Predicting the Effects of Basepair Mutations in DNA-Protein Complexes by Thermodynamic Integration

Frank R. Beierlein,[†‡] G. Geoff Kneale,[§] and Timothy Clark[†‡¶*]

[†]Computer-Chemie-Centrum and [‡]Excellence Cluster "Engineering of Advanced Materials", Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany; and [§]Institute of Biomedical and Biomolecular Sciences, School of Biological Sciences and [¶]Centre for Molecular Design, University of Portsmouth, Portsmouth, United Kingdom

ABSTRACT   Thermodynamically rigorous free energy methods in principle allow the exact computation of binding free energies in biological systems. Here, we use thermodynamic integration together with molecular dynamics simulations of a DNA-protein complex to compute relative binding free energies of a series of mutants of a protein-binding DNA operator sequence. A guanine-cytosine basepair that interacts strongly with the DNA-binding protein is mutated into adenine-thymine, cytosine-guanine, and thymine-adenine. It is shown that basepair mutations can be performed using a conservative protocol that gives error estimates of ~10% of the change in free energy of binding. Despite the high CPU-time requirements, this work opens the exciting opportunity of being able to perform basepair scans to investigate protein-DNA binding specificity in great detail computationally.

INTRODUCTION

Computing the free energies associated with complexation, binding, or solvation processes has been a subject of considerable interest for more than two decades [1,2]. For example, protein-inhibitor binding free energies are of great importance in the drug-design process. Although computationally inexpensive methods like docking and scoring are often so inaccurate that in many cases they are hard put to discriminate a ligand that binds to a target protein from one that does not [3–5], so-called rigorous free energy methods like free energy perturbation (FEP) [6] or thermodynamic integration (TI) [7] in principle allow the exact calculation of binding or solvation free energies [8–10], provided that the Hamiltonians used are of sufficient quality for the given problem and that the potential surfaces of the system are sampled adequately by simulation techniques like Monte Carlo or molecular dynamics (MD). With the increase in computer performance, the availability of high-quality force-field Hamiltonians for many systems of interest, and many algorithmic improvements in free energy algorithms, rigorous free energy methods are currently experiencing a renaissance [11–18]. The availability of dual-topology methods [19] now allows the perturbation of structurally diverse compounds into each other at an acceptable increase of computational cost compared to the earlier, more limited single-topology methods [12,13,15]. Other problems, such as endpoint singularity catastrophes that were frequently observed when atoms or groups were decoupled [20–22], are avoided by using appropriate algorithms like soft-core variants of the Lennard-Jones or Coulomb potentials [13,15,23,24].

In many biochemical/medical/pharmaceutical applications, relative binding free energies of a series of congeneric inhibitors of a given drug target (e.g., an enzyme) are calculated using a thermodynamic-cycle approach. Even complex binding situations, e.g., when water molecules are displaced in the binding process, can now be handled [25], and FEP or TI techniques have been applied to many pharmaceutically important protein-ligand systems. In most cases, classical molecular force fields are used for these studies, but quantum mechanics or quantum mechanics/molecular mechanics approaches that consider polarization effects have also been developed (see e.g. [11,26,27]) and the references cited therein).

Today, most drug targets are proteins, i.e., enzymes or receptors, and less frequently nucleic acids, although RNA as a target for antibiotics is of particular interest. Most promising, however, is the perspective of intervening at a much more fundamental layer of cell life; the control of transcription and translation processes. Many gene switches have been studied extensively [28–35]; in addition to experimental work, computational studies have provided invaluable insight into the molecular mechanisms of transcriptional control [36–39].

Because of the almost identical space requirements of CG and AT basepairs (for an illustration, see e.g. [40]), double-stranded DNA is an ideal object for performing point mutations in silico. Several authors have attempted such perturbations in the past and have proven the feasibility of mutating basepairs using a range of different FEP techniques and Hamiltonians [41–49]. Recently, algorithmic and hardware improvements have allowed the study of far larger systems than was possible earlier by free energy techniques, e.g., DNA-protein complexes, which are relevant in many mechanisms of transcriptional control [41].

The aim of the current work is to establish a robust protocol, using existing algorithms and codes to perform DNA basepair point mutations in large simulation systems. Such a technique has immense potential, for instance for conducting systematic scans of DNA basepair perturbations in DNA-protein complexes with the aim of understanding and even redesigning DNA switching processes. The system we chose for this purpose is the DNA-controller protein (C-protein) complex shown in Fig. 1. The C-protein controls the expression of the restriction enzyme (endonuclease) in the bacterial restriction-modification (R-M) system Esp1396I (28,34). R-M systems (e.g., AhdI (29–33,50,51) and Esp1396I (28,34)) control horizontal gene transfer in bacterial populations by labeling the host DNA at a host-specific sequence by methylation (which is carried out by a specific methyltransferase (M)), thus protecting it from cleavage by a restriction enzyme/endonuclease (R), which destroys foreign DNA when it cleaves the unmethylated sequence. Precise temporal control of endonuclease expression is crucial for the survival of the host cell and is achieved by a simple but elegant genetic switch (28,30). A C-protein dimer binds to an operator sequence (O_L) upstream of the genes that encode the controller protein (C) and the restriction endonuclease enzyme (R). Binding of the C-protein dimer enhances the affinity for binding the σ-RNA polymerase subunit to the adjacent DNA operator sequence O_R. As a consequence, both the C-protein and the endonuclease are transcribed. An increase in the C-protein concentration leads increasingly to displacement of the σ-subunit on the O_R operator sequence by C-protein dimers and thus endonuclease (and C-protein) expression is downregulated. In addition, in the R-M system Esp1396I, the C-protein also regulates the expression of the methyltransferase (M) gene. A detailed description of the system studied here and its exact regulation mechanism can be found in (28). C.Esp1396I has been extensively studied and is currently the only C-protein for which a DNA-complex x-ray structure is available (28).

Protein-DNA recognition is frequently mediated by a combination of direct and indirect mechanisms (dual
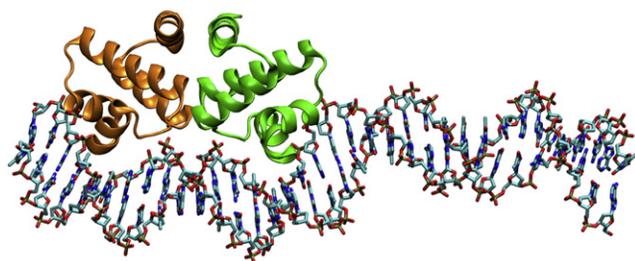
readout): 1), Protein residues can interact directly with DNA atoms (direct readout), e.g., by forming hydrogen bonds between arginine guanidinium groups in the protein and the carbonyl O6 and the ring N7 of guanine bases in DNA, or 2), the protein can also recognize or induce structural features of the DNA backbone, e.g., DNA bending, expansion, or compression of the major and/or minor grooves (indirect readout), which is often achieved through interactions of amino acid side chains with the DNA backbone, notably to the phosphate groups (52–59). In the first case the DNA sequence is recognized directly by hydrogen bonds to specific bases; in the second, the DNA sequence is recognized indirectly by its ability to adopt a specific conformation. Both features are important in C-protein interactions with its various operator sites, and govern the relative affinities for these sites, which is critical for correct regulation of gene expression.

In this work, we investigate the changes in binding free energy of the C-protein-DNA complex C.Esp1396I (28) (we use a complex with only the left DNA operon O_L (the one with a higher affinity to the C-protein dimers) complexed, see Fig. 1) when the evolutionarily conserved DNA basepair GC at position 3 is perturbed to AT, CG, or TA. Guanine at position 3 forms a strong interaction with the guanidinium group of the highly conserved residue Arg-35 of the C-protein, which also shows a π-stacking interaction with the similarly highly conserved thymine T2 of the DNA operator sequence. Our purpose is not only to determine the specificity of the sequence recognition by the C-protein in terms of free energy and to identify its origins, but also to establish a generally applicable computational protocol for performing alchemistic basepair point mutations in double-stranded DNA. Such a protocol represents a powerful tool in the important field of protein-DNA interactions.

Our simulations are designed to 1), test the feasibility of using a modern dual-topology TI implementation to perform arbitrary mutations between all possible basepairs as a tool for scanning basepair selectivity in nucleic acid-protein interactions, without the limitations of former single-topology methods; and 2), to establish a benchmark for the feasibility of this approach using as few CPU-imposed limitations as possible. We have therefore simulated a large, real system of considerable experimental interest using simulation times as long as possible with the most recent force fields for the protein and DNA, using an adequate MD protocol (e.g., with particle mesh Ewald electrostatics and without any restraints in the production phase). Other recent TI or FEP studies of nucleic acid systems were forced to use restraints to obtain qualitative agreement with experiment (49) or used much shorter simulation times and a cutoff for electrostatics (41). We believe that the dual-topology approach chosen in this study, together with a soft-core treatment for nonbonding interactions to avoid endpoint problems, opens exciting opportunities for future perturbation studies on a wide range of systems. Although our simulations still do

FIGURE 1 DNA-protein complex used in this study. Only the higher affinity operator sequence (O_L) is bonded to a C-protein dimer (*orange*, *green*). The perturbation basepair formed by DNA residues 3 and 68 is located on the left hand side of the 35 bp operator sequence. Hydrogen atoms, water molecules, and counterions were omitted for clarity. Molecular graphics were created using VMD 1.8.7 (79,80).

not represent our ideal (simulation and equilibration times are, for instance, still too short), they do provide a realistic idea of the applicability and generality of in silico mutations between Watson-Crick basepairs in signal-transduction systems.

METHOD

To investigate the changes in binding free energy of the C-protein-DNA complex C.Esp1396I (28) upon perturbing the evolutionarily conserved DNA basepair GC at position 3 into AT, CG, or TA (see Fig. 2), we used the thermodynamic cycle shown in Scheme 1. Wild-type (WT) DNA is perturbed into mutant DNA both in the solvated complex formed by DNA and the C-protein, and in a free, solvated DNA fragment. Relative C-protein binding free energies of mutant and WT DNA are calculated using Eq. 1:

$$\Delta G_{\text{bind}}^{\text{mutant}} - \Delta G_{\text{bind}}^{\text{wt}} = \Delta G_{\text{pert}}^{\text{bound}} - \Delta G_{\text{pert}}^{\text{free}}, \qquad (1)$$

where the different free energy changes are defined in Scheme 1.

A TI approach (7) was used to calculate the perturbation free energies from free energy gradients. The transition between WT and mutant DNA was described by a λ-coordinate; MD trajectories were generated at a range of λ-values. Because a dual-topology (19) approach was used, it was possible to perform perturbations between structurally diverse species (purine and pyrimidine nucleobases).

To avoid endpoint singularity problems in cases of atoms that are present in only the WT or mutant DNA, a 3-step perturbation protocol (60) was chosen for these vanishing/appearing atoms. In a first series of simulations, these atoms were made neutral (discharged, $\Delta G_{\text{rem chg}}$), the actual perturbation was then performed using a soft-core (15) potential for the Lennard-Jones interactions of these atoms (ΔG_{pert}), and finally the relevant atoms were recharged (ΔG_{chg}).

COMPUTATIONAL DETAILS

As explained previously, two sets of MD simulations were performed: The free leg, i.e., the 35 basepair (bp) DNA fragment that binds to C.Esp1396I, solvated in water, and the bound leg, i.e., the DNA-C-protein complex, also solvated in water.

For the setup of the free simulations, a 35 bp DNA fragment was created using Amber10 Nucgen (61). Starting coordinate and topology files of the WT and the three mutants were obtained using an iterative procedure using Amber10 Leap (full details are given in the Supporting Material). The starting coordinate and topology files for the bound leg were set up analogously, using the x-ray structure of the C.Esp1396I DNA complex (PDB (62–64) code 3CLC (28)). A 35 bp complex with only the left DNA operon O_L
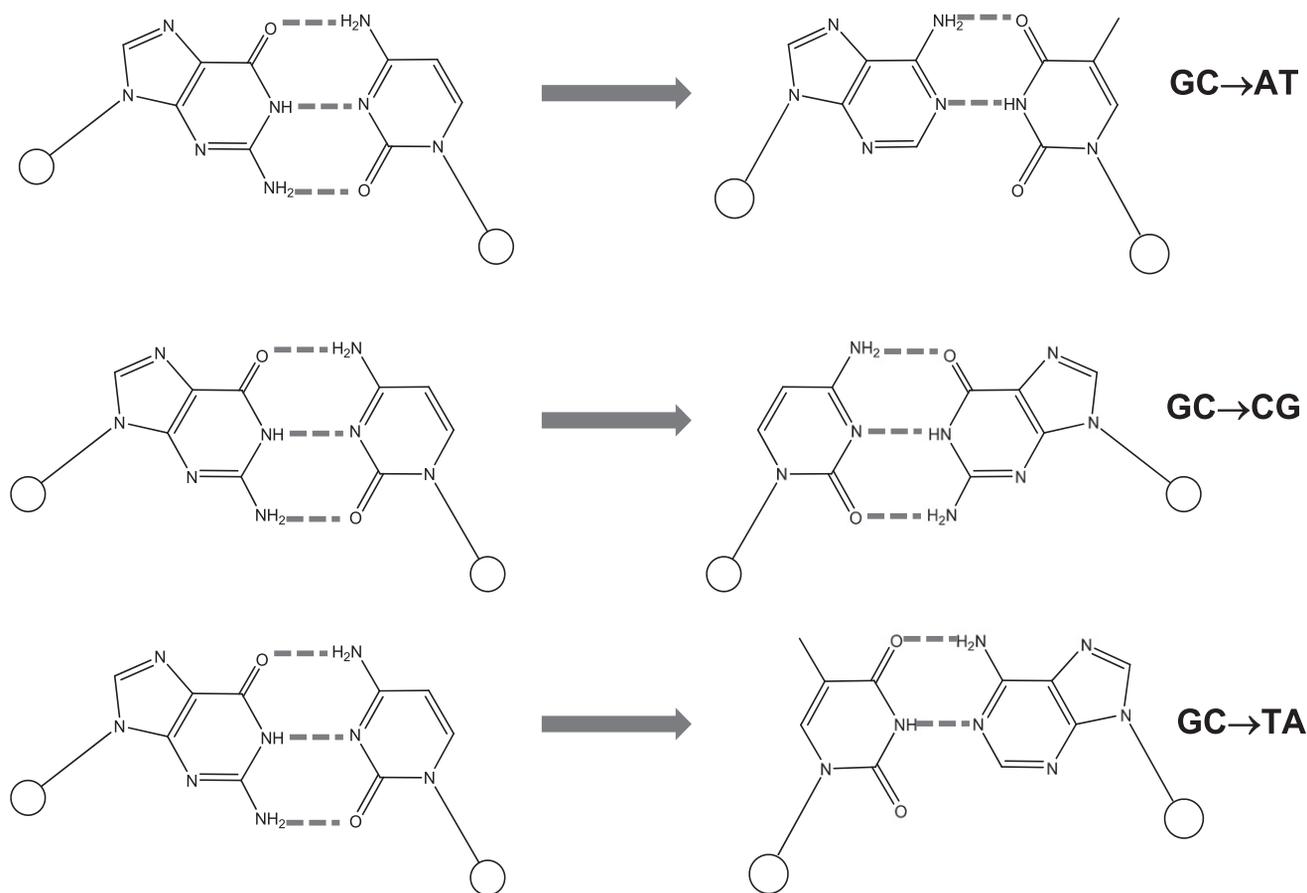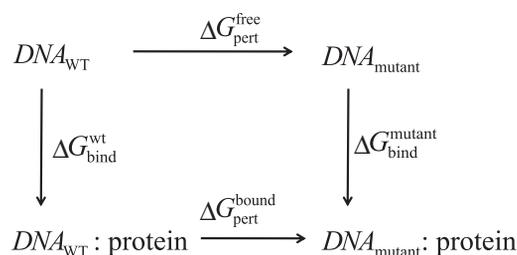

FIGURE 2   Perturbations applied to the basepair formed by DNA residues 3 and 68 in the 35 bp operator sequence: guanine-cytosine to adenine-thymine (GC → AT), guanine-cytosine to cytosine-guanine (GC → CG), and guanine-cytosine to thymine-adenine (GC → TA).

$$DNA_{WT} \xrightarrow{\Delta G_{pert}^{free}} DNA_{mutant}$$

$$\downarrow \Delta G_{bind}^{wt} \qquad\qquad \downarrow \Delta G_{bind}^{mutant}$$

$$DNA_{WT} : protein \xrightarrow{\Delta G_{pert}^{bound}} DNA_{mutant} : protein$$

SCHEME 1   Thermodynamic cycle used to calculate relative DNA-protein binding free energies.

(the one with a higher affinity to the C-protein dimers) complexed by protein chains A and B was obtained by manually deleting protein chains C and D. Solvated coordinate and topology files were obtained using Amber10 Leap, details are given in the Supporting Material. The simulations of the solvated systems were split into 6 lines of simulations for each of the 3 perturbations investigated (GC → AT, GC → CG, GC → TA): Free/discharge, free/perturbation, free/charge, and bound/discharge, bound/perturbation, bound/charge, these simulations will be explained later. All these lines were subjected to energy minimization and dual-topology MD simulations using Amber10 Sander (61), 9 λ-windows (0.1, …, 0.9) were chosen for this initial study. Geometry optimizations and MD simulations were performed using the Stony Brook (ff99SB) (65) and Barcelona (ff99bsc0) (66) modifications of the Amber99 (67) force field, water molecules were described by the TIP3P (68) potential. Although more modern (and probably more accurate) water potentials are available, the ff99SB and ff99bsc0 force fields were designed to be used with TIP3P, therefore we have used it for this work. The interaction between the water model and the protein/DNA force field is critical in studies such as these (65). All λ-windows of all lines of simulations of a given perturbation were treated independently and run in parallel. After initial minimization (see Supporting Material), Langevin dynamics simulations at 298 K using a 2 fs time step were performed (see Supporting Material for full details). System heat up was performed during a 200 ps constant volume (NVT) simulation with weak restraints. The system was then subjected to 200 ps constant pressure (NPT) equilibration at 1 atm without any restraints; this trajectory was not used for data analysis. Results were gathered from a 1200 ps (6 × 200 ps) production phase, using the same simulation parameters as in the NPT equilibration phase. At the end of each (consecutive) MD simulation of 200 ps, the variables required for free energy analysis (ensemble average of dV/dλ, and the corresponding root mean-square (RMS) fluctuation/standard deviation) were obtained from the output files.

To avoid endpoint singularity problems, which are frequently observed in cases of vanishing or appearing atoms, the following approach was chosen, as suggested on the Amber web site (60): In a set of 3 independent simulation lines, atoms vanishing or appearing in the perturbation under study were i), discharged from their initial charge values, ii), annihilated or grown using a van der Waals soft-core for these atoms, and iii), charged to obtain the charges of the target topology. All other atoms were perturbed from their charges/atom types given in the initial topology to the values in the target topology in step (ii). The simplest perturbation under study here is GC → AT. In this case, only substituents of the purine/pyrimidine ring systems are affected. Therefore, only the affected atoms were discharged and charged. These atoms were also the ones treated with the van der Waals soft-core in the actual perturbation step (ii). In the GC → CG and GC → TA perturbations, much larger perturbations are applied: Purines are turned into pyrimidines and vice versa. Here, the whole purine and pyrimidine ring systems were discharged/charged and treated using the soft-core potential in the actual perturbation step. Numerical integration of the dV/dλ gradients was performed according to the midpoint rule, gradients for λ = 0 and λ = 1 were obtained by extrapolation from the two adjacent data points (60).

In principle, precision estimates for the calculated relative binding free energies can be obtained in two ways: A very simple approach is to calculate a standard error from the results obtained from a series of MD runs (batch averaging). Alternatively, it is possible to calculate a standard error from the RMS fluctuation (the standard deviation) of the gradient dV/dλ divided by the square root of the number of independent experiments in the simulation, as discussed in (15). Both approaches were pursued here (see Supporting Material for full details).

RESULTS AND DISCUSSION

Plots of dV/dλ (see Supporting Material, Figs. S1–S18), kinetic/potential/total energy, T, P, V, and system density (not shown) versus time were inspected for the λ = 0.1 and λ = 0.9 windows of each simulation. It was found that in terms of these variables, the simulation system was equilibrated very well after 200 ps NVT and 200 ps NPT equilibration. The root mean-square deviation (RMSD) plots for the DNA backbone and protein C_α atoms (for examples, see Supporting Material, Figs. S19–S21) revealed that the RMSD values stay relatively constant and do not increase much further after ~1000 ps simulation time, which means that the trajectories have sufficiently adapted to the actual topology and force field parameters (we are aware, though, that much longer equilibration times would be desirable, see later). To reveal any drift in values, in the results tables (Tables S1–S6 in the Supporting Material) the integrated free energies are listed for all partial trajectories (200 ps each) between 400 and 1600 ps. Most interestingly, the energy gradients dV/dλ, ensemble averages of which are used to obtain the free energies of interest, do not show any drift in values after 400 ps equilibration (plots are given in the Supporting Material, Figs. S1–S18), which

indicates that the trajectories used for integration (400–1600 ps) can be relied on.

Monitoring the Watson-Crick hydrogen bonds in the terminal basepairs 1–70 and 35–36 revealed that fraying of the ends of the DNA fragment did not play a significant role, which is remarkable as no distance restraints on the terminal basepairs were used in this study. In the perturbation basepair 3–68, monitoring the hydrogen bonds showed the expected result: opening the Watson-Crick hydrogen bonds in the discharge step, no significant hydrogen bonds during the actual perturbation, and reformation of the Watson-Crick hydrogen bonds in the charging step.

Plotting $< dV/d\lambda >$ vs. λ yields smooth curves (Supporting Material Figs. S22–S30) with RMS fluctuations (standard deviations) of ~5 kcal mol^{-1} for the discharge/charge steps of all perturbations investigated here and the perturbation step of the simplest perturbation, GC $\rightarrow$ AT, and ~10 kcal mol^{-1} for the perturbation step of the more challenging perturbations GC $\rightarrow$ CG and GC $\rightarrow$ TA, where purines are perturbed into pyrimidines and vice versa. These curves were used for numerical integration. The perturbation free energies are listed in Table 1; detailed results are shown in the Supporting Material (Tables S1–S6).

The relative binding free energies for the complexes of DNA mutants GC $\rightarrow$ AT, GC $\rightarrow$ CG, and GC $\rightarrow$ TA of C.Esp1396I are all positive compared to the WT (3.95 $\pm$ 0.35, 12.32 $\pm$ 0.57, and 9.30 $\pm$ 0.31 kcal mol^{-1}, respectively). This reflects the fact that only the WT is found in nature (having evolved together with the DNA-binding control protein), and that the three mutations investigated here were not chosen because of their biological relevance, but as a test to check whether the TI protocol described here can be used to investigate DNA point mutations in DNA-protein complexes in general. A referee has pointed out that the choice of the starting geometry (the WT x-ray structure) biases our conformational space toward the WT conformation, which is certainly true. However, our experience with simulations on systems of this type suggests that an equilibration simulation of ~100 ns would be enough to remove this bias in most cases. Thus, although

we have used shorter equilibration periods here, we are confident that the technique described can be made generally applicable by long equilibration simulations. We also note that the starting geometry is not strictly that of the WT complex because we have deleted one C-protein dimer from the experimental structure. Overall, the perturbation free energies for all mutations investigated here and all steps (discharge, perturb, charge) do not change significantly in the course of the six consecutive MD simulations (400–1600 ps, 200 ps each), as indicated by the small standard errors of ~0.1–0.4 kcal mol^{-1} (see Supporting Material Tables S1–S6). This suggests that the MD simulations used are sufficiently well equilibrated for the purpose of calculation of perturbation free energies. The standard deviations for the perturbation free energies obtained from the RMS fluctuations/standard deviations of the gradients of $dV/d\lambda$ for individual λ-windows by weighted summation range from ~4.4 to 6.6 kcal mol^{-1} for the discharge/charge steps of all perturbations investigated here and the perturbation step of the least challenging perturbation, GC $\rightarrow$ AT. It is only in the case of the challenging van der Waals soft-core perturbations of GC $\rightarrow$ CG and GC $\rightarrow$ TA, where purines are perturbed into pyrimidines and vice versa (and hence a larger number of atoms is affected), that the standard deviations are in the range of 8.9–10.7 kcal mol^{-1}. Precision estimates (standard errors) can be obtained by dividing the $dV/d\lambda$ RMS fluctuations obtained for the individual λ-windows by the square root of the actual number of independent experiments in a given simulation run and summing up the weighted errors of the individual λ-windows. This was done here for selected partial trajectories (see Computational Details). Most interestingly, the precision estimates thus obtained for the partial trajectory 400–600 ps are in close agreement with the standard errors obtained from batch averaging (see Supporting Material Tables S1–S6). These standard errors/precision estimates are in the range of ~0.1–0.2 kcal mol^{-1} for the individual perturbation steps ($\Delta G_{rem\ chg}$, ΔG_{pert}, ΔG_{chg}) of the least challenging perturbation GC $\rightarrow$ AT. Slightly larger standard errors/precision values are found for the charge removal/recharging steps ($\Delta G_{rem\ chg}$, ΔG_{chg}) of the purine-to-pyrimidine (and vice versa) perturbations GC $\rightarrow$ CG and GC $\rightarrow$ TA (~0.1–0.3 kcal mol^{-1}). Higher errors (due to more noise in the more challenging perturbations, where whole ring systems are perturbed into each other) are found for the van der Waals soft-core perturbation step (ΔG_{pert}) of GC $\rightarrow$ CG and GC $\rightarrow$ TA (0.2–0.6 kcal mol^{-1}), but these values are still moderate and show that our calculations are sufficiently precise.

As a decomposition of perturbation free energies is not possible in a strict sense and one has to be very careful when interpreting these numbers (an issue that has been the subject of intense discussion) (69–73), it is difficult to correlate the calculated perturbation free energies directly to certain molecular interactions in the four complexes (WT

TABLE 1 Relative binding free energies $\Delta\Delta G = \Delta G_{bind,mut} - \Delta G_{bind,wt}$ [kcal mol^{-1}] for complexes of DNA mutants GC $\rightarrow$ AT, GC $\rightarrow$ CG, and GC $\rightarrow$ TA of the C.Esp1396I protein-DNA complex

$\Delta\Delta G_{bind}$ [kcal mol^{-1}]	400–1600 ps	$\pm$ (Standard error)	$\pm$ (Precision estimate)	1400–1600 ps
GC $\rightarrow$ AT	3.95	0.35	0.36	4.47
GC $\rightarrow$ CG	12.32	0.57	0.79	10.94
GC $\rightarrow$ TA	9.30	0.31	0.86	10.27

$\Delta\Delta G$ values are reported as mean values over six consecutive MD simulations (400–1600 ps); the corresponding standard errors are given. Precision estimates calculated from partial trajectories are given for comparison (see Computational Details). To detect a possible drift in the values, relative binding free energies calculated from the last partial trajectory (1400–1600 ps) are given for reference.

plus three mutants). For the solvated DNA fragment in water (free), the simplest perturbation GC $\rightarrow$ AT, where only substituents of the ring systems are affected, shows a free energy cost for removing the charges ($\Delta G_{\text{rem chg}}$) on the affected atoms (see Computational Details) of 266.93 $\pm$ 0.12 kcal mol^{-1}, whereas the recharging step (ΔG_{chg}) only gives a gain of -101.94 ± 0.08 kcal mol^{-1}. As one referee correctly pointed out, a direct interpretation of these numbers is difficult (and there is no need to do so). The actual perturbation step (ΔG_{pert}), where the topology is changed and van der Waals interactions with the environment are affected, contributes with 8.66 $\pm$ 0.07 kcal mol^{-1}; probably mainly due to the cost of the growth of a methyl group in thymine. In the bound perturbation of GC $\rightarrow$ AT (DNA-protein complex), ΔG_{pert}, ΔG_{chg} remain mainly unaffected (8.33 $\pm$ 0.07 and -103.01 ± 0.15 kcal mol^{-1}). $\Delta G_{\text{rem chg}}$, however, increases to 272.28 $\pm$ 0.19 kcal mol^{-1}, yielding a difference of the charge removal free energies in the bound and free case of 5.35 kcal mol^{-1}. This number certainly reflects the energetic cost of breaking up one hydrogen bond with the guanidinium group of protein residue Arg-35 when the guanine carbonyl ring substituent is discharged. In the mutant, this hydrogen bond is not restored in the charging step and the adenine amino group is responsible for a partial displacement of the Arg-35 guanidinium group because of steric hindrance. The difference of the charge removal free energies in the bound and free case (5.35 kcal mol^{-1}) is the main contribution to the overall relative binding free energy ($\Delta\Delta G$) for GC $\rightarrow$ AT of 3.95 $\pm$ 0.35 kcal mol^{-1}. An illustration of the binding situation in the WT and the mutant is shown in Figs. S31 and S32 in the Supporting Material (WT and mutant after 1600 ps MD) and S35 (superimposition before minimization/MD) in the Supporting Material.

Perturbation GC $\rightarrow$ CG is an ideal test for our method, as in the free case, "nothing" must happen (in the limit of neglecting intra-DNA interactions like π-stacking between basepairs): $\Delta G_{\text{rem chg}}$ and ΔG_{chg} are equal and opposite in value (335.43 $\pm$ 0.27 and -335.56 ± 0.15 kcal mol^{-1}), and ΔG_{pert} is zero (-0.07 ± 0.40 kcal mol^{-1}). In the bound case, ΔG_{pert} and ΔG_{chg} remain mainly unaffected (-0.17 ± 0.19 and -333.43 ± 0.29 kcal mol^{-1}). $\Delta G_{\text{rem chg}}$ increases by 10.3 kcal mol^{-1} to 345.73 $\pm$ 0.10 kcal mol^{-1}, the difference of the charge removal free energies in the bound and free case can be explained by the loss of the two specific hydrogen bonds between guanine and Arg-35 when the guanine ring system is discharged. In the mutant, steric hindrance between the cytosine amino hydrogens and the arginine guanidinium group leads to a displacement of the arginine guanidinium moiety, which is no longer in-plane with cytosine C3. Therefore, the π-stacking interaction between DNA base thymine T2 and Arg-35/cytosine C3 suffers. Again, the difference of the charge removal free energies in the bound and free case (10.30 kcal mol^{-1}) is the main contribution to the overall relative binding free energy of mutant GC $\rightarrow$ CG, which amounts to 12.32 $\pm$

0.57 kcal mol^{-1}. Interestingly, an FEP study of the same mutation CG to GC in a different protein-DNA system, using a different Hamiltonian and a different protocol, also with an interaction between guanine and a protein arginine side chain, gave a similarly positive $\Delta\Delta G$ value of $+7.7 \pm 0.6$ kcal mol^{-1} (41). An illustration of the binding situation in the mutant is shown in Fig. S33 (mutant after 1600 ps MD) and S36 (superimposition before minimization/MD) in the Supporting Material.

The last perturbation investigated here, GC $\rightarrow$ TA, shows, of course, identical values as shown above for the discharge step ($\Delta G_{\text{rem chg}}$, free: 335.78 $\pm$ 0.09, bound: 345.30 $\pm$ 0.10 kcal mol^{-1}). In the free simulation, ΔG_{pert} and ΔG_{chg} are -0.51 ± 0.16 and -163.76 ± 0.26 kcal mol^{-1}; again, a direct interpretation of the difference between the discharge and charge step free energies is difficult. In the bound case, ΔG_{pert} and ΔG_{chg} are not significantly changed (0.34 $\pm$ 0.21 and -164.83 ± 0.22 kcal mol^{-1}), whereas $\Delta G_{\text{rem chg}}$ is increased by 9.52 kcal mol^{-1} compared to the free simulation (345.30 $\pm$ 0.10 kcal mol^{-1}) as a result of the loss of the two hydrogen bonds between guanine and the guanidinium group of Arg-35 when the guanine ring system is discharged. In the mutant, only one of these hydrogen bonds is restored and the thymine methyl group clashes with the Arg-35 side chain, which leads to a partial displacement of the Arg-35 guanidinium group (which is still in-plane with the thymine ring system in this case). This displacement leads to a weakening of the π-stacking interaction between DNA base thymine 2 and Arg-35/thymine 3 in the mutant. Overall, a relative binding free energy $\Delta\Delta G$ of 9.30 $\pm$ 0.31 kcal mol^{-1} is obtained for mutant GC $\rightarrow$ TA compared to the WT. An illustration of the binding situation in the mutant is shown in Fig. S34 (mutant after 1600 ps MD) and S37 (superimposition before minimization/MD) in the Supporting Material.

CONCLUSIONS AND OUTLOOK

The positive relative binding free energies given above for three DNA point mutations in the complex of the C-protein-DNA complex C.Esp1396I (GC $\rightarrow$ AT: 3.95 $\pm$ 0.35 kcal mol^{-1}, GC$\rightarrow$CG: 12.32 $\pm$ 0.57 kcal mol^{-1}, and GC $\rightarrow$ TA: 9.30 $\pm$ 0.31 kcal mol^{-1}) are consistent with the highly conserved nature of the G3-C68 basepair investigated here, although, as noted above, there may be some bias toward the WT introduced by the fact that we used a WT starting structure and relatively short equilibration times. Indeed, while G at this position is highly conserved (frequency ca. 90%), the only other base found at this site is an A (74), which is consistent with the GC $\rightarrow$ AT mutation being the least disruptive. The most likely direction for the errors in our calculated energies is for them to be too large because of the bias introduced by using the WT geometry. At the very least however, the rank order of the calculated basepair mutation energies

seems reasonable on the basis of the sequence information. There will also clearly be situations in which highly specific features of the protein-DNA binding environment will render calculations, such as those reported here, unreliable because of insufficient sampling. Generally, however, these interactions are evident on visual inspection, so that they can hopefully be detected in advance.

Experimentally, it has been shown that mutating the protein to remove the Arg-35 side chain completely abolished DNA binding to the operator (28). However, the energetics of this interaction will also include the stacking interactions of Arg-35 with the adjacent TA basepair in the TG motif. Future applications of the theoretical techniques we have developed will include mutating the adjacent TA basepair, to deconvolute the separate energetic contributions of these two interactions.

All possible mutations bind significantly worse to the control protein than the WT. This is, of course, not surprising, although it need not necessarily be the case, because the precise binding constants for each of the three operator sites have been carefully tuned by evolution—too strong an affinity for a given operator can also be deleterious (34). In fact the binding affinity of the C-protein to the operator under study here, O_L, is an order of magnitude lower than that for the highest affinity site, O_M, which regulates expression of the methyltransferase. At least in part, this is due to the presence of a symmetrically related guanine in the DNA sequence of O_M that can make contacts with the Arg-35 side chain of the second subunit of the dimer.

The aim of the current study was not to engineer new mutations, but rather to establish a general protocol for systematic scans in the search for biologically relevant DNA point mutations in protein-binding DNA sequences. Just exactly how general our protocol is remains to be determined. From other work, we estimate simulation times for complete relaxation of systems such as the one studied here to be of the order of 100 ns, which would remove the bias toward the WT conformation that possibly exists in the calculations reported here. Combining our approach with enhanced conformational sampling of important protein side chains before starting the TI simulations would also help remove bias and make the technique more general.

With ~100,000 CPU hours required for each mutation studied here, this task is an ideal challenge for current supercomputers. We estimate that for a scan of all 35 basepairs in the DNA fragment simulated here, 10.5 million CPU hours on a ca. 3 GHz Intel Xeon machine would be required; examining the effects for only one dimer binding, would only require mutations for the 19 basepairs of O_L (5.7 million CPU hours). Binding to the right operator (O_R) would also be of interest when the second dimer binds. A more realistic goal in the short term is simply to compare the binding energies of the three natural binding sites, because the K_d values for these three sites vary over four orders of magnitude (34).

Algorithmic improvements will help to reduce the computer requirements, e.g., the possibility also to perform perturbations in one step using a soft-core potential for electrostatic interactions will reduce CPU costs to one-third. This option has recently been added to the simulation program (75), and we intend to compare our current three-step protocol with a one-step perturbation scheme using van der Waals and electrostatics soft-core potentials. Future simulations will also include a comparison of TI results with those obtained using the Bennett acceptance ratio technique (76–78).

In the current study, we have focused on direct readout, i.e., the direct, specific interaction between DNA and protein atoms. Additional to this direct mechanism, indirect readout, i.e., modulation of the precise DNA shape and dynamics also plays a major role in protein-DNA recognition. This effect and a detailed analysis of interactions in the endpoints of the perturbations investigated here ($\lambda = 0.0$ and $\lambda = 1.0$, WT and mutant) are currently being analyzed in "normal" (i.e., not TI or FEP), long-time (up to 200 ns) MD simulations.

We note also that we have performed simulations of several hundred nanoseconds in a separate study of mechanistic aspects of regulation by the C-protein in this system. The x-ray structure remains stable with a low RMSD for the entire simulation period in such simulations on the DNA sequence complexed to two C-protein dimers.

The major conclusion of this work is, however, that basepair mutations can be performed using a conservative protocol that gives error estimates of ~10% of the change in free energy of binding. Despite the high CPU time requirements, this work opens the exciting opportunity of being able to perform basepair scans to investigate protein-DNA binding specificity in great detail computationally.

SUPPORTING MATERIAL

REFERENCES

1. Rao, S. N., U. C. Singh, …, P. A. Kollman. 1987. Free energy perturbation calculations on binding and catalysis after mutating Asn 155 in subtilisin. *Nature.* 328:551–554.

2. Jorgensen, W. L. 1989. Free energy calculations: a breakthrough for modeling organic chemistry in solution. *Acc. Chem. Res.* 22:184–189.

3. Leach, A. R., B. K. Shoichet, and C. E. Peishoff. 2006. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* 49:5851–5855.

4. Warren, G. L., C. W. Andrews, …, M. S. Head. 2006. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* 49:5912–5931.

5. Schneider, G. 2010. Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.* 9:273–276.

6. Zwanzig, R. W. 1954. High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* 22:1420–1426.

7. Kirkwood, J. G. 1935. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* 3:300–313.

8. Michel, J., and J. W. Essex. 2010. Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput. Aided Mol. Des.* 24:639–658.

9. Michel, J., N. Foloppe, and J. W. Essex. 2010. Rigorous free energy calculations in structure-based drug design. *Mol. Inf.* 29:570–578.

10. Gilson, M. K., and H.-X. Zhou. 2007. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* 36:21–42.

11. Beierlein, F. R., J. Michel, and J. W. Essex. 2011. A simple QM/MM approach for capturing polarization effects in protein-ligand binding free energy calculations. *J. Phys. Chem. B.* 115:4911–4926.

12. Michel, J., and J. W. Essex. 2008. Hit identification and binding mode predictions by rigorous free energy simulations. *J. Med. Chem.* 51:6654–6664.

13. Michel, J., M. L. Verdonk, and J. W. Essex. 2007. Protein-ligand complexes: computation of the relative free energy of different scaffolds and binding modes. *J. Chem. Theory Comput.* 3:1645–1655.

14. Michel, J., M. L. Verdonk, and J. W. Essex. 2006. Protein-ligand binding affinity predictions by implicit solvent simulations: a tool for lead optimization? *J. Med. Chem.* 49:7427–7439.

15. Steinbrecher, T., D. L. Mobley, and D. A. Case. 2007. Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *J. Chem. Phys.* 127:214108.

16. Krapf, S., T. Koslowski, and T. Steinbrecher. 2010. The thermodynamics of charge transfer in DNA photolyase: using thermodynamic integration calculations to analyse the kinetics of electron transfer reactions. *Phys. Chem. Chem. Phys.* 12:9516–9525.

17. Steinbrecher, T., A. Hrenn, …, A. Labahn. 2008. Bornyl (3,4,5-trihydroxy)-cinnamate—an optimized human neutrophil elastase inhibitor designed by free energy calculations. *Bioorg. Med. Chem.* 16:2385–2390.

18. Steinbrecher, T., D. A. Case, and A. Labahn. 2006. A multistep approach to structure-based drug design: studying ligand binding at the human neutrophil elastase. *J. Med. Chem.* 49:1837–1844.

19. Gao, J., K. Kuczera, …, M. Karplus. 1989. Hidden thermodynamics of mutant proteins: a molecular dynamics analysis. *Science.* 244:1069–1072.

20. Beveridge, D. L., and F. M. DiCapua. 1989. Free energy via molecular simulation: applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Biophys. Chem.* 18:431–492.

21. Simonson, T. 1993. Free energy of particle insertion. An exact analysis of the origin singularity for simple liquids. *Mol. Phys.* 80:441–447.

22. Valleau, J. P., and G. Torrie. 1977. Modern Theoretical Chemistry. Plenum Press, New York.

23. Beutler, T. C., A. E. Mark, …, W. F. van Gunsteren. 1994. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem. Phys. Lett.* 222:529–539.

24. Zacharias, M., T. P. Straatsma, and J. A. McCammon. 1994. Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. *J. Chem. Phys.* 100:9025–9031.

25. Barillari, C., J. Taylor, …, J. W. Essex. 2007. Classification of water molecules in protein binding sites. *J. Am. Chem. Soc.* 129:2577–2587.

26. Shaw, K. E., C. J. Woods, and A. J. Mulholland. 2009. Compatibility of quantum chemical methods and empirical (MM) water models in quantum mechanics/molecular mechanics liquid water simulations. *J. Phys. Chem. Lett.* 1:219–223.

27. Woods, C. J., F. R. Manby, and A. J. Mulholland. 2008. An efficient method for the calculation of quantum mechanics/molecular mechanics free energies. *J. Chem. Phys.* 128:014109.

28. McGeehan, J. E., S. D. Streeter, …, G. G. Kneale. 2008. Structural analysis of the genetic switch that regulates the expression of restriction-modification genes. *Nucleic Acids Res.* 36:4778–4787.

29. McGeehan, J. E., S. D. Streeter, …, G. G. Kneale. 2005. High-resolution crystal structure of the restriction-modification controller protein C.AhdI from *Aeromonas hydrophila. J. Mol. Biol.* 346:689–701.

30. Streeter, S. D., I. Papapanagiotou, …, G. G. Kneale. 2004. DNA footprinting and biophysical characterization of the controller protein C.AhdI suggests the basis of a genetic switch. *Nucleic Acids Res.* 32:6445–6453.

31. McGeehan, J. E., I. Papapanagiotou, …, G. G. Kneale. 2006. Cooperative binding of the C.AhdI controller protein to the C/R promoter and its role in endonuclease gene expression. *J. Mol. Biol.* 358:523–531.

32. Papapanagiotou, I., S. D. Streeter, …, G. G. Kneale. 2007. DNA structural deformations in the interaction of the controller protein C.AhdI with its operator sequence. *Nucleic Acids Res.* 35:2643–2650.

33. Bogdanova, E., M. Djordjevic, …, K. Severinov. 2008. Transcription regulation of the type II restriction-modification system AhdI. *Nucleic Acids Res.* 36:1429–1442.

34. Bogdanova, E., M. Zakharova, …, K. Severinov. 2009. Transcription regulation of restriction-modification system Esp1396I. *Nucleic Acids Res.* 37:3354–3366.

35. Saenger, W., P. Orth, …, W. Hinrichs. 2000. The tetracycline repressor-A paradigm for a biological switch. *Angew. Chem. Int. Ed. Engl.* 39:2042–2052.

36. Beierlein, F. R., O. G. Othersen, …, T. Clark. 2006. Simulating FRET from tryptophan: is the rotamer model correct? *J. Am. Chem. Soc.* 128:5142–5152.

37. Lanig, H., O. G. Othersen, …, T. Clark. 2006. Molecular dynamics simulations of the tetracycline-repressor protein: the mechanism of induction. *J. Mol. Biol.* 359:1125–1136.

38. Lanig, H., O. G. Othersen, …, T. Clark. 2006. Structural changes and binding characteristics of the tetracycline-repressor binding site on induction. *J. Med. Chem.* 49:3444–3447.

39. Seidel, U., O. G. Othersen, …, T. Clark. 2007. Molecular dynamics characterization of the structures and induction mechanisms of a reverse phenotype of the tetracycline receptor. *J. Phys. Chem. B.* 111:6006–6014.

40. Stryer, L. 1995. Biochemistry. W. H. Freeman and Company, New York.

41. Hart, K., and L. Nilsson. 2008. Investigation of transcription factor Ndt80 affinity differences for wild type and mutant DNA: A molecular dynamics study. *Proteins.* 73:325–337.

42. Nam, K., G. L. Verdine, and M. Karplus. 2009. Analysis of an anomalous mutant of MutM DNA glycosylase leads to new insights into the catalytic mechanism. *J. Am. Chem. Soc.* 131:18208–18209.

43. Chandani, S., C. H. Lee, and E. L. Loechler. 2005. Free-energy perturbation methods to study structure and energetics of DNA adducts: results for the major N2-dG adduct of benzo[a]pyrene in two conformations and different sequence contexts. *Chem. Res. Toxicol.* 18:1108–1123.

44. Cubero, E., C. A. Laughton, …, M. Orozco. 2000. Molecular dynamics study of oligonucleotides containing difluorotoluene. *J. Am. Chem. Soc.* 122:6891–6899.

45. Florián, J., M. F. Goodman, and A. Warshel. 2000. Free-energy perturbation calculations of DNA destabilization by base substitutions: the effect of neutral guanine·thymine, adenine·cytosine and adenine·difluorotoluene mismatches. *J. Phys. Chem. B.* 104:10092–10099.

46. Hernández, B., R. Soliva, …, M. Orozco. 2000. Misincorporation of 2′-deoxyoxanosine into DNA: a molecular basis for NO-induced mutagenesis derived from theoretical calculations. *Nucleic Acids Res.* 28:4873–4883.

47. Wunz, T. P. 1992. Nucleoside free energy perturbation calculations: mutation of purine-to-pyrimidine and pyrimidine-to-purine nucleosides. *J. Comput. Chem.* 13:667–673.

48. Gago, F., and W. G. Richards. 1990. Netropsin binding to poly [d(IC)].poly[IC)] and poly[d(GC).poly[d(GC)]: a computer simulation. *Mol. Pharmacol.* 37:341–346.

49. Yildirim, I., H. A. Stern, …, D. H. Turner. 2009. Effects of restrained sampling space and nonplanar amino groups on free-energy predictions for RNA with imino and sheared tandem GA base pairs flanked by GC, CG, iGiC or iCiG base pairs. *J. Chem. Theory Comput.* 5:2088–2100.

50. Callow, P., A. Sukhodub, …, G. G. Kneale. 2007. Shape and subunit organization of the DNA methyltransferase M.AhdI by small-angle neutron scattering. *J. Mol. Biol.* 369:177–185.

51. Marks, P., J. McGeehan, …, G. Kneale. 2003. Purification and characterization of a novel DNA methyltransferase, M.AhdI. *Nucleic Acids Res.* 31:2803–2810.

52. Bouvier, B., and R. Lavery. 2009. A free energy pathway for the interaction of the SRY protein with its binding site on DNA from atomistic simulations. *J. Am. Chem. Soc.* 131:9864–9865.

53. Prévost, C., M. Takahashi, and R. Lavery. 2009. Deforming DNA: from physics to biology. *ChemPhysChem.* 10:1399–1404.

54. Curuksu, J., M. Zacharias, …, K. Zakrzewska. 2009. Local and global effects of strong DNA bending induced during molecular dynamics simulations. *Nucleic Acids Res.* 37:3766–3773.

55. O'Flanagan, R. A., G. Paillard, …, A. M. Sengupta. 2005. Non-additivity in protein-DNA binding. *Bioinformatics.* 21:2254–2263.

56. Deremble, C., and R. Lavery. 2005. Macromolecular recognition. *Curr. Opin. Struct. Biol.* 15:171–175.

57. Paillard, G., C. Deremble, and R. Lavery. 2004. Looking into DNA recognition: zinc finger binding specificity. *Nucleic Acids Res.* 32:6673–6682.

58. Paillard, G., and R. Lavery. 2004. Analyzing protein-DNA recognition mechanisms. *Structure.* 12:113–122.

59. Flatters, D., and R. Lavery. 1998. Sequence-dependent dynamics of TATA-Box binding sites. *Biophys. J.* 75:372–381.

60. Tutorial 9, AMBER web site. 2009. http://ambermd.org/tutorials/advanced/tutorial9/.

61. Case, D. A., T. A. Darden, …, P. A. Kollman. 2008. AMBER 10. University of California, San Francisco, CA.

62. Bernstein, F. C., T. F. Koetzle, …, M. Tasumi. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.

63. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

64. RCSB PDB web page. 2009. http://www.pdb.org.

65. Hornak, V., R. Abel, …, C. Simmerling. 2006. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins.* 65:712–725.

66. Pérez, A., I. Marchán, …, M. Orozco. 2007. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophys. J.* 92:3817–3829.

67. Wang, J., P. Cieplak, and P. A. Kollman. 2000. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* 21:1049–1074.

68. Jorgensen, W. L., J. Chandrasekhar, …, M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

69. Mark, A. E., and W. F. van Gunsteren. 1994. Decomposition of the free energy of a system in terms of specific interactions. Implications for theoretical and experimental studies. *J. Mol. Biol.* 240:167–176.

70. Boresch, S., G. Archontis, and M. Karplus. 1994. Free energy simulations: the meaning of the individual contributions from a component analysis. *Proteins.* 20:25–33.

71. Smith, P. E., and W. F. van Gunsteren. 1994. When are free energy components meaningful? *J. Phys. Chem.* 98:13735–13740.

72. Boresch, S., and M. Karplus. 1995. The meaning of component analysis: decomposition of the free energy in terms of specific interactions. *J. Mol. Biol.* 254:801–807.

73. Brady, G. P., A. Szabo, and K. A. Sharp. 1996. On the decomposition of free energies. *J. Mol. Biol.* 263:123–125.

74. Sorokin, V., K. Severinov, and M. S. Gelfand. 2009. Systematic prediction of control proteins and their DNA binding sites. *Nucleic Acids Res.* 37:441–451.

75. Case, D. A., T. A. Darden, …, P. A. Kollman. 2010. AMBER 11. University of California, San Francisco, CA.

76. Bennet, C. H. 1976. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* 22:245–268.

77. Shirts, M. R., and J. D. Chodera. 2008. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* 129:124105.

78. Shirts, M. R., and V. S. Pande. 2005. Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. *J. Chem. Phys.* 122:144107.

79. VMD web page. 2009. http://www.ks.uiuc.edu/Research/vmd/.

80. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38, 27–28.