Pacific Symposium on Biocomputing 12:484-495(2007)

AB INITIO PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES

L. ANGELA LIU and JOEL S. BADER*

Department of Biomedical Engineering
and High-Throughput Biology Center,
Johns Hopkins University,
Baltimore, MD 21218, USA
**E-mail: joel.bader@jhu.edu*

Transcription factors are DNA-binding proteins that control gene transcription by binding specific short DNA sequences. Experiments that identify transcription factor binding sites are often laborious and expensive, and the binding sites of many transcription factors remain unknown. We present a computational scheme to predict the binding sites directly from transcription factor sequence using all-atom molecular simulations. This method is a computational counterpart to recent high-throughput experimental technologies that identify transcription factor binding sites (ChIP-chip and protein-dsDNA binding microarrays). The only requirement of our method is an accurate 3D structural model of a transcription factor–DNA complex. We apply free energy calculations by thermodynamic integration to compute the change in binding energy of the complex due to a single base pair mutation. By calculating the binding free energy differences for all possible single mutations, we construct a position weight matrix for the predicted binding sites that can be directly compared with experimental data. As water-bridged hydrogen bonds between the transcription factor and DNA often contribute to the binding specificity, we include explicit solvent in our simulations. We present successful predictions for the yeast MAT-α2 homeodomain and GCN4 bZIP proteins. Water-bridged hydrogen bonds are found to be more prevalent than direct protein-DNA hydrogen bonds at the binding interfaces, indicating why empirical potentials with implicit water may be less successful in predicting binding. Our methodology can be applied to a variety of DNA-binding proteins.

Keywords: transcription factor binding sites; free energy; position weight matrix; hydrogen bond

1. Introduction

Transcription factors (TFs) are proteins that exert control over gene expression by recognizing and binding short DNA sequences ($\lesssim$6 base pairs,

Pacific Symposium on Biocomputing 12:484-495(2007)

roughly the width of the major groove).[1–5] The experimental methods used to identify these binding sites,[6,7] including SELEX[8] and the recent high-throughput experiments (ChIP-chip[9] and protein-dsDNA binding microarrays[10]), are often labor-intensive and expensive. Due to the complex molecular recognition mechanism between protein and DNA, there is no simple one-to-one code for protein-DNA recognition,[11,12] which makes theoretical predictions of TF binding sites challenging. As a result, the binding sites for many TFs are still unknown.

There is consequently great interest in methods with the potential to determine binding preferences purely from sequence and 3D structure of a protein-DNA complex, with several methods using energy-motivated scoring functions to compute the possible TF binding sites.[13–17] Position weight matrices are typically generated using the energy differences among different DNA sequences under an additive approximation.[17–19] Good results have been obtained for some families of TFs, such as zinc finger proteins.[14] Several limitations remain, however. First, proteins that require water-bridged contacts with the DNA are poorly modeled by empirical, implicit solvent energy functions. Second, minimized energies often include only enthalpic and no entropic effects. Third, protein and DNA backbones are fixed to favor the conformation of the native DNA sequence, leading to a bias in the computed position weight matrix.

In this work, we present a computational approach that overcomes the above limitations. Our approach uses molecular dynamics simulation and thermodynamic integration[20,21] to calculate binding free energy differences. The only requirement of the method is a starting 3D structural model of the protein-DNA complex, which can be obtained from X-ray/NMR determination or homology modeling. Our method is complementary to experimental methods such as protein binding microarrays and ChIP-chip.

Our approach studies the actual binding free energy of TF-DNA complexes and includes entropic effects by exploring the entire energy surface. Therefore, not only can we produce position weight matrices for binding site representation, but our binding free energy differences can also be directly compared with experimental measurements. Another advantage of our work is that we include explicit solvent molecules (counterions and water) in our simulation, whose importance has been reviewed.[22–24] Our work investigates the role of water in the TF-DNA recognition and binding specificity and accounts for the dynamics of water-bridged contacts. Furthermore, the intrinsic flexibility of protein and DNA backbones is explored in our simulation, which allows weak binders to be discovered. Finally, our

Pacific Symposium on Biocomputing 12:484-495(2007)

method can be modified to estimate non-additivity among DNA base pairs.

2. Model systems

We select homeodomain and bZIP proteins as our model systems in this study. These families are abundant in eukaryotic genomes. Except for a few members of these families, binding sites have not been well-characterized.

The 3D structures of homeodomain and bZIP proteins and their DNA-binding interfaces are highly conserved, making high quality homology modeling possible. Homeodomains contain ∼60 amino acid residues that form three α-helices with a hydrophobic core in the middle. The third helix is often referred to as the "recognition helix" as it binds the DNA in the major groove and forms most of the base-specific contacts. The first 5 residues at the N-terminus of the protein bind the DNA in the minor groove and also form a few base-specific contacts. A typical basic region leucine zipper protein (bZIP) is ∼60 amino acids long and forms a nearly-straight α-helix when bound to DNA.[25] The bZIP domain is composed of two relatively independent regions: the "leucine zipper region" is a dimerization region that helps stabilize the protein secondary structure, and the "basic region" contacts the DNA major groove and determines the DNA-binding specificity.

The yeast mating-type protein α2 (homeodomain) and the yeast general control protein GCN4 (bZIP) are studied in this work due to the availability of their experimental structures and binding sites for comparison and verification. Although the interactions among different monomers of TFs are important in exerting combinatorial controls over gene expression, only the monomers are considered in this work. The crystal structure of MAT-α2 (PDB:1APL) contains two identical binding sites for two isolated monomers of MAT-α2. One site is chosen in our modeling. The bZIP proteins normally bind to the DNA as homodimers or heterodimers. Since the binding interface between the basic region and the DNA is highly conserved,[25–27] we select only the half-site of GCN4 in our study and do not model the dimerization region.

Homeodomain and bZIP monomers typically contact 4-6 DNA base pairs. We include a 10-base pair DNA duplex in our simulation. The DNA sequences are the same as in the corresponding crystal structures of MAT-α2 protein and GCN4 protein. The consensus binding site sequence for MAT-α2 protein is TTACA.[28] The consensus binding site sequence for GCN4 protein is aTGA[C|G] for its monomer.[25,29] The lowercase "a" represents weak selection at that position and the last position can be a C or

Pacific Symposium on Biocomputing 12:484-495(2007)

G.

3. Methods

3.1. *Molecular dynamics simulation and free energy calculation*

Figure 1 illustrates the theoretical foundation of this work. The binding

$$\begin{array}{ccc} \text{DNA (aq)} + \text{protein (aq)} & \rightarrow & \text{DNA-protein (aq)} \quad \Delta G \\ \downarrow \Delta G_{\text{DNA}} & & \downarrow \Delta G_{\text{comp}} \\ \text{DNA' (aq)} + \text{protein (aq)} & \rightarrow & \text{DNA'-protein (aq)} \quad \Delta G' \end{array}$$

$$\Delta\Delta G = \Delta G' - \Delta G = \Delta G_{\text{comp}} - \Delta G_{\text{DNA}}$$

Fig. 1. Thermodynamic cycle used in the relative binding free energy calculation.

free energies of a protein with two different DNA sequences can be measured experimentally. The first horizontal reaction contains the native DNA and TF-DNA complex, whereas the second horizontal reaction contains the mutant DNA and its complex. In computations, it is relatively easy to calculate the free energy change caused by a mutation in the DNA sequence, indicated by the vertical reactions in the figure. The difference in binding free energy in the two experimental measurements, $\Delta G' - \Delta G$, is identical to the computational free energy difference, $\Delta G_{\text{comp}} - \Delta G_{\text{DNA}}$. This difference, $\Delta\Delta G$, will be referred to as the relative binding free energy in this paper. More detailed theoretical background can be found in Refs.[20,21]

The molecular simulation package CHARMM[30] was used to carry out the molecular dynamics simulation, and its BLOCK module was used for free energy calculations. We first established well-equilibrated native protein-DNA complex and DNA-duplex configurations using molecular dynamics simulation. Missing hydrogen atoms were added to the crystal structures of MAT-α2 (PDB:1APL) and GCN4 (PDB:1YSA). Charges of the titratable amino acid residues were assigned to their values at neutral pH. TIP3P water molecules were added and periodic boundary conditions were applied. Counterions (Na^+ ion) were introduced to neutralize the system using the random water-replacement routine developed by Rick Venable.[31] The CHARMM27 force field was used. The positions of the ions and water molecules were minimized followed by full minimizations of the entire system using the adopted basis Newton-Raphson method. The non-bonded cutoff radius was 14 Å. The system was then heated to 300 K and equi-

Pacific Symposium on Biocomputing 12:484-495(2007)

librated for 1.5 ns in the NPT ensemble using a 1 fs time step. The final configurations contained about 7000 water molecules and 25000 atoms for both MAT-α2 and GCN4 protein-DNA complexes. The protein-DNA complex and the DNA duplex were simulated separately.

From the equilibrated native configurations, we used a house-built program to replace each native base pair by multi-copy base pairs.[32,33] In this multi-copy approach, multiple base pairs are superimposed and their contributions to the total energy or force function are scaled by coupling parameters. In this paper, all multi-copy base pairs are a superposition of two physical base pairs. Therefore, there are 6 possible multi-copy base pairs at one position. The standard base geometry[34] was used to build a library of multi-copy base pair equilibrium geometries. Three consecutive rotations were applied to align the multi-copy base with the native base to preserve the orientation with repect to the rest of the DNA duplex. The structure with the multi-copy base pair was minimized first to remove possible bad contacts caused by the introduction of the multi-copy base. It was then heated to 350 K and equilibrated for 15 ps. This heating step helps move the conformation away from the native structure's local minima and may improve sampling of the glassy waters at the protein-DNA interface. The system was then cooled to 300 K and equilibrated for 65 ps. A 100 ps production run was done during which the trajectory was saved every 0.5 ps. The simulation is done in the NVT ensemble using the same periodic boundary condition as in the fully-equilibrated native structure. The free energy analysis on the production trajectory is outlined below.

Thermodynamic integration[20,21] was used to calculate the free energy change for mutating the original base pair into another possible base pair in the multi-copy base pair. The linear coupling scheme in the coupling parameter λ was used in BLOCK for the energy function of the multi-copy structures, which allows analytical solution of the free energy gradient. Typically, multiple values of λ are required for the integration. From preliminary calculations, we have found that the free energy gradient was approximately linear with respect to λ for multi-copy base pairs. Therefore, we used a mid-point approximation ($\lambda = 0.5$) for computational saving.

The binding free energy difference decomposes into separate contributions from DNA, protein, and solvent (ions and water) using the same

notation as Fig. 1:

$$\begin{aligned}
\Delta\Delta G_{\rm total} &= \Delta G_{\rm comp} - \Delta G_{\rm DNA} = \Delta\Delta G_{\rm internal} + \Delta\Delta G_{\rm external} \quad (1)\\
\Delta G_{\rm comp} &= \Delta G^c_{\rm prot} + \Delta G^c_{\rm solvent} + \Delta G^c_{\rm DNA}\\
\Delta G_{\rm DNA} &= \Delta G^f_{\rm solvent} + \Delta G^f_{\rm DNA}\\
\Delta\Delta G_{\rm internal} &= \Delta G^c_{\rm DNA} - \Delta G^f_{\rm DNA}\\
\Delta\Delta G_{\rm external} &= \Delta G^c_{\rm prot} + \Delta G^c_{\rm solvent} - \Delta G^f_{\rm solvent},
\end{aligned}$$

where the superscripts ^{c} and ^{f} represent the protein-DNA complex and the free DNA duplex, respectively. For homeodomains, the contribution of the N-terminus to the binding free energy difference was also calculated using $\Delta\Delta G_{\rm Nterm} = \Delta G^c_{\rm Nterm} - 0$, where the zero represents the corresponding ΔG term in the DNA duplex.

The binding free energy differences in Eq. (1) are converted into Boltzmann factors and position weight matrices as in Ref.[15] using the additive approximation. These matrices are converted into sequence logos[35] using WebLogo.[36] For the TFs considered in this work (Sec. 2), the DNAs remain relatively undeformed upon TF binding, which may make the additive approximation accurate.[14]

3.2. *Hydrogen bond analysis*

The native protein-DNA complex and DNA-duplex trajectories were further analyzed to explore the role of water in the binding specificity. Charmm's Hbond module was used to analyze whether a hydrogen bond (H-bond) exists in a certain frame in the trajectory. A distance cutoff of 2.4 Å was used as the maximum H-bond length (between acceptor and donor hydrogen) with no angle cutoffs. Then a house-built program was used to calculate the lifetime histograms for all occurrences of H-bonds. A 2 ps resolution was used such that any breakage of the H-bond shorter than 2 ps is ignored.[37] The existence of a direct or a water-bridged H-bond between the protein and DNA at each base pair position was also calculated. H-bonds formed by the N-terminal residues of MAT-α2 were considered separately from the rest of the protein.

4. Results and Discussions

Using the methods outlined in Sec. 3, the predicted sequence logos for the free energy terms in Eq. (1) are shown in Fig. 2. Our prediction of MAT-α2 achieves excellent agreement for all 5 positions in the "TTACA" consensus

Pacific Symposium on Biocomputing 12:484-495(2007)

sequence. This agreement verifies that the mid-point approximation for thermodynamic integration (Sec. 3) is valid for this TF. The N-terminus is

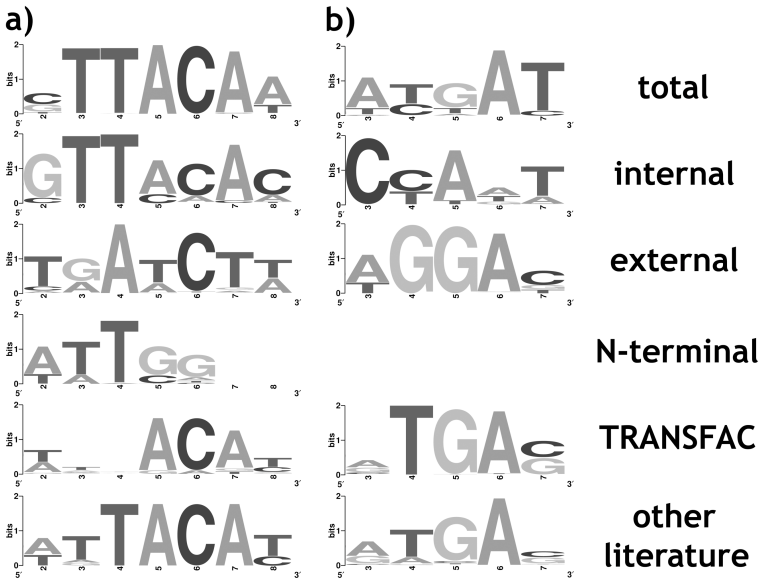

Fig. 2. Predicted sequence logos and experimental logos for yeast proteins MAT-α2 (homeodomain) and GCN4 (bZIP). The base pair positions that have base-specific contacts as either direct or water-bridged H-bonds between the protein and the DNA bases are shown. The total, internal, external, and N-terminal (for MAT-α2) logos are listed in that order. Logos generated from both TRANSFAC[38] and primary experimental publications are listed at the bottom. For MAT-α2, the TRANSFAC logo is for heterodimer MAT-**a**1/MAT-α2,[39] and the literature logo is for heterotetramer MAT-α2/MCM1.[28] For GCN4, the TRANSFAC logo is based on sequences obtained from 4 rounds of affinity column selection and PCR amplification;[40] the literature logo is based on sequences of 15 promoter regions of GCN4 targets from DNA site protection experiments.[41] These two logos were obtained by converting the experimental dimer binding sequences into 2 half-site monomer binding sequences to facilitate comparison with the computational predictions.

responsible for the first two positions in the "TTACA" consensus sequence. A reduced model that considers only the "recognition helix" may fail to identify these positions. The DNA internal energies contribute largely to all five positions. Our GCN4 prediction agrees with the experimental binding sites at 4 out of 5 positions of the aTGA[C|G] consensus sequence, whereas the last position is variable in the experiemental sites. The external free

energies are largely responsible for these positions. The information content of our prediction agrees well with the literature logo, which considered both strong and weak binding sequences.[41] The TRANSFAC logo shows higher information content, possibly because it was constructed from only the strongest binding sequences.[40]

The lifetime histograms of different types of H-bonds for MAT-α2 are shown in Fig. 3. The lifetimes for H-bonds between the DNA-duplex and

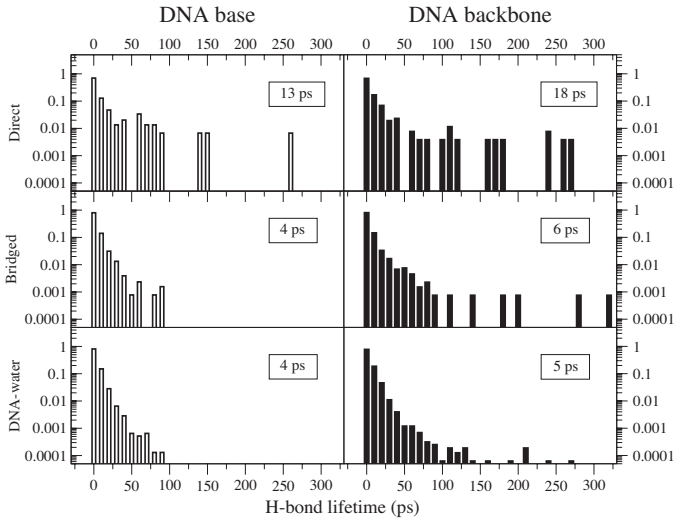

Fig. 3. Histograms of H-bond lifetimes for yeast MAT-α2 homeodomain protein during a 600 ps simulation. The top, middle, and bottom panels represent the direct protein-DNA H-bonds, the water-bridged protein-DNA H-bonds, and the H-bonds between DNA and water, respectively. The left and right panels represent the H-bonds formed by the DNA bases and the DNA backbone, respectively. The insets of the panels show the average lifetimes.

water are similar to a previous simulation study,[37] although the average lifetime is slightly shorter. The histograms for GCN4 (not shown) are similar except for slightly longer average lifetimes for the direct and water-bridged H-bonds.

Since the binding specificity of a TF arises primarily from contacts made with the DNA bases, we now examine the left panels of Fig. 3 further. There are 3 long-lived (> 100 ps) direct protein-DNA H-bonds for MAT-α2 during a 600 ps equilibration. Two of them are between the recognition helix and the major groove bases, which are also found in the crystal structure.[28] One

Pacific Symposium on Biocomputing 12:484-495(2007)

H-bond is between the N-terminal tyrosine and adenine base in the minor groove, which is not present in the crystal structure since the tyrosine side chain was not resolved. For GCN4, all long-lived direct H-bonds are also observed in the crystal structure.[25]

Both the MAT-α2 and GCN4 binding interfaces are highly hydrated in the simulations, with the MAT-α2 interface more hydrated than GCN4 (data not shown). Figure 4 shows the H-bond existence time-series for the native MAT-α2–DNA complex. Base pair positions 1, 9, and 10 have rare occurrences of H-bonds and thus are not shown. Figures 3 and 4 demonstrate that the water-bridged H-bonds are highly dynamic, with H-bonds breaking and forming constantly. This is important because it indicates that the 100 ps production runs for the multi-copy structures provide adequate sampling of the bridging water. Figure 4 shows that bridged H-bonds form a large and extensive contact network at the protein-DNA binding interface that is more prevalent than the direct protein-DNA H-bond network.[42] As a result, the binding specificity arises exclusively from water-bridged H-bonds at base pair positions 3, 7, and 8 for MAT-α2 and at base pair positions 3 and 6 for GCN4 (data not shown), respectively. These results indicate that water-bridged H-bonds contribute more to the binding affinity and specificity than direct H-bonds in these TF-DNA complexes.

5. Conclusion

We present here an all-atom molecular simulation and free energy calculation method that calculates the TF binding sites based on a 3D structural model of the protein-DNA complex. Explicit water molecules are included and are found to form a dynamic and more prevalent H-bond network than direct protein-DNA H-bonds. The predicted position weight matrices of MAT-α2 and the half-site of GCN4 agree well with the experimental binding sites.

We are currently carrying out the following studies that will help establish the scope and limitations of our method. First, we are analyzing the hydration dynamics at the binding interface for multi-copy trajectories. These results serve to evaluate the efficiency in the configrational sampling of our simulation protocol. Second, we are implementing multi-copy base pairs using the most recent AMBER force field[43,44] to investigate sensitivity of our method to the force field parameters. Preliminary studies of homeodomain and bZIP proteins have shown that high quality homology modeling is possible (RMSD smaller than or about 1 Å can be obtained for many family members). We are currently investigating the effects of starting 3D

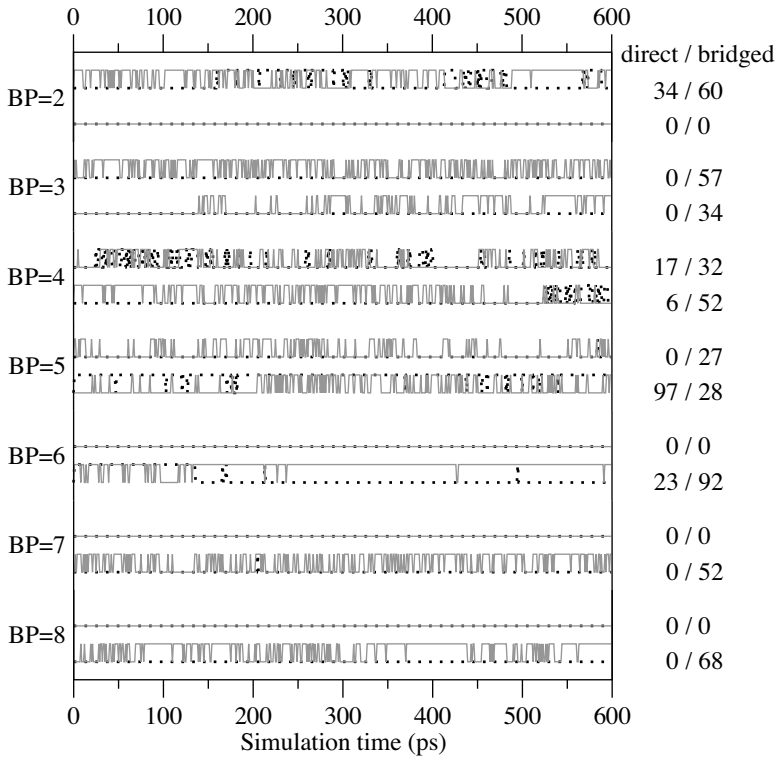

Fig. 4. Time-series for direct (black dotted lines) and water-bridged (gray solid lines) H-bonds between the yeast MAT-α2 homeodomain protein and its native DNA base atoms during a 600 ps simulation. Base pair positions 2 to 8 are plotted. The two possible states being plotted are (i) having at least one direct or water-bridged H-bond (the spikes or steps), and (ii) no direct or bridged H-bond (the base lines). Cooccurring H-bonds at each base pair position are plotted as one H-bond. There are two panels for each base pair position. The upper and lower panels contain the time-series for H-bonds formed by the N-terminus and the rest of the protein with the DNA bases, respectively. The percentages of time that H-bonds exist are listed on the right hand side of the corresponding data series. For example, at base pair 2, 34% of the time during the equilibration, there is at least one direct H-bond between the N-terminus and the DNA base atoms, whereas 60% of the time there is at least one bridged H-bond between the N-terminus and the DNA base atoms.

structural models on the TF binding site predictions. Finally, sensitivity of the results to the starting DNA sequence is also being considered.

Our method is computationally intensive. The prediction of binding sites for one TF requires $\sim$400 CPU-days on a 3.0 GHz Intel processor, which is about \$1500 considering a 3-year lifespan for a CPU. We are currently

developing and testing multiple-multi-copy methods in which two or more base pairs are both multi-copies. These calculations can improve the computational efficiency of our method. Furthermore, the free energy analysis of such structures helps quantify the correlation among the base pairs and provides an estimation of error for the additive approximation.

Acknowledgements

LAL acknowledges funding from the Department of Energy (DE-FG0204ER25626). JSB acknowledges funding from NSF CAREER 0546446, NIH/NCRR U54RR020839, and the Whitaker foundation. We acknowledge a starter grant and an MRAC grant of computer time from the Pittsburgh Supercomputer Center, MCB060010P, MCB060033P, and MCB060056N.

References

1. C. O. Pabo and R. T. Sauer, *Annu Rev Biochem* **53**, 293 (1984).
2. C. O. Pabo and R. T. Sauer, *Annu Rev Biochem* **61**, 1053 (1992).
3. G. Patikoglou and S. K. Burley, *Annu Rev Biophys Biomol Struct* **26**, 289 (1997).
4. N. M. Luscombe, S. E. Austin, H. M. Berman and J. M. Thornton, *Genome Biol* **1**, p. REVIEWS001 (2000).
5. M. D. Biggin, *Nat Genet* **28**, 303 (2001).
6. M. L. Bulyk, *Genome Biol* **5**, p. 201 (2003).
7. G. D. Stormo and D. S. Fields, *Trends Biochem Sci* **23**, 109 (1998).
8. C. Tuerk and L. Gold, *Science* **249**, 505 (1990).
9. B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell and R. A. Young, *Science* **290**, 2306 (2000).
10. S. Mukherjee, M. F. Berger, G. Jona, X. S. Wang, D. Muzzey, M. Snyder, R. A. Young and M. L. Bulyk, *Nat Genet* **36**, 1331 (2004).
11. P. V. Benos, A. S. Lapedes and G. D. Stormo, *Bioessays* **24**, 466 (2002).
12. A. Sarai and H. Kono, *Annu Rev Biophys Biomol Struct* **34**, 379 (2005).
13. J. Ashworth, J. J. Havranek, C. M. Duarte, D. Sussman, J. Monnat, R. J., B. L. Stoddard and D. Baker, *Nature* **441**, 656 (2006).
14. G. Paillard and R. Lavery, *Structure (Camb)* **12**, 113 (2004).
15. A. V. Morozov, J. J. Havranek, D. Baker and E. D. Siggia, *Nucleic Acids Res* **33**, 5781 (2005).
16. R. G. Endres, T. C. Schulthess and N. S. Wingreen, *Proteins* **57**, 262 (2004).
17. R. A. O'Flanagan, G. Paillard, R. Lavery and A. M. Sengupta, *Bioinformatics* **21**, 2254 (2005).
18. M. L. Bulyk, P. L. Johnson and G. M. Church, *Nucleic Acids Res* **30**, 1255 (2002).
19. P. V. Benos, M. L. Bulyk and G. D. Stormo, *Nucleic Acids Res* **30**, 4442 (2002).

20. P. M. King, *Free energy via molecular simulation: a primer*, in *Computational Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, eds. W. van Gunsteren, K. Weiner, P and A. Wilkinson (ESCOM, Leiden, 1993), p. 267.
21. L. Andrew, *Molecular Modelling: Principles and Applications*, 2nd edn. (Prentice Hall, 2001).
22. T. E. Cheatham 3rd and P. A. Kollman, *Annu Rev Phys Chem* **51**, 435 (2000).
23. J. W. Schwabe, *Curr Opin Struct Biol* **7**, 126 (1997).
24. C. Wolberger, *Curr Opin Struct Biol* **6**, 62 (1996).
25. T. E. Ellenberger, C. J. Brandl, K. Struhl and S. C. Harrison, *Cell* **71**, 1223 (1992).
26. C. R. Vinson, P. B. Sigler and S. L. McKnight, *Science* **246**, 911 (1989).
27. T. W. Siggers, A. Silkov and B. Honig, *J. Mol. Biol.* **345**, 1027 (2005).
28. C. Wolberger, A. K. Vershon, B. Liu, A. D. Johnson and C. O. Pabo, *Cell* **67**, 517 (1991).
29. W. Keller, P. Konig and T. J. Richmond, *J Mol Biol* **254**, 657 (1995).
30. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.* **4**, 187 (1983).
31. R. Venable, Ion addition by selective water replacement `http://www.charmm.org/ubbthreads/ubbthreads.php?Cat=0`, CHARMM Community Script Archive.
32. I. Lafontaine and R. Lavery, *Biopolymers* **56**, 292 (2000).
33. I. Lafontaine and R. Lavery, *Biophys J* **79**, 680 (2000).
34. W. K. Olson, M. Bansal, S. K. Burley, R. E. Dickerson, M. Gerstein, S. C. Harvey, U. Heinemann, X. J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C. S. Tung, E. Westhof, C. Wolberger and H. M. Berman, *J Mol Biol* **313**, 229 (2001).
35. T. D. Schneider and R. M. Stephens, *Nucleic Acids Res* **18**, 6097 (1990).
36. G. E. Crooks, G. Hon, J. M. Chandonia and S. E. Brenner, *Genome Res* **14**, 1188 (2004).
37. A. M. Bonvin, M. Sunnerhagen, G. Otting and W. F. van Gunsteren, *J Mol Biol* **282**, 859 (1998).
38. E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhauser, M. Pruss, F. Schacherer, S. Thiele and S. Urbach, *Nucleic Acids Res* **29**, 281 (2001).
39. C. Goutte and A. D. Johnson, *Embo J* **13**, 1434 (1994).
40. A. R. Oliphant, C. J. Brandl and K. Struhl, *Mol. Cell. Biol.* **9**, 2944 (1989).
41. K. Arndt and G. R. Fink, *Proc Natl Acad Sci U S A* **83**, 8516 (1986).
42. M. Billeter, P. Guntert, P. Luginbuhl and K. Wuthrich, *Cell* **85**, 1057 (1996).
43. D. A. Case, T. E. Cheatham 3rd, T. Darden, H. Gohlke, R. Luo, J. Merz, K. M., A. Onufriev, C. Simmerling, B. Wang and R. J. Woods, *J Comput Chem* **26**, 1668 (2005).
44. T. E. Cheatham 3rd and M. A. Young, *Biopolymers* **56**, 232 (2000).