Computational Alanine Scanning Mutagenesis—An Improved Methodological Approach

IRINA S. MOREIRA, PEDRO A. FERNANDES, MARIA J. RAMOS
*REQUIMTE/Departamento de Química, Faculdade de Ciências da Universidade do Porto,
Rua do Campo Alegre 687, 4169-007 Porto, Portugal*

*Received 31 August 2005; Accepted 21 April 2006
DOI 10.1002/jcc.20566
Published online 28 December 2006 in Wiley InterScience (www.interscience.wiley.com).*

Abstract: Alanine scanning mutagenesis of protein–protein interfacial residues can be applied to a wide variety of protein complexes to understand the structural and energetic characteristics of the hot-spots. Binding free energies have been estimated with reasonable accuracy with empirical methods, such as Molecular Mechanics/Poisson-Boltzmann surface area (MM-PBSA), and with more rigorous computational approaches like Free Energy Perturbation (FEP) and Thermodynamic Integration (TI). The main objective of this work is the development of an improved methodological approach, with less computational cost, that predicts accurately differences in binding free energies between the wild-type and alanine mutated complexes ($\Delta\Delta G_{\text{binding}}$). The method was applied to three complexes, and a mean unsigned error of 0.80 kcal/mol was obtained in a set of 46 mutations. The computational method presented here achieved an overall success rate of 80% and an 82% success rate in residues for which alanine mutation causes an increase in the binding free energy > 2.0 kcal/mol (warm- and hot-spots). This fully atomistic computational methodological approach consists in a computational Molecular Dynamics simulation protocol performed in a continuum medium using the Generalized Born model. A set of three different internal dielectric constants, to mimic the different degree of relaxation of the interface when different types of amino acids are mutated for alanine, have to be used for the proteins, depending on the type of amino acid that is mutated. This method permits a systematic scanning mutagenesis of protein–protein interfaces and it is capable of anticipating the experimental results of mutagenesis, thus guiding new experimental investigations.

© 2006 Wiley Periodicals, Inc. J Comput Chem 28: 644–654, 2007

Key words: alanine scanning mutagenesis; computational mutagenesis; MM-PBSA; hot-spot; binding free energy; protein–protein interfaces; molecular mechanics; Poisson-Boltzmann; Generalized Born model

Introduction

A network of protein–protein interactions forms the basis of most cellular function. Understanding of protein–protein interaction is essential to the understanding of molecular recognition and the physical basis of affinity. It also allows the design of new protein–protein interactions, permits one to engineer new functions and adjust cellular behavior in a predictive manner, and enables the rational design of new therapeutic agents.[1,2]

Protein association has been shown to be sensitive to mutational events. Even though protein interfaces are large and complex, single residues, being responsible for the majority of the interaction energy, can still contribute significantly to the binding free energy.[3] The functional epitope, defined by the contact residues that make energetic contributions to binding, constitutes only a small fraction of the structural epitope, defined by the protein amino acid residues in contact with a ligand. One of the biological facts that can explain this situation is the inaccessibility to site-directed mutagenesis of the protein backbone, an important contributor to interfaces as it represents on average about one-fifth of the interface area and contributes nearly two-thirds of the hydrogen bonds.[4] Another justification for the reduced size of the functional epitope is that the atoms that remain partly accessible to solvent constitute three-quarters of the interfaces area, and when deleted, as in an alanine mutant, they can be replaced by water molecules at much less cost than fully buried atoms. It has been demonstrated that inaccessibility to the solvent is a necessary condition for a residue to constitute a binding hot-spot.[4–6] As a result, the critical components in a functional epitope, the hot-spots, have been defined as those sites where alanine mutations cause a significant increase in the binding free energy of at least 4.0 kcal/mol,[7] even though lower values are used for statistical analyses.[7] Warm-spots are those with binding free energy differences between 2.0 and 4.0 kcal/mol and null-spots are the residues with binding free energy differences lower than 2.0 kcal/mol. Those hot-spots tend

Correspondence to: M. J. Ramos; e-mail: mjramos@fc.up.pt

to be correlated with conserved residues at specific locations, and are enriched with tryptophan, tyrosine, and arginine.[8] The binding hot-spots have been detected in numerous protein–protein interfaces but are not randomly spread along them. Instead, they tend to be in dense clusters forming a network of interactions, and contribute cooperatively to the stability of the complex.[8,9]

Alanine-scanning mutagenesis of protein–protein interfacial residues, combined with structural and thermodynamic studies have enabled the discovery of energetically important determinants of specificity at intermolecular interfaces that are critical in determining binding affinity.[10] Unraveling hot-spots in binding interfaces continues to stimulate interest, since reliable prediction of key residues in the interface has immediate applications in protein engineering and it is an attractive alternative therapy for many diseases (Structure Based Drug Design).[11]

In recent years, a number of methods to correctly identify the hot-spots have been developed. Free Energy Perturbation (FEP) and Thermodynamic Integration (TI) yield rigorous and accurate free energy differences but these methods are implemented numerically; thus sufficient statistical sampling must be carried out which makes them extremely time consuming and prevents them from being commonly used in structure-based design.[12] Simple physical models,[13,14] empirical methods,[15] linear interaction energy methods,[16] and Monte Carlo methods[17] have been proposed to identify the residues contributing significantly to the stability of protein associations. A computer algorithm, FOLDEF (for FOLD-X energy function) tested on monomeric proteins and protein complexes was also developed.[18] This method presents an error below 0.81 kcal/mol for 70% of the mutants, and it is of significant interest for the improvement of structure prediction methods, in particular in the field of *ab initio* prediction.[18]

An all-atom method has been developed to probe protein–protein interactions by calculating free energies combining molecular mechanics and continuum solvent and was named MM-PBSA (Molecular Mechanics/Poisson-Boltzmann Surface Area) methodology.[19–25] In this approach, the free energy of a molecule can be written as:

$$G_{\text{molecule}} = E_{\text{internal}} + E_{\text{electrostatic}} + E_{\text{vdw}} + G_{\text{polar solvation}} + G_{\text{nonpolar solvation}} - \text{TS} \qquad (1)$$

with the first term corresponding to the internal energy of the solute (bond, angle and dihedral), the second and the third terms to the electrostatic and the van der Waals interactions respectively, and the last three terms being the polar free energy of solvation, the nonpolar free energy of solvation and the entropic contribution for the solute free energy. To simulate solvation effects on biological macromolecules, both explicit and implicit solvent models have been developed offering a varying degree of microscopic detail in exchange for computational efficiency.[26] Explicit solvation representation, where a biological molecule is embedded in a large number of solvent molecules, is detailed but it is in general very time consuming and computationally demanding because periodic boundary conditions are usually necessary.[27] Alternatively, the discrete water molecules can be replaced by an infinite continuum medium with the dielectric properties of water.[28] The implicit solvent model has become an increasingly popular technique because it is an approach computationally much more affordable, while still providing a reasonable description of an aqueous solvent environment. Moreover, discrete methods do not allow the calculation of free energy of polar solvation ($\Delta G_{\text{solvation}}$), which can only be estimated with continuum models.

These algorithms of varied complexity can be divided essentially into two main types: empirical functions or simple physical methods that use experimentally calibrated knowledge-based simplified models to evaluate the binding free energy, and versatile/universal fully atomistic methods that estimate the free energy of association or changes in the binding free energies as a result of mutating the residues of the interacting molecules based only in the respective Hamiltonians.[19] The two types of methods have each specific advantages and limitations. An equilibrium must be achieved between the use of simple algorithms that permit fast calculations and the inclusion, conservation, and consideration of the important atomic detail of biomolecules.[2] Consequently, when deciding on the computational approach for predicting the binding free energies, it is important to foresee the computational time required, without forgetting that sometimes it is affordable and advantageous to carry out more accurate time-consuming calculations because an atomic-detail description of biomolecules is often important in elucidating their structures and functions.[3] In summary, although it is correct that empirical models can presently be as accurate as fully atomistic models, and computationally faster, they suffer from a number of limitations that make them inadequate to a significant number of biological complexes,[14] and more limited in terms of further development. In this context, it still makes sense to continue to study and improve fully atomistic models, which give more detailed information, have a broader application field, and can be systematically improved by inclusion of more exact Hamiltonians and longer simulation times. This is trivial from a methodological point of view, limited only by the actual state of computer technology.

Since the last decade, an effort has been made in achieving an accurate, predictive methodology for alanine scanning mutagenesis, capable of reproducing the experimental mutagenesis values. Until now the results from atomic level methodologies have neither presented the precision nor the accuracy to achieve the "chemical accuracy," which is ~1 kcal/mol. Chemical accuracy is traditionally used as a standard for good agreement between theoretical and experimental results since it is sufficient to describe van der Waals interactions, the weakest interaction considered to affect chemistry.

The success rates have been rather modest so far. The alanine mutation of charged amino acids (aspartic acid, glutamic acid, lysine, arginine, and histidine) generates mainly values in disagreement with the experimental ones, and the computational time involved is much too high to permit a systematic mutagenesis of protein–protein interfaces.[19–25] However, the experimental systematic scanning mutagenesis of protein–protein interfaces even though more exact is also more difficult to perform because it is a very expensive and time consuming methodology. Thus, a computational approach represents an excellent compromise between accuracy and time necessary to reach the binding free energy differences ($\Delta\Delta G_{\text{binding}}$).

The main objective of this work is the development of an improved methodological approach, with less computational cost and high success rate that reproduces the quantitative free energy differences obtained from experimental mutagenesis procedures. This computational method is transferable to any macromolecular complex and is a predictive model capable of anticipating the experimental results of mutagenesis, thus guiding new experimental investigations.

Computational Details

Model Setup

The structures considered here are three protein–protein complexes. The crystallographic structures with a resolution of 1.95 Å, 3.50 Å, and 1.80 Å respectively were taken from the RCSB Protein Data Bank with PDB entries: 1F47,[29] 1Fcc[30] and 1Vfb.[31] They are complexes that mediate bacterial cell division (1F47), a human immunoglobulin IgG complexed with the C2 fragment of streptococcal protein G (1Fcc), and an immunoglobulin complexed with a Hen Egg lysozyme (1Vfb).

All molecular mechanics simulations presented in this work were performed using the sander module, implemented in the Amber8[32] simulations package, with the *Cornell* force field.[33] In the molecular simulations the solvent was modeled through a modified Generalized Born solvation model.[34] The system was first minimized by 1000 steps of steepest decent followed by 1000 steps of conjugated gradient to release the bad contacts in the crystallographic structure. Subsequently, molecular dynamics (MD) simulations were performed starting from the minimized structures. Bond lengths involving hydrogens were constrained using the SHAKE algorithm.[35] The equations of motion were integrated with a 2-fs time-step and the nonbonded interactions were truncated with a 16 Å cutoff. The temperature of the system was regulated by the Langevin thermostat.[36–38] The total simulation time was 3000 ps for the 1fcc and 1vfb complexes, and 4000 ps for the 1f47 complex.

The MM-PBSA script[25] implemented in Amber8[32] 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 $C\gamma$ with a hydrogen atom, and setting the $C\beta$—H bond direction to that of the former $C\beta$—$C\gamma$. For the binding free energy calculations, a total of 25 snapshots of the complexes were extracted, one every 20 ps in the last 500 ps of the run.

Alanine Scanning Mutagenesis

In this paper we present a new and improved methodological approach, based in the MM-PBSA protocol. In MM-PBSA, the complexation free energy is calculated using the following thermodynamic cycle:

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

Thermodynamic cycle used to calculate the complexation free energy.

Here, ΔG_{gas} is the interaction free energy between the ligand and the receptor in the gas phase and $\Delta G^{lig}{}_{solv}$, $\Delta G^{rec}{}_{solv}$, and $\Delta G^{cpx}{}_{solv}$ are the solvation free energies of the ligand, the receptor, and the complex respectively. The binding free energy difference between the mutant and wild type complexes is defined as:

$$\Delta\Delta G_{\text{binding}} = \Delta G_{\text{binding-mutant}} - \Delta G_{\text{binding-wild type}} \tag{2}$$

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

$$\Delta G_{\text{binding-molecule}} = G_{\text{complex}} - (G_{\text{receptor}} + G_{\text{ligand}}) \tag{3}$$

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 salvation, and the entropic contribution for the molecule free energy, as given by eq. (1). The first three terms were calculated using the *Cornell* force field[33] with no cutoff. The electrostatic solvation free energy was calculated by solving the Poisson-Boltzmann equation with the software Delphi v.4.[39,40] In this continuum method, the protein is modeled 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 <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 is 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[41] vdW radii parameter set. Standard parm94 charges[33] were used in order to be consistent with the energetic of the simulations we are analyzing. These parameters have been shown in an earlier work to constitute a good compromise between accuracy and computing time.[42] For the energy calculations three internal dielectric constant values, exclusively characteristic of the mutated amino acid, were used: 2 for the nonpolar amino acids, 3 for the polar residues and 4 for the charged amino acids The nonpolar contribution to solvation free energy due to van der Waals interactions between the solute and the solvent and cavity formation was modeled as a term that is dependent on the solvent accessible surface area of the molecule. It was estimated using an empirical relation: $\Delta E_{\text{nonpolar}} = o \cdot A + b$, 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.[43] $o\cdot$ and b are empirical constants and the values used were 0.00542 kcal/(Å^2 mol) and 0.92 kcal/mol 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 $\Delta\Delta G_{\text{binding}}$ is negligible.[25]

The apparent dissociation constant for each alanine mutant was estimated from the concentration required for 50% inhibition (IC_{50}). The IC_{50} values were used to calculate the experimental binding free energies differences using the following relationship:

$\Delta\Delta G_{\text{binding}} = RT\ln(\text{IC}_{50}\ \text{mutant}/\text{IC}_{50}\ \text{wild-type})$, where R is the ideal gas constant and T is the temperature in K.

Results

Alanine scanning mutagenesis of protein–protein interfacial residues is a very important process for rational drug design.[4,44,45] Therefore, it is necessary to develop a faster and more reliable computational method to predict the binding affinities of the ligands. Some methods have been developed but they do not present the necessary accuracy. The MM-PBSA method by Massova/Kollman is a fully atomistic method that, although not accurate enough in the original implementation, opened the possibility for the development of a new improved methodology.

We began our work by studying the influence of different computational simulation protocols on $\Delta\Delta G_{\text{binding}}$. Initially we have tried different types of simulations (minimization or dynamics), different solvent representations (explicit or implicit), and different internal dielectric constants for the proteins. Subsequently, we have tried protocols with a different number of dynamic simulation trajectories. The first protocol is a "single mutation protocol"[25] consisting in optimizing or running a molecular mechanics simulation with only the wild-type structures and subjecting them to a post-processing treatment to generate the mutant complexes by a simple truncation of the side chain of the residue we wish to mutate, replacing the C_γ with a hydrogen atom. The monomer structures were generated from the structure of the wild-type and the mutant complexes by deletion of the other partner in the protein–protein complex. Consequently, the free energy of the wild-type and mutant monomers and the mutant complexes are calculated without rearrangement of the surrounding environment. We have also tried a "two mutation protocol" based on running two separate trajectories or optimizing the geometry for the wild type and the mutant. The free energy of the monomers is calculated without optimizing or running MD with these structures, and therefore making only a single energy point calculation. Finally, we have tried a "fourth mutation protocol" by optimizing the wild-type, the mutant complex, the wild-type monomer and the mutant monomer. As one of the monomers was not subjected to mutation, a MD simulation was not performed. This monomer has its effect cancelled in eq. (4).

$$\begin{aligned}\Delta\Delta G_{\text{binding}} = {} & (G_{\text{complex-mutant}} - G_{\text{complex-wild type}}) \\ & - (G_{\text{ligand-mutant}} - G_{\text{ligand-wild type}}) \\ & - (G_{\text{receptor-wild type}} - G_{\text{receptor-wild type}}) \qquad (4)\end{aligned}$$

As can be perceived from the first mutation protocol to the last, there is a passage from 1 MD/system to 1 MD/mutation leading to a simulation time proportional to the mutation number and therefore increasing the CPU cost to enormous values. As the dielectric constant of a protein (ε) is not a universal parameter, and values from 1 to 4 (and even higher) are commonly used, we have calculated $\Delta\Delta G_{\text{binding}}$ with dielectric constants from 1 to 5, and with the solvent described by explicit water molecules (as the Massova/Kollman method implies). Our initial objective was to predict binding affinities within an accuracy of 1 order of magnitude, which corresponds to 1.36 kcal/mol at room temperature, and 1.42 kcal/mol at physiological temperature. Subsequently, we have noticed that all residues that obey such criteria have absolute errors slightly smaller, within 1.3 kcal/mol, and introduced this last value as a success rate indicator.

The results given by this methodology were not very satisfactory (overall success rates of 44% using $\varepsilon = 1$, 62% using $\varepsilon = 2$, 60% using $\varepsilon = 3$, 63% using $\varepsilon = 4$, and 58% using $\varepsilon = 5$, lower if we only consider the warm-spots and hot-spots).

Eventually we developed and settled on a new and improved methodological approach, based in the MM-PBSA protocol, which we present here. The three proteins studied have different dimensions and properties and Figures 1a–1c show representations of these complexes highlighting the hot-spots present in each one. From the complex present in Figure 1a to the one present in Figure 1b there is an increase in size (from 159 to 262 residues). The one in Figure 1c was chosen because it has a substantially different architecture, being formed by three distinct chains, a heavy and light chain that constitute the receptor, and a third chain that constitutes the ligand.

We have found that the use of a single trajectory to calculate $\Delta\Delta G_{\text{binding}}$ is conducive to a much better agreement with the experimental data than the use of multiple trajectories. It is supposed that by using a single trajectory, error cancellation will overcome the insufficient sampling of the conformational space.[25,46] Therefore, the use of multiple trajectories is only benefited if very long trajectories can be generated, something that is not feasible presently, even with modern computational resources. Thus, a single trajectory of the wild-type complexes was run and post-processed to obtain $\Delta\Delta G_{\text{binding}}$. The corresponding MD were performed in an infinite continuum medium with the dielectric properties of water using the Generalized Born solvation model. The preference for this type of solvent representation over the explicit water representation can be justified by several reasons, namely the smaller simulation time necessary compared to those of 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 that of the explicit water simulation, and an easier interpretation of the results since the water degrees of freedom are absent.[47] The continuum solvent is used to calculate the $\Delta G_{\text{solvation}}$ value, and therefore it is coherent to use the same method to generate the dynamic trajectories.

To obtain reliable estimates of the relative binding energy, the average total energy value (total energy = potential energy + kinetic energy) must converge. Figures 2a–2c respectively show these average total energy values for all of the 3 complexes studied. The total energy remaining constant indicates that we have reached equilibrium. Although for the 1Vfb and 1Fcc complexes, the systems are well equilibrated after 1500 ps of MD simulation, for the 1F47 complex equilibration was only achieved after 2000 ps. This can be explained by the reduced size of this complex (159 amino acids) and the high mobility of the ligand which consists of only 15 amino acids. Its flexibility gives rise to higher root-mean-square deviation (RMSD) values after 2 ns, due to a localized conformational transition corresponding to a slight opening of the end of the second turn of the FtsZ α-helix, which anyway lies beyond the region where the binding determinants are located. However, calculation of the RMSD restricted to the interface region shows a stable

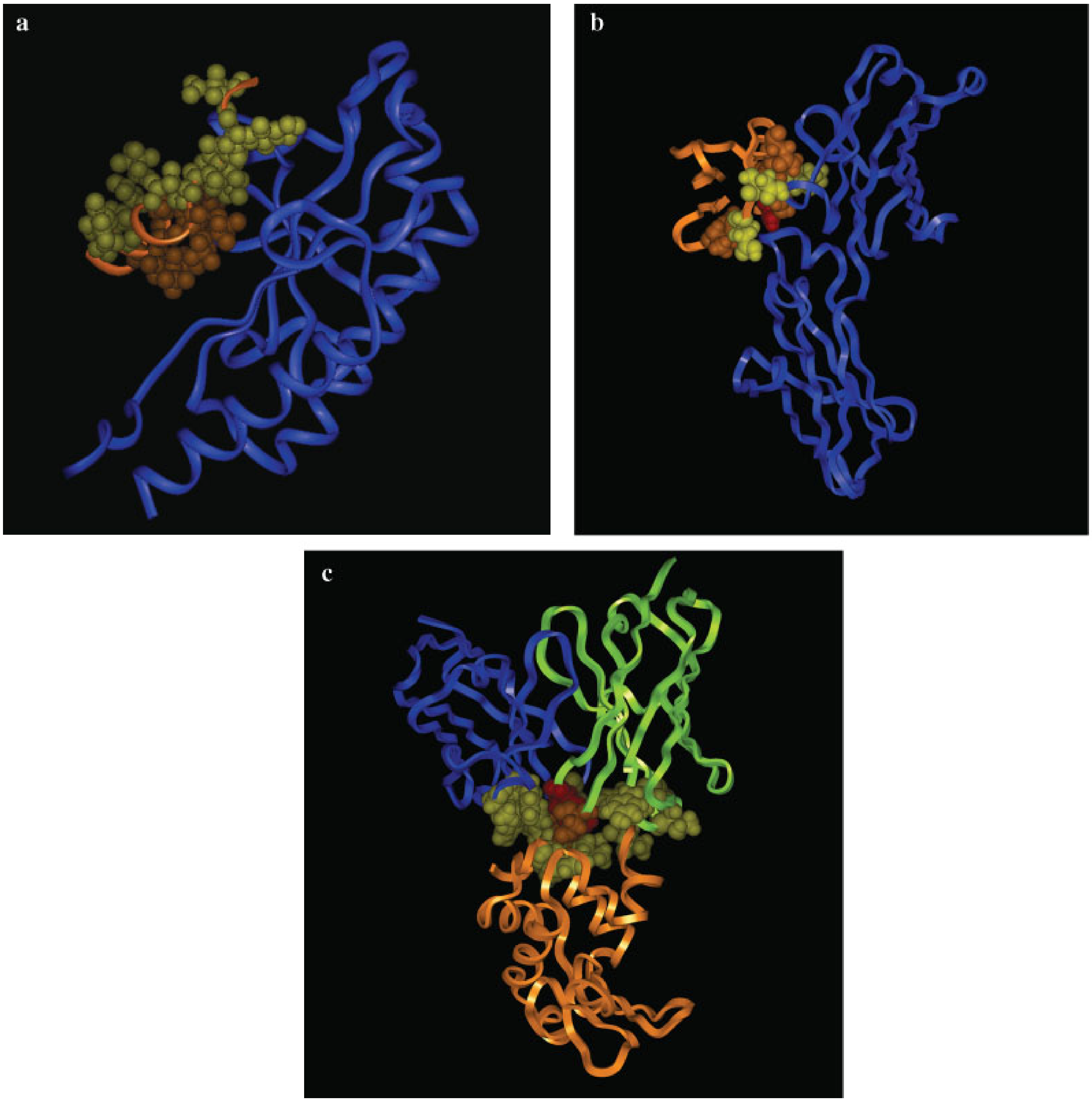


Figure 1. (a) Complex formed between the bacterial cell-division protein ZipA and the FtsZ fragment highlighting the mutated residues by a ball and stick representation; (b) complex formed between the human immunoglobulin IgG and the C2 fragment of streptococcal protein G highlighting the mutated residues by a ball-and-stick representation; (c) complex formed between an immunoglobulin and a hen egg lysozyme highlighting the mutated residues by a ball and stick representation. In yellow are represented the null-spots (relative binding energy < 2.0 kcal/mol), in orange the warm-spots (relative binding energy between 2.0 and 4.0 kcal/mol), and in red the hot-spots (residues with a relative binding energy higher than 4.0 kcal/mol).

trend around 2 Å until the end of the simulation. For the processing analysis the last 500 ps of the trajectories were selected. In Figures 3a–3c we have plotted the time series of RMSD from the X-ray crystal structure of Cα atoms of the complex and the respective separate proteins for the three simulations. After equilibration, the systems are very stable and the RMSD of the main chain reaches a maximum of 2.0–3.5 Å. This value is much smaller if we did consider only the separate monomers.

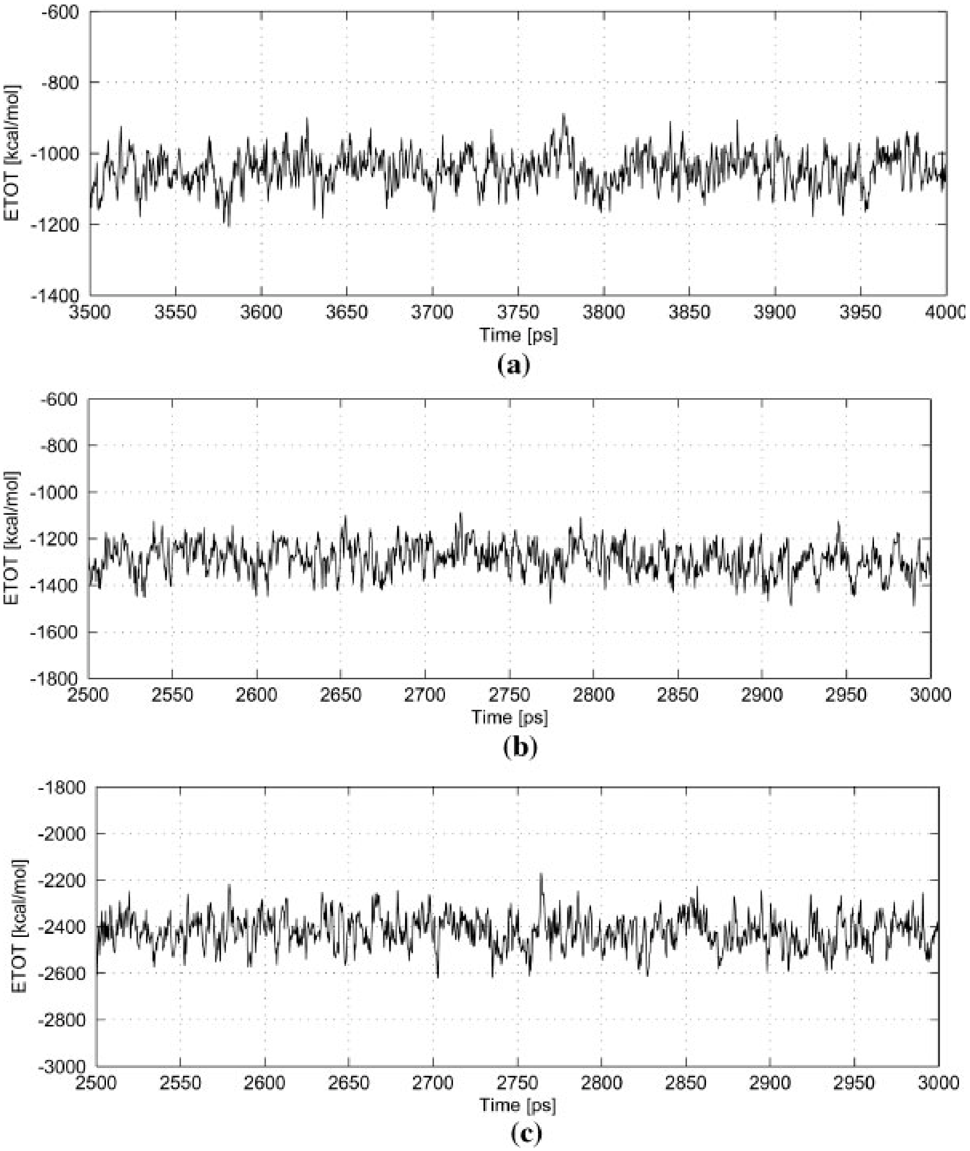


Figure 2. Total energy values as a function of simulation time for the last 500 ps of the dynamic simulation for the complex formed between (a) the bacterial cell-division protein ZipA and the FtsZ fragment; (b) the human immunoglobulin IgG and the C2 fragment of streptococcal protein G; and (c) an immunoglobulin and a hen egg lysozyme.

One of the fundamental limitations of the current approaches to calculate the relative binding free energy is the small success rate when charged residues are involved. Since hot-spots are enriched with this kind of residues, it becomes a huge problem that it is necessary to solve. We have studied several ways to deal with this problem. Our attention was captured by one empirical parameter used in the numerical differentiation of the Poisson-Boltzmann equation as implemented in DelPhi to calculate the free energy of polar solvation: the internal dielectric constant.

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 a parameter that depends on the model and the methodology used.[26,48,49] The internal dielectric constant is a means of accounting for responses to an electric field that are not treated explicitly.[48] This response depends on the constituting amino acid residues, and therefore different protein regions should have different internal dielectric constants. In the absence of these limitations, an internal dielectric constant of 1 should be used. However, it is necessary to use a dielectric constant of at least 2, because the induced dipoles are not included explicitly.[50] As already established, when group reorientation is important and is not included explicitly in the formalism (due to insufficient sampling of the conformational space), the dielectric

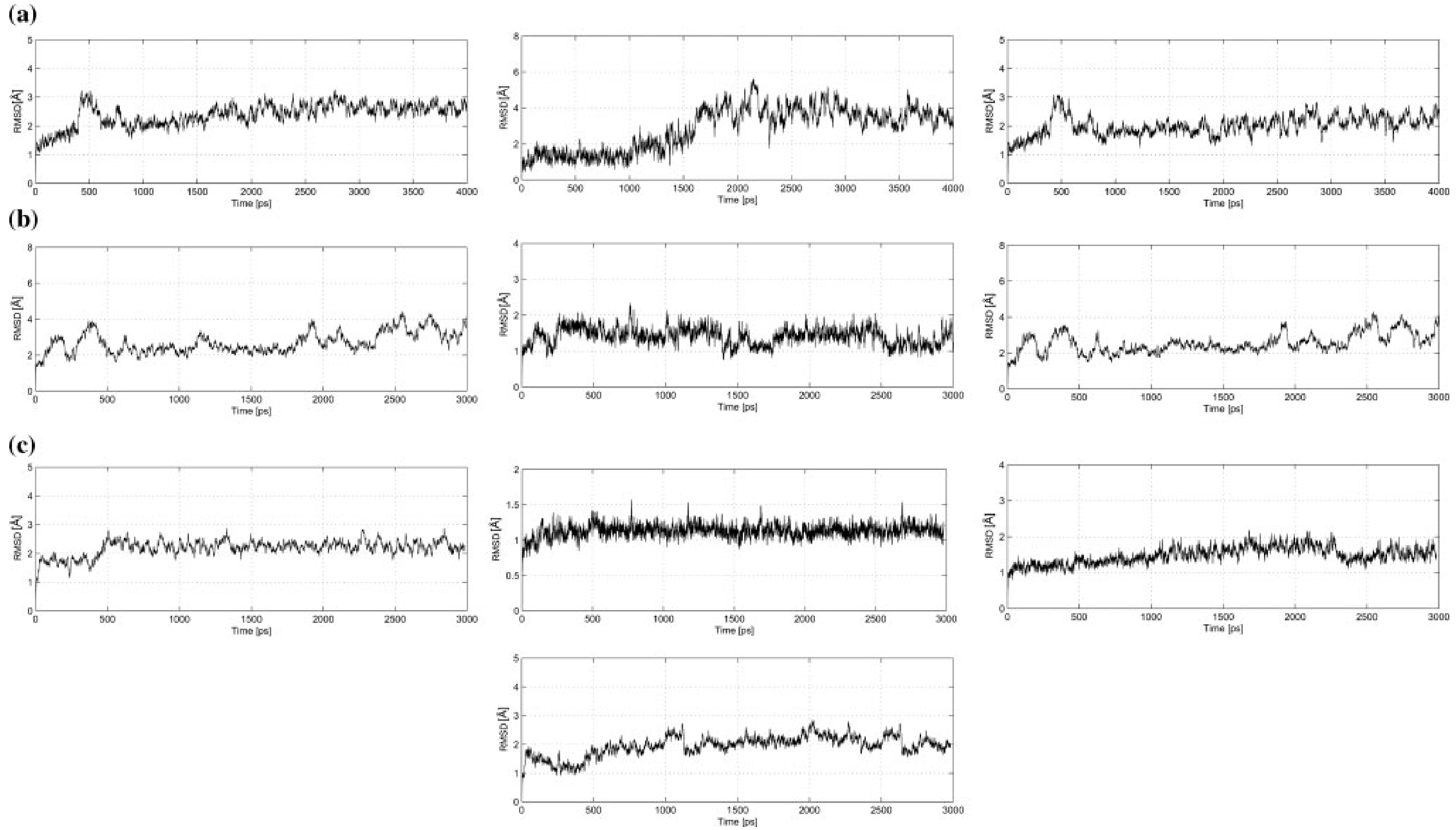


Figure 3. (a) RMSD plots for the protein backbone of the complex formed between the bacterial cell-division protein ZipA and the FtsZ fragment relative to its initial structure. From left to right: the complex, the ligand and the receptor; (b) RMSD plots for the protein backbone of the complex formed between the human immunoglobulin IgG and the C2 fragment of streptococcal protein G relative to its initial structure. From left to right: the complex, the ligand and the receptor; (c) RMSD plots for the protein backbone of the complex formed between an immunoglobulin and a hen egg lysozyme relative to its initial structure From left to right: the complex, one receptor chain, the other receptor chain and the ligand.

Table 1. Absolute Deviation Error (Dev) for the $\Delta\Delta G_{\text{binding}}$ of the Different Amino Acid Types.[a]

	Internal dieletric constant				
Residue type	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	$\varepsilon = 4$	$\varepsilon = 5$
Nonpolar	3.08	0.52	1.18	1.51	1.78
Polar	3.78	1.07	0.78	0.92	1.22
Charged (plus His)	3.63	1.61	1.07	0.92	1.08

[a]Dev = $\langle | \Delta\Delta G_{\text{binding}}$ (calculated) $- \Delta\Delta G_{\text{binding}}$ (experimental) $| \rangle$.

constant of the solute should be raised to values from 2 to 4 or higher.[26,51,52]

To evaluate the influence of the value of the dielectric constant we calculated the $\Delta\Delta G_{\text{binding}}$ values for all studied residues with dielectric constants ranging from 1 to 5, and analyzed the unsigned deviation between the calculated and the experimental $\Delta\Delta G_{\text{binding}}$ values for each of the five internal dielectric constant values. The results are shown in Table 1. From perusal of Table 1 it can be perceived that the value of the dielectric constant that better mimics the experimental values is not universal; instead it increases with the polarity of the residues, being 2 for nonpolar residues, 3 for polar residues, and 4 for charged residues.

Thus, we present here a new idea: the use of different internal dielectric constants for different mutations with values dependent on the type of amino acid, which was mutated to an alanine.

There are 20 alpha amino acids commonly found in proteins and they can be divided into basically four groups according to the structure of the side chain: nonpolar and neutral (valine, alanine, leucine, isoleucine, phenylalanine, proline, glycine, methionine, and tryptophan), polar and neutral (aspargine, glutamine, cysteine, tyrosine, serine, and threonine), acidic and charged (aspartic acid and glutamic acid), and basic and charged (lysine, arginine, and histidine). As histidine can be uncharged or charged at physiological pH we have grouped this residue with lysine and arginine at the basic and charged amino acids.

Recalling that we used only one trajectory for the computational energy analyses, it is important to highlight that side chain reorientation is not included explicitly in the formalism. As amino acid polarity increases, the structural effect beyond the neighbor residues also increases, and the conformational reorganization after alanine mutagenesis should be more extensive. This reorganization is not explicitly taken into account in the single trajectory protocols but its effect can be implicitly included by raising the internal dielectric constant. It is not possible to know the correct internal dielectric constant value that should be used because it depends on the mutated amino acid and the interacting residues. Nevertheless, we have noticed that by using only a three-internal-dielectric-constant set exclusively characteristic of the mutated amino acid (2 for the nonpolar amino acids, 3 for the polar residues, and 4 for the charged amino acids), it was possible to obtain an excellent agreement with the experimental results for the $\Delta\Delta G_{\text{binding}}$ values. This fact appears to indicate that the organization level of the neighbor residues of the mutated amino acid depends essentially of the amino acid mutated, and does not depend on the nature of the interacting residues. A possible explanation is that an amino acid present in the interface is usually surrounded by residues of the other protein with the same polarity. Analysis of protein–protein interfaces show that this complementarity between the individual molecules is a very important factor for the binding. Therefore, the nature of the mutated amino acid is usually representative of the nature of the surrounding environment. Furthermore, this effect can be mimicked by a single macroscopic parameter, the internal dielectric constant.

This approach is supported by other studies like the one performed by Wisz and Hellinga (2003),[53] that applied different internal dielectric values to the analysis of the p*K*a values of ionizable groups. They also noticed that multiple, geometry-dependent dielectric constants assigned separately for each pairwise interaction and determined by location of the two charges relative to the solvent, by the local environment, by the type of interaction between relevant amino acid side chains should be used to mimic the relaxation effects.[53]

Although called an amino acid, proline is in fact an imino acid. When proline is in a peptide bond, it does not have a hydrogen on the α amino group, so it cannot donate a hydrogen bond to stabilize an α helix or a β sheet. Unlike other amino acids that exist almost exclusively in the *trans-* form in polypeptides, proline can exist in the *cis-* configuration in peptides. The proline backbone conformation is significantly different from the one from alanine, and even though the ring is not reactive, it does restrict the geometry of the backbone chain in any protein where it is present. For example, when proline is found in an α helix, the helix will have a slight bend due to the lack of the hydrogen bond. Proline mutations to alanine were not considered here because this type of mutation is disruptive, can produce sometimes abnormal changes to the binding as a result of significant conformational changes, and therefore would masquerade the results.

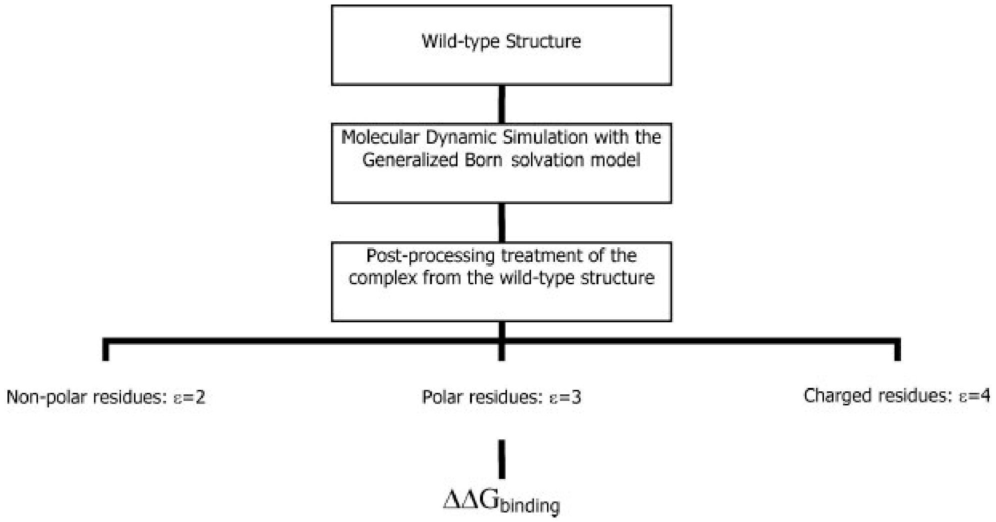
We have accomplished a method that gives a high success rate. This method has been applied to 46 mutations in 3 complexes and the results are shown in Table 1. If we consider a deviation of ±1.3 kcal/mol from the experimental value as an accurate result, we can observe from Table 1 that we have an overall success rate of 80%, an 80% success rate for the null-spots, a 78% achievement of the correct relative binding free energy of the warm-spots, and an 82% success rate concerning the residues with a $\Delta\Delta G_{\text{binding}}$ higher than 2 kcal/mol. It can also be observed that there is only 5% of false positives in the hot-spot detection, and that the residues responsible are warm-spots with $\Delta\Delta G_{\text{binding}}$ between 2 and 4 kcal/mol. 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. It is difficult to estimate how accurate the method is in estimating binding spots with $\Delta\Delta G_{\text{bind}}$ higher than 4 kcal/mol. The problem with the hot-spots (within the >4kcal/mol definition) is that they are rare, and an impractically large number of systems should be simulated to get a very accurate picture. Moreover, the absolute value for $\Delta\Delta G_{\text{bind}}$ in hot-spots is usually not available experimentally, and usually only lower limits can be found in the literature, which does not allow for the evaluation of the absolute error in the computational calculation, and therefore the agreement is made only in a qualitative perspective. Nevertheless, we should

Table 2. Results of the Methodological Approach for Computational Alanine Screening Mutagenesis.[a]

Residue type	Protein	Mutation	%Burial	$\Delta\Delta G_{\mathrm{exp}}$	$\Delta\Delta G_{\mathrm{calculated}}$
Nonpolar	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	−0.72
	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
		Leu129Ala	78.1	0.20	0.00
		Ile124Ala	98.5	1.20	0.56
Polar	1F47	Tyr3Ala	81.1	0.86	3.20
	1Fcc	Thr25Ala	90.3	0.24	0.01
		Asn35Ala	88.5	2.40	1.23
		Thr44Ala	76.4	2.00	2.24
		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
		Tyr101Ala	97.3	>4.0	3.61
		Asn56Ala	65.7	0.20	0.47
	1Vfb-ligand	Ser24Ala	98.4	0.70	2.23
		Tyr23Ala	97.6	0.80	2.45
		Thr118Ala	86.9	0.80	1.07
		Asn19Ala	89.3	0.40	0.21
		Gln121Ala	99.9	2.90	3.49
Charged	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	8.70
		Lys28Ala	97.2	1.30	3.30
		Lys31Ala	99.3	3.50	4.78
		Asp40Ala	72.7	0.30	−0.13
		Glu42Ala	67.0	0.40	−0.07
	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 (%)		null-spots			80
		warm-spots			78
		hot-spots			100
		Overall			80

[a] The units of free energies and potential energies are kcal/mol. The burial percentage of each mutant residue including backbone atoms upon binding is according to:

$$\%\mathrm{Burial} = 100 - \frac{\mathrm{Area}_{\mathrm{complex}}}{\mathrm{Area}_{\mathrm{unbound}}} \times 100.$$


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
ΔΔG binding

Scheme 1. Resume of the methodological approach for computational alanine screening mutagenesis.

stress that the 4 kcal/mol value is not universal. Thus, many authors[1,6,9] have used a cutoff of 2 kcal/mol to define a hot-spot. Within that criterion the method has a success rate of 82%.

The standard deviation of the mean for $\Delta\Delta G_{\text{binding}}$ ranges from 0.64 to 0.98 kcal/mol for all the 46 mutations analyzed. As mentioned previously, the method was applied to three complexes, and a mean and maximum unsigned error of 0.80 and 4.52 kcal/mol was obtained in a set of 46 mutations. It is important to stress that the systems analyzed are diverse (three proteins with different characteristics and sizes) and the universe of mutations studied is substantial. The calculated effects of mutating the 10 uncharged residues are in good agreement with the experimental mutagenesis values as can be observed from Table2. We have also obtained an excellent agreement with the experimental values of the relative binding free energy upon mutation, for the nineteen polar and seventeen charged residues, the most challenging ones because their mutation usually results in errors by over 4–10 kcal/mol.[24,25]

The causes for the deviations obtained in this work can be several, ranging from inherent inaccuracies of the force field, lack of explicit atomic polarization, incomplete exploration of the side-chain rotamer conformational space and the use of a single trajectory protocol. The last three sources of uncertainty have been implicitly corrected in an average way through the consideration of different dielectric constants for different amino acids. This is indeed the conceptual backbone of the methodology. However, the microenvironment around each amino acid varies substantially, and therefore there will always be some situations which deviate from the average, and that cannot be perfectly corrected through an empirical parameter calibrated to reflect the most common scenarios.

Conclusions

In the present study, we have improved the strategy originally proposed by Massova/Kollman[25] to predict the binding affinities of the ligands. The use of the molecular mechanics AMBER1994 force field (Cornell et al.[33]) and a continuum solvation approach with different internal dielectric constant values for different kinds of residues allowed the identification of the hot-spots at protein–protein interfaces with a high success rate.

This fully atomistic computational methodological approach consists in using computational MD simulations with a single trajectory protocol. These simulations are performed in an infinite continuum medium with the dielectric properties of water using the Generalized Born solvation model, and a post-processing treatment of the complex permits to calculate the respective energies for the complex and all interacting monomers from the wild-type structure. A set of different internal dielectric constants has to be used for the proteins, depending on the type of amino acid that is mutated. Therefore, for the charged amino acids (aspartic acid, glutamic acid, lysine, arginine, and histidine) a constant of 4 should be used, for the remaining polar residues (aspargine, glutamine, cysteine, tyrosine, serine, and threonine) not ionized at physiological pH the internal dielectric constant should be 3, and for the nonpolar amino acids (valine, leucine, isoleucine, phenylalanine, methionine, and tryptophan) the internal dielectric constant should be 2. The different internal dielectric constants account for the different degree of relaxation of the interface when different types of amino acids are mutated for alanine; the stronger the interactions these amino acids establish, the more extensive the relaxation should be, and the greater the internal dielectric constant value must be to mimic these effects.

This methodological approach summarized in Scheme 1 gives an overall success rate of 80%, an 80% success rate for the null-spots, 78% achievement of the correct relative binding free energy differences between the wild-type and mutant complexes of the warm-spots, and an 82% rate of success for residues with a $\Delta\Delta G_{\text{binding}}$ higher than 2 kcal/mol.

The standard deviation of the mean, defined as $\sigma/\sqrt{n}$, where n is the number of snapshots, ranges from 0.64 to 0.98 kcal/mol; thus "chemical accuracy" was achieved. The method was applied to three complexes, and a mean and maximum unsigned error of 0.80 and 4.52 kcal/mol was obtained in a set of 46 mutations.

This method is simple, fast, has a low computational cost, and can be applied to a wide range of proteins providing a correct anatomic image of an interface. It can be used prior to an experimental investigation helping in the hot-spot detection and the choice of the amino acids to mutate.

References

1. Kortemme, T.; Baker, D. Curr Opin Chem Biol 2004, 8, 91.
2. Russel, R. B.; Alber, F.; Aloy, P.; Davis, F. P.; Korkin, D.; Pichaud, M.; Topf, M.; Sali, A. Curr Opin Struct Biol 2004, 14, 313.
3. Clackson, T.; Wells, J. A. Science 1995, 267, 383.
4. Lo Conte, L.; Chothia, C.; Janin, J. J Mol Biol 1999, 285, 2177.
5. Arkin, R.; Randal, M.; DeLano, W. L.; Hyde, J.; Luong, T. H.; Oslob, J. D.; Raphael, D. R.; Taylor, L.; Wang, J.; McDowell, R. S.; Wells, J. A.; Braisted, A. C. Proc Natl Acad Sci USA 2003, 100, 1603.
6. Bogan, A. A.; Thorn, K. S. J Mol Biol 1998, 280, 1.
7. Pons, J.; Rajpal, A.; Kirsch, J. Protein Sci 1999, 8, 958.
8. Hu, Z.; Ma, B.; Nussinov, R. Proteins 2000, 39, 331.
9. Keskin, O.; Ma, B.; Nussinov, R. J Mol Biol 2005, 345, 1281.
10. Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Freer, S. T.; Rose, P. M. Proteins 2003, 53, 201.
11. Gao, Y.; Wang, R.; Lia, L. J Mol Model 2004, 10, 44.
12. Lopez, M. A.; Kollman, P. A. Protein Sci 1993, 2, 1975.
13. Kortemme, T.; Baker, D. Proc Natl Acad Sci USA 2002, 99, 14116.
14. Kortemme, T.; Kim, D. E.; Baker, D. Sci STKE 2004, 219, 12.
15. Schapira, M.; Totrov, M.; Abagyan, R. J Mol Recognit 1999, 12, 177.
16. Aqvist, J.; Medina, C.; Samuelsson, J. E. Protein Eng 1994, 7, 385.
17. Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Freer, S. T.; Rose, P. M. Proteins 2002, 48, 539.
18. Guerois, R.; Nielsen, J. E.; Serrano, L. J Mol Biol 2002, 320, 369.
19. Kollman, P. A.; 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, D. A.; Cheatham, T. E., III. Accounts Chem Res 2002, 33, 889.
20. Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A. Annu Rev Biophys Biomol Struct 2001, 30, 211.
21. Massova, I.; Kollman, P. A. J Am Chem Soc 1999, 121, 8133.
22. Wang, J.; Morin, P.; Wang, W.; Kollman, P. A. J Am Chem Soc 2001, 123, 5221.
23. Wang, W.; Kollman, P. A. J Mol Biol 2000, 303, 567.
24. Reyes, C. M.; Kollman, P. A. J Mol Biol 2000, 295, 1.
25. Huo, S.; Massova, I.; Kollman, P. A. J Comput Chem 2002, 23, 15.
26. Wagoner, J.; Baker, N. A. J Comput Chem 2004, 25, 1623.
27. Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L., III. J Comput Chem 2004, 25, 265.
28. Juffer, A.; Argos, P.; Vogel, H. J Phys Chem B 1997, 101, 7664.
29. Mosyak, L.; Zhang, Y.; Glasfeld, E.; Haney, S.; Stahl, M.; Seehra, J.; Somers, W. S. EMBO J 2002, 19, 3179.
30. Sauer-Eriksson, A. E.; Kleywegt, G. J.; Uhlen, M.; Jones, T. A. Structure 1995, 3, 265.
31. Bhat, T. N.; Bentley, G. A.; Boulot, G.; Greene, M. I.; Tello, D.; Dall'Acqua, W.; Souchon, H.; Schwarz, F. P.; Mariuzza, R. A.; Poljak, R. J. Proc Natl Acad Sci USA 1994, 9, 1089.
32. Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A.AMBER 8; University of California, San Francisco, 2004. The Amber biomolecular simulation program.
33. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T., Caldwell, J. W.; Kollman, P. A. J Am Chem Soc 1995, 117, 5179.
34. Tsui, V.; Case, D. A. Biopolymers (Nucleic Acid Sci) 2001, 56, 275.
35. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. J Comput Phys 1997, 23 327.
36. Pastor, I. W.; Brooks, B. R.; Szabo, A. Mol Phys 1998, 65, 1409.
37. Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Biopolymers 1992, 32, 523.
38. Izaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. J Chem Phys 2001, 114, 2090.
39. Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. J Comput Chem 2002, 23, 128.
40. Rocchia, W.; Alexov, E.; Honig, B. J Phys Chem B 2001, 105, 6507.
41. Sitkoff, D.; Sharp, K. A.; Honig, B. J Phys Chem 1994, 98, 1978.
42. Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. J Mol Struct (Theochem) 2005, 729, 11.
43. Connolly, M. L. J Appl Crystallogr 1983, 16, 548.
44. Hinke, S. A.; Manhart, S.; Speck, M.; Pederson, R. A.; Demuth, H. U.; McIntosh, C. H. Life Sci 2004, 75, 1857.
45. Arkin, M. R.; Wells, J. A. Nat Rev Drug Discov 2004, 3, 301.
46. Cohlke, H.; Case, D. A. J Comput Chem 2004, 25, 238.
47. Xia, B.; Tsui, V.; Case, D. A.; Dyson, J.; Wright, P. E. J Biomol NMR 2002, 22, 317.
48. Sheinerman, F. B.; Norel, R.; Honig, B. Curr Opin Struct Biol 2002, 10, 153.
49. Schutz, C. N.; Warshel, A. Proteins 2001, 44, 400.
50. Luo, R.; Hsieh, M. J. Proteins 2004, 56, 475.
51. Gilson, M. K.; Honig, B. Biopolymers 1986, 25, 2097.
52. Simonson, T. J Am Chem Soc 1998, 120, 4875.
53. Wisz, M. S.; Hellinga, H. W. Proteins 2003, 51, 360.