2Si and the metal halide MX, has four equilibrium isomers, the p-complex (**A**), the three-

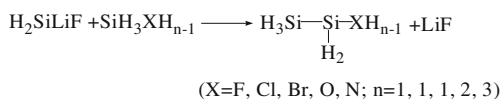
Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1129-x) contains supplementary material, which is available to authorized users.

Y. Qi (✉) · B. Geng · Z. Chen
Shandong Provincial Key Laboratory of Fluorine Chemistry and Chemical Materials, School of Chemistry and Chemical Engineering, University of Jinan,
Jinan 250022, People's Republic of China
e-mail: chm_qiyh@ujn.edu.cn

membered-ring (**B**), the σ -complex (**C**), and the ‘classical’ tetrahedral (**D**) structures [5, 7, 14–16].



Whether in vacuum or in various solvents, the p-complex structure has the lowest energy [17]. So the insertion reactions of the p-complex silylenoid H_2SiLiF with $\text{SiH}_3\text{XH}_{n-1}$ ($\text{X}=\text{F}, \text{Cl}, \text{Br}, \text{O}, \text{N}$; $n=1, 1, 1, 2, 3$) are invested in this paper.



Through this theoretical work, we hope (i) to clarify the reaction mechanisms and to determine the structures and energies of all stationary points, (ii) to investigate the thermodynamics of these insertion reactions, (iii) to estimate their activation barriers, (iv) to establish general trends and predictions for the insertion reactions of silylenoids into Si-X bonds ($\text{X}=\text{F}, \text{Cl}, \text{Br}, \text{O}, \text{N}$), (v) to reveal the solvent effects on the insertion reactions of silylenoids with $\text{SiH}_3\text{XH}_{n-1}$ ($\text{X}=\text{F}, \text{Cl}, \text{Br}, \text{O}, \text{N}$; $n=1, 1, 1, 2, 3$).

Computational methods

Optimized geometries for the stationary points were obtained at the B3LYP/6-311+G (d, p) [18–20] level. The corresponding harmonic vibrational frequency calculations were carried out in order to characterize all stationary points as either local minima (no imaginary frequencies) or transition states (one imaginary frequency). Based on the optimized geometries, energies were obtained and natural bond orbital (NBO) [21–23] analyses were then used to study the nature of different interactions between atoms and groups. The reaction paths were examined by intrinsic reaction coordinate (IRC) [24] calculations. The solvent effects, which were simulated using the self-consistent reaction field (SCRf) method with Tomasi’s polarized continuum model (PCM) [25–33], were investigated at the same level. Gaussian 03 [34] series of programs were employed in all calculations.

Results and discussion

As shown in Fig. 1, the p-complex silylenoid H_2SiLiF (**A**) can be regarded as a singlet complex in which electrons from the F in LiF are donated into the unoccupied p orbital of the Si atom in H_2Si . For the convenience of expression, the Si atom in **A** is marked as Si^1 .

The molecular electrostatic potentials of H_2SiLiF (**A**) and $\text{SiH}_3\text{XH}_{n-1}$ are shown in Fig. 2. There are three maxima in the electrostatic potential of H_2SiLiF , which are situated on the F atom, Si atom (negative) and on the Li atom (positive), respectively. For the conditions of $\text{X}=\text{F}, \text{Cl}$ and Br, there are two maxima in the molecular electrostatic potential of $\text{SiH}_3\text{XH}_{n-1}$, located on the X atoms (negative) and the Si atom (negative), respectively. In the molecular electrostatic potentials of SiH_3OH and SiH_3NH_2 , both the negative and positive maxima lie on the O atom and N atom, respectively. So the calculated electrostatic potentials indicate that the Si atom in H_2SiLiF , X and Si atoms in $\text{SiH}_3\text{XH}_{n-1}$, are active atoms in interacting with other molecules.

When $\text{SiH}_3\text{XH}_{n-1}$ approaches H_2SiLiF with the X and SiH_3 ends of $\text{SiH}_3\text{XH}_{n-1}$ attacking the p orbital and the σ lone pair electrons of Si^1 atom respectively, (see (**B**) in Fig. 1), insertion reactions take place. Figures 3 and 4 show the structures of some stationary points, and Supporting information lists orthogonal coordinates for others. The total energies together with the zero-point energies (ZPEs) and relative energies (relative to the corresponding reactants) of all stationary points are described by Table 1.

Insertion reaction of A into Si-Cl

When SiH_3Cl approaches **A**, the initial formation of the precursor complex **CIM1** is facilitated by the interaction between the p orbital on Si^1 and the negative Cl atom of SiH_3Cl . Compared with the structures of SiH_3Cl and **A** molecules, the SiH_3Cl and **A** moieties in **CIM1** changes little. The long $\text{Si}^1\text{-Cl}$ length (4.157 Å) in **CIM1** and the small relative energy of **CIM1** (-1.5 kJ mol^{-1}) indicate that the $\text{Si}^1\text{...Cl}$ interaction is very weak.

Fig. 1 The p-complex H_2SiLiF (a) and its insertion (b) reaction paths with $\text{SiH}_3\text{XH}_{n-1}$ ($X=\text{F}, \text{Cl}, \text{Br}, \text{O}, \text{N}$; $n=1, 1, 1, 2, 3$)

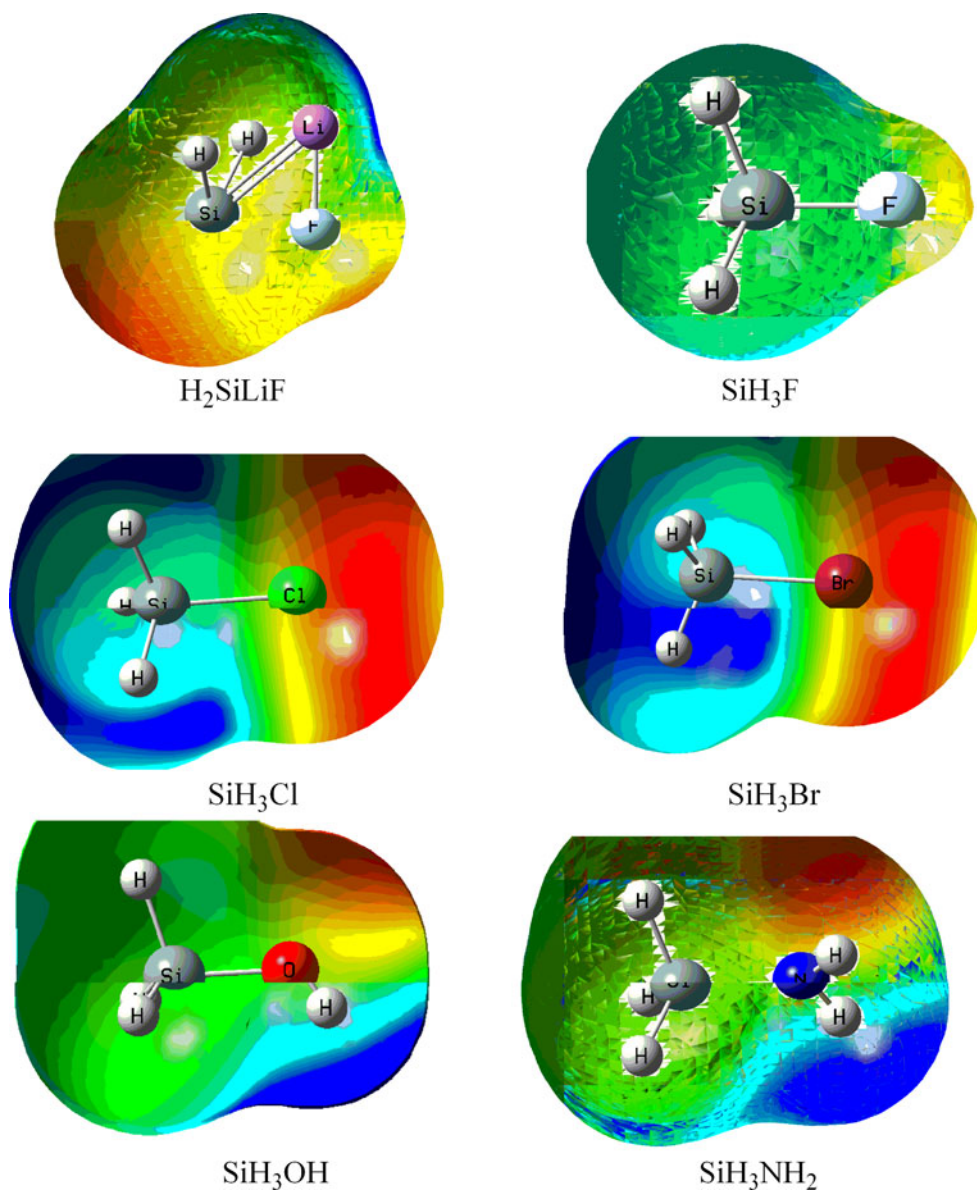
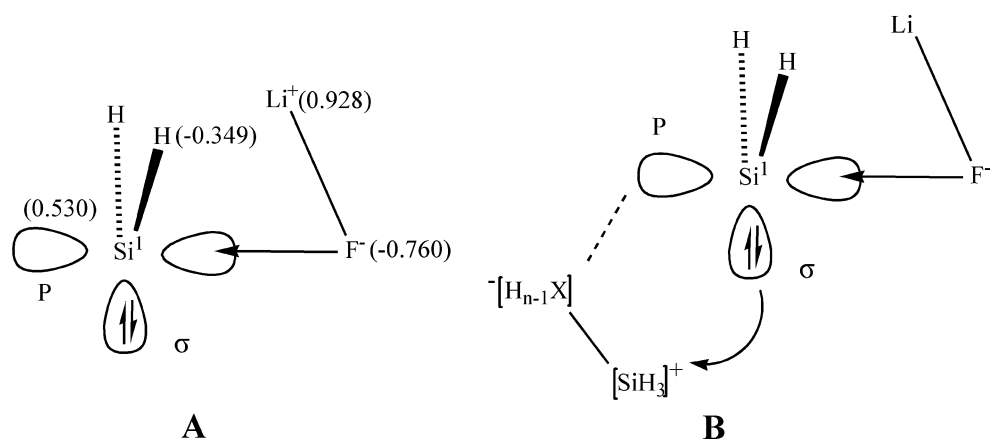
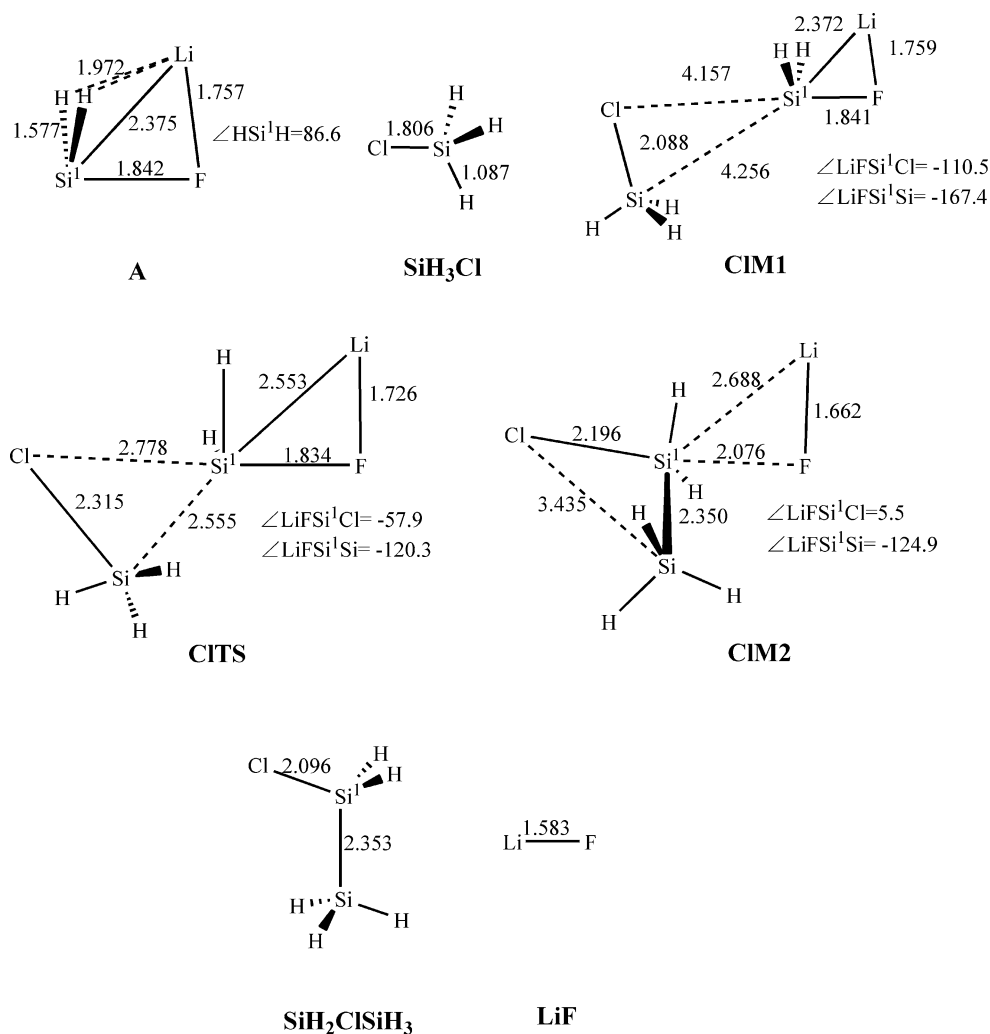


Fig. 2 Molecular electrostatic potential of H_2SiLiF and $\text{SiH}_3\text{XH}_{n-1}$ ($X=\text{F}, \text{Cl}, \text{Br}, \text{O}, \text{N}$; $n=1, 1, 1, 2, 3$) at the B3LYP/6-311+G (d, p) level. Blue denotes maximal molecular electrostatic potential, which is labeled by black arrows; red denotes negative molecular electrostatic potential

Fig. 3 The B3LYP/6-311+G (d, p) geometries (in Å and (°)) for the stationary points in the insertion reaction of H_2SiLiF with SiH_3Cl



As shown in Fig. 1, two electron donation effects contribute to the proceeding of the insertion reaction. One is the donation of the electrons of Cl into the p orbital on the Si^1 atom. The other is the donation of the σ electrons on the Si^1 atom to the positive SiH_3 group. The electron donations make the formation of the transition state **CITS**, whose only one imaginary frequency is 173.61 cm^{-1} . In **CITS**, the natural charge of the Si^1 atom is 0.195 higher than that in **CIM1**(0.517), while the natural charge of the SiH_3Cl moiety decrease from the positive charge (0.005) in **CIM1** to the negative charge (-0.398) in **CITS**. This suggests that the Si^1 atom has denoted electrons to the SiH_3Cl moiety. The insertion reaction path was also fully confirmed by the IRC computations (see Fig. 5). It is obvious that the bond lengths, Si^1 -Si, Si^1 -Cl, and Si-Cl, change strongly in the course of the reaction. The Si^1 -Si and Si^1 -Cl bonds rapidly shorten from the reactant side. The Si-Cl bond lengthens. The relative energy of **CITS** is 53.3 kJ mol^{-1} .

After getting over the transition state **CITS**, **CIM2** are gradually formed with the LiF moiety leaving from the Si^1

atom. In fact, **CIM2** is a complex of $\text{H}_3\text{SiSiH}_2\text{Cl}$ and LiF. The energy of **CIM2** is 44.7 kJ mol^{-1} lower than the sum of the energies of $\text{H}_3\text{SiSiH}_2\text{Cl}$ and LiF molecules.

As shown in Table 1, the insertion reaction is exothermic by 38.6 kJ mol^{-1} for the **A**+ SiH_3Cl system.

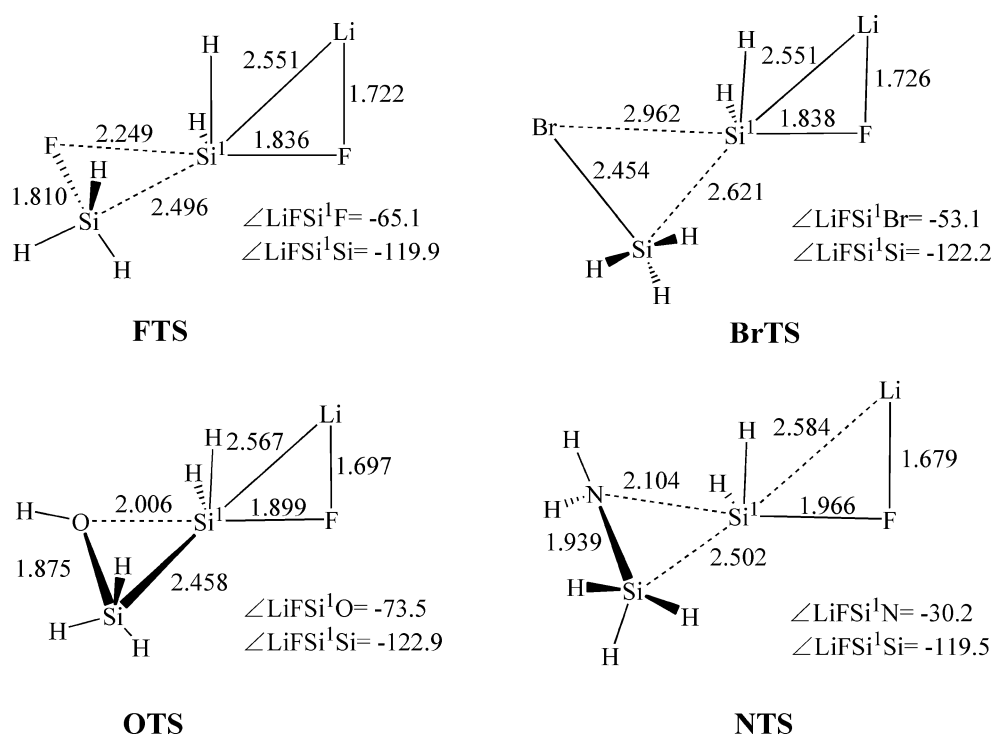
Insertion reactions of **A** into Si-X (X=F, Br, O, N)

The insertion processes of **A** and $\text{SiH}_3\text{XH}_{n-1}$ (X=F, Br, O, N; n=1, 1, 2, 3) are similar to that of **A** and SiH_3Cl .

In the precursor complex **XM1**, the Si^1 -X distances are 3.251 (X=F), 3.916 (X=Br), 2.817 (X=O), and 2.877 Å (X=N), respectively. The energies of **XM1** are lower than their corresponding reactants by 2.1 (X=F), 1.2 (X=Br), 0.9 (X=O), and 1.5 kJ mol^{-1} (X=N), respectively. The long Si^1 -X distances and the small stability energies of **XM1** indicate that there is only weak interaction between the Si atom and the X atom in **XM1**, and **XM1** is instable.

The transition states **XTS** are confirmed by calculation of the energy Hessian. The model calculations estimate that

Fig. 4 The B3LYP/6-311+G (d, p) geometries (in Å and (°)) for the transition states XTS of in the insertion reactions of H_2SiLiF with $\text{SiH}_3\text{XH}_{n-1}$ (X=F, Br, O, N; n=1, 1, 2, 3)



the relative energies of **FTS**, **BrTS**, **OTS**, and **NTS** are 56.9, 47.0, 58.1, and 78.7 kJ mol^{-1} , respectively. That is, the reaction barriers of the insertions into Si-X bonds decrease for the same-row element X from right to left and the same-family element X from up down in the periodic table.

The intermediate **XM2** can further decompose to substituted silane $\text{H}_3\text{SiSiH}_2\text{XH}_{n-1}$ and LiF. The energies of **XM2** are below the sum of the energies of $\text{H}_3\text{SiSiH}_2\text{XH}_{n-1}$ and LiF by 51.9 (X=F), 45.5 (X=Br), 38.5 (X=O), and 37.0 kJ mol^{-1} (X=N), respectively.

It is apparent that the reaction enthalpy for the **A** insertions are 34.6 (Si-F), 39.9 (Si-Br), 35.1 (Si-O), and 34.5 kJ mol^{-1} (Si-N), respectively.

The silylenoid insertions into Si-X (X=F, Cl, Br, O, N) bonds are similar to silylenoid insertions into C-X [12] in the reaction processes and mechanisms. The calculated reaction barriers for the C-X insertions at the B3LYP/6-311+G (d, p) level are 168.0 (X=F), 181.7 (X=Cl), 171.7 (X=Br), 183.8 (X=O) and 219.4 kJ mol^{-1} (X=N), respectively. So the insertions into Si-X bonds are easier than the corresponding insertions into C-X bonds.

Solvent effects on the insertion reactions

Silylenoid reactions often take place in solvents, so solvent effects on the reactions are conducted. Ether is chosen as the solvent. Calculation results can be summed as follows.

First, the insertion processes in solvents are same to that in vacuum, which is concerted reactions, involving the formation of precursor complexes, transition states and insertion products. The geometry structures (see [Supporting information](#)) of stationary points in solvents (ether, THF and acetone, respectively) are correspondingly similar to those in vacuum. Whether in ether or in vacuum, the essence of these reactions are the donations of the electrons of X into the p orbital on the Si^1 atom and the σ electrons on the Si^1 atom to the positive SiH_3 group.

Second, calculated energies of the stationary points are listed in Table 1. Several conclusions can be drawn from these calculations. (1) Energies of all stationary points are in the order of $E_{\text{ether}} < E_{\text{vacuum}}$, indicating that the thermal stabilities of the stationary points are larger in ether than in vacuum. (2) The insertion barriers for the p-complex structures are 36.3 (X=F), 38.3 (X=Cl), 36.1 (X=Br), 54.5 (X=O), and 73.8 (X=N) kJ mol^{-1} at the B3LYP/6-311+G (d, p) level. Compared with those in vacuum, the barrier heights in vacuum are higher than those in ether, showing insertion reactions are easy to occur in ether. For the X=F, Cl, Br conditions, the energy barriers change very little (the maximum difference is 2.2 kJ mol^{-1} , which is between the cases of CH_3Cl and CH_3Br) in ether, whereas, the reaction barriers in vacuum decrease for the same-family element X from up down in the periodic table. For the X=F, O, and N cases both in ether and in vacuum, there is a very clear trend for the same-row element X from right to left in the periodical table. (3) Same with those in vacuum,

Table 1 Total energies (a.u.) and relative energies (kJ mol⁻¹, in parentheses) for reactants, intermediates, transition states and products of the insertion reactions at the B3LYP/6-311+G (d, p) level

Molecules	$E_{\text{in vacuum}}$	$E_{\text{in ether}}$
A+SiH ₃ F	-789.37817(0.0)	-789.40198(0.0)
FM1	-789.37897(-2.1)	-789.40033(4.3)
FTS	-789.35649(56.9)	-789.38816(36.3)
FM2	-789.41111(-86.5)	-789.43794(-94.4)
H ₃ SiSiH ₂ F+LiF	-789.39136(-34.6)	-789.43080(-75.7)
A+SiH ₃ Cl	-1149.72103(0.0)	-1149.74209(0.0)
CIM1	-1149.72161(-1.5)	-1149.74150(1.6)
CITS	-1149.70073(53.3)	-1149.72750(38.3)
CIM2	-1149.75278(-83.3)	-1149.78057(101.0)
H ₃ SiSiH ₂ Cl+LiF	-1149.735734(-38.6)	-1042.77334(-82.0)
A+SiH ₃ Br	-3263.64087(0.0)	-3263.66161(0.0)
BrM1	-3263.64133(-1.2)	-3263.66118(1.1)
BrTS	-3263.62295(47.0)	-3263.64785(36.1)
BrM2	-3263.67340(-85.4)	-3263.70146(-104.6)
H ₃ SiSiH ₂ Br+LiF	-3263.65608(-39.9)	-3263.69345(-83.6)
A+SiH ₃ OH	-765.32781(0.0)	-765.35382(0.0)
OM1	-765.32816(-0.9)	-765.35110(7.2)
OTS	-765.30569(58.1)	-765.33308(54.5)
OM2	-765.35586(-73.6)	-765.38408(-79.4)
H ₃ SiSiH ₂ OH+LiF	-765.34120(-35.1)	-765.38337(-77.6)
A+SiH ₃ NH ₂	-745.42727(0.0)	-745.45095(0.0)
NM1	-745.42785(-1.5)	-745.44961(3.5)
NTS	-745.39729(78.7)	-745.42283(73.8)
NM2	-745.45449(-71.5)	-745.48025(-76.9)
H ₃ SiSiH ₂ NH ₂ +LiF	-745.44043(-34.5)	-745.48059(-77.8)

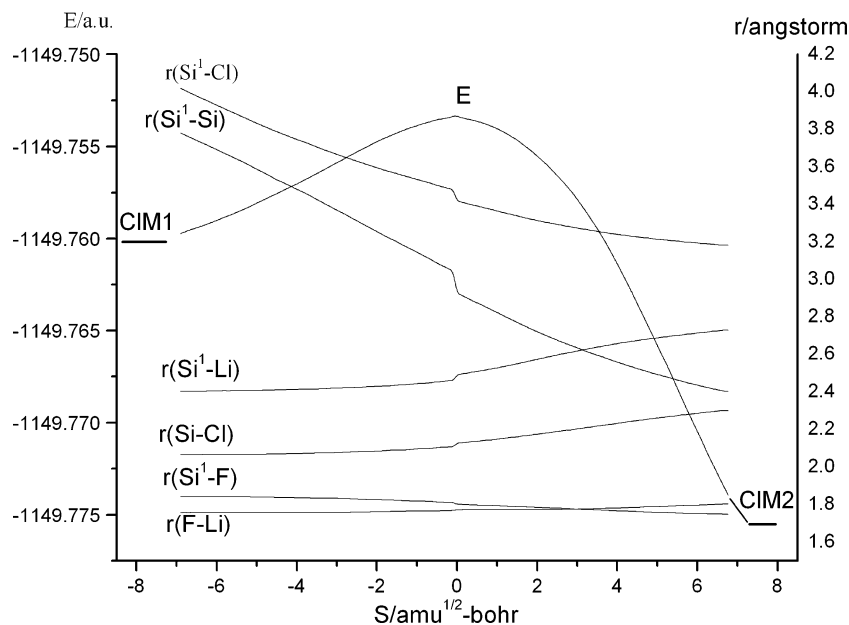
H₃SiSiH₂XH_{n-1} and LiF are expected final products for the insertion reactions in ether. (4) It is apparent that all the silylenoid insertions in ether are thermodynamically exothermic.

Concluding remarks

In the present work, we have studied the reaction mechanisms of silylenoid H₂SiLiF insertions with SiH₃XH_{n-1} (X=F, Cl, Br, O, N; n=1, 1, 1, 2, 3) by DFT theory. It should be mentioned that this study has provided the first theoretical demonstration about the reaction trajectory and theoretical estimation of the activation energy and reaction enthalpy for those processes.

- (1) The theoretical results indicate that the insertion reactions of H₂SiLiF with SiH₃XH_{n-1} occur in a concerted manner, forming silanes H₃SiSiH₂XH_{n-1} and LiF.
- (2) The essence of H₂SiLiF insertion into Si-X bonds, reactions are the donations of the electrons of X into the p orbital on the Si atom in H₂SiLiF and the σ electrons on the Si atom in H₂SiLiF to the positive SiH₃ group.
- (3) For Si-X bonds, the order of reactivity by H₂SiLiF insertion in vacuum indicates the reaction barriers decrease for the same-row element X from right to left and the same-family element X from up down in the periodic table.

Fig. 5 Energy (E) and bond distance (r) vs. reaction coordinate (S) in the insertion reaction of H₂SiLiF with SiH₃Cl at the B3LYP/6-311+G (d, p) level. The Si atom in H₂SiLiF is marked as Si¹



- (4) The insertion reactions in ether are similar to those in vacuum. The energy barriers in vacuum are higher than those in ether, showing insertion reactions are easy to occur in ether.
- (5) The silylenoid insertions are thermodynamically exothermic both in vacuum and in ether.

Acknowledgments This work was supported financially by the PhD Foundation of University of Jinan (No. XBS0924).

References

1. Gilman H, Peterson D (1965) *J Am Chem Soc* 87:2389–2394
2. Nefedow O, Manakow M (1964) *Angew Chem* 76:270–270
3. Tamao K, Kawachi A (1995) *Angew Chem Int Ed Engl* 34:818–820
4. Lee M, Hyeon M, Lim Y, Choi J, Park C, Jeong S, Lee U (2004) *Chem Eur J* 10:377–381
5. Clark T, Schleyer PvR (1980) *J Organomet Chem* 191:347–353
6. Feng S, Ju G, Deng C (1991) *Sci China B* 9:907–914
7. Feng S, Feng D, Deng C (1993) *Chem Phys Lett* 214:97–102
8. Feng D, Feng S, Deng C (1996) *Chem J Chin Univ* 17:1108–1111
9. Feng D, Feng S, Deng C (1995) *Chin J Chem* 13:481–486
10. Feng D, Feng S, Deng C (1998) *Chem J Chin Univ* 19:451–454
11. Feng S, Feng D, Li J (2000) *Chem Phys Lett* 316:146–150
12. Qi Y, Feng D, Feng S (2010) *Struct Chem* 21:879–884
13. Xie J, Feng D, Feng S (2006) *J Organomet Chem* 691:08–223
14. Feng S, Feng D (2001) *J Mol Struct THEOCHEM* 541:171–177
15. Feng D, Xie J, Feng S (2004) *Chem Phys Lett* 396:245–251
16. Feng S, Feng D, Deng C (1995) *Chin J Chem* 13:19
17. Xie J, Feng D, Feng S, Ding Y (2007) *Struct Chem* 18:65–70
18. Beck A (1993) *J Chem Phys* 98:5648–5653
19. Beck A (1988) *Phys Rev A* 38:3098–3100
20. Lee C, Yang W, Parr R (1988) *Phys Rev B* 37:785–789
21. Foster J, Weinhold F (1980) *J Am Chem Soc* 102:7211–7218
22. Reed A, Weinhold F (1983) *J Chem Phys* 78:4066–4074
23. Weinhold F (1998) In: Schleyer PvR (ed) *Encyclopedia of Computational Chemistry*, Vol 3. Wiley
24. Curtiss L, Redfern P, Raghavachari K, Rassolov V, Pople J (1999) *J Chem Phys* 110:4703–4710
25. Miertus S, Scrocco E, Tomasi J (1981) *Chem Phys* 55:117–129
26. Miertus S, Tomasi J (1982) *Chem Phys* 65:239–245
27. Cossi M, Barone V, Cammi R, Tomasi J (1996) *Chem Phys Lett* 255:327–335
28. Cancès MT, Mennucci B, Tomasi J (1997) *J Chem Phys* 107:3032–3042
29. Cossi M, Barone V, Mennucci B, Tomasi J (1998) *Chem Phys Lett* 286:253–260
30. Barone V, Cossi M, Tomasi J (1998) *J Comput Chem* 19:404–417
31. Barone V, Cossi M (1998) *J Phys Chem A* 102:1995–2001
32. Mennucci B, Tomasi J (1997) *J Chem Phys* 106:5151–5159
33. Tomasi J, Mennucci B, Cancès E (1999) *J Mol Struct THEOCHEM* 464:211–226
34. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven NJ, Kudin TK, Burant JC, Millam JM, Iyengar SS, Tomasi JB, Mennucci V, Cossi BM, Scalmani GN, Rega G, Petersson A, 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 J, 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 IR, Martin LD, Fox J, Keith T, AlLaham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PM, Johnson WB, Chen W, Wong MW, Gonzalez C, Pople JA (2003) *Gaussian 03*, revision. Gaussian Inc, Pittsburgh, PA

Predicting the potency of hERG K⁺ channel inhibition by combining 3D-QSAR pharmacophore and 2D-QSAR models

Yayu Tan · Yadong Chen · Qidong You · Haopeng Sun ·
Manhua Li

Received: 5 February 2011 / Accepted: 23 May 2011 / Published online: 10 June 2011
© Springer-Verlag 2011

Abstract Blockade of the hERG K⁺ channel has been identified as the most important mechanism of QT interval prolongation and thus inducing cardiac risk. In this work, an ensemble of 3D-QSAR pharmacophore models was constructed to provide insight into the determinants of the interactions between the hERG K⁺ channel and channel inhibitors. To predict hERG inhibitory activities, the predicted values from the ensemble of models were averaged, and the results thus obtained showed that the predictive ability of the combined 3D-QSAR pharmacophore model was greater than those of the individual models. Also, using the same training and test sets, a 2D-QSAR model based on a heuristic machine-learning method was developed in order to analyze the physico-chemical characters of hERG inhibitors. The models indicated that the inhibitors have certain key inhibitory features in common, including hydrophobicity, aromaticity, and flexibility. A final model was developed by combining

the combined 3D-QSAR pharmacophore with the 2D-QSAR model, and this final model outperformed any other individual model, showing the highest predictive ability and the lowest deviation. This model can not only predict hERG inhibitory potency accurately, thus allowing fast cardiac safety evaluation, but it provides an effective tool for avoiding hERG inhibitory liability and thus enhanced cardiac risk in the design and optimization of new chemical entities.

Keywords 2D-QSAR model · 3D-QSAR pharmacophore model · Cardiac risk · hERG channel inhibitory potency

Introduction

In recent research, many clinical agents (or agents that are currently under investigation) from diverse therapeutic classes have been found to prolong the QT interval, inducing arrhythmia or even sudden death. Examples include the GI stimulant cisapride [1] and the antihistamine terfenadine [2], which have been withdrawn from the market due to the cardiac risk associated with their use. Since the blockade of the human ether-à-go-go-related gene (hERG) K⁺ channel has been identified as the most important mechanism of QT interval prolongation at cellular level [3], assessing hERG inhibitory liability has become a key action during the early safety screening of novel pharmaceuticals [4].

hERG encodes the K⁺ channel responsible for the cardiac rapidly activating delayed rectifier K⁺ current (I_{Kr}), which plays a predominant role in mediating membrane repolarization during the course of the action potential (AP) in ventricular myocytes, and thus influen-

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1136-y) contains supplementary material, which is available to authorized users.

Y. Tan · Q. You · H. Sun · M. Li
Department of Medicinal Chemistry,
China Pharmaceutical University,
24 Tongjiexiang,
Jiangsu 210009, People's Republic of China

Q. You
e-mail: youqidong@gmail.com

Y. Tan · Y. Chen (✉)
School of Basic Science,
China Pharmaceutical University,
24 Tongjiexiang,
Jiangsu 210009, People's Republic of China
e-mail: chenyardong@gmail.com

ces the duration of the QT interval [5]. The hERG K⁺ channel is a homotetramer with axial and rotational symmetry, and consists of four identical subunits each containing six transmembrane domains (S1–S6). The amino acids from the domain of S5–S6 along with four subunits together form the functional pore, which has a funnelform structure, and is the main region for the ligand–receptor interaction [6]. However, the crystal structure of the hERG protein is yet to be determined, so the mechanism of the ligand–receptor interaction is still unknown. On the other hand, mutation experiments have proven that most hERG inhibitors mainly interact with the aromatic residues Y652 and/or F656 within the pore, through hydrophobic effects and/or cation– π or π stacking interactions. Residues T623, S624 and V625, which are adjacent to the selectivity filter, are also involved in some instances. In addition to the fact that this channel has a larger inner cavity than other K⁺ channels, it can accommodate a broad spectrum of compounds [7–9]. The hERG K⁺ channel also presents distinctive dynamic behavior involving slow activation and deactivation but rapid inactivation [10] in order to exert its regulating action, and thus exhibits a highly flexible structure with symmetry. This suggests that the hERG protein and the inhibitors of this channel are likely to adopt distributed sites and various modes of interaction [11], which was also confirmed in recent modeling and docking research by Zachariae et al. [12].

There are currently many biological measurements that can be employed for hERG inhibition assays; for example, the patch-clamp technique is the “gold standard” [13]. However, they are all expensive and time-consuming, so it is more economical and convenient to develop *in silico* models that possess high predictive abilities. Many studies employing various computational techniques have been developed for the prediction of hERG inhibitory potency, such as ligand-based approaches that include common pharmacophore [14–16], two-dimensional quantitative structure–activity relationship (2D-QSAR) [17–19], and three-dimensional quantitative structure–activity relationship (3D-QSAR) [20] approaches, as well as structure-based approaches involving homology modeling and docking [21, 22]. Although these models are different, they share key features, and have all been shown to be feasible.

Several hERG pharmacophore models have already been generated, including one created in our laboratory in an earlier study [23], and different interaction modes have been proposed. Traditionally, the rigid model with the best performance was identified and then applied to predict hERG inhibitory potency. However, since hERG inhibitors are highly diverse in terms of structure and pharmacology,

and the hERG protein is a highly flexible homotetramer, the mechanism of the ligand–protein interaction is so complicated that single, rigid models cannot sufficiently elucidate it [24]. Therefore, in this study, an ensemble of 3D-QSAR pharmacophore models was constructed in order to summarize and combine the possible interaction modes more completely. The PHASE program in the Maestro package [25–27] was employed, which was shown by Evans et al. to give very robust performance [24]. This model ensemble included five different but representative 3D-QSAR pharmacophore models that gave the best performances and high predictive abilities. When applying this ensemble of models to predict hERG inhibitory activity—in order to incorporate more interaction modes and obtain more reliable results than just one rigid model—the average was calculated when combining the predicted values from the ensemble of models [28]. The results thus obtained suggest that the predictive ability of the combined 3D-QSAR pharmacophore model is better than those of the individual models. Moreover, a 2D-QSAR approach was also applied to further analyze the physicochemical characters of hERG inhibitors, and a heuristic machine-learning method was employed in this case using the CODESSA software package [29, 30]. The same training set and test set were used for all of the models, allowing us to compare the results from the different models. In this parallel evaluation, it was apparent that the combined 3D-QSAR pharmacophore model was better at classifying active compounds, while the 2D-QSAR model was better at classifying inactive compounds. Given the complementary characteristics of the two approaches, we then decided to combine these two models. The resulting final model—a combination of both the combined 3D-QSAR pharmacophore model and the 2D-QSAR model—outperforms any single model, exhibiting the best predictive ability, the lowest mean absolute error (MAE) value of 0.45, the highest R_{test}^2 of 0.75, and the highest classification accuracy of 83.33% for the test set. The combined model, with its clearly enhanced predictive ability, also demonstrated that combining model results through averaging is a very feasible and practicable approach. Meanwhile, the key features of hERG inhibitors that were highlighted by the models are consistent, correlating with the crucial interaction residues Y652 and F656 from the hERG protein.

The final model combining both the combined 3D-QSAR pharmacophore model and 2D-QSAR model is the most rational and presents the best predictive ability; it can not only predict hERG inhibitory potency accurately for early cardiac safety screening, but it provides an effective tool for avoiding hERG inhibitory liability and thus enhanced cardiac risk in the design and optimization of new chemical entities.

Materials and methods

Dataset preparation

In the dataset, compounds were collected from published biological studies, taking into account a diversity and broad range of biological activities. Only biological studies using whole-cell patch-clamp measurements and mammalian cell lines of *HEK* or *CHO* were considered, to maintain the variety and consistency of the data set [31]. One hundred thirteen diverse compounds were collected (see Table 1), with a broad range of hERG inhibitory activities (in vitro IC_{50}), spanning from 0.0009 to 4400 μM . The IC_{50} values were converted to pIC_{50} values using the formula $-\log(IC_{50})$.

The dataset was split into a training set and a test set. All of the compounds were classified into seven classes according to the order of magnitude of the IC_{50} value, and then 53 training compounds were selected on the basis of structural diversity and wide coverage of the activity. The remaining 60 compounds were used as the test set.

In the SYBYL package [95], the molecules were constructed, hydrogens were added, and the energy of each molecule was minimized to obtain the optimized structure. When calculating the optimized geometry, the MMFF94s force field and MMFF94 charges were applied, with a gradient of $0.001 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$. The chirality of some molecules such as levacetylmethadol was also examined carefully, based on the information on PubChem.

Development of the 3D-QSAR pharmacophore models

The common pharmacophore approach can be used to find the 3D structural characteristics of hERG inhibitors that are required for them to interact with the target. Considering that hERG inhibitors are highly diverse in terms of structure and pharmacology, and the hERG protein is a rather flexible homotetramer, there are likely to be various sites and modes of interaction, making the mechanism of ligand–protein interaction so complicated that a single, rigid pharmacophore model cannot elucidate it sufficiently. Accordingly, an ensemble of 3D-QSAR pharmacophore models was constructed in order to summarize the possible interaction modes more completely, and thus to achieve more reliable predictions than a single, rigid model can.

The PHASE program in the Maestro package (software for pharmacophore perception, structure alignment, activity prediction and 3D database creation and searches) was employed to develop the 3D-QSAR pharmacophore models. PHASE utilizes fine-grained conformational sampling and a range of scoring techniques to identify a common

pharmacophore hypothesis, which is then combined with known activity data to generate a 3D-QSAR model which identifies aspects of the molecular structure that govern activity. Using PHASE, a number of steps are performed (*Prepare Ligands, Create Sites, Find Common Pharmacophores, Score Hypotheses* and *Build QSAR Model*), during which many parameters can be adjusted, allowing flexibility and versatility.

The dataset was first imported into the PHASE program, and then conformational ensembles were generated. During conformational sampling, the MacroModel torsional sampling method of ConfGen was used, with the MMFFs force field and the distance-dependent dielectric used as the solvation treatment. In order to eliminate high-energy or redundant conformers, the maximum relative energy difference was set to $10.0 \text{ kcal mol}^{-1}$, and the cutoff root mean square deviation (RMSD) was 1 \AA . In addition, the maximum number of conformers for each structure was set to 1000, while the number of conformers for each rotatable bond was set to 100. In the training set, compounds with $pIC_{50} \geq 0$ were designated the most active compounds and those with $pIC_{50} \leq -2$ were designated inactive compounds by changing the *Pharm Set* parameter.

During pharmacophore site creation, all available pharmacophore features were included in the conformational set for each compound: H-bond donor (labeled D), H-bond acceptor (labeled A), hydrophobic group (labeled H), aromatic ring (labeled R), and positively (labeled P) and negatively (labeled as N) charged groups. For hydrophobic group (H) and aromatic ring (R) features, the definition term of *default_aromatic_surface* was also included by clearing the *Ignore* check box; other feature definitions were set to their defaults.

Common pharmacophores were then identified based on the most active compounds by grouping similar pharmacophores according to their intersite distances using a tree-based partitioning technique. Here, the minimum intersite distance was set to 2 \AA , and the maximum tree depth was 5. Initially, the number of matching active compounds (n_{act}) was equal to the total number of most active compounds ($n_{\text{act_tot}}$) in the training set, but because of the high diversity of hERG inhibitors, and even with a minimum number of pharmacophore sites (n_{sites}) of 3, there was no common pharmacophore that could match all of the most active compounds. Therefore, n_{sites} was set to 4, and n_{act} was reset by decreasing $n_{\text{act_tot}}$ one by one until hypotheses were found and scored successfully. Finally, n_{act} was set to 12.

Afterwards, the generated pharmacophore hypotheses were initially scored based on the most active compounds in order to identify the pharmacophores that yielded the best alignment with the actives from each surviving n -

Table 1 The dataset of hERG inhibitors collected from the literature

Number ^a	Molecule name	<i>p</i> IC ₅₀	Reference	Number ^a	Molecule name	<i>p</i> IC ₅₀	Reference
1	Astemizole	3.04	[32]	58*	Cibenzoline	-0.57	[67]
2*	Desmethylastemizole	3.00	[32]	59*	Granisetron	-0.57	[68]
3	Clemastine	2.92	[33]	60*	Sibutramine	-0.58	[59]
4	Pimozide	2.82	[34]	61	Loratadine	-0.59	[34]
5	Chlorobutanol	2.36	[35]	62	Flecainide	-0.59	[69]
6*	Flunarizine	2.24	[36]	63*	Propranolol	-0.59	[70]
7	Terfenadine	2.15	[34]	64*	Citalopram	-0.60	[71]
8	Bepriidil	1.64	[34]	65	Cocaine	-0.64	[58]
9	Haloperidol	1.60	[34]	66	Pentamidine	-0.71	[72]
10	Cisapride	1.59	[34]	67*	Maprotiline	-0.72	[73]
11*	Norastemizole	1.56	[32]	68*	Metoclopramide	-0.73	[74]
12*	Droperidol	1.49	[37]	69*	Quetiapine	-0.76	[44]
13	Lidoflazine	1.43	[38]	70*	Dolasetron	-0.78	[68]
14	Halofantrine	1.40	[39]	71	Pyrilamine	-0.78	[38]
15	Dronedarone	1.23	[40]	72*	Desloratidine	-0.80	[43]
16	Amiodarone	1.15	[40]	73*	Doxepin	-0.81	[75]
17*	<i>N</i> -desbutylhalofantrine	1.14	[41]	74	Lovastatin	-0.85	[38]
18*	Pergolide	0.92	[42]	75*	Disopyramide	-0.86	[76]
19	Ketanserin	0.89	[34]	76*	Diltiazem	-0.87	[59]
20*	Domperidone	0.80	[43]	77	Buprenorphine	-0.88	[62]
21*	Risperidone	0.78	[44]	78*	Perhexiline	-0.89	[77]
22*	Clomiphene	0.74	[45]	79	Methadone	-0.99	[62]
23	Amsacrine	0.69	[46]	80*	Amitriptyline	-1.00	[60]
24	Sertindole	0.68	[38]	81*	Levobupivacaine	-1.01	[78]
25	Ziprasidone	0.62	[47]	82	Digitoxin	-1.05	[79]
26*	Doxazosin	0.49	[48]	83	Chlorpheniramine	-1.11	[38]
27	Thioridazine	0.41	[38]	84*	Mianserin	-1.17	[80]
28	Verapamil	0.35	[34]	85	Terazosin	-1.25	[61]
29*	Cyamemazine	0.33	[49]	86*	Bupivacaine	-1.26	[78]
30*	Mesoridazine	0.26	[50]	87*	Sparfloxacin	-1.26	[81]
31*	Azimilide	0.25	[51]	88*	Ropivacaine	-1.31	[78]
32	Prenylamine	0.23	[38]	89	Pilsicainide	-1.31	[82]
33*	Clebopride	0.21	[52]	90*	Spirolactone	-1.36	[83]
34	Fluoxetine	0.15	[34]	91*	Grepafloxacin	-1.44	[84]
35	Orphenadrine	0.07	[53]	92	Sildenafil	-1.52	[85]
36	Quinidine	0.05	[34]	93*	Roxithromycin	-1.56	[86]
37*	Brompheniramine	0.05	[54]	94*	Digoxin	-1.73	[79]
38*	Perphenazine	0.00	[55]	95	Moxifloxacin	-1.87	[87]
39*	Ajmaline	-0.02	[56]	96	Meperidine	-1.88	[62]
40	Vesnarinone	-0.04	[57]	97*	Ciprofloxacin	-2.00	[84]
41*	Cocaethylene	-0.08	[58]	98	Telithromycin	-2.00	[88]
42*	Imipramine	-0.08	[59]	99*	Canrenoic acid	-2.02	[83]
43*	Trifluoperazine	-0.15	[55]	100*	Procainamide	-2.14	[89]
44*	Chlorpromazine	-0.17	[60]	101*	Metoprolol	-2.16	[70]
45*	Prazosin	-0.2	[61]	102*	Clarithromycin	-2.23	[90]
46	Fentanyl	-0.26	[62]	103	Methylecgonidine	-2.23	[58]
47	Propafenone	-0.30	[38]	104*	Articaine	-2.35	[91]
48*	Olanzapine	-0.30	[59]	105	Lamotrigine	-2.36	[92]
49*	Levacetylmethadol	-0.34	[62]	106	Phenytoin	-2.38	[93]

Table 1 (continued)

Number ^a	Molecule name	pIC_{50}	Reference	Number ^a	Molecule name	pIC_{50}	Reference
50*	Norfluoxetine	-0.36	[63]	107	Codeine	-2.48	[62]
51*	Chloroquine	-0.40	[39]	108*	Oleandomycin	-2.53	[86]
52*	Clozapine	-0.40	[64]	109	Sotalol	-2.91	[34]
53*	Mefloquine	-0.42	[39]	110	Morphine	-3.00	[62]
54	Diphenhydramine	-0.42	[34]	111	Erythromycin	-3.15	[34]
55	Trazodone	-0.46	[65]	112	Phenobarbital	-3.48	[93]
56	Berberine	-0.49	[66]	113	4-Aminopyridine	-3.64	[94]
57*	Ambasilide	-0.56	[51]				

^a Compounds in the test set are labeled with *asterisks*

dimensional box. The hypotheses were then scored using the inactive compounds, to assign restrictive scores in order to distinguish inappropriate pharmacophores that could also align with inactives. When ranking the scores, the larger the difference between the scores of the actives and inactives, the better the hypothesis for discriminating them. All terms were included and the default weights were used in the score formula.

In the end, after the compounds had been aligned with the successfully generated and scored hypotheses (model and non-model ligand alignment), all of the hypotheses were submitted to the final stage of 3D-QSAR model construction. During the stage of 3D-QSAR model generation, the partial least squares (PLS) method was used to build the correlation between hERG inhibitory activity and grid locations that divide the space occupied by the molecules and the aligned pharmacophore hypothesis into uniformly sized 3D cubes. Here, 53 training compounds and 60 test compounds were defined by changing the *QSAR Set* parameter, and then a series of atom based 3D-QSAR pharmacophore models were generated with a random seed of 0, a grid spacing of 1 Å, and PLS factors of 1–4.

Development of the 2D-QSAR model

The 2D-QSAR approach was employed to further analyze the physicochemical characters of hERG inhibitors. The same training set as used for the 3D-QSAR pharmacophore models was employed to develop a regression model that could quantitatively correlate the target inhibitory activity with features of hERG inhibitors by learning from the calculated molecular descriptors.

The CODESSA software package was used for this task, as over 450 molecular descriptors could be calculated using this software, including constitutional descriptors, topological descriptors, geometric descriptors, electrostatic descriptors, quantum-chemical descriptors, and thermodynamic descriptors. The optimized structures in the dataset were

additionally calculated by the MOPAC package to provide quantum-mechanical data for computing quantum-chemical descriptors and thermodynamic descriptors, and the AM1 semiempirical parameter was used in the calculation.

After loading the dataset into the CODESSA program, all available descriptors were calculated. A heuristic statistical method (HM) was then applied, which involves the stepwise selection of the best multiple linear regression model with the most significant molecular descriptors. The advantage of using a heuristic method in CODESSA is that it is then possible to automatically search and find the most significant but lowly inter-correlated descriptors for the best regression model at high computational speed. In the calculation, the descriptors were preselected in advance. Descriptors that have unavailable or constant values were discarded to ensure that the values of each descriptor are useable and variable. After that, one-parameter correlations for each descriptor were computed, and descriptors that were less significant than the preset statistical criteria were eliminated. Next, two-parameter correlations were computed for each pair of descriptors, and descriptors that showed intercorrelation that was above a significant intercorrelation level (r_{sig}) were discarded in order to avoid overfitting. Here, the value of r_{sig} was set to 0.7, and the other parameters remained at their default settings. Finally, among the top ten correlations with the highest F values, the correlation with the highest R^2 value (correlation coefficient) was considered to be the optimal one.

Generally, the statistical parameters R^2 , R_{cv}^2 and F are used to evaluate a 2D-QSAR model; these represent the correlation coefficient of the regression, the average of the leave-one-out cross-validated correlation coefficient, and the Fisher F criterion value, respectively. When the maximum number of descriptors (ND_{max}) in the regression equation increases, the statistical parameters will improve concomitantly until overfitting occurs, so the predictive reliability decreases [96]. It has been suggested that the ratio between ND_{max} and the number of training com-

pounds is about 1:5. In addition, among models with good performances, the simplest model with the least descriptors will be the best choice. Therefore, a series of models were built in advance by varying ND_{\max} by one each time (from 3 to 10). Upon increasing one descriptor, the optimum number was achieved when the value of R^2 did not improve significantly or the value of R_{cv}^2 decreased; this corresponded to the best 2D-QSAR model.

Results and discussion

3D-QSAR pharmacophore models

A total of 47 3D-QSAR pharmacophore models were successfully constructed using the PHASE program. In PHASE, the performances of the 3D-QSAR pharmacophore models were evaluated using a series of statistical parameters, as summarized in Table 2 (the specific formulae used are shown in the “[Electronic supplementary material](#)”). The values of the SD, R^2 , F and P were used to evaluate the training set predictions, while the values of the RMSE, Q^2 and Pearson's R were used for the test set predictions. Generally, models with lower values of the SD and RMSE and higher values of R^2 , Q^2 and Pearson's R give the best performance. In addition, a high value of F indicates statistical significance for a model, while a low value of P indicates a high degree of confidence. The accuracy of the 3D-QSAR pharmacophore model improves as the number of PLS factors increases, until overfitting occurs. For example, if IC_{50} values are accurate to a multiplicative factor of 2, the corresponding $-\log [IC_{50}]$ values are only accurate to $\log(2)$. Accordingly, if the SD is smaller than this experimental uncertainty, then the data are clearly overfitted. Here, models containing three or more PLS factors tended to fit the pIC_{50} values beyond their experimental uncertainty, so only one- and two-factor models were considered.

The generated 3D-QSAR pharmacophore models were categorized into five clusters based on the combination and the spatial arrangement of pharmacophore features. In each cluster, the models were ranked according to the statistical parameters, and then the model with the best performance (lowest values of the SD, RMSE and P , highest values of R^2 , F , Q^2 and Pearson's R) was selected. Finally, the five most representative models were included in the 3D-QSAR pharmacophore model ensemble, which thus summarizes as many of the interaction modes as possible. The models and the corresponding statistical results are listed in Table 2.

The model ensemble exhibited good performance, with values of R^2 ranging from 0.9137 to 0.9410, SDs of between 0.3318 and 0.4077, and RMSEs of <1.0. The HHHP.67 model gave the highest R^2 value (0.9410), the highest F value (327.0) and the highest Pearson's R value (0.6525), while the HHR.967 model gave the lowest RMSE value (0.8131) and the highest Q^2 value (0.3701). The predictions for the training compounds were relatively precise, and the deviations between the experimental and predicted pIC_{50} values were within one log unit.

Three kinds of chemical features were included in the pharmacophore hypotheses: hydrophobic (H), aromatic ring (R) and positively charged (P) groups. The hydrophobic group (H) feature was present in all of the hypotheses, but its spatial arrangement varied somewhat. The pharmacophores are shown in Fig. 1, where they are aligned with highly active compounds. The HHHP.67 (see Fig. 1a) and HHHP.726 (see Fig. 1b) hypotheses incorporate the same chemical features, including one positively charged group (P) at the center that is flanked by three hydrophobic groups (H), but the spatial arrangements of these features are different for the two hypotheses. Compounds aligned with the HHHP.726 pharmacophore exhibit mildly twisted conformers, while in the HHHP.67 hypothesis, the distances between the three hydrophobic groups (H) are smaller, and this hypothesis is usually aligned to a more curved and

Table 2 The statistical results obtained when evaluating the 3D-QSAR pharmacophore model ensemble

Pharmacophore features	SD	R^2	F	P	RMSE	Q^2	Pearson R
HHHP.67	0.3318	0.9410	327.0	6.325E-26	0.8670	0.2836	0.6525
HHHP.726	0.4077	0.9137	185.4	2.381E-19	0.9423	0.2966	0.5718
HHR.281	0.3637	0.9343	234.8	3.056E-20	0.9185	0.2801	0.5338
HHR.967	0.3404	0.9392	309.2	4.692E-25	0.8131	0.3701	0.6404
HHHP.2085	0.3990	0.9180	218.4	6.544E-22	0.8567	0.2832	0.5838

SD standard deviation of the regression, R^2 square of the correlation coefficient of the regression, F ratio of the model's variance to the variance in observed activity for the regression, P significance level of F , RMSE root mean square error in the test set predictions, Q^2 square of the correlation coefficient between the predicted and observed activities in the test set predictions, Pearson R Pearson's R value for the correlation between the predicted and observed activities in the test set predictions

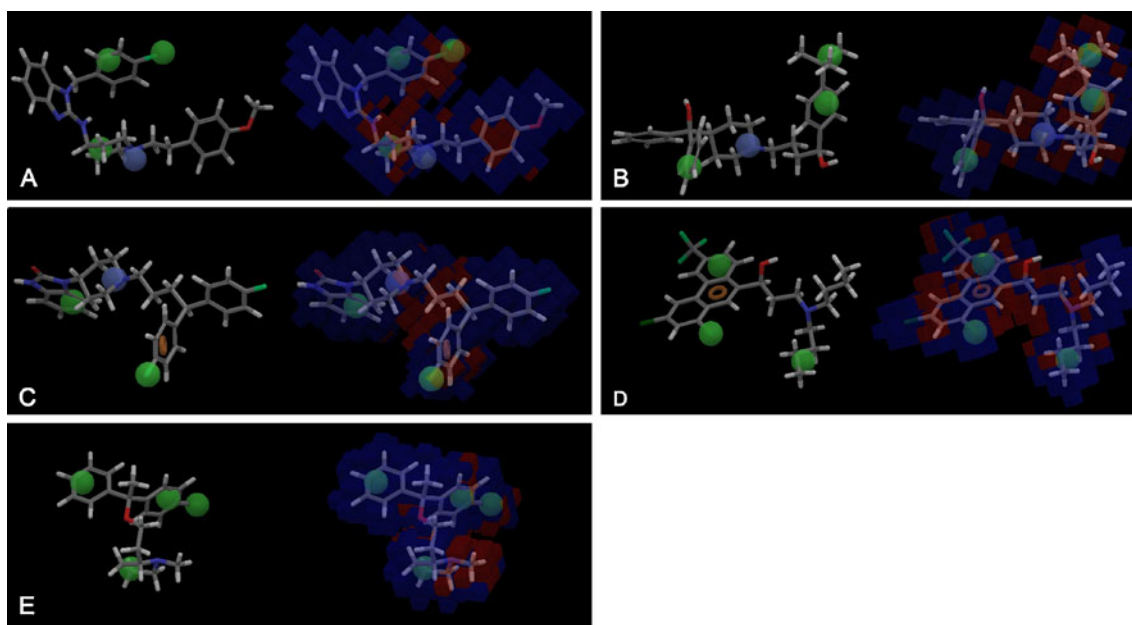


Fig. 1 a–e Pharmacophore hypotheses aligned with highly active inhibitors. Pharmacophore features are denoted by color: *green* hydrophobic group (H), *blue* positively charged group (P), *orange* aromatic ring (R). Corresponding 3D-QSAR pharmacophore model plots (coefficient threshold=0) that depict the hydrophobic/nonpolar (H) effect are also displayed on the *right*; *blue cubes* indicate positive

correlation, while *red cubes* show negative correlation. **a** HHHP.67 hypothesis aligned with astemizole ($pIC_{50}=3.04$). **b** HHHP.726 hypothesis aligned with terfenadine ($pIC_{50}=2.15$). **c** HHPR.281 hypothesis aligned with pimozone ($pIC_{50}=2.82$). **d** HHHR.967 hypothesis aligned with halofantrine ($pIC_{50}=1.40$). **e** HHHH.2085 hypothesis aligned with clemastine ($pIC_{50}=2.92$)

twisted conformer or just an area of the compound. The spatial arrangement of the chemical features in the HHPR.281 hypothesis (see Fig. 1c) is similar to that for the HHHP.726 hypothesis, but one hydrophobic group feature is replaced with an aromatic ring feature. The HHHH.2085 (see Fig. 1e) and HHHR.967 (see Fig. 1d) hypotheses lack a positively charged group (P) feature and contain hydrophobic groups (H) only or with an aromatic ring (R). This is mainly because some of the uncharged hERG inhibitors lack a basic nitrogen center, suggesting that the basic nitrogen center is contributing but not crucial to hERG inhibition. As a whole, when aligned with the hypotheses, compounds with high inhibitory activity such as astemizole and terfenadine commonly exhibit more or less curved conformers. The two adjacent hydrophobic groups (H) (a hydrophobic group and an aromatic ring in the HHHR.967 hypothesis) are rather close—2.686 Å to 3.654 Å apart—most of which is occupied by the two connected components. For example, in the alignment of terfenadine with the HHHP.726 hypothesis, the benzene ring segment and the *tert*-butyl segment are aligned with the hydrophobic group (H) features.

For the inhibitors, the hydrophobic group (H) is usually occupied by an aliphatic carbon chain, an aromatic ring or a halogen atom (mostly a Cl atom), which interacts with the hERG protein through a hydrophobic effect and/or a π stacking interaction. The positively charged group (P)

mostly corresponds to a tertiary amine group, and sometimes a secondary amine group, and these interact with the hERG protein by a cation– π interaction, improving the inhibitory potency. The pharmacophore features of the models are similar, and correlate with the crucial interaction residues Y652 and F656 from the hERG protein, although the spatial arrangements are different for different models. According to the literature, the hERG K^+ channel shows distinctive dynamic behavior of slow activation and deactivation but rapid inactivation, so it exhibits a highly flexible conformation. In addition, the hERG protein consists of four identical subunits with axial and rotational symmetry, but the inhibitors do not necessarily interact with the aromatic residues from each of the subunits equally [97], providing a clue that the ligand–protein interactions are flexible and asymmetric. Like the hypotheses, the components of different inhibitors may interact with adjacent residues that are separated by only a small distance, or with residues that are much further apart. The hypotheses further prove that hydrophobic group (H) and aromatic ring (R) features are the molecular characters that are crucial to hERG inhibitors, and that the hERG protein and inhibitors can adopt flexible modes for interaction, so the flexibility of the inhibitors is advantageous to inhibition.

The 3D-QSAR pharmacophore models can also be visualized as 3D cube plots. Such plots showing the effect of hydrophobic/nonpolar (H) features are displayed in

Fig. 1, in which the blue-colored cubes show positive correlation (indicating an increase in inhibitory activity), while the red cubes show negative correlation (indicating a decrease in inhibitory activity). This allows us to determine which parts of a compound make positive or negative contributions to its inhibitory potency.

Evaluation of the 3D-QSAR pharmacophore models

In order to further evaluate the quality and predictive ability of the 3D-QSAR pharmacophore model ensemble, the test set containing 60 compounds was applied. However, several of the compounds in the test set did not align with the corresponding pharmacophore hypothesis (there was no match to at least three pharmacophore features, and the fitness value was poor). These compounds were therefore considered inactive compounds (designated $pIC_{50} \leq -2$ previously), and in the subsequent statistical calculation, the predicted pIC_{50} values were set to -2 . For the test compounds, the predicted pIC_{50} values from each model and the experimental values are listed in Table 3. In addition, the mean absolute error (MAE) for each prediction set was calculated. The formula is

$$MAE = \frac{1}{k} \sum_{j=1}^k |\hat{y}_j - y_j|, \quad (1)$$

where k refers to the number of molecules in the test set, y_j refers to the experimental activity for test molecule j , and \hat{y}_j is the predicted activity for test molecule j . R_{test}^2 , representing the square of the correlation coefficient between the experimental pIC_{50} and the predicted pIC_{50} in the test compound predictions, was also calculated for each model.

MAE values were found to lie between 0.64 and 0.75, but the values of R_{test}^2 were low (0.28–0.48), indicating that it is difficult to obtain satisfactory predicted results with the single, rigid model, and the prediction of structurally diverse hERG inhibitors appeared to be a rather complicated task.

To get more reliable predicted results and reduce the uncertainty associated with single, rigid model prediction, a compromise method involving averaging was used to generate a combined 3D-QSAR pharmacophore model. For each compound in the test set, the average of the pIC_{50} values predicted by all individual models was calculated. The same weight was used for all models, as it was not clear which model is most accurate for an unknown compound. The formula for calculating the average is

$$\text{average} = \frac{1}{k} \sum_{j=1}^k \hat{y}_j, \quad (2)$$

where k refers to the number of models and \hat{y}_j refers to the predicted value from model j .

Finally, the combined 3D-QSAR pharmacophore model was found to yield the lowest MAE value of 0.53 and the highest R_{test}^2 value of 0.64 (see Table 3), suggesting that the predictive ability of this combination of 3D-QSAR pharmacophore models is significantly better than those of the individual models.

According to previously published studies, the classic hERG pharmacophore model basically contains a basic nitrogen center flanked by hydrophobic or aromatic groups connected by flexible linkers, as generated using the HypoGen module in the Catalyst program. Meantime, Kramer et al. developed a composite model to predict hERG inhibitory activity based on different QSAR models identified by preliminary pharmacophore scanning [14]. Also, Aronov developed a pharmacophore model in the MOE software package for neutral hERG inhibitors that contain hydrophobic/aromatic features and hydrogen bond acceptors but lack the basic nitrogen center [98]. In the model developed by Garg et al., an aromatic group, a hydrophobic group and a hydrogen bond acceptor group were included, based on training molecules with low IC_{50} values (less than $10 \mu\text{M}$) [99]. The pharmacophore model published recently by Durdagi et al. contains aromatic group, hydrogen bond acceptor and hydrogen bond donor features, and was produced with the PHASE program. [100]. However, generally, the rigid pharmacophore model with the best performance is applied in order to predict hERG inhibitory potency.

In our study, we employed the PHASE program, in which structure alignment is performed based on the generated common pharmacophore, and activity prediction incorporates the grid technique of 3D-QSAR. In the pharmacophores, the hydrophobic and aromatic features were still found to be predominant, and while it seems that the basic nitrogen center is not a crucial feature, it may enhance the potency of hERG inhibitors. Moreover, due to the high diversity of hERG inhibitors and the flexibility and symmetry of the hERG protein, they are likely to adopt distributed sites and flexible modes for interaction. Thus, the mechanism of ligand–protein interaction is so complicated that a single, rigid model cannot sufficiently elucidate it. Therefore, the combination of the representative 3D-QSAR pharmacophore models was calculated, as this incorporated as many interaction modes as possible, thus giving more reliable predicted results. This model is able to consider the particular interactions between the hERG K^+ channel and its inhibitors in more depth, and resulting improved predictive ability of the combined 3D-QSAR pharmacophore model indicates that calculating the average result from a combination of models is a very feasible and practicable approach.

Table 3 The predicted pIC_{50} values of all the models versus the experimental pIC_{50} values for the test set

Molecule name	Exp. pIC_{50}	Pred. pIC_{50} of 3D-QSAR pharmacophore model ensemble						Pred. pIC_{50} of combined 3D-QSAR and 2D-QSAR model	Combination of combined 3D-QSAR pharmacophore model and 2D-QSAR model
		HHHP.67	HHHP.726	HHPR.281	HHHR.967	HHHH.2085	Combined model		
Desmethylastemizole	3.00	3.00	2.12	1.57	2.69	2.77	2.43	1.86	2.14
Flunarizine	2.24	0.34	0.11	1.12	0.17	0.24	0.40	1.54	0.97
Norastemizole	1.56	1.39	-0.41	0.15	0.59	0.82	0.51	0.41	0.46
Droperidol	1.49	0.95	1.57	0.66	0.72	-0.10	0.76	0.57	0.66
N-desbutylhalofantrine	1.14	0.67	0.25	-0.20	0.10	0.75	0.31	2.20	1.26
Pergolide	0.92	0.13	0.56	-0.15	-0.35	0.29	0.10	-0.84	-0.37
Domperidone	0.80	0.60	1.48	0.64	0.85	-0.15	0.68	0.63	0.65
Risperidone	0.78	0.51	0.62	0.34	-0.03	0.01	0.29	0.91	0.60
Clomiphene	0.75	0.21	-0.17	-1.23	-0.44	0.33	-0.26	2.38	1.06
Doxazosin	0.49	-1.06	0.88	-0.86	0.31	-1.07	-0.36	0.69	0.17
Cyamemazine	0.33	0.53	≤ -2	≤ -2	0.01	0.21	-0.65	-0.36	-0.51
Mesoridazine	0.26	0.53	-0.15	-0.17	-0.13	0.60	0.14	0.15	0.14
Azimilide	0.25	0.92	0.30	0.13	0.56	0.39	0.46	0.49	0.48
Clebopride	0.21	0.47	0.40	0.80	1.10	-0.57	0.44	0.97	0.70
Brompheniramine	0.05	-0.22	-0.03	0.17	0.01	0.05	0.00	-0.83	-0.42
Perphenazine	0.00	0.21	-0.60	0.16	-0.41	-0.59	-0.25	1.66	0.71
Ajmaline	-0.02	≤ -2	-0.34	-0.42	-0.31	≤ -2	-1.01	-1.24	-1.13
Cocaethylene	-0.08	-0.42	≤ -2	-0.18	-0.01	-0.17	-0.56	-0.21	-0.38
Imipramine	-0.08	-0.07	-0.83	-1.10	-0.46	0.07	-0.48	-0.37	-0.42
Trifluoperazine	-0.15	0.15	-0.47	0.35	-0.08	-0.61	-0.13	0.87	0.37
Chlorpromazine	-0.17	-0.01	-0.01	0.42	-0.67	-0.19	-0.09	1.27	0.59
Prazosin	-0.20	-1.04	-0.20	-0.76	-0.79	-0.88	-0.73	-1.75	-1.24
Olanzapine	-0.30	0.58	-0.69	-0.31	0.34	0.62	0.11	0.09	0.10
Levacetylmethadol	-0.34	0.08	≤ -2	-1.03	-0.62	0.71	-0.57	-0.84	-0.71
Norfluoxetine	-0.36	0.09	-0.88	-0.46	-0.54	0.28	-0.30	0.34	0.02
Chloroquine	-0.40	-0.02	0.25	0.11	-0.11	-0.31	-0.02	-0.46	-0.24
Clozapine	-0.40	-0.28	-0.47	-0.22	-0.15	-0.40	-0.30	0.60	0.15
Mefloquine	-0.42	-0.16	-0.71	-0.87	-0.02	-0.11	-0.37	-0.79	-0.58
Ambasilide	-0.56	0.27	0.90	-0.10	-0.22	0.57	0.28	-0.50	-0.11
Cibenzoline	-0.57	-1.15	≤ -2	≤ -2	-0.38	-0.35	-1.18	-1.06	-1.12
Granisetron	-0.57	-0.60	-0.24	-0.83	0.08	-0.50	-0.42	-0.91	-0.66
Sibutramine	-0.58	-0.14	-0.46	-0.15	0.04	0.26	-0.09	-1.10	-0.59
Propranolol	-0.59	-0.37	-0.65	-0.42	≤ -2	≤ -2	-1.09	-1.41	-1.25
Citalopram	-0.60	0.71	-0.36	0.05	-0.14	0.48	0.15	0.17	0.16
Maprotiline	-0.72	-0.24	-0.65	-1.16	-0.53	-0.25	-0.57	-0.47	-0.52
Metoclopramide	-0.73	-0.31	-0.29	0.36	-0.10	-0.83	-0.23	-1.64	-0.94
Quetiapine	-0.76	0.36	≤ -2	0.05	0.02	0.19	-0.28	0.27	0.00
Dolasetron	-0.78	-0.56	-0.57	-0.54	0.74	-0.60	-0.31	-0.62	-0.46
Desloratidine	-0.80	-0.30	-0.19	-0.32	-0.59	-0.26	-0.33	-1.47	-0.90
Doxepin	-0.81	-0.10	≤ -2	≤ -2	-0.5	0.38	-0.84	0.11	-0.37
Disopyramide	-0.86	0.11	≤ -2	≤ -2	-0.18	-0.38	-0.89	-1.48	-1.18
Diltiazem	-0.87	0.04	-0.56	≤ -2	-0.60	-0.11	-0.65	0.19	-0.23
Perhexiline	-0.89	-0.73	≤ -2	≤ -2	0.23	0.13	-0.87	-2.02	-1.45
Amitriptyline	-1.00	0.04	-0.24	-0.78	≤ -2	-0.56	-0.71	-0.88	-0.80
Levobupivacaine	-1.01	0.37	-0.2	-0.37	0.19	-0.42	-0.09	-1.78	-0.93

Table 3 (continued)

Molecule name	Exp. pIC_{50}	Pred. pIC_{50} of 3D-QSAR pharmacophore model ensemble						Pred. pIC_{50} of 2D-QSAR model	Combination of combined 3D-QSAR pharmacophore model and 2D-QSAR model
		HHHP.67	HHHP.726	HHPR.281	HHHR.967	HHHH.2085	Combined model		
Mianserin	-1.17	0.70	≤ -2	≤ -2	-0.93	-0.30	-0.91	-0.60	-0.75
Bupivacaine	-1.26	-0.29	-0.62	0.16	0.01	-0.32	-0.21	-1.81	-1.01
Sparfloxacin	-1.26	-0.10	-1.06	-1.52	-0.44	-0.49	-0.72	-2.10	-1.41
Ropivacaine	-1.31	0.25	-1.01	≤ -2	-0.18	-0.67	-0.72	-1.93	-1.33
Spironolactone	-1.36	-0.36	≤ -2	≤ -2	≤ -2	-0.07	-1.29	-2.00	-1.64
Grepafloxacin	-1.44	-0.36	-1.46	-0.64	-0.07	0.38	-0.43	-1.36	-0.90
Roxithromycin	-1.56	-0.57	0.74	≤ -2	-0.78	0.32	-0.46	-3.03	-1.74
Digoxin	-1.73	-0.76	-1.10	≤ -2	-1.47	-1.20	-1.31	-2.23	-1.77
Ciprofloxacin	-2.00	≤ -2	≤ -2	-0.93	-0.49	≤ -2	-1.48	-1.22	-1.35
Canrenoic acid	-2.02	≤ -2	≤ -2	≤ -2	≤ -2	≤ -2	-2.00	-2.72	-2.36
Procainamide	-2.14	-0.59	-0.18	≤ -2	-0.79	-1.08	-0.93	-2.90	-1.92
Metoprolol	-2.16	-0.17	-0.41	-0.15	≤ -2	≤ -2	-0.95	-2.48	-1.71
Clarithromycin	-2.23	≤ -2	≤ -2	≤ -2	≤ -2	≤ -2	-2.00	-2.74	-2.37
Articaine	-2.35	-1.03	-0.39	-0.53	-0.57	-1.00	-0.70	-2.70	-1.70
Oleandomycin	-2.53	≤ -2	0.03	-0.06	≤ -2	≤ -2	-1.21	-2.29	-1.75
MAE		0.68	0.75	0.75	0.64	0.69	0.53	0.69	0.45
R_{test}^2		0.46	0.28	0.34	0.48	0.41	0.64	0.64	0.75

The 2D-QSAR model

In the CODESSA program, a total of 293 molecular descriptors were successfully calculated for the dataset. The same training set as used in the 3D-QSAR pharmacophore model was employed here; a heuristic statistical method was employed to develop the regression model that presents the correlation between hERG inhibitory activity and the most significant molecular descriptors.

Generally, in a predictive model, the conditions $R^2 > 0.6$ and $R_{cv}^2 > 0.5$ for the training set and $R_{test}^2 > 0.5$ for the test set should be satisfied, while a high F test value indicates that the model is statistically significant. When determining the optimum number of descriptors to include in the model, a series of models with from 3 to 10 descriptors were built, and the influence of the number of descriptors on the values of R^2 and R_{cv}^2 is shown in Fig. 2. Seven descriptors appeared to be optimal for the model, including two constitutional descriptors ($X1$ and $X2$), one topological descriptor ($X4$), three electrostatic descriptors ($X3$, $X5$ and $X7$), and one quantum-chemical descriptor ($X6$). The corresponding

multiple linear regression equation and the values of the descriptors included in it are presented in Table 4. This model exhibits good performance (as summarized in Table 5), with satisfactory statistical parameters: $R^2 = 0.9121$, $R_{cv}^2 = 0.8724$, and $F = 66.67$.

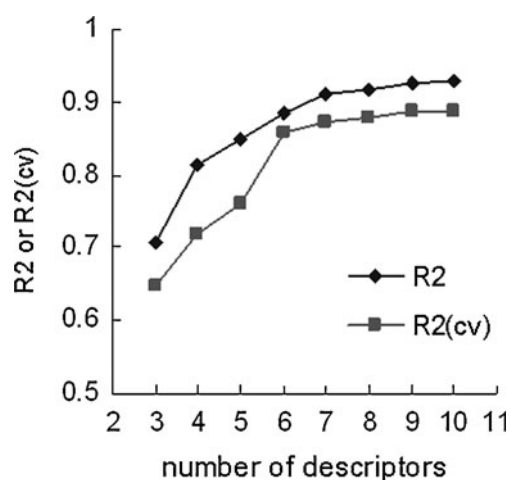


Fig. 2 Number of descriptors versus R^2 or R_{cv}^2 for the 2D-QSAR model

Table 4 Values of the descriptors in the regression equation for the 2D-QSAR model ($pIC_{50} = 265.53 + 0.16675X1 + 28.104X2 - 22.864X3 - 1.6285X4 - 173.75X5 - 67.109X6 - 55.214X7$)

Descriptor	Coefficient	<i>t</i> -test
Intercept	265.53	4.5635
Number of aromatic bonds (<i>X1</i>)	0.16675	9.0751
Relative number of Cl atoms (<i>X2</i>)	28.104	8.9682
Minimum partial charge (Q_{\min}) (<i>X3</i>)	-22.864	-6.6925
Balaban index (<i>X4</i>)	-1.6285	-8.2991
FPSA-3 (fractional PPSA; PPSA-3/TMSA; Zefirov's PC) (<i>X5</i>)	-173.75	-5.5304
Maximum valency of a C atom (<i>X6</i>)	-67.109	-4.5558
HACA-1/TMSA (Zefirov's PC) (<i>X7</i>)	-55.214	-3.6681

Four classes of descriptors are included in the 2D-QSAR model. First of all, constitutional descriptors of the aromatic ring (*X1*) and Cl atom (*X2*) features are the most representative; these have the highest *t*-test values. These features are all positively correlated with the value of pIC_{50} , meaning that higher values will lead to higher inhibitory potency. Topological descriptors describe the atomic connectivity of the molecule, and the Balaban index descriptor—which describes the rigidity (molecules with more or longer flexible carbon chains have lower Balaban index values)—is included among these. This is negatively correlated with the value of pIC_{50} , indicating that higher flexibility is advantageous to inhibitory potency. These features are consistent with those in the 3D-QSAR pharmacophore models, and further confirm that the hydrophobic group, the aromatic ring and flexibility are the crucial characteristics of hERG inhibitors.

Another class of attribute is represented by the electrostatic descriptors, which reflect the charge distribution characteristics of the molecules (responsible for polar interactions), and they are all negatively correlated with the value of pIC_{50} . Here, the minimum partial charge descriptor has a high *t*-test value. The charged partial surface area (CSPA) descriptor of FPSA-3 represents the proportion of the surface area of the whole molecule that is positively charged, and the HACA descriptor represents hydrogen acceptor features. Finally, there is a quantum-chemical descriptor of the maximum valency of a C atom (which is negatively correlated with pIC_{50}). This valency-related descriptor represents the strength of intramolecular

bonding interactions, the stability and flexibility of the molecules, and other valency-related properties.

The same test set was also applied to evaluate the 2D-QSAR model, and the results were compared with those for the 3D-QSAR pharmacophore models. The predicted values are listed in Table 3. On the whole, the MAE value from the 2D-QSAR model is higher (MAE=0.69) than that from the combined 3D-QSAR pharmacophore model (MAE=0.53), but the models give the same R_{test}^2 , 0.64. Although several compounds with moderate activities were not predicted as well by the 2D-QSAR model as the combined 3D-QSAR pharmacophore model, most of the compounds were predicted accurately, especially the inactive compounds.

This different 2D-QSAR approach has the same key inhibitory features as the 3D-QSAR pharmacophore models, and shows robust performance and predictive ability, making it an effective tool for predicting hERG inhibitory potency, and one that is complementary to the combined 3D-QSAR pharmacophore model.

Combination of the combined 3D-QSAR pharmacophore model and the 2D-QSAR model

When analyzing and comparing the test set predictions from the combined 3D-QSAR pharmacophore model and the 2D-QSAR model, an additional statistic—predictive accuracy—was evaluated, which represented the percentage of correctly classified compounds. Using a threshold of $pIC_{50} = -1$, we determined the ability of each model to classify the compounds into two classes: active compounds (with $pIC_{50} > -1$) and inactive compounds (with $pIC_{50} \leq -1$). The results are summarized in Table 6. The formulae for predictive accuracy are

Table 5 Statistical parameters of the 2D-QSAR model

Number of descriptors	R^2	<i>F</i>	R_{cv}^2	S^2	RMSE
7	0.9121	66.67	0.8724	0.2827	0.4899

$$\text{Total accuracy} = \frac{n_{TP} + n_{TN}}{n_P + n_N} \quad (3)$$

Table 6 The predictive accuracies of the combined 3D-QSAR pharmacophore model, the 2D-QSAR model, and the combination of these two models

Thresholds	Predictive accuracy	Combined 3D-QSAR pharmacophore model	2D-QSAR model	Combination of the two models
$pIC_{50} = -1$	Total accuracy	76.67%	81.67%	83.33%
	Accuracy for actives ($pIC_{50} > -1$)	93.02%	79.07%	86.07%
	Accuracy for inactives ($pIC_{50} \leq -1$)	35.29%	88.24%	76.47%

$$\text{Accuracy for actives} = \frac{n_{TP}}{n_P} \quad (4)$$

$$\text{Accuracy for inactives} = \frac{n_{TN}}{n_N}, \quad (5)$$

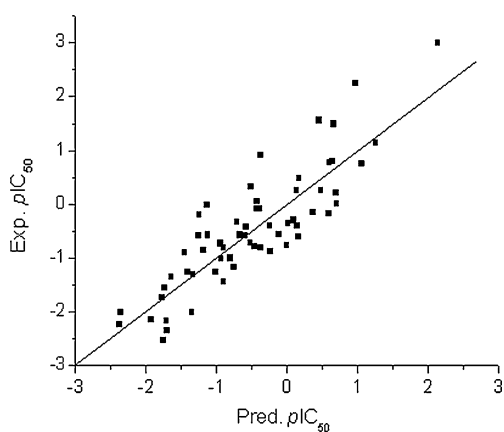
where n_P refers to the number of active compounds with experimental $pIC_{50} > -1$, n_N refers to the number of inactive compounds with experimental $pIC_{50} \leq -1$, while n_{TP} refers to the number of true positive predictions, and n_{TN} refers to the number of true negative predictions.

Both of the models exhibited good classification abilities; however, the combined 3D-QSAR pharmacophore model was better at classifying active compounds, while the 2D-QSAR model was better at classifying inactive compounds. Due to the complementary nature of the two approaches, we decided to combine these two models by averaging the predicted results from them, in order to enhance the reliability of predictions of hERG inhibitory activity. We found that this final combined model exhibited the highest and most robust predictive ability of any of the models we considered in our study (see

Table 3), with the lowest MAE value of 0.45 and the highest R_{test}^2 of 0.75 (as shown in Fig. 3). Active compounds and inactive compounds were largely distinguished accurately by this model, with a predictive accuracy of 83.33% (see Table 6), thus indicating high predictive ability during safety screening.

Conclusions

In this study, we have developed robust models for the prediction of hERG channel inhibitory potency, based on the application of two different and complementary ligand-based approaches: 3D-QSAR pharmacophore and 2D-QSAR. All models exhibited good performance, with the values of R^2 ranging from 0.9121 to 0.9410 for the training set. The final model, which combined both the combined 3D-QSAR pharmacophore model and the 2D-QSAR model, was the most rational and predictive, as it had the lowest MAE value of 0.45, the highest R_{test}^2 value of 0.75, and the highest classification accuracy of 83.33% for the test set. On the other hand, the key inhibitory features were shared by all of the models. The hydrophobic feature, which is usually an aliphatic carbon chain, an aromatic ring or a halogen atom, is the most significant and crucial character, further confirming that the aromatic residues F652 and Y656 within the pore play a crucial role in the hERG inhibitory interaction. The other crucial feature is the flexibility of the inhibitors, as this allows them to adopt compatible modes for interaction with such a flexible and symmetrical protein. The final model combining both the combined 3D-QSAR pharmacophore model and the 2D-QSAR model is able to not only predict hERG inhibitory potency accurately for early cardiac safety screening, but it can also provide further insight into the particular interaction modes that occur between the hERG protein and its inhibitors. Meanwhile, based on our findings, we also expect this final model to provide an effective tool for avoiding hERG inhibitory liability and thus enhanced cardiac risk in the design and optimization of new chemical entities.

**Fig. 3** Experimental pIC_{50} values for the test set versus predicted pIC_{50} values obtained with the combination of the combined 3D-QSAR pharmacophore model and the 2D-QSAR model

Acknowledgments The authors gratefully acknowledge financial support from National Major Science and Technology Project of China (grant no. 2009ZX09501-003), the National Natural Science Foundation of China (grant no. 21072231) and the Qing Lan Project of Jiangsu Province. We also thank Semichem Company for providing the complete CODESSA software for our evaluation.

References

- Wysowski DK, Corken A, Gallo-Torres H, Talarico L, Rodriguez EM (2001) *Am J Gastroenterol* 96:1698–1703
- Woosley RL, Chen Y, Freiman JP, Gillis RA (1993) *JAMA* 269:1532–1536
- Killeen MJ (2009) *Drug Discov Today* 14:589–597
- ICH (2005) ICH Official web site. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Safety/S7B/Step4/S7B_Guideline.pdf
- Sanguinetti MC, Jiang C, Curran ME, Keating MT (1995) *Cell* 81:299–307
- Trudeau MC, Warmke JW, Ganetzky B, Robertson GA (1995) *Science* 269:92–95
- Mitcheson JS, Chen J, Lin M, Culberson C, Sanguinetti MC (2000) *Proc Natl Acad Sci USA* 97:12329–12333
- Sanchez-Chapula JA, Navarro-Polanco RA, Culberson C, Chen J, Sanguinetti MC (2002) *J Biol Chem* 277:23587–23595
- Sanguinetti MC, Chen J, Fernandez D, Kamiya K, Mitcheson J, Sanchez-Chapula JA (2005) *Novartis Found Symp* 266:159–166
- Perrin MJ, Subbiah RN, Vandenberg JI, Hill AP (2008) *Prog Biophys Mol Biol* 98:137–148
- Myokai T, Ryu S, Shimizu H, Oiki S (2008) *Mol Pharmacol* 73:1643–1651
- Zachariae U, Giordanetto F, Leach AG (2009) *J Med Chem* 52:4266–4276
- Neher E, Sakmann B (1992) *Sci Am* 266:44–51
- Kramer C, Beck B, Kriegl JM, Clark T (2008) *Chem Med Chem* 3:254–265
- Leong MK (2007) *Chem Res Toxicol* 20:217–226
- Aronov AM, Goldman BB (2004) *Bioorg Med Chem* 12:2307–2315
- Coi A, Massarelli I, Saraceno M, Carli N, Testai L, Calderone V, Bianucci AM (2009) *Chem Biol Drug Des* 74:416–433
- Thai KM, Ecker GF (2009) *Mol Divers* 13:321–336
- Nisius B, Goller AH, Bajorath J (2009) *Chem Biol Drug Des* 73:17–25
- Ermondi G, Visentin S, Caron G (2009) *Eur J Med Chem* 44:1926–1932
- Imai YN, Ryu S, Oiki S (2009) *J Med Chem* 52:1630–1638
- Du L, Li M, You Q, Xia L (2007) *Biochem Biophys Res Commun* 355:889–894
- Wang X, Yang Q, Yin D, Cheng Y, You Q (2008) *Chin J Chem* 26:2125–2132
- Evans DA, Doman TN, Thorner DA, Bodkin MJ (2007) *J Chem Inf Model* 47:1248–1257
- Schrödinger, LLC (2009) PHASE v.3.1. Schrödinger, LLC, New York
- Dixon SL, Smondyrev AM, Rao SN (2006) *Chem Biol Drug Des* 67:370–372
- Dixon SL, Smondyrev AM, Knoll EH, Rao SN, Shaw DE, Friesner RA (2006) *J Comput Aided Mol Des* 20:647–671
- Hansen K, Rathke F, Schroeter T, Rast G, Fox T, Kriegl JM, Mika S (2009) *J Chem Inf Model* 49:1486–1496
- Katritzky AR, Kulshyn OV, Stoyanova-Slavova I, Dobchev DA, Kuanar M, Fara DC, Karelson M (2006) *Bioorg Med Chem* 14:2333–2357
- Coi A, Massarelli I, Murgia L, Saraceno M, Calderone V, Bianucci AM (2006) *Bioorg Med Chem* 14:3153–3159
- Polak S, Wisniowska B, Brandys J (2009) *J Appl Toxicol* 29:183–206
- Zhou Z, Vorperian VR, Gong Q, Zhang S, January CT (1999) *J Cardiovasc Electrophysiol* 10:836–843
- Ridley JM, Milnes JT, Hancox JC, Witchel HJ (2006) *J Mol Cell Cardiol* 40:107–118
- Kirsch GE, Trepakova ES, Brimecombe JC, Sidach SS, Erickson HD, Kochan MC, Shyjka LM, Lacerda AE, Brown AM (2004) *J Pharmacol Toxicol Methods* 50:93–101
- Kornick CA, Kilborn MJ, Santiago-Palma J, Schulman G, Thaler HT, Keefe DL, Katchman AN, Pezzullo JC, Ebert SN, Woosley RL, Payne R, Manfredi PL (2003) *Pain* 105:499–506
- Trepakova ES, Dech SJ, Salata JJ (2006) *J Cardiovasc Pharmacol* 47:211–220
- Drolet B, Zhang S, Deschenes D, Rail J, Nadeau S, Zhou Z, January CT, Turgeon J (1999) *J Cardiovasc Electrophysiol* 10:1597–1604
- Katchman AN, Koerner J, Tosaka T, Woosley RL, Ebert SN (2006) *J Pharmacol Exp Ther* 316:1098–1106
- Traebert M, Dumotier B, Meister L, Hoffmann P, Dominguez-Estevéz M, Suter W (2004) *Eur J Pharmacol* 484:41–48
- Ridley JM, Milnes JT, Witchel HJ, Hancox JC (2004) *Biochem Biophys Res Commun* 325:883–891
- Mbai M, Rajamani S, January CT (2002) *Cardiovasc Res* 55:799–805
- Hurst RS, Higdon NR, Lawson JA, Clark MA, Rutherford-Root KL, McDonald WG, Haas JV, McGrath JP, Meglasson MD (2003) *Eur J Pharmacol* 482:31–37
- Davie C, Pierre-Valentin J, Pollard C, Standen N, Mitcheson J, Alexander P, Thong B (2004) *J Cardiovasc Electrophysiol* 15:1302–1309
- Kongsamut S, Kang J, Chen XL, Roehr J, Rampe D (2002) *Eur J Pharmacol* 450:37–41
- Yuill KH, Borg JJ, Ridley JM, Milnes JT, Witchel HJ, Paul AA, Kozlowski RZ, Hancox JC (2004) *Biochem Biophys Res Commun* 318:556–561
- Thomas D, Hammerling BC, Wu K, Wimmer AB, Ficker EK, Kirsch GE, Kochan MC, Wible BA, Scholz EP, Zitron E, Kathofer S, Kreye VA, Katus HA, Schoels W, Karle CA, Kiehn J (2004) *Br J Pharmacol* 142:485–494
- Ducroq J, Printemps R, Le Grand M (2005) *J Pharmacol Toxicol Methods* 52:115–122
- Thomas D, Bloehs R, Koschny R, Ficker E, Sykora J, Kiehn J, Schlomer K, Gierten J, Kathofer S, Zitron E, Scholz EP, Kiesecker C, Katus HA, Karle CA (2008) *Eur J Pharmacol* 579:98–103
- Crumb W, Llorca PM, Lancon C, Thomas GP, Garay RP, Hameg A (2006) *Eur J Pharmacol* 532:270–278
- Su Z, Martin R, Cox BF, Gintant G (2004) *J Mol Cell Cardiol* 36:151–160
- Walker BD, Singleton CB, Tie H, Bursill JA, Wyse KR, Valenzuela SM, Breit SN, Campbell TJ (2000) *Cardiovasc Res* 48:44–58
- Kim KS, Shin WH, Park SJ, Kim EJ (2007) *Int J Toxicol* 26:25–31
- Scholz E, Konrad F, Weiss D, Zitron E, Kiesecker C, Bloehs R, Kulzer M, Thomas D, Kathöfer S, Bauer A, Maurer M, Seemann

- G, Katus H, Karle C (2007) *Naunyn Schmiedebergs Arch Pharmacol* 376:275–284
54. Shin WH, Kim KS, Kim EJ (2006) *Pharmacol Res* 54:414–420
55. Kim KS, Kim EJ (2005) *Drug Chem Toxicol* 28:303–313
56. Kiesecker C, Zitron E, Luck S, Bloehs R, Scholz EP, Kathofer S, Thomas D, Kreye VA, Katus HA, Schoels W, Karle CA, Kiehn J (2004) *Naunyn Schmiedebergs Arch Pharmacol* 370:423–435
57. Katayama Y, Fujita A, Ohe T, Findlay I, Kurachi Y (2000) *J Pharmacol Exp Ther* 294:339–346
58. Ferreira S, Crumb WJ Jr, Carlton CG, Clarkson CW (2001) *J Pharmacol Exp Ther* 299:220–226
59. Guo L, Guthrie H (2005) *Pharmacol J Toxicol Methods* 52:123–135
60. Tie H, Walker BD, Valenzuela SM, Breit SN, Campbell TJ (2000) *Lancet* 355:1825
61. Thomas D, Wimmer AB, Wu K, Hammerling BC, Ficker EK, Kuryshev YA, Kiehn J, Katus HA, Schoels W, Karle CA (2004) *Naunyn Schmiedebergs Arch Pharmacol* 369:462–472
62. Katchman AN, McGroary KA, Kilborn MJ, Kornick CA, Manfredi PL, Woosley RL, Ebert SN (2002) *J Pharmacol Exp Ther* 303:688–694
63. Rajamani S, Eckhardt LL, Valdivia CR, Klemens CA, Gillman BM, Anderson CL, Holzem KM, Delisle BP, Anson BD, Makielski JC, January CT (2006) *Br J Pharmacol* 149:481–489
64. Lee SY, Kim YJ, Kim KT, Choe H, Jo SH (2006) *Br J Pharmacol* 148:499–509
65. Zitron E, Kiesecker C, Scholz E, Luck S, Bloehs R, Kathofer S, Thomas D, Kiehn J, Kreye VA, Katus HA, Schoels W, Karle CA (2004) *Naunyn Schmiedebergs Arch Pharmacol* 370:146–156
66. Rodriguez-Menchaca A, Ferrer-Villada T, Lara J, Fernandez D, Navarro-Polanco RA, Sanchez-Chapula JA (2006) *J Cardiovasc Pharmacol* 47:21–29
67. Hiramatsu M, Wu LM, Hirano Y, Kawano S, Furukawa T, Hiraoka M (2004) *Hear Vessel* 19:137–143
68. Kuryshev YA, Brown AM, Wang L, Benedict CR, Rampe D (2000) *J Pharmacol Exp Ther* 295:614–620
69. Paul AA, Witchel HJ, Hancox JC (2002) *Br J Pharmacol* 136:717–729
70. Kawakami K, Nagatomo T, Abe H, Kikuchi K, Takemasa H, Anson BD, Delisle BP, January CT, Nakashima Y (2006) *Br J Pharmacol* 147:642–652
71. Witchel HJ, Pabbathi VK, Hofmann G, Paul AA, Hancox JC (2002) *FEBS Lett* 512:59–66
72. Kuryshev YA, Ficker E, Wang L, Hawryluk P, Dennis AT, Wible BA, Brown AM, Kang J, Chen XL, Sawamura K, Reynolds W, Rampe D (2005) *Pharmacol Exp Ther* 312:316–323
73. Ferrer-Villada T, Navarro-Polanco RA, Rodriguez-Menchaca AA, Benavides-Haro DE, Sanchez-Chapula JA (2006) *Eur J Pharmacol* 531:1–8
74. Claassen S, Zunkler BJ (2005) *Pharmacology* 74:31–36
75. Duncan RS, McPate MJ, Ridley JM, Gao Z, James AF, Leishman DJ, Leaney JL, Witchel HJ, Hancox JC (2007) *Biochem Pharmacol* 74:425–437
76. Paul AA, Witchel HJ, Hancox JC (2001) *Biochem Biophys Res Commun* 280:1243–1250
77. Walker BD, Valenzuela SM, Singleton CB, Tie H, Bursill JA, Wyse KR, Qiu MR, Breit SN, Campbell TJ (1999) *Br J Pharmacol* 127:243–251
78. Gonzalez T, Arias C, Caballero R, Moreno I, Delpon E, Tamargo J, Valenzuela C (2002) *Br J Pharmacol* 137:1269–1279
79. Wang L, Wible BA, Wan X, Ficker E (2007) *J Pharmacol Exp Ther* 320:525–534
80. Tie H (2002) Cellular mechanisms of QT prolongation and proarrhythmia induced by non-antiarrhythmic drugs. University of New South Wales
81. Kang J, Wang L, Chen XL, Triggle DJ, Rampe D (2001) *Mol Pharmacol* 59:122–126
82. Wu LM, Orikabe M, Hirano Y, Kawano S, Hiraoka M (2003) *J Cardiovasc Pharmacol* 42:410–418
83. Caballero R, Moreno I, Gonzalez T, Arias C, Valenzuela C, Delpon E, Tamargo J (2003) *Circulation* 107:889–895
84. Bischoff U, Schmidt C, Netzer R, Pongs O (2000) *Eur J Pharmacol* 406:341–343
85. Dustan Sarazan R, Crumb WJ Jr, Beasley CM Jr, Emmick JT, Ferguson KM, Strnat CA, Sausen PJ (2004) *Eur J Pharmacol* 502:163–167
86. Volberg WA, Koci BJ, Su W, Lin J, Zhou J (2002) *J Pharmacol Exp Ther* 302:320–327
87. Lacroix P, Crumb WJ, Durando L, Ciottoli GB (2003) *Eur J Pharmacol* 477:69–72
88. Lu HR, Vlaminckx E, van de Water A, Rohrbacher J, Hermans A, Gallacher DJ (2007) *Eur J Pharmacol* 577:222–232
89. Ridley JM, Milnes JT, Benest AV, Masters JD, Witchel HJ, Hancox JC (2003) *Biochem Biophys Res Commun* 306:388–393
90. Saenen JB, Paulussen AD, Jongbloed RJ, Marcelis CL, Gilissen RA, Aerssens J, Snyders DJ, Raes AL (2007) *J Mol Cell Cardiol* 43:63–72
91. Siebrands CC, Friederich P (2007) *Eur J Anaesthesiol* 24:148–153
92. Danielsson BR, Lansdell K, Patmore L, Tomson T (2005) *Epilepsy Res* 63:17–25
93. Danielsson BR, Lansdell K, Patmore L, Tomson T (2003) *Epilepsy Res* 55:147–157
94. Ridley JM, Milnes JT, Zhang YH, Witchel HJ, Hancox JC (2003) *J Physiol* 549:667–672
95. Tripos (2010) SYBYL-X 1.1. Tripos, St. Louis
96. Topliss JG, Edwards RP (1979) *J Med Chem* 22:1238–1244
97. Hancox JC, James AF (2008) *Mol Pharmacol* 73:1592–1595
98. Aronov AM (2006) *J Med Chem* 49:6917–6921
99. Garg D, Gandhi T, Gopi Mohan C (2008) *J Mol Graph Model* 26:966–976
100. Durdagi S, Duff HJ, Noskov SY (2011) *J Chem Inf Model* 51:463–474

Molecular docking and structural analysis of cofactor-protein interaction between NAD^+ and 11β -hydroxysteroid dehydrogenase type 2

Hideaki Yamaguchi · Tatsuo Akitaya · Tao Yu ·
Yumi Kidachi · Katsuyoshi Kamiie · Toshiro Noshita ·
Hironori Umetsu · Kazuo Ryoyama

Received: 28 March 2011 / Accepted: 27 May 2011 / Published online: 11 June 2011
© Springer-Verlag 2011

Abstract Molecular docking and structural analysis of the cofactor-protein interaction between NAD^+ and human (h) or mouse (m) 11β -hydroxysteroid dehydrogenase type 2 (11β HSD2) were performed with the molecular operating environment (MOE). 11β HSD1 (PDB code: 3HFG) was selected as a template for the 3D structure modeling of 11β HSD2. The MOE docking (MOE-dock) and the alpha sphere and excluded volume-based ligand-protein docking (ASE-dock) showed that both NAD^+ -h 11β HSD2 and

NAD^+ -m 11β HSD2 models have a similar binding orientation to the template cofactor-protein model. Our present study also revealed that Asp91, Phe94, Tyr232 and Thr267 could be of importance in the interaction between NAD^+ and 11β HSD2. NADP^+ was incapable of entering into the cofactor-binding site of the 11β HSD2 models. The present study proposes the latest models for 11β HSD2 and its cofactor NAD^+ , and to the best of our knowledge, this is the first report of a m 11β HSD2 model with NAD^+ .

H. Yamaguchi (✉) · T. Akitaya
Department of Pharmacy, Faculty of Pharmacy, Meijo University,
150 Yagotoyama, Tenpaku,
Nagoya 468–8503, Japan
e-mail: hyamagu@meijo-u.ac.jp

T. Yu
Graduate School of Medicine, Department of Functional
Diagnostic Science, Osaka University,
1-7 Yamadaoka, Suita,
Osaka 565–0871, Japan

Y. Kidachi · K. Kamiie · K. Ryoyama
Department of Pharmacy,
Faculty of Pharmaceutical Sciences, Aomori University,
2-3-1 Kobata,
Aomori 030–0943, Japan

T. Noshita
Department of Life Sciences, Faculty of Life
and Environmental Sciences, Prefectural University of Hiroshima,
562 Nanatsuka,
Shobara 727–0023, Japan

H. Umetsu
Laboratory of Food Chemistry, Department of Life Sciences,
Junior College, Gifu Shotoku Gakuen University,
1-38 Nakauzura,
Gifu 055–8288, Japan

Keywords 11β HSD2 · Anticancer drug · ASE-dock ·
Cofactor-protein interactions · MOE

Abbreviations

11β HSD	11β -hydroxysteroid dehydrogenase
ASE-dock	Alpha sphere and excluded volume-based ligand-protein docking
HTS	High throughput screening
LBS	Ligand-binding site
MOE	Molecular operating environment
MOE-dock	Molecular operating environment docking
VS	Virtual screening

Introduction

Molecular modeling has found widespread utility in the field of drug development [1–3]. Various computational approaches are employed, and they have the ability to search large compound databases in silico and select a limited number of candidate molecules for in vitro testing to identify biologically desirable novel chemicals [4]. Virtual screening (VS) can be applied to search for potential active hits from a virtual library that represents an existing

compound library. It can also be used to estimate absorption, distribution, metabolism and excretion (ADME) parameters, drug-likeness and toxicity [2, 3, 5–8]. VS may sometimes have advantages over high throughput screening (HTS) in some applications. For example, HTS of 400,000 compounds resulted in 85 hits with IC_{50} values $<100 \mu\text{M}$, while biological testing of 365 proposed compounds derived by molecular docking returned 127 hits with IC_{50} values $<100 \mu\text{M}$ [9].

Glucocorticoids and mineralocorticoids induce a variety of physiological or pharmacological responses in cells, including proliferation, differentiation and apoptosis *via* the classic nuclear glucocorticoid and mineralocorticoid receptors (GR and MR) [10–12]. To activate the receptors, the amount of glucocorticoids and mineralocorticoids is dependent on both the circulating levels and the prereceptor metabolism catalyzed by 11β -hydroxysteroid dehydrogenases (11β HSDs) in the cells [13]. 11β HSD type 1 (11β HSD1) is NADPH-preferring and has been shown to have prominently reductase activity [14–16]. In contrast, 11β HSD type 2 (11β HSD2) is NAD^+ -requiring and shows only dehydrogenase activity for endogenous glucocorticoids [17–19].

Our previous study revealed that glycyrrhetic acid (GA), an 11β HSD2 inhibitor, was selectively toxic toward central nervous system-derived tumor cells [20], which suggests that targeting 11β HSD2 with highly selective inhibitors could be utilized for controlling the development and progression of cancer. Furthermore, in light of the advancement of *in silico* molecular modeling, structural analysis of 11β HSD2 with its possible ligands could be of importance for successful antitumor drug development. Although a few 11β HSD2 models have been reported [21–23], better and the latest models still need to be developed with the advancement of *in silico* modeling. Our strategy for the *in silico* approaches works

in three steps. The first step was to develop a model for 11β HSD2, which has been achieved recently [23]. The second step is to analyze the cofactor-protein interaction between NAD^+ and 11β HSD2. The third step will be, utilizing the results of the second step, to propose an antitumor compound that strongly inhibits 11β HSD2. In the present study, we will report the second step, the molecular docking and structural analysis of the cofactor-protein interaction between NAD^+ and 11β HSD2 by a highly sophisticated software package, the Molecular Operating Environment (MOE) 2009.10 (Chemical Computing Group Inc., Montreal, Canada).

Computational methods

Structural comparisons of the 11β HSD1 models

The crystal structure coordinates of 11β HSD1 (PDB code: 1XU7 [24], 1XU9 [24], 2BEL [25] and 3HFG [26]) were loaded into the MOE for the *in silico* analysis of the cofactor-protein interaction between NAD^+ and 11β HSD1.

Homology modeling of 11β HSD2

Homology modeling of 11β HSD2 was executed as previously reported [23]. In brief, the human (h) 11β HSD2 (NCBI reference sequence: NM_000196.3) [27] and mouse (m) 11β HSD2 (NCBI reference sequence: NM_008289.2) [28] sequences and the crystal structure coordinates of h 11β HSD1 (PDB code: 3HFG) [26] were loaded into the MOE. The primary structures of h 11β HSD1, h 11β HSD2 and m 11β HSD2 were aligned, carefully checked to avoid deletions or insertions in conserved regions and corrected wherever necessary. A series of h 11β HSD2 and m 11β HSD2 models were

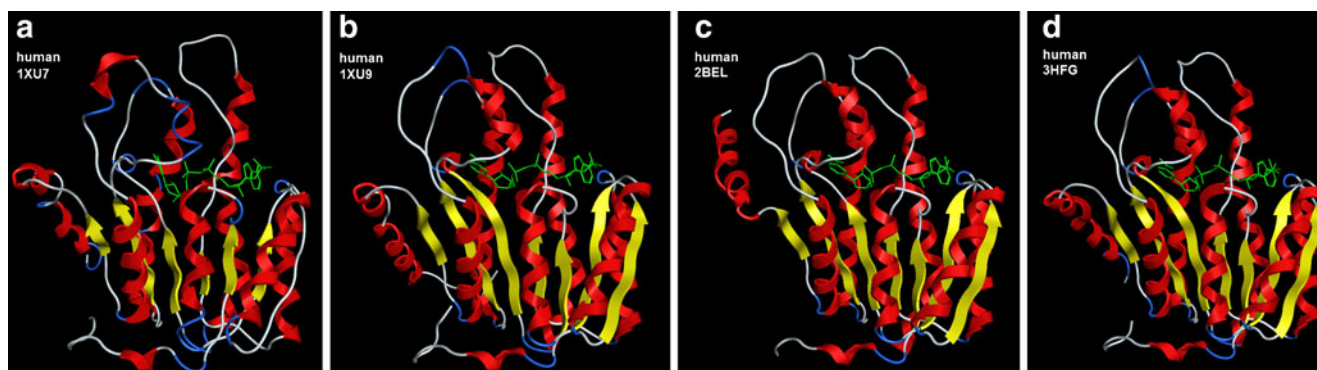


Fig. 1 The structures of the 11β HSD1 models. The crystal structure coordinates of h 11β HSD1 (PDB code: (a) 1XU7, (b) 1XU9, (c) 2BEL and (d) 3HFG) were loaded into the MOE and each model was

independently reconstructed. NAD^+ in the h 11β HSD1 models is located in the LBS near the cofactor-binding motif and the catalytic activity-related Ser170, Tyr183 and Lys187 triad

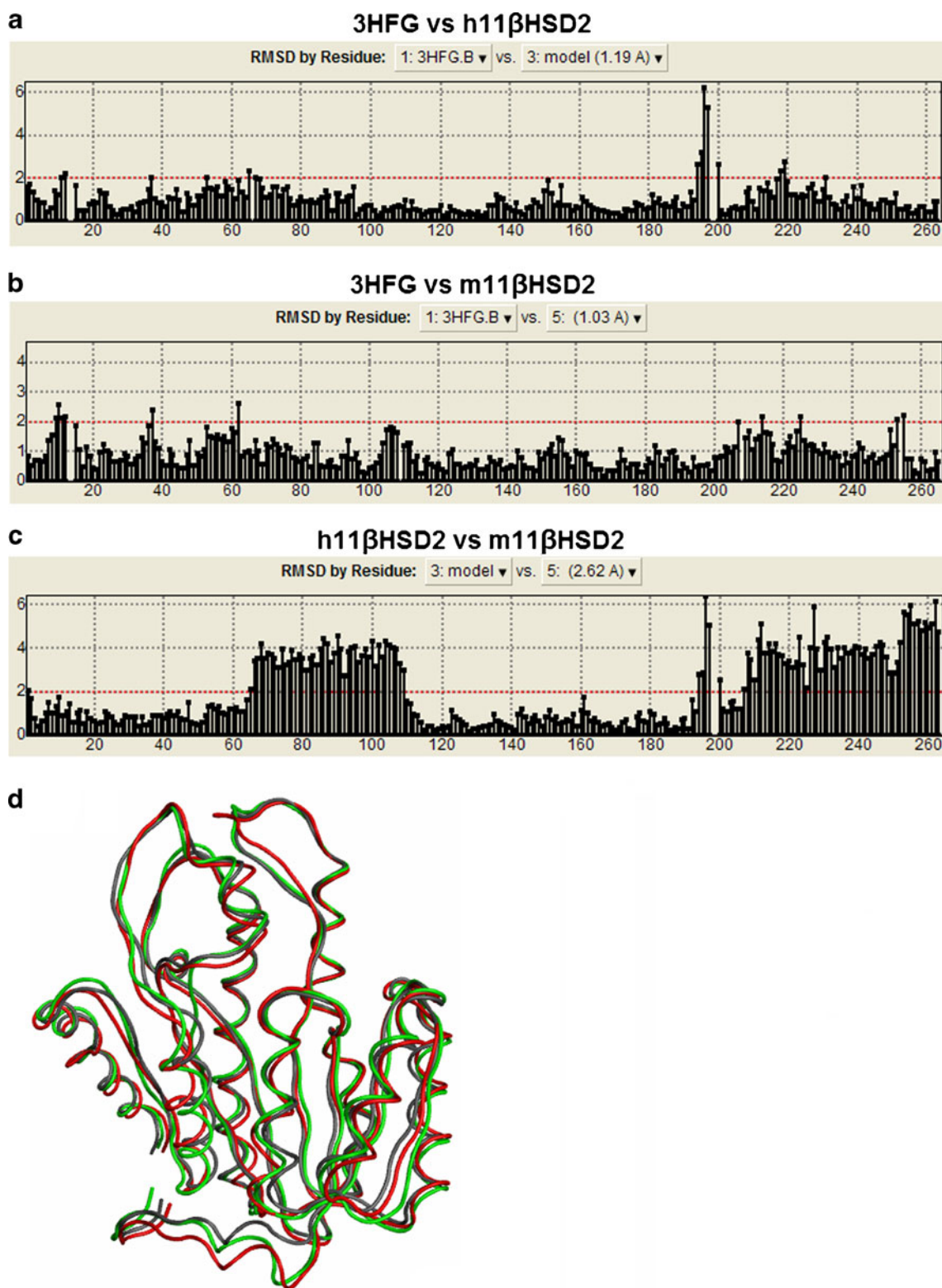


Fig. 2 RMSD values between the main chain atoms of the (a) h11βHSD1 vs h11βHSD2, (b) h11βHSD1 vs m11βHSD2 and (c) h11βHSD2 vs m11βHSD2 after main chain fit. RMSD values for h11βHSD1 vs h11βHSD2, h11βHSD1 vs m11βHSD2 and h11βHSD2 vs m11βHSD2 are 1.19, 1.03 and 2.62 Å, respectively. The positions of the amino acid residues are shown on the x-axis,

while the RMSD values are shown on the y-axis. The RMSD values for the residues located in the LBS near the cofactor-binding motif and the catalytic activity-related triad are always less than 2 Å. (d) A superimposition of the template h11βHSD1 (dark-gray), h11βHSD2 (green) and m11βHSD2 (red) models reveals that the three models exhibit significant 3D similarities

independently constructed with the MOE using a Boltzmann-weighted randomized procedure [29] combined with specialized logic for the handling of sequence insertions and deletions [30]. The models with the best packing quality function were selected for full energy minimization and further inspection.

Binding site selection and exploration

The binding site selection and exploration for 11 β HSD2 was executed as previously reported [23]. In brief, the Site Finder module of the MOE was used to identify possible substrate-binding pockets within the newly generated 3D structures of 11 β HSD2. Hydrophobic or hydrophilic alpha spheres served as probes denoting zones of tight atom packing. These alpha spheres were utilized to define potential ligand-binding sites (LBSs) and as centroids for the creation of dummy ligand atoms [31, 32]. The dummy atoms were matched to the corresponding alpha spheres during the characterization of the LBSs in h11 β HSD2 and m11 β HSD2. This method generates bound conformations that approach crystallographic resolutions [33].

MOE docking (MOE-dock)

The docking and analysis of the cofactor-protein interaction between NAD⁺ and 11 β HSD2 were performed with the MOE-dock system along with the simulated annealing method as a starter system [34]. The simulated annealing is a global optimization technique that is based on the Monte Carlo method [35]. It explores various states of a configuration space by

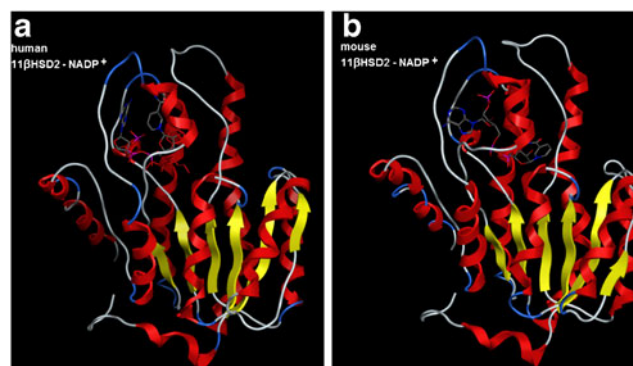


Fig. 4 Docking simulations of NADP⁺ to 11 β HSD2. **(a)** ASE-dock for h11 β HSD2. **(b)** ASE-dock for m11 β HSD2. The cofactor NADP⁺ is incapable of entering into the cofactor-binding site of 11 β HSD2. The capability of NAD⁺ (Fig. 3a–d) and incapability of NADP⁺ (Fig. 4a and b) entering into the cofactor-binding site of the 11 β HSD2 models support the previously reported *in vivo* function of 11 β HSD2 as a NAD⁺-dependent dehydrogenase

generating small random changes in the current state and then accepting or rejecting each new state according to the Metropolis criterion [36]. An LBS was identified by a cluster of hydrophobic and hydrophilic alpha spheres and ligand atoms were matched to the corresponding alpha spheres during the docking process. The ligand, NAD⁺ in the present study, explored the conformational space to locate the most favorable binding orientation and conformation by aligning and matching all triangles of the template points with compatible geometry. A total of 30 possible ligand poses were generated for h11 β HSD2 or m11 β HSD2, and an affinity scoring function, ΔG , was employed to rank candidate poses. Poses for the ligand were also scored based on complementarity with binding pocket alpha spheres.

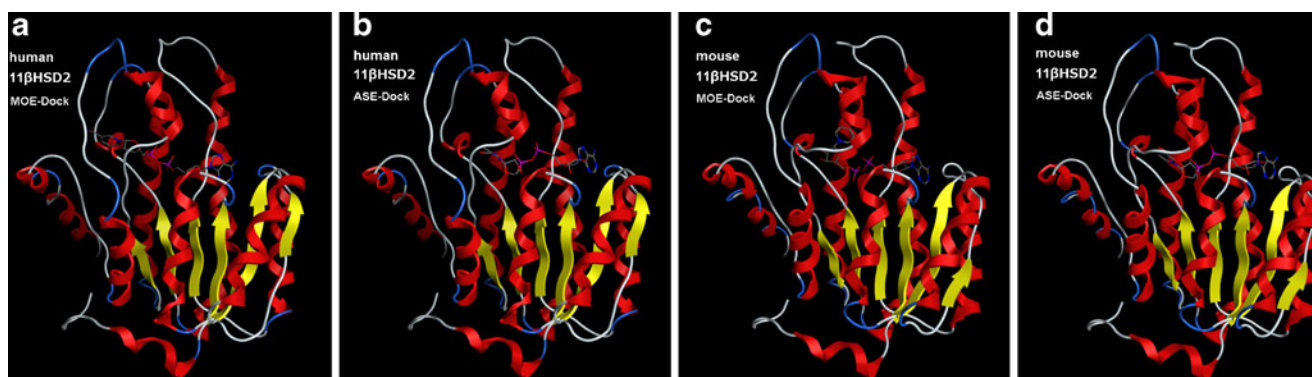


Fig. 3 Docking simulations of NAD⁺ to 11 β HSD2. The MOE-dock and ASE-dock were performed to evaluate the present docking simulation. **(a)** MOE-dock for h11 β HSD2. **(b)** ASE-dock for h11 β HSD2. **(c)** MOE-dock for m11 β HSD2. **(d)** ASE-dock for m11 β HSD2. Both simulations show that the cofactor NAD⁺ has similar binding orientation to the Rossmann fold in the h11 β HSD2

and m11 β HSD2 models. The similarity between the docked NAD⁺-11 β HSD2 poses (Fig. 3a–d) and the NADP⁺-11 β HSD1 models (Fig. 1a–d) suggests that the present methods are capable of generating the NAD⁺-11 β HSD2 models similar to the reported near-native 11 β HSD1 complex

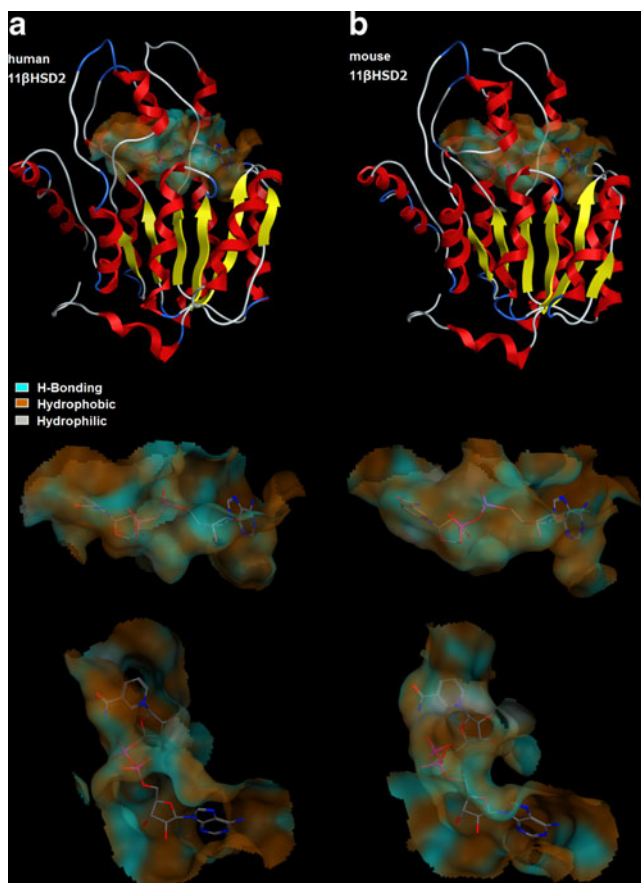


Fig. 5 Analysis of the molecular surfaces of the LBSs. The molecular surfaces of the LBSs (4.5 Å from NAD⁺) for the (a) h11βHSD2 and (b) m11βHSD2 models were analyzed. The hydrogen-bond (sky-blue), hydrophobic (light-brown) and hydrophilic (gray) regions in the LBSs of 11βHSD2 reveal that the adenosine part of NAD⁺ is somewhat surrounded by the hydrophobic regions (the close-up view, middle and lower panels)

Alpha sphere and excluded volume-based ligand-protein docking (ASE-dock)

The docking and analysis of the cofactor-protein interaction between NAD⁺ (or NADP⁺) and 11βHSD2 were also

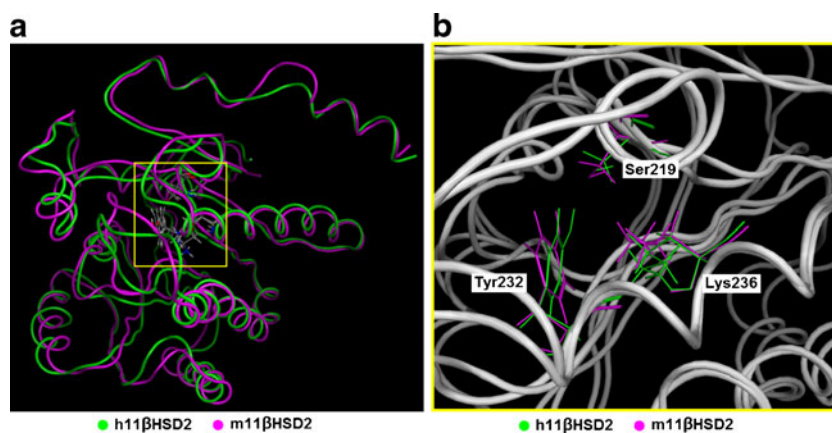
performed with ASE-dock in the MOE [37]. In the ASE-dock module, ligand atoms have alpha spheres within 1 Å. Based on this property, concave models are created and ligand atoms from a large number of conformations generated by superimposition with these points can be evaluated and scored by maximum overlap with alpha spheres and minimum overlap with the receptor atoms. The scoring function used by ASE-dock is based on ligand-protein interaction energies and the score is expressed as a U_{total} value. The ligand conformations were subjected to energy minimization using the MMF94S force field [38], and 500 conformations were generated using the default systematic search parameters. Five thousand poses per conformation were randomly placed onto the alpha spheres located within the LBS in 11βHSD2. From the resulting 500,000 poses, the 200 poses with the lowest U_{total} values were selected for further optimization with the MMF94S force field. During the refinement step, the ligand was free to move within the binding pocket.

Results and discussion

Structural comparisons of the 11βHSD1 models

The crystal structure coordinates of 11βHSD1 (PDB code: 1XU7 [24], 1XU9 [24], 2BEL [25] and 3HFG [26]) were loaded into the MOE and each model was independently reconstructed (Fig. 1a–d). The crystal structure resolution for 1XU7, 1XU9, 2BEL and 3HFG were 1.8, 1.55, 2.11 and 2.3 Å, respectively. The cofactor-binding motif of Gly41-XXX-Gly45-X-Gly47 [39–42] was conserved in all of the 11βHSD1 models. The secondary structures of all the models exhibited a central 6- or 7-stranded all-parallel β-sheet sandwich-like structure, flanked on both sides by 3-helices, and the models also exhibited similar 3D structures. It has been proposed that Ser170 is associated with catalysis by stabilizing the reaction intermediates,

Fig. 6 Structural comparison of the 11βHSD2 models at their LBSs. (a) Superimposition of the h11βHSD2 (green) and m11βHSD2 (magenta) models. The LBSs are enclosed in a yellow rectangle. (b) The LBSs of the h11βHSD2 (green) and m11βHSD2 (magenta) models. Ser219, Tyr232 and Lys236 are shown as stick models. Both models exhibit similar structures at their LBSs

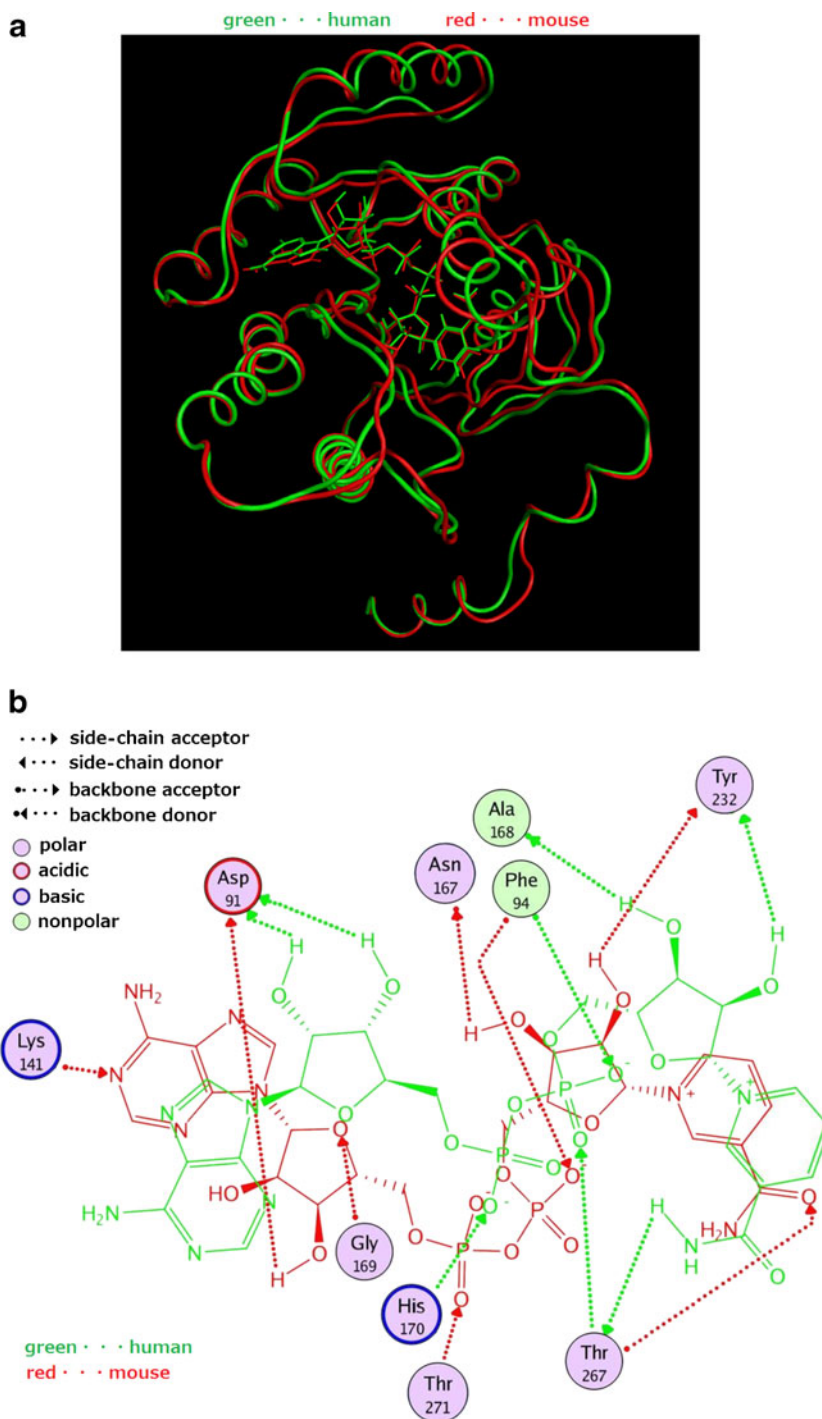


that the Tyr183 hydroxyl group is the proton donor involved in the electrophilic attack on the substrate carboxyl group in a reduction reaction, and that Lys187 facilitates the proton transfer from the hydroxyl oxygen of Tyr183 to the substrate [42]. NADP⁺ in the 11 β HSD1 models was located in the LBS near the cofactor-binding motif and the catalytic activity-related Ser170, Tyr183 and Lys187 triad [39–42].

Homology modeling of the 11 β HSD2 structures

11 β HSD1 (PDB code: 3HFG) was selected as a template (Fig. 1d) for the present 3D structure modeling of 11 β HSD2 because of its good crystal structure resolution (2.3 Å) and its information was the latest (from 2009) [26] among the 11 β HSD1 models (Fig. 1a–d). The sequence alignment of 11 β HSD1 and 11 β HSD2 has previously been reported [23].

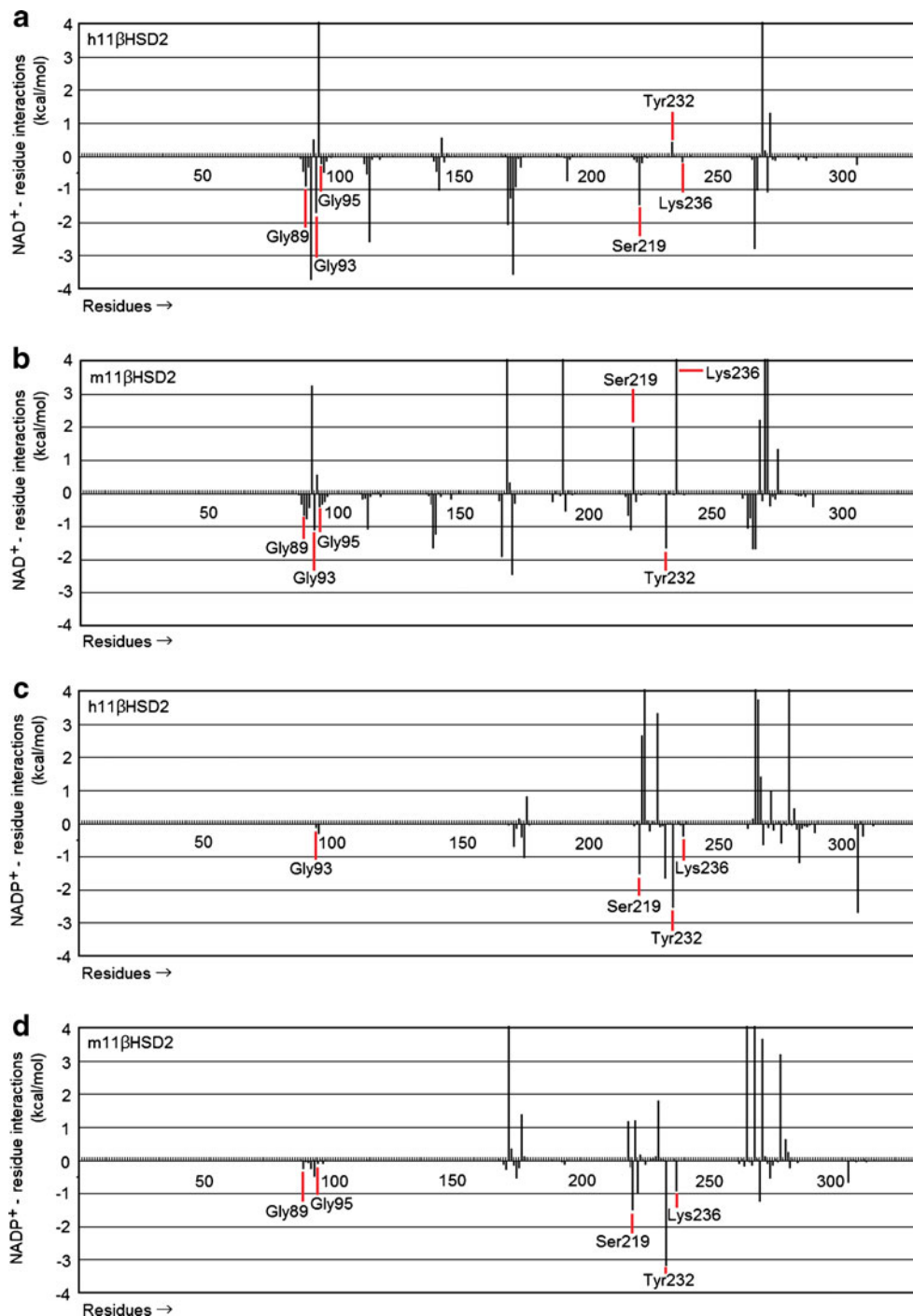
Fig. 7 Cofactor-protein interaction between NAD⁺ and 11 β HSD2. **(a)** A superimposition of the h11 β HSD2 (green) and m11 β HSD2 (red) models with NAD⁺. The two models exhibit significant 3D similarities. **(b)** The cofactor-protein interaction plots for NAD⁺-h11 β HSD2 (green) or NAD⁺-m11 β HSD2 (red). The bound conformation of NAD⁺ present in the LBS suggests that NAD⁺ can form a strong hydrogen bond with Asp91, Phe94, Tyr232 and Thr267. Apart from these common and important interactions for the cofactor NAD⁺, Ala168 and His170 (for h11 β HSD2) and Lys141, Asn167, Gly169 and Thr271 (for m11 β HSD2) reveal possible interactions with the cofactor



The % sequence identity between h11 β HSD1 and h11 β HSD2 was 21.4%, and that between h11 β HSD1 and m11 β HSD2 was 23.7%. For the construction of the 11 β HSD2 models, 100 independent models of the target proteins were built using a Boltzmann-weighted randomized modeling procedure in the MOE that was adapted from reports by Levitt [29] and Fichteler et al. [30]. The intermediate models were evaluated by a residue packing quality function, which is sensitive to the

degrees to which non-polar side-chain groups are buried and hydrogen bonding opportunities are satisfied. The 11 β HSD2 models with the best packing quality function and full energy minimization were utilized in the present study. The secondary structures of the 11 β HSD2 models exhibited a central 6-stranded all-parallel β -sheet sandwich-like structure, flanked on both sides by 3-helices, which was also found in the 11 β HSD1 models. The stereochemical qualities of the

Fig. 8 Cofactor-residue interaction energies between the cofactor and 11 β HSD2. The cofactor-residue interaction energies between (a) NAD⁺ and h11 β HSD2, (b) NAD⁺ and m11 β HSD2, (c) NADP⁺ and h11 β HSD2 and (d) NADP⁺ and m11 β HSD2 were calculated by the methods of Labute [33] using the MOE, assigning energy terms in kcal mol⁻¹ for each residue. Generally, a negative value indicates that the residue attracts the cofactor, while the residue with a positive value repels the cofactor. Gly89, Gly93 and Gly95 in h and m11 β HSD2 appear to attract NAD⁺. Ser219 and Lys236 in h11 β HSD2 and Tyr232 in m11 β HSD2 can also attract NAD⁺ (Fig. 8a and b). In contrast, although the catalytic triad in 11 β HSD2 seem to attract NADP⁺, the cofactor-binding motif shows very little NADP⁺ attracting energies (Fig. 8c and d), which supports the results in Fig. 4 that NADP⁺ could not fit into the LBS in 11 β HSD2



h11 β HSD2 and m11 β HSD2 models were assessed by Ramachandran plots, and only 1.9% for h11 β HSD2 [23] and 1.1% for m11 β HSD2 were in the disfavored region, which indicates that the phi and psi backbone dihedral angles in the h11 β HSD2 and m11 β HSD2 models were reasonably accurate. Root mean square deviation (RMSD) values between the main chain atoms of the h11 β HSD1 (3HFG) vs h11 β HSD2 (Fig. 2a), h11 β HSD1 vs m11 β HSD2 (Fig. 2b) and h11 β HSD2 vs m11 β HSD2 (Fig. 2c) after main chain fit were 1.19, 1.03 and 2.62 Å, respectively. RMSD values for each residue were also analyzed. The RMSD values for the residues located in the LBS near the cofactor-binding motif and the catalytic activity-related triad were always less than 2 Å (Fig. 2a–c). A superimposition of the template h11 β HSD1 (dark-gray), h11 β HSD2 (green) and m11 β HSD2 (red) models revealed that the three models exhibited significant 3D similarities (Fig. 2d).

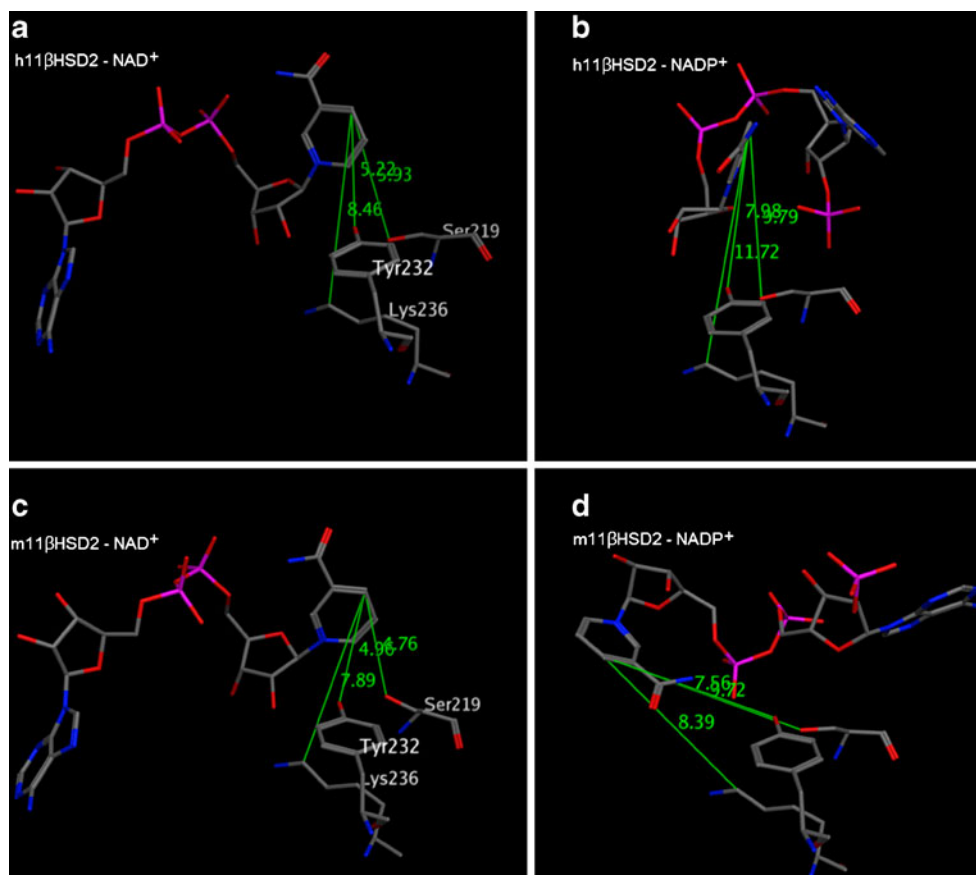
Docking simulations of NAD⁺ to 11 β HSD2

The MOE-dock and ASE-dock were performed to evaluate the present docking simulation. Both simulations showed that the cofactor NAD⁺ had a similar binding orientation to the Rossmann fold in the h11 β HSD2 and m11 β HSD2

models (Fig. 3a–d). The cofactor orientation in 11 β HSD2 was also similar to that in 11 β HSD1 (Fig. 1a–d). The similarity between the present docked NAD⁺-11 β HSD2 poses (especially with the ASE-dock simulations) and the 11 β HSD1 models suggests that the present methods are capable of generating the NAD⁺-11 β HSD2 models similar to the reported near-native 11 β HSD1 complex. Thus, the ASE-dock simulations were used in the subsequent docking analyses. NADP⁺ was incapable of entering into the cofactor-binding site of 11 β HSD2 (Fig. 4a and b). Activity of 11 β HSD1 is NADP(H)-dependent and bi-directional possessing both dehydrogenase and reductase activity, but in vivo the enzyme appears to function almost exclusively as a reductase [14–16]. By contrast, 11 β HSD2 utilizes NAD⁺, functions as a dehydrogenase and serves to protect the MR from corticoids excess [17–19]. Our in silico results in the present study showed the capability of NAD⁺ and incapability of NADP⁺ entering into the cofactor-binding site of the 11 β HSD2 models, which supports the previously reported in vivo function of 11 β HSD2 as a NAD⁺-dependent dehydrogenase.

The 11 β HSD2 models also presented similar structures of their LBSs and the molecular surfaces of the LBSs (Fig. 5a and b). The hydrogen-bond (sky-blue), hydropho-

Fig. 9 Predicted interatomic distances between the cofactor and the catalytic triad. (a) NAD⁺ and h11 β HSD2. (b) NADP⁺ and h11 β HSD2. (c) NAD⁺ and m11 β HSD2. (d) NADP⁺ and m11 β HSD2. For h11 β HSD2, the shortest distance between the C4 atom of NAD⁺ and Tyr232 is 5.22 Å (Fig. 9a), while that between NADP⁺ and Tyr232 is 7.98 Å (Fig. 9b). For m11 β HSD2, the distance between NAD⁺ and Ser219 is 4.76 Å (Fig. 9c), while that between NADP⁺ and Tyr232 is 7.56 Å (Fig. 9d). NADP⁺ is always farther from the catalytic triad



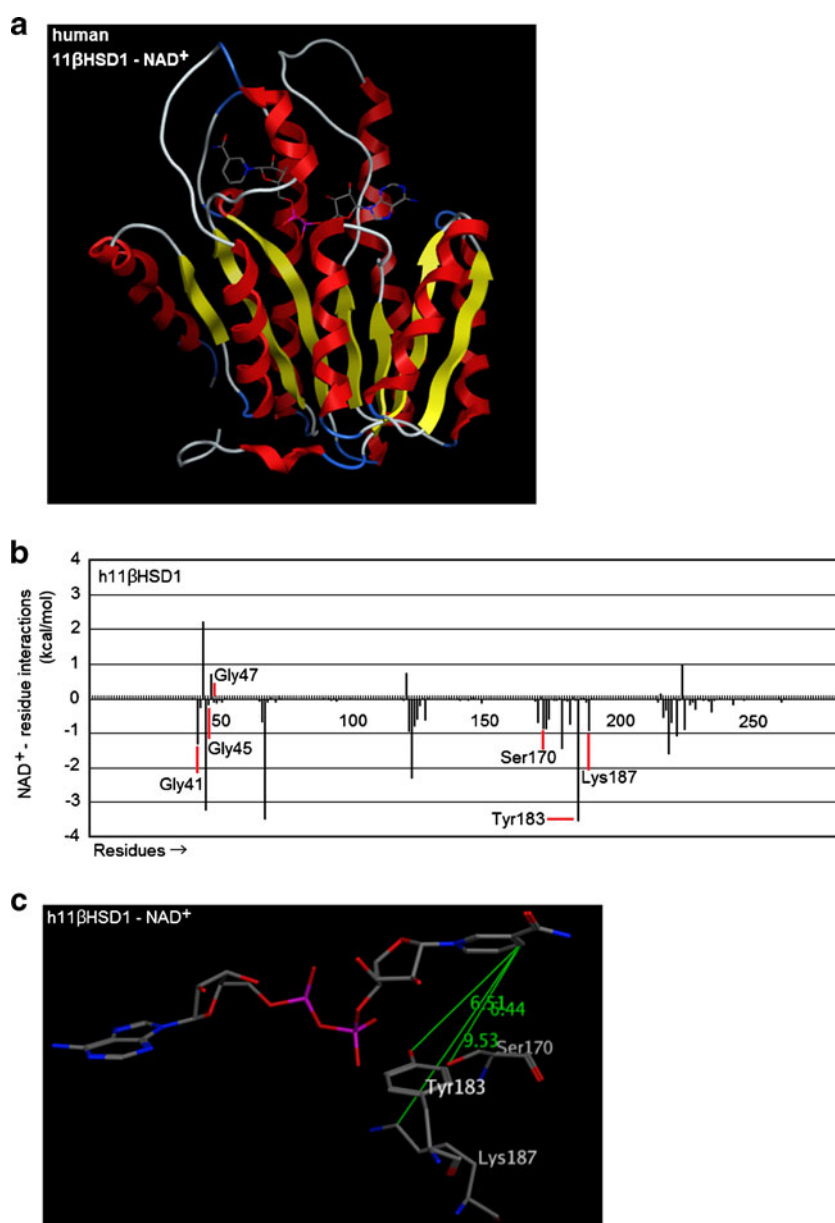
bic (light-brown) and hydrophilic (gray) regions in the LBSs of 11 β HSD2 revealed that the adenosine part of NAD⁺ was somewhat surrounded by the hydrophobic regions (Fig. 5a and b; the close-up view, middle and lower panels). This result is supported by the previous report by Vincent et al. [43].

Cofactor-protein interaction between NAD⁺ and 11 β HSD2

First, structural comparisons of the 11 β HSD2 models without NAD⁺ in the LBSs were executed. A superimposition of the h11 β HSD2 (green) and m11 β HSD2 (magenta) models is shown in Fig. 6a. The two models exhibited significant 3D similarities. They also presented similar

structures of their LBSs (Fig. 6b). A superimposition of the h11 β HSD2 (green) and m11 β HSD2 (red) models with NAD⁺ also demonstrated that the two models exhibited significant 3D similarities (Fig. 7a). Furthermore, to create the cofactor-protein interaction plots for NAD⁺-11 β HSD2, the Ligand Interactions module of the MOE was used, which provided a clearer arrangement of putative key intermolecular interactions that aid in interpretation of the 3D juxtaposition of the cofactor and the LBS in 11 β HSD2 (Fig. 7b). Asp91 has been reported to play an important role in the binding of the cofactor in h11 β HSD2 models [21, 44], and our present results exhibited that Asp91, Phe94, Tyr232 and Thr267 could be of importance in both the h11 β HSD2 and m11 β HSD2

Fig. 10 Docking simulation of NAD⁺ to 11 β HSD1 and interaction between them. (a) ASE-dock. (b) Cofactor-residue interaction energies. (c) Predicted interatomic distances. NAD⁺ has a similar binding orientation to the Rossmann fold (Fig. 10a) in the h11 β HSD1-NADP⁺ model (PDB code: 3HFG). The conserved cofactor-binding motif (Gly41, Gly45 and Gly47) and the catalytic triad (Ser170, Tyr183 and Lys187) could attract NAD⁺ (Fig. 10b). However, the distance between NAD⁺ and the catalytic triad in h11 β HSD1 is rather far (6.44 Å; Fig. 10c) compared to the h11 β HSD2-NAD⁺ model (5.22 Å; Fig. 9a) and the m11 β HSD2-NAD⁺ model (4.76 Å; Fig. 9c), which may be the reason why 11 β HSD1 is NADPH-preferring (not NAD⁺) [14–16] and 11 β HSD2 is NAD⁺-requiring [17–19]



models. The bound conformation of NAD^+ present in the LBS suggests that NAD^+ can form a strong hydrogen bond with Asp91, Phe94, Tyr232 and Thr267. Apart from these common and important interactions for the cofactor NAD^+ , Ala168 and His170 (for h11 β HSD2) and Lys141, Asn167, Gly169 and Thr271 (for m11 β HSD2) revealed possible interactions with the cofactor. The pentose bound to the nicotinamide part of NAD^+ had two interactions (Ala168 and Tyr232 for h11 β HSD2; Asn167 and Tyr 232 for m11 β HSD2) with 11 β HSD2. The pentose in the adenosine part of NAD^+ had two interactions (Asp91 and Asp91 for h11 β HSD2; Asp91 and Gly169 for m11 β HSD2) with 11 β HSD2. Further, the phosphoric acid part of NAD^+ had two interactions (His170 and Thr267 for h11 β HSD2; Phe94 and Thr271 for m11 β HSD2) with 11 β HSD2. These interactions perhaps contribute to the stable binding of NAD^+ to 11 β HSD2. Only Lys141 in m11 β HSD2 was found to have an interaction with the N1 position of the adenosine part of NAD^+ . An interaction with the N1 position of adenosine can be found in the famous Watson-Crick DNA base pairing, and the NAD^+ -Lys141 interaction may function as a further stabilizer of the NAD^+ -m11 β HSD2 binding.

Cofactor-residue interaction energies between NAD^+ and 11 β HSD2

Further, the cofactor-residue interaction energies were calculated by the methods of Labute [33] using the MOE, assigning energy terms in kcal mol^{-1} for each residue. Generally, a negative value indicated that the residue attracted the cofactor, while the residue with a positive value repelled the cofactor. Among the conserved cofactor-binding motif of Gly89-XXX-Gly93-X-Gly95 and the Ser219, Tyr232 and Lys236 catalytic triad, Gly89, Gly93 and Gly95 in h and m11 β HSD2 appeared to attract NAD^+ . Ser219 and Lys236 in h11 β HSD2 and Tyr232 in m11 β HSD2 could also attract NAD^+ (Fig. 8a and b). In contrast, although the catalytic triad in 11 β HSD2 seemed to attract NADP^+ , the cofactor-binding motif showed very little NADP^+ attracting energies (Fig. 8c and d), which supports the results in Fig. 4 that NADP^+ could not fit into the LBS in 11 β HSD2. These results were also supported by the analysis of the distances between the C4 atom of NAD^+ (or NADP^+) and the nearest atom in the catalytic triad (Fig. 9a–d). For h11 β HSD2, the shortest distance between NAD^+ and Tyr232 was 5.22 Å (Fig. 9a), while that between NADP^+ and Tyr232 was 7.98 Å (Fig. 9b). For m11 β HSD2, the distance between NAD^+ and Ser219 was 4.76 Å (Fig. 9c), while that between NADP^+ and Tyr232 was 7.56 Å (Fig. 9d). NADP^+ was always farther from the catalytic triad.

Docking simulation of NAD^+ to 11 β HSD1 and interaction between them

ASE-dock was performed to examine if NAD^+ can fit into the LBS in h11 β HSD1. Interestingly, the simulation

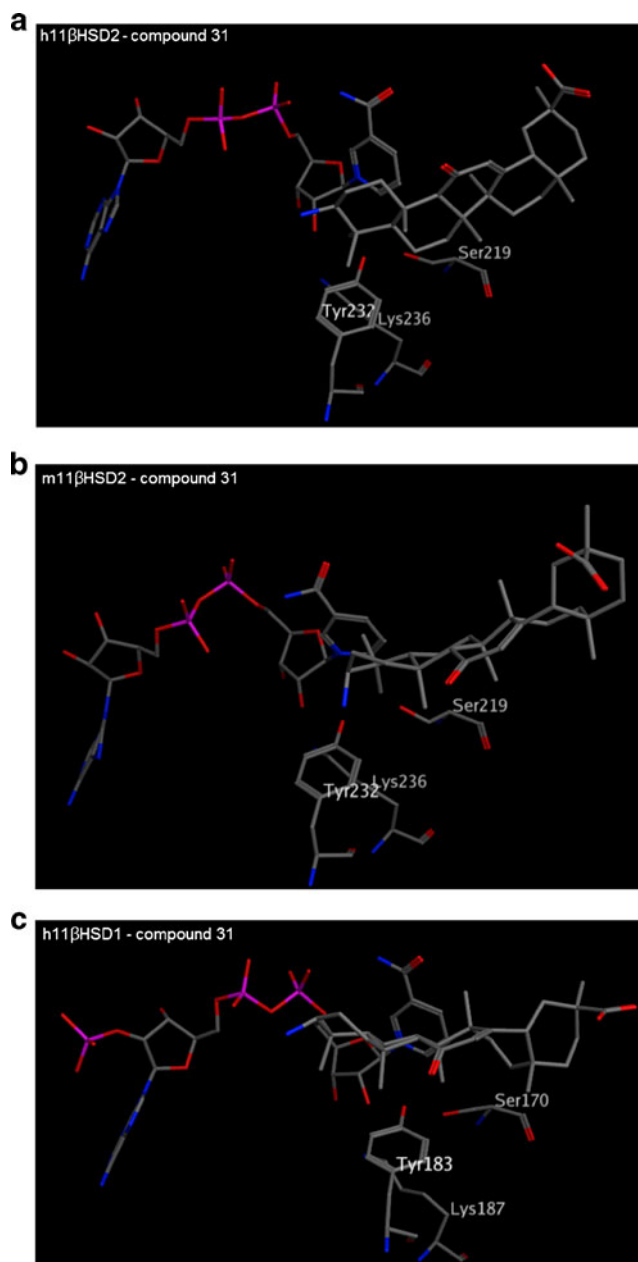


Fig. 11 Docking simulation of a GA derivative and 11 β HSD2. The analyses of the ligand-receptor docking between a GA derivative (Compound 31; OH at C3 of GA was replaced by NH_2 [45].) and the modeled 11 β HSD2 (or 11 β HSD1) were performed. (a) Compound 31 and h11 β HSD2. (b) Compound 31 and m11 β HSD2. (c) Compound 31 and h11 β HSD1. The docking simulation shows that the compound 31 has a similar binding orientation to the LBS (Fig. 11a–c) in the reported 11 β HSD1-compound 31 complex [45]. This result suggests that the homology modeling of 11 β HSD2 and the docking simulations in the present study are performed reasonably well

showed that NAD^+ had a similar binding orientation to the Rossmann fold (Fig. 10a) in the h11 β HSD1-NADP⁺ model (PDB code: 3HFG). The cofactor-residue interaction energies revealed that the conserved cofactor-binding motif (Gly41, Gly45 and Gly47) and the catalytic triad (Ser170, Tyr183 and Lys187) could attract NAD^+ (Fig. 10b). However, the distance between NAD^+ and the catalytic triad in h11 β HSD1 was rather far (6.44 Å; Fig. 10c) compared to the h11 β HSD2-NAD⁺ model (5.22 Å; Fig. 9a) and the m11 β HSD2-NAD⁺ model (4.76 Å; Fig. 9c), which may be the reason why 11 β HSD1 is NADPH-preferring (not NAD^+) [14–16] and 11 β HSD2 is NAD^+ -requiring [17–19].

Docking simulation of a GA derivative and 11 β HSD2

The analyses of the ligand-receptor docking between a GA derivative (Compound 31; OH at C3 of GA was replaced by NH_2 [45].) and the modeled 11 β HSD2 (or 11 β HSD1) were also performed. Compound 31 has been reported to inhibit both 11 β HSD1 and 2 [45]. The docking simulation showed that compound 31 had a similar binding orientation to the LBS (Fig. 11a–c) in the reported 11 β HSD1-compound 31 complex [45]. This result suggests that the homology modeling of 11 β HSD2 and the docking simulations in the present study were performed reasonably well.

Conclusions

The analysis of the cofactor-binding region in 11 β HSD2 revealed that a subtle change of the NAD^+ binding orientation significantly altered its enzymatic activity [44], which indicates that the location of the cofactor in the enzyme is very important. Consequently, the location of the cofactor possibly influences the physico-chemical properties of the LBS in the enzyme and has some effects on the binding of the inhibitor to the enzyme. Thus, detailed analysis of the cofactor-protein interaction is of great significance in designing in silico 11 β HSD2-inhibitor models for successful development of antitumor drugs. The main objective in the present study was to analyze the cofactor-protein interaction between NAD^+ and 11 β HSD2. Two docking simulations were performed to analyze the interaction between NAD^+ and 11 β HSD2. The similarity between the NAD^+ -11 β HSD2 poses and the previously reported NADP⁺-11 β HSD1 models suggests that the present methods are capable of generating the NAD^+ -11 β HSD2 docking models similar to the near-native NADP⁺-11 β HSD1 complex. Our present results also revealed that Asp91, Phe94, Tyr232 and Thr267 could be of importance in both the NAD^+ -h11 β HSD2 and NAD^+ -m11 β HSD2 interactions. Consequently, it is proposed that

the h11 β HSD2 model, as well as the m11 β HSD2 model, with NAD^+ in the present study will be suitable for further in silico structure-based de novo drug design. Furthermore, to the best of our knowledge, this is the first report of a m11 β HSD2 model with NAD^+ .

Acknowledgments This study was partially supported by a grant-in-aid from the Promotion and Mutual Aid Corporation for Private Schools of Japan.

References

- Kurogi Y, Guner OF (2001) *Curr Med Chem* 8:1035–1055
- Ekins S (2004) *Drug Discovery Today* 9:276–285
- Jorgensen WL (2004) *Science* 303:1813–1818
- Bajorath J (2002) *Nat Rev Drug Discovery* 1:882–894
- Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) *Adv Drug Delivery Rev* 23:3–25
- van de Waterbeemd H, Gifford E (2003) *Nat Rev Drug Discovery* 2:192–204
- Lipinski CA (2004) *Drug Discovery Today: Technologies* 1:337–341
- Shoichet BK (2004) *Nature* 432:862–865
- Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, Connolly DT, Shoichet BK (2002) *J Med Chem* 45:2213–2221
- Funder JW (2005) *Heart Fail Rev* 10:15–22
- Hu GX, Lian QQ, Lin H, Latif SA, Morris DJ, Hardy MP, Ge RS (2008) *Steroids* 73:1018–1024
- Revollo JR, Cidowski JA (2009) *Ann NY Acad Sci* 1179:167–178
- Rabbitt EH, Lavery GG, Walker EA, Cooper MS, Stewart PM, Hewison M (2002) *FASEB J* 16:36–44
- Kotelevtsev Y, Holmes MC, Burchell A, Houston PM, Schmol D, Jamieson P, Best R, Brown R, Edwards CR, Seckl JR, Mullins JJ (1997) *Proc Natl Acad Sci USA* 94:14924–14929
- Holmes MC, Kotelevtsev Y, Mullins JJ, Seckl JR (2001) *Mol Cell Endocrinol* 171:15–20
- Morton NM, Holmes MC, Fievet C, Staels B, Tailleux A, Mullins JJ, Seckl JR (2001) *J Biol Chem* 276:41293–41300
- Walker BR, Campbell JC, Williams BC, Edwards CR (1992) *Endocrinology* 131:970–972
- Albiston AL, Obeyesekere VR, Smith RE, Krozowski ZS (1994) *Mol Cell Endocrinol* 105:R11–R17
- Agarwal AK, Mune T, Monder C, White PC (1995) *Endocr Res* 21:389–397
- Yamaguchi H, Kidachi Y, Kamiie K, Noshita T, Umetsu H, Ryoyama K (2010) *Biol Pharm Bull* 33:321–324
- Carvajal CA, Gonzalez AA, Romero DG, González A, Mosso LM, Lagos ET, Hevia Mdel P, Rosati MP, Perez-Acle TO, Gomez-Sanchez CE, Montero JA, Fardella CE (2003) *J Clin Endocrinol Metab* 88:2501–2507
- Koch MA, Wittenberg LO, Basu S, Jeyaraj DA, Gourzoulidou E, Reinecke K, Odermatt A, Waldmann H (2004) *Proc Natl Acad Sci USA* 101:16721–16726
- Yamaguchi H, Akitaya T, Yu T, Kidachi Y, Kamiie K, Noshita T, Umetsu H, Ryoyama K (2011) *Eur J Med Chem* 46:1325–1330
- Hosfield DJ, Wu Y, Skene RJ, Hilgers M, Jennings A, Snell GP, Aertgeerts K (2005) *J Biol Chem* 280:4639–4648
- Ogg D, Elleby B, Norström C, Stefansson K, Abrahmsén L, Oppermann U, Svensson S (2005) *J Biol Chem* 280:3789–3794

26. Wan ZK, Chenail E, Xiang J, Li HQ, Ipek M, Bard J, Svenson K, Mansour TS, Xu X, Tian X, Suri V, Hahm S, Xing Y, Johnson CE, Li X, Qadri A, Panza D, Perreault M, Tobin JF, Saiah E (2009) *J Med Chem* 52:5449–5461
27. Ni XT, Duan T, Yang Z, Guo CM, Li JN, Sun K (2009) *Placenta* 30:1023–1028
28. Huang Y, Li X, Lin H, Chu Y, Chen B, Lian Q, Ge RS (2010) *Biochem Biophys Res Commun* 391:1752–1756
29. Levitt M (1992) *J Mol Biol* 226:507–533
30. Fechteler T, Dengler U, Schomberg D (1995) *J Mol Biol* 253:114–131
31. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S (1998) *Proteins* 33:1–17
32. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S (1998) *Proteins* 33:18–29
33. Goto J, Kataoka R, Hirayama N (2004) *J Med Chem* 47:6804–6811
34. Chang DT, Oyang YJ, Lin JH (2005) *Nucleic Acids Res* 33(suppl 2):W233–W238
35. Metropolis N, Ulam S (1949) *J Am Stat Assoc* 44:335–341
36. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH (1953) *J Chem Phys* 21:1087–1092
37. Goto J, Kataoka R, Muta H, Hirayama N (2008) *J Chem Inf Model* 48:583–590
38. Halgren TA (1996) *J Comput Chem* 17:490–519
39. Obeid J, White PC (1992) *Biochem Biophys Res Commun* 188:222–227
40. Oppermann UC, Filling C, Berndt KD, Persson B, Benach J, Ladenstein R, Jornvall H (1997) *Biochemistry* 36:34–40
41. Filling C, Berndt KD, Benach J, Knapp S, Prozorovski T, Nordling E, Ladenstein R, Jornvall H, Oppermann UC (2002) *J Biol Chem* 277:25677–25684
42. Hoffmann F, Maser E (2007) *Drug Metab Rev* 39:87–144
43. Vincent SJF, Zwahlen C, Post CB, Burgner JW, Bodenhausen G (1997) *Proc Natl Acad Sci USA* 94:4383–4388
44. Atanasov AG, Tam S, Rochen JM, Baker ME, Odermatt A (2003) *Biochem Biophys Res Commun* 308:257–262
45. Schuster D, Maurer EM, Laggner C, Nashev LG, Wilckens T, Langer T, Odermatt A (2006) *J Med Chem* 49:3454–3466

Determination of best-fit potential parameters for a reactive force field using a genetic algorithm

Poonam Pahari · Shashank Chaturvedi

Received: 9 February 2011 / Accepted: 9 May 2011 / Published online: 11 June 2011
© Springer-Verlag 2011

Abstract The ReaxFF interatomic potential, used for organic materials, involves more than 600 adjustable parameters, the best-fit values of which must be determined for different materials. A new method of determining the set of best-fit parameters for specific molecules containing carbon, hydrogen, nitrogen and oxygen is presented, based on a parameter reduction technique followed by genetic algorithm (GA) minimization. This work has two novel features. The first is the use of a parameter reduction technique to determine which subset of parameters plays a significant role for the species of interest; this is necessary to reduce the optimization space to manageable levels. The second is the application of the GA technique to a complex potential (ReaxFF) with a very large number of adjustable parameters, which implies a large parameter space for optimization. In this work, GA has been used to optimize the parameter set to determine best-fit parameters that can reproduce molecular properties to within a given accuracy. As a test problem, the use of the algorithm has been demonstrated for nitromethane and its decomposition products.

Keywords Genetic algorithm · Force field · Decomposition products · Potential parameters

PACS numbers 34.20.Cf · 83.10.Mj · 71.15.Pd · 31.50.-x · 33.15.-e

P. Pahari (✉) · S. Chaturvedi
Computational Analysis Division,
Bhabha Atomic Research Centre,
Visakhapatnam, India
e-mail: poonam.pahari@gmail.com

S. Chaturvedi
e-mail: shashankvizag@gmail.com

Introduction

Molecular dynamics (MD) simulations involve the use of a potential function to determine the forces acting on individual particles. The trajectories of the particles under the influence of this self-consistent force yield the temporal evolution of the system [1–4]. MD methods can be divided into ab initio and classical methods. Ab initio methods, although very accurate and general, are computationally extremely demanding. Their application is thus restricted to relatively small systems and short simulation times [5]. On the other hand, the use of classical force fields allow the dynamical simulation of millions of atoms, which makes them applicable to the study of a wide variety of physical processes (e.g., shock waves, dislocation dynamics, fracture and oxidation) that require larger system sizes and simulation times. However, the force fields are very difficult to develop, and their accuracy must be established for each application. The main challenge is to develop methodologies that retain the accuracy of quantum mechanics while allowing large-scale simulations.

The interaction between atoms in a polyatomic molecule, or in a solid, can be described in terms of the potential energy surface (PES). This specifies the potential energy of a system of atoms in terms of the coordinates of all the atoms. There are two techniques for obtaining the PES. The first involves the generation of potential energy data using ab initio electronic structure methods, followed by a variety of fitting procedures [6, 7]. The second technique involves fitting a functional form for the PES to data obtained from high-resolution spectroscopy [8] or scattering experiments [9, 10], such as rotational spectra or vibrational energy levels. In the second technique, it is a major task to gather available information to construct a functional representation which can be used for MD simulations [11–13]. Hence, other options have also been explored, such as neural networks and interpolative

moving least squares [14–16]. The use of genetic algorithms (GA) for determining best-fit potential parameters for nickel, by matching solid-state properties, is reported in [17].

For complex, polyatomic molecules, particularly those involving CHNO atoms, the potentials must allow for bond formation and breaking, and for the influence of close-neighbor effects on molecular structures. Potentials that take these effects into account are called “reactive potentials”—examples include Tersoff, Brenner, REBO, BEBO, Valbond, and so on [2, 18–23]. An empirical interatomic potential has been proposed by Tersoff for covalent systems. This potential has the form of the Morse pair potential, with the bond strength parameter depending upon the local environment, and is the first attempt to explain the structural chemistry of covalent systems. The Brenner potential [3] is based on the Tersoff potential, but includes correction terms that account for the overbinding of radicals and nonlocal effects. Hence, it can be applied to hydrocarbons, graphite and diamond lattices [3]. The Brenner potential can describe bond breaking, but its formalism does not include nonbonding contributions like Coulomb and van der Waals forces, which are important in predicting the structures and properties of many systems. Thus, this potential does not accurately predict the potential energy curves for hydrocarbons and graphite. Certain generalized forms of the Brenner potential include nonbonding forces, but are still not able to accurately predict the shapes of dissociation and reactive potential curves [5].

The reactive empirical bond order (REBO) potential [19] has also been developed by modifying the Tersoff potential. The initial form of the potential was only dependent on interatomic distances and did not include any dependence on molecular shape. A more advanced version has eliminated this limitation, but it still ignores long-range interactions and partial charges [24, 25]. The bond energy bond order (BEBO) [20, 21] and ValBond potentials [22, 23] also suffer from the limitation that they cannot describe the fully bonded equilibrium geometries of complex molecules [5].

The reactive force field ReaxFF appears to address all of these problems; details are available from [5]. It uses a general relationship between bond distance and bond order on the one hand, and between bond order and bond energy on the other. Valence terms, including contributions from torsion and

valence angles, are defined in terms of the same bond orders, so that all of these terms smoothly go to zero as the bonds break. ReaxFF has Coulomb and van der Waals potentials to describe nonbonding interactions. This potential was initially developed with hydrocarbons in mind, and later extended to more complex systems consisting of carbon, hydrogen, nitrogen and oxygen [26]. It requires a total of 611 parameters for CHNO systems. Best-fit values of these parameters need to be determined for a particular system of atoms.

We are interested in MD studies of the pressure- and temperature-induced decompositions of CHNO materials [27]. Hence, it is necessary to obtain the best-fit parameters of ReaxFF for representative CHNO materials. That is the topic of the present work.

In the present work, a best-fit form of the ReaxFF potential was obtained by attempting to match experimentally known molecular parameters (such as bond lengths, valence angles and torsion angles) for the molecular species of interest and for their decomposition products. This work has two novel features. The first is the application of the GA minimization technique to a complex potential (ReaxFF) with a very large number of adjustable parameters, which implies a large parameter space for optimization. The second novel feature is the use of a parameter reduction technique to determine which subset of the 611 parameters plays a significant role for the species of interest; this is necessary to reduce the optimization space to manageable levels.

In the “Computational technique” section, we describe the principles behind the parameter reduction technique and the GA algorithm. In “Results for nitromethane and its decomposition products,” we identify the relevant subset of ReaxFF parameters for a sample CHNO molecule and its decomposition products. The “Results based on the genetic algorithm” section describes the use of a genetic algorithm to determine the best-fit values for this reduced set of ReaxFF parameters. The limitations of the GA study are then summarized, and conclusions are presented.

Computational technique

The ReaxFF potential makes use of more than 600 parameters for molecules containing C, H, N and O atoms. In the present

Table 1 Numbering for the atoms in the molecular species considered

Molecular species	Atom type and number	Atom type and number	Atom type and number	Atom type and number	Atom type and number	Atom type and number	Atom type and number
CH ₃ N O ₂	C-1	H-2	H-3	H-4	N-5	O-6	O-7
CH ₃ N O	C-1	H-2	H-3	H-4	N-5	O-6	
CH ₂ O	C-1	H-2	H-3	O-4			
OH	O-1	H-2					
NO	N-1	O-2					

work, our objective is to illustrate a new method for determining the best-fit values of these parameters for a given reactive molecule and its decomposition products. In order to reduce the computational cost, we choose a simple CHNO molecule: nitromethane (CH_3NO_2) and its decomposition products. Table 1 shows the molecular species considered, and the numbering schemes for the atoms of each of these molecular species.

Overall procedure

For a given set of ReaxFF parameters, we can perform geometry optimization separately for each species based on molecular mechanics (MM). The minimum energy state yielded by MM gives the geometric properties for a given species, such as bond lengths, bond energies, and valence and torsion angles. These properties are then compared with

Table 2 Errors obtained in single parameter optimization. $T_{\text{err}}=0.193$ when single-parameter optimization was performed with 611 parameters

Name of species	Property	Atom number	Atom number	Atom number	Atom number	V_N	V_R	V_M	E_i
CH_3NO_2	Bond distance	1	2			1.109	1.109	1.095	0.0002
CH_3NO_2	Bond distance	1	3			1.108	1.1086	1.0939	0.0002
CH_3NO_2	Bond distance	1	4			1.108	1.1086	1.0939	0.0002
CH_3NO_2	Bond distance	1	5			1.545	1.5456	2.1954	0.1769
CH_3NO_2	Bond distance	5	6			1.209	1.2098	1.2649	0.0021
CH_3NO_2	Bond distance	5	7			1.209	1.2098	1.2649	0.0021
CH_3NO_2	Valence angle	3	1	2		220.0	110.3	95.8	0.0043
CH_3NO_2	Valence angle	4	1	3		218.0	109.8	96.1	0.0040
CH_3NO_2	Valence angle	5	1	2		214.0	107.8	116.4	0.0016
CH_3NO_2	Valence angle	1	5	6		238.0	119.2	153.2	0.0204
CH_3NO_2	Valence angle	1	5	7		238.0	119.2	153.3	0.0204
CH_3NO_2	Torsion angle	2	3	1	4	242.0	121.0	96.5	0.0102
CH_3NO_2	Torsion angle	3	2	1	5	238.0	119.3	131.6	0.0027
CH_3NO_2	Torsion angle	2	1	5	6	178.0	89.9	74.0	0.0079
CH_3NO_2	Torsion angle	4	1	5	7	300.0	150.8	169.2	0.0037
CH_2O	Bond distance	1	2			1.106	1.1061	1.1373	0.0008
CH_2O	Bond distance	1	3			1.106	1.1061	1.1373	0.0008
CH_2O	Bond distance	1	4			1.208	1.208	1.3046	0.0064
CH_2O	Valence angle	3	1	2		235.0	117.463	105.243	0.0027
CH_2O	Valence angle	4	1	2		242.5	121.267	127.378	0.0006
CH_2O	Valence angle	4	1	3		242.5	121.271	127.379	0.0006
CH_2O	Torsion angle	4	1	3	2	236.0	118.45	107.37	0.0022
CH_3NO	Bond distance	1	2			1.094	1.094	1.047	0.0018
CH_3NO	Bond distance	1	3			1.092	1.092	1.048	0.0016
CH_3NO	Bond distance	1	4			1.1101	1.1101	1.0493	0.0030
CH_3NO	Bond distance	1	5			1.482	1.482	1.2934	0.0162
CH_3NO	Bond distance	6	5			1.211	1.211	2.431	1.0156
CH_3NO	Valence angle	4	1	3		218.0	109.3	102.1	0.0011
CH_3NO	Valence angle	5	1	2		222.0	111.1	116.5	0.0006
CH_3NO	Valence angle	5	1	3		214.0	107.3	115.3	0.0014
CH_3NO	Valence angle	6	5	1		226.0	113.3	115.6	0.0001
CH_3NO	Valence angle	2	1	3		216.0	108.8	118.4	0.0003
CH_3NO	Torsion angle	4	1	3	2	236.0	118.45	107.37	0.0022
CH_3NO	Torsion angle	5	1	2	3	230.0	115.75	128.80	0.0032
CH_3NO	Torsion angle	6	5	1	2	244.0	122.22	136.50	0.0034
OH	Bond distance	1	2			0.9396	0.9396	1.1456	0.0481
NO	Bond distance	1	2			1.1223	1.1223	1.2296	0.0091

the experimentally known (“reference”) values. The error in each geometrical quantity is defined as

$$E_i = [(V_R - V_M)/V_N]^2, \quad (1)$$

where V_R is the reference value, V_M is the value yielded by energy minimization, and V_N is a suitable normalization factor. The reference values were taken from [28, 29] or computed using semiempirical molecular orbital calculations with the MOPAC code [30].

The first two columns of Table 2 list the molecular species and the geometrical properties that we seek to match by adjusting values of the ReaxFF parameters. A total of 37 molecular properties, including bond lengths, valence and torsion angles are taken into account when computing the deviation. The atoms involved in calculating these 37 properties are also listed in columns 3–6 of the table.

The normalization factors V_N are chosen as follows. According to [5], the allowed deviations in the bond lengths and angles were 0.01 Å and 2° respectively. Since bond lengths are of the order of 1 Å, and bond angles are typically of the order of 100°, this corresponds to allowed deviations of 1% and 2%, respectively. Hence, in Eq. 1, the normalizing factor for the bond length is taken to be the same as its reference value, and as twice its reference value for bond angles. For the bond energy, the reference value is taken as V_N . The values of V_N and V_R are given in columns 7 and 8 of Table 2.

Given the error in each geometrical quantity, we define the overall objective function as:

$$T_{\text{err}} = \left(\left[\sum_{i=1}^{N_o} E_i \right] / N_o \right)^{1/2}$$

where N_o is the total number of properties being matched and T_{err} is the function to be minimized.

Determination of significant parameters

The purpose of the present work is to illustrate a new best-fit procedure by applying it to a single reactive molecule and its decomposition products. For this restricted group of species, it is expected that only a subset of the 611 parameters would play a significant role, with the other parameters playing a relatively minor role. We first need to determine the significant parameters, and then to determine their best-fit values. This identification of significant parameters is performed in two stages, as described below.

The first stage, which involves preliminary screening, is a single-parameter search [5]. The parameters are varied one at a time, with the rest being held constant. Each parameter is allowed to take on three different values, and the best-fit value for the parameter is chosen by locating the minimum of a parabola; this is somewhat similar to Brent’s method.

We observe that there are significant changes in T_{err} during optimization with respect to certain parameters, while other parameters yield relatively small changes. The parameters that lead to significant changes in T_{err} can thus be identified.

If the set of parameters identified in this way is still too large, we resort to a second stage. This involves calculating the equivalent of the “cross-correlation” between the changes in each of these input parameters and the resulting changes in molecular properties. This second stage consists of the following steps:

1. Start with a nominal set of parameters P_{j0} .
2. Apply a set of randomly-chosen mutations to this vector, to generate, say, a set of 20,000 mutated vectors. Mutated values of each parameter are generated using the relation

$$P_j = P_{j0} \times (1 \pm A_j \times R_n),$$

where A_j is the amplitude of the perturbation in the j -th parameter and R_n is a random number lying between 0 and 1. Larger values of A_j give access to a larger search space around the nominal point. On the other hand, using very large values of A_j may lead to unphysical choices of some parameters, especially those that are used as exponents in the ReaxFF potential. Hence, the second stage should be repeated for different values of the amplitude A_j . In principle, the amplitudes A_j could be different for each parameter. However, since the amplitudes are normalized, we have opted to use one value (A_{j1}) for parameters that appear as exponents in ReaxFF, and another value (A_{j2}) for all other parameters.

3. Using a process called “mating” in GA theory, generate combinations of these 20,000 vectors, yielding a total of 40,000 final vectors. These vectors are stored as a matrix A with dimensions of 40,000 × number of parameters. Details of the mating process are explained in the “Genetic algorithm procedure.”
4. For each of these vectors, perform molecular mechanics (MM) calculations for nitromethane and its product

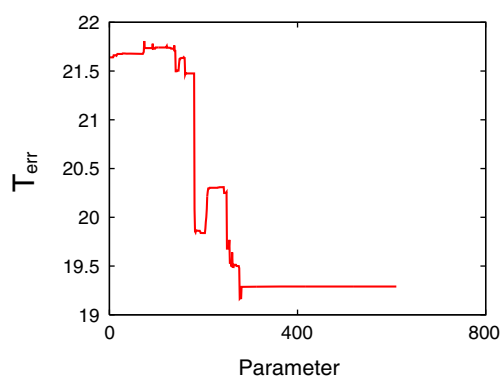


Fig. 1 Evolution of T_{err} during single-parameter optimization with respect to ReaxFF parameters. Ordinate shows $100 \times T_{\text{err}}$

species, and obtain the minimum energy state of each species. This yields the equilibrated values of the molecular properties listed in Table 2. The corresponding deviations from the reference values are stored as a matrix B with dimensions of $40,000 \times$ number of properties.

Each column in matrix A now contains 40,000 values of a particular parameter, while each column in matrix B contains the deviations in one molecular property.

- Each element a_{ij} in a column of matrix A is then normalized as follows:

$$a_{\text{normalized}} = \frac{a_{ij} - \bar{a}}{\sigma},$$

where \bar{a} is the arithmetic mean and σ is the standard deviation of the elements in that column. The same process is applied to elements in matrix B. This normalization is required because matrix A, consisting

Table 3 Errors obtained after single-parameter optimization with 190 parameters. $T_{\text{err}}=0.0513$

Name of molecule	Property	Atom type	Atom type	Atom type	Atom type	Weight	Literature/exp. value	ReaxFF computed value	Error
CH ₃ NO ₂	Bond distance	1	2			1.109	1.109	1.090	0.3×10^{-3}
CH ₃ NO ₂	Bond distance	1	3			1.108	1.1086	1.089	0.3×10^{-3}
CH ₃ NO ₂	Bond distance	1	4			1.108	1.1086	1.088	0.3×10^{-3}
CH ₃ NO ₂	Bond distance	1	5			1.545	1.5456	1.643	0.4×10^{-2}
CH ₃ NO ₂	Bond distance	5	6			1.209	1.2098	1.233	0.4×10^{-3}
CH ₃ NO ₂	Bond distance	5	7			1.209	1.2098	1.233	0.4×10^{-3}
CH ₃ NO ₂	Valence angle	3	1	2		220.0	110.3	95.2	0.47×10^{-2}
CH ₃ NO ₂	Valence angle	4	1	3		218.0	109.8	95.5	0.43×10^{-2}
CH ₃ NO ₂	Valence angle	5	1	2		214.0	107.8	120.6	0.36×10^{-2}
CH ₃ NO ₂	Valence angle	1	5	6		238.0	119.2	153.8	0.211×10^{-1}
CH ₃ NO ₂	Valence angle	1	5	7		238.0	119.2	152.4	0.193×10^{-1}
CH ₃ NO ₂	Torsion angle	2	3	1	4	242.0	121.0	95.8	0.109×10^{-1}
CH ₃ NO ₂	Torsion angle	3	2	1	5	238.0	119.3	131.9	0.28×10^{-2}
CH ₃ NO ₂	Torsion angle	2	1	5	6	178.0	89.9	93.6	0.4×10^{-3}
CH ₃ NO ₂	Torsion angle	4	1	5	7	300.0	150.8	155.9	0.3×10^{-3}
CH ₂ O	Bond distance	1	2			1.106	1.1061	1.101	0.000
CH ₂ O	Bond distance	1	3			1.106	1.1061	1.101	0.000
CH ₂ O	Bond distance	1	4			1.208	1.208	1.226	0.2×10^{-3}
CH ₂ O	Valence angle	3	1	2		235.0	117.463	113.55	0.3×10^{-3}
CH ₂ O	Valence angle	4	1	2		242.5	121.267	123.225	0.1×10^{-3}
CH ₂ O	Valence angle	4	1	3		242.5	121.271	123.226	0.1×10^{-3}
CH ₂ O	Torsion angle	4	1	3	2	360.0	179.99	180.0	0.000
CH ₃ NO	Bond distance	1	2			1.094	1.094	1.148	0.24×10^{-2}
CH ₃ NO	Bond distance	1	3			1.092	1.092	1.158	0.36×10^{-2}
CH ₃ NO	Bond distance	1	4			1.1101	1.1101	1.143	0.9×10^{-3}
CH ₃ NO	Bond distance	1	5			1.482	1.482	1.443	0.2×10^{-3}
CH ₃ NO	Bond distance	6	5			1.211	1.211	1.23	0.2×10^{-3}
CH ₃ NO	Valence angle	4	1	3		218.0	109.3	102.9	0.8×10^{-3}
CH ₃ NO	Valence angle	5	1	2		222.0	111.1	115.4	0.4×10^{-3}
CH ₃ NO	Valence angle	5	1	3		214.0	107.3	113.0	0.7×10^{-3}
CH ₃ NO	Valence angle	6	5	1		226.0	113.3	125.7	0.3×10^{-2}
CH ₃ NO	Valence angle	2	1	3		216.0	108.8	96.4	0.33×10^{-2}
CH ₃ NO	Torsion angle	4	1	3	2	236.0	118.45	108.32	0.18×10^{-2}
CH ₃ NO	Torsion angle	5	1	2	3	230.0	115.75	119.26	0.2×10^{-3}
CH ₃ NO	Torsion angle	6	5	1	2	244.0	122.22	134.98	0.27×10^{-2}
OH	Bond distance	1	2			0.9396	0.9396	0.9438	0.000
NO	Bond distance	1	2			1.1223	1.1223	1.190	0.37×10^{-2}

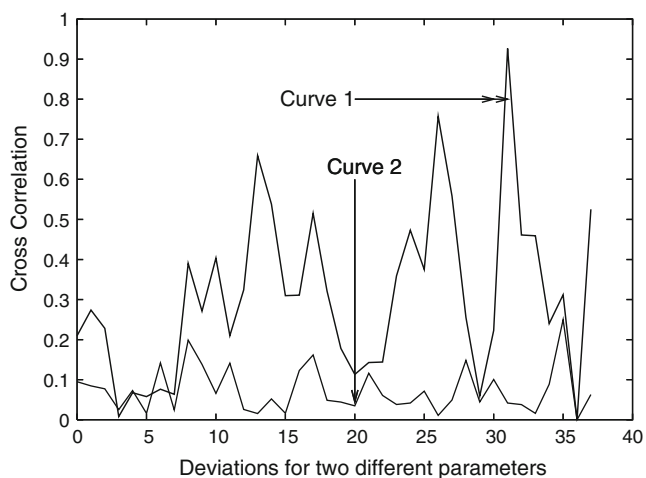


Fig. 2 Cross-correlation of the 41st (*curve 1*) and 37th (*curve 2*) parameters with each of 37 molecular deviations

of the input parameters, and matrix B , consisting of the deviations in the molecular quantities, physically represent different quantities with very different numerical values.

6. Compute matrix C , given by:

$$C = A^T B,$$

where C has the dimensions of number of input parameters \times number of deviations in output properties.

Each row in C corresponds to the cross-correlation of one parameter with each of the deviations in molecular properties. If all of the values in a row are small, it is reasonable to claim that the parameter does not significantly affect molecular properties.

This method is, in a sense, similar to calculating the cross-correlation (CC) between two random variables X and Y [31]. The CC gives an estimate of the linkage between changes in the two variables. If the two variables are only weakly coupled to each other, their cross-correlation is small. Here, instead of X and Y being scalar, we have vectors with dimensions corresponding to the number of input parameters and the number of deviations in output properties, respectively. The peak value of the CC in a given row is the quantity of interest, since a parameter must be retained in our optimization if it significantly affects even one deviation in a molecular quantity. We specify some cutoff value and select only those parameters that have a peak CC higher than the cutoff value. The cutoff must lie between 0 and 1. The lower the value, the larger the number of parameters that will be retained, and the greater the cost of subsequent optimization.

As the value approaches unity, most parameters would be rejected, which would reduce the optimization cost, but at the risk of eliminating important physics from the ReaxFF model. Cutoff values of 0.2, 0.4 and 0.6 have been examined in this study.

The above procedure yields a subset of ReaxFF parameters that significantly affect at least one of the molecular properties listed in Table 2. Following this parameter reduction, an optimization study is performed based on GA. The GA procedure is described in the next subsection.

Genetic algorithm procedure

The concept of genetic algorithms was inspired by Darwin's theory of evolution [32]. The idea is to perform natural selection for some group of parameters G which accurately describes the real system. The group of parameters are allowed to breed by mating and mutation, after which natural selection is carried out. Natural selection kills the poorest adapted species. The selected species are then allowed to breed by mating and mutation again, and so on for each genetic iteration.

GA requires a starting population consisting of a certain number of parameter vectors, such as 256 or 1024. The starting population is generated as follows. Half the required number (e.g., 128 or 512) is generated by adding random fluctuations to the reference vector, yielding an additional set of 128 (or 512) parameter vectors. This methodology is adapted from the concept of mutation in GA. The remaining half is created by adapting the concept of mating: the parameter vectors generated above are "cut" at random positions and

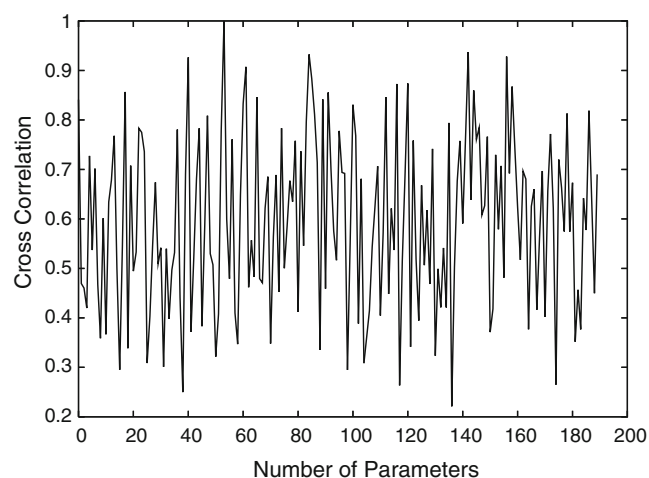
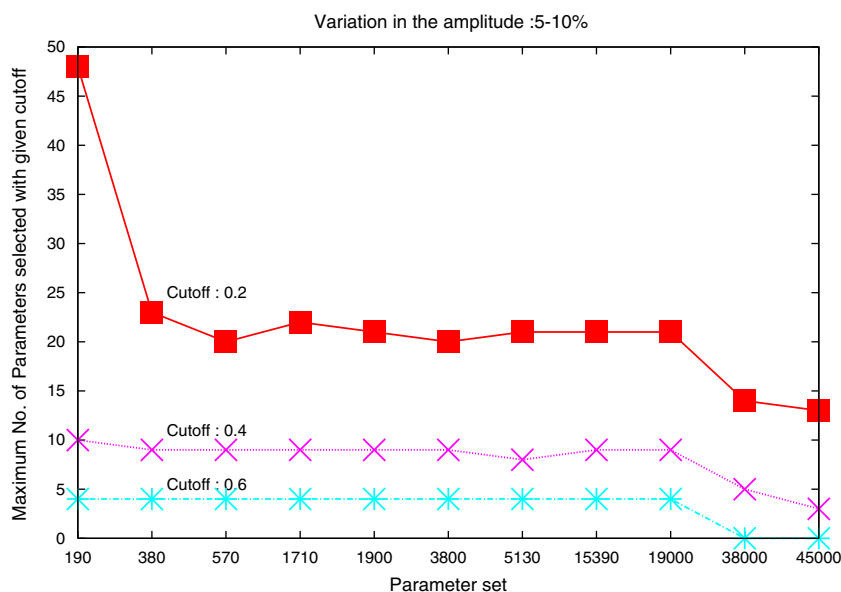


Fig. 3 Maximum cross-correlation value for each of the 190 parameters

Fig. 4 Number of parameters selected as a function of the number of vectors for the amplitude set (0.05, 0.1). Results are shown for cross-correlation cutoff levels of 0.2, 0.4 and 0.6



joined to yield a new pair of vectors. The procedure is best illustrated with an example. Let the starting pair of vectors be denoted by $a1(i)$ and $a2(i)$, $1 \leq i \leq 51$. Suppose the random position is 20. The new set of vectors $b1(i)$ and $b2(i)$ is then calculated as follows:

$$b1(i) = a1(i_1 \dots 20) + a2(i_{21} \dots 51)$$

$$b2(i) = a2(i_1 \dots 20) + a1(i_{21} \dots 51)$$

For each of the vectors generated above, corresponding to a parameter set, molecular mechanics simulations are carried out using the ReaxFF potential for each of the species listed in Table 1. The steepest descent algorithm

is used to locate the energy minimum for each species. Once the energy is minimized, we determine the bond lengths, valence angles and torsion angles for the molecules in the equilibrium configuration. This then yields the values of the deviations listed earlier. Finally, we get a single value of the function T_{err} for this parameter set.

Each GA trial thus yields a vector of T_{err} values, with each element of the vector corresponding to one parameter set. Half of the parameter vectors—those that yield the lowest values of T_{err} —are then selected for the next GA trial, with the rest being eliminated.

Fig. 5 Number of parameters selected as a function of the number of vectors for the amplitude set (0.1, 0.2). Results are shown for cross-correlation cutoff levels of 0.2, 0.4 and 0.6

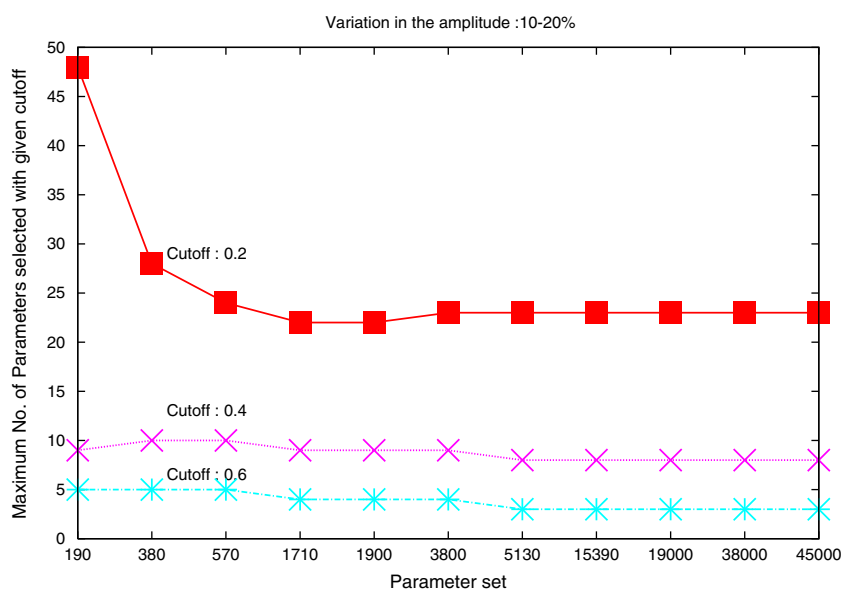


Table 4 Cutoff=0.2

No. of vectors	Amplitude variation 0.05–0.1	Amplitude variation 0.1–0.2	Amplitude variation 0.4–0.5
28	190	190	190
190	48	48	30
380	23	28	0
570	20	24	0
1710	22	22	0
1900	21	22	0
3800	20	23	0
5130	21	23	0
15390	21	23	0
19000	21	23	0
38000	14	23	0
45000	13	23	0

Note that, in the first trial, the starting point is a single reference vector from which the desired population is generated by a combination of mutation and mating, as described above. However, in successive trials, 50% of the vector population is yielded by the last step. For example, if the total desired population has a size of 256, the last trial yields 128. The remaining 50% (128 vectors) are generated as follows. We extract half (64) of the vectors yielded by the last trial—those corresponding to the lowest T_{err} values. Sixty-four vectors are then created by random mutations of these vectors, and the remaining 64 are created by randomly mating these 64. Thus, we have a total of 128 vectors obtained from the previous trial and 128 newly created vectors produced using the mutation–mating procedure.

Results for nitromethane and its decomposition products

Single-parameter search for parameter reduction

All 611 parameters of the Reaxff potential are used in this step. Figure 1 shows the variation of T_{err} during this optimization process. Columns 9 and 10 of Table 2 show the results obtained at the end of this 611-parameter optimization.

We observe that there are significant changes in T_{err} with respect to certain parameters, while other parameters yield relatively small changes. This study yields a set of 190 parameters which significantly affect T_{err} . We then repeat the single-parameter search with this reduced set of 190 parameters, yielding the final result given in Table 3. The

error falls further to 0.05. The 190 parameter values thus obtained form the starting point for the next stage of optimization: cross-correlation calculation.

Cross-correlation calculation for parameter reduction

The set of 190 significant parameters identified above still defines a rather large search space. In order to cut down on the number of input parameters even further, we need to calculate the equivalent of “cross-correlation” between each of these 190 input parameters and the 37 deviations, using the procedure defined in the “Determination of significant parameters” section.

Figure 2 shows the variation in the CC of two parameters with the 37 molecular deviations. The highest values observed are 0.92 and 0.24 for the 41st and 37th parameters, respectively. This means that the 41st parameter is highly significant while the 37th parameter is not. In the same way, we determine the highest CC for all 190 parameters, which are shown in Fig. 3. Imposing a cutoff of 0.7 reduces the number of significant parameters to 51.

As explained in the “Determination of significant parameters” section, cross-correlation results are affected by three choices:

1. Amplitudes A_{j1} and A_{j2} . In order to determine the sensitivity of the results, we have examined the sets $(A_{j1}, A_{j2}) = (0.05, 0.1)$, $(0.1, 0.2)$ and $(0.4, 0.5)$, respectively. For higher amplitudes, many combinations of parameters are likely to yield unphysical results for molecular properties. Such parameter sets are rejected altogether in this study.

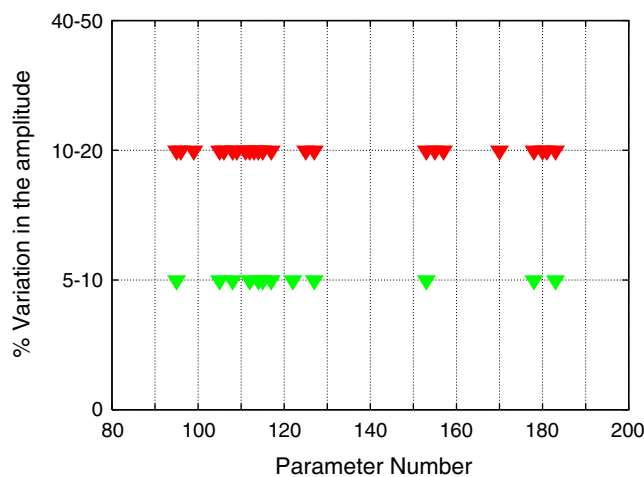


Fig. 6 Serial numbers of parameters selected for a vector size of 45,000. Cutoff level is 0.2. Results are shown for two different amplitude levels

Table 5 Parameter vector set=28

Amplitude variation	No. of significant parameters for a cutoff of 0.2	No. of significant parameters for a cutoff of 0.4	No. of significant parameters for a cutoff of 0.6	No. of significant parameters for a cutoff of 0.7
0.05–0.1	190	184	87	51
0.1–0.2	190	188	135	86
0.4–0.5	190	179	82	48

2. The cutoff used for the peak cross-correlation value: we have examined 0.2, 0.4, 0.6 and 0.7.
3. The number of vectors generated. This sensitivity is examined below.

The number of vectors should naturally be much larger than the number of parameters (190) so that we sample various combinations. The solution is to progressively increase the number of vectors until the number of parameters accepted after the cross-correlation study becomes constant. Hence, we have varied the number of vectors progressively from 28 (1/7 the number of parameters) to 45,000 (fifty times the number of parameters).

For an amplitude set of (0.05, 0.1), Figure 4 shows the maximum number of parameters selected as a function of the number of vectors. The following points may be noted:

1. As expected, the number of parameters retained is a sensitive function of the cutoff.
2. For a given cutoff, we would expect the number of parameters retained to become constant beyond a certain number of vectors. While this is true up to 19,000 vectors, there is a surprising fall in the number of parameters retained beyond that point. The probable explanation for this is that a higher number of vectors

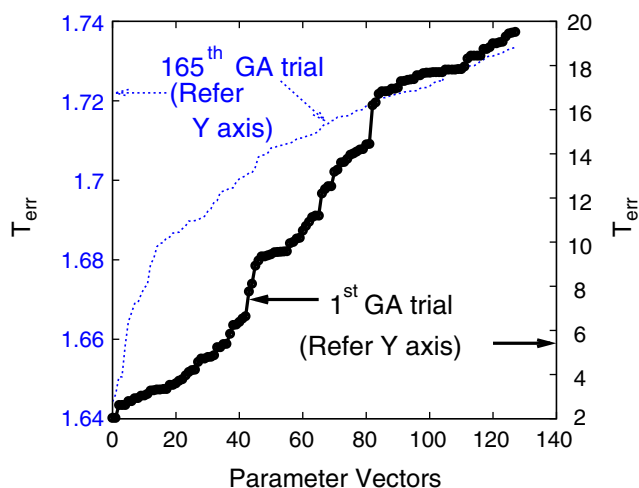


Fig. 7 T_{err} as a function of vector number after the 1st and 165th GA trials. Ordinate shows $100 \times T_{err}$

gives access to a larger number of combinations of ReaxFF parameters. This allows the significances of parameters to be determined separately. The conclusion is that a minimum number of vectors is necessary for the CC study.

Figure 5 shows the same results for an amplitude set of (0.1, 0.2). The results become constant as we increase the number of parameter sets beyond 5,000. The results between 5,000 and 19,000 are close to those in Fig. 4. However, there is no fall beyond 19,000. This is probably because, due to the higher-amplitude perturbations, we require a larger number of vectors to properly sample the search space.

The conclusion is that for a given perturbation amplitude, the number of vectors must be progressively increased to a high enough value until the result becomes constant. Also, the study must be repeated for different amplitudes to get a reasonably large search space.

Table 4 shows the number of parameters selected as a function of the number of vectors for different amplitudes and for a cutoff level of 0.2. For the case of an amplitude of 0.4–0.5, only the case with 190 vectors leads to the selection of any parameters. This is because, due to the large amplitude, most parameter sets lead to unphysical results, and are thus rejected altogether.

So far, we have seen that using a larger amplitude for a given cutoff leads to the acceptance of a slightly higher number of parameters. It is then necessary to check if this process has converged (i.e., whether the parameters selected are mostly the same for different amplitudes). Figure 6 shows the results for a fixed cutoff of 0.2 for three amplitudes. The abscissa lists the serial numbers of the parameters. We see that amplitude sets of (0.05–0.1) and (0.1–0.2) lead to the selection of essentially the same parameters, with only one exception. This shows that the process has converged.

This result shows that, after convergence, only 5–15 parameters are selected for a cutoff threshold of 0.6. At this point, it is necessary to make a choice between different options before starting GA optimization. The first is to continue with this cutoff threshold and number

of vectors; GA would then proceed with a rather small number of parameters, which is computationally attractive but may not yield a good optimum. The second is to reduce the cutoff, retaining the same number of vectors, but we might then include parameters that are known to be irrelevant. The third option is to retain a high cutoff

but reduce the number of vectors. The third option ensures that only parameters that are highly correlated are included in GA, although the vector space has not been adequately sampled.

In the present study, the purpose is to illustrate the overall technique rather than to make a judgement about the

Table 6 Deviations in molecular quantities corresponding to GA trial 165

Name of molecule	Property	Atom type	Atom type	Atom type	Atom type	Weight	Literature/exp. value	ReaxFF computed value	Square error
CH ₃ NO ₂	Bond distance	1	2			1.109	1.109	1.1055	0.9416 × 10 ⁻⁵
CH ₃ NO ₂	Bond distance	1	3			1.108	1.10856	1.10372	0.1905 × 10 ⁻⁴
CH ₃ NO ₂	Bond distance	1	4			1.108	1.10862	1.10376	0.1917 × 10 ⁻⁴
CH ₃ NO ₂	Bond distance	1	5			1.545	1.5456	1.5446	0.4066 × 10 ⁻⁶
CH ₃ NO ₂	Bond distance	5	6			1.209	1.20976	1.2224	0.1093 × 10 ⁻³
CH ₃ NO ₂	Bond distance	5	7			1.209	1.20976	1.2224	0.1095 × 10 ⁻³
CH ₃ NO ₂	Valence angle	3	1	2		220.0	110.3	110.1	0.8646 × 10 ⁻⁶
CH ₃ NO ₂	Valence angle	4	1	3		218.0	109.8	110.2	0.3124 × 10 ⁻⁵
CH ₃ NO ₂	Valence angle	5	1	2		214.0	107.8	108.1	0.1477 × 10 ⁻⁵
CH ₃ NO ₂	Valence angle	1	5	6		238.0	119.2	119.2	0.1564 × 10 ⁻⁶
CH ₃ NO ₂	Valence angle	1	5	7		238.0	119.2	119.3	0.1407 × 10 ⁻⁷
CH ₃ NO ₂	Torsion angle	2	3	1	4	242.0	121.0	121.6	0.5859 × 10 ⁻⁵
CH ₃ NO ₂	Torsion angle	3	2	1	5	238.0	119.3	119.2	0.1741 × 10 ⁻⁶
CH ₃ NO ₂	Torsion angle	2	1	5	6	178.0	89.9	89.5	0.4585 × 10 ⁻⁵
CH ₃ NO ₂	Torsion angle	4	1	5	7	300.0	150.8	151.2811	0.2572 × 10 ⁻⁵
CH ₂ O	Bond distance	1	2			1.106	1.1061	1.09007	0.2078 × 10 ⁻³
CH ₂ O	Bond distance	1	3			1.106	1.1061	1.0901	0.2063 × 10 ⁻³
CH ₂ O	Bond distance	1	4			1.208	1.208	1.2558	0.1566 × 10 ⁻²
CH ₂ O	Valence angle	3	1	2		235.0	117.463	111.034	0.7483 × 10 ⁻³
CH ₂ O	Valence angle	4	1	2		242.5	121.267	124.474	0.175 × 10 ⁻³
CH ₂ O	Valence angle	4	1	3		242.5	121.271	124.49	0.176 × 10 ⁻³
CH ₂ O	Torsion angle	4	1	3	2	360.0	179.99	180.0	0.7714 × 10 ⁻⁹
CH ₃ NO	Bond distance	1	2			1.094	1.094	1.1046	0.9395 × 10 ⁻⁴
CH ₃ NO	Bond distance	1	3			1.092	1.092	1.095	0.9525 × 10 ⁻⁵
CH ₃ NO	Bond distance	1	4			1.1101	1.1101	1.0853	0.4988 × 10 ⁻³
CH ₃ NO	Bond distance	1	5			1.482	1.482	1.487	0.9668 × 10 ⁻⁵
CH ₃ NO	Bond distance	6	5			1.211	1.211	1.214	0.7618 × 10 ⁻⁵
CH ₃ NO	Valence angle	4	1	3		218.6	109.3	108.3	0.2128 × 10 ⁻⁴
CH ₃ NO	Valence angle	5	1	2		222.0	111.1	109.9	0.2558 × 10 ⁻⁴
CH ₃ NO	Valence angle	5	1	3		214.0	107.3	113.0	0.7 × 10 ⁻³
CH ₃ NO	Valence angle	6	5	1		226.0	113.3	117.7	0.38107 × 10 ⁻³
CH ₃ NO	Valence angle	2	1	3		216.0	108.8	105.962	0.1726 × 10 ⁻³
CH ₃ NO	Torsion angle	4	1	3	2	236.0	118.45	115.34	0.1734 × 10 ⁻³
CH ₃ NO	Torsion angle	5	1	2	3	230.0	115.75	118.67	0.1612 × 10 ⁻³
CH ₃ NO	Torsion angle	6	5	1	2	244.0	122.22	139.58	0.50626 × 10 ⁻²
OH	Bond distance	1	2			0.9396	0.9396	0.9369	0.2572 × 10 ⁻⁵
NO	Bond distance	1	2			1.1223	1.1223	1.1223	0.3182 × 10 ⁻⁹

$T_{\text{err}}=0.0164$

best choice for CC calculation. Hence, we have chosen to use the results for the smallest-sized vector set (28), but with a high cutoff of 0.7. Table 5 shows the CC results. A cutoff of 0.7 yields 51 useful parameters for an amplitude set of 0.05–0.1.

This parameter set is used in the next stage: GA optimization.

Results based on the genetic algorithm

The GA process is now started with the reference vector yielded by single-parameter optimization, corresponding to $T_{\text{err}}=0.0513$, as shown in Table 3. The optimization varies only the 51 significant parameters determined by the CC study.

We have performed a large number of GA trials following the procedure given in the “Genetic algorithm procedure” section, using a population of 256 vectors. For the 1st and 165th GA trials, Fig. 7 shows the lowest 128 T_{err} values as a function of the vector number. The minimum T_{err} value yielded by the first trial is 0.02, and this goes down to 0.0164 after the 165th trial. The individual deviations in molecular properties yielded by the 165th trial are shown in Table 6.

The best vector obtained from 165 trials with a population size of 256 was then used to initiate a new GA study, with a population size of 1024. Results are shown in Fig. 8. After the completion of 74 and 134 GA trials, we obtain minimum T_{err} values of 0.01578 and 0.01556, respectively, indicating that the optimization has approached its best result.

The best result of 0.015 is obtained after 1250 GA trials. Table 7 presents the deviations at this point. The individual

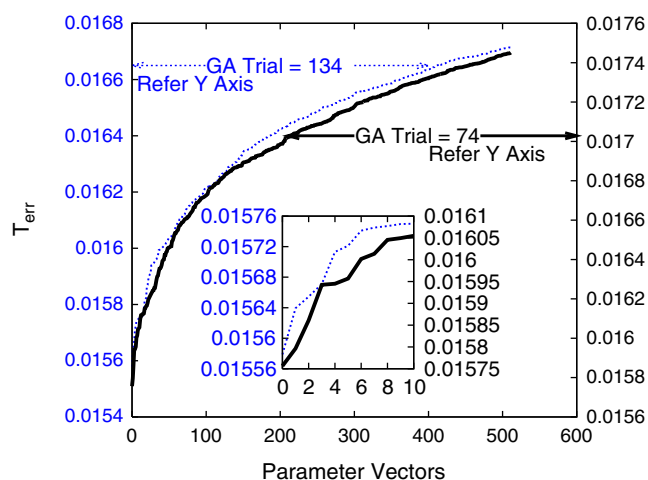


Fig. 8 T_{err} as a function of vector number after the 74th and 134th GA trials

deviations are generally below acceptable limits, except for a torsion angle in CH_3NO .

Limitations of this study

1. The literature on genetic algorithms describes a number of ways of performing mutation and mating operations [7, 17, 33]. Only one of these schemes has been applied in this work. A more detailed study, trialing different schemes, could yield better results.
2. The cross-correlation study yields a reduced parameter set. The number of parameters in that set depends upon the amplitudes, cutoff level and the number of vectors chosen. The final optimization results should be determined for different choices in the CC process, in order to increase the probability of finding a “global” minimum.
3. Chemical reactions such as molecular decomposition involve transition states far from equilibrium. Therefore, a reduced parameter set optimized for fits to equilibrium molecular structures may not necessarily represent reactions properly. We plan to extend this methodology to match sections of the potential energy surface corresponding to distorted/extended states of these species.
4. The genetic algorithm converges slowly. The RMS error reduces from 0.02 to 0.0164 in the 165th trial, and later on reduces to 0.015 after 1250 trials. One explanation for this slow rate of convergence is that the present GA algorithm retains the vectors yielding the lowest values of the objective function. This means that the vectors get progressively closer to a single minimum, reducing the rate of convergence and also restricting the result to a single minimum which may turn out to be a local minimum. There are alternate GA schemes where vectors yielding higher T_{err} values are also retained, although with a probability that is inversely related to the value of T_{err} [33]. That technique is more likely to locate the global minimum.

Conclusions

We have obtained a best-fit form of the ReaxFF potential for nitromethane and its decomposition products by matching experimentally known molecular parameters, such as bond lengths, valence angles and torsion angles. This work has two novel features. The first is the use of a parameter reduction technique to determine which subset of the 611 ReaxFF parameters plays a significant

Table 7 Results corresponding to GA trial 1250

Name of molecule	Property	Atom type	Atom type	Atom type	Atom type	Weight	Literature/exp. value	ReaxFF computed value	Error
CH ₃ NO ₂	Bond distance	1	2			1.109	1.109	1.1086	0.9921×10^{-7}
CH ₃ NO ₂	Bond distance	1	3			1.108	1.10856	1.106	0.5037×10^{-5}
CH ₃ NO ₂	Bond distance	1	4			1.108	1.10862	1.1063	0.4153×10^{-5}
CH ₃ NO ₂	Bond distance	1	5			1.545	1.5456	1.538	0.224×10^{-4}
CH ₃ NO ₂	Bond distance	5	6			1.209	1.20976	1.214	0.1385×10^{-4}
CH ₃ NO ₂	Bond distance	5	7			1.209	1.20976	1.214 0	0.12335×10^{-4}
CH ₃ NO ₂	Valence angle	3	1	2		220.0	110.3	110.042	0.13647×10^{-5}
CH ₃ NO ₂	Valence angle	4	1	3		218.0	109.8	110.0347	0.11594×10^{-5}
CH ₃ NO ₂	Valence angle	5	1	2		214.0	107.8	108.0077	0.94234×10^{-6}
CH ₃ NO ₂	Valence angle	1	5	6		238.0	119.2	119.2	0.2259×10^{-8}
CH ₃ NO ₂	Valence angle	1	5	7		238.0	119.2	119.2	0.8956×10^{-7}
CH ₃ NO ₂	Torsion angle	2	3	1	4	242.0	121.0	121.3	0.22617×10^{-5}
CH ₃ NO ₂	Torsion angle	3	2	1	5	238	0 119.3	119.3	0.434×10^{-7}
CH ₃ NO ₂	Torsion angle	2	1	5	6	178.0	89.9	89.7	0.18777×10^{-5}
CH ₃ NO ₂	Torsion angle	4	1	5	7	300.0	150.8	151.11	0.1346×10^{-5}
CH ₂ O	Bond distance	1	2			1.106	1.1061	1.0961	0.82295×10^{-4}
CH ₂ O	Bond distance	1	3	1		106	1.1061	1.0961	0.81514×10^{-4}
CH ₂ O	Bond distance	1	4			1.208	1.208	1.184	0.40849×10^{-3}
CH ₂ O	Valence angle	3	1	2		235.0	117.463	110.58	0.85778×10^{-3}
CH ₂ O	Valence angle	4	1	2		242.5	121.267	124.7	0.2004×10^{-3}
CH ₂ O	Valence angle	4	1	3		242.5	121.271	124.719	0.2021×10^{-3}
CH ₂ O	Torsion angle	4	1	3	2	360.0	179.99	180.0	0.7714×10^{-9}
CH ₃ NO	Bond distance	1	2	1		1.094	1.094	1.102	0.5723×10^{-4}
CH ₃ NO	Bond distance	1	3			1.092	1.092	1.0954	0.9851×10^{-5}
CH ₃ NO	Bond distance	1	4			1.1101	1.1101	1.0803	0.72236×10^{-3}
CH ₃ NO	Bond distance	1	5			1.482	1.482	1.485	0.50418×10^{-5}
CH ₃ NO	Bond distance	6	5			1.211	1.211	1.203	0.48259×10^{-4}
CH ₃ NO	Valence angle	4	1	3		218.6	109.3	109.2	0.166302×10^{-4}
CH ₃ NO	Valence angle	5	1	2		222.0	111.1	109.2	0.77436×10^{-4}
CH ₃ NO	Valence angle	5	1	3		214	0 107	3 108	$7 0.40815 \times 10^{-4}$
CH ₃ NO	Valence angle	6	5			1 226.0	113.3	118.1	0.45154×10^{-3}
CH ₃ NO	Valence angle	2	1	3		216.0	108.8	109.2	0.3112×10^{-5}
CH ₃ NO	Torsion angle	4	1	3	2	236.0	118.45	115.34	0.1734×10^{-3}
CH ₃ NO	Torsion angle	5	1	2	3	230.0	115.75	118.67	0.1612×10^{-3}
CH ₃ NO	Torsion angle	6	5	1	2	244.0	122.22	139.58	0.50626×10^{-2}
OH	Bond distance	1	2			0.9396	0.9396	0.937	0.763739×10^{-5}
NO	Bond distance	1	2			1.1223	1.1223	1.1223	0.2717×10^{-9}

$T_{\text{err}} = 0.015$

role for the species of interest; this is necessary to reduce the optimization space to manageable levels. The second is the application of GA techniques to a complex potential (ReaxFF) with a very large number of adjustable parameters, which implies a large parameter space for optimization.

Using a subset of 51 ReaxFF parameters, we have obtained a reasonably good match to 37 molecular properties for nitromethane and its decomposition products, with a root mean square deviation of 1.5%. It is expected that the use of more sophisticated GA algorithms would yield an even better match to reference data.

Acknowledgment It is a pleasure to acknowledge the help given by Dr. A.C.T. van Duin in providing the Reaxff MD code and also helping with the use of that code.

References

1. Rapaport DC (1995) The art of molecular dynamics simulation. Cambridge University Press, Cambridge
2. Tersoff J (1986) Phys Rev Lett 56:632
3. Brenner DW (1990) Phys Rev B 42:9458
4. Stuart SJ, Tutein B, Harrison JA (2000) J Chem Phys 112:6472
5. van Duin A, Dasgupta S, Lorant F, Goddard WA (2001) J Phys Chem A 105:9396
6. Smeyers YG, Bellido MN (2004) Int J Quant Chem 23:507
7. Makarov DE, Metiu H (1998) J Chem Phys 108:590
8. Melandri S, Favero G, Caminati W, Favero B, Esposti AD (1997) J Chem Soc Faraday Trans 93:2131
9. Graham AP, Hofmann F, Toennies JP, Chen LY, Ying SC (1997) Phys Rev Lett 78:3900
10. Carlson AF, Madix RJ (2000) Surf Sci 470:62
11. Kryachko ES, Lwdin O, Brndas E (2004) Fundamental world of quantum chemistry: a tribute to the memory of Per-Olov Lowdin, vol II. Kluwer, Dordrecht
12. Hutson JM, Ernesti A, Law MM, Roche CF, Wheatley RJ (1996) J Chem Phys 105:9130
13. Atkins KM, Hutson JM (1996) J Chem Phys 105:440
14. Prudente FV, Acioli H, Neto JJS (1998) J Phys Chem 109:8801
15. Saad D, Rattray M (1997) Phys Rev Lett 79:2578
16. Thompson DL, Wagner AF, Minkoff M (2006) J Phys Conf Ser 46:234
17. Xu YG, Liu GR (2003) J Micromech Microeng 13:254
18. Tersoff J (1988) Phys Rev Lett 61:2879
19. Haskins PJ, Cook M, Fellows J, Wood A (1998) In: Proc 11th Int Symp on Detonation, Aspen, CO, USA, 30 Aug–4 Sept 1998, p 897
20. Johnston HS, Parr C (1963) J Am Chem Soc 85:2544
21. Johnston HS (1963) J Am Chem Soc 85:2544
22. Root DM, Landis CM (1993) J Am Chem Soc 115:4201
23. Cleveland T, Landis CM (1996) J Am Chem Soc 118:6020
24. Brenner DW, Shendrova OA, Harrison A, Stuart SJ, Boris N, Sinnott SB (2002) J Phys Condens Matter 14:783
25. Boris N, Lee KH, Sinnott SB (2004) J Phys Condens Matter 16:7261
26. Strachan A, van Duin A, Chakraborty D, Dasgupta S, Goddard WA III (2003) Phys Rev Lett 91:098301
27. Strachen A, Kober EM, van Duin A, Oxaggard J, Goddard WA III (2005) J Chem Phys 122:054502
28. University of Waterloo Webpage (2011) <http://www.science.uwaterloo.ca/~cchieh/cact/c120/bondel.html>
29. DOlgov EY, Batev VA, Godunov IA (2004) Int J Quant Chem 96:193
30. SoftWare (2011) SoftWare: 64-bit operating system (webpage). <http://www.cachesoftware.com/mopac/index.shtml>
31. Croxton FE, Cowden DJ, Klein S (1939) Applied general statistic. Prentice Hall Inc, New York
32. Holland JH (1975) Adaptation in natural and artificial systems. The University of Michigan Press, Ann Arbor
33. Deaven DM, Ho KM (1995) Phys Rev Lett 75:288

Virtual screening for potential inhibitors of bacterial MurC and MurD ligases

Tihomir Tomašič · Andreja Kovač · Gerhard Klebe ·
Didier Blanot · Stanislav Gobec · Danijel Kikelj ·
Lucija Peterlin Mašič

Received: 6 May 2011 / Accepted: 25 May 2011 / Published online: 12 June 2011
© Springer-Verlag 2011

Abstract Mur ligases are bacterial enzymes involved in the cytoplasmic steps of peptidoglycan biosynthesis and are viable targets for antibacterial drug discovery. We have performed virtual screening for potential ATP-competitive inhibitors targeting MurC and MurD ligases, using a protocol of consecutive hierarchical filters. Selected compounds were evaluated for inhibition of MurC and MurD ligases, and weak inhibitors possessing dual inhibitory activity have been identified. These compounds represent new scaffolds for further optimisation towards multiple Mur ligase inhibitors with improved inhibitory potency.

Keywords ATP · Inhibitor · Multiple ligand · Mur ligase · Pharmacophore · Virtual screening

Introduction

The increasing emergence of Gram-positive and Gram-negative bacterial strains resistant to most of the currently available antibiotics has compromised the treatment of

bacterial infections and led to increased morbidity and mortality worldwide. In a quest for new antibacterial drugs for combating bacterial drug-resistance, the biochemical machinery involved in peptidoglycan biosynthesis remains a viable source of unexploited targets [1]. Peptidoglycan is an essential cell-wall polymer unique to prokaryotic cells that preserves cell integrity by withstanding high internal osmotic pressure and maintaining a defined cell shape [2, 3]. The biosynthesis of peptidoglycan is a multi-step process comprising intracellular assembly of the UDP-MurNAc-pentapeptide, which is subsequently translocated through the cytoplasmic membrane and incorporated into the growing peptidoglycan layer. ATP-dependent Mur ligases (MurC to MurF) catalyze a series of reactions leading to UDP-MurNAc-pentapeptide by sequentially adding L-Ala (MurC), D-Glu (MurD), a diamino acid which is generally L-Lys in Gram-positive or *meso*-diaminopimelic acid in Gram-negative bacteria (MurE) and D-Ala-D-Ala (MurF), to the starting MurC substrate UDP-MurNAc [4]. Mur ligases are essential for the survival of bacteria, which makes the discovery of their inhibitors an important challenge [4, 5].

Mur ligases catalyze the formation of a peptide or amide bond between the carboxyl group of the UDP-substrate and the amino group of the condensing amino acid. They operate by similar chemical mechanisms [6, 7] (Fig. 1) and, as shown for MurC and MurF, an ordered kinetic mechanism [8, 9]. The initial step of the enzymatic reaction is the binding of ATP and the corresponding UDP-substrate to the free enzyme, which is followed by ATP-promoted activation of the carboxyl group of the UDP-substrate. The generated acylphosphate intermediate is then attacked by the amino group of the incoming amino acid or dipeptide. The resulting tetrahedral intermediate breaks down with elimination of inorganic phosphate and concomitant formation of a peptide or amide bond (Fig. 1).

T. Tomašič · A. Kovač · S. Gobec · D. Kikelj · L. P. Mašič (✉)
Faculty of Pharmacy, University of Ljubljana,
Aškerčeva 7,
1000 Ljubljana, Slovenia
e-mail: lucija.peterlin@ffa.uni-lj.si

G. Klebe
Institut für Pharmazeutische Chemie,
Philipps Universität Marburg,
Marbacher Weg 6,
35032 Marburg, Germany

D. Blanot
Enveloppes Bactériennes et Antibiotiques, IBBMC,
UMR 8619 CNRS, Univ Paris-Sud,
91405 Orsay, France

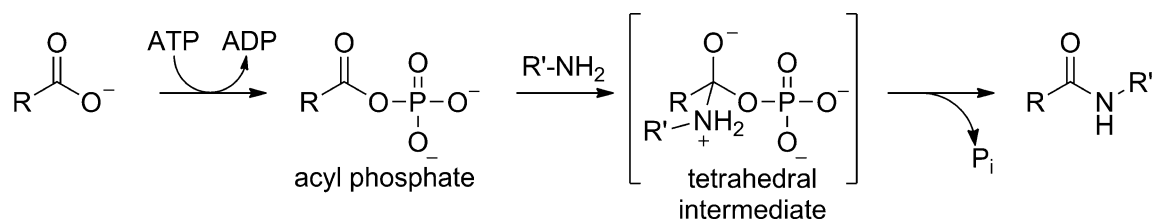


Fig. 1 Catalytic mechanism of Mur ligases

It is now widely recognized that multiple ligands – compounds designed to bind to more than one target – can be more beneficial therapeutically than highly target-specific ligands. Design of the multiple ligands is usually a demanding challenge, with the need to appropriately balance affinities for different targets while preserving their drug-like properties [10]. It was shown, using network models of antibacterial drugs, that multitarget attacks perturb complex systems more effectively than focused attacks, even if the number of targeted interactions is the same [11]. Targeting multiple bacterial enzymes that are structurally related or recognize common structural motifs with designed multiple ligands could not only lead to potent antibacterial activity but also reduce the frequency of target-mediated resistance to such a compound, since mutations conferring resistance would have to occur in at least two different target genes in a single generation [12]. Since Mur ligases share the same catalytic mechanism and possess several conserved amino acid residues in their active sites, it should be possible to design multiple ligands inhibiting more than one Mur ligase. Multiple inhibition of Mur ligases could be achieved, either by mimicking the UDP-MurNAc part of the UDP-substrates or by targeting the ATP-binding site of the enzymes. We have recently reported trihydroxybenzylidenethiazolidin-4-ones as multiple Mur ligase inhibitors and demonstrated by high-resolution NMR experiments that they interact mainly with the residues flanking the UDP-MurNAc-L-Ala-binding site of MurD ligase [13].

ATP-binding pockets of Mur ligases, which appear to be highly conserved throughout the enzyme family [14–16], are another possible target for the discovery of multiple ligands. However, targeting the ATP-binding site of bacterial enzymes is associated with several problems. First, an ATP-competitive inhibitor must be able to compete with the high ATP concentration in the bacterial cell (0.6 – 18 mM) [17]. This ATP concentration is similar to that in human cells (1 – 10 mM) [18] and it was shown successfully with protein kinases that it is possible to design competitive and selective inhibitors of the ATP-binding site. Secondly, inhibitor binding to the ATP-binding site must be selective for the targeted bacterial enzyme over human ATP-dependent enzymes, particularly kinases. Nevertheless, recent successful examples of ATP-

competitive bacterial enzyme inhibitors possessing antibacterial activity and displaying good selectivity profiles with respect to human enzymes show that these challenges can be overcome [19].

We studied the ATP-binding site of MurD ligase (PDB entry: 3UAG) [6] using ProBiS [20, 21], a Web server for detecting protein binding sites based on local structural alignments, and found that the MurD ATP-binding site is very similar to those of the other members of the three substrate amide ligase superfamily (MurC, MurE and MurF) [15] and that it is not closely related to those of ATP-utilizing human enzymes [19]. These results make the ATP-binding site of Mur ligases a promising target for the design of multiple ligands not interacting with human ATP-binding enzymes.

Ligand- or structure-based virtual screening (VS), which involves computational analysis of large libraries of compounds and subsequent selection of a smaller subset for biological testing, is an alternative to experimental high-throughput screening (HTS) [22]. VS has already led to the discovery of Mur ligase inhibitors designed to target the UDP-substrate-binding site [23, 24]. In the present paper, we describe a VS campaign aimed to discover the first ATP-competitive multiple Mur ligase inhibitors. This campaign has identified some compounds showing weak MurC and MurD inhibitory activity which feature new scaffolds for the design of ATP-competitive inhibitors of Mur ligases.

Methods

Sequence alignment

Amino acid sequences of all four Mur ligases from *E. coli* were retrieved from the UniProt archive (UniProt accession numbers: P17952 for MurC, P14900 for MurD, P22188 for MurE and P11880 for MurF) [25]. The amino acid sequences were aligned using the Align Multiple Sequences protocol with slow pair-wise sequence alignment available in Accelrys Discovery Studio 2.5 [26] running on a workstation with Intel Core i7 860 CPU processor, 8 GB RAM, two 750 GB hard drives and an Nvidia GT220 GPU graphic card, running Centos 5.5.

Virtual screening

All structural handling was performed using SYBYL [27] running on a Silicon Graphics O2 (R5000) workstation. As database engine to perform all initial searches, the UNITY [28] system of Tripos was used. For all ligands considered, a 3D structure has been generated from the 2D chemical formula with the program CORINA [29]. Protonation states have been assumed in the standard setting as suggested by CORINA. Hot spot analysis was performed using GRID [30] and graphical display was achieved by Pymol [31]. Definition of a search pharmacophore was accomplished through the facilities implemented in UNITY. Parallel docking was done with FlexX [32, 33] under the FlexX-Pharm pharmacophore type constraints running on a cluster of Linux PCs. Final docking was done with GOLD [34].

Colorimetric enzyme inhibition assay

The target compounds were tested for their ability to inhibit the addition of L-Ala (d-Glu) to UDP-MurNAc (UDP-MurNAc-L-Ala) catalyzed by MurC (MurD) from *Escherichia coli* [35, 36]. Detection of the orthophosphate generated during the reaction was based on the colorimetric Malachite green method, as described [37], with slight modifications. A mixture with a final volume of 50 μ L contained 50 mM Hepes, pH 8.0, 5 mM $MgCl_2$, 10 mM $(NH_4)_2SO_4$, 0.01% Triton X-100, 120 μ M UDP-MurNAc (80 μ M UDP-MurNAc-L-Ala), 120 μ M L-Ala (100 μ M D-Glu), 450 μ M ATP (400 μ M ATP), purified MurC (MurD) from *E. coli* (diluted with 20 mM Hepes, pH 7.2, 1 mM dithiothreitol), and 500 μ M or 250 μ M of the tested compound dissolved in DMSO. The final concentration of DMSO was 5% (v/v). The reaction mixture was incubated at 37 $^{\circ}C$ for 15 min, then quenched with 100 μ L Biomol[®] reagent. Absorbance at 650 nm was measured after 5 min. Residual activity (RA) was calculated relative to control assays without the compounds and with DMSO. Results are presented as % inhibition, calculated as 100% - RA (Table 1).

Table 1 MurC and MurD inhibitory potencies of inhibitors discovered by VS

Compd.	c [μ M]	MurC % inhibition	MurD % inhibition
1	500	26	50
2	250	7	29
3	250	32	11
4	500	30	33
5	500	36	30

Results and discussion

In our ongoing efforts to discover inhibitors of Mur ligases, we have already designed several series of inhibitors targeting the UDP-substrate-binding site of MurD [13, 38, 39]. An alternative strategy to inhibit Mur ligases would be the rational design of compounds blocking the ATP-binding site of the enzymes. To this end, we decided to perform virtual screening to discover multiple Mur ligase inhibitors binding to the ATP-binding site of two or more Mur enzymes. First, sequence alignment of MurC-MurF from *E. coli* confirmed the existence of several conserved amino acid residues, particularly in the ATP-binding site [14–16] (Fig. 2a). Next, the superimposed crystal structures of *Haemophilus influenzae* MurC (PDB entry: 1P3D) [40] and *E. coli* MurD (PDB entry: 3UAG) [6], co-crystallized with adenosine 5'-(β,γ -imido)triphosphate (AMPPNP), a non-hydrolysable ATP analogue, and ADP, respectively, revealed almost identical conformations of the AMPPNP and ADP and similar interactions with the ATP-binding site residues (Fig. 2b). Superposition of the crystal structures of MurC [41], MurD [6], MurE [42] and MurF [43] ligases from *E. coli* and visual inspection of their ATP-binding sites, especially with respect to amino acid residues interacting with the ATP molecule in the *E. coli* MurD, revealed that these common features for the binding of the ATP molecule (Fig. 2b) are also characteristic for the *E. coli* MurC, MurE and MurF. In detail, the adenine ring atoms N-6 and N-7 of ATP form two hydrogen bonds with the side chain carboxamide group of an asparagine residue (Asn295 in *H. influenzae* MurC, Asn271 in *E. coli* MurD), which anchor the adenine moiety in its central domain pocket. According to the sequence alignment, the same hydrogen bonds are also formed in the case of *E. coli* MurE and MurF, which indicates that the interaction of the Asn residue with the N-6 and N-7 of the adenine ring is highly conserved in Mur ligases. Furthermore, the hydroxyl groups of the ribose moiety are in contact with Asp or Glu residues (Glu352 in *H. influenzae* MurC, Asp317 in *E. coli* MurD), while the ATP phosphate groups interact with P-loop residues comprising Gly-Lys-Thr/Ser-Thr. These observations led to a good prospect for the design of multiple ATP-binding site targeting inhibitors of Mur ligases.

First, we systematically analyzed the ATP-binding pocket with the following probe functional groups using GRID [30], to analyze those areas of the MurC and MurD ATP-binding sites, where a putative ligand functional group can favorably interact with the active site residues: (i) an amide NH group as a typical hydrogen bond donor group, (ii) carbonyl as a hydrogen bond acceptor group and (iii) DRY probe to describe hydrophobic interactions. The results of this analysis for MurD (Fig. 3) show the binding

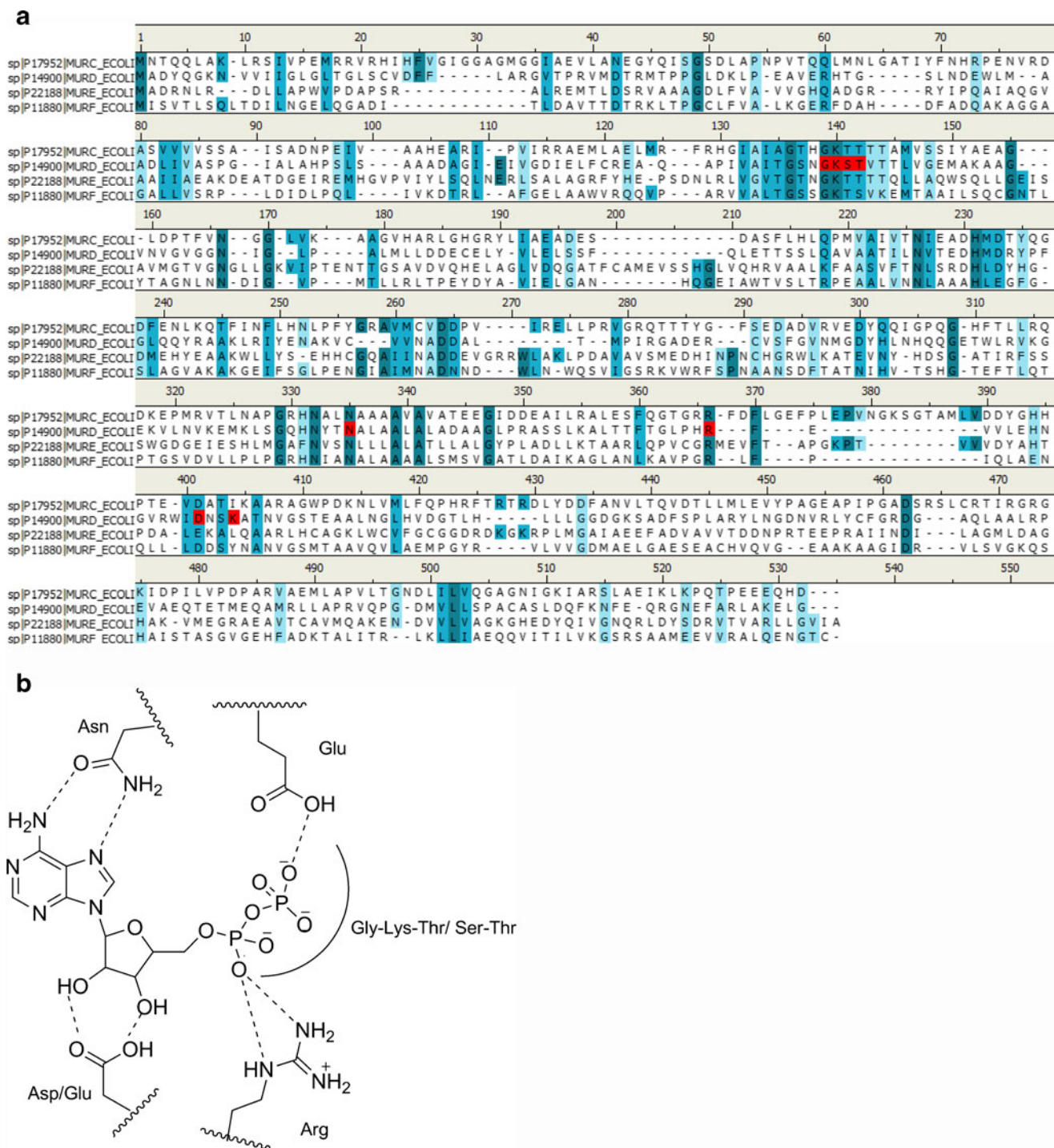


Fig. 2 (a) Sequence alignment of MurC–MurF from *E. coli*. Residues marked red form hydrogen bonds with ADP according to the crystal structure of MurD–ADP complex (PDB entry: 3UAG). (b) Common

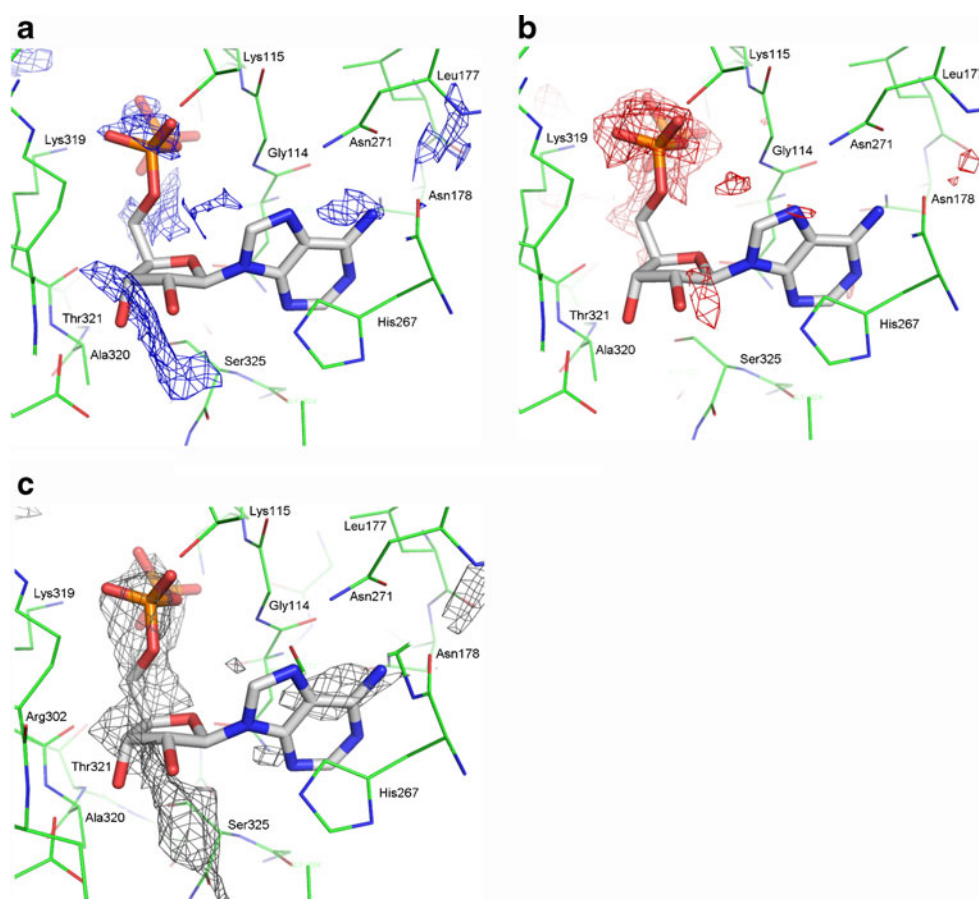
amino acid residues involved in the binding of ADP in the binding sites of MurC, MurD, MurE and MurF

mode of ADP, together with contour maps for hydrogen bond donor and acceptor groups and hydrophobic moieties. The diphosphate-binding pocket is a favorable region for binding both the hydrogen bond acceptor and hydrogen bond donor groups. The ribose-binding pocket particularly favors hydrogen bond donor groups of the ligand, while the

adenine-binding pocket can accommodate both the hydrogen bond donor groups in the area occupied by the N-6 amino group and the hydrogen bond acceptor groups in the area around the N-7 atom.

Considering similar ATP conformation and interaction pattern in the active sites of Mur ligases (Fig. 2b), together

Fig. 3 Hot spots of binding in the MurD ATP-binding site. GRID maps using (a) amide NH as a hydrogen bond donor probe (in blue); (b) carbonyl group as a hydrogen bond acceptor probe (in red); (c) hydrophobic DRY probe (in grey)



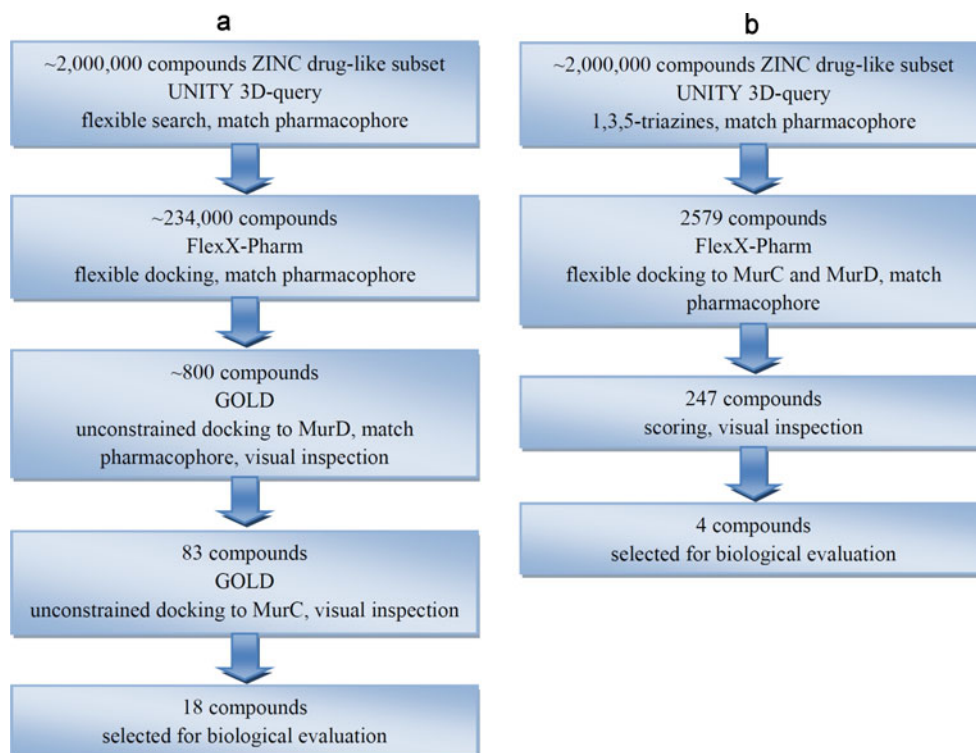
with similar results from the GRID analyses of MurC and MurD ATP-binding sites, the crystal structure of *E. coli* MurD was selected as a target protein for VS in which we applied a protocol of consecutive hierarchical filters that were selected to find compounds binding to the ATP-binding pocket of MurC and MurD ligases (Fig. 4). The initial database contained a subset of about 2,000,000 compounds from the ZINC drug-like database [44], pre-filtered by Lipinski's rule of five [45]. In the first step, the database was reduced to approximately 234,000 molecules (Fig. 4a) by selecting compounds matching a 3D pharmacophore based on GRID analysis of the MurD ATP-binding site and on the binding conformation of ADP in the MurD active site (Fig. 5). In detail, the database was filtered using a 3D flexible search, as available in UNITY, for compounds containing one aromatic feature (corresponding to the adenine ring), one hydrogen-bond acceptor (corresponding to the adenine ring N-7 amino group) and two hydrogen-bond donors (corresponding to the adenine N-6 group and the ribose 3'-hydroxyl group) with their positions defined in space as in the ADP-bound conformation in MurD active site, but with up to 2.0 Å tolerance.

In the following selection step, the remaining putative ligands were docked flexibly into the binding site of MurD considering the pharmacophore-type constraints, using

FlexX. The pharmacophore was again defined based on the GRID analysis (Fig. 3), using the FlexX-Pharm module and following pharmacophore constraints: (i) formation of two hydrogen bonds with the carboxamide side chain group of Asn271 (selecting ligands containing one hydrogen bond acceptor and one hydrogen bond donor), and (ii) formation of one of the two defined hydrogen bonds with the side chain carboxylate group of Asp317. Docking conformations of about 12,600 compounds matched this pharmacophore constraint.

The database of putative binders was further reduced to 800 by employing additional pharmacophore-type constraints imposed by GRID analysis of the active site: (i) aromatic/hydrophobic moiety in the adenine-binding pocket, and (ii) formation of hydrogen bonds with amino acid residues interacting with phosphate groups of ADP, namely Gly114, Lys115, Ser116 and Thr117. The remaining 800 molecules were re-docked into the MurD ATP-binding site using GOLD, a genetic algorithm-based docking program, without applying a predefined pharmacophore, to study whether the FlexX pharmacophore-guided binding conformation could be reproduced. Docking solutions were ranked by the scoring function GOLD-score [46] scoring function and only the top five ranked docking solutions of each compound were considered

Fig. 4 The protocols of consecutive filters applied in virtual screening



further in the validation step. This step was used to assess whether the best-ranked GOLD-calculated poses accommodate the candidate molecules in a way such that the previously defined pharmacophore hypothesis could be satisfied. Again, a reduced UNITY pharmacophore comprising an aromatic moiety with hydrogen bond acceptor

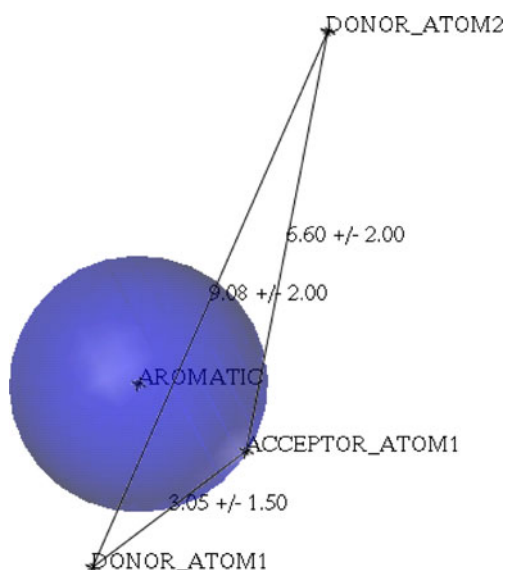


Fig. 5 UNITY 3D pharmacophore based on the binding conformation of ADP in the active site of *E. coli* MurD (PDB entry: 3UAG)

and donor groups interacting with the Asn271 side chain were considered. This search resulted in 83 candidate molecules which, in the next step, were docked into the MurC ATP-binding pocket by GOLD, since our goal was identification of inhibitors targeting MurC and MurD. The calculated conformations of the putative dual inhibitors in both, the MurC and MurD ATP-binding sites, were scored with DrugScore [47] and then inspected visually. The selection of candidates for biological testing was based on their DrugScore predicted binding affinity and, warranted by our multi-target approach, on the similarity of their binding modes in the active sites of MurC and MurD. Following this protocol, 18 compounds (structures not shown) were selected for experimental evaluation of their inhibition of MurC and MurD ligases from *E. coli*.

Inhibitory activity of the tested compounds was monitored using the colorimetric Malachite green assay for detection of orthophosphate generated during the enzymatic reaction. To exclude possible non-specific (promiscuous) inhibition due to aggregate formation, the compounds were tested in the presence of detergent (Triton X-100, 0.01%). Only compounds **1** and **2** exhibited weak inhibition of Mur ligases (Table 1): 1,3,5-triazine-based compound **1** (Fig. 6) displayed 26% inhibition of MurC and 50% inhibition of MurD at 500 μ M, while quinolin-2(1*H*)-one-based compound **2** (Fig. 6) showed 29% inhibition of MurD at 250 μ M but did not inhibit MurC.

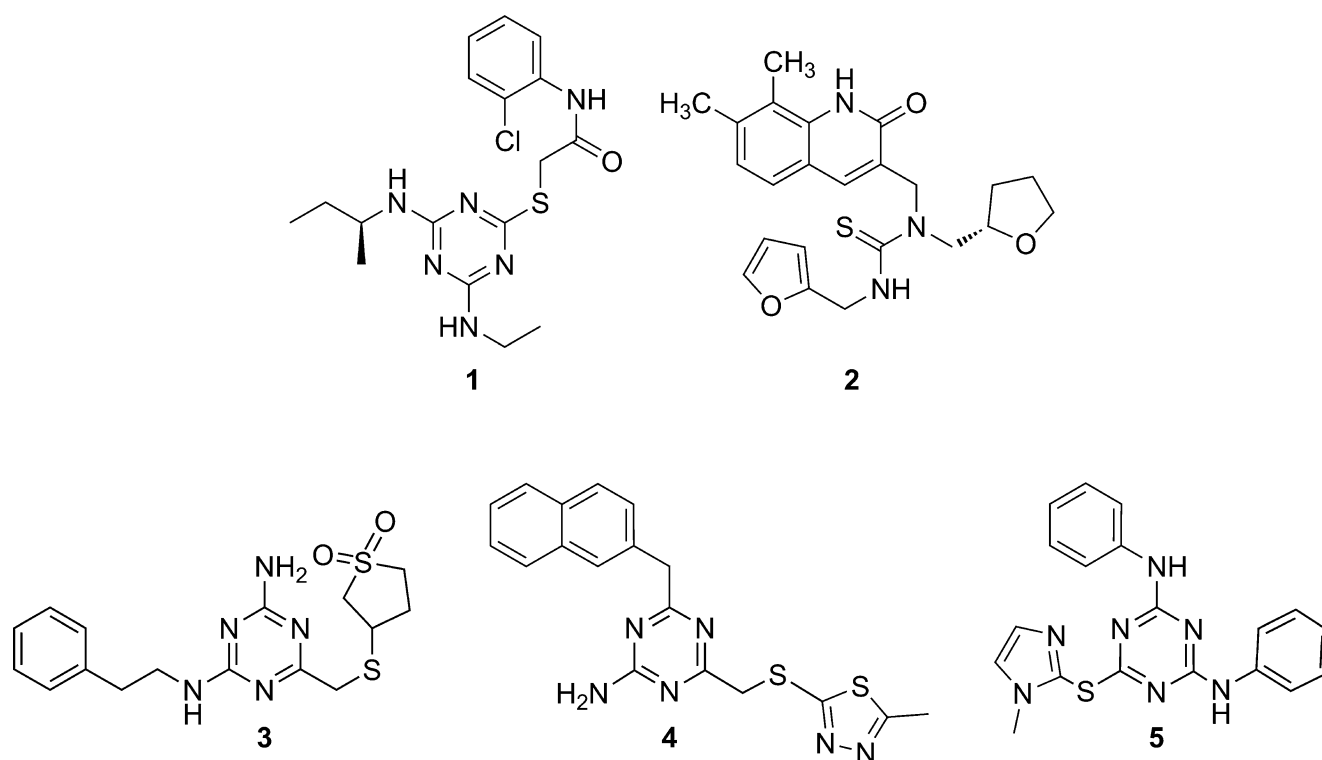


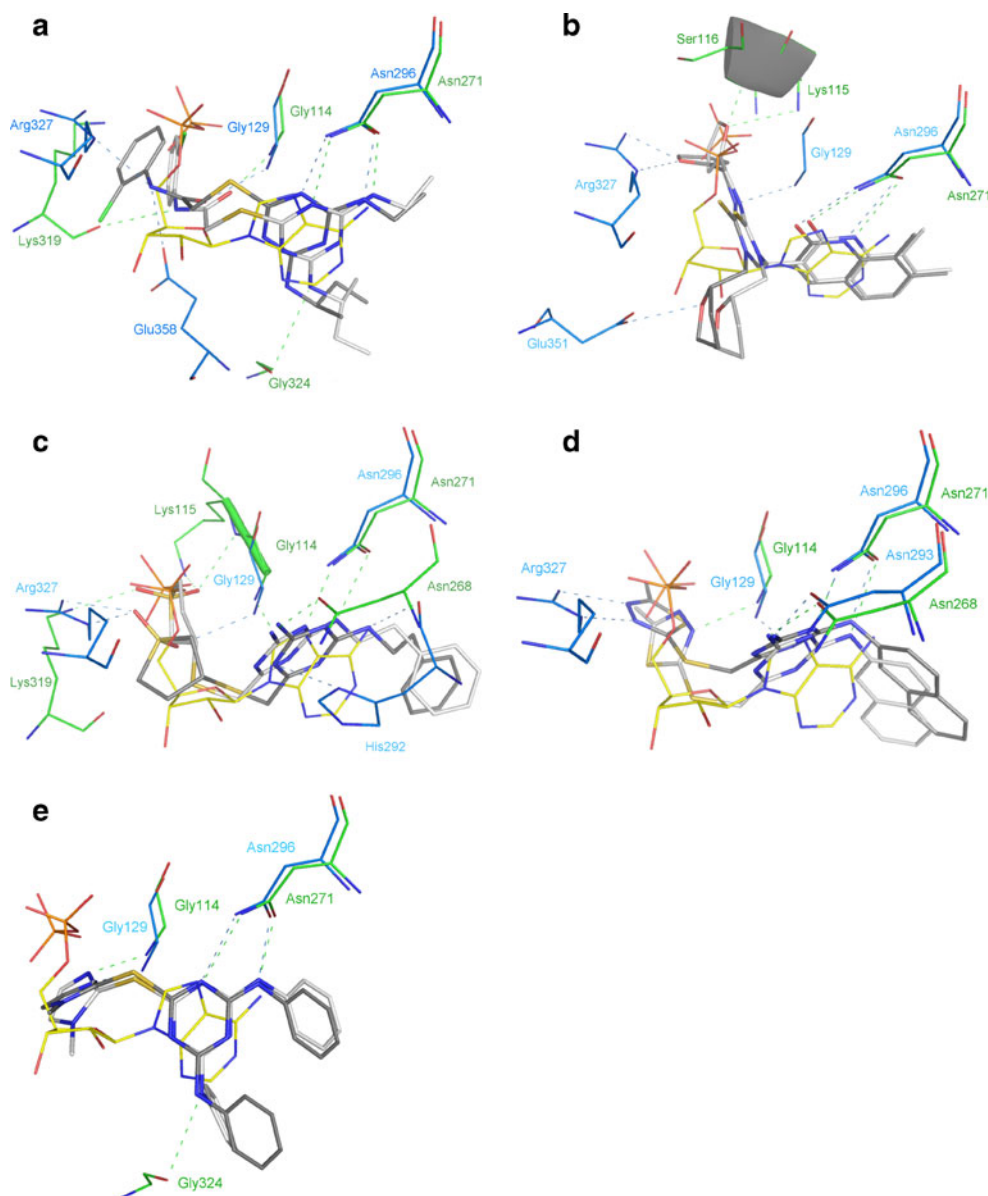
Fig. 6 Structures of the discovered inhibitors of MurC and MurD

The quinolin-2(1*H*)-one moiety of compound **2** forms putative hydrogen bonds with the side chain of Asn296 (*E. coli* MurC) or Asn271 (*E. coli* MurD), while the oxygen atom of the furanyl ring forms two hydrogen bonds with the guanidino group of Arg327 in MurC or with the backbone NH groups of Lys115 and Ser116 in the MurD diphosphate-binding pockets (Fig. 7b).

1,3,5-Triazine-based inhibitors of MurF ligase have already been discovered by virtual screening [23] and by subsequent optimization of the VS hit [48]. Moreover, the 1,3,5-triazine ring present in compound **1** has also been described to bind to the ATP-binding pocket of DNA gyrase, where it occupies the part of the binding site usually accommodated by the adenine ring [49]. Such a binding mode of the 1,3,5-triazine ring is in agreement with our docked geometry of **1**, where the NH group of the ethylamino substituent and the ring nitrogen atom between the ethylamino and (*S*)-butan-2-amino groups of the 1,3,5-triazine moiety interact with Asn296 and Asn271 in the adenine-binding pockets of *E. coli* MurC and MurD ligases, respectively, while the 2-chloroaniline ring points towards the diphosphate-binding pocket (Fig. 7a). Since compound **1** was the only weak dual inhibitor identified from the selected 18 compounds, we decided to extract all 1,3,5-triazine-based compounds present in the database matching the first UNITY pharmacophore (Fig. 5), aiming to find among them compounds with improved potency (Fig. 4b).

The 2579 1,3,5-triazines obtained were docked into the MurC and MurD active sites using the above-described FlexX-Pharm pharmacophore constraints. The best ranked docking solutions of 247 compounds matching the pharmacophore were scored with DrugScore, then inspected visually, and four candidates selected for biological testing. Two of them inhibited MurC and MurD (compounds **4** and **5**), and one only MurC (compound **3**) (Fig. 6). However, the inhibitory potency of the dual acting compounds **4** and **5** was not superior to that of **1** (Table 1). The predicted binding modes of the 1,3,5-triazine ring of compounds **3–5** in the adenine-binding pockets of MurC and MurD ligases are similar to that of inhibitor **1** (Fig. 7). In detail, two putative hydrogen bonds are formed with the side chain of Asn296 in MurC or Asn271 in the MurD active site, which is in agreement with the pharmacophore model. In the case of compound **3**, additional hydrogen bonds can possibly be formed with the amide NH group of Gly129 and the side chains of His292 and Arg327 in the MurC active site, while in the MurD active site putative hydrogen bonds are formed with the backbone NH groups of Gly114 and Lys115 and the side chains of Asn268 and Lys319 (Fig. 7c). Compound **4** could further interact with the amide NH group of Gly129 and the side chain of Asn293 in MurC or the amide NH group of Gly114 and the side chain of Asn268 in the MurD active site. In the active site of MurC ligase, additional hydrogen bonds with the side chain of Arg327 are also

Fig. 7 Superposition of the ADP (from PDB code: 3UAG, in yellow lines) and the best-ranked docking pose of inhibitor (a) **1**; (b) **2**; (c) **3**; (d) **4**; (e) **5** in the *E. coli* MurC active site in dark grey sticks and in the *E. coli* MurD active site in light grey sticks. For clarity, only active site residues interacting with the inhibitors are shown (MurC active site residues in blue lines, PDB code: 2F00, and MurD active site residues in green lines, PDB code: 3UAG). Potential hydrogen bonds between enzyme active site residues and inhibitors are shown as dashed lines



possible (Fig. 7d). The calculated binding mode of the dual inhibitor **5** predicts the formation of one additional hydrogen bond with the amide NH group of Gly129 in the MurC active site, while two possible hydrogen bonds could be formed with the backbone NH groups of Gly114 and Gly324 in the MurD active site (Fig. 7e). In general, differences in the MurC and MurD inhibitory activities of compounds **1–5** cannot be well rationalized by the calculated binding modes, since similar interactions are predicted to be formed in the case of both enzymes.

Conclusions

We have performed a virtual screening study enumerating the ZINC drug-like database for potential ATP-competitive Mur ligase inhibitors possessing multitarget activity, using a protocol of consecutive hierarchical filters. Selected candidates were tested for MurC and MurD inhibition, but only weak dual MurC and MurD inhibitors were identified. There may be several reasons for the low hit rate. First, the analysis of the chemical properties of known antibacterial

drugs shows that they populate a unique property space that is different from that of the drugs in other therapeutic areas [50], which makes the use of compound libraries designed to target eukaryotic enzymes difficult. Supposedly, chemical libraries that are better suited for finding antibacterial compounds are thus needed. Further, only a few compounds were evaluated in the Mur ligase inhibition assays. Even in HTS campaigns, where large collections of compounds were evaluated, only a few or no hits were discovered against several bacterial targets [51]. Nevertheless, the new scaffolds for the design of multiple Mur ligase inhibitors targeting the ATP-binding site, that were discovered in the present VS, provide starting points for further optimization.

Acknowledgments This work was supported by the Sixth Framework Programme (FP6) Integrated Project Inhibition of New TArgets for Fighting Antibiotic Resistance (EUR-INTAFAR) (Project No. LSHM-CT-2004-512138), by the Slovenian Research Agency (Grant No. P1-0208) and by the World Federation of Scientists. The authors thank Professor Roger Pain for critical reading of the manuscript.

References

- Livermore DM (2003) Bacterial resistance: origins, epidemiology, and impact. *Clin Infect Dis* 36:11–23
- van Heijenoort J (2001) Recent advances in the formation of the bacterial peptidoglycan monomer unit. *Nat Prod Rep* 18:503–519
- Vollmer W, Blanot D, de Pedro MA (2008) Peptidoglycan structure and architecture. *FEMS Microbiol Rev* 32:149–167
- Barreteau H, Kovač A, Boniface A, Sova M, Gobec S, Blanot D (2008) Cytoplasmic steps of peptidoglycan biosynthesis. *FEMS Microbiol Rev* 32:168–207
- El Zoeiby A, Sanschagrin F, Levesque RC (2003) Structure and function of the Mur enzymes: development of novel inhibitors. *Mol Microbiol* 47:1–12
- Bertrand JA, Auger G, Martin L, Fanchon E, Blanot D, Le Beller D, van Heijenoort J, Dideberg O (1999) Determination of the MurD mechanism through crystallographic analysis of enzyme complexes. *J Mol Biol* 289:579–590
- Bouhss A, Dementin S, van Heijenoort J, Parquet C, Blanot D (2002) MurC and MurD synthetases of peptidoglycan biosynthesis: borohydride trapping of acyl-phosphate intermediates. *Methods Enzymol* 354:189–196
- Anderson MS, Eveland SS, Onishi HR, Pompliano DL (1996) Kinetic mechanism of the *Escherichia coli* UDPMurNac-tripeptide D-alanyl-D-alanine-adding enzyme: use of a glutathione *S-transferase fusion*. *Biochemistry* 35:16264–16269
- Emanuele JJ, Jin HY, Yanchunas J, Villafranca JJ (1997) Evaluation of the kinetic mechanism of *Escherichia coli* uridine diphosphate-*N*-acetylmuramate:L-alanine ligase. *Biochemistry* 36:7264–7271
- Morphy R, Rankovic Z (2009) Designing multiple ligands - medicinal chemistry strategies and challenges. *Curr Pharm Des* 15:587–600
- Csermely P, Agoston V, Pongor S (2005) The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol Sci* 26:178–182
- Silver LL (2007) Multi-targeting by monotherapeutic antibacterials. *Nat Rev Drug Discov* 6:41–55
- Tomašić T, Zidar N, Kovač A, Turk S, Simčič M, Blanot D, Müller-Premru M, Filipič M, Grdadolnik SG, Zega A, Anderluh M, Gobec S, Kikelj D, Peterlin Mašič L (2010) 5-Benzylidenethiazolidin-4-ones as multitarget inhibitors of bacterial Mur ligases. *ChemMedChem* 5:286–295
- Bouhss A, Mengin-Lecreux D, Blanot D, van Heijenoort J, Parquet C (1997) Invariant amino acids in the Mur peptide synthetases of bacterial peptidoglycan synthesis and their modification by site-directed mutagenesis in the UDP-MurNac:L-alanine ligase from *Escherichia coli*. *Biochemistry* 36:11556–11563
- Eveland SS, Pompliano DL, Anderson MS (1997) Conditionally lethal *Escherichia coli* murein mutants contain point defects that map to regions conserved among murein and folyl poly-gamma-glutamyl ligases: identification of a ligase superfamily. *Biochemistry* 36:6223–6229
- Bouhss A, Dementin S, Parquet C, Mengin-Lecreux D, Bertrand JA, Le Beller D, Dideberg O, van Heijenoort J, Blanot D (1999) Role of the ortholog and paralog amino acid invariants in the active site of the UDP-MurNac-L-alanine:D-glutamate ligase (MurD). *Biochemistry* 38:12240–12247
- Chappelle EW, Levin GV (1968) Use of the firefly bioluminescent reaction for rapid detection and counting of bacteria. *Biochem Med* 2:41–52
- Traut TW (1994) Physiological concentrations of purines and pyrimidines. *Mol Cell Biochem* 140:1–22
- Škedelj V, Tomašić T, Peterlin Mašič L, Zega A (2011) ATP-binding site of bacterial enzymes as a target for antibacterial drug design. *J Med Chem* 54:915–929
- Konc J, Janežič D (2010) ProBiS: a web server for detection of structurally similar protein binding sites. *Nucleic Acids Res* 38:W436–W440
- Konc J, Janežič D (2010) ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* 26:1160–1168
- Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* 3:935–949
- Turk S, Kovač A, Boniface A, Bostock JM, Chopra I, Blanot D, Gobec S (2009) Discovery of new inhibitors of the bacterial peptidoglycan biosynthesis enzymes MurD and MurF by structure-based virtual screening. *Bioorg Med Chem* 17:1884–1889
- Perdih A, Kovač A, Wolber G, Blanot D, Gobec S, Šolmajer 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
- The UniProt Consortium (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* 39:D214–D219
- Accelrys Discovery Studio is available from Accelrys Inc, San Diego, California 92121, USA
- SYBYL Molecular modelling package 7.3. (2006) St. Louis, MO Tripos Inc
- UNITY Chemical Information Software (2006) St. Louis, MO Tripos Inc
- Gasteiger J, Rudolph C, Sadowski J (1990) Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput Method* 3:537–547
- Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 28:849–857
- Pymol is available from Delano Scientific LLC, San Francisco, CA. <http://pymol.sourceforge.net>
- Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* 261:470–489

33. Rarey M, Wefing S, Lengauer T (1996) Placement of medium-sized molecular fragments into active sites of proteins. *J Comput Aided Mol Des* 10:41–54
34. Gold v4.1 is available from The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK. www.ccdc.cam.ac.uk
35. Liger D, Masson A, Blanot D, van Heijenoort J, Parquet C (1995) Over-production, purification and properties of the uridine-diphosphate-*N*-acetylmuramate:L-alanine ligase from *Escherichia coli*. *Eur J Biochem* 230:80–87
36. Auger G, Martin L, Bertrand J, Ferrari P, Fanchon E, Vaganay S, Petillot Y, van Heijenoort J, Blanot D, Dideberg O (1998) Large-scale preparation, purification, and crystallization of UDP-*N*-acetylmuramoyl-L-alanine:D-glutamate ligase from *Escherichia coli*. *Protein Expr Purif* 13:23–29
37. Lanzetta PA, Alvarez LJ, Reinach PS, Candia OA (1979) Improved Assay for Nanomole Amounts of Inorganic-Phosphate. *Anal Biochem* 100:95–97
38. Tomašić T, Zidar N, Rupnik V, Kovač A, Blanot D, Gobec S, Kikelj D, Peterlin Mašič L (2009) Synthesis and biological evaluation of new glutamic acid-based inhibitors of MurD ligase. *Bioorg Med Chem Lett* 19:153–157
39. Zidar N, Tomašić T, Šink R, Rupnik V, Kovač A, Turk S, Patin D, Blanot D, Contreras Martel C, Dessen A, Müller Premru M, Zega A, Gobec S, Peterlin Mašič L, Kikelj D (2010) Discovery of novel 5-benzylidenerhodanine and 5-benzylidenethiazolidine-2,4-dione inhibitors of MurD ligase. *J Med Chem* 53:6584–6594
40. Mol CD, Brooun A, Dougan DR, Hilgers MT, Tari LW, Wijnands RA, Knuth MW, McRee DE, Swanson RV (2003) Crystal structures of active fully assembled substrate- and product-bound complexes of UDP-*N*-acetylmuramic acid:L-alanine ligase (MurC) from *Haemophilus influenzae*. *J Bacteriol* 185:4152–4162
41. Deva T, Baker EN, Squire CJ, Smith CA (2006) Structure of *Escherichia coli* UDP-*N*-acetylmuramoyl:L-alanine ligase (MurC). *Acta Crystallogr D Biol Crystallogr* 62:1466–1474
42. Gordon E, Flouret B, Chantalat L, van Heijenoort J, Mengin-Lecreux D, Dideberg O (2001) Crystal structure of UDP-*N*-acetylmuramoyl-L-alanyl-D-glutamate:meso-diaminopimelate ligase from *Escherichia coli*. *J Biol Chem* 276:10999–11006
43. Yan Y, Munshi S, Leiting B, Anderson MS, Chrzas J, Chen Z (2000) Crystal structure of *Escherichia coli* UDPMurNAc-tripeptide D-alanyl-D-alanine-adding enzyme (MurF) at 2.3 Å resolution. *J Mol Biol* 304:435–445
44. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 45:177–182
45. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Delivery Rev* 46:3–26
46. 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
47. Gohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* 295:337–356
48. Sosič I, Štefane B, Kovač A, Turk S, Blanot D, Gobec S (2010) The synthesis of novel 2,4,6-trisubstituted 1,3,5-triazines: a search for potential MurF enzyme inhibitors. *Heterocycles* 81:91–115
49. Ward WHJ, Holdgate GA (2001) 7 Isothermal Titration Calorimetry in Drug Discovery. In: King FD, Oxford AW (eds) *Progress in Medicinal Chemistry*, vol 38. Elsevier, pp 309–376
50. O'Shea R, Moser HE (2008) Physicochemical properties of antibacterial compounds: implications for drug discovery. *J Med Chem* 51:2871–2878
51. Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL (2007) Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat Rev Drug Discov* 6:29–40

Substrate binding and translocation of the serotonin transporter studied by docking and molecular dynamics simulations

Mari Gabrielsen · Aina Westrheim Ravna ·
Kurt Kristiansen · Ingebrigt Sylte

Received: 20 December 2010 / Accepted: 16 May 2011 / Published online: 14 June 2011
© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract The serotonin (5-HT) transporter (SERT) plays an important role in the termination of 5-HT-mediated neurotransmission by transporting 5-HT away from the synaptic cleft and into the presynaptic neuron. In addition, SERT is the main target for antidepressant drugs, including the selective serotonin reuptake inhibitors (SSRIs). The three-dimensional (3D) structure of SERT has not yet been determined, and little is known about the molecular mechanisms of substrate binding and transport, though such information is very important for the development of new antidepressant drugs. In this study, a homology model of SERT was constructed based on the 3D structure of a prokaryotic homologous leucine transporter (LeuT) (PDB id: 2A65). Eleven tryptamine derivatives (including 5-HT) and the SSRI (*S*)-citalopram were docked into the putative substrate binding site, and two possible binding modes of the ligands were found. To study the conformational effect that ligand binding may have on SERT, two SERT–5-HT and two SERT–(*S*)-citalopram complexes, as well as the SERT apo structure, were embedded in POPC lipid bilayers and comparative molecular dynamics (MD) simulations were performed. Our results show that 5-HT in the SERT–5-HT^B complex induced larger conformational changes in the cytoplasmic parts of the transmembrane helices of SERT than any of the other ligands. Based on these results, we suggest that the formation and breakage of ionic interactions with amino acids in transmembrane helices 6

and 8 and intracellular loop 1 may be of importance for substrate translocation.

Keywords SERT · Homology modeling · (*S*)-citalopram binding · Substrate binding · Molecular dynamics · Substrate transport

Introduction

The serotonin [5-hydroxytryptamine (5-HT)] transporter (SERT) is located in the membrane of presynaptic neurons and plays an important role in the termination of serotonergic neurotransmission by transporting 5-HT from the synaptic cleft into the presynaptic neuron. SERT, and the closely related dopamine and noradrenaline (norepinephrine) transporters (DAT and NET, respectively), are located in limbic areas of the CNS that are involved in mood, emotion and reward processes, and are important targets of therapeutic drugs as well as psychoactive illicit drugs. Among the compounds that act on SERT are drugs belonging to the two main groups of antidepressants—the classic tricyclic antidepressants (TCAs) and the newer selective serotonin reuptake inhibitors (SSRIs)—and well-known drugs of abuse such as cocaine and amphetamines, including 3,4-methylenedioxymethamphetamine (MDMA, commonly known as “ecstasy”).

SERT, DAT and NET belong to the neurotransmitter/sodium symporter (NSS) transporter family (Transporter Classification code 2.A.22 [1]), also known as the SLC6 family [2]. This transporter family constitutes a large number of secondary transporters that use Na⁺ electrochemical gradients to transport extracellular solutes across membranes. At least 177 eukaryotic and 167 prokaryotic transporters have been classified as belonging to this family

Mari Gabrielsen is a fellow of the Ph.D. school in Molecular and Structural Biology (MSB) at the University of Tromsø, Norway.

M. Gabrielsen · A. W. Ravna · K. Kristiansen · I. Sylte (✉)
Medical Pharmacology and Toxicology, Department of Medical
Biology, Faculty of Health Sciences, University of Tromsø,
N-9037 Tromsø, Norway
e-mail: ingebrigt.sylte@uit.no

[3], transporting a large number of solutes. In addition to the biogenic amines, amino acids such as γ -aminobutyric acid (GABA), glycine, tryptophan, tyrosine and leucine (the GAT-1, GlyT, TnT, Tyl1 and LeuT transporters, respectively) are transported by NSS transporters [1].

The three-dimensional (3D) structure of SERT (or, indeed, that of any eukaryotic NSS family member) has not been experimentally determined; however, the first X-ray crystal structure of a prokaryotic NSS family member, the *Aquifex aeolicus* leucine transporter (LeuT), was published in 2005 [4]. Since then, several crystal structures of LeuT have been published, and 3D structures of LeuT in an occluded conformation [5–7] and in an outward-facing conformation [8] are now available. These crystal structures can be used as templates for the generation of 3D models of SERT and other NSS transporters using the homology modeling approach, taking advantage of the fact that 3D structure is more conserved than the sequence [9]. Several SERT models have been generated based on the occluded LeuT crystal structure [10–12] and a published comprehensive alignment of NSS family members by Beuming et al. [3].

In 1966, transporter proteins were suggested to operate through an alternating-access mechanism [13] in which a central substrate binding site is alternately exposed to either the extracellular environment or the cytoplasm through conformational changes of the protein. The 3D crystal structures of LeuT thus fit this proposed transport mechanism, as they are in open-to-out and occluded conformations [4–8]. In the latter conformation, leucine is bound in the substrate binding site of LeuT, and the side chains of two phenylalanine residues (corresponding to Y176 and F335 in SERT) and one arginine and glutamate residue (corresponding to R104 and E493 in SERT) block access from the extracellular environment to the substrate binding site [4–7]. In the outward-facing conformation, the competitive inhibitor L-tryptophan displaces leucine from the substrate binding site and causes LeuT to stabilize in an outward-facing conformation, where the distance between the side chains of Y176 and F335 increases [8]. In all of the LeuT 3D structures, however, approximately 20 Å of tightly packed helical regions effectively separate the substrate binding site from the cytoplasmic environment [4–8]. Thus, neither the crystal structures of LeuT nor the SERT homology models based on these structures reveal much information about how substrates are transported from the extracellular environment into the interiors of the cells. One possible way to gain more insight into the conformational mechanisms that take place in a transporter following the binding of either substrate or inhibitor may be by performing long molecular dynamics (MD) simulations.

To study ligand binding and SERT conformational changes upon ligand binding, the LeuT occluded structure (PDB id 2A65) [4] was used to generate a homology model

of SERT, and 5-HT and ten other tryptamine derivatives, as well as the SSRI (*S*)-citalopram, were docked into the putative substrate binding pocket detected in the SERT model. Analysis of the docking results revealed two putative binding modes of the tryptamine derivatives and (*S*)-citalopram in SERT. Based on these docking results, one representative complex of SERT and 5-HT and (*S*)-citalopram in both binding modes was selected for MD simulations, in addition to the apo-SERT. The MD simulations were performed after embedding the SERT–ligand complexes in palmitoyl-oleoyl-phosphatidylcholine (POPC) lipid bilayers. The results from the MD simulations of the five SERT–(ligand)–POPC complexes showed that the putative substrate binding site had started to extend towards the intracellular parts of SERT during the MD simulation in one of the SERT–5-HT complexes (namely, the SERT–5-HT^B complex). In the same complex, a vestibule extending from the cytoplasm towards the substrate binding site had started to form. Based on these results, we identified several amino acids that may play a role in the opening and closing of a vestibule reaching from the substrate binding site to the cytoplasm.

Methods

Homology modeling of SERT

The SERT (UniProtKB/Swiss-Prot accession number P31645 [14]) and the LeuT (PDB id 2A65) [4] amino acid sequences were aligned using ICM software (version 3.5) [15], and the alignment was adjusted to fit the published comprehensive alignment of NSS family members [3]. Based on this alignment, the homology model of SERT was constructed using the BuildModel macro of ICM [15]. The macro constructs the backbone of the target protein using the backbone conformation of the template in the aligned regions using core sections defined by the average C_{α} atom positions in these regions. The conformations of the side chains of amino acids that were identical for the template and the target structures were then transferred from the template to the target, whereas nonidentical side chains were assigned their most likely rotamer. For the loops with insertions or deletions between the template and target sequences, the macro performs a loop search of the PDB database, selecting loops with matching loop ends and a loop sequence that is as closed as possible. The loops are inserted into the model and the side chains are modified according to the model sequence and steric interactions with the surroundings of the model.

The SERT amino acids E78-T192 and W220-I608 were included in the homology model. These amino acids comprise the 12 putative transmembrane helices (TMs) and the

intracellular and extracellular loops (ILs and ELs, respectively) connecting the transmembrane helices, except for parts of the large EL2 (amino acids 193–219). This loop segment was not included in the model as it is lacking in the LeuT template. Amino acids corresponding to the N-terminal (amino acids 1–77) and C-terminal (amino acids 609–630) regions of SERT were also not included in the model for the same reason.

The two sodium ion binding sites and one chloride binding site in the LeuT crystal structure [4] were copied to SERT after superimposing the LeuT crystal structure and the SERT model. A chloride ion was also added to the SERT homology model such that it occupied a position corresponding to the carboxylate carbon coordinates of LeuT glutamic acid at position 290 (corresponding to S372 in SERT), as suggested by Forrest [11] and Zomot [16].

Energy refinement of the SERT homology model was performed using the ICM RefineModel macro. This three-step macro performs (1) a side-chain conformational sampling using “Montecarlo fast” [17], (2) iterative annealing with tethers provided, and (3) a second side-chain sampling. The program module Montecarlo fast [17] samples the conformational space by performing iterations that consist of a random move followed by a local energy minimization. The complete energy is then calculated, and the iteration is accepted or rejected based on the energy and the temperature. In the annealing of the backbone (step 2), the tethers included are harmonic restraints that pull an atom in the model to a static point in space represented by a corresponding atom in the template.

The energy-refined SERT homology model was uploaded to the SAVES server for a structure quality check (http://nihserver.mbi.ucla.edu/Saves_3/). The Ramachandran plot provided by Procheck showed that the SERT homology model was a good-quality model; 96.6% of the non-glycine and non-proline amino acids were in the favored regions, whereas 3.4% (12 amino acids) were in additional allowed regions. Of these 12 amino acids, one amino acid, D98, was located in the putative substrate binding area. This amino acid is important for substrate and inhibitor binding to SERT [10, 12, 18–20], and was located in an unwound region of TM1. However, this location gives D98 more freedom to rotate, and hence explains its location in additionally allowed regions of the Ramachandran plot.

Ligand docking

To detect possible binding pockets in the SERT structure, the ICM PocketFinder macro was used (default tolerance level of 4.6). The algorithm uses a transformation of the Lennard–Jones potential calculated from a three-dimensional protein structure and does not require any knowledge about a

potential ligand molecule; i.e., it is based solely on protein structure [21].

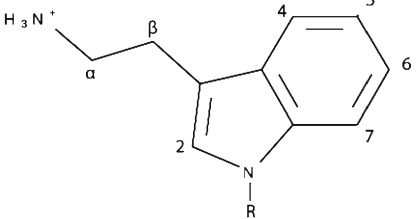
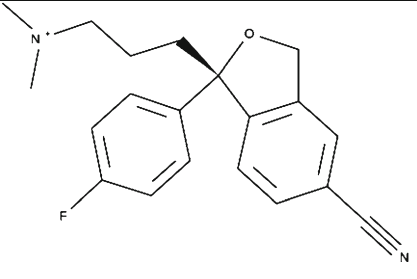
5-HT and ten other tryptamine derivatives (tryptamine, 4-hydroxytryptamine (4-HT), 7-methyltryptamine (7-MT), 2-methylserotonin (2-MT), 5-methoxy-3-(1,2,5,6-tetrahydro-4-pyridinyl)-1 *H*-indole (RU24969), *N*-isopropyltryptamine (NIT), 5-methoxy-*N*-isopropyltryptamine (5MNIT), 7-benzyloxytryptamine (7-BT), 5,6,7-trihydroxytryptamine and serotonin *o*-sulfate (Table 1) were constructed using the ChemDraw option of ICM. Default ECEPP/3 partial charges were assigned to the protonated forms of the ligands [22], and the compounds were docked using the batch docking method of ICM. RU24969 was also docked in its unprotonated state. The SSRI [(*S*)-1-[3-(dimethylamino)propyl]-1-(4-fluorophenyl)-1,3-dihydroisobenzofuran-5 carbonitrile; (*S*)-citalopram] (Table 1) was constructed using ChemDraw and docked into the same binding site as the tryptamine derivatives, as experimental studies indicate that (*S*)-citalopram is a competitive 5-HT inhibitor [18]. The ligands were docked using a semi-flexible docking protocol where SERT was kept rigid but the ligands flexible.

The poses of each ligand were clustered and compared with the clusters of the other ligands. This analysis led to the identification of two putative ligand positions for both the tryptamine derivatives and (*S*)-citalopram. One representative from each of the two clusters of 5-HT (representing the tryptamine derivatives) and (*S*)-citalopram were selected for MD simulations.

Molecular dynamics simulation

The automated CHARMM-GUI membrane builder tool [23] was used for the generation of a palmitoylcholine (POPC) lipid bilayer around the five SERT–(ligand) complexes selected after docking. The pre-orientated LeuT structure [4] from the Orientations of Proteins in Membranes (OPM) database [24] was used to orient the SERT model in the membrane by superimposing the LeuT and SERT. An unequilibrated lipid bilayer was generated using the replacement method, in which SERT was packed with lipid-like spheres whose positions then were used to place randomly chosen POPC lipid molecules from a lipid library composed of 2000 different conformations of lipids generated by MD simulations of pure lipid bilayers. The dimensions of the entire SERT–(ligand)–POPC molecular system was approximately 100×100×100 Å, including 1 Å extra added in each direction in order to introduce space between the boundary of the system and the boundary atoms of the simulation cell. One hundred fifteen lipids were included in the outer bilayer and 121 in the inner bilayer. Water molecules (TIP3) and K⁺ and Cl[−] ions were then added by the membrane builder tool to fully

Table 1 The structures of tryptamine derivatives and (*S*)-citalopram docked into the putative substrate binding site in SERT. Positions of substitutions in the tryptamine derivatives are shown

Tryptamine derivatives	
Tryptamine	-
5-Hydroxytryptamine (5-HT, serotonin)	5
4-Hydroxytryptamine (4-HT)	4
7-Methyltryptamine (7-MT)	7
2-Methyltryptamine (2-MT)	2
7-Benzyloxytryptamine (7-BT)	7
<i>N</i> -Isopropyltryptamine (NIT)	R
5-Methoxy- <i>N</i> -isopropyltryptamine (5-MNIT)	5, R
5-Methoxy-3-(1,2,5,6-tetrahydro-4-pyridinyl)-1 <i>H</i> -indole (RU24969)	β
Trihydroxytryptamine	5, 6, 7
Serotonin <i>o</i> -sulfate	5
(<i>S</i>)-Citalopram	

solvate the system. In total, each of the five complexes consisted of approximately 98,000 atoms.

The NAMD scalable MD simulator (versions 2.6 and 2.7b1) [25] was used to equilibrate the systems and perform the production runs. The MD simulations were run using 64 processors on the Stallo supercomputer at the University of Tromsø, Norway, using Chemistry at HARvard Molecular Mechanics (CHARMM) force fields. The CHARMM par_all27_prot_lipidNBFIX parameter file, which includes the CHARMM22/CMAP force field [26, 27]

for the protein and the CHARMM27 force field [28, 29] for lipids, was used. For the complexes containing 5-HT or (*S*)-citalopram, the CHARMM36 general force field for small molecule drug design (CGenFF v. 2a3 [30]) was included, manually adding force field angle and dihedral parameters that are not included in CGenFF v. 2a3 [30]. To allow the large volume fluctuations that are typical of the initial dynamics of a new system in an NPT ensemble, a margin of 5 was used during the equilibration steps, which was reduced to 2 during the production runs [25]. During the simulations,

Nosé–Hoover–Langevin dynamics were used to simulate the NPT ensemble. This method combines the Nosé–Hoover constant pressure method with piston fluctuation control implemented using Langevin dynamics by coupling the piston to a heat bath. A damping constant of $10/\text{langevinPistonDecay}$ was used during the equilibration steps, which was reduced to $1/\text{langevinPistonDecay}$ during the production runs. The $\text{langevinPistonDecay}$ (50 fs) was set to be smaller than $\text{langevinPistonPeriod}$ (200 fs) to ensure that harmonic oscillations in the periodic cell were overdamped. The target pressure was set at 1.01325 bar (atmospheric pressure at sea level), and group-based pressure (useGroupPressure) was used to control the periodic cell fluctuations, as the atom-based pressure has more high-frequency noise. In addition, a flexible cell (useFlexibleCell) was used, allowing the height, length, and width of the cell to fluctuate independently during the simulation, which is very useful for anisotropic systems such as membranes.

The equilibration of the five SERT–(ligand)–POPC complexes consisted of three steps during which the system was gradually released. During steps (1) and (2), harmonic constraints of $1 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ were specified in the PDB beta field of each atom to be constrained. In order to induce the appropriate order of the fluid-like bilayer, all atoms except the lipid tail atoms were constrained during step (1), and lipids, water and ions were permitted to adapt to the structure of the protein. During step (2), only protein atoms were constrained, whereas the whole system was released during step (3). During step (1), 10,000 steps of conjugate gradient energy minimization were performed, followed by 10,000 steps (10 ps) of system heating to 300 K under constant temperature control and 500,000 steps (0.5 ns) of MD. During steps (2) and (3), only 10,000 steps of conjugate gradient minimization followed by 500,000 steps (0.5 ns) of MD were performed. In total, 30,000 steps of conjugate gradient minimization, 10 ps of heating and 1.5 ns of MD simulations were run to equilibrate the system. To confirm that the systems stabilized during equilibration, the RMSD from the starting structure was monitored during each simulation using the molecular dynamics (VMD) viewer version 1.8.6 [31]. Finally, the equilibration phases of the SERT–5-HT binding modes A and B and the (*S*)-citalopram binding modes A and B, as well as SERT alone, were followed by 22, 21, 32, 23 and 25 ns MD simulations, respectively. The production simulations were performed at 300 K. Following the production runs, VMD [31] was used to generate average structures of each complex based on the last 10 ns of each simulation, and ICM PocketFinder [21] was used to detect possible pockets in the average structures. Based on these analyses, the SERT–5-HT^B complex MD simulation was prolonged to 49 ns.

Results

Homology modeling

The constructed homology model consisted of 12 TMs, among which TMs 1–5 and 6–10 were arranged with a pseudo-twofold axis in the membrane plane, as for LeuT [4]. Three possible binding pockets were identified by ICM PocketFinder in the SERT homology model: one in the region corresponding to the LeuT substrate binding site, and two extracellular pockets which were separated from the putative substrate binding pocket by the side chains of Y176 and F335, the aromatic amino acids of the extracellular gate. In LeuT [4], only one pocket was detected in this extracellular region, as EL4 in LeuT is missing three amino acids at the tip of EL4 as compared to SERT [3] (results not shown).

ICM PocketFinder [21] identified a binding pocket that corresponded to the substrate binding site of LeuT [4]. Experimental data on SERT and the X-ray structure of LeuT also suggest that the substrate binding site of SERT and LeuT are in the same region [10, 12, 20, 32–34], halfway across the membrane bilayer within the TMs. This location is also consistent with the alternate access theory [13]. Amino acids from four TMs contribute to the binding pocket detected by ICM PocketFinder, namely from TM1 (Y95, D98, G100), TM3 (I172, A173, Y176), TM6 (F335, S336, G338, F341, V343) and TM8 (S438, T439, G442). An important feature of the detected binding pocket is the deviation from regular helical structure in the unwound regions of TM1 (A96–D98) and TM6 (G338–G342). A similar deviation is observed in corresponding regions of the X-ray structure of LeuT. In the unwound regions, the main-chain carbonyl oxygen and amide nitrogen atoms are exposed such that they can easily take part in direct hydrogen-bonding interactions with ligands and coordinate ions.

The substrate binding pocket detected by ICM PocketFinder could be divided into three subpockets based on the main properties of amino acids involved. The first subpocket, the hydrophobic subpocket, was located towards the intracellular end of the binding site and was surrounded by the side chains of A169 (TM3), A173 (TM3), V343 (TM6), and G442 (TM8). The side chain of I172 (TM3) was positioned such that it could participate in forming the hydrophobic subpocket but also separate the hydrophobic subpocket from an aromatic. The aromatic subpocket consisted of the side chains of the two aromatic amino acids of the extracellular gate, Y176 (TM3) and F335, and F341 located in the unwound region of TM6. The third subpocket, the ionic subpocket, was located in the vicinity of D98 (TM1).

Analysis of the docking results

The docking of 5-HT and ten other tryptamine derivatives and (*S*)-citalopram indicated two possible binding modes of the compounds, designated SERT–5-HT^A, SERT–5-HT^B, SERT–(*S*)-citalopram^A and SERT–(*S*)-citalopram^B, respectively (Fig. 1). The SERT–5-HT binding modes represent the binding poses of all tryptamine derivatives. In both the SERT–5-HT^A and SERT–5-HT^B binding modes, 5-HT occupied the ionic and hydrophobic—but not the aromatic—subpockets of the binding site. The protonated amine of 5-HT was located near the D98 carboxyl side chain in both modes, which is in accordance with experimental

data [10, 12, 19, 20]. The two binding modes of 5-HT differ in the orientation of the indole ring nitrogen and the orientation of the 5 position (Fig. 1). In the SERT–5-HT^A binding mode, the indole ring nitrogen was found between Y95 and F341, whereas the 5 position was pointing towards Y176, S438 and T439. In the SERT–5-HT^B binding mode, however, the indole ring was flipped 180° compared to binding mode A, and the indole nitrogen group was pointing towards the aromatic side chains of Y176 and S438, and the 5 position towards A169 and F341 (Fig. 1). Interestingly, similar binding modes of 5-HT to the SERT–5-HT^A and SERT–5-HT^B binding modes have also been described by other groups [10, 12, 35].

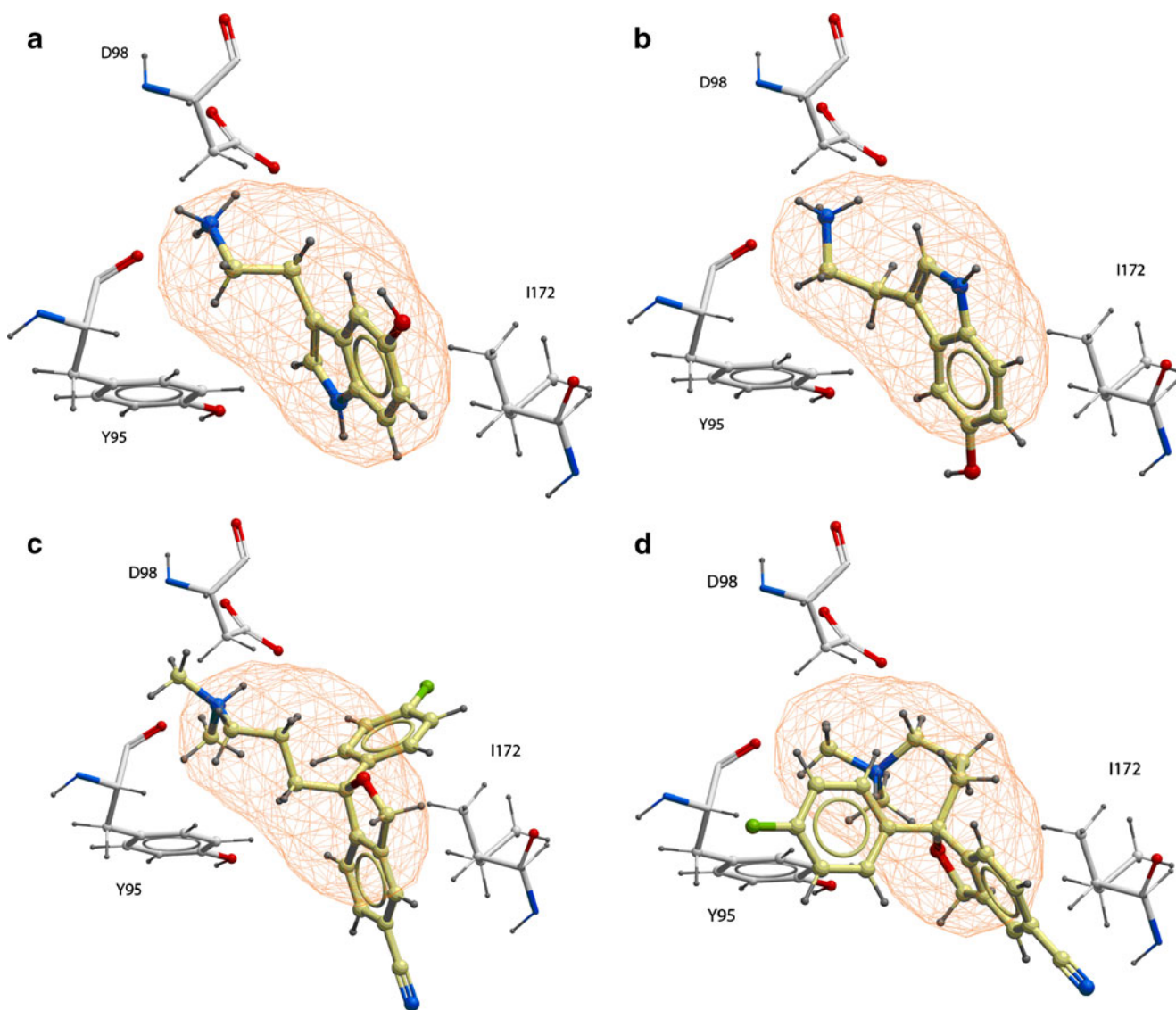


Fig. 1 Ligand binding modes detected through docking. **a** SERT–5-HT^A binding mode, **b** SERT–5-HT^B binding mode, **c** SERT–(*S*)-citalopram^A binding mode, and **d** SERT–(*S*)-citalopram^B binding mode. The side chains of amino acids Y95, D98 and I172 and the binding

pocket detected by ICM PocketFinder (red wire representation) are shown. Color coding of atoms in amino acids: red oxygen, blue nitrogen, gray carbon and hydrogen. Color coding of ligands: red oxygen, blue nitrogen, yellow carbon, gray hydrogen

Predictions of the 5-HT–SERT binding energies for the two binding modes using the calcBindingEnergy macro of ICM [36] showed that the poses represented by the SERT–5-HT^A complex had binding energies in the range -5.7 to -13.8 kcal mol⁻¹ (average -10.0 kcal mol⁻¹), while poses represented by the SERT–5-HT^B complex had binding energies in the range -4.8 to -10.7 kcal mol⁻¹ (average -8.1 kcal mol⁻¹).

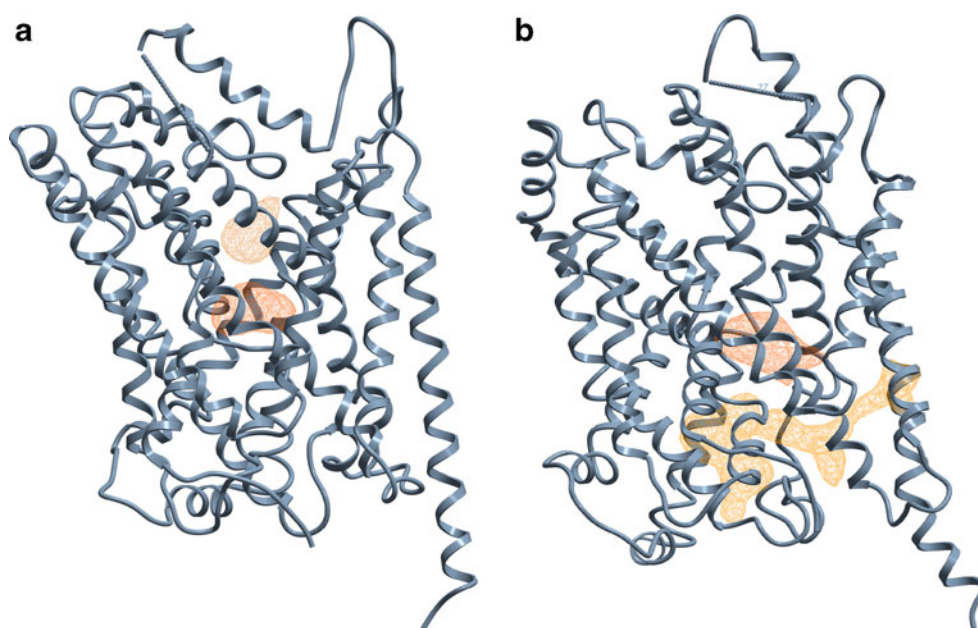
In the SERT–(*S*)-citalopram^A binding mode (Fig. 1), (*S*)-citalopram occupied all three subpockets of the putative substrate binding site. The amine moiety of (*S*)-citalopram was located in the ionic subpocket close to D98, whereas the cyanophthalane and fluorophenyl moieties were located in the hydrophobic (in close proximity to A169, A173, V343 and G442) and aromatic subpockets (pointing towards F335), respectively. The oxygen moiety of (*S*)-citalopram was pointing in the direction of Y95 (Fig. 1). In comparison, the cyanophthalane and amine moieties of (*S*)-citalopram in the SERT–(*S*)-citalopram^B binding mode were also found in the hydrophobic and ionic subpockets, respectively, in a very similar location to that in the SERT–(*S*)-citalopram^A binding mode. However, the fluorophenyl moiety of (*S*)-citalopram in this binding mode was found to be juxtaposed in-between the side chains of Y95 and S438, and the oxygen moiety was pointing in the direction of Y176 (Fig. 1). The prediction of binding energies using the calcBindingEnergy macro of ICM [36] showed that poses represented by the SERT–(*S*)-citalopram^A complex had binding energies in the range -7.4 to -19.1 kcal mol⁻¹ (average -14.7 kcal mol⁻¹), while those represented by the SERT–(*S*)-citalopram^B complex had binding energies in the range -12.7 to -19.7 kcal mol⁻¹ (average -16.4 kcal mol⁻¹).

Molecular dynamics simulations

In order to study possible conformational changes of SERT upon the binding of 5-HT (substrate) and (*S*)-citalopram (inhibitor), more than 20 ns of MD simulations were performed for each system: one representative SERT–ligand complex from each of the binding modes detected as well as apo-SERT were embedded in POPC lipid bilayers, followed by system equilibration and longer MD simulations. The average structures of each of the five complexes were then generated based on the last 10 ns of the production runs, and ICM PocketFinder was used to detect possible pockets that had formed in SERT during the production runs.

Interestingly, in the average structure of the SERT–5-HT^B binding mode, the substrate binding pocket began to elongate towards the cytoplasm, and another pocket started to form that extended from the cytoplasm up towards the elongated substrate binding pocket during the MD simulation (Fig. 2). Our results showed that in the average structure of SERT–5-HT^B, only a narrow stretch of TMs 6 and 8, in addition to intracellular loop 1 (IL1), separated the two pockets and prevented access from the substrate binding site to cytoplasm (Fig. 3). The other simulations also changed the size of the substrate binding site and induced other pockets to form; however, intracellular vestibules similar to that generated in the SERT–5-HT^B complex were not observed in any of the other average structures (results not shown). Based on these observations, the simulation of the SERT–5-HT^B complex was prolonged to 49 ns. The prolongation indicated that the pocket extending from the cytoplasm up towards the elongated substrate binding pocket was also maintained during 21 to 49 ns of the MD simulation.

Fig. 2 SERT structures. **a** Initial SERT structure and **b** the average SERT–5-HT^B structure generated based on the last 10 ns of the MD simulation. “Intra-structural” pockets detected by ICM PocketFinder are shown. The putative substrate binding pocket is represented as *red wire*



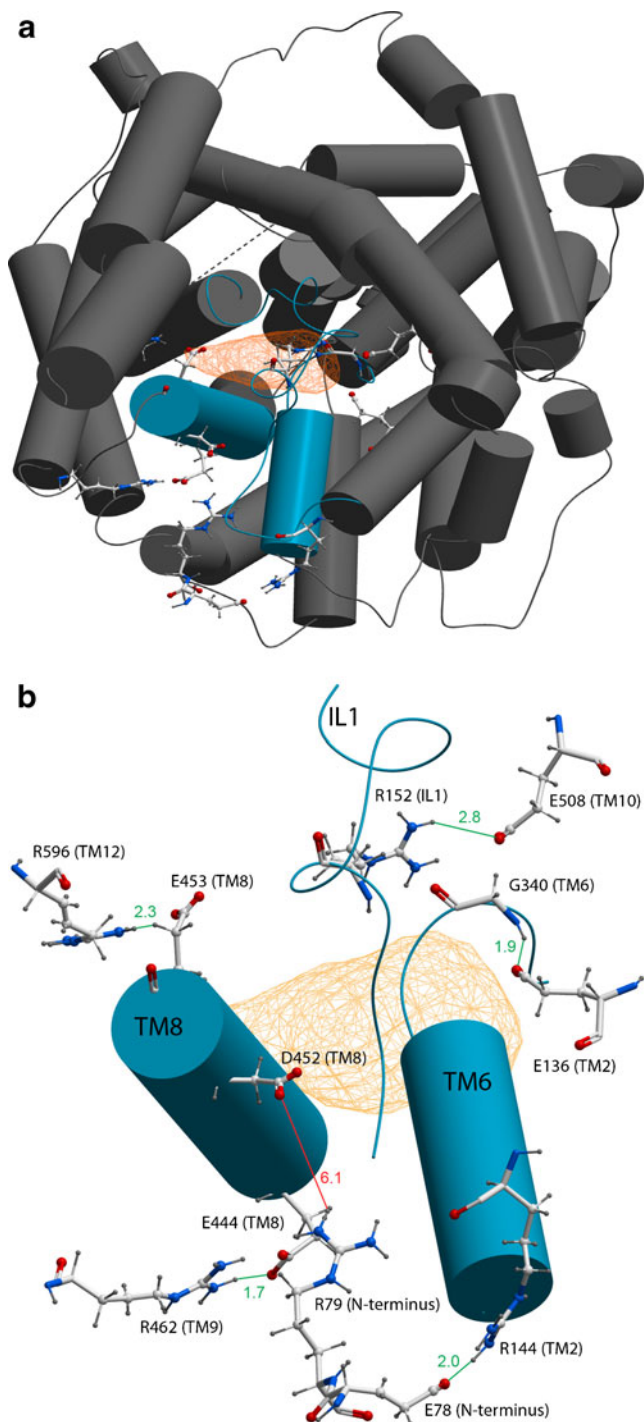


Fig. 3 **a** Intracellular view of the average SERT–5-HT^B structure. SERT C α carbon atoms are shown in *gray cylindrical representation*. For clarity, amino acids 148–160, 338–350 and 444–453 are shown in *blue*. The putative substrate binding site is displayed as *red wire*. Amino acids that are proposed to play a role in the opening of a vestibule extending from the putative substrate binding site (*red wire representation*) to the cytoplasm are shown as *xstick*. **b** Close-up of **a** with residues in *xstick*. *Green lines* show interactions formed during the simulation; *red line* shows an interaction broken during simulation

The 5-HT in the average SERT–5-HT^B structure (12–21 ns) was slightly shifted compared with the initial structure (Fig. 4). Superimposition of the structure of SERT prior to MD and the average structure of the SERT–5-HT^B complex showed that the hydroxyl oxygen atom of 5-HT was located closer to the Y95 (TM1) hydroxyl group. The distance before MD was 4.1 Å, while the distance in the average structure was 3.4 Å (range 1.9–5.5 Å). 5-HT was also located 1.7 Å closer to the cytoplasmic side than before MD. The distance between the G338 (TM6) backbone oxygen and the Y95 (TM1) hydroxyl group also increased slightly, from 1.8 Å to 2.1 Å in the average structure (range 2.0–3.0 Å), indicating that TMs 1 and 6 had begun to move further apart as well (Fig. 4). Prolongation of the MD indicated that these distances did not change much during 21–49 ns of MD. The distance between the 5-HT hydroxyl group and the hydroxyl group of Y95 varied between 2.3 and 5.3 Å, while the distance between the G338 backbone oxygen and the Y95 hydroxyl group varied between 1.8 and 2.7 Å.

The observation that only some residues block the access from the putative substrate binding site to the cytoplasm prompted us to look for amino acids in the unwound region of TM6, in TM8, and in IL1 of SERT that may have interacted with amino acids in other regions of SERT and contributed to the formation of the emerging vestibule. We found G340 in TM6 and E444, D452 and E453 in TM8, as well as R152 and K153 in IL1 very interesting

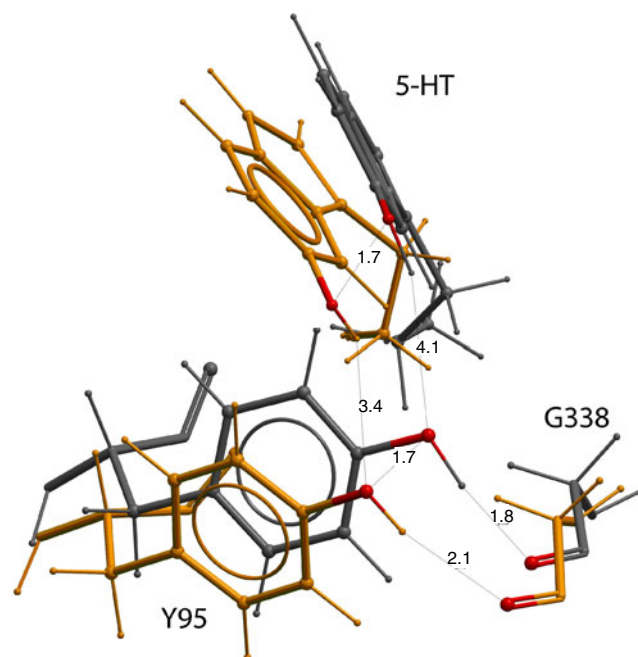


Fig. 4 Comparison of the 5-HT binding mode in the initial SERT–5-HT^B complex (*gray*) and that in the average SERT–5-HT^B structure generated based on the last 10 ns of MD (*orange*). Atomic distances (Å) are shown as *dotted lines*. For clarity, selected hydroxyl oxygen atoms on 5-HT, Y95 and G338 are colored *red*

Table 2 Atomic distances [Å] between amino acids that were proposed to play a role in the opening of a vestibule from the SERT substrate binding site to the cytoplasm. Locations of amino acids are shown in parentheses

Distance [location]	Initial SERT	SERT (no ligand)	SERT-5-HT ^A	SERT-5-HT ^B	SERT-(S)-citalopram ^A	SERT-(S)-citalopram ^B
E78–R144 [N-terminus: TM2/IL1]	7.0	9.3	13.7	1.9	14.6	6.3
R79–D452 [N-terminus: TM8]	1.7	1.7	1.7	6.1	1.8	8.2
R152–E453 [IL1–TM8]	1.7	4.3	1.8	9.0	3.0	2.0
R152–E508 [IL1–TM10]	17.2	11.5	8.6	2.7	11.4	14.7
E136–G340 [TM2–TM6]	1.8	2.7	1.8	1.9	2.1	1.9
E444–R462 [TM8–TM9]	6.6	1.7	1.7	1.7	1.7	1.8

in this respect. The distances between these residues and their interaction partners in the structure of SERT before the MD simulations and in the average structures generated following the MD simulations were thus measured and compared (Table 2).

We also noted that the cytoplasmic part of TM3 (K159–I168) had unwound during the MD simulation and had thus become more flexible. The unwinding may have played a role in the opening of the vestibule; however, this unwinding was seen in all average structures and may be an artifact of poor force field representation of protein–protein, protein–solvent and solvent–lipid interactions. Using CHARMM force fields and NPT for simulations in a tensionless ensemble may lead to the condensation of the bilayer to a near gel-like state, which may influence the protein structure and result in incorrect predictions if the lateral density of lipids increases beyond a liquid crystalline state [37]. The unwinding may also be a result of structural differences between SERT and LeuT in this region [3].

Structural differences between SERT and LeuT in IL1 may explain the unwinding of the α -helical structure in IL1 that was present in the initial structure of SERT, just as in LeuT [4], but not in any of the average structures generated following the MD simulations. The homology between SERT and LeuT in this region is very low, with only one identical amino acid (I154, SERT numbering) [3], and the presence of an α -helical structure in IL1 of SERT is thus questionable.

Discussion

Homology modeling and docking

The homology modeling approach is a valuable tool for investigating protein structures when experimental structures are lacking. Homology models are useful for predicting ligand potency and specificity through the use of different docking approaches, and high-quality homology models have also been used in the study of conformational changes using MD

simulations [38]. In the present study, 5-HT and ten other tryptamine derivatives (SERT substrates) and the SSRI (S)-citalopram were docked into the putative substrate binding site of a SERT homology model, and possible conformational changes of SERT upon ligand binding were studied by MD simulations.

The accuracy of homology models depends on three factors: the sequence identity and functional similarity between the template and target proteins; the amino acid sequence alignments between the template and the targets; and the resolution at which the crystal structure of the template protein was resolved. For membrane proteins in general, sequence identities between template and target proteins of 50% have been found to yield membrane homology models with a C α -RMSD of approximately 1 Å from the template structure in the transmembrane regions, assuming that the template structure has been solved at a resolution of 3.5 Å or better [39]. Sequence identities of 30% or more are, for most membrane proteins, predicted to yield acceptable homology models with a C α -RMSD of approximately 2 Å in the TM regions [39].

The sequence identity between LeuT and SERT is approximately 50% in the putative substrate binding site detected by the ICM PocketFinder. In contrast, the overall sequence identity between the transporters is less than 20%, but it rises to approximately 35% in TMs that are predicted to be directly involved in substrate binding (i.e., TMs 1, 3, 6 and 8). LeuT is considered a good template for generating homology models of SERT that can be used for ligand docking and molecular dynamics. Actually, due to the topological restrictions provided by the hydrophobic membrane environment surroundings, membrane proteins such as SERT actually have more limited ways of folding than water-soluble proteins, which may suggest that membrane protein homology models are more accurate than homology models of water-soluble proteins at the same level of sequence identity [39]. This also thus supports the generation of acceptable homology models of not only the SERT substrate binding site but the whole structure using LeuT as a template.

Our docking results suggest two different ways 5-HT and the other tryptamine derivatives may bind in SERT: the SERT–5-HT^A and SERT–5-HT^B binding modes (Fig. 1). In both of these binding modes, the positively charged amine moiety of 5-HT was in the vicinity of the negatively charged D98 side chain, and the C6 position of the indole ring was located close to A173 at the other end of the molecule; however, the indole nitrogen moiety pointed in different directions in the two binding modes. Interestingly, similar binding modes of 5-HT to the SERT–5-HT^A and SERT–5-HT^B binding modes have also been obtained through docking and experimental studies by other groups [10, 12, 35]. Celik et al. [10] found that the C5 and C7 positions of 5-HT should be located in hydrophilic and hydrophobic pockets of SERT, and that the 5 hydroxyl moiety of 5-HT was in the vicinity of T439 (TM8) [10]. Though the C5 and C7 moieties of 5-HT in both the SERT–5-HT^A and SERT–5-HT^B binding modes described here are located in such regions, only the localization of C5 of 5-HT in the SERT–5-HT^A binding mode was found in the vicinity of T439. In another study, however, 5-HT in a similar binding mode to the SERT–5-HT^B binding mode showed good correlation with experimental data and was also found to best describe the cross-species sensitivities reported in support vector machine (SVM) sensitivity maps generated for the human and *Drosophila melanogaster* serotonin transporters [12]. This binding mode was also suggested by Jørgensen et al. [35].

Our results show that the size of the putative substrate binding site detected in this structure of SERT was relatively small and not optimal for the docking of larger compounds such as (*S*)-citalopram. Nonetheless, the binding mode of (*S*)-citalopram has recently been studied by docking into occluded SERT homology models and by experimental site-directed mutagenesis [18]. Andersen et al. [18] found that the fluorophenyl moiety of (*S*)-citalopram was located near I172, A173 and N177, whereas the cyanophthalane moiety was in proximity to V343. Though the cyanophthalane moiety of (*S*)-citalopram in both binding modes in the present study was in the vicinity of V343, only the fluorophenyl of (*S*)-citalopram in the SERT–(*S*)-citalopram^A binding mode was in the vicinity of I172 (Fig. 1). A similar (*S*)-citalopram binding mode to the SERT–(*S*)-citalopram^A binding mode has also been used as initial binding mode in another MD study in SERT [35].

Our docking indicated that the tryptamine derivatives do not interact with SERT in the aromatic subpocket of the binding pocket, whereas (*S*)-citalopram does. A possible mechanism of action of inhibition by (*S*)-citalopram may therefore be that (*S*)-citalopram interferes with the closure of the extracellular gating residues Y176 and F335, stabilizing SERT in an outward-facing conformation,

thereby hindering conformational changes needed for transport to occur. A similar mechanism of inhibition has recently been suggested for TCAs [40].

Molecular dynamics simulations

In order to gain insights into SERT conformational changes that may take place upon ligand binding, one representative ligand orientation from each of the two possible binding modes of 5-HT (representing the tryptamine derivatives) and (*S*)-citalopram, as well as the apo-SERT structure, were selected for MD simulations in POPC lipid bilayers. The simulations were run for 22 ns (SERT–5-HT^A), 49 ns (SERT–5-HT^B), 32 ns (SERT–(*S*)-citalopram^A), 23 ns (SERT–(*S*)-citalopram^B) and 25 ns (apo-SERT), and average structures of each of the five MD simulations were generated and used to analyze the results. Average structures may represent unphysical states of SERT that may not exist. However, the present average structures were based on the last 10 ns of the MD simulation, where energetically favorable and structural stable SERT–(ligand)–POPC complexes were obtained. The average structures used were thus considered to be representative of the most densely populated conformations during this period of the simulation.

The substrate 5-HT is expected to cause a different conformational change of SERT than inhibitors such as (*S*)-citalopram, as the former compound is transported whereas the latter inhibits transport. In order to visualize such conformational changes, the ICM PocketFinder was used to detect pockets in the five average structures. In the average structure from SERT–5HT^B binding mode simulation, the pockets detected showed that a vestibule had started to emerge that extended from the putative substrate binding site towards the cytoplasm (Fig. 2). The results suggested that the continued rearrangement of the unwound regions of TM6, TM8 and IL1 relative to one another may open a pathway from the substrate binding site to the cytoplasm (Fig. 3). A similar vestibule was not observed in any of the other simulations (results not shown).

A pocket extending from the cytoplasm up towards the substrate binding pocket was formed during the MD simulation of the SERT–5-HT^B complex. A corresponding pocket was not formed during MD of the SERT–5-HT^A complex. Based on these observations, we also examined whether the position of 5-HT changed during the simulation of the SERT–5-HT^B complex. By superimposing the initial structure of SERT on the average SERT–5-HT^B structure (12–21 ns), we found that the 5-HT hydroxyl group was located closer to the Y95 (TM1) hydroxyl group at the cytoplasmic end of the binding pocket in the average SERT–5-HT^B structure. In addition, the atomic distance between Y95 (TM1) and G338 (TM6) was slightly

increased (Fig. 4). Prolonging the MD simulation up to 49 ns showed that these distances were maintained between 21 and 49 ns of MD simulation, and additional changes in SERT structure or in 5-HT position were not seen.

The hydroxyl group of Y95 (TM1) and the backbone oxygen atom of G338 in the unwound region of TM6 were within hydrogen-bonding distance in the initial structure of SERT, and this interaction might play a role in keeping the translocation pathway closed. Our results thus suggest that one of the first steps in 5-HT translocation is the formation of a hydrogen bond between the 5-OH of 5-HT and Y95 (TM1), which may sever the hydrogen bond between Y95 (TM1) and G338 (TM6). In another study, the mutation of G338 to cysteine (G338C) was shown to stabilize SERT in an outward-facing conformation [33]. The transport activity of the G338C mutant was less than 5% of the wild-type transport activity; however, transport could partially be restored by simultaneously mutating Y95 to phenylalanine (Y95F), which indicates that Y95 (TM1) and G338 (TM6) cannot be hydrogen bonded for 5-HT transport to occur [33].

The amino acids in TM6 that separated the putative substrate binding site from the cytoplasmic vestibule were located in the unwound region of TM6, which in the initial SERT structure consisted of G338, P339, G340, F341 and G342, but in the average SERT–5-HT^B structure also contained two more amino acids, S336 and L337. The unwinding of the latter amino acids is in agreement with a study suggesting that amino acids 334–337 in SERT are in an unwound region based on aqueous accessibility data [33]. This region contains several glycine residues [3] and is thus expected to be very flexible: one study shows that even the conservative mutations of G338 and G342 to alanine (G338A and G342A, respectively) cause reductions in 5-HT transport of approximately 28% and 10%, respectively, as compared to the wild type [33].

The transmembrane helix closest to TM6 in the model was TM2. Thus, an interaction between the unwound region of TM6 and amino acids in TM2 might contribute to opening up the binding site towards the intracellular region by pulling the flexible unwound part of TM6 towards TM2. We observed that a hydrogen bond was present between the backbone of G340 (unwound region of TM6) and the side chain of E136 (TM2), as in LeuT [4]. Our results show that the distance between the backbone nitrogen of G340 and the E136 side chain did not change significantly during the MD simulation of the SERT–5-HT^B complex (Table 2); however, superimposing the average structure on the initial SERT structure showed that the G340 backbone nitrogen atom and the E136 carboxyl carbon atoms shifted 2.5 Å during the simulation (results not shown). Hence, though the distance between G340 and E136 remains constant during the MD simulation, the unwound TM6 region

and TM2 had moved 2.5 Å in the same direction, away from the putative substrate binding site. An ionic interaction between another TM2 amino acid, R144, and E78 in the N-terminus also formed, and may have contributed to the joint movement of TMs 2 and 6. E136 (TM2) is conserved among the Na⁺-dependent NSS transporters [3], and has been shown to be very important for transport in SERT: a conservative mutation of this glutamic acid to aspartic acid (E136D) causes a reduction in SERT transport, and mutations to alanine or glutamine (E136A, E136Q) inhibit transport [41]. The atomic distance between R144 (TM2) and E78 (N-terminus) decreased from 7 Å in the initial structure of SERT to 1.9 Å in the average structure of SERT–5-HT^B (Table 2).

In TM8, three amino acids were found to be particularly interesting with respect to opening an intracellular vestibule from the putative substrate binding site to the cytoplasm: namely E444, D452 and E453. E444 (TM8) was located in close proximity to the substrate binding site, and during all MD simulations an ionic interaction between E444 (TM8) and R462 (TM9) was formed (Table 2). D452 and E453 were located at the cytoplasmic end of the TM8. During the MD simulation of the SERT–5-HT^B complex, we observed that the distance between E453 (TM8) and R152 (IL1) increased whereas the distance between D452 (TM8) and K153 (IL1) decreased, thus changing the conformation of this long loop. The importance of R152 for transport is in agreement with a recent study in mouse SERT showing that the G39/K152 phenotype has reduced transport in comparison with the wild type (E39/R152 phenotype) [42].

Very interestingly, we observed that during the MD simulation of SERT–5-HT^B, an interaction between R152 (IL1) and E508 (TM10) developed. In the initial structure of SERT, the atomic distance between these residues was >17 Å, while the distance decreased to only 2.7 Å in the average SERT–5-HT^B structure (Table 2). Furthermore, this interaction was not formed in any of the other MD simulations (Table 2). E508 is one of a few amino acids in TM10 that are fully conserved between SERT and LeuT [3]. Interestingly, E508 (TM10) was also located in the region of E136 (TM2) in SERT, and it is suggested that this amino acid interacts with G340 in the unwound region of TM6 (see above); it is also known to be important for transport in SERT [41].

Summary

Our MD simulations indicate that the SERT–5-HT^B binding mode and not the SERT–5-HT^A binding mode induces conformational changes in SERT that may be associated with substrate translocation. The simulations suggest that substrate translocation may involve forming and breaking ionic interactions between TM6, TM8 and IL1 and their

interaction partners. Although our observations are in agreement with experimental studies, the suggested mechanism is hypothetical, as it is based solely upon theoretical calculations using a homology-based model.

The simulations may indicate that the formation of a hydrogen bond between Y95 in TM1 and 5-HT causes a hydrogen bond between Y95 and G338 in TM6 to be broken, enabling the unwound region of TM6 to move away from the substrate binding site and transport to begin. The formation of an ionic interaction between R144 (TM2) and E78 (N-terminus) and the interaction between G340 (unwound region of TM6) and E136 (TM2) then cause TM6 to move away from the putative substrate binding site. The movements of E136 (TM2) also affect the nearby amino acid E508 (TM10), causing it to interact with R152 in IL1, thus changing the conformation of this loop. Simultaneously, an ionic interaction between E444 (TM8) and R462 (TM9), located close to the putative substrate binding site, is formed. The interaction between E453 in the cytoplasmic part of TM8 and R596 in TM12 may also contribute to relocating TM8 away from the vestibule. The formation of an ionic interaction between E78 in the N-terminus and R144 in TM2, and the subsequent movement of TM2, may also weaken the interaction between the N-terminus and TM8, as illustrated by the increase in the R79–D452 distance (Table 2).

Acknowledgments This work was supported by a grant from the Neuron program of the Research Council of Norway (project 176956), by the Polish–Norwegian Research Fund (grant PNRF-103-A1-1/07), and by CPU hours from NOTUR (Norwegian Metacenter for Computational Science). NAMD was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Saier MH Jr (2000) A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol Mol Biol Rev* 64:354–411
- Chen NH, Reith ME, Quick MW (2004) Synaptic uptake and beyond: the sodium- and chloride-dependent neurotransmitter transporter family SLC6. *Pflugers Arch* 447:519–531
- Beuming T, Shi L, Javitch JA, Weinstein H (2006) A comprehensive structure-based alignment of prokaryotic and eukaryotic neurotransmitter/Na⁺ symporters (NSS) aids in the use of the LeuT structure to probe NSS structure and function. *Mol Pharmacol* 70:1630–1642
- Yamashita A, Singh SK, Kawate T, Jin Y, Gouaux E (2005) Crystal structure of a bacterial homologue of Na⁺/Cl⁻-dependent neurotransmitter transporters. *Nature* 437:215–223
- Singh SK, Yamashita A, Gouaux E (2007) Antidepressant binding site in a bacterial homologue of neurotransmitter transporters. *Nature* 448:952–956
- Zhou Z, Zhen J, Karpowich NK, Goetz RM, Law CJ, Reith ME, Wang DN (2007) LeuT-desipramine structure reveals how antidepressants block neurotransmitter reuptake. *Science* 317:1390–1393
- Zhou Z, Zhen J, Karpowich NK, Law CJ, Reith ME, Wang DN (2009) Antidepressant specificity of serotonin transporter suggested by three LeuT-SSRI structures. *Nat Struct Mol Biol* 16:652–657
- Singh SK, Piscitelli CL, Yamashita A, Gouaux E (2008) A competitive inhibitor traps LeuT in an open-to-out conformation. *Science* 322:1655–1661
- Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* 5:823–826
- Celik L, Sinning S, Severinsen K, Hansen CG, Moller MS, Bols M, Wiborg O, Schiott B (2008) Binding of serotonin to the human serotonin transporter. Molecular modeling and experimental validation. *J Am Chem Soc* 130:3853–3865
- Forrest LR, Tavoulari S, Zhang YW, Rudnick G, Honig B (2007) Identification of a chloride ion binding site in Na⁺/Cl⁻-dependent transporters. *Proc Natl Acad Sci USA* 104:12761–12766
- Kaufmann KW, Dawson ES, Henry LK, Field JR, Blakely RD, Meiler J (2009) Structural determinants of species-selective substrate recognition in human and *Drosophila* serotonin transporters revealed through computational docking studies. *Proteins* 74:630–642
- Jardetzky O (1966) Simple allosteric model for membrane pumps. *Nature* 211:969–970
- Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 32:D115–119
- Abagyan RT, Kuznetsov D (1994) ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J Comput Chem* 15:488–506
- Zomot E, Bendahan A, Quick M, Zhao Y, Javitch JA, Kanner BI (2007) Mechanism of chloride interaction with neurotransmitter: sodium symporters. *Nature* 449:726–730
- Abagyan R, Totrov M (1994) Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol* 235:983–1002
- Andersen J, Olsen L, Hansen KB, Taboureau O, Jorgensen FS, Jorgensen AM, Bang-Andersen B, Egebjerg J, Stromgaard K, Kristensen AS (2010) Mutational mapping and modeling of the binding site for (S)-citalopram in the human serotonin transporter. *J Biol Chem* 285:2051–2063
- Andersen J, Taboureau O, Hansen KB, Olsen L, Egebjerg J, Stromgaard K, Kristensen AS (2009) Location of the antidepressant binding site in the serotonin transporter: importance of Ser-438 in recognition of citalopram and tricyclic antidepressants. *J Biol Chem* 284:10276–10284
- Barker EL, Moore KR, Rakhshan F, Blakely RD (1999) Transmembrane domain I contributes to the permeation pathway for serotonin and ions in the serotonin transporter. *J Neurosci* 19:4705–4717
- An J, Totrov M, Abagyan R (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol Cell Proteomics* 4:752–761
- Nemethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA (1992) Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the Ecepp/3 algorithm, with application to proline-containing peptides. *J Phys Chem* 96:6472–6484
- Jo S, Kim T, Im W (2007) Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PLoS ONE* 2:e880

24. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI (2006) OPM: orientations of proteins in membranes database. *Bioinformatics* 22:623–625
25. 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
26. Mackerell AD Jr, Bashford D, Bellot M, Dunbrack RL Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Leu FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reither WE III, 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
27. Mackerell AD Jr, Feig M, Brooks CL III (2004) Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* 25:1400–1415
28. Feller SE, MacKerell AD (2000) An improved empirical potential energy function for molecular simulations of phospholipids. *J Phys Chem B* 104:7510–7515
29. Feller SE, Gawrisch K, MacKerell AD Jr (2002) Polyunsaturated fatty acids in lipid bilayers: intrinsic and environmental contributions to their unique physical properties. *J Am Chem Soc* 124:318–326
30. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, Mackerell AD Jr (2009) CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem* 31:671–690
31. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14(33–38):27–38
32. Adkins EM, Barker EL, Blakely RD (2001) Interactions of tryptamine derivatives with serotonin transporter species variants implicate transmembrane domain I in substrate recognition. *Mol Pharmacol* 59:514–523
33. Field JR, Henry LK, Blakely RD (2010) Transmembrane domain 6 of the human serotonin transporter contributes to an aqueously accessible binding pocket for serotonin and the psychostimulant 3,4-methylene dioxymethamphetamine. *J Biol Chem* 285:11270–11280
34. Walline CC, Nichols DE, Carroll FI, Barker EL (2008) Comparative molecular field analysis using selectivity fields reveals residues in the third transmembrane helix of the serotonin transporter associated with substrate and antagonist recognition. *J Pharmacol Exp Ther* 325:791–800
35. Jorgensen AM, Tagmose L, Jorgensen AM, Bogeso KP, Peters GH (2007) Molecular dynamics simulations of Na⁺/Cl⁻-dependent neurotransmitter transporters in a membrane-aqueous system. *Chem Med Chem* 2:827–840
36. Schapira M, Totrov M, Abagyan R (1999) Prediction of the binding energy for small molecules, peptides and proteins. *J Mol Recognit* 12:177–190
37. Benz RW, Castro-Roman F, Tobias DJ, White SH (2005) Experimental validation of molecular dynamics simulations of lipid bilayers: a new approach. *Biophys J* 88:805–817
38. Vashisth H, Abrams CF (2010) All-atom structural models for complexes of insulin-like growth factors IGF1 and IGF2 with their cognate receptor. *J Mol Biol* 400:645–658
39. Forrest LR, Tang CL, Honig B (2006) On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys J* 91:508–517
40. Sinning S, Musgaard M, Jensen M, Severinsen K, Celik L, Koldso H, Meyer T, Bols M, Jensen HH, Schiott B, Wiborg O (2009) Binding and orientation of tricyclic antidepressants within the central substrate site of the human serotonin transporter. *J Biol Chem* 285:8363–8374
41. Korkhov VM, Holy M, Freissmuth M, Sitte HH (2006) The conserved glutamate (Glu136) in transmembrane domain 2 of the serotonin transporter is required for the conformational switch in the transport cycle. *J Biol Chem* 281:13439–13448
42. Carneiro AM, Airey DC, Thompson B, Zhu CB, Lu L, Chesler EJ, Erikson KM, Blakely RD (2009) Functional coding variation in recombinant inbred mouse lines reveals multiple serotonin transporter-associated phenotypes. *Proc Natl Acad Sci USA* 106:2047–2052

Replica exchange molecular dynamics simulation of structure variation from $\alpha/4\beta$ -fold to 3α -fold protein

Raudah Lazim · Ye Mei · Dawei Zhang

Received: 20 December 2010 / Accepted: 19 May 2011 / Published online: 14 June 2011
© Springer-Verlag 2011

Abstract Replica exchange molecular dynamics (REMD) simulation provides an efficient conformational sampling tool for the study of protein folding. In this study, we explore the mechanism directing the structure variation from $\alpha/4\beta$ -fold protein to 3α -fold protein after mutation by conducting REMD simulation on 42 replicas with temperatures ranging from 270 K to 710 K. The simulation began from a protein possessing the primary structure of GA88 but the tertiary structure of GB88, two G proteins with “high sequence identity.” Albeit the large C α -root mean square deviation (RMSD) of the folded protein (4.34 Å at 270 K and 4.75 Å at 304 K), a variation in tertiary structure was observed. Together with the analysis of secondary structure assignment, cluster analysis and principal component, it provides insights to the folding and unfolding pathway of 3α -fold protein and $\alpha/4\beta$ -fold protein respectively paving the way toward the understanding of the ongoings during conformational variation.

Keywords Cluster analysis · Conformational variation · Principal component analysis · Replica exchange molecular dynamics · Root mean square deviation · Secondary structure assignment

Introduction

Protein folding intrigues many and this has led to a multitude of studies, both experimental and theoretical, to decode the “protein folding problem” [1–4]. Comprehending and predicting the tertiary structure of proteins from information encompassed within the primary structure has been one of the many obstacles faced in structural biology which a great deal of researchers had tried to overcome [3, 4]. Rose et al. offered an alternative method of conquering the folding problem by questioning the specificity aspect in terms of amino acid compositions which influences a protein’s propensity for one fold over the other [5]. This insight had prompted numerous researchers to engineer a pair of protein with “high amino acid sequence identity” but different native folds with the highest sequence identity documented at 95% [2, 4, 6]. The main purpose of designing these proteins is to determine the minimum number of amino acids responsible for the protein’s preference to exist in one conformation over the other. At 95% sequence identity, Alexander et al. (2009) had established the correlation between single point mutation and conformational variation [2]. This ability of protein to undergo variation from one conformation to the other are keynotes in many research efforts due to vital associations to the understanding of protein misfolding which forms the root of numerous diseases such as Alzheimer’s disease [7–9].

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1147-8) contains supplementary material, which is available to authorized users.

R. Lazim · D. Zhang (✉)
Division of Chemistry and Biological Chemistry,
School of Physical and Mathematical Sciences,
Nanyang Technological University,
Singapore 637371, Singapore
e-mail: zhangdw@ntu.edu.sg

Y. Mei
State Key Laboratory of Precision Spectroscopy,
East China Normal University,
Shanghai 200062, China
e-mail: ymei@phy.ecnu.edu.cn

Efforts to decipher protein folding with the help of molecular dynamics (MD) simulation had gained a fair amount of scrutiny since the start of MD study of biological macromolecules in 1977 [10, 11]. MD simulation provides a strong platform for the study of protein dynamics and enables one to apprehend the conformational transformation leading to the folding or unfolding of the protein of interest [3, 12, 13]. MD simulations performed to scrutinize protein folding thus far had reached an accuracy of $<1 \text{ \AA}$ C_α root mean square deviation (RMSD) relative to crystal structure which was achieved through a folding study of villin headpiece subdomain through replica exchange molecular dynamics (REMD) simulation by Duan et al. [11].

REMD simulation, compared to regular MD simulation, is a more efficient conformational sampling method which allows protein replicas to perform random walks across temperature space without being trapped in the abundance of energy local minima at low temperatures [11, 14, 15]. In this work, our study focuses on the utilization of REMD simulations to gain insights on the conformational variation from $\alpha/4\beta$ -fold protein to 3α -fold protein. Two proteins of “high sequence identity” but different native folds namely GA88 and GB88, which have a sequence identity of 88% through 24 mutations and 17 mutations of the wild-type proteins respectively, were used for this experiment [4]. The protein GA88 which has a native 3α -fold configuration is the main point of our study and GB88 was merely used as a template for the construction of GA88 with a non-native $\alpha/4\beta$ -fold [4]. GA88 is comprised of three α -helices; H1 (residue 9-23), H2 (residue 27-34), and H3 (residue 39-51) [4]. On the other hand, GB88 comprises of one α -helix (HB, residue 23-26) and four β -strands (B1, residue 1-8; B2, residue 13-20; B3, residue 42-46 and B4, residue 51-55) [4].

REMD simulation was performed to ferret out the mechanism dictating the folding of GA88 into the 3α -fold configuration from the $\alpha/4\beta$ -fold. Instead of the usual *ab initio* folding and thermal unfolding studies of proteins, the REMD simulation was executed starting from $\alpha/4\beta$ -fold configuration which was derived from GB88. Even though the tertiary conformation of the starting structure is an exact parallel of the tertiary structure of GB88, the amino acid sequence is analogous to GA88. This is to ensure unerring conformational transition from a protein with one α -helix domain and four β -strand domains to a protein with three α -helix domains thus capacitating the study of the folding and unfolding mechanism involved during the conformational variation from $\alpha/4\beta$ -fold to 3α -fold. In this paper, the protein with the tertiary structure of GB88 but the primary sequence of GA88 will be coined as $\alpha/4\beta$ -GA88 and the protein with the wild-type fold of GA88 as 3α -GA88.

Methodology

REMD simulation of 42 replicas across 42 temperatures ranging from 270 to 710 K was run using “multisander” in AMBER 10 simulation package over total simulation time of 75 ns per replica [16]. Temperature distribution across the range of 270 to 710 K was carefully selected to attain a targeted acceptance ratio of 0.20. AMBER03 force field and generalized Born (GB) model were applied to describe the protein and the solvation effect respectively [17, 18]. Non-polar solvation term, which is often calculated proportional to the surface area, was not included because it is thought to over stabilize β structure in both Poisson-Boltzmann and generalized Born models [19, 20]. The starting structure was prepared by obtaining the NMR structure of GB88 from Protein Data Bank (PDB) with PDB ID of 2JWU and the following mutations, A24G, T25I, F30I, Y33I, Y45L, T49I and K50L, were done using LEaP module in AmberTools 1.2 to obtain the starting structure for the REMD simulation [4, 21]. The initial protein structure was minimized with the initial 500 cycles in steepest descent method and thereafter via conjugate gradient. The time step applied for the simulation is 2 fs. SHAKE algorithm was implemented to constrain all bonds with hydrogen atoms and non-bonded interactions were curtailed at 12 \AA [22]. During the REMD simulation, the replicas are initially heated for 100 ps to their intended temperatures using Langevin thermostat with collision frequency of 4 ps^{-1} [23]. Replica exchange was attempted every 10,000 steps.

The calculation of C_α – RMSD with NMR structure of GA88 (PDB id of 2JWS) as reference and cluster analysis was carried out using the “ptraj” program in AmberTools [4]. Similarly, principal components (PCs) were generated using “ptraj” and each PCs were visualized using Visual Molecular Dynamics (VMD) with the aid of interactive essential dynamics (IED) [24, 25].

Results and discussion

From the REMD trajectories, an acceptance ratio of more than 0.20 were observed indicating the absence of local energy minima trapping in the system and all the temperatures were explored numerous times by each replica during the course of simulation. The trajectories obtained from the REMD simulation were analyzed to reveal the mechanism governing the respective folding and unfolding of 3α -fold protein and $\alpha/4\beta$ -fold protein during conformational variation.

To evaluate whether global folding of simulated protein to a conformation analogous to GA88 had occurred during the simulation, the disparity in space between the simulated

structures and the NMR structure of GA88 when aligned were obtained by the calculation of C_{α} – RMSD over residues 9 to 53 (Fig. 1). The eight amino acid residues at the N terminus and the three amino acid residues at the C terminus were not included in the calculation as the NMR ensemble of the structures of GA88, as highlighted by He et al., are disarrayed at the aforementioned residual positions [4]. The lowest C_{α} – RMSD observed was 4.34 Å at 270 K and as seen from Fig. 1a, the C_{α} – RMSD calculated for the simulated structure at 270 K over the 75 ns simulation displays an overall decrease with time indicating the global folding of the simulated protein. This is within apprehension as GA88 had been reported to completely fold at 298 K [26]. A look at the C_{α} – RMSD calculation at 304 K showed a similar nonetheless less obvious downhill C_{α} – RMSD to that noted for 270 K with best C_{α} – RMSD of 4.75 Å (Fig. 1a), indicating that folding was still occurring at this temperature.

Large variations in C_{α} – RMSD and high probability distribution at large C_{α} – RMSDs were also evident in Fig. 1a. Therefore, to further corroborate the conformational variation from $\alpha/4\beta$ -GA88 to 3α -GA88, cluster analysis for the trajectory at 270 K was carried out to observe the conformational transformation that occurred during the 75 ns simulation. Based on the five clusters acquired, more than 50% of the trajectory comprise of structures with the

individual α -helical domains of 3α -GA88 folded (cluster 2, 4 and 5 in Fig. 2), confirming the variation in secondary structure from β -sheet to α -helix during the simulation.

A closer look at the folded structures at 270 K and 304 K revealed that additional α -helix were formed at the first eight residues (TTYKLILN) of the simulated structures (Fig. 1b and c) [4]. In addition, the first four residues within H3 (VEGV) forms a random coil instead of α -helix (Fig. 1b and c) [4]. From the prospect of helix propensity scale derived by Pace and Scholtz, the supra observations are valid as the average helical propensity of the first eight residue is 0.45 kcal mol⁻¹, comparable to part of H3 which folds into a helix (residue 43-51) with average helical propensity of 0.44 kcal mol⁻¹ [27]. However, it was highlighted by He et al. that the following mutations A6I, N7L and S8N which were made when designing GA88 from the parent protein had led to a net reduction in the helical propensity of GA88 at residues 6 to 8 which were originally an α -helix in the parent protein, thus favoring the random coil structure [4]. These contradicting observations imply a possible underlying bias of the force field toward helical structure. This was supported by studies conducted by Wang and Wade who underlined the favoring of AMBER03 force field toward helical structures upon unfolding of protein with β -sheet domains [17, 28, 29]. On the other hand, part of H3 which fails to

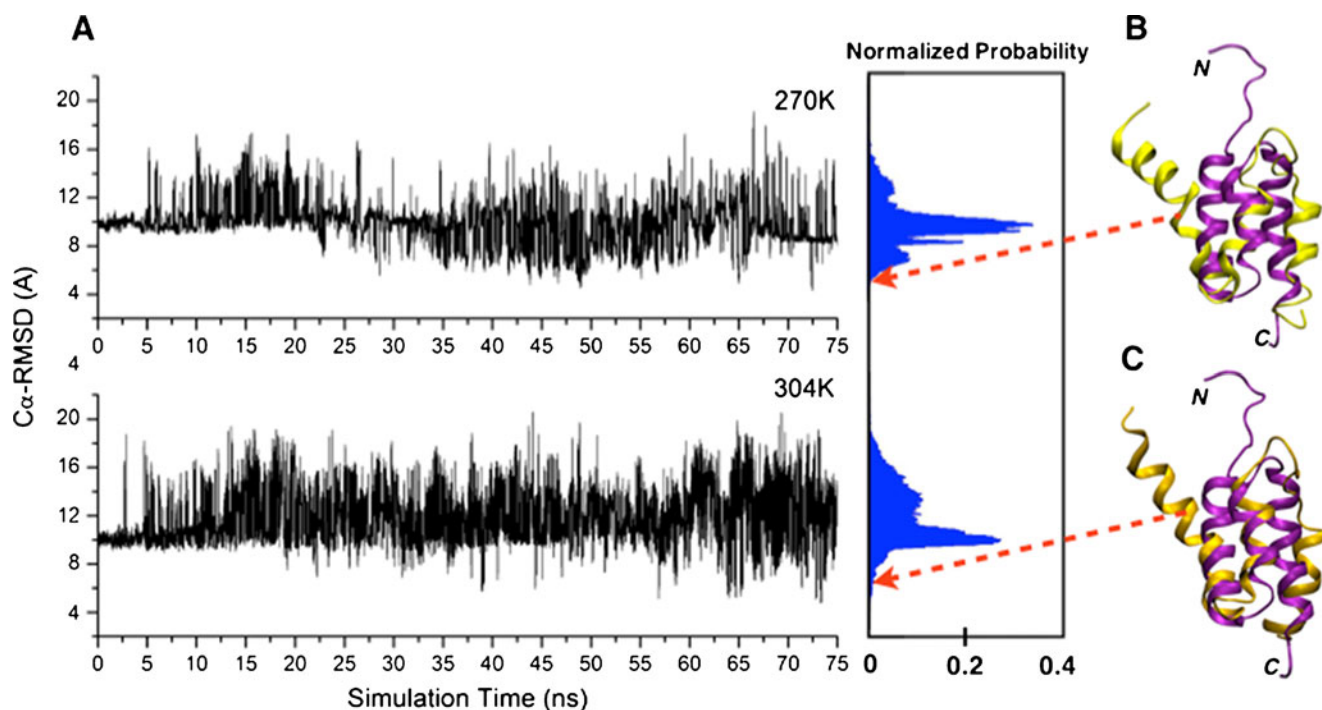
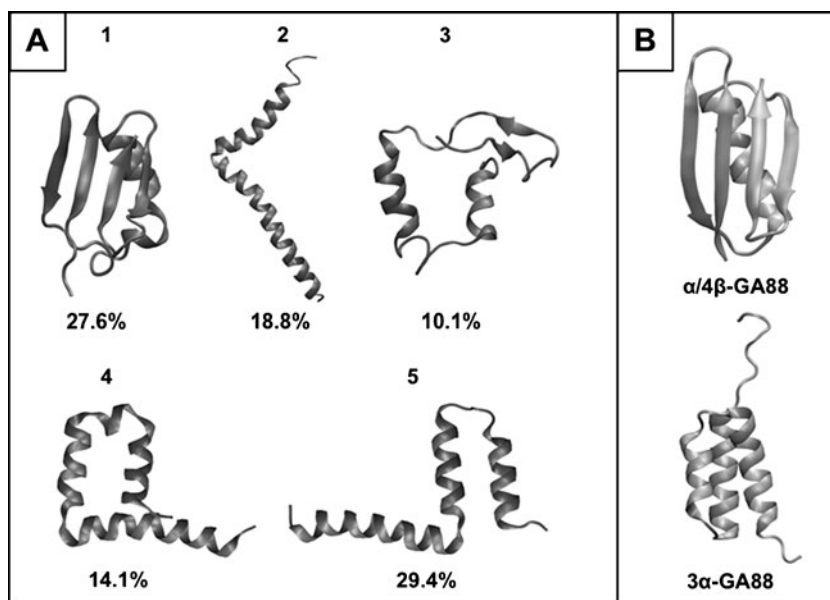


Fig. 1 (a) C_{α} – RMSD and probability distribution of C_{α} – RMSD of simulated structures (residue 9-53) at 270 K and 304 K. (b) Overlap between folded simulated structure (yellow) at 270 K and the NMR structure of GA88 (purple) (PDB id 2JWS) with best C_{α} – RMSD

(residue 9-53) of 4.34 Å [4]. (c) Overlap between folded simulated structure (orange) at 304 K and the NMR structure of GA88 (purple) with best C_{α} – RMSD (residue 9-53) of 4.75 Å

Fig. 2 (a) Cluster analysis. Representative structures of each cluster are represented using cartoon representation with the percentage stated denoting the percentage occurrences of these structures during the 75 ns simulation. (b) Cartoon representation of the NMR structure of GA88 (PDB id 2JWS) and GB88 (PDB id 2JWU) [4]



fold into a helix (residue 39–42) has an average helical propensity of $0.65 \text{ kcal mol}^{-1}$ with the presence of glycine (helix propensity $\sim 1.00 \text{ kcal mol}^{-1}$) highlighting the propensity of the short sequence to form a random coil instead of α -helix [27].

It was also apparent that the helix bundle as seen in GA88 was not accurately aggregated in the folded protein (Fig. 1b–c). One of the reasons contemplated was, implicit water solvation was not able to accurately account for the desolvation of the hydrophobic core [30]. Especially, in this study, surface area term was not included to avoid the biasing of β structure. Consequently, entropic cost incurred during the exclusion of water molecules from the hydrophobic core was not fully accounted hence favoring a lower energy conformation different from that of the native state [30]. Another is based on observations made by others which points out the over stabilization of salt bridges compared to hydrophobic interactions by implicit solvation model [15, 31, 32]. Based on the folded protein highlighted in Fig. 1b and c, charged amino acid residues especially K13 and K46 are observed to be fully exposed to solvent as opposed to the orderly packing of the aforementioned residues within the hydrophobic core in the NMR structure of GA88. Similarly, clusters 4 and 5 (Fig. 2) which encompassed a total of 43.5% of the trajectory based on the cluster analysis performed, demonstrated similar bias of charged residues to be either fully exposed to solvent or engaged in the formation of salt bridges. As mentioned *vide supra* with regards to the folded protein, K13 and K46 of the representative structures of clusters 4 and 5 alike, are also fully exposed to solvent (Fig. 3). In addition, salt bridges formed between K28 and E48 and between K31 and E48 in cluster 5 were also discerned (Fig. 3). These

events possibly impede the aggregation process of the helix bundle during the simulation. Overestimation of salt bridges over hydrophobic interactions possibly steered the formation of these redundant salt bridges between solvent and charged amino acids and between amino acids [31]. Hydrophobic interactions being key interactions in the assembly of the helix bundle in GA88 may thus be underestimated resulting in a folded structure away from the desired protein ensemble [4, 15, 30–32].

Moreover, Wang and Wade studied the effects various force fields had on the intermediates observed during the thermal denaturation of β -sheets [28]. Based on this study, Wang and Wade concluded that even though biasing of AMBER03 force field toward helical structures was observed, the combination of AMBER03 force field with explicit solvent model permitted the observation of the unfolding pathway of the β -sheets close to that observed experimentally and therefore is one of the useful tools in the study of the unfolding of α -helix and β -sheet structures in proteins [17, 28]. Explicit solvent model has also been substantiated to represent the solvation effect of hydrophobic residues more accurately than implicit solvation by others ergo being more precise in the modeling of hydrophobic core [15, 31–33]. Hence, a structure refinement by means of regular MD simulation of one of the prominent conformations in the trajectory was performed using explicit solvent model. The swinging of H1 and H3 into a position relatively close to its native positions was observed albeit slowly (data not shown).

To further fathom the folding mechanism of 3α -GA88, we scrutinized the protein by looking at separate helical domains namely H1 (residues 9–23), H2 (residues 27–34) and H3 (residues 39–51) to observe the folding of these individual

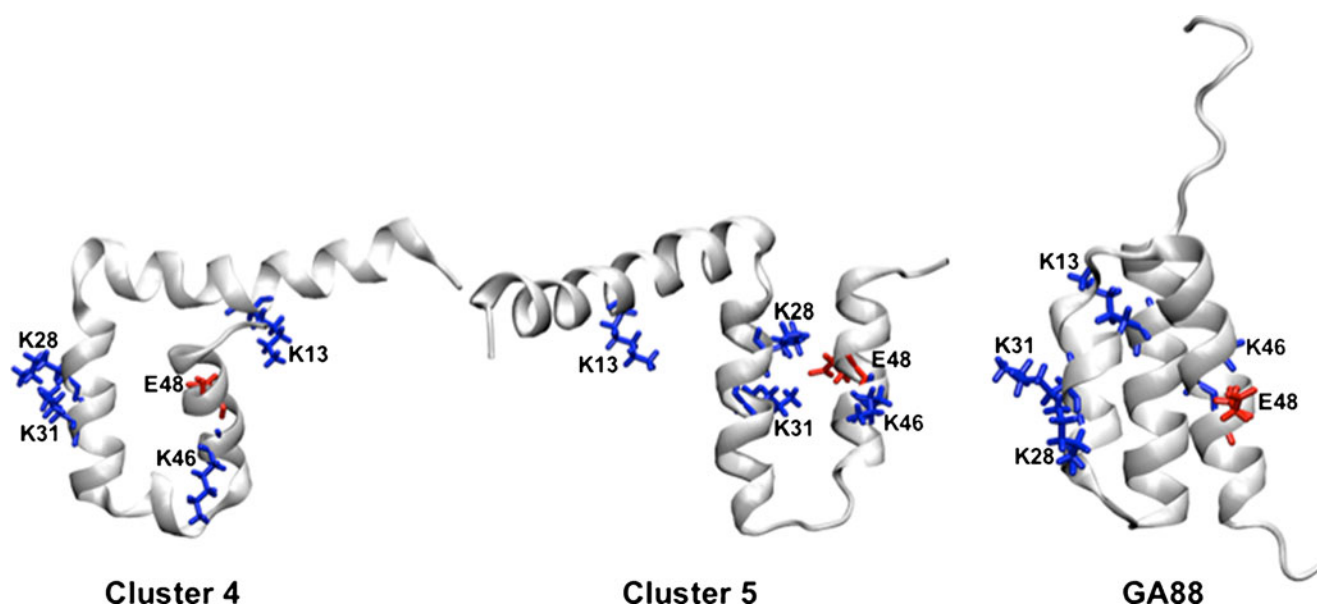


Fig. 3 Amino acid residues K13, K28, K31, K46 and E48 of cluster 4, cluster 5 and GA88 (PDB id 2JWS) are represented using licorice representation, viewed using VMD [4, 24]

helical domains. The calculation of C_{α} – RMSDs of the three helical domains of the simulated protein at 270 K and 304 K (Fig. 4a for 270 K and Fig. S1 in supporting information for 304 K) with respect to the three helical domains of the NMR structure of GA88 reveal the following: (i) The C_{α} – RMSDs of H1 and H3 show a general decrease with time implying the folding of H1 and H3. (ii) The C_{α} – RMSD of H2 remains approximately constant evincing the conservation of the helix domain at residues 27 to 34 from HB of $\alpha/4\beta$ -GA88 to H2 of 3α -GA88 over the entire simulation. Even though the C_{α} – RMSDs of H1 and H3 shows a general decrease with time, fluctuations manifest throughout the 75 ns simulation indicating the persistent folding and unfolding of these helical domains to β -strands which was corroborated by DSSP plot of the protein in Fig. 5 showing the interspersed presence of α -helix and β -strands at residues 9 to 23 and residues 39 to 51 [34, 35].

To reduce the numerous dimensionality contained within the MD trajectories to two dimensional, free energy landscape was plotted to observe the populations encompassed within the trajectories. Free energy landscapes of H1, H2 and H3 were plotted with C_{α} – RMSD and radius of gyration (R_g) as the reaction coordinates (Fig. 4b). The free energy landscapes of H1 and H3 displays a noteworthy population at the lower left-hand side of the landscape which corresponds to folded states and the notable single population observed for H2 connotes the conservation of H2. Furthermore, the fluctuations observed for the C_{α} – RMSD of H1 and H3 (Fig. 4a) which corresponds to the folding and unfolding of H1 and H3 during the simulation was supported by the presence of two prominent popula-

tions corresponding to helical domains and β -strands in the energy landscape of H1 and H3 (Fig. 4b) thus further attesting to the interpretation put together based on the C_{α} – RMSD of the three helical domains of the folded protein. The conservation of H2 is further supported by observing the temporal change in secondary structures by means of DSSP (Fig. 5) whereby residues 27 to 34 show considerable conservation of helical structures over the 75 ns simulation [34, 35]. Furthermore, the DSSP plot at both 270 K and 304 K demonstrates the occurrence of the conformational transitions from B1-loop-B2 (residues 1-20) to a structure incorporating H1 and from B3-loop-B4 (residues 42-55) to a structure incorporating H3 as the simulation progresses authenticating the conformational variation between $\alpha/4\beta$ -GA88 to 3α -GA88.

Other than gaining insights on the mechanism governing the folding of 3α -GA88, an appreciation of the unfolding pathway of $\alpha/4\beta$ -GA88 during the conformational variation is also of great importance. Here, the unfolding pathway was probed by means of principal component analysis (PCA). PCA is a well established mathematical technique employed by many in the study of protein folding and unfolding [3, 11, 36, 37]. This technique aids in the study of protein dynamics through the reduction of the numerous dimensions present in MD trajectories thus curbing the $3N$ (N = number of atoms in the protein) degrees of freedom of the protein to key degrees of freedom which are of great importance in the description of functionally crucial motions leading to the folding or unfolding of proteins [3, 36–38]. In this study, PCA of the first 15 ns of the trajectory was conducted to filter out the

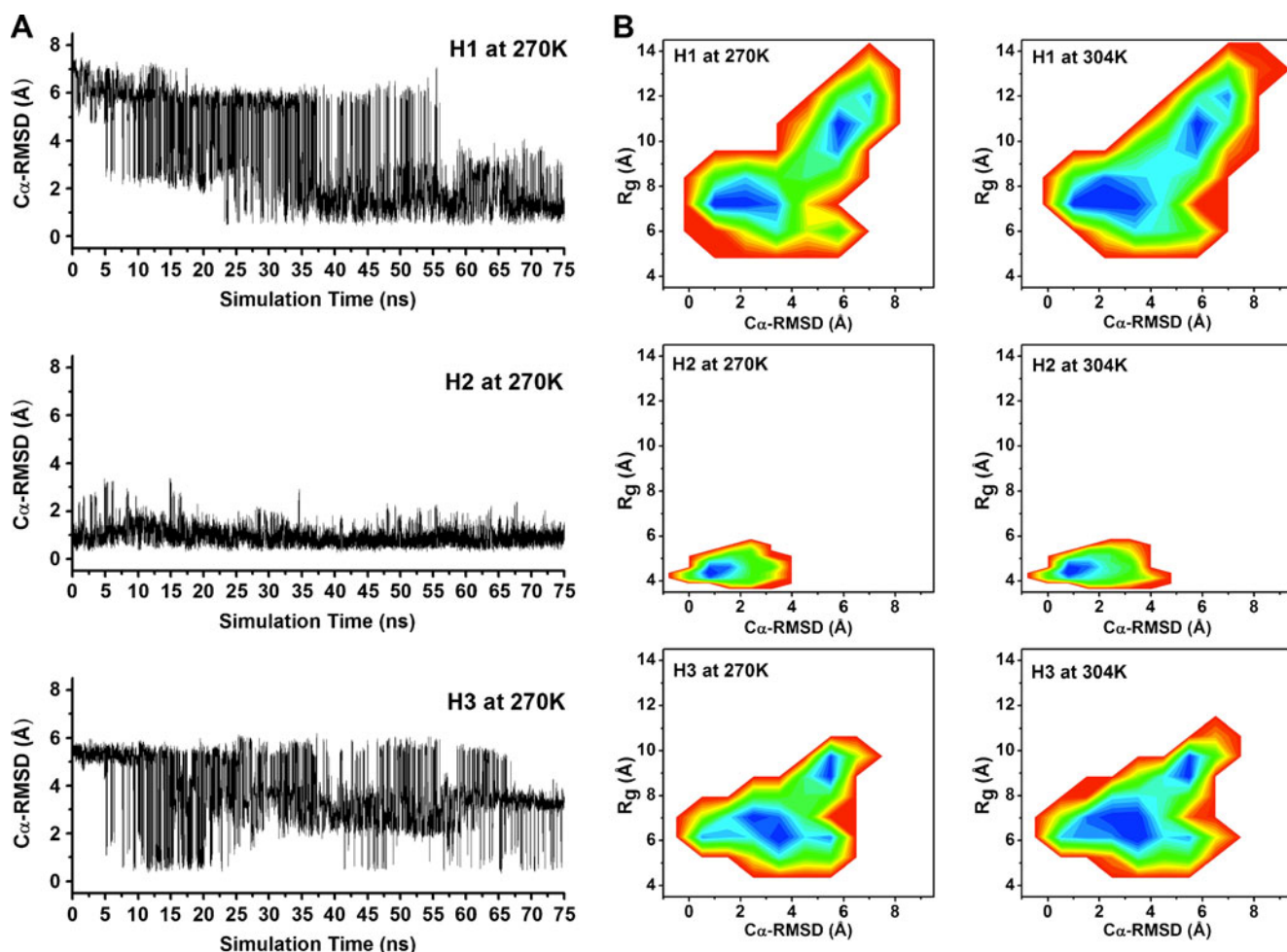


Fig. 4 (a) C_{α} – RMSD of H1 (residue 9 to 23), H2 (residue 27 to 34) and H3 (residue 39 to 51) domain of the simulated protein with reference to the NMR structure of GA88 (PDB code: 2JWS) at 270 K

[4]. (b) Free-energy landscape of H1, H2 and H3 domains of simulated protein at 270 K and 304 K. Energy level increases from blue to red

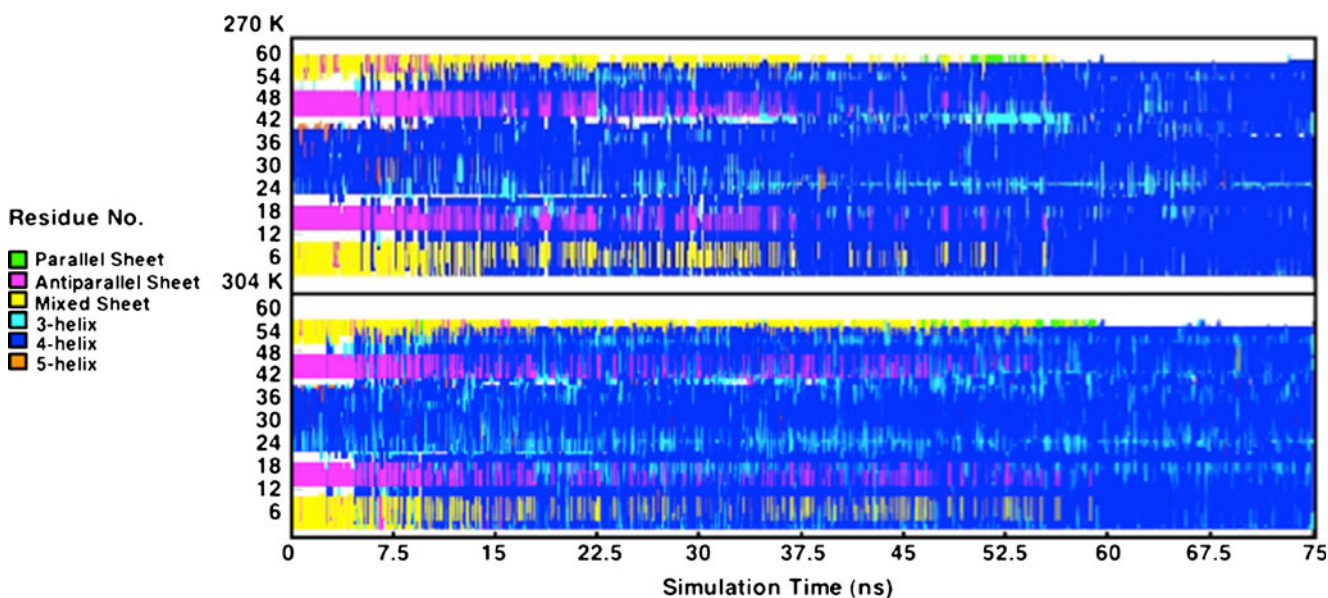


Fig. 5 Secondary structure assignment (DSSP) of folded protein during REMD simulation at 270 K (top) and 304 K (bottom) [34, 35]

dominant motion modes leading to the unfolding of $\alpha/4\beta$ -GA88 via eigen decomposition of the overall motion of the protein during the unfolding process [3, 36–38]. It was highlighted by Das et al. and many others that a large portion (~90%) of the structural fluctuations observed by means of PCA are described by a small set of degrees of freedom (~5%) of the protein [36, 37, 39]. Therefore, five PCs were generated and based on these five PCs, the first two PCs, PC1 and PC2, at both 270 K and 304 K show major contributions to the unfolding of $\alpha/4\beta$ -GA88 as illustrated by the plot of PCs versus time included in Fig. S2 in the supporting information.

The visual analysis of the crucial motion modes describing the unfolding pathway of $\alpha/4\beta$ -GA88 were made possible with the help of the interactive essential dynamics (IED) program [25]. IED program enables one to control the addition or the removal of eigenvectors describing the motion modes of the protein and the projection of the protein along an eigenvector [25]. In this study, the IED program, together with visual molecular dynamics (VMD) as display interface, assist in the comprehension of the functionality of each eigenvector in detailing the unfolding of $\alpha/4\beta$ -GA88 [24, 25]. PC1 and PC2 were noted to contribute dominant motions leading up to the unfolding of $\alpha/4\beta$ -GA88 by the separation of the two β -sheets, B1-loop-B2 and B3-loop-B4, of the protein to form structures inclusive of H1 and H3 respectively. PC1 accounts for the pulling motion separating B1 and B4 while PC2 accounts for the bending of the HB in $\alpha/4\beta$ -GA88 (see Fig. 6). These motions lead to the unpacking of the hydrophobic core of $\alpha/4\beta$ -GA88 which is comprised of Y3, L5 and L7 in B1, A26 and A34 in HB, W43 in B3 and F52 and V54 in B4 [4]. This unpacking of the hydrophobic core is crucial to drive the folding of H1 and H3 to form 3α -GA88.

Even though PC1 and PC2 are major contributors to the unfolding of $\alpha/4\beta$ -GA88 at both 270 K and 304 K, it was

discerned that a higher temperature allows the protein to explore more motion modes leading to unfolding. At 304 K, the amplitudes of the modal activity of the protein described by PC3 to PC5 are greatly enhanced compared to PC3 to PC5 at 270 K, with PC1 and PC2 still being the major contributor to the unfolding of $\alpha/4\beta$ -GA88 for both temperatures. (Supporting information, Fig. S2) A plot of PCs against time for PC1 and PC2 shown in Fig. 6, also suggest that PC2 plays a more prominent role than PC1 in the unfolding of $\alpha/4\beta$ -GA88 in the initial part of the trajectory. Hence, through PCA, one is not only able to identify crucial motions leading to the folding or unfolding of proteins but also able to determine when these motions occur.

Conclusions

REMD simulation was carried out to comprehend the mechanism influencing the conformational variation from $\alpha/4\beta$ -fold to 3α -fold originating from a protein with the primary sequence of GA88 in the guise of GB88. The conformational variation from $\alpha/4\beta$ -GA88 to 3α -GA88 was successfully observed albeit the large C_α -RMSD of the folded structure when compared to the NMR structure of GA88 at 270 K (4.34 Å) and 304 K (4.75 Å). Notwithstanding the large C_α -RMSD, analysis of the REMD trajectory aids in the understanding of the folding and unfolding pathway of 3α -GA88 and $\alpha/4\beta$ -GA88 respectively. Trajectories from the REMD simulation suggest an underlying bias of force field toward helix by the formation of additional helix in the first eight residues of 3α -GA88 instead of the intended random coil. The failure of the folded protein to assume the correct helix bundle conformation was attributed to the over stabilization of salt bridges and the imprecise desolvation of the

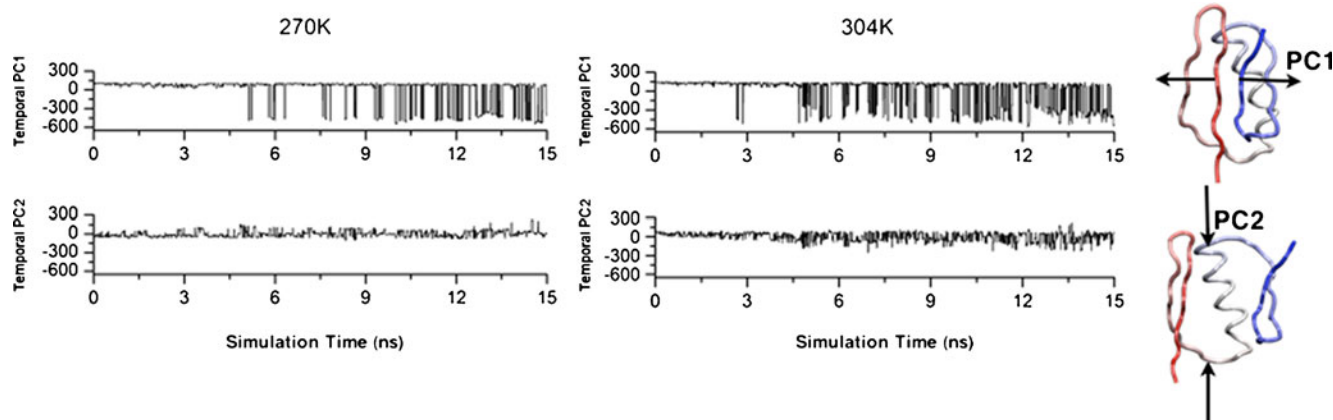


Fig. 6 Temporal activity of the protein based on the first two PCs at 270 K and 304 K for the first 15 ns of the trajectory together with illustrations of motion modes corresponding to PC1 and PC2 visualized using VMD with the aid of IED [24, 25]

hydrophobic core leading to the inaccurate representation of hydrophobic interactions crucial in the aggregation of the helix bundle. By means of C_{α} –RMSD, free energy landscape and secondary structure assignment (DSSP) of the separate helical domains of 3α -GA88 namely H1, H2 and H3, we are able to identify the conservation of the helix domain in $\alpha/4\beta$ -GA88 during the conformational variation and the folding of H1 and H3 close to the respective helical domains in GA88 [34, 35]. PCA aids in the study of the unfolding pathway of $\alpha/4\beta$ -GA88 by disclosing the motion modes crucial for the unfolding of this protein. PCA conducted reveals PC1 and PC2 as the main contributor to the unfolding of $\alpha/4\beta$ -GA88 with PC2 being more prominent during the initial part of the trajectory although PC1 contributes largely to the overall unfolding of the protein.

Acknowledgments DWZ is supported in part by Nanyang Technological University (NTU) start-up grant, and in part by Singapore Academic Research Fund (AcRF) Tier 1 Grant of M52110095. DWZ would like to acknowledge and thank NTU High Performance Computing (HPC) support and resources. YM is supported by National Natural Science Foundation of China (Grants No. 20803034).

References

- Onuchic JN, Luthey-Schulten Z, Wolynes PG (1997) Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem* 48:545–600
- Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2009) A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci* 106:21149–21154
- Maaß A, Tekin ED, Schüller A, Palazoglu A, Reith D, Faller R (2010) Folding and unfolding characteristics of short beta strand peptides under different environmental conditions and starting configurations. *Biochim Biophys Acta* 1804:2003–2015
- He Y, Chen Y, Alexander P, Bryan PN, Orban J (2008) NMR structures of two designed proteins with high sequence identity but different fold and function. *Proc Natl Acad Sci* 105:14412–14417
- Rose GD, Creamer TP (1994) Protein folding: predicting predicting. *Proteins Struct Funct Genet* 19:1–3
- Dalal S, Balasubramanian S, Regan L (1997) Protein alchemy: changing β -sheet into α -helix. *Nat Struct Biol* 4:548–552
- Yang W-Z, Ko T-P, Corselli L, Johnson RC, Yuan HS (1998) Conversion of a β -strand to an α -helix induced by a single-site mutation observed in the crystal structure of Fis mutant Pro²⁶Ala. *Protein Sci* 7:1875–1883
- Dobson CM (2003) Protein folding and misfolding. *Nature* 426:884–890
- Zhou R (2003) Trp-cage: folding free energy landscape in explicit water. *Proc Natl Acad Sci* 100:13280–13285
- Karplus M (1987) Molecular dynamics simulations of proteins. *Phys Today* 40:68–72
- Lei H, Wu C, Liu H, Duan Y (2007) Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc Natl Acad Sci* 104:4925–4930
- Freddolino PL, Liu F, Gruebele M, Schulten K (2008) Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophys J* 94:L75–L77
- Pande VS, Rokhsar DS (1999) Molecular dynamics simulations of unfolding and refolding of a β -hairpin fragment of protein G. *Proc Natl Acad Sci USA* 96:9062–9067
- Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* 314:141–151
- Zhou R, Berne BJ (2002) Can a continuum solvent model reproduce the free energy landscape of a β -hairpin folding in water? *Proc Natl Acad Sci* 99:12777–12782
- Case DA, Darden TA et al (2008) AMBER 10. University of California, San Francisco
- 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
- Tsui V, Case DA (2001) Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolym (Nucleic Acid Sci)* 56:275–291
- Levy RM, Zhang LY, Gallicchio E, Felts AK (2003) On the nonpolar hydration free energy of proteins: Surface area and continuum solvent models for the solute-solvent interaction energy. *J Am Chem Soc* 125:9523–9530
- Lwin TZ, Zhou R, Luo R (2006) Is Poisson-Boltzmann theory insufficient for protein folding simulations? *J Chem Phys* 124:34902–34907
- www.pdb.org, Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. *Nucl Acids Res* 28:235–242
- 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
- Uberuaga BP, Anghel M, Voter AF (2004) Synchronization of trajectories in canonical molecular-dynamics simulations: observation, explanation, and exploitation. *J Chem Phys* 120:6363–6374
- Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14:33–38
- Mongan J (2004) Interactive essential dynamics. *J Comput Aided Mol Des* 18:433–436
- Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2007) The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci* 104:11963–11968
- Pace NC, Scholtz MJ (1998) A helix propensity scale based on experimental studies of peptides and proteins. *Biophys J* 75:422–427
- Wang T, Wade RC (2006) Force field effects on a β -sheet protein domain structure in thermal unfolding simulations. *J Chem Theor Comput* 2:140–148
- Mittal J, Best RB (2010) Tackling force-field bias in protein folding simulations: Folding of villin HP35 and pin WW domains in explicit water. *Biophys J* 99:L26–L28
- Cheung MS, Garcia AE, Onuchic JN (2002) Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc Natl Acad Sci* 99:685–690

31. Geney R, Layten M, Gomperts R, Hornak V, Simmerling C (2006) Investigation of salt bridge stability in a generalized Born solvent model. *J Chem Theor Comput* 2:115–127
32. Jang S, Kim E, Pak Y (2007) Direct folding simulation of α -Helices and β -hairpins based on a single all-atom force field with an implicit solvation model. *Proteins Struct Funct Bioinf* 66:53–60
33. Nymeyer H, García AE (2003) Simulation of the folding equilibrium of α -helical peptides: a comparison of the generalized Born approximation with explicit solvent. *Proc Natl Acad Sci* 100:13934–13939
34. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637
35. Mezei M (2010) Simulaid: a simulation facilitator and analysis program. *J Comput Chem* 31:2658–2668
36. Lou H, Cukier RI (2006) Molecular dynamics of apo-adenylate kinase: a principal component analysis. *J Phys Chem B* 110:12796–12808
37. Das A, Mukhopadhyay C (2007) Application of principal component analysis in protein unfolding: an all-atom molecular dynamics simulation study. *J Chem Phys* 127(1-8):165103
38. Maisuradze GG, Liwo A, Scheraga HA (2009) Principal component analysis for protein folding dynamics. *J Mol Biol* 385:312–329
39. Palazoglu A, Gursoy A, Arkun Y, Erman B (2004) Folding dynamics of proteins from denatured to native state: principal component analysis. *J Comput Biol* 11:1149–1168

Improving the desolvation penalty in empirical protein pK_a modeling

Mats H. M. Olsson

Received: 13 January 2011 / Accepted: 27 May 2011 / Published online: 14 June 2011
© Springer-Verlag 2011

Abstract Unlike atomistic and continuum models, empirical pK_a predicting methods need to include desolvation contributions explicitly. This study describes a new empirical desolvation method based on the Born solvation model. The new desolvation model was evaluated by high-level Poisson-Boltzmann calculations, and discussed and compared with the current desolvation model in PROPKA—one of the most widely used empirical protein pK_a predictors. The new desolvation model was found to remove artificial erratic behavior due to discontinuous jumps from man-made first-shell cutoffs, and thus improves the desolvation description significantly.

Keywords Solvation · Desolvation · pK_a · Protein modeling · PROPKA

Introduction

The desolvation penalty makes an important contribution when calculating ionization energies for protein ionizable groups. It is the primary driving force explaining why charged residues—primarily Asp, Glu, Lys, and Arg—are found mainly on protein surfaces and only rarely buried in the commonly more hydrophobic protein interior. When they do occur in protein interiors, such residues often form part of an active site or otherwise have an important

function in the protein, e.g., Glu 35 in lysozyme, Asp 25 in HIV-protease, and Glu 78 and Glu 172 in *Bacillus circulans* xylanase. Thus, since ionizable residues can be found in protein interiors, it is crucial to include the desolvation penalty correctly when modeling protein pK_a predictions.

Most pK_a predicting approaches calculate protein pK_a values starting from a thermodynamic cycle where an ionization process is considered in a reference water solution and in the protein [1, 2]. This eventually requires calculating the change in solvation free energy for the charged and uncharged form of the residue/solute. In most approaches, the desolvation is typically included explicitly; for instance, in all-atom molecular dynamics (MD) simulations the explicit water solvent molecules are prevented from occupying areas next to a buried ionizable residue by the presence of intervening atoms, provided that the local folding energy is larger than the desolvation penalty. Similarly, in Poisson-Boltzmann (PB) and Generalized-Born (GB) models the desolvation comes from using much lower values for the internal protein dielectric constant, $\epsilon_{in} \approx 4$, compared to the external solvent dielectric constant, $\epsilon_{ext} \approx 80$ (the desolvation contribution is thus also intimately connected with atomic radii). In empirical approaches, however, there are typically no explicit solvent or atomic radii a priori that can serve as the desolvation penalty; instead it needs to be added as a specific contribution. Two common ways of including solvation or desolvation in empirical protein modeling are through the solvent accessible surface area (ASA) [3] or the contact model.

Recently, we have found that the current desolvation contact model used in PROPKA—one of the most popular empirical protein pK_a predictors—is fundamentally incorrect and needs to be revised since it has discontinuous jumps and exhibits unphysical behavior. In this study, I present an improved desolvation model that take its

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1141-1) contains supplementary material, which is available to authorized users.

M. H. M. Olsson (✉)
Department of Chemistry, University of Copenhagen,
Universitetsparken 5,
2100 Copenhagen, Denmark
e-mail: Mats.Olsson@hotmail.com

inspiration from the Born solvation model [4] and is therefore also closely related to the GB and PB solvation models. The model is quite simple, but in the end provides a more reasonable description of desolvation than the previous PROPKA contact model. The following sections sketch out the theoretical justification of the desolvation model, validate the behavior of the model and obtain approximate parameters using PB calculations. Finally, the new model, which has been implemented in a new version of PROPKA [2], is compared with the current PROPKA desolvation model for a few residues with experimental pK_a values. As we move to justifying and describing the model in the following section, we should keep in mind that we are not looking for an exact theory but for a simple and computationally fast approximation that has the most important features of an idealized system. We first briefly consider some of the fundamental relationships and practical details of pK_a predictions.

pK_a desolvation models

We start by noting that the vast majority of protein pK_a values are very similar to their corresponding reference pK_a values in water, and the effect of the protein is comparatively small even for a significantly shifted residue. In fact, using the relationship between pK_a value and solvation free energy differences

$$\Delta\Delta G_{solv}^{\circ} = 2.30RT \cdot \Delta pK_a, \quad (1)$$

where R is the gas constant and T is the temperature, we find that, for each pH unit of shift ($pK_a^{\text{protein}} - pK_a^{\text{water}}$) the corresponding shift in solvation free energy is only $1.36 \text{ kcal mol}^{-1}$. This is more than an order of magnitude smaller than the absolute solvation energy change of several tens of kcal mol^{-1} in water solvent between the protonated and unprotonated residue (the solvation energy for acetic acid for instance is close to 80 and 10 kcal mol^{-1} for its charged and uncharged form, respectively). We therefore formulate the pK_a value of a protein residue in terms of its known water reference as

$$pK_a^{\text{protein}} = pK_a^{\text{water}} + \Delta pK_a^{\text{water} \rightarrow \text{protein}} \quad (2)$$

and, thus, see the effect of the protein as a perturbation to the reference water value ($\Delta pK_a^{\text{water} \rightarrow \text{protein}}$) [1]. For pK_a predictions, this perturbation is subsequently divided into an intrinsic “self-energy” term and a Coulomb charge–charge interaction term. The former corresponds to transferring the residue from its reference state in the solvent to a state in its protein position where all other titratable residues are in their neutral form. In a second step the interactions between titratable residues are turned on and either solved iteratively by a Tanford-Roxby [5] scheme or

the titration curve is calculated by a Monte Carlo scheme [6]. For our present purpose, however, we are not concerned by this detailed analysis and conclude that the solvation free energy of transferring an ionizable residue from water to the protein can, like its pK_a value, be divided into a desolvation, protein resolution (electrostatic interactions with polar protein groups, e.g., $\text{COO}^- \cdots \text{HN}$) and Coulomb charge–charge contribution.

The desolvation term describes the solvation penalty or the loss of solvation energy exerted by the protein as protein atoms replace ambient water. The electrostatic terms then describe the substituting resolvating effect of those atoms and charge–charge interactions. Before we continue and consider desolvation models in more detail we note that the desolvation and the resolvating electrostatic contributions have opposite effects on the pK_a value and, in many cases, balance each other to give a rather subtle total pK_a shift. In some cases, however, apolar hydrophobic residues surrounding the ionizable residue overthrow this balance, resulting in significantly shifted pK_a values.

Since calculating the desolvation energy is the reverse of calculating the solvation energy, and solvation energy is a more abundant topic in the literature [7, 8], we start by briefly considering two of the more commonly used simplified approaches to calculating large molecule or protein solvation energies. These solvation methods are especially relevant for the desolvation pK_a contribution since, in principle, both decompose the molecule into solvating fragments or atoms, calculate the desolvation from nearby fragments and obtain the solvation energy as the residual solvation. For approaches based on ASA the solvation energy is given by

$$\Delta G_{solv}^{\circ} = \sum_i \sigma_i ASA_i \quad (3)$$

where σ_i is an atom-based per-unit-area solvation parameter and ASA_i is its solvent accessible surface area in its protein or large-molecule position [3]. The contact model in its basic form is even simpler and can be written as

$$\Delta G_{solv}^{\circ} = \sum_i \alpha \cdot CW = \sum_i \alpha \cdot (C - CP) \quad (4)$$

where α is a scale factor, C is the number of possible nearest-neighbor contacts, and CW and CP are the corresponding contacts from water solvent and protein atoms, respectively [9]. Effectively, both these methods consider solvent effects from only the first coordination sphere, either through the contact area or through a number of nearest neighbor contacts. The solvation of a charged residue, however, depends on the electrostatic potential of the charge, which is proportional to r^{-1} and should therefore be considered long range. Thus, it seems reasonable to try to extend this picture.

The desolvation model in PROPKA versions 1 [10] and 2 [11] makes use of two types of contact desolvation contributions: local and global. The first term is akin to the basic ASA or contact model in that it reports on atoms in the immediate neighborhood that prevent solvent molecules from coming into contact with the residue in question. The global term is an extension that corresponds to a “depth of burial” energy (i.e., the distance of a residue from the protein surface). The difference in desolvation energy between the charged and uncharged residue is thus calculated as

$$\Delta\Delta G_{desolv} = 2.30RT(C_{local} \cdot N_{local} + C_{global} \cdot (N_{global} - 400)) \tag{5}$$

where the local contribution is a product of an empirical constant, $C_{local}=0.07$, and the number of non-hydrogen atoms (N_{local}) within a radius, R_{local} , of the residue ionizable center. R_{local} depends on the residue type, but typically its value is between 3.5 and 6 Å. The global contribution is calculated in a similar way using a global empirical constant, $C_{global}=0.01$, and a residue-independent radius, $R_{global}=15.5$ Å. Although this model includes contributions beyond the first coordination, it represents the desolvation penalty inappropriately (as will be discussed later).

In order to extend these first-shell models, we start by considering the solvation energy of residue i , which is defined as

$$\Delta G_{solv}^{(i)} = \Delta G_{env}[\rho_i] - \Delta G_{vac}[\rho_i] \tag{6}$$

where the first and second terms are the energy of its charge density, ρ_i , in a solvating environment and in vacuum, respectively. These terms are formally defined as charging processes where the charge densities are turned on, $0 \rightarrow \rho_i$, in the environment and in vacuum. Following on from many previous studies [12, 13], we next consider the solvation environment as a continuum and write the solvation energy as the integral

$$\Delta G_{solv}^{(i)} = \int_V f[\rho_i, r, \varepsilon(r)] dV \tag{7}$$

Here, $f[\rho_i, r, \varepsilon(r)]$ is the solvation free-energy density at a given point, r , that depends on the residue charge distribution and the dielectric medium, $\varepsilon(r)$, of the protein+water system. The dielectric medium is, in this context, clearly non-homogeneous and is therefore denoted $\varepsilon(r)$ rather than ε . V indicates that we integrate over the volume surrounding the residue. Thus, unlike several other studies concerned with simplified large-molecule solvation models [12, 14], we do not include the solute, or in this case residue, volume in the integral since its contribution is identical in both terms of Eq. 6 and therefore does not contribute to the solvation energy. Following classical electrodynamics [15] and the long-range

limit of Schaefer and Froemmel [12], we choose $f[\rho, r, \varepsilon(r)]$ to be proportional to a functional of the dielectric medium, $\alpha[\varepsilon(r)]$ and to the inverse distance according to

$$f[\rho, r, \varepsilon(r)] \propto \alpha[\varepsilon(r)] \cdot r^{-4} \quad \text{for } r \geq r_0 \tag{8}$$

(again for volumes outside the solute and accessible to the solvent, $r \geq r_0$). Thus, for a spherical system with a unit point charge in a homogeneous environment that can be described by a dielectric constant, ε , we can rewrite Eq. 7 in spherical coordinates

$$\Delta G_{solv} = \int_{r_0}^{\infty} f[\rho, r, \varepsilon(r)] 4\pi r^2 dr \tag{9}$$

and carry out the integration from r_0 to infinity and get

$$\Delta G_{solv} = 4\pi \alpha(\varepsilon) \int_{r_0}^{\infty} r^{-2} dr = -\alpha(\varepsilon) \frac{4\pi}{r_0} \tag{10}$$

Thus, we see that when we choose an appropriate form of $\alpha(\varepsilon)$, we recover the well-known Born equation [4]

$$\Delta G_{solv}^{Born} = -\left(1 - \frac{1}{\varepsilon}\right) \frac{Q^2}{2a} \tag{11}$$

where Q is the point charge, a is its corresponding radius (r_0 in Eq. 8), and ε is the homogeneous dielectric constant. This can be considered “exact” for an ideal system.

One way to extend the above model might be to represent the charge distribution with atom-based residual charges, $\rho = \sum q_i$. In this case, both the charge distribution and the solute/solvent boundary are better described, but for it to make sense a desolvation correction to the radii from nearby solute atoms is required [16]. This more detailed description would eventually lead to a pair-wise self-energy approximation [13]. However, this seems excessive in PROPKA since we in any case use a single point-charge model for Coulomb interactions and generally a simplistic model of intra-protein interactions. The desolvation term should also preferably be as simple as possible since calculating the desolvation penalty is the most time-consuming step in PROPKA. Though the single-charge approximation might seem crude we should keep in mind that our task eventually is to calculate the *desolvation* energy, which is the decrease in solvation free energy as a small portion of the solvation density is removed due to nearby protein atoms. In this sense the “exact” solvation energy for the free amino acid, for which a more detailed description would probably be necessary, is already included in the model through the water-reference pK_a value. Moreover, pK_a predictions require the desolvation *difference* between the charged and uncharged residue, which corresponds to a difference in the difference in solvation free energy, and we therefore assume that only the *excess* charge distribution can be modeled as a

spherical difference. Thus, given these circumstances, a point-charge model should suffice.

For calculating the desolvation, it seems feasible to reverse the situation and to integrate the excluded contributions from the solvation density volumes that are now displaced by those atoms. This is of course the idea introduced in reference [9]. However, a more convenient definition, which we will adopt here, is to retain parts of the polarization, e.g., the electronic polarization, in the excluded volume and integrate the difference. Again, we take the Born model as inspiration and use the function

$$\alpha(\varepsilon) = \left(\frac{1}{\varepsilon_w} - \frac{1}{\varepsilon} \right) \quad (12)$$

where ε_w is the dielectric constant of ambient water and ε is an effective dielectric constant of the residue occupying the excluded volume without the contribution from its permanent dipoles. Thus, we obtain our expression as

$$\Delta\Delta G_{desolv}^{(i)} = \int_{V'} \alpha(\varepsilon) r^{-4} dV' \quad (13)$$

Here, $\Delta\Delta G_{desolv}^{(i)}$ refers to the difference in desolvation energy between the charged and uncharged form of residue i , and V' indicates that we restrict the integration to the space not occupied by the solvent, i.e., the volume occupied by the protein but not residue i . Now, since we are looking for a simple and computationally fast approach, and the integral of Eq. 13 would be time consuming and maybe also non-trivial, we replace the integral with a summation over point-volumes according to

$$\Delta\Delta G_{desolv}^{(i)} = \sum_{j, j \neq i}^N \alpha(\varepsilon) \frac{V_j}{r_{ij}^4} \quad (14)$$

where the sum runs over all protein atoms not belonging to residue i , and V_j is the volume occupied by atom j at distance r_{ij} from the charge center of residue i . This approximation might seem crude compared to, e.g., PB approaches that in principle solve the corresponding solvation integral numerically, but this point-volume model is much faster to evaluate and much more in line with the remaining non-rigorous electrostatic terms of PROPKA. The important point for us at present is that the desolvation contribution is directly proportional to the volume of displaced solvent and that the distance dependence comes out as r^{-4} .

So far we have considered a static or average-structure model of the desolvation. In real-life protein titrations, however, the protein typically reorganizes as the residue changes ionization state. This reorganization can be anything from smaller protein-dipole reorientation to water penetration and, in extreme cases, local protein unfolding. This structural change significantly complicates a rigorous

treatment, making at least the static picture approximate, and marks one of the biggest challenges to both empirical and non-empirical contemporary pK_a predictions. In the case of most PB approaches this is modeled by the protein dielectric constant, $\varepsilon_p=4$, which generally includes “protein effects not treated explicitly” [17]. In PROPKA this effect is also treated implicitly, but in this case it is instead parameterized into the effective perturbations (the $\Delta pK_a^{\text{water} \rightarrow \text{protein}}$ contributions) and therefore not seen directly. In the particular case of the desolvation, this simply means that we fit the $\alpha(\varepsilon)$ coefficient in Eq. 14 to a set of experimentally determined protein pK_a values to include the effective contribution, which includes protein reorganization.

As it turns out, this effective protein response is different on the protein surface compared to in the protein interior since water penetration and reorganization is easier on the surface (this is not rigorous, but comes from the empirical observation that pK_a values are significantly better reproduced by two parameters compared to one [2]). Thus, we fit two global constants (c_{surface} and c_{buried}) to the experimental data through the function

$$\Delta G_{desolv}^{(i)} = c \cdot \sum_{j, j \neq i}^N \frac{V_j}{r_{ij}^4} \quad (15)$$

where c is the interpolation between the two surface and buried extremes

$$c = c_{\text{surface}} - (c_{\text{surface}} - c_{\text{buried}}) \cdot w_i(N) \quad (16)$$

$w_i(N)$ is the buried ratio of residue i and defined elsewhere [2]. The fact that we fit the c values to experiments also relieves us from defining a protein dielectric constant, and therefore we evade an open and much debated question in protein modeling.

Before concluding this section and moving on to validating our model, we need to clarify that the solute Born radius that appears in the integral of Eq. 9, and is an integral part of all PB and GB approaches, lies for our derivation in the solvent contribution to the water reference pK_a value. The radius determines how close the solvation free-energy density approach the ion, the effective size of the ion, whereas the integral of Eq. 13 involves only excluded-solvent regions that are outside this limit. Since the reference value is a tabulated experimental value we do not need to concern ourselves with what type of solute radii are appropriate, i.e., whether Connolly radii are better than van der Waals radii, etc. Note, however, that we have to define volumes for the desolvating atoms according to Eq. 15, but since these volumes appears as a product with c we effectively have to define only the relative volumes and let the fitting procedure scale these to obtain the appropriate excluded-solvent volumes.

Another attractive feature of this model is that there is also no need to define the end of the first solvation-shell radius since the desolvation from all surrounding atoms is included with the appropriate distance-dependent weight r^{-4} . Conversely, the local desolvation contribution in the old PROPKA contact model, and contact models in general, requires this type of radius to define which atoms are within the first coordination sphere and therefore contribute to excluded nearest neighbor contacts. This is obviously an attractive feature of the new model since it avoids unnecessary, and to some extent arbitrary, parameters.

Methods

Test and evaluation

Before we continue with evaluating this model, we should point out that pK_a calculations are not at a stage where their individual contributions can be calculated with accuracy. Instead, pK_a values depend on two comparatively large and opposing quantities—the desolvation penalty and the protein electrostatic resolution—that to a large extent balance each other out, often resulting in a rather subtle total effect. Therefore, there is little purpose in obtaining and cementing the necessary parameters without defining the remaining parameters. Instead, this section focuses on illustrating the model's strengths and functional form, keeping in mind that the paper's main objective is to provide a better desolvation model than the previous PROPKA contact model.

Validation using Poisson-Boltzmann

The best approach to validating our new desolvation model would obviously be to compare our model with mutation experiments where the environment in the vicinity of an ionizable residue is replaced by apolar groups. However, since such experiments are scarce and difficult to assess, we have instead calculated the desolvation energy using a PB approach to further probe our model from a physical perspective. In this part of the study, we calculate how the solvation energy decreases as a fixed small “solvation volume” is excluded from an otherwise fully solvated solute “test residue” at various distances from the solute. Our objective is to verify our model by fitting calculated desolvation energies to Eq. 14 and to obtain a rough estimate of the coefficient $\alpha(\epsilon) \cdot V_j$ and validate the r^{-4} distance dependence.

Our solute test residue is a sphere with a radius of 2.0 \AA that is surrounded by a dielectric medium with $\epsilon=80$. This system has a solvation energy of 82 kcal mol^{-1} when it is completely surrounded by the high-dielectric medium and

can thus serve as a united-atom model of lysine or a simplified carboxylic acid. The desolvation energy is then obtained as the solvation-energy difference between the fully solvated test residue and when a small sphere of solvation volume with radius 1.5 \AA is replaced by a low-dielectric medium, $\epsilon=1, 2$, or 4 , representing, for instance, an oxygen atom, see Fig. 1. We used the program suite MEAD [18] to calculate the solvation energy by solving the PB equation with the finite-difference method.

Before presenting the results we should point out that we are calculating very small changes in the solvation free energy and therefore have to set up our calculations such that we minimize errors, and utilize error cancellation to the largest extent possible. Thus, we used four consecutive grids for our PB solver, 39×1.000 , 75×0.500 , 147×0.250 , and $289 \times 0.125 \text{ \AA}$, with a final grid finer than that commonly used extending 18 \AA from the solute center. To reduce any remaining artificial grid dependencies, we first tried to average the calculated solvation energies over random solute displacements (between 0 and 0.1 \AA from the grid center) for each solute to excluded-volume distance. Generating these displacements independently for each distance was, however, found to be insufficient to obtain well-behaved data since random numerical noise results in small and negative desolvation energies for longer distances. Instead, we generated 52 random solute positions and calculated the desolvation energy profiles for distances from 3 to 15 \AA along a random vector with an increment of 0.25 \AA , and averaged these. This enhanced cancellation of errors and resulted in more well-behaved data. The resulting desolvation profile is presented in Fig. 2 for $\epsilon=1, 2$ and 4 . The solid error bars represent standard error ($2 \times \text{SE}$) and the dotted error bars represent standard deviation ($2 \times \text{SD}$). In other words, we can say with 95% confidence that the average value is within the solid error bars, but any one particular calculation falls with 95% certainty within the dashed error bars.

The reference solvation free energy for the three calculations (with $2 \times \text{SE}$ in parenthesis) was found to be -82.26 (0.02), -82.26 (0.01), and -82.25 (0.02) kcal mol^{-1} . Even though these numbers seem, and should be, identical, the standard error shows that the solvation calculations converge to only $0.02 \text{ kcal mol}^{-1}$ with respect to random displacements, whereas the desolvation energies can be significantly smaller. Figure 3 depicts how the standard error decreases as we include an increasing number of random displacements in our averaging. The black lines and inset histograms represent the reference solvation energies, whereas the red, green, and blue lines represent the desolvation energies at distances 3 , 3.75 , and 10 \AA , respectively. From the three black lines, corresponding to different sets of random numbers from desolvation calculations with different dielectric constants, we see that the solvation free-energy reference has a standard error of $0.01 \text{ kcal mol}^{-1}$ after 52 random

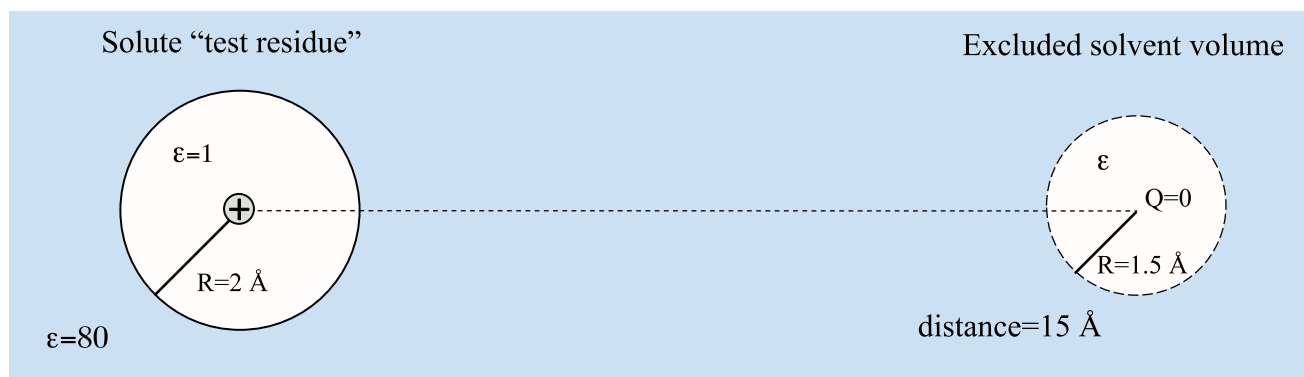


Fig. 1 Idealized system used to study desolvation energy where desolvation is calculated as the difference between the solvation energy of a fully solvated solute “test residue” and a small “excluded

displacements. In the case of the desolvation energy, the standard error is much smaller and we obtain 1.0, 0.13, and 0.03 cal mol⁻¹ for distances 3, 3.75, and 10 Å respectively. Thus, we see that our desolvation calculations converge much faster than the total reference solvation since we set up the calculations to maximize cancelation of errors. If instead we made random displacements independently for each distance, thereby not utilizing error cancelation, we would also obtain a standard error of 0.01 kcal mol⁻¹ for the desolvation calculations, which makes the standard error larger than the calculated desolvation energy already from distances of 4.25 Å and greater. Our calculations show that this limit is shifted to longer distances and that the calculated desolvation is

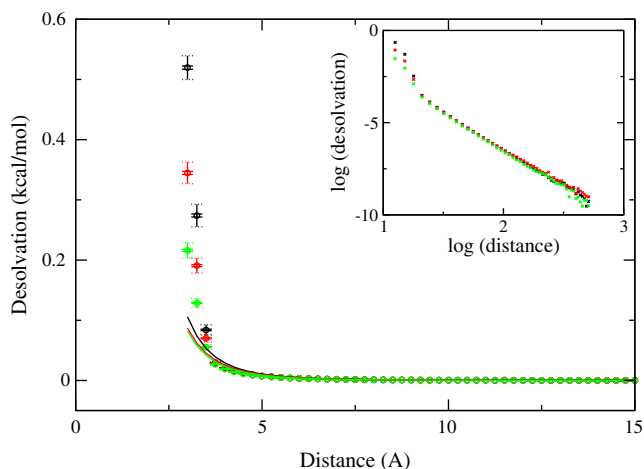


Fig. 2 Desolvation profile for the system defined in Fig. 1 calculated with the Poisson-Boltzmann (PB) approach using $\epsilon=1, 2,$ and 4 (black, red, and green respectively) and averaged over 52 random initial origins and directions. Circles show the calculated averaged points, the lines show the fitted power equations ($\Delta G_{\text{desolv}}=C \cdot r_{ij}^{-n}$), and the error bars show double standard error (solid bars) and double standard deviation (dotted bars). Inset Log-log plot showing the trusted range and quality of the regression fit; desolvation is best described by a power function with exponent 4 as expected

solvent volume” in the high dielectric medium is replaced by a vacuum or low dielectric. The figure does not show the entire grid; the dielectric medium extends 18 Å along all axes from the solute center

larger than the standard error also at distances of 15 Å (0.1 cal mol⁻¹ versus 0.06 cal mol⁻¹). For the more stringent requirement, $\Delta G_{\text{desolv}} > 2 \times \text{SE}$, we find that these limits become approximately 3.75 and 13.5 Å. Using a coarser grid (147 × 0.250 Å) increases the uncertainties by a factor of more than 3 and an even coarser grid (75 × 0.500 Å) by a factor close to 16 (see Figs. S1–S4 in the supplementary information).

The resulting averaged desolvation profile and fit to Eq. 14 are presented in Fig. 2 and Table 1. The first, and possibly most important property we want to verify is whether we can represent the calculated data with a power function, how good this fit is, and if we obtain the expected exponent -4 of the distance dependence. Fitting the averaged energies for all 49 distances to a power function ($\Delta G_{\text{desolv}}=C \cdot r^{-n}$) gives an exponent in the range -4.47

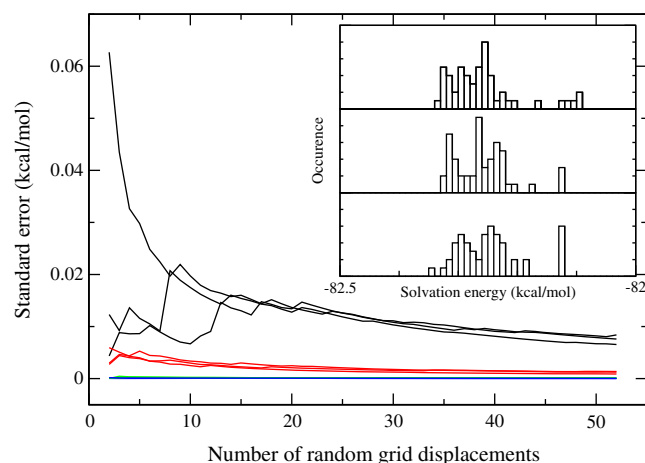


Fig. 3 Convergence of the reference solvation energy (black lines and inset histograms) and the desolvation energy presented in Fig. 2 for distances 3, 3.75, and 10 Å (red, green, and blue respectively). Although the solvation reference has not converged to more than 0.01 kcal mol⁻¹, the desolvation energies converge much faster because of efficient use of error cancellation

Table 1 Regression parameters for the desolvation energy fitted to a power function ($\Delta G_{\text{desolv}} = C \cdot r_{ij}^n$) given in kcal mol⁻¹ Å⁻ⁿ or dimensionless

ϵ	Points from 3.00 to 15.00 Å			Points from 3.75 to 10.00 Å		
	C	n	r	C	n	r
1	14.37	-4.47	-0.98	5.95	-4.12	-1.00
2	9.10	-4.23	-0.99	4.81	-4.00	-1.00
4	9.78	-4.35	-0.99	5.56	-4.13	-1.00

to -4.23 for the different dielectric constants, with a correlation coefficient close to 1. However, as can be seen from the inset in Fig. 2, points corresponding to shorter distances (3.00 to 3.50 Å) deviate from the remaining, more linear, points, and points from distances further than 10 Å show more noise. Discarding these points and redoing the fit improves the correlation coefficient and we obtain an exponent in the range -4.13 to -4.00 , which is closer to the expected -4 .

Determining the coefficient of the power-function turns out to be more difficult than the exponent. Using the regression coefficients fitted to all 49 points, and the volume of the 1.5 Å sphere (14.14 Å³), we obtain $c(\epsilon=1)=1.02$, $c(\epsilon=2)=0.64$, and $c(\epsilon=4)=0.69$ kcal mol⁻¹ Å⁻¹. Overall, however, we have to conclude that these values are associated with large uncertainties, and it is possible to determine α only somewhere in the range 0.3 to 1.4 kcal mol⁻¹ Å⁻¹, which corresponds to 0.2 to 1 units Å in terms of pK_a desolvation (the range quoted here is slightly larger than the three numbers imply since we also take into account additional fits not presented in this article). The fact that we are not able to determine the coefficient more precisely is not a problem since c is a parameter and its value is eventually obtained by fitting to experimental pK_a values. Nevertheless, since the biggest challenge in these calculations has to do with numerical stabilities for very small energies, we have also doubled the excluded volumes by adding a mirrored excluded volume in the opposite direction, thus getting larger effects and thereby improve numerical stabilities. The results are found to give similar values and the same conclusions (the resulting graph and table can be found in the supplementary information).

Even though the theoretical framework in this study suggests a power function, it is conceivable that other functions could also represent the desolvation reasonably well; for instance a Gaussian-shaped solvent exclusion such as the “effective energy function 1” (EEF1) [19] used in CHARMM. It is clearly possible to find such a function to represent the desolvation profile calculated in this study; however, a regression fit to the data points does not give a Gaussian-shaped function since the exponent becomes

positive, see Fig. 4. This is because the fit places larger weight on the long-range regime, whereas the Gaussian description is really only justified in the overlap regime [12]. In our case, however, we are effectively interested only in the non-overlapping region since van der Waals interactions prevent any significant overlap between non-bonded atoms. Fitting the data points to an exponential function gives a clear underestimate of the desolvation at close distances. Thus, we can conclude that the power function provided by Eq. 14 indeed represents the best option.

This section has so far considered an idealized system within an overall static picture to validate the form of our desolvation model. However, as mentioned above, embedding the system in a protein gives a more complicated response and therefore a different numerical value of the desolvation parameter. In this respect it is better by far to combine the desolvation model with functions for the other interaction types (resolvation and Coulomb contributions), and to determine the effective desolvation constants by fitting to experimental pK_a values. This also has the added advantage of compensating for parameter uncertainties such as the protein dielectric constant and absolute atom volumes. This was done in a recent study by fitting 6 empirical parameters to 85 experimentally determined Asp and Glu pK_a values [2], where we obtained the desolvation parameters $c_{\text{buried}}=1.27$ and $c_{\text{surface}}=0.32$ kcal mol⁻¹ Å⁻¹ when combined with the van der Waals radii defined in Table S1 in the electronic supplementary information. While we realize that there is not necessarily any numerical correspondence between these more reliable pK_a-fitted values and those obtained from the PB calculations above, we note that they are, for all practical purposes, within reasonable agreement.

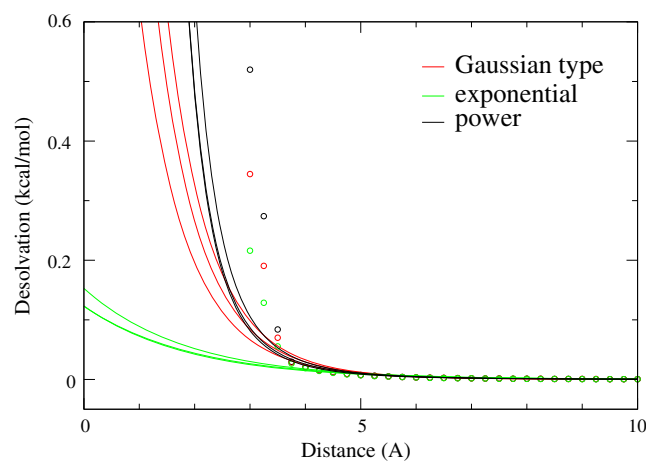


Fig. 4 Regression fit of our calculated desolvation energies to a power-, exponential-, and Gaussian-type function (black, green, and red respectively); of these three types of functions, the power function best represents the calculated points

Results and discussion

At this point, we have defined and evaluated our new desolvation model from a theoretical perspective; the next step is to evaluate how it behaves in the context of a real protein, and to compare the new with the previous model and show why the new model is more reasonable.

Comparison with the previous PROPKA desolvation model

Although the original PROPKA desolvation model, which is based on a contact scheme, reproduces desolvation effects for a number of buried residues in various proteins reasonably well, it has some conceptual errors that lead to failure of the model in more challenging cases. This problem, and the more general problem with the contact model, can best be summarized by considering the pK_a value of a significantly buried residue with a large desolvation penalty and scrutinizing its origin. As a test case, we chose residue Glu66 in the staphylococcal nuclease mutant V66E/P117G/H124L/S128A [20]. Here, an ionizable residue is introduced into a hydrophobic region of the protein and, consequently, its pK_a value is raised significantly compared to its water reference because of desolvation. In particular, a large desolvation contribution is likely to come from hydrophobic side-chain atoms from the residues T62, V23, L14, V99, I92, and L36, which are all oriented towards Glu66 (see Fig. 5, Table 2).

Unfortunately, PROPKA2 does not give a reliable estimate of this pK_a value, but experimentally we know that it is raised by 4 units compared to its reference water value, 8.5 versus 4.5, and since the program does not give any additional terms, we assume that the majority of this

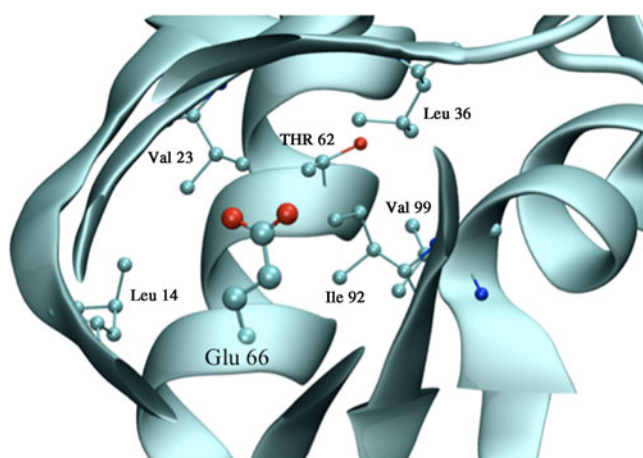


Fig. 5 Residue Glu 66 in the mutant V66E/P117G/H124L/S128A of staphylococcal nuclease is buried in a hydrophobic patch of the protein. The figure highlights some of the neighboring hydrophobic residues that increase its desolvation energy and therefore raise the pK_a value

Table 2 Closest atoms surrounding Glu 66 in the mutant V66E/P117G/H124L/S128A of staphylococcal nuclease

Residue	Atom	Distance (Å)
Thr 62	CG2	2.95
Val 23	CG2	3.59
Leu 14	CD2	5.15
Val 99	CG1	5.25
Ile 92	CG2	5.65
Leu 36	CD1	5.85

comes from the desolvation contribution in the PROPKA framework. This corresponds to a desolvation energy of $5.4 \text{ kcal mol}^{-1}$. Using the parameters defined previously, we obtain a desolvation energy close to $3.5 \text{ kcal mol}^{-1}$ for our new model compared to 2.9 for the old; this is already slightly better compared to the experiment value. Figure 6 depicts the desolvation contribution for 1-Å spherical segments as a function of the distance from the Glu 66 residue center for the old (red line) and new (green line) desolvation models. We see that the largest individual contributions to the old model come from distances 3–4 and 4–5 Å from the residue (0.22 and $0.23 \text{ kcal mol}^{-1}$, respectively), and that they originate mainly from the local desolvation term. However, we note that the third largest contribution comes from distances of 13–14 Å, and in fact, 53 % comes from the region 9 Å and outwards. For the new desolvation model, on the other hand, the largest individual contribu-

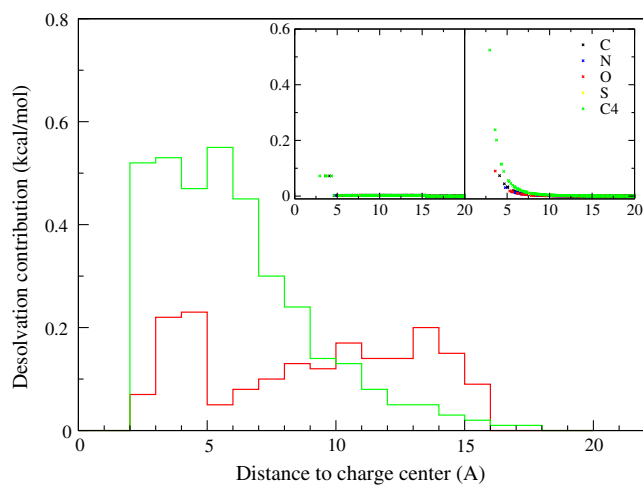


Fig. 6 Desolvation contributions accumulated for 1-Å spherical segments as a function of distance to the Glu 66 charge center; for the new volume desolvation model (green) it can be seen that the majority of the desolvation comes from a region 2–7 Å from the residue, whereas from the PROPKA2 contact model (red) there are significant contributions from the regions at 3–5 and 10–16 Å from the residue. *inset* Profiles showing the corresponding contribution for each atom for the new volume desolvation model (right) and the PROPKA2 contact model (left) for each atom type

tion comes from distances 5–6 Å ($0.55 \text{ kcal mol}^{-1}$), and only 15 % comes from the region 9 Å and outwards. In this case, it might not seem such a big problem, but staphylococcal nuclease is a comparatively small enzyme, which is also why the red line eventually falls off at 14–16 Å in the graph. For deeply buried residues in larger enzymes, the majority of desolvation accumulates close to the global desolvation cutoff ($r \approx 15.5 \text{ Å}$) since the number of atoms increases rapidly with distance. The two pictures above obviously conflict, and even though the origin of the desolvation depends on the shape of the protein and position of the residue, we turn back to the Born model to guide us to the correct asymptotic behavior as the distance increases.

We find from the Born model that the solvation contribution from 1 Å segments close to 15 Å should be vanishingly small compared to those close to r_0 , $\frac{1}{14} - \frac{1}{15} \ll \frac{1}{2} - \frac{1}{3}$, and this must also be true for the desolvation. Clearly, the volume desolvation behaves much more appropriately in this respect. Undeniably, the old desolvation model seems also to be counter-intuitive since, if we extend the radius for the heavy-atom count in Eq. 5 to be analogous to the infinite integration of Eq. 13, the sum would quickly approach infinity and the major desolvation contributions would come from extreme distances. It can also be seen from the inset, which depicts the individual atomic contribution for the contact model (left) and the volume model (right), that the specific radius chosen for local desolvation provides a discontinuity where an atom just outside R_{local} contributes very little, whereas an atom just inside contributes almost an order of magnitude more. This arbitrariness of the model and increased number of parameters is obviously an undesirable feature that reduces the applicability and generality of the model.

A more complete validation of the new desolvation model alone in terms of energy is not feasible with existing experimental data since pK_a values also contain contributions from other terms (i.e., protein resolution and Coulomb interactions). The second best option is to complete the pK_a -predicting model and validate the complete model against experimental values. This has been found to reduce the root mean squares deviation (rmsd) from 0.91 for PROPKA2 (which includes the old desolvation model) to 0.79 for PROPKA3 (which includes the new desolvation model) for a set of 201 Asp and Glu pK_a values [2]. The most notable improvement is seen for Asp 75 in barnase for which PROPKA predicts a pK_a shift of -5.3 and 1.0 (versions 2 and 3, respectively), whereas the experimental pK_a shift is close to -0.1 . This residue is found to be 77 % buried in the protein and surrounded by two nearby Arg residues. Since the experimental pK_a value is similar to its reference value it means that the electrostatic interactions and desolvation is almost balanced

and gives a small total shift (-0.1). The old desolvation model seems to significantly underestimate the desolvation effect and therefore predict an extremely low pK_a value, whereas the new model treats the two effects in a more balanced way, and predicts a pK_a value closer to experiment. Although this is by far the most extreme case among the 201 pK_a values, we found additional examples where the problem is less severe. We also note that the overestimate of PROPKA3 for this residue is probably related to an underestimation of Coulomb interactions rather than overestimation of desolvation. The discontinuity in the contact model provided by R_{local} is also found to create problems, and is aggravated in cases of π -stacking interactions. Interactions where r is slightly larger than R_{local} give virtually no desolvation contribution ($6 \times 0.01 = 0.06$ units for a phenyl group), whereas the contribution is non-negligible when r is slightly smaller than R_{local} ($6 \times 0.08 = 0.48$ units). At any rate, the new desolvation model seems to result both in a more physical description of the behavior of desolvation contributions and better pK_a predictions.

Conclusions

This study has shown that the desolvation model in PROPKA2 is fundamentally incorrect in that it does not properly take into account the solvent volume that is displaced by surrounding protein fragments, and it has a faulty distance dependence. This is remedied by a new volume desolvation model that uses atom volumes to quantify the desolvation penalty by relating the solvent inaccessible volume and the size of intervening protein atoms, i.e., large atoms such as sulfur atoms have a larger effect on desolvation than small atoms. The model also includes these contributions with the correct r^{-4} distance dependence, which provides an especially attractive feature since it removes all boundaries and unnecessary radii parameters that are not only often difficult to define, but essentially artificial.

The properties of the new volume desolvation model clearly represents an improvement over the previous PROPKA contact model, but the question of whether this model also provides a general improvement over the popular solvent ASA model remains. This is beyond the scope of the present study, but before addressing the question it should be realized that the answer might be more complicated than first perceived. Following the derivation outlined here, the ASA model seems quite archaic since it does not take into account the solvation from solvent behind a shielding contact group, and artificial crystal contacts and other crystallization effects therefore represent a severe problem because of these first-sphere

properties. However, the model has withstood the test of time, and for empirical pK_a predictions we need to be able to describe the free energy of charging the residue and capturing all the structural rearrangement that entails. If we want to model this rearrangement implicitly using single structures from pdb files, it is not obvious that the rearrangement is similar throughout the protein. It might be that surface regions of a protein respond differently compared to interior regions and the ASA model implicitly manages to capture that difference. All the same, it seems clear that the new volume desolvation model in PROPKA provides a better starting point for further investigation than the previous model, and indeed has rectified a number of both conceptual and practical problems while reducing the number of parameters.

Acknowledgment This work was supported by the Danish Council for Strategic Research by a research grant from the Program Commission on Strategic Growth Technologies (2106-07-0030).

References

1. Warshel A (1981) Calculations of enzymatic-reactions—calculations of pK_a , proton-transfer reactions, and general acid catalysis reactions in enzymes. *Biochemistry* 20:3167–3177
2. Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH (2011) PROPKA3: consistent treatment of internal and surface residues in empirical pK_a predictions. *J Chem Theor Comput* 2:525–537
3. Lee B, Richards FM (1971) Interpretation of protein structures—estimation of static accessibility. *J Mol Biol* 55:379
4. Born M (1920) Volumes and hydration warmth of ions. *Z Phys* 1:45–48
5. Tanford C, Roxby R (1972) Interpretation of protein titration curves—application to lysozyme. *Biochemistry* 11:2192
6. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21:1087–1092
7. 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
8. Cramer CJ, Truhlar DG (1999) Implicit solvation models: equilibria, structure, spectra, and dynamics. *Chem Rev* 99:2161–2200
9. Colonnacesari F, Sander C (1990) Excluded volume approximation to protein–solvent interaction—the solvent contact model. *Biophys J* 57:1103–1107
10. Li H, Robertson AD, Jensen JH (2005) Very fast empirical prediction and rationalization of protein $pK(a)$ values. *Proteins* 61:704–721
11. Bas DC, Rogers DM, Jensen JH (2008) Very fast prediction and rationalization of $pK(a)$ values for protein–ligand complexes. *Proteins* 73:765–783
12. Schaefer M, Froemmel C (1990) A precise analytical method for calculating the electrostatic energy of macromolecules in aqueous-solution. *J Mol Biol* 216:1045–1066
13. Gilson MK, Honig B (1991) The inclusion of electrostatic hydration energies in molecular mechanics calculations. *J Comput Aided Mol Des* 5:5–20
14. Schaefer M, Karplus M (1996) A comprehensive analytical treatment of continuum electrostatics. *J Phys Chem* 100:1578–1599
15. Jackson JD (1975) *Classical electrodynamics*. Wiley, New York, pp 146–159
16. Still WC, Tempezyk A, Hawley RC, Hendrickson T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc* 112:6127–6129
17. Schutz CN, Warshel A (2001) What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins* 44:400–417
18. Bashford D, Gerwert K (1992) Electrostatic calculations of the pK_a values of ionizable groups in bacteriorhodopsin. *J Mol Biol* 224:473–486
19. Lazaridis T, Karplus M (1999) Effective energy function for proteins in solution. *Protein Struct Funct Genet* 35:133–152
20. Denisov VP, Schlessman JL, Garcia-Moreno B, Halle B (2004) Stabilization of internal charges in a protein: water penetration or conformational change? *Biophys J* 87:3982–3994

Probing the structural requirements of A-type Aurora kinase inhibitors using 3D-QSAR and molecular docking analysis

Hui-xiao Zhang · Yan Li · Xia Wang · Yong-hua Wang

Received: 6 December 2010 / Accepted: 14 March 2011 / Published online: 28 April 2011
© Springer-Verlag 2011

Abstract Aurora-A, the most widely studied isoform of Aurora kinase overexpressed aberrantly in a wide variety of tumors, has been implicated in early mitotic entry, degradation of natural tumor suppressor p53 and centrosome maturation and separation; hence, potent inhibitors of Aurora-A may be therapeutically useful drugs in the treatment of various forms of cancer. Here, we report an *in silico* study on a group of 220 reported Aurora-A inhibitors with six different substructures. Three-dimensional quantitative structure–activity relationship (3D-QSAR) studies were carried out using comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) techniques on this series of molecules. The resultant optimum 3D-QSAR models exhibited an r_{cv}^2 value of 0.404–0.582 and their predictive ability was validated using an independent test set, ending in r_{pred}^2 0.512–0.985. In addition, docking studies were employed to explore these protein–inhibitor interactions at the molecular level. The results of 3D-QSAR

and docking analyses validated each other, and the key structural requirements affecting Aurora-A inhibitory activities, and the influential amino acids involved were identified. To the best of our knowledge, this is the first report on 3D-QSAR modeling of Aurora-A inhibitors, and the results can be used to accurately predict the binding affinity of related analogues and also facilitate the rational design of novel inhibitors with more potent biological activities.

Keywords Aurora-A · Inhibitor · 3D-QSAR · CoMFA · CoMSIA · Molecular docking

Introduction

Mammalian Aurora kinases comprise a family of three highly homologous serine/threonine kinases, namely Aurora-A, -B, and -C, which are involved in regulating multiple steps of mitosis, including centrosome duplication, formation of a bipolar mitotic spindle, alignment of chromosomes on the mitotic spindle, establishment and maintenance of the spindle checkpoint, and cytokinesis [1–5]. Since their discovery in 1995 [6], and the first observation of their expression in human cancer tissue in 1998 [7], these kinases have been the subject of intense research in both the academic and industrial oncology communities as novel attractive targets for anticancer therapy [8]. The biology of the three isoforms of Aurora kinase (Aurora-A, -B, and -C) has been reviewed extensively [2, 3]. It is found that, although they are very closely related in kinase domain sequence—Aurora B and C are 75% and 72% identical to Aurora A—certain discrepancies still exist in amino acid length and sequence at the N-terminal domain, and in the cellular localization, regulation, and substrate specificity of these kinases [5, 9].

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1042-3) contains supplementary material, which is available to authorized users.

H.-x. Zhang · X. Wang · Y.-h. Wang (✉)
Center of Bioinformatics, Northwest A&F University,
Yangling, Shaanxi 712100, China
e-mail: yh_wang@nwsuaf.edu.cn

Y. Li
School of Chemical Engineering,
Dalian University of Technology,
Dalian, Liaoning 116012, China

Y. Li
Lab of Pharmaceutical Resource Discovery,
Dalian Institute of Chemical Physics,
Graduate School of the Chinese Academy of Sciences,
Dalian, Liaoning 116023, China

Aurora A localizes to the centrosome and the mitotic spindle from prophase to telophase, and plays a critical role in regulating many early mitotic events, including entry into mitosis [7, 10, 11]. Depletion of Aurora-A results in delayed entrance into mitosis and formation of numerous monopolar spindles due to defects in centrosome maturation and separation and in the organization of the microtubules that form the spindle [8]. Aurora-A can phosphorylate Cdc25b, a direct regulator of the cyclin B1-Cdk1 complex whose activation is an essential requirement for mitotic entry [12]. In addition, phosphorylation of the kinesin motor protein HsEg5 (KSP)—a crucial driver of centrosome separation—by Aurora-A is associated with the later process of centrosome separation as the bipolar spindle forms [13]. Aurora-A is critical to the regulation of the EXTAK multiprotein complex comprised of the proteins Eg5, XMAP2154, TPX-2, Aurora-A, and HURP, which together act to bundle, crosslink, and stabilize the growing microtubule network [8]. Disruption of any component in the complex would perturb spindle formation and lead to mono- and multi-polar spindles [12]. Moreover, Aurora A can promote mdm2-mediated degradation of the natural tumor suppressor p53 and inhibition of its transcriptional activity [14, 15].

The Aurora-A gene lies within a region of chromosome 20q13 that is frequently amplified in many human cancers [7], and is also associated with the chromosomal instability phenotype in colorectal cancers [16]. Overexpression of Aurora-A has been reported to be transforming in some cell types [7, 10], and appears to associate with a wide variety of tumors, including those from colon [7], breast [10], ovary [17], pancreas [18], head, and neck [19]. In addition, transgenic mice overexpressing Aurora-A in the mammary gland develop mammary tumors at a high incidence rate [20]. These results provide compelling evidence that Aurora-A acts as an oncogene and plays a key role in cell cycle progression and carcinogenesis—an area that is emerging as a promising molecular targeted cancer treatment option.

A number of small molecule inhibitors of Aurora kinases have been developed, and more than ten such inhibitors have entered early clinical assessment [8]. ZM447439, a quinazoline derivative and the first Aurora kinase inhibitor to be developed in 2003, inhibits both Aurora-A and -B (IC_{50} values of 110 and 130 nM, respectively) [21]. VX-680/MK-0457, which is a 4,6-diaminopyrimidine that inhibits all three Aurora kinases (A, B, and C) with K_i values of 0.6, 1.8, and 4.6 nM, respectively, and was first demonstrated in 2004 to show potent antitumor activity in vivo [22]. Hesperadin is an indolinone inhibitor of Aurora-B (IC_{50} of 250 nM) with significant cross-reactivity against six other kinases (no data on Aurora-A or -C are reported) [23]. Examples of Aurora selective inhibitors include AZD1152 (the first Aurora-B selective inhibitor to enter clinical trials) [24], MLN8054 (the first reported Aurora-A

selective inhibitor) [25], and the most recently developed inhibitor, MK-5108 (Aurora-A selective) [26]. These Aurora inhibitors, which have diverse structures and biological activities, offer the potential to improve the treatment of cancer by helping to develop new drugs as well as by defining optimal therapeutic strategies.

In silico modeling has been demonstrated as one of the most widely used and effective tools in reducing costs and speeding up the drug discovery process. Nowadays, it has become an urgent task to design more potent Aurora inhibitors in order to present new strategies to identify therapeutics for cancer treatment. In order to understand the function–structure relationships of Aurora inhibitors, simple explorations based on the derivatives of some effective inhibitors have been carried out [8, 9, 27, 28]. Furthermore, crystallography studies have shown that the Aurora kinases can adopt a number of different conformations that represent distinct drug targets with alternative opportunities to derive potency and selectivity [8]. For instance, the crystal structure of VX-680 with Aurora-A kinase showed that the compound is bound to an “inactive-closed” conformation of the enzyme, and that the cyclopropyl group of the amide occupies a small hydrophobic pocket capped by a phenylalanine residue (Phe275). However, the crystal structures of activated “open” Aurora-A show that this pocket is not available in this conformation [8]. The crystal structure of Aurora-A and TPX2 illustrated that TPX2 makes two contacts with the Aurora-A kinase domain. The interactions between TPX2 and Aurora-A help mold the activation loop into a conformation that is ready for substrate binding, and also provide a lever arm-like mechanism that causes the rotation of phosphorylated T288 away from the solvent-exposed position found in free Aurora-A, thus protecting it from dephosphorylation by PP1 [29]. Recently, a structural study revealed the potential importance of Thr217 by revealing a hydrogen-bonding interaction with pyrazole compounds that exhibit specificity for Aurora A over Aurora B [28]. Despite the many co-complex structures that have been solved, in most cases a clear explanation for the observed inhibitory activity against Aurora kinases is still unclear. To date, a comprehensive review of the structural requirements of Aurora-A inhibitors based on quantitative structure–activity relationship (QSAR) has not been reported, highlighting the urgency of undertaking such studies to fill the blank in this field. Thus, in this work, two widely used QSAR methods, i.e., comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) [30, 31], were exploited to derive 3D-QSAR models for six different chemical series of Aurora-A inhibitors. These techniques were applied successfully in the past to various therapeutic areas in our laboratory [32–36]. In addition to 3D-QSAR analyses, docking

simulations were also performed to explore the molecular interactions between ligands and their receptors at the active site. To the best of our knowledge, this is the first attempt toward the establishment of 3D-QSAR for A-type Aurora inhibitors, which may help in designing and forecasting the Aurora-A inhibitory activity of novel molecules.

Materials and methods

Dataset

In order to build as large a dataset as possible, while still maintaining consistency of structure and bioactivity in generating QSAR models, a total of 220 molecules reported as Aurora-A inhibitors were collected from recently published data [8, 9, 27, 28, 37–41]. These chemicals have diverse structures, and the main skeletons of these molecules can be divided into six main groups (Table 1): Groups GI–GVI, comprise 37, 36, 25, 54, 24, and 44 molecules, respectively. The in vitro inhibitory activity, K_i or IC_{50} (μM) against Aurora-A was converted to pK_i or pIC_{50} in developing 3D-QSAR models. For each group, the molecules of the test set represent nearly 25% of the whole dataset. The strategy for selection of training and test sets was to ensure that test compounds represented a similar structural diversity and range of biological activities as the training set. To illustrate this, a principal component analysis (PCA) was performed on the dataset as follows: (1) more than 600 structural descriptors, including the topological, constitutional, walk and path counts, atom-centered fragments and connectivity indices for each molecule, were calculated for all the compounds using Dragon software (http://www.taletе.mi.it/help/dragon_help/); (2) PCA was then performed within the calculated structure descriptor space for the whole dataset, giving three significant principal components (PCs) that explain more than 70% of the variation in the data [42]. The structures and inhibitory activity data of the training and test set molecules are described in Tables S1, S2, S3, S4, S5 and S6, and details of distribution of the compounds over the three PCs for each class are depicted in Figs. S1, S2, S3, S4, S5 and S6 (Supporting Information).

Molecular modeling

The 3D-QSAR and molecular docking computations were carried out using Sybyl 6.9 (<http://tripos.com/>) on a Redhat Linux platform. The 3D structures of the training and test set compounds were built using the Sketch Molecule function in Sybyl. Optimization of the 3D structures was carried out using TRIPOS force field with the Gasteiger Hückel charges, and repeated minimization was performed

using Powell conjugate gradient method until a root-mean-square (rms) deviation of $0.001 \text{ kcal mol}^{-1}$ was achieved. Compound alignment was performed separately for each dataset with each respective common structure (Table 1, shown in bold). In each dataset, the most active compound was chosen as the template molecule and all compounds were aligned to a common substructure using the “align database” command in Sybyl software. The corresponding alignment results of the six groups are shown in Figs. S7, S8, S9, S10, S11 and S12 (Supporting Information).

CoMFA and CoMSIA analyses

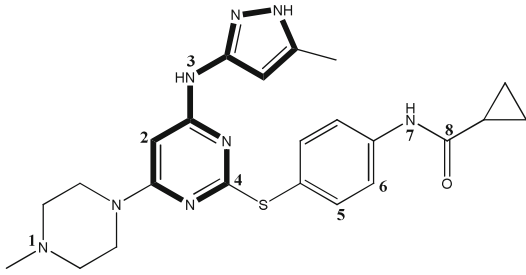
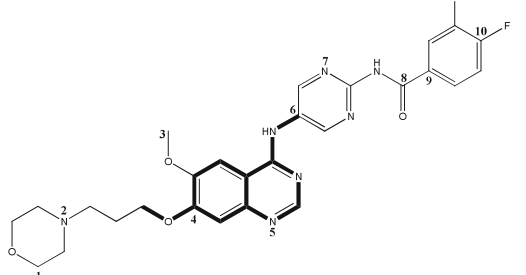
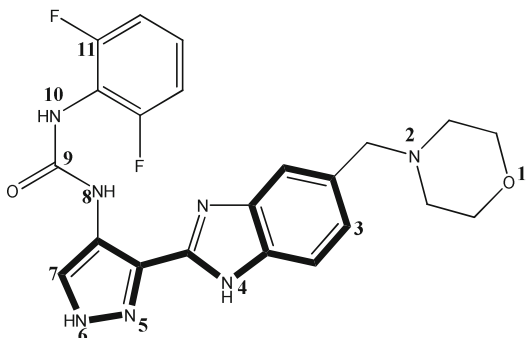
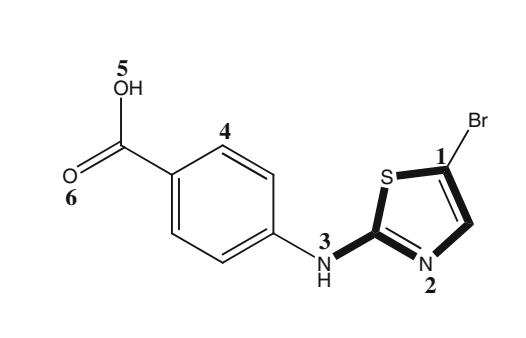
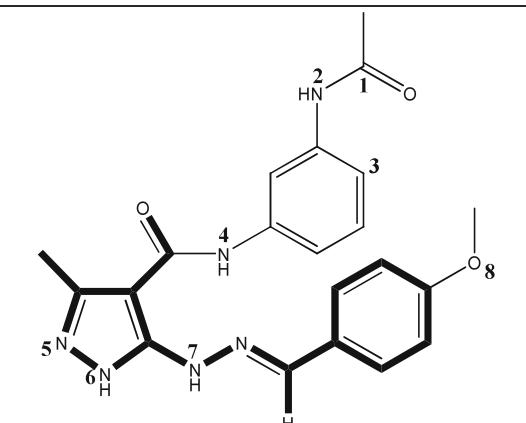
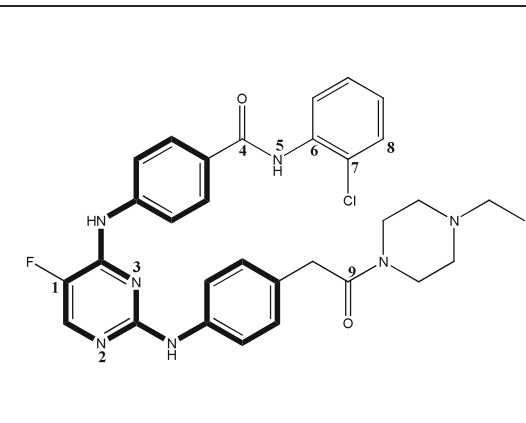
In order to derive the CoMFA and CoMSIA descriptor fields, a 3D cubic lattice with grid spacing of 2 \AA in x , y and z coordinates, was created to encompass the aligned molecules. CoMFA descriptors were calculated using an sp^3 carbon probe atom with a van der Waals radius of 1.52 \AA and a charge of $+1.0$ to generate steric (Lennard-Jones 6–12 potential) field energies and electrostatic (Coulombic potential) fields with a distance-dependent dielectric at each lattice point. The steric and electrostatic cutoff values were set to 35 kcal mol^{-1} for group II and 30 kcal mol^{-1} (default value) for the remaining groups, which are optimal parameters for the respective models. In CoMSIA analyses, the steric, electrostatic, hydrophobic, and hydrogen-bond donor and acceptor descriptors were calculated using the probe atom $C_{sp^3}^+$ with a radius of 1 \AA and a $+1.0$ charge placed at the lattice points of the same region of the grid as used for CoMFA calculations.

Partial least-square (PLS) regression analyses was used to evaluate the predictive values of models using the leave-one-out (LOO) cross validation method. The number of components leading to the highest cross-validated r^2 and lowest standard error of prediction was set as the optimum number of components in the PLS analyses; F value and standard error of estimates (SEE) were then calculated. The models were also evaluated for their ability to predict the activity of compounds in the test set. A detailed description of this method can be found in many previous works [32, 35, 36].

Molecular docking

Molecular docking analysis was carried out using the Surflex module of the Sybyl package to explore the interaction mechanism and to illustrate the accurate binding model for the active site of Aurora-A with its ligands [43]. Up to now, 38 various Aurora-A crystal structures complexed with different inhibitors have been reported in the RCSB Protein Data Bank (<http://www.pdb.org>). To ensure reasonable docking models, the selection of Aurora-A crystal structures was made according to the following criteria:

Table 1 Main skeletons (shown in bold) and template molecular structures in each group with corresponding inhibitory activities (pK_i or pIC_{50}) for Aurora-A kinase

Group I		Group II	
			
Compound	pK_i (μM)	Compound	pIC_{50} (μM)
37	3.222	68	3.824
Group III		Group IV	
			
Compound	pIC_{50} (μM)	Compound	pIC_{50} (μM)
81	2.824	119	1.102
Group V		Group VI	
			
Compound	pIC_{50} (μM)	Compound	pIC_{50} (μM)
176	1.481	220	2.469

- (1) The ligand in the crystal structure to be applied should share a common structure with certain group compounds; also, the most active compound in the corresponding dataset should have a reasonable docking score (total score of 5.96 on average) in obtaining the models. Therefore, the following PDB files were used: 3E5A for G-I [44], 2C6E for G-II [40], 2W1E for G-III [38], 3FDN for G-V [28], and 2NP8 for G-VI [45].
- (2) For the remaining group, G-IV, no common structure was observed with any of the ligands in the X-ray complexes. In this case, the file 1MUO.pdb [46] was selected as the co-crystallized ligand shares several highly topological similarities, such as molecular size, shape, distribution of H-bond donors/acceptors with the most active compound **119** in G-IV.

In Surflex-docking, protomol construction was based on protein residues proximal to the native ligand and on parameter settings to produce a small and buried docking target. Two parameters, i.e., `protomol_bloat` and `protomol_threshold`, which determine how far from a potential ligand the site should extend, and how deep into the protein the atomic probes used to define the protomol can penetrate, were adjusted to produce reasonable docking results (for detailed values, see section on [Docking analysis and comparison with 3D contour maps](#) below). For receptor preparation, all ligands were first removed and the polar hydrogen atoms were added. Water molecules in 3E5A, 2C6E, 2W1E, 3FDN and 2NP8 crystal structures were not removed for the reason that co-crystallized water molecules were found in the active site and could be involved in ligand–protein interactions by forming mediating H-bonds between the ligand and the protein. No water molecules were considered for docking with G-IV, since this protein receptor 1MUO.pdb has no co-crystallized water molecules in the active site. Automatic Mode was adopted to generate the protomol, and other parameters used the default values of this software.

Results and discussion

All combinations of CoMFA and CoMSIA models for the 220 compounds were calculated and analyzed; only the optimal 3D-QSAR models for each class are listed in Table 2. The best models were selected primarily on the basis of better cross-validated r^2 and predictive r^2 values and the chosen models were then exploited to generate 3D contour maps. In addition, a parameter, r_m^2 , was included to validate the external predictability of QSAR models, and a value of $r_m^2 > 0.5$ could be taken as an indicator of good external predictability [47]. The plot of actual versus

predicted activities for the training and test set molecules for each class is depicted in Fig. 1, where the data points are rather uniformly distributed around the regression line, indicating that the obtained models are reasonable.

In 3D-QSAR analyses, one of the major obstacles lies with the ‘congeners’, which misfit the final equation and are termed as outliers. In our study, several factors may account for the outliers: (1) unique structural differences such as compounds **20** and **29**, which have a $-t\text{Bu}$ substituent in the GI series; (2) different binding conformations like compounds **145** and **166** that have very low binding affinity in docking analysis (2.72 and 3.32, respectively); and (3) a higher residual between the observed and predicted biological activity, as in the case of compounds **71** and **199**, which have residuals more than 1 log unit. All these compounds were deleted from the data set, and the 3D-QSAR models were derived from the remaining compounds; the resulting models served as the basis for further assessment and discussion.

Graphical interpretation of the 3D-QSAR models

One of the attractive features of 3D-QSAR modeling is that the results can be visualized as 3D coefficient contour plots. To aid the visualization, the most potent molecule in each group of compounds is displayed and discussed as the reference compound. In order to select appropriate contour levels for each feature, the resulting histograms of actual field values were analyzed, and a contour level was chosen interactively as that producing the best interpretable contour map.

Group-I

In Fig. 2a, the yellow contours near position 2 indicate that bulky substituents at this position are not favorable for inhibitory activity. This is in accordance with the findings of Pollard et al. [8], showing that improvements in bioactivity were obtained upon replacement of the quinazoline with 6-heterocyclic substituted pyrimidines (compounds **21** and **32–34**). A large, sterically unfavorable, yellow polyhedron is seen near positions 4 and 5. In the CoMSIA electrostatic field, the blue contour observed near positions 5 and 6 indicates that a negatively charged group at these positions would have a detrimental effect on biological activity. The red contour near position 8 suggests the favorability of electronegative groups for inhibitory activity (compounds **16**, **17** and **37**). Figure 2c shows the contour map of the hydrophobic field with compound **37** overlaid. In the CoMSIA hydrophobic field, a large white contour seen in the vicinity of positions 7 and 8 indicates that a hydrophilic substituent at these positions is favored for inhibitory activity. There is also a small yellow region

Table 2 Summary of statistical results of the optimal three-dimensional quantitative structure–activity relationship (3D-QSAR) models for each of the six groups GI–GVI. r_{cv}^2 Cross-validated correlation coefficient using the leave-one-out (LOO) methods, N_C optimal number of components, SEP standard error of prediction, r_{ncv}^2 non-cross-validated correlation coefficient, SEE standard error of estimate, r_{pred}^2 predicted correlation coefficient for the test set of compounds, S steric, E electrostatic, H hydrophobic, D H-bond donor, A H-bond acceptor

	G-I	G-II	G-III	G-IV	G-V	G-VI
	CoMSIA	CoMFA	CoMSIA	CoMSIA	CoMSIA	CoMSIA
r_{cv}^2	0.501	0.404	0.582	0.432	0.549	0.454
N_C	7	6	6	6	4	6
SEP	0.135	0.333	0.241	0.288	0.048	0.280
r_{ncv}^2	0.982	0.973	0.982	0.809	0.986	0.964
SEE	0.089	0.216	0.168	0.240	0.083	0.151
F value	147.609	119.974	110.034	21.838	205.669	110.175
r_{pred}^2	0.946	0.809	0.928	0.512	0.985	0.719
r_m^2	0.890	0.552	0.838	0.507	0.975	0.662
Contribution (%)						
S	25.6	45.4	43.1	21.3	12.0	13.7
E	28.0	54.6	56.9	25.8	58.6	44.9
H	46.4	-	-	-	-	41.4
D	-	-	-	-	29.4	-
A	-	-	-	52.9	-	-

near the white contour, suggesting that a hydrophobic substituent around this yellow region would also enhance inhibitory activity. In addition, a white contour observed near position 1 signifies that the introduction of hydrophilic group at this position would improve inhibitory effects on the enzyme (compounds **36** and **37**).

Group-II

The steric contour plot of the best model with the template molecule (compound **68**) is shown in Fig. 3a. The green contour observed near position 6 suggests that bulky substituents may favor activity, yet the yellow contour near

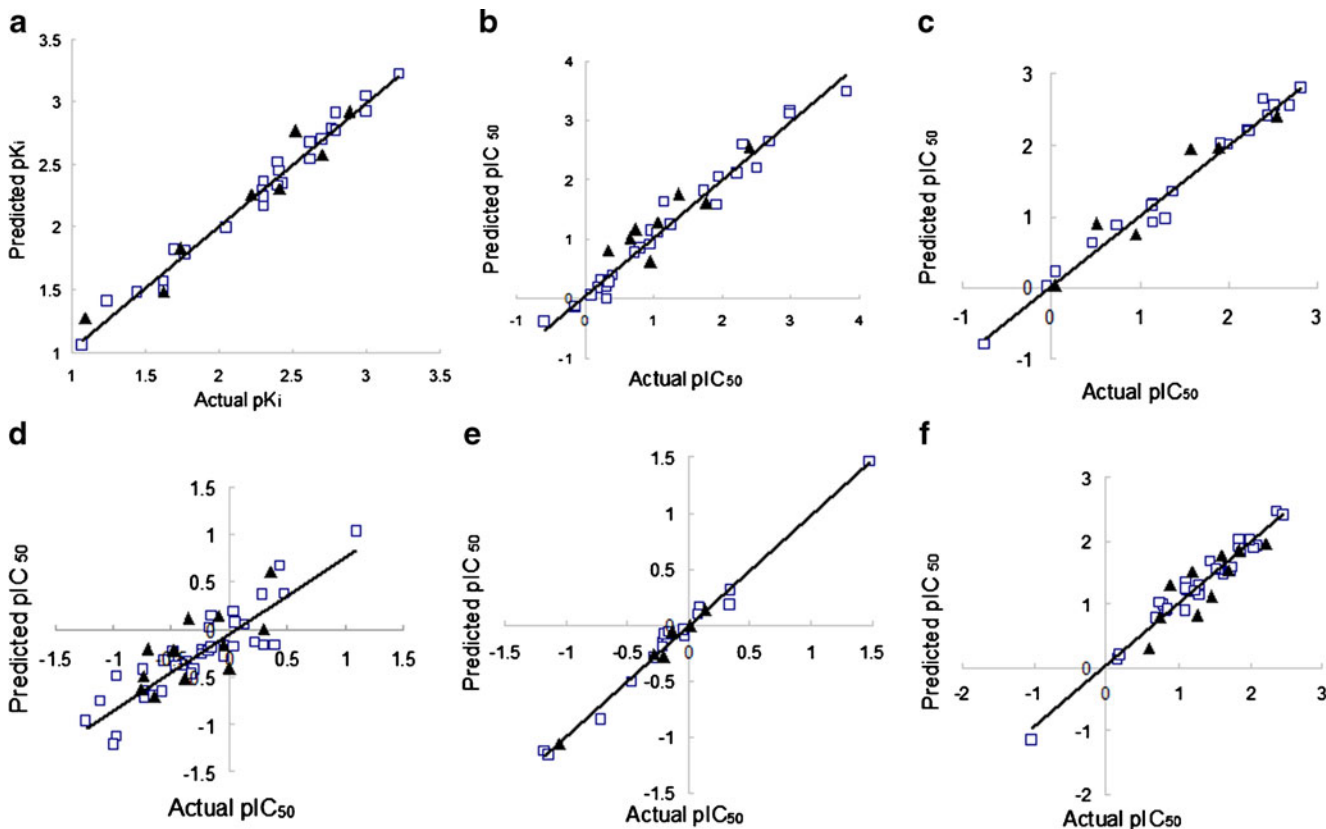


Fig. 1 Plots of the predicted versus experimental activity data of the optimal three-dimensional quantitative structure–activity relationship (3D-QSAR) model in each group (GI–GVI) for the training and the test set compounds. □ Training set, ▲ test set

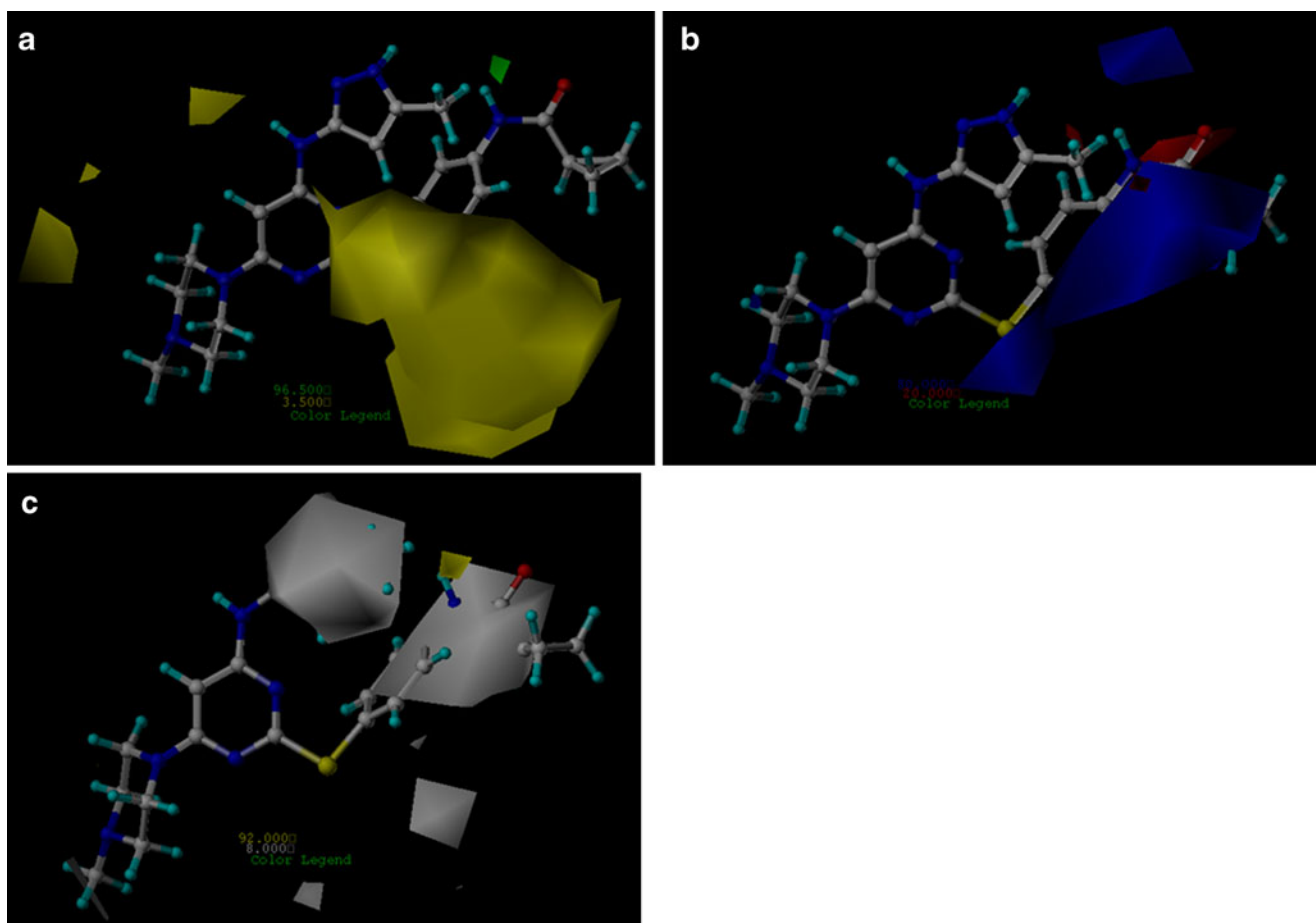


Fig. 2 Comparative molecular similarity indices analysis (CoMSIA) stdev*coeff (a) steric, (b) electrostatic and (c) hydrophobic contour maps for Group I. Color code: a green and yellow contours favorable and unfavorable bulky groups, respectively; b blue and red contours

favorable and unfavorable electropositive groups, respectively; c yellow and white contours favorable and unfavorable hydrophobic groups, respectively. Compound 37 in ball and stick is displayed as a reference

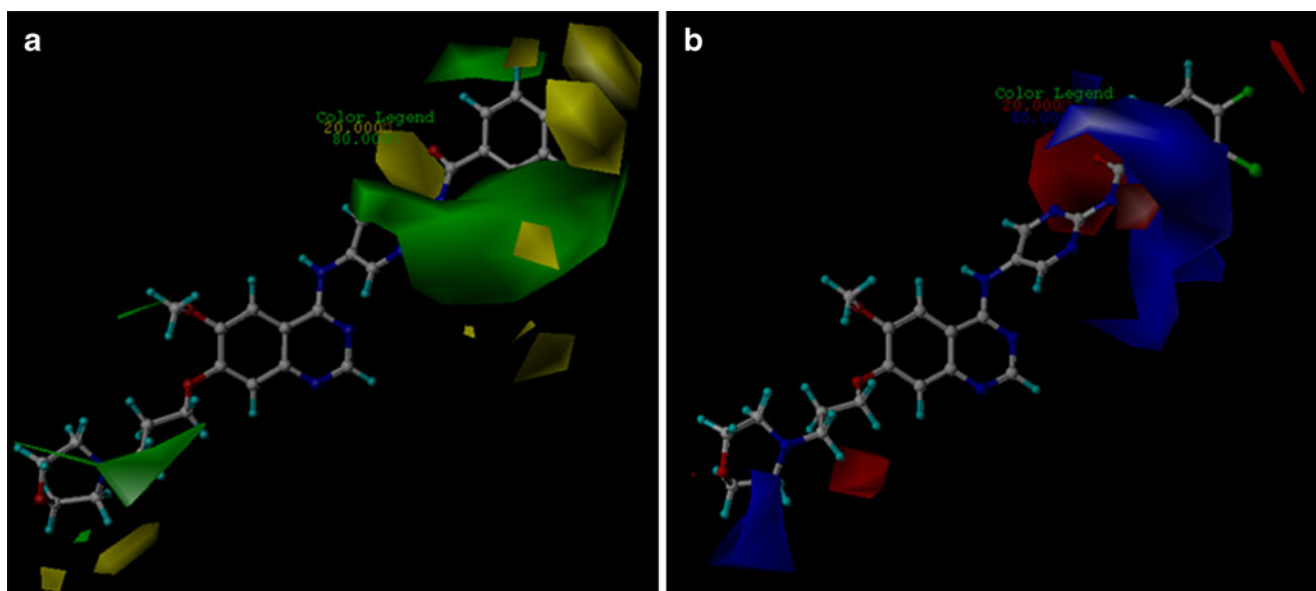


Fig. 3 Comparative molecular field analysis (CoMFA) stdev*coeff (a) steric and (b) electrostatic contour maps for Group II. Color codes of a and b as in Fig. 2. The compound 68 in ball and stick is displayed as a reference

position 10 indicates that a bulky substituent would decrease biological activity. A small green contour near position 3 indicates that a sterically bulky group is favored in this region (compounds **45**, **47** and **48**). The electrostatic contour map with the reference compound **68** is described in Fig. 3b. The red contour near position 10 indicates the significance of a negatively charged group for biological activity. The positively charged blue contour near position 1 suggests that a compound activity might be decreased by an electronegative group at this position (compounds **46** and **49**).

Group-III

The CoMSIA model of steric contribution is shown in Fig. 4a, with compound **81** overlaid on the map. A large yellow contour near position 7 indicates that compounds like **97** with bulky substituents (–COOEt) entering this yellow region will be less active than those unsubstituted or with small substituents like compounds **94** and **98** (–CH₂OH). A small green contour at position 10 suggests the requirement for a bulky substituent in this area to enhance biological activity. The CoMSIA electrostatic map is displayed in Fig. 4b. Clearly, a blue region is observed near position 7, suggesting a high demand for positively charged substituent in this region to improve inhibitor activity. The red contour near position 1 indicates that its occupancy by negatively charged groups would favor inhibitory activity, as revealed by compounds **74** and **78**. Another small red contour near position 11 suggests an electronegative group is preferred in this region (compounds **79** and **80**).

Group-IV

For the CoMSIA steric model (Fig. 5a, with compound **119**), a large green contour at position 4 suggests that

occupancy of this sterically favorable region with a bulky substituent would lead to an increase in bioactivity. The green contour located near position 1 indicates that a bulky substituent is preferred at this position (compounds **145** and **146**). Figure 5b showed the CoMSIA electrostatic contour plot with compound **119** overlaid on the map. The blue contour plot near position 4 indicates that an electropositive group is favorable. This is consistent with the experimental results that compound **130** shows higher activity than **129** and **131** since **130** has a more electropositive group (3-Me) than **129** (3-F) and **131** (3-CF₃) in this region. In the H-bond acceptor contour map (Fig. 5c), the red contour near position 5 indicates that an H-bond donor group is favored as supported by the fact that compound **119** is more active than **118** since **119** has an H-bond donor group (–OH) herein while **118** does not (–OEt). The magenta contour observed near position 4 suggests that the H-bond acceptor enhanced molecular activity.

Group-V

The graphical representation of the CoMSIA steric field with reference compound **176** is displayed in Fig. 6a. The green contour near position 1 suggests that a bulky substituent may be necessary to increase the inhibitory potency of the compound. A large yellow contour located at position 3 indicates that bulky substituents have unfavorable steric interactions (compounds **169**, **173** and **175**). In CoMSIA electrostatic field (Fig. 6b with compound **176**), a positive charge favored blue contour is observed near position 4, which is in accordance with the finding of Coumar et al. [28], that replacement of O with NH at this position would enhance compound activity. The graphical representation of CoMSIA H-bond donor field is shown in Fig. 6c. A small cyan contour near position 4 indicates that the H-bond donor group is favorable for activity. This is verified by experiment results that compound **157** exhibits

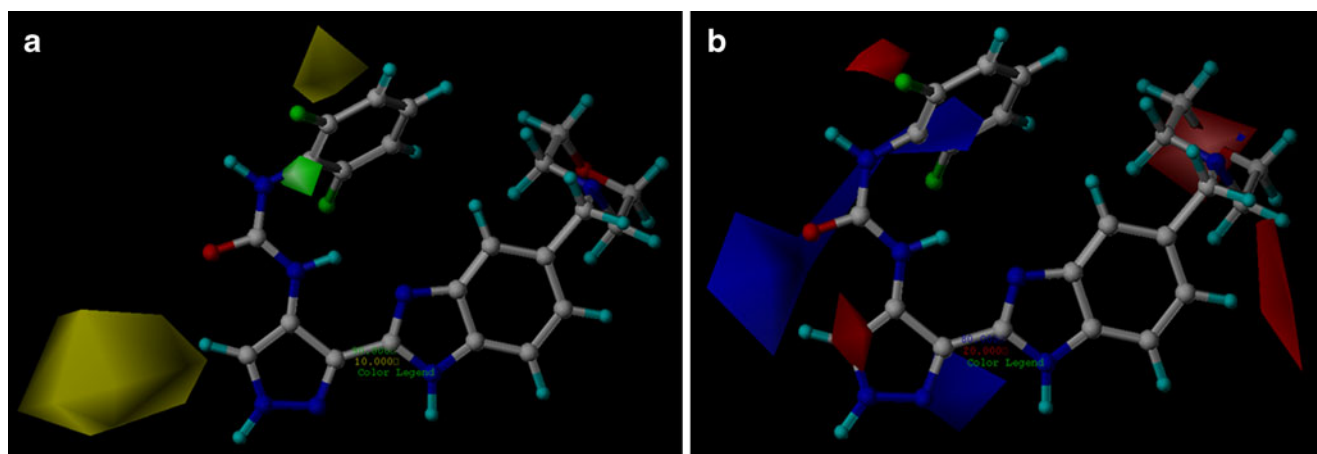


Fig. 4 CoMSIA stdev*coeff (a) steric and (b) electrostatic contour maps for Group III. Color code of a and b as in Fig. 2. The compound **81** in ball and stick is displayed as a reference

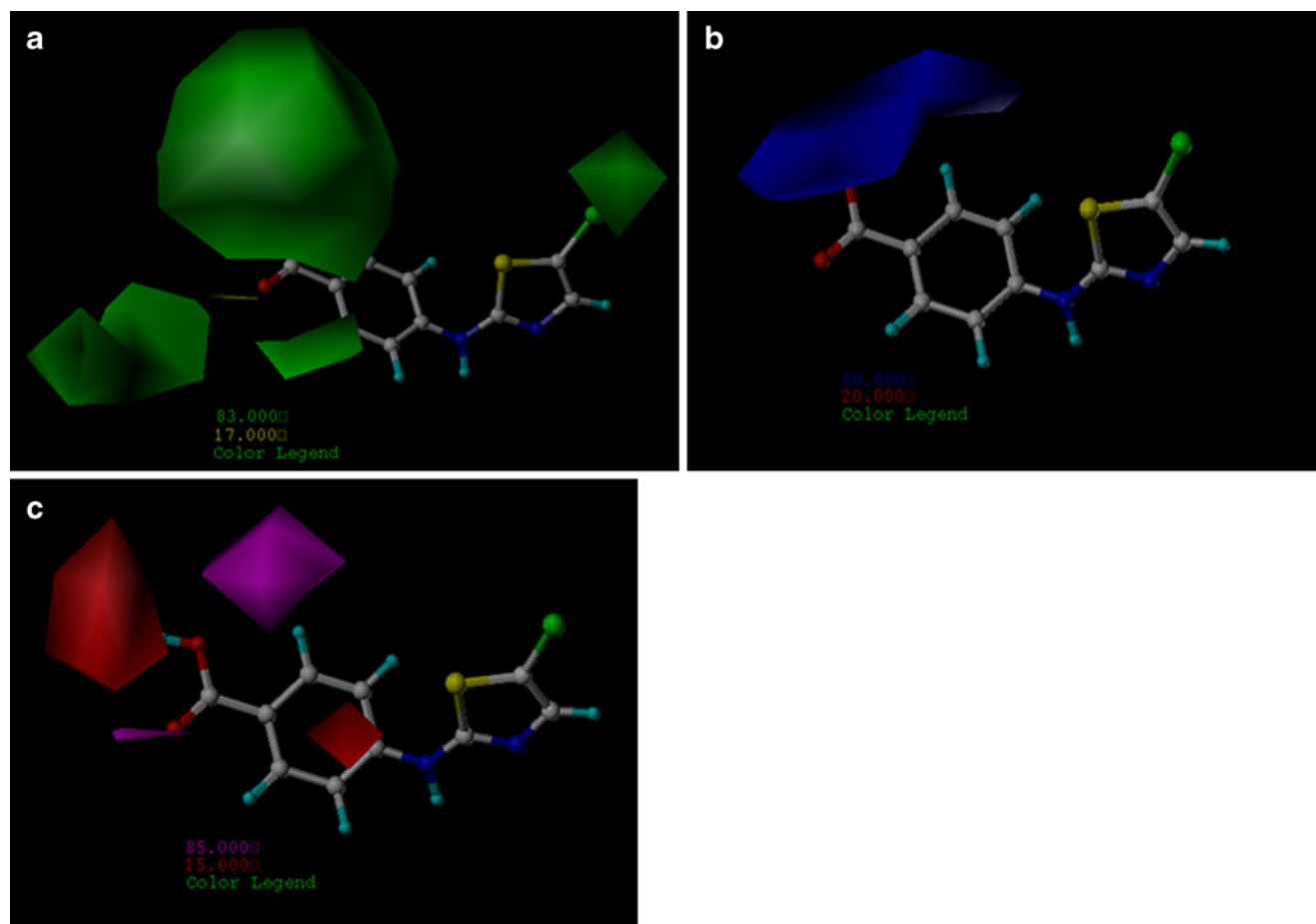


Fig. 5 CoMSIA stdev*coeff (a) steric, (b) electrostatic and (c) H-bond acceptor contour maps for Group IV. Color code for a and b as in Fig. 2; c magenta and red contours favorable and unfavorable

H-bond acceptor groups, respectively. The compound **119** in ball and stick is displayed as a reference

higher activity than **156** since **157** has an H-bond donor group (–NH–) at this position while **156** does not (–O–). A purple contour near position 1 suggests there would be a positive effect on biological activity by having an H-bond acceptor replaced in this region.

Group-VI

The steric contour map of the CoMSIA model with compound **220** is displayed in Fig. 7a. The yellow contour observed near position 1 indicates that a bulky substituent may decrease biological activity, which agrees partly with the finding by Aliagas-Martin et al. [37] that smaller aliphatic groups are preferred at this position (compounds **188**, **189** and **190**). A large green contour is seen near position 6, suggesting that a bulky substituent is favorable in this region, as confirmed by the fact that compounds **200** and **201**, with substituents *i*-Pr and Ph, respectively, show higher activity than unsubstituted analogue **197**. A small green contour near position 9 signifies that occupation of this area by a bulky group would have a positive effect on activity. The electrostatic contour map of the

CoMSIA model with compound **220** is shown in Fig. 7b. A small red contour near position 1 indicates the requirement for increased electron density in this area, which is in accordance with the findings of Aliagas-Martin et al. [37] that electron-withdrawing substituents, especially halogens, are preferred in this region (compounds **181**, **185**, **187** and **189**). The blue contour map observed near position 5 suggests that electro-negative groups are not favored for inhibitory activity (compounds **211** vs **217** and **218** vs **220**). The hydrophobic contour map of the CoMSIA model with compound **220** is shown in Fig. 7c. The white contour near position 8 indicates that its occupancy by a hydrophilic group would enhance activity. A medium size yellow contour located near position 7 suggests that a hydrophobic group is favorable for inhibitory activity (compounds **202**, **205** and **208**).

Docking analysis and comparison with 3D contour maps

All the 220 molecules in six different groups were docked into the active site of Aurora-A protein. Prior to docking the inhibitors with the protein crystal structure, a redocking of the

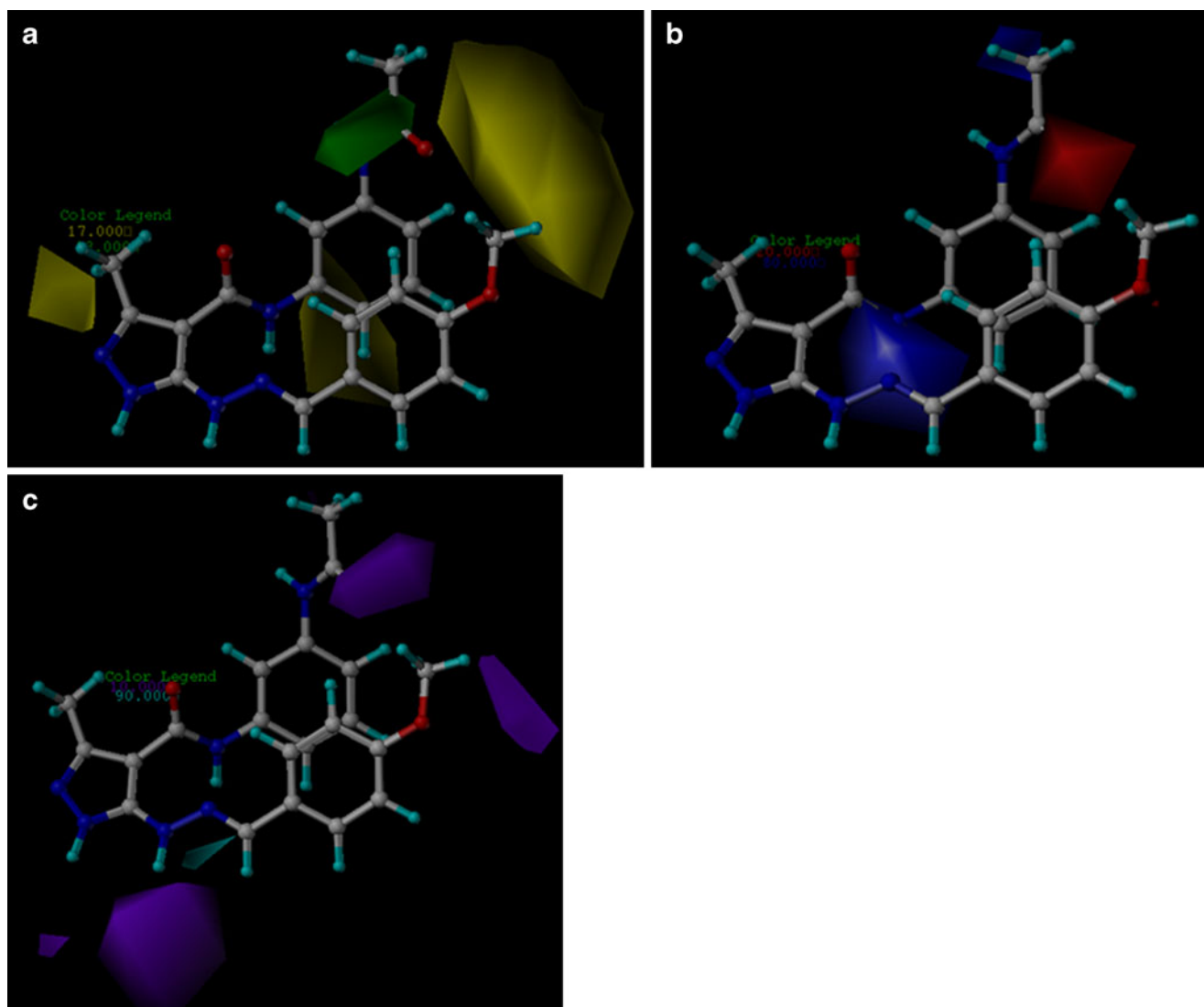


Fig. 6 CoMSIA stdev*coeff (a) steric, (b) electrostatic and (c) H-bond donor contour maps for Group V. Color code of **a** and **b** as in Fig. 2; **c** cyan and purple contours favorable and unfavorable

the co-crystallized ligand was performed by removing the ligand from the binding site and redocking it to the binding site of Aurora-A kinase. Our analysis suggests good agreement between the localization of the inhibitor observed upon docking and that from the crystal structure as evidenced by the result that RMSD values in each group (I–VI) were 0.87 Å, 1 Å, 0.34 Å, 0.02 Å, 0.27 Å and 1.5 Å, respectively. The low RMSD values suggest the high docking reliability of Surflex-Dock in reproducing the experimentally observed binding mode for Aurora A kinase inhibitor and the parameter set for Surflex-docking reproduces X-ray structures with reasonable accuracy.

Group-I

The protomol bloat and threshold applied the default values (0 and 0.5, respectively) and the binding mode of

H-bond donor groups, respectively. The compound **176** in ball and stick is displayed as a reference

compound **37** is displayed in Fig. 8. The ligand is anchored in the binding site via three H-bonds and one water-mediated contact with the protein. Pyrazole –N– and –NH ring atoms form H-bonds with the backbone at Ala213 (–N⋯HN, $d_1=2.08$ Å, $\theta_1=146.8^\circ$) and Glu211 (–NH⋯O, $d_2=2.34$ Å, $\theta_2=77.1^\circ$), respectively. The –NH– nitrogen atom at position 3 forms a H-bond with the carbonyl oxygen atom on the backbone at Ala213 (–NH⋯O, $d_3=2.24$ Å, $\theta_3=169.2^\circ$). The oxygen atom at position 8 forms a H-bond (2.68 Å, 144.9°) with water16, which itself forms H-bonds to the backbone –NH of Phe275, side chain –OH of Glu181 and carbonyl oxygen atom of Gln185. Substituents like phenyl directly linked to position 4 would potentially have a steric clash with residue Phe144, as is evident from the presence of a CoMSIA large sterically unfavorable yellow contour. The side chain –NH– of the

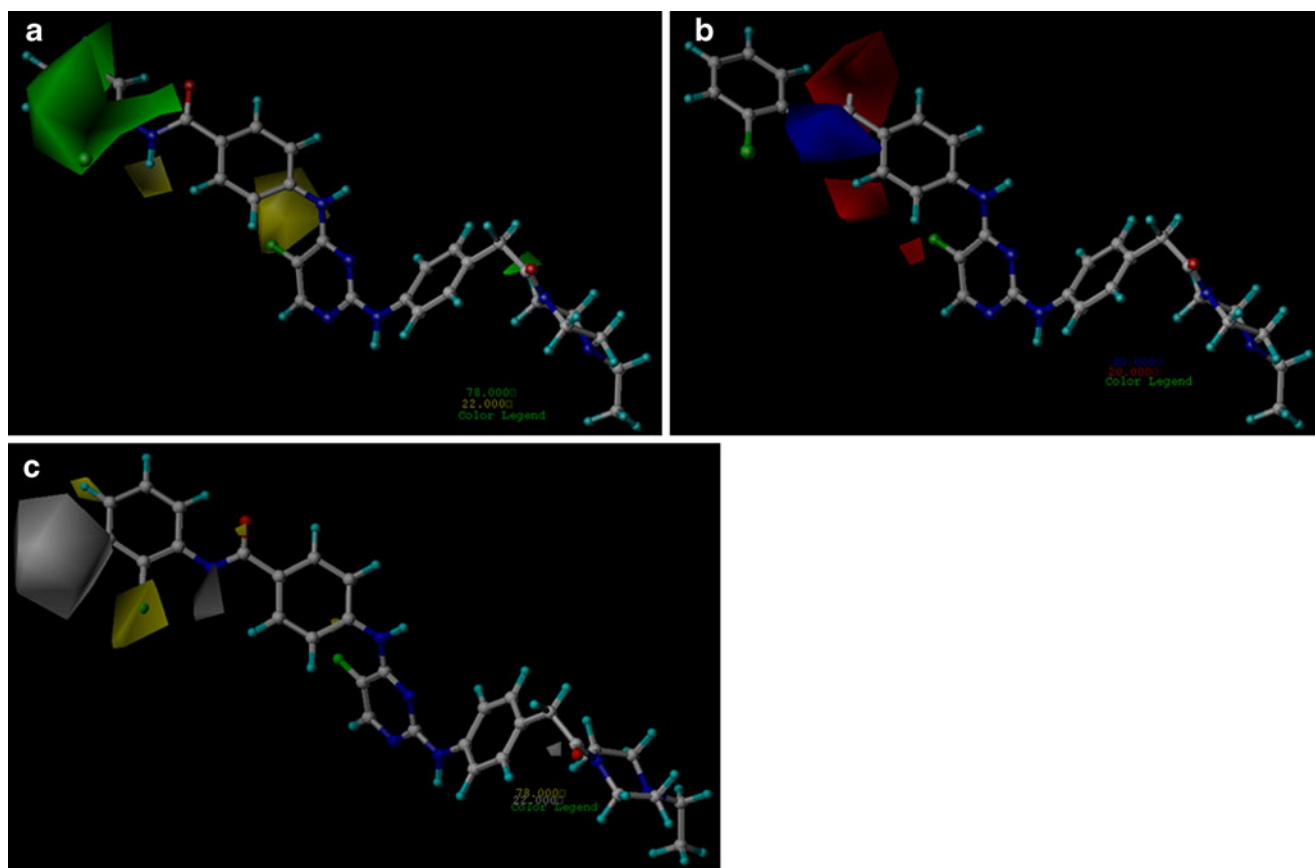


Fig. 7 CoMSIA stdev*coeff (a) steric, (b) electrostatic and (c) hydrophobic contour maps for Group VI. Color code for a, b and c as in Fig. 2. The compound **220** in ball and stick is displayed as a reference

Gln185 residue and water molecule (w16, 3E5A.pdb) near position 8 suggests a requirement for an electronegative group like carbonyl, which is in accordance with the CoMSIA red contour observed herein. The presence of

the white contour for the pyrazole ring indicates a hydrophilic favorable region, as confirmed by the docking results that two H-bonds exist in this region between the pyrazole ring and residues Ala213 and Glu211, respectively. According to the docked structure, the small white contour observed near position 1 suggests that the substituent at this position is exposed to the solvent.

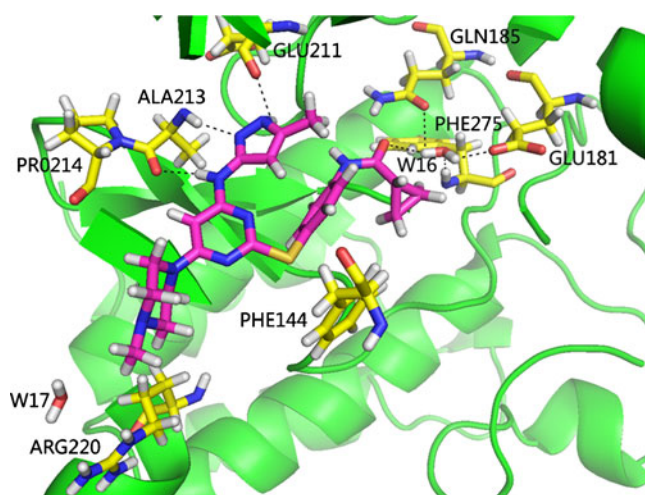
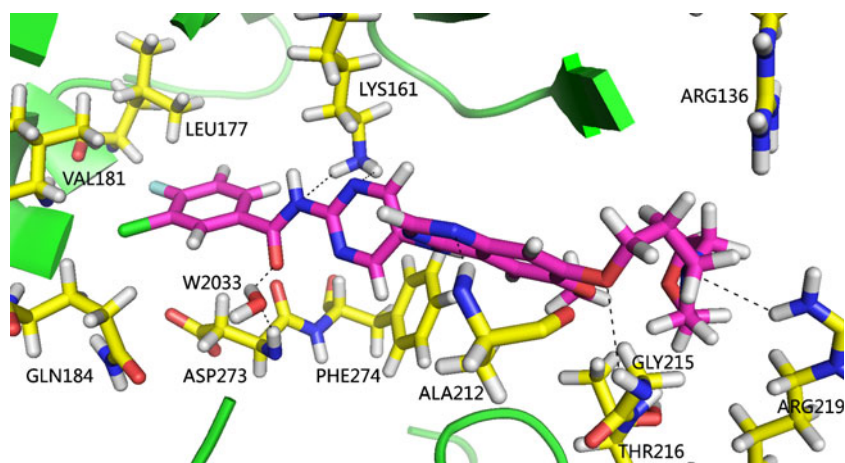


Fig. 8 Docked conformation derived for compound **37** with the binding site of Aurora-A kinase. H-bonds are shown as dotted black lines. Active site amino acid residues and the inhibitor are represented in stick model. W16 and W17 represent water molecules

Group-II

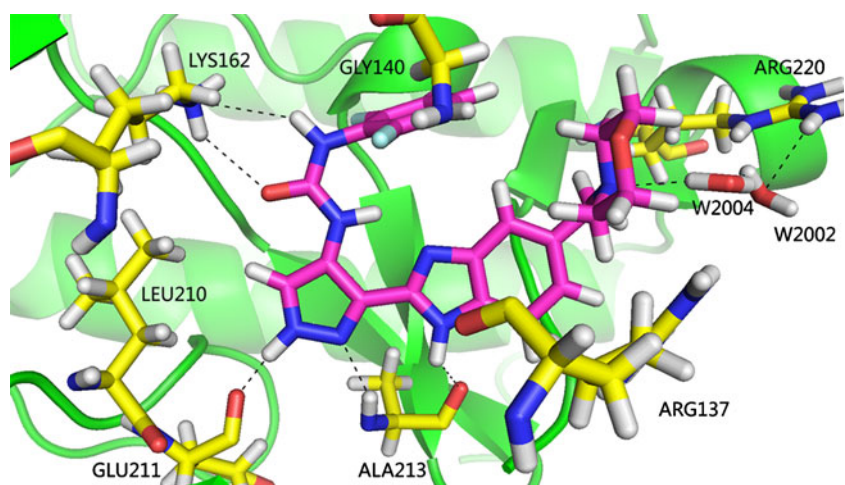
The protomol bloat and threshold values were set to 0 and 0.6, respectively, and compound **68** is shown in Fig. 9. Five H-bonds and one water-mediated interaction exist in the active site of the protein structures. For example, a quinazoline ring nitrogen at position 5 interacts through H-bonding with the backbone of the Ala212 amino acid residue ($-N\cdots HN$, $d_1=2.25\text{ \AA}$, $\theta_1=160.7^\circ$). The nitrogen atom of the morpholine ring at position 2 forms a H-bond with the guanidino group of Arg219 ($-N\cdots HN$, $d_2=3.09\text{ \AA}$, $\theta_2=127.3^\circ$). The pyrimidine ring N atom at position 7 is located within H-bonding distance (2.97 \AA) from the Lys161 side-chain amino function. The water-mediated interaction (i.e., forming H-bonds with proteins through water molecules) is observed between the carbonyl oxygen

Fig. 9 Docked conformation derived for compound **68** with the binding site of Aurora-A kinase. H-bonds are shown as *dotted black lines*. Active site amino acid residues and the inhibitor are represented as *stick model*. W2033 represents a water molecule



at position 8 and the backbone $-NH$ of Asp273. The 4-substituent of the quinazoline ring binds to the solvent-exposed pocket, where it interacts with Arg136, Thr216 and Arg219 amino acid residues. Comparing the docked structure and the 3D contour plots reveals that the yellow contour is present in the regions of Leu177 and Val181. Hence, a bulky substitution at position 10 would have an unfavorable steric interaction, which may also explain the lowest activity of compound **73**. Another sterically unfavorable region (yellow contour) is located near the carbonyl oxygen atom at position 8. Our docked model shows that a bulky substituent at this position would have an unfavorable steric clash with the backbone of residue Asp273. The carbonyl group at position 8 is observed near the backbone $-NH$ group of Asp273. This may explain the increased activity of compounds with electronegative groups at this position and is consistent with the CoMFA red contour presented in this region. A large blue contour seen in the vicinity of position 9 suggests a favorable electropositive region, as corroborated by the presence of several amino acid carbonyl groups of Phe274, Asp273 and Gln184 in this region.

Fig. 10 Docked conformation derived for compound **81** with the binding site of Aurora-A kinase. H-bonds are shown as *dotted black lines*. Active site amino acid residues and the inhibitor are represented as *stick model*. W2002 and W2004 represent water molecules, respectively



Group-III

The default values of protomol bloat and threshold were applied and compound **81** is described in Fig. 10. A total of five H-bonds and one water-mediated interaction are formed between compound **81** and Aurora-A kinase. The pyrazole ring nitrogen at positions 5 and 6 forms H-bonds with the backbone of Ala213 ($-N\cdots HN$, 2.23 Å, 152.3°) and Glu211 ($-NH\cdots O$, 1.88 Å, 157.7°), respectively. The N atom at position 4 enters into a H-bonding interaction with the carbonyl group of Ala213 ($-NH\cdots O$, 2.38 Å, 117.8°). The carbonyl oxygen atom at position 9 and nitrogen atom at position 10 form H-bonds with the side chain of Lys162 ($-O\cdots HN$, 3.00 Å, 131.1° and $-NH\cdots N$, 3.35 Å, 143.9°), respectively. Interaction between the morpholine ring N at position 2 and the side chain guanidino group of Arg220 is mediated by a water bridge formed by water 2002 and water 2004 (2W1E.pdb, Fig. 10). The presence of residue Leu210 near position 7 of the pyrazole ring indicates that a bulky substituent is not favored in this region, which is in agreement with the 3D contour plots showing that a large sterically unfavorable yellow contour is located at this

position. Those binders with larger substituents at position 10 are generally better because the space in the receptor binding site is relatively large. The red contour near position 3 suggests a negative charge favorable region, as verified by the $-\text{NH}_2$ of the guanidino group of Arg137 located herein. The presence of the $-\text{NH}$ group on the backbone of Gly140 near position 11 indicates the preference of electronegative groups at this position, which can also be inferred from the CoMSIA red contour map. The blue contour observed near position 10 shows the region favorable for electropositive groups, which corresponds to interaction with the $-\text{NH}$ group of Lys162.

Group-IV

The protomol bloat and threshold values were 0 and 0.43, respectively. Figure 11 depicts the interacting model of compound **119** with the kinase. Four H-bonds anchor the ligand into the binding site of Aurora-A. The thiazole ring nitrogen at position 2 acts as an acceptor to form an H-bond with the backbone $-\text{NH}$ of Ala213 ($-\text{N}\cdots\text{HN}$, 2.00 Å, 162.6°). The N atom at position 3 forms another H-bond with the backbone of Ala213 ($-\text{NH}\cdots\text{O}$, 1.87 Å, 142.1°). The carbonyl oxygen and $-\text{OH}$ atoms of the carboxyl group at the para-position of the phenyl ring form H-bonds with the guanidino group of Arg137 ($-\text{O}\cdots\text{HN}$, 1.97 Å, 154.9°) and the backbone O of Leu139 ($-\text{OH}\cdots\text{O}$, 2.47 Å, 62.5°), respectively. The substituent at position 1 can bind to a relatively shallow hydrophobic pocket formed by Val147, Ala160, Lys162 and Leu210 residues, which is in agreement with the CoMSIA small green contour present at this position. The yellow contour observed near position 6 indicates a sterically unfavorable region at this position. This is confirmed by docking results showing that bulkier groups at position 6 can lead to a steric clash with the side

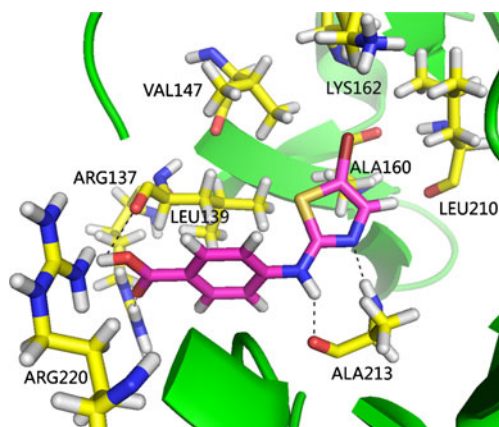


Fig. 11 Docked conformation derived for compound **119** with the binding site of Aurora-A kinase. H-bonds are shown as *dotted black lines*. Active site amino acid residues and the inhibitor are represented as *stick model*

chain of residue Arg137. The presence of a blue contour near position 5 suggests the requirement of electropositive groups at this site and expects to have a favorable interaction with electronegative groups like the carbonyl backbone of Leu139. The magenta contour seen at position 4 indicates an H-bond acceptor favored region as verified by the H-bond donor group of $-\text{NH}_2$ of Arg220 presented herein. The carbonyl group of Leu139 located at position 5 suggests the importance of H-bond donor groups at this position, which is also supported by the presence of an H-bond donor favorable red contour (CoMSIA model).

Group-V

The protomol bloat and threshold values were 0 and 0.43, respectively, and compound **176** is depicted in Fig. 12. There are four H-bonds and one water-mediated interaction between the inhibitor and binding site residues. The carbonyl oxygen at position 1 forms an H-bond with the backbone of Lys141 ($-\text{O}\cdots\text{HN}$, 2.01 Å, 162.6°). The pyrazole ring $-\text{N}-$ and $-\text{NH}$ atoms at positions 5 and 6 form H-bonds with the backbone atoms of Ala213 ($-\text{N}\cdots\text{HN}$, 2.16 Å, 155.9° and $-\text{NH}\cdots\text{O}$, 1.60 Å, 144.4°), respectively. The O atom at position 8 is located within H-bonding distance (3.45 Å) of the backbone of Thr217. Interaction between the nitrogen atom at position 2 and the side chain hydroxyl group of Thr217 is glued by a structural water molecule (w25, 3FDN.pdb). Docking results show that space around the substituent at position 1 is relatively large, and that this moiety seems to be exposed to the solvent, which is in line with the sterically favorable green contour presented herein. A large yellow contour at position 3 suggests a preference for small groups at this position, which is also validated by the docked

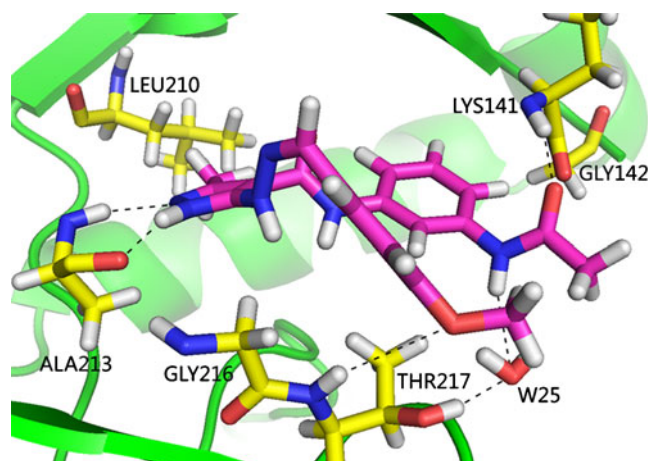


Fig. 12 Docked conformation derived for compound **176** with the binding site of Aurora-A kinase. H-bonds are shown as *dotted black lines*. Active site amino acid residues and the inhibitor are represented as *stick model*. W25 represents water molecule

structure showing that substitution with bulky groups will have an unfavorable steric clash with the backbone atoms of Lys141 and Gly142. The electronegative favorable red contour near position 1 corresponds to the –NH– backbone of Lys141, which explains the increased activity of compounds with electronegative groups in this region. A small cyan contour near position 4 suggests H-bond donor groups are favored at this position as confirmed by the –NH group of Thr217 located nearby. The purple contour observed near position 7 indicates an H-bond acceptor favorable region, which is further supported by the presence of a backbone –NH group of Gly216 in this location.

Group-VI

The protomol bloat and threshold were set to 1 and 0.43, respectively, and compound **220** is displayed in Fig. 13. A total of three H-bonds and two water-mediated interactions exist between the ligand and the active site of Aurora-A kinase. The F atom at position 1 forms a H-bond with the side chain –NH of Lys162 (–F⋯HN, 2.35 Å, 136.1°). The pyrimidine ring nitrogen atoms at positions 2 and 3 form H-bonds with the side chain of Lys143 (–N⋯HN, 2.49 Å, 117.0° and –N⋯HN, 2.47 Å, 102.3°), respectively. The carbonyl oxygen at position 4 and the backbone –NH of Thr217 is linked by a water-mediated H-bond bridge (w455, 2NP8.pdb). Another water-mediated interaction is formed between the F atom at position 1 and the backbone carbonyl group of Ala273 through the water molecule w489. The side chain of Lys162 and Asp274 located near position 1 indicates that analogues with bulky substituents at the 1 position of the pyrimidine ring would have an unfavorable steric interaction. This is in accordance with

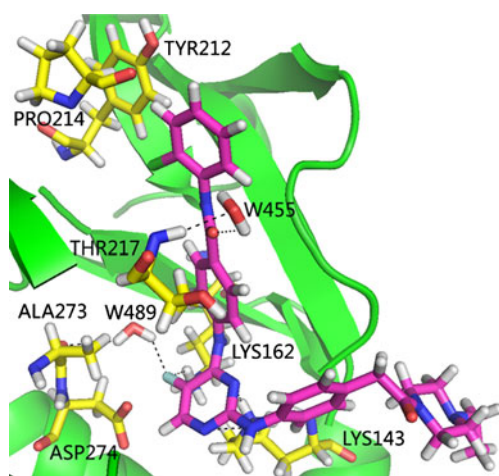


Fig. 13 Docked conformation derived for compound **220** with the binding site of Aurora-A kinase. H-bonds are shown as *dotted black lines*. Active site amino acid residues and the inhibitor are represented as stick model. W455 and W489 represent water molecules, respectively

the 3D contour maps showing that a CoMSIA sterically unfavorable yellow polyhedron is observed at this position. The green contour seen near position 6 suggests a sterically favorable region, as corroborated by the docking structure showing that this space is relatively large, and extends even outside to the solvent. The red contours near positions 1 and 4 indicate that compounds with electronegative groups at these positions may have a favorable interaction with surrounding residues as shown by the side chain –NH of Lys162 and backbone –NH of Thr217 present nearby. The side chain hydroxyl group of Tyr212 and the backbone carbonyl group of Pro214 near position 8 suggests that hydrophilic groups are favored in this region, which is in line with the presence of a CoMSIA hydrophilic favorable white contour. A small white contour along with a small green contour is observed near position 9, suggesting that hydrophilic and bulky substituents are both favorable at this position, as demonstrated by the docking model showing that this moiety is located in the lower lobe of the solvent-exposed binding area.

In order to explore the similarities and differences in binding modes, the six docked complexes were superimposed together using PYMOL software (www.pymol.org). The common big ligand binding pocket was found to be constructed by 34 residues: Arg137, Pro138, Leu139, Gly140, Lys141, Gly142, Lys143, Val147, Ala160, Lys162, Leu164, Leu178, Glu181, Val182, Gln185, Leu194, Leu196, Leu208, Leu210, Glu211, Tyr212, Ala213, Pro214, Leu215, Gly216, Thr217, Tyr219, Arg220, Glu260, Asn261, Leu263, Ala273, Asp274 and Phe275 (residue numbering according to 3E5A.pdb).

Residues Arg137, Lys141, Lys143, Lys162, Glu181, Gln185, Glu211, Ala213, Thr217, Arg220, Ala273, Asp274 and Phe275 produced mainly H-bonds with the ligand, and amino acids Arg137, Lys141, Gly142, Val147, Ala160, Lys162, Leu178, Val182, Leu210, Thr217, Arg220 and Asp274 formed steric interactions or hydrophobic interactions (Val147, Ala160, Lys162, Leu210, Glu211, Tyr212, Ala213, Pro214 and Arg220) with inhibitors. Interestingly, Phe144 belongs only to G-I (3E5A.pdb) and cannot be found in other binding models, indicating that this residue might be more specific for G-I derivatives. It was also found that residue Ala213 (Ala212 in 2C6E) formed important H-bonding interactions in the five models G-I to G-V, but not in G-VI, suggesting that this residue plays a critical role in the recognition of Aurora-A by inhibitors.

We conclude that the results obtained from molecular docking and those from 3D-QSAR modeling can complement and validate each other, suggesting that the 3D-QSAR models generated in the present study are reasonable and could be utilized to derive useful information in the design of novel Aurora-A inhibitors.

Conclusions

3D-QSAR studies using CoMFA and CoMSIA techniques were performed on six different chemical series of Aurora-A inhibitors. These studies yielded stable and statistically significant predictive models indicated by good performance with both internal and external validations. The 3D contour maps obtained from the optimal QSAR models in each group correlated well with the structural and functional features of the active binding sites identified from docking studies. One common residue, i.e., Ala213 (Ala212 in G-II), was found in the kinase active site that played a significant role in recognition of the inhibitors by presenting H-bonding interactions in five groups (not in G-VI). Other notable findings are listed in detail for each individual group as follows:

Group-I

Hydrophobic interaction was found to govern the inhibitory activity of group I compounds by making the highest contribution of 46.4% in the optimal CoMSIA model. At position 4, a linker consisting of hetero atoms such as O and S between quinazoline and aromatic ring can enhance kinase activity. In addition, electronegative and hydrophilic substituents at position 8 can also improve the Aurora-A inhibitory activity of a compound.

Group-II

Electrostatic interaction is more important in G-II molecules, showing a higher contribution of 54.6% in the best CoMFA model. A bulky substituent at position 6, and small and electronegative substituents at positions 8 and 10 would improve biological activity.

Group-III

Electrostatic field contributes more than steric field (56.9% and 43.1%, respectively) in the best CoMSIA model, suggesting electrostatic interactions are more critical to G-III compounds. Substitution with small and electropositive groups at position 7, and relatively large and electropositive groups at position 10 might increase compound activity.

Group-IV

The H-bond acceptor field exhibits a prominent contribution of 52.9% in the optimal CoMSIA model, which indicates the importance of H-bonding interactions to this kind of molecule. Bulky and H-bond donor substituents at position 5, and H-bond acceptor group at positions 4 and 6 would have a positive effect on bioactivity.

Group-V

Electrostatic interactions were found to have a determinant effect on inhibitory potency by making a contribution of 58.6% in the best CoMSIA model. Bulky, electronegative and H-bond acceptor substituents at position 1, and electropositive and H-bond donor substituents at position 4 are favorable for biological activity.

Group-VI

The hybrid effect of electrostatic and hydrophobic interactions is more crucial to the inhibitory activity of G-VI compounds. Substitution with small and electronegative groups at position 1, and bulky and hydrophilic groups at positions 8 and 9 may lead to an increase in compound activity.

To the best of our knowledge, this is the first study aimed at deriving predictive 3D-QSAR models for A-type Aurora kinase inhibitors. Moreover, the docking studies provided good insights into inhibitor–protein interactions at the molecular level. The good correlation between experimental and predicted pK_i or pIC_{50} values for test set compounds further indicated the robustness of the 3D-QSAR models. Thus, the derived models can be utilized in predicting the affinity of related analogues, guiding future structural modifications and synthesizing novel potent Aurora-A inhibitors.

Acknowledgments This work is supported financially by the National Natural Science Foundation of China (Grant No. 10801025) and the Fund of Northwest A&F University. We thank Dr. Ming Hao for helping with PCA analysis.

References

1. Fu J, Bian M, Jiang Q, Zhang C (2007) *Mol Cancer Res* 5:1–10
2. Keen N, Taylor S (2004) *Nat Rev Cancer* 4:927–936
3. Andrews PD (2005) *Oncogene* 24:5005–5015
4. Jackson JR, Patrick DR, Dar MM, Huang PS (2007) *Nat Rev Cancer* 7:107–117
5. Giet R, Petretti C, Prigent C (2005) *Trends Cell Biol* 15:241–250
6. Glover DM, Leibowitz MH, McLean DA, Parry H (1995) *Cell* 81:95–105
7. Bischoff JR, Anderson L, Zhu Y, Mossie K, Ng L, Souza B, Schryver B, Flanagan P, Clairvoyant F, Ginther C, Chan CSM, Novotny M, Salomon DJ, Plowman GD (1998) *EMBO J* 17:3052–3065
8. Pollard JR, Mortimore M (2009) *J Med Chem* 52:2629–2651
9. Andersen CB, Wan Y, Chang JW, Riggs B, Lee C, Liu Y, Sessa F, Villa F, Kwiatkowski N, Suzuki M, Nallan L, Heald R, Musacchio A, Gray NS (2008) *ACS Chem Biol* 3:180–192
10. Zhou H, Kuang J, Zhong L, Kuo WL, Gray JW, Sahin A, Brinkley BR, Sen S (1998) *Nat Genet* 20:189–193
11. Dutertre S, Descamps S, Prigent C (2002) *Oncogene* 21:6175–6183

12. Barr AR, Gergely F (2007) *J Cell Sci* 120:2987–2996
13. Giet R, Uzbekov R, Cubizolles F, Le Guellec K, Prigent C (1999) *J Biol Chem* 274:15005–15013
14. Bolanos-Garcia VM (2005) *Int J Biochem Cell Biol* 37:1572–1577
15. Liu Q, Kaneko S, Yang L, Feldman RI, Nicosia SV, Chen J, Cheng JQ (2004) *J Biol Chem* 279:52175–52182
16. Nishida N, Nagasaka T, Kashiwagi K, Boland CR, Goel A (2007) *Cancer Biol Ther* 6:525–533
17. Gritsko TM, Coppola D, Paciga JE, Yang L, Sun M, Shelley SA, Fiorica JV, Nicosia SV, Cheng JQ (2003) *Clin Cancer Res* 9:1420–1426
18. Li DH, Zhu JJ, Firozi PF, Abbruzzese JL, Evans DB, Cleary K, Friess H, Sen S (2003) *Clin Cancer Res* 9:991–997
19. Reiter R, Gais P, Jutting U, Steuer-Vogt MK, Pickhard A, Bink K, Rauser S, Lassmann S, Höfler H, Werner M, Walch A (2006) *Clin Cancer Res* 12:5136–5141
20. Wang X, Zhou YX, Qiao W, Tominaga Y, Ouchi M, Ouchi T, Deng CX (2006) *Oncogene* 25:7148–7158
21. Ditchfield C, Johnson VL, Tighe A, Ellston R, Haworth C, Johnson T, Mortlock A, Keen N, Taylor SS (2003) *J Cell Biol* 161:267–280
22. Harrington EA, Bebbington D, Moore J, Rasmussen RK, Ajose-Adeogun AO, Nakayama T, Graham JA, Demur C, Hercend T, Diu-Hercend A, Su M, Golec JMC, Miller KM (2004) *Nat Med* 10:262–267
23. Hauf S, Cole RW, LaTerra S, Zimmer C, Schnapp G, Walter R, Heckel A, van Meel J, Rieder CL, Peters JM (2003) *J Cell Biol* 161:281–294
24. Mortlock AA, Foote KM, Heron NM, Jung FH, Pasquet G, Lohmann JJ, Warin N, Renaud F, Savi CD, Roberts NJ, Johnson T, Dousson CB, Hill GB, Perkins D, Hatter G, Wilkinson RW, Wedge SR, Heaton SP, Odedra R, Keen NJ, Crafter C, Brown E, Thompson K, Brightwell S, Khatri L, Brady MC, Kearney S, McKillop D, Rhead S, Parry T, Green S (2007) *J Med Chem* 50:2213–2224
25. Manfredi MG, Ecsedy JA, Meetze KA, Balani SK, Burenkova O, Chen W, Galvin KM, Hoar KM, Huck JJ, Leroy PJ, Ray ET, Sells TB, Stringer B, Stroud SG, Vos TJ, Weatherhead GS, Wysong DR, Zhang M, Bolen JB, Claiborne CF (2007) *Proc Natl Acad Sci USA* 104:4106–4111
26. Shimomura T, Hasako S, Nakatsuru Y, Mita T, Ichikawa K, Kodera T, Sakai T, Nambu T, Miyamoto M, Takahashi I, Miki S, Kawanishi N, Ohkubo M, Kotani H, Iwasawa Y (2010) *Mol Cancer Ther* 9:157–166
27. Bebbington D, Binch H, Charrier JD, Everitt S, Fraysse D, Golec J, Kay D, Knegt R, Mak C, Mazzei F, Miller A, Mortimore M, O'Donnell M, Patel S, Pierard F, Pinder J, Pollard J, Ramaya S, Robinson D, Rutherford A, Studley J, Westcott J (2009) *Bioorg Med Chem Lett* 19:3586–3592
28. Coumar MS, Leou JS, Shukla P, Wu JS, Dixit AK, Lin WH, Chang CY, Lien TW, Tan UK, Chen CH, Hsu JT, Chao YS, Wu SY, Hsieh HP (2009) *J Med Chem* 52:1050–1062
29. Anderson K, Yang J, Koretke K, Nurse K, Calamari A, Kirkpatrick RB, Patrick D, Silva D, Tummino PJ, Copeland RA, Lai Z (2007) *Biochemistry* 46:10287–10295
30. Cramer RD III, Patterson DE, Bunce JD (1988) *J Am Chem Soc* 110:5959–5967
31. Klebe G, Abraham U, Mietzner T (1994) *J Med Chem* 37:4130–4146
32. Wang YH, Li Y, Yang SL, Yang L (2005) *J Comput Aided Mol Des* 19:137–147
33. Li Y, Wang YH, Yang L, Zhang SW, Liu CH (2006) *Internet Electron J Mol Des* 5:1–12
34. Xu X, Yang W, Li Y, Wang YH (2010) *Expert Opin Drug Discov* 5:21–31
35. Li Y, Wang YH, Yang L, Zhang SW, Liu CH, Yang SL (2005) *J Mol Struct* 733:111–118
36. Ai CZ, Wang YH, Li Y, Li YH, Yang L (2008) *QSAR Comb Sci* 27:1183–1192
37. Aliagas-Martin I, Burdick D, Corson L, Dotson J, Drummond J, Fields C, Huang OW, Hunsaker T, Kleinheinz T, Krueger E, Liang J, Moffat J, Phillips G, Pulk R, Rawson TE, Ultsch M, Walker L, Wiesmann C, Zhang B, Zhu BY, Cochran AG (2009) *J Med Chem* 52:3300–3307
38. Howard S, Berdini V, Boulstridge JA, Carr MG, Cross DM, Curry J, Devine LA, Early TR, Fazal L, Gill AL, Heathcote M, Maman S, Matthews JE, McMenamin RL, Navarro EF, O'Brien MA, O'Reilly M, Rees DC, Reule M, Tisi D, Williams G, Vinkovi M, Wyatt PG (2009) *J Med Chem* 52:379–388
39. Rawson TE, Rütth M, Blackwood E, Burdick D, Corson L, Dotson J, Drummond J, Fields C, Georges GJ, Goller B, Halladay J, Hunsaker T, Kleinheinz T, Krell HW, Li J, Liang J, Limberg A, McNutt A, Moffat J, Phillips G, Ran Y, Safina B, Ultsch M, Walker L, Wiesmann C, Zhang B, Zhou A, Zhu BY, Ru"ger P, Cochran AG (2008) *J Med Chem* 51:4465–4475
40. Heron NM, Anderson M, Blowers DP, Breed J, Eden JM, Green S, Hill GB, Johnson T, Jung FH, McMiken HH, Mortlock AA, Pannifer AD, Pauptit RA, Pink J, Roberts NJ, Rowsell S (2006) *Bioorg Med Chem Lett* 16:1320–1323
41. Jung FH, Pasquet G, Lambert-van der Brempt C, Lohmann JJ, Warin N, Renaud F, Germain H, De Savi C, Roberts N, Johnson T, Dousson C, Hill GB, Mortlock AA, Heron N, Wilkinson RW, Wedge SR, Heaton SP, Odedra R, Keen NJ, Green S, Brown E, Thompson K, Brightwell S (2006) *J Med Chem* 49:955–970
42. Leonard JT, Roy K (2006) *QSAR Comb Sci* 25:235–251
43. Jain AN (2003) *J Med Chem* 46:499–511
44. Zhao B, Smallwood A, Yang J, Koretke K, Nurse K, Calamari A, Kirkpatrick RB, Lai Z (2008) *Protein Sci* 17:1791–1797
45. Tari LW, Hoffman ID, Bensen DC, Hunter MJ, Nix J, Nelson KJ, McRee DE, Swanson RV (2007) *Bioorg Med Chem Lett* 17:688–691
46. Cheetham GM, Knegt RM, Coll JT, Renwick SB, Swenson L, Weber P, Lippke JA, Austen DA (2002) *J Biol Chem* 277:42419–42422
47. Roy PP, Roy K (2008) *QSAR Comb Sci* 27:302–313

Transformation of the θ -phase in Mg-Li-Al alloys: a density functional theory study

Caili Zhang · Peide Han · Zhuxia Zhang ·
Minghui Dong · Lili Zhang · Xiangyang Gu ·
Yanqing Yang · Bingshe Xu

Received: 9 November 2010 / Accepted: 18 March 2011 / Published online: 15 June 2011
© Springer-Verlag 2011

Abstract In Mg-Li-Al alloys, θ -phase MgAlLi_2 is a strengthening and metastable phase which is liable to be transformed to the equilibrium phase AlLi on overaging. While the structural details of the θ -phase MgAlLi_2 and the microscopic transformation are still unknown. In this paper, the structure of MgAlLi_2 unit cell was determined through X-ray powder diffraction simulation. Microscopic transformation process of θ -phase MgAlLi_2 was discussed in detail using first principles method.

Keywords Alloy · Density functional theory · Magnesium · Transformation

Introduction

As the lightest engineering alloys, Mg-Li alloys have attracted increasing interests in transportation industries as they have many advantages, such as high specific strength,

good formability, good damping ability and high energetic particle penetration resistance [1–3]. The most common and typical Mg-Li based alloys are Mg-Li-Al based alloys such as LA141, MA18 and MA21 in which the Li content is always larger than 8 wt% [4]. The stability of these Mg-Li-Al alloys is relatively poor since Li is a very active element. In Mg-Li-Al alloys, θ -phase MgAlLi_2 is a strengthening and metastable phase which is liable to be transformed to the equilibrium phase AlLi on overaging [4–7]. While the structural details of the θ -phase MgAlLi_2 and the microscopic transformation are still unknown, so, we attempt to reveal both of them in this paper.

We organize the paper as follows. Section 2 elaborates the computational details of our first-principles calculations. It is followed by a section presenting our determination of the structure of θ -phase MgAlLi_2 through X-ray powder diffraction simulation. In Sect. 4, we further investigate the microscopic transformation process. A brief summary and statement of conclusions are presented in Sect. 5.

Methodology

Cambridge serial total energy package (CASTEP) [8], first-principles pseudopotential plane-wave method based on density functional theory (DFT) is used in the present study. In the process of solving the Schrödinger equation, Ultrasoft pseudopotentials in reciprocal space represented by a generalized gradient approximation as described by Perdew, Burke and Ernzhofner (GGA-PBE) [9] was adopted for all elements in our models to describe the exchange-correction energy and the computationally expensive electron-ion interaction, respectively. The values of kinetic energy cutoff E_{cut} and the k-points number are increased until the calculated energy converges within the required

C. Zhang · P. Han · Z. Zhang · M. Dong · L. Zhang · Y. Yang ·
B. Xu (✉)

Key laboratory of Interface Science and Engineering in Advanced
Materials, Ministry of Education,
Taiyuan University of Technology,
Taiyuan 030024, People's Republic of China
e-mail: xubs@tyut.edu.cn

C. Zhang · P. Han · Z. Zhang · M. Dong · L. Zhang · Y. Yang ·
B. Xu

College of Materials Science and Engineering,
Taiyuan University of Technology,
Taiyuan 030024, People's Republic of China

X. Gu

College of Mechanical Engineering,
Taiyuan University of Technology,
Taiyuan 030024, People's Republic of China

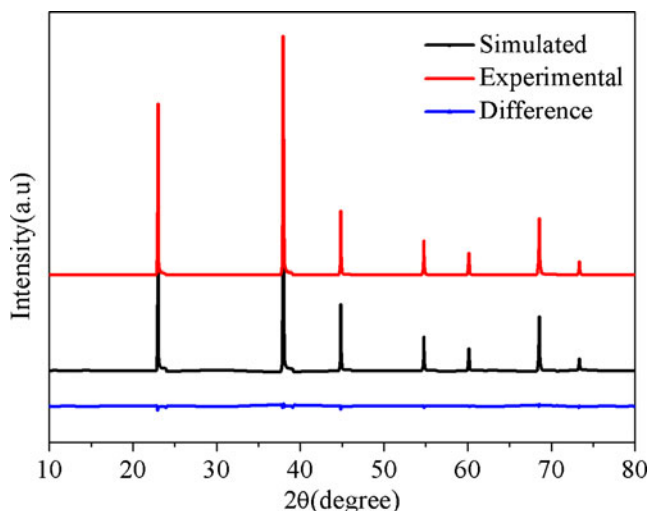


Fig. 1 Comparison between experimental [10] and calculated X-ray powder diffraction patterns for MgAlLi_2 unit cell. The difference between the experimental and calculated patterns is shown below the data

tolerance, where E_{cut} determines the number of plane waves and k points determines the sampling of the irreducible wedge of the Brillouin zone. E_{cut} is set at 380 eV, the k -point meshes for Brillouin zone sampling are constructed using the Monkhorst-Pack scheme and a $4 \times 4 \times 4$ k -point mesh are used for all cells, which are found to be sufficient to give fully converged results. The calculation of elastic constant and electronic structure is followed by cell optimization with the convergence tolerance of energy of

5.0×10^{-6} eV/atom, maximum displacement of 5.0×10^{-4} Å, maximum force of 0.01 eV/Å.

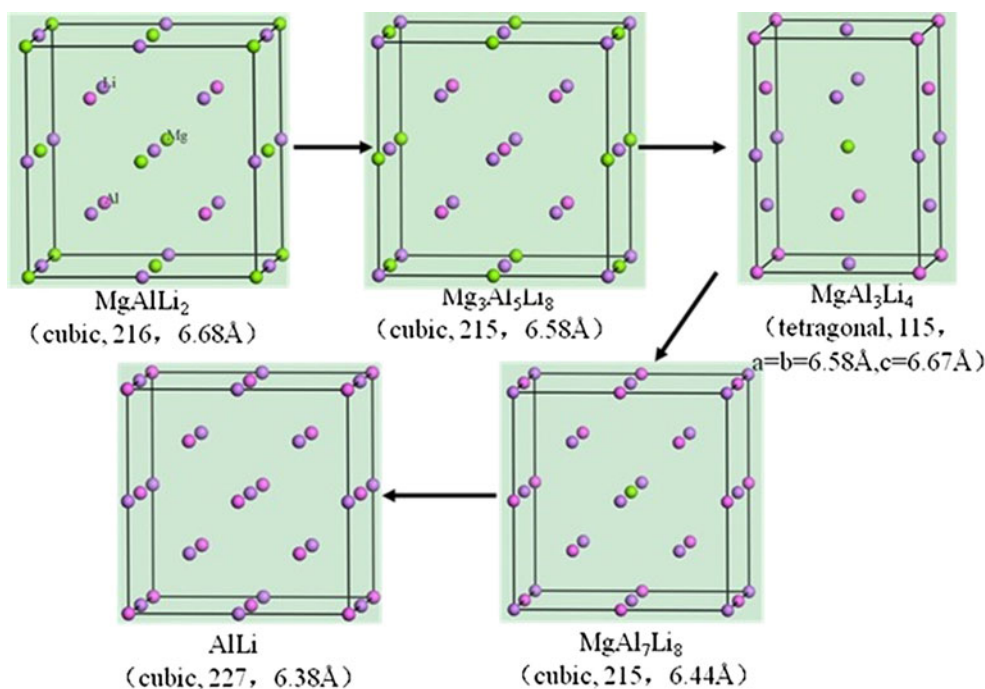
Crystal structure of the θ -phase

Up to now, there have been no published structural details of the metastable θ -phase. In this work, we attempt to reveal the characteristics of the θ -phase using X-ray diffraction simulation.

Experiments indicate the θ -phase is MgAlLi_2 [4–7], and the initial structure of MgAlLi_2 was taken from the data reported previously by Levinson et al. [10]. According to it, the X-ray powder diffraction pattern of MgAlLi_2 is known and its unit cell is a cubic structure with the lattice parameter of 6.7 Å, while, the atomic position is unknown. The atomic position was guessed by replacing the same position in the structure of ABC_2 as Al_2CO_2 , Be_2CoSi , Li_2PdSb and so on. On the basis of repeated comparison of the known X-ray powder diffraction pattern of MgAlLi_2 and the simulating X-ray powder diffraction patterns of these structures after performing an energy minimization calculation, the final structure of MgAlLi_2 unit cell was determined according to the following procedure.

A combination of Pawley [11] and Rietveld [12] refinement methods was used for the optimum structure of the MgAlLi_2 unit cell. To refine the structure, it was necessary to obtain a first approximation of the profile

Fig. 2 Structural details including structures, space group numbers and lattice parameters for the five intermetallic compounds during phase transformation



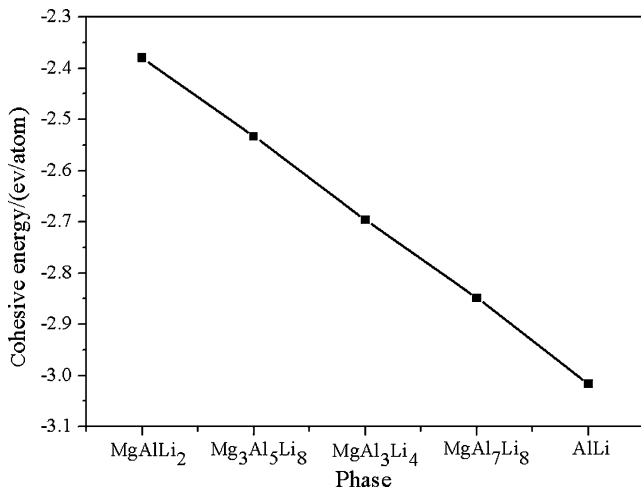


Fig. 3 Cohesive energies of the five intermetallic phases

parameters, such as peak profiles, background parameter, line shift parameters, lattice parameters, asymmetry parameters, and a good starting model of the atomic positions. The Pawley refinement was applied to the calculation of these values without details of the atomic position beforehand, for it can refine the intensities of Bragg reflections as freely varying refinement parameters, irrespective of the atomic structure. It was applied repetitively, with only a few parameters being refined at a time, up to obtaining a best-valued refinement, which gets an optimum agreement between simulated and experimental data. Once the initial guesses of the profile parameters were obtained, Rietveld method was applied to refine the structure of the MgAlLi₂ unit cell. Profile and lattice parameters obtained from the Pawley refinement were used for calculations in the Rietveld method. The calculated powder pattern which was compared to the observed one is shown in Fig. 1. It provides a measure of similarity by means of the weighted profile factor (Rwp). The optimum structure was selected based on the achieved minimum Rwp 5.00%.

The procedure described above gave information about the structure of the θ-phase. Its space group number is 216, and the lattice parameter is 6.7 Å. The atomic positions are as follows, Mg (0 0 0), Li (0.5 0.5 0.5) (0.75 0.75 0.75) and Al (0.25 0.25 0.25). In order to get a clear

understanding of phase transition, we use the above structure of MgAlLi₂ as the initial structure to do further research that is described in section 4

Microscopic transformation process of the θ-phase

Change in structure

The structural details and stabilities of complex intermetallic phases produced during transformation process from the metastable θ-phase to stable AlLi will be discussed in this section. These structures have been calculated from MgAlLi₂ derived through the refinement method by replacing the Mg atoms with some Al atoms, and further minimizing the energy of the structure. The structural transformation is as follows, MgAlLi₂ → Mg₃Al₅Li₈ → MgAl₃Li₄ → MgAl₇Li₈ → AlLi. The structures, the space group numbers and the lattice parameters are all shown in Fig. 2.

It is noteworthy that the cohesive energy of a material is a fundamental property which is descriptive in studying the phase stabilities of different structures of the same material [13]. Cohesive energy is often defined as the work which is needed when crystal is decomposed into the single atom. Hence, when stabilities of different structures of the same material are compared, the smaller the absolute value of cohesive energy is, the more unstable the crystal structure is. In this work, cohesive energies of per atom (*E*) for the five crystal cells were calculated by using the following expression:

$$E = \frac{1}{x + y + z} (E_{tot} - xE_{atom}^{Mg} - yE_{atom}^{Al} - zE_{atom}^{Li}) \quad (1)$$

where *E* refers to the total energy of crystal used in the present calculation, *x*, *y* and *z* refer to the numbers of Mg, Al and Li atoms, respectively. *E_{atom}^{Mg}*, *E_{atom}^{Al}* and *E_{atom}^{Li}* are the energies for isolated Mg atom, Al atom and Li atom. The energies of isolated Mg, Al and Li atoms are -977.0508 eV, -52.6554 eV and -188.4079 eV, respectively. Cohesive energies of per atom of all crystals are calculated from Eq. (1), and the results were illustrated visually in Fig. 3. It can be concluded that the values of cohesive energy decrease gradually from MgAlLi₂ to AlLi, that is to say, the stability

Table 1 The data of the elastic constants (GPa), bulk modulus B₀ (GPa), shear modulus G (GPa) and Young’s modulus E (GPa)

Phase	C ₁₁	C ₁₂	C ₄₄	C ₁₃	C ₃₃	C ₆₆	B ₀	G	E
MgAlLi ₂	44.32	24.61	38.95	—	—	—	31.18	27.31	63.42
Mg ₃ Al ₅ Li ₈	58.73	24.48	32.40	—	—	—	35.90	26.29	63.40
MgAl ₃ Li ₄	44.32	24.61	45.07	24.61	62.57	42.48	37.32	31.11	73.04
MgAl ₇ Li ₈	50.10	38.15	58.03	—	—	—	42.13	37.21	86.24
AlLi	58.56	38.48	39.41	—	—	—	45.17	27.66	68.92

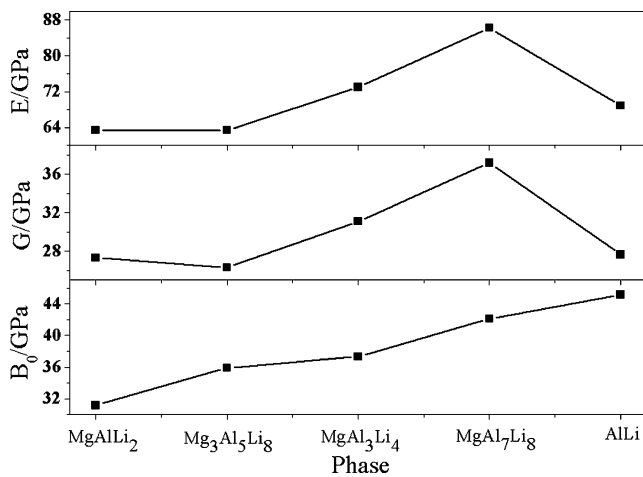


Fig. 4 The curves of Young's modulus E , shear modulus G and bulk modulus B_0 of the five intermetallic phases

of the five intermetallic phases improves gradually, which is also in good accordance with experiment.

Change in mechanical properties

To get a clear understanding of the change of mechanical properties during the phase transformation, we calculated the elastic constants C_{ij} , bulk modulus B_0 , shear modulus G and Young's modulus E of the five intermetallic phases. C_{ij} of solids enclose a great deal of the important information on their mechanical and dynamical properties. Crystal structure cannot exist in a stable phase unless its elastic constants obey certain relationships. They also determine the response of a crystal under external strain and provide key information of the strength of the material, as characterized by the B_0 , G and E . Hence, they play an

important role in the estimation of the material's stiffness and can be used to check the phase stability of proposed compounds.

The macroscopically measurable quantities obtained for materials are G that represents the isotropic response for shearing, E corresponding to the stress–strain ratio in the case of tensile forces, B_0 which is important for technological and engineering applications. For a cubic material, there are three independent elastic constants: C_{11} , C_{12} and C_{44} , and for a tetragonal material, five elastic constants: C_{11} , C_{12} , C_{33} , C_{44} and C_{66} are all independent. These macroscopic parameters are related to the microscopic elastic constants by means of the following Eqs. 2, 3, 6 for a cubic material and Eqs. 4, 5, 6 for a tetragonal material:

$$B_0 = (C_{11} + 2C_{12})/3 \quad (2)$$

$$G = (3C_{44} + C_{11} - C_{12})/5 \quad (3)$$

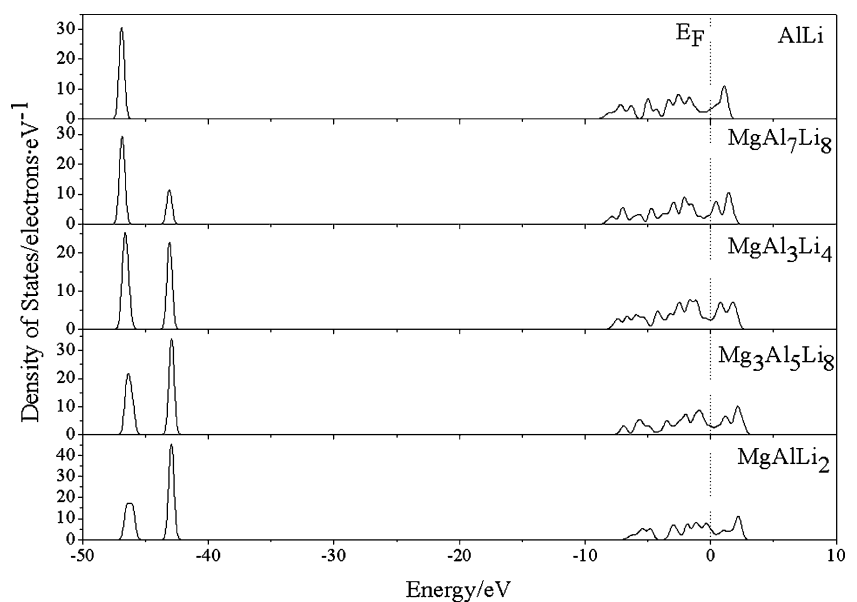
$$B_0 = (2C_{11} + C_{33} + 6C_{12})/9 \quad (4)$$

$$G = (2C_{11} + C_{33})/15 + (2C_{44} + C_{66} - C_{12})/5 \quad (5)$$

$$E = 9B_0G/(3B_0 + G). \quad (6)$$

The values of C_{ij} , B_0 , G and E are given in Table 1. The mechanical stability of a crystal, which means that the strain energy must be positive, is checked by the whole set of elastic constants C_{ij} which satisfies special restrictions. To remain mechanically stable, the following restrictions on

Fig. 5 Total density of states of the five intermetallic phases



the elastic constants must be required, Eq. 7 and 8 for a cubic system and tetragonal system, respectively.

$$(C_{11}-C_{12}) > 0, C_{11} > 0, C_{44} > 0, (C_{11} + 2C_{12}) > 0 \quad (7)$$

$$C_{11} > |C_{12}|, (C_{11}+C_{12}) C_{33} > 2C_{13}^2, C_{44} > 0, C_{66} > 0 \quad (8)$$

As we can see in Table 1, the whole set of C_{ij} calculated for the five intermetallic phase all obey well the above conditions, which is an important requisite for materials' stability, indicating that all of them are mechanically stable.

It is acknowledged that B_0 is a measure of resistance to volume change by applied pressure, that is, materials with high bulk modulus are likely to be hard materials. The values of B_0 increase gradually from $MgAlLi_2$ to $AlLi$. Young's modulus E is defined as the ratio between stress and strain, and is used to provide a measure of the stiffness of the solid, i.e., the larger the value of E , the stiffer the material. Shear modulus G is a measure of resistance to reversible deformations upon shear stress. From Fig. 4, we can see that the values of E and G have no significant change from $MgAlLi_2$ to $Mg_3Al_5Li_8$, and the values of E and G increase obviously from $Mg_3Al_5Li_8$ to $MgAl_7Li_8$, while, the values decrease sharply from $MgAl_7Li_8$ to $AlLi$. Thus, the values of B , E and G are not key factor for the experimental observation that the metastable and strengthen phase $MgAlLi_2$ is liable to transfer to the softening phase $AlLi$.

Change in electronic properties

Based on the above discussion, we did further research on the change in electronic properties which is reflected by total density of states (DOS). The results are shown in Fig. 5. The DOS curves of the five intermetallic phases mainly consist of three parts: the two peaks which are from -47 to -45 eV and from -44 to -42 eV in the low energy interval are mainly due to the electrons of Al and Mg, respectively. The DOS curve above the Fermi level is due to the electrons of Li. The peak value at -46 eV increases gradually with the Al content increasing from $MgAlLi_2$ to $AlLi$, at the same time, the peak value at -43 eV decreases gradually with the Mg content decreasing from $MgAlLi_2$ to $AlLi$. That is to say, the chemical bonding increases gradually at lower energy level with the Al content increasing from $MgAlLi_2$ to $AlLi$, correspondingly, the structural stability increases gradually. Thus, the trends in the cohesive energy can be understood in terms of different

chemical bonding below Fermi level for the five intermetallic phases.

Conclusions

The phase transformation of the metastable θ -phase $MgAlLi_2$ in Mg-Li-Al alloys was discussed in detail from microscopic perspective in this paper. The structural details of $MgAlLi_2$ were determined through X-ray powder diffraction simulation. Its space group number is 216 with the cubic structure and the lattice parameter is 6.7 \AA . The atomic positions are as follows, Mg (0 0 0), Li (0.5 0.5 0.5) (0.75 0.75 0.75) and Al (0.25 0.25 0.25). Three intermediate phases $Mg_3Al_5Li_8$, $MgAl_3Li_4$ and $MgAl_7Li_8$ exist during transformation from the metastable $MgAlLi_2$ to $AlLi$. The stability of the five intermetallic phases improves gradually and some changes on mechanical properties exist during transformation, which are in good accordance with experiment. In the end, the densities of states (DOS) of the five intermetallic phases were calculated and analyzed.

Acknowledgments This research was supported by the National Natural Science Foundation of China (Grant No.50874079, 51002102), the Natural Science Foundation of Shanxi Province (Grant No. 2009021026), Education Department of Shanxi Province (Grant No. 20080010), Taiyuan Science and Technology Project (Grant No 100115105).

References

- Counts WA, Friák M, Raabe D, Neugebauer J (2009) Acta Mater 57:69–76
- Yang CW, Lui TS, Chen LH, Hung HE (2009) Scr Mater 61:1141–1144
- Kim WJ (2009) Scr Mater 61:652–655
- Wu RZ, Zhang ML (2009) Mater Sci Eng A 520:36–39
- Wu HU, Lin JY, Gao ZW, Chen HW (2009) Mater Sci Eng A 523:7–12
- Wu HU, Gao ZW, Lin JY, Chiu CH (2009) J Alloys Compd 474:158–163
- Song GS, Staiger M, Kral M (2004) Mater Sci Eng A 371:371–376
- Clark SJ, Segall MD, Pickard CJ, Hasnip PJ, Probert MJ, Refson K, Payne MC (2005) Z Kristallogr 220:567–570
- Perdew JP, Burke K, Ernzerhof M (1996) Phys Rev Lett 77:3865–3868
- Levinson DW, Mcpherson DJ (1956) Trans Am Soc Met 48:689–696
- Pawley GS (1981) J Appl Cryst 14:357–361
- Rietveld HM (1969) J Appl Cryst 2:65–71
- Zhang CL, Han PD, Yan X, Wang C, Xia LQ, Xu BS (2009) J Phys D Appl Phys 42:125403

Controlling the aggregation and rate of release in order to improve insulin formulation: molecular dynamics study of full-length insulin amyloid oligomer models

Workalemahu Mikre Berhanu · Artëm E. Masunov

Received: 13 January 2011 / Accepted: 9 May 2011 / Published online: 15 June 2011
© Springer-Verlag 2011

Abstract Insulin is a hormone that regulates the physiological glucose level in human blood. Insulin injections are used to treat diabetic patients. The amyloid aggregation of insulin may cause problems during the production, storage, and delivery of insulin formulations. Several modifications to the C-terminus of the B chain have been suggested in order to improve the insulin formulation. The central fragments of the A and B chains (LYQLENY and LVEALYL) have recently been identified as β -sheet-forming regions, and their microcrystalline structures have been used to build a high-resolution amyloid fibril model of insulin. Here we report on a molecular dynamics (MD) study of single-layer oligomers of the full-length insulin which aimed to identify the structural elements that are important for amyloid stability, and to suggest single glycine mutants in the β -sheet region that may improve the formulation. Structural stability, aggregation behavior and the thermodynamics of association were studied for the wild-type and mutant aggregates. A comparison of the oligomers of different sizes revealed that adding strands enhances the internal stability of the wild-type aggregates. We call this “dynamic cooperativity”. The secondary structure content and clustering analysis of the MD trajectories show that the largest aggregates retain the fibril conformation, while the monomers

and dimers lose their conformations. The degree of structural similarity between the oligomers in the simulation and the fibril conformation is proposed as a possible explanation for the experimentally observed shortening of the nucleation lag phase of insulin with oligomer seeding. Decomposing the free energy into electrostatic, van der Waals and solvation components demonstrated that electrostatic interactions contribute unfavorably to the binding, while the van der Waals and especially solvation effects are favorable for it. A per-atom decomposition allowed us to identify the residues that contribute most to the binding free energy. Residues in the β -sheet regions of chains A and B were found to be the key residues as they provided the largest favorable contributions to single-layer association. The positive $\Delta\Delta G_{\text{mut}}$ values of 37.3 to 1.4 kcal mol⁻¹ of the mutants in the β -sheet region indicate that they have a lower tendency to aggregate than the wild type. The information obtained by identifying the parts of insulin molecules that are crucial to aggregate formation and stability can be used to design new analogs that can better control the blood glucose level. The results of our simulation may help in the rational design of new insulin analogs with a decreased propensity for self-association, thus avoiding injection amyloidosis. They may also be used to design new fast-acting and delayed-release insulin formulations.

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1123-3) contains supplementary material, which is available to authorized users.

W. M. Berhanu
NanoScience Technology Center and Department of Chemistry,
University of Central Florida,
Orlando, FL 32826, USA

A. E. Masunov (✉)
NanoScience Technology Center, Department of Chemistry,
and Department of Physics, University of Central Florida,
Orlando, FL 32826, USA
e-mail: amasunov@mail.ucf.edu

Keywords Amyloid fibril · Insulin · β -Sheet · Aggregation · Oligomer · Secondary structure · LYQLENY · LVEALYL · Molecular dynamics simulations · Cluster · MM-GBSA · Per-residue decomposition

Introduction

Insulin is a 51-residue protein hormone consisting of two polypeptide chains, chain A (comprising 21 residues) and

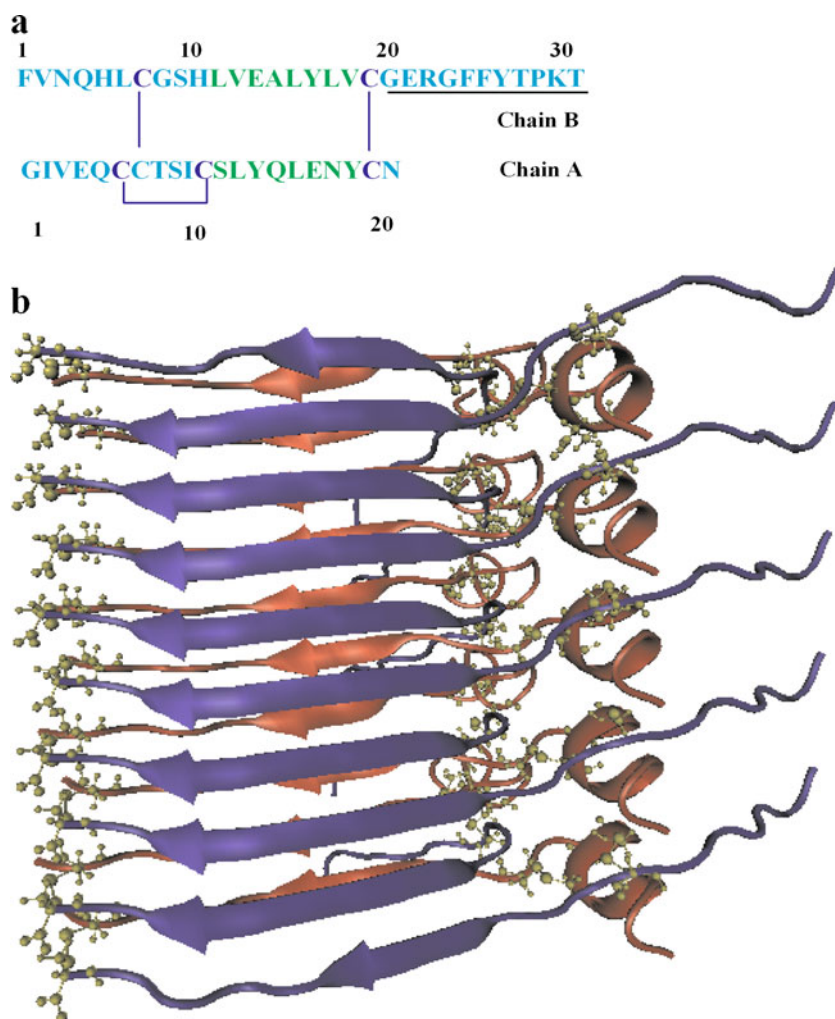
the chain B (comprising 30 residues), linked by three disulfide bonds (Fig. 1a). Deficiency in the secretion or action of insulin in response to the level of glucose may cause an abnormally high concentration of sugar, which can have a profound impact on metabolism in the human body [1, 2]. Insulin is stored in the body in the secretory vesicles of the pancreas as a zinc-containing hexamer. When in the bloodstream, insulin is present in its biologically active monomeric form [3, 4]. The underproduction of insulin or a lack of receptor sensitivity to insulin is known to cause diabetes, which affects 171 million people worldwide [5]. Insulin is the mainstay of drug therapy for patients with type I diabetes, and it can reduce morbidity in the long term. The disease is caused by the autoimmune destruction of insulin-secreting β cells of the pancreas. Without sufficient levels of insulin, these patients cannot properly utilize glucose, and they typically have markedly elevated blood glucose levels (hyperglycemia) while their intracellular glucose levels are generally low. The chronic complications of a consistently high blood sugar level are

serious and include retinopathy (diabetes is the most common cause of blindness), neuropathy, nephropathy (diabetes is a leading cause of chronic renal failure), cardiovascular disease, peripheral vascular disease (diabetes is the leading cause of limb amputation), and it makes the patient more susceptible to infection [6].

Similar to many other proteins, insulin can misfold and form highly ordered fibrillar amyloid aggregates. Insulin fibrils have been observed in vivo following continuous subcutaneous insulin infusion [7] and repeated insulin injections [8]; they are the main factor in the pathogenesis called injection amyloidosis [9, 10]. These insulin fibrils that form in vivo display the defining characteristics of amyloid aggregates associated with neurodegenerative diseases [11]: they bind to the dye Congo red with “apple-green” birefringence, they show an elongated, unbranched fibrillar morphology [10], they exhibit nucleation-dependent polymerization, and they present a cross- β X-ray diffraction pattern [9]. Recently, serum samples from patients with Parkinson’s disease have been

Fig. 1a–b Amino acid sequences and structural models of double-layer insulin oligomers.

a Amino acid sequence of insulin (chain B at the *top* and chain A at the *bottom*). Segments LVEAYLV of chain B and SLYQLENT of chain A are colored *green*. Disulfide bonds are colored *blue*. The C-terminal region of chain B (*underlined and italicized*) is not involved in amyloid fibrillization. The *underlined residues* are missing from the insulin model used in this study; only 40 amino acids are taken into account in the fibril model. **b** Single-layered structural models of insulin oligomers (ten-stranded). Two chains are associated via an interdigitated pair of LYQLENTY molecules of chain A and LVEALYL molecules of chain B, which interlock tightly to form the dry steric zipper. Chain A is *red* and chain B is shown in *blue*. Disulfide bonds are indicated in *yellow*



found to display an autoimmune response to insulin oligomers and fibrils [12], possibly indicating the presence of insulin aggregates in this disease too. Insulin also forms amyloid-like fibrils *in vitro*, a process that is promoted by elevated temperatures, low pH, and increased ionic strength [11, 13]. In addition, insulin fibril formation has been a limiting factor in the long-term storage of insulin to treat diabetes. Amyloid fibrillation may cause problems during the production, storage and delivery of protein-based pharmaceuticals [14, 15]. In the case of commercial insulin, fibril formation is a problem during some of the isolation and purification steps when the pH is lowered to 1–3 [11]. The agitation of insulin solutions during transportation and in portable delivery systems may also induce fibrillar aggregation [14–16]. Moreover, during the therapeutic use of protein drugs, it is essential to avoid fibril formation, since amyloid fibrillated protein is biologically inactive [11, 17] and may cause immunological responses in patients [17, 18]. Future drug development may aim to either stabilize native structure, inhibit the formation of crucial intermediates on the pathway to fibril formation, or prevent interactions between fibrillation intermediates such as the partially unfolded monomer and oligomers [15].

The tendency of insulin monomers to aggregate is also of fundamental importance to other physiologically relevant questions [19]. Recent experimental work by Maji et al. [20] showed that mammalian cells store a large quantity of the hormone in the form of amyloid fibrils in the secretory granules of cells until a signal triggers its release, at which point they can secrete hormones much faster than their rates of synthesis would permit. The amyloid aggregates have the properties required of a long-acting drug because they are stable depots that guarantee controlled release of the active peptide drug from the amyloid termini [21, 22]. This concept was tested by Maji et al. [21] with a family of short- and long-acting analogs of gonadotropin-releasing hormone, and it was shown that amyloids can act as a source for the sustained release of biologically active peptides. Modifications of the amino acid sequence of insulin, such as single point mutations, influence both its activity and protein aggregation [19]. The newer insulin analogs have several improvements due to their modified action profiles [23]. The main advantages of short-acting preparations include a faster onset of action and a shorter duration time. Long-acting analogs have structural changes that delay the onset of action, allow slow and continuous absorption into the systemic circulation, and prolong the duration, thus producing a time–concentration profile that imitates the normal insulin basal level and leads to physiological basal glycemic control with fewer nocturnal hypoglycemia [24].

Upon aggregation, the molecule of insulin undergoes structural changes from a predominantly α -helical state to a

β -sheet-rich conformation, and many models of insulin fibrils have been suggested [11, 15, 24]. The fibrillar β -sheets have been described as being either parallel [25–27] or antiparallel [28], and being flat [29], β -helical [30], or having β -roll-type structure [31]. Previous biophysical studies suggest that the B chain, or a segment of it, may be the primary determinant of insulin fibrillation. For example, equimolar amounts of the peptide RRRRLVEALYLV (containing residues B11–B17) can attenuate insulin fibrillation [32]. The segment B11–B17 with the sequence LVEALYL is the smallest peptide that can both nucleate and inhibit the fibrillation of full-length insulin, depending on the molar ratio. This activity suggests that this segment is central to the cross-beta spine of the insulin fibril [14]. In addition, the point mutations H10D and L17Q in chain B of insulin prolong the lag phase of insulin fibrillation, further supporting the idea that this segment is important in fibril formation [33]. Also, exposing this fibril-prone segment by truncating the five residues of the C-terminal of the B chain increases the propensity of insulin to form fibrils [34].

Recent studies have shown that the A chain also contributes to insulin fibrillation. Both the A and B chains can form fibrils on their own [35, 36], and seeds of these chains can nucleate the fibrillation of full-length insulin [35]. In addition, it was reported that segments as short as six residues from either chain A (residues A13–A18) or chain B (residues B12–B17) can form fibrils by themselves [37]. The same segments were found to be protected against hydrogen exchange when insulin was incubated under conditions favorable to fibril formation [38].

The first atomic-resolution view of the fibrillar spine came from single-crystal structures of the segments LYQLEN (residues A13–A18) and VEALYL (residues B12–B17) [28]. The combination of several complementary techniques (including X-ray fiber diffraction of insulin fibrils and scanning-transmission electron microscopy analysis of the morphology of insulin fibrils) allowed a highly reliable structure of full-length insulin amyloid fibrils to be constructed [14, 29–31]. This model has a β -solenoid structure consisting of repeated structural units of similar but not identical peptides that are covalently connected by two disulfide bonds [14, 31]. The solenoids are linked by a dry steric zipper formed by mating the two central LVEALYL (residues B11–B17) strands. Because LYQLENY contains a Tyr residue at the second position, this side chain is superimposed on a Tyr from LVEALYL, preserving the “kissing tyrosine” interaction observed across the wet interface of the crystal of LVEALYL (Fig. 1b).

Computational studies have been performed as complements to experimental studies in order to provide insights into insulin aggregation. All-atom molecular dynamics

(MD) simulations have been used to study amyloid oligomer stability by testing different candidate β -sheet arrangements of preformed oligomers that mimic possible nucleus seeds at the very early stage of fibril formation [39–41]. Mark et al. [42] performed a series of short molecular dynamics simulations to investigate the structures of monomeric insulin molecules and their dimers in aqueous solution. Their simulations showed that, both monomeric and dimeric insulin have high degrees of intrinsic flexibility in the absence of crystal contacts. Monomer MD simulations [43, 44] established that the proposed binding site for glucose is stable, both statically and dynamically [45]. Other MD simulations of the insulin dimer have also been published [46, 47]. They reveal details concerning the dynamics of the dimer during the simulation, including the hydrogen-bond pattern and correlated motions.

In this contribution, we report on an MD study of single-layer insulin aggregates based on the high-resolution model of insulin fibrils that aimed to elucidate the nature of insulin self-assembly. We present information on the energetics of the insulin association at the atomic level that could be used to design new short- and long-acting insulin analogs. Mutant forms of insulin with altered aggregation properties that could potentially be used in slow- or fast-acting therapeutic formulations are suggested on the basis of the observed contacts at the aggregate interface. There has been no previous systemic study of how mutation affects the stability of the insulin oligomer aggregates. Our MD simulation of different sizes of insulin oligomer may contribute to a better understanding of the nucleation process and conformational changes during the very early stages of fibril formation. This study aims to answer the following questions:

1. Which regions of the wild-type insulin oligomer aggregate are flexible?
2. How do the single point mutations influence the structures and flexibilities of these regions?
3. What are the effects of single glycine mutations of the side chains involved in the steric zipper?
4. What are the conformational differences among the aggregates of various sizes?

Computational details

We conducted a total of ~ 0.35 μ s of explicit-solvent molecular dynamics (MD) simulation of the insulin single-layer oligomer of wild type and mutated sequences with intact disulfide bridges, using a temperature of 330 K to emulate the experimental conditions of in vitro insulin fibrillization [48, 49].

System setup

In this study, we rely on the insulin fibrillar model constructed by Ivanova et al. [14] using the crystal structures of the LVEALYLV, SLYQLENY and fiber diffraction patterns. The C-terminal region of chain B (residues 20–30) is not involved in amyloid fibrillization [31], and was omitted. A comparison of the amino acid sequences of the insulin sequences from five different mammalian species (porcine, bovine, sheep, mouse and rat) for residues 20–30 shows that nine of the amino acids residues are conserved and that B30 Tyr in the human sequence is replaced with Ala in those of the other species [50]. These residues are missing from the insulin model used in this study. Therefore, only 40 amino acids are taken into account in the fibril model used here [14]. The starting coordinates (Fig. 1) for the MD simulations were taken from the web page <http://people.mbi.ucla.edu/sawaya/jmol/fibrilmodels>. An interesting feature of insulin is that its three disulfide bridges are retained in the in vitro and in vivo fibrillar forms [14]. Thus, these disulfide bonds must constrain the possible conformational rearrangements during the α -helix-to- β -sheet transition [14]. This conformational constraint makes insulin a unique model system for studying protein misfolding and subsequent amyloid fibrillization [14].

Molecular dynamics simulations

The molecular dynamics (MD) simulation was performed using the AMBER11 [51] package with an all-atom Amber99SB force field and explicit TIP3P water models. Each of the amyloid peptides and the corresponding mutants were solvated by explicit water molecules that extend 10 Å from any edge of the octahedral box to the protein atoms. Counterions were added to the box by randomly replacing water molecules in order to neutralize the system. The energy of each system was initially minimized using the conjugate gradient method in order to remove bad contacts. The peptide atoms were first constrained, and then relaxed without positional constraints. The system was then subjected to 50 ps of a gradual heating procedure while constraining the backbone atoms of the protein to allow the relaxation of water and ions, followed by a 1 ns equilibration run without positional constraints. Constant pressure (1 atm) and constant temperature (330 K) were maintained in the system by an isotropic Langevin barostat and a Langevin thermostat. The temperature 330 K was selected as a compromise which ensured that the amyloid fibrils are still experimentally stable [48, 49, 52, 53], but the molecular system evolves faster in the limited simulation time and possible kinetic traps are avoided. Electrostatic interactions were calculated using the particle

mesh Ewald (PME) method. The cutoff radius for Lennard–Jones interactions was set to 12 Å. The SHAKE algorithm [54] was used for bond constraints, and the time step was 2 fs for all simulations. Each system was simulated for 20 ns and the trajectories were saved at 4.0 ps intervals for further analysis. The VMD (Visual Molecular Dynamics) program was used to visualize the trajectories [55]. The MM-PBSA single-trajectory approach, implemented as a script (MMPBSA.py) in AMBER11, was used to calculate the binding energy. Solute entropic contributions were not calculated in this study. The entropy term was estimated using normal mode and harmonic methods for qualitative comparisons, rather than to quantitatively reproduce binding free energies [56].

In silico mutagenesis

Ten different single point glycine mutants were studied to examine the effects of the steric zipper. In chain A, the following three single point glycine mutations were performed: (a) tyrosine (Y) at position 14 was replaced with glycine (G), (b) leucine (L) at position 16 was replaced with glycine (G), and (c) asparagine (N) at position 18 was replaced with glycine (G). In chain B, a total of seven mutations were realized: (d) leucine (L) at position 11 was replaced with glycine (G), (e) valine (V) at position 12 was replaced with glycine (G), (f) glutamic acid (E) at position 13 was replaced with glycine (G), (g) alanine (A) at position 14 was replaced with glycine (G), (h) leucine (L) at position 15 was replaced with glycine (G), (i) tyrosine (Y) at position 16 was replaced with glycine (G), (j) leucine (L) at position 16 was replaced with glycine (G). The three mutants in chain A will henceforth be termed Y14G_A, L16G_A and N18G_A, respectively. The other seven mutants in chain B will be termed L11G_B, V12G_B, E13G_B, A14G_B, L15G_B, Y16G_B and L17G_B, respectively. All the starting structures for the mutants were built from the wild type

structure [57] by replacing the side chains of the targeted residues with glycine using VMD [55]. Such analogs could potentially lead to the development of more potent insulin-based medicines with extended durations of action, the ability to control this duration using prodrugs, as well as enhanced medicine bioavailability. Insulin analogs were developed in order to try to replace more physiological insulin through injection at a subcutaneous site.

Binding free-energy calculation

The insulin single-layer oligomer aggregates studied here contain multiple protein–protein interfaces, and the calculation of the free energy of association of monomers in single-layer oligomers requires a suitable interface. In order to assess the stability of the insulin oligomer as the number of strands increases (the longitudinal growth), and the effect of mutations of amino acids involved in the intra chain, we measured the interaction energy between the terminal strands and the central dimer, marked A and B, respectively in Fig. 2. A molecular mechanics–generalized Born surface area (MM-GBSA) method was used to calculate the binding free energies. The free energy analyses were done using a single trajectory approach, where whole complex and the fragments A and B snapshots were extracted from the MD trajectory. According to the MM-GBSA method [58, 59], the binding free energy was calculated using the following equation:

$$\Delta G_{\text{bind}} = \langle G_C \rangle - \langle G_A \rangle - \langle G_B \rangle, \quad (1)$$

where the bracket $\langle \rangle$ indicates the average of the energy terms over 2500 snapshots extracted from the MD simulation. The free energy of each system X (= A, B, or C) was computed as the sum of the three terms [60, 61]

$$\langle \Delta G_X \rangle = \langle E_{\text{MM}} \rangle + \langle \Delta G_{\text{solv}} \rangle - T \langle S \rangle. \quad (2)$$

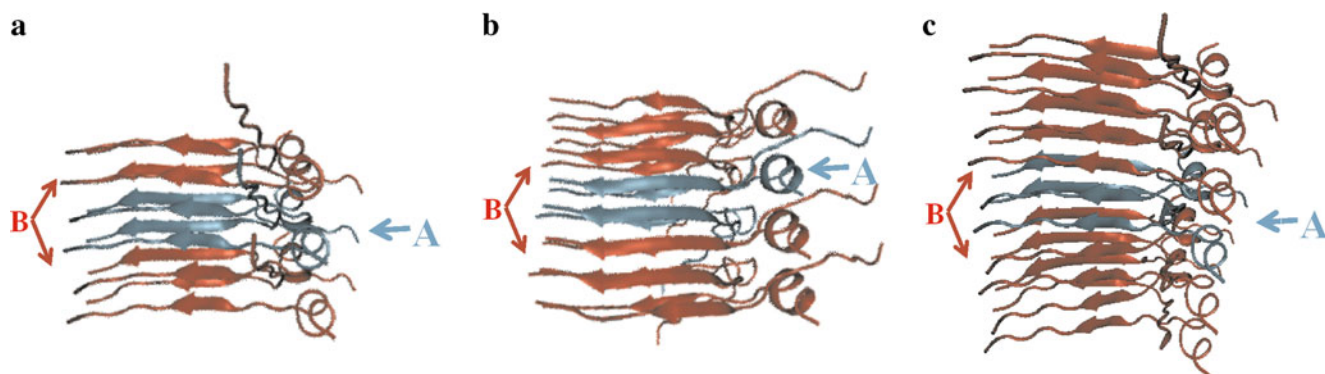


Fig. 2 Schematic of the setup used to estimate the internal stability of the insulin single-layer aggregates and mutants. Free energies of interaction were calculated between the central chains A (cyan) and

the terminal strands B (red), reflecting the strength with which chain A clamps the insulin stack in the β -solenoid structure

Here, E_{MM} is the molecular mechanical energy of the molecule expressed as the sum of the internal energy (bonds, angles and dihedrals) (E_{int}), electrostatic energy (E_{ele}) and the van der Waals term (E_{vdw}):

$$E_{MM} = E_{int} + E_{ele} + E_{vdw}. \quad (3)$$

ΔG_{solv} accounts for the solvation energy, which can be divided into polar and nonpolar parts:

$$\Delta G_{solv} = \Delta G_{GB} + \Delta G_{SA}. \quad (4)$$

The polar part ΔG_{GB} accounts for the electrostatic contribution to the solvation and is obtained from generalized Born (GB) calculations in a continuum model of the solvent. The second term, ΔG_{SA} , is the nonpolar contribution to the solvation free energy, which is linearly dependent on the solvent-accessible surface area (SASA):

$$\Delta G_{SA} = \gamma SASA + b. \quad (5)$$

The parameters γ and b were set to their default values in AMBER11. The entropic contribution was not calculated in this study, since it is only crudely estimated using normal mode analysis [59, 62].

Results and discussion

Relative structural stabilities of insulin oligomers

The conformational changes of the oligomers and the conservation of their stability were monitored by observing the time evolution of the root mean square (RMSD) and the root mean square fluctuation (RMSF) of the backbone. The RMSDs provide useful information on the relative stabilities of the oligomers, and were previously used in stability analyses of amyloid oligomers with β -sheet structure [39, 63–66]. Figure 3 plots the RMSDs of the wild-type and mutant oligomer aggregates relative to the corresponding initial structure as a function of simulation time.

RMSD

The conformational changes of the wild-type insulin oligomers of different sizes and the conservation of their stability were monitored by watching the variations in the RMSD over time. Figure 3 plots the RMSDs of the main-chain heavy atoms of the insulin oligomers relative to the corresponding initial structure as a function of simulation time. The RMSD time profiles evolve into reasonable plateaux during the course of the 10 ns production run, indicating that statistical convergence was attained in these simulations. The average main-chain RMSDs between the

MD simulation and the initial structure were found to be 4.3–4.9 Å for WT and 3.75–4.75 Å for the mutants. Along the trajectories, the systems tended to retain their original conformations.

High conformational flexibilities were observed for the wild-type monomer and dimer, as indicated by the RMSDs, RMSFs (Figs. 3 and 4), average secondary structure contents (Table 1) and a cluster analysis (Fig. S2). The RMSFs and the cluster analysis presented in Fig. 3 and Fig. S2 for the monomer indicate that it undergoes significant conformational changes as it forms a globular structure instead of its initial solenoid form. The C-terminus of the monomer bends into the central region and forms an antiparallel β -sheet between residues 12–16 and residues 35–40. The dimer largely preserves its solenoid conformation, but exhibits large per-residue fluctuation values in the β -sheet region in chains A (${}_{11}\text{SLYQLENY}_{19}$) and B (${}_{12}\text{VEALYL}_{17}$); these values are twice as large as the RMSFs in other cases (Fig. 4).

RMSF

The residue-based root mean square fluctuations (RMSFs) of the backbones were used to assess the local dynamics and flexibilities of the residues using the Ptraj tool in AMBER11. A detailed analysis of the RMSFs of the C_{α} , C, and N atoms versus the residue number for wild-type and mutant insulin oligomer aggregates is shown in Fig. 4. Large oligomers such as SH1-ST8 and SH1-ST10 are more flexible at their N- and C-termini than smaller oligomers (except for SH1-ST2). The relatively large RMSFs per residue in the β -sheet regions of SH1-ST1 and SH1-ST2 indicate that they are relatively unstable and that the initial fibril conformation is lost (Fig. 4A). For the other oligomers (SH1-ST4 to SH1-ST10), the β -sheet region exhibits a much smaller structural fluctuation from the fibril conformation. Figure 4B and C show the RMSF values for each residue, as computed throughout the simulation for wild type insulin (SH1-ST10) and its corresponding single point glycine mutants. The RMSFs of the single point mutants were found to be larger than those of the wild type. The smallest fluctuations in average RMSF for chains A and B were found in the segments LYQLENY and LVEALYL, respectively. The RMSF results for the wild type and the mutants indicate that all of the chains show only small variations for the residues located within the β -sheet region but large variations for the residues in the termini regions. The greater flexibility of the two termini residues is due to a reduction in the hydrogen bonds between the peptides. The side chains of the termini residues are more exposed to the water and tend to form hydrogen bonds with water molecules [66].

Fig. 3 Backbone RMSDs of the single-layered insulin A1-21 and B1-19 models with 1, 2, 4, 6, 8 and 10 β -solenoids and single point mutants (SH1-ST10). The RMSD curves are shown for all residues

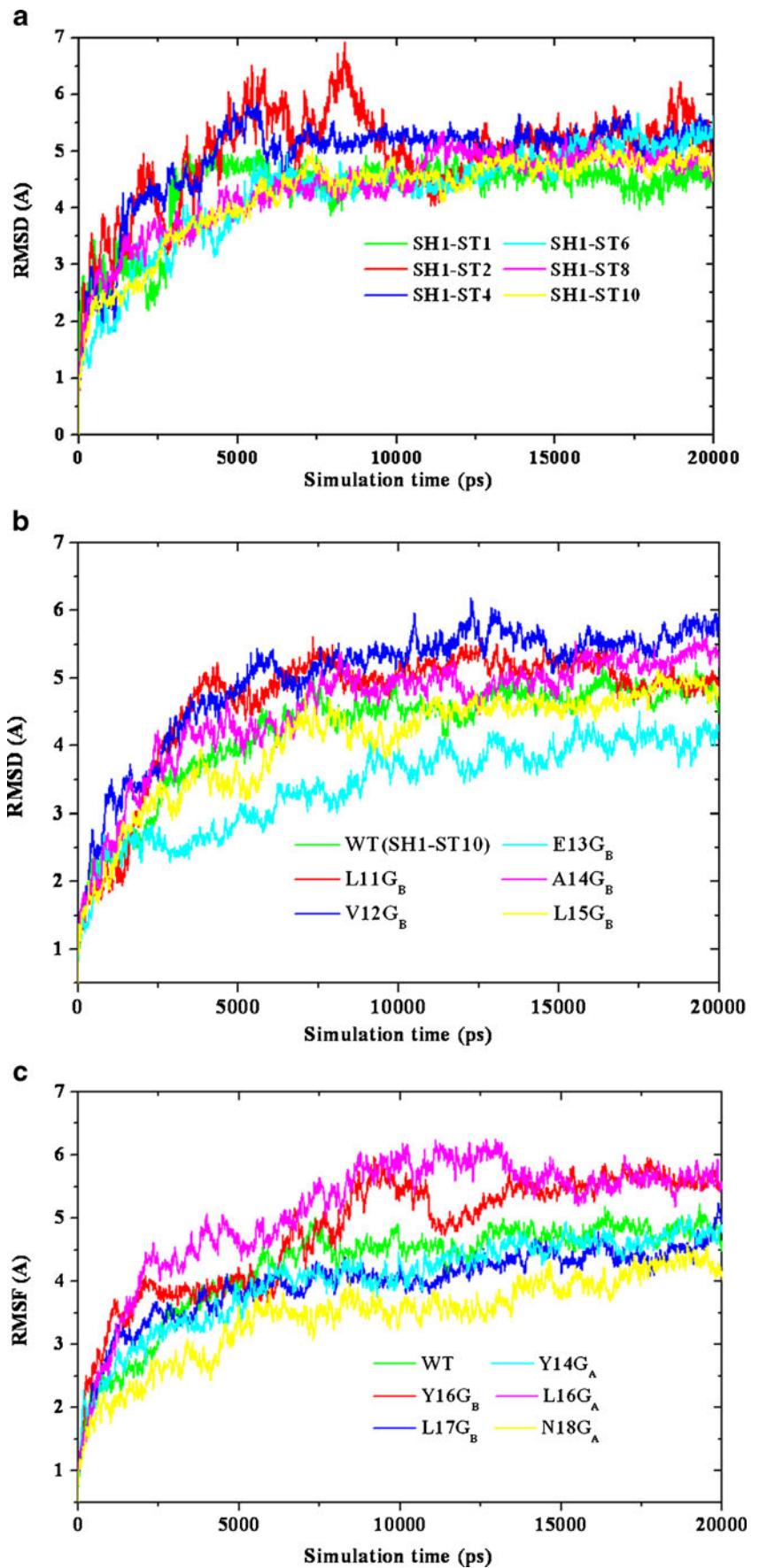
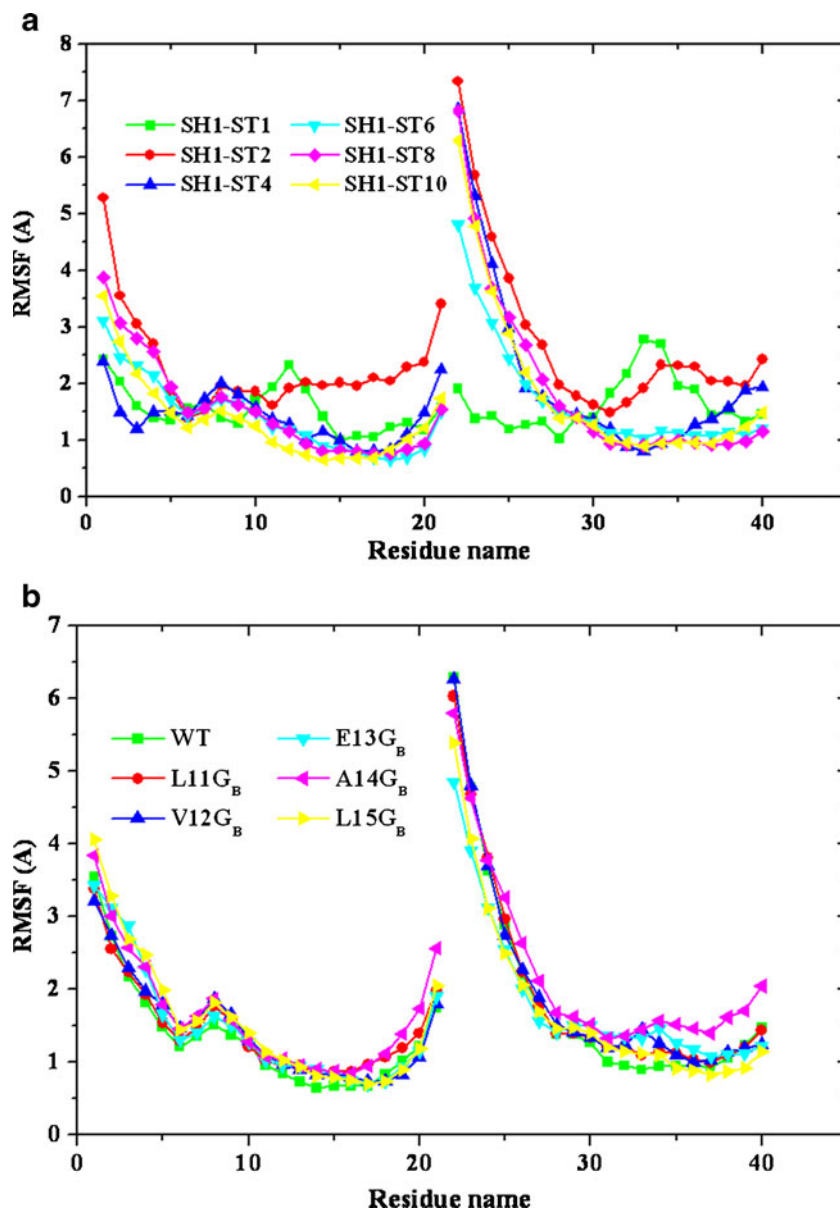


Fig. 4 Plot of the average root mean square fluctuations (RMSFs) from the 20 ns simulation with respect to the initial structure. Residues of chain A are numbered 1–21 and those of chain B are numbered 22–40



Secondary structure content

We carried out secondary structure analysis using the dssp tool in AMBER11 [67]. Table 1 reports the average number of residues in a given secondary structure as a function of simulation time and the corresponding initial structure (Fig. S3). When the average secondary structure content over time is considered, differences between the smaller and larger oligomers become evident from the simulations. The single- and double-stranded aggregates exhibit lower β -sheet contents, and more residues in helices and coil-like conformations. The larger oligomers (such as SH1-ST4 to SH1-ST10) exhibit more β -sheet contents and fewer residues in helices and coil-like conformations. The larger aggregates retain the fibril conformation mainly due to an increased number of backbone hydrogen bonds [66].

Cluster analysis

Cluster analysis (“clustering”) places similar samples of data into groups called clusters, such that an ensemble of data (for example the different structures obtained from an MD trajectory) is partitioned into groups of similar objects. Structural clustering is useful for understanding the molecular motion within conformational space [68]. Conventional clustering algorithms can reduce any large MD trajectory to a set of conformational basins. To identify the most populated conformations sampled, clustering was applied to all snapshots from the trajectories using the Ptraj program of AMBER11. The standard approach, which has been used with considerable success, is to cluster the configurations in terms of the RMSD. To perform the clustering, we utilized the average linkage algorithm implemented in Ptraj [69]. The

Table 1 Average secondary structure contents^a of the wild-type insulin oligomers of different sizes and the single point glycine mutants (of SH1-ST10) at the steric zipper between the A and B chains, as well as those of the corresponding starting structures

	Starting	Average	Starting	Average	Starting	Average	Starting	Average	Starting	Average	Starting	Average						
	WT (SH1-ST1)	0 (0)	0 (0)	WT (SH1-ST2)	0 (0)	0 (0)	WT (SH1-ST4)	0 (0)	0 (0)	WT (SH1-ST6)	0 (0)	0 (0)	WT (SH1-ST8)	0 (0)	0 (0)	WT (SH1-ST10)	0 (0)	0 (0)
β-Sheet	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
β-Bridge	0 (0)	0.33 (3.4), [2.1]	23 (63.9)	15.3 (56.1), [3.6]	47 (63.5)	42.3 (62.3), [3.1]	83 (68)	69.9 (79.6), [5.2]	115 (68.1)	114 (69.2), [5.7]	158 (73.8)	135.6 (70.1), [7.6]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Coil	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Turn	7 (100)	5.35 (54.8), [2.4]	2 (5.5)	6.4 (23.5), [2.4]	9 (12.2)	10.3 (15.2), [3.4]	25 (20.5)	16.4 (18.7), [4]	35 (20.7)	28.1 (17.1), [6.1]	45 (21)	28.1 (17.1), [6.1]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
α-Helix	0 (0)	1.47 (15.06), [0.3]	11 (30.6)	3.9 (14.3), [2.9]	14 (18.9)	12.1 (17.8), [4.6]	8 (6.6)	14.4 (8.7), [5.9]	12 (7.1)	14.4 (8.7), [5.9]	8 (3.7)	11.2 (5.8), [4.3]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
3 ₁₀ -Helix	0 (0)	2.6 (26.64), [2.0]	0 (0)	1.6 (5.9), [2.0]	2 (2.7)	2.7 (4.0), [2.8]	6 (4.9)	8.4 (9.6), [3.9]	7 (4.1)	5.7 (3.5), [4.1]	3 (1.4)	11.8 (6.1), [4.6]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
π-Helix	0 (0)	0.01 (0.2), [0.3]	0 (0)	0.05 (0.2), [0.1]	2 (2.7)	0.5 (0.7), [1.2]	0 (0)	1.3 (1.5), [1.9]	0 (0)	2.4 (1.5), [2.7]	0 (0)	0.5 (0.3), [1.0]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	WT (SH1-ST10)			L11G _B (SH1-ST10)			V12G _B (SH1-ST10)			E13G _B (SH1-ST10)			A14G _B (SH1-ST10)			L15G _B (SH1-ST10)		
β-Sheet	0 (0)	0 (0), [0]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
β-Bridge	158 (73.8)	135.6 (70.1), [7.6]	135 (71)	127.7 (68.3), [7.9]	146 (72.3)	128 (68.8), [6.4]	131 (68.6)	130.5 (69.6), [7.9]	134 (69.1)	120.9 (69.2), [7.7]	157 (70.7)	142.5 (70.2), [7.6]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Coil	0 (0)	0 (0), [0]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Turn	45 (21)	28.1 (17.1), [6.1]	28 (14.7)	31.5 (16.8), [4.9]	36 (17.8)	37.6 (20.2), [5.0]	45 (23.6)	34.9 (18.6), [5.7]	46 (23.7)	38.8 (21.6), [5.2]	40 (18)	32.6 (16.1), [6.7]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
α-Helix	8 (3.7)	11.2 (5.8), [4.3]	16 (8.4)	16.5 (8.8), [4.9]	9 (4.5)	11.5 (6.2), [4.6]	4 (2.1)	12.6 (6.7), [5.3]	8 (4.1)	7.0 (3.9), [3.6]	18 (8.1)	17.5 (8.6), [6.2]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
3 ₁₀ -Helix	3 (1.4)	11.8 (6.1), [4.6]	9 (4.7)	8.8 (4.7), [4.7]	6 (3.0)	8.4 (4.5), [4.3]	9 (4.7)	9.5 (5.1), [4.2]	6 (3.1)	9.8 (5.4), [4.1]	7 (3.2)	8.2 (4.0), [4.5]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
π-Helix	0 (0)	0.5 (0.3), [1.0]	2 (1)	2.5 (1.4), [2.2]	5 (2.5)	0.5 (0.3), [1.3]	4 (2.1)	0.3 (0.2), [1.0]	0 (0)	1.4 (0.8), [2.4]	0 (0)	2.3 (1.1), [3.0]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	WT (SH1-ST10)			Y16G (SH1-ST10)			L17G _A (SH1-ST10)			Y14G _A (SH1-ST10)			L16G _A (SH1-ST10)			N18G _A (SH1-ST10)		
β-Sheet	0 (0)	0 (0), [0]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
β-Bridge	158 (73.8)	135.6 (70.1), [7.6]	133 (72.3)	127.7 (68.3), [6.0]	146 (72.6)	135.1 (71.5), [6.7]	143 (69.1)	121.1 (66.0), [7.9]	139 (67.8)	127.8 (65.8), [6.8]	151 (70.9)	122.9 (68.7), [8.2]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Coil	0 (0)	0 (0), [0]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Turn	45 (21)	28.1 (17.1), [6.1]	39 (21.2)	31.5 (16.8), [4.9]	23 (11.4)	30.6 (16.2), [5.4]	29 (14)	30.0 (16.4), [5.9]	38 (18.5)	39.0 (20.1), [5.8]	36 (16.9)	36.4 (20.4), [4.4]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
α-Helix	8 (3.7)	11.2 (5.8), [4.3]	10 (5.4)	16.5 (8.8), [4.9]	24 (12)	17.9 (9.5), [5.2]	24 (11.6)	21.0 (11.4), [7.6]	13 (6.3)	16.3 (8.4), [5.1]	10 (4.7)	7.3 (4.1), [5.6]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
3 ₁₀ -Helix	3 (1.4)	11.8 (6.1), [4.6]	0 (0)	8.8 (4.7), [4.7]	3 (1.5)	3.1 (1.6), [2.8]	6 (2.9)	10.2 (5.6), [5.6]	3 (1.5)	9.8 (5.0), [4.5]	9 (4.2)	9.0 (5.0), [4.1]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
π-Helix	0 (0)	0.5 (0.3), [1.0]	2 (1.1)	2.5 (1.4), [2.2]	5 (2.5)	2.5 (1.2), [2.2]	5 (2.4)	1.2 (0.6), [2.0]	12 (5.8)	1.2 (0.6), [2.0]	7 (3.3)	3.2 (3.2), [2.4]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

^a Percentages and standard deviations are given in parentheses and square brackets, respectively

uniqueness or equivalence of different clusters was assessed based on a visual comparison of representative structures. The clustering was performed on a 5000-frame reference set (4 ps sampling rate). Table 2S shows the clustering metric values, and Fig. S2 shows the superposition of the initial structure and the most populated cluster structure for the single-layer insulin aggregates of different sizes. The analysis of the structures indicated that the most populated clusters were associated with the smaller oligomers (single- and double-stranded), indicating that these underwent greater structural rearrangements from the initial conformation taken from the fibril model. The conformation was preserved for the larger aggregates (SH1-ST8 and SH1-ST10).

Free-energy calculation

A detailed characterization of the individual energy terms of the calculated binding free energies of the studied insulin oligomer aggregates is shown in Table 2. An inspection of the free energy components for the wild types and mutants reveals that the electrostatic component of the free energy of binding (ΔE_{ele}) contributes unfavorably to the binding ($\Delta G > 0$). The nonpolar component contributes favorably ($\Delta G < 0$), as expected, since the formation of complexes reduces the solvent-accessible surface area. In most cases, the electrostatic component of the solvation free energy ΔG_{GS} is consistently favorable. The energy due to electrostatic

interactions (ΔE_{ele}) between strands led to unfavorable binding. These observations are consistent with previous calculations of the electrostatic component of the free energy of solvation. However, the less favorable electrostatics in each case are compensated for by the highly favorable nonpolar component of the free energy. In each case, the favorable nature of the nonpolar interaction originates from the nonpolar component of solvation (ΔG_{GB}) and the van der Waals interaction energy (ΔE_{vdw}).

The results of the binding free-energy calculations (Table 2) indicate that the structurally stable models have the lowest binding free energies, while the models that are structurally unstable were found to have the highest binding free energies. The difference in binding free energy between the unmutated (wild-type) and mutated complex is defined as:

$$\Delta\Delta G_{\text{mut}} = \Delta G_{\text{mut}} - \Delta G_{\text{wild}}. \quad (6)$$

Positive and negative $\Delta\Delta G_{\text{mut}}$ values indicate unfavorable and favorable contributions, respectively. The positive $\Delta\Delta G_{\text{mut}}$ values of 37.3–1.4 kcal mol⁻¹ of the mutants in the β -sheet region (except for Y14G_A and L15G_B) indicate lower tendencies to aggregate than the wild type. This finding could be used in the rational design of new insulin analogs with decreased propensities for self-association, thus reducing the risk of injection amyloidosis of insulin.

Table 2 Individual energy components for the calculated binding free energies of insulin amyloid aggregate peptides

MM-GBSA binding energy components of the single-layer insulin amyloid aggregates of different sizes							
System	ΔE_{vdw}	ΔE_{ele}	ΔG_{GB}	ΔG_{GS}	ΔG_{solv}	ΔG_{total}	$\Delta\Delta G_{(6-n)}$
WT(SH1-ST4)	-163.96±0.18	576.65±1.02	-502.40±0.95	-21.35±0.02	-523.75±0.95	-111.06±0.17	11.43
WT(SH1-ST6)	-177.21±0.28	1149.66±0.91	-1071.50±0.86	-23.43±0.03	-1094.94±0.86	-122.49±0.24	0.0
WT(SH1-ST8)	-149.64±0.19	1514.62±2.27	-1433.42±2.22	-25.81±0.02	-1459.27±2.22	-142.29±0.22	-19.8
WT(SH1-ST10)	-196.31±0.14	1827.75±	-1742.52±1.00	-25.20±0.02	-1767.28±0.13	-136.28±0.13	-13.79
MM-GBSA binding energy components of chain-A mutants of single-layer insulin amyloid aggregates (SH1-ST10)							
System	ΔE_{vdw}	ΔE_{ele}	ΔG_{GB}	ΔG_{GS}	ΔG_{solv}	ΔG_{total}	$\Delta\Delta G_{\text{mut}}$
Y14G _A (chain A)	-165.16±0.12	1974.31±0.84	-1887.76±0.84	-20.36±0.02	-1908.11±0.84	-98.96±0.16	37.3
L16G _A (chain A)	-208.14±0.14	1710.42±1.74	-1628.03±1.73	-27.10±0.03	-1655.15±1.73	-152.87±0.16	-16.59
N18G _A (chain A)	-201.272±0.12	1623.99±0.93	-1527.81±0.91	-26.02±0.01	-1553.82±0.91	-131.11±0.15	5.17
MM-GBSA binding energy components of chain-B mutants of single-layer insulin amyloid aggregates (SH1-ST10)							
System	ΔE_{vdw}	ΔE_{ele}	ΔG_{GB}	ΔG_{GS}	ΔG_{solv}	ΔG_{total}	$\Delta\Delta G_{\text{mut}}$
L11G _B (chain B)	-171.7±0.23	1598.8±3.4	-1522.2±3.4	-22.0±0.02	-1544.2±3.4	-117.1±0.2	19.2
V12G _B (chain B)	-167.1±0.2	1789.9±1.1	-1715.79±1.0	-21.5±0.01	-1737.2±1.0	-114.4±0.2	21.9
E13G _B (chain B)	-186.7±0.3	981.4±1.0	-893.7±1.0	-25.3±0.02	-918.9±1.06	-124.2±0.25	12.1
A14G _B (chain B)	-191.2±0.2	1620.8±4.0	-1533.3±4.0	-25.4±0.03	-1558.7±4.0	-129.2±0.25	7.1
L15G _B (chain B)	-209.3±0.2	1622.9±2.4	-1529±2.4	-26.6±0.02	-1555.6±2.4	-142.03±	-5.7
Y16G _B (chain B)	-194.8±0.2	1752.6±1.9	-1663.1±1.9	-23.3±0.03	-1686.4±1.9	-128.622±0.2	7.6
L17G _B (chain B)	-204.4±0.3	1823.9±2.3	-1728.42±2.2	-26.0±0.02	-1754.43±2.3	-134.904±0.2	1.4

E_{vdw} and E_{elec} are the van der Waals and electrostatic binding terms, ΔG_{GB} , ΔG_{GS} , ΔG_{solv} are the polar, nonpolar and total solvation energies. Data are shown as mean ± SD. $\Delta G_{\text{total}} = \Delta E_{\text{vdw}} + \Delta E_{\text{ele}} + \Delta G_{\text{sol}}$; $\Delta G_{\text{sol}} = \Delta G_{\text{GB}} + \Delta G_{\text{GS}}$; $\Delta G_{\text{mut}} = \Delta G_{\text{mut}} - G_{\text{wild}}$ (the difference in binding free energy between the mutant and the wild type). $\Delta\Delta G_{(n-6)}$ is the oligomer free energy expressed relative to the hexamer state for β -sheet oligomers

The relatively large positive values of $\Delta\Delta G_{\text{mut}}$ for the mutants (Y14G_A, L11G_B, V12G_B and E12G_B) indicate that they are less likely to associate than the wild type. In general, substituting the β -sheet regions of chains A and B for a small, short Gly disrupts the shape complementarity of the steric zipper and weakens hydrophobic interactions (see Table 2, Tables 3S–5S). The single point glycine mutation reduces the unfavorable electrostatic interaction. The mutation of the negatively charged glutamate (E) to G in the mutant E12G_B reduces the electrostatic repulsion that occurs in the wild type, as shown by the significantly reduced unfavorable electrostatic interaction (Table 2). The mutation of Tyr14 in chain A to glycine eliminates the hydrogen bond between Tyr14 of chain A and Tyr16 of chain B, so the calculated binding free energy was high in this case. The negative values of $\Delta\Delta G_{\text{mut}}$ for the mutants are due to the increased hydrophobic interactions in the steric zipper between the chains (Tables 4S and 5S). The trend in the calculated binding free energy is in agreement with the observed instabilities based on RMSD and RMSF. The aggregate oligomer models that showed structural instability were found to have unfavorable binding energies compared to the stable models. This is also in agreement with experiment, which finds that complete substitution of the hydrophobic side chain for Gly impedes fibril growth [70].

Free energy decomposition on a per-residue basis

Free energy decomposition not only identifies ‘hotspots’ in the binding energy but also provides insight into the nature of the key interactions [45]. To provide basic information on the intermolecular interactions contributed by individual residues in the single-layer insulin aggregate interaction, the free energy (the per-residue combined side chain and backbone binding free energy) was decomposed using MM-GBSA module in AMBER11. The calculation was performed over the 2500 MD snapshots taken from the 20 ns simulation. The per-residue interaction free energies were separated into those for the residue backbone (ΔG for backbone binding) and the side chain (ΔG for side chain binding). The energy contributions from the selected residues are summarized in Fig. 5.

The results from the energy decomposition show that the major contribution to the binding energy of the insulin oligomer aggregate derives from key amino acid residues (those with $\Delta G_{\text{binding}} \leq -0.50$ kcal mol⁻¹) that occur mainly in the β -sheet region. These residues are in chain A (Q5, L13, Y14, Q15, L16, N18 and Y19) and in chain B (S9, L11, V12, L15, L17 and V18). The results of the per-residue decomposition indicate the importance of these particular residues in the β -sheet region in the formation and stabilization of insulin, which is in agreement with experimental

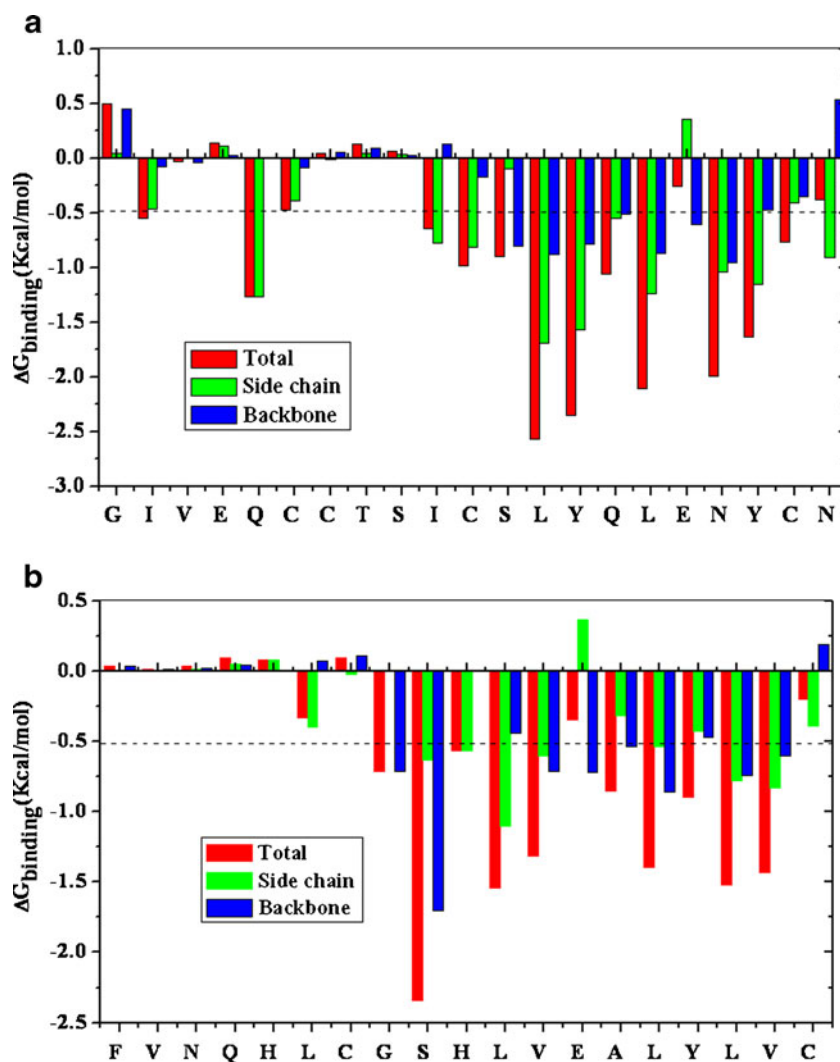
observations [14]. To establish the interactions associated with these residues, their electrostatic, van der Waals and solvation energy terms are shown in Table 2. Table 2 shows that favorable contributions to the binding free energy arising from these residues relate to E_{elec} , E_{vdw} , and ΔG_{GS} , while unfavorable contributions come from ΔG_{GB} . The favorable E_{elec} terms from the residues in the β -sheet region are compensated for by highly unfavorable repulsion from three glutamates between the adjacent insulin strands.

Fibril nucleation and the structures of insulin oligomers

Understanding the process of amyloid fibril formation is an important goal of protein aggregation studies [71]. Amyloids grow in a nucleation-dependent manner [71, 72]. Fibrillation kinetics are typically characterized by an initial apparent lag phase related to the formation of oligomers, protofibrils, and aggregation nuclei [73]. No detectable fibers are formed during this lag phase. This is followed by an elongation phase in which the fiber is formed over a time period that is often shorter than the lag phase. Eventually, the process reaches equilibrium when the most soluble proteins are converted into fibrils [74]. On the other hand, if fibers (oligomers, protofibrils, and fibrils) have already formed, they grow extremely quickly, with very short lag times. Recent experimental works performed in various labs on the capacity of oligomers to stabilize fibril nucleation activity on A β amyloid [75], on amylin [74], and on insulin [76–81] have indicated that oligomers and fibrils have different capacities to act as seeds. Anselm et al. [40] used the degree of structural similarity of each oligomer to the fibril conformation found in their simulation as a possible reason for differences among oligomers of various sizes in terms of their effectiveness as nucleation seeds. The degree of structural similarity between the fibril conformation and the conformation of the oligomer after MD simulation could be used to explain the shortening of the lag phase in the presence of oligomers of different sizes. The trend in relation to retaining the initial fibril conformation will help us to elucidate an atomic-level explanation for the observed difference in the seed effects of oligomers of various sizes. The results from our simulation show that a single-layer insulin oligomer as small as the trimer is capable of preserving the conformation present in the fibril (see Fig. S2). The dimer only retains some of the properties of the mature fibril, while the monomer adopts a structure that differs significantly from that of the fibril.

The secondary structure contents and the clustering analysis of the trajectories for the single-layer insulin oligomers of differing sizes show that the larger aggregates retain the fibril conformation but the smaller ones (SH1-ST1 and SH1-ST2) lose this conformation. This observation can be used to explain the shortening of the nucleation

Fig. 5a, b Decomposition of the free energy on a per-residue basis for chain A (a) and chain B (b) of the ten-strand single-layer insulin aggregate



lag phase of insulin aggregation with oligomer seeds. Insulin, like other amyloid peptides, appears to follow nucleation-dependent polymerization kinetics [71, 82] whereby a small number of monomers associate through a free energy barrier corresponding to a critical nucleus size; beyond this size there is a gradient of favorable free energy or “downhill” polymerization. Based on the results for the secondary structure contents and the cluster analysis, we propose that SH1-ST4 is a critical nucleus for single-layer insulin fibril oligomer growth. To characterize the critical nucleation, we computed the difference in association energy between our proposed minimum nuclei and the larger oligomers (SH1-ST6, SH1-ST8 and SH1-ST10) using the equation

$$\Delta\Delta G(n) = G(n) - (G_4); n = 4, 6, 8, 10. \quad (7)$$

The results are shown in Table 2 and are plotted in Fig. S3. Our calculations show that, for a high number of strands, the oligomer is stable and its free energy is favorable for the addition of new chains. The results of

our semi-quantitative approach for insulin single layers of limited size are in agreement with those obtained from previous extensive simulations done on the critical nucleus and mechanism of fibril elongation for A β amyloid [83, 84].

Conclusions

The results from this work provide valuable insight into the forces that drive the stability of the peptide–peptide complexes in the single-layer aggregate oligomer models of insulin and those that lead to unstable complexes. The study of the wild type and mutants in an explicit solvent may prove valuable to future efforts aimed at the design of short- and long-acting insulin analogs. The major findings of this study can be summarized as follows:

- (i) The stabilities of the single-layer insulin peptide oligomers increase as the number of strands increases (dynamic cooperative effect).

- (ii) The binding energy calculated by the MM-GBSA method shows that hydrophobic interactions play an important role in stabilizing the structural organization of the single-layer insulin. Per-residue decomposition shows that the key amino acid residues (those with $\Delta G_{\text{binding}} \leq -1.00 \text{ kcal mol}^{-1}$) occur mainly in the β -sheet regions of chains A and B. Due to the electrostatic repulsion between the three negatively charged glutamates in adjacent insulin strands, electrostatic contribution to the binding energy is unfavorable.
- (iii) A single glycine substitution at the steric zipper interface disrupts the hydrophobic contacts and reduces the van der Waals interactions in the mutants, thus reducing the binding free energy. The results of the binding free-energy calculation indicated that the wild type is more structurally stable than most of the mutants. A comparison of the binding free energy between the wild type and the chain-A mutants (Y14G_A, L16G_A and N18G_A) indicated that shape complementarity between neighboring strands plays a key role in stabilizing the entire oligomeric structure.
- (iv) The secondary structure contents and the clustering analysis of the trajectories of the single-layer insulin oligomers of various sizes showed that the larger aggregates retain the fibril conformation but the smaller ones (SH1-ST1 and SH1-ST2) lose this conformation. This observation could explain the observed shortening of the nucleation lag phase of insulin aggregation with oligomer seeds. Based on the secondary structure contents and the cluster analysis, we propose that SH1-ST4 is a critical nucleus for single-layer insulin fibril oligomer growth.

Our simulations provide detailed insight into the structural stability and aggregation behavior of wild-type and mutant single-layer insulin aggregates (obtained from the high-resolution insulin fibril model) at the atomic level. In search for clinically advantageous fast-acting insulin analogs, several approaches were found to be useful for altering the monomer/monomer interface. These include the disruption of β -sheet interactions in the β -chain through charge repulsion, and changes in hydrophobic interactions at the C-terminus of chain B [50]. Our simulations of wild-type and single glycine mutants at the steric zipper region show that other parts of the insulin molecule can be targeted in the design of both short- and long-acting insulin analogs as well. Aside from the design of such insulin analogs, the present study may prove useful in the rational design of insulin aggregation inhibitors that can be used to stabilize insulin formulations, leading to safer handling and more cost-effective storage of such formulations, especially in developing countries.

Acknowledgments This work was supported in part by the National Science Foundation (CCF/CHE 0832622). This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231.

References

- Zierath JR, Krook A, Wallberg-Henriksson H (2000) *Diabetologia* 43:821–835
- Shepherd PR, Kahn BB (1999) *N Engl J Med* 341:248–257
- Nystrom FH, Quon MJ (1999) *Cell Signal* 11:563–574
- Ottensmeyer FP, Beniac DR, Luo RZT, Yip CC (2000) *Biochemistry* 39:12103–12112. doi:10.1021/bi0015921
- Wild S, Roglic G, Green A, Sicree R, King H (2004) *Diabetes Care* 27:1047–1053
- Rewers M (2008) *Diabetes Care* 31:830–832. doi:10.2337/dc08-0245
- Storkel S, Schneider HM, Muntefering H, Kashiwagi S (1983) *Lab Invest* 48:108–111
- Dische FE, Wernstedt C, Westermark GT, Westermark P, Pepys MB, Rennie JA, Gilbey SG, Watkins PJ (1988) *Diabetologia* 31:158–161
- Greenwald J, Riek R (2010) *Structure* 18:1244–1260. doi:10.1016/j.str.2010.08.009
- Sipe JD, Benson MD, Buxbaum JN, Ikeda S, Merlini G, Saraiva MJM, Westermark P (2010) *Amyloid J Protein Fold Disord* 17:101–104. doi:10.3109/13506129.2010.526812
- Brange J, Andersen L, Laursen ED, Meyn G, Rasmussen E (1997) *J Pharm Sci* 86:517–525
- Wilhelm KR, Yanamandra K, Gruden MA, Zamotin V, Malisaukas M, Casate V, Darinskas A, Forsgren L, Morozova-Roche LA (2007) *Eur J Neurol* 14:327–334. doi:10.1111/j.1468-1331.2006.01667.x
- Ahmad A, Uversky VN, Hong D, Fink AL (2005) *J Biol Chem* 280:42669–42675. doi:10.1074/jbc.M504298200
- Ivanova MI, Sievers SA, Sawaya MR, Wall JS, Eisenberg D (2009) *Proc Natl Acad Sci USA* 106:18990–18995. doi:10.1073/pnas.0910080106
- Groenning M, Frokjaer S, Vestergaard B (2009) *Curr Protein Pept Sci* 10:509–528
- Sluzky V, Klibanov AM, Langer R (1992) *Biotechnol Bioeng* 40:895–903
- Grillo AO, Edwards KLT, Kashi RS, Shipley KM, Hu L, Besman MJ, Middaugh CR (2001) *Biochemistry* 40:586–595
- Onoue S, Ohshima K, Debari K, Koh K, Shioda S, Iwasa S, Kashimoto K, Yajima T (2004) *Pharm Res* 21:1274–1283
- Valla V (2010) *Exp Diabetes Res* 14:178372. doi:10.1155/2010/178372
- Maji SK, Perrin MH, Sawaya MR, Jessberger S, Vadodaria K, Rissman RA, Singru PS, Nilsson KPR, Simon R, Schubert D, Eisenberg D, Rivier J, Sawchenko P, Vale W, Riek R (2009) *Science* 325:328–332. doi:10.1126/science.1173155
- Maji SK, Schubert D, Rivier C, Lee S, Rivier JE, Riek R (2008) *PLoS Biol* 6:240–252. doi:10.1371/journal.pbio.0060017
- Zhao F, Ma ML, Xu B (2009) *Chem Soc Rev* 38:883–891
- Bell DSH (2007) *Drugs* 67:1813–1827
- Geddes AJ, Parker KD, Atkins EDT, Beighton E (1968) *J Mol Biol* 32:343–344
- Bouchard M, Zurdo J, Nettleton EJ, Dobson CM, Robinson CV (2000) *Protein Sci* 9:1960–1967
- Burke MJ, Rougvié MA (1972) *Biochemistry* 11:2435–2439
- Nettleton EJ, Tito P, Sunde M, Bouchard M, Dobson CM, Robinson CV (2000) *Biophys J* 79:1053–1065
- Sawaya MR, Sambashivan S, Nelson R, Ivanova MI, Sievers SA, Apostol MI, Thompson MJ, Balbirnie M, Wiltzius JJW, McFarlane

- HT, Madsen AO, Riekkel C, Eisenberg D (2007) *Nature* 447:453–457. doi:10.1038/nature05695
29. Jimenez JL, Nettleton EJ, Bouchard M, Robinson CV, Dobson CM, Saibil HR (2002) *Proc Natl Acad Sci USA* 99:9196–9201. doi:10.1073/pnas.142459399
 30. Vestergaard B, Groenning M, Roessle M, Kastrup JS, van de Weert M, Flink JM, Frokjaer S, Gajhede M, Svergun DI (2007) *PLoS Biol* 5:1089–1097. doi:10.1371/journal.pbio.0050134
 31. Choi JH, May BCH, Wille H, Cohen FE (2009) *Biophys J* 97:3187–3195. doi:10.1016/j.bpj.2009.09.042
 32. Gibson TJ, Murphy RM (2006) *Protein Sci* 15:1133–1141. doi:10.1110/ps.051879606
 33. Nielsen L, Frokjaer S, Brange J, Uversky VN, Fink AL (2001) *Biochemistry* 40:8397–8409
 34. Brange J, Dodson GG, Edwards DJ, Holden PH, Whittingham JL (1997) *Proteins* 27:507–516
 35. Devlin GL, Knowles TPJ, Squires A, McCammon MG, Gras SL, Nilsson MR, Robinson CV, Dobson CM, MacPhee CE (2006) *J Mol Biol* 360:497–509. doi:10.1016/j.jmb.2006.05.007
 36. Hong DP, Fink AL (2005) *Biochemistry* 44:16701–16709. doi:10.1021/bi051658y
 37. Ivanova MI, Thompson MJ, Eisenberg D (2006) *Proc Natl Acad Sci USA* 103:4079–4082. doi:10.1073/pnas.0511298103
 38. Tito P, Nettleton EJ, Robinson CV (2000) *J Mol Biol* 303:267–278
 39. Zheng J, Jang H, Ma B, Tsai CJ, Nussinov R (2007) *Biophys J* 93:3046–3057. doi:10.1529/biophysj.107.110700
 40. Horn AHC, Sticht H (2010) *J Phys Chem B* 114:2219–2226. doi:10.1021/jp100023q
 41. Tsai HH, Reches M, Tsai CJ, Gunasekaran K, Gazit E, Nussinov R (2005) *Proc Natl Acad Sci USA* 102:8174–8179
 42. Mark AE, Berendsen HJC, Vangunsteren WF (1991) *Biochemistry* 30:10866–10872
 43. Zoete V, Meuwly M, Karplus M (2004) *Proteins Struct Funct Bioinf* 55:568–581. doi:10.1002/prot.20071
 44. Zoete V, Meuwly M (2006) *J Comput Chem* 27:1843–1857. doi:10.1002/jcc.20512
 45. Zoete V, Meuwly M, Karplus M (2005) *Proteins Struct Funct Bioinf* 61:79–93. doi:10.1002/prot.20528
 46. Falconi M, Cambria MT, Cambria A, Desideri A (2001) *J Biomol Struct Dyn* 18:761–772
 47. Lu BZ, Chen WZ, Wang CX, Xu XJ (2002) *Proteins* 48:497–504. doi:10.1002/prot.10172
 48. Sasahara K, Naiki H, Goto Y (2005) *J Mol Biol* 352:700–711. doi:10.1016/j.jmb.2005.07.033
 49. Meersman F, Dobson CM (2006) *BBA Proteins Proteomics* 1764:452–460. doi:10.1016/j.bbapap.2005.10.021
 50. Mayer JP, Zhang F, DiMarchi RD (2007) *Biopolymers* 88:687–713. doi:10.1002/bip.20734
 51. Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, Duke RE, Luo R, Walker RC, Zhang W, Merz KM, Roberts B, Wang B, Hayik S, Roitberg A, Seabra G, Kolossváry I, Wong KF, Paesani F, Vanicek J, Liu J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Cai Q, Ye X, Wang J, Hsieh MJ, Cui G, Roe DR, Mathews DH, Seetin MG, Sagui C, Babin V, Luchko T, Gusarov S, Kovalenko A, Kollman PA (2010) *AMBER 11*. University of California, San Francisco
 52. Mauro M, Craparo EF, Podesta A, Bulone D, Carrotta R, Martorana V, Tiana G, San Biagio PL (2007) *J Mol Biol* 366:258–274. doi:10.1016/j.jmb.2006.11.008
 53. Arora A, Ha C, Park CB (2004) *Protein Sci* 13:2429–2436. doi:10.1110/ps.04823504
 54. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) *J Comput Phys* 23:327–341
 55. Humphrey W, Dalke A, Schulten K (1996) *J Mol Graph* 14:33–38
 56. Jouaux EM, Timm BB, Arndt KM, Exner TE (2009) *J Pept Sci* 15:5–15. doi:10.1002/psc.1078
 57. Wiltzius JJW, Sievers SA, Sawaya MR, Cascio D, Popov D, Riekkel C, Eisenberg D (2008) *Protein Sci* 17:1467–1474. doi:10.1110/ps.036509.108
 58. 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) *Acc Chem Res* 33:889–897. doi:10.1021/ar000033j
 59. Gohlke H, Case DA (2004) *J Comput Chem* 25:238–250. doi:10.1002/jcc.10379
 60. Chong LT, Duan Y, Wang L, Massova I, Kollman PA (1999) *Proc Natl Acad Sci USA* 96:14330–14335
 61. Massova I, Kollman PA (1999) *J Am Chem Soc* 121:8133–8143
 62. Chong LT, Pitera JW, Swope WC, Pande VS (2009) *J Mol Graph Model* 27:978–982. doi:10.1016/j.jmgm.2008.12.006
 63. Buchete NV, Hummer G (2007) *Biophys J* 92:3032–3039. doi:10.1529/biophysj.106.100404
 64. Huet A, Derreumaux P (2006) *Biophys J* 91:3829–3840. doi:10.1526/biophysj.106.090993
 65. Berhanu WM, Masunov AE (2010) *Biophys Chem* 149:12–21. doi:10.1016/j.bpc.2010.03.003
 66. Berhanu WM, Masunov AE (2011) *J Mol Model*. doi:10.1007/s00894-010-0912-4
 67. Kabsch W, Sander C (1983) *Biopolymers* 22:2577–2637
 68. Keller B, Daura X, van Gunsteren WF (2010) *J Chem Phys* 132:16. doi:10.1063/1.3301140
 69. Shao JY, Tanner SW, Thompson N, Cheatham TE (2007) *J Chem Theor Comput* 3:2312–2334. doi:10.1021/ct700119m
 70. Takeda T, Klimov DK (2009) *Biophys J* 96:4428–4437. doi:10.1016/j.bpj.2009.03.015
 71. Bhak G, Choe YJ, Paik SR (2009) *BMB Rep* 42:541–551
 72. Harper JD, Lansbury PT (1997) *Annu Rev Biochem* 66:385–407
 73. Soto C, Estrada L, Castilla J (2006) *Trends Biochem Sci* 31:150–155. doi:10.1016/j.tibs.2006.01.002
 74. Padrick SB, Miranker AD (2002) *Biochemistry* 41:4694–4703. doi:10.1021/bi0160462
 75. Ono K, Condrón MM, Teplow DB (2009) *Proc Natl Acad Sci USA* 106:14745–14750. doi:10.1073/pnas.0905127106
 76. Sorci M, Grassucci RA, Hahn I, Frank J, Belfort G (2009) *Proteins* 77:62–73. doi:10.1002/prot.22417
 77. Heldt CL, Sorci M, Posada D, Hirska A, Belfort G (2011) *Biotechnol Bioeng* 108:237–241. doi:10.1002/bit.22902
 78. Nayak A, Lee CC, McRae GJ, Belfort G (2009) *Biotechnol Prog* 25:1508–1514. doi:10.1002/btpr.255
 79. Nayak A, Sorci M, Krueger S, Belfort G (2009) *Proteins* 74:556–565. doi:10.1002/prot.22169
 80. Manno M, Giacomazza D, Newman J, Martorana V, San Biagio PL (2010) *Langmuir* 26:1424–1426. doi:10.1021/la903340v
 81. Fodera V, Cataldo S, Librizzi F, Pignataro B, Spiccia P, Leone M (2009) *J Phys Chem B* 113:10830–10837. doi:10.1021/jp810972y
 82. Xue WF, Homans SW, Radford SE (2008) *Proc Natl Acad Sci USA* 105:8926–8931. doi:10.1073/pnas.0711664105
 83. Fawzi NL, Okabe Y, Yap EH, Head-Gordon T (2007) *J Mol Biol* 365:535–550. doi:10.1016/j.jmb.2006.10.011
 84. Fawzi NL, Kohlstedt KL, Okabe Y, Head-Gordon T (2008) *Biophys J* 94:2007–2016. doi:10.1529/biophysj.107.121467

Analysis of surface cavity in serpin family reveals potential binding sites for chemical chaperone to reduce polymerization

Poonam Singh · Mohammad Sazzad Khan ·
Asma Naseem · Mohamad Aman Jairajpuri

Received: 24 March 2011 / Accepted: 26 April 2011 / Published online: 17 June 2011
© Springer-Verlag 2011

Abstract Serpin constitute about 10% of blood protein and are associated with mutations that results in aberrant intermolecular linkages which leads to polymer formation. Studies with short peptides have shown promise in depolymerization of serpins however a reactive center loop based peptide also makes the serpin inactive. A chemical chaperone based approach is a better option in terms of maintaining activity and retarding polymerization but not much is known about its binding and mechanism. Specific target for chemical chaperones and its effectiveness across many serpin is not known. We did an analysis of serpin cavity using CASTp and show that cavities are distributed throughout the molecule where the largest cavities are generally present in areas of major conformational change like shutter region, helix D and helix F. An analysis of different conformational states of serpins showed that this large cavity undergoes increase in size in latent and cleaved states as compared to native state. We targeted serpins with a variety of carbohydrate, methylamine and amino acid based chemical chaperones and selected those that have highest binding energy across different serpins to assess their ability to bind large cavities. The results show that carbohydrate based chemical chaperone like sorbitol, sucrose, arabitol and trehalose and amino acid based chaperones like dopamine, phenylalanine, arginine and glutamic acid are the most effective in binding serpins. Most of these chemical chaperone interacted with residues in the shutter region and the helix D arm at the C-terminal which are part of the largest cavities. We selected the

carbohydrate based chemical chaperone with best binding energies and did experimental study under the condition that induce polymerization and show that indeed they were able to retard polymer formation with moderate effect on inhibition rates. However a fluorometric study with native antithrombin showed that chemical chaperone may effect the conformation of the proteins. Our study shows that chemical chaperones have the best binding affinities for the cavities around shutter region and helix D and that a cavity targeting based approach seems to be a better option for retarding polymerization in serpins, but a thorough analysis of its effect on folding, inhibition and cofactor binding is required.

Keywords Autodock · CastP · Protein polymerization · Reactive center loop · Serine protease inhibitor

Abbreviations

Serpin	Serine protease inhibitors
ATIII	Antithrombin
HCFII	Heparin cofactor II
PAI	Plasminogen activator inhibitor
PC1	Protein C1-inhibitor
ADT-Vina	Autodock tools vina
CASTp	Computed atlas of surface topography of proteins
RCL	Reactive center loop
TMAO	Trimethylamine N-oxide

Introduction

Serine protease inhibitors (serpins) like neuroserpin, antithrombin, α -1antitrypsin, α -antichymotrypsin and plasminogen activator inhibitor are a unique superfamily

P. Singh · M. S. Khan · A. Naseem · M. A. Jairajpuri (✉)
Protein Conformation and Enzymology Lab,
Department of Biosciences, Jamia Millia Islamia University,
New-Delhi 110025, India
e-mail: m_jairajpuri.bi@jmi.ac.in

of protease inhibitors that are involved in important biological processes like blood coagulation, fibrinolysis, inflammation, cell migration and complement activation [1, 2]. Serpins have a common secondary fold, which is defined by at least 30% sequence identity and constitutes about 7–9 α -helices and three β sheets. An exposed reactive center loop (RCL) is an important structural component of serpin that is needed for targeting and inhibition of cognate serine proteases [3, 4]. Serpin undergoes a remarkable conformational transition on interactions with proteases, which is translocated to more than 70 Å away on the opposite site rendering it functionless by distorting the geometry of serine protease catalytic triad and conversion of the reactive center loop to strand 4A of β -sheet A [5]. This significant transition during the course of inhibition through the shutter region and breach region is also the cause of many serpin based polymerization defects. Naturally occurring variants of serpin forms the basis of several familial heredity diseases due to conformational deformation linked polymerization [1]. The majority of serpinopathy-linked mutations cluster in the center of the serpin molecule, underneath β -sheet A, in a region termed as shutter [6]. This portion of the molecule is the point of initial RCL insertion. It is suggested that destabilization of β -sheet A in either the shutter or the breach is sufficient to favor the transition to a polymeric or latent state over maintenance of the monomeric metastable native state [7]. Serpin polymerization is postulated to occur via a domain-swapping event whereby the RCL of one molecule docks into β -sheet A of another to form an inactive long-chain serpin polymer [8]. In addition to promoting polymerization, several serpin mutations have been identified that promote formation of a disease-linked latent state [9].

Serpin polymerization is a significant problem and devising a cure has been cumbersome owing to their complex mechanism of inhibition, metastable nature, cofactor binding ability and large scale conformational change. A reactive center loop peptide based approach has been successful in retarding the polymer growth however it also renders the serpin benign for further inhibitory use. Chemical chaperones such as glycerol and trimethylamine N-oxide (TMAO) mediate increase secretion of mutant α -1 antitrypsin and act as effective pharmacological strategy for prevention of liver injury and emphysema. Phenyl butyric acid (PBA) was also shown to have a similar effect on secretion of α -1 antitrypsin [10]. Previous studies have shown that glycerol is able to bind β sheet A of antithrombin [11] and increases the secretion of Z α -1 antitrypsin [10]. Glycerol, erythritol and trehalose (a disaccharide) reduce the rate of polymerization of wild and mutant type neuroserpin [12]. Overall chemical chaperone seems to be an attractive option because it can be

administered orally, can cross blood brain barriers and restore proper trafficking to the lysosome and dissociates. However, chemical chaperones require high concentrations for effective folding of mutant proteins and might be toxic in in-vivo applications.

It was shown recently that filling of cavities around strand 2A of β -sheet A might retard polymerization in antitrypsin [13]. Comprehensive analysis of cavities in serpin will reveal the dimension of cavities in areas involved in conformational change, cofactor binding and inhibition. Especially important will be to detect variation in cavity size in the process of protease inhibition. Binding small osmolyte to either retard polymerization or to enhance inhibitory activity of serpins hold promise, but it is also important to know the nature of such interactions. In this study we first did a comprehensive cavity based analysis of different conformational states of various serpins. Next docking and experimental studies confirmed that indeed chemical chaperone bind effectively in the shutter and helix D region to retard polymerization with minimum loss of activity.

Methodology

Materials

Hi-Trap heparin column was from GE Biosciences. Amicon Ultra-15 centrifugal filters (M_r 30,000 cutoff) were used for buffer exchange and concentration of protein solutions. Human thrombin and S-2238 were from American Diagnostic. All the other chemicals were purchased either from Sigma or Merck.

Cavity analysis using CASTp

CASTp was used to study surface features, functional regions and roles of important residues of different serpin conformations like native, latent and cleaved. It also gives an interactive visualization of computed pockets [14]. In CASTp, a pocket, which is a local spatial surface pattern, is regarded as an empty concavity on a protein surface into which solvent can gain access. The pockets were obtained by a geometric computation method, which can capture the physicochemical texture and the shape of a surface around functional residues, from the protein structures in PDB [15]. CASTp 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. It measures analytically the area and volume of each pocket and cavity, both in solvent accessible surface (SA, Richards' surface) and molecular surface (MS, Connolly's surface) [15]. All

hetero atoms treated as ligand are automatically removed from calculation, which includes solvent water molecules.

Molecular docking studies

Autodock Vina was used for molecular docking and virtual screening of chemical chaperone binding to serpins [16]. Autodock Vina is a newly developed program for molecular docking and virtual screening. It achieves an approximately two orders of magnitude speed-up in comparison with the molecular docking software AutoDock4.0 [17] and can achieve significantly improved accuracy of the binding mode predictions. The program automatically calculates the grid maps and clusters and uses a sophisticated gradient optimization method in its local optimization. In our present work we took 27 ligands like amino acid, carbohydrate and methylamines that were collected from PubChem database. Ligands like proline, arabitol, taurine and γ -amino butyric acid were collected from NCBI (PubChem substances) and their coordinate files were generated using Online Smile Translator. Polar hydrogen was added and Kollman charges were assigned to all atoms. Affinity grids were centered on and encompassing the active site were calculated with 0.375\AA spacing. Gasteiger charges were assigned to all atoms and rotatable bonds were assigned using AutoDock tools. Autodock was used to evaluate ligand binding energies over the conformational search space using Lamarckian genetic algorithm. Default docking parameters were used with some exceptions. In the output log file, we have considered the minimum energy conformation state of each ligand showing binding affinity in kcal mol^{-1} . RMSD values are calculated relative to the best mode and use only movable heavy atoms. Finally images of ligand and serpin bound complexes were prepared in PyMOL program and polar contacts between them were noted down.

Purification of antithrombin from human plasma

Large-scale purification of antithrombin from human plasma was achieved by using Hi-Trap heparin affinity column which was eluted with a 300 ml 0.15 – 2.50 M NaCl gradient [18, 19]. A Biorad Econopac integrated protein purification unit was used to achieve the purification. Human plasma was obtained from Rotary Blood Bank (New-Delhi) and kept under freezing condition until it was used for purification. 100 ml of human plasma was diluted 1:1 with 20 mM phosphate buffer containing 100 mM NaCl, 0.1 mM EDTA (PNE) and having a pH of 7.4 and ionic strength of 0.15. 2% sodium azide was added to the diluted plasma to avoid bacterial growth. Diluted solution was filtered on a 0.22 μm filter under cold condition and loaded on to a 5 ml Hi-trap heparin column equilibrated with PNE. After washing with 5 column volume of column

with PNE buffer, protein was eluted with 0.15 M, 0.25 M, 0.50 M, 0.75 M, 1 M, 1.25 M 1.50 M, 1.75, 2 M, 2.25 M and 2.5 M NaCl 1x-PNE gradient. After, elution fractions containing single band of purified ATIII with thrombin inhibitory activity were pooled, concentrated and buffer exchanged in tangential flow Amicon Ultra-15 centrifugal concentrators having a 30,000 molecular weight cut-off. SDS-PAGE was used to assess the purity of the ATIII. Concentration of purified ATIII was determined from absorbance at 280 nm using molar extinction coefficient of plasma ATIII.

Conditions that induce polymerization

Long chain polymer of ATIII was prepared by heating under specific buffer and pH condition. $100\ \mu\text{g ml}^{-1}$ each of native antithrombin in total of 1 ml was incubated at $60\ ^\circ\text{C}$ in 50 mM Tris buffer and 50 mM KCL, 40 % glycerol at pH 6.0 in the absence and presence of chemical chaperone at different time interval. Aliquots were removed and rapidly added to the ice-cold non-denaturing loading buffer and analyzed for native PAGE. Any hindrance/prevention of polymerization process by chemical chaperones can be detected by the combination of providing the above conditions in which polymer can form. The decrease in polymer formation and increased intensity of the monomeric band was detected using Native PAGE.

Kinetics of polymer transition

$100\ \mu\text{g ml}^{-1}$ each of native antithrombin in total of 1 ml was incubated at $60\ ^\circ\text{C}$ in 50 mM Tris buffer and 50 mM KCL, pH 6.0, in the absence and presence of chemical chaperone at different time interval. Samples were removed at indicated times and snap frozen and stored at $-70\ ^\circ\text{C}$. These aliquots were assayed for thrombin progressive activity (in PNE buffer) to assess the loss of ATIII inhibitory activity due to transition to polymeric ATIII with time. Reaction for the measurements of activity was set up under pseudo first order condition and contained ATIII and thrombin in a 10:1 ratio. ATIII and thrombin were reacted in microplates, and following the E+I incubations, S-2238 substrate was added and measured at 405 nm. Appropriate thrombin and S-2238 controls with chemical chaperone in the absence of antithrombin were taken.

Results and discussion

Analysis of cavities in the native state of serpin

Native states of serpin like antitrypsin, neuroserpin, antithrombin, antichymotrypsin, plasminogen activator in-

hibitor and PCI were analyzed for the cavities to find their area and volume and to assess if they are part of functionally and structurally important regions of protein. The results are shown in Fig. 1, which includes information about residues that are part of the largest cavity. Analysis of the native state showed that with the exception of antithrombin the number of cavities in each serpin were consistent. Largest cavity in the native state is centered around the shutter region in most of the serpin with the

exception of antithrombin. Largest cavity in antithrombin was around helix D in a region that transforms the conformation change to the reactive center loop on account of heparin binding. Volume of the largest cavity varied in each serpin where protein C inhibitor cavity has a volume of 1603 \AA^3 as compared to a volume of 283 \AA^3 in the antichymotrypsin. Antitrypsin had a total of 66 cavities where the volume of the largest cavity was 526 \AA^3 and it has an area of about 459 \AA^2 . This cavity predominantly

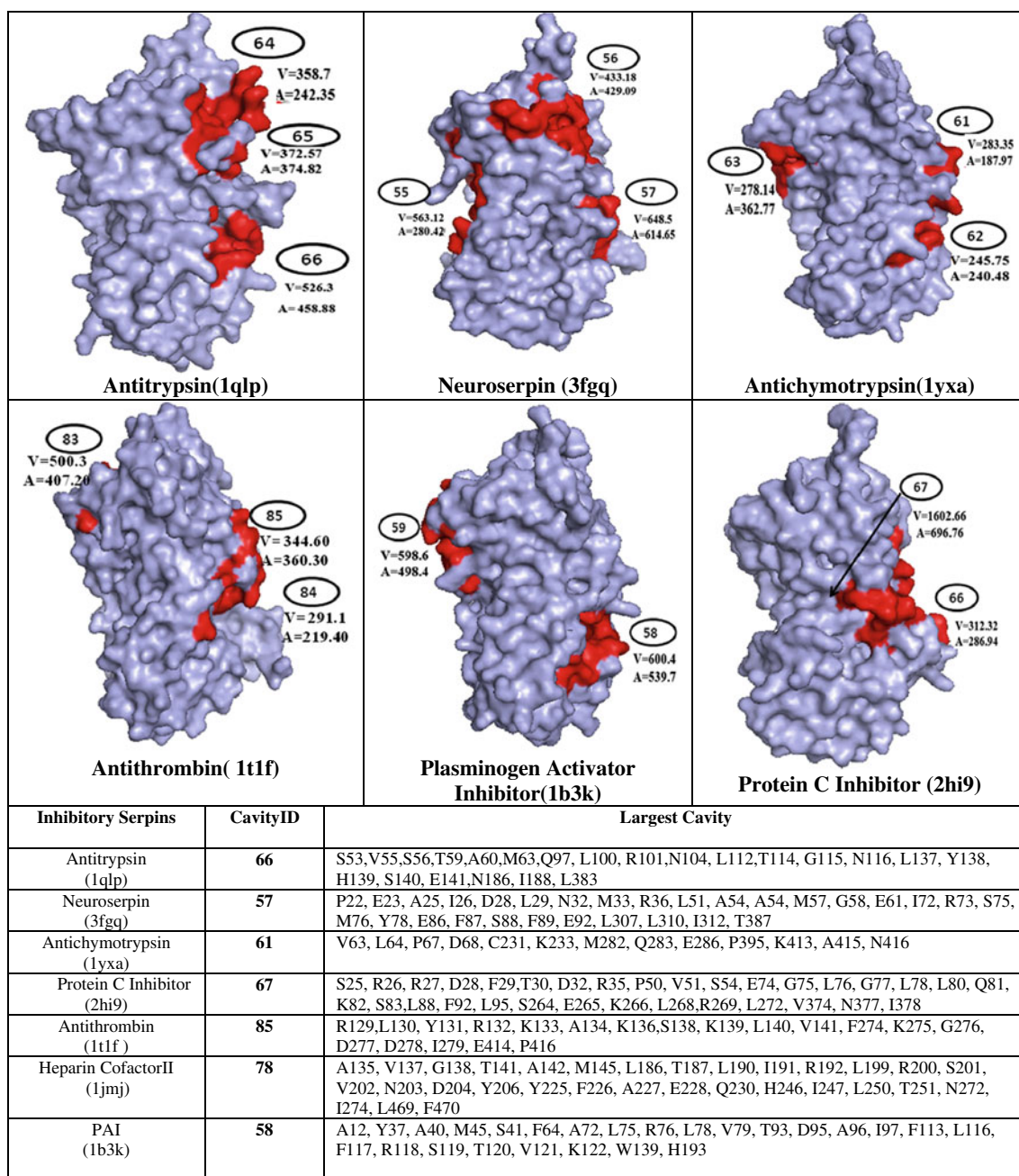


Fig. 1 Cavities were identified at a cutoff of 250 \AA^3 in different native structures of serpins using CASTp program. Amino acids that are part of a particular cavity were also identified. The cavity

identification numbers are arbitrary; they are numbered according to the largest to the smallest consecutively as identified by the program. Arrow indicates the cavity at the backside

included residues, which are part of helix B (Ser53, Val55, Ser56 Thr59 and Ala60), along with residues at N-terminal and C terminal end of helix D. Shutter region contains F-helix, B-helix and parts of strand s2A, s3A, s5A and s6A of β -sheet A, it plays an important role in stability and function of serpins [1, 20, 21]. It is interesting to note that an overwhelming majority of the polymerization variants of serpin are part of the shutter region. Shutter region mutations in α 1-antichymotrypsin (Leu55Pro), α 1-antitrypsin (Phe51Leu, Ser53Phe and Val55Pro), protein C1-inhibitor (Ser52Phe and Ser54Leu), antithrombin (Pro80Ser/Thr, Thr85Met/Lys, Cys95Arg and Leu99Phe) and neuroserpin (Ser56Arg and Ser52Arg) are linked to several pathological conditions due to polymerization [22–24]. It is quite likely that increased polymerization propensity of the serpin shutter region variants is due to their presence in the large cavity in an area that is involved in conformational change. Interestingly as shown in Fig. 1 several polymerization variants are in or around the largest cavity in the shutter region. Indicating that cavity size and its variations may have a critical role to play in the serpin inhibition and polymerization mechanism. Indeed increase in size of a cavity by introduction of bulkier group (by mutation) in antitrypsin was shown to retard the polymerization [25].

Comparison of cavities in different conformational states of serpin

We did cavity analysis in different states of serpins like native, latent and cleaved conformations and the analysis is shown in Table 1. The analysis showed that the native to cleaved transition leads to an increase in the size of the largest cavity in almost all the serpins with the exception of antithrombin. Antitrypsin showed an increase of 10 folds where the size of the largest cavity increased from 526 Å³ in native to a volume of 5339 Å³ in the cleaved state. Similarly neuroserpin largest cavity size was increased from 649 Å³ in native to 5316 Å³ in the cleaved conformation. In the latent structure of antitrypsin the biggest cavity had a volume (701 Å³) which is slightly greater than the native structure. Latent state is a loop inserted state where the reactive center loop inserts as s4A without cleavage, our analysis did not show large variation in the cavity volume of latent as compared to the native. This indicates that the loop insertion mechanism in latent is different from that of the cleaved and role of cavity in latent loop insertion might be limited. Surface cavity contributes to metastability of antitrypsin and cavities near the β -sheet A have been shown to be important in regulating the inhibitory activity [13, 26]. It is possible

Table 1 CASTp analysis of Connolly surface area and volume of the largest cavities in different conformational states of antitrypsin: Table shows the cavities which were identified (at the cutoff of 250 Å³ in different monomeric native, cleaved and latent crystal structures of inhibitory serpins

Serpins	Native		Cleaved		Latent	
	Cavity ID ^b	Cavity Volume ^a	Cavity ID	Cavity Volume	Cavity ID	Cavity Volume
Antitrypsin	66	459	55	5339	58	701
	65	375	54	480	57	627
	64	242	52	379	56	438
Neuroserpin	57	649	50	5316	–	–
	56	433	–	–	–	–
	54	563	–	–	–	–
Antichymotrypsin	63	278	37	4578	–	–
	62	246	–	–	–	–
	61	283	–	–	–	–
Antithrombin	85	500	62	303	60	508
	84	345	–	–	59	432
	83	291	–	–	61	457
	–	–	–	–	82	317
Plasminogen activator inhibitor	59	599	51	3450	77	401
	58	600	–	–	74	359
	–	–	–	–	75	292
Protein C1-inhibitor	–	–	–	–	73	254
	67	1602	51	3450	63	710
	66	312	49	273	62	542
–	–	–	–	60	452	

^a Table shows the volume of the three biggest cavities in different conformational states of serpin

^b The cavity identification numbers are arbitrary; they are numbered according to the largest to the smallest consecutively as identified by the CASTp program

that changes in the cavity volume may be part of the RCL translocation mechanism to thermodynamically trap in loop inserted conformation during inhibition. Natural variants can introduce local destabilization in these cavities which might make this area polymerization prone. Destabilized residues inside surface cavities may be stabilized by targeting with small molecule which can counter local deformation and help reduce polymerization. However targeting molecules to specific cavity will be difficult especially since the cross specificity may affect the functional properties of serpin.

Docking of chemical chaperone to serpin

We choose carbohydrate, methylamine and amino acid based chemical chaperone to target serpins with an aim of determining the binding energies, interacting residues and also if they are part of cavity. The binding energy computed using Autodock is shown in Table 2 and represents the binding affinities of 27 chemical chaperone to serpin. The results show that most of the chemical

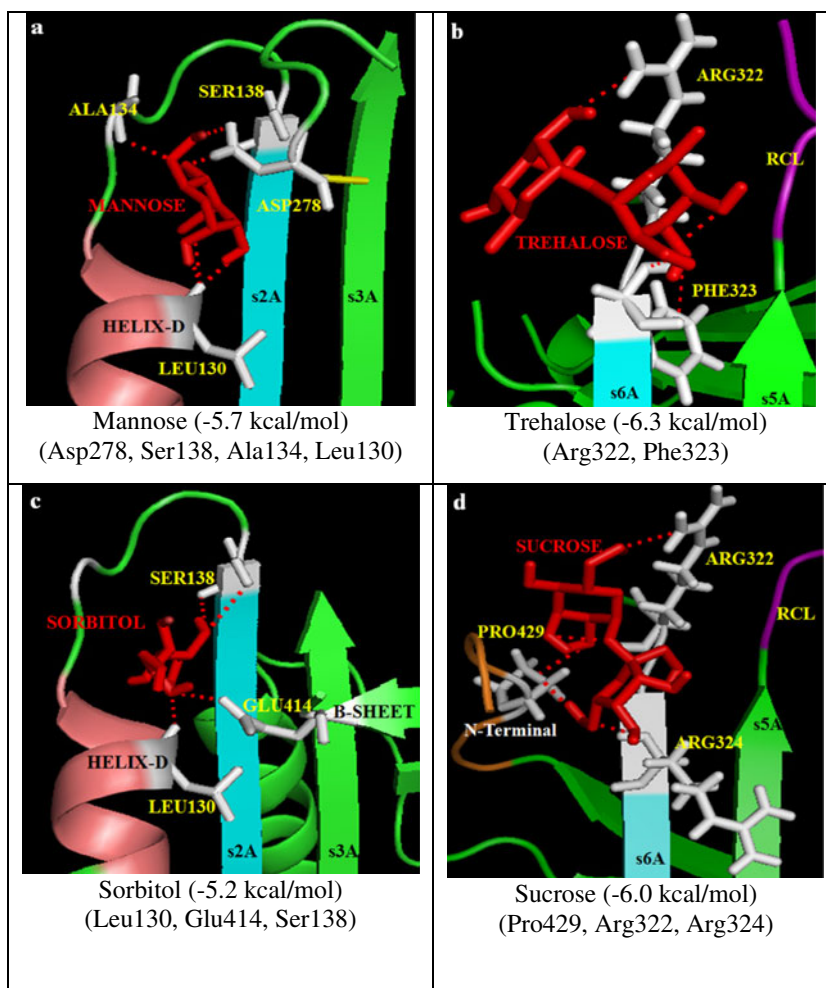
chaperone interacted with residues in the shutter region and the helix D arm at the C-terminal end. The data indicates a wide array of binding energies, more specifically our result shows that in native state most of the chaperones show a preference of binding in the cavity around shutter region whereas few like taurine, serine, glycerol, mannitol, TMAO seems to prefer cavities around F and D helix. The results show that carbohydrate based chemical chaperone like sorbitol, sucrose, arabitol and trehalose, and amino acid based chaperones like dopamine, phenylalanine, arginine, glutamic acid are most effective in binding serpins. Figure 2 shows the interaction of mannose, trehalose, sorbitol and sucrose with antithrombin. The results show that mannose interacts with Asp278 (strand 2A), Leu130 (helix D) and Ala134 (helix D), whereas sorbitol interacts with Glu414 (strand 3B), Leu130 (helix D) and Ser138 (strand 2A). Interestingly all these residues are part of the largest cavity in the antithrombin as shown in Fig. 1. Trehalose showed the best binding energy among all the chemical chaperones analyzed in our study. Trehalose forms hydrogen bond

Table 2 Binding affinity of chemical chaperones to native serpin serpin like antithrombin, antitrypsin, neuroserpin, anti-chymotrypsin, HCFII, PAI and PC1-Inhibitor. Predicted free energies of binding (kcal mol^{-1}) of antithrombin with corresponding chemical chaperone were calculated using Autodock Vina

S.NO.	Chaperones	ATIII	ANT	NEU	ACH	PC1-I	HCFII	PAI
1	<i>Alanine</i>	-3.9	-3.9	-4	-3.7	-3.7	-3.8	-4
2	<i>Arginine</i>	-5.4	-5.4	-5	-5.9	-5.3	-5.5	-5
3	<i>Betaine</i>	-3.8	-4	-3.8	-3.5	-4	-3.7	-3.8
4	<i>Dopamine</i>	-5.9	-6.2	-6.2	-5.7	-6.2	-6.4	-6.2
5	<i>Gaba</i>	-3.9	-4	-4.4	-3.9	-4	-4	-4.4
6	<i>Glutamic-acid</i>	-5	-5	-5.8	-4.6	-4.9	-5	-5.8
7	<i>Glycine</i>	-3.5	-3.6	-3.5	-3.6	-3.5	-3.6	-3.5
8	<i>Isoleucine</i>	-4.8	-4.8	-4.8	-4.4	-4.5	-5	-4.8
9	<i>Lysine</i>	-4.7	-4.8	-4.7	-4.1	-4.7	-4.7	-4.7
10	<i>Phenylalanine</i>	-5.4	-5.7	-5.3	-5.2	-5.6	-6	-5.3
11	<i>Proline</i>	-4.4	-4.6	-4.8	-4.6	-4.5	-4.6	-4.8
12	<i>Serine</i>	-4.1	-4.1	-4.1	-4.1	-4.2	-4	-4.1
13	<i>Taurine</i>	-4	-4	-3.8	-4.1	-3.9	-3.8	-3.8
14	<i>Threonine</i>	-4.5	-4.4	-4.4	-4.2	-4.7	-4.3	-4.4
15	<i>Tyrosine</i>	-5.6	-6	-6.1	-5.4	-5.9	-6.2	-6.1
16	<i>Valine</i>	-4.4	-4.5	-4.5	-4.5	-4.3	-4.6	-4.5
17	<i>Arabitol</i>	-5.1	-4.8	-4.9	-4.9	-4.4	-4.7	-4.9
18	<i>Erythritol</i>	-4.3	-4.2	-4.5	-4.2	-4.1	-4.4	-4.5
19	<i>Glycerol</i>	-3.8	-4.1	-4	-4	-3.9	-3.6	-4
20	<i>Sorbitol</i>	-5.2	-5	-5.5	-4.5	-5.1	-4.9	-5.5
21	<i>Sucrose</i>	-6.0	-5.9	-6.1	-6.0	-6.5	-5.8	-6.1
22	<i>Mannitol</i>	-4.8	-5	-4.7	-4.7	-5.2	-4.7	-4.7
23	<i>Mannose</i>	-5.7	-5.2	-5.6	-4.1	-5.3	-5.3	-5.6
24	<i>Trehalose</i>	-6.3	-6.6	-6.4	-6.1	-3	-6.6	-6.4
25	<i>Glycerophosphocholine</i>	-5.1	-4.6	-5.1	-5.1	-5.2	-4.6	-5.1
26	<i>Sarcosine</i>	-3.8	-3.7	-3.9	-3.8	-3.7	-3.9	-3.9
27	<i>Trimethylamine N-oxide</i>	-3	-3	-2.9	-2.8	-3	-3	-2.9

Fig. 2 Binding affinity and hydrogen bond interactions of mannose, trehalose, sorbitol and sucrose with native monomeric antithrombin.

Structures of antithrombin native state (1TIF) showing the autodock analysis with mannose, trehalose, sorbitol and sucrose. AutodockVina analysis and hydrogen bond distances were calculated as detailed in *Materials and methods*



S.No.	Chemical Chaperone	Residues with polar contacts with chaperones	Interaction type	Distance (Å)
1.	Mannose	Asp278 (s2B)	Side chain	3.14
		Leu130 (Helix D)	Main chain	2.70
		Ala134 (Upper edge of Helix D)	Main chain	3.13
		Ser138 (s2A)	Main chain	3.13
2.	Trehalose	Arg322 (at the edge of s6A)	Side chain	3.11
		Phe323(s6A)	Main chain	3.02
3.	Sorbitol	Glu414 (s3B)	Side chain	2.82
		Leu130 (Helix D)	Main chain	2.97
		Ser138 (s2A)	Main chain, Side chain	2.88, 3.14
4.	Sucrose	Pro429 (N terminal)	Main chain, Side chain	2.82, 3.07
		Arg322 (at the edge of s6A)	Side chain	3.09
		Arg324 (s6A)	Main chain	3.04

interaction with Phe321 and Arg322 (strand 6A), sucrose formed hydrogen bond interaction with Arg322 and Arg324 of strand 6A and Pro429 at the N-terminal end. These residues are at the upper portion of the shutter region very near to the region where RCL inserts as strand 4A during inhibition. Helix D is involved in transformation of the conformational

change in antithrombin for full exposure of RCL on account of heparin binding [27]. Interaction of carbohydrate based chemical chaperone in areas important for the translocation of conformational change and inhibitory mechanism allowed us to test these chemical chaperone for reducing polymerization and its effect on inhibition.

Effect of chemical chaperones on the rate of polymerization

We purified antithrombin from human plasma as detailed in the *Materials and methods* section and conditions were provided to the purified ATIII to induce polymer formation in the absence and presence of chemical chaperone at different time intervals. Figure 3 summarizes the results showing screening of chemical chaperone with antithrombin. Figure 3a with native antithrombin without chemical chaperone showed that antithrombin band is clearly visible at the 0 time period point. However as soon as the heating is done polymer bands start to appear and there is a progressive increase in the strength of the polymerization band with increase in the incubation temperature. Under the same condition at appropriate concentration of sucrose, mannose, sorbitol and trehalose (Figs. 3b to e), we observed a single band at 5 min interval, and that high molecular weight polymer bands have almost completely disappeared due to hindrance in the process of polymerization.

Kinetics of polymer transition in the presence and absence of chemical chaperone was assessed under the condition that forms polymers. It is clear from the graph (Fig. 3f) of residual antithrombin activity *versus* time that

single band observed in the polymer transition experiments shows appreciable antithrombin inhibitory activity. Native antithrombin in the absence of chemical chaperone almost completely loses the ability to inhibit thrombin when incubated for 20 min in polymerization condition. Sorbitol, trehalose and sucrose when incubated with antithrombin maintained its native inhibitory activity even at 90 min in polymerization condition, but in the presence of mannose antithrombin inhibitory activity was reduced by 40 %. These results are a clear indication that retardation of polymerization leads to increased inhibitory activity when incubated with chemical chaperone. Isolating a lead compound that can effectively bind serpins can provide a structural scaffold that may be used for designing organic compounds that can effectively hinder polymerization without modulating the inhibition rates and cofactor binding abilities.

Conclusions

Previous evidence suggests that chemical chaperones can be promising in reducing the rate of polymerization in

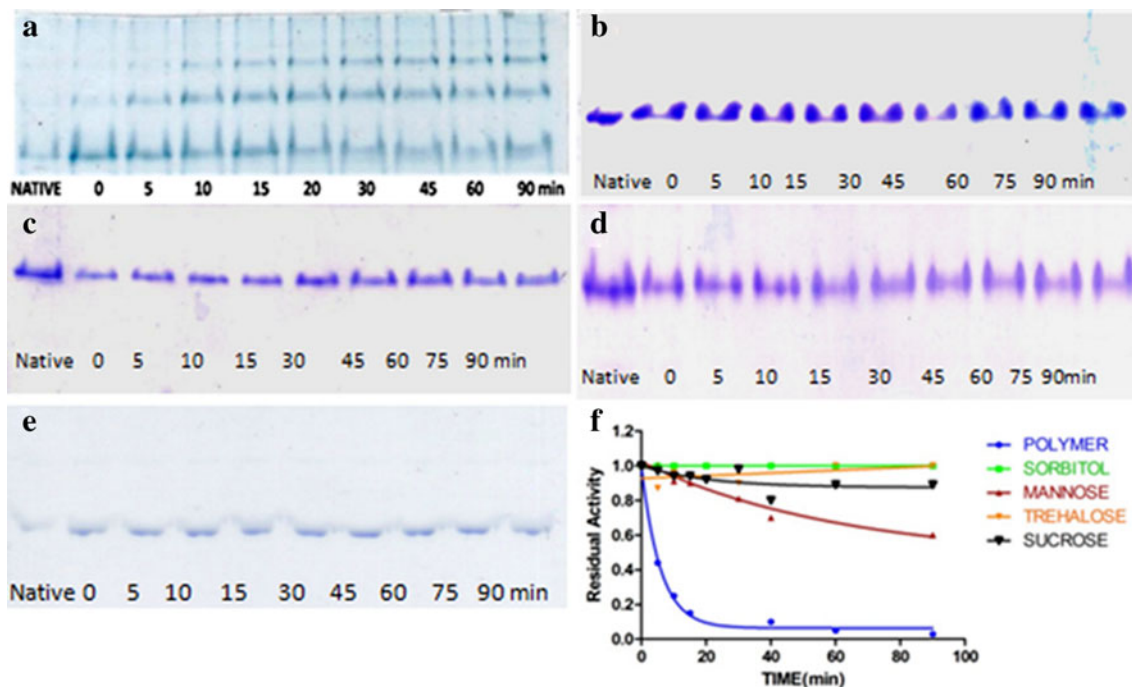


Fig. 3 Polymer formation in the presence and absence of chemical chaperones: Polymer formation in the native antithrombin in the presence and absence of chemical chaperone was determined by heating at 2 μ M in 50 mM Tris buffer, pH 6.0 at 60 $^{\circ}$ C for the time shown in the figures. Aliquots were removed and frozen prior to assessment on non-denaturing PAGE. Each gel picture has a native control along with the 0 min where the aliquot was withdrawn immediately after incubation at 60 $^{\circ}$ C. (a) Represents 2 μ M native antithrombin without chemical chaperone. (b) 1.5 M sorbitol, (c)

1.0 M trehalose, (d) 1.5 M mannose and (e) 1.25 M sucrose were incubated with native antithrombin. Rate of loss of inhibitory activity of antithrombin was measured under the same conditions as described in *Material and methods* section. Native antithrombin was completely inactive at 20–25 min, whereas under the same conditions, in the presence of sorbitol, trehalose and sucrose antithrombin was 97% active at 90 min incubation, in the presence of mannose it was 60% active. Each graph was average of two independent experiments

serpins but its effect on the structure function remains largely unknown. In the present work we have studied the surface cavities to identify targets for chemical chaperone. A shutter region cavity which invariably is the largest cavity in the native state of many serpins may be the ideal target to block polymerization. Carbohydrate based chemical chaperones seems to be most effective in binding many serpin with high affinity and reduces polymerization without affecting the inhibition rates.

Acknowledgments This research was supported by grants from Department of Biotechnology and University Grant Commission, Government of India. PS is supported by a grant from Rajiv Gandhi National Fellowship. AN is supported by an Innovation in Science Pursuit for Inspired Research (INSPIRE) fellowship from Department of Science and Technology, Government of India. SK is supported by a fellowship from University Grant Commission.

References

- Devlin GL, Bottomley SP (2005) A protein family under 'stress'-serpin stability, folding and misfolding. *Front Biosci* 10:288–299
- Stein PE, Carrell RW (1995) What do dysfunctional serpins tell us about molecular mobility and disease? *Struct Biol* 2:96–113
- Huntington JA, Read RJ, Carrell RW (2000) Structure of a serpin-protease complex shows inhibition by deformation. *Nature* 407:923–926
- Mast AE, Enghild JJ, Pizzo S, Salvesen G (1991) Analysis of the plasma elimination kinetics and conformational stabilities of native, proteinase-complexed, and reactive site cleaved serpins: comparison of alpha 1-proteinase inhibitor, alpha 1-antichymotrypsin, antithrombin III, alpha 2-antiplasmin, angiotensinogen and ovalbumin. *Biochemistry* 30:1723–1730
- Silverman GA, Bird PI, Carrell RW, Church FC, Coughlin PB, Gettins PG, Irving JA, Lomas DA, Luke CJ, Moyer RW, Pemberton PA, Remold-O'Donnell E, Salvesen GS, Travis J, Whisstock JC (2001) The serpins are an expanding superfamily of structurally similar but functionally diverse proteins. *J Biol Chem* 276:33293–33296
- Carrell RW, Tavis J (1985) Alpha 1 antitrypsin and the serpins variation and counter variation. *Trends Biochem Sci* 8:20–24
- Lomas DA, Evans DL, Finch JT, Carrell RW (1992) The mechanism of Z α -1 antitrypsin accumulation in the liver. *Nature* 357:605–607
- Mast AE, Enghild JJ, Salvesen G (1992) Conformation of the reactive site loop of alpha 1-proteinase inhibitor probed by limited proteolysis. *Biochemistry* 31:2720–2728
- Beauchamp NJ, Pike RN, Daly M, Butler L, Makris M, Dafforn TR, Zhou A, Fitton HL, Preston FE, Peake IR, Carrell RW (1998) Antithrombin Wibble and Wobble (T85M/K): archetypal conformational diseases with in vivo latent-transition, thrombosis, and heparin activation. *Blood* 92:2696–706
- Burrows JA, Willis LK, Perlmutter DH (2000) Chemical chaperones mediate increased secretion of mutant α -1 antitrypsin (a1-AT) Z: a potential pharmacological strategy for prevention of liver injury and emphysema in α -1 antitrypsin deficiency. *Proc Natl Acad Sci* 97:1796–1801
- Zhou A, Stein PE, Huntington JA, Carrell RW (2003) Serpin polymerisation is prevented by a hydrogen-bond network which is centred on His 334 and stabilised by glycerol. *J Biol Chem* 278:15116–15122
- Sharp LK, Mallya M, Kinghorn KJ, Wang Z, Crowther DC, Huntington JA, Belorgey D, Lomas DA (2006) Sugar and alcohol molecules provide a therapeutic strategy for the serpinopathies that cause dementia and cirrhosis. *FEBS J* 273:2540–2552
- Lee C, Maeng JS, Kocher JP, Lee B, Yu MH (2001) Cavities of α -1 antitrypsin that play structural and functional roles. *Protein Sci* 10:1446–1453
- Binkowski TA, Naghibzadeh S, Liang J (2003) CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res* 31:3352–3355
- Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S (1998) Analytical shape computing of macromolecules I: molecular area and volume through alpha shape. *Proteins* 33:1–17
- Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31:455–461
- 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
- Jairajpuri MA, Lu A, Desai U, Olson ST, Bjork I, Bock SC (2003) Antithrombin III phenylalanines 122 and 121 contribute to its high affinity for heparin and its conformational activation. *J Biol Chem* 278:15941–15950
- Jairajpuri MA, Lu A, Bock SC (2002) Elimination of P1 arginine 393 interactions with underlying glutamic acid 255 partially activates antithrombin III for thrombin inhibition but not factor Xa inhibition. *J Biol Chem* 277:24460–24465
- Krem MM, Di Cera E (2003) Conserved Ser residues, the Shutter Region, and Speciation in Serpin Evolution. *J Biol Chem* 278:3810–3814
- Lee KN, Park SD, Yu MH (1996) Probing the native strain in alpha1-antitrypsin. *Nat Struct Biol* 3:497–500
- Potempa J, Korzus E, Travis J (1994) The serpin superfamily of proteinase inhibitors: structure, function and regulation. *J Mol Biol* 269:15957–1560
- Whisstock JC, Bottomley SP (2006) Molecular gymnastics: serpin structure, folding and misfolding. *Curr Opin Struct Biol* 16:761–768
- Gooptu B, Lomas DA (2009) Conformational pathology of the serpins: themes, variations, and therapeutic strategies. *Annu Rev Biochem* 78:147–176
- Parfrey H, Mahadeva R, Ravenhill NA, Zhou A, Dafforn TR, Foreman RC, Lomas DA (2003) Targeting a surface cavity of α 1-antitrypsin to prevent conformational disease. *J Biol Chem* 278:33060–33066
- Im H, Seo EJ, Yu MH (1999) Metastability in the inhibitory mechanism of human α -1 antitrypsin. *J Biol Chem* 274:11072–11077
- Belzar KJ, Zhou A, Carrell RW, Gettins PG, Huntington JA (2002) Helix D elongation and allosteric activation of antithrombin. *J Biol Chem* 277:8551–8558

Insight into the binding interactions of CYP450 aromatase inhibitors with their target enzyme: a combined molecular docking and molecular dynamics study

Roberta Galeazzi · Luca Massaccesi

Received: 22 December 2010 / Accepted: 31 May 2011 / Published online: 18 June 2011
© Springer-Verlag 2011

Abstract CYP450 aromatase catalyzes the terminal and rate-determining step in estrogen synthesis, the aromatization of androgens, and its inhibition is an efficient approach to treating estrogen-dependent breast cancer. Insight into the molecular basis of the interaction at the catalytic site between CYP450 aromatase inhibitors and the enzyme itself is required in order to design new and more active compounds. Hence, a combined molecular docking–molecular dynamics study was carried out to obtain the structure of the lowest energy association complexes of aromatase with some third-generation aromatase inhibitors (AIs) and with other novel synthesized letrozole-derived compounds which showed high in vitro activity. The results obtained clearly demonstrate the role of the pharmacophore groups present in the azaheterocyclic inhibitors (NSAIs)—namely the triazolic ring and highly functionalized aromatic moieties carrying H-bond donor or acceptor groups. In particular, it was pointed out that all of them can contribute to inhibition activity by interacting with residues of the catalytic cleft, but the amino acids involved are different for each compound, even if they belong to the same class. Furthermore, the azaheterocyclic group strongly coordinates with the Fe(II) of heme cysteinate in the most active NSAI complexes, while it prefers to adopt another orientation in less active ones.

Keywords Molecular docking · Molecular dynamics · Aromatase inhibitors · Binding interactions

Introduction

Aromatase is a CYP450 enzyme involved in the production of estrogens that acts by catalyzing the conversion of testosterone (an androgen) to estradiol (an estrogen). Aromatase is located in estrogen-producing cells in the adrenal glands, ovaries, placenta, testicles, adipose (fat) tissue and brain. The growth of some breast cancers is promoted by estrogens as, upon the binding of an estrogen to the estrogen receptor (ER), the receptor activates the transcription of its target genes, which are responsible for cancer cell proliferation in estrogen-dependent breast tumors. Inhibiting aromatase is an efficient approach to treating estrogen-dependent breast cancer because the aromatization of androgen is the terminal and rate-determining step in estrogen synthesis [1, 2]. Aromatase inhibitors (AIs) can be classified in terms of both their structures and mechanisms of action. Two types of AIs can be distinguished: irreversible steroidal inhibitors such as exemestane, which forms a permanent bond with the aromatase enzyme complex, and nonsteroidal inhibitors (NSAIs)—mostly triazole derivatives such as anastrozole (Arimidex[®]) and letrozole (Femara[®])—that inhibit the enzyme by reversible competition [3–7]. These new-generation compounds appear to be better tolerated and produce fewer collateral effects than the commonly used tamoxifen. All of these compounds have been demonstrated to be very potent and specific, but until now the structural basis of drug recognition and association has not been completely elucidated, since the 3D structure of aromatase has only recently been reported in complex with its natural ligand androstenedione (ASD) [6] (pdb code: 3eqm). While some papers on the molecular basis of interactions between aromatase and some of its inhibitors have already been reported [8–14], most published docking studies deal with

R. Galeazzi (✉) · L. Massaccesi
Dipartimento I.S.A.C, Università Politecnica delle Marche,
via Breccia Bianche,
60131 Ancona, Italy
e-mail: r.galeazzi@univpm.it

the 3D structure modeled for homology [15–17] (PDB code 1tqa). Among these studies, only a few papers refer to the automated rigid docking of some steroid-based inhibitors to the crystallographic 3D structure of aromatase [10, 18]. Very recently, Roy et al. performed a 3D-QSAR study of various classes of CYP19 aromatase inhibitors [18]; that study employed rigid docking and did not minimize the energy of ligand–enzyme complexes. Other studies were based upon indirect approaches such as QSAR [14]. Detailed knowledge of the intermolecular interaction at the catalytic site between the inhibitors and the enzyme is essential if we are to better rationalize the strengths of known drugs, and it would provide the starting point for the proper rational design of new and more active compounds. With this in mind, we carried out a combined molecular docking/molecular dynamics study to model the molecular Michaelis complexes of aromatase with a number of currently used drugs, starting from the crystallized structure of the enzyme (3eqm). The compounds considered in that study—reported here—are some third-generation aromatase inhibitors (AIs), the steroidal inhibitor exemestane, the NSAIs letrozole, anastrozole, and other letrozole-derived compounds that show high in vitro activity [19]. The final goal was to compare their binding energies and their intermolecular interactions with their experimental inhibition activities, in order to identify those with the highest affinity for aromatase.

Computational details

Preparation of inhibitors and complexes of them with the enzyme

The association complexes described in the present article were constructed starting from the crystallographic structure of human aromatase bound to its natural ligand androstenedione (pdb code 3eqm) available at the Brookhaven Protein Data Bank. Missing hydrogens were added to the X-ray crystallographic structure using the CHIMERA software package [20]. The pdb structure was protonated assuming a pH of 7.4 and using the following physical conditions: salinity 0.15, internal dielectric 6, external dielectric 80 [21]. The crystallographic water molecules present were retained, as it has already been pointed out [22–25] that the hydroxylation mechanism (the first two reaction steps catalyzed by aromatase) directly involves the catalytic water molecules, the binding of which is promoted by dioxygen binding as in P450cam [26–28], yielding a C19-aldehyde derivative of androstenedione via 19,19-gem-diol formation with retention of the pro-S hydrogen. At this point, for the sake of clarity, we will focus on current knowledge of the human aromatase mechanism.

This enzyme is a cytochrome P-450 which functions in association with an NADPH-dependent reductase. The overall process of androgen to estrogen conversion consists of several steps (five), the first two of which (the hydroxylation of C-19) are “classical” cytochrome P-450 hydroxylations, as confirmed by the experimental work of Beunsen [29]. The details of the third step (leading to the aldehydic intermediate) and the last two (the real aromatization of the D ring) are currently unclear, even though an attempt to provide an explanation at the DFT level has been made [30] (a detailed QM/MM study of the last two steps of the aromatization reaction are currently being investigated by our group in collaboration with Prof. Bottoni’s group at Bologna University).

Keeping this knowledge in mind, the docking MD protocol to localize the Michaelis association complexes for ASD, exemestane and all the other NSAIs considered in this study was carried out with explicit water inside the catalytic cleft and without it, in order to assess the role of the solvent during the association process. As a result, for ASD, exemestane and anastrozole, no differences in the geometries of the complexes were found, demonstrating that water does not interfere with the association between the steroid ligand and aromatase itself; it is only directly involved in the subsequent reaction steps according to literature data [22–29]. In contrast, as will be discussed widely in the “Results” section, some differences were observed for letrozole and its derivatives when the docking was performed in the presence and absence of explicit water molecules inside the cleft.

Furthermore, for all of the MD simulations carried out in the absence of explicit water, the overall effect of the solvent was accounted for by using the GB/SA model for water ($\epsilon=78.5$) to emulate the solvent electrostatic effect [31]. We chose an implicit solvation model instead of an explicit one to reduce the computation time.

The 3D structures of the studied inhibitors were built using the Macromodel 5.5 software package [32], and were fully minimized using the AMBER force field (AMBER*) [33]. Subsequently, all of the ligands were placed inside the catalytic site, and the natural ligand ASD was substituted at random orientations and random torsion values, thus allowing a search for flexible conformations of the compounds during the docking process.

Molecular docking

The docking program Autodock 4.0 was used to perform the automated molecular docking [34–36].

A Lamarckian genetic algorithm (LGA) was applied to deal with the inhibitor–enzyme interactions. A grid map with $70 \times 70 \times 70$ points spaced equally at 0.375 \AA was generated using the Autogrid program to evaluate the

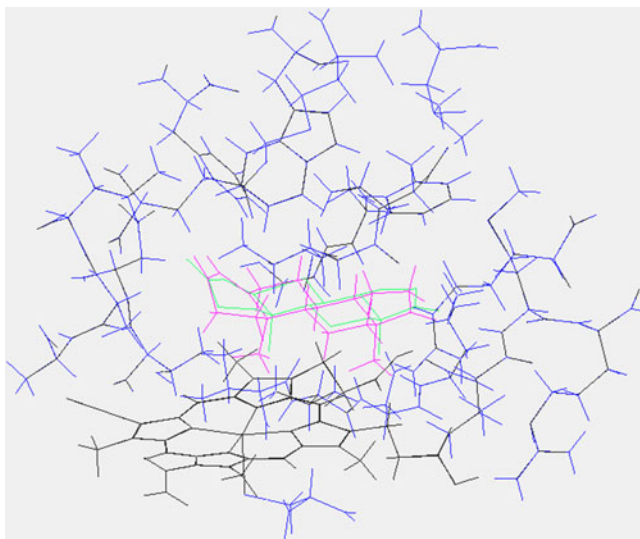


Fig. 1 Superimposition of the 3D crystallographic (3eqm) structure and the same molecular structure after MD simulation and minimization (the Michaelis complex). ASD in the original 3eqm is shown in *green*, while ASD in the Michaelis complex is shown in *pink*. The heme is shown in *black*. Only residues within 6.0 Å from ASD are displayed

binding energies between the compounds and the protein. Docking parameters were set to default, except for the number of GA runs (200), the energy evaluations (25,000,000), the maximum number of top individuals that automatically survive (0.1), and the step size for translation (0.2 Å). The docked inhibitor–enzyme complexes were ranked according to the predicted binding energy and arranged into clusters according to the RMS values. For each compound, the most representative structures of each cluster, which also included the lowest energy complex, were used as the starting point for further MD simulations. This allowed the binding pocket and the inhibitor to relax and arrange themselves into the best Michaelis association complex.

MD simulation protocol: reaching the final Michaelis complex

After the cluster analysis, the best representative structures for each substrate–inhibitor complex were then used for MD simulations as starting geometries after complete minimization performed by the AMBER force field [33], as implemented in MMOD 5.5 [32], in which the heme parameters were added according to literature data [37]. Minimization and MD simulations were carried out on a core of unconstrained atoms around the active site (8 Å), and on a shell of constrained atoms [energy penalty force constant of 100 kJ/(Å² mol⁻¹)] surrounding the core (6 Å). An initial minimization (2000 steps, steepest descent) and a subsequent constant-temperature MD simulation (2 ns,

298 K, 1.0 fs time step) were carried out. An equilibration time of 40 ps was allowed before data collection was initiated. The SHAKE algorithm was used to constrain stretching bonds involving hydrogen atoms [38]. The coordinates of the system were saved on a trajectory file every 10 ps, giving a total of 200 structures for further analysis. Each obtained structure was fully minimized first by steepest descent and then by conjugate gradient with a derivative convergence criterion of 0.05 kJ/(Å² mol⁻¹). The lowest-energy structure was considered to be representative of the Michaelis complex or the lowest-energy association complex. This molecular modeling protocol has already been shown to yield reliable results when studying other inhibitor complexes [39–43].

The computational effort described above was necessary to build a reasonable association complex, which is key information that cannot be obtained from crystallographic data, as it cannot reproduce the dynamics of the active site.

Results and discussion

Molecular interaction of the natural steroidal ligand androstenedione

Starting from the crystallographic 3D structure of aromatase in complex with its natural ligand androstenedione (ASD) (pdb code 3eqm), the structure of the initial Michaelis complex was modeled. A 2 ns MD simulation was performed at 298 K (see “Computational details”), which allowed the residues to relax and find their lowest-energy structural conformations along the PES. This computational protocol is necessary in order to remove the rigid constriction which occurs in the crystal, thus leading to a better optimized structure after the addition of hydrogens.

In the optimized molecular complex structure, a catalytic cleft was observed that appears to be particularly rich in both polar and apolar hydrophobic residues such as Ile305,

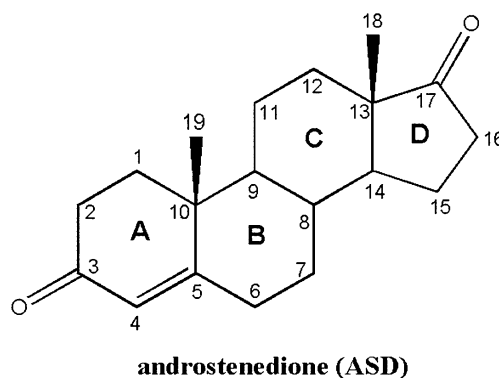


Fig. 2 Structure and atom numbering system of androstenedione (ASD)

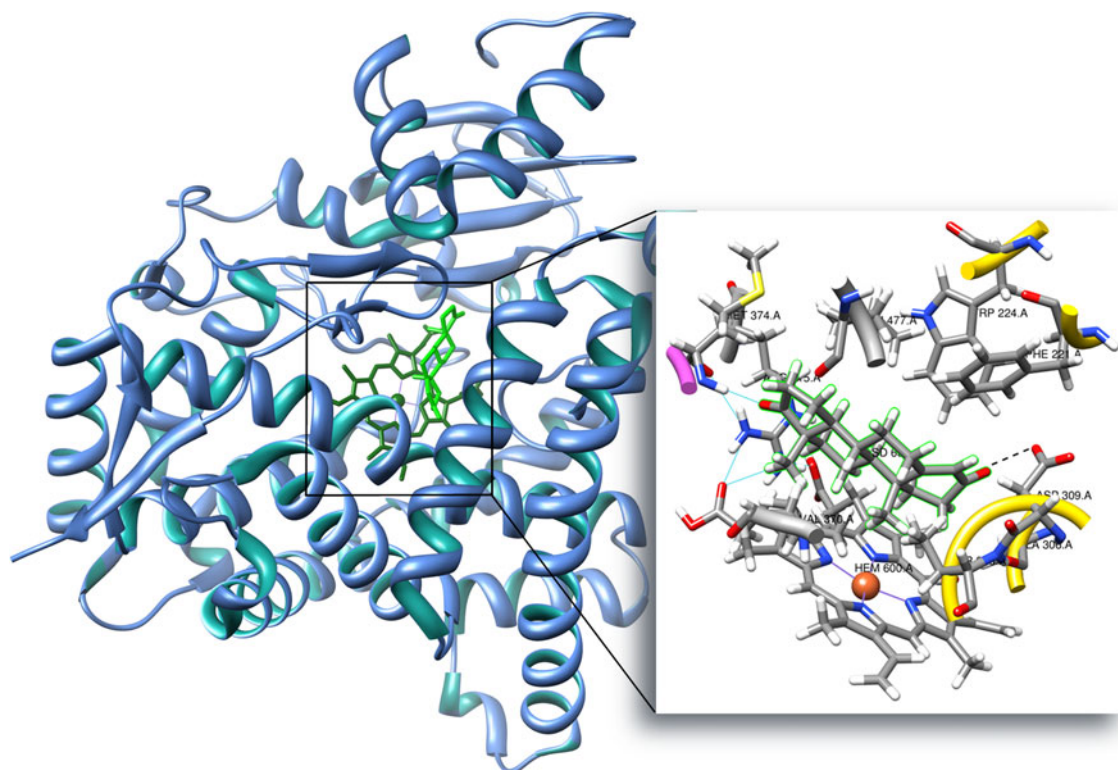


Fig. 3 Interactions with ASD at the catalytic site. The ligand is highlighted in *green*. Only those residues involved in ASD stabilization are shown

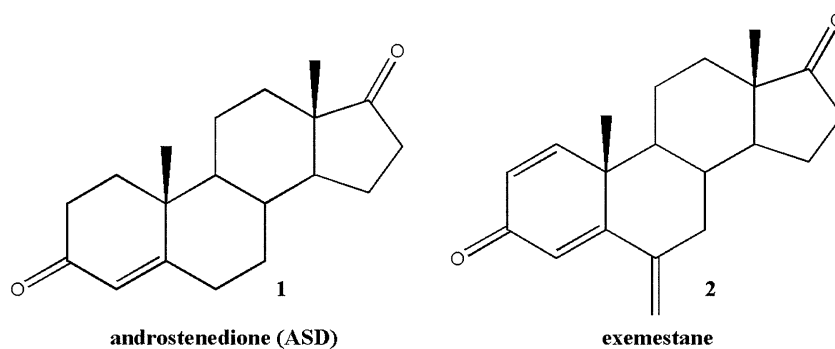
Ala306, Asp309, Arg115 (from the I helix), Phe221, Trp224 (from the F helix), Ile133, Phe134 (B_C loop), Val370, Leu372, Val373 (from the K helix–b3 loop), Met374 (from b3), Leu477 and Ser478 (from the b8–b9 loop). This amino acid assembly is particularly effective at stabilizing both polar and apolar ligands, particularly those containing aromatic moieties. Those findings mostly confirm the crystallographic observations of Gosh et al. [6], even though we noticed a difference in the interaction distances ascribed to both the relaxation of the lateral chains of the residues in the catalytic cleft and the relative position of the natural ligand (Fig. 1). It is worth mentioning that that after MD stabilization, the H-bond between the C-17 carbonyl oxygen (acceptor) and the Met374 backbone N–H (donor) strengthens ($d=1.82$ Å vs. 2.8 Å). All of these interactions

were observed in both the association complex containing water inside the catalytic cleft and that without it, due to the same orientation of ASD at the binding site.

This observation confirms that water does not interfere in the process of ligand association; it only becomes active in the subsequent reaction steps, as already reported in the literature [22–29]. Furthermore, the C-3 carbonylic group lies in the pocket formed by the residues Asp309, Thr310 and Ala306. Asp309 has been suggested to be involved in the aromatization step involving the abstraction of H2 β from the A ring of ASD (Fig. 2).

The residues Thr310 and Ala306 are stabilized and locked into their respective positions by an inter-residue H-bond between Thr310–NH and O=C–Ala306 ($d_{H...O} = 2.15$ Å), while the carboxylate of the lateral chain of the

Fig. 4 Molecular structures of ASD and exemestane



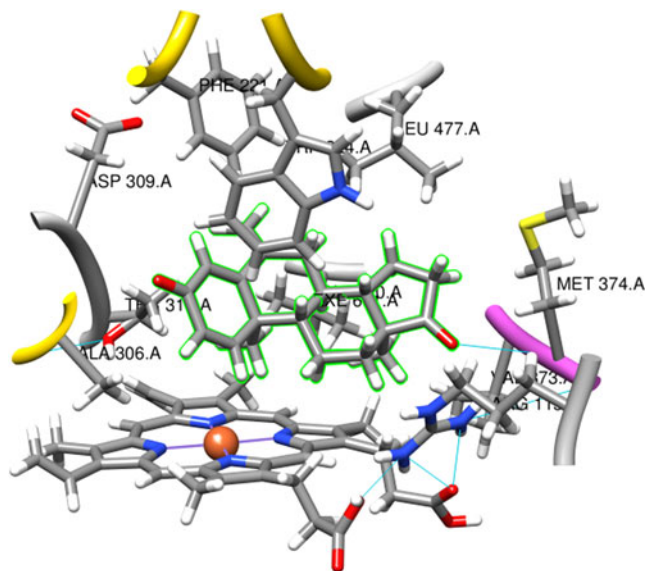


Fig. 5 Interaction between exemestane and the catalytic residues in the minimized association complex. The ligand is highlighted in green and only those amino acids that interact directly with it are shown

catalytic residue Asp309 is oriented towards the C-3 carbonylic group (A ring) ($d_{O...O}=3.74$ Å) (Fig. 3). As shown in Fig. 3, the B and C rings of ASD are also stabilized by hydrophobic interactions with the lateral chains of the residue Val370 and the heme moiety (the latter directly involving the C-19 methyl group) on the one side and Trp224 on the other.

Molecular interaction of the steroidal inhibitor exemestane (SAIs)

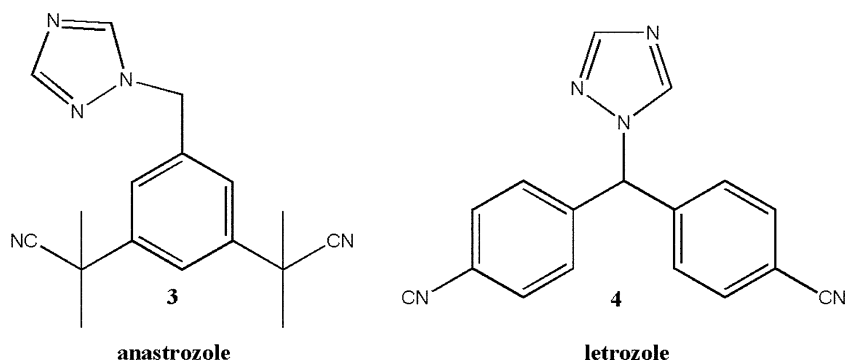
In order to focus on the binding of the irreversible steroidal inhibitor exemestane (Fig. 4), the association complex starting from 3eqm was again modeled. The ligand androstenedione was replaced with exemestane and automated docking with Autodock was initially performed, followed by MD simulation and minimization in order to identify the lowest-energy conformer (see “Computational details”). Since this molecule is classified as an irreversible inhibitor that prevents the aromatization reaction, we

expected it to bind deeper inside the catalytic cleft than ASD. Once again we performed the docking/MD procedure for the complex containing water inside the binding cleft and the complex without it. As obtained for ADS, the docked positioning of exemestane and its interactions with catalytic residues were not affected by the presence of explicit catalytic water molecules. Thus, herein we report the results obtained for the association complex with explicit solvent.

Ghosh et al. [6], in their study of the crystallographic structure of 3eqm, also attempted to discuss the binding interactions of exemestane by simply replacing the natural ligand ASD inside the catalytic site. However, no modeling of the complex was carried out, and discrepancies were found for some distances, which were less than the van der Waals contact radii. We observed that the main interactions observed for androstenedione are conserved, namely H-bonding involving C-3 carbonyl oxygen and Met374N–H (1.82 Å), the heme stabilization of the C-19 methyl, the hydrophobic stabilization of the A ring by residues Thr310 and Ala306 (which in addition interact together through an H-bond involving Ala306 C=O and Thr310 O–H; $d_{H...O}=1.88$ Å), and the hydrophobic interaction between the B and C rings of the steroid and the lateral chains of Val370 and Val373. Finally, some additional stabilization interactions were also observed that directly involve the exocyclic C-6 methylidene group, which is buried deeply in a shallow hydrophobic cavity formed by the lateral chains of residues Thr 310, Ser 478, Phe221, Val269 and Val370 (Fig. 5).

The distance between the Phe221 C–H and the C-methylidene is 2.92 Å, while the distance from C γ -Thr310 to the same exemestane carbon atom is 4.25 Å. This is longer than that observed in the crystallographic structure (3.0 Å), which is in fact lower than the van der Waals contact distance, as already pointed out [6]. This improved adjustment of the structure is due to the relaxation of the lateral chains during the MD simulation. Another aspect that must be taken into account is that in this association complex, the lateral chain of Asp309 is directed far from the C-17 carbonyl group. This is particularly interesting, since this residue seems to be directly implicated in the

Fig. 6 Molecular structures of anastrozole and letrozole



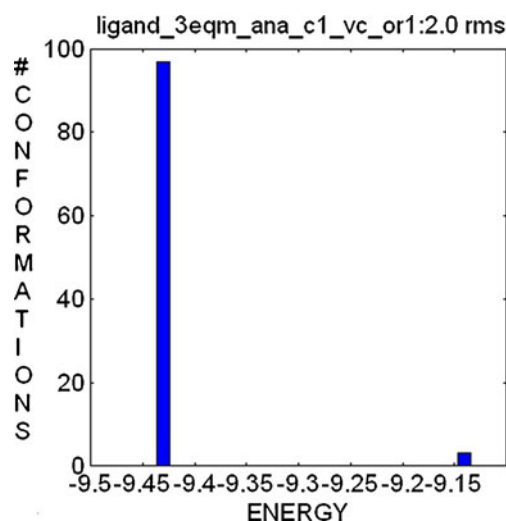


Fig. 7 Cluster analysis for anastrozole after automated docking

catalytic process (namely proton abstraction), so this different orientation, together with the high stabilization arising from the fact it is buried inside the hydrophobic site, may play an important role in explaining the irreversible binding of exemestane (a DFT/MM investigation of the catalytic mechanism is currently being performed by our group).

Docking of some new-generation competitive non-steroidal inhibitors (NSAIs): letrozole, anastrozole, and some derived azaheterocycle-containing compounds

Anastrozole and letrozole are two of the most important non-steroidal aromatase inhibitors (NSAIs), which exhibit significantly improved efficacies compared to tamoxifen, the drug most commonly used in this context [44–46]. A common substructure in NSAIs is a nitrogen-containing heterocycle that is a very effective pharmacophore group which interacts directly and coordinates with Fe(II) of the heme group [47–50]. In particular, anastrozole and letrozole can also be classified as type II inhibitors, in contrast to type I inhibitors, which include the natural ligand androstenedione and steroid-derived compounds (such as exemestane), since their binding to aromatase usually induces a bathochromic change in the Soret UV band compared to type I compounds. This bathochromic shift has been tentatively ascribed to the resulting coordination of the heme iron with an heteroatom such as the nitrogen of the triazolic ring [49, 50]. However, there have been no structural, crystallographic or computational studies to definitively assess this behavior. A hydrogen-bonding acceptor group has been considered another relevant pharmacophore element; in anastrozole and letrozole, this element is provided by the butyronitrile and benzonitrile groups, respectively (Fig. 6), as well as the aromatic phenyl moieties.

Recently, Jackson et al. [12] reported both the synthesis and the activities of some new NSAIs based on a biphenyl scaffold [(5-triazolyl methyl-2-cyano)-biphenyl], and they carried out a molecular automated docking study in order to confirm the suggested pharmacophores. However, they used the 3D model of aromatase obtained from homology modeling [15, 16], directly orienting the triazole ring towards heme and constraining the NSAIs inhibitor position by imposing a distance constraint between the heterocyclic nitrogen and Fe(II) without checking for the presence of any other energetically favored orientation. The same protocol was also applied by the group [51] when docking letrozole into the homology model of aromatase. Furthermore, it must be pointed out that the theoretical model of aromatase (1tga) shows many discrepancies with respect to the experimental model 3eqm, not in relation to the backbone but mainly to the orientations of the lateral chains of residues. This fact can lead to misunderstanding of the real molecular interactions between the ligand and the residues in the active cleft. Finally, the analysis of the rigid docking carried out by Roy et al. [18] showed a tendency for the azaheterocycle to orient towards heme, even if it is not properly coordinated with iron. In addition, they also pointed out the presence of several steric bumps which were ascribed to the low activities of the inhibitors considered, but which may also be due to a different preferential pose of the ligand and to fact that the complex geometries were not optimized. All of this computational evidence sheds light on some aspects of the association between the AI and aromatase, but it is not exhaustive as it does not consider the catalytic residues and ligand

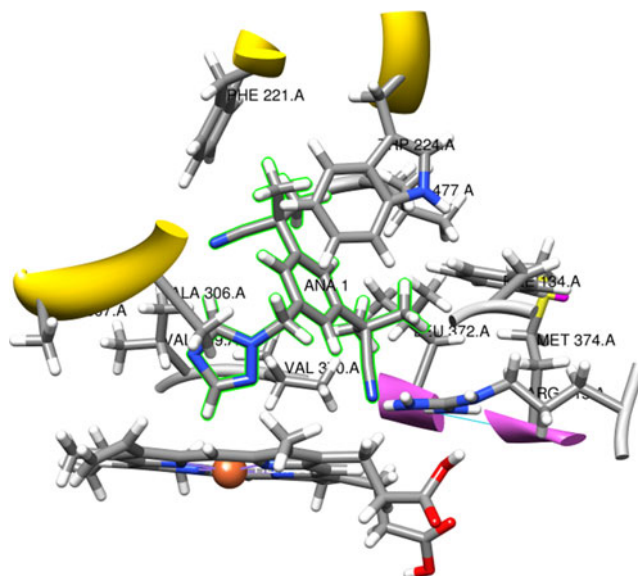


Fig. 8 Molecular interactions for the lowest-energy anastrozole–aromatase complex. The ligand is highlighted in *green*, and only those residues that interact directly with it are shown

relaxation during and after the complexation process. Therefore, in our study, docking was carried out by considering all possible orientations of the two ligands at the active site as starting points and, as already pointed out in the “Computational details” section, in both the presence and absence of water molecules.

Initially, automated docking (Autodock 4.0) was performed without imposing any distance restraint, which results in the clustering of conformers generated. Then, for each representative structure of each cluster obtained, and independent of their predicted binding energies, MD simulation/MM stabilization of the association complexes was carried out. This computational protocol is particularly important, since the energy calculated by Autodock has an intrinsic error which could be associated with the partial charges calculated (according to Gasteiger–Marsili) for the ligand [52] and with the lack of amino acid lateral chain relaxation. This latter information can be obtained from a further MD simulation. However, we also noticed that this energetic discrepancy is particularly important when the energy differences of the docked clusters are $<0.5 \text{ kcal mol}^{-1}$. For anastrozole, we did not observe any differences when explicit solvent was included inside the cleft in the cluster analysis (Fig. 7) and in the association complexes compared to when it was not, while for letrozole and its derivatives some discrepancies were found.

In fact, for anastrozole, whether in the presence of water or not, the cluster analysis showed the existence of two possible clusters, and thus two association complex geometries were identified: one with the azaheterocycle pointing away from the heme and the other with the same ring oriented towards Fe

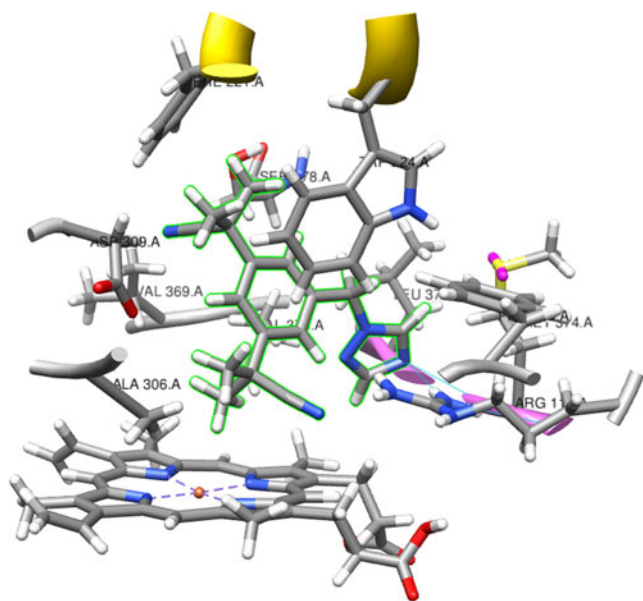


Fig. 9 Molecular interactions for the highest-energy anastrozole-aromatase complex

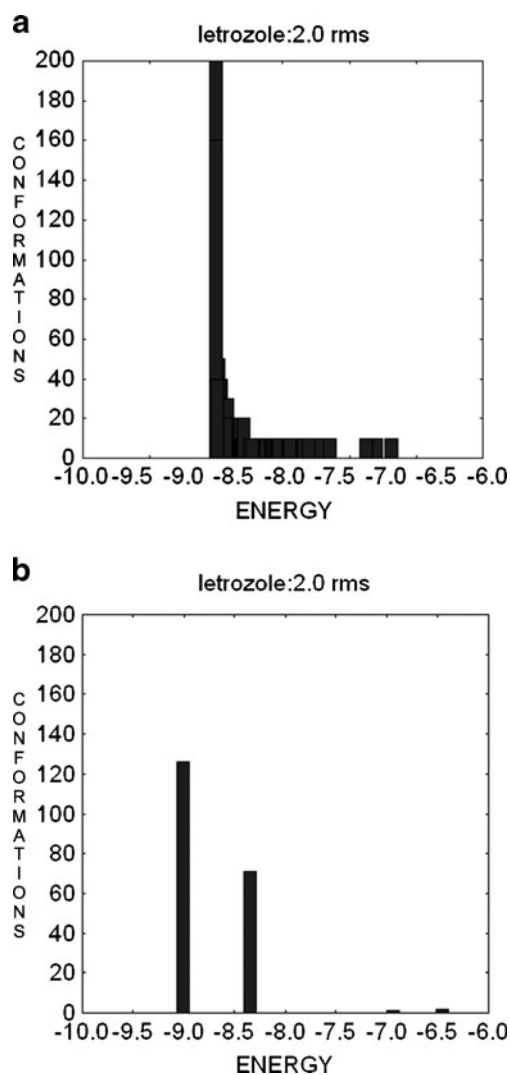


Fig. 10 **a** Cluster analysis for letrozole (with catalytic water) in automated docking. **b** Cluster analysis for letrozole (with water removed) in automated docking

(II). Thus, a representative structure of each cluster was chosen as the initial MD association complex. The results obtained strongly confirm the previously suggested pharmacophores [18], since the energy of the complex with the ligand oriented towards heme is much lower than the other one. Furthermore, $\Delta E_{\text{binding}}$ between these two association geometries increased significantly after molecular dynamical stabilization. These results confirm the need to perform an accurate MD simulation and trajectory analysis in order to evaluate the real interactions and energies that occur at the active site [53].

Upon thoroughly analyzing the molecular interactions that occur at the binding cleft, it becomes apparent that in the lowest-energy complex, the heterocyclic group points towards the heme moiety coordinating the Fe(II) ($d_{\text{Naza...Fe}} = 3.3 \text{ \AA}$), and one of the two cyanobutyl groups forms an H-bond (as acceptor) with the lateral chain of Arg115 ($d_{\text{NH...NC}} = 1.88 \text{ \AA}$).

Table 1 Distances of interest between the triazolic ring and water molecules inside the catalytic cleft (HOH 604, 605, 607, 621, 624, 630) during MD simulations of letrozole

Distance	After docking (Å)	MD 2 ns (Å)	MD 4 ns (Å)
$d(\text{N}(\text{triazole})\dots\text{Fe})$	7.2	6.9	5.1
$d(\text{O}(604)\dots\text{Fe})$	13.1	14.8	15.0
$d(\text{O}(605)\dots\text{Fe})$	11.3	11.4	12.5
$d(\text{O}(607)\dots\text{Fe})$	13.1	17.6	18.5
$d(\text{O}(621)\dots\text{Fe})$	20.1	20.3	20.5
$d(\text{O}(624)\dots\text{Fe})$	8.7	10.3	11.1
$d(\text{O}(630)\dots\text{Fe})$	18.4	18.6	18.7

Moreover, the same group is also stabilized by strong hydrophobic interactions involving the two methyls and the lateral chains of residues Phe134, Val370, Leu372, and Leu477. The other cyanobutyryl group does not form any H-bond but is stabilized by the residues Phe221 and Thr310. Furthermore, the phenyl moiety of anastrozole forms a π -stacking interaction with Trp224 (Fig. 8).

In the other highest-energy orientation, weaker stabilizing interactions were observed. The azaheterocycle group points away from the heme and is surrounded by the hydrophobic residues Leu372, Leu477, Met374 and Arg115, which is also involved in H-bonding with one cyanobutyryl group ($d_{\text{N-H}\dots\text{NC}}=2.26$ Å) and the azaheterocycle, which in turn is also involved in H-bonding with Met374. Finally, one cyanobutyryl group is also stabilized by the lateral chain of Ala306 while the other is instead surrounded by Val369 and Phe221, which are unable to stabilize the cyano moiety. The central phenyl moiety is once again stabilized by a π -stacking interaction with Trp224 (Fig. 9).

Besides anastrozole, letrozole was also docked to 3eqm, together with some recently synthesized letrozole-derived compounds which have shown interesting activity. Eventually, we considered some letrozole-derived compounds [19] that have been well characterized and evaluated. In their research work, however, the authors did not perform any modeling study to assess the binding interactions. Some of the most representative compounds were therefore chosen for a docking study using our model and the same

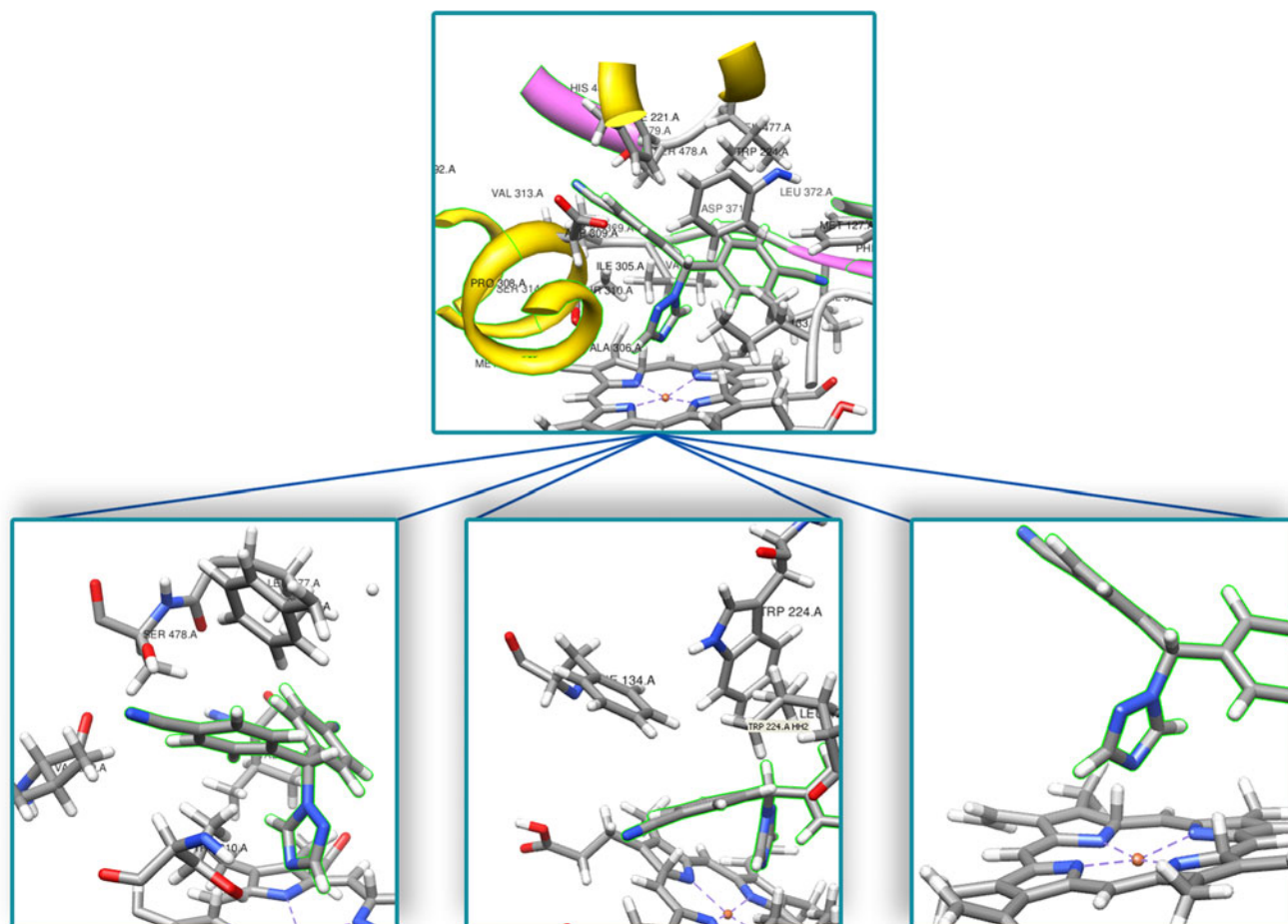
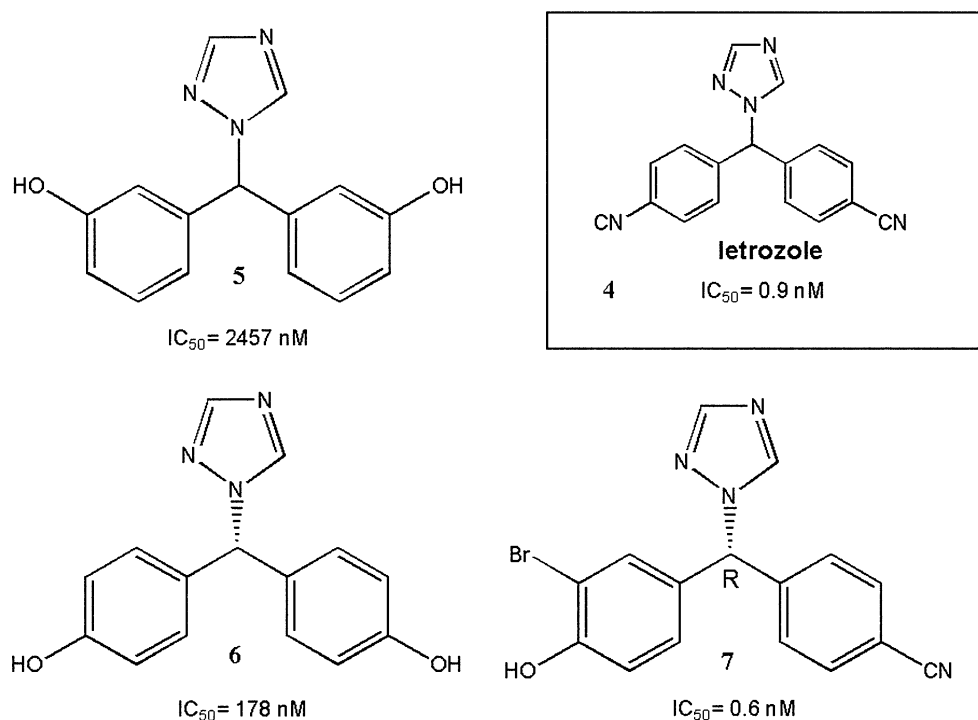


Fig. 11 Letrozole-stabilizing interactions in the association complex. The ligand is highlighted in green

Fig. 12 Structures and activity data for the letrozole-derived inhibitors 5–7 that were docked into our model [19]



computational protocol as described above. The aim was to definitively assess the pharmacophore requirements of azaheterocycle-containing compounds.

In particular, for letrozole, slightly different results were obtained when the docking process was conducted in the presence and in the absence of catalytic solvent molecules. In fact, docking results obtained in the presence of water showed the existence of several clusters, with the first two being almost isoenergetic ($\Delta E = 0.15 \text{ kcal mol}^{-1}$) (Fig. 10a). In the first cluster, letrozole is oriented with the triazolic ring pointing far from heme iron, while the triazolic ring coordinates with it in the second cluster. The latter orientation is in agreement with the experimental bathochromic shift observed. In the absence of water, the results of the cluster analysis show the existence of only one significant populated cluster, which has the heterocyclic ring coordinated with the heme (Fig. 10b).

In order to clarify this peculiar behavior, we carried out further MD simulations, starting from representative structures for each cluster resulting from docking in the presence of water, and at the end of the simulation we observed a significant displacement of the solvent molecules located inside the cleft from the ligand. Thus, we decided to go further and perform a longer MD run monitoring the positions of the water molecules. Indeed, the MD trajectories in both cases show water molecules moving towards the opening of the cleft (Table 1) as well as stronger binding of the triazolic ring to the heme iron. In particular, at the beginning, the heterocyclic ring is far from the heme, but as the simulation progressed it reoriented towards the iron

while the water moved away from it (Table 1), meaning that the ligand oriented itself as in the cluster 1 association complex (RMSD = 0.12 Å). This behavior is not particularly unexpected; as already mentioned, NSAIs are competitive inhibitors, and the steric hindrance associated with them is much greater than that afforded by SAIs. Thus, during the formation of the association complex, water molecules can initially interfere with the correct positioning of the inhibitor inside the cleft, but during ligand binding they move further away to allow the inhibitors accommodate themselves more snugly inside the binding cavity. This influence of water, which was not observed for anastrozole, may be ascribed to the smaller steric molecular dimensions of anastrozole than letrozole and its derivatives, which have two phenyl moieties instead of one (Fig. 6). Finally, we should also point out that in the presence of water, the lowest-energy association complex after MD simulations has a geometry that is fully in agreement with experimental data, with the triazolic ring coordinated to heme. This position of the ligand is perfectly superimposable on that

Table 2 Experimental IC_{50} and calculated E_{binding} values for compounds 4–7

Compound	E_{binding} (kcal/mol)	IC_{50} [19]
4	−9.01	0.9 nM
5	−7.74	2457 nM
6	−7.76	178 nM
7	−9.12	0.6 nM

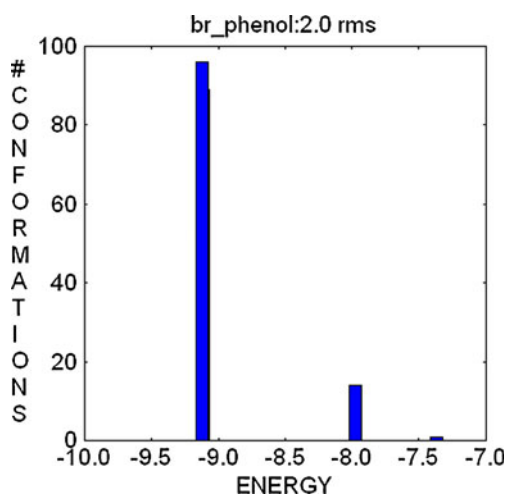


Fig. 13 Cluster analysis for compound **7** after automated docking

obtained in the absence of water. Thus, the molecular interactions observed are reported here.

First of all, as already stated in the paper, the orientation of the ligand agrees with experimental findings [49, 50], since the triazole ring is directly involved in strong coordination

with the heme iron ($d_{N...Fe}=2.80 \text{ \AA}$), thus explaining the observed bathochromic shift of the cytochrome Soret UV band [47–50].

One of the two cyanophenyl groups points towards the hydrophobic pocket surrounded by residues Val369, Val373, Leu477 and Thr310, and is stabilized by a π -stacking interaction with the lateral chain of Phe221 and by an H-bond with Ser478 ($d_{OH...NC}=2.08 \text{ \AA}$) (Fig. 11). Instead, the other cyanophenyl moiety is involved in an H-bonding interaction with Met374 ($d_{N-H...NC}=1.89 \text{ \AA}$) and in hydrophobic and π -stacking stabilization of the phenyl ring through interactions with the residues Ile133, Phe134, Trp224, Val370 and Leu477.

All of these findings were subsequently compared with the geometries of the association complexes of three letrozole-derived compounds (**5–7**) (Fig. 12). These compounds were chosen according to their activity data, since compound **7** is more active than letrozole, while the other two (**5–6**) are much less active. The aim was to correlate their activity with the interactions occurring in the stabilized lowest-energy complex. E_{binding} was calculated, and it is reported together with the experimental IC_{50} in

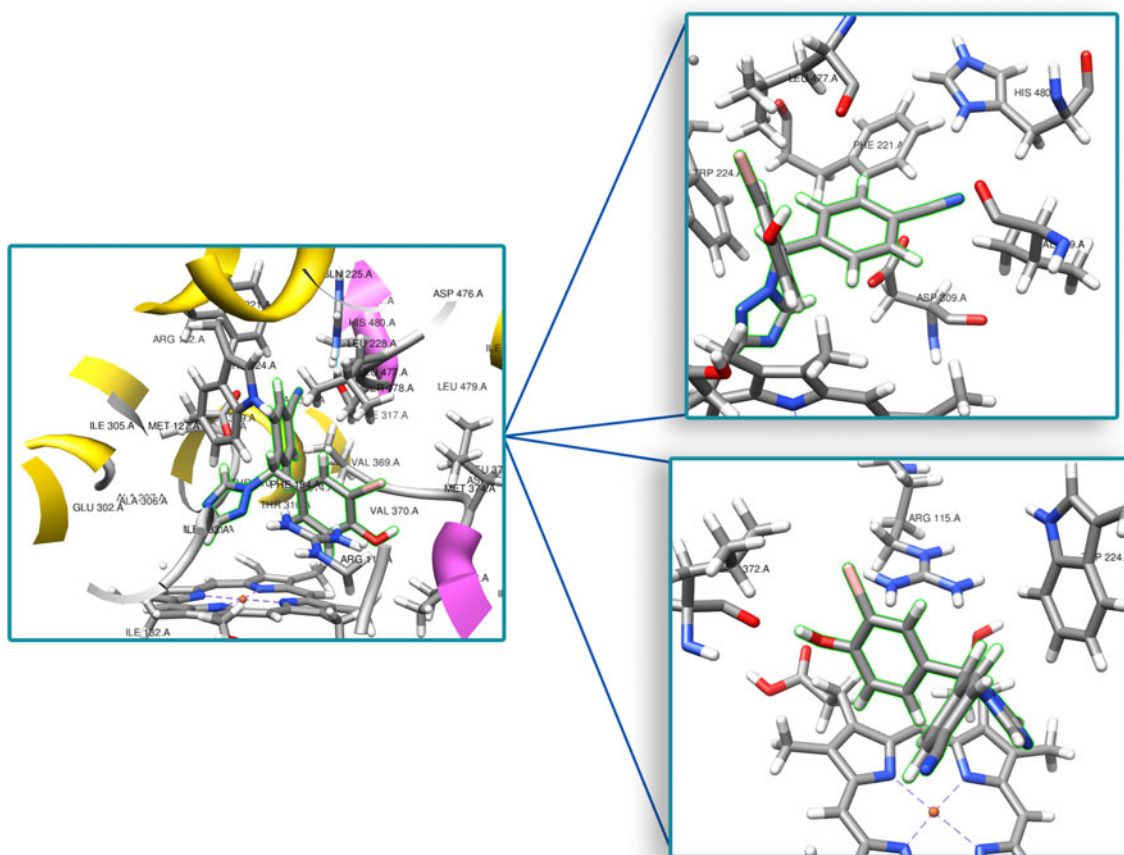


Fig. 14 Interactions for the bromophenol derivative **7** association complex. The ligand is highlighted in green. Only those residues interacting with letrozole are shown

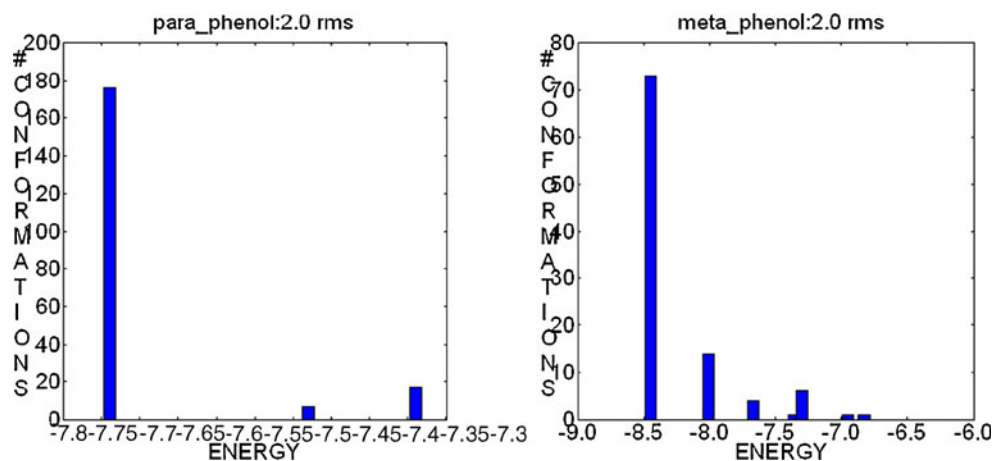
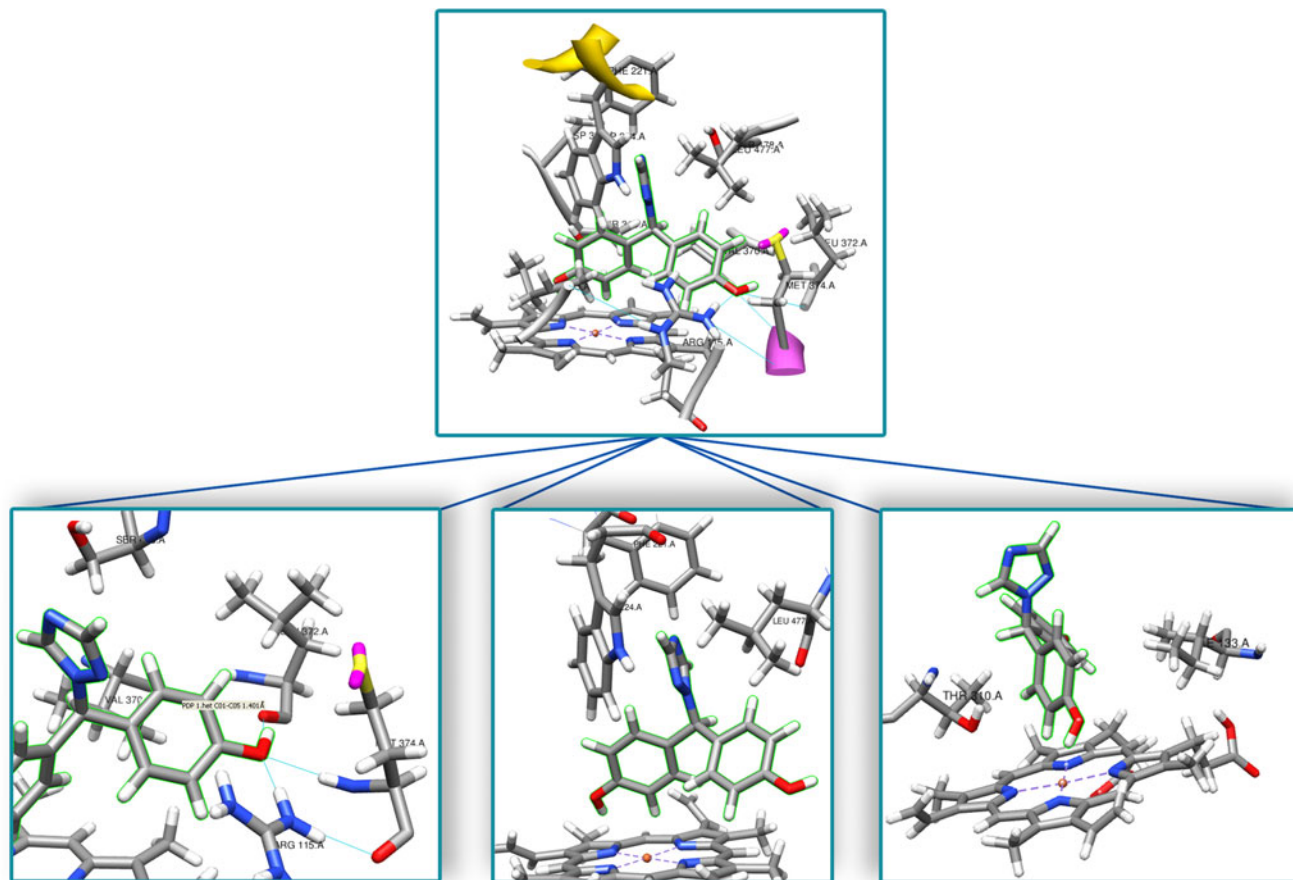
Fig. 15 Cluster analysis for compounds 6 and 7

Table 2. As can be observed, the computed order of activity for compounds 4–7 is in good agreement with that for the experimental data.

However, these four compounds show different binding preferences. For the most active bromophenyl derivative (7), the cluster analysis shows the presence of four groups of orientations (Fig. 13), with the first two being the closest in energy, and highly populated. Due to their very small

difference in energy ($\Delta E_{\text{binding c11-c12}}=0.1 \text{ kcal mol}^{-1}$), MD stabilization of the representative structures of both clusters was performed, and in this case the MD data analysis clearly demonstrated a strong energetic preference for the first of the two.

If we consider the molecular interactions at the binding cleft, in this association complex (as it is for letrozole), compound 7 is oriented with the triazolic ring towards the

**Fig. 16** Interactions of the association complex of the *p*-diphenol derivative 6. The ligand is highlighted in green

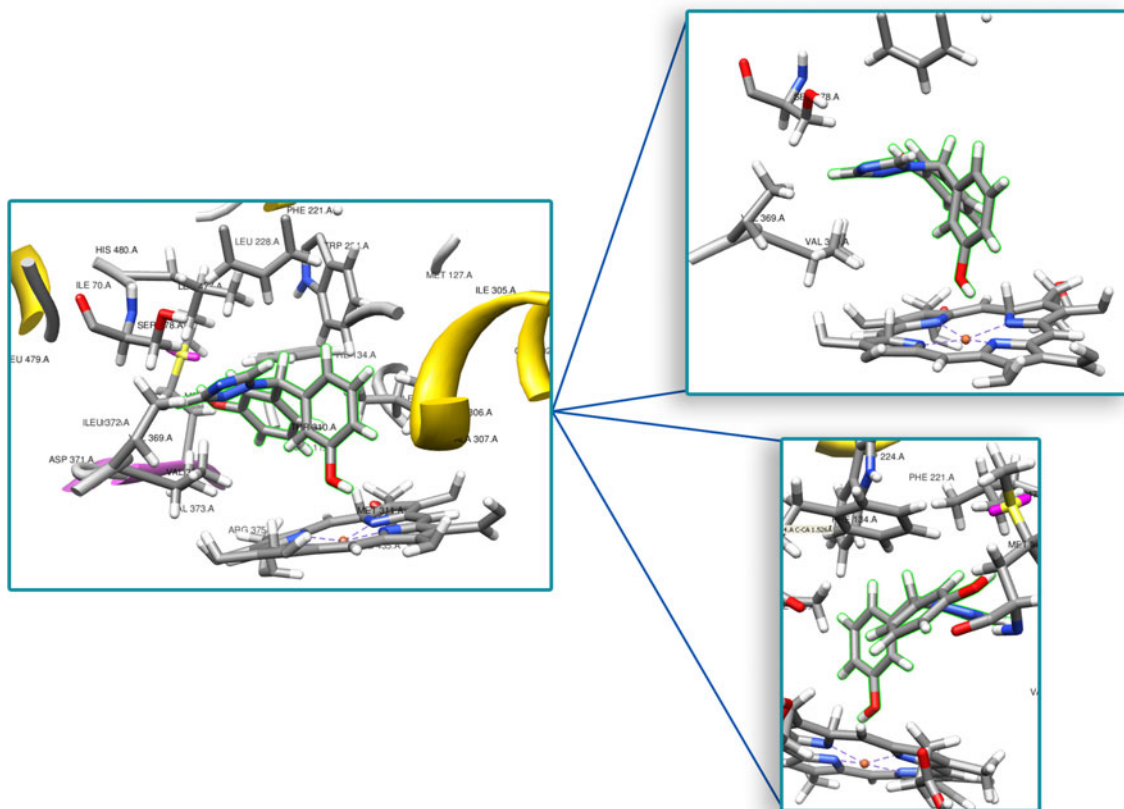


Fig. 17 Interactions of the association complex of the *meta*-diphenol derivative **7**. The ligand is highlighted in *green*

heme iron ($d_{N...Fe}=5.2$ Å), although this interaction appears weaker than that observed for letrozole. The cyanophenyl moiety is again oriented towards the hydrophobic cleft surrounded by residues Phe221, Val369, Thr224 and Leu477. In addition, the lateral chain of Asp309 is oriented perpendicular to the phenyl stabilizing it by π -stacking interactions. Moreover, His480 is involved in an H-bonding interaction with the cyano group itself ($d_{N-H...NC}=2.58$ Å).

The bromophenyl moiety interacts with Leu372, forming a strong H-bond ($d_{C=O...HO}=1.74$ Å), and with the lateral chain of Arg115 (Fig. 14). On analyzing the poses of the other two (less active) inhibitors **5** and **6**, it is surprisingly to observe that they do not show a preference in terms of the coordination of the triazolic ring with the heme Fe(II). In particular, for derivative **6**, clustering into three families that are well separated in energy is observed, with the first cluster being the most populous (Fig. 15).

In the lowest-energy association complex, the triazolic ring is oriented far from the heme, pointing instead towards the hydrophobic residues Leu477, Phe221 and Trp224 (the latter is involved in a π -stacking stabilization with the heterocyclic ring). Another strong stabilization concerns one of the two *p*-phenyl moieties, since the hydroxylic group forms three H-bonds with residues Leu372 ($d_{C=O...HO}=1.78$ Å), Met374 ($d_{N-H...OH}=2.04$ Å) and Arg115 ($d_{NH...OH}=1.91$ Å). Further-

more, the lateral chain of Val370 is oriented towards the aromatic ring. The other *p*-phenol group is perpendicular to the heme and involved in a weak H-bonding interaction with one of its nitrogens ($d_{OH...N}=2.30$ Å), while the aromatic ring is surrounded by the hydrophobic chains of Ile133 and Thr310 (Fig. 16).

Finally, the results obtained for *meta*-phenol (**5**) show complex clustering into eight families (Fig. 15). However,

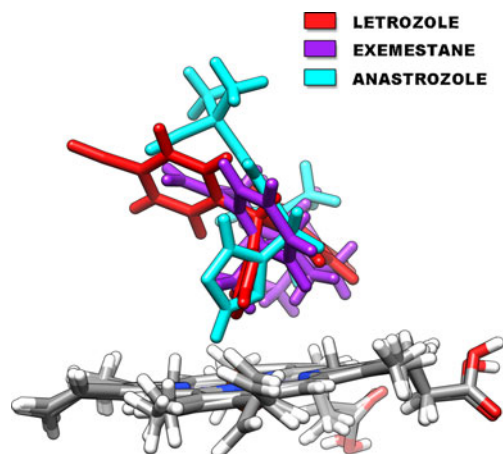


Fig. 18 RMS superimposition of triazole-containing NSAIs in their lowest-energy association complexes

the first cluster is very different in energy and characterized by a very high population density. The lowest-energy association complex found for **5** belongs to this cluster family (see Fig. 15), and the molecular interactions observed in it are even more surprising than those observed for compound **6**. In fact, the most important structural feature is the presence of coordination between one of the two *m*-phenol oxygens and Fe(II) ($d_{O...Fe}=2.50$ Å), while the triazole ring preferentially orients itself towards the residues Phe221, Ser478, Val369 and Val370 (Fig. 17). The other *m*-phenol group is instead surrounded by Ile133 and Leu477 and stabilized by stacking interactions with residues Phe134 and Trp224. Its hydroxyl group is oriented towards the sulfur of Met374 ($d_{OH...S}=3.72$ Å), thus yielding dipolar electrostatic stabilization.

Conclusions

From all of these observations, it is clear that the aza moiety is an important pharmacophore group, as it always shows notable stabilizing interactions in the association complex, and is involved directly in the coordination with the Fe(II) of heme, as suggested by experimental data. At the same time, we have demonstrated that aromatic functionalized groups compete with the aza moiety to stabilize the association complexes, and together these determine the activities of the NSAIs (Fig. 18). This aspect is not secondary, since the catalytic cleft is rich in both π -stacking stabilizing residues (Phe134, Phe221, Trp224) and H-donor or -acceptor amino acids (Arg115, Met 374, Ser478). However, it is worth mentioning that all of the inhibitors with nanomolar activity considered show a direct interaction of the heterocyclic group with Fe(II), while less-active NSAIs do not. Thus, in order to better assess the relevance of this coordination, and to better identify the role of the heterocyclic ring, we are currently studying the docking of other NSAIs containing azaheterocyclic groups that are different from the triazole.

References

- Simpson ER, Mahendroo MS, Means GD, Kilgore MW, Hinshelwood MM, Graham-Lorence S, Amarnah B, Ito Y, Fisher CR, Michael MD, Mendelson CR, Bulun SE (1994) Aromatase cytochrome P450, the enzyme responsible for estrogen biosynthesis. *Endocr Rev* 15:342–355
- Brueggmeir RW, Hackett JC, Diaz-Cruz ES (2005) Aromatase inhibitors in the treatment of breast cancer. *Endocr Rev* 26:331–345
- Plourde PV, Dyroff M, Dowsett M, Demers L, Yates R, Webster A (1995) Arimidex: a new oral, once-a-day aromatase inhibitor. *J Steroid Biochem Mol Biol* 53:175–179
- Lipton A, Demers LM, Harvey HA, Kambic KB, Grossberg H, Brady C et al (1995) Letrozole (CGS 20267). A phase I study of a new potent oral aromatase inhibitor of breast cancer. *Cancer* 75:2132–2138
- Chen S, Zhang F, Sherman MA, Kijima I, Cho M, Yuan YC, Toma Y, Osawa Y, Zhou D, Eng ET (2003) Structure-function studies of aromatase and its inhibitors: a progress report. *J Steroid Biochem Mol Biol* 86:231–237
- Gosh D, Griswold J, Erman M, Pangborn W (2009) Structural Basis for androgen specificity and estrogen synthesis in human aromatase. *Nature* 458:219–223
- Gosh D, Griswold J, Erman M, Pangborn W (2010) X-ray structure of human aromatase reveals an androgen-specific active site. *J Steroid Biochem Mol Biol* 118:197–202
- Hong Y, Li H, Yuan YC, Chen S (2009) Molecular characterization of aromatase. *Ann NY Acad Sci* 1155:112–120
- Paoletta S, Stevenson GB, Wildeboer D, Eherman TM, Hylands PJ, Barlow DJ (2008) Screening of herbal constituents for aromatase inhibitory activity. *Bioorg Med Chem* 16:8466–8470
- Takahashi M, Yamashita K, Numazawa M (2010) Probing the binding pocket of the active site of aromatase with 2-phenylaliphatic androsta-1,4-3,17-dione steroids. *Steroids* 75:330–337
- Cassidy CE, Setzer WN (2010) Cancer-relevant biochemical targets of cytotoxic *Lonchocarpus* flavonoids: a molecular docking analysis. *J Mol Model* 16:311–326
- Jackson T, Lawrence LW, Trusselle MN, Purhoit A, Reed MJ, Potter BVL (2008) Non-steroidal aromatase inhibitors based on a biphenyl scaffold: synthesis, in vitro SAR and molecular modeling. *ChemMedChem* 3:603–618
- Chen S, Kao YC, Laughton CA (1997) Binding characteristics of aromatase inhibitors and phytoestrogens to human aromatase. *J Steroid Biochem Mol Biol* 61:107–115
- Oliveira AA, Ramalho TC, da Cunha EFF (2009) QSAR study of androstenedione analogs as aromatase inhibitors. *Lett Drug Des Discov* 6:554–562
- Favia AD, Cavalli A, Masetti M, Carotti A, Recanatini M (2006) Three dimensional model of the human aromatase enzyme and density functional parametrization of the iron containing protoporphyrin IX for a molecular dynamics study of heme-cysteinato cytochromes. *Proteins* 62:1074–1087
- Graham-Lorence S, Amarnah B, White R.E, Peterson JA, Simpson ER (1995) A three dimensional model of aromatase cytochrome P450*. *Protein Sci* 4:1065–1080
- Williams PA, Cosme J, Ward A, Angove HC, Matak Vinkovic D, Jhoti H (2003) Crystal structure of human cytochrome P450 2C9 with bound warfarin. *Nature* 424:464–468
- Roy PP, Roy K (2010) Docking and 3D-QSAR studies of diverse classes of human aromatase (CYP19) inhibitors. *J Mol Model* 16:1597–1616
- Wood PM, Lawrence Woo LW, Labrosse JR, Trusselle MN, Abbate S, Longhi G, Castiglioni E, Lebon F, Purohit A, Reed MJ, Potter BVL (2008) Chiral aromatase and dual aromatase-steroid sulfatase inhibitors from the letrozole template: synthesis, absolute configuration, and in vitro activity. *J Med Chem* 51:4226–4238
- Petersen 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* 13:1605–1612
- Onufriev A et al. (2011) H++ server homepage. <http://biophysics.cs.vt.edu/>
- Thomson EA, Siiteri PK (1974) Utilization of oxygen and reduced nicotinamide adenine dinucleotide phosphate by human placental microsomes during aromatization of androstenedione. *J Biol Chem* 249:5364–5372
- O'Neal Johnston J (1998) Aromatase inhibitors. *Crit Rev Biochem Mol Biol* 33:375–405

24. Aktar M, Calder DL, Corina DL, Wright JN (1982) Mechanistic studies on C-19 demethylation in estrogen biosynthesis. *Biochem J* 201:569–580
25. Aktar M, Njar VC, Wright JN (1993) Mechanistic studies on aromatase and related C-C bond cleaving P-450 enzymes. *J Steroid Biochem Mol Biol* 44:375–387
26. Silgar SG, Murray RI (1986) In: de Montellano PR Ortiz (ed) *Cytochrome P-450: structure, mechanism and biochemistry*, 3rd edn. Plenum, New York, p 429
27. Poulos TL (1986) In: de Montellano PR Ortiz (ed) *Cytochrome P-450: structure, mechanism and biochemistry*, 3rd edn. Plenum, New York, p 505
28. Waxman DJ (1986) In: de Montellano PR Ortiz (ed) *Cytochrome P-450: structure, mechanism and biochemistry*, 3rd edn. Plenum, New York, p 525
29. Beunssen DD, Carrell HL, Covey DF (1987) Metabolism of 19-methyl-substituted steroids by human placental aromatase. *Biochemistry* 26:7833
30. Hackett JC, Brueggenmeier RW, Hadad CM (2005) The final catalytic step of cytochrome P450 aromatase: a density functional theory study. *J Am Chem Soc* 127:5244–5237
31. Still WC, Tempczyk A, Hawley RC, Hendrickson T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc* 112:6127–6129
32. Mohamadi F, Richards NGJ, Guida WC, Liskamp R, Lipton M, Caufied C, Chang G, Hendrickson T, Still WC (1990) Macro-model—an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J Comput Chem* 11:440–467
33. Weiner JS, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner PA Jr (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. *J Am Chem Soc* 106:765–784
34. 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* 162:785–791
35. 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
36. Huey R, Goodsell DS, Morris GM, Olson AJ (2004) Grid-based hydrogen bond potentials with improved directionality. *Lett Drug Des Discov* 1:178–183
37. 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
38. Leimkuhler B, Skeel R (1994) Symplectic numerical integrators in constrained Hamiltonian systems. *J Comput Phys* 112:117–125
39. Brunetti L, Galeazzi R, Orena M, Bottoni A (2008) Catalytic mechanism of L,L-diaminopimelic acid with diaminopimelate epimerase by molecular docking simulations. *J Mol Graph Model* 26:1082–1090
40. Melchiorre C, Andrisano V, Bolognesi ML, Budriesi R, Cavalli A, Cavrini V, Rosini M, Tumiatti V, Recanatini M (1998) Acetylcholinesterase noncovalent inhibitors based on a polyamine backbone for potential use against Alzheimer's disease. *J Med Chem* 41:4186–4189
41. Rampa A, Bisi A, Valenti P, Recanatini M, Cavalli A, Andrisano V, Cavrini V, Fin L, Buriani A, Giust P (1998) Acetylcholinesterase inhibitors: synthesis and structure–activity relationships of omega-[N-methyl-N-(3-alkylcarbamoyloxyphenyl)-methyl]aminoalkoxy-heteroaryl derivatives. *J Med Chem* 41:3976–3986
42. Calvaresi M, Garavelli M, Bottoni A (2008) Computational evidence for catalytic mechanism of glutamine cyclase, a DFT investigation. *Proteins* 73:527–538
43. Stenta M, Calvaresi M, Altoè P, Spinelli D, Garavelli M, Galeazzi R, Bottoni A (2009) The catalytic mechanism of DAP epimerase: a QM/MM investigation. *J Chem Theory Comput* 5:1915–1930
44. Thurlimann B, Keshaviah A, Coates AS, Mouridsen H, Mauriac L, Forbes JF, Paridaens M, Castiglione-Gertsch M, Gelber RD, Rabaglio M, Smith I, Wardely A, Price KN, Goldhirsh A (2005) A comparison of letrozole and tamoxifen in postmenopausal women with early breast cancer. *N Engl J Med* 353:2747–2757
45. Jakesz R, Jonet W, Gnant M, Mittleboeck M, Grail R, Tausch C, Hilfrich J, Kwasny W, Menzel C, Samonigg H (2005) Switching of postmenopausal women with endocrine-responsive early breast cancer to anastrozole after 2 years' adjuvant tamoxifen: combined results of ABCSG trial 8 and ARNO 95 trial. *Lancet* 366:455–462
46. Furet P, Batzl C, Bathnagar A, Francotte E, Rihs G, Lang M (1993) Aromatase inhibitors: synthesis, biological activity, and binding mode of azole-type compounds. *J Med Chem* 36:1393–1400
47. Kottler T, Lawrence Woo LW, Trusselle MN, Purohit A, Reed MJ, Potter BVL (2008) Non-steroidal aromatase inhibitors based on a biphenyl scaffold: synthesis, in vitro SAR and molecular modeling. *ChemMedChem* 3:603–618
48. Neves MAC, Dinis TCP, Colombo G, Sá e Melo ML (2009) Fast three dimensional pharmacophore virtual screening of new potent nonsteroid aromatase inhibitors. *J Med Chem* 52:143–150
49. Cole PA, Robinson CH (1990) Mechanism and inhibition of cytochrome P-450 aromatase. *J Med Chem* 33:2933–2942
50. Hong Y, Cho M, Yuan YC, Chen S (2008) Molecular basis for the interaction of four different classes of substrates and inhibitors with human aromatase. *Biochem Pharm* 75:1161–1169
51. Lawrence Woo LW, Bubert C, Sutcliffe OB, Smith A, Chander SK, Mahon MF, Purohit A, Reed MJ, Potter BVL (2007) Dual aromatase–steroid sulfatase inhibitors. *J Med Chem* 50:3540–3560
52. Bikadi Z, Hazai E (2009) Application of the PM6 semi-empirical method to modeling proteins enhances docking accuracy of AutoDock. *J Chem Inf* 1:15–31
53. Galeazzi R (2009) Molecular dynamics as a tool in rational drug design: current status and some major applications. *Curr Comput Aided Drug Des* 5:225–240

Pressure-induced phase transition in wurtzite ZnTe: an ab initio study

Sebahaddin Alptekin

Received: 15 March 2011 / Accepted: 6 June 2011 / Published online: 18 June 2011
© Springer-Verlag 2011

Abstract A constant pressure ab initio MD technique and density functional theory with a generalized gradient approximation (GGA) was used to study the pressure-induced phase transition in wurtzite ZnTe. A first-order phase transition from the wurtzite structure to a *Cmcm* structure was successfully observed in a constant-pressure molecular dynamics simulation. This phase transformation was also analyzed using enthalpy calculations. We also investigated the stability of wurtzite (WZ) and zinc-blende (ZB) phases from energy–volume calculations, and found that both structures show quite similar equations of state and transform into a *Cmcm* structure at 16 GPa using enthalpy calculations, in agreement with experimental observations. The transition phase, lattice parameters and bulk properties we obtained are comparable with experimental and theoretical data.

Keywords Ab initio calculation · High pressure · Phase transformation · Semiconductors

Introduction

The structural and electronic properties of II–VI semiconductor compounds have been extensively studied in the last 30 years because such compounds are of technical and scientific importance. The general phase transition property of II–VI semiconductor compounds is that they transform from the zinc-blende (ZB) or wurtzite (WZ) phase to the

rocksalt (RS) before transforming to the β -Sn phase. However, a significant alteration to this generally accepted series of structural changes has been reported recently. Pressure-induced polymorphism requires theoretical and experimental studies to understand the observed changes, including the semiconductor–metallic transformation exhibited by many II–VI materials under hydrostatic pressure. These first-order structural transitions increase the zero-pressure metal coordination in the lattice and thus narrow the band gaps.

ZnTe crystallizes under ambient conditions in the hexagonal wurtzite (WZ) and the zinc blende (ZB) structures with space groups $P6_3mc$ and $F\bar{4}3m$, respectively. The high-pressure behavior of ZnTe has been the subject of a few experimental and theoretical studies [1–10], but it will require additional studies to be fully understood. Raman spectra showed evidence for a phase transition around a pressure of 94 GPa in ZnTe [1]. Côté et al. reported that the transformation pressure range for the ZnSe cinnabar was 10.2–13.4 GPa, based on pseudopotential calculations for ZnSe and ZnTe [2]. Recently, Nelmes and colleagues used angle-dispersive techniques and image-plate detectors and found that ZnTe also has an unusual orthorhombic structure with *Cmcm* symmetry under an applied pressure of 16 GPa [3]. This result is rather important for both experimentalists and theorists studying the structural stability of II–VI semiconductors under pressure. Lee and colleagues applied the ab initio pseudopotential plane-wave method within the local density approximation (LDA) to study the structural phase transitions of ZnTe, and found that in the orthorhombic phase of ZnTe with space group *Cmcm*, the primitive unit cell consists of eight basis atoms [4]. The cinnabar and *Cmcm* structures can each be regarded as a distorted NaCl structure, and the transition from the cinnabar to the *Cmcm*

S. Alptekin (✉)
Department of Physics, Faculty of Science,
Cankiri Karatekin University,
18100 Cankiri, Turkey
e-mail: salptekin@karatekin.edu.tr

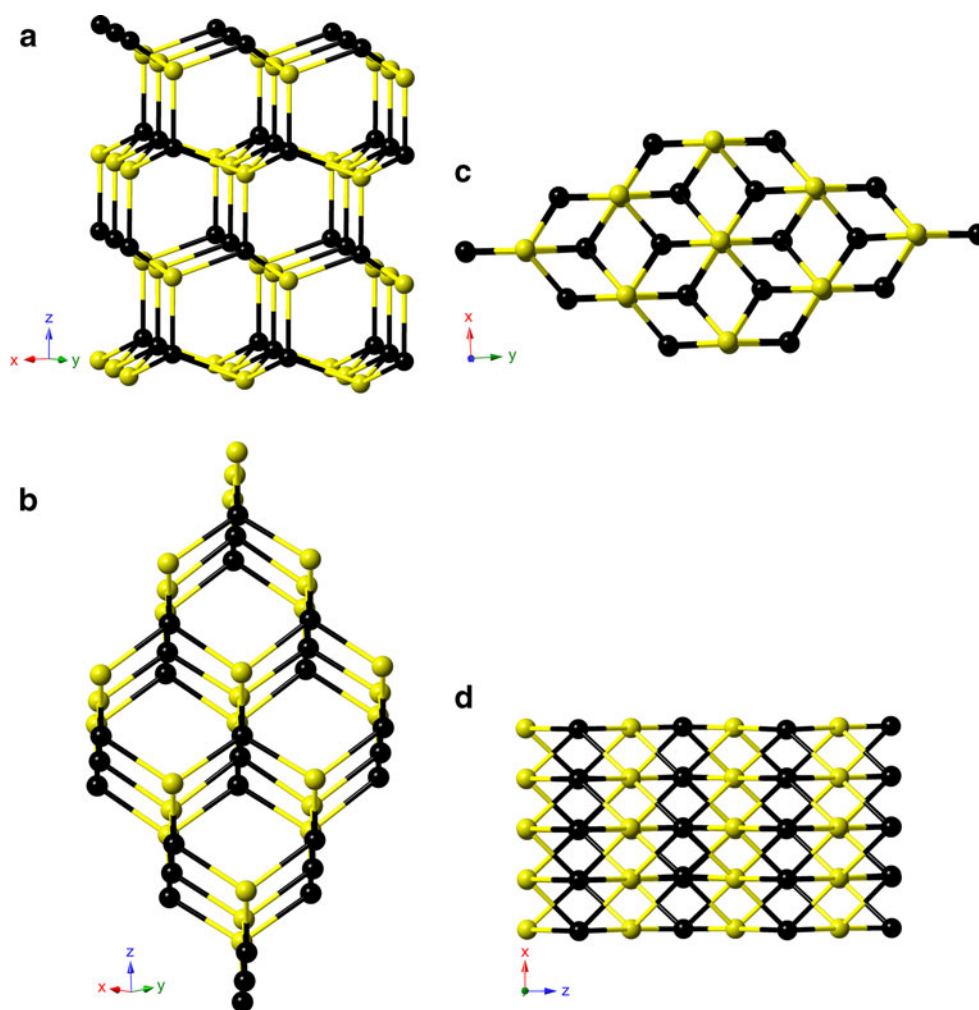
is weakly first order [4]. Wurtzite GaAs is found to be stable at ambient pressure for temperatures up to 473 K, with a structure that is only slightly distorted from ideal. On recompression, the c/a ratio is constant with pressure, and wurtzite GaAs transforms to the orthorhombic $Cmcm$ phase at 18.7 GPa [6]. ZnTe has been studied at high pressures up to 76 GPa and at room temperature in a diamond anvil cell using angle-dispersive X-ray diffraction with synchrotron radiation and an imaging plate detector. The equation-of-state parameters of the high-pressure phase of ZnTe were derived for the first time, with $B_0=134$ GPa for the $Cmcm$ -type phase [11]. Calculations in previous studies have shown that the zinc-blende phase of ZnTe is the most stable, and the structure transforms to the cinnabar phase and again to the $Cmcm$ phase as the pressure increases [4]. It was also shown that the phase of $Cmcm$ ZnTe has a site-ordered orthorhombic structure with space group $Cmcm$ and unusual fivefold coordination, which can be understood as a strong distortion of the NaCl structure [3]. HgTe and CdTe convert to $Cmcm$ at high pressure as ZnTe. However, ZnTe differs only in that it has no intermediate NaCl phase.

In this paper, we use a constant-pressure ab initio MD technique to study the pressure-induced phase transition in ZnTe. We show that wurtzite ZnTe undergoes a first-order phase transition (Fig. 1) into a $Cmcm$ structure. We also investigate the stabilities of the wurtzite (WZ) and zinc-blende (ZB) phases from energy–volume calculations.

Methods

The calculation was carried out with the ab initio program SIESTA [12]. We used the density functional theory (DFT) with the generalized gradient approximation (GGA) of Perdew, Burke and Ernzerhof for the exchange–correlation energy [13]. The norm-conservative Troullier–Martins pseudopotentials [14] were employed for core electrons, and valence electrons were described with a split-valence double- ξ basis set expanded with polarized functions. A uniform mesh with a plane wave cut-off of 150 Ry was used to represent the electron density, the local part of the pseudopotential, and the Hartree and exchange–correlation

Fig. 1 Crystal structures of ZnTe: **a** the wurtzite phase at zero pressure, **b** the zinc-blende phase at zero pressure, **c**, **d** the $Cmcm$ phase formed at 50 GPa (the atoms Zn and Te are shown in black and yellow, respectively)



potentials. The simulation cell consisted of 72 atoms with periodic boundary conditions. We used Γ -point sampling for the supercell's Brillouin zone integration. The molecular dynamics (MD) simulations were performed using the NPH (constant number of atoms, constant pressure, and constant enthalpy) ensemble. The reason for choosing this ensemble was to remove thermal fluctuations, which makes it easier to examine the structure during the phase transformation. The system was first equilibrated at zero pressure, and then the pressure was gradually increased in increments of 10.0 GPa. For each value of the applied pressure, the structure was allowed to relax and find its equilibrium volume and the lowest energy by optimizing its lattice vectors and atomic positions together until the stress tolerance was less than 0.5 GPa and the maximum atomic force was smaller than $0.01 \text{ eV } \text{\AA}^{-1}$. To optimize the geometries, a variable cell shape conjugate-gradient method under a constant pressure was used. For the energy volume calculations, we considered the unit cell for ZnTe phases. The Brillouin zone integration was performed with an automatically generated $10 \times 10 \times 10$ k -point mesh for the phases following the convention of Monkhorst and Pack [15]. In order to determine the symmetries of the high-pressure phases formed in the simulations, we used the KPLLOT program [16], which provides detailed information on the space group, the cell parameters and the atomic positions of a given structure. In the symmetry analysis, we used tolerances of 0.2 \AA , 4° and 0.7 \AA for bond lengths, bond angles and inter-plane spacing, respectively.

Results

Enthalpy calculations

Transition pressures in constant pressure simulations are generally overestimated, just as in superheating molecular dynamics simulations. This implies a high intrinsic activation barrier for transforming one solid phase into another in simulations. When particular conditions such as the finite size of simulation cells and the lack of any defects and surfaces in simulated structures are considered, such overestimated transition pressures are anticipated. Structural phase transformations in simulations do not proceed by nucleation and growth; they occur across all of the simulation cells. This means that the systems have to cross a significant energy barrier to transform from one phase to another one, and hence simulated structures have to be overpressurized in order to obtain a phase transition. Additionally, the absence of thermal motion (relaxation of the structure at constant pressure) in our simulations shifts the transitions to a higher pressure. On the other hand, to determine the most stable structure at finite pressure and

temperature, the free energy $G = E_{\text{tot}} + PV - TS$ should be used. Our density functional calculations were essentially performed at zero temperature, and entropic contributions could be neglected. Therefore, the enthalpy values, $H = E + pV$, including pressure–volume effects, were calculated. We performed energy–volume calculations to study the stability of the WZ, ZB and $Cmcm$ phases. The structures were equilibrated at several volumes, and their energy–volume relations were fitted to the third-order Birch–Murnaghan equation of states. The energy–volume curves for the structures are presented in Fig. 2. The ZB crystal has the lowest energy. The total energy difference between the ZB and WZ phases is, however, rather small—about 0.58 eV/atom . Such a small energy difference between these phases was anticipated, because both structures have similar tetrahedral bonds up to the second-neighbor distances. This behavior is compatible with a phase transition between these structures, which is also clearly reflected in the enthalpy calculation.

A simple comparison of the static lattice enthalpies of the wurtzite state, zinc-blende state and the $Cmcm$ state leads to the transition pressure between them. The point at which the three enthalpy curves cross indicates a pressure-induced phase transition between these phases. The computed enthalpy curves for the WZ, ZB and $Cmcm$ phases are plotted as a function of pressure in Fig. 3. As can be seen from the figure, the enthalpy curves of the WZ phase and ZB phase have the same enthalpy and cross that of the $Cmcm$ phase at 16 GPa, indicating a first-order phase transition between these phases. Figures 2 and 3 are similar to those seen in [5, 17]. This transition pressure agrees with the experimental value of 16 GPa [3]. On the other hand, the previous theoretical result was 13.9 GPa [5]. From the energy–volume data, we also calculated the bulk moduli of these phases. For the WZ state, the calculated bulk modulus was 63.83 GPa, which is relatively close to the theoretical

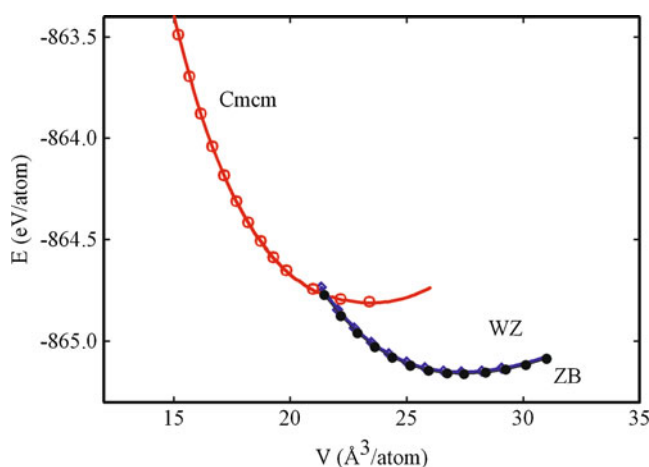


Fig. 2 The calculated energies for the WZ, $Cmcm$ and ZB phases as a function of volume

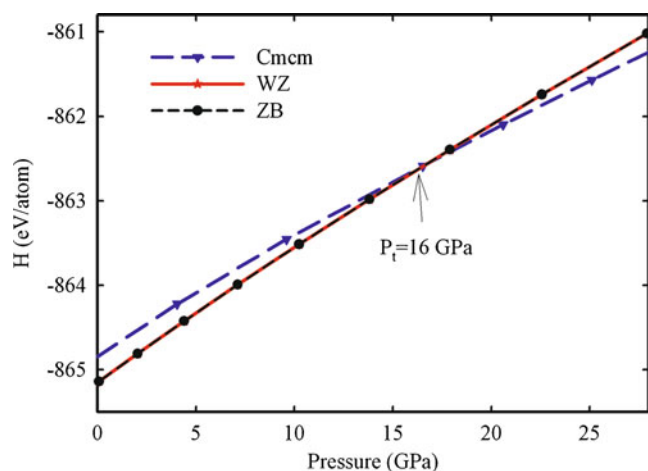


Fig. 3 The calculated enthalpy curves for the WZ, *Cmc* and ZB phases of ZnTe

value of 55.3 GPa [7]. The bulk modulus can vary a great deal according to the methodology of the study. The bulk modulus of the *Cmc* phase was calculated to be 75.3 GPa and the bulk modulus of the ZB phase was calculated to be 63.4 GPa. The experimental and theoretical results for different structures of ZnTe are also given in Table 1 [2, 5, 7, 10, 11, 18, 19]. In general, our results agree with other reported experimental and theoretical results.

Constant-pressure simulation

The pressure–volume relation for ZnTe obtained through constant-pressure simulation can be seen in Fig. 4. The figure shows that the volume monotonically decreases with increasing pressure up to 40 GPa. When the pressure is increased further, from 40 to 50 GPa, the structural phase transition begins, and the volume decreases noticeably,

which is typical of a first-order phase transition. The structural analysis indicates that wurtzite ZnTe converts into a *Cmc* structure. The transition pressure obtained from constant-pressure simulation is, on the other hand, considerably larger than the experimental result of 16 GPa [3] and the static enthalpy result of 16 GPa calculated in the previous section. This overestimate was anticipated considering some of the aspects of the simulations: the use of an ideal structure, the size of the simulated structure, etc. Consequently, simulated systems have to cross a significant energy barrier to transform from one phase to another one. The high energy barrier can be only crossed when the simulation box is overpressurized in order to achieve such a phase transition [20, 21].

In this study, we were particularly interested in understanding the transformation mechanism that controls structural phase transitions. Therefore, as a next step, we studied the movements of atoms during phase transformation by analyzing the changes in the simulation cell and plotting the simulation cell lengths and angles at 50 GPa as a function of minimization step in Fig. 5. The simulation cell vectors A, B, and C were originally oriented along the [100], [010] and [001] directions, respectively. The magnitudes of these vectors are plotted in the figure. It is clear from the figure that the mechanism of transformation from the WZ to the *Cmc* structure in ZnTe is straightforward, and involves noticeable decreases in |B| and |A|, a noticeable increase in |C|, and a change in the α -angle (between the A and B lattice vectors) from 120° to 126° , which occur simultaneously. Structural analysis using the KPLOTT program [16] indicates that this new state has an orthorhombic structure with *Cmc* symmetry. The lattice constants of the *Cmc* phase are $a=3.221$, $b=6.235$ Å and $c=6.634$ Å. When the *Cmc* structures are compared, the wurtzite phase is

Table 1 Lattice parameters (a , b , c), atomic positions u (u_x , u_y , u_z), c/a values and bulk moduli ($B_0 = -\{\text{change in pressure}/\text{fractional change in volume}\}$) for ZnTe structures

Phase	B_0 (GPa)	a (Å)	b (Å)	c (Å)	c/a	u	Reference
WZ	63.83	4.254	4.254	6.989	1.643	0.373	This study
	55.3	4.234			1.648	0.373	[7]
ZB	63.4	6.047	6.047	6.047			This study
	55.4	6.002					[7]
	47.7	6.158	6.158	6.158			[5]
	54.7	6.013					[2]
	50.54	6.063					[18]
	51.2	6.174					[19]
	45.25	6.17					[10]
	76.4 ^{exp}						[11]
<i>Cmc</i>	75.3	3.221	6.235	6.634(sixfold)			This study
	62.2	5.655	6.277	5.267(fivefold)			[5]
	58.79	5.739	6.318	5.265(fivefold)			[10]
	134 ^{exp}						[11]

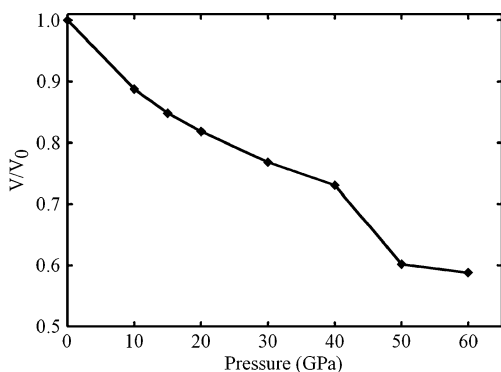


Fig. 4 Pressure–volume curve of wurtzite ZnTe as a function of pressure in the constant-pressure simulation

initially fourfold coordinated, while the resultant *Cmcm* phase is sixfold coordinated. For ZB, on the other hand, the ZB phase is fourfold coordinated at the beginning, while the resultant *Cmcm* phase is fivefold coordinated [5, 10]. Figure 6 shows the variations in simulation cell length as a function of pressure.

Discussion

The WZ-to-*Cmcm* phase change is a reconstructive phase transformation that involves large displacements of atoms. Therefore, a WZ-structured system can transform from one phase to another by passing through various closely related paths during the transformation. In other words, the

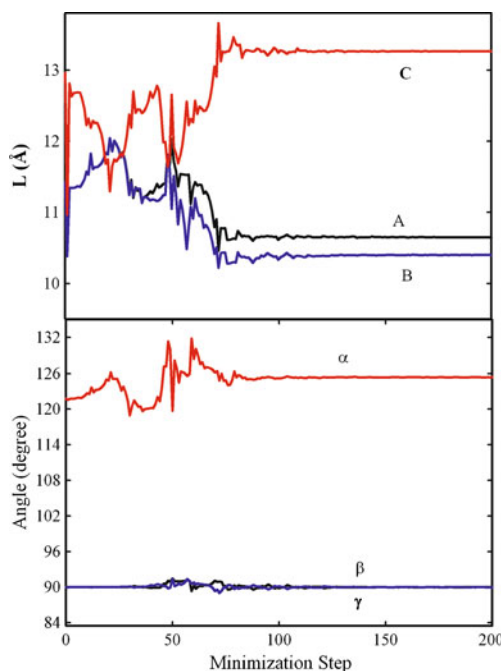


Fig. 5 Changes in the simulation cell lengths and angles as a function of minimization step at 50 GPa

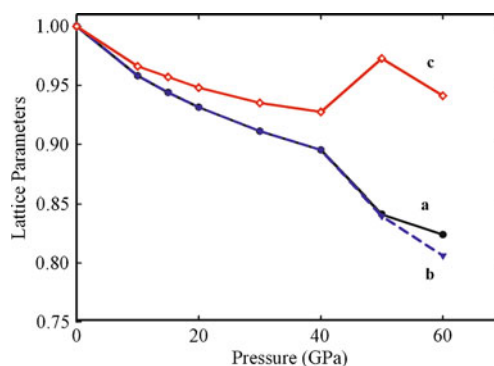


Fig. 6 Variations in simulation cell lengths as a function of pressure (the *y*-axis contains normalized values)

transformation mechanism might follow various paths or involve several intermediate states. Some structures, such as wurtzite or zinc-blende (ZnSe, BeO, ZnS, etc.), transform into the RS structure [8, 9, 22]. In previous studies, the transformation of ZnTe into the RS structure was not observed. Unlike other similar species, ZnTe transforms into the *Cmcm* structure. However, the limited number of structures considered so far has not yielded predictions of stable phases of ZnTe beyond *Cmcm* [23]. The fact that ZnTe is the only member of the IIB–VIA family for which the NaCl phase has not been observed at room temperature is related to the fact it possesses the lowest ionicity of that family [23]. In our previous study, we found that SnSe undergoes a second-order structural phase transition from threefold-coordinated *Pnma* to fivefold-coordinated *Cmcm* in the constant-pressure simulation [24]. In our current study, the lattice parameters calculated for the wurtzite phase were $a=4.254 \text{ \AA}$ and $c=6.989 \text{ \AA}$, in agreement with the theoretical value $a=4.234 \text{ \AA}$, while c/a is 1.643 [7]. The calculated lattice parameters were $a=6.047 \text{ \AA}$ for the zinc-blende phase, in agreement with the experimental value $a=6.089 \text{ \AA}$ [9]. The transition volumes at which the transition occurs were $V_i \sim 0.73V_0$ and $V_f \sim 0.60V_0$, with the equilibrium volume $V_0=27.7 \text{ \AA}^3/\text{atom}$. The volume reduction ($\Delta V/V_0$) during the phase transition from the WZ to *Cmcm* phases was found to be about 12.17%, comparable with the experimental result for the volume compression at the transition, which is on the order of 13% at the zinc blende to cinnabar transition, and 13% at the cinnabar to *Cmcm* transition [25]. The bulk modulus we calculated for the WZ state was 63.83 GPa, and its pressure derivative B'_0 was 4.4; these parameters were calculated to be 75.3 GPa and 3.9 for the *Cmcm* phase, and 63.4 GPa and 4.2 for the ZB phase ($B'_0=3.0$ in [11] and $B'_0=5.0$ in [25]). In general, our results agree with previous experimental and theoretical results. Our study shows that the initial wurtzite phase of ZnTe is fourfold coordinated, while the resultant *Cmcm* phase is sixfold coordinated.

Conclusions

We have used an *ab initio* constant-pressure MD technique within a generalized gradient approximation (GGA) to study the pressure-induced phase transition of wurtzite ZnTe. A first-order phase change into a *Cmcm* structure was successfully reproduced in the constant-pressure simulation. The WZ-to-*Cmcm* transformation mechanism of ZnTe is different from previously proposed mechanisms. Additionally, we investigated the stabilities of the wurtzite and zinc-blende phases based on energy–volume calculations. We also found that both structures have quite similar equations of state and transform into a *Cmcm* structure at 16 GPa using enthalpy calculations. Our theoretical calculations agree with the structural phase transformations of ZnTe observed experimentally. Our calculated transition phase, lattice parameters and bulk moduli are in agreement with experimental and theoretical data. We obtained the sixfold coordinated *Cmcm* structure of ZnTe in this study.

Acknowledgments We are grateful to Dr. Murat Durandurdu for his help. We are also grateful to the SIESTA group for making their code publicly available.

References

1. Frogley MD, Dunstan DJ, Palosz W (1998) *Phys Rev B* 107:537–541
2. Côté M, Zakharov O, Rubio A, Cohen ML (1997) *Phys Rev B* 55:13025–13031
3. Nelves RJ, McMahon MI, Wright NG, Allan DR (1994) *Phys Rev Lett* 73:1805–1808
4. Lee GD, Hwang C, Lee MH, Ihm J (1997) *J Phys Condens Matter* 9:6619–6631
5. Franco R, Mori-Sánchez P, Recio JM (2003) *Phys Rev B* 68:195208
6. McMahon MI, Nelves RJ (2005) *Phys Rev Lett* 95:215505
7. Yang JH, Chen S, Yin WJ, Gong XG, Walsh A, Wei SH (2009) *Phys Rev B* 79:245202
8. Durandurdu M (2009) *J Phys Condens Matter* 21:125403
9. Alptekin S, Durandurdu M (2009) *Solid State Commun* 149:345–348
10. Gupta SK, Kumar S, Auluck S (2009) *Physica B* 404:3789–3794
11. Onodera A, Ohtani A, Tsuduki S, Shimomura O (2008) *Solid State Commun* 145:374–378
12. Ordejón P, Artacho E, Soler JM (1996) *Phys Rev B* 53:10441–10444
13. Perdew JP, Burke K, Ernzerhof M (1996) *Phys Rev Lett* 77:3865–3868
14. Troullier N, Martins JM (1991) *Phys Rev B* 43:1993–2006
15. Monkhorst HJ, Pack JD (1976) *Phys Rev B* 13:5188–5192
16. Hundt R, Schön JC, Hannemann A, Jansen M (1999) *J Appl Crystallogr* 32:413–416
17. Mujica A, Needs RJ, Muñoz A (1995) *Phys Rev B* 52:8881–8892
18. Gangadharan R, Jayalakshmi V, Kalaiselvi J, Mohan S, Murugana R, Palanivel B (2003) *J Alloys Compd* 359:22–26
19. Christensen NE, Christensen OB (1986) *Phys Rev B* 33:4739–4746
20. Mizushima K, Yip S, Kaxiras E (1994) *Phys Rev B* 50:14952–14959
21. Martoňák R, Laio A, Parrinello M (2003) *Phys Rev Lett* 90:75503
22. Durandurdu M (2009) *J Phys Chem Solids* 70:645–649
23. Mujica A, Rubio A, Muñoz A, Needs RJ (2003) *Rev Mod Phys* 75:863–912
24. Alptekin S (2011) *J Mol Model*. doi:10.1007/s00894-011-1019-2
25. San-Miguel A, Polian A, Gautier M, Itié JP (1993) *Phys Rev B* 48:8683–8693

Quantitative structure activity relationships of some pyridine derivatives as corrosion inhibitors of steel in acidic medium

El Sayed H. El Ashry · Ahmed El Nemr · Safaa Ragab

Received: 10 January 2011 / Accepted: 3 June 2011 / Published online: 22 June 2011
© Springer-Verlag 2011

Abstract Quantum chemical calculations using the density functional theory (B3LYP/6-31G* DFT) and semi-empirical AM1 methods were performed on ten pyridine derivatives used as corrosion inhibitors for mild steel in acidic medium to determine the relationship between molecular structure and their inhibition efficiencies. Quantum chemical parameters such as total negative charge (*TNC*) on the molecule, energy of highest occupied molecular orbital (E_{HOMO}), energy of lowest unoccupied molecular orbital (E_{LUMO}) and dipole moment (μ) as well as linear solvation energy terms, molecular volume (*V*_i) and dipolar-polarization (π^*) were correlated to corrosion inhibition efficiency of ten pyridine derivatives. A possible correlation between corrosion inhibition efficiencies and structural properties was searched to reduce the number of compounds to be selected for testing from a library of compounds. It was found that theoretical data support the experimental results. The results were used to predict the corrosion inhibition of 24 related pyridine derivatives.

Keywords Pyridine derivatives · AM1 · B3LYP · Corrosion inhibition · DFT · Linear solvation energy

Introduction

The study of corrosion processes and their inhibition by organic inhibitors is a very active field of research [1–21]. Many researchers report that the inhibition effect mainly depends on some physicochemical and electronic properties of the organic inhibitor which relate to its functional groups, steric effects, electronic density of donor atoms, and orbital character of donating electrons, and so on [22, 23]. The effect of concentrations, functional groups and halide ions of quaternary ammonium inhibitors as well as the effect of some nitrogen- and sulfur-containing organic compounds such as substituted benzothiazoles and various types of organic sulfur-containing compounds on the corrosion of iron and steel have been studied [8–10]. The inhibition efficiency of 3,5-bis(*N*-pyridyl)-4-amino-1,2,4-triazoles [11–14], thiophenol, phenol and aniline [15], 2,5-bis(*N*-pyridyl)-1,3,4-thiadiazole [16], 2,5-bis(4-dimethylaminophenyl)-1,3,4-thiadiazole [17], 2-mercaptothiazoline and cetyl pyridinium chloride [18] have been reported. The influence of heterocyclic anils on corrosion inhibition of metals has also been reported [19–21]. The inhibiting mechanism is generally explained by the formation of a physically and/or chemically adsorbed film on the metal surface [24, 25]. It is well known that organic compounds which act as inhibitors are rich in heteroatoms, such as sulfur, nitrogen, and oxygen [26, 27]. These compounds and their derivatives are excellent corrosion inhibitors in a wide range of media and are selected essentially from empirical knowledge based on their macroscopic physico-chemical properties. The efficiency of an organic inhibitor of metallic corrosion does not only depend on the structure characteristics of the inhibitor, but also on the nature of the metal and environment. The selection of a suitable inhibitor for a

E. S. H. El Ashry · S. Ragab
Department of Chemistry, Faculty of Science,
Alexandria University,
Alexandria, Egypt

E. S. H. El Ashry
e-mail: eelashry60@hotmail.com

A. El Nemr (✉)
Environmental Division,
National Institute of Oceanography and Fisheries, Kayet Bey,
Alexandria, Egypt
e-mail: ahmedmoustafaelnemr@yahoo.com

particular system is a difficult task because of the selectivity of the inhibitors and wide variety of environments.

Recently, theoretical prediction of the efficiency of corrosion inhibitors has become very popular in parallel with the progress in computational hardware and the development of efficient algorithms which assisted the routine development of molecular quantum mechanical calculations [28].

Quantitative structure activity relationships (QSAR) has been a subject of intense interest in the field of medicinal chemistry in determining the molecular structure as well as elucidating the electronic structure and reactivity [29], but to a less extent in the field of corrosion [30–65]. The concept of assessing the efficiency of a corrosion inhibitor with the help of computational chemistry is to search for compounds with desired properties using chemical intuition, experience and a mathematically quantified and computerized form. Once a correlation between the structure and activity or property is found, any number of compounds, including those not yet synthesized, can be readily predicted employing computational methodology [66] via a set of mathematical equations which are capable of representing accurately the chemical phenomenon under study [67, 68].

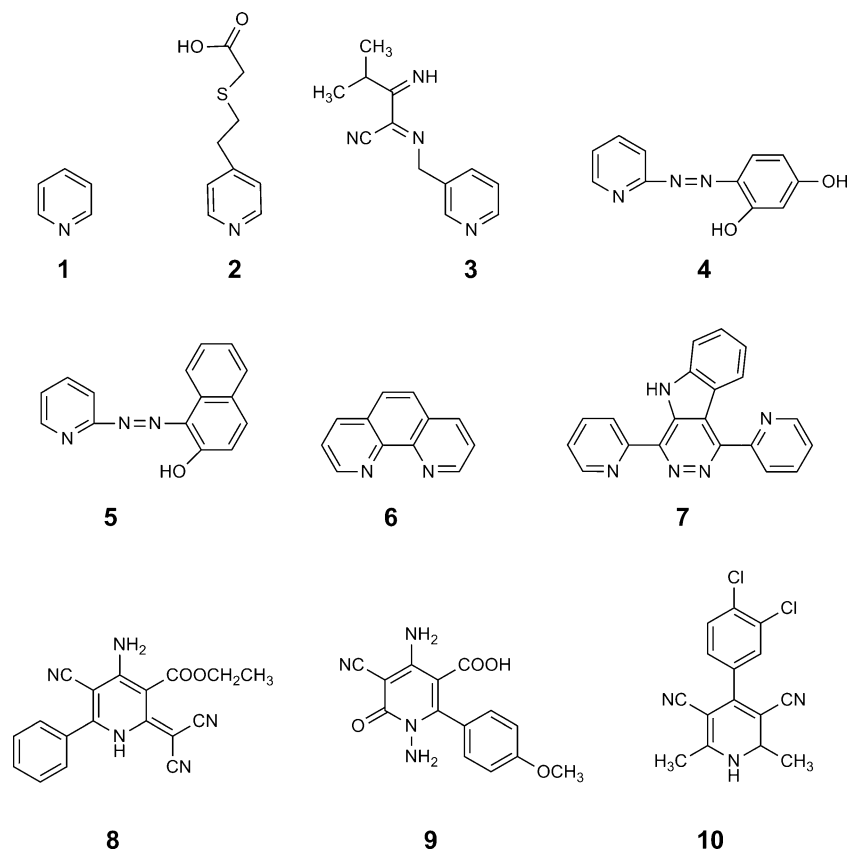
Since, the development of new corrosion inhibitors is a necessity; the aim of this work is to correlate the structural characteristics of some pyridine derivatives with their corrosion inhibition efficiencies at different concentrations of

inhibitors in aqueous acidic solutions. The development of equations for calculating the corrosion inhibition efficiencies may lead to a prediction of the inhibition efficiencies of some related derivatives in order to help in selecting compounds for testing from the large number of compounds that can be developed by the concept of combinatorial chemistry and constructed libraries of compounds. For this purpose relation between the inhibition efficiencies and quantum chemical calculation parameters, E_{HOMO} , E_{LUMO} , dipole moment, total negative charge on molecules and linear solvation energy relationship have been investigated.

Methods of calculations

Quantum calculations were carried out using wave function restricted-closed-shell AM1 semi-empirical SCF-MO methods and DFT (B3LYP/6-31G*) single point and structure optimized in the Gaussian 2003 program implemented in CS ChemOffice packet program version 11 for windows [69]. Calculations were performed on IBM compatible which is implemented on PC, Intel Pentium IV 2.8 GHz computer. AM1 and B3LYP quantum theoretical calculations were started without any geometry constraints for full geometry optimizations using program default calculation setting. Single point calculation was obtained for the full

Fig. 1 Compounds 1–10 which have experimental inhibition efficiencies



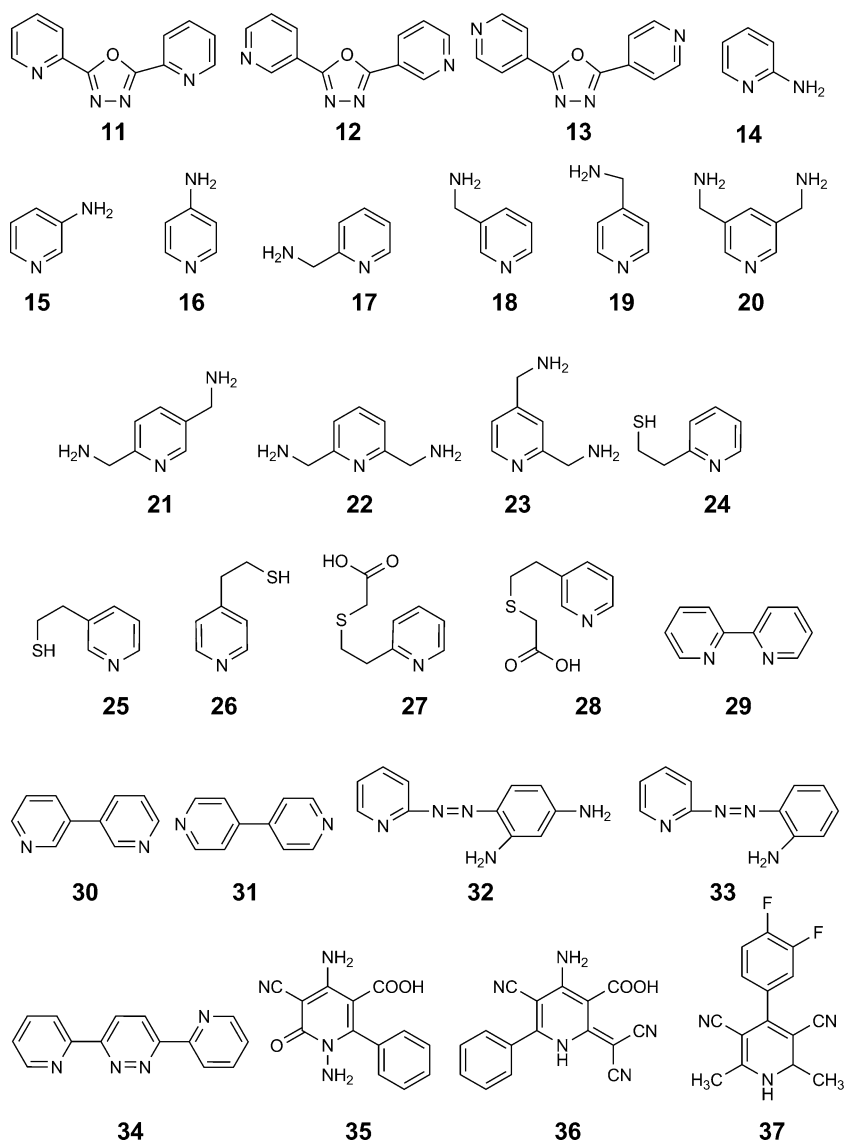
optimized AM1 structure using B3LYP/6-31G* method. The study of the effects of protonation and the effect of both intramolecular H-bonding and tautomerism on the inhibitor compounds are avoided in order to decrease the required time for calculation. The following quantum chemical indices, depending on the trial and error in the SPSS program in solving the non-linear equations: the energy of the highest occupied molecular orbital (E_{HOMO}), the energy of the lowest unoccupied molecular orbital (E_{LUMO}), the dipole moment (μ), and total negative charge (TNC : obtained by summation of negative charge on the atoms of the molecule) on the molecule, as well as the linear solvation energy relationships (LSER) parameters: intrinsic molecular volume ($V_i/100$) and dipolar-polarizability factor (π^*) were considered [70]. Surface area (\AA^2), volume (\AA^3), partition coefficient ($\log P$, the partition coefficient is a ratio of concentrations of un-ionized compound between the two solutions. To measure the partition

coefficient of ionizable solutes, the pH of the aqueous phase is adjusted such that the predominant form of the compound is un-ionized. The logarithm of the ratio of the concentrations of the un-ionized solute in the solvents is called $\log P$, refractivity (\AA^3) and polarizability (\AA^3) were calculated using QSAR calculation by HyperChem program version 8.07. Statistical analyses were performed using SPSS program version 15.0 for windows. Non-linear regression analyses were performed by unconstrained sum of squared residuals for loss function and estimation methods of Levenberg-Marquardt using SPSS program version 15.0 for windows.

Results and discussion

The pyridine derivatives 1–10 (Fig. 1) were reported as corrosion inhibitors for mild steel in acidic medium. Their

Fig. 2 Structure of compounds 11–13, which used for models validations and compounds 14–37, which used for prediction of inhibition efficiencies using models 6–9



inhibition efficiencies were reported as E_{exp} (%) based on the weight loss methods for compounds 1, 2 [71], 3 [72], 4 [73], 5 [74] and 6 [75], 7 [76], 8, 9, 10 [72], and 11, 12, 13 [77].

Quantum chemical parameters, of compounds 1–10, validated compounds 11–13 and the selected compounds 14–37 for predicting their efficiencies (Figs. 1 and 2; Tables 1, 2, 3), surface area (\AA^2), volume (\AA^3), $\log P$, refractivity (\AA^3), polarizability, (\AA^3) E_{HOMO} (eV), E_{LUMO}

(eV), μ (Debye), TNC and linear solvation energy parameters (LSER), the intrinsic (van der Waals) molecular volume V_i ($\text{cm}^3 \text{mol}^{-1}$) and dipolar-polarizability term (π^*) in the LSER model that scales the solute electrostatic stabilization of molecular charge by using methods of Hickey and Passino-Reader [70].

Correlation analysis between quantum parameters obtained and E_{exp} (%) for compounds 1–10 show signifi-

Table 1 Quantum chemical parameters (QSAR) of compounds 1 to 37 in gas phase using AM1 semi-empirical calculation and LSER parameters (V_i , π^*) calculation

compound	V_i cm^3/M	π^*	Vol \AA^3	$\log P$	Sur. Area \AA^2	Ref. \AA^3	Polariz. \AA^3	E_{HOMO} eV	E_{LUMO} eV	$E_{\text{L-H}}$ eV	μ Debye	TNC	χ	η	ΔN
1	0.472	0.87	317.66	0.12	199.03	27.39	9.730	-8.4320	0.4890	8.921	2.709	-1.165	4.897	5.035	0.209
2	1.551	1.83	616.23	-0.21	383.28	55.21	20.79	-9.0973	-0.1191	8.978	2.520	-2.043	4.608	4.489	0.266
3	1.361	1.82	679.33	2.44	379.23	65.71	24.45	-10.021	-0.5356	9.486	1.953	-1.675	5.278	4.743	0.181
4	1.068	1.87	634.63	0.14	329.20	69.08	22.72	-8.9068	-0.4298	8.477	1.519	-1.834	4.668	4.238	0.275
5	1.325	2.35	734.41	1.24	345.37	85.67	28.26	-8.5779	-0.8891	7.689	2.375	-1.817	4.734	3.844	0.295
6	0.942	1.85	559.10	0.52	244.51	61.83	21.38	-9.0414	-0.7186	8.323	3.024	-1.229	4.880	4.161	0.255
7	1.621	2.41	875.72	-0.98	364.58	113.4	37.16	-8.5433	-1.2186	7.325	3.557	-2.247	4.881	3.662	0.289
8	1.701	3.63	862.16	-1.88	434.11	98.16	34.58	-8.6522	-1.4600	7.192	6.976	-3.125	5.056	3.596	0.270
9	1.446	2.68	749.22	-3.29	376.48	83.33	29.47	-9.2214	-0.9111	8.310	5.951	-3.324	5.066	4.155	0.233
10	1.526	2.35	796.06	0.12	462.81	86.03	31.03	-8.7091	-0.8579	7.851	6.920	-1.785	4.784	3.926	0.282
11	1.089	2.61	664.22	2.97	323.49	66.62	24.53	-9.5820	-1.3910	8.191	1.672	-1.805	5.489	4.096	0.185
12	1.089	2.61	660.76	-0.04	323.70	66.08	24.31	-9.4550	-1.2910	8.164	0.172	-2.756	5.373	4.082	0.199
13	1.089	2.61	660.58	-0.04	322.36	66.08	24.01	-9.9110	-1.4090	8.501	3.100	-2.853	5.660	4.251	0.158
14	0.535	1.00	352.57	-0.16	195.66	30.66	11.08	-8.5707	0.4867	9.057	1.892	-1.077	4.042	4.529	0.327
15	0.535	1.05	350.78	-1.47	196.54	31.01	11.08	-8.0888	0.8249	8.914	2.862	-0.979	3.632	4.457	0.378
16	0.535	1.10	350.74	-1.59	196.37	30.94	11.08	-8.9040	0.4036	9.308	3.282	-1.093	4.250	4.654	0.295
17	0.631	1.19	408.65	-0.98	227.54	36.38	12.91	-9.0983	0.1915	9.290	2.511	-1.291	4.453	4.645	0.274
18	0.631	1.24	408.89	-1.07	226.74	35.10	12.91	-9.0046	-0.0952	8.909	1.820	-1.292	4.550	4.455	0.275
19	0.631	1.29	404.44	-1.07	227.38	35.10	12.91	-9.0341	0.0544	9.089	0.808	-1.106	4.490	4.544	0.276
20	0.795	1.61	491.32	-2.14	257.91	42.89	16.16	-9.4857	-0.0527	9.433	2.573	-1.394	4.769	4.716	0.236
21	0.795	1.56	496.28	-2.05	264.57	44.16	16.10	-9.5528	-0.0704	9.482	1.405	-1.393	4.812	4.741	0.231
22	0.795	1.51	498.97	-0.93	259.92	44.21	16.10	-9.5340	0.0030	9.537	3.506	-1.392	4.765	4.769	0.234
23	0.795	1.61	494.78	-1.02	257.61	42.94	16.10	-9.5756	-0.0141	9.562	1.024	-1.409	4.795	4.781	0.231
24	0.771	1.33	479.84	1.30	297.30	44.39	16.40	-8.7358	-0.0597	8.676	3.599	-1.087	4.398	4.338	0.300
25	0.771	1.27	477.00	0.11	297.40	44.32	16.40	-9.0026	0.0918	9.094	2.878	-1.057	4.455	4.547	0.280
26	0.771	1.32	477.83	0.11	297.67	44.32	16.40	-9.1275	0.0860	9.214	1.570	-1.071	4.521	4.607	0.269
27	1.014	1.83	618.92	0.97	382.76	42.89	16.10	-9.0016	-0.1391	8.862	3.290	-2.016	4.570	4.431	0.274
28	1.014	1.88	615.10	-0.21	378.38	55.21	20.79	-9.0587	-0.1266	8.932	2.028	-2.029	4.593	4.466	0.270
29	0.853	1.74	520.60	0.85	262.77	53.88	18.68	-9.0656	-0.3862	8.679	2.982	-1.229	4.726	4.340	0.262
30	0.853	1.84	508.99	-0.62	260.46	52.41	18.68	-9.3645	-0.6214	8.743	3.228	-1.222	4.993	4.372	0.230
31	0.853	1.94	509.08	-0.74	259.29	52.33	18.68	-9.9397	-0.7484	9.191	0.040	-1.225	5.344	4.596	0.180
32	1.124	1.92	655.74	-1.50	309.50	72.85	24.15	-8.2730	-0.1650	8.108	3.002	-2.205	4.219	4.054	0.343
33	1.056	1.83	622.65	0.21	302.39	69.30	22.80	-8.5093	-0.5067	8.003	1.468	-1.718	4.508	4.001	0.311
34	1.212	2.89	705.17	-0.72	345.45	83.60	26.92	-9.1577	-0.3751	8.783	2.994	-1.588	4.766	4.391	0.254
35	1.310	2.55	679.96	-2.30	316.23	76.95	27.00	-9.4564	-1.0920	8.364	3.978	-2.994	5.274	4.182	0.206
36	1.509	3.23	773.45	-2.25	387.81	88.65	30.91	-8.7110	-1.5336	7.177	7.662	-2.871	5.122	3.589	0.262
37	1.383	2.45	729.30	-0.64	415.99	76.86	26.99	-8.7400	-0.8591	7.881	7.668	-1.897	4.800	3.940	0.279

Table 2 Quantum chemical parameters of compounds using gas phase single point B3LYP/6-31G*

Compound	E_{HOMO} Hartree	E_{LUMO} Hartree	$E_{\text{H-L}}$ Hartree	μ Debye	TNC	χ	η	ΔN
1	-0.25831	-0.0395	-0.219	2.6520	-1.468	0.149	0.109	0.495
2	-0.24167	-0.0479	-0.194	1.2867	-5.347	0.145	0.097	0.580
3	-0.25801	-0.0932	-0.165	3.0845	-5.136	0.176	0.082	0.495
4	-0.21627	-0.0746	-0.142	2.0801	-4.653	0.145	0.071	0.789
5	-0.18866	-0.0991	-0.090	2.8240	-7.207	0.144	0.045	1.266
6	-0.24425	-0.0651	-0.179	3.9414	-7.298	0.155	0.090	0.573
7	-0.20094	-0.0871	-0.114	4.2983	-8.421	0.144	0.057	0.994
8	-0.23384	-0.1055	-0.128	7.1264	-3.817	0.170	0.064	0.682
9	-0.24061	-0.0811	-0.160	7.6375	-11.639	0.161	0.080	0.604
10	-0.23601	-0.0896	-0.146	9.4887	-14.452	0.163	0.073	0.645
11	-0.25816	-0.1221	0.136	0.2166	-1.906	0.190	0.068	0.493
12	-0.25430	-0.1042	0.150	1.4703	-3.992	0.179	0.075	0.520
13	-0.27452	-0.1009	0.174	1.4002	-4.517	0.188	0.087	0.401
14	-0.22564	-0.0365	-0.189	2.2977	-2.246	0.131	0.095	0.667
15	-0.22516	-0.0235	-0.202	3.6637	-1.582	0.124	0.101	0.659
16	-0.25947	-0.0105	-0.249	4.2917	-2.999	0.135	0.124	0.491
17	-0.26361	-0.0098	-0.254	3.0229	-1.748	0.137	0.127	0.475
18	-0.27538	-0.0122	-0.263	2.4606	-2.422	0.144	0.132	0.431
19	-0.26210	-0.0094	-0.253	1.5900	-3.457	0.136	0.126	0.481
20	-0.23641	-0.0016	-0.235	3.2194	-2.524	0.119	0.117	0.589
21	-0.23108	-0.0056	-0.226	1.4693	-3.913	0.118	0.113	0.616
22	-0.23191	-0.0129	-0.219	3.9556	-3.822	0.122	0.109	0.616
23	-0.22966	-0.0137	-0.216	0.9234	-4.358	0.122	0.108	0.628
24	-0.23613	-0.0434	-0.193	4.0462	-2.967	0.140	0.096	0.610
25	-0.23108	-0.0296	-0.202	1.4693	-3.913	0.130	0.101	0.629
26	-0.25228	-0.0305	-0.222	2.1890	-3.276	0.141	0.111	0.522
27	-0.24255	-0.0433	-0.199	3.6457	-4.878	0.143	0.100	0.574
28	-0.24716	-0.0482	-0.199	2.9455	-5.044	0.148	0.100	0.551
29	-0.24483	-0.0592	-0.186	3.8579	-2.475	0.152	0.093	0.567
30	-0.25636	-0.0242	-0.232	4.2370	-1.963	0.140	0.116	0.504
31	-0.26686	-0.0030	-0.264	0.0545	-2.468	0.135	0.132	0.464
32	-0.20567	-0.0646	-0.141	3.9109	-5.257	0.135	0.071	0.866
33	-0.21255	-0.0810	-0.132	1.8735	-4.165	0.147	0.066	0.840
34	-0.21314	-0.0902	-0.123	3.6516	-4.535	0.152	0.061	0.859
35	-0.25481	-0.0885	-0.166	5.0664	-11.612	0.172	0.083	0.515
36	-0.23601	-0.1079	-0.128	7.9844	-5.168	0.172	0.064	0.666
37	-0.22641	-0.0882	-0.138	9.6869	-7.612	0.157	0.069	0.723

cant correlations ($p < 0.001$) for some of the studied quantum parameters. In some correlation analysis we omit one or two points, in order to get better correlation, due to incompatibility of these points with the other points. Trends for inhibition efficiency for compounds 1–10 is not as simple as we think because there are many parameters with and against the correlation but there are relations which can be seen in Figs. 3, 4, 5, some have positive and some have negative correlations with experimental inhibition efficiency. The experimental inhibitions exhibit good correlations with the LSER and QSAR parameters obtained by Hyper-

Chem program (Fig. 3), for example, E_{exp} (%) has $R^2 = 0.975$ with $V_i/100$, $R^2 = 0.910$ with volume, $R^2 = 0.878$ with surface area, $R^2 = 0.851$ with polarizability, $R^2 = 0.785$ with π^* and $R^2 = 0.770$ with refractivity. Compounds with high values of $V_i/100$, π^* , volume, polarizability, surface area and refractivity show high corrosion inhibition efficiencies (Table 1). The inhibition efficiency increases if the compound can easily donate electrons from its HOMO level to the LUMO level of the metal whereby chelation on the metal surface occurs. The parameters obtained by AM1 calculation show good correlation with TNC ($R^2 = 0.867$), μ

Table 3 Quantum chemical parameters of compounds 1 to 37 using gas phase optimized B3LYP/6-31G*

Compound	E_{HOMO} Hartree	E_{LUMO} Hartree	$E_{\text{L-H}}$ Hartree	μ Debye	TNC	χ	η	ΔN
1	-0.2571	-0.0528	0.204	2.663	-1.269	0.155	0.102	0.501
2	-0.2507	-0.0580	0.193	2.865	-4.114	0.154	0.096	0.534
3	-0.2661	-0.1169	0.149	2.550	-4.843	0.192	0.075	0.440
4	-0.2283	-0.0964	0.132	2.054	-3.365	0.162	0.066	0.720
5	-0.2143	-0.0960	0.118	2.832	-3.867	0.155	0.059	0.863
6	-0.2423	-0.0662	0.176	3.849	-5.283	0.154	0.088	0.585
7	-0.2009	-0.0871	0.114	4.298	-8.421	0.144	0.057	0.994
8	-0.2390	-0.0963	0.143	4.895	-7.447	0.168	0.071	0.628
9	-0.1457	-0.1049	0.041	10.081	-11.363	0.125	0.020	3.235
10	-0.2360	-0.0896	0.146	9.504	-11.867	0.163	0.073	0.645
11	-0.2565	-0.0940	0.163	3.379	-1.904	0.175	0.081	0.505
12	-0.2646	-0.1160	0.149	3.540	-3.845	0.190	0.074	0.451
13	-0.2652	-0.1402	0.125	3.427	-4.417	0.203	0.063	0.436
14	-0.1190	-0.0423	0.077	2.197	-2.398	0.081	0.038	2.301
15	-0.1793	-0.0296	0.150	3.876	-1.943	0.104	0.075	1.021
16	-0.2315	-0.0211	0.210	4.656	-2.762	0.126	0.105	0.623
17	-0.2274	-0.0382	0.189	2.994	-3.063	0.133	0.095	0.658
18	-0.2354	-0.0422	0.193	2.461	-2.422	0.139	0.097	0.613
19	-0.2423	-0.0352	0.207	2.552	-3.004	0.139	0.104	0.572
20	-0.2320	-0.0365	0.195	3.615	-4.211	0.134	0.098	0.629
21	-0.2311	-0.0456	0.186	1.469	-3.913	0.138	0.093	0.641
22	-0.1412	-0.0026	0.139	4.166	-2.353	0.072	0.069	1.337
23	-0.2173	0.0407	0.258	1.322	-4.011	0.088	0.129	0.655
24	-0.2240	-0.0508	0.173	4.966	-2.720	0.137	0.087	0.692
25	-0.2526	-0.0405	0.212	3.415	-3.159	0.147	0.106	0.522
26	-0.2339	-0.0711	0.163	1.955	-3.037	0.153	0.081	0.643
27	-0.2309	-0.0640	0.167	2.098	-3.012	0.147	0.083	0.658
28	-0.2400	-0.0415	0.199	2.756	-2.946	0.141	0.099	0.587
29	-0.2502	-0.0365	0.214	3.570	-2.033	0.143	0.107	0.533
30	-0.2976	-0.0053	0.292	4.229	-1.969	0.151	0.146	0.362
31	-0.2654	-0.0761	0.189	0.001	-4.156	0.171	0.095	0.457
32	-0.2057	-0.0646	0.141	3.911	-5.257	0.135	0.071	0.866
33	-0.2125	-0.0810	0.132	1.878	-3.372	0.147	0.066	0.840
34	-0.2131	-0.0902	0.123	3.650	-2.912	0.152	0.061	0.859
35	-0.2448	-0.1094	0.135	5.082	-11.093	0.177	0.068	0.592
36	-0.2604	-0.0893	0.171	4.261	-9.539	0.175	0.086	0.482
37	-0.3083	-0.0166	0.292	10.591	-8.163	0.162	0.146	0.325

($R^2=0.853$), $E_{\text{LUMO}} - E_{\text{HOMO}}$ ($R^2=0.838$), E_{HOMO} ($R^2=0.802$), the number of transferred electrons (ΔN , $R^2=0.791$) and E_{LUMO} ($R^2=0.703$) (Fig. 4). The ΔN was calculated depending on the quantum chemical method [78, 79] as represent in Eq. 1:

$$\Delta N = \frac{(\chi_{\text{Fe}} - \chi_{\text{inh}})}{2(\eta_{\text{Fe}} + \eta_{\text{inh}})} \quad (1)$$

where χ_{Fe} and χ_{inh} denote the absolute electronegativity of iron and the inhibitor molecule, respectively; η_{Fe} and η_{inh} denote the absolute hardness of iron and the inhibitor

molecule, respectively. These quantities are related to electron affinity (A) and ionization potential (I)

$$\chi = (I + A)/2 \quad (2)$$

$$\eta = (I - A)/2 \quad (3)$$

I and A are related in turn to E_{HOMO} and E_{LUMO}

$$I = -E_{\text{HOMO}}$$

$$A = -E_{\text{LUMO}}$$

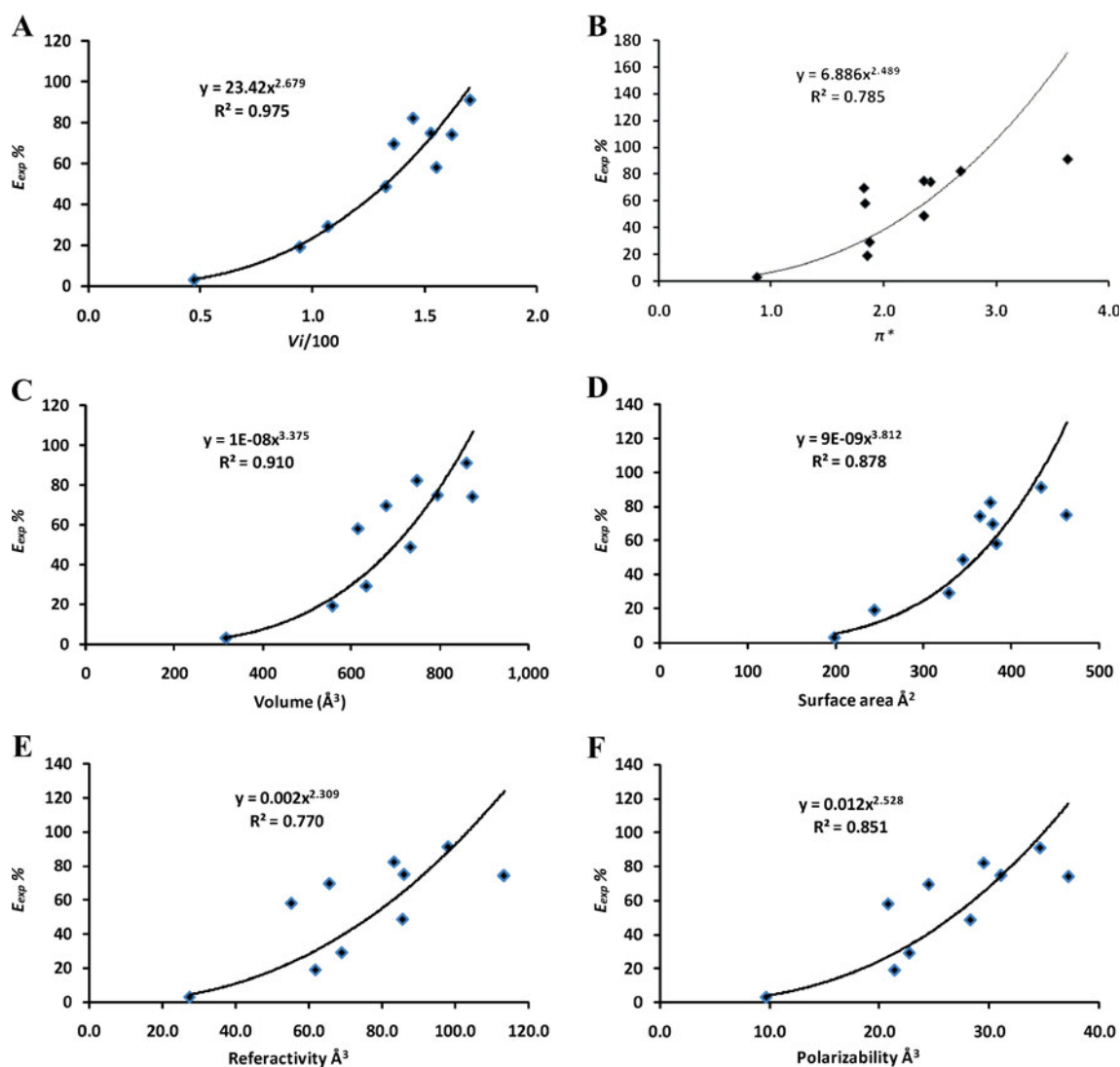


Fig. 3 Plotted of QSAR calculated parameters using HyperChem program version 8.07 versus E_{exp} %

Values of χ and η were calculated by using the values of I and A obtained from quantum chemical calculation. Using a theoretical χ value of 7 eV mol^{-1} ($0.30435 \text{ Hartree mol}^{-1}$) and η value of 0 eV mol^{-1} ($0.0 \text{ Hartree mol}^{-1}$) for iron atom [79] and ΔN , the fraction of electrons transferred from inhibitor to the iron molecule was calculated. Plot of E_{exp} % vs ΔN (Fig. 4f) clearly shows that the inhibition efficiency increased with the ΔN increase. According to other reports [78, 79], values of ΔN showed inhibition effect resulted from electrons donation. Agreeing with published study [78], the inhibition efficiency increased with increasing electron-donating ability at the metal surface. In this study, compounds 1–10 were the donors of electrons, and the iron surface was the acceptor. These compounds were bound to the iron surface and thus formed inhibition adsorption layer against corrosion.

The efficiency increases with the increase of μ , TNC and ΔN . Consequently, the compounds with lower energy gap ($E_{LUMO} - E_{HOMO}$), lower E_{LUMO} and higher E_{HOMO} have good efficiency (Fig. 4a-f).

Figure 5 shows the correlations between the experimental inhibition efficiency of compounds 1–10 and their quantum chemical parameters obtained by single point calculation using DFT (B3LYP/6-31G*). Strong correlation coefficient with experimental inhibition efficiency are obtained for E_{LUMO} ($R^2=0.879$), $E_{LUMO} - E_{HOMO}$ ($R^2=0.856$), μ ($R^2=0.847$), TNC ($R^2=0.843$), E_{HOMO} ($R^2=0.793$) and ΔN ($R^2=0.745$). Inhibition efficiency increases with the increase of E_{HOMO} and decrease of E_{LUMO} (Fig. 5a and b). On the other hand, the inhibition % decreases with the increase of the energy gap ($E_{LUMO} - E_{HOMO}$) (Fig. 5c), while increase in μ , TNC and ΔN led to increase of the inhibition efficiency (Fig. 5d, e, f). The correlation results

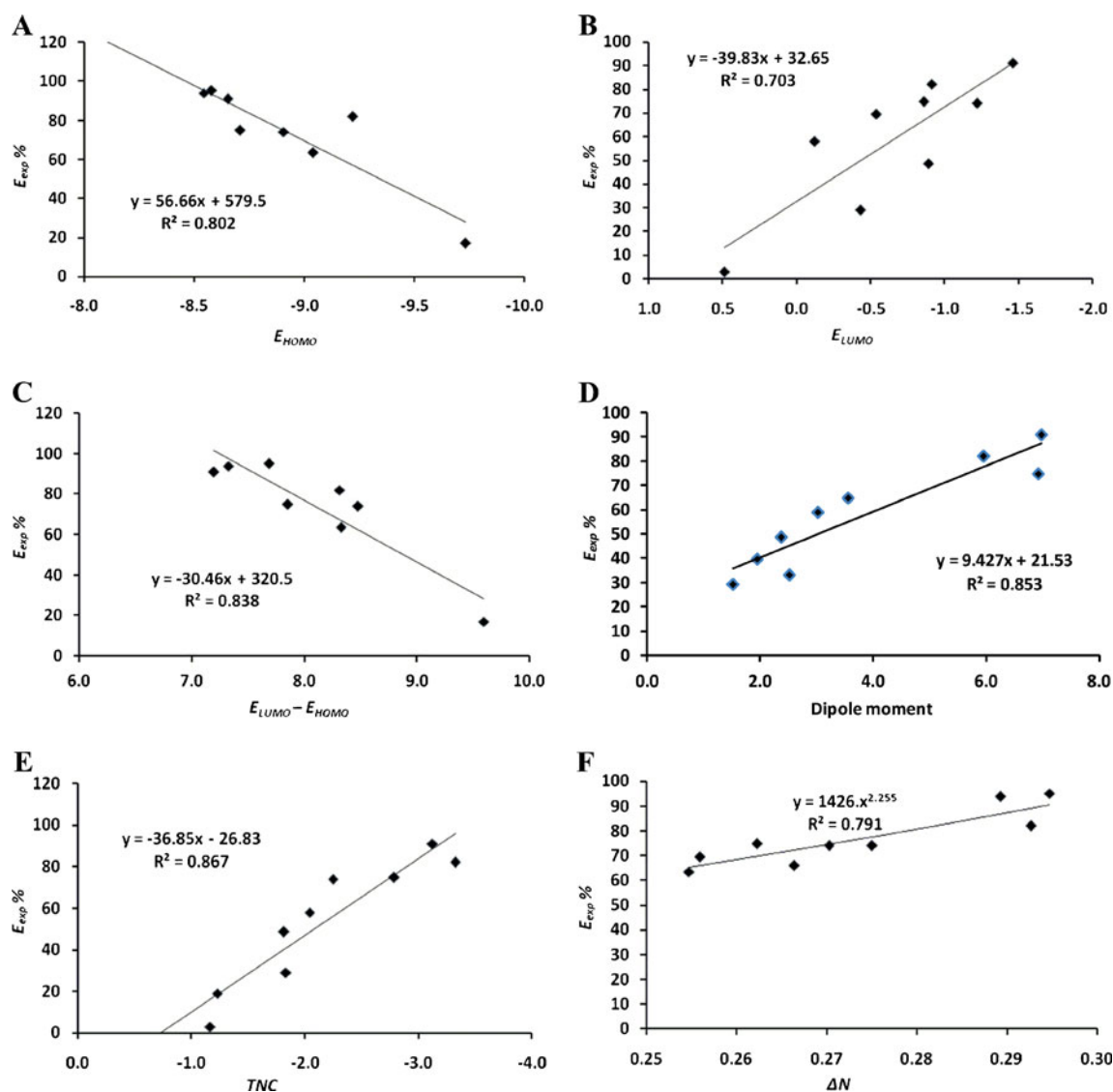


Fig. 4 Plotted of AM1 calculated parameters using ChemOffice program version 11 versus $E_{exp} \%$

obtained between the experimental inhibition for compounds 1–10 and the calculation parameters obtained by DFT (B3LYP/6-31G*) single point calculation are in agreement with that obtained by the AM1 calculation.

Figure 6 represents the results of the correlations between the experimental efficiency of compounds 1–10 and their quantum chemical parameters obtained by geometry optimized DFT (B3LYP/6-31G*) calculation. Significant correlation with the efficiency are obtained for $E_{LUMO} - E_{HOMO}$ ($R^2 = 0.911$), TNC ($R^2 = 0.900$), μ ($R^2 = 0.897$), E_{LUMO} ($R^2 = 0.831$), E_{HOMO} ($R^2 = 0.737$) and ΔN ($R^2 = 0.640$). Similarly as mentioned above, the efficiency is increased with the increase of E_{HOMO} and decrease of E_{LUMO} (Fig. 6a and b). On the other hand, the efficiency decreased with the increase of the energy gap ($E_{LUMO} - E_{HOMO}$) (Fig. 5c), while increased with the increase in μ ,

TNC and ΔN (Fig. 5d, e, f). These results for DFT (B3LYP/6-31G*) geometry optimized calculation are similar to the results obtained from single point and AM1 calculation, which may make the AM1 calculation preferred due to its short time calculation comparing with single point and optimized DFT calculations, which required much longer time to complete.

The effectiveness of 5, 7, 8, 9 and 10 compared to others as corrosion inhibitors may be attributed to the presence of an additional $-C=C-$ and/or $-N=N-$ groups in conjugation with the pyridine ring and $-C=N-$ beside high functionality group in 8, 9 and 10 that may play an important role in increasing molecular adsorption to metal surface through the extensive delocalization of the π -electrons of aromatic ring, $-C=C-$ group and lone pair of electrons on N and O atoms.

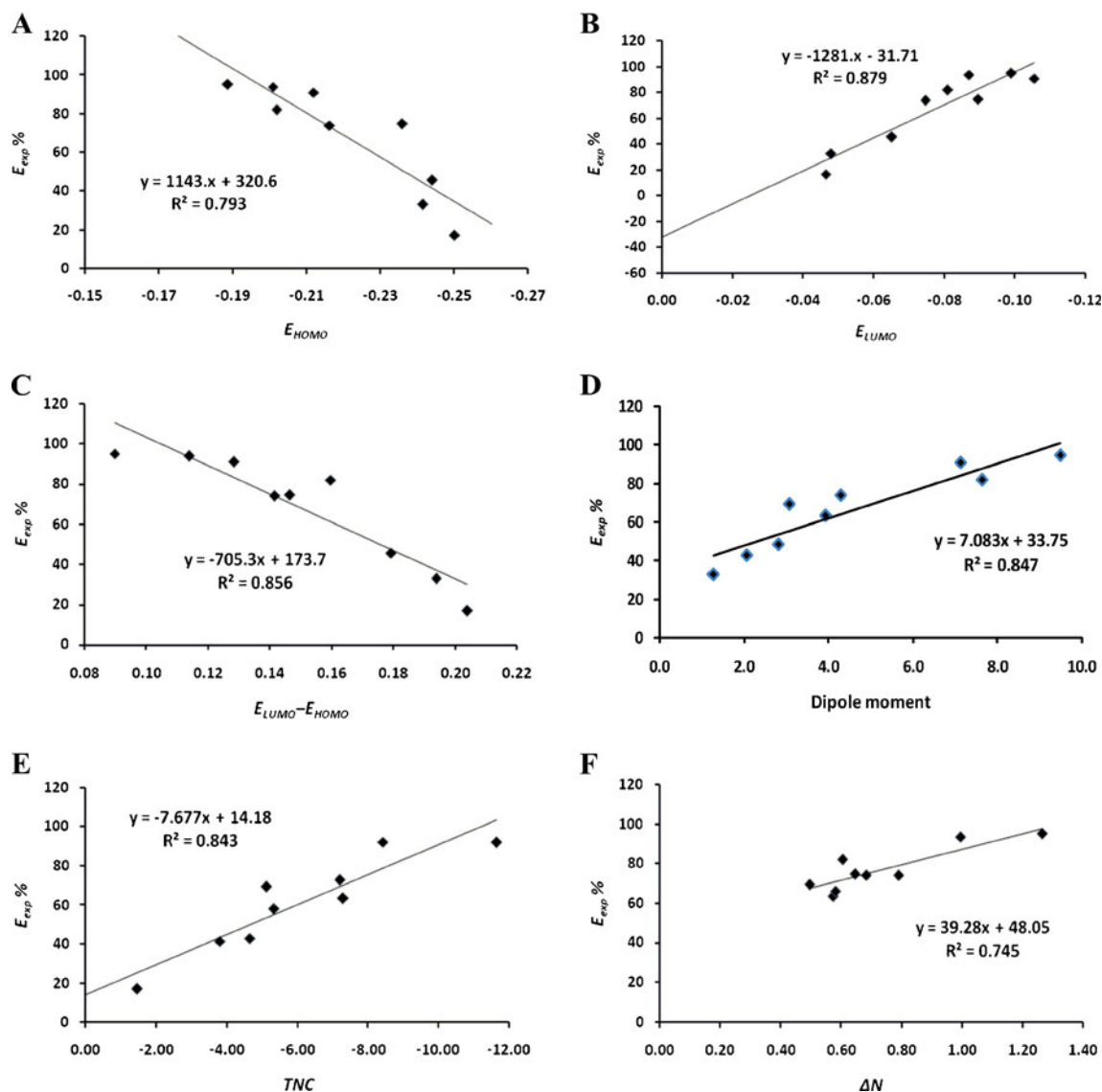


Fig. 5 Plotted of single point B3LYB-6-31G* calculated parameters using Gaussian 03 program versus $E_{exp} \%$

Although a number of satisfactory correlations [78, 80–88] have been reported for the inhibition efficiency of various inhibitors and selected quantum chemical parameters, no simple relation or direct trend relationship can be derived for such classes of inhibitors. A non-linear regression analysis was used to correlate quantum chemical parameters (E_{HOMO} , E_{LUMO} , μ , TNC , Volume, Surface area, $\log P$, polarizability, Refractivity), LSER (\bar{V}_i , π^*) and inhibitor concentrations (C_i) with the experimental inhibition efficiencies obtained by weight loss methods for compounds 1–10. Thus, a composite index of more than one quantum parameter which might affect the inhibition efficiency of molecules has been correlated with the experimental corrosion inhibition efficiencies.

The nonlinear equation has been derived from the linear model [88] which approximates corrosion inhibitor efficiency ($E_{cal} \%$):

$$E_{cal}(\%) = \text{Exp}^{A x_j C_i + B} \quad (4)$$

where A and B are constants obtained by regression analysis; x_j a quantum chemical index characteristic for the molecule j ; C_i denotes the experiment's concentration. Eq. 4 are used to derive Eq. 5 which is the non-linear model (NLM) proposed by Lukovits and co-worker [89] for studying the interaction of corrosion inhibitors with metal surface in acidic medium.

In the non-linear method of analysis, multiple regressions were performed on inhibition efficiencies for compounds 1–10 at concentrations range from 1 to

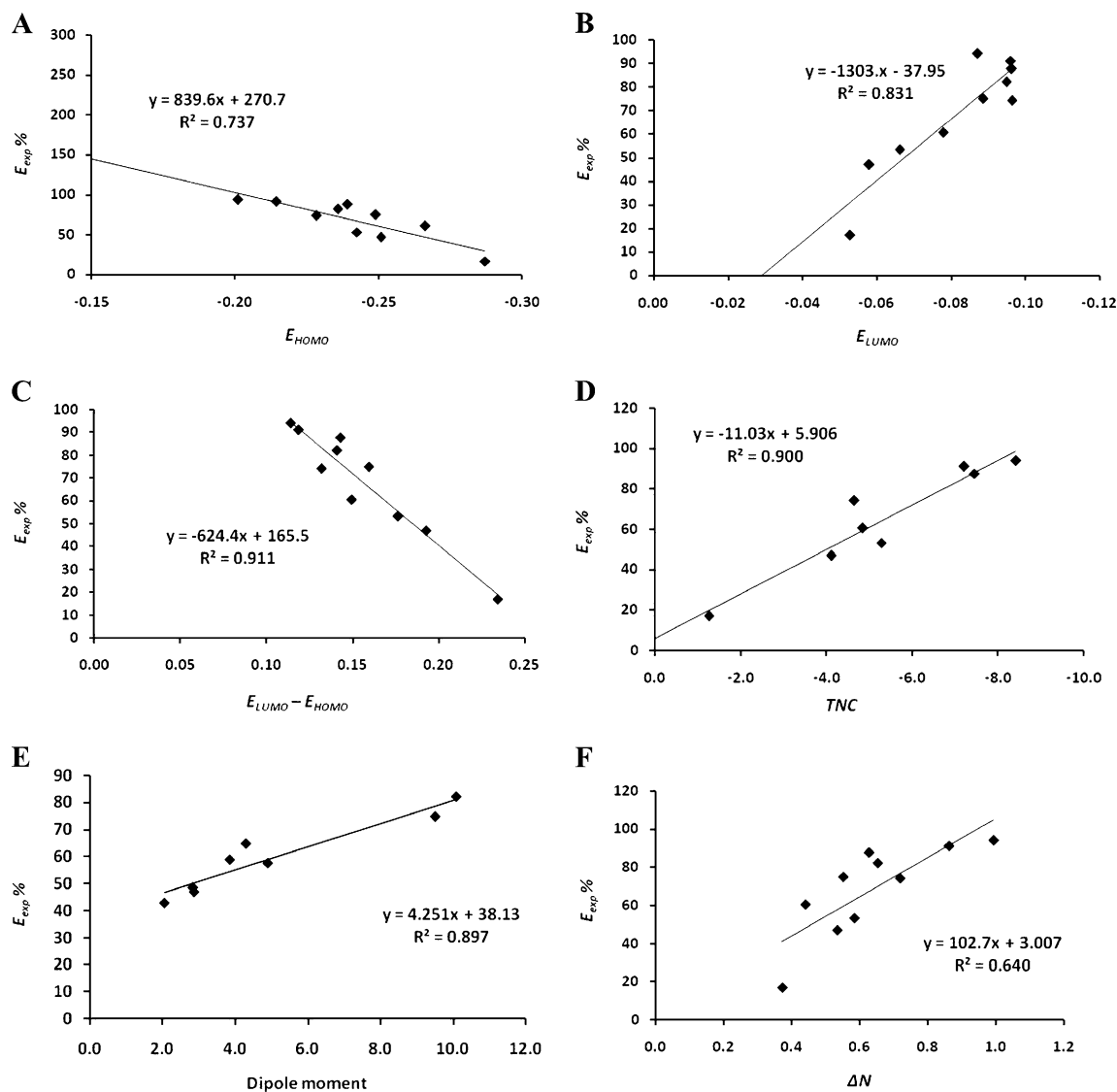


Fig. 6 Plotted of B3LYB/6-31G* optimized calculated parameters using Gaussian 03 Program versus E_{exp} %

500,000 μM . Trial and error methods were used to obtain the best model. Non-linear Eqs. 6 and 7 were obtained for AM1 where x_j is a composite index of selected quantum chemical parameters E_{HOMO} , E_{LUMO} , TNC , $\log P$, polarizability, surface area and $V_i/100$ in Eq. 6, and E_{HOMO} , E_{LUMO} , TNC , $\log P$, polarizability and $V_i/100$ in Eq. 7. Calculated efficiencies from such equations at different concentrations of compounds 1–10, illustrated good correlation with experimental efficiencies (E_{exp} %) with correlation coefficients $R^2=0.967$ and 0.962 , respectively (Fig. 7a and b).

The non-linear model Eq. 8 proposed for single point DFT (B3LYP/6-31G*) calculation on compounds 1–10 show also significant correlation between E_{exp} (%) and E_{cal} (%), $R^2=0.966$ (Fig. 7c). The x_j represented a composite index of selected quantum parameters E_{HOMO}/E_{LUMO} , π^* , Surface area, $\log P$, $TNC/V_i/100$ and polarizability/refractivity in Eq. 8.

The non-linear model for geometry optimized DFT (B3LYP/6-31G*) are represented in Eq. 9 with correlation coefficient $R^2=0.966$ (Fig. 7d) between the E_{exp} (%) and E_{cal} (%), where x_j is the quantum parameters E_{HOMO}/E_{LUMO} , π^* , surface area, $\log P$, $TNC/V_i/100$, polarizability/refractivity and volume/ μ .

Table 4 represents the E_{cal} % obtained at different concentrations from the four predicted models, Eqs. 6–9. The results obtained from the different methods of calculations are found to be very close to the E_{exp} % which is very clear in Table 4 and Fig. 8. On the other hand, significant correlation coefficient are obtained between the E_{exp} % and the average E_{cal} % obtained from Eqs. 6–9 ($R^2=0.970$) (Fig. 9).

$$E_{cal}(\%) = \frac{\text{Exp}^{(Ax_i+B)C_i}}{1 + \text{Exp}^{(Ax_i+B)C_i}} \times 100 \quad (5)$$

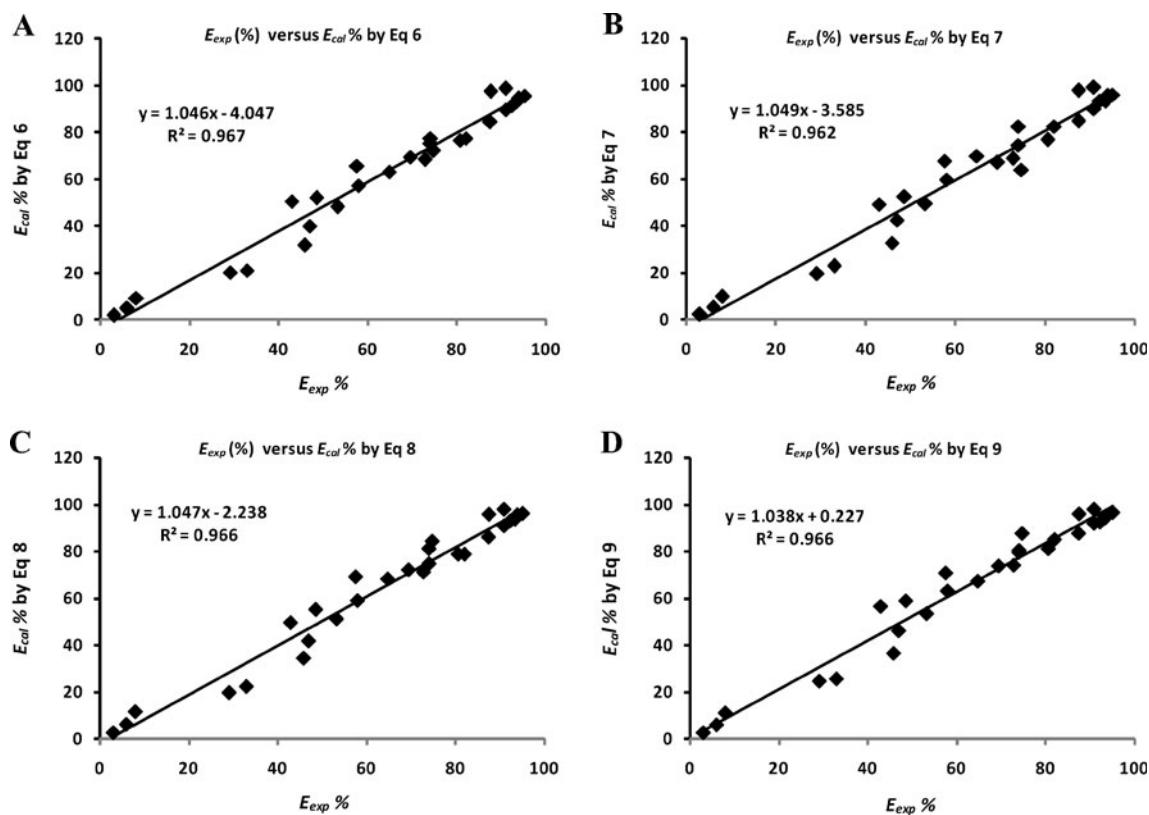


Fig. 7 Plot of E_{exp} % versus E_{cal} % produced from predicted Eqs. 6–9

AM1 Calculation

$$E_{cal}\% = \frac{\text{Exp}^{(2.94 \times E_{HOMO} + 7.083 \times E_{LUMO} - 5.525 \times \frac{w}{100} + 2.746 \times \log P + 0.687 \times \text{Polariz} + 0.005 \times \text{Sur.Area} - 8.175 \times TNC - 2.09) \times C_i}}{1 + \text{Exp}^{(2.94 \times E_{HOMO} + 7.083 \times E_{LUMO} - 5.525 \times \frac{w}{100} + 2.746 \times \log P + 0.687 \times \text{Polariz} + 0.005 \times \text{Sur.Area} - 8.175 \times TNC - 2.09) \times C_i}} \times 100 \quad (6)$$

$$E_{cal}\% = \frac{\text{Exp}^{(-4.869 \times \frac{w}{100} + 2.746 \times \log P + 0.708 \times \text{Polariz} + 2.923 \times E_{HOMO} + 7.38 \times E_{LUMO} - 8.407 \times TNC - 2.122) \times C_i}}{1 + \text{Exp}^{(-4.869 \times \frac{w}{100} + 2.746 \times \log P + 0.708 \times \text{Polariz} + 2.923 \times E_{HOMO} + 7.38 \times E_{LUMO} - 8.407 \times TNC - 2.122) \times C_i}} \times 100 \quad (7)$$

B3LYP/6-31G* single point calculation

$$E_{cal}\% = \frac{\text{Exp}^{(38.769 \times \frac{E_{LUMO}}{E_{HOMO}} - 4.543 \times \pi^* + 0.028 \times \text{Sur.Area} - 1.406 \times \log P + 0.482 \times \frac{TNC}{100} + 38.719 \times \frac{\text{Polariz}}{\text{Ref}} - 29.166) \times C_i}}{1 + \text{Exp}^{(38.769 \times \frac{E_{LUMO}}{E_{HOMO}} - 4.543 \times \pi^* + 0.028 \times \text{Sur.Area} - 1.406 \times \log P + 0.482 \times \frac{TNC}{100} + 38.719 \times \frac{\text{Polariz}}{\text{Ref}} - 29.166) \times C_i}} \times 100 \quad (8)$$

$R^2=0.966$

Table 4 Experimental inhibition efficiency obtained using weight loss of compounds 1 to 13 and calculated inhibition efficiency obtained by the proposed equations for AM1, semiempirical calculation and DFT (B3LYP/6-31G*)

Comp. No.	Conc. (μM)	E_{exp} (%)	E_{cal} (%):		B3LYP/31-6G*		Average of E_{cal} %	SD-1	SD-2
			AM1		Single point	Optimized			
			Eq 6	Eq 7	Eq 8	Eq 9			
1	100	3.00	2.00	2.09	2.54	2.38	2.25	0.25	0.53
	250	6.00	4.85	5.07	6.12	5.75	5.45	0.59	0.39
	500	8.00	9.26	9.65	11.54	10.87	10.33	1.06	1.65
2	100	33.00	21.05	22.74	22.27	25.44	22.88	1.85	7.16
	250	47.00	40.00	42.39	41.74	46.04	42.54	2.54	3.15
	500	58.00	57.15	59.54	58.89	63.05	59.66	2.48	1.17
3	100	69.50	69.50	66.85	72.04	73.76	70.54	3.02	0.73
4	25	29.11	20.24	19.43	19.68	24.46	20.95	2.36	5.77
	100	42.95	50.37	49.10	49.50	56.43	51.35	3.42	5.94
	300	74.10	75.28	74.32	74.62	79.53	75.94	2.43	1.30
5	5	48.62	52.01	52.19	55.11	58.81	54.53	3.19	4.18
	10	72.89	68.43	68.58	71.06	74.06	70.53	2.64	1.67
	15	80.68	76.48	76.61	78.65	81.07	78.20	2.16	1.75
	25	87.50	84.42	84.51	85.99	87.71	85.66	1.55	1.30
	40	91.00	89.66	89.72	90.76	91.95	90.52	1.08	0.34
	60	93.47	92.86	92.91	93.64	94.48	93.47	0.76	0.00
	100	95.20	95.59	95.62	96.09	96.62	95.98	0.48	0.55
6	50000	45.87	31.79	32.62	34.32	36.37	33.78	2.03	8.55
	100000	53.28	48.25	49.19	51.10	53.34	50.47	2.25	1.99
7	10	64.80	63.22	69.74	68.14	67.24	67.09	2.78	1.62
	20	74.10	77.47	82.17	81.05	80.41	80.28	2.01	4.37
	60	92.30	91.16	93.26	92.77	92.49	92.42	0.90	0.08
	80	93.50	93.22	94.86	94.48	94.26	94.20	0.70	0.50
	100	94.00	94.50	95.84	95.53	95.36	95.31	0.57	0.92
8	5	57.60	65.78	67.52	69.04	70.74	68.27	2.12	7.55
	100	87.60	97.47	97.65	95.71	96.03	96.71	0.99	6.44
	200	91.00	98.72	98.81	97.81	97.97	98.33	0.51	5.18
9	100	82.10	77.35	82.32	78.72	85.09	80.87	3.51	0.87
10	100	74.80	72.16	63.79	84.23	87.68	76.97	11.02	1.53
11	200	79.20	63.31	57.30	71.47	73.52	66.40	7.50	9.05
	400	82.80	77.53	72.85	83.36	84.74	79.62	5.49	2.25
	800	83.80	87.35	84.30	90.93	91.74	88.58	3.43	3.38
	1200	85.50	91.19	88.95	93.76	94.34	92.06	2.48	4.64
	12	200	83.00	72.90	72.72	85.24	76.38	76.81	5.87
12	400	87.00	84.32	84.21	92.03	86.61	86.79	3.66	0.15
	800	89.00	91.50	91.43	95.85	92.82	92.90	2.07	2.76
	1200	91.60	94.16	94.12	97.19	95.10	95.14	1.44	2.51
	13	200	37.00	35.20	34.90	41.88	30.51	35.62	4.69
13	400	48.30	52.08	51.74	59.03	46.75	52.40	5.05	2.90
	800	54.00	68.49	68.20	74.24	63.72	68.66	4.31	10.37
	1200	76.20	76.53	76.28	81.21	72.48	76.63	3.57	0.30

Average: The average of the calculated inhibition efficiency obtained by the proposed models; SD 1: standard deviation between the results obtained by the proposed equations; SD 2: standard deviation between the experimental inhibition efficiency obtained by weight loss and average of the calculated inhibition efficiency obtained by the proposed equations

B3LYP/6-31G* Optimized structure calculation

$$E_{\text{cal}}\% = \frac{\text{Exp} \left((29.401 \times \frac{E_{\text{LUMO}}}{E_{\text{HOMO}}} - 2.283 \times \pi^* + 0.056 \times \text{Sur.Area} - 0.093 \times \log P + 1.687 \times \frac{\text{TNC}}{100} - 86.367 \times \frac{\text{Polariz}}{\text{Ref}} - 0.047 \frac{V_{\text{ol}}}{\mu} + 17.321) \times C_i \right)}{1 + \text{Exp} \left((29.401 \times \frac{E_{\text{LUMO}}}{E_{\text{HOMO}}} - 2.283 \times \pi^* + 0.056 \times \text{Sur.Area} - 0.093 \times \log P + 1.687 \times \frac{\text{TNC}}{100} - 86.367 \times \frac{\text{Polariz}}{\text{Ref}} - 0.047 \frac{V_{\text{ol}}}{\mu} + 17.321) \times C_i \right)} \times 100 \quad (9)$$

$$R^2 = 0.966$$

The average of E_{cal} (%) obtained for each concentration is correlated with the E_{exp} (%) with correlation coefficient $R^2 = 0.970$ (Fig. 9) with standard deviations ranged from ± 0.00 to ± 8.55 . Moreover, the standard deviations between the E_{cal} obtained from different proposed model are ranged from ± 0.25 to ± 11.02 (Table 4). The high correlation coefficients ($R^2 = 0.970$) obtained from the four proposed QSAR Eqs. 6–9 are strong evidence for the participation of quantum parameters E_{HOMO} , E_{LUMO} , dipole moment, TNC, surface area and polarizability and LSER in the inhibition efficiency for compounds 1–10.

The higher the value of E_{HOMO} of the inhibitor indicates the ability of the molecules to offer electrons to d orbitals of metallic steel and the higher inhibition efficiency of the inhibitor for steel in acidic medium. The coefficients of E_{LUMO} in Eqs. 6–9 are negative indicating that d orbitals of steel gave electrons to the d orbital of the pyridine derivative compounds leading to the presence of a feedback bond. The presence of feedback bond leads to an increase in chemical adsorption of inhibitor molecules on the steel surface and so increases the inhibition efficiency of these compounds (Tables 1 and 3).

Compounds 11–13 were computed using the same methods applied to compounds 1–10 and the computational data obtained (Table 1–3) are used to test the validity

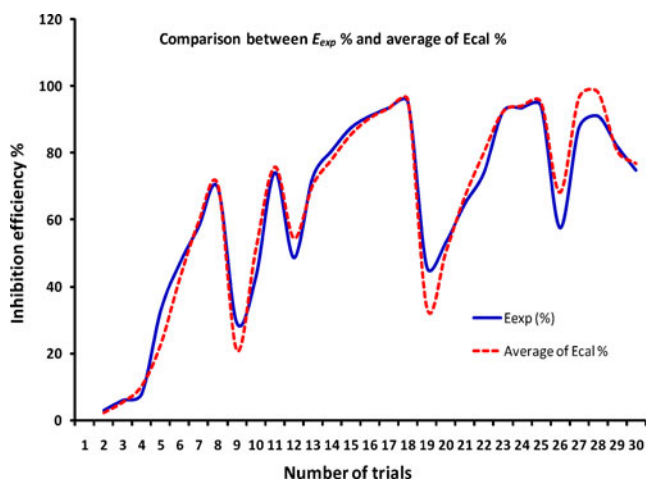


Fig. 8 Comparison between the predicted inhibition efficiencies obtained from Eqs. 6–9

of Eqs. 6–9 to predict inhibition efficiency (Table 4). Surprising, the inhibition efficiency results obtained are in good agreements with the experimental data obtained by weight loss and EIS methods [77]. The standard deviations between the results obtained for compounds 11–13 by Eqs. 6–7 are ranged between ± 1.44 and ± 7.5 as well as the standard deviations between the average of the results obtained and the experimental data are ranged between ± 0.15 and ± 10.37 , which represent the applicability of all the proposed models (Table 4).

The above results have attracted our attention to predict the corrosion inhibition of some analogues of the pyridine derivatives 1–10 in order to reduce the number of tested compounds for inhibition efficiency. Consequently, the inhibition efficiency can be treated as a controlled property via the change of electronic properties of compounds by changing their functional groups. Thus, 24 proposed pyridine derivatives 11–34 (Fig. 2) were subjected to similar methods of quantum calculations (Tables 1–3). The four proposed models (Eqs. 6–9) were applied at concentration 50 mM and the results are reported in Table 5. All four equations give a very near corrosion inhibition for most of the 24 proposed compounds with standard deviation range from ± 0.11 to ± 11.18 (Table 5, Fig. 10). Pyridine derivatives 11–20 show E_{cal} % lower than pyridine even at higher concentration except compounds 11 and 12 which have E_{cal} % more than pyridine. However, introduction of NH_2 or CH_2NH_2 substituted group to pyridine

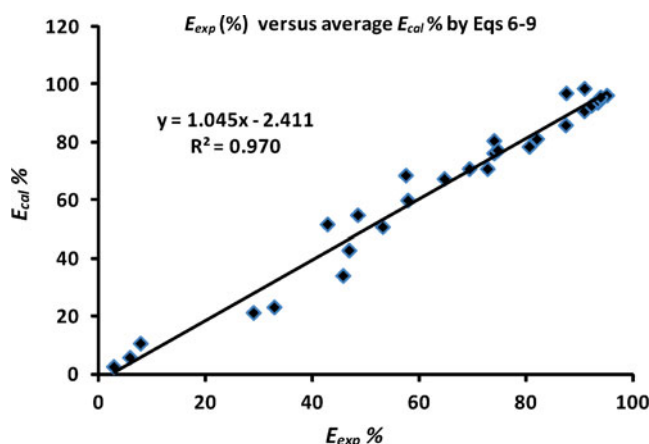


Fig. 9 Plot of E_{exp} % versus average E_{cal} % of the predicted inhibition from Eqs. 6–9

Table 5 Calculated inhibition efficiency [E_{cal} (%)] for compounds 14–37 obtained by the proposed models for AM1 semiempirical calculation and B3LYP/6-316G*

Comp. No.	Conc. (mM)	E_{cal} (%):		B3LYP/6-31G*		Average	SD-1
		AM1		Single point	Optimized		
		Eq 6	Eq 7	Eq 8	Eq 9		
14	50	72.71	74.90	74.18	72.24	73.51	1.24
15	50	59.76	63.97	71.21	62.37	64.33	4.90
16	50	1.23	1.35	5.32	2.35	2.56	1.90
17	50	9.31	9.54	10.31	11.32	10.12	0.91
18	50	1.37	1.29	11.71	6.44	5.20	4.96
19	50	0.80	0.76	3.39	0.70	1.41	1.32
20	50	0.24	0.24	10.87	1.59	3.24	5.13
21	50	0.22	0.22	7.82	0.00	2.07	3.84
22	50	7.68	7.78	6.46	10.02	7.99	1.48
23	50	5.50	5.63	5.57	0.00	4.17	2.78
24	50	97.28	96.41	84.28	99.73	94.42	6.90
25	50	58.74	52.74	68.88	76.09	64.11	10.40
26	50	51.47	45.45	69.05	65.25	57.80	11.18
27	50	99.96	99.94	86.10	89.60	93.90	7.13
28	50	99.96	99.95	97.68	90.25	96.96	4.60
29	50	76.10	74.82	78.81	90.74	80.12	7.27
30	50	0.41	0.36	10.03	22.65	8.36	10.55
31	50	0.02	0.02	0.25	0.00	0.07	0.12
32	50	99.99	99.99	99.84	99.77	99.90	0.11
33	50	99.77	99.78	99.92	93.68	98.29	3.07
34	50	97.55	97.58	99.81	100.00	98.73	1.36
35	50	98.58	98.99	99.28	92.49	97.34	3.25
36	50	98.77	98.80	100.00	96.99	98.64	1.24
37	50	97.55	96.71	99.99	98.13	98.09	1.39

decrease the inhibition efficiency. While, addition of $\text{CH}_2\text{CH}_2\text{SH}$ function group to pyridine increase the inhibition efficiency and the substitution at *ortho*- and *meta*-position gave enhanced efficiency more than *para*-position. Change the position of substitution in compound 2 to have 27 and 28 which led to an enhanced calculated

inhibition efficiency. Converting compound 6 to 29 led to an enhanced efficiency while converting 6 to 30 and 31 led to loss of efficiency. Substituting two OH groups in 4 by two NH_2 groups or by only one NH_2 group enhanced the calculated efficiency. Replacing two Cl atoms in 10 by two F atoms (compound 34) led to loss of efficiency at low concentration while it enhanced at high concentration.

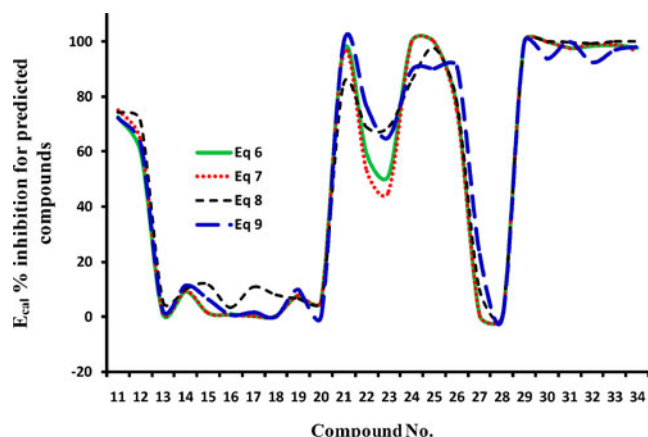


Fig. 10 Comparison between the results obtained from Eqs. 6–9 for compounds 14–37

Conclusions

A comparison of the inhibition effectiveness of the pyridine derivatives indicated that their inhibition effect has been closely related to orbital energies (E_{HOMO} and E_{LUMO}), dipole moment, polarizability, surface area, TNC and LSER parameters V_i and π^* . Inhibition efficiency of pyridine derivatives mainly increases when the E_{HOMO} of inhibitors increases, and E_{LUMO} and energy gap decrease. Increase in dipole moment, TNC and ΔN led to increase of the inhibition efficiency. A composite index of more than five quantum chemical parameters and one of LSER parameter should be included in the proposed models. Highly

significant multiple correlation coefficient ($R^2 > 0.96$) was obtained between experimental and calculated efficiencies for all the proposed models.

Correlation between experimental efficiencies obtained by weight loss and the QSAR parameter gave good correlation coefficient for some of the parameters obtained by AM1 and DFT calculation.

These correlations may be useful in designing new inhibitors by selecting the substitutions on the parent molecules. QSAR approach may be used to find the optimal group of parameters that might predict the structure and molecule suitability to be an inhibitor. The quantum mechanical approach may well be able to foretell molecule structures that are better for corrosion inhibition purposes if it is taken into account that the effect depends only on the inhibitor molecule properties.

References

- Eddy NO, Ebenso EE, El Nemr A, El Ashry ESH (2009) *J Mol Model* 15:1085–1092
- El Ashry ESH, El Nemr A, Essawy SA, Ragab S (2006) *Electrochim Acta* 51:3957–3968
- El Ashry ESH, El Nemr A, Essawy SA, Ragab S (2006) *Chem Phys An Ind J* 1:41–62
- El Ashry ESH, El Nemr A, Essawy SA, Ragab S (2006) *ARKIVOC* 11:205–220
- El Ashry ESH, El Nemr A, Essawy SA, Ragab S (2008) *Prog Org Coat* 61:11–20
- Khamis E, Bellucci F, Latanision RM, El Ashry ESH (1991) *Corros* 47:677–672
- Khamis E, El Ashry ESH, Ibrahim AK (2000) *Brit Corros J* 35:150–154
- Schweinsberg DP, Ashworth V (1988) *Corros Sci* 28:539–545
- Raicheva SN, Aleksiev BV, Sokolova EI (1993) *Corros Sci* 34:343–350
- Cheng XL, Ma HY, Chen SH, Yu R, Chen X, Yao ZM (1998) *Corros Sci* 41:321–333
- Quraishi MA, Khan MAW, Jamal D, Ajmal M, Muralidharan S, Iyer SVK (1996) *J Appl Electrochem* 26:1253–1258
- Quraishi MA, Khan MAW, Jamal D, Ajmal M, Muralidharan S, Iyer SVK (1997) *Brit Corros J* 32:72–81
- Mernari B, Attari HE, Traisnel M, Bentiss F, Lagrenée M (1998) *Corros Sci* 40:391–399
- Hluchan V, Wheeler BL, Hackerman N (1988) *Werkst Korros* 39:512
- Bouayed M, Rabaa H, Schiri A, Saillard JY, Ben Bachir A, Le Beuze A (1998) *Corros Sci* 41:501–517
- El Azhar M, Mernari B, Traisnel M, Gengembre L, Bentiss F, Lagrenée M (2001) *Corros Sci* 43:2229–2238
- Bentiss F, Traisnel M, Lagrenée M (2001) *J Appl Electrochem* 31:41–48
- Wang L, Yin GJ, Yin G (2001) *Corros Sci* 43:1197–1202
- Quraishi MA, Khan MAW, Ajmal M, Muralidharan S (1995) *Port Electrochim Acta* 13:63–71
- Quraishi MA, Khan MAW, Ajmal M, Muralidharan S, Iyer SVK (1997) *Corros* 53:475–481
- Quraishi MA, Khan MAW, Ajmal M (1996) *Methods Mater* 43:5–8
- Quraishi MA, Sardar R (2002) *Mater Chem Phys* 78:425–431
- Stupnišek-Lisac E, Podbršček S, Soric T (1994) *J Appl Electrochem* 24:779–784
- Touhami F, Aouniti A, Abed Y, Hammouti B, Kerit S, Ramdani A, Elkacemi K (2000) *Corros Sci* 42:929–940
- Tang L, Li X, Li L, Mu G, Liu G (2006) *Coat Technol* 201:384–388
- Hosseini M, Mertens SFL, Ghorbani M, Arshadi MR (2003) *Mater Chem Phys* 78:800–808
- Subramanyam NC, Sheshardi BS, Mayanna SA (1993) *Corros Sci* 34:563–571
- Domenicano A, Hargittai I (1992) *Accurate molecular structures, their determination and importance*. Oxford University Press, New York
- Kraka E, Cremer D (2000) *J Am Chem Soc* 122:8245–8264
- Bentiss F, Traisnel M, Vezin H, Lagrenée M (2003) *Corros Sci* 45:371–380
- Martinez S, Stagljar I (2003) *J Mol Struct THEOCHEM* 640:167–174
- Emreguel KC, Hayvali M (2006) *Corros Sci* 48:797–812
- Yurt A, Bereket G, Kivrak A, Balaban A, Erk B (2005) *J Appl Electrochem* 35:1025–1032
- Emreguel KC, Abduelkadir AA, Atakol O (2005) *Mater Chem Phys* 93:325–329
- Talati JD, Desai MN, Shah NK (2005) *Anti-Corros Methods Mater* 52:108–117
- Baghaei F, Sheikhshoae I, Dadgarnezhad A (2005) *Asian J Chem* 17:224–232
- Aytac A, Oezmen U, Kabasakaloglu M (2005) *Mater Chem Phys* 89:176–181
- Dadgarnezhad A, Sheikhshoae I, Baghaei F (2004) *Asian J Chem* 16:1109–1116
- Dadgarnezhad A, Sheikhshoae I, Baghaei F (2004) *Anti-Corros Methods Mater* 51:266–273
- Yurt A, Balaban A, Kandemir SU, Bereket G, Erk B (2004) *Mater Chem Phys* 85:420–426
- Emregul KC, Atakol O (2004) *Mater Chem Phys* 83:373–379
- Sudha P, Menaka S, Elango KP (2004) *Trans SAEST* 39:17–21
- Emregul KC, Kurtaran R, Atakol O (2003) *Corros Sci* 45:2803–2817
- Emregul KC, Atakol O (2003) *Mater Chem Phys* 82:188–193
- Bilgic S, Caliskan N (2001) *J Appl Electrochem* 31:79–83
- Grigorev VP, Shpanko SP, Nassar AF, Dymnikova OV (2000) *Russ J Electrochem (Translation of Elektrokimiya)* 36:1157–1162
- Shokry H, Sekine I, Yuasa M et al. (1998) *Zairyo to Kankyo*:47–451
- Donya AP, Bratchun VI, Pakter MK, Shalimova MA (1997) *Prot Metals (Translation of Zashchita Metallov)* 33:377–381
- Quraishi MA, Ajmal M, Shere S (1996) *Bull Electrochem* 12:523–527
- Mohamed AK, Bekheit MM, Fouda AS (1991) *Bulletin de la Societe Chimique de France (May-June)* 331–340
- Desai MN, Desai MB, Shah CB, Desai SM (1986) *Corros Sci* 26:827–837
- Cruz J, Martínez R, Genesca J, García-Ochoa E (2004) *J Electroanal Chem* 566:111–121
- Gece G, Bilgic S (2009) *Corros Sci* 51:1876–1878
- Jamalizadeh E, Hosseini SMA, Jafari AH (2009) *Corros Sci* 51:1428–1435
- Arslan T, Kandemirli F, Ebenso EE, Love I, Alemu H (2009) *Corros Sci* 51:35–47
- Jamalizadeh E, Jafari AH, Hosseini SMA (2008) *J Mol Struct THEOCHEM* 870:23–30
- Yana Y, Li W, Caia L, Houb B (2008) *Electrochim Acta* 53:5953–5960
- Roque JM, Pandiyan T, Cruz J, García-Ochoa E (2008) *Corros Sci* 50:614–624

59. Ju H, Kai ZP, Li Y (2008) *Corros Sci* 50:865–871
60. Lebrini M, Lagrenée M, Vezin H, Traisnel M, Bentiss F (2007) *Corros Sci* 49:2254–2269
61. Tang Y, Yang X, Yang W, Chen Y, Wan R (2010) *Corros Sci* 52:242–249
62. Bereket G, Öğretir C, Özşahin Ç (2003) Quantum chemical studies on the inhibition efficiencies of some piperazine derivatives for the corrosion of steel in acidic medium. *J Mol Struct THEOCHEM* 663:39–46
63. Kandemirli F, Sagdinc S (2007) *Corros Sci* 49:2118–2130
64. Zhang SG, Lei W, Xia MZ, Wang FY (2005) *J Mol Struct THEOCHEM* 732:173–182
65. Gece G (2008) *Corros Sci* 50:2981–2992
66. Karelson M, Lobanov V (1996) *Chem Rev* 96:1027–1043
67. Hinchliffe A (1994) *Modelling Molecular Structures*. John Wiley & Sons, New York
68. Hinchliffe A (1999) *Chemical Modelling From Atoms to Liquids*. Wiley, New York
69. CS Chemoffice Pro for Microsoft Windows, Cambridge Scientific Computing Inc, 875 Massachusetts Avenue, Suite 61, Cambridge MA 2139, USA
70. Hickey JP, Passino-Reader DR (1991) *Environ Sci Technol* 25:1753–1760
71. Bouklah M, Ouassini A, Hammouti B, El Idrissi A (2005) *Appl Surf Sci* 250:50–56
72. Abd El-Maksoud SA, Fouda AS (2005) *Mater Chem Phys* 93:84–90
73. Tang L, Li X, Li L, Mu G, Liu G (2006) *Surf Coat Technol* 201:384–388
74. Tang L, Li X, Li L, Qua Q, Mua G, Liu G (2005) *Mater Chem Phys* 94:353–359
75. Li X, Tang L, Li L, Mu G, Liu G (2006) *Corros Sci* 48:308–321
76. Bentiss F, Gassama F, Barbry D, Gengembre L, Vezin H, Lagrenee M, Traisnel M (2006) *Appl Surf Sci* 252:2684–2691
77. Lebrini M, Bentiss F, Vezin H, Lagrenee M (2005) *Appl Surf Sci* 252:950–958
78. Lukovits I, Kalman E, Zucchi F (2001) *Corros* 57:3–8
79. Sastri VS, Perumareddi JR (1997) *Corros* 53:617–622
80. Lukovits I, Palfi K, Bako I, Kalman E (1997) *Corros* 53:915–919
81. Li SL, Wang YG, Chen SH, Yu R, Lei SB, Ma HY, Liu DX (1999) *Corros Sci* 41:1769–1782
82. Bereket G, Hur E, Öğretir C (2002) *J Mol Struct THEOCHEM* 578:79–88
83. Fang J, Li J (2002) *J Mol Struct THEOCHEM* 593:179–184
84. Xiao-Ci Y, Hong Z, Ming-Dao L, Hong-Xuan R, Lu-An Y (2000) *Corros Sci* 42:645–653
85. Cruz J, Garcia-Ochoa E, Castro M (2003) *J Electrochem Soc* 150:B26
86. Khalil N (2003) *Electrochim Acta* 48:2635–2640
87. Khaled KF, Babic-Samardzija K, Hackerman N (2004) *J Appl Electrochem* 34:697–705
88. Khaled KF, Babic-Samardzija K, Hackerman N (2005) *Electrochim Acta* 50:2515–2520
89. Lukovits I, Kalman E, Palinkas G (1995) *Corros* 51:201–205

Homology modeling of the structure of acyl coA:isopenicillin N-acyltransferase (IAT) from *Penicillium chrysogenum*. IAT interaction studies with isopenicillin-N, combining molecular dynamics simulations and docking

Liliana Moreno-Vargas · Jose Correa-Basurto ·
Rachid C. Maroun · Francisco J. Fernández

Received: 7 March 2011 / Accepted: 30 May 2011 / Published online: 22 June 2011
© Springer-Verlag 2011

Abstract In the last step of penicillin biosynthesis, acyl-CoA:isopenicillin N acyltransferase (IAT) (E.C. 2.3.1.164) catalyzes the conversion of isopenicillin N (IPN) to penicillin G. IAT substitutes the α -amino adipic acid side chain of IPN by a phenylacetic acid phenolate group (from phenylacetyl-CoA). Having a three-dimensional (3D) structure of IAT helps to determine the steps involved in side chain exchange by identifying the atomic details of substrate recognition. We predicted the IAT 3-D structure (α - and β -subunits), as well as the manner of IPN and phenylacetyl-CoA bind to the mature enzyme (β -subunit). The 3D IAT prediction was achieved by homology modeling and molecular docking in different snapshots, and refined by molecular dynamic simulations. Our model can reasonably interpret the results of a number of experi-

ments, where key residues for IAT processing as well as strictly conserved residues most probably involved with enzymatic activity were mutated. Based on the results of docking studies, energies associated with the complexes, and binding constants calculated, we identified a site located in the region generated by $\beta 1$, $\beta 2$ and $\beta 5$ strands, which forms part of the central structure of β -subunit, as the potential binding site of IPN. The site comprises the amino acid residues Cys103, Asp121, Phe122, Phe123, Ala168, Leu169, His170, Gln172, Phe212, Arg241, Leu262, Asp264, Arg302, Ser309, and Arg310. Through hydrogen bonds, the IPN binding site establishes interactions with Cys103, Leu169, Gln172, Asp264 and Arg310. Our model is also validated by a recently revealed crystal structure of the mature enzyme.

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1143-z) contains supplementary material, which is available to authorized users.

L. Moreno-Vargas · F. J. Fernández (✉)
Laboratorio de Ingeniería Genética y Metabolismo Secundario,
Departamento de Biotecnología,
Universidad Autónoma Metropolitana-Iztapalapa,
D. F., México 09340, Mexico
e-mail: fjpg@xanum.uam.mx

J. Correa-Basurto
Laboratorio de Modelado Molecular y Bioinformática,
Escuela Superior de Medicina, Instituto Politécnico Nacional,
D. F., México 11340, Mexico

R. C. Maroun
Laboratoire de Neurobiologie et Pharmacologie Moléculaire,
Centre de Psychiatrie et de Neurosciences Broca-Sainte Anne
(INSERM U894),
Paris 75014, France

Keywords Penicillin biosynthesis · Acyl-CoA: isopenicillin-N acyltransferase · Ligand binding-site · Protein-ligand interaction · Homology modeling · Molecular docking · Long-term molecular dynamic simulations

Introduction

Isopenicillin N (IPN) conversion to penicillin G is the only step in fungi like *Penicillium chrysogenum* that is capable of producing hydrophobic penicillins. Acyl-CoA:isopenicillin-N acyltransferase (IAT) substitutes an α -amino adipic acid side chain from IPN for a phenylacetic acid phenolate group (from phenylacetyl-CoA). This substitution takes place either directly or in two steps. In the latter case, 6-animopenicillanic acid (6-APA) is formed as an intermediary [1–4]. In vitro, IAT not only converts IPN to penicillin

G, but can also perform several other transfer reactions, such as the exchange of acyl side chains among several penicillins, due to its transacylase penicillin activity [5]. IAT uses a wide variety of hydrophobic and hydrophilic acyl derivatives of coenzyme A (CoA) as substrates, as well as non-CoA related thioesters [6–8]. In some exceptional cases, it also accepts some precursors with modifications in the acetyl group and its benzene ring [9].

The gene for IAT, *penDE*, encodes a 357 residue proenzyme with a molecular weight of 40 kDa, which is processed autocatalytically to generate two polypeptides of 102 residues (α -subunit) and 255 residues (β -subunit) [10–14]. The α and β subunits constitute the mature form of the heterodimeric enzyme. The β -subunit contains in its extreme C-terminus a typical peroxisomal targeting sequence, known as ARL-COOH. If this sequence is removed, IAT is not fully compartmentalized in the peroxisome, with no penicillin production as a consequence [15–17]. This implies that a peroxisomal location is apparently essential for penicillin biosynthesis. Several co-expression experiments indicate that both subunits are necessary to maintain proper protein folding and enzymatic activity [17]. However, several studies [2–4, 10, 18, 19] indicate that the β -subunit has the acyltransferase activity, whereas the α -subunit does not play a major role in the catalytic activity. All these data expose the need to determine unambiguously the role that each subunit plays in IAT activity.

Site-directed mutagenesis experiments on IAT have identified several important residues required for autoprocessing capacity and enzymatic activity. In this context, residues Ser227 and Ser309 have been determined as very important for IAT enzymatic activity; however, the former is also required during the autoprocessing reaction [20]. Another important residue for the IAT processing reaction and preservation of its catalytic activity is Cys103 [21]. Additionally, a drastic reduction in enzymatic activity has been reported when introducing a non-polar group like Val at position 150, replacing a glycine (Gly150Val). The same is observed when placing a basic polar residue like Lys at position 258, originally Glu—a residue that stabilizes the structure of its conformational neighbors—because of its acid characteristics [18].

Due to the difficulty of obtaining higher protein concentrations, and the low solubility of the mature form of the enzyme (which translates into severe aggregation problems), efforts to elucidate the three-dimensional (3D) structure of IAT have been minimal. In this sense, Hensgens et al. [22] reported the crystallization and X-ray diffraction of the Cys103Ala IAT mutant. In a subsequent study, the latter group also reported an experimental technique to solve the aggregation problems that occur during IAT purification [23], but it was not possible to solve the 3D

structure in any case. When our manuscript was in preparation, three crystal structures (PDB entry 2X1C, resolution=1.85 Å; 2X1D, resolution=1.64 Å; 2X1E, resolution=2.00 Å) of the mature enzyme were revealed by Bokhove et al. [24]. To identify the atomic details of how IAT recognizes its substrates, and to determine how the enzyme catalyzes the side chain exchange among them, a 3D IAT model is required. It is known that for an enzyme to perform its function, it has to be properly folded [25]. This folded form has a 3D structure that is representative of a collection of folded forms with the same value of minimum energy. Anfinsen et al. [26] were the first to establish that unfolded proteins are inactive, and that their activity can be restored when they are refolded. This means that a linear amino acid sequence contains all the necessary information to maintain the correct folded form and its enzymatic activity. When a 3D structure of a protein cannot be obtained using experimental methods, it is possible to use computational tools in order to generate a model (from information contained in the primary structure) that is generally trustworthy [27–30]. To contribute to the knowledge of events governing the process of enzyme-substrate recognition, we generated a 3D IAT structure using homology modeling, and refined this structure by molecular dynamics (MD) simulations, as well as docking simulations to show its recognition by IPN and phenylacetyl-CoA. The structure was validated using the structures recently deposited into the Protein Data Bank (PDB).

Methods

40 kDa proenzyme sequence search

We searched for the linear IAT amino acid sequence (ID P15802) using the biological database Swiss-Prot (<http://www.expasy.org/sprot/>) [31].

Search for α and β subunit homologue structures

After amino acid sequences were retrieved, a search for homologous structures for both subunits, α and β , was performed using the PHYRE server (<http://www.imperial.ac.uk/phyre>) [32]. The sequences were sent separately to identify templates to be used in the homology modeling of each subunit.

Multiple alignment and template selection

Once sequences and templates from the target proteins were obtained, alignment was performed using CLUSTALW2 (<http://www.ebi.ac.uk/clustalw/>) [33] and BLAST (basic

local alignment search tool) [34] to determine percentage identity and similarity. Finally, to select the template for building the target protein, the criterion was to use high percentage sequence identity, and a high crystallographic resolution of the structure [35, 36].

Determination of the tendency of the α - and β -subunits to form secondary structures

The tendency of both IAT subunits to form secondary structures was evaluated at each amino acid residue using the servers APSSP (Advanced Protein Secondary Structure Prediction Server, <http://www.imtech.res.in/raghava/apssp2>) [37], Predict Protein (<http://www.predictprotein.org/>) [38], and JPred (<http://www.compbio.dundee.ac.uk/~www-jpred/>) [39–41].

Molecular modeling by homology

Both subunits were modeled separately using molecular modeling by homology. As templates, we used 3D structures from the PDB (<http://www.rcsb.org/pdb>). The α -subunit was modeled using the spatial coordinates of the atomic structure of a transcriptional regulator from the *tetR* family of *Streptomyces coelicolor* (PDB code: 2REK, 1.86 Å resolution) [42]. The β -subunit was solved using a *Clostridium perfringens* hydrolase (PDB code: 2BJG, 2.10 Å resolution) [43]. Molecular modeling was performed using three different programs: Swiss-Model (<http://swissmodel.expasy.org/>) [44], EsyPred3D (<http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/>) [45], and Modeller 8.2 (<http://salilab.org/modeller>) [46].

Geometry optimization of the proposed models

Once the 3D models were prepared, hydrogen atoms were added and side chain orientations optimized through energy minimization using the steepest descent method, employing 2,000 cycles using the CHARMM27 parameters found in NAMD [47–49].

Molecular dynamics simulation

The system was embedded in a water box with 0.2 M NaCl to relax the 3D IAT model. All water molecules reached every protein atom that was not hydrogen, with a distance of 3.8 Å. The entire system was subjected to an equilibration process before the MD. Equilibration consisted of an initial minimization of water molecules with the fixed atoms of the polypeptide backbone, followed by a minimization with an α carbon restriction, and finally short MD simulations (10 ps) to reduce the initial irregular contacts and fill up the empty ones. Next, the entire system,

under periodic boundary conditions in all three directions, was simulated at 310 K along 5 ns (Langevin dynamic and restricted constant pressure). From that moment on, the simulation was continued in the NTP ensemble for 70 ns. The trajectory was stored every picosecond, and analyzed using the VMD program [50]. To study the recognition energetics, and the manner of ligand binding, we took snapshots every 0.5 ns from the MD simulations. All MD simulations were performed using NAMD [48, 49], with the CHARMM27 force field [47]. The cut-off used for the long-term interaction was 10 Å.

Stereochemical quality evaluation of the models

Before running MD simulations, the coordinate files of the 3D models were sent to iMolTalk- Structural Bioinformatics Toolkit (<http://i.moltalk.org>) to produce a Ramachandran plot (φ and ψ angles), reflecting polypeptide chain distortion in the non-allowed region [51]. We also sent the coordinate files to MolProbity (<http://kinemage.biochem.duke.edu>) to identify side chains with less common conformations, possibly as a result of local protein tension [52]. The quality of the models was further validated using two additional tools: ANOLEA (<http://protein.bio.puc.cl/cardex/servers/anolea/>) [53, 54] and PROCHECK (<http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/>) [55, 56].

In silico generation of two β -subunit mutants, Gly150Val and Glu258Lys

In silico mutations of residues at positions Gly150 and Glu258 were introduced using HYPERCHEM (Version 8.0, Hypercube, <http://www.hyper.com>). We then searched for a low-energy geometry for the new residue by local minimization, using a second order method known as the Newton-Raphson Method, which imposes a finalization condition of a root mean square (RMS) gradient of less than 0.001 kcal mol⁻¹ Å⁻¹ [57].

Identification of normal vibrational modes in the β -subunit

To analyze intrinsic movements in the β -subunit structure, we determined the normal low-frequency vibrational modes using the Elnemo server. This tool allows visualization of the associated conformational changes in the molecule's natural mobility (<http://igs-server.cnrs-mrs.fr/elnemo/index.html>) [58–60]. The key parameters used were: DQMIN=-100, DQMAX=100, DQSTEP=20 and NRBL=auto. We requested a total of 100 low-frequency normal modes, and examined the essential characteristics of the low frequency normal modes of a protein, with collectivity of atomic movements, and the observed overlapping conformational changes.

Validation of the docking procedure

We used AutoDock 3.0.5, a computational simulation method that has rendered successful results [61]. This is a method for molecular recognition or docking used widely in the identification of binding sites of protein structures [62–65]. For our work, we validated the docking method using three different crystallographic complexes: streptavidin and biotin (PDB retried 1STP), triosephosphate isomerase (TPI) from *Trypanosoma brucei* and 3-phosphoglycerate (3-PA) (PDB retried 1III), TPI from *Trypanosoma cruzi* and C8 (PDB retried 1SUX).

Simulation of molecular recognition of IPN on the β -subunit

Once the method was standardized and the parameters defined, we worked with the complex of interest: the β -subunit and one of the natural substrates of IAT: isopenicillin-N (IPN). Docking studies were performed using AutoDock 3.0.5, employing the Lamarckian Genetic algorithm. The search space was restricted with a rectangular parallelepiped, which covers the entire protein. A rectangular grid ($126 \times 126 \times 126 \text{ \AA}$) with separated points at 0.375 \AA was generated. The docking parameters were of 100 tests, with 10 million energy evaluations per each test, and a population size of 100 individuals.

Preparation of the α -subunit

The β -subunit files used were those obtained when using molecular homology modeling. Before initiating docking evaluations, the Kollman charges for all protein atoms and polar hydrogens were assigned using AutoDockTools 1.5.0 (<http://autodock.scripps.edu>).

Preparation of the ligand

We used AutoDockTools 1.5.0 to add atomic charges assigned via the Gasteiger-Marsili formalism and hydrogen atoms located in polar atoms. The ligand structure with the minimum energy was obtained by density functional theory (DFT) calculations, using the B3LYP/6-31 G+(d, p) base, aided by Gaussian98 software [66].

Results and discussion

Sequence alignment and molecular homology modeling

According to PHYRE, the precision of template estimation for the α -subunit was over 45%, whereas for the β -subunit it was 100%. The percentages of identity and similarity of

the templates with the test sequences provided by BLAST were 33.3% and 51% for the α -subunit, respectively, and 37% and 50% for the β -subunit, respectively. Sequence alignment by CLUSTALW2 is shown in Fig. 1 [67, 68]. According to the literature, if protein sequences share more than 30% identity, one can be confident that they are structurally similar [36]. Since sequence identity percentages were over 30%, the quality of the models was quite high, as shown by the results of several stereochemical quality evaluations as mentioned below.

α -Subunit

The α -subunit model shows three α -helices connected by highly flexible loops (Fig. 2). Previous estimations of the tendency to form secondary structure, based only on sequence information, matched the proposed structure in 83%. The analysis of structural characteristics of the α -subunit model determined that 82.2% of its residues were at the nuclear region of each secondary structure of the Ramachandran plot (α -helices, folded parallel and anti-parallel β -sheet). In addition, 12.2% of the residues were in allowed zones, and 4.4% were in zones considered “generous” (Phi and Psi angle values meaning that they do not have much tension on the side chain of the residue, and neither generates the most acceptable conformations). The stereochemical quality evaluation of the model identified a single side chain at a non-allowed zone of the plot, which was Ser52 (data not shown). No tension was observed on the polypeptide backbone, or in side chain residues. Pro63 is part of a helix, although its Phi and Psi angles are not within unstable regions. Generally, α -helices are composed of charged residues, like Lys and Arg, which confer stability on α -helices. Loops in this subunit are highly flexible, as revealed in an analysis of normal vibrational modes of the molecule (data not shown). In general terms, we have an energetically favored structure ($-7,412.87 \text{ kcal mol}^{-1}$), with a preserved planarity character of its peptide bonds, and with a suitable environment for its hydrophobic or hydrophilic residues, and no inadequate atom–atom contacts. Considering these results, we can trust our model, so it can be used in studies to determine the role that certain residues play in the structure and, hence, in the enzymatic function.

β -Subunit

The template used for molecular modeling of the β -subunit structure is from a 329-amino-acid residue hydrolase, a conjugated bile salt acid hydrolase (CBAH or glycosyl hydrolase) that belongs to the C-N linear amide hydrolase family. This family is a member of the clan of the N-terminal nucleophile aminohydrolases (NTN). All the clan enzymes

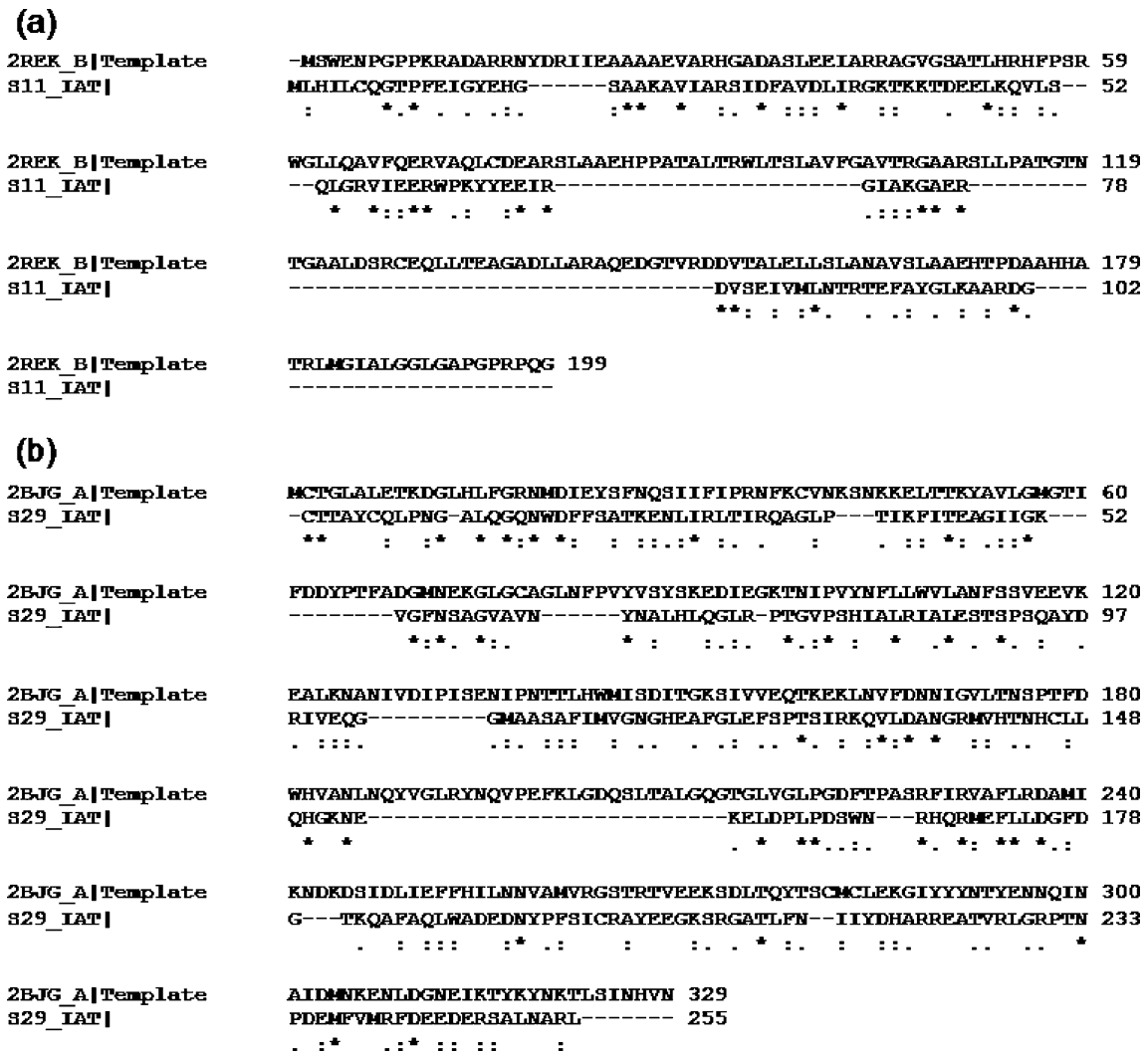


Fig. 1 Sequence alignment of the acyl-CoA:isopenicillin-N acyltransferase (IAT) subunits with other fungal sequences. **a** Transcriptional regulator sequence of the *tetR* family from *Streptomyces*

coelicolor (PDB code: 2REK, chain B), aligned against the α -subunit. **b** *Clostridium perfringens* hydrolase (PDB code: 2BJG, chain A) aligned against β -subunit

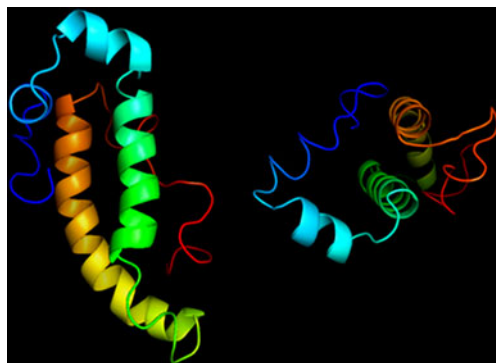


Fig. 2 Two views of the IAT α -subunit model obtained with molecular modeling by homology, after a 2 ns simulation with molecular dynamics (MD) using NAMD, with the CHARMM27 force field. *Blue* N-terminus, *red* C-terminus

show an unusual folding, where the nucleophile and the other catalytic group occupy equivalent sites, facilitating nucleophilic attack and the possibility of autocatalytic processing [69, 70]. Since several other structures were evaluated during molecular modeling by homology of the β -subunit (the PHYRE server located five other hydrolase structures: penicillin V acylase, *Bifidobacterium longum* bile salt hydrolase, glutamyl acylase, penicillin amide hydrolase, and glutaryl 7-aminocephalosporanic acid acylase—all members of the NTN superfamily), it is important to consider data regarding active sites and their catalytic residues for which there is experimental evidence to propose a potential active site in the IAT β -subunit structure.

We observed four regions in the proposed model, two α -helices and two β -strands (anti-parallel β -strands) with an α/β folding, and a multilayer architecture (4 layers, α - β - β - α) [71] consisting of a central part with nine β -strand

members, and six α -helices packed on both sides of the β -sheet (Fig. 3). In this case, previous estimations of the tendency to form secondary structure matched the model of the structure in 91% of residues. Evaluation of this conformation revealed no tensions on the polypeptide backbone, or on the side chains of the residues. Besides, we observed that the α -helices are composed of charged residues like Lys, Arg and His, which confers more stability. This model also maintains the planarity of the peptide bonds, and has an appropriate environment for hydrophobic and hydrophilic residues, without showing inadequate atom–atom contacts, which determines that the conformation is energetically favored ($-9,479.2 \text{ kcal mol}^{-1}$).

Stereochemical quality evaluation revealed that 82.4% of the residues are in zones that correspond to nuclear regions representing physically accessed conformations (α -helices, parallel and anti-parallel β -sheets), 12.7% of the residues are in allowed zones in the plot (regions adjacent to the nucleus), and 3.9% are in zones considered as generous. Thanks to this evaluation, we were able to determine that only two residues in the model (Ser195 and Glu325) are in non-allowed zones of the plot. Ser195 is located on a loop that connects α -helices 1 and 2, and it apparently forms part of a cavity, whereas Glu325 is located at the terminal carbonyl in a highly flexible loop (data not shown). It is important to mention that, as a consequence of the geometry optimization of the obtained models, the exposed hydrophobic surface was minimized, increasing the hydrophobic residue packing. The hydrophilic surface solvent and the number of hydrogen bonds were exposed, which, in general, resulted in good stereochemistry.

To evaluate the packing degree related to the number and volumes of the cavities in the protein structure, we used CASTp (Computed Atlas of Surface Topography of Proteins) (<http://cast.engr.uic.edu/cast/>) [72, 73]. We did not observe any substantial difference in the volume values

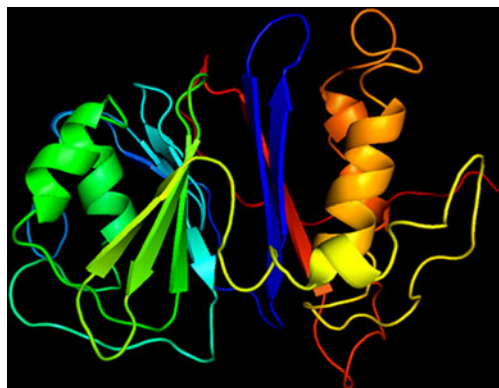


Fig. 3 View of the β -subunit model, obtained by molecular modeling by homology, after 10 ns of MD simulation using NAMD with the CHARMM27 force field. *Blue* N-terminus, *red* C-terminus

of the cavities in the β -subunit structure, indicating a level of packing of the proposed model near to the found in experimentally determined structures. The spatial location of such cavities provided a first approximation of the availability of several potential binding sites of molecules as small as ions or water, or larger molecules like substrates or inhibitors. On the other hand, analysis of the normal vibrational modes of the molecule showed that this subunit is highly flexible. We were able to identify, in a very approximate manner, movements of the protein structure. Visualizing the vibrational modes altogether provides with an approximate idea about how zones are displaced within the protein. Particularly, we observed in one of the normal vibrational modes (mode 11) that the α -helix 3 (residues Met242, Val243, His244, Thr245, Asn246 and His247) and a fragment of the β -strand 9 (Gly311, Ala312, Thr313, Leu314 and Phe315) have extended collective movements that could facilitate the access of a ligand to a binding site in these protein zones. This was demonstrated, as it is described later, using docking to identify potential binding sites in the β -subunit structure.

Molecular recognition between β -subunit and IPN

Due to its location, and determination of the potential binding site characteristic of the β -subunit structure, we were able to estimate details of the recognition process between the β -subunit and IPN describing the preliminary mechanism of the reaction during catalysis. When docking was performed on the β -subunit surface, we identified five potential binding sites. The first was located in the interface region of the β -sheet that constitutes the central part of the β -subunit (A site, 33% occupation). The second was found in a region between β -strand 1 and α -helices 4 and 5 (B site, 22% occupation), the third was found surrounded by α -helices 1 and 2 (C site, 20% occupation), the fourth between α -helices and β -strand 9 (D site, 14% occupation), and the fifth at a more superficial region between β -strands 2 and 9 (E site, 11% occupation) (Fig. 4). Composition of the binding sites is detailed in Table 1. We can observe that they are composed primarily of polar charged and non-charged residues, as well as aromatic systems from hydrophobic residues. In general, the distribution of functional groups in the binding sites confers the ability to set hydrogen bonds and hydrophobic-type interactions. Polar residues establish interactions mediated by hydrogen bonds or by van der Waals forces, while hydrophobic residues participate in π - π interactions, where the π system of aromatic chains from Phe, Tyr and Trp interacts with the π system of the ligand, as reported in previous publications [74–78]. Thus, π - π interactions are due to the aromatic moiety of the Phe residue, which interacts

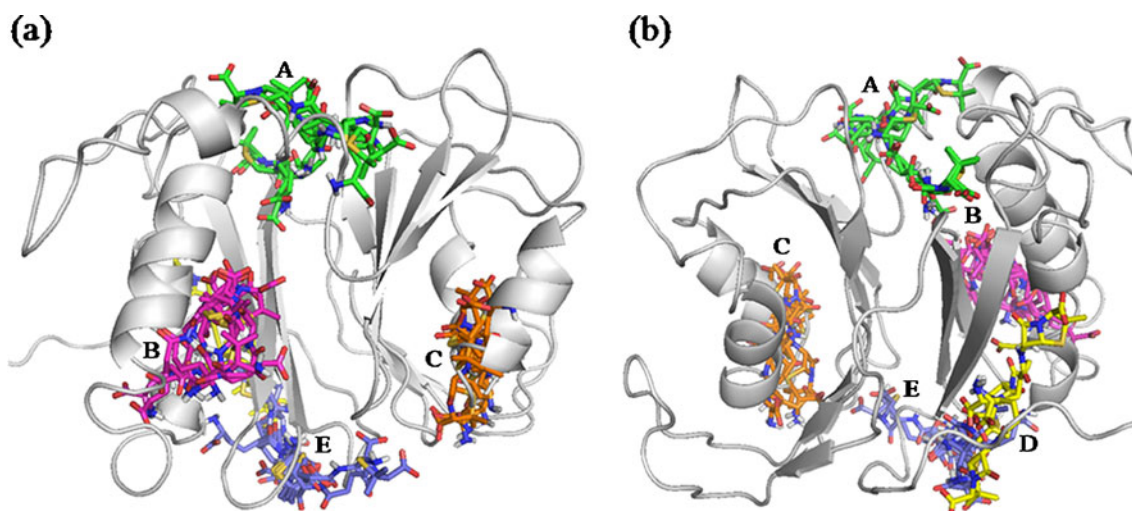


Fig. 4 Spatial location of potential binding sites to isopenicillin N (IPN) on the β -subunit surface. IPN molecules are shown in different colors to highlight the five binding sites A–E: *green A, magenta B,*

orange C, yellow D, violet E. β -subunit back (a) and front (b) views enable the visualization of the distribution of IPN molecules on the structure

with the β -lactam-thiazolidine ring of IPN. This could explain why complexes with a very negative docking energy generally present this type of interaction. It is clear that, besides the hydrophobic effect and the effect of the hydrogen and saline bonds, π -cation and π - π interactions must be considered to evaluate the force of ligands binding to a receptor.

Identification of the binding site in IPN in the β -subunit structure

Based on the results of our docking studies, the energies associated to the complexes, and the binding constants calculated, we identified the site located in the interfacial region generated by the β_1 , β_2 and β_5 strands, which form part of the central structure of the β -subunit, as the potential binding site of IPN. At time zero, the site comprises the amino acid residues Cys103, Phe122, Phe123, Leu169, His170, Gln172, Arg241, Leu262, Asp264, Arg302, Ser309, Arg310, and has a volume of 987 Å (site A). Using hydrogen bonds, the IPN binding site establishes interactions with Cys103, Leu169, Gln172,

Asp264 and Arg310. It is worth mentioning that Ser309, located in the IPN binding site, is also considered important for IAT activity (red residue, Fig. 5a) besides Cys103. Some in vitro experiments indicate that substitution of Ser309 for a Cys does not cause any observable change in its catalytic activity. However, when replacing it by an Ala, the enzyme activity is completely lost. This would indicate that a nucleophilic side chain is necessary at this position in order to keep catalytic activity. In this context, it has been suggested that this residue is involved in the substrate acylation process. Ser309 is part of an amino acid sequence that contains a G-X-S-X-G consensus motif, which is similar to the active site consensus sequences of multiple thioesterases (and of several of the templates used in the molecular modeling process for this subunit). Presumably, Ser is one of the catalytic residues, because acyl thioester activated groups are transferred to a hydroxyl group in its side chain. Regarding the reaction catalyzed by IAT, the acyl group from phenylacetyl-CoA, or other acyl thioesters, could be transferred to Ser309, and even to Cys, with a mechanism of action analogous to that of other thioesterases [20].

Table 1 Summary of the characteristics of the five potential binding sites (A–E) located on the β -subunit surface, predicted by AutoDock 3.0.5 (<http://autodock.scripps.edu>) at time zero

Site	Residues that integrate with the binding site
A	Cys103, Phe122, Phe123, Leu169, His170, Gln172, Arg241, Leu262, Asp264, Arg302, Ser309, Arg310.
B	T105, Y107, C108, Q109, L110, M272, E273, D280, Q284, Q288, D292, Y296.
C	E148, A188, L189, S191, T192, S193, S195, A196, Y197, G206, G216, N217.
D	D294, F298, Y304, A312, T313, L314, N316, I318, L330, G331, P333, N335, P336.
E	N112, G113, A114, L115, W290, I318, Y319, D320, H321, A322, R323.

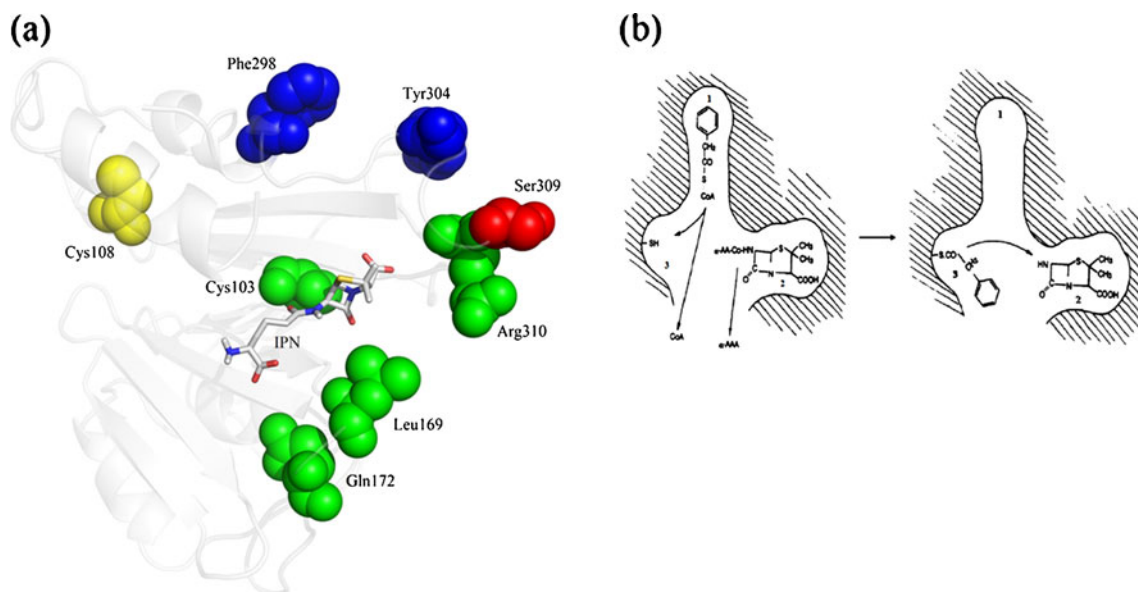


Fig. 5 **a** Spatial location of residues in the β -subunit interacting with IPN (site A, in *green*). Hydrophobic residues that could interact with phenylacetyl-CoA, the other natural IAT substrate, through π - π type contacts, are highlighted (site D, *blue*), as well as a cysteine with a sulfhydryl group that could play a role in the recognition of previously activated acyl groups (site B, *yellow*). Ser309 (*red*) is one of the residues considered extremely important in maintaining IAT enzymatic activity. **b** Proposed model for binding sites of two natural IAT

substrates: phenylacetyl-CoA and IPN. The biosynthesis of penicillin G could either proceed directly, or in a two-step process, in which case, 6-aminoadipic acid (6-APA) is formed as an intermediary. In the latter case, the first step involves the formation of an IPN-enzyme complex, followed by IPN hydrolysis with the release of α -aminoadipic acid (site 2). In the second step, transfer of an acyl moiety from phenylacetyl-CoA, to generate benzylpenicillin or penicillin G, involves the loss of CoA (modified from [5])

Spatial location of residues involved in the interaction mechanism between two natural IAT substrates: IPN and phenylacetyl-CoA

Site A is adjacent to sites B and D. Site B contains one of the four conserved Cys of the IAT sequence, Cys108. As mentioned above, this residue could provide a sulfhydryl group that, after hydrolysis of the thioester bond (yellow residue, Fig. 5a), would react with acyl groups of activated precursors. On the other hand, site D has the aromatic residues Phe298 and Tyr304. Previous results from our group indicate that there is a high possibility that phenylacetyl-CoA binds to this site, interacting with its hydrophobic part with the aromatic rings in Phe298 and Tyr304 through π - π type contacts (blue residues, Fig. 5a). These three sites could be involved in IAT substrate fixation for further catalytic processing. In this context, the research group of J.F. Martín proposed a mechanism of action for IAT, IPN, and phenylacetyl-CoA that states the existence of three binding sites [5]. A first site would receive acyl groups and some activated forms like phenylacetyl, phenoxyacetyl, and other acyl forms that would interact in it, which could be used to form penicillin (site 1, Fig. 5b). A second site would bind an IPN, and probably it would be the same site that binds the intermediary 6-APA (site 2, Fig. 5b). Finally, a third site with a sulfhydryl group would receive acyl groups from their activated precursors after

thioester bond hydrolysis (site 3, Fig. 5b). This model proposes the formation of an IPN-enzyme complex at site 2, followed by IPN hydrolysis with the release of α -L-aminoadipic acid, as the first step in the conversion of IPN to benzylpenicillin.

The next step in the IAT-catalyzed reaction is less known, although it has been postulated that an IAT activity catalyzes acyl group transfer from phenylacetyl-CoA to produce benzylpenicillin, with the release of CoA. In 1982, Queener and Neuss [79] suggested the existence of a binding site for acyl-CoA derivatives, which would necessarily implicate the formation of a second complex, phenylacetyl-CoA-enzyme (a hypothesis supported by experiments performed by the group of Martín in 1993) [5]. It is important to consider that, in the case of the enzymes used as templates for the β -subunit modeling process, a Cys, Ser, or Thr residue of the side chain is used, often as part of a β -sheet (located towards the amino terminal), as nucleophile in the catalytic attack of a carbonyl carbon [69, 70]. This information, together with the already existing IAT information, enables us to propose a potential active site for the β -subunit structure of IAT, probably delimited by sites A, B and D, identified through docking. Considering experimental data from point mutations on the IAT structure, and the homology between the enzyme and other hydrolases (enzymes with a folding that facilitates a nucleophilic attack), mutation of residues not

previously considered to form part of the active site could provide information regarding its location and mechanism of action. As mentioned before, IAT, IPN and phenylacetyl-CoA molecular recognition simulation studies allowed us to infer additional details regarding the nature of the binding sites, with the ability to establish interactions with IAT natural substrates. In the case of IAT/IPN, the established interactions were mediated predominantly by hydrogen bonds, whereas in the case of IAT/phenylacetyl-CoA, interactions were established through π - π type contacts (data not shown). Besides those previously determined *in vitro* by several research groups, seven residues were identified *in silico* as relevant to the protein–ligand interaction, and therefore for the enzymatic activity: Leu169, Gln172, Asp264 and Arg310 (IAT/IPN), Phe298 and Tyr304 (IAT/phenylacetyl-CoA), and also Cys108. *In vitro* studies of mutants of some of these residues could provide information regarding the role they play in the location of the active site on one hand, and in the molecular recognition mechanism of small ligands on the other hand, as well as in determining the β -subunit stability.

MD simulations identify conformational changes in the β -subunit

We were able to identify and study many conformational changes using data from MD simulations, as well as to evaluate the time evolution of the system's behavior; specifically, movements and structural displacements. In order to explore a significant number of low-energy structures and to extract information about the zones with larger fluctuations in the β -subunit, we performed a MD simulation along 74 ns. From this trajectory, a group of representative conformations were recovered every 0.5 ns- and analyzed, observing a fundamental difference among structures at the level of compaction and resolution of some secondary structures, basically α -

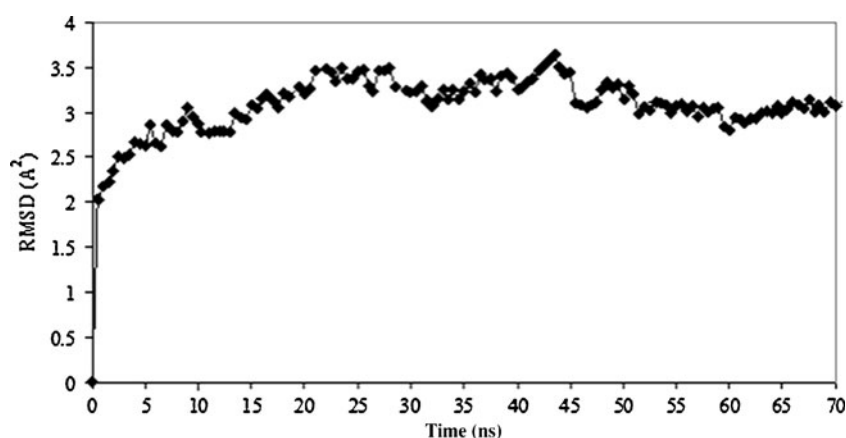
helices. All structures were stereochemically validated, finding a high percentage of interatomic distances, binding angles, and acceptable torsion angles. Likewise, using a structural alignment, we established equivalencies between recovered β -subunit conformations from the trajectory of the MD, and a structure from molecular modeling by homology (0 ns). The root mean square deviation (RMSD) shows the degree of divergence among them, revealing a significant rearrangement in the backbone of the protein and side chains during the first 15 ns. Since RMSD values reflect not only protein conformation but also the state of the angular rotations in side chains, we can assume that the most significant structural rearrangement of collective movements happens during the first few nanoseconds of MD simulations. After 15 ns, β -subunit conformations do not differ significantly among the observed structures, maintaining an acceptable compaction level (Fig. 6).

Docking studies employing snapshots from MD

The reason for recovering β -subunit structures every 0.5 ns was to use them for docking, to provide protein flexibility, and to identify how the binding site conformation influences the recognition process. With this approach, we were able to sample the complex protein–ligand conformation space, reaching a flexible receptor–ligand system. When docking was performed in each one of the β -subunit structures from the MD trajectory, it showed that the ligand reached site A in most cases (88.5%) in the initial structure (0 ns). The protein–ligand complex at 10 ns had particularly high binding energetics ($-14.49 \text{ kcal mol}^{-1}$) (Fig. 7), indicating that when the structure of the receptor acquires this conformation, the ligand can participate in more appropriate chemical interactions.

We analyzed the binding sites of complexes with very negative docking energies, and we were able to determine the type of residues that constitute that zone, as well as the

Fig. 6 Root mean square deviation (RMSD) calculation of the β -subunit conformations, recovering information every 0.5 ns from MD performed for 74 ns. For clarity, only the α -carbon positions are considered as the peptide bond has a minimally varying planar conformation



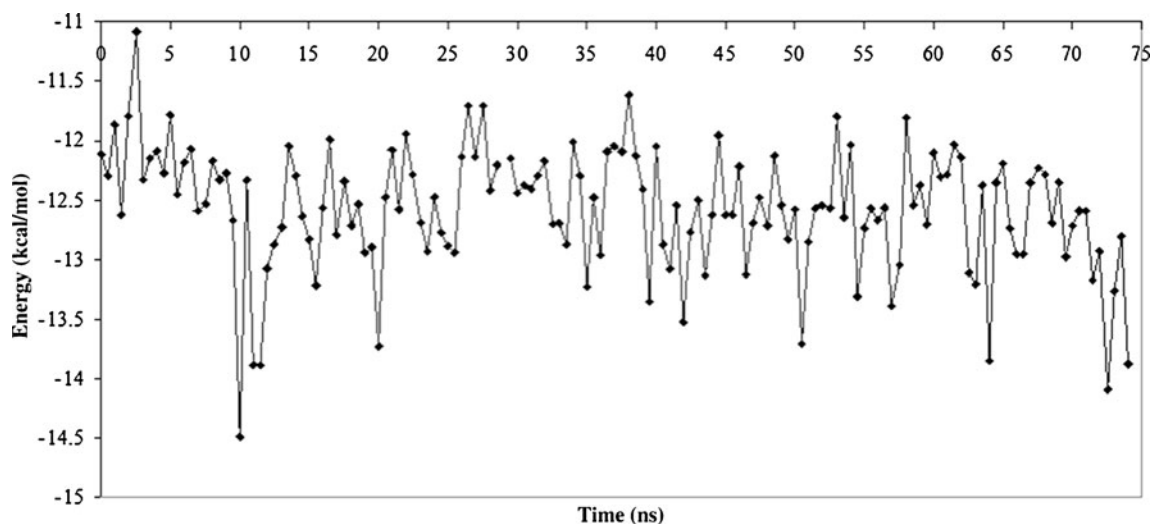


Fig. 7 Docking energy of complexes formed between β -subunit and IPN. The β -subunit three-dimensional structures recovered every 0.5 ns from MD simulations with a duration of 74 ns. Docking was made using AutoDock 3.0.5 (flexible ligand) (<http://autodock.scripps.edu>)

chemical interactions that are established within it. We observed that the binding site residue composition alters when the protein undergoes different conformational changes, but in all cases the ligand reaches site A (0 ns, Fig. 4a, Table 1). Considering that the amount and type of atomic interactions depend on the characteristics of the residues that form the site, we can say that, in our case, the binding force of the ligand was higher in those with more residues capable of forming hydrogen bonds (Table 2).

Comparison of the physical and chemical characteristics of the binding site of the structures recovered every 0.5 ns provided information regarding to the possible conformational changes associated with intrinsic movements of the β -subunit, and their influence on the molecular recognition between the β -subunit and IPN. We observed a rearrangement of the residues involved in binding. For example, in the β -subunit structure after 10 ns, we observed a spatial rearrangement, with only residues Cys103, Phe122, Phe123, Leu169, His170, Gln172, Arg241, Leu262, Asp264, Arg302, Ser309, Arg310 maintaining the integration of Asp121, Ala168 and Phe212. We determined that IPN established contact by a hydrogen bond with Cys103, Leu169, Gln172, Asp264 and Arg310, with a binding energy of $-14.49 \text{ kcal mol}^{-1}$. According to the docking energy computed, these changes optimize the interaction of

the ligand with the β -subunit, since the number of established contacts increases. In the case of the other three complexes with higher energy (and in most of our results), we observed interactions mediated predominantly by Arg residues (Arg232, Arg268, Arg302, Arg308 and Arg310), which carry a guanidine group with a very strong basic character. It is worth mentioning that Arg310 was always found with IPN binding residues. In vitro studies analyzing the effect on the enzyme activity with point mutations of this residue will provide information regarding the location and conformation of the active site in IAT.

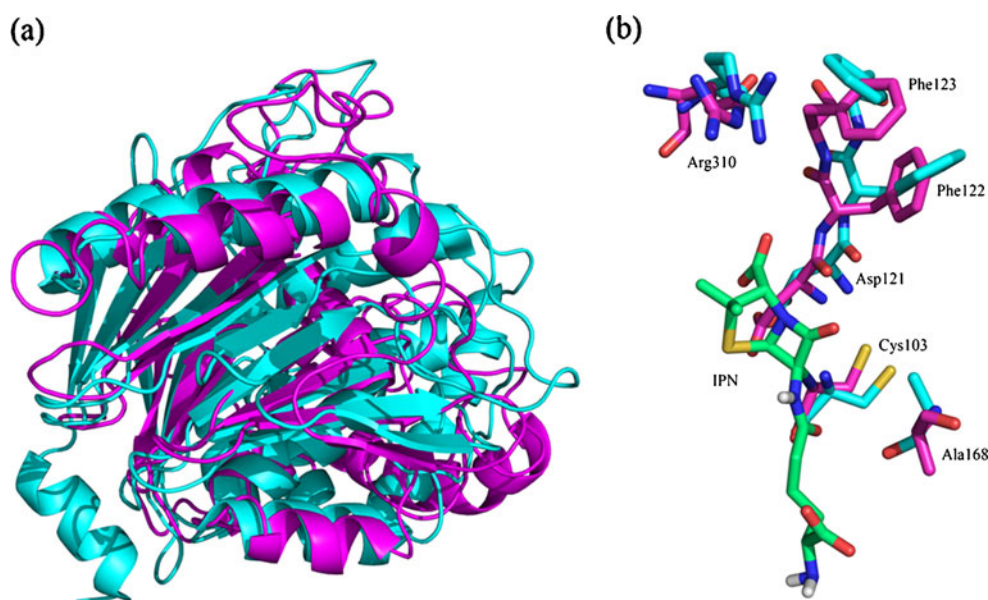
While our manuscript was in preparation, three crystal structures (PDB entry 2X1C, resolution=1.85 Å; 2X1D, resolution=1.64 Å; 2X1E, resolution=2.00 Å) of the mature enzyme were published by Bokhove et al. [24]. Our model can thus be validated by comparison with these structures. As can be seen in Fig. 8a, the overall tertiary structure of the IAT in our model agrees well with Bokhove's crystal structure. The only major difference is observed in the loop region. Because our model was refined by MD simulation we assume that the origin of this discrepancy is that the flexibility of the macromolecule was not completely reproduced.

On the other hand, besides the correct tertiary structure, it is encouraging to observe that our model also provides

Table 2 Identification of complexes with higher docking energy in site A, and residues that interact with isopenicillin N (IPN) at different times of molecular simulation (ns)

Time (ns)	Docking energy (kcal/mol)	Residues interacting with IPN
0	-12.11	Leu169, His170
10.0	-14.49	Cys103, Leu169, Gln172, Asp264, Arg310
11.0	-13.89	Cys103, Asp264, Arg302, Arg310
20.0	-13.73	Arg302, Lys308, Arg310
50.5	-13.71	Arg232, Gly240

Fig. 8 **a** Structure alignment of IAT. The backbones of Bokhove's crystal structure [24] and our model are rendered as cyan and magenta ribbons, respectively. **b** Substrate binding pocket of IAT and illustration of the key residues involved in the binding of IPN



sufficient detail of the substrate binding pocket (Fig 8b). For example, after 10 ns of MD simulations, a spatial rearrangement was observed and residues Cys103, Asp121, Phe122, Phe123, Ala168, Leu169, His170, Gln172, Phe212, Arg241, Leu262, Asp264, Arg302, Ser309 and Arg310 were conserved. Co-crystallization of mature IAT with the β -lactam core 6-APA revealed that 6-APA binds near Cys103, that its thiazole ring makes Van der Waals interactions with Phe122 and Phe123, and its carboxylate group makes a salt bridge with Arg310 (see Fig. 4A in [24]) allowing allocation of the substrate binding site. In our case, the natural substrate IPN was found in the same place but establishing interactions with C103, Asp264, and R310.

Effect on β -subunit structure of mutations at Gly150Val and Glu258Lys residues

In 1994, Fernández-Perrino et al. [18] reported that introducing an apolar group like Val at position Gly150 can abolish enzymatic activity. The same was observed when placing a basic polar residue (Lys258), whose acid characteristics would stabilize the structure at a position originally occupied by Glu. These results have two possible explanations: (1) both residues are part of the IAT active site, or (2) both may participate in conformational stabilization of the enzyme. Upon obtaining the β -subunit structure (wild type enzyme), we evaluated the effect of point mutations (Gly150Val and Glu258Lys) to determine the role of these residues in IAT function and structure. As with the wild type enzyme, we performed a series of local minimization experiments on the replaced side chain. In the case of the Gly150Val mutant, the presence of a branched amino acid such as Val, whose isopropyl side chain not only has a distinct reactivity to the hydrogen atom of Gly, but also flexibility and an adequate

molecular size, caused a reduction in the spatial conformation, with concomitant spatial restriction. This space reduction caused steric impediments and conformational tensions that destabilized the β -subunit structure locally. Mutated residues did not modify the conformation of the neighbors significantly. However, structural neighbors with more reactive side chains, such as Glu190 and Arg132, suffer considerable changes. Of these, Glu132 had the most significant effect. Rearrangement in this side chain with negative charge altered the chemical environment, influencing the conformation of many of its neighbors, such as Ser191 and Thr192 (Fig. 9). Adding up the effects of this derived fluctuation in local conformation led to a loss of structured packing. The effect of the mutation propagated not only to neighboring residues—sequence and structure—but also to other zones far from the structure. We were interested in determining the conformational changes in residues of site A (Fig. 10a). The thiol group in Cys103 was displaced a few degrees, trying to avoid overlapping of its electronic cloud with adjacent side chains. In this way, its neighbor Asp121 reoriented the carboxyl group in response to the rearrangement of local chemical conditions. Ala168, with a methyl group in its side chain, did not suffer any significant change, probably due to the size of the substituted side chain. However, the next residue, Leu169, suffered a displacement in space, trying to compensate for the effect of the change in its neighbors, buffering in this way any possible change in its conformation. On the contrary, the imidazole group in His170 changed the orientation plane of its ring, causing a rearrangement in the Gln172 side chain, which had a very strong effect on its next neighbor, Phe212, with a phenyl group that suffered a modification in its plane. In the case of Arg241, the guanidine group rearranged slightly, probably due to its surrounding space, which did not perceive

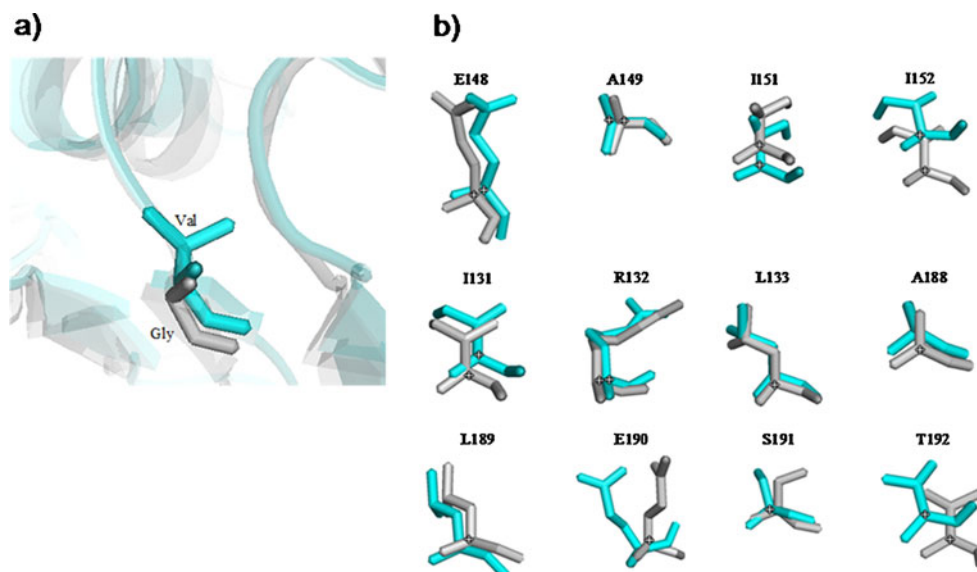


Fig. 9 **a** Spatial location of Gly150 and Val150 residues on the β -subunit surface. The β -subunit structures are visualized simultaneously, which allows to observe the effect of the point mutation. *Gray* Wild

β -subunit structure, *cyan* Gly150Val mutant. **b** Effects on side chain residues close to position 150 as structural neighbors. A significant effect was the 3.3 Å displacement suffered by the Glu190 side chain

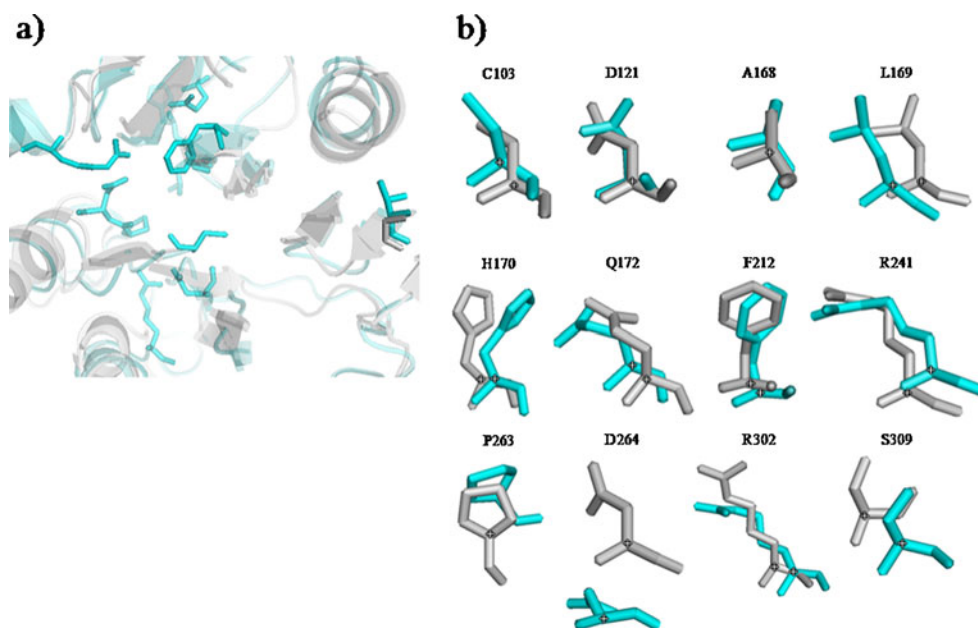
significant changes. The sum of the conformational rearrangements and the chemical conditions close to Pro263 resulted in an observable change in its five-member ring. This change directly affects Asp264, which must not only rearrange its conformation, but also buffer spatial tension, being displaced by a few Ångstroms (with respect to the wild subunit), which apparently will absorb the conformational cost. Finally, the Arg302 and Ser309 side chains, at the most external part of site A, suffer slight modifications to their conformations (Fig. 10b).

From the *in silico* side to this study, another four residues considered as important for IAT structure and function also

exhibited important conformational modifications. Arg310, a residue that played an important role in the IAT interaction with IPN in all the cases analyzed, suffered considerable modification in its guanidine group. Aromatic residues located on site D, Phe298 and Tyr304, had slight modifications in the plane of their aromatic rings, although the most important change was the displacement of residues. Such displacements could affect the establishment of π - π type interactions with the natural substrate of IAT. Finally, Cys108 in site B suffered a displacement of 3.8 Å regarding the wild protein (Fig. 11).

The effect of the mutation at position 150 resulted in changes at different points in the β -subunit structure. The

Fig. 10 **a** Spatial location of Gly150, Val150 and residues of binding site A on the β -subunit surface. Structures are visualized simultaneously; *gray* wild β -subunit, *cyan* Gly150Val mutant. **b** Effects suffered by site A side chains. One of the most significant effects was a displacement and conformational change suffered by the side chain of Asp264, which is also one of the residues that interacts with IPN, through a hydrogen bond



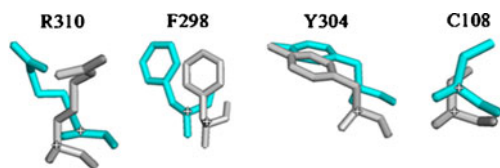


Fig. 11 Spatial location of four residues considered important in establishing interactions between IPN substrates and phenylacetyl CoA, and IAT, Arg310 (site A), Phe298, Tyr304 (site D), as well as Cys108 (site B). Conformations are visualized simultaneously, allowing the effect of the Gly for Val change at position 150 on side chains of these residues to be observed. *Gray* Wild β -subunit structure, *cyan* Gly150Val mutant

sum of the conformational rearrangements and the chemical condition variations had a considerable impact on docking studies, because the binding energy of the complexes became less negative (data not shown), indicating reduced ligand affinity towards the receptor. Evidently, side chain conformation, and therefore binding site surface changes, influence the recognition mechanism acting on both molecules. Regarding the Glu258Lys mutant effect, this mutation was much more severe from the structural point of view. There was a charged-mediated destabilization due to the substitution of a negatively charged residue (Glu) for a positively charged residue (Lys). This destabilization was so

strong that it generated local conformational fluctuations that propagated globally, thus destabilizing the functional conformation of the β -subunit. We were able to observe a series of modifications in the structure, which optimized necessary spaces to avoid overlapping of atomic electronic clouds that constitute adjacent residues of the mutated residue (Fig. 12a, b). Due to the substitution of Glu for Lys, the generated environment caused not only a redistribution of conformations in neighboring sequence residues, but also in structural neighbors, and even on residues far from the active site (Fig. 12c). The most significant conformational changes were at residues considered as first neighbors. In the wild structure, this region is stabilized by a set of positive and negative charges, and free electrons (Lys254⁽⁺⁾, Asn255⁽⁻⁾, Glu256⁽⁻⁾, Lys257⁽⁺⁾, Glu258⁽⁻⁾, Leu259^(non-polar residue) and Asp260⁽⁻⁾), which interact by equilibrating the conformational structure of the site. The effect of a new positive charge from a newly mutated Lys significantly alters the equilibrium, triggering events that initiate a repulsion of similar charges. Lys257 has a positively charged chain and literally avoids contact with the newly substituted Lys side chain, directly affecting its neighbor Glu256 by changing the carboxyl orientation of its side chain. Leu259, Asp260 and Pro262 also experience conformational alterations. This

Fig. 12 a Spatial location of Glu258 and Lys258 on the β -subunit surface. Structures from this subunit are simultaneously visualized to observe the effect of the punctual mutation on the configuration. *Gray* Wild β -subunit structure, *green* Glu258Lys mutant. **b** Structural neighbors at position 258. **c** Effect on side chains of each of the residues shown in **b**. One significant event is the displacement and reconfiguration of the Asp264 side chain—a residue establishing a hydrogen bond

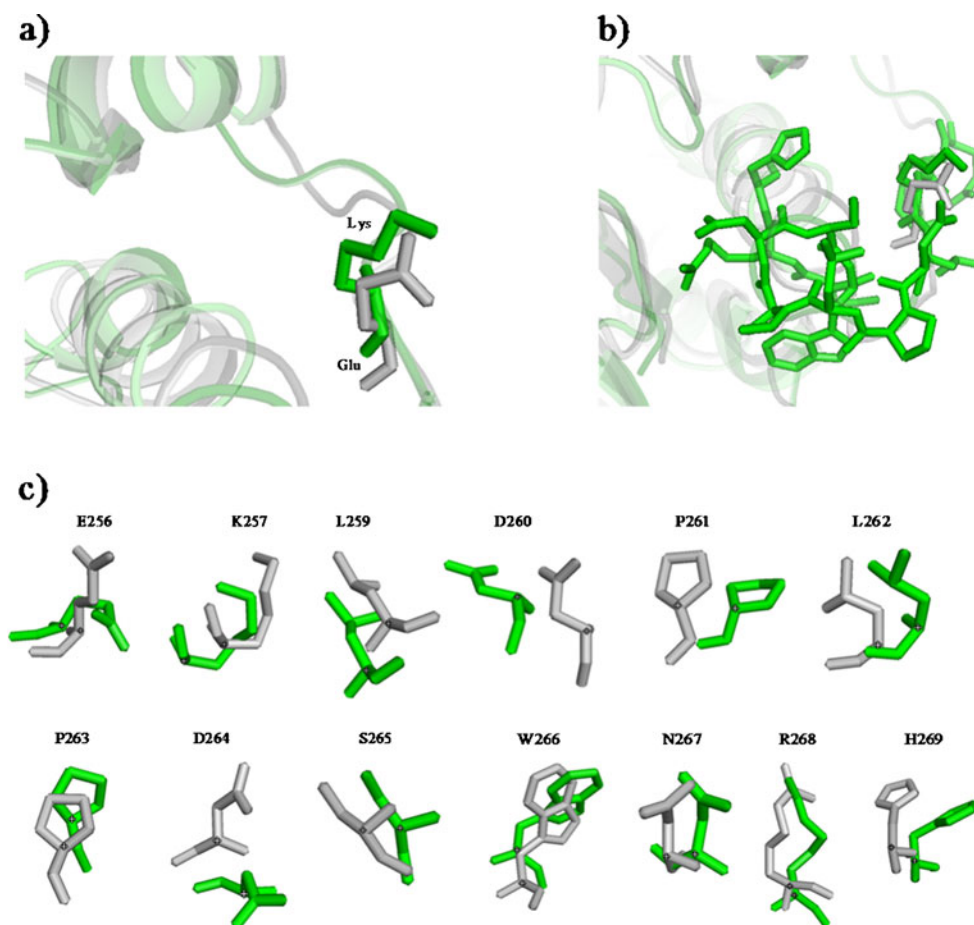
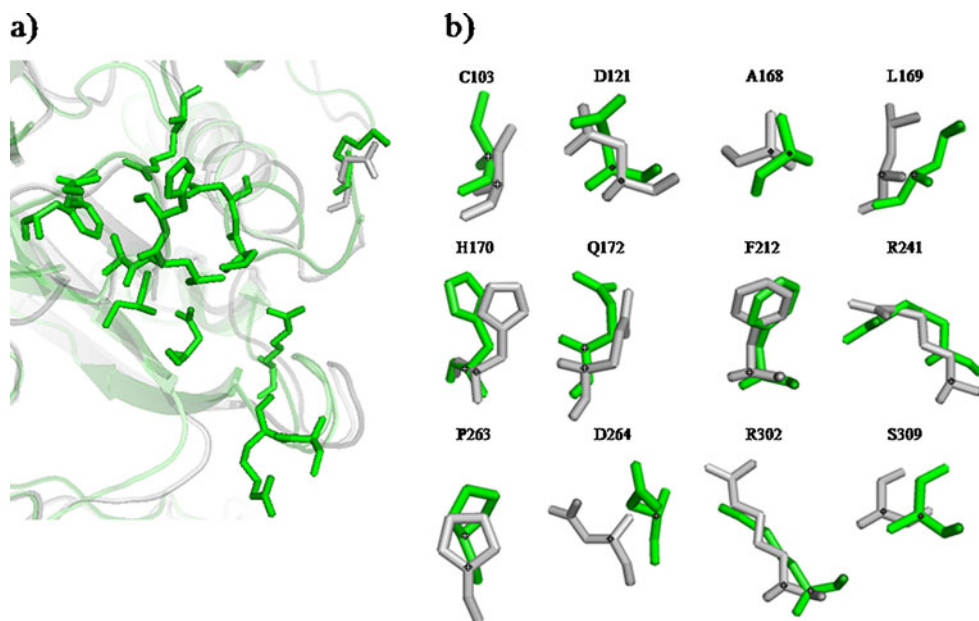


Fig. 13 **a** Spatial location of Glu258 and Lys258, and residues that are part of site A on the β -subunit surface. **b** Effect on site A side chain residues of the Glu for Lys mutation. *Gray* Wild β -subunit structure, *green* Glu258Lys mutant



latter residue changes the plane of the five-membered ring drastically, and the first two are displaced with slight modifications of their side chain conformation. This affects nearby residues, propagating the effect of the mutation. Thus, Asp264 displaces its carboxyl group to avoid an electronic density overlap of oxygen with the Ser265 hydroxyl group. As a consequence, the Trp266 indole group changes its plane and the neutral amide group in Asn267 reorients by a few degrees, which, together with the reigning chemical environment, is enough to cause a conformational change in Arg268, causing further changes in the orientation of the plane of His269, a weakly basic imidazole (Fig. 12c).

Possibly, some major role of β -subunit binding site residues, and even phenylacetyl-CoA, are involved in this breakdown pathway. In this study, we were able to establish a direct impact on the affinity of IPN towards the receptor due to the effect of modification on its spatial conformation. It is important to mention that the effects of the Glu for Lys mutation were observed not only in sequence neighbors, but also in residues of the side chains of site A (Fig. 13a). Arg310, Phe298, Tyr304 and Cys108 are residues marked as important in the molecular recognition of both natural substrates of IAT, which suffer modifications in their side chains. Besides the displacement of the complete residue with respect to the wild enzyme, Arg310 suffered a significant modification in its guanidine conformation. Phe298 and Tyr304, like Gly150Val, had slight modifications in their aromatic ring planarity, with residue displacement as the most important change. Thus, it is again possible to think that these displacements could affect the establishment of π - π type interactions with phenylacetyl-CoA, the other IAT natural substrate. On the other hand, the thiol group in Cys108 did not experience substantial

modifications. It seems as if Arg310 should be more susceptible to modification than other residues (Fig. 14), since recent reports have revealed that the native state is a conformational ensemble defined by multiple partially unfolded forms [80]. The present work has shown that the Gly for Val, or the Glu for Lys mutations could destabilize the structure, causing the functional collapse of the enzyme if these positions were very closer to zones where the protein suffers folding/unfolding reactions. This would explain why its mutation significantly affects IAT enzymatic activity, although it is not part of the active site of both natural substrates, which is located on the surface of the β -subunit.

Conclusions

In conclusion, we obtained some reasonable structural models of the 3D structure of IAT α and β subunits, as well as the manner at which IPN binds to the IAT- β subunit. It is encouraging to find that our models agree well with a

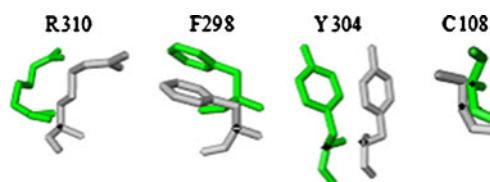


Fig. 14 Spatial location of four residues important in the establishment of interactions between IPN substrates and phenylacetyl-CoA and IAT, Arg310 (site A), Phe298 and Tyr304 (site D), and Cys108 (site B). Conformations are simultaneously visualized to observe the effect on the side chains of the Glu for Lys change at position 258. *Gray* Wild β -subunit structure, *green* Glu258Lys mutant

recently released crystal structure of IAT [24]. In addition, our model can reasonably interpret the results of a number of experiments where mutations were carried out on key residues for IAT processing, as well as in strictly conserved residues, most probably involved with the enzymatic activity. Based on the results of docking studies, associated energies to the complexes, and calculated binding constants, we have identified the site located in the interfacial region generated by the β_1 , β_2 and β_5 strands, which are part of the central structure of the β -subunit, as the potential binding site in IPN. The site comprises the amino acid residues Cys103, Asp121, Phe122, Phe123, Ala168, Leu169, His170, Gln172, Phe212, Arg241, Leu262, Asp264, Arg302, Ser309, and Arg310. Using hydrogen bonds, the IPN binding site establishes interactions with Cys103, Leu169, Gln172, Asp264 and Arg310. Finally, according to the results of molecular recognition by docking studies, we believe that the Gly for Val mutation (position 150), or Glu for Lys (position 258) causes destabilization of the structure and a functional collapse of the β -subunit because these positions are adjacent, and even form part of the zone that suffers folding/unfolding reactions. This would explain why this mutation significantly affects IAT enzymatic activity, although it is not part of the binding site of either of the identified natural substrates on the β -subunit surface.

Acknowledgments The investigation was supported in part by grants from the Consejo Nacional de Ciencia y Tecnología (CONACyT), Mexico (Grant No. 132353), and Instituto Politécnico Nacional (Secretaría de Investigación y Posgrado and Comisión de Operación y Fomento de Actividades Académicas) and Instituto de Ciencia y Tecnología del Distrito Federal.

References

1. Fawcett PA, Usher JJ, Abraham EP (1975) Behavior of tritium labeled isopenicillin N and 6-aminopenicillanic acid acyltransferase from *Penicillium chrysogenum*. *Biochem J* 151:741–746
2. Álvarez E, Cantoral JM, Barredo JL, Díez B, Martín JF (1987) Purification to homogeneity and characterization of acyl-coenzyme A:6-aminopenicillanic acid acyltransferase of *Penicillium chrysogenum*. *Antimicrob Agents Chemother* 31:1675–1682
3. Alonso MJ, Bermejo F, Reglero A, Fernández-Cañón JM, González de Buitrago G, Luengo JM (1988) Enzymatic síntesis of penicillins. *J Antibiot* 41:1074–1084
4. Whiteman PA, Abraham EP, Baldwin JE, Fleming MD, Schofield CJ, Sutherland JD, Willis AC (1990) Acyl-coenzyme A:6-aminopenicillanic acid acyltransferase from *Penicillium chrysogenum* and *Aspergillus nidulans*. *FEBS Lett* 262:342–344
5. Álvarez E, Meesschaert B, Montenegro E, Gutiérrez S, Díez B, Barredo JL, Martín JF (1993) The isopenicillin N acyltransferase of *Penicillium chrysogenum* has isopenicillin N amidohydrolase, 6-aminopenicillanic acid acyltransferase and penicillin amidase activities, all of which are encoded by the single *penDE* gene. *Eur J Biochem* 215:323–332
6. Martín-Villacorta J, Reglero A, Luengo JM (1990) Acyl CoA: 6-APA acyltransferase from *Penicillium chrysogenum*. Studies on its hydrolytic activity. *J Antibiot* 44:108–110
7. Martín-Villacorta J, Reglero A, Luengo JM (1990) Biosynthesis of methoxy benzylpenicillins. *Biotechnol Forum Europe* 8:60–62
8. Ferrero MA, Reglero A, Martínez-Blanco H, Fernández-Valverde M, Luengo JM (1990) In vitro synthesis of new penicillins containing keto acids as side chains. *Antimicrob Agents Chemother* 35:1931–1932
9. Martínez-Blanco H, Reglero A, Luengo JM (1991) In vitro synthesis of different naturally-occurring, semisynthetic and synthetic penicillins using a new and effective enzymatic coupled system. *J Antibiot* 44:1252–1258
10. Barredo JL, Van Solingen P, Díez B, Álvarez E, Cantoral JM, Kattavilder A, Smaal EB, Groenen MAM, Veenstra AE, Martín JF (1989) Cloning and characterization of the acyl-coenzyme A:6-aminopenicillanic acid acyltransferase gene of *Penicillium chrysogenum*. *Gene* 83:291–300
11. Montenegro E, Barredo JL, Gutiérrez S, Díez B, Álvarez E, Martín JF (1990) Cloning, characterization of the acyl-CoA:6-aminopenicillanic acid acyltransferase gene of *Aspergillus nidulans* and linkage to the isopenicillin N synthase gene. *Mol Gen Genet* 221:322–330
12. Muller WH, Bovenberg RA, Groothuis MH, Kattavilder F, Smaal EB, Van der Voort LH, Verkleij AJ (1992) Involvement of microbodies in penicillin biosynthesis. *Biochim Biophys Acta* 1116:210–213
13. Muller WH, Essers J, Humbel BM, Verkleij AJ (1995) Enrichment of *Penicillium chrysogenum* microbodies by isopycnic centrifugation in nycodenz as visualized with immuno-electron microscopy. *Biochim Biophys Acta* 1245:215–220
14. Theilgaard HB, Kristiansen KN, Henriksen CM, Nielsen J (1997) Purification and characterization of delta-(L-alpha-amino adipyl)-L-cysteiny-D-valine synthetase from *Penicillium chrysogenum*. *Biochem J* 327:185–191
15. Aplin RT, Baldwin JE, Cole SCJ, Sutherland JD, Tobin MB (1993) On the production of α , β -heterodimeric acyl-coenzyme A: isopenicillin N acyltransferase of *Penicillium chrysogenum*: studies using a recombinant source. *FEBS Lett* 319:166–170
16. Aplin RT, Baldwin JE, Roach PL, Robinson CV, Schofield CJ (1993) Investigations into the posttranslational modification and mechanism of isopenicillin N:acyl-CoA acyltransferase using electrospray mass spectrometry. *Biochem J* 294:357–363
17. Tobin MB, Baldwin JE, Cole SCJ, Miller JR, Skatrud PL, Sutherland JD (1993) The requirement for subunit interaction in the production of *Penicillium chrysogenum* acyl-coenzyme A: isopenicillin N acyltransferase in *Escherichia coli*. *Gene* 132:199–206
18. Fernández FJ, Gutiérrez S, Velasco J, Montenegro E, Marcos AT, Martín JF (1994) Molecular characterization of three loss-of-function mutations in the isopenicillin N-acyltransferase gene (*penDE*) of *Penicillium chrysogenum*. *J Bacteriol* 176:4941–4948
19. Fernández FJ, Cardoza RE, Montenegro E, Velasco J, Gutiérrez S, Martín JF (2003) The isopenicillin N acyltransferases of *Aspergillus nidulans* and *Penicillium chrysogenum* differ in their ability to maintain the 40-kDa alpha beta heterodimer in an undissociated form. *Eur J Biochem* 270:1958–1968
20. Tobin MB, Cole SCJ, Kovacevic S, Miller JR, Baldwin JE, Sutherland JD (1994) Acyl-coenzyme A:isopenicillin N acyltransferase from *Penicillium chrysogenum*: effect of amino acid substitutions at Ser²²⁷, Ser²³⁰ and Ser³⁰⁹ on proenzyme cleavage and activity. *FEMS Microbiol Lett* 121:39–46
21. Tobin MB, Cole SCJ, Miller JR, Baldwin JE, Sutherland JD (1995) Amino-acid substitutions in the cleavage site of acyl-coenzyme A:isopenicillin N acyltransferase from *Penicillium chrysogenum*: effect on proenzyme cleavage and activity. *Gene* 162:29–35

22. Hensgens CM, Kroezinga EA, van Montfort BA, van der Laan JM, Sutherland JD, Dijkstra BW (2002) Purification, crystallization and preliminary X-ray diffraction of Cys103Ala acyl coenzyme A: isopenicillin N acyltransferase from *Penicillium chrysogenum*. *Acta Crystallogr D* 58:716–718
23. Yoshida H, Hensgens CMH, van der Laan JM, Sutherland JD, Hart DJ, Dijkstra BW (2005) An approach to prevent aggregation during the purification and crystallization of wild type acyl coenzyme A:isopenicillin N acyltransferase from *Penicillium chrysogenum*. *Protein Expr Purif* 41:61–67
24. Bokhove M, Yoshida H, Hensgens CM, van der Laan JM, Sutherland JD, Dijkstra BW (2010) Structures of an isopenicillin N converting Ntn-hydrolase reveal different catalytic roles for the active site residues of precursor and mature enzyme. *Structure* 18:301–308
25. Mirky AE, Pauling L (1936) On the structure native, denatured and coagulated proteins. *Proc Natl Acad Sci USA* 22:439–447
26. Anfinsen CB, Haber E, Sela M, White FH (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci USA* 47:1309–1314
27. Havel TF, Snow ME (1991) A new method for building protein conformations from sequence alignments with homologues of known structure. *J Mol Biol* 217:1–7
28. Hilbert M, Bohm G, Jaenicke R (1993) Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins* 17:138–151
29. Sali A, Blundell TL (1993) Comparative protein modeling by satisfaction of spatial restraints. *J Mol Biol* 234:779–815
30. Srinivasan S, March CJ, Sudarsanam S (1993) An automated method for modeling proteins on known templates using distance geometry. *Protein Sci* 2:277–289
31. Bairoch A, Apweiler R (1997) The Swiss-Prot protein sequence database: its relevance to human molecular medical research. *J Mol Med* 75:312–316
32. Bennett-Lovsey RM, Herbert AD, Sternberg JE, Kelley LA (2008) Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. *Proteins* 70:611–625
33. Labarga A, Valentin F, Anderson M, Lopez R (2007) Web services at the European bioinformatics institute. *Nucleic Acids Res* 35:W6–W11
34. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
35. Bowie JU, Lüthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253:164–170
36. Flores TP, Orengo CA, Moss DS, Thornton JM (1993) Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci* 2:1811–1826
37. Raghava GPS (2002) APSSP2: A combination method for protein secondary structure prediction based on neural network and example based learning. *CASP5 A*–132
38. Rost B, Yachdav G, Liu J (2004) The PredictProtein Server. *Nucleic Acids Res* 32:W321–W326
39. Cole C, Barber JD, Barton GJ (2008) The Jpred 3 secondary structure prediction server. *Nucleic Acids Res* 36:W197–W201. doi:10.1093/nar/gkn238
40. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) Jpred: a consensus secondary structure prediction server. *Bioinformatics* 14:892–893
41. Cuff JA, Barton GJ (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Protein Struct Funct Genet* 34:508–519
42. Dong A, Xu X, Gu J, Edwards AM, Joachimiak A, Savchenko A (2007) Crystal structure of *tetR*-family transcriptional regulator. doi:10.2210/pdb2rek/pdb
43. Rossocha M, Schultz-Heienbrock R, von Moeller H, Coleman JP, Saenger W (2005) Conjugated bile acid hydrolase is a tetrameric N-terminal thiol hydrolase with specific recognition of its cholyl but not of its tauryl product. *Biochemistry* 44:5739–5748
44. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res* 31:3381–3385
45. Lambert C, Leonard N, De Bolle X, Depiereux E (2002) ESyPred3D: prediction of proteins 3D structures. *Bioinformatics* 18:1250–1256
46. Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. *Protein Sci* 9:1753–1773
47. MacKerell AD, Bashford M, Bellot M, Dunbrack RL, Evanseck JD, Field MJ, Fisher 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, Wiorcikiewicz-Kuczera J, Yin D, Karplus MJ (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102:3586–3616
48. Kale L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, Schulten K (1999) NAMD2: greater scalability for parallel molecular dynamics. *J Comput Phys* 151:283–312
49. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten KL (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* 26:1781–1802
50. Humphrey W, Dalke A, Schulten K (1996) VMD-Visual Molecular Dynamics. *J Mol Graph* 14:33–38
51. Diemand AV, Scheib H (2004) MolTalk, a programming library for protein structures and structure analysis. *BMC Bioinformatics* 5:1–7
52. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, Snoeyink J, Richardson JS, Richardson DC (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* 35:W375–W383
53. Melo F, Feytmans E (1997) Novel knowledge-based mean force potential at atomic level. *J Mol Biol* 267:207–222
54. Melo F, Feytmans E (1998) Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol* 277:1141–1152
55. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM (1992) Stereochemical quality of protein structure coordinates. *Proteins* 12:345–364
56. 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
57. Leach AR (2001) Energy minimisation and related methods for exploring the energy surface. In *Molecular modelling. Principles and applications*. Prentice Hall, Englewood Cliffs, NJ
58. Tama F, Gadea FX, Marques O, Sanejouand YH (2000) Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins* 41:1–7
59. Tama F, Sanejouand YH (2001) Conformational change of proteins arising from normal mode calculations. *Protein Eng* 14:1–6
60. Delarue M, Sanejouand YH (2002) Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. *J Mol Biol* 320:1011–1024
61. 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

62. Hetényi C, Van der Spoel D (2002) Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci* 11:1729–1737
63. Hetényi C, van der Spoel D (2006) Blind docking of drug-sized compounds to proteins with up to a thousand residues. *FEBS Lett* 580:1447–1450
64. Alfonso P, Pampín S, Estrada J, Rodríguez-Rey JC, Giraldo P, Sancho J, Pocovi M (2005) Miglustat (NB-DNJ) works as a chaperone for mutated acid beta-glucosidase in cells transfected with several Gaucher disease mutations. *Blood Cells Mol Dis* 35:268–276
65. Rosenfeld RJ, Goodsell DS, Musah RA, Morris GM, Goodin DB, Olson AJ (2003) Automated docking of ligands to an artificial active site: augmenting crystallographic analysis with computer modeling. *J Comput Aided Mol Des* 17:525–536
66. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA Jr, Stratmann RE, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN, Strain MC, Farkas O, Tomasi J, Barone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Ochterski J, Peterson GA, Ayala P, Cui Q, Morokuma K, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Baboul AG, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Andres JL, Gonzalez C, Head-Gordon M, Replogle ES, Pople JA (1998) Gaussian 98, Revision A.9. Gaussian Inc, Pittsburgh, PA
67. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89:10915–10919
68. Altschul SF (1991) Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol* 219:555–565
69. Brannigan JA, Dodson G, Duggleby HJ, Moody PC, Smith JL, Tomchick DR, Murzin AG (1995) A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. *Nature* 378:416–419
70. Pei J, Grishin NV (2003) Peptidase family U34 belongs to the superfamily of N-terminal nucleophile hydrolases. *Protein Sci* 12:1131–1135
71. Manavalan P, Curtis-Johnson W (1983) Sensitivity of circular dichroism to protein tertiary structure class. *Nature* 305:831–832
72. Liang J, Edelsbrunner H, Woodward C (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci* 7:1884–1897
73. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S (1998) Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. *Proteins* 33:18–29
74. Kisselev OG, Kao J, Ponder JW, Fann YC, Gautam N, Marshall GR (1998) Light-activated rhodopsin induces structural binding motif in G protein alpha subunit. *Proc Natl Acad Sci USA* 95:4270–4275
75. Li HL, Galue A, Meadows L, Ragsdale DS (1999) A molecular basis for the different local anesthetic affinities of resting versus open and inactivated states of the sodium channel. *Mol Pharmacol* 55:134–141
76. Marshall GR, Ragno R, Makara GM, Arimoto R, Kisselev O (1999) Bound conformations for ligands for G-protein coupled receptors. *Lett Pept Sci* 6:283–288
77. Okada A, Miura T, Takeuchi H (2001) Protonation of histidine and histidine-tryptophan interaction in the activation of the M2 ion channel from influenza A virus. *Biochemistry* 40:6053–6060
78. Zacharias N, Dougherty DA (2002) Cation-pi interactions in ligand recognition and catalysis. *Trends Pharmacol Sci* 23:281–287
79. Quenner SW, Neuss N (1982) In: Morin EB, Morgan M (eds) *The chemistry and biology of β -lactam antibiotics*, vol 3. Academic, London, pp 1–81
80. Cremades N, Sancho J, Freire E (2006) The native-state ensemble of proteins provides clues for folding, misfolding and function. *Trends Biochem Sci* 31:494–496

Molecular docking and 3D-QSAR study on 4-(1*H*-indazol-4-yl)phenylamino and aminopyrazolopyridine urea derivatives as kinase insert domain receptor (KDR) inhibitors

Xiaoyun Wu · Shuguang Wu · Wen-Hua Chen

Received: 3 April 2011 / Accepted: 3 June 2011 / Published online: 22 June 2011
© Springer-Verlag 2011

Abstract Vascular endothelial growth factor (VEGF) and its receptor tyrosine kinase VEGFR-2 or kinase insert domain receptor (KDR) have been identified as new promising targets for the design of novel anticancer agents. It is reported that 4-(1*H*-indazol-4-yl)phenylamino and aminopyrazolopyridine urea derivatives exhibit potent inhibitory activities toward KDR. To investigate how their chemical structures relate to the inhibitory activities and to identify the key structural elements that are required in the rational design of potential drug candidates of this class, molecular docking simulations and three-dimensional quantitative structure-activity relationship (3D-QSAR) methods were performed on 78 4-(1*H*-indazol-4-yl)phenylamino and aminopyrazolopyridine urea derivatives as KDR inhibitors. Surflex-dock was used to determine the probable binding conformations of all the compounds at the active site of KDR. As a result, multiple hydrophobic and hydrogen-bonding interactions were found to be two predominant factors that may be used to modulate the inhibitory activities. Comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) 3D-QSAR models were developed based on the docking conformations. The

CoMFA model produced statistically significant results with the cross-validated correlation coefficient q^2 of 0.504 and the non-cross-validated correlation coefficient r^2 of 0.913. The best CoMSIA model was obtained from the combination of steric, electrostatic and hydrophobic fields. Its q^2 and r^2 being 0.595 and 0.947, respectively, indicated that it had higher predictive ability than the CoMFA model. The predictive abilities of the two models were further validated by 14 test compounds, giving the predicted correction coefficients r_{pred}^2 of 0.727 for CoMFA and 0.624 for CoMSIA, respectively. In addition, the CoMFA and CoMSIA models were used to guide the design of a series of new inhibitors of this class with predicted excellent activities. Thus, these models may be used as an efficient tool to predict the inhibitory activities and to guide the future rational design of 4-(1*H*-indazol-4-yl)phenylamino and aminopyrazolopyridine urea derivatives-based novel KDR inhibitors with potent activities.

Keywords Aminopyrazolopyridine ureas · CoMFA · CoMSIA · 4-(1*H*-Indazol-4-yl)phenylamino derivatives · KDR inhibitor · Surflex-dock

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1146-9) contains supplementary material, which is available to authorized users.

X. Wu (✉) · S. Wu · W.-H. Chen (✉)
School of Pharmaceutical Sciences, Southern Medical University,
Guangzhou 510515, China
e-mail: xywugz@163.com

W.-H. Chen
e-mail: whchen@smu.edu.cn

Introduction

Angiogenesis is a normal process for organ development during embryogenesis, wound healing and female reproductive cycling, in which new blood vessels are formed from the pre-existing vasculatures [1, 2]. On the other hand, it has been shown that angiogenesis is a rate-limiting step in tumor development. That is, tumors cannot grow beyond 2–

3 mm in the absence of new vasculatures [3]. This is because tumors need new blood capillaries to create their own nutrient supply, to remove metabolic wastes and to facilitate metastasis of tumor cells to other sites [4, 5]. Among the many pro-angiogenic factors, vascular endothelial growth factor (VEGF) and its receptor tyrosine kinase VEGFR-2 or kinase insert domain receptor (KDR) have been identified as one of the most important regulators for tumor angiogenesis [6, 7]. As shown in many preclinical and clinical studies [8–12], inhibiting of KDR may lead to the development of molecules that are capable of inhibiting angiogenesis, tumor progression, and dissemination. These molecules, namely KDR inhibitors, may have high potentials for the treatment of various diseases. For example, Sutent (sunitinib) [13] and Nexavar (sorafenib tosylate) [14] that have recently been approved by FDA as KDR inhibitors, can be used in the treatment of gastrointestinal stromal tumors and advanced renal cell carcinoma. Therefore, considerable interest has been attracted in the development of novel KDR inhibitors.

Recently, Dai et al. reported a library of potent KDR inhibitors based on 4-(1*H*-indazol-4-yl) phenylamino and aminopyrazolopyridine urea derivatives **1–78** (Table 1) [15, 16]. Among them, compound **19**, i.e., ABT-869, is currently under phase II trials. However, there is a lack of quantitative structure-activity relationship (QSAR) study on these compounds so that exactly how their chemical structures relate to the inhibitory activities remains to be established. To address this as well as to identify the key structural elements that are required in the rational design of potential drug candidates of this class, we performed molecular modeling studies on these 78 4-(1*H*-indazol-4-yl)phenylamino and aminopyrazolopyridine urea derivatives as KDR inhibitors, using molecular docking and 3D-QSAR approaches, including comparative molecular field analysis (CoMFA) [17] and comparative molecular similarity indices analysis (CoMSIA) [18]. Molecular docking was carried out to clarify the binding mode and to identify the bioactive conformation of these compounds with KDR, whereas CoMFA and CoMSIA were applied to gain insights into how steric, electrostatic, hydrophobic and hydrogen-bonding interactions modulated the inhibitory activities.

Materials and methods

Data sets

Compounds **1–78** selected for the present study were taken from the literatures [15, 16], and served as the database in the molecular modeling. Their structures and inhibitory activities are listed in Tables 1 and 2. Among

them, 14 compounds that are asterisk labeled in Table 1 served as the test set, and the rest as the training set. The IC_{50} values (M) were converted to the corresponding pIC_{50} ($=-\log IC_{50}$) and used as dependent variables in the CoMFA and CoMSIA analyses.

Molecular modeling

All the molecular modeling and calculations were performed by using Sybyl 7.3 molecular modeling package [19]. All 78 molecules were initially built in Sybyl 7.3. Structural energy minimization process was performed using the Tripos force field with a distance-dependent dielectric and Powell gradient algorithm with a convergence criterion of 0.001 kcal mol⁻¹. Partial atomic charges were calculated using Gasteiger-Hückel method.

Molecular docking

To identify the probable bioactive conformations of these inhibitors, the Surflex-Dock program [20, 21] that is interfaced with Sybyl 7.3, was used to dock all the compounds into the active site of the KDR (PDB code: 1YWN). This program docks a ligand automatically into the binding site of a receptor, using a protomol-based method and an empirically derived scoring function. The protomol is a unique and important factor of the docking algorithm and means a computational representation of a proposed ligand that interacts with the binding site. Surflex-Dock's scoring function contains the factors that play a crucial role in the ligand-receptor interaction, that is, hydrophobic, polar, repulsive, entropic and solvation terms, and it is a well-recognized method in the field [22, 23]. All the default parameters, as implemented in the Sybyl 7.3, were used. The highest-scored conformation based on the Surflex-Dock scoring functions, was selected as the final bioactive conformation. The alignment (Fig. 1) of these compounds, except compounds **3** and **43** (*vide infra*), based on the docked conformations, was used for CoMFA and CoMSIA.

CoMFA and CoMSIA studies

Standard CoMFA and CoMSIA procedures were performed. A 3D cubic lattice was created automatically by extending at least 4 Å beyond all the investigated molecules in all three axes (X, Y and Z directions) with 2.0 Å grid spacing. The CoMFA steric (Lennard-Jones potential) and electrostatic (Coulomb potential) fields at each lattice were calculated using the standard Tripos force field method. A distance dependent dielectric constant of 1.0 was used, and an sp³ hybridized carbon atom with one positive charge and a radius of 1.52 Å served as a probe atom to calculate the

steric and electrostatic fields. The default cutoff value of $30.0 \text{ kcal mol}^{-1}$ was adopted.

The CoMSIA method defines hydrophobic (H), hydrogen bond donor (D), and hydrogen bond acceptor (A) descriptors, in addition to the steric (S) and electrostatic (E) fields used in CoMFA. The CoMSIA fields were

derived, according to Klebe et al. [18], from the same lattice box that was used in the CoMFA calculations, with a grid spacing of 2 \AA and a probe carbon atom with one positive charge and a radius of 1.0 \AA as implemented in Sybyl. The default value of 0.3 was used as the attenuation factor.

Table 1 Structures of compounds 1-78

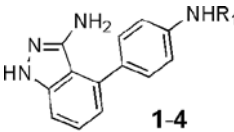
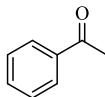
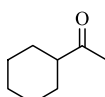
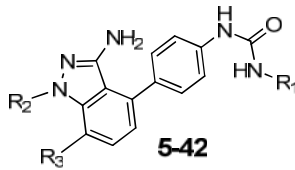


Structure	Compound	Substituent		
	1	R ₁		
	2	H		
	3 ^a			
	4			
	5*	R ₁	R ₂	R ₃
	6	Ph	H	H
	7	<i>m</i> -CH ₃ -Ph	H	H
	8		H	H
	9		H	H
	10*	<i>o</i> -F-Ph	H	H
	11	<i>m</i> -F-Ph	H	H
	12*	<i>p</i> -F-Ph	H	H
	13	<i>o</i> -CH ₃ -Ph	H	H
	14	<i>p</i> -CH ₃ -Ph	H	H
	15	<i>m</i> -Et-Ph	H	H
	16	<i>m</i> -Cl-Ph	H	H
	17	<i>m</i> -Br-Ph	H	H
	18	<i>m</i> -CF ₃ -Ph	H	H
	19	<i>m</i> -OH-Ph	H	H
	20	2-F-5-CH ₃ -Ph	H	H
	21*	4-F-3-CH ₃ -Ph	H	H
	22	3-F-4-CH ₃ -Ph	H	H
	23	2-F-5-CF ₃ -Ph	H	H
	24	<i>m</i> -CH ₃ -Ph	Me	H
		<i>m</i> -CH ₃ -Ph	(CH ₂) ₂ OH	H

Table 1 (continued)

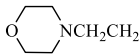

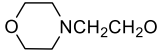



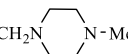
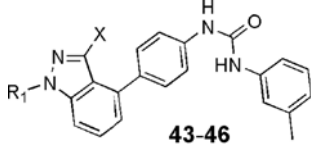
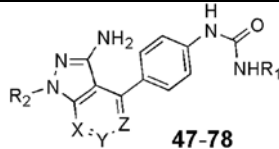



Structure	Compound	Substituent				
	25	<i>m</i> -CH ₃ -Ph 	H			
	26	<i>m</i> -CH ₃ -Ph (CH ₂) ₂ OMe	H			
	27	<i>m</i> -CH ₃ -Ph	Me			
	28	<i>m</i> -CH ₃ -Ph	CH ₃ O			
	29	<i>m</i> -CH ₃ -Ph	F			
	30	<i>m</i> -CH ₃ -Ph	Br			
	31*	<i>m</i> -CH ₃ -Ph	O(CH ₂) ₂ NMe ₂			
	32	<i>m</i> -CH ₃ -Ph	O(CH ₂) ₂ NEt ₂			
	33	<i>m</i> -CH ₃ -Ph	OCH ₂ CH ₂ N 			
	34	<i>m</i> -CH ₃ -Ph	H 			
	35	<i>m</i> -CH ₃ -Ph	H 			
	36	<i>m</i> -CH ₃ -Ph	O(CH ₂) ₂ OMe			
	37	<i>m</i> -Cl-Ph	O(CH ₂) ₂ OMe			
	38	2-F-5-CH ₃ -Ph	O(CH ₂) ₂ OMe			
	39	2-F-5-CH ₃ -Ph	O(CH ₂) ₂ NEt ₂			
	40*	2-F-5-CH ₃ -Ph	OCH ₂ CH ₂ N 			
	41	<i>m</i> -CH ₃ -Ph	H 			
	42*	<i>m</i> -CH ₃ -Ph	H 			
		X	R ₁			
	43^a	NHCOCH ₃	Me			
	44	N(CH ₃) ₂	Me			
	45	H	Me			
	46	H	H			
		R ₁	R ₂	X	Y	Z
	47	<i>m</i> -CH ₃ -Ph	H	CH	N	CH
	48	<i>m</i> -Cl-Ph	H	CH	N	CH
	49*	2-F-5-CH ₃ -Ph	H	CH	N	CH
	50	2-F-5-CF ₃ -Ph	H	CH	N	CH
	51	4-F-3-CH ₃ -Ph	H	CH	N	CH
	52	<i>m</i> -CH ₃ -Ph	Me	CH	N	CH
	53*	<i>m</i> -Cl-Ph	Me	CH	N	CH
	54	2-F-5-CH ₃ -Ph	Me	CH	N	CH
	55	4-F-3-CF ₃ -Ph	Me	CH	N	CH
	56*	Ph	H	N	CH	CH
	57	<i>o</i> -CH ₃ -Ph	H	N	CH	CH

Table 1 (continued)

Structure	Compound	Substituent				
	58	<i>m</i> -CH ₃ -Ph	H	N	CH	CH
	59	<i>p</i> -CH ₃ -Ph	H	N	CH	CH
	60	<i>m</i> -Cl-Ph	H	N	CH	CH
	61	<i>m</i> -CF ₃ -Ph	H	N	CH	CH
	62	<i>m</i> -OCH ₃ -Ph	H	N	CH	CH
	63*	3,5-diF-Ph	H	N	CH	CH
	64	3,5-diMe-Ph	H	N	CH	CH
	65	2-F-5-CH ₃ -Ph	H	N	CH	CH
	66	2-F-5-CF ₃ -Ph	H	N	CH	CH
	67	4-F-3-CH ₃ -Ph	H	N	CH	CH
	68*	4-F-3-CF ₃ -Ph	H	N	CH	CH
	69	3-Cl-4-F-Ph	H	N	CH	CH
	70		H	N	CH	CH
	71		H	N	CH	CH
	72		H	N	CH	CH
	73*	CH ₃ (CH ₂) ₃	H	N	CH	CH
	74*	<i>m</i> -Cl-Ph	H	CH	CH	N
	75	<i>m</i> -CF ₃ -Ph	H	CH	CH	N
	76	2-F-5-CH ₃ -Ph	H	CH	CH	N
	77	3,5-diCH ₃ -Ph	H	CH	CH	N
	78	4-F-3-CH ₃ -Ph	H	CH	CH	N

^a Outlier

* Test set

PLS regression analysis and validation of QSAR models

Partial least squares (PLS) approach was used to derive the 3D QSAR models. The CoMFA and CoMSIA descriptors were used as independent variables and the pIC₅₀ values were used as dependent variables. CoMFA and CoMSIA column filtering was set to 2.0 kcal mol⁻¹ to improve the signal-to-noise ratio. The leave-one-out (LOO) cross-validation was carried out to obtain the optimal number of components (N) and the correlation coefficient q². The obtained N was then used to derive the final QSAR model and to obtain the non-cross-validation correlation coefficient r², standard error of estimate (SEE), and F-value. To assess the predictive power of the derived 3D-models, a set of test compounds that had known biological activities and

that were not included in the model generation, was used to validate the obtained models.

Results and discussion

Dock analysis

To determine the probable binding conformations of these compounds, Surflex-Dock was used to dock all the compounds into the active site of KDR. The docking reliability was validated using the known X-ray structure of KDR complexed with a small ligand **79** (Fig. 2). The co-crystallized **79** was re-docked into the binding site, and the docked conformation with the highest total score was

Table 2 The experimental pIC₅₀ values, predicted pIC₅₀ and their residuals of the training and test set compounds

Compound	Exp pIC ₅₀	CoMFA		CoMSIA	
		Pred	Resid	Pred	Resid
1	5.320	5.285	0.035	5.399	-0.079
2	4.903	5.091	-0.188	4.629	0.274
3 ^a	4.665	–	–	–	–
4	7.066	6.729	0.337	7.262	-0.196
5*	7.194	7.815	-0.621	7.891	-0.697
6	8.523	8.329	0.194	8.508	0.015
7	7.824	7.466	0.358	7.524	0.300
8	6.815	6.563	0.253	6.950	-0.134
9	7.086	7.513	-0.426	7.579	-0.493
10*	7.638	7.676	-0.037	7.548	0.090
11	7.174	7.546	-0.373	7.511	-0.337
12*	7.061	7.700	-0.640	7.640	-0.580
13	7.921	8.099	-0.179	8.276	-0.355
14	8.222	8.297	-0.075	8.399	-0.177
15	8.097	8.073	0.024	7.824	0.272
16	7.444	7.975	-0.531	7.680	-0.236
17	8.000	7.781	0.219	7.615	0.385
18	7.260	7.213	0.047	7.256	0.003
19	8.398	7.766	0.632	7.742	0.656
20	8.398	7.945	0.453	8.375	0.023
21*	7.444	7.563	-0.120	7.529	-0.085
22	7.046	7.437	-0.392	7.500	-0.454
23	7.959	7.586	0.373	7.645	0.313
24	6.222	6.283	-0.061	6.189	0.033
25	5.886	6.154	-0.268	5.883	0.003
26	5.510	6.110	-0.600	5.645	-0.135
27	8.523	7.907	0.616	8.215	0.308
28	7.585	7.614	-0.029	7.919	-0.334
29	8.301	8.098	0.203	7.969	0.332
30	8.000	7.921	0.079	8.090	-0.090
31*	7.420	7.930	-0.509	8.036	-0.616
32	7.456	7.545	-0.089	7.211	0.245
33	7.509	7.739	-0.230	7.704	-0.195
34	7.678	7.795	-0.117	7.782	-0.104
35	7.602	7.809	-0.207	7.563	0.039
36	7.678	7.335	0.343	7.677	0.001
37	7.886	7.874	0.012	7.747	0.139
38	7.678	7.168	0.510	7.278	0.400
39	7.131	7.238	-0.107	7.402	-0.272
40*	7.208	7.597	-0.389	7.863	-0.656
41	6.409	6.409	0.000	6.422	-0.013
42*	5.921	6.403	-0.483	6.400	-0.479
43 ^a	4.079	–	–	–	–
44	5.495	5.040	0.455	5.618	-0.124
45	6.333	6.681	-0.349	6.479	-0.147
46	7.745	7.702	0.043	7.949	-0.204

Table 2 (continued)

Compound	Exp pIC ₅₀	CoMFA		CoMSIA	
		Pred	Resid	Pred	Resid
47	9.155	9.255	-0.100	9.133	0.022
48	8.921	8.915	0.006	8.959	-0.038
49*	8.678	9.172	-0.494	9.067	-0.389
50	9.000	9.028	-0.028	9.097	-0.097
51	9.155	9.173	-0.018	9.315	-0.16
52	8.337	8.282	0.055	8.200	0.137
53*	8.824	8.308	0.516	8.114	0.710
54	8.114	8.342	-0.228	8.321	-0.208
55	8.432	8.715	-0.283	8.396	0.036
56*	8.347	8.246	0.101	8.813	-0.466
57	7.456	7.957	-0.501	7.758	-0.302
58	8.699	8.773	-0.074	8.657	0.042
59	8.538	8.480	0.058	8.451	0.087
60	9.000	8.715	0.285	8.754	0.246
61	8.770	8.681	0.088	8.676	0.094
62	8.420	8.422	-0.002	8.415	0.005
63*	8.201	8.051	0.150	7.990	0.211
64	9.000	8.951	0.049	8.987	0.013
65	8.699	8.092	0.607	8.420	0.278
66	8.620	8.555	0.065	8.624	-0.004
67	8.959	9.095	-0.137	8.785	0.174
68*	8.699	9.434	-0.735	9.114	-0.415
69	8.745	8.770	-0.025	8.710	0.035
70	7.896	7.759	0.138	7.964	-0.068
71	6.939	7.479	-0.540	6.816	0.123
72	7.194	7.464	-0.270	7.135	0.059
73*	6.620	6.496	0.124	6.685	-0.066
74*	7.921	7.836	0.085	7.600	0.321
75	8.149	8.070	0.079	8.021	0.128
76	7.745	7.469	0.276	7.700	0.045
77	7.921	8.337	-0.416	8.325	-0.404
78	7.699	7.746	-0.048	7.603	0.096

^a Outlier

* Test set

selected as the most probable binding conformation (Fig. 3). The low root mean-square deviation (RMSD) of 0.63 Å between the docked and the crystal conformations indicated the high reliability of Surflex-dock in reproducing the experimentally observed binding mode for these KDR inhibitors. As shown in Fig. 3a, redocked **79** was almost in the same position with co-crystallized **79** at the active site of KDR. Therefore, Surflex-Dock method was used in search for the binding conformations of the whole dataset. All the compounds, except **3** and **43**, were successfully docked into the active site of KDR. Compounds **3** and **43** were outliers, possibly because they had different struc-

Fig. 1 Superimposition of 76 successfully docked compounds



tures from the other compounds. That is, compound **3** was a sulfonamide and compound **43** had an acetylated 3-amino group.

It should be noted that the overall conformations of the successfully docked 76 compounds at the active site of KDR were similar to that of ligand **79** in the X-ray structure co-crystallized with KDR. Compound **22** (Fig. 2) having the same 2'-fluoro-5'-trifluoromethylphenyl group with compound **79**, was selected for detailed analysis. Its most probable binding mode and the main residues involved in the interaction are displayed in Fig. 3b and c, respectively. It can be seen that compound **22** has multiple H-bonding and hydrophobic interactions with the hinge-binding region of the kinase. Specifically, the 3-NH₂ of the 3-aminoindazole forms an H-bond with Glu915 at the angle of 177.1° and the distance of 2.948 Å. The 1-NH and 2-nitrogen of the 3-aminoindazole form two H-bonds with Cys917-CO and Cys917-NH at the angles of 132.4° and 159.6°, and the

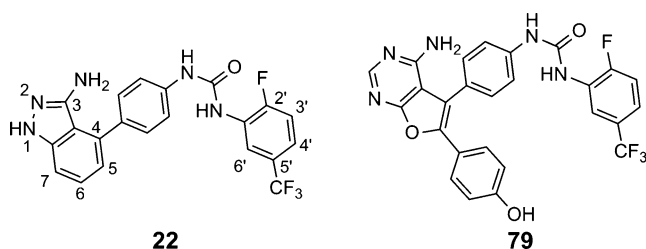


Fig. 2 Structures of atom-numbered compound **22** and ligand **79** for 1YWN

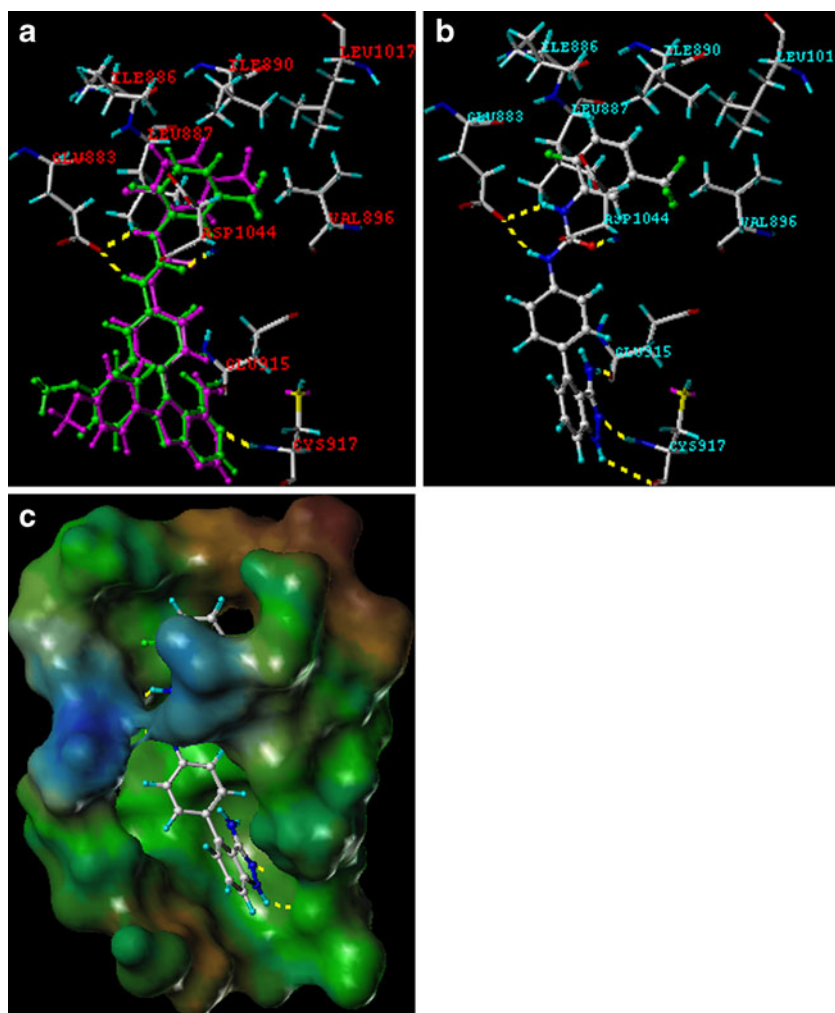
distance of 3.438 Å and 3.059 Å, respectively. The NH moieties of the urea form H-bonding interaction with the backbone of Asp1044 at the angle of 122.2° and the distance of 2.847 Å, respectively, and CO moieties of the urea form two H-bonding interactions with the carboxylic acid residue of Glu883 at the angles of 134.2° and 134.9° and the distances of 2.874 Å and 2.839 Å, respectively. In addition, the terminal 2'-fluoro-5'-trifluoromethylphenyl group is accommodated into the hydrophobic pocket that is composed of residues Ile886, Leu887, Ile890, Val896 and Leu1017.

CoMFA and CoMSIA results

Based on the docked alignment, CoMFA and CoMSIA were performed. Because compounds **3** and **43** were outliers, they were removed from the dataset. Thus, totally 76 molecules were used to build the models. On the basis of the diversity in both the structures and activities, these compounds were divided into two groups, 62 compounds as the training set and 14 compounds as the test set. CoMFA and CoMSIA 3D-QSAR models were obtained using standard procedures. The statistical results of CoMFA and CoMSIA PLS analysis are presented in Table 3. The CoMFA model gave a cross-validated correlation coefficient q^2 of 0.504 with an optimal number of principal components (N) of 5 and a non-cross-validated correlation coefficient r^2 of 0.913. The corresponding contributions of steric and electrostatic fields were 57.4% and 42.6%, respectively. The model was satisfactory from the viewpoint of statistical significance. The predicted pIC_{50} values were in good agreement with the experimental values (Table 2 and Fig. 4a), indicating the strong predictive ability of the obtained model.

In CoMSIA analysis, the five different descriptor fields, that is, the hydrophobic (H), hydrogen bond donor (D) and acceptor (A), steric (S) and electrostatic (E) fields, are not totally independent of each other. Such dependency on individual field usually decreases the statistical significance of the CoMSIA models. Studies on the possible combinations of different fields (Table S1) indicated that the best CoMSIA model with the highest cross-validated correlation coefficient q^2 was obtained from the combination of steric, electrostatic and hydrophobic fields (entry 9, Table S1). Thus, the corresponding model was selected for further analysis and for the prediction of the test compounds. From the cross-validation results, it can be seen that the CoMSIA model has better predictive ability than CoMFA model, possibly because of the importance of the hydrophobic field in the activity of these compounds. The statistical details are summarized in Table 3. The CoMSIA model gave a cross-validated correlation coefficient q^2 of 0.595,

Fig. 3 Binding conformations at the active site of KDR. Key residues are displayed and hydrogen bonds are displayed in dotted lines. (a) Co-crystallized (magenta) and re-docked (green) **79**. (b) Docked compound **22**. (c) MOLCAD lipophilic potential surface of the binding pockets with the docked compound **22**



an optimal number of principal components of eight and a non-cross validated correlation coefficient r^2 of 0.947, indicating that an acceptable CoMSIA model was success-

fully constructed. The corresponding contributions of steric, electrostatic and hydrophobic fields were 25.1%, 42.2% and 32.7%, respectively. Table 2 lists the experimental and

Table 3 Statistical parameters for the CoMFA and CoMSIA models

	Field	N	q^2	r^2	SEE	F	r_{pred}^2	Field contribution		
								S	E	H
CoMFA	S, E	5	0.504	0.913	0.305	116.991	0.727	0.574	0.426	-
CoMSIA	S, E, H	8	0.595	0.947	0.243	119.023	0.624	0.251	0.422	0.327

q^2 : Cross-validated correlation coefficient

r^2 : non-cross-validated correlation coefficient

r_{pred}^2 : predictive correlation coefficient r^2

SEE: standard error of estimate

F: the Fischer ratio

N: optimal number of principal components

S: steric field

E: electrostatic field

H: hydrophobic field

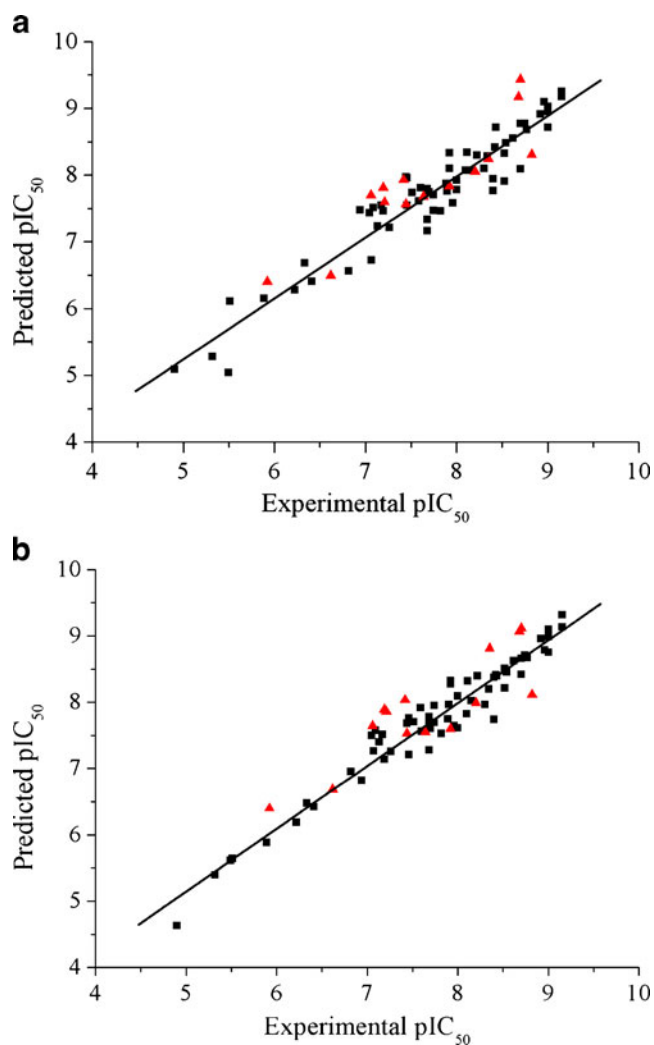


Fig. 4 Graph of experimental versus predicted pIC_{50} of the training (■) and test (▲) sets from the CoMFA (a) and CoMSIA (b) models

predicted activities, and the residual values of the training set. Figure 4b shows the relationship between the predicted and experimental pIC_{50} values from the CoMSIA model.

Validation of the 3D-QSAR models

The predictive powers of the CoMFA and CoMSIA models were validated by the 14 test compounds. The predicted activity of each compound is listed in Table 2 and shown in Fig. 4. It can be seen that the predicted pIC_{50} values of the test compounds were in good agreement with the experimental data within an acceptable error range, with the r_{pred}^2 being 0.727 and 0.624 for CoMFA and CoMSIA models, respectively. This result indicates that the CoMFA and CoMSIA models could be used to predict the inhibitory activities and to guide the design of 4-(1*H*-indazol-4-yl) phenylamino and aminopyrazolopyridine urea derivatives-based novel KDR inhibitors with potent activities.

Contour analysis

To visualize the results of the CoMFA and CoMSIA models, 3D coefficient contour maps were generated. The CoMFA and CoMSIA results were graphically interpreted by field contribution maps using the STDEV*COEFF field type. Compound 22 was displayed in the map in aid of visualization. All the contours represented the default 80% and 20% level contributions for favorable and unfavorable regions, respectively.

CoMFA contour maps

The CoMFA contour maps of steric and electrostatic fields are shown in Fig. 5. In the map of steric field, the green contours represent regions in which bulky groups confer an increased activity, whereas the yellow ones represent regions where bulky groups may lead to a decreased activity. Similarly, in the map of electrostatic field, the blue contours indicate the regions where the electropositive substitution increases the inhibitory activity, whereas the red contours indicate the regions where the electronegative substitution increases the activity.

It can be seen from Fig. 5 that the contours having different physicochemical fields are mainly distributed near the regions of the 3-aminoindazolyl and the terminal 2'-fluoro-5'-trifluoromethylphenyl of the reference compound 22. This suggests that the functional groups located in these two regions may play a crucial role in modulating the affinity toward KDR. In the CoMFA steric contour map (Fig. 5a), a large green contour near the 3'-, 4'-, and 5'-positions of terminal 2'-fluoro-5'-trifluoromethylphenyl

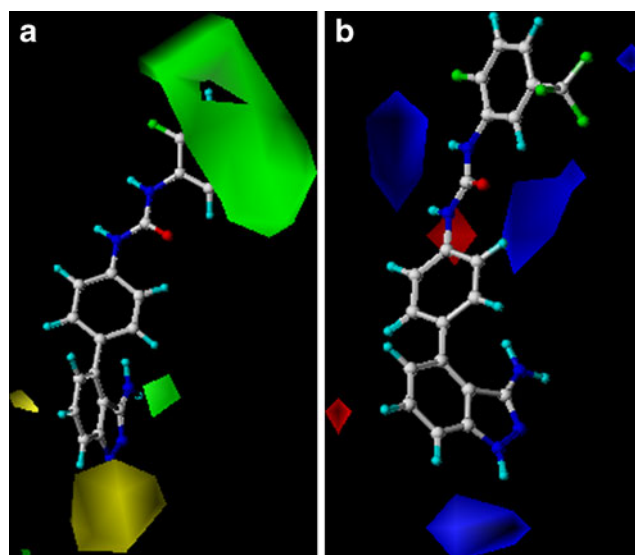


Fig. 5 CoMFA STDEV*COEFF contour maps. (a) Favorable (green) and unfavorable (yellow) steric fields. (b) Electropositive (blue) and electronegative (red) fields. Compound 22 was overlaid in each map

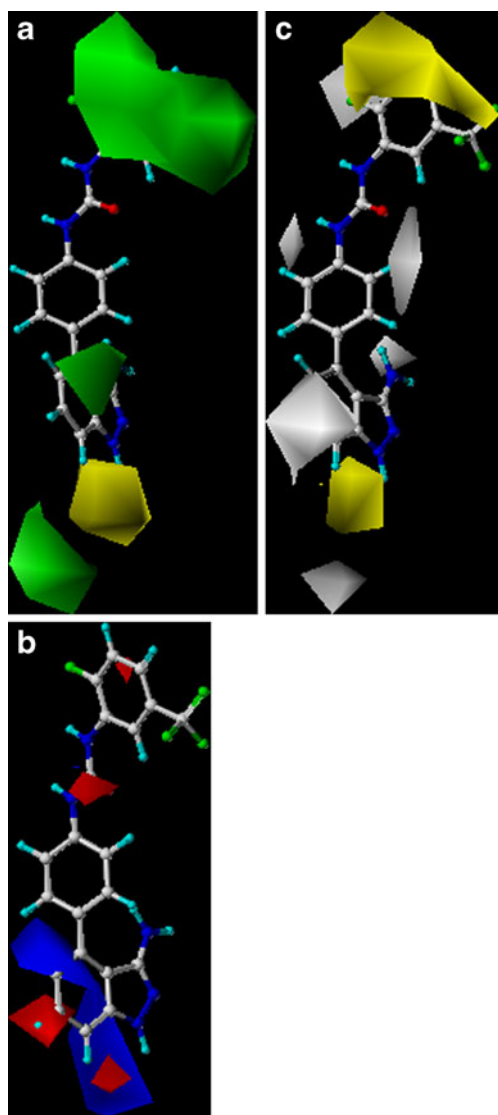
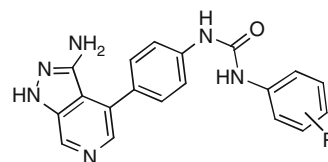


Fig. 6 CoMSIA STDEV*COEFF contour maps. (a) Favorable (green) and unfavorable (yellow) steric fields. (b) Electropositive (blue) and electronegative (red) fields. (c) Favorable (yellow) and unfavorable (white) hydrophobic fields. Compound **22** was overlaid in each plot

group of compounds **22** suggests that introducing of bulky groups at these positions would increase the activity. Moreover, in the CoMFA electrostatic contour map, a small blue contour near the 3'-position of 2'-fluoro-5'-trifluoromethylphenyl group of **22** indicates that introducing of electropositive groups around this position would increase the inhibitory activity. This, together with the green contour discussed above, suggests that electropositive and bulky groups in the vicinity of the 3'-position of the terminal phenyl group are favored by the CoMFA model. In CoMFA steric contour map, a big yellow contour near the 1- and 7-positions suggests that steric bulkiness at these positions is unfavorable by the model. For example, compound **53** bearing methyl group at the 1-position has low inhibitory

activity. A large blue contour near the 1- and 7-positions indicates that introducing electropositive groups at these positions would increase the inhibitory activity. Thus, compounds **6** and **27** having electropositive hydrogen or methyl group at the 7-position are more active than compounds **31** and **42** having electronegative groups at the same position. This, together with the yellow contour discussed above, indicates that any bulky or electronegative substitution at the 1- or 7-position may lead to significant loss of the inhibitory activity. In addition, the small yellow and red contours near the 6-position suggest that small and electronegative groups are favorable in improving the inhibitory activity, which is in agreement with the fact that compound **49** is more active than its corresponding compound **19**.

Table 4 Structures and predicted pIC_{50} values of newly designed derivatives



Compound	R	Predicted pIC_{50}	
		CoMFA	CoMSIA
D1	3-CF ₃	9.855	8.934
D2	3-Et	9.455	9.334
D3	4-Me	9.241	8.997
D4	4-Et	9.342	9.229
D5	3,4-2Me	9.396	9.312
D6	3,4-2Et	9.444	9.746
D7	3,5-2Me	9.385	9.228
D8	3,5-2Et	9.606	9.569
D9	3,5-2Me	9.385	9.228
D10	3-Cl-4-F	9.049	8.631
D11	3,5-2Cl	9.192	8.845
D12	3,4,5-3Me	9.611	9.396
D13	3,4,5-3Et	9.635	10.026

CoMSIA contour maps

The steric and electrostatic contour maps of CoMSIA are shown in Fig. 6a and b, respectively. These contours are quite similar to those of CoMFA. In addition to the CoMFA contours, large green and blue contours near the 5-position suggest that bulky and electropositive groups around this position would increase the activity. This is supported by the fact that incorporating of less bulky and electronegative N into the 5-position leads to a decrease in the activity.

Figure 6c shows the hydrophobic contour map of CoMSIA in which yellow and gray contours indicate the regions where hydrophobic and hydrophilic groups are favored by the model, respectively. A large yellow contour near the 3'-, 4'- and 5'-positions of the terminal 2'-fluoro-5'-trifluoromethylphenyl group of compound 22 indicates that hydrophobic substitution at these positions would increase the activity. This hydrophobic interaction may be very important for improving of the binding affinity, since it is also observed in the CoMFA and CoMSIA steric contour maps. The gray contour near the 2'-position indicates that hydrophilic substitution at this position is favorable. For example, compounds 5 and 56 with hydrogen at the 2'-position are more active than compounds 9 and 57 with F or methyl group at the same position. The gray contour near the 6-position indicates that hydrophilic groups located in this place are favored by the model, which can be seen from the fact that incorporating of hydrophilic N into the 6-position leads to an increase in the activity. A yellow contour near the 1-position indicates that hydrophobic groups that are located at this place are favored by the model. This is in agreement with the fact that compounds 6 and 27 having methyl or hydrogen at the 7-position are more active than compounds 31 and 42 having hydrophilic groups at the same position. A gray contour near the 3-amino group reveals the importance of the hydrophilic amino group on the indazolyl ring in the enhancement of the inhibitory activities.

Design of new inhibitors

As shown above, molecular docking and 3D-QSAR analyses provided detailed insight into the structural requirements for potent activity of the inhibitors of this class. That is, bulky or electronegative substitutions at the 1- or 7-position may lead to significant loss of the activity. Incorporating of small and electronegative nitrogen atom into the 5-position of indazolyl ring may lead to a decrease in the activity, whereas incorporating of a nitrogen atom into the 6-position may lead to an increase in the activity. In addition, appropriately bulky and strongly hydrophobic groups at the 3'-, 4'- and 5'-positin of the terminal phenyl group may greatly increase the activity. To show the practical values of these structure-activity relationships,

we designed a series of new inhibitors and predicted their pIC_{50} values by the established CoMFA and CoMSIA models (Table 4). It can be seen that all the designed derivatives showed better activities than compound 19, and most of the designed derivatives showed higher activities than compounds 47 and 51, both of which were the most active in the database. These results obtained from the developed models serve as computational predictions which can be used to guide the design of new potent inhibitors.

Conclusions

Molecular docking and 3D-QSAR analyses have been successfully applied to a set of recently synthesized KDR inhibitors based on 4-(1*H*-indazol-4-yl)phenylamino and aminopyrazolopyridine urea derivatives. The binding mode of these KDR inhibitors was clarified by Surflex-dock. The results suggest that multiple hydrophobic and hydrogen-bonding interactions are two predominant factors that may be used to modulate the inhibitory activities. Based on the docked alignments, highly predictive CoMFA and CoMSIA models were developed. These two models showed statistically significant results in terms of cross-validated coefficient q^2 and conventional coefficient r^2 , and their predictive capabilities were verified by the test compounds. Based on the obtained structure-activity relationships, a series of new inhibitors with excellent activities predicted by the developed CoMFA and CoMSIA models were designed. Thus, these models can be used as a tool to guide the future rational design of 4-(1*H*-indazol-4-yl)phenylamino and aminopyrazolopyridine urea derivatives-based novel KDR inhibitors with potent activities.

References

1. Risau W (1997) Mechanisms of angiogenesis. *Nature* 386:671–674
2. Ribatti D, Vacca A, Nico B, Roncali L, Dammacco F (2001) Postnatal vasculogenesis. *Mech Dev* 100:157–163
3. Folkman J (1971) Tumor angiogenesis: therapeutic implications. *N Engl J Med* 285:1182–1186
4. Hanahan D, Folkmann J (1996) Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *Cell* 86:353–364
5. Liotta LA, Steeg PS, Stetler-Stevenson WG (1991) Cancer metastasis and angiogenesis: an imbalance of positive and negative regulation. *Cell* 64:327–336
6. Hicklin DJ, Ellis LM (2005) Role of the vascular endothelial growth factor pathway in tumor growth and angiogenesis. *J Clin Oncol* 23:1011–1027
7. Ferrara N (2004) Vascular endothelial growth factor: basic science and clinical progress. *Endocr Rev* 25:581–611
8. Baka S, Clamp AR, Jayson GC (2006) A review of the latest clinical compounds to inhibit VEGF in pathological angiogenesis. *Expert Opin Ther Targets* 10:867–876

9. Sepp-Lorenzino L, Thomas KA (2002) Antiangiogenic agents targeting vascular endothelial growth factor and its receptors in clinical development. *Expert Opin Invest Drugs* 11:1447–1465
10. Klebl BM, Müller G (2005) Second-generation kinase inhibitors. *Expert Opin Ther Targets* 9:975–993
11. Supuran CT, Scozzafava A (2004) Protein tyrosine kinase inhibitors as anticancer agents. *Expert Opin Ther Pat* 14:35–53
12. Holmes K, Roberts OL, Thomas AM, Cross MJ (2007) Vascular endothelial growth factor receptor-2: structure, function, intracellular signaling and therapeutic inhibition. *Cell Signal* 19:2003–2012
13. Sakamoto KM (2004) SU-11248 (SUGEN). *Curr Opin Invest Drugs* 5:1329–1339
14. Ahman T, Eisen T (2004) Kinase inhibition with BAY43-9006 in renal cell carcinoma. *Clin Cancer Res* 10:6388s–6392s
15. Dai YJ, Hartandi K, Ji ZQ, Ahmed AA, Albert DH, Bauch JL, Bouska JJ, Bousquet PF, Cunha GA, Glaser KB, Harris CM, Hickman D, Guo J, Li J, Marcotte PA, Marsh KC, Moskey MD, Martin RL, Olson AM, Osterling DJ, Pease LJ, Soni NB, Stewart KD, Stoll VS, Tapang P, Reuter DR, Davidsen SK, Michaelides MR (2007) Discovery of N-(4-(3-amino-1H-indazol-4-yl)phenyl)-N'-(2-fluoro-5-methylphenyl)urea (ABT-869), a 3-aminoindazole-based orally active multitargeted receptor tyrosine kinase inhibitor. *J Med Chem* 50:1584–1597
16. Dai YJ, Hartandi K, Soni NB, Pease LJ, Reuter DR, Olson AM, Osterling DJ, Doktor SZ, Albert DH, Bouska JJ, Glaser KB, Marcotte PA, Stewart KD, Davidsen SK, Michaelides MR (2008) Identification of aminopyrazolopyridine ureas as potent VEGFR/PDGFR multitargeted kinase inhibitors. *Bioorg Med Chem Lett* 18:386–390
17. Cramer RD 3rd, 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
18. 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
19. SYBYL 7.3 is available from Tripos Associates Inc, 1699 S Hanley Rd, St Louis, MO 631444, USA
20. Jain AN (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem* 46:499–511
21. Jain AN (2007) Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J Comput Aided Mol Des* 21:281–306
22. Pan J, Liu GY, Cheng J, Chen XJ, Ju XL (2010) CoMFA and molecular docking studies of benzoxazoles and benzothiazoles as CYP450 1A1 inhibitors. *Eur J Med Chem* 45:967–972
23. Sun JY, Cai SX, Yan N, Mei H (2010) Docking and 3D-QSAR studies of influenza neuraminidase inhibitors using three-dimensional holographic vector of atomic interaction field analysis. *Eur J Med Chem* 45:1008–1014

Homology modeling, molecular dynamics and QM/MM study of the regulatory protein PhoP from *Corynebacterium pseudotuberculosis*

Gleiciane Moraes · Vasco Azevedo · Marcília Costa · Anderson Miyoshi · Artur Silva · Vivian da Silva · Diana de Oliveira · Maria Fátima Teixeira · Jerônimo Lameira · Cláudio Nahum Alves

Received: 20 January 2011 / Accepted: 1 June 2011 / Published online: 24 June 2011
© Springer-Verlag 2011

Abstract *Corynebacterium pseudotuberculosis* is a facultatively intracellular Gram-positive bacterium that causes caseous lymphadenitis, principally in sheep and goats, though sometimes in other species of animals, leading to considerable economic losses. This pathogen has a TCS known as PhoPR, which consists of a sensory histidine kinase protein (PhoR) and an intracellular response regulator protein (PhoP). This system is involved in the regulation of proteins present in various processes, including virulence. The regulation is activated by PhoP protein phosphorylation, an event that requires a magnesium (Mg^{2+}) ion. Here we describe the 3D structure of the regulatory response protein (PhoP) of *C. pseudotuberculosis* through molecular modeling by homology. The model generated provides the first

structural information on a full-length member of the OmpR/PhoP subfamily. Classical molecular dynamics was used to investigate the stability of the model. In addition, we used quantum mechanical/molecular mechanical techniques to perform (internal, potential) energy optimizations to determine the interaction energy between the Mg^{2+} ion and the structure of the PhoP protein. Analysis of the interaction energy residue by residue shows that Asp-16 and Asp-59 play an important role in the protein- Mg^{2+} ion interactions. These results may be useful for the future development of a new vaccine against tuberculosis based on genetic attenuation via a point mutation that results in the polar residue Asp-16 and/or Asp-59 being replaced with a nonpolar residue in the DNA-binding domain of PhoP of *C. pseudotuberculosis*.

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1145-x) contains supplementary material, which is available to authorized users.

G. Moraes · J. Lameira · C. N. Alves
Laboratório de Planejamento e Desenvolvimento de Fármacos,
Universidade Federal do Pará,
Belém, PA, Brazil

G. Moraes · A. Silva · J. Lameira (✉)
Instituto de Ciências Biológicas, Universidade Federal do Pará,
Belém, PA, Brazil
e-mail: lameira@ufpa.br

V. Azevedo · A. Miyoshi · V. da Silva
Laboratório de Genética Celular e Molecular,
Universidade Federal de Minas Gerais,
Belo Horizonte, MG, Brazil

M. Costa · D. de Oliveira · M. F. Teixeira
Núcleo de Genômica e Bioinformática,
Universidade Estadual do Ceará,
Fortaleza, CE, Brazil

Keywords Two-component regulatory system · PhoPR · Molecular dynamics · Quantitative mechanics/molecular mechanics · Molecular homology

Introduction

The bacterium *Corynebacterium pseudotuberculosis* is a facultative intracellular parasite in the actinobacteria class that is Gram-positive [1] and provokes disease in animals such as sheep, goats and horses [2]. In small ruminants, it is the causative agent of caseous lymphadenitis (CL), a chronic disease that forms suppurative abscesses, which are normally superficial, though they also frequently disseminate to visceral organs [3]. Currently, the most common therapy for CL is drainage of the superficial lymph nodes, followed by antibiotic therapy. However, this form of treatment has limited effectiveness, as it does not eliminate all of the bacteria and is not viable when internal lymph nodes and

organs are affected [4]. Another disadvantage is the high cost of antibiotic treatments, as well as the difficulty involved in getting the antibiotic into the abscess capsule [5]. There are still no completely effective diagnostic methods, vaccines or treatments for CL [1]. In the search for successful prophylactic alternatives for sheep and goats, various strategies are being tested, such as live recombinant vaccines and DNA vaccines [6]. Currently, the commercially available vaccines are based on the inactivation of phospholipase D (PLD) [1], a potent exotoxin with sphingomyelinase activity that aids the dissemination of this pathogen in the host; however, revaccination is necessary every six months in goats [7, 8]. Therefore, new genomic targets that could be candidates for CL vaccines are needed.

The gene *phoP* has been found in the genome of *C. pseudotuberculosis*, which is currently well characterized in *Mycobacterium tuberculosis*. It is known that in *M. tuberculosis*—another actinobacteria—the gene *phoP* plays an important role in pathogenicity, as it is involved in the secretion of proteins involved in virulence [9, 10]. Consequently, this gene is a potential candidate for the development of a CL vaccine. The gene *phoP* is part of the two-component regulatory system (TCS) PhoPR; this double system is composed of the histidine kinase sensor (PhoR) and a regulatory protein (PhoP) [9]. This system is capable of detecting, responding to and adapting to changes in the environment, favoring bacterial survival in inhospitable environments [11].

The basic biochemical events involved in two-component transduction systems were first established by Ninfa and Magasanik [12] for the regulatory system (NR system) that responds to available nitrogen sources in *Escherichia coli*. The mechanism of operation of this TCS involves the phosphorylation of the transmembrane protein (PhoR) in a conserved histidine residue in response to external changes. This signal is then transferred to the receptor domain (N-terminal domain of PhoP) of the regulatory response protein (PhoP) through the transfer of a phosphoryl group to a conserved aspartate residue, resulting in regulatory domain binding (C-terminal domain of PhoP) to DNA, which controls transcription [13]. According to Buckler et al. [13], phosphorylation of the aspartate residue (Asp-53) that is conserved in PhoP occurs through autocatalysis, and requires Mg^{2+} ions. Reversibility of phosphorylation and dephosphorylation is a key mechanism through which extracellular signals are translated into cellular responses. Phosphorylation of the response regulator changes its conformation and allows it to interact with other components, resulting in changes that can regulate response through binding to DNA, influencing transcription [14].

In other bacterial systems, the development of PhoP-attenuated strains as vaccine candidates has been investigated, for example in *Salmonella*. Hohmann et al. [15]

reported that a PhoP/PhoQ-deleted *Salmonella typhi* mutant was safe and immunogenic when delivered as a single dose in humans. This has been developed further as an oral vaccine expressing heterologous antigens [16]. Similar experiments could be carried out with PhoP-attenuated *C. pseudotuberculosis* to test its potential as a candidate live vaccine against pseudotuberculosis.

Little is known about the three-dimensional (3D) structure of PhoP. Full-length structures of transcription regulators homologous to PhoP are difficult to obtain, due to a highly flexible interdomain that is composed of four beta sheets; this interdomain is located between the domain receptor and the domain regulator, which makes protein resolution difficult [13]. However, in the Protein Data Bank (PDB) there are crystallographic structures for homologous proteins, including the full-length response regulator from *M. tuberculosis* (access code 2OQR in PDB) and the response regulator from *Thermotoga maritima* (access code 1KGS in PDB), both of which belong to the OmpR/PhoB subfamily. Using the available structures, new models of homologous proteins can be constructed using homology modeling techniques.

Homology modeling allows the construction of the secondary structure of a protein based on the primary structure. This technique is only possible because the 3D structures of homologous proteins are conserved during the evolutionary process [17]—especially functional residues, since preserving the structure is crucial to the maintenance and performance of specific functions [17].

In this report, we provide the first structural information on a full-length member of the OmpR/PhoP subfamily of the two-component system. The model of regulatory protein PhoP of *C. pseudotuberculosis* was obtained by molecular homology methods. We have also investigated the stability of the model using molecular dynamic (MD) simulation, where the structural flexibility of the interdomain consisting of four beta sheets was exploited. In addition, a hybrid quantum mechanics/molecular mechanics (QM/MM) approach was used to determine the interaction energy between Mg^{2+} ion and PhoP protein, in order to quantify the interactions of this ion with catalytic protein residues. Finally, calculations of surface electrostatic potential maps were performed to explore the interactions of the DNA–PhoP complex.

Methods

Modeling the PhoP protein of *Corynebacterium pseudotuberculosis*

The primary structure of the PhoP protein was obtained from the sequence generated with the complete genome sequence of line CP1002 biovar *ovis* of *C. pseudotuberculosis* isolated from goats, provided by Dr. Roberto Meyer of

the Immunology Laboratory of the Federal University of Bahia [18].

Through alignment in the server *P-fam* [19], it was found that the study sequence was part of a family of transcription regulatory proteins that, together with a histidine kinase sensor, composes a TCS [20]. The model was obtained using homology modeling with the program Modeller, taking into consideration special restrictions [21]. These restrictions can involve distances, angles, dihedral angles, and pairs of dihedral angles. Automated construction of models by structural homology uses geometric distance optimization techniques in order to accommodate special restrictions, based on known homologous structures.

The template was selected through a search of the Protein Data Bank (PDB) [22]. The known 3D structure that was selected as a template was transcription factor DrrD from *T. maritima* (access code 1KGS in PDB), which was found to have 39% identity and 60% sequence similarity with the target.

The Modeller program was applied to generate 20 satisfactory models for the PhoP protein of *Corynebacterium pseudotuberculosis*. The model with the lowest energy and the lowest restraint violation was selected to construct the system. The models obtained by homology modeling were also validated considering the root mean square deviation (RMSD), which evaluates how much the model deviates from the template (1KGS). The stereochemical qualities of the models obtained were checked with PROCHECK [23]. Furthermore, the quality of the model was evaluated using Verify3D [24], ProSA [25] and ANOLEA [26].

Molecular dynamics simulation (MD)

We have recently carried out MD simulations to study protein–inhibitor interaction energies [27, 28]. In this report, molecular dynamic simulations were carried out to investigate the stability of the PhoP model obtained by homology modeling. The computational model for MD simulation was taken to be the best structure generated through molecular modeling by homology, which we termed “PhoPCp.” The Mg^{2+} ion was added to the structure of PhoPCp because this ion is present in many two-component regulatory systems [29].

Since the standard pK_a values of ionizable groups can be shifted by the local protein environment, accurate assignment of the protonation states of all of these residues at pH 7 was carried out. The pK_a values of the amino acids residues were determined with PROPKA 2.0, considering the pH to be 7 [30]. Except for Asp-15, which is protonated at this pH, most of the residues were in their standard protonated or unprotonated states. Asp-15 is located at the active center of the protein.

After adding the hydrogen atoms to the structure, a series of optimization algorithms (steepest descent conjugated gradient and L-BFGS-B) were applied [31]. To avoid denaturation and artificial configurations of the protein structure, all heavy atoms of the protein and the inhibitor were restrained by means of a Cartesian harmonic umbrella with a force constant of $1000 \text{ kJ mol}^{-1} \text{ \AA}^{-2}$. Afterward, the system was fully relaxed, but the peptidic backbone was restrained with a lower constant of $100 \text{ kJ mol}^{-1} \text{ \AA}^{-2}$. The optimized protein was placed in a cubic box of pre-equilibrated waters (80 \AA on a side) using the principal axis of the protein– Mg^{2+} complex as the geometrical center.

For the system, 5 ns of MD simulation were performed at temperature of 300 K. The time evolution of the root mean square deviation (RMSD) of the MD trajectory from the model was computed to analyze the stability of the structure. The simulation was performed using Langevin–Verlet molecular dynamics at 300 K and a canonical thermodynamic assembly (NVT), which fixes the number of atoms (N), the volume (V), and the temperature (T) of the system [32] using the Dynamo library [33]. The system was described using molecular mechanics (MM) with the OPLS-AA [32] and TIP3P [34] force fields for protein and water molecules.

QM/MM calculations

QM/MM potential energy optimization was carried out on the structure obtained after 5 ns of MD simulation, using the B3LYP function together with the 6-31+G(d,p) basis set to describe the quantum region in the hybrid QM/MM scheme. In these calculations, the Mg^{2+} , two water molecules and the side chains of Asp-16, Asp-59 and Met-61 were selected for treatment with QM, whereas the OPLS-AA force field was used for the MM part. In addition, the B3LYP functional within the 6-31+G(d,p) basis set was employed to perform QM/MM internal energy calculations in order to compute interactions between the Mg^{2+} and the protein.

Herein, we have used the potential energy derived from the standard QM/MM formulation (Eqs. 1 and 2) to determine the interaction energy between Mg^{2+} and the environment.

$$E_{\text{QM/MM}} = \langle \Psi | \hat{H}_0 | \Psi \rangle + \left(\sum \langle \Psi | \frac{q_{\text{MM}}}{r_{e,\text{MM}}} | \Psi \rangle + \sum \sum \frac{Z_{\text{QM}} q_{\text{MM}}}{r_{\text{QM,MM}}} \right) + E_{\text{QM/MM}}^{\text{vdW}} + E_{\text{MM}} \quad (1)$$

$$E_{\text{QM/MM}} = E_{\text{vac}} + E_{\text{QM/MM}}^{\text{elect}} + E_{\text{QM/MM}}^{\text{vdW}} + E_{\text{MM}} \quad (2)$$

Here, E_{MM} is the energy of the MM subsystem, $E_{\text{QM/MM}}^{\text{vdW}}$ is the van der Waals interaction energy between the QM and MM subsystems, E_{vac} is the gas phase energy of the polarized QM subsystem, and $E_{\text{QM/MM}}^{\text{elect}}$ includes both the coulombic interaction of the QM nuclei (Z_{QM}) and

the electrostatic interaction of the polarized electronic wavefunction (Ψ) with the charge on the protein (q_{MM}).

The interaction energy between Mg^{2+} and the environment, computed by residue, was evaluated as the difference between the QM/MM energy and the energies of the

separated, noninteracting QM and MM subsystems with the same geometry. Considering that the MM part is described using a nonpolarizable potential, the contribution of each residue (i) of the protein to the interaction energy is given by

$$E_{QM/MM,i}^{Int} = E_{QM/MM,i}^{elect} + E_{QM/MM,i}^{vdW} = \sum_{MM \in i} \left[\langle \Psi | \frac{q_{MM}}{r_{e,MM}} | \Psi \rangle + \sum_{QM} \frac{Z_{QM} q_{MM}}{r_{QM,MM}} \right] + \sum_{MM \in i} \sum_{QM} 4 \epsilon_{QM,MM} \left[\left(\frac{\sigma_{QM,MM}}{r_{QM,MM}} \right)^{12} - \left(\frac{\sigma_{QM,MM}}{r_{QM,MM}} \right)^6 \right] \quad (3)$$

In this calculation, the contribution of the interaction energy of each residue of the enzyme takes into consideration the effect of the polarization of the orbital, as the initial optimization of the molecular orbitals of the quantum portion is affected by the surrounding environment (protein and solvent); this is held constant in order to calculate the electrostatic interaction of each residue individually. More details can be found in the “Electronic supplementary material” (ESM).

Results and discussion

Figure 1 shows the best model obtained by the Modeller program; this model (PhoPCp) presents the largest number of residues within favorable regions (94.6%) of the

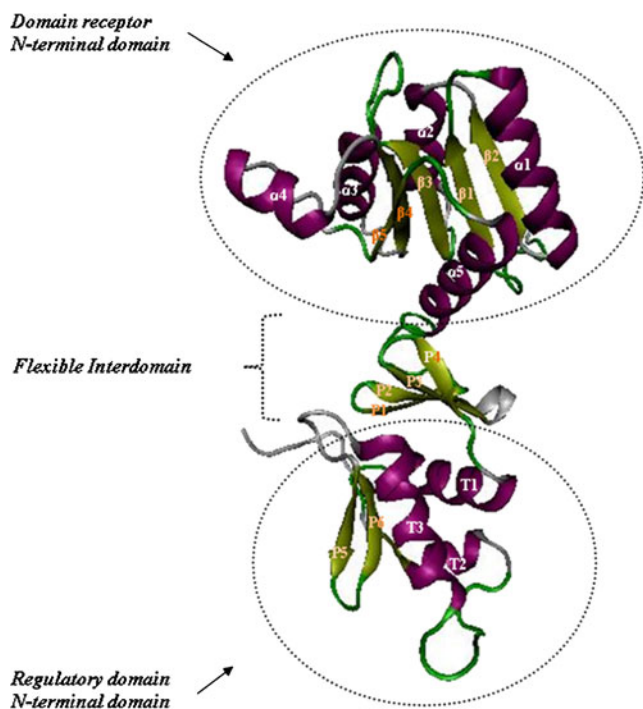


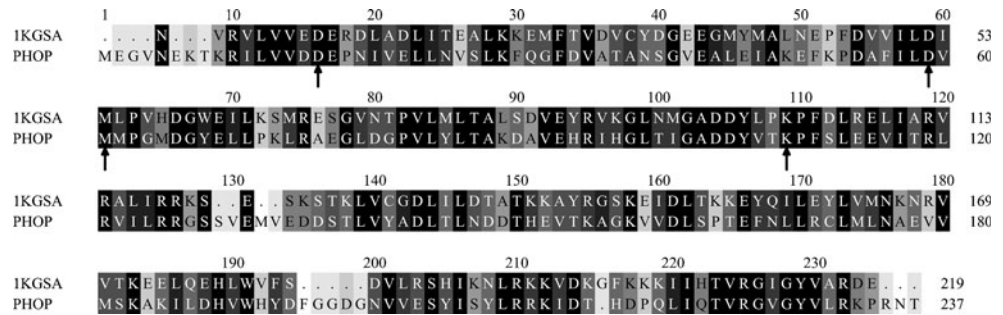
Fig. 1 PhoPCp model generated with the program Modeller through molecular modeling by homology

Ramachandran graph (ESM), which means that this model has the best stereochemical quality among those that were evaluated. The RMSD of only 0.53 obtained from the superposition of PhoPCp and the template signifies that the target and template structures are well aligned. The alignment between PhoP of *C. pseudotuberculosis* (PhoPCp) and 1KGS is shown in Fig. 2. The degree of superposition of the target with the template is shown in Fig. 3. Based on these results, we can conclude that the structures of the target and the template are conserved, as they show 39% identity and 60% similarity.

Furthermore, the quality of the structures obtained by homology modeling and MD were validated by calculating the ProSA Z-score [25] (ESM). If the score was outside the range characteristic of native proteins, the input structure contains errors. Verify3D [24] (ESM) was used to analyze the compatibility of an atomic model (3D) with its own amino acid sequence (1D). The Z-scores of the structures from homology modeling and MD were -8.18 and -7.88 , respectively, indicating that the structures are within the range normally found for proteins of a similar size. The PhoPCp model barely changed after MD, but the 3D structure in the region that binds the receptor domain and the flexible interdomain and the regulatory domain improved after MD. The atomic empirical mean force potential ANOLEA [26] was used to assess the packing quality of the model. This program performs energy calculations on a protein chain, evaluating the nonlocal environment of each heavy atom in the molecule. Negative energy values represent a favorable energy environment whereas positive values show an unfavorable energy environment for a given amino acid (ESM). According to ANOLEA, molecular dynamics considerably improved the model, as only a few amino acid residues of the α -4 helix and the loop that binds to RNA polymerase had unfavorable energy environments.

The aspartate receptor of the activation signal is located at the domain or N-terminal receptor (Asp-53 in 1KGS, Asp-59 in PhoPCp) and at the domain or C-terminal regulator, which are the sites that bind to the DNA,

Fig. 2 Alignment between the mold 1KGS and the sequence of the regulatory protein PhoP. Residues that are identical in terms of sequences are shown in *black* and catalytic residues are indicated by *arrows*



regulating gene transcription. The domain receptor comprises five α -helices (α 1, α 2, α 3, α 4, α 5) and five β -sheets (β 1, β 2, β 3, β 4, β 5); Asp59 is located in β -sheet β 3, as has been described for other homologous crystallographic structures deposited in the PDB (2PKX, 1MVO, 1IDO, 2OQR, 1KGS, 1YS6). The flexible interdomain consists of a group of four β -sheets (P1, P2, P3, P4), which correspond to a highly flexible region that connects both the domain and regulatory receptors; a small α -helix formed between β -sheets P1 and P2, comprising the amino acid residues Thr-150, Asp-149 and Asp-148. However, after MD, this region was corrected such that it formed a loop, as found in structures homologous to PhoP protein. Three α -helices (T1, T2, T3), and two β -sheets (P5, P6) are found below



Fig. 3 3D alignment of 1KGS and PhoPCp, as performed with the program Pymol to examine the structural divergence between the target (*purple*) and the mold (*light blue*)

this flexible region, located at the C-terminal region of the regulatory domain. The region consisting of the sequence P5-loop-P6 forms the β -hairpin motif. According to Ryndak et al. [9], this motif is important because it attaches to the minor groove of the DNA through an arginine residue in the β -hairpin loop. The equivalent arginine residue in this region in PhoPCp is Arg-225.

The loop that connects the α -helices T2 and T3 located in the regulatory domain is very important for this protein,

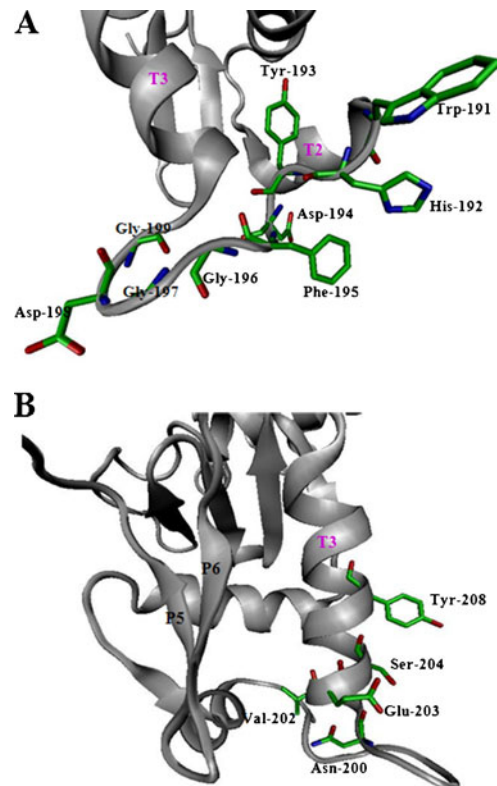
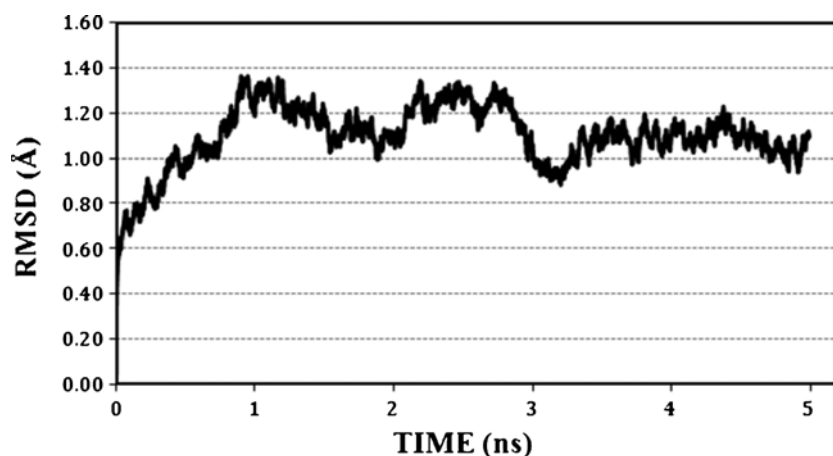


Fig. 4 a–b Conserved structures in the regulatory domain PhoP of *C. pseudotuberculosis*. **a** Residues that comprise the loop that connects the α -helices T3 and T4; this loop interacts with RNA polymerase. **b** α -Helix T4 (responsible for bonding to the minor groove of DNA), emphasizing the recognition residues of this helix (Asn-200, Val-202, Glu-203, Ser-204, Tyr-208)

Fig. 5 Graph of RMSD (in Å), showing the stability of the structure over 5 ns of MD simulation



as there is an indication in the literature that this loop interacts with RNA polymerase at the beginning of transcription [35]. In PhoPCp, this loop contains nine residues: Trp-191, His-192, Tyr-193, Asp-194, Phe-195, Gly-196, Gly-197, Asp-198 and Gly-199. As it is a loop region, there is a notable presence of glycine residues, which are highly flexible. On the other hand, residues Trp-191 and His-192 are basic, so they have the capacity to interact with positive charges (Fig. 4a).

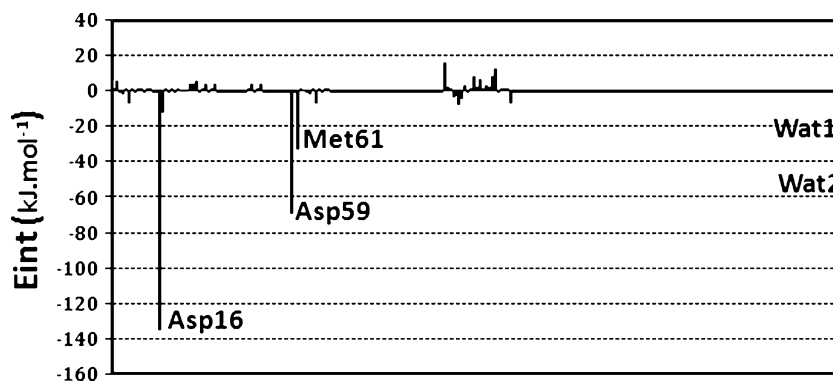
The α -helix T3 of PhoPCp corresponds to α -3 of 1KGS; both proteins are homologous to the linkage domain PhoPC of *M. tuberculosis* (access code 2PMU in PDB). This is a recognition helix for attachment to DNA; it allows attachment to the major groove of DNA during transcription regulation. In 2PMU, the recognition residues for interaction with DNA are Asn-212, Val-213, Glu-215, Ser-216 and Tyr-220 [36], which correspond to the conserved residues Asn-200, Val-202, Glu-203, Ser-204 and Tyr-208 in PhoPCp (Fig. 4b).

According to Solà et al. [37], transfer of the phosphate (signal transmitted by PhoR) occurs through autocatalysis and is necessarily dependent on Mg^{2+} . The structure of PhoPCp has a cavity that allows the entry of magnesium at the active site; this same cavity is found at the other two-component protein-receptor domain.

During the MD simulation, the RMSD drift of the $C\alpha$ atoms in the initial PhoPCp protein structure was determined (Fig. 5). When we examine the deviation of the protein during MD simulation, it is clear that PhoP has considerable stability. The PhoPCp structure reaches a plateau after 3 ns of simulation. The B factor (atomic displacement parameter) for an atom in the protein structure reflects the fluctuation of the atom about its average position. The distribution of B factors along a protein sequence is regarded as an important indicator of the protein's structure, reflecting its flexibility and dynamics [38]. The distribution of B factors was obtained during the last 2 ns of MD simulation. Analysis of B factors reveals that the interdomain, loop polymerase and β -hairpin are flexible regions of PhoPCp. These observations are in accordance with experimental results which suggest that the region that binds the receptor and regulatory domains (flexible interdomain) in an active state is intrinsically flexible [13]. The deviation of the residues near the magnesium ion was smaller (ESM), so we can conclude that the magnesium stabilizes the signal activation receptor cavity.

By calculating the interaction energy for each residue, it becomes possible to determine which residues of the active center of the domain receptor have the strongest interaction with Mg^{2+} (Fig. 6). The interaction energy values for each

Fig. 6 Interaction energy by residue (in kJ mol^{-1}) for the model PhoPCp with magnesium, after MD simulation. Values below zero indicate attraction and those above zero indicate repulsion



residue demonstrate attractive interactions of the residues Asp-16, Asp-59, and Met-61 with the magnesium. Two water molecules that interact with and stabilize the complex were also found. Through MD simulation, we found that the magnesium ion forms an octahedral structure, with the magnesium atom at its geometric center (Fig. 7). This structure is also observed in the homolog PhoB (PDB 2IYN) [37], thus indicating that this methodology correctly describes the system. For this reason, it was possible to find the correct structure at the active site. The residues Asp-16 ($-135.03 \text{ kJ mol}^{-1}$) and Asp-59 ($-69.15 \text{ kJ mol}^{-1}$) show the strongest interaction with magnesium. This suggests that these two residues are key residues to the catalytic enzyme of PhoP. Asp-59 is the activation signal receptor [13]; however, the function of Asp-16 is still unknown. It could be involved in the stability of the octahedral coordination mediated by magnesium. As Asp-16 and Asp-59 are key residues to the catalytic enzyme, the virulence of *C. pseudotuberculosis* can be affected by the mutation of Asp-59 or Asp-16 in PhoP. These results could be useful in the design of new vaccines based on genetic attenuation to inactivate the two-component regulatory system through a point mutation in the PhoP gene.

Ionic interactions play a critical role in biological systems, so a detailed analysis of the formation of the salt bridge between PhoPCp and Mg^{2+} was carried out for the MD trajectories. The mean distances (Table 1) between magnesium and its closest atoms were calculated in order to study the stability of the complex. The mean distances were calculated as a function of the different positions that the two atoms can have in relation to each other during the MD; for this reason, a deviation is associated with this deviation. The distances as a function of simulation time can be found in the ESM. Considering the mean distance values, it can be concluded that the two

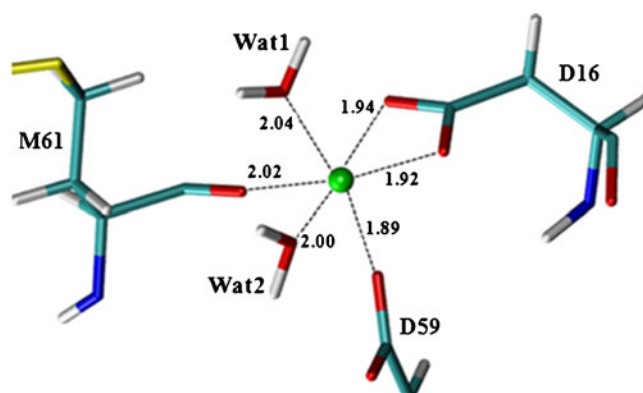


Fig. 7 Octahedral coordination exhibited by magnesium and its closest residues. The dotted lines represent the interactions and the numerical values indicate the mean distances (in Å) during the dynamics. Green magnesium, red oxygen, white hydrogen, cyan carbon, blue nitrogen, yellow sulfur

Table 1 Mean distances (Å) of the residuals for last 5 ns of MD

Residuals	Mean distance	Deviation
D16 and Mg (OD1...Mg)	1.94	0.05
D16 and Mg (OD2...Mg)	1.92	0.05
D59 and Mg (OD1...Mg)	1.89	0.04
D59 and Mg (OD2...Mg)	3.43	0.13
M61 and Mg (O...Mg)	2.02	0.07
Water 1 and Mg (OH ₂ ...Mg)	2.04	0.07
Water 2 and Mg (OH ₂ ...Mg)	2.00	0.06

oxygens (OD1 and OD2) of the side chain of Asp-16 are strongly linked to the Mg^{2+} ion at mean distances of 1.94 and 1.92 Å, respectively, and show very little variation during the MD simulation (ESM); this explains the strong interaction of this residue as observed in the graph of interaction energy by residue. On the other hand, the oxygen OD1 of the side chain of Asp-59 is bonded to magnesium at a mean distance of 1.89 Å (ESM).

It is well known that water molecules play important roles in biological systems, including in catalysis and in ligand binding. The X-ray structure of the PhoB in the metal ion complex [37] revealed that four interfacial water molecules were located close to Asp-16, Asp-59, Met-61 and Mg^{2+} . The mean distances between the waters and the Mg^{2+} ion as a function of simulation time were plotted (ESM). These water molecules, as well as the side-chain oxygen atoms of both aspartic residues and the oxygen atom of the methionine residue were found to coordinate octahedrally with the Mg^{2+} ion. The mean distance between the oxygen atom of the methionine residue and the Mg^{2+} ion is 2.02 Å (ESM).

The distances between the negatively charged group of Asp-59 and the ammonium group of Lys-109 were plotted

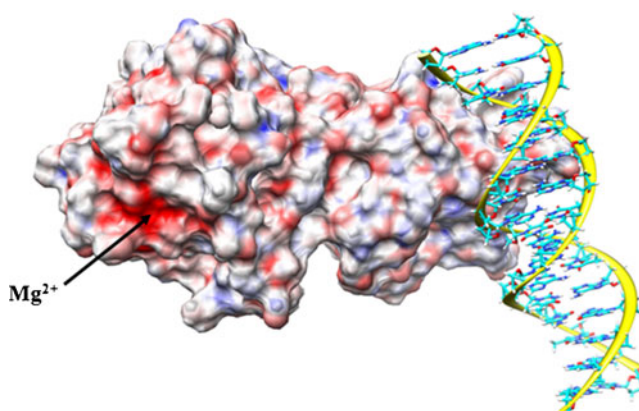


Fig. 8 Map of electrostatic potential for PhoPCp in contact with a double helix of DNA. Regions of high electron density are shown in red, and those with a lower electron density are shown in blue

as a function of simulation time (ESM). The average distances between the oxygen OD2 of the side chain of Asp-59 and the ammonium nitrogen (NZ) of Lys-109 is 2.69 Å, implying that this strong salt bridge was observed throughout the simulation period. In addition, the average distances between the Mg²⁺ ion and key residues were compared with the distances obtained from B3LYP/6-31+G(d)/MM optimization and experimental data [37] (ESM).

Electrostatic potential density maps can be used to pick out electrophilic and nucleophilic centers, which govern bond strengths, the strengths of nonbonded interactions, and molecular reactivity. Regions of higher electron density (nucleophilic regions) and those of lower electron density (electrophilic regions) correspond to positively and negatively charged regions in the molecule. Figure 8 shows a map of electrostatic potential for PhoPCp after 5 ns of MD simulation. In the more negative region, there are some residues at the active site that are important in interactions with Mg²⁺, such as Asp-16 and Asp-59, which coordinate with magnesium. On the other hand, the electrophilic character (the more positive region) is important in relation to the binding interaction with DNA in the process of gene regulation [36]. The regulatory domain corresponds to regions of lower electron density, leading to a good interaction of PhoP with DNA. In this region, there are some residues that are important for interactions with cofactor DNA, such as Asn-200, Val-202, Glu-203, Ser-204 and Tyr-208.

Conclusions

We have applied homology modeling, molecular dynamics and QM/MM techniques to determine the first structural information for a full-length member of the OmpR/PhoP subfamily of the two-component system. The MD results show that the magnesium of the PhoP protein stabilizes the receptor cavity, decreasing the deviations in residues that are important for protein activation. Based on energy calculations for each residue, the most important residues for interaction with magnesium were found to be Asp-16, Asp-59, Met-61 and two water molecules, which together form an octahedral structure. Exploration of the electrostatic potential density maps shows that Asp-16 and Asp-59 are in the more negative region, which is important in interactions with Mg²⁺. These results may be used to identify possible point mutations in the PhoP gene in order to design a new vaccine based on genetic attenuation that inactivates the two-component regulatory system. Attenuations can be made at residues Asp-59 (which is the activation signal receptor residue) and Asp-16 (which can overcome a lack of Asp-59 in the reception

of the activation signal, due to its strong interaction with the Mg²⁺ ion in the receptor cavity).

Acknowledgments The authors would like to thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazilian Agencies) for their financial support of this work.

References

- Dorella FA, Pacheco LG, Oliveira SC, Miyoshi A, Azevedo V (2006) *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. *J Vet Res* 37:1–18
- Comer FI, Hart GW (2000) *O*-glycosylation of nuclear and cytosolic proteins. Dynamic interplay between *O*-GlcNAc and *O*-phosphate. *J Biol Chem* 275:29179–29182
- Stoops SG, Renshaw HW, Thilsted JP (1984) Ovine caseous lymphadenitis: disease prevalence, lesion distribution, and thoracic manifestation in a population of mature culled sheep from Western United States. *J Vet Res* 45:557–561
- Alves FSF, Pinheiro RR, Pires PC (1997) Linfadenite caseosa: patogenia, diagnóstico e controle (Embrapa Caprinos Doc 27). EmbrapaCaprinos, Sobral
- Olson ME, Ceri H, Morck DW, Buret AG, Read RR (2002) Biofilm bacteria: formation and comparative susceptibility to antibiotics. *Can J Vet Res* 66:86–92
- Chaplin PJ, De Rose R, Boyle JS, Mcwaters P, Kelly J, Tennent JM, Lew AM, Scheerlinck JPY (1999) Targeting improves the efficacy of a DNA vaccine against *Corynebacterium pseudotuberculosis* in sheep. *Infect Immun* 67:6434–6438
- Songer JG (1997) Bacterial phospholipases and their role in virulence. *Trends Microbiol* 5:156–160
- Mckean SC, Davies JK, Moore RJ (2007) Expression of phospholipase D, the major virulence factor of *Corynebacterium pseudotuberculosis*, is regulated by multiple environmental factors and plays a role in macrophage death. *Microbiol* 153:2203–2211
- Ryndak Wang MS, Smith I (2008) PhoP, a key player in *Mycobacterium tuberculosis* virulence. *Trends Microbiol* 19:877–885
- Gonsalo-Asensio J, Mostowy S, Harders-Westerveen J, Huygen K, Hernández-Pando R, Thole J, Behr M, Gicquel B, Martin C (2008) PhoP: a missing piece in the intricate puzzle of *Mycobacterium tuberculosis* virulence. *PLoS One* 3:e3496
- Stock AM, Robinson VL, Goudreau PN (2000) Two-component signal transduction. *Annu Rev Biochem* 69:183–215
- Ninfa AJ, Magasanik B (1986) Covalent modification of the ginG product, NRI, by the glnL product, NRII, regulates the transcription of the glnALG operon in *Escherichia coli*. *Proc Natl Acad Sci USA* 83:5909–5913
- Buckler DR, Zhou Y, Stock AM (2002) Evidence of intradomain and interdomain in an OmpR/PhoP homolog from *Thermotoga maritima*. *Structure* 10:153–164
- Kurosu M, Begari E (2010) Bacterial protein kinase inhibitors. *Drug Dev Res* 71:168–187
- Hohmann EL, Oletta CA, Miller SI (1996) Evaluation of a phoP/phoQ deleted. *aroA* deleted live oral *S. typhi* vaccine strain in human volunteers. *Vaccine* 14:9–24
- Angelakopoulos H, Hohmann EL (2000) Pilot study of phoP/phoQ-deleted *Salmonella enterica* serovar typhimurium expressing *Helicobacter pylori* urease in adult volunteers. *Infect Immun* 68:2135–2141

17. Höltje HD, Sippl W, Rognan D, Folkers G (eds)(2003) Introduction to comparative protein modeling. In: Molecular modeling: basic principles and applications, 3rd edn. Wiley-VCH, Weinheim, pp 111–124
18. Dorella FA, Fachin MS, Billault A, Dias Neto E, Soravito C, Oliveira SC, Meyer R, Miyoshi A, Azevedo V (2006) Construction and partial characterization of a *Corynebacterium pseudotuberculosis* bacterial artificial chromosome library through genomic survey sequencing. *Genet Mol Res* 5:653–663
19. Finn R, Mistry J, Tate J, Coghill P, Heger A, Pollington J, Gavin OL, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy S, Bateman A (2010) The Pfam protein families database. *Nucleic Acids Res* 38:D211–D222
20. Hoch JA, Silhavy TJ (1995) Two-component signal transduction. ASM, Washington, DC
21. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234:779–815
22. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. *Nucl Acids Res* 28:235–242
23. The UniProt Consortium (2008) The Universal Protein Resource (UniProt). *Nucl Acids Res* 36:D190–D195
24. Eisenberg D, Luthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol* 277:396–404
25. 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
26. Melo F, Feytmans E (1998) Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol* 277:1141–1152
27. Alves CN, Martí S, Castillo R, Andrés J, Moliner V, Tuñón I, Silla E (2008) A quantum mechanic/molecular mechanic study of the wild-type and N155S mutant HIV-1 integrase complexed with diketo acid. *Biophys J* 94:2443–2451
28. Lameira J, Alves CN, Moliner V, Martí S, Castillo R, Tuñón I (2010) Quantum mechanical/molecular mechanical molecular dynamics simulation of wild-type and seven mutants of *CpNagJ* in complex with PUGNAc. *J Phys Chem B* 114:7029–7036
29. Nowak E, Panjikar S, Konarev P, Svergun D, Tucker P (2006) The structural basis of signal transduction for the response regulator PrrA from *Mycobacterium tuberculosis*. *J Biol Chem* 281:9659–9666
30. Delphine CB, David MR, Jan HJ (2008) Very fast prediction and rationalization of $p\mu_a$ values for protein–ligand complexes. *Proteins* 73:765–783
31. Byrd RH, Lu PH, Nocedal J, Zhu CY (1995) A limited memory algorithm for bound constrained optimization. *SIAM J Sci Comput* 16:1190–1208
32. 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
33. Field MJ, Albe M, Bret C, Martin FP, Thomas A (2000) The dynamo library for molecular simulations using hybrid quantum mechanical and molecular mechanical potentials. *J Comput Chem* 21:1088–1100
34. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) *J Chem Phys* 79:926–935
35. Martinez-Hackert E, Stock AM (1997) Structural relationships in the OmpR family of winged helix transcription factors. *J Mol Biol* 269:301–312
36. Wang S, Engohang-Ndong J, Smith I (2007) Structure of the DNA-binding domain of the response regulator PhoP from *Mycobacterium tuberculosis*. *Biochem* 46:14751–14761
37. Solà M, Drew DL, Blanco AG, Gomis-Ruth FX, Coll M (2006) The cofactor-induced pre-active conformation in PhoB. *Acta Cryst* 62:1046–1057
38. Yuan Z, Bailey TL, Teasdale RD (2005) Prediction of protein B-factor profiles. *Proteins* 58:905–912

Mechanism of intermolecular hydroacylation of vinylsilanes catalyzed by a rhodium(I) olefin complex: a DFT study

Qingxi Meng · Wei Shen · Ming Li

Received: 17 January 2011 / Accepted: 10 June 2011 / Published online: 29 June 2011
© Springer-Verlag 2011

Abstract Density functional theory (DFT) was used to investigate the Rh(I)-catalyzed intermolecular hydroacylation of vinylsilane with benzaldehyde. All intermediates and transition states were optimized completely at the B3LYP/6-31G(d,p) level (LANL2DZ(f) for Rh). Calculations indicated that Rh(I)-catalyzed intermolecular hydroacylation is exergonic, and the total free energy released is -110 kJ mol^{-1} . Rh(I)-catalyzed intermolecular hydroacylation mainly involves the active catalyst **CA2**, rhodium–alkene–benzaldehyde complex **M1**, rhodium–alkene–hydrogen–acyl complex **M2**, rhodium–alkyl–acyl complex **M3**, rhodium–alkyl–carbonyl–phenyl complex **M4**, rhodium–acyl–phenyl complex **M5**, and rhodium–ketone complex **M6**. The reaction pathway **CA2** + **R2** → **M1b** → **T1b** → **M2b** → **T2b1** → **M3b1** → **T4b** → **M4b** → **T5b** → **M5b** → **T6b** → **M6b** → **P2** is the most favorable among all reaction channels of Rh(I)-catalyzed intermolecular hydroacylation. The reductive elimination reaction is the rate-determining step for this pathway, and the dominant product predicted theoretically is the linear ketone, which is consistent with Brookhart's experiments. Solvation has a significant effect, and it greatly decreases the free energies of all species. The use of the ligand Cp' (Cp' = C₅Me₄CF₃) decreased the free energies in general, and in this case the rate-determining step was again the reductive elimination reaction.

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1151-z) contains supplementary material, which is available to authorized users.

Q. Meng · W. Shen · M. Li (✉)
College of Chemistry and Chemical Engineering,
Southwest University,
Chongqing 400715, People's Republic of China
e-mail: liming@swu.edu.cn

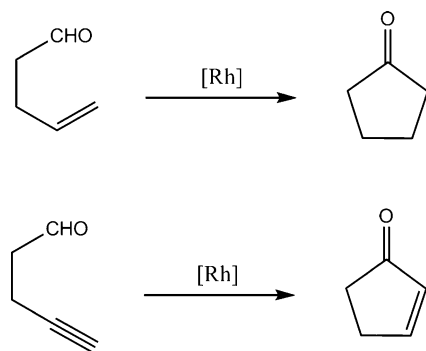
Keywords Rh(I)-catalyzed intermolecular hydroacylation · Vinylsilane · Benzaldehyde · Reaction mechanism · DFT

Introduction

Transition metal-catalyzed C–H bond activation has received considerable attention in synthetic organic chemistry, as the cleavage of an unreactive C–H bond and the subsequent addition of the C–H unit to an unsaturated substrate such as an alkene or alkyne can lead to the formation of a new C–C bond [1–7]. The formation of a C–C bond is one of the most fundamental aims of organic chemistry, so much effort has naturally been devoted to developing more convenient and efficient strategies for the formation of C–C bonds. During the last two decades, many successful applications of catalytic C–H bond activation with a view to creating new C–C bonds have been reported in synthetic communities [8]. C–C bond-forming reactions based on C–H bond activation have been a major focus of study in the fields of organic and organometallic chemistry [7–10].

Rhodium(I)-catalyzed intramolecular (Scheme 1) and intermolecular (Scheme 2) hydroacylation of alkenes or alkynes are two of the most useful C–H bond activation processes [11–24]. In these hydroacylation reactions, monophosphorus (e.g., PPh₃) or biphosphorus (e.g., BINAP, dppe, DUPHOS, etc.) ligands are applied. As shown in Scheme 3, the rhodium(I) bisolefin catalyst CpRh(vinylsilane)₂ can also be used [23], which does not contain phosphorus.

Morehead and Sargent [25] studied the mechanism of the rhodium-catalyzed intramolecular hydroacylation of alkenes theoretically using the software package DMol3

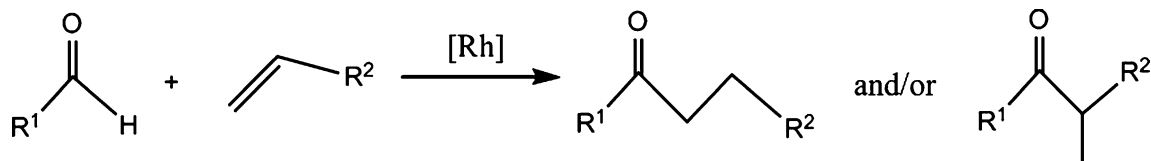


Scheme 1 Rhodium(I)-catalyzed intramolecular hydroacylation of alkenes or alkynes

and the two-layer ONIOM approach [(B3LYP/LANL2DZ:UFF method). The calculations supported the mechanism illustrated in Scheme 4, and indicated that reductive elimination is the rate-limiting step (Scheme 4, **D** → **A**) and that complex **C** is the less dominant route for certain solvent and substrate concentrations. Wu et al. [26] studied the rhodium-catalyzed intramolecular hydroacylation of 4-alkynals for a model system using MP2 calculations. The reaction mechanisms of 4-alkynals are similar to those presented in Scheme 4. They speculated that the activation of the aldehydic C–H bond (i.e., the oxidative addition of aldehydes) is the rate-determining step.

Brookhart [23] studied intermolecular hydroacylation catalyzed by rhodium(I) bisolefin complexes and suggested a likely mechanism (Scheme 5). He also suggested that reductive elimination was the turnover-limiting step (**J** → **F**), and showed that judicious functionalization of the cyclopentadienyl ligand with electron-withdrawing groups can promote the rate of reductive elimination. The utilization of vinylsilane in hydroacylation gave a yield of 85%, and only the linear ketone was obtained.

In order to understand the mechanism of intermolecular hydroacylation catalyzed by the Rh(I) olefin complex in detail, the CpRh(I)-catalyzed intermolecular hydroacylation of vinylsilane with benzaldehyde was studied in the present work. Specifically, the present study focused on: (1) the energetics of the overall catalytic pathways in intermolecular hydroacylation and the rearrangement processes involved; (2) the structural features of the intermediates and transition states involved; (3) the carbonylation versus decarbonylation reaction; (4) solvation and ligand effects in the reaction mechanism.



Scheme 2 Rhodium(I)-catalyzed intermolecular hydroacylation of alkenes

Computational details

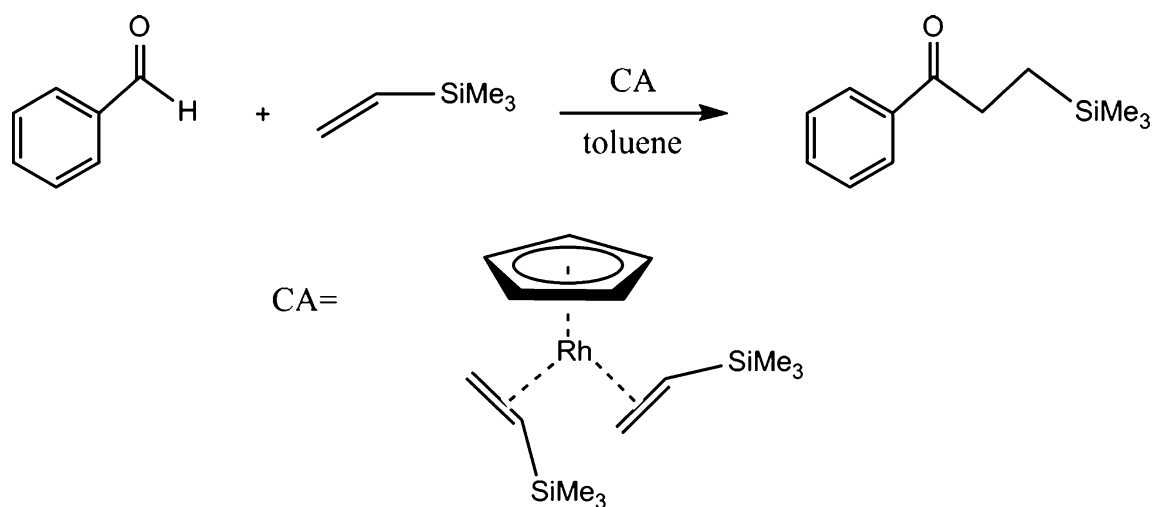
All calculations were carried out with the Gaussian 03 program package [27]. Density functional theory (DFT) methods [28] have been widely applied to various molecular systems with great success because of their efficiency and accuracy [29–31]. This is especially true of the B3LYP method [32, 33], which includes Becke's three-parameter-exchange functionals and the nonlocal Lee, Yang, and Parr correlation functional, as it generally provides better results. The basis set 6-31G(d, p) is used for C, O, Si, and H, and LANL2DZ is used for Rh, adding one set of *f* polarization functions with an exponent of 1.35 [34]. The transition states were verified by intrinsic reaction coordinate (IRC) [35] calculations and by animating the negative eigenvector coordinates with a visualization program (Molekel 4.3) [36, 37]. In addition, based on the gas phase optimized geometry for each species, the solvent effects of toluene were studied by applying a self-consistent reaction field (SCRF) of the polarizable continuum model (PCM) [38] approach at the same computational level.

Furthermore, the bonding characteristics were analyzed by using the “atoms in molecules” (AIM) theory [39], which is based on a topological analysis of the electron charge density and its Laplacian. The magnitude of the electron density, $\rho(r)$, at the bond critical points (BCPs) depends on the interatomic distance and the degree of coordination of the atoms, and it is often used as a measure of the bond strength or the similarity of bonds [40]. Further analysis was performed using the natural bond orbital (NBO) theory [41–44]. The AIM analysis was carried out with the AIM2000 code [45] using the B3LYP/6-31G(d,p) [LANL2DZ(f) for Rh] wavefunctions as input. The NBO analysis was performed by utilizing the NBO5.0 code [46] with the optimized structures.

Molecular orbital compositions and the overlap populations were calculated with the AOMix program [47, 48]. The analysis of the MO compositions in terms of occupied and unoccupied fragment molecular orbitals (OFOs and UFOs, respectively), the charge decomposition analysis (CDA), and the construction of orbital interaction diagrams were all performed using AOMix-CDA [49].

Results and discussion

The relative free energies $\Delta G_{(\text{sol})}$, including solvent energies, and the relative gas-phase free energies ΔG , enthalpies ΔH ,



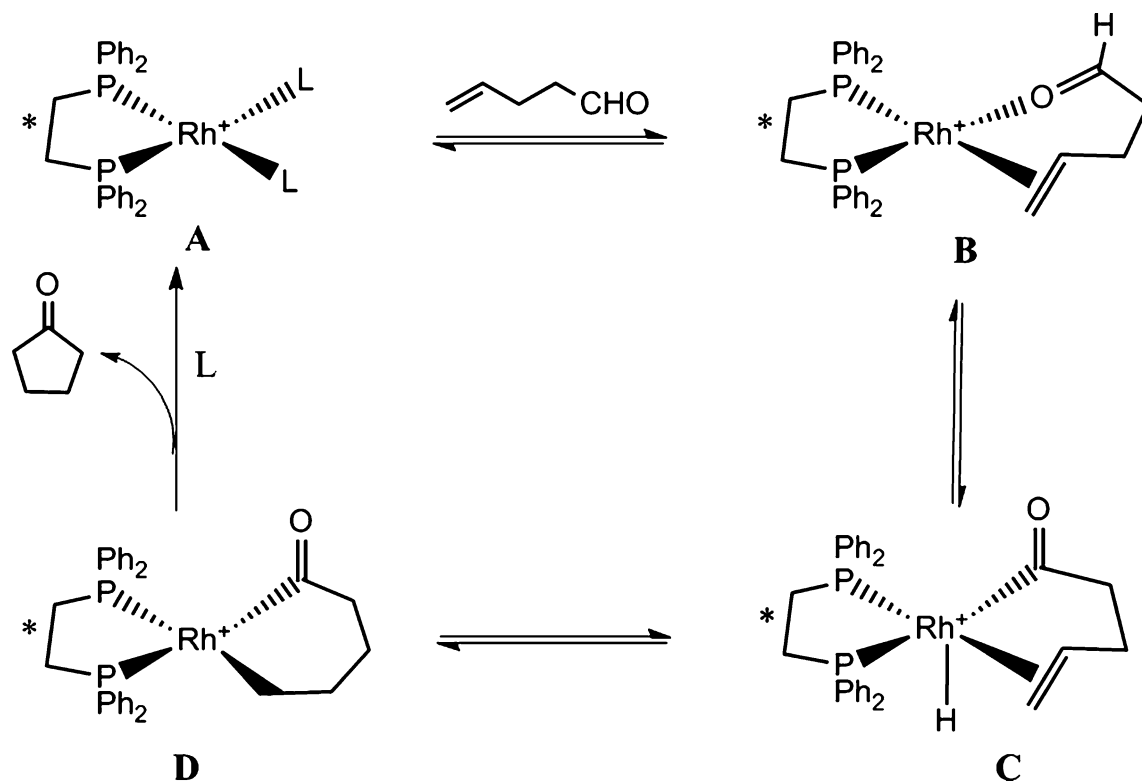
Scheme 3 CpRh(I)-catalyzed intermolecular hydroacylation of vinylsilane with benzaldehyde

and ZPE-corrected electronic energies ΔE are summarized in Table S1 and S2. Unless otherwise noted, the discussed energies are the relative free energies $\Delta G_{(\text{sol})}$ in the following discussions. The most favorable reaction channel predicted by our calculations is outlined in Scheme 6.

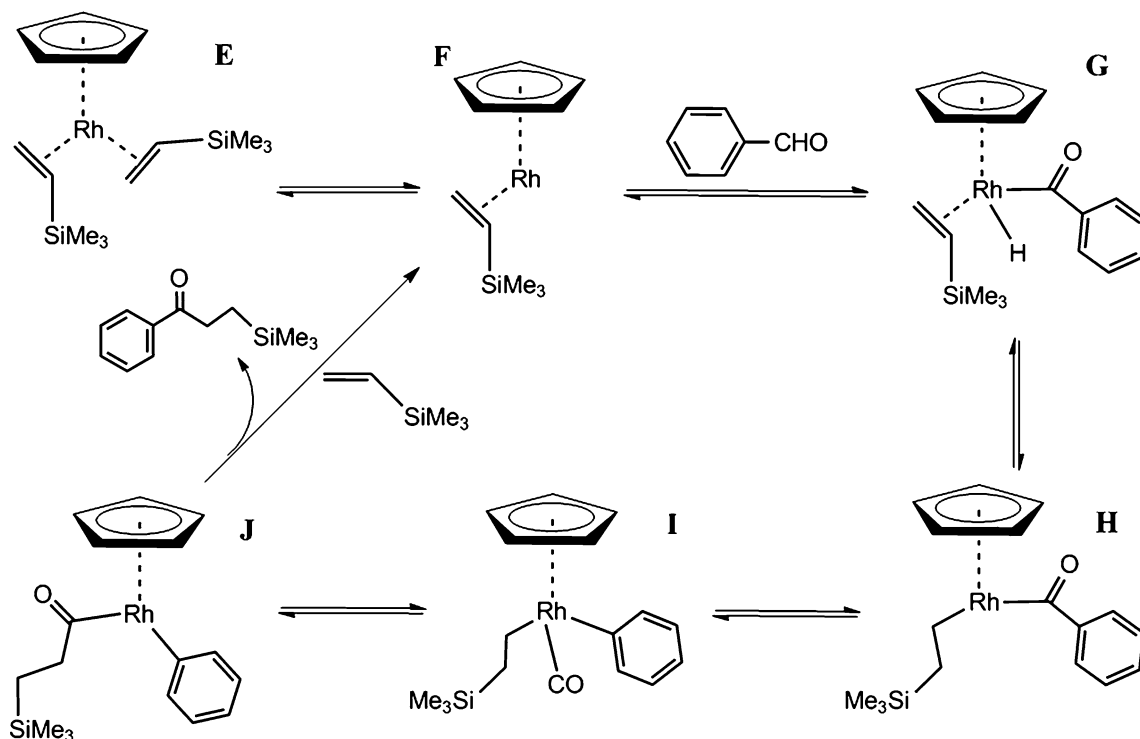
The coordination reaction of the catalysts and benzaldehyde

The optimized structures of two catalysts are shown in Fig. S1. The coordination reaction of CA2 with benzaldehyde

(R2) leads to three possible complexes: **M1a**, **M1b**, and **M1c** (Fig. S2). The first two are formed through π backdonation bond between rhodium and the C=O double bond of benzaldehyde, while the latter is formed by a coordinate bond between rhodium and the oxygen of benzaldehyde. The occupied $\pi_{\text{C3-O1}}$ orbital of benzaldehyde acts on the empty hybrid orbital of rhodium, leading to the σ coordinate bond; on the other hand, the occupied d orbital (d_{xy} , d_{xz} , d_{yz}) of rhodium acts on the empty $\pi^*_{\text{C3-O1}}$ orbital of benzaldehyde, leading to the π



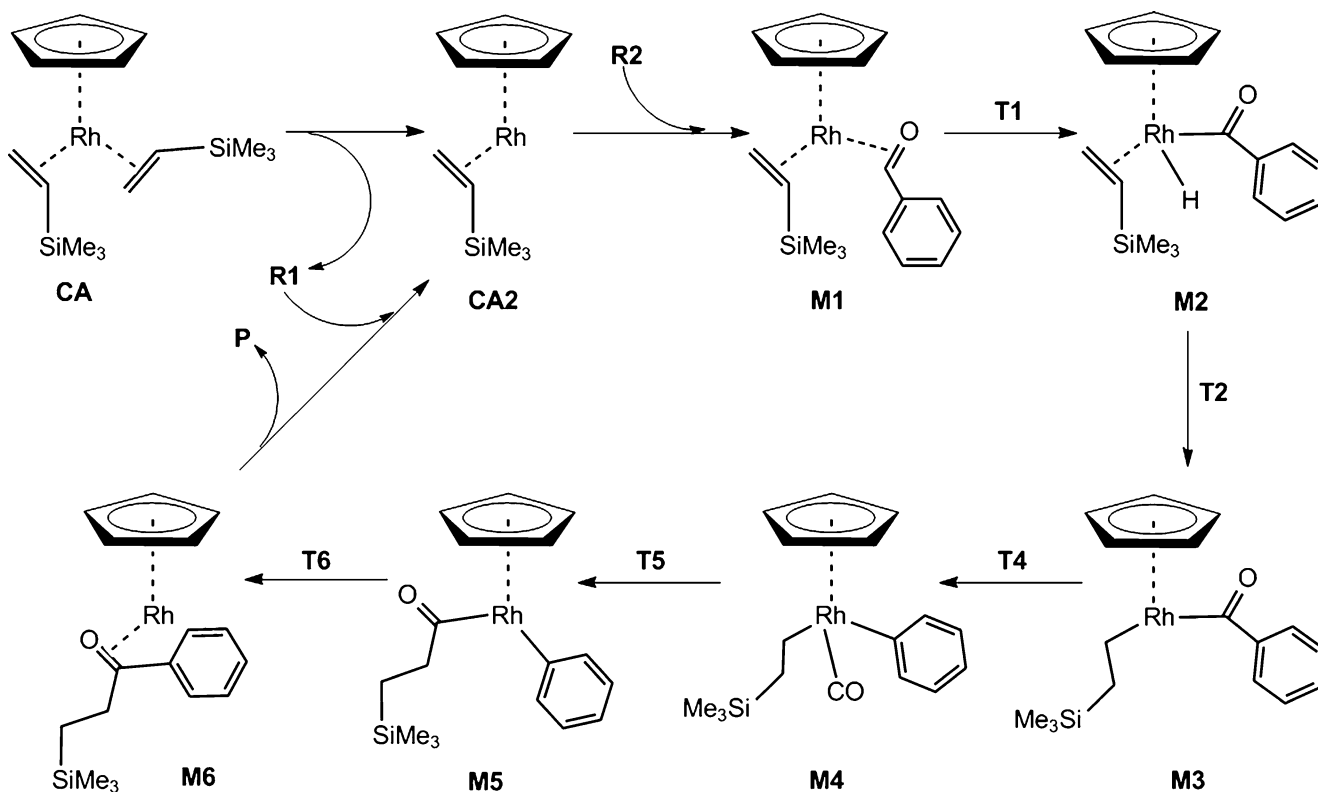
Scheme 4 Possible reaction mechanism of the rhodium(I)-catalyzed intramolecular hydroacylation of alkenes. (Those of alkynes are similar to alkenes)



Scheme 5 Possible reaction mechanism of the CpRh(I)-catalyzed intermolecular hydroacylation of vinylsilane with benzaldehyde

backdonation bond. Obviously, the formation of the π backdonation bond lowers the system's energy and makes **M1a** and **M1b** more stable.

Of the three complexes, **M1b** is the most stable, because the relative free energy of **M1b** is the lowest (Table S1), suggesting that it is the most likely to exist. In **M1a** and



Scheme 6 The most favorable reaction channel of CpRh(I)-catalyzed intermolecular hydroacylation of vinylsilane with benzaldehyde

M1b, the NBO energy of the C3–H1 bond is lower than that of benzaldehyde by 30 kJ mol⁻¹, while in **M1c**, the NBO energy of the C3–H1 bond is higher by 40 kJ mol⁻¹. Hence, in **M1a** and **M1b**, the formation of the π backdonation bond weakens and activates the C3–H1 bond, which results in the oxidative addition of benzaldehyde.

The formation of a branched ketone

Figure 1 shows the potential energy hypersurface for the pathway leading to the formation of the branched ketone **P1**. The C–H activation transition state **T1a** leading to the complex **M2a** is the rate-determining step and, with a free energy of 44.3 kJ mol⁻¹, the highest stationary point in the formation process of **P1**. Intermediate **M2a** then undergoes a *trans* addition to the alkene (intramolecular hydrometalation) through the transition state **T2a1** with a free energy of activation of 0.5 kJ mol⁻¹ to give the complex **M3a1**. Next, intermediate **M3a1** undergoes a carbonyl elimination reaction via the transition state **T4a** with a free energy of activation of 30.7 kJ mol⁻¹, resulting in the CpRh–alkyl–carbonyl–phenyl complex **M4a**. Then intermediate **M4a** undergoes a carbonyl insertion reaction through the transition state **T5a** with a free energy of activation of 13.4 kJ mol⁻¹ to generate the complex **M5a**. Finally, intermediate **M5a** undergoes a reductive elimination reaction via transition state **T6a** with a free energy of activation of 25.6 kJ mol⁻¹ to form Rh–ketone complex **M6a**, leading to the branched ketone **P1**.

In $\sigma_{(\text{Rh}-\text{H1})}$ bond formation, the distance $d_{(\text{C3}-\text{H1})}$ between C3 and H1 increases, $d_{(\text{Rh}-\text{H1})}$ decreases, and Rh shifts to C3. It is clear that a significant interaction between Rh and H1 occurs, and the C3–H1 bond is weakened considerably, as demonstrated by analyzing the changes in the bond orders P_{ij} and electron density ρ at the BCPs (Table S5; e.g., Rh–H1

bond, P_{ij} , **M1a**: 0.005 \rightarrow **T1a**: 0.234 \rightarrow **M2a**: 0.411; ρ , **M1**: 0.000 \rightarrow **T1a**: 0.104 \rightarrow **M2a**: 0.158 e \AA^{-3}). NBO analysis of **M2a** indicates that the Rh–H1 and Rh–C3 bonds show strong single-bond character, and the NBO energies of the bonding orbitals $\sigma_{\text{Rh}-\text{H1}}$ and $\sigma_{\text{Rh}-\text{C3}}$ are -742 and -1015 kJ mol⁻¹, respectively. In the intramolecular hydrometalation, because of the steric resistance of **M2a**, hydrogen migration has only one possible reaction pathway (a *trans* addition to an alkene). **T2a1** involves a Rh–H1–C1–C2 four-membered ring, and the electron density of the ring critical point (RCP) is 0.08 e \AA^{-3} . In **M3a1**, the Rh–C2 and Rh–C3 bonds are 2.109 and 1.919 \AA , respectively (Fig. S4). NBO analysis of **M3a1** indicates that the Rh–C2 and Rh–C3 bonds show strong single-bond character, and the NBO energies are -785 and -1054 kJ mol⁻¹, respectively. Obviously, the Rh–C3 bond is stronger than the Rh–C2 bond. Because C1 is sp^3 hybridized, the $\pi_{\text{C1}-\text{C2}}$ bond is broken, and then the π backdonation bond between rhodium and the $\pi_{\text{C1}-\text{C2}}$ bond is also broken. In **M4a**, the Rh–C3 and Rh–C4 bonds are 1.850 and 2.064 \AA , respectively (Fig. S6). The NBO energies of the Rh–C3 and Rh–C4 bonds are -1519 and -961 kJ mol⁻¹, respectively. Clearly, the NBO energy of the Rh–C3 bond is much lower than in **M3a1** (by 465 kJ mol⁻¹), due to the formation of the π backdonation bond between rhodium and the $\pi_{\text{C3}-\text{O1}}$ bond of the carbonyl. In **M5a**, the Rh–C3 bond is 1.927 \AA , which is longer than in **M4a** by 0.08 \AA . The NBO energy of the Rh–C3 bond is -1024 kJ mol⁻¹, which is higher than in **M4a** by 495 kJ mol⁻¹. The Rh–C3 bond is markedly weakened due to the disruption of the π backdonation bond between rhodium and the $\pi_{\text{C3}-\text{O1}}$ bond of carbonyl.

The formation of a linear ketone

Figure 2 shows the potential energy hypersurface for the pathway leading to the formation of the linear ketone **P2**.

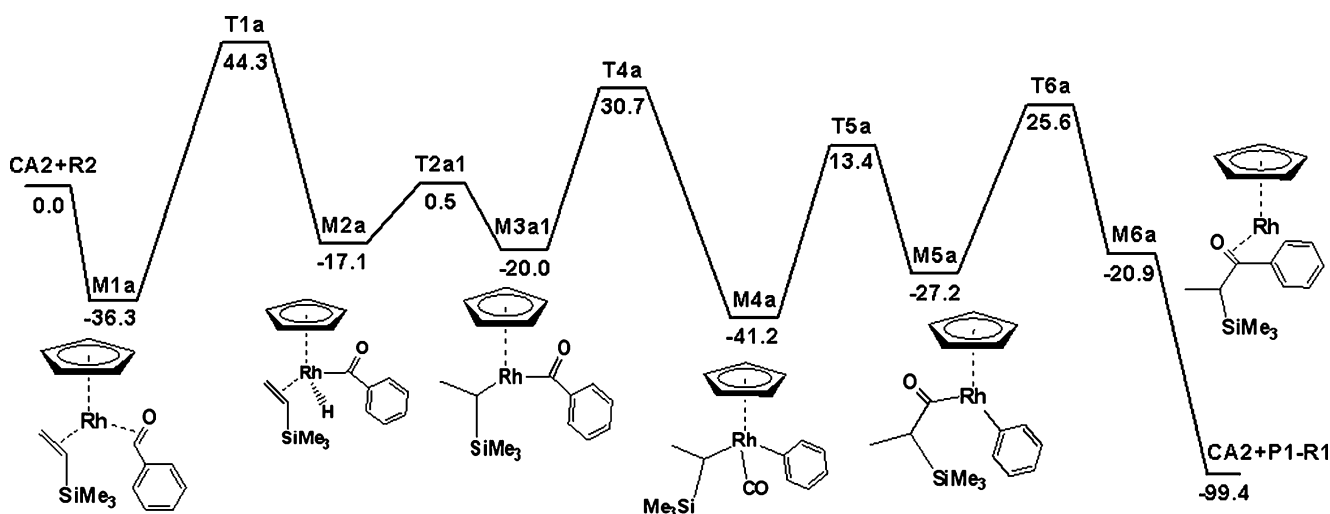


Fig. 1 Free-energy profile for the proposed formation pathway for branched ketone **P1**

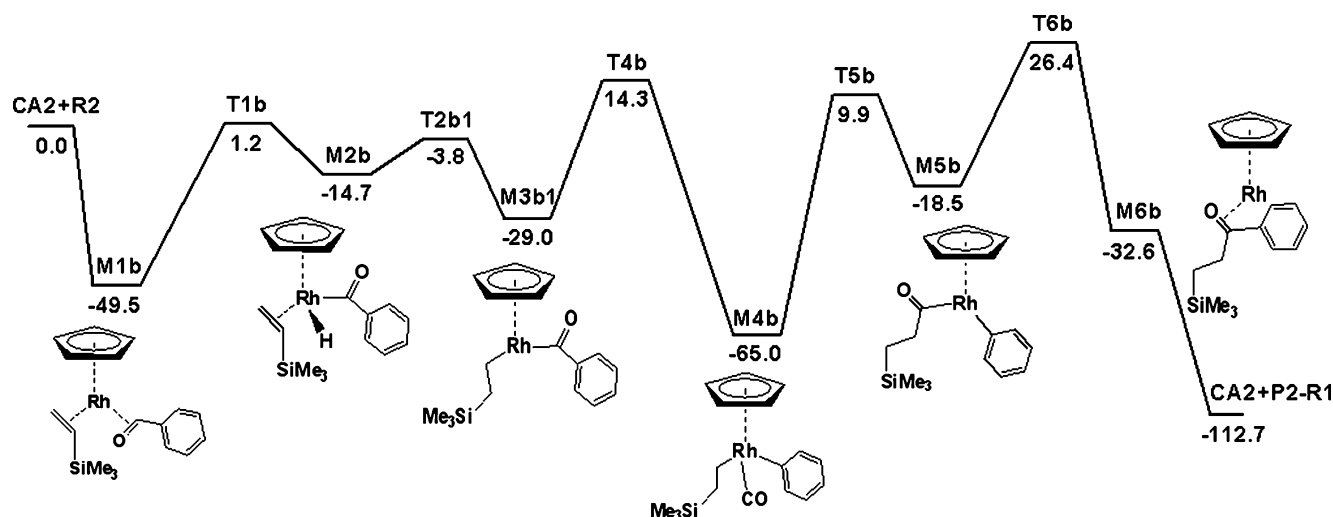


Fig. 2 Free-energy profile for the proposed formation pathway of the linear ketone **P2**

The oxidative addition reaction of benzaldehyde passes through the C–H activation transition state **T1b** with a free energy of 1.2 kJ mol^{-1} , leading to the complex **M2b**. Intermediate **M2b** then undergoes a *cis* addition to the alkene through the transition state **T2b1** with a free energy of activation of -3.8 kJ mol^{-1} to give the complex **M3b1**. Next, intermediate **M3b1** undergoes a carbonyl elimination reaction via the transition state **T4b** with a free energy of activation of 14.3 kJ mol^{-1} , leading to the complex **M4b**. Then intermediate **M4b** undergoes a carbonyl insertion reaction through the transition state **T5b** with a free energy of activation of 9.9 kJ mol^{-1} to generate the complex **M5b**. Finally, intermediate **M5b** undergoes a reductive elimination reaction via the transition state **T6b** with a free energy of activation of 26.4 kJ mol^{-1} to form Rh–ketone complex **M6b**, leading to the linear ketone **P2**. Obviously, the transition state **T6b** is the highest stationary point in the process of forming the linear ketone **P2**. Hence, the reductive elimination reaction is the rate-determining step for this pathway, which is different from that for the formation of the branched ketone **P1**. Comparing and contrasting Fig. 1 with Fig. 2, we find that the free energy of **T5b** is relatively close to that of **T5a**, but the free energy of **T1b** is lower than that of **T1a** by 43.1 kJ mol^{-1} . The steric and structural resistances between $-\text{SiMe}_3$ and $-\text{Ph}$ of **T1a** are much stronger than in **T1b**, and there are two hydrogen bonds, $\text{O}\cdots\text{H}-\text{C1}$ (2.244 \AA) and $\text{O}\cdots\text{H}-\text{CSi}$ (2.470 \AA), in **T1b**.

In the intramolecular hydrometallation, because of the steric resistance of **M2b**, there is only one possible pathway for hydrogen migration (*cis* addition to alkene). As illustrated in Fig. 3, the HOMO-1 for **T2b1** is a mixture of 6.6% HOFO-0 for vinylsilane (fragment 1) and 10.9% LUFO + 0 and 67.4% HOFO-1 for the rhodium–hydrogen fragment (fragment 2). The LUMO + 2 for **T2b1** is a mixture of 15.1% HOFO-0 for vinylsilane and 55.8%

LUFO + 0 and 11.4% LUFO + 4 for the rhodium–hydrogen fragment. It is clear that the reaction between vinylsilane and the rhodium–hydrogen fragment occurs dominantly between HOFO-0 of fragment 1 and HOFO-1, LUFO + 0, and LUFO + 4 of fragment 2. The net charge donation, which includes both charge donation and electronic polarization contributions, is 0.19 electrons. Obviously, this fact suggests that in the intramolecular hydrometallation, vinylsilane donates electrons to rhodium and hydrogen, which results in the formation of the Rh–C2 bond.

NBO analysis of **M3b1** indicates that the Rh–C1 and Rh–C3 bonds show strong single-bonded character, and the π backdonation bond between rhodium and the $\pi_{\text{C1}-\text{C2}}$ bond is also broken. Figure 4 shows that HOMO-2 for **T4b** is a mixture of 22.9% HOFO-0 for the phenyl (fragment 1) and 42.7% HOFO-1 for the rhodium–alkyl–carbonyl fragment (fragment 2); HOMO-0 is a mixture of 4.4% HOFO-2 for the phenyl and 11.2% HOFO-1, 36.8% HOFO-0, and 38.9% LUFO + 0 for the rhodium–alkyl–carbonyl fragment; LUMO + 0 is a mixture of 28.7% LUFO + 0 for the phenyl and 11.6% LUFO + 0 and 36.5% LUFO + 1 for the rhodium–alkyl–carbonyl fragment; LUMO + 3 is a mixture of 31.9% LUFO + 0 for the phenyl and 12.5% LUFO + 0 and 22.4% LUFO + 2 for the rhodium–alkyl–carbonyl fragment. It is clear that the reaction between the phenyl and the rhodium–alkyl–carbonyl fragment occurs dominantly between LUFO + 0, HOFO-0 of fragment 1 and HOFO-1, HOFO-0, LUFO + 0, LUFO + 1, and LUFO + 2 of fragment 2. The net charge donation, which includes both charge donation and electronic polarization contributions, is 1.07 electrons. Clearly, this fact suggests that in the carbonyl elimination reaction, phenyl donates electrons to carbonyl and rhodium, which results in the formation of the Rh–C4 bond and π backdonation bond between rhodium and carbonyl.

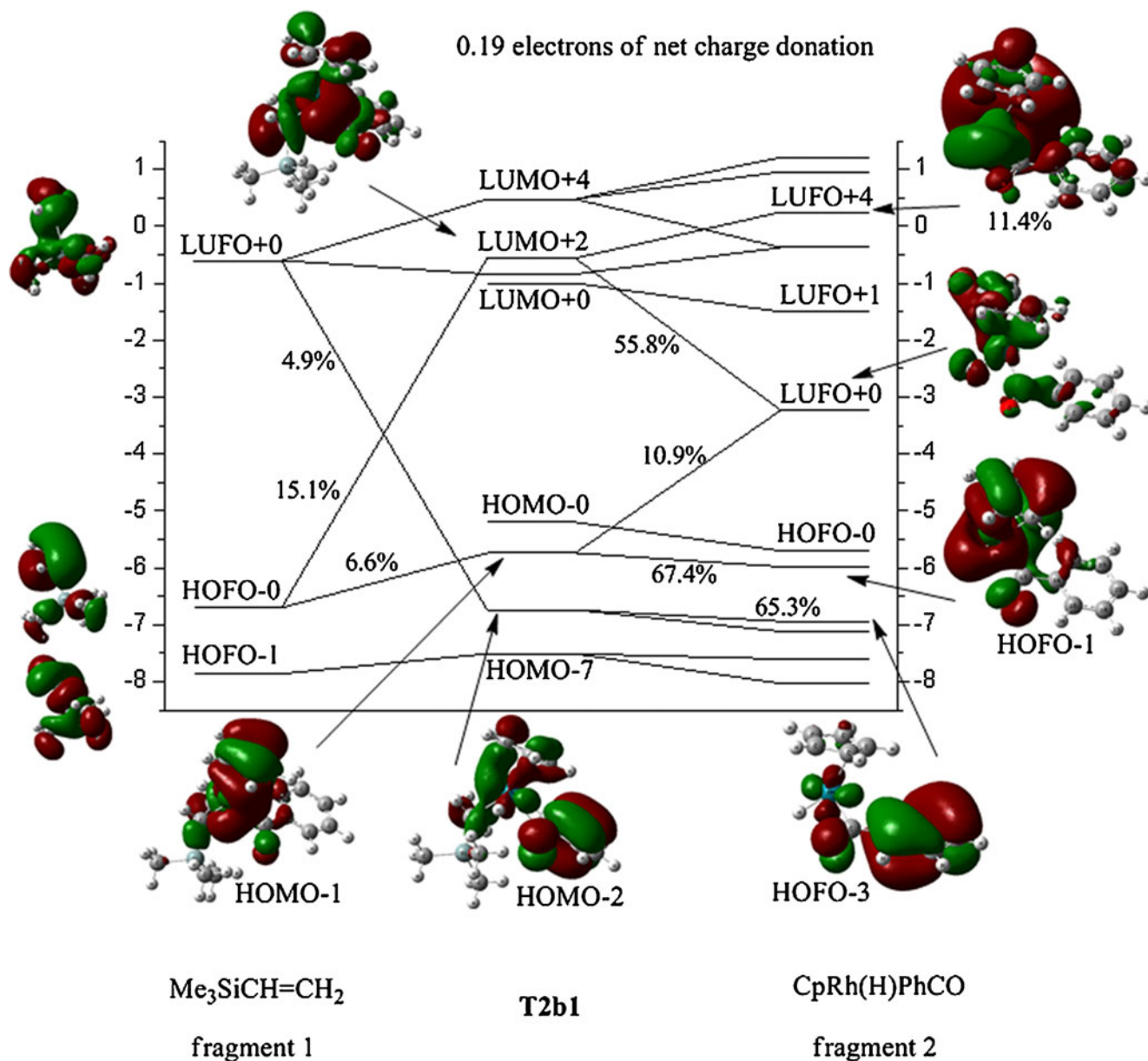


Fig. 3 Orbital interaction diagram for **T2b1**, which is formed from $\text{Me}_3\text{SiCH}=\text{CH}_2$ and $\text{CpRh}(\text{H})\text{PhCO}$ [the AOMix-CDA calculation, based on B3LYP/6-31G(d,p) results (LANL2DZ(f) for Rh)]. The net charge donation $\text{CT}(1 \rightarrow 2) - \text{CT}(2 \rightarrow 1)$ is 0.19 electrons)

The formation of an alkane: the decarbonylation reaction

Experimental studies [23] show that the decarbonylation reaction will generate an alkane in $\text{CpRh}(\text{I})$ -catalyzed intermolecular hydroacylation. Figure 5 shows the potential energy profiles for two pathways leading to the formation of the branched or the linear alkane. The optimized structures of the two transition states **T7a** and **T7b** are shown in Fig. S7. The organometallic product of the reductive decarbonylation have been identified as the dimer $[\text{Rh}(\text{C}_5\text{H}_5)(\text{CO})]_2$ of **CA3**, which ultimately results in catalyst death.

These results show that the branched alkane is formed from **M4a** via a decarbonylation. Although the formation of the $\text{CpRh}-\text{CO}$ -alkane complex **M7a** is exergonic by 85.1 kJ mol^{-1} , the barrier to the formation of the branched alkane is $118.8 \text{ kJ mol}^{-1}$ (prohibitively high). Contrasting with Fig. 1, the activation free energy of **T7a** is higher than that of **T1a** by 33.3 kJ mol^{-1} . Hence, the branched alkane will not be formed. The results also show that the barrier to the formation of the linear alkane is $114.9 \text{ kJ mol}^{-1}$ (again, prohibitively high). Comparing with Fig. 2, the free energy of activation of **T7b** is higher than that of **T6b** by 23.5 kJ mol^{-1} . It is therefore clear that the linear alkane will

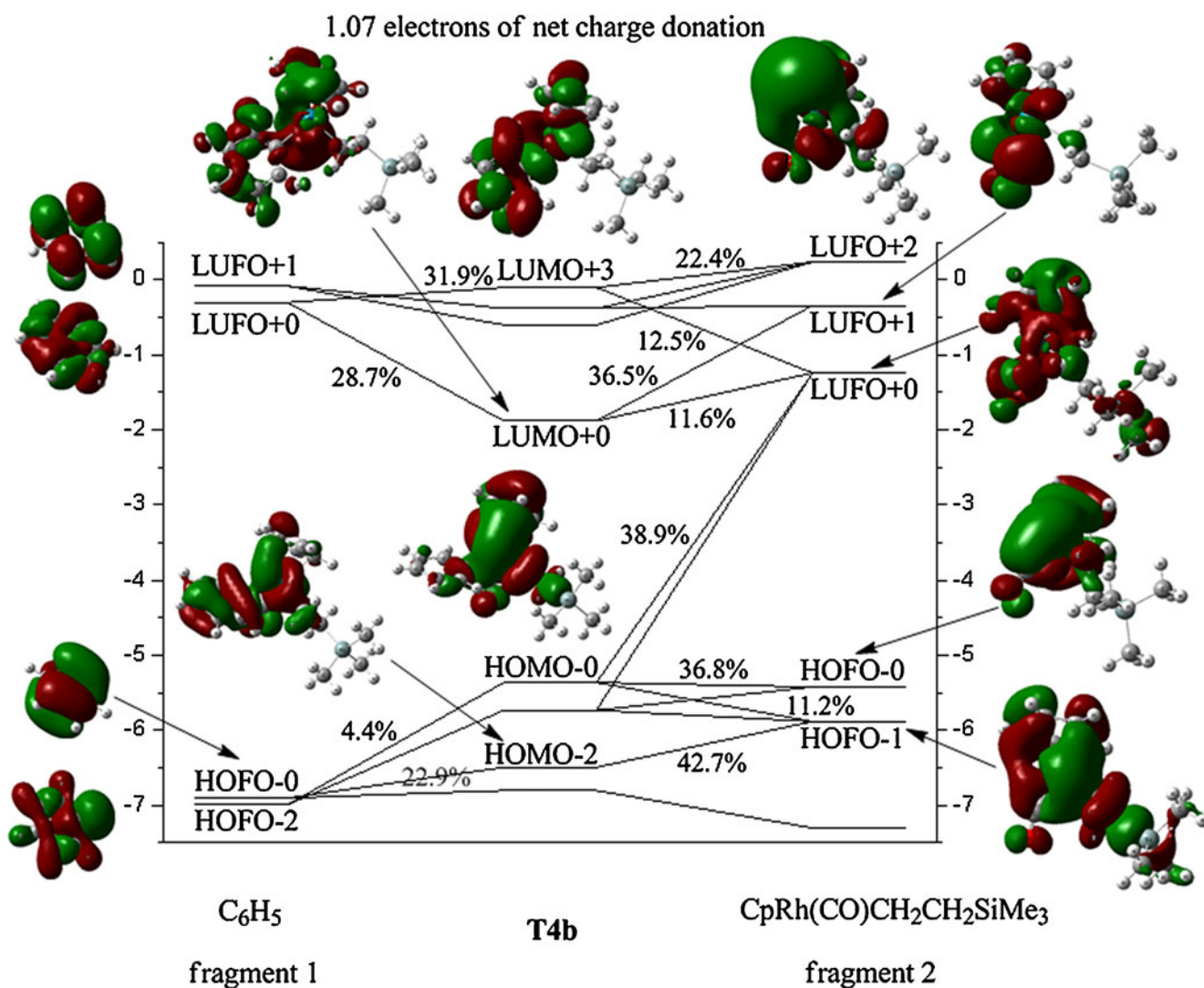


Fig. 4 Orbital interaction diagram for **T4b**, which is formed from phenyl and $CpRh(CO)CH_2CH_2SiMe_3$ [the AOMix-CDA calculation, based on B3LYP/6-31G(d,p) results (LANL2DZ(f) for Rh)]. The net charge donation $CT(1 \rightarrow 2) - CT(2 \rightarrow 1)$ is 1.07 electrons)

not be formed, because there is an energetically more favorable pathway for the formation of the linear ketone **P2**.

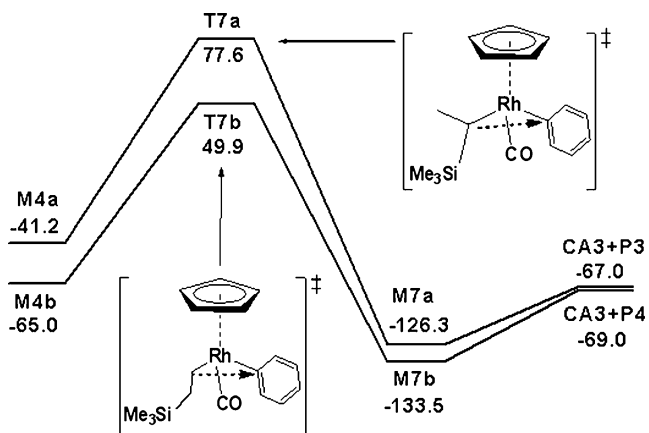


Fig. 5 Free-energy profile for the proposed formation pathways of the alkanes **P3** and **P4**: the decarbonylation reaction

Other possible reaction pathways

The Rh–ketone complexes **M6a** and **M6b** will form directly from **M3a1** and **M3b1** via a reductive elimination reaction. The corresponding transition states are, respectively, **T3a1** with 65.4 kJ mol^{-1} of free energy of activation and **T3b1** with 54.9 kJ mol^{-1} . Calculated results show that the free energy of activation of **T3a1** is higher than that of **T1a** by 21.1 kJ mol^{-1} , and that of **T3b1** is higher than that of **T6b** by 28.5 kJ mol^{-1} . It is therefore clear that the reductive elimination reaction from **M3a1** and **M3b1** via **T3a1** and **T3b1** will be difficult to achieve.

As shown in Fig. S3, **M2a** and **M2b** have three possible reaction pathways: in **M2a**, H1 attacking C1 is denoted “a1,” while C3 attacking C1 or C2 is denoted “a2” or “a3;” in **M2b**, H1 attacking C2 is denoted “b1,” and C3 attacking C1 or C2 is denoted “a2” or “a3.” The optimized structures are shown in Figs. S4 and S5. The reaction pathways “a1”

and “b1” have been discussed above. In intermediates **M3a2**, **M3a3**, **M3b2**, and **M3b3**, as well as transition states **T3a2**, **T3a3**, **T3b2**, and **T3b3**, there is one Rh–C2–C1–C3–O1 or Rh–C1–C2–C3–O1 five-membered ring, and the electron densities of the RCPs are 0.03 for **M3** and 0.02 $e \text{ \AA}^{-3}$ for **T3**. Several results due to the possible reaction channels in **M2a** and **M2b** can be summarized as follows. (1) The formation of intermediate **M3** is exergonic by 3~32 kJ mol^{-1} . (2) Hydrogen migration occurs prior to the C–C bond-forming reaction, because the free energies of activation of transition states of hydrogen migration are much lower than those of C–C bond-forming by 60~90 kJ mol^{-1} . (3) In the C–C bond-forming reaction, the carbonyl carbon atom finds it easier to attack the terminal carbon atom of the alkene.

Overview of the reaction mechanism

The dominant reaction pathways discussed above are outlined in Scheme 6. The Rh(I)-catalyzed intermolecular hydroacylation of vinylsilane with benzaldehyde mainly involves the rhodium–alkene–benzaldehyde complex **M1**, rhodium–alkene–hydrogen–acyl complex **M2**, rhodium–alkyl–acyl complex **M3**, rhodium–alkyl–carbonyl–phenyl complex **M4**, rhodium–acyl–phenyl complex **M5**, and rhodium–ketone complex **M6**. Calculated results indicate that the Rh(I)-catalyzed intermolecular hydroacylation is exergonic, and that the total released free energy is -110 kJ mol^{-1} .

In **M2a** and **M2b**, a hydrogen migration reaction occurs prior to the C–C bond-forming reaction. In **M3a1** and **M3b1**, the carbonyl elimination reaction is more dominant than the reductive elimination reaction. In **M4a** and **M4b**, because of the high free energies of activation, the decarbonylation reaction is prohibited, so an alkane will

not be formed. Therefore, among the reaction channels for the Rh(I)-catalyzed intermolecular hydroacylation of vinylsilane with benzaldehyde, the reaction pathway **CA2** + **R2** → **M1b** → **T1b** → **M2b** → **T2b1** → **M3b1** → **T4b** → **M4b** → **T5b** → **M5b** → **T6b** → **M6b** → **P2** is the most favorable. The reductive elimination reaction is the rate-determining step for this pathway, and the dominant product predicted theoretically is the linear ketone, which is consistent with Brookhart’s experiments [23].

The effect of the solvent

To evaluate the solvent effect for toluene ($\epsilon = 2.379$), single-point computations were performed at the B3LYP/6-31G(d,p) level [LANL2DZ(f) for Rh] using the PCM model with default parameters, except for the temperature (373.15 K is used) [23].

The schematic reaction profiles of the most favorable reaction pathways of Rh(I)-catalyzed intermolecular hydroacylation are shown in Fig. S8. Comparing and contrasting the blue line and red line in Fig. S8, we find that the solvation effect greatly decreases the free energies of all the intermediates and transition states. Clearly, the solvation effect is considerable. Furthermore, Fig. S8 also shows that the formation of the linear ketone **P2** is the energetically most favorable pathway (the blue line).

The effect of the ligand

Brookhart [23] show that the judicious functionalization of the cyclopentadienyl ligand with electron-withdrawing groups can promote the rate of reductive elimination. To understand the effect of the ligand, these calculations were performed on the stationary points of the most favorable reaction pathway when using Cp' (Cp' = $\text{C}_5\text{Me}_4\text{CF}_3$) instead

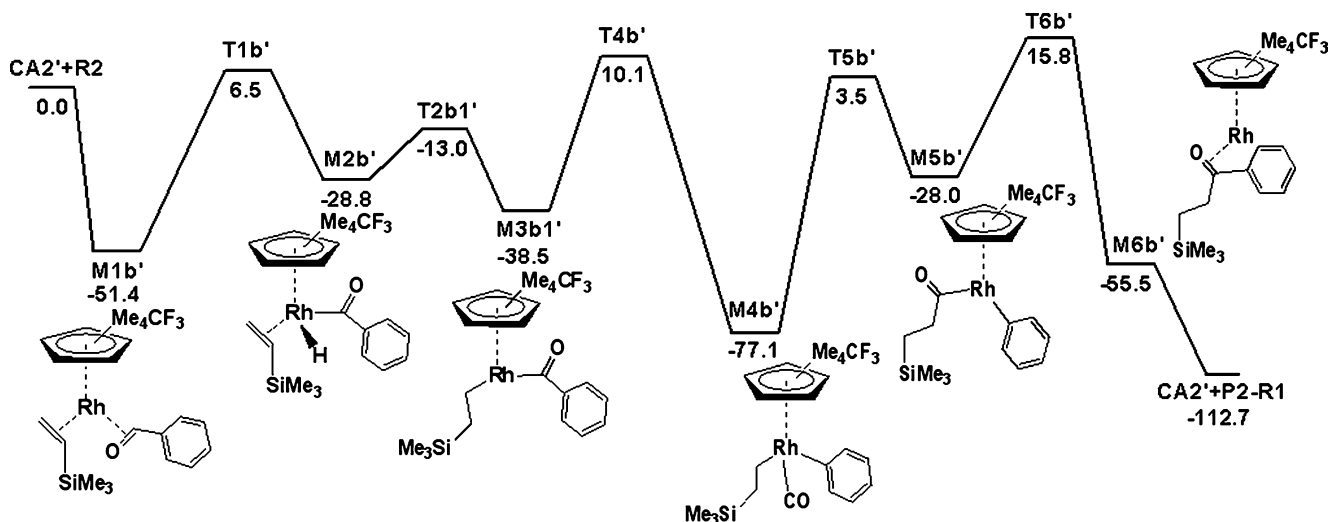


Fig. 6 Free-energy profile for the most favorable pathway when using Cp' (Cp' = $\text{C}_5\text{Me}_4\text{CF}_3$) instead of Cp

of Cp, and the resulting relative free energies $\Delta G_{(\text{sol})}$ including solvent energies are summarized in Table S6. The intermediates and transition states are shown in Fig. S9.

Comparing and contrasting Fig. 6 with Fig. 2, we find that the free energies generally decreased when using Cp' instead of Cp, and the free energy of **T6b'** is lower than that of **T6b** by 10.6 kJ mol⁻¹. The transition state **T6b'** is the highest stationary point in the process of forming the linear ketone **P2**. Hence, the reductive elimination reaction is the rate-determining step for this pathway, which is in good agreement with the results discussed above for Cp.

Comparison to our related computational work

We have studied the intermolecular hydroacylation of the ethene–aldehyde– or propylene–aldehyde–[Rh(PH₃)₂]⁺ model system at the B3LYP/6-311++G(d,p) [LANL2DZ(f) for Rh and P] level [50, 51]. Formic acid or formaldehyde were used instead of benzaldehyde.

We obtained some similar conclusions in our present work and in previous works. Firstly, Rh(I)-catalyzed hydroacylation is exergonic. Secondly, hydrogen migration reaction occurs prior to the C–C bond-forming reaction. Thirdly, the theoretically predicted dominant product is linear.

However, the reaction mechanism is different: the mechanism found in the previous work is similar to that shown in Scheme 4, and the oxidative addition of aldehyde is the rate-determining step for the most favorable pathways. Carbonyl elimination and insertion, decarbonylation, the solvation effect, and the ligand effect—which were all studied in the present work—were not studied in our previous works.

Conclusions

The reaction mechanism of the Rh(I)-catalyzed intermolecular hydroacylation of vinylsilane with benzaldehyde has been explored computationally using DFT [at the B3LYP/6-31G(d,p) level; LANL2DZ(f) for Rh]. Calculated results indicate that the Rh(I)-catalyzed intermolecular hydroacylation is exergonic, and that the total released free energy is -110 kJ mol⁻¹. In the intermediates **M2a** and **M2b**, a hydrogen migration reaction occurs prior to the C–C bond-forming reaction. In **M3a1** and **M3b1**, the carbonyl elimination reaction is dominant over the reductive elimination reaction. In **M4a** and **M4b**, the decarbonylation reaction is prohibited, so an alkane will not be formed.

The reaction pathway **CA2** + **R2** → **M1b** → **T1b** → **M2b** → **T2b1** → **M3b1** → **T4b** → **M4b** → **T5b** → **M5b** → **T6b** → **M6b** → **P2** is the most favorable among all of the reaction channels of Rh(I)-catalyzed intermolecular

hydroacylation. The reductive elimination reaction is the rate-determining step for this pathway, and the theoretically predicted dominant product is the linear ketone, which is consistent with Brookhart's experiments.

The solvation effect is considerable, and it greatly decreases the free energies of all of the species. It also indicates that the formation of the linear ketone **P2** is the energetically most favorable pathway.

The use of the ligand Cp' (Cp'=C₅Me₄CF₃) generally decreased the free energies of the complexes. In that case, the rate-determining step was the reductive elimination reaction, which is in good agreement with the results found for the ligand Cp.

Acknowledgments This work was supported by the Key Project of Science and Technology of the Ministry of Education, P.R. (grant no. 104263), Natural Science Foundation of Chongqing City, P.R. (grant no. CSTC-2004BA4024).

References

1. Jun CH, Lee JH (2004) Pure Appl Chem 76:577–587
2. Jia C, Kitamura T, Fujiwara Y (2001) Acc Chem Res 34:633–639
3. Kakiuchi F, Murai S (1999) Top Organomet Chem 3:47–79
4. Guari Y, Sabo-Etienne S, Chaudret B (1999) Eur J Inorg Chem 1047–1055
5. Arndtsen BA, Bergman RG, Mobley A, Peterson TH (1995) Acc Chem Res 28:154–162
6. Shilov AE, Shul'pin GB (1997) Chem Rev 97:2879–2932
7. Dyker G (1999) Angew Chem Int Ed 38:1698–1712
8. Ritleng V, Sirlin C, Pfeffer M (2002) Chem Rev 102:1731–1770
9. Labinger JA, Bercaw JE (2002) Nature 417:507–514
10. Kakiuchi F, Murai S (2002) Acc Chem Res 35:826–834
11. Inoue SI, Takaya H, Tani K, Otsuka S, Sato T, Noyori R (1990) J Am Chem Soc 112:4897–4905
12. Bergens SH, Bosnich B (1991) J Am Chem Soc 113:958–967
13. Barnhart RW, Wang XQ, Noheda P, Bergens SH, Whelan J, Bosnich B (1994) J Am Chem Soc 116:1821–1830
14. Jun CH, Hong JB, Lee DY (1999) Synlett 1–12
15. Aloise AD, Layton ME, Shair MD (2000) J Am Chem Soc 122:12610–12611
16. Jun CH, Chung JH, Lee DY, Loupy A, Chatti S (2001) Tetrahedron Lett 42:4803–4805
17. Tanaka K, Fu GC (2001) J Am Chem Soc 123:11492–11493
18. Willis MC, Sapmaz S (2001) Chem Commun 2558–2559
19. Jun CH, Moon CW, Lee DY (2002) Chem Eur J 8:2422–2428
20. Tanaka K, Fu GC (2003) J Am Chem Soc 125:8078–8079
21. Takeishi K, Sugishima K, Sasaki K, Tanaka K (2004) Chem Eur J 10:5681–5688
22. Jun CH, Jo EA, Park JW (2007) Eur J Org Chem 1869–1881
23. Roy AH, Lenges CP, Brookhart M (2007) J Am Chem Soc 129:2082–2093
24. Moxham GL, Randell-Sly H, Brayshaw SK, Weller AS, Willis MC (2008) Chem Eur J 14:8383–8397
25. Hyatt IFD, Anderson HK, Morehead AT, Sargent AL (2008) Organometallics 27:135–147
26. Chung LW, Wiest O, Wu YD (2008) J Org Chem 73:2649–2655
27. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven JT, 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 PW, 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.03. Gaussian, Inc., Pittsburgh
28. Parr RG, Yang W (1989) Density-functional theory of atoms and molecules. Oxford University Press, New York
29. Pisano L, Farriol M, Asensio X, Gallardo I, González-Lefont A, Lluch JM, Marquet J (2002) *J Am Chem Soc* 124:4708–4715
30. Pierini AB, Vera DMA (2003) *J Org Chem* 68:9191–9199
31. Pratt DA, Heer ML, Mulder P, Ingold KU (2001) *J Am Chem Soc* 123:5518–5526
32. Becke AD (1993) *J Chem Phys* 98:5648–5652
33. Lee C, Yang W, Parr RG (1988) *Phys Rev B* 37:785–789
34. Ehlers AW, Böhme M, Dapprich S, Gobbi A, Höllwarth A, Jonas V, Köhler KF, Stegmann R, Veldkamp A, Frenking G (1993) *Chem Phys Lett* 208:111–114
35. Gonzalez C, Schlegel HB (1990) *J Phys Chem* 94:5523–5527
36. Flükiger P, Lüthi HP, Portmann S, Weber J (2000–2002) MOLEKEL 4.3. Swiss Center for Scientific Computing, Manno
37. Portmann S, Lüthi HP (2000) *Chimia* 54:766–770
38. Miertus S, Tomasi J (1982) *Chem Phys* 65:239–245
39. Bader RFW (1990) *Atoms in molecules—a quantum theory* (Int Ser Monogr Chem vol 22). Oxford University Press, Oxford
40. Bader RFW, Popelier PLA, Keith TA (1994) *Angew Chem Int Ed Engl* 33:620–631
41. Carpenter JE, Weinhold F (1988) *J Mol Struct (THEOCHEM)* 169:41–50
42. Foster JP, Weinhold F (1980) *J Am Chem Soc* 102:7211–7218
43. Reed AE, Weinstock RB, Weinhold F (1985) *J Chem Phys* 83:735–746
44. Reed AE, Curtiss LA, Weinhold F (1988) *Chem Rev* 88:899–926
45. Biegler-König F, Schönbohm J, Derdau R, Bayles D, Bader RFW (2002) AIM 2000, version 2.0. McMaster University, Hamilton
46. Glendening ED, Badenhoop JK, Reed AE, Carpenter JE, Bohmann JA, Morales CM, Weinhold F (2001) NBO 5.0. Theoretical Chemistry Institute, University of Wisconsin, Madison
47. Gorelsky SI, Lever ABP (2001) *J Organomet Chem* 635:187–196
48. Gorelsky SI (1997) AOMix: program for molecular orbital analysis. York University, Toronto. <http://www.sg-chem.net/>
49. Gorelsky SI, Ghosh S, Solomon EI (2006) *J Am Chem Soc* 128:278–290
50. Gao JG, Wang F, Meng QX, Li M (2009) *Mol Simulat* 35:419–427
51. Wang F, Meng QX, Li M (2010) *Int J Quantum Chem* 110:850–859

In silico screening of epidermal growth factor receptor (EGFR) in the tyrosine kinase domain through a medicinal plant compound database

Orathai Sawatdichaikul · Supa Hannongbua ·
Chak Sangma · Peter Wolschann ·
Kiattawee Choowongkamon

Received: 26 October 2010 / Accepted: 23 May 2011 / Published online: 29 June 2011
© Springer-Verlag 2011

Abstract The unregulated epidermal growth factor receptor tyrosine kinase (ErbB1-TK or EGFR-TK) protein is involved in the proliferation of more than 50% of all cancer types. The reduction of EGFR-TK activity by small or medium-sized molecules has been proven to be an effective treatment for cancer. There is a widespread belief that Chinese medicinal herbs are active against several diseases, including various types of cancer. In this study, 29,960 compounds from the Chemiebase medicinal compound database were virtually screened against the EGFR-TK using AutoDock4.0, GOLD and GLIDE (XP). The results revealed eight potential hits: CAS nos. 104096-45-9, 112649-21-5, 113866-89-0, 142608-98-8, 142608-99-9, 144761-33-1, 155233-17-3 and 80510-05-0. These compounds have been reported to show anticancer activities in the literature. With the help of SiMMap and MOE interaction analysis, the protein–ligand interaction patterns between the functional groups of these compounds and the binding pocket residues were analyzed. Hydrogen bonding and hydrophobic forces are the main components of the

interactions of these hits, similar to those observed for the known inhibitors erlotinib, gefitinib and AEE. The physicochemical filter indicates that compounds CAS nos. 104096-45-9 and 144761-33-1 are likely to be potential leads in the drug discovery process.

Keywords Chemiebase database · EGFR · Tyrosine kinase · Virtual screening · Structure-based drug design

Introduction

Aberrant activity of the protein tyrosine kinases (PTKs) is one of the factors known to lead to cancer development. PTKs play pivotal roles in cell regulation, taking part in proliferation, differentiation, cell cycle progression, angiogenesis and the inhibition of apoptosis [1]. Thus, it is unsurprising that abnormal activity of the PTKs can lead to the development and maintenance of various cancers. Among the PTKs, the human epidermal receptor (HER) and epidermal growth factor receptor (EGFR, ErbB) family is one of the most widely studied. The overexpression, amplification and mutation of EGFR and others within the ErbB receptor family (ErbB2, ErbB3, and ErbB4) are implicated in a range of malignancies, and thus hold particular appeal as a molecular target for cancer drug discovery [2]. There are various reports that EGFR is a cause of mostly non-small cell lung cancer (NSCLC), while ErbB2 is implicated in breast cancer [3].

Inhibition of EGFR-TK has been proven to be an effective cancer therapy [4]. Such inhibitors of EGFR-TKI can be classified into chemical categories: (1) 4-anilinoquinazolines, (2) 4-(ar(alk)ylamino) pyridopyrimidines, (3) 4-phenylaminopyrrolo-pyrimidines [5], and (4)

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1135-z) contains supplementary material, which is available to authorized users.

O. Sawatdichaikul · K. Choowongkamon (✉)
Department of Biochemistry, Kasetsart University,
Bangkok, Thailand 10900
e-mail: fscikt@ku.ac.th

S. Hannongbua · C. Sangma
Department of Chemistry, Kasetsart University,
Bangkok, Thailand 10900

P. Wolschann
Institute for Theoretical Chemistry, University of Vienna,
Vienna, Austria 1090

oxime inhibitors [6, 7]. Examples of ErbB-TKIs (anticancer drugs) that are on the market or under investigation include erlotinib (Tarceva®) [8], lapatinib (Tykerb®) [9], gefitinib (Iressa®) [10, 11], vandetanib (Zactima®) [12], neratinib (HKI-272) [13], BIBW 2992 (Tovok®) [14–16],

pelitinib (EKB-569) [17], canertinib [18], AEE788 [19] and PKI166 [20], as well as the oxime inhibitors POX [7] and HYZ [6], as shown in Fig. 1.

Even though there are several effective anticancer drugs and active inhibitors against a number of protein targets,

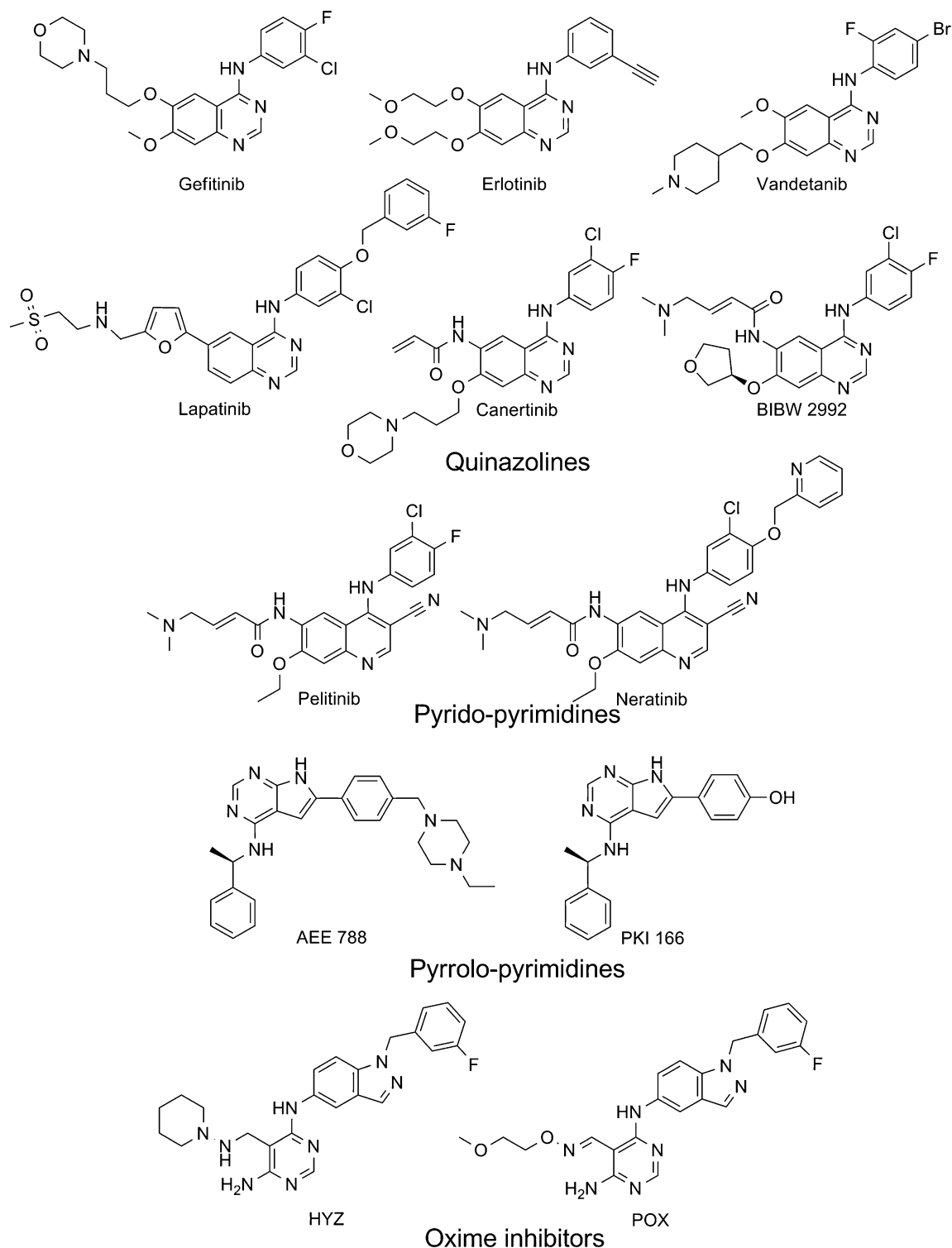


Fig. 1 EGFR-TK inhibitors classified into four chemical categories

increasing resistance coupled with many side effects mean that there is a need for new, improved treatments [21, 22]. Plants are the most important sources of the active ingredients used in modern medicines. More than half of the drugs approved since 1994 are based on natural products [23–25]. They are used in traditional medicines around the world, especially in Asia.

The availability of EGFR-TK structures [6, 7, 26–35] implies that virtual screening (VS) could be used as a tool to search for potential active compounds from medicinal herbs. The structures of the active compounds found in the folk-medicinal herbs have been collected in the Chemiebase database [36]. We developed a virtual screening protocol using three docking programs, AutoDock4 [37–39], GOLD [40, 41], and GLIDE [42]. They were applied to screen the Chemiebase database, which contains approximately 30,000 compounds. The resulting hits were analyzed to gain insights into the key structural features required for good protein–ligand interactions.

Materials and methods

Ligand preparation

The Chemiebase database contains a collection of Thai natural compounds ($N=29,960$) from 79 plant species. This database was obtained from the Cheminformatics Research Unit, Department of Chemistry, Faculty of Science, Kasetsart University. Further details are available at <http://chemiebase.ku.ac.th>. Two-dimensional (2D) structures were converted to 3D using CORINA [43], and optimized using semi-empirical PM3 calculations performed by the GAMESS package [44].

Virtual screening procedures

Filter

The 29,960 compounds in Chemiebase were filtered using the FILTER program (from Openeye Scientific Software) [45] and drug-likeness principles in order to eliminate unwanted compounds. The parameters in the FILTER program were modified to select the compounds that possess molecular weights of less than 500, $clogP$ values of between -5 and 5 , five or less hydrogen-bond donors, not more than ten hydrogen-bond acceptors, and fewer than ten rotational bonds [46].

AutoDock4

The atomic protein–ligand complex (PDB ID 1M17) was separated. Its geometry was optimized with the Tripos

forcefield in Sybyl 7.3 (Tripos Associates, St. Louis, MO, USA) [47]. Rigid receptor docking was performed using the AutoDock4 program. The rotational bonds of all proteins were regarded as being rigid, while the rotational bonds of the ligands were treated as flexible in a Python script (`prepare_ligand4.py`). The Python script for ligand preparation was embedded in the AutoDock program as script commands. The hydrogen atoms, the Kollman united-atom charges and the solvent parameter were applied to the proteins, and the set-up process for the grid was performed in AutoDockTools v.1.5.2. The grid boxes were fixed around the catalytic cleft (as shown in Fig. 2), and the dimensions of the box were set to 70, 60 and 60, while the grid point spacing was 0.375 Å. The grid affinity maps for the A (aromatic carbon), C, HD, N, NA (hydrogen-bond-accepting N), O, OA (hydrogen-bond-accepting O), S, SA (hydrogen-bond-accepting S), Cl, F, Br, I, P, and e (electrostatic) atom types were calculated using AutoGrid 4.0. The Lamarckian genetic algorithm search parameter was activated to simulate protein–ligand docking with 50 trial runs. The population size was set to 150. Docked conformations were clustered using a root mean square deviation (RMSD) tolerance of 2.0 Å by employing the clustering python script (`summarize_results4.py`). The structure of erlotinib (anticancer drug, complex ligand from 1M17) was also considered as a control. The docking results were sorted by the lowest binding energy (AutoDock4 score) as well as OEchemscore in the FRED program [48]. The intersection results from AutoDock4 and FRED revealed 993 selected compounds.

GOLD 4.0.1 kinase ChemScore (KCS)

All hydrogen atoms were added into the atomic coordinates of the protein structure using the *Protonation and Tautomers* function in the *GOLD Setup* window. The other configuration parameters were set to their defaults. The three-dimensional coordinates of the ligand, water and ion molecules were separated from the structure of the protein via the *Set Up and Run a Docking* wizard in the GOLD program. Automatic GA parameter setting was used in all GOLD docking calculations. One hundred percent search efficiency was applied, with between 10,000 and 125,000 GA operations per ligand. The binding site was defined to include all amino acid residues within a radius of 7 Å from the center of erlotinib; all water molecules were removed. The kinase scoring function, modified from the *ChemScore Fitness Function* (KSC), was applied in all docking calculations. The KCS is embedded in the GOLD package. This scoring function includes the contributions of the weak $\text{CH}\cdots\text{O}$ hydrogen bond, which are mostly found in kinase

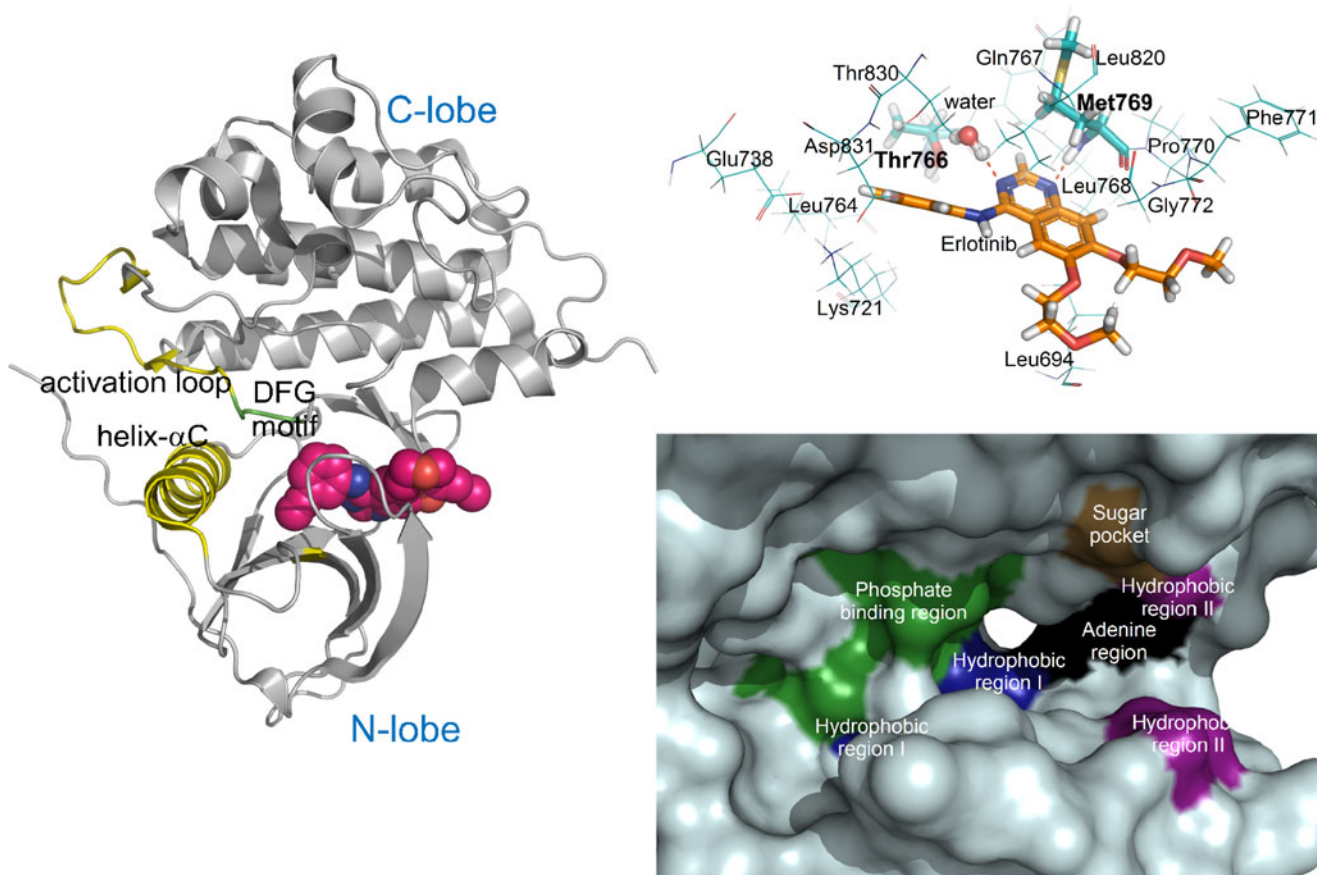


Fig. 2 The binding mode of erlotinib and the binding residue of IM17

proteins [49]. The functions *Allow Early Termination*, *Generate Diverse Solutions* and the internal ligand energy offset were also activated. The 527 compounds that presented KCS scores that were better than that of erlotinib were selected for further studies.

Glide extra precision mode (XP mode)

Protein Preparation Wizard Workflow implemented in Maestro 8.5 was used to prepare the protein using the default settings [50]. The 340 intersected structures from the GOLD (KCS) and AutoDock4 results were prepared further by LigPrep 2.2 [51]. Both protein and ligands were parameterized with the OPLS force field. The *Receptor Grid Generation* panel generated the grid map for the receptor. The center of the grid was located in the catalytic cleft. Docking calculations were performed in *Extra Precision (XP)* mode using the *Ligand Docking* panel. The *Receptor Grid Generation* and *Ligand Docking* panels are functions in the Glide (Grid-based Ligand Docking with Energetics) module. The XP mode combines a powerful sampling protocol with the use of an XP scoring function that is designed to specify only good ligand poses [52]. The compounds that had XP scores

better than that of erlotinib were screened as the candidate compounds.

Post-docking analysis

We used several tools to analyze the interactions between the EGFR and the hit compounds. The most important interactions of the protein and ligand complex from the docked results were calculated using the *Ligand Interactions* module embedded in the MOE package (Chemical Computing Group Inc.) [46]. The relationships between the functional sites (on the protein) and the compound moieties were extracted for the virtual screening compounds. The hit compounds were further analyzed with the SiMMap server [53] to identify binding modes with the target protein.

A new approach to predicting the absorption, distribution, metabolism, excretion and toxicity properties (ADMET or physicochemical properties) of drugs provides an effective tool for filtering weak compounds [54, 55]. This tool allows us to remove molecules with poor pharmacokinetic properties, so it helps to save limited resources. The selected molecules were applied in further studies and MD calculations. All figures were produced using Pymol v.0.99 [56]

Molecular dynamics simulations

The molecular dynamics (MD) simulations were carried out to check whether the hit compounds remain bound in the ATP-binding pocket of EGFR-TK. Three complexes, EGFR-TK with gefitinib (WT-IRE), EGFR-TK with compound A (WT-CPA) and EGFR-TK with compound F (WT-CPF), were constructed using the crystal structure of EGFR-TK (PDB code: 1M17) as the protein. The conformation of each ligand inside EGFR-TK was obtained from the docking results. The protein was capped at the N- and C-termini by ACE and NME respectively in order to avoid the tail charge problem when using the Sybyl 7.3 program [47]. All of the classical simulations were performed with the AMBER 03 force field [57] and the AMBER 10 package. The general AMBER force field (GAFF) parameters [58] and AM1-BCC charges were assigned using the Antechamber module. Each protein ligand complex was immersed in an isomeric truncated-octahedron box of TIP3P water molecules [59] (12 Å from the solute surface) and neutralized by the addition of Cl⁻ anions. All hydrogen atoms were added by using the TLeap module of the AMBER package. The MD strategy consisted of a three-step energy minimization of the solvated complexes. First of all, the hydrogen atoms were optimized using the steepest descent method with 3500 steps, and then the conjugate gradient method was employed for another 3500 steps while other atoms were kept frozen. Secondly, water molecules were minimized using the same protocol as used in the first step, keeping the protein and ligand frozen. Then, a third stage of minimization—using the same approach as in the two steps above—was performed with all atoms relaxed. Periodic boundary conditions (PBC) were also applied. Complexes were heated for 200 ps up to 300 K. They were then maintained in the isothermal–isobaric ensemble (NPT) at the target temperature and target pressure at 1 bar using a Langevin thermostat and Berendsen barostat, respectively. After the equilibration stage, MD simulations were carried out for 18 ns. The time steps in the equilibration and production runs were set to 2 fs, and bonds involving hydrogen atoms were constrained with the SHAKE algorithm [60]. MD simulation analyses were mainly performed with the PTRAJ module of the AMBER10 package.

MM-PBSA calculations

The energy calculations were carried using the MM-PBSA Perl script embedded in AMBER10. The atomic coordinates of EGFR-TK and each ligand were extracted from a single trajectory of the molecular dynamics simulation with explicit water molecules for equilibration. Each energy term

of MM-PBSA was evaluated for 2 ns of the trajectory, from 16 ns to 18 ns, within 250 snapshots.

The binding free energy ΔG_{BIND} was estimated as

$$\Delta G_{\text{MM/PBSA}} = \Delta G_{\text{complex}} - (\Delta G_{\text{protein}} + \Delta G_{\text{ligand}}) \quad (1)$$

$$\Delta G_{\text{MM/PBSA}} = \Delta E_{\text{GAS}} + \Delta\Delta G_{\text{SOLV}} - T\Delta S \quad (2)$$

$$\Delta E_{\text{GAS}} = \Delta E_{\text{ELE}} + \Delta E_{\text{VDW}} + \Delta E_{\text{INT}} \quad (3)$$

$$\Delta\Delta G_{\text{SOLV}} = \Delta\Delta G_{\text{SA}} + \Delta\Delta G_{\text{PB}} \quad (4)$$

ΔE_{GAS} is the interaction energy between the EGFR-TK and ligand(s) in the gas phase, as shown in Eq. 3, while ΔE_{ELE} , ΔE_{VDW} and ΔE_{INT} represent the receptor–ligand electrostatic and van der Waals interactions, and the internal energy, respectively. The solvation free energy ($\Delta\Delta G_{\text{SOLV}}$) consists of two parts: the polar/electrostatic solvation free energy ($\Delta\Delta G_{\text{PB}}$) and the nonpolar/hydrophobic solvation free energy ($\Delta\Delta G_{\text{SA}}$). All energies are averaged along the MD trajectories.

Results and discussion

Validation of the docking methodology

There are two conformations for the X-ray crystal structures of EGFR-TK: the active and inactive conformations [26]. The focus of this study has been on virtual screening Thai natural product compounds against the active conformation of EGFR-TK because most inhibitors have been found to bind with the active form of EGFR-TK [26, 28, 30]. Therefore, the active conformation of EGFR, 1M17 [26], was chosen as the target protein structure for virtual screening. In our previous study, we developed a VS protocol for the kinase domain of EGFR using AutoDock3.0.5 and GOLD 4.0.1 [61] in order to screen the NCI diversity database. In this study, the newer AutoDock4 program and FRED (OEchemscore) for rescoring, as well as GOLD 4.0.1 with three scores—Goldscore (GS), ChemScore (CS) and KCS—were used. Furthermore, the Glide program used in standard precision (SP) and XP as well as high-throughput virtual screening (HTVS) modes was applied. The docking approaches were validated with three known inhibitors from crystal structure complexes; erlotinib, gefitinib and AEE788 (PDB code: 1M17, 2ITO and 2J6M) [26, 30]. The atomic coordinates of three inhibitors were retrieved after the structures were superimposed. These were used as the standard orientation to calculate the RMSDs for the docked conformations. The binding

energies in Table 1 show that the experimental values correlate well with the computational values obtained from AutoDock and GOLD (KCS). Even though they were not as well correlated with the results from GLIDE docking, the overall binding energies calculated from docking programs were still in the same range. In the tests presented in Table 1, Glide XP gave the lowest docking RMSD, which indicates the high quality of this docking method. Although its scoring ability was not as good as AutoDock and GOLD, which may be due to our rather small validation test, we consider Glide XP to be the most generally reliable scoring function, based on previous extensive enrichment studies performed on the EGFR receptor [62].

All 29,960 compounds from the Thai herbal compound database were filtered using Lipinski's rule of five in the FILTER program [45]. The 15,758 compounds that passed through the filter process were prepared and docked into the pocket site of the EGFR-TK structure using combination methods including AutoDock4 and GOLD with KCS, similar to those used in our previous publication [61]. The docked results from AutoDock4 and GOLD (KCS) revealed 993 and 527 compounds, respectively, that were better scored than known inhibitors. A total of 304 consensus compounds were identified by both docking methods according to the flowchart (Fig. 3). We then applied the Glide docking method in the XP mode to these hits. Finally, eight unique compounds that gave XP scores that were better than the reference inhibitors were selected (Tables 2 and 3).

The eight hit compounds show anticancer activities

The eight hit compounds are found in plants, and they have already been reported to possess anticancer activities. **A** was first specified by Loder and group in 1957 [63]. This compound was found in a plant from the genus *Diospyros*, commonly known as the ebony tree (Table 2), and has been reported to show anticancer activity [64]. The derivatives of **B** were reported to possess anticancer activity in relation to several types of cancer, such as hepatocellular carcinoma [65], human leukemia cell line HL60 [66], and oral human epidermoid carcinoma (KB) [67]. **C**, which is found in custard apple and makrut lime, has been reported to inhibit 5 α -reductase enzyme activity, which is involved in prostate cancer [68]. The small molecules exiguaflavone **A** and exiguaflavone **B** (**D** and **E**) are found in *Sophora exigua* (or legume) [69] and *Artemisia indica* Willd (or "kho-hia," its local name) [70, 71]. The analogs of curcumin (**F**; CAS no. 144761-33-1), which is found in wild turmeric plants, show cytotoxicity towards lymphocytes and promising tumor-reducing activity towards Dalton's lymphoma ascites tumor cells [72]. **G** (lespedezaflavanon **B**) is an antimutagenic constituent of neem (*Azadirachta indica*) [73]. Although

Table 1 The validated docking methods from known inhibitors

Compound (CAS no.)	Experimental data		AutoDock 4		GOLD (KCS)			Glide (XP)			
	K_i or K_d values in nM	ΔG^* in kcal/mol	ΔG^* in kcal/mol	AutoDock 4 score in kcal/mol	OEchemscore in kcal/mol	RMSD in Å (core RMSD)	KCS score fitness value	ΔG in kcal/mol	RMSD in Å (core RMSD)	Glide score in kcal/mol	RMSD in Å (core RMSD)
Gefitinib (184475-35-2)	35.3 ^a	-10.17	-10.17	-9.18	-7.64	5.38 (2.49)	30.05	-7.41	3.67 (0.81)	-8.70	3.98 (0.84)
Erlotinib (183321-74-6)	17.5 ^b	-10.58	-10.58	-7.70	-7.86	1.86 (0.91)	28.65	-7.42	1.74 (0.34)	-8.32	1.54 (0.31)
AEE788 (497839-62-0)	5.3 ^a	-11.29	-11.29	-9.91	-9.23	2.05	32.55	-8.14	1.37 (0.68)	-8.32	1.21 (0.57)

^a K_d values (nM) from [32]

^b K_i value (nM) from [76]

* ΔG_{exp} were calculated from the thermodynamic equation $\Delta G_b = -RT \ln K_{\text{binding}}$, $\Delta G_i = RT \ln K_i$, where $K_i = K_d$

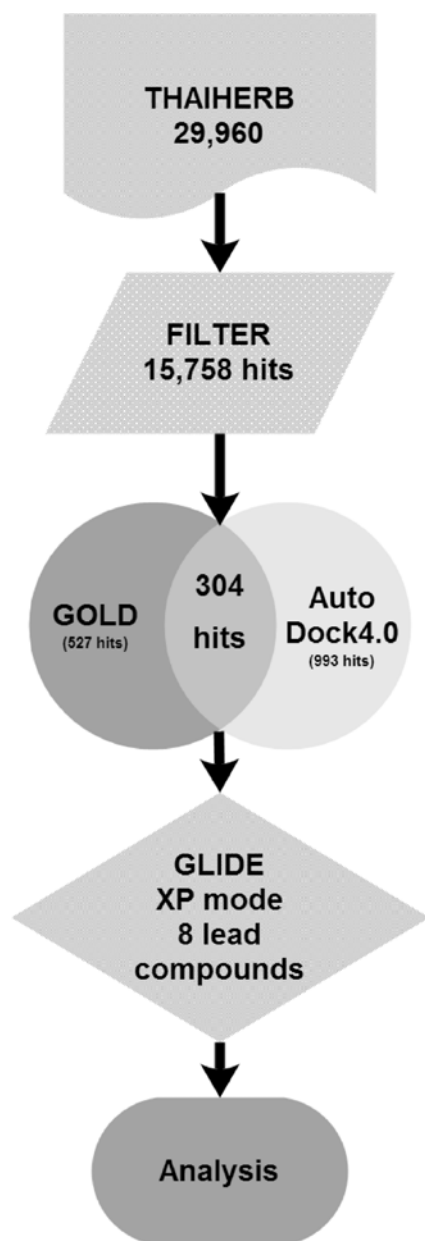


Fig. 3 Schematic diagram of the computational virtual screening workflow

the inhibitory mechanisms of these compounds are unknown, it is possible that this mechanism may be related to the EGFR pathway.

The atomic interactions of the eight candidates with the EGFR kinase domain

The interactions of the eight candidates with the kinase domain are shown in the two- and three-dimensional diagrams (2D and 3D) in Fig. 4. These reveal hydrogen-bond interactions and cation– π interactions between pocket residue(s) and ligand(s). These interactions were also seen for known inhibitors. A

strong hydrogen bond is consistently formed with Met769. The quinazoline rings of the candidates also show lipophilic stacking interactions with the hydrophobic residues in hydrophobic pocket region I: Val702, Lys721, Met742, Leu764 and Thr766 [74]. The interaction patterns of the eight compounds were analyzed based on the MOE–ligand interaction [46] and a site-moiety map that was statistically derived using several anchors by the SiMMap server [53]. The main interactions of the docked hit compounds with EGFR-TK were a combination of hydrogen bonding and hydrophobic interactions with the residues around the catalytic cleft. The polar moieties of seven of the eight compounds formed strong hydrogen bonds with the residue Met769 (Fig. 4a–f, h), which was also noted in the SiMMap analysis; see anchor H1 (Fig. 5). This hydrogen bonding was also seen in all three reference inhibitors (erlotinib, gefitinib and AEE), indicating that this is the most important hydrogen bonding associated with the TK inhibitors.

In addition, Lys721 plays a significant role, because it interacts with seven of the eight candidates in various ways: (1) through hydrogen bonding (cation H-bond; Fig. 4a–b, g–h), (2) through a weak π –cation interaction between the N ζ -atom of Lys721 and the benzene ring of the compound (Fig. 4d–f), and (3) through a hydrophobic interaction between the side chain of Lys721 and the aromatic moiety of the hit compound (Fig. 4c). The side chain of this residue also participates in a hydrophobic interaction with the aniline rings (aromatic moiety) of known inhibitors (erlotinib and gefitinib as well as AEE788). This interaction was also found in the SiMMap analysis; Lys721 represents the van der Waals (VDW) anchor (V2), which forms a hydrophobic interaction with the aromatic moiety, as presented in Fig. 5. Furthermore, the residues Asp831, Pro770, Thr766 and Arg817 also formed hydrogen bonds with the hit compound, which were similar to the hydrogen bonds that the reference inhibitors formed with Thr766 and Thr830 through water molecules.

These eight hits can be classified into five groups: xanthenes (B, G), flavanones (D–E, H), phenylated flavone (C), binaphthalene (A), and diphenylheptanoid or curcumin analog (F). The compounds in the xanthone group showed similar binding modes: hydrogen bonds between the N ζ atom of Lys721 and the O atom of the hydroxyl group on the xanthone backbone, and between the carboxy O atom of Asp831 and the H atom of the hydroxyl group on the xanthone backbone (Fig. 4b, g). Conserved residues that interact with compounds in flavanone and phenylate flavone include Lys721 and Met769, both of which participate in π –cation and hydrogen-bond interactions (Fig. 4c–e, h).

We also analyzed all eight compounds based on physicochemical property predictions [75]. There are only two hits, A and F, that are likely to provide good oral bioavailability, absorption and permeation. Therefore, we

Table 2 Chemical name(s) and plant source(s)

CAS no. (compound)	Structure name	Scientific name(s) of plant source(s)	Common name(s) of plant source(s)
104096-45-9 (A)	N/A	<i>Diospyros mollis</i> Griff, <i>Diospyros ehretioides</i> wall.ex G.Don	Ebony tree
112649-21-5 (B)	Garcinone E	<i>Garcinia mangostana</i> L.	Mangosteen
113866-89-0 (C)	AC 5-1	<i>Annona squamosa</i> L., <i>Citrus hystrix</i> DC.	Custard apple and makrut lime
142608-98-8 (D)	Exiguaflavanone A	<i>Artemisia indica</i> Willd, <i>Sophora exigua</i>	Legume
142608-99-9 (E)	Exiguaflavanone B	<i>Artemisia indica</i> Willd, <i>Sophora exigua</i>	Legume
144761-33-1 (F)	1,6-Heptadiene-3,5-dione	<i>Curcuma aromatica</i> Salisb, <i>Curcuma xanthorrhiza</i>	Wild turmeric
155233-17-3 (G)	Caloxanthone B	<i>Calophyllum inophyllum</i>	Ballnut
80510-05-0 (H)	Euchrestaflavanone A; lespedezaflavanone B	<i>Azadirachta indica</i> A.Juss	Neem

N/A not available

Table 3 Score of each docking method and total electrostatic energy values

CAS no. (Compound)	Structure name	ADT4		GOLD		GLIDE
		ΔG^a_{ADT} (kcal/mol)	Oechemscore (kcal/mol)	KCSscore	ΔG^a_{GOLD} (kcal/mol)	XPscore (kcal/mol)
183321-74-6	Erlotinib	-7.71	-7.86	28.65	-7.42	-8.32
104096-45-9 (A)	N/A*	-8.93	-9.62	32.81	-8.23	-9.22
112649-21-5 (B)	Garcinone E	-10.07	-10.24	34.18	-8.41	-9.13
113866-89-0 (C)	AC 5-1	-8.27	-9.97	34.18	-9.25	-8.85
142608-98-8 (D)	Exiguaflavanone A	-8.05	-10.18	31.20	-7.60	-9.43
142608-99-9 (E)	Exiguaflavanone B	-8.38	-10.38	33.43	-8.23	-9.53
144761-33-1 (F)	1,6-Heptadiene-3,5-dione	-8.24	-8.30	33.02	-9.05	-8.48
155233-17-3 (G)	Caloxanthone B	-10.24	-9.60	35.07	-8.48	-8.57
80510-05-0 (H)	Euchrestaflavanone A; Lespedezaflavanone B	-9.24	-8.65	35.07	-8.68	-8.61

N/A* not available

^a Estimated binding energy from docking programs

further investigated these two compounds in complex with EGFR-TK, WT-CPA and WT-CPF, using molecular dynamics, and compared the results with the WT-IRE complex. As shown in Fig. 6, the RMSD plots for A were more stable

than those for F. This can be explained by their different ligand structures; the binaphthalene group of compound A seems to be more rigid than the middle part of F. Moreover, in the WT-CPF complex, the structure of F changes

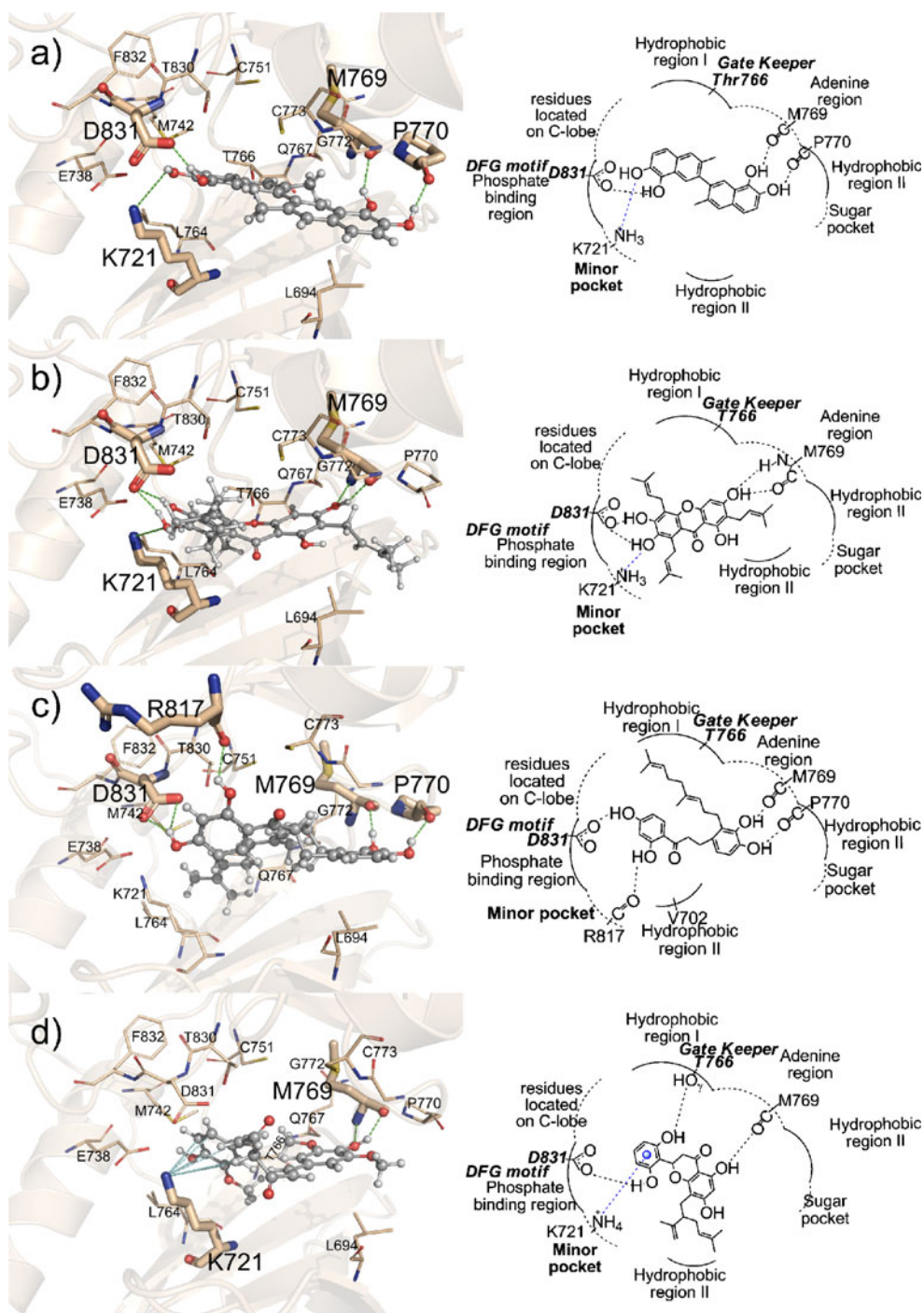


Fig. 4 2D (right panels) and 3D (left panels) diagrams showing protein–ligand interactions between EGFR-TK and the eight hit compounds (a–h). The hit compounds, the amino acid residue involved in the interaction with the hit compound, and the other residues around the binding pocket are represented in ball and stick

and stick and line forms, respectively. The hydrogen bond, π - π , and cation- π interactions between the compounds and the binding residues are shown as green dashed and light blue dashed lines, respectively

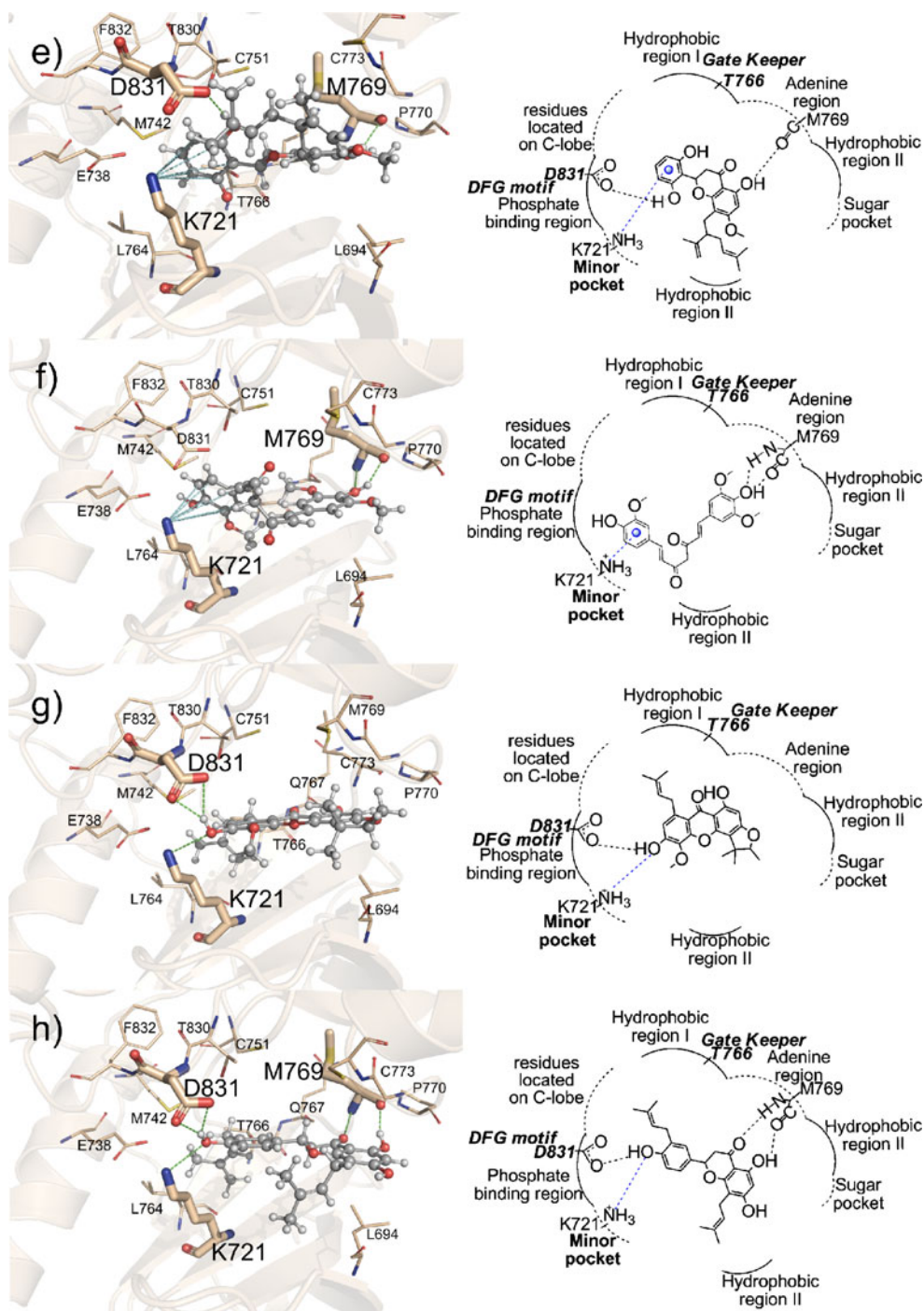


Fig. 4 (continued)

dramatically after 10 ns and then appears to converge after 14 ns. Furthermore, binding site analyses showed that there are differences in the residue(s)–compound interactions between these two complexes (Fig. 4a and f). The MD results also revealed that the hydrogen bond between the carbonyl oxygen atom of Met 769 and the H atom of the hydroxyl group of A (see Fig. 4a) existed approximately

77% of the time during the MD simulation. At the same time, the H atom of the hydroxyl group from A was able to form hydrogen bonds with the O δ 1 atom and the O δ 2 atom of Asp 831, which existed approximately 50% and 10% of the time during the MD simulation, respectively (Fig. 4a, Table S1). On the other hand, no hydrogen bond between Pro770 and A was found. In addition, the probability that

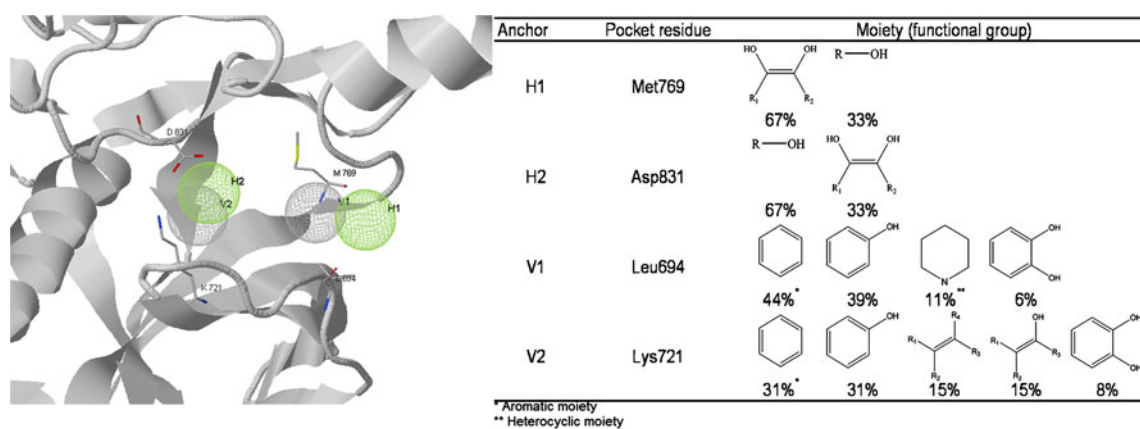


Fig. 5 The site-moiety map anchors in the binding pocket of EGFR-TK that interact with the eight candidate compounds

Fig. 6 Backbone root mean square deviation (RMSD) plots for the simulated complexes WT-IRE, WT-CPA and WT-CPF from top to bottom, respectively

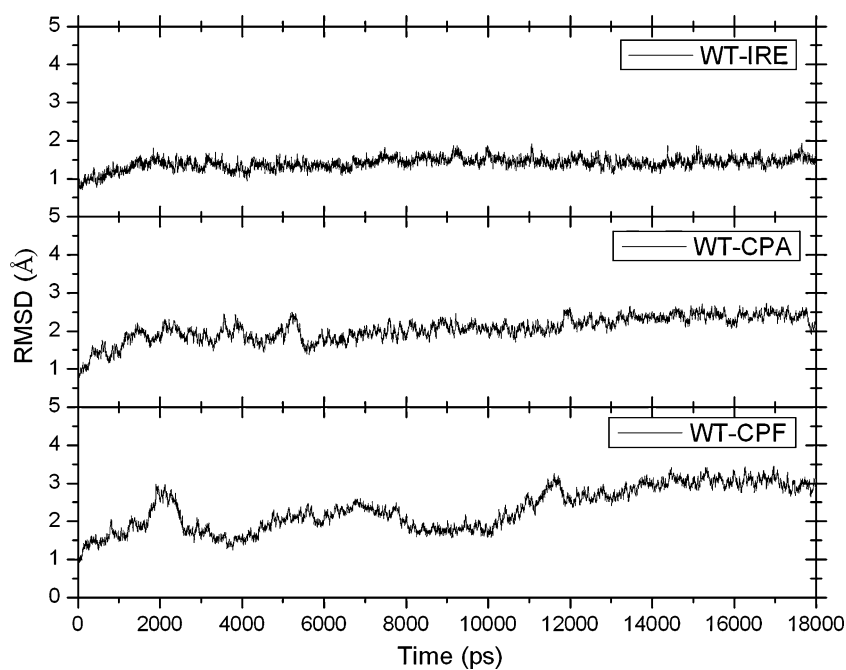


Table 4 Binding free energies (kcal/mol) resulting from MM-PBSA analyses of the WT-IRE, WT-CPA and WT-CPF complexes

Complex	ΔE_{ELE}	ΔE_{VDW}	ΔE_{INT}	$\Delta\Delta G_{\text{SA}}$	$\Delta\Delta G_{\text{PB}}$	$\Delta G_{\text{MM/PBSA}}$	Nonpolar/hydrophobic	Polar/electrostatic
WT-IRE	9.16	-49.41	0.00	-6.61	19.26	-27.61	-56.02	28.42
WT-CPA	-55.31	-32.59	-4.79	-5.71	75.47	-22.92	-43.09	20.16
WT-CPF	-14.33	-45.02	0.00	-6.11	44.19	-21.27	-51.13	29.86

Nonpolar contribution = $\Delta E_{\text{VDW}} + \Delta E_{\text{INT}} + \Delta\Delta G_{\text{SA}}$; polar contribution = $\Delta E_{\text{ELE}} + \Delta\Delta G_{\text{PB}}$

these two hydrogen bonds were present in the WT-CPF complex (Fig. 4f, Table S1) was only around 7–9%. According to Table 4, van der Waals interactions play an important role in the simulations, contributing more than the electrostatic interactions to the total energy in all three calculations.

Conclusions

Our results showed that the eight chemical compounds obtained as hits using our screening method could all be inhibitors of EGFR-TK. This suggests that our docking procedure can be reliably used to predict the binding modes of Chemiebase compounds. These small molecules were obtained from fruits and vegetables that are available locally, and have been used as components of ancient medicinal recipes. Anticancer activities were reported for all eight of the hit compounds based on traditional medicinal sources, which means they could be potential leads in the drug discovery process in the future. In addition, these results suggested that Asian medicines extracted from plants might contain potential compounds against many types of cancer.

Acknowledgments This work was supported by the Thailand Research Fund [TRF project code MRG4980061, TRF Senior Research Scholars (RTA 5380010)], BRC 13/2551, Faculty of Science, Graduate School Research Fund, Kasetsart University, and Kasetsart University Research and Development Institute. We would like to thank the Interdisciplinary Graduate Program in Genetic Engineering, Kasetsart University, for providing SYBYL 7.3, as well as the Institute for Theoretical Chemistry, University of Vienna, Austria, for generously providing the MOE and Schrödinger programs as well as computing time and research facilities. The usage of the GOLD program was allowed by the National Center of Excellence in Petroleum, Petrochemical Technology and Advanced Materials. Thanks are due to Dr. Anton Bayer and Dr. Witcha Treesuwan for their suggestions.

References

- Paul MK, Mukhopadhyay AK (2004) Tyrosine kinase—role and significance in cancer. *Int J Med Sci* 1:101–115
- Harari PM, Huang SM (2002) Epidermal growth factor receptor modulation of radiation response: preclinical and clinical development. *Semin Radiat Oncol* 12:21–26
- Choong NW, Ma PC, Salgia R (2005) Therapeutic targeting of receptor tyrosine kinases in lung cancer. *Expert Opin Ther Targets* 9:533–559
- Cavasotto CN, Ortiz MA, Abagyan RA, Piedrafita FJ (2006) In silico identification of novel EGFR inhibitors with antiproliferative activity against cancer cells. *Bioorg Med Chem Lett* 16:1969–1974
- Janmaat ML, Giaccone G (2003) Small-molecule epidermal growth factor receptor tyrosine kinase inhibitors. *Oncologist* 8:576–586
- Xu G, Abad MC, Connolly PJ, Neeper MP, Struble GT, Springer BA, Emanuel SL, Pandey N, Gruninger RH, Adams M, Moreno-Mazza S, Fuentes-Pesquera AR, Middleton SA (2008) 4-Amino-6-arylamino-pyrimidine-5-carbaldehyde hydrazones as potent ErbB-2/EGFR dual kinase inhibitors. *Bioorg Med Chem Lett* 18:4615–4619
- Xu G, Searle LL, Hughes TV, Beck AK, Connolly PJ, Abad MC, Neeper MP, Struble GT, Springer BA, Emanuel SL, Gruninger RH, Pandey N, Adams M, Moreno-Mazza S, Fuentes-Pesquera AR, Middleton SA, Greenberger LM (2008) Discovery of novel 4-amino-6-arylaminopyrimidine-5-carbaldehyde oximes as dual inhibitors of EGFR and ErbB-2 protein tyrosine kinases. *Bioorg Med Chem Lett* 18:3495–3499
- Moyer JD, Barbacci EG, Iwata KK, Arnold L, Boman B, Cunningham A, DiOrio C, Doty J, Morin MJ, Moyer MP, Neveu M, Pollack VA, Pustilnik LR, Reynolds MM, Sloan D, Theleman A, Miller P (1997) Induction of apoptosis and cell cycle arrest by CP-358,774, an inhibitor of epidermal growth factor receptor tyrosine kinase. *Cancer Res* 57:4838–4848
- Rusnak DW, Lackey K, Affleck K, Wood ER, Alligood KJ, Rhodes N, Keith BR, Murray DM, Knight WB, Mullin RJ, Gilmer TM (2001) The effects of the novel, reversible epidermal growth factor receptor/ErbB-2 tyrosine kinase inhibitor, GW2016, on the growth of human normal and tumor-derived cell lines in vitro and in vivo. *Mol Cancer Ther* 1:85–94
- Arteaga CL, Johnson DH (2001) Tyrosine kinase inhibitors—ZD1839 (Iressa). *Curr Opin Oncol* 13:491–498
- Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci USA* 98:10037–10041
- Hennequin LF, Stokes ES, Thomas AP, Johnstone C, Ple PA, Ogilvie DJ, Dukes M, Wedge SR, Kendrew J, Curwen JO (2002) Novel 4-anilinoquinazolines with C-7 basic side chains: design and structure activity relationship of a series of potent, orally active, VEGF receptor tyrosine kinase inhibitors. *J Med Chem* 45:1300–1312
- Rabindran SK, Discifani CM, Rosfjord EC, Baxter M, Floyd MB, Golas J, Hallett WA, Johnson BD, Nilakantan R, Overbeek E, Reich MF, Shen R, Shi X, Tsou HR, Wang YF, Wissner A (2004) Antitumor activity of HKI-272, an orally active, irreversible inhibitor of the HER-2 tyrosine kinase. *Cancer Res* 64:3958–3965
- Li D, Ambrogio L, Shimamura T, Kubo S, Takahashi M, Chiriac LR, Padera RF, Shapiro GI, Baum A, Himmelsbach F, Rettig WJ, Meyerson M, Solca F, Greulich H, Wong KK (2008) BIBW2992, an irreversible EGFR/HER2 inhibitor highly effective in preclinical lung cancer models. *Oncogene* 27:4702–4711
- Minkovsky N, Berezov A (2008) BIBW-2992, a dual receptor tyrosine kinase inhibitor for the treatment of solid tumors. *Curr Opin Investig Drugs* 9:1336–1346
- Riely GJ (2008) Second-generation epidermal growth factor receptor tyrosine kinase inhibitors in non-small cell lung cancer. *J Thorac Oncol* 3:S146–S149
- Bayes M, Rabassada X, Prous JR (2005) Gateways to clinical trials. *Methods Find Exp Clin Pharmacol* 27:49–77
- Smaill JB, Rewcastle GW, Loo JA, Greis KD, Chan OH, Reyner EL, Lipka E, Showalter HD, Vincent PW, Elliott WL, Denny WA (2000) Tyrosine kinase inhibitors. 17. Irreversible inhibitors of the epidermal growth factor receptor: 4-(phenylamino)quinazoline- and 4-(phenylamino)pyrido[3,2-d]pyrimidine-6-acrylamides bearing additional solubilizing functions. *J Med Chem* 43:1380–1397
- Traxler P, Allegrini PR, Brandt R, Brueggen J, Cozens R, Fabbro D, Grosios K, Lane HA, McSheehy P, Mestan J, Meyer T, Tang C, Wartmann M, Wood J, Caravatti G (2004) AEE788: a dual family epidermal growth factor receptor/ErbB2 and vascular endothelial growth factor receptor tyrosine kinase inhibitor with antitumor and antiangiogenic activity. *Cancer Res* 64:4931–4941

20. Bruns CJ, Solorzano CC, Harbison MT, Ozawa S, Tsan R, Fan D, Abbruzzese J, Traxler P, Buchdunger E, Radinsky R, Fidler IJ (2000) Blockade of the epidermal growth factor receptor signaling by a novel tyrosine kinase inhibitor leads to apoptosis of endothelial cells and therapy of human pancreatic carcinoma. *Cancer Res* 60:2926–2935
21. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, Louis DN, Christiani DC, Settleman J, Haber DA (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med* 350:2129–2139
22. Pao W, Miller VA (2005) Epidermal growth factor receptor mutations, small-molecule kinase inhibitors, and non-small-cell lung cancer: current knowledge and future directions. *J Clin Oncol* 23:2556–2568
23. Cragg GM, Newman DJ (2005) Plants as a source of anti-cancer agents. *J Ethnopharmacol* 100:72–79
24. Newman DJ, Cragg GM (2007) Natural products as sources of new drugs over the last 25 years. *J Nat Prod* 70:461–477
25. Butler MS (2008) Natural products to drugs: natural product-derived compounds in clinical trials. *Nat Prod Rep* 25:475–516
26. Stamos J, Sliwkowski MX, Eigenbrot C (2002) Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *J Biol Chem* 277:46265–46272
27. Wood ER, Truesdale AT, McDonald OB, Yuan D, Hassell A, Dickerson SH, Ellis B, Pennisi C, Horne E, Lackey K, Alligood KJ, Rusnak DW, Gilmer TM, Shewchuk L (2004) A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells. *Cancer Res* 64:6652–6659
28. Zhang X, Gureasko J, Shen K, Cole PA, Kuriyan J (2006) An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* 125:1137–1149
29. Blair JA, Rauh D, Kung C, Yun CH, Fan QW, Rode H, Zhang C, Eck MJ, Weiss WA, Shokat KM (2007) Structure-guided development of affinity probes for tyrosine kinases using chemical genetics. *Nat Chem Biol* 3:229–238
30. Yun CH, Boggon TJ, Li Y, Woo MS, Greulich H, Meyerson M, Eck MJ (2007) Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell* 11:217–227
31. Zhang X, Pickin KA, Bose R, Jura N, Cole PA, Kuriyan J (2007) Inhibition of the EGF receptor by binding of MIG6 to an activating kinase domain interface. *Nature* 450:741–744
32. Yun CH, Mengwasser KE, Toms AV, Woo MS, Greulich H, Wong KK, Meyerson M, Eck MJ (2008) The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc Natl Acad Sci USA* 105:2070–2075
33. Red Brewer M, Choi SH, Alvarado D, Moravcevic K, Pozzi A, Lemmon MA, Carpenter G (2009) The juxtamembrane region of the EGF receptor functions as an activation domain. *Mol Cell* 34:641–651
34. Jura N, Endres NF, Engel K, Deindl S, Das R, Lamers MH, Wemmer DE, Zhang X, Kuriyan J (2009) Mechanism for activation of the EGF receptor catalytic domain by the juxtamembrane segment. *Cell* 137:1293–1307
35. Zhou W, Ercan D, Chen L, Yun CH, Li D, Capelletti M, Cortot AB, Chiriac L, Jacob RE, Padera R, Engen JR, Wong KK, Eck MJ, Gray NS, Janne PA (2009) Novel mutant-selective EGFR kinase inhibitors against EGFR T790M. *Nature* 462:1070–1074
36. Sangma C, Chuakheaw D, Jongkon N, Saenbandit K, Nunriem P, Uthayopas P, Hannongbua S (2005) Virtual screening for anti-HIV-1 RT and anti-HIV-1 PR inhibitors from the Thai medicinal plants database: a combined docking with neural networks approach. *Comb Chem High Throughput Screen* 8:417–429
37. 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
38. Huey R, Goodsell DS, Morris GM, Olson AJ (2004) Grid-based hydrogen bond potentials with improved directionality. *Lett Drug Des Discovery* 1:178–183
39. 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
40. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Improved protein-ligand docking using GOLD. *Proteins* 52:609–623
41. Cambridge Crystallographic Data Centre (2008) GOLD 4.0. Cambridge Crystallographic Data Centre, Cambridge
42. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT (2006) Extra precision Glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem* 49:6177–6196
43. Gasteiger J, Rudolph C, Sadowski J (1990) Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput Methodol* 3:537–547
44. Schmidt MW, Baldrige KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH, Koseki S, Matsunaga N, Nguyen KA, Su S, Windus TL (1993) The general atomic and molecular electronic structure system. *J Comput Chem* 14:1347–1363
45. OpenEye Scientific Software Inc. (2011) FILTER program. OpenEye Scientific Software Inc., Santa Fe. <http://www.eyesopen.com>
46. Clark AM, Labute P (2007) 2D depiction of protein-ligand complexes. *J Chem Inf Model* 47:1933–1944
47. Tripos Associates (2006) Sybyl 7.3. Tripos Associates, St. Louis
48. OpenEye Scientific Software Inc. (2011) FRED program. OpenEye Scientific Software Inc., Santa Fe. <http://www.eyesopen.com>
49. La Motta C, Sartini S, Tuccinardi T, Nerini E, Da Settimo F, Martinelli A (2009) Computational studies of epidermal growth factor receptor: docking reliability, three-dimensional quantitative structure-activity relationship analysis, and virtual screening studies. *J Med Chem* 52:964–975
50. Schrödinger, LLC (2008) Maestro, v.8.5. Schrödinger, LLC, New York
51. Schrödinger, LLC (2008) Ligprep, v.2.2. Schrödinger, LLC, New York
52. Schrödinger, LLC (2008) Glide, v.5.0. Schrödinger LLC, New York
53. Chen YF, Hsu KC, Lin SR, Wang WC, Huang YC, Yang JM (2010) SiMMap: a web server for inferring site-moiety map to recognize interaction preferences between protein pockets and compound moieties. *Nucleic Acids Res* 38:W424–430
54. Hodgson J (2001) ADMET—turning chemicals into drugs. *Nat Biotechnol* 19:722–726
55. Clark DE, Grootenhuis PD (2002) Progress in computational methods for the prediction of ADMET properties. *Curr Opin Drug Discov Devel* 5:382–390
56. DeLano WL (2002) The PyMOL Molecular Graphics System. DeLano Scientific, Palo Alto. <http://www.pymol.org>
57. 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
58. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general Amber force field. *J Comput Chem* 25:1157–1174

59. 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
60. Ryckaert JP, Hinton DP, Byrd RA (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
61. Choowongkorn K, Sawatdichaikul O, Songtawee N, Limtrakul J (2010) Receptor-based virtual screening of EGFR kinase inhibitors from the NCI Diversity Database. *Molecules* 15:4041–4054
62. Zhou Z, Felts AK, Friesner RA, Levy RM (2007) Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. *J Chem Inf Model* 47:1599–608
63. Loder JW, Mongolsuk S, Robertson A, Whalley WB (1957) Diospyrol, a constituent of *Diospyros mollis*. *J Chem Soc* 2233–2237
64. Mallavadhani UV, Panda AK, Rao YR (1998) Pharmacology and chemotaxonomy of *Diospyros*. *Phytochemistry* 49:901–951
65. Ho CK, Huang YL, Chen CC (2002) Garcinone E, a xanthone derivative, has potent cytotoxic effect against hepatocellular carcinoma cell lines. *Planta Med* 68:975–979
66. Matsumoto K, Akao Y, Kobayashi E, Ohguchi K, Ito T, Tanaka T, Iinuma M, Nozawa Y (2003) Induction of apoptosis by xanthenes from mangosteen in human leukemia cell lines. *J Nat Prod* 66:1124–1127
67. Suphavanich K, Maitarad P, Hannongbua S, Sudta P, Suksamrarn S, Tantirungrotechai Y, Limtrakul J (2009) CoMFA and CoMSIA studies on a new series of xanthone derivatives against the oral human epidermoid carcinoma (KB) cancer cell line. *Monatsh Chem* 140:273–280
68. Shimizu K, Kondo R, Sakai K, Buabarn S, Dilokkunanant U (2000) 5 α -Reductase inhibitory component from leaves of *Artocarpus altilis*. *J Wood Sci* 46:385–389
69. Ruangrunsi N, Iinuma M, Tanaka T, Ohyama M, Yokoyama J, Mizuno M (1992) Three flavanones with a lavandulyl group in the roots of *Sophora exigua*. *Phytochemistry* 31:999–1001
70. Chanphen R, Thebtaranonth Y, Wanauppathamkul S, Yuthavong Y (1998) Antimalarial principles from *Artemisia indica*. *J Ethnopharmacol* 61:1146–1147
71. Tsuchiya H, Sato M, Miyazaki T, Fujiwara S, Tanigaki S, Ohyama M, Tanaka T, Iinuma M (1996) Comparative study on the antibacterial activity of phytochemical flavanones against methicillin-resistant *Staphylococcus aureus*. *J Ethnopharmacol* 50:27–34
72. Venkateswarlu S, Ramachandra MS, Subbaraju GV (2005) Synthesis and biological evaluation of polyhydroxycurcuminoids. *Bioorg Med Chem* 13:6374–6380
73. Nakahara K, Roy MK, Ono H, Maeda I, Ohnishi-Kameyama M, Yoshida M, Trakoontivakorn G (2003) Prenylated flavanones isolated from flowers of *Azadirachta indica* (the neem tree) as antimutagenic constituents against heterocyclic amines. *J Agric Food Chem* 51:6456–6460
74. Liu B, Bernard B, Wu JH (2006) Impact of EGFR point mutations on the sensitivity to gefitinib: insights from comparative structural analyses and molecular dynamics simulations. *Proteins* 65:331–346
75. Accelrys Inc. (2009) Discovery Studio 2.5. Accelrys Inc., San Diego. <http://www.accelrys.com>
76. Carey KD, Garton AJ, Romero MS, Kahler J, Thomson S, Ross S, Park F, Haley JD, Gibson N, Sliwkowski MX (2006) Kinetic analysis of epidermal growth factor receptor somatic mutant proteins shows increased sensitivity to the epidermal growth factor receptor tyrosine kinase inhibitor, erlotinib. *Cancer Res* 66:8163–8171