Theor Chem Acc (2007) 117: 99–113
DOI 10.1007/s00214-006-0151-z

REGULAR ARTICLE

Irina S. Moreira · Pedro A. Fernandes
Maria J. Ramos

Unravelling Hot Spots: a comprehensive computational mutagenesis study

Received: 2 March 2006 / Accepted: 12 April 2006 / Published online: 11 July 2006

Abstract As protein–protein interactions are critical for all biological functions, representing a large and important class of targets for human therapeutics, identification of protein–protein interaction sites and detection of specific amino acid residues that contribute to the specificity and strength of protein interactions is very important in the biochemistry field. Alanine scanning mutagenesis has allowed the discovery of energetically crucial determinants for protein association that have been defined as hot spots. Systematic experimental mutagenesis is very laborious and time-consuming to perform, and thus it is important to achieve an accurate, predictive computational methodology for alanine scanning mutagenesis, capable of reproducing the experimental mutagenesis values. Having as a basis the MM–PBSA approach first developed by Massova et al., we performed a complete study of the influence of the variation of different parameters, such as the internal dielectric constant, the solvent representation, and the number of trajectories, in the accuracy of the free energy binding differences. As a result, we present here a very simple and fast methodological approach that achieved an overall success rate of 82% in reproducing the experimental mutagenesis data.

Keywords MM–PBSA · Molecular dynamics · Alanine scanning mutagenesis · Mutagenesis · Free binding energy · Hot spots

0 Introduction

Protein–protein interactions and complex formation are a key matter in understanding cellular functions [1]. One of the most important characteristics of proteic complexes are the residues that constitute the functional binding site. These residues are dominant for the complex function, and determinants for the specificity at intermolecular protein interfaces [2]. These energetically determinant residues are grouped in structural clusters and have been defined as hot spots [3,4]. Hot spots are generally located near the center of protein–protein interface, and one of its most important features is their exclusion from the solvent [2]. Hot spots show high functional and structural adaptability, with different protein partners binding to the same hot spots, which adapt to present the same residues in different structural contexts [4].

Therefore, a hot spot has been defined as a site where alanine mutations cause an increase in the binding free energy ($\Delta\Delta G_{\text{binding}}$) larger than 4.0 kCal/mol, even though lower values are used for statistical analyses [5,6]. The warm-spots are those with binding free energy differences ranging from 2.0 to 4.0 kCal/mol, and the null-spots are the residues with binding free energy differences lower than 2.0 kCal/mol [6].

Alanine-scanning mutagenesis of protein–protein interfacial residues is an important method to determine hot spots permitting the systematic analysis of individual residues. The reliable prediction of important residues to complex formation is fundamental to protein engineering and to rational drug design [7]. Therefore, the development of a quantitative model for the determination of the binding free energy differences and the understanding of the physical basis of affinity and specificity of the complex interaction are vital in computational biochemistry [8].

A number of computational methods with different levels of rigor and speed to identify the hot spots have been developed. Free energy perturbation (FEP) and thermodynamic integration (TI) yield rigorous and accurate free energy differences but are extremely time-consuming preventing them from being commonly used in the systematic protein–protein interface alanine scanning mutagenesis, which involves dozens of residues [9]. Simple physical models [10,11], empirical methods [12], linear interaction energy methods [13], and Monte Carlo methods [14] have been proposed to identify the residues contributing significantly to

Electronic Supplementary material Supplementary materials available to authorised users in the online version of this article at http://dx.doi.org/10.1007/s00214-006-0151-z

I. S. Moreira · P. A. Fernandes · M. J. Ramos (✉)
REQUIMTE/Departamento de Química,
Faculdade de Ciências da Universidade do Porto,
Rua do Campo Alegre 687, 4169-007 Porto, Portugal
E-mail: mjramos@fc.up.pt

the stability of protein associations. Another methodological approach, which is becoming a technique highly used is the MM–PBSA method (Molecular Mechanics/Poisson–Boltzmann Surface Area) [15–21]. This method is a fully atomistic approach that combines a molecular mechanics proteic complex and a continuum solvent model.

As systematic experimental mutagenesis is very laborious and time-consuming to perform, it is important to develop an accurate, predictive computational methodology for alanine scanning mutagenesis, capable of reproducing the experimental mutagenesis values. An important advantage of computer simulations over experiments is not only to provide fast quantitative estimates, but also, and mainly, to enhance our understanding of the nature of complex formation in terms of the biophysical features of the process because they add molecular insight into the macroscopic properties measured therein.

Hence, having as a basis the MM–PBSA approach, we focussed our attention in ways to decrease the computational time involved to permit a systematic alanine scanning mutagenesis of protein–protein interfaces, as well as in techniques that enable the achievement of the chemical accuracy of roughly 1 kCal/mol. Thus, as computational approaches should represent a good compromise between accuracy and time necessary to reach the correct $\Delta\Delta G_{\text{binding}}$ value we tried to find solutions to increase the present success rates, which have been rather modest so far, and therefore achieve a fast and reliable procedure that provides theoretical results capable of reproducing the quantitative free energy differences obtained from experimental methods.

In the next sections we report our attempts and also the result, i.e. a simple and fast methodological approach to carry out computational alanine scanning mutagenesis.

1 Computational methodology

1.1 Models setup

The models used to performed this work are three protein–protein complexes. The first one, represented in Fig. 1a, is expressed in *Escherichia coli* and formed between two essential components of the septal ring structure that mediates bacterial cell division. The proteins presented in this complex are a cell division protein ZipA (a 36 kDa membrane-anchored protein) and a cell division protein FtsZ. The second model considered, represented in Fig. 1b, is a human immunoglobulin IgG complexed with the C2 fragment of streptococcal protein G. The streptococcal protein G comprises two or three domains that bind to the constant Fc region of most mammalian immunoglobulins IgG mainly by charge and polar contacts. The third complex, shown in Fig. 1c, is an immunoglobulin of mouse, antibody D1.3 complexed with a Hen Egg lysozyme. The crystallographic structures with a resolution of 1.95, 3.50 and 1.80 Å, respectively, were taken from the RCSB Protein Data Bank with the PDB entry: 1F47 [22], 1FCC [23] and 1VFB [24]. About 1,243 hydro-

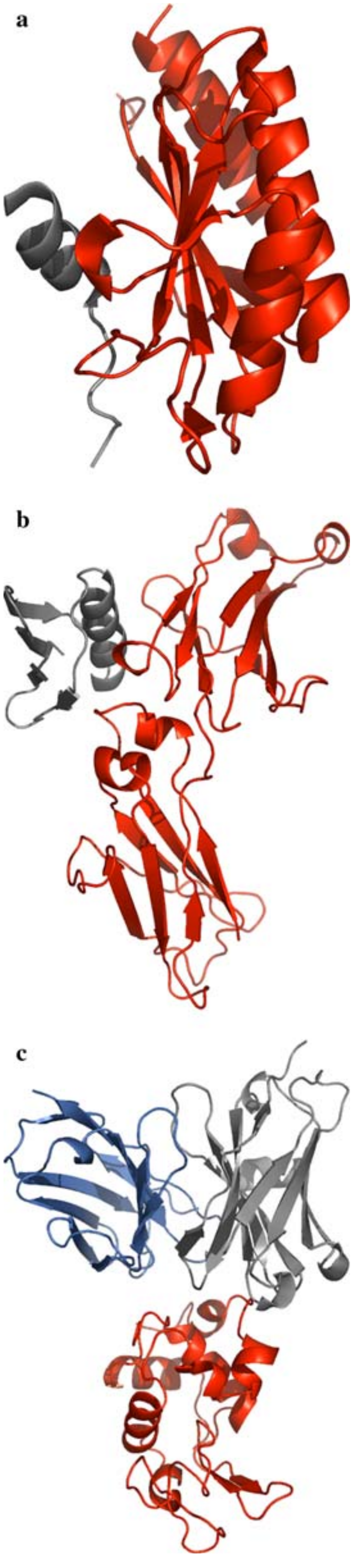


Fig. 1 **a** The bacterial cell-division protein ZipA (in *red*) and its interaction with an FtsZ fragment (in *grey*); **b** the human immunoglobulin IgG (in *red*) complexed with the C2 fragment of streptococcal protein G (in *grey*); **c** an immunoglobulin (in *blue* and *grey*) complexed with a hen egg lysozyme (in *red*)

gen atoms were added to the first complex using the software Protonate from the Amber8 package [25]. The whole system comprised a total of 159 aminoacids, 15 of which in the FtsZ protein (248 atoms) and 144 in the ZipA protein (2,251 atoms). About 2,047 hydrogen atoms were also added to the second complex, and the whole system became a total of 262 amino acids, 206 in protein IgG1 (3,288 atoms) and 56 in protein G (849 atoms). Similarly, 1,965 hydrogen atoms were added to the third complex, and the whole system totalled 352 aminoacids, 107 from VL domain of the antibody D1.3 (1,624 atoms), 116 in VH domain of the antibody D1.3 (1,776 atoms) and 129 belonging to the lysozyme (1,960 atoms). All residues were included in their physiological protonation states (charged Glu, Asp, Lys and Arg, all other residues neutral). All molecular mechanics simulations presented in this work were performed using the Sander module, implemented in the Amber8 [25] simulations package, with the *Cornell* force field [26] and with the TIP3P water model [27].

1.2 Molecular dynamics

There are various methods for treating solvation, ranging from a explicit description at the molecular level to reaction field models where the solvent is modelled as a continuum.

The systems were first minimized with 1,000 steps of steepest descent followed by 1,000 steps of conjugate gradient to release the bad contacts in the crystallographic structures or between the protein and the solvent. Subsequently, molecular dynamics (MD) simulations were performed starting from the minimized structures. In all dynamical simulations the bond lengths involving hydrogens were constrained using SHAKE [28], and the equations of motion were integrated with a 2 fs time-step.

1.2.1 Method 1 (explicit solvent simulations)

In the molecular dynamics simulations we have used two different explicit solvent representations: a water cap and a water box. The nonbonded interactions were truncated with a 12 Å cutoff.

1.2.1.1 Water cap

For each complex, a 22 Å water cap was added centered on the carbon α of the residue we wished to mutate to an alanine. The temperature of the system was regulated by the Berendsen algorithm [29] and it was performed with a 2 ns MD simulation using the sander classic model present in Amber6 package [30].

1.2.1.2 Water box

Periodic boundary conditions were applied using the particle mesh Ewald (PME) method [31] to treat long-range electrostatic interactions. Counterions were added to keep the whole system neutral. The box of solvent molecules must be large enough to minimize electrostatic interactions between periodic images of the solute, thus the value of 10 Å was used between each edge of the box and the closest solute atom. The temperature of the system was regulated by the Langevin algorithm [32–34] and it was performed with a 5 ns NVT simulation for the second and third complexes. To the first complex we carried out a 10 ns NVT simulation.

1.2.2 Method 2 (implicit solvent simulations)

In the molecular dynamics simulations, the solvent was modelled through a modified Generalized Born solvation model [35]. The nonbonded interactions were truncated with a 16 Å cutoff, and the temperature of the system was regulated by the Langevin thermostat [32–34]. The total simulation time was 3 ns for the complex 1VFB, 4 ns for the 1F47 complex, and 5 ns for the 1FCC complex.

1.3 Energy minimization

During the energy minimizations performed, a 15 Å nonbonded cutoff was applied.

1.3.1 Method 3

The geometry of the systems was optimized in vacuum until the convergence criterion for the energy gradient was reached, and the root-mean-square of the cartesian elements of the gradient was less than 0.001 kCal/(mole Å).

1.3.2 Method 4

The geometry of the systems were first optimized in vacuum until the convergence criterion for the energy gradient was reached, and the root-mean-square of the cartesian elements of the gradient was less than 0.001 kCal/(mole Å). Additionally, the solvent was modelled with a modified Generalized Born solvation model [35] and the systems were minimized with 1,000 steps of steepest decent followed by 1,000 steps of conjugated gradient.

1.3.3 Method 5

The solvent was modelled with a modified Generalized Born solvation model [35] and the systems were minimized until the convergence criterion for the energy gradient was reached, and the root-mean-square of the cartesian elements of the gradient was less than 0.001 kCal/mole Å.

1.3.4 Method 6

For each complex, a 45 Å water cap was added centered on the geometric centre of the molecule. The geometry of the system was optimized with a reaction field model that allows the calculation of the reaction field energy for a nonperiodic solute in a spherical cap of water, using a numerical Poisson–Boltzmann solver. The system was then minimized with 3,000 steps of steepest decent.

The MM_PBSA script [21] implemented in Amber8 [25] was used to calculate the binding free energies for the complex and for the alanine mutants. To generate the structure of the mutant complex a simple truncation of the mutated side chain was made, replacing C_γ with a hydrogen atom, and setting the C_β–H bond direction to that of the former C_β–C_γ. In the molecular dynamics simulations for the binding free energy calculations, 25 snapshots of the complexes were extracted, one every 20 ps in the last 500 ps of the run.

1.4 Alanine scanning mutagenesis

In this paper, we present a new and improved methodological approach, based in the MM–PBSA protocol first developed by Massova and Kollman [21]. In MM–PBSA, the complexation free energy is calculated using the following thermodynamic cycle (Scheme 1).

In which ΔG_{gas} is the interaction free energy between the ligand and the receptor in the gas phase and ΔG^{lig}_{solv}, ΔG^{rec}_{solv} and ΔG^{cpx}_{solv} are the solvation free energies of the ligand, the receptor and the complex, respectively. The binding free energy difference between an alanine mutant and wild-type complexes is defined as:

$$\Delta\Delta G_{binding} = \Delta G_{binding-mutant} - \Delta G_{binding-wildtype} \quad (1)$$

The binding free energy of two molecules is the difference between the free energy of the complex and the respective monomers (the receptor and the ligand).

$$\Delta G_{binding-molecule} = G_{complex} - (G_{receptor} + G_{ligand}) \quad (2)$$

The MM–PBSA method is based on partitioning the free energy into a sum of enthalpic and entropic contributions. Typical contributions to the free energy of binding include the internal energy (bond, angle and dihedral), the electrostatic and the van der Waals interactions, the free energy of polar solvation, the free energy of nonpolar solvation and the entropic contribution for the molecule's free energy:

$$\begin{aligned} G_{molecule} = {} & E_{internal} + E_{electrostatic} + E_{vdw} \\ & + G_{polarsolvation} + G_{nonpolarsolvation} - TS \quad (3) \end{aligned}$$

The first three terms were calculated using the *Cornell* force field [26] with no cutoff. The electrostatic solvation free energy was calculated by solving the Poisson-Boltzmann equation with the software Delphi v.4 [36, 37]. In this method, the protein is modelled as a dielectric continuum of low polarizability embedded in a dielectric medium of high polarizability. We used a scale (the reciprocal of the grid spacing) of 2.5 grids/Å, a convergence criterion of 0.001 kT/e (the maximum change in potential should be less than 0.001 kT/e) and the molecule filled 90% of the grid box. Potentials at the boundaries of the finite-difference grid were set using the *coulombic* method (based in the sum of the Debye-Huckel potentials generated by all the charges). The dielectric boundary was taken as the molecular surface defined by a 1.4 Å probe sphere and by spheres centred on each atom with radii taken from the PARSE [38] vdW radii parameter set. Standard parm94 charges [26] were used in order to be consistent with the energetics of the simulations performed. These parameters have been shown in an earlier work to constitute a good compromise between accuracy and computing time [39]. The nonpolar contribution to solvation free energy due to van der Waals interactions between the solute and the solvent and cavity formation, was modelled as a term that is dependent on the solvent accessible surface area of the molecule. It was estimated using an empirical relation: $\Delta G_{nonpolar} = \sigma A + \beta$, where A is the solvent-accessible surface area that was estimated using the molsurf program, which is based on the idea primarily developed by Mike Connolly [40]. The σ and β are empirical constants, and the values used were 0.00542 kCal $Å^{-2}$ mol^{-1} and 0.92 kCal mol^{-1}, respectively. The entropy term, obtained as the sum of translational, rotational, and vibrational components, was not calculated because it was assumed, based in previous work, that its contribution to $\Delta\Delta G_{binding}$ is negligible [21].

$$\begin{tikzcd}
\text{ligand}_{aq} \arrow[d, "-\Delta G^{lig}_{solv}"'] & +\,\text{receptor}_{aq} \arrow[d, "-\Delta G^{rec}_{solv}"'] \arrow[r, "\Delta G_{bind}"] & \text{complex}_{aq} \\
\text{ligand}_{gas} & +\,\text{receptor}_{gas} \arrow[r, "\Delta G_{gas}"] & \text{complex}_{gas} \arrow[u, "+\Delta G^{cpx}_{solv}"']
\end{tikzcd}$$

Scheme 1 Thermodynamic cycle used to calculate the complexation free

2 Results

Alanine scanning mutagenesis of protein–protein interfacial residues to unravel hot spots in binding interfaces continues to stimulate interest, since reliable prediction of key residues in the interface has immediate applications in Structure Based Drug Design [41]. The MM–PBSA method by Massova/Kollman [21] is a fully atomistic method that, although not accurate enough in the original implementation, unlocked the possibility for the development of a new improved methodology. In this study we are going to focus our attention essentially in five aspects that we consider of great importance to a higher success rate in predicting correctly the free energy differences upon alanine mutation of the protein–protein interface amino acid residues.

Biological processes are *perforce* dynamical in nature. Therefore, we have initiated our work by performing MD simulations of the complexes under study. An accurate description of solvation effects is indispensable in computer simulations, especially involving biological macromolecules because molecular properties are very sensitive to the environment. Therefore, we have applied Method 2 described in the Computational Methodology section complex 1F47

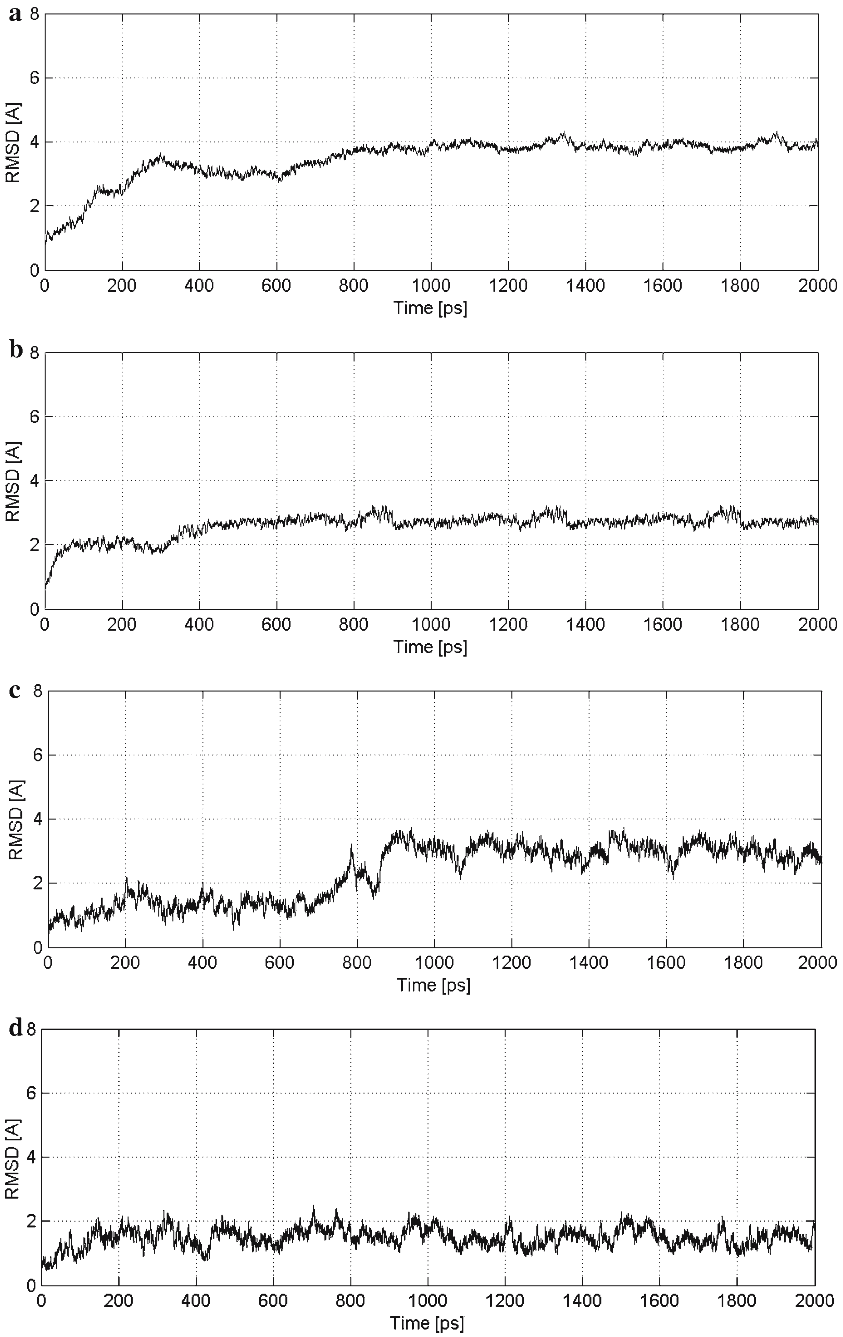


Fig. 2 RMSD plots for the protein backbone of the complex formed between the bacterial cell-division protein ZipA and the FtsZ fragment relative to its initial structure. **a** Wild-type complex MD simulation; **b** Mutant complex MD simulation; **c** Wild-type ligand MD simulation; **d** Mutant ligand MD simulation

represented in Fig. 1a to achieve a good set of structures to analyse. We used two different explicit solvent descriptions: a water cap (method 1.2.1) and a water box (method 1.2.1.2) to study their influence on the precision of the free binding differences. Dynamics simulation in a water box requires thousands of water molecules in the system turning such MD calculation very time consuming and computationally demanding.

Either performing MD or just energy minimizations, we have tried three protocols that have as a major difference the number of structures which have their geometrical arrangement optimized. Therefore, the first protocol named a "single simulation protocol" [21] consisted of optimizing the geometry or running a molecular mechanics simulation with only the wild-type structures and subjecting them to a postprocessing treatment to generate the mutant complexes by a simple truncation of the side chain of the residue we wished to mutate, replacing the C_γ with a hydrogen atom. From the structure of the wild type and the mutant complexes, the monomer structures were subsequently generated just by deletion of the other partner in the complex. Consequently, the free energy of the wild type and mutant monomers and the mutant complexes were calculated without rearrangement of the surrounding environment. We have also tried a "two simulation protocol" based on running two separate trajectories or optimizing the geometry for the wild type and the mutant complexes. The free energy of the monomers was subsequently calculated without optimising or running MD with these structures, and therefore making only a single energy calculation. Finally we have tried a "fourth simulation protocol" named as such because of optimising of the structures or running a MD simulation of four species, the wild-type complex, the mutant complex, the wild-type ligand and the mutant ligand. It is not necessary to run a simulation for the receptor because this monomer was not subjected to mutation, being similar in the two situations, and therefore having its effect cancelled in Eq. (1). As can be noticed from the first mutation protocol to the last, there is a change from 1 MD/system to 1 MD/mutation leading to simulations times proportional to two times or four times the number of mutations and therefore increasing the CPU cost tremendously.

Proteins are complex molecules containing a mixture of neutral, polar and charged amino acids. While the choice of the external dielectric constant depends on the solvent media, the choice of the internal dielectric constant has been the subject of discussion and controversy because the dielectric constant is not a universal constant but simply an adjustable parameter that depends on the model and the methodology used [42–45]. In fact, proteins are very heterogeneous in terms of dielectric response, due to the varying polarity and mobility of the constituent amino acids, and different regions of a protein should be represented by different dielectric constants. The internal dielectric constant is a means of accounting for responses to an electric field that are not treated explicitly [43]. Hence, we have tried different internal dielectric values from 1 to 15 to observe the effect in the binding free energy differences.

Table 1 Summary of the constituents in the water cap simulations for each of the four simulations made for each of the nine mutations of the complex FtsZ: ZipA with a known experimental value

Mutation	Wild-type complex	Mutant complex	Wild-type ligand	Mutant ligand
Asp2Ala	865	866	1,217	1,219
Tyr3Ala	748	750	1,210	1,215
Leu4Ala	681	681	1,198	1,189
Asp5Ala	727	726	1,199	1,185
Ile6Ala	621	621	1,197	1,197
Phe9Ala	738	739	1,194	1,199
Leu10Ala	728	728	1,202	1,192
Arg11Ala	871	879	1,208	1,200
Lys12Ala	914	920	1,207	1,212

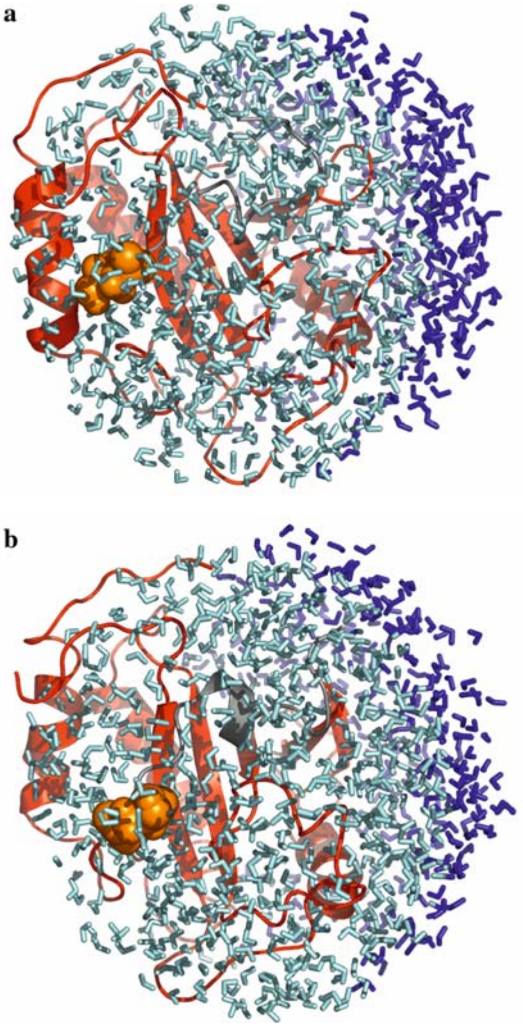


Fig. 3 Two different snapshots of the dynamic simulation in a water cap of the bacterial cell-division protein ZipA and its interaction with an FtsZ fragment. In a cyan stick representation are the water molecules within 22 Å of the Cα of the mutated residue (in *orange*) and in *deep blue* the water molecules outside the 22 Å water cap

2.1 Molecular dynamics

A necessary condition for producing a representative ensemble is that the system is in equilibrium. We have performed four MD simulations with a water cap (see Fig. 3) for each of the nine mutations in the FtsZ:ZipA complex with available experimental values. Although we have ensured that each of the simulations had reached equilibrium, in Fig. 2 the root-mean-square deviations (rmsd) from the X-ray crystal structure of $C\alpha$ atoms of the complex is shown for a representative simulation (from the wild-type and the mutant complex) with a water cap as a function of time.

First, we shall analyse the results from the water cap protocol in which (as it can be observed in Fig. 2) a 22 Å water cap was added centered on the carbon α of the residue we wished to mutate to an alanine. In Table 1, the number of water residues added in each complex subject to computer

Table 2 Results of the Methodological Approach for Computational Alanine Screening Mutagenesis with a water cap explicit solvation protocol

Mutation	$\Delta\Delta G_{exp}$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$
		$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	$\varepsilon = 4$	$\varepsilon = 5$
One simulation protocol						
Asp2AlA	0.69	*−0.20*	*0.14*	*−0.11*	*−0.06*	*−0.03*
Tyr3Ala	0.86	*0.41*	*1.01*	*1.35*	*1.42*	*1.54*
Leu4Ala	0.92	*1.56*	3.61	4.29	4.64	4.82
Asp5Ala	1.73	−1.48	0.23	*0.8*	*1.08*	*1.25*
Ile6Ala	2.50	*3.19*	4.76	5.27	5.34	5.70
Phe9Ala	2.44	4.51	5.59	5.95	6.13	6.23
Leu10Ala	2.29	−0.04	0.56	0.76	0.85	0.91
Arg11Ala	0.00	1.80	3.56	4.14	4.43	4.60
Lys12Ala	0.00	2.02	2.34	2.43	2.46	2.48
Two simulation protocol						
Asp2AlA	0.69	−3.68	−4.63	−4.35	−4.68	−4.86
Tyr3Ala	0.86	15.46	11.44	10.04	9.21	8.84
Leu4Ala	0.92	13.21	7.24	5.22	4.20	3.56
Asp5Ala	1.73	0.40	−9.58	−9.89	−9.93	−15.50
Ile6Ala	2.50	8.72	−0.16	−3.14	−4.64	−5.54
Phe9Ala	2.44	−1.95	4.65	6.84	7.93	8.57
Leu10Ala	2.29	−26.10	−29.13	−32.55	−34.27	−37.30
Arg11Ala	0.00	22.15	34.18	39.14	41.88	42.56
Lys12Ala	0.00	−1.60	4.20	6.75	7.80	8.44
Four simulation protocol						
Asp2AlA	0.69	−8.98	−43.92	−55.51	−53.54	−64.71
Tyr3Ala	0.86	7.9	−2.53	−6.06	−8.41	−8.96
Leu4Ala	0.92	−10.56	−15.87	−17.43	−18.52	−19.45
Asp5Ala	1.73	−50.13	−20.74	−17.79	−10.89	−5.91
Ile6Ala	2.50	−12.81	−17.35	−18.83	−19.53	−19.93
Phe9Ala	2.44	32.49	50.06	55.89	58.8	60.55
Leu10Ala	2.29	−145.14	−131.26	−113.5	−104.6	−99.26
Arg11Ala	0.00	−60.67	−28.05	−17.16	−8.68	−8.45
Lys12Ala	0.00	−35.93	−42.65	−44.86	−45.97	−46.6

The units of free energies are kCal/mol. Numbers in italics are the correct calculated free energy binding values

Table 3 Results of the Methodological Approach for Computational Alanine Screening Mutagenesis with a water box and one simulation protocol for the FtsZ: ZipA complex

Mutation	$\Delta\Delta G_{exp}$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$
		$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	$\varepsilon = 4$	$\varepsilon = 5$
Asp2AlA	0.69	*0.22*	*0.29*	*0.30*	*0.31*	*0.31*
Tyr3Ala	0.86	3.00	4.74	5.31	5.60	5.78
Leu4Ala	0.92	*1.71*	3.16	3.64	3.89	4.03
Asp5Ala	1.73	*0.56*	0.34	0.25	0.21	0.18
Ile6Ala	2.50	*3.66*	4.34	4.25	4.67	4.74
Phe9Ala	2.44	*1.41*	3.93	4.76	5.18	5.43
Leu10Ala	2.29	*1.39*	*1.54*	*1.59*	*1.62*	*1.64*
Arg11Ala	0.00	*0.24*	*0.40*	*0.44*	*0.47*	*0.49*
Lys12Ala	0.00	*−0.06*	*0.36*	*0.50*	*0.58*	*0.64*

The units of free energies are kCal/mol. Numbers in italics the correct calculated free energy binding values

Table 4 Results of the Methodological Approach for Computational Alanine Screening Mutagenesis for the FtsZ: ZipA in a water box and with the different number of trajectories protocols

Protocol	Mutation									
	Asp5Ala					Leu10Ala				
	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	$\varepsilon = 4$	$\varepsilon = 5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	$\varepsilon = 4$	$\varepsilon = 5$
I	*0.56*	0.34	0.25	0.21	0.18	*1.39*	*1.54*	*1.59*	*1.62*	*1.64*
II	142.16	71.46	47.88	*2.84*	*2.69*	140.67	66.37	41.6	–4.01	–4.51
III	395.77	416.03	422.76	425.51	427.11	387.13	412.88	423.28	427.19	429.48

Method I to III correspond to the one simulation, two simulation and four simulation protocol, respectively. The units of free energies are kCal/mol. Numbers in italics are the correct calculated free energy binding values

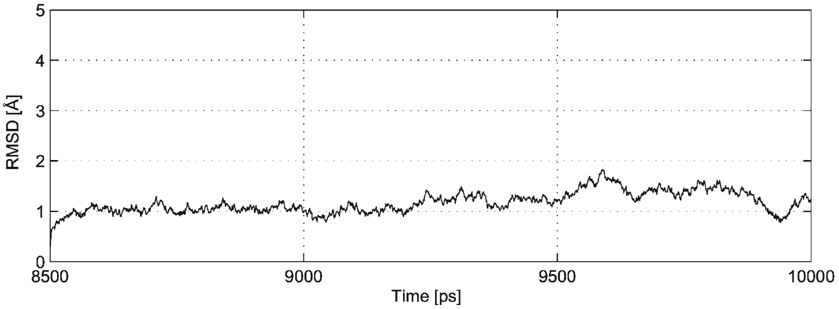


Fig. 4 RMSD plots for the protein backbone of the complex formed between the bacterial cell-division protein ZipA and the FtsZ fragment inserted in a water box relative to its initial structure (the last 1.5 ns of the MD simulation)

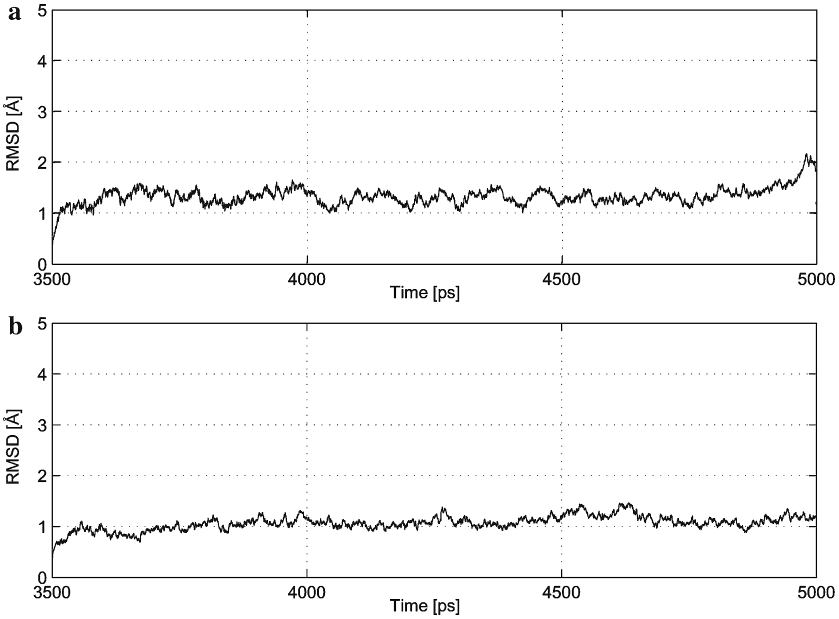


Fig. 5 **a** RMSD plots for the protein backbone of the complex in a water box formed between the human immunoglobulin IgG and the C2 fragment of streptococcal protein G relative to its initial structure; **b** RMSD plots for the protein backbone of the complex formed between an immunoglobulin and a hen egg lysozyme relative to its initial structure (the last 1.5 ns of the MD simulation)

Table 5 Molecular description of the systems subjected to water box MD simulation

	1F47	1FCC	1VFB
# aa residues	159	262	352
# Cl− ions	0	0	11
# Na+ ions	6	1	0
#Water molecules	8,008	3,312	7,306

simulation is represented. If we consider a deviation of ±1.4 kCal/mol from the experimental value as an accurate result (to predict binding affinities within an accuracy of 1 order of magnitude) from Table 2 we can notice that the use of a water cap protocol does not give high success rates (success rates lower than 44%). Different snapshots have a different number of water molecules around the amino acid residues as can be seen in Fig. 2. This might be explained with protein's restriction by a soft half-harmonic potential and by the protein flexibility, which can generate a water cap drift over the protein that is especially important near the boundaries because in some snapshots it can include an amino acid residue and in other snapshots it can leave it desolvated. In fact, even though the water molecules are removed before calculating the free energy differences, their presence/absence influences the geometry of the residues differently in the floating border of the 22 Å water cap, and therefore they can be responsible for the small success rates. It can be also observed that the use of a single trajectory with a water cap as solvent representation to calculate the $\Delta\Delta G_{\text{binding}}$ is conducive to a much better agreement with the experimental data than the use of multiple trajectories. Even though the use of multiple trajectories should in principle give results in closer agreement with the experimental ones, it can be observed as the number of computer simulation grows the accuracy of the results diminishes. Although this seems counter-intuitive, we believe that the use of a single trajectory is more efficient due to the fact that the region of the conformational space accessed by the wild-type complex and the mutant is the same, leading to

Table 6 Results of the Methodological Approach for Computational Alanine Screening Mutagenesis with a water box protocol for all the three complexes studied

Residue type	Protein	Mutation	$\Delta\Delta G_{exp}$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$
				$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	$\varepsilon = 4$	$\varepsilon = 5$
Non-polar residues	1F47	Leu4Ala	0.92	*1.71*	3.16	3.64	3.89	4.03
		Ile6Ala	2.50	*3.66*	4.34	4.25	4.67	4.74
		Phe9Ala	2.44	*1.41*	3.93	4.76	5.18	5.43
		Leu10Ala	2.29	*1.39*	*1.54*	*1.59*	*1.62*	*1.64*
	1Fcc	Trp43Ala	3.80	−2.31	0.35	1.24	1.68	1.95
	1Vfb-L chain	Trp92Ala	1.71	3.30	4.93	5.48	5.75	5.93
	1Vfb-H chain	Trp52Ala	1.23	3.33	5.39	6.09	6.43	6.64
	1Vfb-ligand	Val120Ala	0.90	*0.06*	*0.53*	*0.68*	*0.75*	*0.80*
		Ile124Ala	1.20	*0.03*	*0.69*	*0.91*	*1.02*	*1.09*
Polar residues	1F47	Tyr3Ala	0.86	3.00	4.74	5.31	5.60	5.78
	1Fcc	Thr25Ala	0.24	−2.31	*−0.52*	*0.08*	*0.37*	*0.56*
		Asn35Ala	2.40	−3.35	−0.14	*0.94*	*1.48*	*1.80*
		Thr44Ala	2.00	*1.24*	*1.99*	*2.27*	*2.41*	*2.49*
		Tyr45Ala						
	1Vfb-L chain	Ser93Ala	0.11	*0.01*	*0.32*	*0.40*	*0.47*	*0.55*
		Tyr32Ala	1.30	−2.03	*2.09*	3.47	4.15	4.57
		Tyr49Ala	0.80	−0.96	*0.27*	*0.69*	*0.99*	*1.01*
		Tyr50Ala	0.40	−2.28	*0.50*	*1.44*	1.91	2.20
		Thr53Ala	−0.23	*−0.67*	*−0.14*	*0.04*	*0.13*	*0.19*
	1Vfb-H chain	Thr30Ala	0.09	*1.06*	*0.59*	*0.45*	*0.36*	*0.32*
		Tyr32Ala	0.50	1.98	*1.56*	*1.41*	*1.32*	*1.26*
		Asn56Ala	0.20	*0.36*	*0.28*	*0.26*	*0.24*	*0.24*
		Tyr101Ala	>4.0	2.12	*4.03*	*4.66*	*4.95*	*5.15*
	1Vfb-ligand	Asn19Ala	0.40	−2.61	*0.13*	*1.04*	*1.49*	1.77
		Tyr23Ala	0.80	*0.52*	*1.37*	*1.67*	*1.81*	*1.90*
		Ser24Ala	0.70	7.76	3.71	2.23	*1.54*	*1.13*
		Thr118Ala	0.80	*0.74*	*1.03*	*1.19*	*1.19*	*1.22*
		Gln121Ala	2.90	8.84	8.49	8.36	8.28	8.24
Charged residues	1F47	Asp2Ala	0.69	*0.22*	*0.29*	*0.30*	*0.31*	*0.31*
		Asp5Ala	1.73	*0.56*	0.34	0.25	0.21	0.18
		Arg11Ala	0	*0.24*	*0.40*	*0.44*	*0.47*	*0.49*
		Lys12Ala	0	*−0.06*	*0.36*	*0.50*	*0.58*	*0.64*
	1Fcc	Glu27Ala	>4.90	−16.42	*17.75*	*12.88*	*10.39*	*8.88*
		Lys28Ala	1.30	−5.78	*0.86*	2.98	3.99	4.56
		Lys31Ala	3.50	−5.83	0.86	*2.98*	*3.99*	*4.56*
		Asp40Ala	0.30	−1.72	*−0.61*	*−0.22*	*−0.02*	*0.11*
		Glu42Ala	0.40	*1.09*	*0.76*	*0.66*	*0.61*	*0.59*
	1Vfb-L chain	His30Ala	0.80	*0.76*	*0.82*	*0.86*	*0.87*	*0.89*
	1Vfb-H chain	Asp58Ala	−0.20	*0.62*	*1.08*	1.23	1.29	1.33
		Glu98Ala	1.10	6.06	4.76	4.07	3.63	3.33
		Arg99Ala	0.47	−1.70	−1.18	−0.96	*−0.85*	*−0.75*
		Asp100Ala	3.10	11.06	7.97	6.72	6.23	5.50
	1Vfb-ligand	Asp18Ala	0.30	1.71	*0.50*	*0.11*	*−0.09*	*−0.21*
		Lys116Ala	0.70	4.39	2.50	*1.90*	*1.61*	*1.45*
		Asp119Ala	1.00	9.11	5.83	4.62	3.95	3.52
		Arg125Ala	1.80	−3.62	*0.73*	*2.18*	*2.89*	3.32
% success rate				44	62	60	63	58

The units of free energies are kCal/mol. Numbers in italics the correct calculated free energy binding values

error cancellation, but in a two/four trajectory protocol, the region explored by each system is not necessarily the same. Therefore, the use of a single mutation protocol gives better results due to error cancellation. The mentioned error comes from an incomplete exploration of the rotamer conformational space, which may need hundreds of nanoseconds of simulation time to be complete. Although this is unfeasible presently we are convinced that results should improve with multiple trajectories if calculations times are long enough to explore more substantially the conformational space.

The introduction of a protein–protein complex in a water box, conduces to an increased CPU time. In addition complications also arise from the need to fully equilibrate these molecules and any counterions in the system, which turns these simulations lengthy and costly. We have plotted in Fig. 4 the rmsd from the initial structure of the backbone atoms of the complex for the water box wild-type simulation as a function of time. With the use of Particle-Mesh Ewald (PME) to treat long-range electrostatics, we have obtained stable trajectories for these macromolecules. In Table 3 it can be seen that

Table 7 Results of the Methodological Approach for Computational Alanine Screening Mutagenesis with an implicit solvation protocol for all the three complexes studied

Residue type	Protein	Mutation	$\Delta\Delta G_{exp}$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$
				$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	$\varepsilon = 4$	$\varepsilon = 5$
Non-polar residues	1F47	Leu4Ala	0.92	−1.98	*1.01*	*1.99*	2.47	2.74
		Ile6Ala	2.50	−1.47	*2.42*	3.82	4.02	4.67
		Phe9Ala	2.44	−1.20	*2.48*	*3.69*	4.12	4.67
		Leu10Ala	2.29	*2.32*	*2.73*	*2.98*	*3.06*	*3.10*
	1Fcc	Trp43Ala	3.80	−1.08	1.12	1.84	2.20	2.42
	1Vfb-L chain	Trp92Ala	1.71	−1.88	*2.37*	3.77	4.20	4.90
	1Vfb-H chain	Trp52Ala	1.23	−4.78	*1.25*	3.21	4.89	5.12
	1Vfb-ligand	Val120Ala	0.90	*1.14*	*0.84*	*0.73*	*0.70*	*0.66*
		Ile124Ala	1.20	−1.23	*0.56*	*1.12*	*1.33*	*1.54*
Polar residues	1F47	Tyr3Ala	0.86	−1.63	*1.98*	3.20	3.83	4.55
		Thr25Ala	0.24	*−0.32*	*0.03*	*0.31*	*0.39*	*0.44*
		Asn35Ala	2.40	−2.67	−0.40	*1.19*	*1.22*	*1.28*
		Thr44Ala	2.00	*1.40*	*1.94*	*2.28*	*2.38*	*2.45*
		Tyr45Ala						
	1Vfb-L chain	Ser93Ala	0.11	*−1.06*	*−0.44*	*−0.22*	*−0.10*	*−0.04*
		Tyr32Ala	1.30	−7.34	−0.56	*1.70*	*1.78*	3.45
		Tyr49Ala	0.80	−4.49	−1.36	*−0.32*	*0.21*	*0.51*
		Tyr50Ala	0.40	−7.94	−1.31	*0.91*	*1.05*	2.68
		Thr53Ala	−0.23	−2.09	*−0.67*	*−0.19*	*0.01*	*0.10*
	1Vfb-H chain	Thr30Ala	0.09	*0.53*	*0.35*	*0.29*	*0.26*	*0.23*
		Tyr32Ala	0.50	−6.53	*0.43*	2.75	3.91	4.59
		Asn56Ala	0.20	*0.39*	*0.45*	*0.47*	*0.49*	*0.49*
		Tyr101Ala	>4.0	−12.06	−0.29	*3.61*	*5.54*	*6.70*
	1Vfb-ligand	Asn19Ala	0.40	−5.18	−1.15	*0.21*	*0.91*	*1.31*
		Tyr23Ala	0.80	−1.81	*1.39*	2.45	2.49	2.50
		Ser24Ala	0.70	7.27	3.50	2.23	*1.60*	*1.21*
		Thr118Ala	0.80	*−0.06*	*0.27*	*1.07*	*1.08*	*1.09*
		Gln121Ala	2.90	−4.06	*1.60*	*3.49*	4.46	5.03
Charged residues	1F47	Asp2Ala	0.69	*−0.32*	*0.00*	*0.10*	*0.16*	*0.18*
		Asp5Ala	1.73	−1.21	−0.86	−0.72	−0.64	−0.40
		Arg11Ala	0	*0.72*	*0.97*	*1.04*	*1.08*	*1.08*
		Lys12Ala	0	*0.98*	*1.00*	*1.02*	*1.01*	*1.03*
	1Fcc	Glu27Ala	>4.90	*40.69*	*21.06*	*14.50*	*10.58*	*9.15*
		Lys28Ala	1.30	−8.21	−1.79	*0.48*	*1.53*	*2.14*
		Lys31Ala	3.50	−7.13	1.50	*2.79*	*4.17*	5.92
		Asp40Ala	0.30	−2.65	*−0.74*	*0.08*	*0.30*	*0.70*
		Glu42Ala	0.40	*0.84*	*0.87*	*0.60*	*0.40*	*0.32*
	1Vfb-L chain	His30Ala	0.80	−2.40	*0.20*	*1.59*	*2.10*	2.39
	1Vfb-H chain	Asp58Ala	−0.20	*−0.14*	*0.55*	*0.79*	*0.93*	*1.00*
		Glu98Ala	1.10	*0.80*	*1.06*	*1.12*	*1.31*	*1.37*
		Arg99Ala	0.47	−5.31	−2.35	−1.12	−0.82	*−0.51*
		Asp100Ala	3.10	10.65	7.09	5.83	5.20	4.79
	1Vfb-ligand	Asp18Ala	0.30	4.47	2.79	2.43	1.92	1.83
		Lys116Ala	0.70	3.67	2.50	2.10	*1.62*	*1.50*
		Asp119Ala	1.00	*0.82*	*1.53*	*1.78*	*1.92*	*1.99*
		Arg125Ala	1.80	−6.54	−0.77	*1.11*	*2.03*	*2.64*
% success rate				36	64	*69*	*71*	*62*

The units of free energies are kCal/mol. Numbers in italics are the correct calculated free energy binding values

the results are closer to the real value with the use of this type of protocol, and much higher resulting success rates. Even though by using water caps we have obtained results that allowed us to conclude that the success rate decreases with the number of trajectories, we have tried with a water box the three protocols to some of the mutant complexes (Asp5Ala, Leu10Ala), and we present the results in Table 4. The accuracy of the results in relation to the use of a water cap got better, but the cost of these protocols turns the alanine scanning mutagenesis of the all protein–protein interface difficult to perform. As we were trying to achieve a methodological approach not too laborious and with high success rates, we decided based on these multiple evidences that the use of a single mutation protocol would be the best alternative.

Thus, we have extended this study to the three complexes, and a description of the systems for which we have applied only the "one simulation protocol" in a water box is shown in Table 5. The rmsd from the X-ray crystal structure of Cα atoms of the complex for the other two complexes (1FCC and 1VFB) as a function of time is represented in Fig. 5. In Table 6 are the results from the alanine scanning mutagenesis study for this protocol for all three complexes. Recalling

Table 8 Results of the Methodological Approach for Computational Alanine Screening Mutagenesis with a higher success rate

Residue type	Protein	Mutation	% burial	$\Delta\Delta G_{exp}$	$\Delta\Delta G$
Non-polar residues	1F47	Leu4Ala	91.3	0.92	*1.01*
		Ile6Ala	97.5	2.50	*2.42*
		Phe9Ala	95.7	2.44	*2.48*
		Leu10Ala	94.2	2.29	*2.73*
	1Fcc	Trp43Ala	97.6	3.80	1.12
	1Vfb-L chain	Trp92Ala	90.3	1.71	*2.37*
	1Vfb-H chain	Trp52Ala	96.3	1.23	*1.25*
	1Vfb-ligand	Val120Ala	98.6	0.90	*0.84*
		Ile124Ala	98.5	1.20	*0.56*
Polar residues	1F47	Tyr3Ala	81.1	0.86	3.20
	1Fcc	Thr25Ala	90.3	0.24	*0.31*
		Asn35Ala	88.5	2.40	*1.19*
		Thr44Ala	76.4	2.00	*2.28*
		Tyr45Ala	87.2		
	1Vfb-L chain	Ser93Ala	84.5	0.11	*−0.22*
		Tyr32Ala	98.0	1.30	*1.70*
		Tyr49Ala	93.0	0.80	*−0.32*
		Tyr50Ala	91.1	0.40	*0.91*
		Thr53Ala	86.4	−0.23	*−0.19*
	1Vfb-H chain	Thr30Ala	70.7	0.09	*0.29*
		Tyr32Ala	91.5	0.50	2.75
		Asn56Ala	65.7	0.20	*0.47*
		Tyr101Ala	97.3	>4.0	*3.61*
	1Vfb-ligand	Asn19Ala	89.3	0.40	*0.21*
		Tyr23Ala	97.6	0.80	2.45
		Ser24Ala	98.4	0.70	2.23
		Thr118Ala	86.9	0.80	*1.07*
		Gln121Ala	99.9	2.90	*3.49*
Charged residues	1F47	Asp2Ala	63.0	0.69	0.16
		Asp5Ala	78.8	1.73	−0.64
		Arg11Ala	54.3	0	*1.08*
		Lys12Ala	55.0	0	*1.01*
	1Fcc	Glu27Ala	99.8	>4.90	*10.58*
		Lys28Ala	97.2	1.30	*1.53*
		Lys31Ala	99.3	3.50	*4.17*
		Asp40Ala	72.7	0.30	*0.30*
		Glu42Ala	67.0	0.40	*0.40*
	1Vfb-L chain	His30Ala	83.5	0.80	*2.10*
	1Vfb-H chain	Asp58Ala	83.5	−0.20	*0.93*
		Glu98Ala	98.0	1.10	*1.31*
		Arg99Ala	82.3	0.47	*−0.82*
		Asp100Ala	87.5	3.10	5.20
	1Vfb-ligand	Asp18Ala	87.8	0.30	1.92
		Lys116Ala	83.6	0.70	*1.62*
		Asp119Ala	86.1	1.00	*1.92*
		Arg125Ala	80.3	1.80	*2.03*
% success rate					*82*

The units of free energies are kCal/mol. Numbers in italics the correct calculated free energy binding values

that we want to achieve a computational method fast and accurate that has a good agreement with the experimental we decided to try other different approaches. The first alternative tried was to substitute the discrete water molecules by a continuum representation of the solvent. The corresponding MD simulations were performed using the Generalized Born solvation model. Figure 6 shows the time series of rmsd from the X-ray crystal structure of Cα atoms of the complex for all simulations, and in Table 7 are collected the alanine scanning mutagenesis results. These prove to be better than those obtained with the explicit solvent, as can be seen by inspection of Tables 6 and 7.

Taking into account the extensive computational time of explicit solvent representation, and the fact that the success rate is smaller, we have opted to use the Generalized Born model to obtain a good set of structures. The preference for this type of solvent representation can be justified by several reasons: the smaller simulation time necessary compared to the explicit solvent methods; the more complete exploration of the conformational space due to the lack of the viscous damping forces of the water; the reduced lengthy equilibration of water compared to the explicit water simulation; and an easier interpretation of the results since the water degrees of freedom are absent [35]. The continuum solvent is used to

Table 9 Success rates for the minimization protocols tested

Success rate	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta G$
	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	$\varepsilon = 4$	$\varepsilon = 5$	$\varepsilon = 10$	$\varepsilon = 15$
Method 3	44%	49%	49%	51%	47%	49%	53%
Method 4	27%	49%	47%	51%	49%	49%	51%
Method 5	38%	49%	47%	47%	55%	60%	60%
Method 6	40%	55%	67%	62%	58%	58%	55%

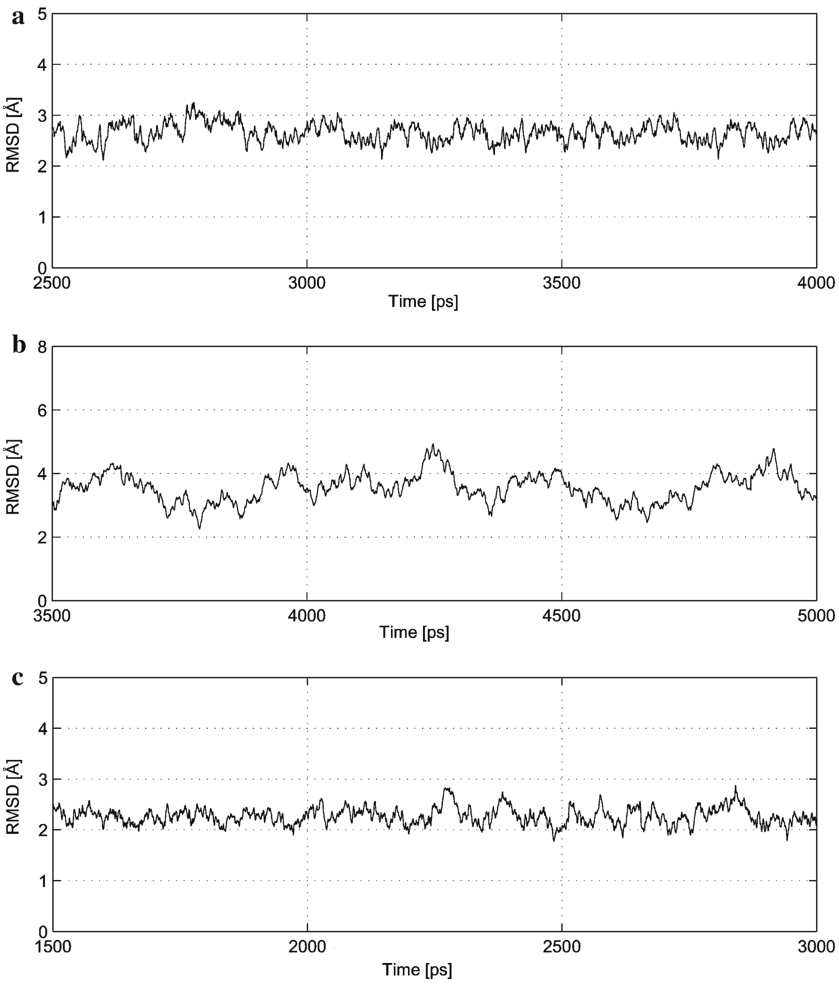


Fig. 6 **a** RMSD plots for the protein backbone of the complex formed between the bacterial cell-division protein ZipA and the FtsZ fragment in an implicit solvation model relative to its initial structure; **b** RMSD plots for the protein backbone of the complex in a water box formed between the human immunoglobulin IgG and the C2 fragment of streptococcal protein G in an implicit solvation model relative to its initial structure; **c** RMSD plots for the protein backbone of the complex formed between an immunoglobulin and a hen egg lysozyme in an implicit solvation model relative to its initial structure (for the last 1.5 ns of the MD simulation)

calculate the $\Delta G_{\text{solvation}}$ value, and therefore it is coherent to use the same method to generate the MD trajectories.

At this stage we have opted to use only one trajectory for the computational energy analyses. Hence, it is important to highlight that side chain reorientation and dipolar reorganization arising from conformational transitions upon binding is not included explicitly in the formalism. Although the conformational reorganization after alanine mutagenesis is not explicitly taken into account in the formalism, it can be mimicked by a single factor: the scaling of the internal dielectric constant to large values when larger re-organizations are expected. However, this implementation is not trivial. From the observation of all the results obtained in this study (Tables 2, 3, 4, 5, 7, 8), we have concluded that the internal dielectric constant should not be implemented as a homogeneous constant, an idea that makes sense if we remember that protein environments are highly inhomogeneous. By using a three internal dielectric constant set, exclusively characteristic of the mutated amino acid (2 for the non-polar amino acids, 3 for the polar residues and 4 for the charged amino acids plus

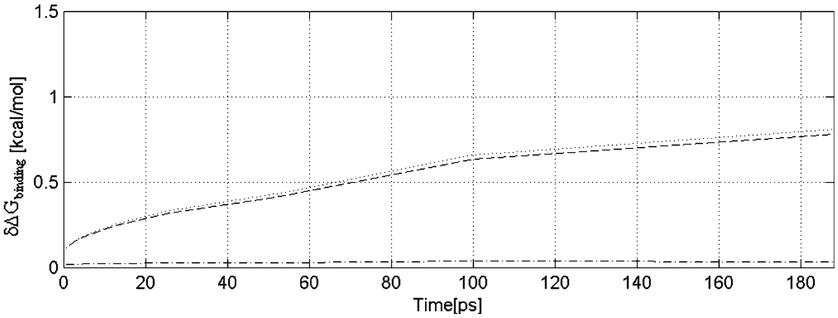


Fig. 7 The standard deviation of the mean value of the $\Delta G_{\text{binding}}$ as a function of the size of the blocks used

histidine), it was possible to obtain an excellent agreement with the experimental results for the $\Delta\Delta G_{\text{binding}}$ values. At this point, we have achieved a computational method with a success rate of 82% as it can be observed in Table 9. If we consider a deviation of ±1.4 kCal/mol from the experimental value as an accurate result, we have an overall success rate of 82%, a 82% success rate for the null-spots, a 78% achievement of the correct relative binding free energy of the warm-spots, and we have correctly detected both the hot spots present. To use a 4.0 kCal/mol cutoff value to define hot spots reduces the set of hot spot residues to the ones that are really important for complex binding. However, the problem with the hotspots (within the $>$ 4 kCal/mol definition) is that they are rare, and the absolute value for $\Delta\Delta G_{\text{bind}}$ in hotspots is usually not available. This way, other cutoff values such as 2.0 kCal/mol are used to define hot spots to make possible a statistical analysis. Although the data set used here has only two hot spots it is important to highlight that these residues were correctly identified especially because other computational methods tend to have a lower success rate for these kind of residues. Nevertheless, we can also highlight that this method has a high success rate of 82% (9/11) for the warm and hot spots residues (residues with a $\Delta\Delta G_{\text{binding}}$ higher than 2.0 kCal/mol).

It is important to ensure that the averaging over configurations is uncorrelated, which it is not the case for nearby points in an MD trajectory. Therefore we have performed a coarse-grained or two-stage sampling to calculate the statistical inefficiency or the correlation time. This procedure consists in dividing the ensemble of structures obtained with the MD simulation in n_b blocks, each one containing a subset of t_b structures, and calculating a block average value for the desired property, using all the points in each block. Subsequently we use these values for averaging over blocks and obtain the ensemble average value of property A. The block averages are given by:

$$\langle A\rangle_b = \frac{1}{t_b}\sum_{i=b_{\text{start}}}^{b_{\text{end}}} A_{\text{i}}, \tag{4}$$

where $b_{\text{start}} = (b-1)n_b + 1$, $b_{\text{end}} = bt_{\text{b}}$, and b can assume values from 1 to n_b. The mean square deviation of averages taken over the blocks is:

$$\left\langle \delta^2 A\right\rangle_b = \frac{1}{n_b}\sum_{b=1}^{n_b}\left(\langle A\rangle_b - \langle A\rangle\right)^2 \tag{5}$$

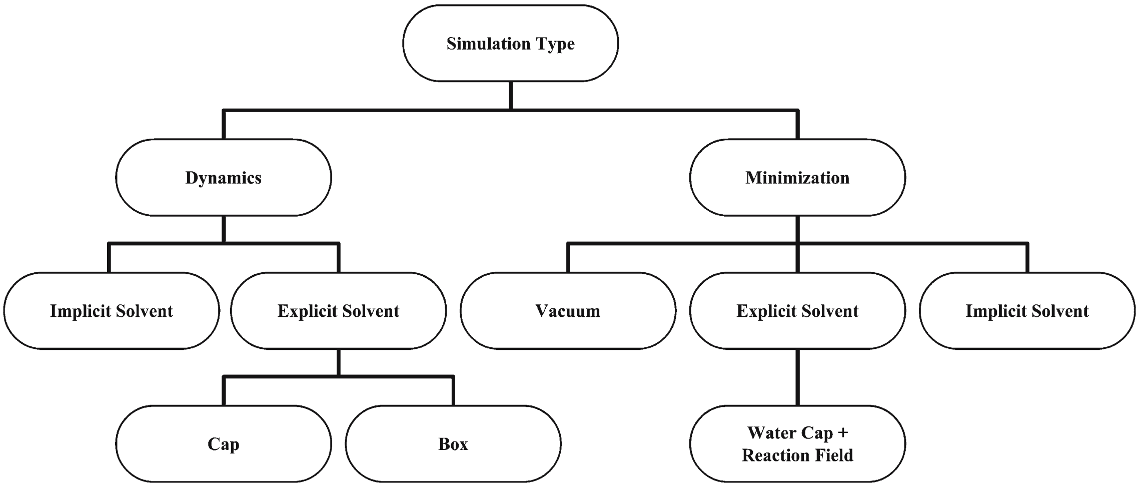
The standard deviation in the mean values is proportional to the inverse of the square root of the number of snapshots. In Fig. 7 we have plotted the standard deviation of the mean values for the $\Delta G_{\text{binding}-\text{molecule}}$for the 1FCC-Thr25Ala mutant and for the 1FCC wild-type as well as the$\Delta\Delta G_{\text{binding}}$ as a function of the block size.

We can observe in Fig. 7 that the values for $\Delta G_{\text{binding}}$ have reached a plateau. This means that the average energies calculated from blocks of data in the plateau region are uncorrelated. The distance between uncorrelated blocks has the dimension of time and is called the correlation time. As it can be seen, $\Delta\Delta G_{\text{binding}}$ has a lower correlation time than $\Delta G_{\text{binding}-\text{molecule}}$, which once more emphasises that error cancellation is of great importance for the enhancement of the success rate of a computational alanine scanning mutagenesis study. From this graphic it can also be perceived that in ergodic conditions the standard deviation in the ensemble average is lower than 1.0 kCal/mol, and therefore the 500 ps sampling used is adequate, not being necessary to enlarge the sampling.

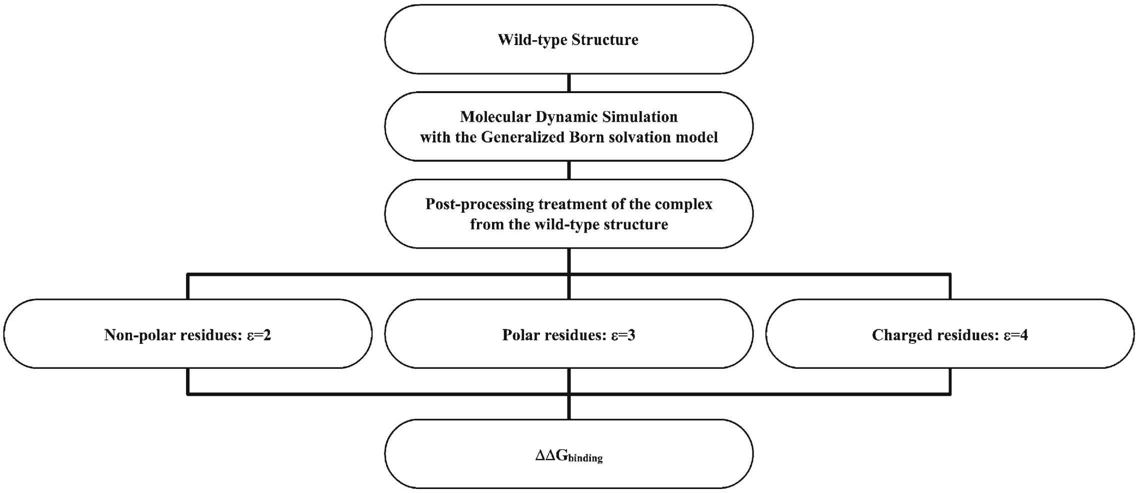
2.2 Energy minimization

The minimization algorithm should give the minimum with the lowest energy, the global energy minimum. However, most minimization algorithms can only go downhill on the energy surface and so they can only locate the minimum that is nearest to the starting point. Moreover, averaging between thermally accessible minima can be important to obtain an accurate value for $\Delta\Delta G_{\text{binding}}$, as the potential energy surface of these systems are very complex, with multiple minima populated at room temperature. Even though we have tried four types of minimization protocols (described in the Sect. 1) with success rates shown in Table 9 and the detailed results in supplementary information. The minimization approach is successful, giving results higher than the usually reported in the literature [22].

It is important to analyse the different rms from the minimized structures in relation to their respective X-ray



Scheme 2 Resume of the complete study made to achieve a fast and simple computational alanine screening mutagenesis methodological approach


Wild-type Structure
Molecular Dynamic Simulation with the Generalized Born solvation model
Post-processing treatment of the complex from the wild-type structure
Non-polar residues: ε=2
Polar residues: ε=3
Charged residues: ε=4
ΔΔGbinding

Scheme 3 Resume of the methodological approach for computational alanine screening mutagenesis

crystallographic structures. The rms values for the three complexes in method 3 are 0.95, 0.41 and 0.79 Å. In method 4 they are 0.94, 0.47 and 0.73 Å. In method 5 we have as rms values 0.22, 0.50 and 0.32 Å. Finally in method 6 we have rms of 0.29, 0.38 and 0.26 Å, respectively. Method 5 and method 6 present the lower rms values as well as the higher success rates. This success rate is especially higher in method 6 where a reaction-field force was applied. In this method the solute intramolecular interactions are computed by the usual molecular mechanics force field terms, while the solute–solvent and solvent–solvent interactions are computed by a mean-field approximation with the PB electrostatic theory. From all the minimization protocols this one seems the most capable to reproduce the free energy binding differences.

In Scheme 2 is presented a resume of this computational alanine scanning study. Initially we have tried different solvent representations (explicit or implicit), and different internal dielectric constants for the protein. Subsequently, we have tried protocols with a different number of dynamics simulation trajectories. Finally, we have decided to try a less expensive method, a minimization approach.

In summary, after a complete study we have achieved a simple, fast computational methodological approach that is summarized in Scheme 3. This method has a low computational cost and can be applied prior to an experimental investigation to a wide range of proteins providing important information concerning protein–protein interface amino acid residues.

3 Conclusion

Alanine-scanning mutagenesis of protein–protein interfacial residues is an important method to determine hot spots allowing for the systematic analysis of individual residues, and the understanding of the physical and chemical properties of protein–protein interfaces of complexes to determine their unique features.

As experimental determination is very laborious it is important to achieve a fast and accurate computational method that can provide quantitative estimates, but also, and mainly, that can enhance our understanding of the nature of complex formation in terms of the biophysical features of the process.

We have studied the influence of the variation of different parameters and simulations, such as the internal dielectric constant, the solvent representation, and the number of trajectories on the accuracy of the free energy binding differences.

As a result, based on the MM–PBSA method [17], we have achieved a methodological approach that uses the molecular mechanics AMBER force field and a continuum solvation approach with different internal dielectric constant values for different kinds of residues with an overall performance rate of 82%, an 82% success rate for the null-spots, 78% achievement of the correct relative binding free energy of the warm-spots, and an 82% rate of success for the warm and hot-spots within the database.

References

1. Arkin MR, Wells AJ (2004) Drug Discov 3:301–317
2. Sharma SK, Ramsey TM, Bair KW (2002) Curr Med Chem Anticancer Agents 2:311–330
3. Bogan AA, Thorn KS (1998) J Mol Biol 280:1–9
4. Delano WL, Ultsch MH, de Vos AM, Wells JA (2000) Science 287:1279–1283
5. Pons J, Rajpal A, Kirsch J (1999) Protein Sci 8:958–968
6. Keskin O, Ma B, Nussinov R (2005) J Mol Biol 345:1281–1294
7. Arkin MR, Randal M, DeLano WL, Hyde J, Luong TN, Oslob JD, Raphael DR, Taylor L, Wang J, McDowell RS, Wells JA, Braisted A (2003) Proc Natl Acad Sci USA 100:1603–1608
8. Gao Y, Wang R, Lia L (2004) J Mol Model 10:44–54
9. Lopez MA, Kollman PA (1993) Protein Sci 2:1975–1986
10. Kortemme T, Baker D (2002) Proc Natl Acad Sci USA 99:14116–14121
11. Kortemme T, Kim DE, Baker D (2004) Sci STKE 219:12–15
12. Schapira M, Totrov M, Abagyan RJ (1999) Mol Recognit 12:177–190
13. Aqvist J, Medina C, Samuelsson JE (1994) Protein Eng 7:385–391
14. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Freer ST, Rose PM (2002) Proteins 48:539–557
15. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE III (2000) Acc Chem Res 33:889–897
16. Wang W, Donini O, Reyes CM, Kollman PA (2002) Annu Rev Biophys Biomol Struct 30:211–243
17. Massova I, Kollman PA (1999) J Am Chem Soc 121:8133–8143
18. Wang J, Morin P, Wang W, Kollman PA (2001) J Am Chem Soc 123:5221–5230
19. Wang W, Kollman PA (2002) J Mol Biol 303:567–582
20. Reyes CM, Kollman PA (2000) J Mol Biol 295:1–6
21. Huo S, Massova I, Kollman PA (2002) J Comput Chem 23:15–27
22. Mosyak L, Zhang Y, Glasfeld E, Haney S, Stahl M, Seehra J, Somers WS (2000) EMBO J 19:3179–3191
23. Sauer-Eriksson AE, Kleywegt GJ, Uhlen M, Jones TA (1995) Structure 3:265–278
24. Bhat TN, Bentley GA, Boulot G, Greene MI, Tello D, Dall'Acqua W, Souchon H, Schwarz FP, Mariuzza RA, Poljak RJ (1994) Proc Natl Acad Sci USA 9:1089–1093
25. Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, Wang B, Pearlman DA, Crowley M, Brozell S, Tsui V, Gohlke H, Mongan J, Hornak V, Cui G, Beroza P, Schafmeister C, Caldwell JW, Ross WS, Kollman PA (2004) AMBER 8 University of California, San Francisco
26. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) J Am Chem Soc 117:5179–5197
27. Jorgensen WL, Chandrasekhar J, Madura J, Impey RW, Klein ML (1983) J Chem Phys 79:926–935
28. Ryckaert JP, Ciccotti G, Berendsen HJ (1977) J Comput Phys 23:327–335
29. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) J Chem Phys 81:3684–3690
30. Case DA, Pearlman DA, Caldwell JW, Cheatham III TE, Ross WS, Simmerling CL, Darden TA, Merz KM, Stanton RV, Cheng AL, Vincent JJ, Crowley M, Tsui V, Radmer R J, Duan Y, Pitera J, Massova I, Seibel GL, Singh UC, Weiner PK, Kollman PA (1999) AMBER 6 University of California, San Francisco
31. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) J Chem Phys 103:8577–8593
32. Pastor RW, Brooks BR, Szabo A (1988) Mol Phys 65:1409–1419
33. Loncharich RJ, Brooks BR, Pastor RW (1992) Biopolymers 32:523–535
34. Izaguirre JA, Catarello DP, Wozniak JM, Skeel RD (2001) J Chem Phys 114:2090–2098
35. Tsui V, Case DA (2001) Biopolymers (Nucl Acid Sci) 56:275–291
36. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B (2002) J Comput Chem 23:128–137
37. Rocchia W, Alexov E, Honig B (2001) J Phys Chem B 105:6507–6514
38. Sitkoff D, Sharp KA, Honig BJ (1994) Phys Chem 98:1978
39. Moreira IS, Fernandes PA, Ramos MJ (2005) J Mol Struct (Theochem) 729:11–18
40. Connolly ML (1983) J Appl Cryst 16:548–558
41. Gao Y, Wang R, Lia L (2004) J Mol Model 10:44–54
42. Xia B, Tsui V, Case DA, Dyson J, Wright PE (2002) J Biomol NMR 22:317–331
43. Sheinerman FB, Norel R, Honig B (2000) Curr Opin Struct Biol 10:153–159
44. Schutz CN, Warshel A (2001) Proteins 44:400–417
45. Hsieh MJ, Luo R (2004) Proteins 56:475–486