Structure-based modeling of protein: DNA specificity

Adam P. Joyce, Chi Zhang, Philip Bradley and James J. Havranek
Advance Access publication date 19 November 2014

Abstract

Protein:DNA interactions are essential to a range of processes that maintain and express the information encoded in the genome. Structural modeling is an approach that aims to understand these interactions at the physicochemical level. It has been proposed that structural modeling can lead to deeper understanding of the mechanisms of protein:DNA interactions, and that progress in this field can not only help to rationalize the observed specificities of DNA-binding proteins but also to allow researchers to engineer novel DNA site specificities. In this review we discuss recent developments in the structural description of protein:DNA interactions and specificity, as well as the challenges facing the field in the future.

Keywords: Protein:DNA interactions; Molecular modeling; Binding specificity; Structure prediction; Structural modeling; Transcription factor binding sites

Sequence-specific protein:DNA interactions are critical for proper cellular functioning; consequently, there is substantial interest in predicting and/or reengineering their specificity. Amino acid changes in DNA-binding proteins can act as driving alterations that lead to disease [1–3] or evolutionary adaptation [4]. Changes in the affinities of transcription factors for mutated binding sites can also alter the occupancy and identity of bound proteins in gene regulatory regions, resulting in phenotypic consequences that may fuel evolutionary change [5–8]. Scientists have applied tools from structural biology to achieve an atomic-level understanding of binding mechanisms for a number of protein:DNA complexes [9]. The structures of these complexes have shed considerable light on the determinants of DNA sequence readout [10] (Figure 1), effectively refuting the idea of a simple and general ‘code’ for protein:DNA recognition [14], while at the same time enabling rational structure-guided engineering of DNA interaction specificity for certain families [15].

Structure-based computational approaches to binding prediction seek to rationalize observed specificity patterns and predict new interactions. Broadly speaking, these approaches proceed by constructing three-dimensional models of protein:DNA complexes (Figure 2) and deriving estimates of binding affinity and/or specificity from them. Structure-based approaches vary in their degree of computational and physical rigor, ranging from relatively low-resolution statistically based potentials to all-atom molecular dynamics simulation. In comparison, nonstructural approaches are often far less computationally intensive, require little or no knowledge of physical interactions and frequently yield models of equal or greater quality than state-of-the-art structure-based calculations when provided with sufficient experimental binding data for training.

Corresponding author. James J. Havranek, Campus Box 8232, 4566 Scott Avenue, St. Louis, MO 63110. Tel.: 314-362-2535; Fax: 314-362-7855; E-mail: havranek@genetics.wustl.edu

A. P. J. and C. Z. contributed equally

Adam Joyce holds a bachelor’s degree in Biology and Biotechnology from Tufts University. His research interests are focused on the evolution and design of tools for synthetic biology.

Chi Zhang’s research interests are in the computational design and directed evolution of molecular tools for genome engineering. Phil Bradley’s research is focused on the prediction and design of protein structures and interactions. He has been a faculty member at the Fred Hutchinson Cancer Center since 2007.

Jim Havranek’s research interests include the development of algorithms for computational protein design, quantitative modeling of protein–DNA interactions and the application of protein engineering to enable novel proteomics assays.

© The Author 2014. Published by Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com
However, structural modeling of the protein:DNA interface can provide substantial information beyond predictions of binding specificity. First, the physical forces that govern protein:DNA interactions are generalizable to any protein:DNA complex; therefore, advances in structure-based modeling can have immediate and significant impact on our ability to model thousands of individual genomic interactions. Second, structural models of protein:DNA complexes are highly useful to model secondary binding events, such as interactions with a protein cofactor or allosteric regulator. Third, structural models facilitate the \textit{in silico} exploration of mutations or covalent modifications to protein or DNA (e.g. CpG methylation, DNA damage, or protein phosphorylation). Lastly, energy functions and sampling methods developed for binding site prediction have the potential to drive innovation in the engineering and design of genomic tools, such as synthetic transcription factors and site-specific nucleases [16–18].

This review focuses on recent developments and future challenges for the prediction and design of protein:DNA structures and interactions using the techniques of macromolecular modeling, ranging in resolution from coarse-grained statistical potentials to all-atom molecular dynamics simulation. While the protein:DNA interface likely represents a challenging application area for structure-based methods due to its highly solvated and polar character, the wealth of high-throughput experimental binding data that has become available in the past few years suggests that this field is poised for substantial progress.

\textbf{IMPROVEMENTS IN STATISTICAL POTENTIALS}

The accuracy of structure-based binding predictions depends critically on the quality of the potential energy functions used to estimate binding affinities from modeled complexes. The potential energy functions that have been used for this purpose can be roughly classified as being either physics- or knowledge-based. The functional form of the energy terms in physics-based potentials is derived from a physicochemical model of the underlying interactions, and as a result these potentials can be quite sensitive to the atomic coordinates: small changes in atomic position can lead to large changes in computed energy due to steric or electrostatic clashes. In knowledge-based statistical potentials, on the other hand, the interaction potentials are derived from experimentally determined protein:DNA structural information. The probabilities of observing different kinds of interactions in crystal structures are calculated and converted into potential energies, for
example, by using the inverse Boltzmann approach. Statistical potentials can model any previously observed behavior even if the underlying physical phenomena are poorly understood. However, they cannot predict atomic interaction patterns absent from the training set of available protein:DNA structures [19, 20]. The resolution of statistical potentials can vary from atom-level to residue-level; in general they do not have the sensitivity of molecular mechanics potentials.

The moderate spatial resolution of statistical potentials makes them a good match for scoring the approximate structural models generated by homology modeling or by the docking of unbound structures (Figure 2). In contrast, molecular mechanics potentials may be less forgiving in these cases, due to the steric clashes often present in these complexes. Chen et al. [21] used structural alignment to generate synthetic protein:DNA complexes from structures of unbound proteins, and applied a statistical potential to predict position weight matrices (PWMs) for these proteins. Although PWMs generated using this approach were less accurate than those generated from native complexes, results were comparable with those obtained from complexes generated by docking, and were better than those obtained from homologous complexes generated by bound structural templates from the same protein family. Their analysis demonstrated the utility of statistical potentials for predicting PWMs given

Figure 2: Modeling protein:DNA complexes. The choice of protocol depends on the structural 'template' available for constructing the model. If a bound structure is available for the protein of interest ('Native complex', top left), the modeling needed for binding predictions involves primarily base pair mutations ('Gua→Ade: template in cyan/dark gray and model in yellow/light gray) and side chain rearrangements (gray arrow). Building a model using a homologous complex as a template will require protein ('R→A, E→N') as well as base pair mutations, and may require protein and DNA backbone relaxation. If the unbound structure of the native protein is known, a DNA-bound model can be constructed by superimposing this unbound structure onto the structure of a homologous factor in a bound structure (bottom left), or by de novo 'docking' onto DNA (bottom right, multiple candidate docked conformations shown). (A colour version of this figure is available online at: http://bfg.oxfordjournals.org)
approximate models, and also indicated that correctly capturing the conformational changes of proteins on binding DNA will be important for future improvements. A number of alternate approaches exist for generating synthetic complexes, and it remains to be seen whether they yield improved models for predicting protein:DNA specificity [22].

Atomic resolution is usually preferred when using statistical potentials to predict protein:DNA binding specificity, yet atomistically detailed statistical potentials have very large numbers of parameters which can make them challenging to train robustly (for example, a pairwise atomic potential with 30 atom types and 10 distance bins has 4650 free parameters). Recently, improvements have been made to train the potentials more efficiently. Xu et al. [23] developed an energy function that was trained to include the target structure templates themselves in recognizing transcription factor binding sites. This development led to increased prediction accuracy and robustness compared with their previous potential, vcFIRE [24]. Their method also outperformed sequence-based approaches in prediction accuracy in cases for which limited experimental data was available. In another approach, the training incorporated experimentally determined PWMs. Traditionally, statistical potentials count the number of times a given interaction is observed across protein:DNA complexes and assume that each complex is equally likely. However, the occurrence frequencies can also be weighted proportionally to the binding affinity of the protein for different DNA sequences. AlQuraishi & McAdams trained their potentials by weighting DNA sequences differently according to their experimental probability of occurrence specified by their corresponding PWMs [25]. Although this approach did not significantly improve PWM predictions, it was a novel step in the long-term goal of combining structural data with biochemical data for protein:DNA binding site prediction.

In contrast to atomistic potentials, coarse-grained residue-level potentials do not generally have sufficient resolution to make predictions for PWMs. However, they are well-suited for generating protein:DNA complexes by docking unbound structures. Although residue-level potentials have far fewer parameters than atom-level potentials and require less computing power, docking with large decoy sets can still be computationally intensive. Parisien et al. [26] applied machine-learning techniques to reduce the number of parameters required in their residue-level potential function to 15. Their rigid body docking protocol performed well at rebuilding native protein:DNA contacts for both bound and unbound structures, although it was still a challenge to achieve root-mean-square deviations below 5Å when using unbound structures as the starting point. Besides reducing parameters, efforts have been made to make statistical potentials more accurate. Most statistical potentials are distance-based and thus may benefit from including an angular term. Takeda et al. derived a novel orientation-dependent residue-level potential for protein:DNA docking [27]. Their potential performed significantly better than their previous multi-body potential in docking accuracy. Its binding affinity prediction was also greatly improved and was on par with some atom-level statistical potentials, though it was still less accurate than others (e.g. vcFIRE) [24].

Finally, because statistical potentials usually require much less computational power than physics-based potentials, they can easily be adapted to run on web servers. Three web servers for predicting PWMs using protein:DNA complexes have been constructed in the past few years, making these statistical potentials easily accessible to researchers without a computational background: 3D-footprint [28], 3DTF [29] and PiDNA [30].

MODELING WATER IN PROTEIN–DNA INTERFACES

Modeling the role of water is likely to be more important for protein–DNA interfaces than for other macromolecular calculations. Biochemical and structural data indicate that water-mediated interactions play a key role in protein–DNA recognition (Figure 3A and B) [31, 32]. This is in contrast to the modeling tasks of protein folding and docking, which have achieved notable successes without incorporating explicit water molecules [33, 34]. In addition, the polyanionic nature of nucleic acids suggests that electrostatics, also commonly omitted from protein modeling, will figure prominently in any energetic description of protein:DNA complexes. Water plays an important role in quantitative models for electrostatic phenomena by virtue of its high dielectric constant. Finally, protein:DNA interfaces possess many polar and charged amino acids that are sequestered from bulk solvent, yet must still satisfy their hydrogen bonding potential. Water can serve this role by filling voids in the interface and
providing hydrogen bond donors or acceptors for polar groups in both the protein and DNA.

The effects of water on the energetics of a protein–DNA complex can be treated at several levels of detail. At one extreme is the complete neglect of explicit water molecules, perhaps partially compensated by the inclusion of an implicit solvation potential [35]. In one study, the ability of water to attenuate hydrogen bonds, but not to participate in them, was considered [36]. At the other end of the spectrum is the explicit treatment of water molecules that fully solvate a macromolecule or complex using molecular mechanics [37–39]. Computational protocols also differ in where and how explicit water molecules are introduced into a model. For instance, water networks have been constructed en masse, with the goal of optimizing hydrogen bonding across an entire interface [40]. Water molecules have also been attached to polar groups in amino or nucleic acids at optimal geometries for hydrogen bonding, giving rise to the ‘solvated rotamer’ strategy [41]. In some approaches, the locations of water molecules are determined simultaneously along with the conformational sampling that optimizes the protein:DNA interface [42, 43]. The specific choices of how water molecules are modeled, and where and when they enter into the calculation, are based on trade-offs between the accuracy of the physical potential used, the scale of conformational sampling that is to be considered and the computational resources that are available.

Unsurprisingly, given the extra computational requirements and significant uncertainty in the optimal approach for modeling water in protein:DNA interfaces, few studies include water in the calculation of protein–DNA binding specificity. Nevertheless, some conclusions can be drawn regarding the impact of including explicit water. Van Dijk et al. [43] incorporated explicit water into the protein:DNA docking capabilities of HADDOCK (High Ambiguity Driven protein-protein DOCKing). While they were not explicitly calculating DNA-binding preferences, the methods they describe are readily transferable to the protein:DNA homology modeling problem. Water molecules were first placed on unbound models for both protein and DNA based on the results of molecular dynamics simulations. During a subsequent docking step, water was removed or added from the developing complex using a Monte Carlo approach. The inclusion of water molecules led to modest but significant improvements in the docked complex geometries. In particular, they were able to recover specific water-mediated hydrogen bonds in the Engrailed homeodomain:DNA interface [44]. They found most consistent success in those cases where the bound and unbound conformations of the protein were very similar, which is expected

Figure 3: Water molecules at the protein:DNA interface participate in hydrogen bonding networks. (A) The trp repressor protein achieves recognition of its operator sequence through multiple water-mediated contacts, involving both protein side chain and mainchain atoms. (B) The EcoRI restriction enzyme interacts with its cognate cleavage site with both and water-mediated contacts. Failure to model water molecules explicitly leads to a relaxed DNA specificity profile reminiscent of ‘star activity’, which has been attributed to the loss of bound interfacial water. (A colour version of this figure is available online at: http://bfg.oxfordjournals.org)
for the homology modeling calculations required to estimate PWMs.

Li and Bradley directly studied the effect of explicit water molecules on predicting protein:DNA recognition specificity [42]. Their method considered water molecules only at the consensus minor and major groove locations that have been determined from crystallographic studies [45]. Water occupancy at these locations was allowed to vary during the course of the structural optimization. Similar to Van Dijk et al., they observed limited but significant improvement over a large test set of protein:DNA complexes. Notably, the inclusion of explicit water led to improvements in the description of water-mediated hydrogen bonds that are known to be important for the specificity of the EcoRI restriction enzyme (Figure 3B). Interestingly, neglecting explicit water molecules yielded a specificity profile consistent with EcoRI ‘star activity’. Star activity has been linked experimentally to the release of bound interfacial waters thought to participate in the formation of the cognate protein:DNA complex [46]. Of particular interest for the calculation of PWMs, their method was able to predict correctly that in the case of one experimentally determined protein:DNA complex, a higher affinity DNA sequence than the one in the crystal structure could be found. This demonstrates that it is possible for structure-based calculations to use an experimental structure as a homology modeling template to accurately describe water-mediated protein:DNA interactions not found in the original complex.

In summary, the consideration of explicit water molecules can lead to a more faithful description of protein:DNA recognition specificity. The improvements have been found to be clear, if modest in effect [42, 43]. However, in certain cases key water-mediated interactions appear to be crucial for describing specificity, and approaches that neglect explicit water may not generate useful PWMs. In the near future, we are likely to witness improvements in the placement and scoring of water molecules and their interactions, as well as in the computational efficiency of calculating these effects.

FLEXIBILITY AT THE PROTEIN:DNA INTERFACE

The Protein Data Bank contains representative structures for the majority of known DNA-binding protein families in complex with DNA (~3000 total structures, with substantial redundancy), and predictions based on these homologous ‘template’ structures have the potential to expand our knowledge of sequence-specificity to thousands of uncharacterized proteins. However, homologous template complexes present a single static conformation that is unique to the crystallized protein and DNA molecules, and sequence changes to either partner often result in steric clashes or, conversely, novel low-energy states. In these cases, it is necessary to sample and evaluate any deviations from template coordinates within a set of allowable conformations reflecting the total ‘flexibility’ of the protein backbone, amino acid side chains, bases or base pairs and the sugar–phosphate backbone. Physically, flexibility is integral to the process of protein:DNA recognition. Within a single protein:DNA complex, both inter- and intramolecular contacts vary according to DNA sequence, and individual side chains freely adopt alternative conformations in specific and nonspecific binding modes [47]. Comparison of protein:DNA interfaces in the free and DNA-bound states has revealed greater intrinsic structural variation in protein:DNA interfaces than other surface areas [48–50]. Additionally, crystallographic studies have shown that extensive contact with proteins can induce significant deviation from the canonical B-form DNA backbone and standard base pair geometry [51]. Collectively, these findings demonstrate that both protein and DNA can exhibit conformational changes relative to their unbound structures.

The incorporation and conformational sampling of new side chains are essential for the prediction of sequence specificity using homologous proteins or unbound structures as templates. Typically, this search is discretized using libraries of torsion-angle rotamerized side chains [52, 53]. Using Monte Carlo optimization of rotamer selection, Havranek et al. [53] demonstrated recovery of both identity and native conformation for DNA-contacting residues in the presence of DNA, with accuracy comparable with modeling of monomeric proteins. This model was further extended to include a simplified representation of DNA strain; however, compared with full conformational relaxation of both protein side chains and DNA in a single native complex, a ‘static model’ allowing neither side chain nor DNA motion reproduced experimental PWMs more accurately in most cases [35]. In this study,
conformational sampling was least accurate when water molecules were omitted from the structural templates. Parisien and colleagues also found that side chain reorganization in unbound structures significantly reduced the recovery of native protein:DNA contacts in 47 protein:DNA structures using the rigid-body docking tool FTDock [26]. Together, these studies illustrate that additional degrees of freedom in interfacial side chains, in the absence of appropriate constraints can reduce the accuracy of structural and specificity prediction.

Currently, most homology-based predictions of protein:DNA specificity rely on the assumption that the target and template structures possess sufficiently similar, if not identical, backbone coordinates. Violations of this assumption can have dramatic functional consequences [54], and, given that increased backbone flexibility has been commonly observed in protein:DNA interfaces [48, 49], this assumption is likely to be inappropriate for modeling many DNA-binding proteins (Figure 4A). Moreover, polar amino acids with long side chains, which are enriched at protein:DNA interfaces, commonly form distance- and orientation-constrained contacts with specific DNA bases, and will experience large deviations in torsional sampling space following subtle backbone movements [55]. Correct backbone placement is therefore essential for an accurate depiction of protein:DNA contacts. Using a novel fragment insertion protocol to improve backbone torsional sampling, Yanover and Bradley generated homology models of C2H2 zinc fingers that recapitulated near-native docking conformations, base-specific contacts and experimentally generated models of sequence specificity [56]. Havranek and Baker introduced structure-guided backbone flexibility using a motif library of observed side chain:base contacts, termed ‘inverse rotamers’ [55]. In this approach, after incorporating a motif into the DNA template, the adjacent protein backbone was allowed to sample nearby positions; changes were accepted if the backbone could accommodate the motif in an energetically favorable conformation.

The protein-bound DNA backbone frequently displays both local and global deformation from the standard B-form helix, and resultant changes in the positions of phosphate atoms and base parameters can substantially impact binding conformation and sequence recognition (Figure 4B). Siggers and Honig developed a torsional sampling approach in which mutated base pairs were introduced with coplanarity to the template bases and subsequently conformationally diversified by means of small compensating rotations about four DNA backbone torsion angles [22, 57]. This increase in DNA flexibility substantially improved specificity prediction, especially for templates with low similarity to the target structure. Yanover and Bradley introduced conformational diversity into both protein and DNA backbones
simultaneously by insertion of fragments from multiple template structures of protein:DNA complexes [58]; the DNA backbone sampling procedure of Siggers and Honig was then applied to minimize the local impact of fragment insertion [22]. Full-atom simulation of bound and unbound DNA using molecular mechanics force fields is another powerful, though computationally intensive, approach to modeling DNA deformation [58]. Steady gains in computing power and optimization of nucleic acid force field parameters have improved the speed and accuracy of molecular dynamics (MD)-based methods [59, 60], but the combinatorial challenge of minimizing all possible DNA sequences that a protein may bind has been a major barrier to the application of MD to specificity prediction. Using the ADAPT methodology, Deremble and colleagues developed a technique to subdivide the DNA interface into overlapping pentanucleotide segments, which are independently evaluated and summed to yield the total, sequence-dependent energy of the protein:DNA complex [61]. The substantial reduction in computing time permitted the simultaneous conformational relaxation of both protein and DNA, and achieved accurate structural predictions for proteins bound to highly deformed DNA [62].

EVALUATING IMPROVEMENTS IN PROTEIN:DNA MODELING

A wealth of experimental data on protein:DNA interactions is now available for training and testing structure-based approaches. High-throughput in vitro [63–67] and in vivo [68] experimental methods have been developed that can produce rich binding affinity profiles for multiple DNA binding proteins relatively rapidly. These methods enable the mapping of affinity landscapes for individual DNA binding proteins with unprecedented depth and resolution, facilitating the detection of subtle binding features such as secondary motifs [69], correlations between target site positions [67], higher-order binding interactions [64] and DNA-shape-mediated readout [70]. In addition, these methods have been applied to survey large families of homologous factors, providing valuable data on the mapping between protein sequence and DNA binding specificity within families [71, 72].

The standard approach to benchmarking a structure-based algorithm has been to reduce the reference experimental data set to a PWM, to similarly condense the output of the prediction algorithm, and then to assess the agreement between the two PWMs by aligning them and scoring the strength of the alignment using one of a number of established PWM comparison metrics [73, 74]. This approach ignores the richness of deep binding affinity data sets, and it also overlooks the potential of structure-based approaches to rationalize exactly those higher-order effects that are neglected by the PWM representation. Historically, it has been a challenge to recapitulate even the first order, position-independent binding profile, and this remains a valuable assessment for benchmarking, particularly in template-based approaches. We anticipate, however, that as structural modeling methods continue to improve, it will be increasingly informative to directly compare predicted and experimentally measured relative affinities for large sets of full-length target site sequences (rather than PWM columns or consensus sequences), particularly for target proteins with a bound, high-resolution, crystal structure. This comparison should be particularly enlightening when applied across families for which multiple experimental binding profiles and co-crystal structures are available, giving insight into the origins of binding specificity divergence among related proteins.

PROSPECTS FOR THE FUTURE

The structure-based prediction of protein:DNA specificity will be affected by several ongoing trends. First, we can expect that high-throughput experimental techniques will continue to provide a wealth of protein:DNA affinities useful in both training and testing the robustness of structure-based prediction algorithms. Second, the number of experimentally determined crystal structures of protein:DNA complexes will continue to grow. The availability of examples of additional structural families will expand the number of DNA-binding proteins that are amenable to structural modeling of specificity. The availability of complexes with different DNA sequence specificities, altered binding modes, and diversified backbone conformations will provide more appropriate starting templates for homology modeling, lessening the need to incorporate protein or DNA flexibility in modeling calculations. Examples of novel protein:DNA complexes will also add to the set of training data for statistical potentials. Finally, the steady increase in computing power will facilitate improvements in scoring potentials and conformational sampling.
previously described. Furthermore, the nature of specificity calculations (involving evaluations of a protein bound to multiple DNA sequences) make them an ideal fit for the parallel architectures increasingly available to individual researchers at reasonable costs.

**Key points**

- A range of scoring potentials from knowledge-based models to molecular dynamics have been applied to the structural modeling problem. Choosing a scoring function involves trade-offs between physical rigor and computational resources.
- Rarely do researchers have experimental models for all of the complexes in which they are interested; changes to both nucleic acid and protein sequences must be modeled, with the potential for error.
- Current scoring functions may lack contributions from crucial phenomena such as water-mediated hydrogen bonding and electrostatic damping.

**Acknowledgements**

Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R01GM088277 (to P.B.) and R01GM101602 (to J.J.H.). The National Institutes of Health under Award Number R01GM101602 (to J.J.H.).

**References**

15. Wolfe SA, Ramm EI, Pabo CO. Combining structure-based design with phase display to create new Cys(2)His(2) zinc finger dimers. *Structure* 2000;8:739–50.


