Published in final edited form as:

Physica D. 2006 December ; 224(1-2): 174–181.

Decoding transcriptional regulatory interactions

L. Angela Liu and Joel S. Bader

Department of Biomedical Engineering and High-Throughput Biology Center, Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21215 USA

Abstract

Transcription factor proteins control the temporal and spatial expression of genes by binding specific regulatory elements, or motifs, in DNA. Mapping a transcription factor to its motif is an important step towards defining the structure of transcriptional regulatory networks and understanding their dynamics. The information to map a transcription factor to its DNA binding specificity is in principle contained in the protein sequence. Nevertheless, methods that map directly from protein sequence to target DNA sequence have been lacking, and generation of regulatory maps has required experimental data. Here we describe a purely computational method for predicting transcription factor binding. The method calculates the free energy of binding between a transcription factor and possible target DNA sequences using thermodynamic integration. Approximations of additivity (each DNA basepair contributes independently to the binding energy) and linear response (the DNA-protein and DNA-solvent couplings are linear in an effective reaction coordinate representing the basepair character at a specific position) make the computations feasible and can be verified by more detailed simulations. Results obtained for MAT- α 2, a yeast homeodomain transcription factor, are in good agreement with known results. This method promises to provide a general, computationally feasible route from a genome sequence to a gene regulatory network.

Keywords

computational biology; bioinformatics; gene regulation; transcription factor; homeodomain

1 Introduction

Given a genome sequence, it is increasingly easy to identify gene and protein components that compose biological networks and pathways. Predicting the network edges — the physical interactions between proteins and the specific binding between transcription factors and DNA regulatory elements — is a crucial step towards predicting and designing the properties of living systems. Transcription factor proteins regulate genes by binding to specific DNA sequences, termed regulatory elements, in the upstream promoter regions of target genes (Fig. 1). Multiple transcription factors can regulate a single gene, with complicated interactions yielding combinatorial control of expression. Predicting the binding preferences of individual transcription factors is a first step towards understanding the full details of transcriptional regulation.

The importance of transcriptional regulation is reflected by the large fraction of genes that perform this function in multicellular organisms. Over 10% of human genes encode transcription factors that recognize specific DNA sequences. Most transcription factors belong to one of a handful of major families that are represented in other eukaryotic lineages, including yeast, nematode, and insect (Fig. 2). Some transcription factor families appear in only a subset

of lineages, for example the AP2 family which occurs in plants but not in animals or fungi, and the nuclear hormone receptor family which arose after the yeast –animal split. Other families, such as homeodomain proteins and leucine zipper transcription factors, are present in all major eukaryotic kingdoms.

While several DNA-binding motifs can recognize specific base sequences [1,2], the majority of characterized transcription factors contain an α -helix that makes base-specific contacts with the DNA major groove. Homeodomain, MYB, leucine zipper domain (bZIP), helix-loop-helix (HLH), winged-helix, zinc finger, nuclear hormone receptor, and helix-turn-helix (HTH, specific to prokaryotes) families all have α -helix motifs. The large classic zinc finger family encodes three sequential α -helices held by a scaffold that makes contact with three DNA major groove regions. Homeodomain transcription factors are three-helix bundles in which one helix binds to the DNA major groove and the other two are solvent-exposed. The specificity of a homeodomain protein arises primarily from contacts of a single α -helix with about four to six contiguous basepairs. The other two helices can make additional contacts with the DNA backbone but do not confer specificity. Leucine zippers are encoded by two separate α -helices that form homodimers or heterodimers when bound to DNA. Each chain recognizes a half site, and the half sites can compose a single continuous recognition region. The nuclear hormone receptor family contains two α -helices on a single scaffold. The two helices recognize DNA sequences with variable spacing. These differences in how the α -helices are presented to DNA has made it difficult to infer general recognition or complementarity rules between specific protein residues and cognate DNA basepairs.

Recognition of DNA by β -sheets is less common, but is observed in families such as AP2 in plant and ApiAP2 in malaria [3]. Some malarial proteins have multiple ApiAP2 domains, suggesting a β -sheet analog to the α -helix nuclear hormone receptor motif. TATA-box binding proteins (TBP) also use β -sheets in DNA binding. However, TBP binds DNA in the minor groove and causes a large bend in the DNA structure, which is thought to facilitate the opening of the DNA duplex for transcription.

Intense interest in understanding transcriptional regulation has spurred development of experimental methods to characterize transcription factor binding sites. Screening by SELEX is possible but requires purified protein and costly rounds of enrichment. SELEX has been eclipsed by two recent methods that are beginning to yield high-throughput data: chromatin immunoprecipitation followed by microarray readout (ChIP/chip) [4–6] and protein binding to double-stranded DNA microarrays [7].

In the ChIP/chip approach, a cell or organism is engineered to express a tagged version of a transcription factor. If the tagged protein is expressed and binds DNA, it can be isolated to enrich the bound DNA. The DNA is then hybridized to identify the genomic region where the transcription factor binds. ChIP/chip methods require access to a cellular state that produces a transcription factor of interest. This can be difficult if the cellular state is unknown or difficult to access, such as a precise developmental state. Certain organisms, for example pathogens, are impractical to culture. The rapid evolution of regulatory networks makes it difficult to infer binding specificity cross-species even for orthologous proteins. Finally, it may not be clear whether the tagged protein, or some other component of a transcription factor complex, provides the observed specificity.

Binding of protein to a dsDNA array requires only purified protein, which can be generated by a protein expression system. These experiments can be expensive, however, due to the need to generate an organism-specific hybridization chip. This is because the dsDNA chips do not contain an array of motifs, but rather an array of the intergenic regions for a particular genome.

When promoter regions are difficult to delineate, as in human, these arrays are even more challenging to fabricate and expensive to use.

The ChIP/chip and dsDNA methods provide only indirect evidence of binding preference by identifying DNA regions that are apparently enriched for transcription factor binding. Transcription factors have general affinity for the DNA backbone, and the signal from the specifically-bound DNA regions is accompanied by a background from non-specifically bound DNA. Identifying the specifically-bound genomic regions requires methods analogous to analysis of gene expression data to identify differentially expressed transcripts.

Once the enriched regions are identified, bioinformatics methods are required to infer the actual short motifs bound by the transcription factor. These methods have their origin in motif-finding methods, such as *MEME* and *Biopro prospector*, that were developed to identify regulatory elements in co-expressed genes [8–13]. The statistical search does not always converge to the correct binding site. A recent study failed to identify known specificities for 138 of 203 yeast proteins tested [6].

As an alternative to experimental characterization, recognition codes are hypothetical specific pairing rules between amino acid side-chains and the DNA basepairs they specifically recognize. The theoretical possibility for recognition codes arises from different hydrogen-bonding opportunities in the major and minor DNA groove depending on the basepair. Unfortunately, it has been impossible to identify general rules that span different transcription factor families. The best performance has been obtained for the classic zinc finger family [14]. Improved recognition codes parameterized using expectation maximization have now been used to predict binding for *Drosophila* zinc fingers [15]. There has been little or no progress with other transcription factor families, other than observation of core binding regions for certain sub-families. Even weak knowledge is useful because binding motifs identified from computational methods can combine with experimental data to provide more precise predictions of transcriptional regulation. This has already been achieved with existing binding motif databases combined with ChIP/chip data [16].

The lack of success of recognition codes has prompted efforts towards structure-based predictions of protein-DNA interactions. Published studies suggest that molecular force-fields are sufficiently accurate to model interactions between DNA, protein, and solvent molecules [17–20]. Water molecules are crucial for many transcription factor families where water-mediated contacts are responsible for specificity recognition [21,22]. Nonetheless, studies aimed at predicting binding sites have generally used implicit solvent and fixed DNA backbones for expediency. These failings present a general problem because water-bridged hydrogen bonds and distortion of the DNA backbone are common features of transcription factor binding.

Baker and coworkers adapted methods used for protein-protein interaction prediction to study zinc finger binding preferences [23] based on data from experimental screens of variants of a murine zinc finger, Zif268 [24]. A brute force search through possible DNA binding regions used fixed backbones, a rotamer side-chain library, and implicit solvent. The authors concluded that this method was questionable for predicting binding specificity, but suitable for the easier goal of designing a protein to recognize a pre-specified DNA sequence. Including hydrated rotamers in the protein sidechain library has been suggested [25] but requires additional parameterization and has yet to be tested. One of their recent work [26] extended the study to other transcription factors with limited success. Again their work does not include the effects of explicit water molecules, and their binding energies excluded any entropic contributions.

Lavery and coworkers studied nucleic acids [27] and their interactions with proteins [28–30] with a multi-copy approach that superimposes all four possible basepairs at each DNA position,

again with implicit solvent. After generating structures with the fictitious basepairs, binding sites are ranked by a scoring function. This method yielded accurate predictions for 18 proteins tested, but primarily for complexes that did not contain water-mediated protein-DNA interactions. Furthermore, this analysis method does not properly account for the fictitious nature of the multi-copy base, which induces protein conformations that might not occur naturally.

A simplifying assumption in analysis of protein-DNA interactions is that the energetic contributions of basepairs are additive. This assumption yields the standard position-weight-matrix representation of a binding site. Recent work continues to point to the accuracy of the additive approximation [26], particularly for sequences similar to the most favored sequence [31]. The main cause for non-additive contributions may be DNA backbone deformations [29], pointing again to the important of incorporating DNA flexibility in simulations.

The methods introduced here are motivated by Lavery's multi-copy model for DNA. As described in the methods (Sec. 2), a multi-copy basepair is used to represent a transition state between two physical basepairs. Statistics collected during an equilibrium simulations of a DNA-protein complex with a multi-copy basepair and the same DNA sequence solvated by water are used to calculate a difference in binding free energy for two physical basepairs. Repeating this calculation for each position along a DNA sequence yields predictions for the binding specificity of the transcription factor. Results of applying these methods are presented for MAT- $\alpha 2$, a yeast homeodomain transcription factor (Sec. 3). The predicted binding motif is in good agreement with the motif described in the original literature. Finally, future plans and extensions are discussed (Sec. 4).

2 Methods

2.1 Statistical mechanics and bioinformatics formulations

The specificity of a transcription factor (TF) may be represented as

$$\Pr(X \mid \text{bound}) / \Pr(Y \mid \text{bound}) = \exp[-\beta \Delta G(X)] / \exp[-\beta \Delta G(Y)], \quad (1)$$

where X and Y are two DNA sequences, and β^{-1} is the thermal energy $k_B T$. Sequences X and Y are assumed to be equally abundant in a genome for simplicity; this assumption could easily be lifted by including an additional factor of their relative abundance. The binding free energy for a DNA sequence is the difference in free energies of the solvated TF-DNA complex and the solvated isolated components,

$$\Delta G(X) = G(\text{TF} - X) - G(\text{TF}) - G(X), \quad (2)$$

The solvation energy for the TF-DNA complex is $G(\text{TF} - X)$, and the solvation energies of the isolated components are $G(\text{TF})$ and $G(X)$. An analogous expression gives $\Delta G(Y)$. Combining Eqs.1 and 2 indicates that the $G(\text{TF})$ terms cancel, and

$$\begin{aligned} \Delta \Delta G(X \leftarrow Y) &= -k_B T \log \Pr(X \mid \text{bound}) / \Pr(Y \mid \text{bound}) \\ &= [G(\text{TF} - X) - G(\text{TF} - Y)] - [G(X) - G(Y)] \\ &= \Delta G(X \leftarrow Y, \text{ TF complex}) - \Delta G(X \leftarrow Y, \text{ water}). \end{aligned} \quad (3)$$

Under an additive approximation, the energetic contributions are additive over DNA positions,

$$\Delta \Delta G(X \leftarrow Y) = \sum_{i=1}^W \Delta \Delta G(X_i \leftarrow Y_i), \quad (4)$$

where the index i runs over the W basepairs in the TF-DNA interface. Under this approximation, the probability that position i of bound sequence X , denoted x_i , is basepair α is

$$\Pr(x_i = \alpha) = \exp[-\beta \Delta \Delta G(x_i = \alpha \leftarrow x_i = \delta)] / \sum_{\gamma=A,C,G,T} \exp[-\beta \Delta \Delta G(x_i = \gamma \leftarrow x_i = \delta)], \quad (5)$$

where δ is an arbitrary reference state. Since the identity of a basepair is defined by the nucleotide on one of the strands, the basepairs AT, CG, GC, and TA are abbreviated as the nucleotides A, C, G, and T. Although there are $C(4, 2) = 6$ distinct pairwise combinations possible, the energy spacing between basepairs is fully specified by only 3 values, for example the $\Delta \Delta G$ values using the most favored basepair as the nominal zero of energy. Therefore only 3 pairwise comparisons are required to establish the position weight matrix at any position.

The additive approximation permits a direct comparison with the position-weight-matrix (PWM) of a binding site. This representation is obtained by aligning known binding sites for a transcription factor and tabulating $n(x_i = \alpha)$, the number of times that basepair α occurs at position i . The PWM is then defined as

$$\Pr(x_i = \alpha) = n(x_i = \alpha) / \sum_{\gamma=A,C,G,T} n(x_i = \gamma) \quad (6)$$

PWMs are conveniently visualized as logos [32,33]. The information at position i is summarized by a stack of letters signifying each nucleotide. The height of the stack is the information content (IC) at the position,

$$\text{IC} = 2 + \sum_{\alpha=A,C,G,T} \Pr(x_i = \alpha) \log_2 \Pr(x_i = \alpha), \quad (7)$$

and the height of each letter is proportional to $\Pr(x_i = \alpha)$ with the letters sorted from top to bottom in decreasing order of probability.

2.2 Computational strategy

Free energy differences $\Delta G(X \leftarrow Y)$ for DNA sequences X and Y in either the TF complex or in water can be calculated using thermodynamic integration [34],

$$\Delta G(X \leftarrow Y) = \int_0^1 dt \langle dH[\lambda(t)] / dt \rangle_{\lambda(t)}, \quad (8)$$

where t parameterizes a path $\lambda(t)$ that switches from state Y to state X . Our calculations use a Hamiltonian $H[\lambda]$ where the switching involves a superposition of basepairs at a single position,

$$H[\lambda] = T_0 + \sum_{\alpha=A,C,G,T} T_\alpha + U_0 + \sum_{\alpha=A,C,G,T} \lambda_\alpha U_\alpha, \quad (9)$$

with T_0 and U_0 representing the kinetic and potential energy of the entire system except for the position that is being switched, and T_α and U_α representing the kinetic and potential energy of the switching position. The components λ_α represent the fractional representation of basepair α . For example, $\lambda_A = 1$ with $\lambda_C = \lambda_G = \lambda_T = 0$ represents a physical state with basepair AT, and $\lambda_A = \lambda_T = 0.5, \lambda_C = \lambda_G = 0$ represents a 50-50 mixture of AT and TA with no CG or GC character. The kinetic energy in Eq. 9 need not be weighted by λ because classical configurational properties should be independent of these terms. For a linear switch between two states α and γ , with $\lambda_\alpha(t) = t$ and $\lambda_\gamma(t) = 1 - t$, the free energy change is

$$\Delta G(a \leftarrow \gamma) = \int_0^1 dt \langle H_a - H_\gamma \rangle_{\lambda(t)}. \quad (10)$$

In principle, calculation of ΔG for this change requires a series of equilibrium calculations at multiple values of t , also known as multi-configurational thermodynamic integration. These calculations become increasingly difficult as t approaches 0 and 1, requiring much longer simulations in order to achieve comparable standard deviations as at other t values. Suppose, however, that the interaction between the switching position and the rest of the system can be modeled as an effective linear coupling to a bath. In this case, the Hamiltonian for state α is

$$H_\alpha = T + (1/2)(y - y_\alpha)' \cdot k \cdot (y - y_\alpha) + E_\alpha, \quad (11)$$

where y , which may be a vector, represents bath modes, y_α is the expected value of y in state α , E_α is the energy at this minimum value, k is the coupling between bath modes, and T is the kinetic energy of the bath. If the coupling k is independent of α , then the free energy difference between states is

$$\Delta G(a \leftarrow \gamma) = E_\alpha - E_\gamma. \quad (12)$$

The quantity that enters into thermodynamic integration is

$$\begin{aligned} \langle H_a - H_\gamma \rangle_\lambda &= E_\alpha - E_\gamma + (1/2)y_\alpha' \cdot k \cdot y_\alpha - (1/2)y_\gamma' \cdot k \cdot y_\gamma \\ &\quad - \sum_\delta \lambda_\delta y_\delta' \cdot k \cdot (y_\alpha - y_\gamma). \end{aligned} \quad (13)$$

At the midpoint value $\lambda_\alpha = \lambda_\gamma = 0.5$, $\langle H_a - H_\gamma \rangle$ is exactly equal to $E_\alpha - E_\gamma$. Our strategy, then, is to use the midpoint calculation as an estimate for the full integral. This estimator has reduced variance with a tradeoff of possibly introducing bias. Provided that the bias is smaller than the energy differences to be measured, however, this approximation will be successful in predicting specificity. The bias will be small for an effective linear system. The bias may still be small for non-linearities if the non-linearities cancel systematically. Cancellations might be anticipated when the ΔG 's of the TF-DNA complex and of DNA in water are subtracted. From preliminary calculations, we found that this is indeed the case for a model TF studied here (results not shown), confirming that the midpoint simplification is a good approximation.

2.3 Computational implementation

Coordinates for a complex of MAT- $\alpha 2$ with a DNA 10-mer were obtained from the crystal structure (PDB entry: 1APL) [35] with the DNA sequence of ATT-TACACGC. This sequence differs from the consensus sequence in Transfac [36], ACATG, although Transfac indicates a weak preference for ACATG over ACACG. The tleap module of Amber was used to add hydrogen atoms to the macromolecule. The charge states of the titratable residues are as follows: ASP -1, GLU -1, LYS +1, ARG +1, HIS +1. Since the single histidine is exposed at the protein surface, the nitrogens at both δ and ϵ positions are protonated. The complex was solvated in a truncated octahedron with 4632 water molecules and 9 Na⁺ ions to maintain a neutral system. Periodic boundary condition and Ewald summation were applied and the system was equilibrated for 1 ns using Amber. The 399 force field of Amber was used. The resulting macromolecular geometry as well as the positions of the ions were imported into Charmm [37]. Water molecules were re-introduced in a cubic box. This system was then equilibrated for 700 ps at constant pressure (1 atm) and constant temperature (300 K) using the c27 force field of Charmm. Spherical cutoff of 14 Å was used to evaluate the non-bonded

interactions including the electrostatics and van der Waals forces and potentials. This final system contains 8320 water molecules and 9 Na⁺ ions (26366 total atoms). The same protocol was applied to the DNA duplex from the crystal structure. The final solvated DNA system contains 18 Na⁺ ions and 7117 water molecules in a cubic box (22003 atoms in total).

Calculations of $\Delta\Delta G$ values were organized as a single-elimination tournament at each position along the DNA 10-mer. The λ switching functions between pairs of basepairs are implemented using the Blocks procedure in Charmm. In an effort to minimize systematic effects due to favorable solvation of GC over AT, the first round of the tournament ran AT vs. TA and GC vs. CG. The two basepairs with lower energies were then compared in the second round free energy calculation. Our rationale for a single-elimination tournament was to permit the most favored basepair to serve as the reference, rather than a higher-energy state.

The tournaments were conducted by replacing a basepair with a multi-copy superposition. Since we use the midpoint approximation ($\lambda = 0.5$), two base-pairs are superimposed according to the tournament. The new DNA with the multi-copy basepair was energy minimized, heated to 350 K and equilibrated at 350 K for 30 ps, cooled to 300 K and equilibrated at 300 K for 120 ps, and then run for 100 ps production in the NVT ensemble. We checked that energy differences showed no drift over the 100 ps production run. This procedure was used for both the TF-DNA complex and the free DNA in water.

Frames from the production run were collected every 0.5 ps yielding $F = 200$ frames per run. The energy difference $\Delta H_f = H_\alpha - H_\gamma$ was calculated for each frame. The mean $\Delta \hat{H}$ and its standard error were calculated as

$$\Delta \hat{H} = F^{-1} \sum_{f=1}^F \Delta H_f \quad (14)$$

$$\text{StdErr}[\Delta H] = \left[\sum_{f=1}^F (\Delta H_f^2 - \Delta \hat{H}^2) / F(F-1) \right]^{1/2} \times [(1+c)/(1-c)]^{1/2}. \quad (15)$$

The factor $(1+c)/(1-c)$ corrects the standard error for correlation c between neighboring frames. Standard error propagation was used to obtain the standard error of $\Delta\Delta G$, i.e.

$$\text{StdErr}[\Delta\Delta G] = \{ \text{StdErr}[\Delta G(\text{DNA in complex})]^2 + \text{StdErr}[\Delta G(\text{DNA in water})]^2 \}^{1/2}. \quad (16)$$

After the first round of the TF-DNA complex and DNA-water simulations were completed at each position, the winning pair of basepairs were determined and competed against each other using the same protocol as in the first round. The $\Delta\Delta G$ values were converted to position-weight-matrices using Eq. 5 for visualization and comparison.

3 Results

Homeodomain proteins are important transcriptional regulators for developmental processes. This family was first identified in fruit fly as responsible for the proper development of body plan and segmentation. Homeodomain proteins regulate region-specific expression in flowers. Feedback loops involving homeodomain genes and proteins are an integral part of the cellular memory that maintains the expression patterns through development.

Specific binding to DNA sequences of homeodomains is achieved by contacts between an α -helix and the DNA major groove. Several structures of homeodomain-DNA complexes have been published by Wolberger and coworkers, including homeodomain mating type protein complexes in yeast [38–41,35] corresponding to PDB identifiers 1APL, 1YRN, 1AKH, 1LE8, and 1K61.

The MAT- α 2 complex was simulated starting from the 1APL coordinates as described above, using a single-elimination tournament to calculate the three energy differences that fully specify the position weight matrix entries at each position (Sec. 2). Simulation results and error bars are provided for the ΔG calculations for the TF-DNA complex and the DNA in water, together with the final $\Delta\Delta G$ for each comparison (Fig. 3).

The logo extracted from the computer simulation compares favorably with logos for the same protein from Transfac and from the primary literature [35], as shown in Fig. 4. The Transfac logo was based on a consensus DNA motif extracted from targets of a heterodimer between MAT- α 2 and MAT- α 1. The literature logo is from work by Wolberger and coworkers on targets of a heterotetramer with MAT- α 2 and MCM1 targeting a different set of genes and yielding a somewhat different logo.

Both the previous logos include sequence specificity at basepairs that lack specific contacts with MAT- α 2. Presumably, this specificity is due to contacts with the other transcription factor complexed with MAT- α 2. The basepairs that contact MAT- α 2 are numbered 3 through 8 inclusive. We restrict attention to predictions at these positions where the two experimental logos make strong predictions defined as 1 or more bits of information.

The Transfac logo has only 3 high information positions, 5 through 7. These are all predicted correctly by simulation. The probability for this to occur by chance is $(1/4)^3$, or a p-value of 0.016. The Wolberger logo has high information content for all 6 positions in contact. The simulation correctly predicts 5 of these. The p-value for correctly predicting 5 of 6 positions is approximately $C(6, 5)(1/4)^5(3/4)^1 + (1/4)^6$, or 0.005. The results at the final position are in questionable agreement, A or T in simulation and T or C in the Wolberger logo, both with relatively low information contents in the logos. This position was found to contain a water-mediated contact between a serine side chain and the DNA [41], which might contribute to the lower information content, as water bridges allow more promiscuous recognition between protein side chains and DNA bases.

4 Discussion and Conclusion

We have presented and tested a simulation method for calculating the DNA-binding specificity of a transcription factor. The method has no requirement for experimental data, other than the ability to construct an initial homology model for a DNA-protein complex. As most transcription factor families are represented by one or more bound complexes in PDB (Fig. 2), this is not a practical limitation.

This method has the benefit of permitting full main chain motions of both the DNA and the protein. DNA deformation contributions are known to be important contributors to binding specificity and are included in this method. Explicit solvent molecules permit water-bridged hydrogen bonding between transcription factor sidechains and DNA bases that are also known to contribute to specificity.

The simulation design as a single-elimination tournament, together with the additive approximation, make the computational implementation trivially parallelizable by running each position on a different node of a high-performance cluster.

Ongoing work is investigating the sensitivity of predictions to the background DNA sequence used to start the calculation and the quality of the initial homology model. The calculations described here used high-quality structural data for the transcription factor under study. We are now repeating the calculations using homology models built from related family members. Superposition of homeodomain proteins MAT- α 2, Ubx, and Pbx1 present in PDB suggests that homology modeling will introduce little error due to the strong conservation of protein fold (Fig. 5). Furthermore, the critical DNA-binding region is an α -helix that is highly amenable to homology modeling, as opposed to a loop that would be difficult to predict.

Other retrospective analysis of trajectories is being conducted to investigate sampling efficiency. The 100 ps production runs must be sufficiently long to sample relevant motions of protein, DNA, and solvent. Motions that are slower than this timescale may lead to inefficient sampling and simulation error. An important timescale that might be slow is the residence time of water molecules in the DNA-protein interface.

Knowledge gained through these simulations will provide new information for guiding dynamical models of gene regulation. One of the most difficult aspects of modeling is defining the structure of a network. Network structure is defined by the set of interactions, ideally causal interactions, between network components. This topology often must be inferred from observed data by Bayesian network approaches. As the number of possible network topologies is combinatorially large, inference is computationally and practically challenging. The set of gene regulatory relationships defined by predicted protein-DNA interactions will be useful for constraining or biasing which network edges are considered in a topology search. This will permit the development of more detailed, quantitative models of gene regulation.

Acknowledgements

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

References

1. Pabo CO, Sauer RT. Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem* 1992;61:1053–1095. [PubMed: 1497306]URL <http://dx.doi.org/10.1146/annurev.bi.61.070192.005201>
2. Garvie CW, Wolberger C. Recognition of specific DNA sequences. *Mol Cell* 2001;8(5):937–946. [PubMed: 11741530]
3. Balaji S, Babu MM, Iyer LM, Aravind L. Discovery of the principal specific transcription factors of apicomplexa and their implication for the evolution of the ap2-integrase dna binding domains. *Nucleic Acids Res* 2005;33(13):3994–4006. [PubMed: 16040597]1362–4962 (Electronic) Journal Article
4. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. Transcriptional regulatory networks in *saccharomyces cerevisiae*. *Science* 2002;298(5594):799–804. [PubMed: 12399584]1095–9203 Journal Article
5. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. *Nature* 2001;409(6819):533–8. [PubMed: 11206552] 0028–0836 Journal Article
6. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004;431(7004):99–104. [PubMed: 15343339]1476–4687 (Electronic) Journal Article

7. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML. Rapid analysis of the dna-binding specificities of transcription factors with dna microarrays. *Nat Genet* 2004;36(12):1331–9. [PubMed: 15543148]1061–4036 Journal Article
8. Bailey TL, Elkan C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* 1995;21:51–83.
9. Roth FP, Hughes JD, Estep PW, Church GM. Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nat Biotechnol* 1998;16(10):939–45. [PubMed: 9788350]1087–0156 Journal Article
10. Beer MA, Tavazoie S. Predicting gene expression from sequence. *Cell* 2004;117(2):185–98. [PubMed: 15084257]0092–8674 Journal Article
11. Liu, JS. Monte Carlo strategies in scientific computing, Springer series in statistics. Springer; New York: 2001. jun S. Liu. ill.; 24 cm.
12. Liu X, Brutlag DL, Liu JS. Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* 2001:127–38. [PubMed: 11262934]Journal Article
13. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science* 2003;301(5629):71–6. [PubMed: 12775844]1095–9203 Journal Article
14. Choo Y, Klug A. Physical basis of a protein-dna recognition code. *Current Opinion in Structural Biology* 1997;7(1):117. [PubMed: 9032060]
15. Kaplan T, Friedman N, Margalit H. Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Computational Biology* 2005;1(1):5–13.
16. Macisaac KD, Gordon DB, Nekudova L, Odom DT, Schreiber J, Gifford DK, Young RA, Fraenkel E. A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. *Bioinformatics* :1367–4803.(Print) Journal article
17. Gutmanas A, Billeter M. Specific DNA recognition by the Antp homeodomain: MD simulations of specific and nonspecific complexes. *Proteins* 2004;57(4):772–782. [PubMed: 15468320]URL <http://dx.doi.org/10.1002/prot.20273>
18. Duan J, Nilsson L. The role of residue 50 and hydration water molecules in homeodomain DNA recognition. *Eur Biophys J* 2002;31(4):306–316. [PubMed: 12122477]URL <http://dx.doi.org/10.1007/s00249-002-0217-3>
19. Iurcu-Mustata G, Belle DV, Wintjens R, Prévost M, Rooman M. Role of salt bridges in homeodomains investigated by structural analyses and molecular dynamics simulations. *Biopolymers* 2001;59(3):145–159. [PubMed: 11391564]URL <http://dx.doi.org/3.0.CO;2-Z>
20. Billeter M, Güntert P, Luginbühl P, Wüthrich K. Hydration and DNA recognition by homeodomains. *Cell* 1996;85(7):1057–1065. [PubMed: 8674112]
21. Wolberger C. Homeodomain interactions. *Curr Opin Struct Biol* 1996;6(1):62–68. [PubMed: 8696974]
22. Jayaram B, Jain T. The role of water in protein-DNA recognition. *Annu Rev Biophys Biomol Struct* 2004;33:343–361. [PubMed: 15139817]URL <http://dx.doi.org/10.1146/annurev.biophys.33.110502.140414>
23. Havranek JJ, Duarte CM, Baker D. A simple physical model for the prediction and design of protein-dna interactions. *J Mol Biol* 2004;344(1):59–70. [PubMed: 15504402]0022–2836 Journal Article
24. Bulyk ML, Huang X, Choo Y, Church GM. Exploring the dna-binding specificities of zinc fingers with dna microarrays. *Proc Natl Acad Sci U S A* 2001;98(13):7158–63. [PubMed: 11404456]0027–8424 Journal Article
25. Jiang L, Kuhlman B, Kortemme T, Baker D. A solvated rotamer approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins* 2005;58(4):893–904. [PubMed: 15651050]1097–0134 Journal Article
26. Morozov AV, Havranek JJ, Baker D, Siggia ED. Protein-dna binding specificity predictions with structural models. *Nucleic Acids Res* 2005;33(18):5781–98. [PubMed: 16246914]1362–4962 (Electronic) Evaluation Studies Journal Article
27. Lafontaine I, Lavery R. Optimization of nucleic acid sequences. *Biophys J* 2000;79(2):680–685. [PubMed: 10920002]

28. Lafontaine I, Lavery R. ADAPT: a molecular mechanics approach for studying the structural properties of long DNA sequences. *Biopolymers* 2000;56(4):292–310. [PubMed: 11754342]URL <http://dx.doi.org/3.0.CO;2-9>
29. O'Flanagan RA, Paillard G, Lavery R, Sengupta AM. Non-additivity in protein-dna binding. *Bioinformatics* 2005;21(10):2254–63. [PubMed: 15746285]journal Article
30. Paillard G, Lavery R. Analyzing protein-dna recognition mechanisms. *Structure (Camb)* 2004;12(1): 113–22. [PubMed: 14725771]journal Article
31. Benos PV, Bulyk ML, Stormo GD. Additivity in protein-dna interactions: how good an approximation is it? *Nucl Acids Res* 2002;30(20):4442–4451. [PubMed: 12384591]
32. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 1990;18(20):6097–100. [PubMed: 2172928]0305–1048 (Print) Journal Article
33. Crooks GE, Hon G, Chandonia JM, Brenner SE. Weblogo: a sequence logo generator. *Genome Res* 2004;14(6):1188–90. [PubMed: 15173120]1088–9051 (Print) Journal Article
34. Frenkel, D.; Smit, B. Understanding molecular simulation: from algorithms to applications. 2nd Edition. Academic Press; San Diego: 2002. daan Frenkel, Berend Smit. ill.; 24 cm. Computational science series; 1.
35. Wolberger C, Vershon AK, Liu B, Johnson AD, Pabo CO. Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell* 1991;67(3):517–528. [PubMed: 1682054]
36. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 2006;34(Database issue):D108–10. [PubMed: 16381825]1362–4962 (Electronic) Journal Article
37. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J Comp Chem* 1983;4:187–217.
38. Li T, Stark MR, Johnson AD, Wolberger C. Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. *Science* 1995;270(5234):262–269. [PubMed: 7569974]
39. Ke A, Mathias JR, Vershon AK, Wolberger C. Structural and thermodynamic characterization of the DNA binding properties of a triple alanine mutant of MATalpha2. *Structure* 2002;10(7):961–971. [PubMed: 12121651]
40. Li T, Jin Y, Vershon AK, Wolberger C. Crystal structure of the MATa1/MATalpha2 homeodomain heterodimer in complex with DNA containing an A-tract. *Nucleic Acids Res* 1998;26(24):5707–5718. [PubMed: 9838003]
41. Aishima J, Wolberger C. Insights into nonspecific binding of homeodomains from a structure of MATalpha2 bound to DNA. *Proteins* 2003;51(4):544–551. [PubMed: 12784213]URL <http://dx.doi.org/10.1002/prot.10375>

Fig. 1. (left) The crystal structure of the homeodomain transcription factor yeast MAT- α 2 (1APL) demonstrates α -helix recognition of the major groove. (right) The AP2/ERF domain in plants binds with 3 chains of a β -sheet contacting the major groove.

Family	Yeast	Arabi- dopsis	Dictyo- stelium	Malaria	Worm	Fly	Human	Total	PDB	Bound
Classic zinc finger	49	145	18	5	161	494	1710	2582	45	6
Homeodomain	8	136	15	0	119	190	412	880	43	6
MYB	15	562	36	10	19	37	125	804	24	5
Helix-loop-helix	10	239	0	0	56	139	240	684	9	4
Nuclear hormone reptr	0	0	0	0	364	38	178	580	118	3
AP2/ERF	0	236	0	27	0	0	0	263	3	1
Basic leucine zipper	11	85	17	0	18	16	82	229	39	8
Winged Helix / Forkhead	5	0	0	0	22	35	89	151	6	1
Winged Helix / ETS	0	0	0	0	12	14	65	91	14	4

Fig. 2.

A census across representative eukaryotic species shows the most populous transcription factor families. The column PDB lists the number of protein structures, bound and unbound to DNA, in the Protein Data Bank. The column Bound provides the number of DNA complexes that are unique at 50% sequence identity.

Position	1	2	3	4	5	6	7	8	9	10
Round 1: AT → TA										
protein-DNA complex	0.6±0.9	-0.7±0.7	-2.8±0.9	-2.3±0.9	2.1±1.2	1.0±1.0	1.6±0.8	1.1±1.1	1.3±1.0	2.9±1.0
DNA in water	-1.4±0.8	0.9±1.1	0.2±0.9	0.3±0.9	-4.1±1.0	0.1±1.0	-2.5±0.9	0.4±0.9	-3.9±0.8	0.6±1.0
$\Delta\Delta G$	2.0±1.3	-1.6±1.3	-2.9±1.3	-2.7±1.3	6.2±1.5	0.9±1.4	4.1±1.2	0.7±1.4	5.3±1.3	2.2±1.4
Round 1: CG → GC										
protein-DNA complex	-0.1±1.0	-0.9±1.3	4.3±1.2	-1.3±0.8	-3.1±1.0	4.7±1.2	-1.9±1.0	0.8±1.0	1.3±1.0	2.9±1.1
DNA in water	-2.0±1.0	-1.1±1.0	1.1±1.0	1.7±1.0	-5.4±0.9	-3.7±1.1	-3.4±1.1	-4.1±1.0	2.2±1.1	1.3±0.9
$\Delta\Delta G$	1.9±1.4	0.2±1.6	3.2±1.6	-3.0±1.2	2.3±1.4	8.5±1.6	1.5±1.4	5.0±1.4	-0.9±1.5	1.7±1.4
Round 2										
	AT → CG	CG → TA	CG → TA	GC → TA	AT → CG	AT → CG	AT → CG	AT → CG	AT → GC	AT → CG
protein-DNA complex	-81.7±0.8	85.1±0.9	81.3±0.9	78.0±1.1	-76.7±1.1	-83.6±0.8	-78.8±0.8	-78.8±0.8	-81.7±1.0	-83.4±1.1
DNA in water	-82.1±1.0	84.1±0.9	84.6±0.8	84.3±1.0	-82.5±0.8	-79.9±1.0	-81.1±1.0	-81.0±1.2	-84.3±1.0	-82.2±0.9
$\Delta\Delta G$	0.4±1.3	1.0±1.2	-3.3±1.2	-6.2±1.4	5.8±1.4	-3.7±1.3	2.3±1.3	2.2±1.4	2.6±1.4	-1.2±1.5
Final winner	AT	CG	TA	TA	AT	CG	AT	AT	AT	CG

Fig. 3.

Free energy results are provided for the single-elimination tournament for MAT- $\alpha 2$. Energies are in kcal/mol $\pm$ standard error.

Fig. 4.

Sequence logos are presented for MAT- α 2: (top) simulation; (middle) Transfac database based on MAT- α 1 / MAT- α 2 heterodimer; (bottom) primary literature based on MAT- α 2 / MCM1 tetramer [35].

Fig. 5. Strong conservation of the folds of homeodomain proteins for yeast MAT- α 2 (1APL, red), fly Ultrabithorax (Ubx, 1B8I, blue) and human Pbx1 (1B72, green) suggests that homology models will provide a suitable starting point for calculating binding preferences of homeodomain transcription factors. The corresponding DNA bound by each homeodomain protein is in lighter color than that of the protein.