Published in final edited form as:

J Phys Chem B. 2017 May 25; 121(20): 5151–5161. doi:10.1021/acs.jpcc.6b12450.

Deficiencies in Molecular Dynamics Simulation-Based Prediction of Protein-DNA Binding Free Energy Landscapes

Morteza Khabiri[†] and Peter L. Freddolino^{*,†,‡}

[†]Department of Biological Chemistry, University of Michigan Medical School, Ann Arbor, MI USA

[‡]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI USA

Abstract

Transcriptional regulation allows cells to match their gene expression profiles to their current requirements based on environment, cellular physiological state, and extracellular signals. DNA binding transcription factors are major agents of transcriptional regulation, and bind to DNA with a factor-specific sequence preference to exert regulatory effects. A crucial step in unraveling the logic of a regulatory network is determining the sequence-specific binding affinity landscapes for the transcription factors in it. While such landscapes can be measured experimentally, the ability to predict them computationally would both reduce the effort required to obtain the needed data, and provide additional insight into the key interactions shaping protein-DNA interactions. Here we apply free energy calculations based on all-atom molecular dynamics simulations to predict the changes in binding free energy for all single base pair perturbations of the binding sites for four eukaryotic transcription factors for which high-quality experimental data exist. We find that the simulated results both vastly overestimate the magnitude of changes in binding free energy, and frequently predict the incorrect signs. These simulations will nevertheless serve as a jumping-off point for refining our current representation of protein-DNA interactions to allow quantitative reproduction of experimental data on such systems in the future.

petefred@umich.edu, Phone: +1 (734) 647-5839.

Supporting Information Available

- ELK_DDA.xlsx: Table of calculated and experimental free energy parameters for ELK1
- GCM_DDA.xlsx: Table of calculated and experimental free energy parameters for GCM1
- MAX_DDA.xlsx: Table of calculated and experimental free energy parameters for MAX
- PPR1_DDA.xlsx: Table of calculated and experimental free energy parameters for PPR1
- ener_landscapes_orig.pdf: Plot of binding free energy changes similar to Fig. 3, for a separate set of simulations where the ends of the DNA were not constrained, and the beginning conformation of the DNA in each case was standard B-DNA aligned to the crystallographic binding site instead of the crystallographic structure itself. Note that two of the targets used here differ from those in the main text: in place of PPR1 and MAX, we used the additional human transcription factors TEAD4 and SPIB. Results from these alternate simulations are qualitatively very similar to those discussed in the main text, in terms of an overall poor reproduction of experimental binding free energy landscapes. All binding data from CisBP.⁶
- RMSD_plots.pdf: RMSD traces similar to those shown in Fig. 6 for selected residues from the GCM1, MAX, and PPR1 simulations. Note that the higher backbone RMSDs observed in some of the PPR1 cases reflect motion of a loop on the side of the protein opposite the DNA binding site.

Graphical Abstract

Introduction

Transcription factors (TFs) and other DNA-binding proteins shape the behavior of all cells, coordinating appropriate gene expression patterns in response to internal or external cues. Our ability to understand and manipulate cellular behavior thus relies on our understanding of their transcriptional regulatory networks, including the locations and occupancies of transcription factor binding sites. A crucial determinant of the binding locations of a given TF is the landscape of binding affinities of that TF for different DNA sequences.

At the highest possible resolution, a binding affinity landscape could be represented as a list of binding free energies for every possible DNA sequence up to the length of the TF's recognition site. TFs typically recognize sequences approximately 10 base pairs (bp) in length,¹ and thus a comprehensive representation of the binding landscape of a typical TF would require a table of 4^{10} binding free energies. In practice, these landscapes are typically approximated using simplified models, such as position weight matrices (PWMs), which provide a basepair-level decomposition and allow approximation of the binding free energy for an arbitrary sequence.² More complex models including interaction terms between different portions of the target sequence have also been applied by Clifford and Adami³. They showed that the binding sites of Dorsal transcription factor (involved in *Drosophila* development) contain on average around 0.5 bits information content of adjacent sequences.

A variety of experimental methods exist to measure the binding free energy landscapes of DNA-binding proteins, including protein binding microarrays (PBMs)⁴ and high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX) experiments.⁵ While powerful in allowing PWMs or other binding landscape approximations to be built for target transcription factors, these experiments remain expensive, labor intensive, and provide little structural insight into the nature of the protein-DNA complex or the specific interactions governing affinity landscapes. While the necessarily heroic resources have been marshalled to perform such experiments for a wide variety of transcription factors from model eukaryotes,⁶ performing similar experiments for all transcription factors of the hundreds of other species relevant to human health (most notably, pathogenic and commensal micro-organisms) is completely impractical. In addition, the dramatic effects of transcription factor binding on DNA structure (*e.g.*, Finkel and Johnson⁷) likely plays a major role in the regulatory effects of these proteins. The needs for a more efficient method to obtain sequence affinity landscapes, insight into the structural basis for how changes to the DNA sequence affect binding affinity, and information on the structural effects of

different transcription factors on DNA all argue for the large-scale application of molecular modeling to study the affinity landscapes of DNA-binding proteins. A wide variety of approximate models have been applied to calculate the binding affinity landscapes of transcription factors. While these models have performed well in identifying the native binding site for a particular protein, the approximations taken by most approaches cause them to perform poorly in efforts to more broadly map the binding affinity of a given protein for a variety of sequences (reviewed by Liu and Bradley ⁸).

Pilot applications of all-atom molecular dynamics (MD) simulations with explicit solvent, while expensive, have been particularly successful in reproducing experimental binding free energy landscapes. Seeliger and co-workers applied such simulations for all possible single base pair perturbations of the native target site for the zinc-finger transcription factor Zif268,⁹ and reproduced experimental results to near-chemical accuracy (average absolute error of 0.31 kcal/mol to experimental values). Proper rank ordering was also observed in a similar study calculating all perturbations of a single base pair in the C-protein of the restriction-modification system Esp1396I.¹⁰

In principle, the application of molecular dynamics simulations to calculate binding free energy landscapes of TFs and other DNA-binding proteins has several potential benefits: computational predictions (i) would be easily scalable with computing resources, (ii) provide biophysical insight into the nature of the interactions governing specificity landscapes, and (iii) through the information obtained, would pave the way to simplified methods for structure-based prediction that would achieve similar predictive accuracy without the need for full atomistic free energy calculations.

To assess the current readiness of all-atom molecular dynamics simulations for large-scale calculation of TF binding free energy landscapes, we applied computational protocols similar to those applied by Seeliger and co-workers⁹ to predict the binding free energy changes for all single-bp perturbations of the recognition site for four eukaryotic transcription factors: ELK1, GCM1, MAX, and PPR1 (the first three from humans and the last from *S. cerevisiae*). Comparing our results to recent HT-SELEX data sets, we observe generally poor correlations between the computational predictions and experimental data, with the magnitude of free energy changes often over-estimated and frequently with incorrect signs. Analysis of the results suggests that changes in DNA structure upon binding of different sequences, including the balance of that changing structure with nonbonded protein-DNA interactions, play a key role in defining the binding free energy landscape. Our findings illustrate potential directions for future improvements of MD force fields for better treatment of these interactions, and also raise a cautionary note that relaxation times far beyond those of typical free energy simulations may be required to obtain converged conformational ensembles for protein-DNA complexes.

Computational Methods

Structures

To understand the biophysical basis for TF selectivity of a TFs binding site, it is necessary to have high-resolution data at the atomistic level of the TF in complex with various DNA

sequences. Thus, the first step is obtaining the initial structure of the target DNA and the protein bound to a DNA target site. The structure of all selected TF-DNA complex were taken from Protein Data Bank, human ELK1 (PDB code 1DUX), human GCM1 (PDB code 1ODH), human MAX (PDB code 1HLO) and *S. cerevisiae* PPR1 (PDB code 1PYI); the structures and their corresponding position weight matrices are shown in Figure 1. The double stranded DNA in each crystal structure was extended on both sides such that the length of oligonucleotide reached 25–26 bp. In order to have a compatible structure for added oligomers relative to the DNA in crystal structure the consensus protocol was performed as discussed below. The end extensions for the complex structures, and the isolated DNA systems with their complementary strands, were built with the right-handed canonical Bform parameters using the Nucleic Acid Builder (NAB) program¹¹ in the AmberTools suite (mutated sequences were built using equivalent procedures); the DNA sequences used in each case are shown in Table 1. The magnesium ions were modeled near the DNA in positions that minimize the electrostatic potential energy using the cionize plug-in tool in VMD;¹² in each case we added sufficient magnesium ions to neutralize the DNA. We used cubic simulation boxes for each TF-DNA complex and also for the DNA-only simulations; initial conformations used 24 Å of padding from the solute to the nearest periodic boundary, and we confirmed that in no case did a solute come within the short-range nonbonded cutoff of its periodic image.

Computational Details

All MD simulations were performed using the pmemd program of Amber14. We used the AMBER ff14sb force field¹³ to represent the protein. For DNA, we used a recently corrected variant of the parmbsc1 force field¹⁴ and for monovalent ions we used the Joung-Cheatham modified ion parameters.¹⁵ The water molecules from the initial crystal structures were kept while preparing the system. TIP3P water molecules and 150 mM potassium chloride were randomly placed around the solute to achieve an electrically neutral solvated system. In the case of Mg^{2+} , specialized divalent cationic parameters were used as described in Saxena and Sept¹⁶. To begin MD simulation on the consensus recognition sequence of each TF, energy minimization was performed followed by slow heating from 0 K to 300 K under constant pressure over a period of 200 ps using a harmonic restraint of 100 kcal mol⁻¹ Å⁻² on all heavy atoms. To match the structure of the added DNA segments in TF-DNA complex crystal structure, we define two harmonic restrains, one on the added DNA and the other one on the TF-DNA complex in the crystal structure. The restraint on the added DNA segments was slowly relaxed from 100 to 0 kcal mol⁻¹ Å⁻² during 2 ns using constant temperature (300 K) and pressure (1 bar) conditions via the Berendsen algorithm. After relaxing, the restraint was removed from the TF-DNA complex with the same procedure. In order to avoid flipping of bases in both ends of the DNA structure, as well as the rotation of solute in the box, a position restraint was applied on the P atoms of nucleotides at one end and a distance restraint on the other end. The final system, with no harmonic restrains, was equilibrated for 20 ns in order to reach a proper density and system size. After that, the average box size from the last 5 ns was chosen as a starting point for NVT simulation. Once the box size was set based on the NPT equilibration for the consensus sequence of each transcription factor, it remained fixed for all sequence variants that we studied.

Temperature was controlled using Langevin dynamics with a coupling constant of 5 ps^{-1} . Short-range nonbonded interactions were cut off at $9 \text{ \AA}$. Simulations employed periodic boundary conditions, using the particle mesh Ewald method for long-range electrostatic interactions. During the MD calculation, hydrogen vibrations were removed using SHAKE bond constraints, allowing a longer time step of 2 fs . Post-processing analysis of the trajectories was carried out by using cpptraj module¹⁷ and custom R and python scripts.

Construction of hybrid topologies

After 150 ns equilibrium simulations of the wildtype DNA and protein-DNA complexes, all possible single base pair mutations of the presumed recognition site were generated and subjected to a separate 150 ns equilibrium simulation. After separate equilibrium simulations of the original and mutant sequence, we constructed hybrid topologies for each mutation to be considered by using the tleap and parmed programs in Amber. Before hybridization, each residue was mutated using custom python scripts. Equivalent methods were applied for both the protein-DNA complexes and free DNA duplexes.

Free energy calculations

To investigate the changes in binding free energy of the selected TFs and their DNA model systems in the presence of mutations, we used the thermodynamic cycles as shown in Figure 2a. Each ΔA value was calculated using a non-equilibrium Fast Growth Thermodynamic Integration (FGTI) approach similar to the CGI method employed by Seeliger et al.⁹: we chose a series of snapshots from the equilibrium trajectory of each endpoint simulation, and performed non-equilibrium thermodynamic integration (TI) calculations for the transition between the base pair from the equilibrium simulation and the target mutation (Fig. 2b). As the work distributions were frequently non-normal, we then employed the Bennett Acceptance Ratio (BAR, as described for non-equilibrium FGTI calculations in Shirts et al.¹⁸) instead of the CGI estimator¹⁹ for the overall change in free energy between the two states. For each of the endpoint equilibrium trajectories, we selected 32 evenly-spaced snapshots (64 for ELK1) taken from the last 50 ns of the trajectory.

The transition between each pair of DNA sequences was described using a single step, dual-topology approach and soft-core potential for both electrostatics and Lennard-Jones interactions as implemented in Amber14 with $\alpha=0.2$ and $\beta=9.0$ in Steinbrecher et al.²⁰ the actual perturbation step. Before running TI calculation, each hybrid systems were minimized for 2000 steps and then heated up for 20 ps . The λ value increment by 0.0001 in every 15 steps out of 150000 steps between 0.0 and 1.0 . The $dV/d\lambda$ gradients were collected in every steps. The work distribution estimate by integration over $dV/d\lambda$ and the free energy difference between wildtype and mutant states $dV/d\lambda$ gradients were collected in every steps.

To examine the uncertainties associated with our free energy estimates, we calculated the correlation times of the work values obtained from subsequent snapshots of the equilibrium simulations, and employed block-bootstrap estimates for the standard error and confidence intervals with a block size at least as large as the correlation time. In addition, we calculated

the overlap between the forward and reverse work distributions by numerical integration, and flag as unreliable any cases with less than 10% overlap in the discussions below.

Notation

For the sake of simplicity, throughout the text we will frequently refer to the DNA sequence or mutations to it referring only to the sequence of one strand (which we will synonymously refer to as the forward strand or 5'→3' direction), but in all cases mutations are made using appropriate base pairs (that is, a C→T mutation at position 10 indicates replacement of a CG base pair at that position with a TA base pair). Typically only the sequence of the forward strand will be listed in figures and tables. When indicating specific nucleotides, we refer to nucleotides on the forward strand with subscript numbering (*e.g.*, C₁₀ for a cytosine at position 10 on the forward strand), and nucleotides on the reverse strand by appending a prime to the numbering of their partner on the forward strand (*e.g.*, G₁₀' to refer to the guanine base paired with the cytosine of the previous example). All sequence numbering for DNA sequences are numbered from the beginning of the forward strand in the actual simulated DNA molecule (thus, the portion of the sequence being considered for mutation typically begins between position 8 and 10).

Results and discussion

We applied the free energy calculations described above to determine the binding free energy landscapes for each of the four transcription factors listed in Table 2 to all single base pair perturbations of their recognition sites. The results are shown in Fig. 3, and compared with experimental data. We provide details of the correlations between binding free energy landscapes obtained using various methods in Table 3. It is readily apparent that the MD based binding free energy changes are very poorly correlated with the experimental data, with Spearman correlations typically on the order of 0.2. In addition, even the magnitude of the free energy changes observed based on the MD calculations is substantially larger than the experimentally observed changes.

Binding free energy changes do not correlate strongly with salt bridge or hydrogen bonding patterns

Given the surprisingly poor performance of our calculations in reproducing experimental results on binding free energies changes, we sought to identify the key factors contributing to the discrepancies. As described above in the Computational Methods section, we confirmed that the vast majority of our free energy calculations showed good overlap between the forward and reverse work distributions, and performed careful error analysis using a block-bootstrap based approach to rule out technical sources of error being responsible for the observed discrepancies. Instead, we turned to the biophysics of the protein-DNA interactions as a likely source of difficulty. Given that these interactions are driven largely by hydrogen bonding and salt bridges formed between the protein and DNA, we examined the location and occupancy of all such polar interactions across the sequences considered in our study.

As seen in Figure 4, however, in general there is no systematic relationship between the number of polar protein-DNA interactions present in different mutant sequences, and the calculated changes in binding free energy. In particular, it is not possible to do something so simple as count salt bridges to the DNA backbone or hydrogen bonds to the mutated base and determine the resulting ΔA .

Effects of DNA structure on binding free energy changes

Given that the variations that we observed in binding free energies typically could not be explained solely on the basis of polar interactions between protein and DNA, it is useful to consider how properties of the DNA itself correlate with the calculated ΔA values. We thus calculated the values of the 16 DNA structural parameters (as defined by Lu and Olson²¹ and implemented in `cpptraj`¹⁷) and identified those showing a Spearman correlation of magnitude greater than 0.2 with the ΔA profiles for each protein (shown in Table 4). As observed both there and in Fig. 5, in no case is there a clear correlation between DNA structural characteristics and the observed binding free energy differences (since all q-values following Benjamini-Hochberg correction, corresponding to the FDR threshold at which a given correlation would be considered significant, are greater than 0.1). At present it is unclear whether this in fact reflects a general lack of such structural changes as a component in the thermodynamics of protein-DNA interactions, a deficiency of modern force fields in not properly modeling such changes, or a lack of sufficient sampling to allow any such structural changes to be observed.

Technical Issues in Free Energy Calculation

The substantial disagreement in both sign and magnitude between our free energy calculations and experimentally observed values compels us to consider whether systematic errors in our simulation setup give rise to the observed discrepancies. Indeed, free energy calculations are notoriously difficult, subject to errors in system setup and convergence which can be very difficult to detect. By all typical metrics, however, no such concerns are generally apparent in our simulations. In particular, the forward and reverse work distributions show good overlaps (we censor all cases in Fig. 3 with less than 10% overlap); examination of RMSD plots for the endpoint equilibrium simulations show good convergence without drift in the window being sampled for fast growth calculations, both for the protein/DNA backbones and for binding site residues (see, *e.g.*, Fig. 6); the correlation times of the work values obtained from subsequent snapshots are low, on the order of 1–2 frames typically (and no more than 8), indicating good sampling; and our block bootstrap based error calculations will explicitly account for any systematic drift present due to undersampling on the observed timescales. We also note that calculations using a similar setup without the constraints applied to the DNA here, and using a different equilibration protocol, yield similar results (see Supporting Information file `ener_landscapes_orig.pdf`) in terms of the degree of disagreement and magnitude of MD values. We also tested a small subset of mutations using a standard slow growth thermodynamic integration instead of the fast growth/Crooks fluctuation theorem protocol employed here, but obtained similar results (data not shown).

While we have excluded many possible sources of error in the free energy calculation, there remains one crucial “unknown unknown”: whether the relaxation of the DNA-protein interface upon introduction of a mutation occurs over much longer timescales than those sampled here. Given the absence, in general, of major slow structural rearrangements observed in our equilibrium trajectories, such processes would have to occur far slower than the 100 ns timescale to escape our notice, but could in principle account for the errors observed here if we are in fact not sampling the true, physically relevant conformations of the protein-DNA complexes in the presence of altered DNA sequences. The equilibration of water around and within the DNA-protein interface may also play a key role in the equilibration time of the system. In examining water occupancy around the complexes throughout our equilibrium simulations, we observe that there are few well-ordered waters near the binding site except in the GCM1 system, with most that are present arranged along the phosphate backbone of the DNA (see Fig. 7a). Notably, water occupancy does not change strongly in the presence of different mutations, even in a case showing a very large change in binding free energy (*e.g.*, Fig. 7b). Again, however, changes on longer timescales than those sampled here cannot be ruled out.

Conclusions

Building on recent optimism regarding the ability of molecular dynamics-based free energy calculations to model accurately the thermodynamic landscapes of protein-DNA interactions, we have performed what is, to our knowledge, the broadest survey to date of all-atom MD based predictions of transcription factor binding free energy landscapes. In contrast with previous studies (*e.g.*, the highly successful predictions of Seeliger et al.⁹, on which we based our simulation protocols), we observed disappointing performance on each of the four targets examined, with no case showing a significant correlation between the predicted and experimental binding free energy profiles, and all cases showing unrealistically high magnitudes of calculated changes in binding free energy. While disappointing, we view these findings as an opportunity for more careful examination of the physics underlying protein-DNA interactions and identification of effects which are not being treated sufficiently in the current generation of force fields. Indeed, we may view the current state of force fields with regard to balancing protein-DNA interactions as being akin to that of protein force fields roughly ten years ago (reviewed by Freddolino et al.²²). In that case, attempted protein folding simulations lead to the identification of systematic biases in the balance of secondary structure stabilities (*e.g.*,^{23–25}), followed by substantial efforts at force field improvements^{26,27} that have lead more recently to far more reliable and transferable simulation of protein folding.^{28–30}

In the case of the protein-DNA interactions considered here, we are able to identify several potential trouble spots which will need to be carefully investigated to improve the performance of MD-based efforts to predict protein-DNA binding free energies. In general we did not observe substantial changes in the number of polar contacts between the protein and DNA in consensus vs. experimentally disfavored positions, which may reflect an inability of the force field to properly stabilize the particular interactions used in nature to stabilize specific interactions. We also observed no substantial correlations between DNA structure changes and observed free energy profiles. The role of the DNA structure itself in

DNA-protein interactions has recently received increasing attention,^{31,32} and likely an improved balance of terms describing DNA conformation with those involved in the protein-DNA nonbonded interaction will be crucial to successful quantitative simulation of protein-DNA interactions.

A recent surge of interest and effort has arisen in refinement and testing of DNA force fields, resulting in substantial improvements to both AMBER (Parmbsc1¹⁴ and OL15,³³ both recently tested by Galindo-Murillo et al.³⁴) and CHARMM (CHARMM36,³⁵ tested further by Galindo-Murillo et al.³⁶) DNA parameters. In both cases recent refinements have yielded clear improvements in the reproduction of physically realistic conformations of duplex DNA. Here we make use of one very recent set of refinements to the AMBER family DNA forcefields. However, compared with the recent reparameterizations, which have focused on DNA in isolation, here we also contend with the additional complication of balancing protein-protein, protein-DNA, and DNA-DNA interaction terms to obtain accurate answers. We strongly suspect that proper treatment of such interactions will require careful consideration of both the high quality binding free energies now available through methods such as HT-SELEX, and more detailed correlation of molecular dynamics results with NMR and similar data (as in Esadze et al.³⁷), in additional parameterization efforts to improve the overall treatment of the protein-DNA interface in modern MD simulations.

Based on the magnitude of the errors observed, and in particular the physically unfeasible differences between our calculated changes in binding free energies and those observed in experiment (with the calculated changes an order of magnitude larger than the experimental values), we suspect that the difficulties encountered here go far beyond parameterization problems, and extend to sampling problems that were not even detectable over the timescales of our simulations. Recent simulations of DNA alone have illustrated that full convergence of the conformational ensemble of duplex DNA in solution occurs on timescales of a few microseconds.^{34,36} While the present work was predicated on the assumption that the presence of a high-quality crystal structure of the protein-DNA interface, along with only limited local perturbations of the DNA sequence, would be sufficient to ensure faster convergence of the bound structure ensembles, the results observed here suggest examination of longer timescale relaxation of the protein-DNA complex as one likely avenue to improve the predictive performance of MD-based predictions of protein-DNA interaction free energies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank all members of the Freddolino laboratory for many useful discussions, Prof. Gary Stormo for encouragement and for providing early access to a re-analysis of HT-SELEX data, and Prof. Tom Cheatham and Dr. Daniel Roe for technical assistance and advice on simulation setup. Funding was provided by an NIH R00 (GM097033) to PLF. Computer time on Blue Waters was provided via the Great Lakes Consortium for Petascale Computation under project IDs glcpc_jt5 and baco. PLF is forever grateful to Prof. Klaus Schulten for years of mentorship, advice, and support.

References

1. Stewart AJ, Hannonhalli S, Plotkin JB. Why transcription factor binding sites are ten nucleotides long. *Genetics*. 2012; 192:973–985. [PubMed: 22887818]
2. Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics*. 2000; 16:16–23. [PubMed: 10812473]
3. Clifford J, Adami C. Discovery and information-theoretic characterization of transcription factor binding sites that act cooperatively. *Phys Biol*. 2015; 12:056004. [PubMed: 26331781]
4. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription factor binding site specificities. *Nat Biotech*. 2006; 24:1429–1435.
5. Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. DNA-binding specificities of human transcription factors. *Cell*. 2013; 152:327–339. [PubMed: 23332764]
6. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. 2014; 158:1431–1443. [PubMed: 25215497]
7. Finkel S, Johnson R. The Fis protein: it's not just for DNA inversion anymore. *Mol Microbiol*. 1992; 6:3257–3265. [PubMed: 1484481]
8. Liu LA, Bradley P. Atomistic modeling of protein-DNA interaction specificity: progress and applications. *Curr Opin Struct Biol*. 2012; 22:397–405. [PubMed: 22796087]
9. Seeliger D, Buelens FP, Goette M, de Groot BL, Grubmüller H. Towards computational specificity screening of DNA-binding proteins. *Nucl Acids Res*. 2011; 39:8281–8290. [PubMed: 21737424]
10. Beierlein FR, Kneale GG, Clark T. Predicting the effects of basepair mutations in DNA-protein complexes by thermodynamic integration. *Biophys J*. 2011; 101:1130–1138. [PubMed: 21889450]
11. Macke, TJ., Case, DA. *Molecular Modeling of Nucleic Acids*. Vol. 682. American Chemical Society; 1998. Modeling unusual nucleic acid structures; p. 373-393. Chapter 24
12. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph*. 1996; 14:33–38. [PubMed: 8744570]
13. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theo Comp*. 2015; 11:3696–3713.
14. Ivani I, Dans PD, Noy A, Pérez A, Faustino I, Hospital A, Walther J, Andrio P, Goñi R, Balaceanu A, et al. Parmbsc1: a refined force field for DNA simulations. *Nat Meth*. 2016; 13:55–58.
15. Joung IS, Cheatham TE III. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J Phys Chem B*. 2008; 112:9020–9041. [PubMed: 18593145]
16. Saxena A, Sept D. Multisite ion models that improve coordination and free energy calculations in molecular dynamics simulations. *J Chem Theor Comp*. 2013; 9:3538–3542.
17. Roe DR, Cheatham TE III. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J Chem Theor Comp*. 2013; 9:3084–3095.
18. Shirts MR, Pitera JW, Swope WC, Pande VS. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J Chem Phys*. 2003; 119:5740–5761.
19. Goette M, Grubmüller H. Accuracy and convergence of free energy differences calculated from nonequilibrium switching processes. *J Comp Chem*. 2009; 30:447–456. [PubMed: 18677708]
20. Steinbrecher T, Mobley DL, Case DA. Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *J Chem Phys*. 2007; 127:214108. [PubMed: 18067350]
21. Lu XJ, Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucl Acids Res*. 2003; 31:5108–5121. [PubMed: 12930962]
22. Freddolino PL, Harrison CB, Liu Y, Schulten K. Challenges in protein folding simulations: Timescale, representation, and analysis. *Nat Phys*. 2010; 6:751–758. [PubMed: 21297873]

23. Best RB, Buchete N, Hummer G. Are current molecular dynamics force fields too helical? *Biophys J*. 2008; 95:L07–L09. [PubMed: 18456823]
24. Freddolino PL, Liu F, Gruebele M, Schulten K. Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophys J*. 2008; 94:L75–L77. [PubMed: 18339748]
25. Freddolino PL, Park S, Roux B, Schulten K. Force field bias in protein folding simulations. *Biophys J*. 2009; 96:3772–3780. [PubMed: 19413983]
26. Mittal J, Best RB. Tackling force-field bias in protein folding simulations: folding of Villin HP35 and Pin WW domains in explicit water. *Biophys J*. 2010; 99:L26–L28. [PubMed: 20682244]
27. Piana S, Lindorff-Larsen K, Shaw DE. How robust are protein folding simulations with respect to force field parameterization? *Biophys J*. 2011; 100:L47–L49. [PubMed: 21539772]
28. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. *Science*. 2011; 334:517–520. [PubMed: 22034434]
29. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE. Systematic validation of protein force fields against experimental data. *PloS one*. 2012; 7:e32131. [PubMed: 22384157]
30. Piana S, Klepeis JL, Shaw DE. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr Opin Struct Biol*. 2014; 24:98–105. [PubMed: 24463371]
31. Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, Bussemaker HJ, Gordân R, Rohs R. Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc Natl Acad Sci*. 2015; 112:4654–4659. [PubMed: 25775564]
32. Stormo G, Roy B. DNA structure helps predict protein binding. *Cell Systems*. 2016; 3:216–218. [PubMed: 27684185]
33. Zgarbová M, Sponer J, Otyepka M, Cheatham TE III, Galindo-Murillo R, Jurek P. Refinement of the sugar–phosphate backbone torsion beta for AMBER force fields improves the description of Z- and B-DNA. *J Chem Theor Comp*. 2015; 11:5723–5736.
34. Galindo-Murillo R, Robertson JC, Zgarbová M, Šponer J, Otyepka M, Jurek P, Cheatham TE III. Assessing the current state of AMBER force field modifications for DNA. *J Chem Theor Comp*. 2016; 12:4114–4127.
35. Hart K, Foloppe N, Baker CM, Denning EJ, Nilsson L, MacKerell AD Jr. Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium. *J Chem Theor Comp*. 2011; 8:348–362.
36. Galindo-Murillo R, Roe DR, Cheatham TE. Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d (GCACGAACGAACGAACGC). *Biochem Biophys Acta*. 2015; 1850:1041–1058. [PubMed: 25219455]
37. Esadze A, Chen C, Zandarashvili L, Roy S, Pettitt B, Iwahara J. Changes in conformational dynamics of basic side chains upon protein-DNA association. *Nucl Acids Res*. 2016; 44:6961–6970. [PubMed: 27288446]

Figure 1.

Overview of the transcription factors considered in the present study. (a) Initial structures of the protein-DNA complexes considered for each case, with each TF's recognition site in the DNA highlighted using blue and red strands. (b) PWMs for each of the transcription factors considered here, obtained from CisBP.⁶ Numbering follows the convention used throughout this paper (described in detail below); numbering begins from one terminus of the full simulated DNA duplex. A purple line in each case indicates the extent of the region studied by mutation.

Figure 2.

Schematic of computational setup for calculating binding free energy changes. (a) Thermodynamic cycle for calculation of the change in binding free energy (ΔG_{bind}) for a mutation of the target DNA sequence from DNA to DNA*; our notation and cycle follow that of Seeliger *et al.*⁹ Calculation of $\Delta G1$ and $\Delta G2$ are sufficient to obtain the needed information. (b) Computational procedure for FGTI based free energy calculation of the effects of a mutation. Two long simulations are performed in the presence and absence of the mutation, and then a series of short nonequilibrium trajectories in which the mutation is rapidly applied through alchemical perturbation allows calculation of the associated free energy change.

Figure 3.

Comparison of changes in binding free energy for all single base pair perturbations of each of the four transcription factors considered. The reference sequence is shown on the top of each panel and the changes in binding free energy are plotted for the new sequences shown along the bottom. Values are compared with experimental data from human or (in some cases) mouse versions of the transcription factor (from CisBP⁶) for ELK1, GCM1, and MAX, or to yeast data in the case of PPR1. Note the different axes for the MD results on the left vs. CisBP results on the right. Error bars for the MD data show a 95% confidence interval based on a blocked bootstrap. 'x' symbols are plotted for the MD where there was poor overlap (< 10%) between the forward and reverse work distributions, which might lead to unexpectedly high errors.

Figure 4.

Comparison of hydrogen bonding rates at each position for the consensus sequence and all mutant sequences considered. In each case we show the average number of hydrogen bonds present between the protein and either the base or backbone (as noted) of each nucleotide, with interactions to the “top” strand of the DNA (that with the sequence shown) above the x axis, and interactions with its reverse complement below the axis. Values for the consensus sequence are shown as a dashed green line; all of the mutated sequences are shown colored by the calculated $\Delta\Delta A$ values, scaled from red (most positive $\Delta\Delta A$) to blue (most negative $\Delta\Delta A$). The region of the DNA considered in the mutational calculations is highlighted with a purple bar. Interactions were calculated using the AmberTools *cpptraj* tool with default settings.

Figure 5.

Structural characteristic of each target TF with the highest Spearman correlation to the observed binding free energy profiles for each target protein (see also Table 4; as noted there, none of the correlations are statistically significant). Structural parameters are average deviations observed in the protein-DNA complex structure from the values observed for the DNA alone in solution, calculated as described in the text. We show a dashed green line for the reference sequence, the three mutations with the most positive and most negative $\Delta\Delta A$ values as colored lines (matching the displayed color scale), and the average of all other mutations shown as a solid gray line. The absence of systematic differences between mutations with very positive and very negative $\Delta\Delta A$ values is further evidence against a simple explanation for the observed free energy profiles in terms of structural deviations in the DNA.

Figure 6.

RMSDs relative to the starting structure (before equilibration) for protein and DNA backbone atoms (Backbone; considering the binding site on the DNA and all protein residues excluding the first and last five) or for all heavy atoms of residues involved in the protein-DNA interface (Binding site; using the DNA bases in the defined binding site and all protein residues within 3 Å in the initial structure) over the course of the indicated simulations; in all cases the structures were aligned based on the backbone residues. Note that snapshots for fast growth thermodynamic integration were taken in all cases from the 100–150 ns range of the simulation. Here we show results for mutations of G₁₇ and T₁₈ in ELK1; similar plots are provided as supplementary material in RMSD_plots.pdf for cases drawn from the GCM1, PPR1, and MAX systems. In all cases we observe that the traces are usually stationary, show low RMSDs (almost always below 3 Å) for both the backbone and interface, and do not show substantial deviations even for cases with very positive or very negative ΔA values.

Figure 7.

Water occupancy in the vicinity of protein-DNA complexes of interest, calculated using the VMD volmap plugin with a 1 Å grid spacing. Red, green, and black isosurfaces are shown at the median bulk occupancy over the entire unit cell, 1.5 times the median, and two times the median, respectively (the black isosurface represents slightly over 90% occupancy). (a) Water occupancy around each of the four target protein-DNA complexes during simulations of the original sequences. (b) Water occupancies upon each possible mutation of the G₁₇ position in the ELK1 system, which showed highly anomalous changes in binding free energy.

Table 1

Complete sequences of double stranded DNA sequences used in our simulations for each TF. The italicized region is considered the binding site; the consensus sequence is shown, and changes in binding free energy were calculated for all single-basepair perturbations of that sequence.

TF name	Target sequence
ELK1	ACGTACGTG <i>ACCGGAAGT</i> AGCATGCA
GCM1	AGTCAGTCAG <i>ATGCGGGT</i> ACGACTGA
PPR1	AGTCAG <i>TCGGCATTCTCCG</i> AGACTGA
MAX	CAGTCAGCA <i>CCACGTGGT</i> GACTGACT

Table 2

Details of the TF structures used in our simulations.

TF Name	TF family	PDB code
ELK1	ETS	1DUX
GCM1	GCM	1ODH
PPR1	Zinc finger	1PYI
MAX	bHLH	1HLO

Table 3

Spearman correlations (p-values in parentheses) between the indicated sets of binding free energy changes. “Human”, “Mouse”, and “Yeast” refer to the same experimental data sets referenced in Figure 3.

TF name		vs Human	vs Mouse
ELK1	MD	0.257 (0.196)	0.166 (0.408)
	Human	–	0.297 (0.133)
GCM1	MD	–0.038 (0.849)	0.169 (0.429)
	Human	–	0.578 (0.003)
MAX	MD	0.014 (0.945)	0.384 (0.048)
	Human	–	0.423 (0.028)
		vs Yeast	
PPR1	MD	0.278 (0.222)	–

Table 4

DNA characteristics showing the largest-magnitude Spearman correlations with the observed spectrum of free energy changes. For each transcription factor, we consider the changes in each of 16 characteristics of the DNA structure between the bound and unbound forms, and calculate the Spearman correlation between the sums of those differences along the length of the sequence and the A profiles. All cases with Spearman correlation $|\rho| > 0.2$ are shown here, along with p-values and Benjamini-Hochberg corrected q-values testing the null hypothesis no correlation. As is apparent from the q-values column, none of the correlations should be considered statistically significant.

TF name	Characteristic	Spearman ρ	p-value	q-value
ELK1	Rise	0.420	0.029	0.467
	Stagger	-0.371	0.057	0.581
	Roll	-0.245	0.217	0.695
	Stagger	-0.574	0.002	0.110
	HB	0.339	0.084	0.609
GCM1	Shift	-0.324	0.100	0.638
	Rise	-0.307	0.119	0.648
	Opening	-0.297	0.133	0.648
	Stretch	-0.260	0.190	0.648
	Minor	0.259	0.192	0.648
	Twist	0.565	0.008	0.244
	Minor	-0.509	0.018	0.393
	Shear	0.422	0.057	0.581
	Opening	0.347	0.124	0.648
	Major	-0.316	0.163	0.648
PPR1	Buckle	0.249	0.276	0.784
	Tip	0.247	0.281	0.784
	Slide	0.240	0.294	0.784
	Tilt	0.362	0.064	0.581
	Buckle	-0.337	0.086	0.609
	Tip	-0.287	0.147	0.648
	Shear	-0.269	0.175	0.648
	Shift	0.264	0.184	0.648
	Propeller	-0.215	0.282	0.784
MAX				