JCTC Journal of Chemical Theory and Computation

Article

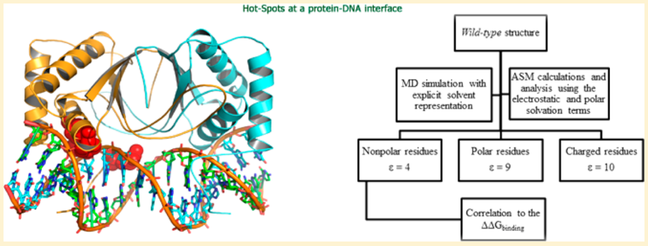
pubs.acs.org/JCTC

Computational Alanine Scanning Mutagenesis—An Improved Methodological Approach for Protein–DNA Complexes

Rui M. Ramos and Irina S. Moreira*

REQUIMTE/Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal

S *Supporting Information*


Hot-Spots at a protein-DNA interface
Wild-type structure
MD simulation with explicit solvent representation
ASM calculations and analysis using the electrostatic and polar solvation terms
Nonpolar residues $\varepsilon = 4$
Polar residues $\varepsilon = 9$
Charged residues $\varepsilon = 10$
Correlation to the $\Delta\Delta G_{\text{binding}}$

ABSTRACT: Proteins and protein-based complexes are the basis of many key systems in nature and have been the subject of intense research in the last decades, in an attempt to acquire comprehensive knowledge of reactions that take place in nature. Computational Alanine Scanning Mutagenesis approaches have been extensively used in the study of protein interfaces and in the determination of the most important residues for complex formation, the Hot-spots. However, as it is usually applied to the study of protein–protein interfaces, we tried to modify and apply it to the study of protein–DNA interfaces, which are also crucial in nature but have not been the subject of as much research. In this work, we carry out MD simulations of seven protein–DNA complexes and tested the influence of the variation of different parameters on the determination of the binding free energy terms ($\Delta\Delta G_{\text{binding}}$) of 78 mutations: solvent representation, internal dielectric constant, Linear and Nonlinear Poisson–Boltzmann equation, Generalized Born model, simulation time, number of structures analyzed, number of MD trajectories, force field used, and energetic terms involved. Overall, this new approach gave an average error of 1.55 kcal/mol, and *P*, *R*, F1, accuracy, and specificity values of 0.78, 0.50, 0.61, 0.77, and 0.92, respectively. This improved computational alanine scanning mutagenesis approach may serve as a tool to explore the behavior of this important class of complexes.

■ INTRODUCTION

Proteins and nucleic acids are the basis of all biological systems, which by their turn are the basis of Nature as we know it.[1] They can act as carriers and catalysts and can provide mechanical support and immune protection among other aspects.[2] Proteins have the tendency to associate with other macromolecules such as other proteins, nucleic acids, or small ligands, forming stable complexes whose interactions (protein–protein interactions (PPI), protein–DNA interactions (PDI), or protein–ligand interactions (PLI)) are essential for the correct function of the molecular system. The understanding of these interactions allows engineering new functions and adjusting cellular behavior in a predictive manner as well as it enables the rational design of new therapeutic agents.[3]

Although proteins are usually large with complex interfaces, their association is governed by single residues with high binding affinity, the Hot-Spots (HS).[4] HS, which are considered the most important residues for complex formation and for its stability, are defined as residues that upon alanine mutation generate a binding free energy difference ($\Delta\Delta G_{\text{binding}}$) higher than 2.0 kcal/mol; Null-Spots (NS) are defined as residues that cause a binding free energy difference lower than 2.0 kcal/mol.[5] The HS are conserved residues and have a specific location, since they usually form clusters in the central or core region of the interface.[6]

In the last years, there has been an attempt to accurately reproduce the experimental mutagenesis values using various computational methodologies, which can be generally classified as (i) empirical functions, which are methods that make use of simple knowledge-based models to access complex association; (ii) feature-based approaches, usually being more qualitative than quantitative; and (iii) fully atomistic methods that perform mutations of the interfacial residues to estimate binding free energies.[7] Traditional methods such as thermodynamic integration (TI) or free energy perturbation (FEP)[8] are the most accurate methods to calculate the binding strength of protein complexes. On the other hand, they are also the most demanding computational methods, which limit the screening of a large number of structural perturbations. Therefore, they are commonly replaced by faster methods such as the Molecular Mechanics-Poisson–Boltzmann/Generalized Born Surface Area (MM-PB/GBSA) method.[6b,9]

Our improvement of this method has proved before[6,9b,c,10] to be a valuable resource in the computational determination and accurate prediction of mutagenesis values, with the focus on HS. However, to the best of our knowledge, the ASM methodology has been applied almost exclusively in PPI. The protein–DNA interface, for example, has as much biological interest as the protein–protein, but the application of the ASM methodology to this type of interface is insufficient. This may be explained due to the inherent difficulty in characterizing the energetics of these highly charged systems.

Received: May 13, 2013

dx.doi.org/10.1021/ct400387r | *J. Chem. Theory Comput.* XXXX, XXX, XXX–XXX

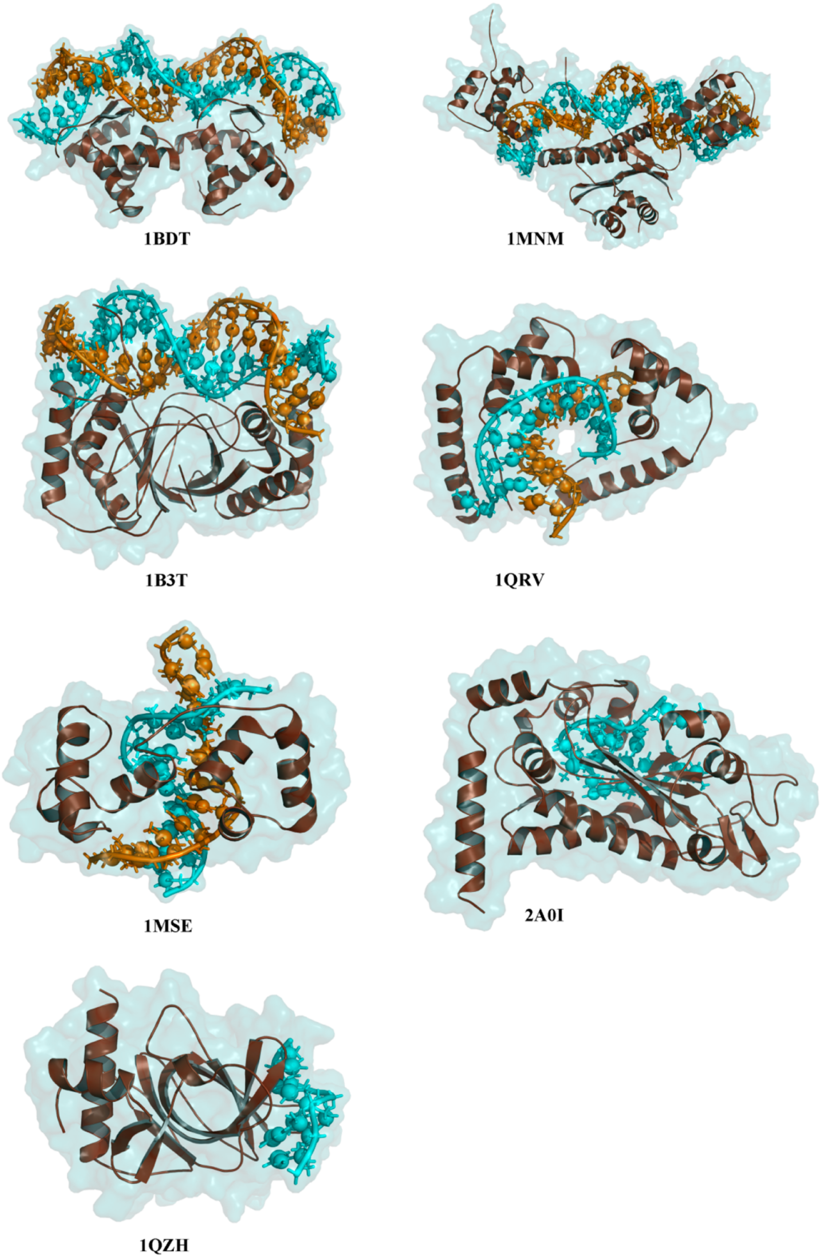


Figure 1. Representation of the seven protein−DNA complexes studied in this work. Protein and DNA are shown in brown and blue/orange, respectively.

In this work, we studied the influence of the variation of several parameters in the calculation of the binding free energy values for a set of protein−DNA complexes. In particular, we focused our attention in the influence of (i) the ionic concentration; (ii) the energy terms included in the calculation; (iii) the various types of solvent representation; (iv) the MD simulation length; (v) the number of snapshots analyzed; (vi) the force field used; (vii) the number of MDs; (viii) the PB solvers used; (ix) the MM-PBSA vs MM-GBSA approach; and (x) the internal dielectric constant. Therefore, we explored and extended the computational alanine scanning mutagenesis to a different class of interfaces, namely the protein−DNA interface.

■ METHODOLOGY

1. System Setup. The PRONIT[11] database was used to select the protein−DNA complexes with a known X-RAY structure, as well as a meaningful number of alanine mutations and their corresponding $\Delta\Delta G_{\text{binding}}$ values. We selected seven different systems: four for our training-set and three for our test-set. Our training-set was formed by the following complexes: (i) nuclear protein EBNA1 and DNA (PDB ID: 1B3T[12]); (ii) gene-regulating protein arc and DNA (PDB ID: 1BDT[13]); (iii) C-Myb DNA-binding domain and DNA (PDB ID: 1MSE[14]); and (iv) the high-mobility group protein D and DNA (PDB ID: 1QRV[15]). Our test-set was formed by (i) the complex between the MCM1 transcriptional regulator and MAT α-2 transcriptional repressor and DNA (PDB ID: 1MNM[16]); (ii) the Pot1 protein and DNA (PDB ID: 1QZH[17]); and (iii) the TraI protein and DNA (PDB ID: 2A0I[18]) (Figure 1). The composition of the two sets and their mutants is listed in Table 1. The protonation state of the different residues at the physiological range was determined using the PROPKA methodology.[19]

Table 1. Description of the 78 Residues That Constitute Our Data Set, Evidencing the Respective System, PDB ID, Amino Acid Type, and Experimental $\Delta\Delta G_{\text{binding}}$[a]

PDB ID	#AA PDB	#AA mutated	$\Delta\Delta G_{\text{binding}}$, kcal mol^{-1}	ref.
1B3T	469	R	3.41	12
1B3T	518	Y	2.62	
1B3T	522	R	4.40	
1MSE	116	S	0.06	14
1MSE	139	N	0.60	
1MSE	141	E	−0.10	
1MSE	187	S	0.10	
1QRV	9	L	0.02	15
1QRV	13	M	1.20	
1QRV	32	V	−0.30	
1BDT	4	M	1.10	13
1BDT	5	S	1.30	
1BDT	6	K	1.30	
1BDT	7	M	1.90	
1BDT	9	Q	1.80	
1BDT	11	N	2.00	
1BDT	13	R	6.00	
1BDT	23	R	0.20	
1BDT	29	N	−1.00	
1BDT	31	R	−0.30	
1BDT	32	S	−2.30	
1BDT	33	V	−2.30	
1BDT	34	N	3.20	
1BDT	35	S	−0.20	
1BDT	39	Q	−0.40	
1MNM	17	E	0.24	16
1MNM	21	I	0.88	
1MNM	22	E	0.20	
1MNM	23	I	0.86	
1MNM	24	K	−0.13	
1MNM	25	F	4.05	
1MNM	26	I	4.05	
1MNM	27	E	0.41	
1MNM	28	N	1.75	
1MNM	29	K	−0.13	
1MNM	32	R	2.76	
1MNM	33	H	−0.81	
1MNM	34	V	0.58	
1MNM	35	T	0.64	
1MNM	36	F	4.05	
1MNM	37	S	0.43	
1MNM	38	K	4.05	
1MNM	40	K	2.83	
1MNM	41	H	−1.24	
1MNM	43	I	3.24	
1MNM	45	K	4.05	
1MNM	46	K	4.05	
1MNM	48	F	−0.13	
1MNM	51	S	0.00	
1MNM	52	V	0.11	
1MNM	53	L	4.05	
1MNM	66	T	0.64	
1QZH	62	T	1.53	17
1QZH	64	D	1.46	
1QZH	88	F	3.96	
1QZH	91	Q	1.26	
1QZH	115	Y	0.66	
1QZH	122	L	1.00	
1QZH	123	S	−0.70	
2A0I	3	S	4.15	18
2A0I	8	R	1.70	
2A0I	88	K	5.57	
2A0I	147	D	0.30	
2A0I	148	T	0.30	
2A0I	149	S	3.41	
2A0I	150	R	2.92	
2A0I	153	E	1.70	
2A0I	155	Q	2.20	
2A0I	158	T	0.90	
2A0I	187	E	2.10	
2A0I	220	K	0.80	
2A0I	221	H	2.82	
2A0I	223	M	2.70	
2A0I	237	R	4.39	
2A0I	241	I	3.91	
2A0I	242	R	1.20	
2A0I	254	R	3.17	
2A0I	265	K	1.80	

[a] Experimental $\Delta\Delta G_{\text{binding}}$ values were measured using different methods, such as isothermal titration calorimetry, gel shift assays, fluorescence intensity and anisotropy measurements, and others. Detailed information about this subject can be consulted in the references.

2. Molecular Dynamics Simulations. The MD simulations were performed using the AMBER9[20] package with the AMBER force field *ff99SB* for the proteins and ff94 for the nucleic acids. For one particular system (PDB ID: 1B3T), in order to study the influence of the force field over the ASM results, we have also performed a MD simulation of the complex using parmbsc0.[21] For the same system, to investigate the influence of the number of trajectories analyzed, we have also performed the MD simulation of the three mutant complexes. The mutations were introduced with the software PYMOL.[22] Two different types of simulations were performed: in implicit and in explicit solvent. In the implicit solvent simulations, we used the GB solvent method (GB^{OBC})[23] due to its capacity to reproduce accurately the experimental binding free energy values, in previous works made with protein−protein systems.[5b,9c] The explicit solvent simulations were also used due to the highly charged and polar character of the typical protein−DNA interface. Each system was solvated with a box of TIP3P water molecules that extended 10 Å from any edge of the box to the protein atoms.[24] Following typical procedures, the necessary number of counterions (Na^+ ions) was added to achieve global system neutrality.[21,25] In each of the simulations, we started with a minimization stage, to remove bad contacts, by steepest descent followed by conjugated gradient. Periodic boundary conditions were applied using the particle mesh Ewald (PME) method[26] to treat long-range electrostatic interactions, and the nonbonded interactions were truncated with a 10-Å cutoff. In all MD simulations, the bond lengths involving hydrogens were constrained using SHAKE[27] and the equations of motion were integrated with a 2-fs time step. The systems were subjected to 2 ns of heating procedure, gradually raised from 0 to 300 K (NVT ensemble), followed by 6 ns of production stage in NPT ensemble. The Langevin[28] algorithm was used to regulate the temperature of the system.

3. Computational Alanine Scanning Mutagenesis. In this work, we used the MM-PB/GBSA (Molecular Mechanics-Poisson−Boltzmann/Generalized Born Surface Area) script[10h] integrated into the AMBER9 software for the calculations of the binding free energy difference ($\Delta\Delta G_{\text{binding}}$) upon alanine mutation. The MM-PB/GBSA methodology combines a molecular mechanics approach with continuum solvent models for the calculation of the binding free energy. This method excels by its speed, accuracy, and low computational time required for the calculation, when compared with other methods used for binding free energy calculation such as TI,[8a] and it was successfully used in the last years with protein–protein interfaces.[5b,9a,c,29] The MM-PB/GBSA approach first developed by Massova et al.[10h] was improved by Moreira et al.[9c] and can now be applied with an accuracy of 1 kcal/mol with protein–protein interfaces. This methodology provided a good starting point for the application to protein–DNA interfaces. In the ASM methodology, the mutant complexes are generated by a single truncation of the mutated side chain, replacing Cγ with hydrogen atom and setting the Cβ-H direction to that of the former Cβ-Cγ. The $\Delta\Delta G_{\text{binding}}$ is defined as the difference between the mutant and wild type complexes, which can be defined as

$$\Delta\Delta G_{\text{binding}} = \Delta G_{\text{cpx-mutant}} - \Delta G_{\text{cpx-wild type}} \tag{1}$$

The free energy calculations include enthalpic and entropic contributions, being a sum of the internal energy (bond, dihedral and angle), the electrostatic and the van der Waals interactions, the free energy of polar solvation, the free energy of nonpolar solvation, and the entropic contribution (eq 2).

$$G_{\text{molecule}} = E_{\text{internal}} + E_{\text{elec}} + E_{\text{vdW}} + G_{\text{polar solvation}} + G_{\text{nonpolar solvation}} - TS \tag{2}$$

For the calculations of relative free energies between closely related complexes, it is assumed that the total entropic term in eq 2 is negligible, as the partial contributions essentially cancel each other.[9a] The internal, electrostatic, and van der Waals energy terms were calculated with no cutoff. The $G_{\text{nonpolar solvation}}$ that results from the van der Waals interaction between the solute and the solvent is proportional to the Solvent Accessible Surface Area (SASA) of the molecule and is estimated by 0.00542 × SASA + 0.92 using the molsurf program developed by Mike Connolly,[30] in which 0.00542 and 0.92 are empirical constants with units kcal Å^{-2} mol^{-1} and kcal mol^{-1}, respectively.

In this work, we determined the polar contribution to the solvation free energy by various methods, which distinguish our approach from the traditional one used with protein–protein interfaces. The $G_{\text{polar solvation}}$ was computed using the software Delphi[31] and was calculated by solving the Linear Poisson–Boltzmann (LPB) equation, the traditional method, and the Nonlinear Poisson–Boltzmann (NLPB) equation, which accounts for the importance of salt concentration in the medium. This factor is particularly important in protein–DNA interfaces due their highly charged and polar character.[32] With this in mind, we used a value of 2.5 grids/Å for scale (the reciprocal of the grid spacing); a value of 0.001 kT/c for the convergence criterion; a 90% for the fill of the grid box; and the Coulombic method to set the potentials at the boundaries of the finite-difference grid. The dielectric boundary was taken as the molecular surface defined by a 1.4 Å probe sphere and by spheres centered on each atom with radii taken from the Parse[33] vdW radii parameter set. The salt concentrations used were 0.010 M and 0.145 M, which are in the physiological range.[34] We have also calculated the electrostatic solvation energy term using PB solver implemented in the *pbsa* module from the AMBER package. The Poisson–Boltzmann (MM-PBSA) was also substituted by the GB model (MM-GBSA) with salt concentrations of 0.010 and 0.145 M. We tested a set of 10 different dielectric constants, from 1 to 10, to mimic the expected rearrangement upon alanine mutation and to assess the importance of each dielectric constant in the determination of $\Delta\Delta G_{\text{binding}}$. Since the systematic mutation of residues is a difficult and time-consuming process, whenever possible, we used a VMD plugin (CompASM[35]) that offers an easy to use graphical interface for the input files preparation, to run the calculations, and to analyze the final result.

4. Analysis. Throughout this work we used different methods to analyze the results, such as RMSDs (root mean square deviations) and different statistical tests. The majority of this analysis was carried out with VMD[36] and the PTRAJ module from the AMBER9 package. RMSDs were calculated for each system, for both protein and DNA, to ensure their equilibration throughout the simulation. All of the seven complexes were stable with variations lower than 1 Å. The overall performance of the method can be evaluated by a series of statistical concepts and tests, such as the F1 score (eq 3), which is defined as a function of Precision (P, eq 4) and Recall (R, eq 5), in which TP stands for true positive (predicted HS that are actual HS), FP stands for false positive (predicted HS that are not an actual HS), and FN stands for false negative (nonpredicted HS that are an actual HS). Accuracy is defined as the ratio of number of correctly predicted residues to number of all predicted residues (eq 6) in which TN stands for true negatives (correctly predicted NS). Specificity (eq 7) is another measure of performance, especially for NS. The statistical parameter called Precision is very useful, as it indicates the reliability of the predictions and the outcome of alanine mutations. On the other hand, Recall is related with the number of HS correctly predicted and therefore crucial in this kind of methods. Specificity is related with the nonhot-spots and their predictive calculations. F1 and Accuracy give the overall performance of the methods. The ideal method would have these values as close to 1 as possible. However, the existent methods are still far from the ideal situation presenting values in the range 0.50–0.70 (reviewed by Fernandez-Recio[7b]).

$$\text{F1} = \frac{2PR}{P + R} \tag{3}$$

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{6}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{7}$$

■ RESULTS

The training set was composed of four different complexes, in which a total of 25 residues were mutated to alanine. The test-set was composed of three complexes, and a total of 53 residues were mutated to an alanine (Table 1). In our data set, we have the following distribution in terms of amino acids and group type: Asp (3%), Glu (8%), Phe (5%), His (4%), Ile (6%), Lys (13%), Leu (4%), Met (5%), Asn (6%), Gln (5%), Arg (14%),

Table 2. Results of Average Errors and Statistical Tests Obtained with the LPB Equation for 25 Structures of the Last 2 ns of the MD Simulation in Explicit Solvent

	$\|\Delta\Delta G_{\text{MM−PBSA}} - \Delta\Delta G_{\text{Experimental}}\|$, kcal mol^{-1}									
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
all	72.88	80.56	26.18	20.35	16.87	14.56	12.90	11.66	19.14	9.92
charged	247.19	278.46	85.51	65.33	53.22	45.13	39.36	35.01	61.98	28.92
polar	4.93	3.13	2.40	2.09	1.93	1.85	1.79	1.75	1.62	1.68
nonpolar	5.41	4.55	4.52	4.41	4.34	4.30	4.26	4.24	4.20	4.20
	$\|\Delta\Delta G_{\text{MM−PBSA}} - \Delta\Delta G_{\text{Experimental}}\| < 1$/%									
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
all	8.0	16.0	12.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0
charged										
polar	16.7	33.3	12.0	41.7	41.7	41.7	41.7	41.7	41.7	41.7
nonpolar										
	statistical tests/all									
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
P	0.25	0.24	0.29	0.31	0.33	0.33	0.38	0.38	0.38	0.38
R	0.83	0.67	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
F1	0.38	0.35	0.44	0.46	0.48	0.48	0.53	0.53	0.53	0.53
accuracy	0.36	0.40	0.48	0.52	0.56	0.56	0.64	0.64	0.64	0.64
specificity	0.21	0.32	0.37	0.42	0.47	0.47	0.58	0.58	0.58	0.58

Ser (13%), Thr (6%), Val (5%), and Tyr (3%). In this group, 41% are charged residues, 33% polar and 26% nonpolar; 36% are HS, and 64% are NS. This group of residues represents the general constitution of a protein−DNA interface, with the abundance of charged residues, especially Lys and Arg.[37]

In this work, we attempt to establish an adequate methodology for the determination of HS and NS in protein−DNA interfaces, starting from the ASM methodology developed for protein−protein interfaces.[6b,9b,c] The influence of the variation of various parameters in the determination of accurate $\Delta\Delta G_{\text{binding}}$ values was assessed:

1. Influence of the Salt Concentration. Initially, we decided to test the applicability of the ASM methodology in its original formulation,[6b,9c] which is using the LPB equation to calculate the $G_{\text{polar solvation}}$ term (salt concentration equal to zero). For that purpose, we calculated the $\Delta\Delta G_{\text{binding}}$ value for a set of 25 structures generated from the last 2 ns of MD simulations in explicit solvent. Throughout this work, we performed the various calculations with dielectric constants ranged from 1 to 10 to access the importance of the dielectric constant in the determination. These will be presented in all the points discussed in the paper. The average error between the calculated and the experimental values for all the mutations analyzed as well as for charged, polar, and nonpolar groups of residues is listed in Table 2. A detailed list of the calculated values can be found in Supporting Information (Table SI-1). We have obtained average errors ranging from 72.88 (ε = 1) to 9.92 (ε = 10) kcal/mol for all the mutations. We have obtained an overestimation of $\Delta\Delta G_{\text{binding}}$ for charged residues; as for polar residues, we obtained lower mean error values and managed to correctly predict some HS and NS. This overestimation that also affects the polar and nonpolar values, although to a lesser extent, was common to the various determinations performed, meaning it was a problem in the core of the computational methodology. Indeed, the LPB equation, normally used in the traditional ASM methodology for the determination of binding free energies is not the most appropriate for dealing with highly charged systems, such as the protein−DNA systems under study, where electrostatic and ionic interactions prevail and whose value of binding free energy is dependent on the salt concentration. Taking this into account, we decided to use the NLPB equation instead of the LPB equation, which application for the treatment of systems involving DNA has been previously described.[34]

We then used the NLPB equation implemented in Delphi program for the calculation of the binding free energy upon alanine mutation for two different salt concentrations (0.010 M and 0.145 M). The complete results can be found in Supporting Information (Tables SI-II and SI-III) and the summarized results in Tables 3 and 4, respectively. The principal, and most important, difference we observed when comparing the results obtained through the LPB and the NLPB equations is the behavior of the charged residues. No longer are the $\Delta\Delta G_{\text{binding}}$ values overestimated. To the contrary, we observed an approximation to the experimental values. The average error of the calculated value for these residues decreased to the range from 7.76 (ε = 1) to 7.29 (ε = 10) kcal/mol and 7.51 to 7.12 kcal/mol for a salt concentrations of 0.010 M and 0.145 M, respectively. Previously, the average error varied from 247.19 (ε = 1) to 28.92 (ε = 10) kcal/mol. It is an obvious difference and more relevant when we compared the average errors for all the 25 mutations in study: previously, we had 72.88 (ε = 1) and 9.92 (ε = 10) kcal/mol, and now, we obtained 5.58 (ε = 1) and 3.84 (ε = 10) for a salt concentration of 0.010 M and 5.57 and 3.79 kcal/mol for a salt concentration of 0.145 M. As both values of salt concentration give similar results, we will opt for 0.145 M, which is the typical value in the interior of a biological cell.[34] This way, the advantages of using the NLPB equation to calculated the $G_{\text{polar-solvation}}$ term for highly charged systems such as the ones involving DNA become clear.

2. Influence of the Energy Terms Included in the Calculation. Although the results were greatly improved by taking into account the salt concentration, they are still far from being perfect. We have to keep in mind that we are dealing with very charged systems, composed by dozen of charged phosphates groups and with a high number of positively charged amino acids at their interfaces. Therefore, we have decided to separate the electrostatics contributions of each residue for the

Table 3. Results of Average Errors and Statistical Tests Obtained with the NLPB Equation for 25 Structures of the Last 2 ns of the MD Simulation in Explicit Solvent, and with a Salt Concentration of 0.010 M

	$\lvert\Delta\Delta G_{\text{MM-PBSA}} - \Delta\Delta G_{\text{Experimental}}\rvert$, kcal mol^{-1}									
	$\varepsilon 1$	$\varepsilon 2$	$\varepsilon 3$	$\varepsilon 4$	$\varepsilon 5$	$\varepsilon 6$	$\varepsilon 7$	$\varepsilon 8$	$\varepsilon 9$	$\varepsilon 10$
all	5.58	3.65	3.76	3.79	3.80	3.82	3.83	3.84	3.84	3.84
charged	7.76	4.15	5.65	6.33	6.70	6.93	7.07	7.17	7.24	7.29
polar	4.66	2.93	2.35	2.06	1.88	1.81	1.76	1.72	1.69	1.66
nonpolar	4.90	4.50	4.35	4.29	4.25	4.22	4.20	4.18	4.17	4.16
	$\lvert\Delta\Delta G_{\text{MM-PBSA}} - \Delta\Delta G_{\text{Experimental}}\rvert < 1$/%									
	$\varepsilon 1$	$\varepsilon 2$	$\varepsilon 3$	$\varepsilon 4$	$\varepsilon 5$	$\varepsilon 6$	$\varepsilon 7$	$\varepsilon 8$	$\varepsilon 9$	$\varepsilon 10$
all	8.0	12.0	12.0	24.0	24.0	24.0	20.0	20.0	20.0	20.0
charged	14.3	28.6								
polar	8.3	16.7	25.0	50.0	50.0	50.0	50.0	41.7	41.7	41.7
nonpolar										
	statistical tests/all									
	$\varepsilon 1$	$\varepsilon 2$	$\varepsilon 3$	$\varepsilon 4$	$\varepsilon 5$	$\varepsilon 6$	$\varepsilon 7$	$\varepsilon 8$	$\varepsilon 9$	$\varepsilon 10$
P	0.19	0.29	0.29	0.31	0.36	0.36	0.38	0.38	0.38	0.38
R	0.50	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
F1	0.27	0.44	0.44	0.46	0.50	0.50	0.53	0.53	0.53	0.53
accuracy	0.36	0.48	0.48	0.52	0.60	0.60	0.64	0.64	0.64	0.64
specificity	0.32	0.37	0.37	0.42	0.53	0.53	0.58	0.58	0.58	0.58

Table 4. Results of Average Errors and Statistical Tests Obtained with the NLPB Equation for 25 Structures of the Last 2 ns of the MD Simulation in Explicit Solvent, and with an Ionic Concentration of 0.145 M

	$\lvert\Delta\Delta G_{\text{MM-PBSA}} - \Delta\Delta G_{\text{Experimental}}\rvert$, kcal mol^{-1}									
	$\varepsilon 1$	$\varepsilon 2$	$\varepsilon 3$	$\varepsilon 4$	$\varepsilon 5$	$\varepsilon 6$	$\varepsilon 7$	$\varepsilon 8$	$\varepsilon 9$	$\varepsilon 10$
all	5.57	3.70	3.66	3.73	3.74	3.76	3.78	3.79	3.79	3.79
charged	7.51	4.27	5.40	6.09	6.47	6.71	6.87	6.98	7.06	7.12
polar	4.70	2.94	2.36	2.07	1.89	1.81	1.76	1.72	1.69	1.67
nonpolar	5.02	4.56	4.21	4.32	4.27	4.24	4.22	4.20	4.19	4.17
	$\lvert\Delta\Delta G_{\text{MM-PBSA}} - \Delta\Delta G_{\text{Experimental}}\rvert < 1$/%									
	$\varepsilon 1$	$\varepsilon 2$	$\varepsilon 3$	$\varepsilon 4$	$\varepsilon 5$	$\varepsilon 6$	$\varepsilon 7$	$\varepsilon 8$	$\varepsilon 9$	$\varepsilon 10$
all	8.0	8.0	8.0	24.0	24.0	24.0	20.0	20.0	20.0	20.0
charged	14.3	14.3								
polar	8.3	8.3	16.7	50.0	50.0	50.0	41.7	41.7	41.7	41.7
nonpolar										
	statistical tests, % (all)									
	$\varepsilon 1$	$\varepsilon 2$	$\varepsilon 3$	$\varepsilon 4$	$\varepsilon 5$	$\varepsilon 6$	$\varepsilon 7$	$\varepsilon 8$	$\varepsilon 9$	$\varepsilon 10$
P	0.19	0.29	0.29	0.31	0.36	0.36	0.38	0.38	0.38	0.38
R	0.50	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
F1	0.27	0.44	0.44	0.46	0.50	0.50	0.53	0.53	0.53	0.53
accuracy	0.36	0.48	0.48	0.52	0.60	0.60	0.62	0.64	0.64	0.64
specificity	0.32	0.37	0.37	0.42	0.53	0.53	0.56	0.58	0.58	0.58

binding free energy and measure the sum of the two major energetic terms at play $\Delta\Delta E_{\text{ele}}$ and $\Delta\Delta G_{\text{polar solvation}}$. Comparison of Tables 5 and SI-IV to Tables 4 and SI-III clearly shows an overall improvement. The average error decreased from 5.57 ($\varepsilon = 1$) and 3.79 ($\varepsilon = 10$) kcal/mol to 5.11 ($\varepsilon = 1$) and 2.32 ($\varepsilon = 10$) kcal/mol. Also, the increased reliability of the method for nonpolar residues and to charged residues is notorious. Besides the decrease of the average errors, there is an increase in the amount of values with errors below 1%, especially for nonpolar residues, as well as better results in regard to the statistical tests. The use of the electrostatics terms to measure the contribution of specific residues at an interface was already introduced by Sheinerman et al.[38] and Koskolff et al.[39] Here, we applied it to protein–DNA systems.

3. Influence of Solvent Representation. Until now, we have observed that the best computational method to accurately calculate the $\Delta\Delta G_{\text{binding}}$ values upon alanine mutation was to perform an explicit water MD simulation while calculating the $\Delta\Delta E_{\text{ele}} + \Delta\Delta G_{\text{polar solvation}}$ values for 25 structures of the last 2 ns of the simulation. Therefore, we have also tested the influence of the MD solvent representation in the accurate detection of HS (explicit water vs implicit solvent). Results are presented in Table 6 and SI-V. There is a loss of accuracy with the average errors passing from 5.11 ($\varepsilon = 1$) and 2.32 ($\varepsilon = 10$) kcal/mol to 5.35 ($\varepsilon = 1$) and 2.83 ($\varepsilon = 10$) kcal/mol, which is more notable with charged and polar residues. The decrease of accuracy is also noted when we compare the results for the statistical tests. The highly polar and charged character of these interfaces explains once more the preference for an explicit water MD simulation in which the water influence in the hot-spots microenvironment can be assessed.

Table 5. Results of Average Errors and Statistical Tests Obtained with the NLPB Equation for 25 Structures of the Last 2 ns of the MD Simulation in Explicit Solvent, with a Salt Concentration of 0.145 M, Using the $\Delta\Delta E_{\text{ele}} + \Delta\Delta G_{\text{polar solvation}}$ Terms

$\lvert\Delta\Delta E_{\text{ele}} + \Delta\Delta G_{\text{polar solvation}} - \Delta\Delta G_{\text{Experimental}}\rvert$, kcal mol^{-1}										
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
all	5.11	2.96	2.47	2.23	2.25	1.96	2.28	2.29	2.30	2.32
charged	9.10	5.07	4.28	3.98	4.34	3.45	4.73	4.85	4.93	4.98
polar	3.86	2.24	1.75	1.57	1.47	1.41	1.35	1.32	1.30	1.30
nonpolar	2.95	1.93	1.81	1.48	1.39	1.33	1.29	1.26	1.25	1.23
$\lvert\Delta\Delta E_{\text{ele}} + \Delta\Delta G_{\text{polar solvation}} - \Delta\Delta G_{\text{Experimental}}\rvert < 1$/%										
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
all	20.0	28.0	36.0	44.0	44.0	48.0	40.0	36.0	36.0	36.0
charged	14.3		14.3	28.6		14.3				
polar	16.7	33.3	41.7	50.0	66.7	66.7	58.3	50.0	50.0	50.0
nonpolar	33.3	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
statistical tests, % (all)										
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
P	0.18	0.30	0.33	0.40	0.50	0.50	0.50	0.50	0.50	0.50
R	0.33	0.50	0.50	0.67	0.50	0.50	0.50	0.50	0.50	0.50
F1	0.24	0.38	0.40	0.50	0.50	0.50	0.50	0.50	0.50	0.50
accuracy	0.48	0.60	0.64	0.68	0.76	0.76	0.76	0.76	0.76	0.76
specificity	0.53	0.63	0.68	0.68	0.84	0.84	0.84	0.84	0.84	0.84

Table 6. Results of Average Errors and Statistical Tests Obtained with the NLPB Equation for 25 Structures of the Last 2 ns of the MD Simulation in Implicit Solvent, and with a Salt Concentration of 0.145 M

$\lvert\Delta\Delta E_{\text{ele}} + \Delta\Delta G_{\text{polar solvation}} - \Delta\Delta G_{\text{Experimental}}\rvert$, kcal mol^{-1}										
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
all	5.35	3.93	3.51	3.29	3.14	3.16	2.97	2.92	2.87	2.83
charged	5.66	6.24	6.40	6.45	6.45	6.85	6.40	6.37	6.34	6.31
polar	5.47	3.12	2.44	2.09	1.89	1.76	1.66	1.59	1.53	1.49
nonpolar	4.72	2.87	2.27	1.97	1.80	1.68	1.60	1.54	1.49	1.45
$\lvert\Delta\Delta E_{\text{ele}} + \Delta\Delta G_{\text{polar solvation}} - \Delta\Delta G_{\text{Experimental}}\rvert < 1$/%										
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
all	28.0	20.0	20.0	16.0	20.0	28.0	28.0	28.0	28.0	32.0
charged	28.6									
polar	16.7	25.0	25.0	16.7	25.0	41.7	41.7	41.7	41.7	50.0
nonpolar	50.0	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
statistical tests, % (all)										
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
P	0.29	0.25	0.25	0.25	0.30	0.38	0.43	0.50	0.50	0.50
R	0.67	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
F1	0.40	0.33	0.33	0.33	0.38	0.43	0.46	0.50	0.50	0.50
accuracy	0.52	0.52	0.52	0.52	0.60	0.68	0.72	0.76	0.76	0.76
specificity	0.47	0.53	0.53	0.53	0.63	0.74	0.79	0.84	0.84	0.84

4. Influence of the Number of Snapshots Analyzed. In order to investigate the influence of the number of structures analyzed on the computational predictions, we conducted a comparative study by using two different sets: 25 snapshots or 50 snapshots taken from the last 2 ns of the explicit MD simulations. Tables 7 and SI-VI list all the results obtained. There is a slight decreased of accuracy, especially concerning the nonpolar residues that is, however, not statistically meaningful. The computational time involved in solving the nonlinear form of the Poisson–Boltzmann equation with Delphi is considerably high, and therefore, the use of a higher number of snapshots for the calculation of the various $\Delta\Delta G_{\text{binding}}$ values is not worth it. This fact was also observed by Hou et al. in protein–ligand systems.[40]

5. Influence of the MD Simulation Length. We also tested the influence of the length of the MD simulation time on the correct prediction of the $\Delta\Delta G_{\text{binding}}$ values. The values listed in Tables 8 and SI-VII are not statistically different from the ones obtained for a longer MD simulation. However, it seems that the charged residues are more accurately detected in the longer MD simulation.

We have also investigated how a higher number of snapshots taken from a larger range of conformational sampling would influence the computational performance of ASM. For that purpose we have analyzed 100 snapshots from the 2–8 ns time scale and the results are presented in Tables 9 and SI-VIII. As mentioned in the Methodological Part (Analysis section) the RMSDs of all the complexes were very low (under 1 Å), and therefore, data could be retrieved from the MD as soon as at the 2 ns time-scale. This analysis enforces our conclusions of the previous point, that a higher number of snapshots taken

Table 7. Results of Average Errors and Statistical Tests Obtained with the NLPB Equation for 50 Structures of the Last 2 ns of the MD Simulation in Explicit Solvent, and with a Salt Concentration of 0.145 M

	$|\Delta\Delta E_{\mathrm{ele}} + \Delta\Delta G_{\mathrm{polar\ solvation}} - \Delta\Delta G_{\mathrm{Experimental}}|$, kcal mol^{-1}									
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
all	4.23	2.64	2.51	2.48	2.45	2.25	2.42	2.41	2.40	2.40
charged	5.25	3.71	4.37	4.75	4.95	4.42	5.14	2.19	5.23	5.25
polar	3.94	2.18	1.74	1.57	1.46	1.39	1.34	1.31	1.31	1.30
nonpolar	3.59	2.30	1.87	1.66	1.53	1.43	1.38	1.34	1.30	1.28
	$|\Delta\Delta E_{\mathrm{ele}} + \Delta\Delta G_{\mathrm{polar\ solvation}} - \Delta\Delta G_{\mathrm{Experimental}}| < 1$/%									
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
all	12.0	32.0	32.0	36.0	40.0	44.0	36.0	32.0	32.0	32.0
charged	14.3	28.6	14.3			14.3				
polar	16.7	41.7	41.7	58.3	66.7	66.7	58.3	50.0	50.0	50.0
nonpolar		16.7	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
	statistical tests, % (all)									
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
P	0.27	0.33	0.40	0.33	0.43	0.50	0.50	0.50	0.50	0.50
R	0.50	0.67	0.67	0.50	0.50	0.50	0.50	0.50	0.50	0.50
F1	0.35	0.44	0.50	0.40	0.46	0.50	0.50	0.50	0.50	0.50
accuracy	0.56	0.60	0.68	0.64	0.72	0.76	0.76	0.76	0.76	0.76
specificity	0.58	0.58	0.68	0.68	0.79	0.84	0.84	0.84	0.84	0.84

Table 8. Results of Average Errors and Statistical Tests Obtained with the NLPB Equation for 25 Structures Taken from the 4 to 6 ns from the MD Simulations in Explicit Solvent, and with a Salt Concentration of 0.145 M

	$|\Delta\Delta E_{\mathrm{ele}} + \Delta\Delta G_{\mathrm{polar\ solvation}} - \Delta\Delta G_{\mathrm{Experimental}}|$, kcal mol^{-1}									
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
all	4.21	2.62	2.51	2.49	2.46	2.26	2.43	2.42	2.41	2.42
charged	5.24	3.77	4.47	4.84	5.03	4.48	5.22	5.26	5.28	5.32
polar	4.06	2.17	1.73	1.56	1.45	1.38	1.33	1.30	1.29	1.29
nonpolar	3.32	2.17	1.78	1.61	1.49	1.41	1.36	1.33	1.31	1.29
	$|\Delta\Delta E_{\mathrm{ele}} + \Delta\Delta G_{\mathrm{polar\ solvation}} - \Delta\Delta G_{\mathrm{Experimental}}| < 1$/%									
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
all	12.0	40.0	32.0	36.0	32.0	40.0	36.0	36.0	32.0	32.0
charged	28.6	28.6								
polar	8.3	50.0	50.0	58.3	50.0	66.7	58.3	58.3	50.0	50.0
nonpolar		33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
	statistical tests, % (all)									
	ε1	ε2	ε3	ε4	ε5	ε6	ε7	ε8	ε9	ε10
P	0.33	0.33	0.36	0.40	0.50	0.50	0.50	0.50	0.50	0.50
R	0.67	0.67	0.67	0.67	0.50	0.50	0.50	0.50	0.50	0.50
F1	0.44	0.44	0.47	0.50	0.50	0.50	0.50	0.50	0.50	0.50
accuracy	0.60	0.60	0.64	0.68	0.76	0.76	0.76	0.76	0.76	0.76
specificity	0.58	0.58	0.63	0.68	0.84	0.84	0.84	0.84	0.84	0.84

from a small (or larger ensemble) does not increase the accuracy of the values obtained. Therefore, the analysis of 25 snapshots from the last 2 ns window of the MD simulations seems the right choice to ensure the best possible accuracy in the lowest computational time.

6. Influence of the Force Field. Parm94 and parm99 versions of the AMBER force field were shown to perform very well in a 10 ns MD window. However, as they can potentially lead to DNA distortions in large scale MD simulations, new force fields versions such as parmbsc0 have emerged.[21] To investigate the influence of the force field in the accuracy of our calculations, we have performed a MD simulation of the Nuclear Protein EBNA/DNA complex (PDB ID: 1B3T[12]) using parmbsc0. The results (Table 10) showed us that for our particular case, the previous force field version works slightly better as the average error ($|\Delta\Delta E_{\mathrm{ele}} + \Delta\Delta G_{\mathrm{polar\ solvation}} - \Delta\Delta G_{\mathrm{Experimental}}|$) is lower for the three mutations studied. This in agreement with Xu et al. that stated that the latest force field does not necessarily performs better than the newest as the fitting for new force field parameters focus in the conformation and dynamics of proteins and not in the binding free energies.[41]

7. Number of Analyzed MD Trajectories. New sampling was collected for the three mutants of the system, Nuclear Protein EBNA/DNA complex (PDB ID: 1B3T[12]). The results are presented in Table 11. We tried a "two MD simulation protocol" based on running two separate trajectories, one for the wild-type and one for the mutant. This changes the protocol from 1 MD/system to 1 MD/mutant, and therefore, it leads to a considerable increase of the computational time involved. Table 11 shows an increase of the computational

Table 9. Results of Average Errors and Statistical Tests Obtained with the NLPB Equation for 100 Structures Taken from the 2 to 8 ns from the MD Simulations in Explicit Solvent, and with a Salt Concentration of 0.145 M

$\|\Delta\Delta E_{\text{ele}} + \Delta\Delta G_{\text{polar solvation}} - \Delta\Delta G_{\text{Experimental}}\|$, kcal mol^{-1}										
	$\varepsilon 1$	$\varepsilon 2$	$\varepsilon 3$	$\varepsilon 4$	$\varepsilon 5$	$\varepsilon 6$	$\varepsilon 7$	$\varepsilon 8$	$\varepsilon 9$	$\varepsilon 10$
all	4.18	2.62	2.48	2.46	2.44	2.15	2.41	2.31	2.40	2.40
charged	5.18	3.69	4.35	4.75	4.95	4.10	5.16	4.88	5.25	5.27
polar	3.98	2.21	1.73	1.55	1.45	1.38	1.33	1.30	1.29	1.29
nonpolar	3.42	2.21	1.81	1.61	1.50	1.41	1.36	1.32	1.29	1.26
$\|\Delta\Delta E_{\text{ele}} + \Delta\Delta G_{\text{polar solvation}} - \Delta\Delta G_{\text{Experimental}}\| < 1$/%										
	$\varepsilon 1$	$\varepsilon 2$	$\varepsilon 3$	$\varepsilon 4$	$\varepsilon 5$	$\varepsilon 6$	$\varepsilon 7$	$\varepsilon 8$	$\varepsilon 9$	$\varepsilon 10$
all	16.0	36.0	32.0	32.0	36.0	36.0	36.0	32.0	32.0	32.0
charged	14.3	28.6	14.3							
polar	16.7	50.0	41.7	50.0	58.3	58.3	58.3	50.0	50.0	50.0
nonpolar	16.7	16.7	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
statistical tests, % (all)										
	$\varepsilon 1$	$\varepsilon 2$	$\varepsilon 3$	$\varepsilon 4$	$\varepsilon 5$	$\varepsilon 6$	$\varepsilon 7$	$\varepsilon 8$	$\varepsilon 9$	$\varepsilon 10$
P	0.27	0.33	0.36	0.40	0.43	0.50	0.50	0.40	0.50	0.50
R	0.50	0.67	0.67	0.67	0.50	0.50	0.50	0.33	0.50	0.50
F1	0.35	0.44	0.47	0.50	0.46	0.50	0.50	0.36	0.50	0.50
accuracy	0.56	0.60	0.64	0.68	0.72	0.76	0.76	0.72	0.76	0.76
specificity	0.58	0.58	0.63	0.68	0.79	0.84	0.84	0.84	0.84	0.84

Table 10. Results of Average Errors Obtained for the Complex PDB ID 1B3T (with the NLPB Equation for 25 Structures Taken from the 6 to 8 ns from the MD Simulations in Explicit Solvent, and with a Salt Concentration of 0.145 M) for Two Different Force Fields[a]

$\|\Delta\Delta E_{\text{ele}} + \Delta\Delta G_{\text{polar solvation}} - \Delta\Delta G_{\text{Experimental}}\|$, kcal mol^{-1}										
	$\varepsilon 1$	$\varepsilon 2$	$\varepsilon 3$	$\varepsilon 4$	$\varepsilon 5$	$\varepsilon 6$	$\varepsilon 7$	$\varepsilon 8$	$\varepsilon 9$	$\varepsilon 10$
all	7.82	6.90	4.49	3.23	3.67	1.92	4.17	4.30	4.38	4.44
charged	10.01	9.97	6.73	4.62	5.14	2.41	5.70	5.84	5.92	5.97
polar	3.43	0.77	0.01	0.47	0.75	0.95	1.09	1.22	1.31	1.40
nonpolar										
$\|\Delta\Delta E_{\text{ele}} + \Delta\Delta G_{\text{polar solvation}} - \Delta\Delta G_{\text{Experimental}}\|$, kcal mol^{-1}										
	$\varepsilon 1$	$\varepsilon 2$	$\varepsilon 3$	$\varepsilon 4$	$\varepsilon 5$	$\varepsilon 6$	$\varepsilon 7$	$\varepsilon 8$	$\varepsilon 9$	$\varepsilon 10$
all	9.73	5.40	5.63	5.69	5.62	4.27	5.49	5.42	5.37	5.31
charged	13.72	8.05	8.19	8.09	7.88	5.77	7.54	7.40	7.28	7.16
polar	1.75	0.11	0.51	0.87	1.10	1.25	1.37	1.47	1.55	1.61
nonpolar										

[a] The ff94 and the parmbsc0 force field versions are on top and bottom, respectively.

error. We believe that a single trajectory benefits from error cancelation as the region of the conformational space accessed by the mutant and wild-type is the same. Maybe if the calculations were run for long enough to better explore the conformational space, it would be possible to obtain better results. The same was already observed for protein−protein systems.[42]

8. Influence of Using Different PB Solvers. We have also investigated how the use of the *pbsa* module in AMBER package instead of Delphi would influence the accuracy of the results. The results are presented in Tables 12 and SI-IX. The overall performance measured by the statistical tests is very similar to the one obtained using Delphi (Table 5). However, the $|\Delta\Delta E_{\text{ele}} + \Delta\Delta G_{\text{polar solvation}} - \Delta\Delta G_{\text{Experimental}}|$ values attained for the charged residues are higher than the ones calculated with Delphi.

9. MM-PBSA vs MM-GBSA. To reduce the computational effort, we have also replaced the PB continuum calculation, one of the most time-consuming steps of the MM-PBSA approach, by the Generalized Born (GB) solvent model. Results are presented in Tables 13 and SI-X. In this work, we observed that MM-GBSA performances slightly better than the theoretically more rigorous MM-PBSA. The improvement achieved is higher for the charged and polar residues, typical HS residues on PDI. As GB is computational less expensive than PB, we believe that it should be used in the final formulation of our method. To test if our previous conclusions were independent of the PB/GB use for the calculation of the $\Delta\Delta G_{\text{polar solvation}}$ term, we have also perform various tests, and the results can be consulted in SI: (i) GB for 25 structures taken from the last 2 ns from the MD simulations in explicit solvent, with salt concentration 0.145 M and all energetic terms included (Table SI-XI), (ii) the same conditions but without adding the salt effect (Table SI-XII), (iii) the same conditions but with salt concentration of 0.010 M (Table SI-XIII), (iv) GB for 25 structures taken from the last 2 ns from the MD simulations in explicit solvent and with a salt concentration of 0 M using the $\Delta\Delta E_{\text{ele}}$ + $\Delta\Delta_{\text{Gpolar solvation}}$ terms (Table SI-XIV), (v) the same conditions but with a salt concentration of 0.010 M (Table SI-XV), (vi) GB for 25 structures taken from the last 2 ns from the MD

Table 11. Results of Average Errors Obtained for the Complex PBD ID 1B3T with the NLPB Equation for 25 Structures Taken from the 6 to 8 ns from the MD Simulations in Explicit Solvent, and with a Salt Concentration of 0.145 M and Multiple Trajectories

	$|\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation} - \Delta\Delta G_{Experimental}|$, kcal mol^{-1}									
	$\varepsilon1$	$\varepsilon2$	$\varepsilon3$	$\varepsilon4$	$\varepsilon5$	$\varepsilon6$	$\varepsilon7$	$\varepsilon8$	$\varepsilon9$	$\varepsilon10$
all	88.91	79.55	63.72	49.20	40.31	34.38	30.10	26.71	24.34	22.28
charged	83.50	71.09	54.50	43.05	36.07	31.38	27.97	25.40	23.35	21.71
polar	99.74	96.49	82.16	61.50	48.80	40.39	34.36	29.33	26.30	23.43
nonpolar										

Table 12. Results of Average Errors and Statistical Tests Obtained with the NLPB Equation for 25 Structures Taken from the Last 2 ns from the MD Simulations in Explicit Solvent, with a Salt Concentration of 0.145 M Using the *pbsa* Module of the AMBER Package

	$|\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation} - \Delta\Delta G_{Experimental}|$, kcal mol^{-1}									
	$\varepsilon1$	$\varepsilon2$	$\varepsilon3$	$\varepsilon4$	$\varepsilon5$	$\varepsilon6$	$\varepsilon7$	$\varepsilon8$	$\varepsilon9$	$\varepsilon10$
all	9.84	42.44	37.58	31.54	26.86	23.29	20.53	18.34	16.64	15.19
charged	28.57	145.11	128.64	107.81	91.50	79.04	69.37	61.70	55.71	50.63
polar	2.69	2.79	2.38	2.05	1.84	1.71	1.63	1.56	1.52	1.48
nonpolar	2.30	1.94	1.77	1.56	1.47	1.40	1.36	1.31	1.29	1.26
	$|\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation} - \Delta\Delta G_{Experimental}| < 1$/%									
	$\varepsilon1$	$\varepsilon2$	$\varepsilon3$	$\varepsilon4$	$\varepsilon5$	$\varepsilon6$	$\varepsilon7$	$\varepsilon8$	$\varepsilon9$	$\varepsilon10$
all	8.0	24.0	24.0	24.0	24.0	24.0	32.0	32.0	28.0	32.0
charged										
polar	16.7	25.0	25.0	25.0	25.0	33.3	50.0	50.0	41.7	50.0
nonpolar		50.0	50.0	50.0	50.0	33.3	33.3	33.3	33.3	33.3
	statistical tests, % (all)									
	$\varepsilon1$	$\varepsilon2$	$\varepsilon3$	$\varepsilon4$	$\varepsilon5$	$\varepsilon6$	$\varepsilon7$	$\varepsilon8$	$\varepsilon9$	$\varepsilon10$
P	0.42	0.38	0.33	0.44	0.44	0.44	0.38	0.43	0.43	0.50
R	1	1	0.80	0.80	0.80	0.80	0.60	0.60	0.60	0.60
F1	0.59	0.56	0.47	0.57	0.57	0.57	0.46	0.50	0.50	0.54
accuracy	0.72	0.68	0.64	0.76	0.76	0.76	0.72	0.76	0.76	0.80
specificity	0.65	0.60	0.60	0.75	0.75	0.75	0.75	0.80	0.80	0.85

Table 13. Results of Average Errors and Statistical Tests Obtained with GB for 25 Structures Taken from the Last 2 ns from the MD Simulations in Explicit Solvent, with a Salt Concentration of 0.145 M

	$|\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation} - \Delta\Delta G_{Experimental}|$, kcal mol^{-1}									
	$\varepsilon1$	$\varepsilon2$	$\varepsilon3$	$\varepsilon4$	$\varepsilon5$	$\varepsilon6$	$\varepsilon7$	$\varepsilon8$	$\varepsilon9$	$\varepsilon10$
all	11.14	5.47	3.61	2.77	2.28	1.98	1.78	1.65	1.54	1.47
charged	23.59	11.98	8.14	6.22	5.06	4.29	3.74	3.33	3.01	2.75
polar	7.69	3.60	2.11	1.54	1.24	1.06	0.96	0.93	0.89	0.90
nonpolar	3.54	1.63	1.33	1.19	1.12	1.13	1.13	1.13	1.13	1.13
	$|\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation} - \Delta\Delta G_{Experimental}| < 1$/%									
	$\varepsilon1$	$\varepsilon2$	$\varepsilon3$	$\varepsilon4$	$\varepsilon5$	$\varepsilon6$	$\varepsilon7$	$\varepsilon8$	$\varepsilon9$	$\varepsilon10$
all	16.0	40.0	52.0	60.0	56.0	56.0	64.0	64.0	64.0	64.0
charged	14.3	28.6	28.6	28.6	28.6	28.6	28.6	28.6	28.6	28.6
polar	16.7	33.3	50.0	66.7	66.7	66.7	83.3	83.3	83.3	83.3
nonpolar	16.7	66.7	83.3	83.3	66.7	66.7	66.7	66.7	66.7	66.7
	statistical tests, % (all)									
	$\varepsilon1$	$\varepsilon2$	$\varepsilon3$	$\varepsilon4$	$\varepsilon5$	$\varepsilon6$	$\varepsilon7$	$\varepsilon8$	$\varepsilon9$	$\varepsilon10$
P	0.35	0.43	0.54	0.54	0.60	0.50	0.57	0.80	0.80	0.80
R	1	1	1	1	1	0.67	0.67	0.67	0.67	0.67
F1	0.52	0.60	0.71	0.71	0.75	0.57	0.61	0.73	0.73	0.73
accuracy	0.56	0.68	0.80	0.80	0.84	0.76	0.80	0.88	0.88	0.88
specificity	0.42	0.58	0.74	0.74	0.79	0.79	0.84	0.95	0.95	0.95

simulations in implicit solvent and with a salt concentration of 0.145 M, using the $\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation}$ terms (Table SI-XVI), and (vii) GB for 100 structures taken from the 2−8 ns from the MD simulations in explicit solvent and with a salt concentration of 0.145 M, using the $\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation}$ terms (Table SI-XVII). Analyses of these tables demonstrated that for the

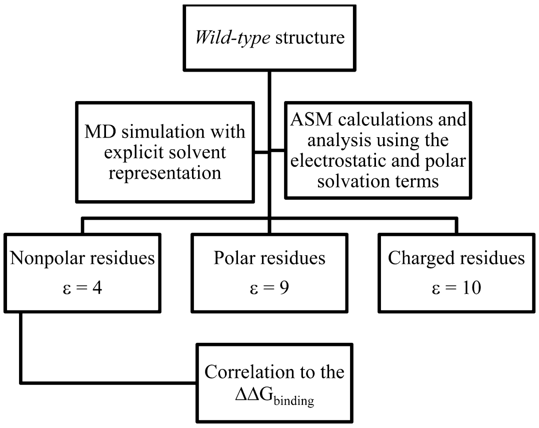

Wild-type structure
MD simulation with explicit solvent representation
ASM calculations and analysis using the electrostatic and polar solvation terms
Nonpolar residues ε = 4
Polar residues ε = 9
Charged residues ε = 10
Correlation to the $\Delta\Delta G_{binding}$

Figure 2. Schematic representation of the final method formulation for the determination of $\Delta\Delta G_{binding}$.

MM-GBSA approach the same conclusions hold true: the use of only the $\Delta\Delta E_{ele}$ + $\Delta\Delta G_{polar\ solvation}$ terms, a salt concentration of 0.145 M, performing MD in explicit solvent and using 25 snapshots of the last 2 ns conduces to more accurate results.

10. Influence of the Internal Dielectric Constant Value. The uniform dielectric continuum model is a very crude description of the highly heterogeneous protein interior. Therefore, the dielectric coefficient becomes a model dependent scaling factor without a transparent physical significance. So, the internal dielectric constant ($\varepsilon_{internal}$) is a means of accounting for responses to an electric field that are not treated explicitly. This response is not universal and depends on the constituting amino acid residues. As side chain reorientation and dipolar reorganization arising from conformational transitions upon mutation is not included explicitly in the formalism, we have tried different internal dielectric values from 1 to 10. Like the ASM of PPI, we observed that the internal dielectric constant should not be implemented as a homogeneous constant. Therefore, by inspection of Table 13, we try to obtain the lowest average error $|\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation} - \Delta\Delta G_{Experimental}|$ and the highest statistical values for the three different groups of residues. We concluded that three different internal dielectric constants should be used, dependent on the type of amino acid mutated to an alanine. By using $\varepsilon_{internal}$ = 4 for the nonpolar residues, $\varepsilon_{internal}$ = 9 for the polar amino-acids and $\varepsilon_{internal}$ = 10 for the charged residues, it was possible to obtain an excellent agreement with the experimental results. The scaling of this macroscopic parameter to larger values when larger reorganizations are expected mimics the lack of conformational sampling and relaxation due to the mutation for an alanine. So, overall we found that an accurate calculation of the free binding energy difference upon alanine mutation of protein−DNA systems could be achieved by performing 8 ns explicit MD simulations and analyzed 25 snapshots of the last 2 ns, using the Generalized Born model with a salt concentration of 0.145 M, summing only the $\Delta\Delta E_{ele}$ + $\Delta\Delta G_{polar\ solvation}$ energetic contributions and using a set of three dielectric constant values (Figure 2).

The performance of the method can be assessed by using different quantitative measures such as Precision, Recall, Specificity, F1-measure, and Accuracy defined at the methodology section. Recall and Specificity are important since they emphasize the predicting power of both HS and NS, respectively. Precision determines how accurate the positive predictions are. F1-measure and Accuracy give an overall performance of the method. For our data set we have obtained a value of 0.76, 0.57, 0.67, 0.50, and 0.79 for Accuracy, F1-measure, Recall, Precision, and Specificity, respectively, which is higher than the values reported (reviewed in ref 7b). Our average error $|((\Delta\Delta E_{ele}+\Delta\Delta G_{polar\ solvation}) - \Delta\Delta G_{binding\text{-}experimental})|$ for the training set is 1.48 kcal/mol. 64% of the values have errors under 1 kcal/mol.

■ TEST SET

We have applied the final method formulation to the test-set. The results are listed in Tables 14 and SI-XVIII. For the test-set our average error $|((\Delta\Delta E_{ele}+\Delta\Delta G_{polar\ solvation}) - \Delta\Delta G_{binding\text{-}experimental})|$ is 1.58 kcal/mol with 50.9% of the values having errors under 1 kcal/mol. We have obtained a value of 0.74, 0.59, 0.46, 0.83, and 0.94 for Accuracy, F1-measure, Recall, Precision, and Specificity, respectively.

Table 14. Results of Average Errors and Statistical Tests for the Test Set, Using the Final Formulation of the Developed Method

	all	charged	polar	nonpolar
$\lvert\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation} - \Delta\Delta G_{Experimental}\rvert$, kcal mol^{-1}	1.58	1.71	1.21	1.71
$\lvert\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation} - \Delta\Delta G_{Experimental}\rvert < 1/\%$	50.9	44.0	64.3	50.0
P	0.83	0.80		1
R	0.46	0.73		0.25
F1	0.59	0.76		0.40
accuracy	0.74	0.80	0.79	0.57
specificity	0.94	0.86	1	1

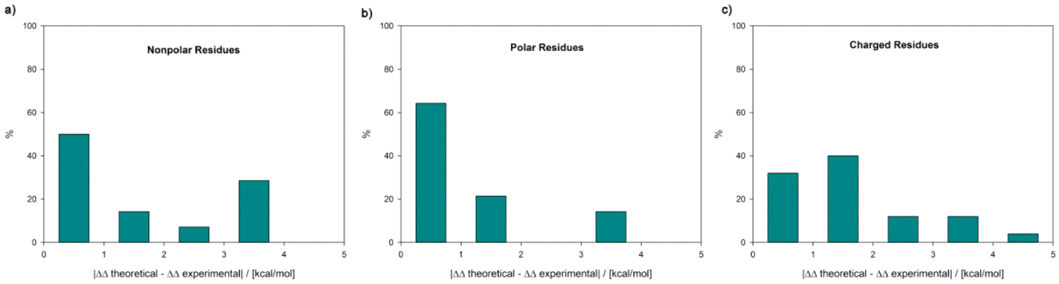


Figure 3. Probability (%) of the $|\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation} - \Delta\Delta G_{Experimental}|$ values for the nonpolar, polar and charged residues are under 5.0 kcal/mol.

Table 15. Results of Average Errors and Statistical Tests for the Data Set, Using the Final Formulation of the Developed Method

	all	charged	polar	nonpolar
$\|\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation} - \Delta\Delta G_{Experimental}\|$, kcal mol^{-1}	1.55	1.94	1.06	1.55
$\|\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation} - \Delta\Delta G_{Experimental}\| < 1/\%$	56.4	40.6	73.1	60.0
P	0.78	0.77	1	0.67
R	0.50	0.77	0.17	0.25
F1	0.61	0.77	0.29	0.36
accuracy	0.77	0.81	0.81	0.65
specificity	0.92	0.83	1	0.92

Figure 3 illustrates the probability (%) for the $\Delta\Delta E_{ele}$+ $\Delta\Delta G_{polar\ solvation}$ value for (a) the nonpolar residues, (b) the polar residues, and (c) the charged residues plus histidine, for the test-set. It shows an increase of the maximum difference between the $|\Delta\Delta E_{ele} + \Delta\Delta G_{polar\ solvation} - \Delta\Delta G_{Experimental}|$ from the polar residues to the nonpolar ones and charged ones. For the nonpolar residues 64% have values between 0 and 2 kcal/mol and 36% higher than 2.0 kcal/mol. For the polar residues we observe that the majority 85% range between 0 and 2 kcal/mol and 15% between 2.0 and 4.0 kcal/mol. Finally, for the charged residues we observe that 72% range between 0 and 2.0 kcal/mol, 24% between 2.0 and 4.0 kcal/mol, and 4% have values higher than 4.0 kcal/mol.

We increased the number of mutations analyzed from 25 to 78, which we believe provide us enough data to validate our method. Table 15 summarizes our results for the data set. As it can be perceived there are very similar to the ones obtained for the test-set, which increases our confidence on the method.

CONCLUSION

For the last years there has been an attempt to unveil and to fully understand the processes and interactions that support the formation of biological complexes, such as protein-based complexes. The formation of these complexes is generally accomplished due to the presence of single residues with high binding affinity, the Hot-spots. Several computational techniques and methods have been developed attempting to accurately reproduce the experimental binding energies of these key residues. The Alanine-Scanning Mutagenesis approach based on the MM-PB/GBSA method was shown to be a reliable method for the study of protein−protein complexes. Its applicability to other types of interfaces is still scarce. Actually, for protein−DNA interfaces the computational mutagenesis information available is insufficient and lacks a comprehensive study on the subject.

With that in mind, we attempt to improve the original ASM methodology and use it to calculate the binding free energy of several residues in seven protein−DNA complexes. The use different internal dielectric constant values for different classes of residues, as well as the application of Generalized Born model with a salt concentration of 0.145 M, allowed the identification of both HS and NS at protein−DNA interfaces with good success rate. This computational approach makes use of the known benefits of performing MD simulations to correctly reproduce the movements occurring in the medium as well as the use of different internal dielectric constants to mimic the behavior of each type of amino acid that is mutated. We studied different parameters that could influence the determination of the binding free energy values of the 25 mutations present in our training set and the 53 mutations in the test-set. The results that we have obtained and the final formulation of the method, summarized in Figure 2, indicate that it is better to run explicit solvent MD simulations and to use the GB model with a 0.145 M salt concentration, which is the typical value inside the cell. We also noted that we could get more accurate values if we only considered the two major energetic terms at play $|\Delta\Delta E_{ele}$ and $\Delta\Delta G_{polar\ solvation}$. Regarding the internal dielectric constants, we found that for the charged amino acids (aspartic acid, glutamic acid, lysine, arginine, and histidine) a constant of 10 should be used, for the polar residues (aspargine, glutamine, cysteine, tyrosine, serine, and threonine) not ionized at physiological pH the internal dielectric constant should be 9, and for the nonpolar amino acids (valine, leucine, isoleucine, phenylalanine, methionine and tryptophan), the internal dielectric constant should be 4. Overall our method yields for the data set an average error $|((\Delta\Delta E_{ele}+\Delta\Delta G_{polar\ solvation}) - \Delta\Delta G_{binding\text{-}experimental})|$ of 1.55 kcal/mol, 0.77 and 0.92 for Accuracy and Specificity, respectively, as well as 0.78 and 0.50 on Precision and Recall. We have to highlight that the charged residues, the ones with higher importance on PDI, are the ones with higher average errors. This type of behavior was already encountered on the HS determination in PPI.[9c,10h]

This methodology was only tested in protein−DNA systems. In the future, a data set of various protein−RNA systems should also be built to check the transferability of the protocol. This improved methodological approach to the traditional ASM method is still simple, has a relatively low computational cost associated, especially when compared to methods such as TI or FEP, and may help improve the current methods for the analysis of protein−DNA interfaces.

ASSOCIATED CONTENT

Supporting Information

Table SI-I: Results obtained with the LPB equation for 25 structures of the last 2 ns of the MD simulation in explicit solvent. Table SI-II: Results obtained with the NLPB equation for 25 structures of the last 2 ns of the MD simulation in explicit solvent, and with a salt concentration of 0.010 M. Table SI-III: Results obtained with the NLPB equation for 25 structures of the last 2 ns of the MD simulation in explicit solvent, and with an ionic concentration of 0.145 M. Table SI-IV: Results obtained with the NLPB equation for 25 structures of the last 2 ns of the MD simulation in explicit solvent, with a salt concentration of 0.145 M, using the $\Delta\Delta E_{ele}$ + $\Delta\Delta G_{polar\ solvation}$ terms. Table SI-V: Results obtained with the NLPB equation for 25 structures of the last 2 ns of the MD simulation in implicit solvent, and with a salt concentration of 0.145 M, using the $\Delta\Delta E_{ele}$ + $\Delta\Delta G_{polar\ solvation}$ terms. Table SI-VI: Results obtained with the NLPB equation for 50 structures of the last 2 ns of the MD simulation in explicit solvent, and with a salt concentration of 0.145 M, using the $\Delta\Delta E_{ele}$ + $\Delta\Delta G_{polar\ solvation}$ terms. Table SI-VII: Results obtained with the NLPB equation for 25 structures taken from the 4 to 6 ns from the MD simulations in explicit solvent, and with a salt concentration of 0.145 M, using the $\Delta\Delta E_{\text{ele}}$ + $\Delta\Delta G_{\text{polar solvation}}$ terms. Table SI-VIII: Results obtained with the NLPB equation for 100 structures taken from the 2 to 8 ns from the MD simulations in explicit solvent, and with a salt concentration of 0.145M, using the $\Delta\Delta E_{\text{ele}}$ + $\Delta\Delta_{\text{Gpolar solvation}}$ terms. Table SI-IX: Results obtained with the NLPB equation for 25 structures taken from the last 2 ns from the MD simulations in explicit solvent, with a salt concentration of 0.145 M, using the *pbsa* module in the AMBER package and using the $\Delta\Delta E_{\text{ele}}$ + $\Delta\Delta G_{\text{polar solvation}}$ terms. Table SI-X: Results obtained with GB for 25 structures taken from the last 2 ns from the MD simulations in explicit solvent and with a salt concentration of 0.145 M, using the $\Delta\Delta E_{\text{ele}}$ + $\Delta\Delta_{\text{Gpolar solvation}}$ terms. Table SI-XI: Results obtained with GB for 25 structures taken from the last 2 ns from the MD simulations in explicit solvent and with a salt concentration of 0.145 M with all energetic terms. Table SI-XII: Results obtained with GB for 25 structures taken from the last 2 ns from the MD simulations in explicit solvent and with a salt concentration of 0 M with all energetic terms. Table SI-XIII: Results obtained with GB for 25 structures taken from the last 2 ns from the MD simulations in explicit solvent and with a salt concentration of 0.010 M with all energetic terms. Table SI-XIV: Results obtained with GB for 25 structures taken from the last 2 ns from the MD simulations in explicit solvent and with a salt concentration of 0 M, using the $\Delta\Delta E_{\text{ele}}$ + $\Delta\Delta G_{\text{polar solvation}}$ terms. Table SI-XV: Results obtained with GB for 25 structures taken from the last 2 ns from the MD simulations in explicit solvent and with a salt concentration of 0.010 M, using the $\Delta\Delta E_{\text{ele}}$ + $\Delta\Delta G_{\text{polar solvation}}$ terms. Table SI-XVI: Results obtained with GB for 25 structures taken from the last 2 ns from the MD simulations in implicit solvent and with a salt concentration of 0.145 M, using the $\Delta\Delta E_{\text{ele}}$ + $\Delta\Delta G_{\text{polar solvation}}$ terms. Table SI-XVII: Results obtained with GB for 100 structures taken from the last 2 ns from the MD simulations in explicit solvent and with a salt concentration of 0.145 M, using the $\Delta\Delta E_{\text{ele}}$ + $\Delta\Delta G_{\text{polar solvation}}$ terms. Table SI-XVIII: Results obtained for the data set using the final methodological formulation. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: irina.moreira@fc.up.pt.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

I.S.M. is supported by FCT Ciência 2008 (Hiring of PhDs for the SCTN, financed by POPH, QREN, Typology 4.2, Promoting Scientific Employment, cofinanced by MES national funding and The European Social Fund).

■ REFERENCES

(1) Tsai, C. J.; Ma, B. Y.; Nussinov, R. Protein–protein interaction networks: How can a hub protein bind so many different partners? *Trends Biochem. Sci.* **2009**, *34*, 594–600.

(2) (a) Chothia, C.; Janin, J. Principles of protein–protein recognition. *Nature* **1975**, *256*, 705–708. (b) Janin, J. Elusive affinities. *Proteins Struct. Funct. Genet.* **1995**, *21*, 30–39. (c) Jones, S.; Thornton, J. M. Principles of protein–protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13–20.

(3) (a) Kortemme, T.; Baker, D. Computational design of protein–protein interactions. *Curr. Opin. Chem. Biol.* **2004**, *8*, 91–97. (b) Russell, R. B.; Alber, F.; Aloy, P.; Davis, F. P.; Korkin, D.; Pichaud, M.; Topf, M.; Sali, A. A structural perspective on protein–protein interactions. *Curr. Opin. Struct. Biol.* **2004**, *14*, 313–324.

(4) (a) Clackson, T.; Ultsch, M. H.; Wells, J. A.; de Vos, A. M. Structural and functional analysis of the 1:1 growth hormone–receptor complex reveals the molecular basis for receptor affinity. *J. Mol. Biol.* **1998**, *277*, 1111–1128. (b) DeLano, W. L.; Ultsch, M. H.; de Vos, A. M.; Wells, J. A. Convergent solutions to binding at a protein–protein interface. *Science* **2000**, *287*, 1279–1283. (c) DeLano, W. L. Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.* **2002**, *12*, 14–20.

(5) (a) Thorn, K. S.; Bogan, A. A. ASEdb: A database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* **2001**, *17*, 284–285. (b) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Hot spots—A review of the protein–protein interface determinant amino acid residues. *Proteins Struct. Funct. Bioinf.* **2007**, *68*, 803–812.

(6) (a) Martins, J. M.; Ramos, R. M.; Moreira, I. S. Structural determinants of a typical leucine-rich repeat protein. *Commun. Comp. Phys.* **2013**, *13*, 238–255. (b) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Hot spot occlusion from bulk water: A comprehensive study of the complex between the lysozyme HEL and the antibody FVD1.3. *J. Phys. Chem. B* **2007**, *111*, 2697–2706.

(7) (a) Grosdidier, S.; Fernandez-Recio, J., Identification of hot-spot residues in protein–protein interactions by computational docking. *BMC Bioinf.* **2008**, *9*; (b) Fernandez-Recio, J. Prediction of protein binding sites and hot spots. *WIREs Comput. Mol. Sci.* **2011**, *1*, 680–698.

(8) (a) Martins, S. A.; Perez, M. A. S.; Moreira, I. S.; Sousa, S. F.; Ramos, M. J.; Fernandes, P. A. Computational alanine scanning mutagenesis: MM-PBSA vs TI. *J. Chem. Theory Comput.* **2013**, *9*, 1311–1319. (b) Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A. Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein–ligand, protein–protein, and protein–nucleic acid noncovalent interactions. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 211–243.

(9) (a) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889–897. (b) Massova, I.; Kollman, P. A. Computational alanine scanning to probe protein–protein interactions: A novel approach to evaluate binding free energies. *J. Am. Chem. Soc.* **1999**, *121*, 8133–8143. (c) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Computational alanine scanning mutagenesis—An improved methodological approach. *J. Comput. Chem.* **2007**, *28*, 644–654.

(10) (a) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Accuracy of the numerical solution of the Poisson-Boltzmann equation. *J. Mol. Struct. THEOCHEM* **2005**, *729*, 11–18. (b) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Detailed microscopic study of the full ZipA: FtsZ interface. *Proteins Struct. Funct. Bioinf.* **2006**, *63*, 811–821. (c) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Unraveling the importance of protein–protein interaction: Application of a computational alanine-scanning mutagenesis to the study of the IgG1 streptococcal protein G (C2 fragment) complex. *J. Phys. Chem. B* **2006**, *110*, 10962–10969. (d) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Hot spot computational identification: Application to the complex formed between the hen egg white lysozyme (HEL) and the antibody HyHEL-10. *Int. J. Quantum Chem.* **2007**, *107*, 299–310. (e) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Backbone importance for protein–protein binding. *J. Chem. Theory Comput.* **2007**, *3*, 885–893. (f) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Protein–protein recognition: A computational mutagenesis study of the MDM2-P53 complex. *Theor. Chem. Acc.* **2008**, *120*, 533–542. (g) Moreira, I. S.; Martins, J. M.; Ramos, M. J.; Fernandes, P. A.; Ramos, M. J. Understanding the importance of the aromatic amino-acid residues as

hot-spots. *Biochem. Biophys. Acta* **2013**, *1834*, 401−414. (h) Huo, S.; Massova, I.; Kollman, P. A. Computational alanine scanning of the 1:1 human growth hormone−receptor complex. *J. Comput. Chem.* **2002**, *23*, 15−27.

(11) (a) Kumar, M. D. S.; Bava, K. A.; Gromiha, M. M.; Prabakaran, P.; Kitajima, K.; Uedaira, H.; Sarai, A. ProTherm and ProNIT: Thermodynamic databases for proteins and protein−nucleic acid interactions. *Nucleic Acids Res.* **2006**, *34*, D204−D206. (b) Prabakaran, P.; An, J.; Gromiha, M. M.; Selvaraj, S.; Uedaira, H.; Kono, H.; Sarai, A. Thermodynamic database for protein−nucleic acid interactions (ProNIT). *Bioinformatics* **2001**, *17*, 1027−1034.

(12) Bochkarev, A.; Bochkareva, E.; Frappier, L.; Edwards, A. M. The 2.2 angstrom structure of a permanganate-sensitive DNA site bound by the Epstein-Barr virus origin binding protein, EBNA1. *J. Mol. Biol.* **1998**, *284*, 1273−1278.

(13) Schildbach, J. F.; Karzai, A. W.; Raumann, B. E.; Sauer, R. T. Origins of DNA-binding specificity: Role of protein contacts with the DNA backbone. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 811−817.

(14) Ogata, K.; Morikawa, S.; Nakamura, H.; Sekikawa, A.; Inoue, T.; Kanai, H.; Sarai, A.; Ishii, S.; Nishimura, Y. Solution structure of a specific DNA complex of the MYB DNA-binding domain with cooperative recognition helices. *Cell* **1994**, *79*, 639−648.

(15) Murphy, F. V.; Sweet, R. M.; Churchill, M. E. A. The structure of a chromosomal high mobility group protein−DNA complex reveals sequence-neutral mechanisms important for non-sequence-specific DNA recognition. *EMBO J.* **1999**, *18*, 6610−6618.

(16) Tan, S.; Richmond, T. J. Crystal structure of the yeast MATα2/MCM1/DNA ternary complex. *Nature* **1998**, *391*, 660−666.

(17) Lei, M.; Podell, E. R.; Baumann, P.; Cech, T. R. DNA self-recognition in the structure of Pot1 bound to telomeric single-stranded DNA. *Nature* **2003**, *426*, 198−203.

(18) Larkin, C.; Datta, S.; Harley, M. J.; Anderson, B. J.; Ebie, A.; Hargreaves, V.; Schildbach, J. F. Inter- and intramolecular determinants, of the specificity of single-stranded DNA binding and cleavage by the F factor relaxase. *Structure* **2005**, *13*, 1533−1544.

(19) (a) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very fast prediction and rationalization of pK_a values for protein−ligand complexes. *Proteins Struct. Funct. Bioinf.* **2008**, *73*, 765−783. (b) Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and rationalization of protein pK_a values. *Proteins Struct. Funct. Bioinf.* **2005**, *61*, 704−721. (c) Olsson, M. H. M.; Sondergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent treatment of internal and surface residues in empirical pK_a predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525−537.

(20) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; R. Luo; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; S. Hayik; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; H. Gohlke; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; C. Schafmeister; Ross, W. S.; Kollman, P. A. *AMBER 9*, University of California: San Francisco, 2006.

(21) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., III; Laughton, C. A.; Orozco, M. Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers. *Biophys. J.* **2007**, *92*, 3817−3829.

(22) *The PyMOL Molecular Graphics System*; Schrödinger: Cambridge, MA. http://www.pymol.org/.

(23) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins Struct. Funct. Bioinf.* **2004**, *55*, 383−394.

(24) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(25) Hancock, S. P.; Ghane, T.; Cascio, D.; Rohs, R.; Di Felice, R.; Johnson, R. C. Control of DNA minor groove width and Fis protein binding by the purine 2-amino group. *Nucleic Acids Res.* **2013**, *41* (3), 6750−6760.

(26) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald—An *N* log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(27) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of cartesian equations of motion of a system with constraints—Molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.

(28) (a) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Langevin dynamics of peptides—The frictional dependence of isomerizaton rates of N-acetyllalanyl-N′-methylamide. *Biopolymers* **1992**, *32*, 523−535. (b) Izaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. Langevin stabilization of molecular dynamics. *J. Chem. Phys.* **2001**, *114*, 2090−2098.

(29) Bradshaw, R. T.; Patel, B. H.; Tate, E. W.; Leatherbarrow, R. J.; Gould, I. R. Comparing experimental and computational alanine scanning techniques for probing a prototypical protein−protein interaction. *Protein Eng., Des. Sel.* **2011**, *24*, 197−207.

(30) Connolly, M. L. Analytical molecular surface calculation. *J. Appl. Crystallogr.* **1983**, *16*, 548−558.

(31) (a) Rocchia, W.; Alexov, E.; Honig, B. Extending the applicability of the nonlinear Poisson−Boltzmann equation: Multiple dielectric constants and multivalent ions. *J. Phys. Chem. B* **2001**, *105*, 6507−6514. (b) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. *J. Comput. Chem.* **2002**, *23*, 128−137.

(32) (a) Bertonati, C.; Honig, B.; Alexov, E. Poisson−Boltzmann calculations of nonspecific salt effects on protein−protein binding free energies. *Biophys. J.* **2007**, *92*, 1891−1899. (b) Talley, K.; Kundrotas, P.; Alexov, E. Modeling salt dependence of protein−protein association: Linear vs non-linear Poisson−Boltzmann equation. *Commun. Comp. Phys.* **2008**, *3*, 1071−1086.

(33) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate calculation of hydration free-energies using macroscopic solvent models. *J. Phys. Chem.* **1994**, *98*, 1978−1988.

(34) Rohs, R.; West, S. M.; Sosinsky, A.; Liu, P.; Mann, R. S.; Honig, B. The role of DNA shape in protein−DNA recognition. *Nature* **2009**, *461*, 1248−U81.

(35) Ribeiro, J. V.; Cerqueira, N. M. F. S. A.; Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. CompASM: An AMBER-VMD alanine scanning mutagenesis plug-in. *Theor. Chem. Acc.* **2012**, *131*, 1271−1278.

(36) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33−&.

(37) (a) Ahmad, S.; Keskin, O.; Sarai, A.; Nussinov, R. Protein−DNA interactions: Structural, thermodynamic, and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Res.* **2008**, *36*, 5922−5932. (b) Ramos, R. M.; Fernandes, L. F.; Moreira, I. S. Extending the applicability of the O-ring theory to protein−DNA complexes. *Comput. Biol. Chem.* **2013**, *44*, 31−9.

(38) Sheinerman, F. B.; Al-Lazikani, B.; Honig, B. Sequence, structure, and energetic determinants of phosphopeptide selectivity of SH2 domains. *J. Mol. Biol.* **2003**, *334*, 823−841.

(39) Kosloff, M.; Travis, A. M.; Bosch, D. E.; Siderovski, D. P.; Arshavsky, V. Y. Integrating energy calculations with functional assays to decipher the specificity of G protein/RGS protein interactions. *Nat. Struct. Mol. Biol.* **2011**, *18*, 846−853.

(40) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* **2011**, *51*, 69−82.

(41) Xu, L.; Sun, H.; Li, Y.; Wang, J.; Hou, T. Assessing the performance of MM/PBSA and MM/GBSA methods. 3. The impact of force fields and ligand charge models. *J. Phys. Chem. B* **2013**, *117*, 8408−8421.

(42) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Unravelling hot spots: A comprehensive computational mutagenesis study. *Theor. Chem. Acc.* **2007**, *117*, 99−113.