TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 03/10 pp568-577
Volume 19, Number 6, December 2014

Structure-Based Prediction of Transcription Factor Binding Sites

Jun-tao Guo*, Shane Lofgren, and Alvin Farrel

Abstract: Transcription Factors (TFs) are a very diverse family of DNA-binding proteins that play essential roles in the regulation of gene expression through binding to specific DNA sequences. They are considered as one of the prime drug targets since mutations and aberrant TF-DNA interactions are implicated in many diseases. Identification of TF-binding sites on a genomic scale represents a critical step in delineating transcription regulatory networks and remains a major goal in genomic annotations. Recent development of experimental high-throughput technologies has provided valuable information about TF-binding sites at genome scale under various physiological and developmental conditions. Computational approaches can provide a cost-effective alternative and complement the experimental methods by using the vast quantities of available sequence or structural information. In this review we focus on structure-based prediction of transcription factor binding sites. In addition to its potential in genome-scale predictions, structure-based approaches can help us better understand the TF-DNA interaction mechanisms and the evolution of transcription factors and their target binding sites. The success of structure-based methods also bears a translational impact on targeted drug design in medicine and biotechnology.

Key words: transcription factor binding site; structure-based predictions; knowledge-based potential; physics-based potential

1 Introduction

One of the grand challenges in post-genomic bioinformatics is to discover the dynamic regulatory networks embedded in static genomic sequences. Transcription Factors (TFs) regulate gene expression through interactions with specific DNA sequences, called Transcription Factor Binding Sites (TFBSs)[1,2]. Identification of TFBSs on a genomic scale represents a critical step in delineating transcription regulatory networks and remains a major goal in genomic annotations. Traditionally, experimental techniques such as DNase I footprinting and gel-mobility shift assay have been used to identify TF-binding sites. However, these methods are time-consuming and not suitable for large-scale studies[3]. Current High-Throughput (HT) experimental approaches are more efficient in determining the binding specificity on a large scale. These methods include Systematic Evolution of Ligands by EXponential enrichment (SELEX)[4-6], Chromatin ImmunoPrecipitation (ChIP)-based technology, such as ChIP-chip[7] and ChIP-seq[8], and Protein Binding Microarrays (PBM)[9].

The high-throughput experimental technologies for TFBS characterization consist of *in vitro* and *in vivo* methods. SELEX and PBM represent two widely used *in vitro* approaches. SELEX applies an iterative selection procedure, in which a target transcription factor and a large library of combinatorial double-stranded oligonucleotides are first mixed together. The solution is then passed through a filter to remove weak or non-binding oligonucleotides. The bound sequences are amplified by Polymerase Chain Reaction

• Jun-tao Guo, Shane Lofgren, and Alvin Farrel are with Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA. E-mail: jguo4@uncc.edu.
∗ To whom correspondence should be addressed.
Manuscript received: 214-06-17; accepted: 2014-06-24

(PCR) for subsequent rounds of selection with more stringent elution conditions at every iteration. After a certain number of rounds of selection, the sequences with high binding affinity are identified. The main strength of this method is its ability to effectively determine high affinity TFBSs. The newly developed HT-SELEX methods improve the detection accuracy for low-specificity TFs[10]. PBM is a microarray-based technology for characterizing TFs' binding specificities in a high-throughput manner[9]. Original PBM used random sequences as probes, but a later design called universal protein binding microarrays uses a de Bruijn sequence that contains every possible combination of oligonucleotides, up to a given length[11]. PBM is able to assess the binding affinity of a TF with any given nucleotide sequence. For both SELEX and PBM, a TF-binding specificity is usually represented as a Position Weight Matrix (PWM)[10]. A recent comparative analysis revealed that SELEX- and PBM-derived binding models are in agreement for most TFs[12].

While these *in vitro* methods have the benefit of directly measuring the TF-DNA binding specificity, they do not resemble the actual cellular environment that includes the presensce of cofactors, nucleosome occupancy, and epigenetic factors such as DNA methylation. The widely-used *in vivo* high-throughput approaches are ChIP-based methods, including ChIP-chip and ChIP-seq, in which the binding sequences are identified with either traditional microarray or massively parallel DNA sequencing technology after the sequences are pulled down with antibodies of targeted transcription factors. The reads are then aligned to the reference genome of the organism being studied. Statistical analysis software is generally used to align clusters of reads, called peaks, to determine the signals of true TF binding from the artifacts[13, 14]. The main strength of these ChIP-based methods is their ability to probe locations of the TF-binding sites in living cells. As a result, these *in vivo* methods may not capture all the binding events that are regulated in different physiological conditions and at different developmental stages, as they are the results of snapshots under certain conditions. In addition, they only provide the approximate locations of TF-binding sites, as ChIP-Seq peaks are generally at least 100 bp in width, much larger than any typical TF-binding sites. Therefore, motif prediction programs are needed to identify the binding sequences within the peak regions[15-19]. How to tell direct interactions from indirect interactions also represents a challenge[13, 20].

2 Computational Methods

To take advantage of the rapidly increasing genomic sequences, sequence-based computational methods have been developed and proved to be valuable in predicting TF-binding sites[21]. These computational techniques generally rely on the similarity of binding sequences for deriving the binding preference of a transcription factor, which then serves as a "query profile" for genome-scale scanning for additional binding sites. A number of different algorithms have been made available for TF-binding site prediction using promoter sequences. Examples include probabilistic models, such as Expectation-Maximization (EM)[22] and Gibbs sampling[23]. Recent algorithms also integrate phylogenetic footprinting or orthologous sequences into the traditional prediction scheme[3]. These sequence-based methods have also been applied to the experimental ChIP-based methods for the identification of binding motifs[13]. Since the DNA binding sequences are usually short (typically 5-15 base-pairs in length) and degenerate, sequence-based approaches generally suffer from a high number of false positive predictions and are at disadvantage if the signal embedded in the binding sites is weak, especially when the TF binding sequences significantly deviate from the known consensus sequences. Several recent studies have also demonstrated that some transcription factors can recognize multiple distinct sequence motifs[24-26]. There is clearly a need to investigate alternative methods that are not constrained by sequence conservation.

Structure-based prediction methods that focus on protein-DNA interactions rather than sequence conservation represent one such approach. *In vivo*, the specific binding between a TF and its binding sites relies on their biophysical interactions. Therefore structure-based methods closely mimic the real binding and recognition events. While experimental technology and sequence-based computational methods can answer the *where* (genome location) and *what* (the binding sequence) questions, structure-based approaches can also provide explanations to *why* and *how* they bind at these locations. More importantly, structure-based methods can help explain the possible effects of mutations on gene expression and guide the rational

design of therapeutic drugs[27]. Although research into protein-DNA recognition focusing on the "recognition code" was first attempted in the 1970s[28], interaction-based protein-DNA recognition did not receive renewed attention until several years ago, when a large number of high-resolution protein-DNA complex structures in Protein Data Bank (PDB) became available[29].

Another advantage of structure-based TF-binding site prediction is its capability of considering the position interdependence of TFs and the contribution of flanking sequences that are not conserved to the binding specificity[26,30]. For example, many nuclear receptors, a special class of transcription factors, recognize highly similar hexanucleotide sequences, e.g., AGGTCA[31]. From a pure sequence point of view it would be difficult to address how different nuclear receptors distinguish their native binding sites from many highly similar subsequences. Structural analysis of the nuclear receptor-DNA complex structures suggests that the DNA sequences flanking the recognition site, though not conserved, are important in TF-DNA binding specificity[31]. The flanking sequences contribute to the binding specificity through making contacts with the DNA binding domain of the nuclear receptors[30]. It has been shown that the C-Terminal Extension (CTE) of Rev-ERBα (NR1D1), a heme sensor involved in metabolic and circadian pathways, interacts with the 5 flanking sequence in the minor groove[32,33]. Our analysis of eight nuclear receptors revealed that the highly conserved region folds as an α-helix and interacts with the conserved response element AGGTCA in the major groove while the less conserved CTE region makes contacts with the minor groove (Fig. 1). These examples highlight the need to consider structural interactions between transcription factors and DNA sequences for accurate description of binding specificity and prediction of transcription factor binding sites.

A typical structure-based TF-binding site prediction method has the following steps: The query DNA sequence is "threaded" onto the DNA structure in a TF-DNA complex and the compatibility between the DNA sequence and the transcription factor is then evaluated using energy functions. Virtually all structure-based methods for TF-binding site prediction require a TF-DNA interaction model. Most studies use experimentally solved protein-DNA complex structures[34-36]. Homologous TF-DNA models have also been used for TF-binding site prediction based on the observation that proteins from the same family, in general, interact with DNA in a similar fashion[37]. For example, Kaplan and coworkers[38] constructed a model for a target protein from the Cys2His2 zinc finger family and its DNA binding sites through analyzing homologous protein-DNA complex structures. The prediction reliability is demonstrated in a genome-wide scan for targets of Cys2His2 transcription factors in *Drosophila melanogaster*. Siggers and Honig[39] used a few members of the Cys2His2 zinc-finger family to predict the binding specificity for the entire protein family.

One immediate question in structure-based TFBS prediction concerns the divergence of DNA structures of various cognate binding sequences for one TF. For this method to work, these DNA structures should be very similar. We searched the PDB for a set of TF-DNA complexes in which the same transcription factor (or highly homologous transcription factor

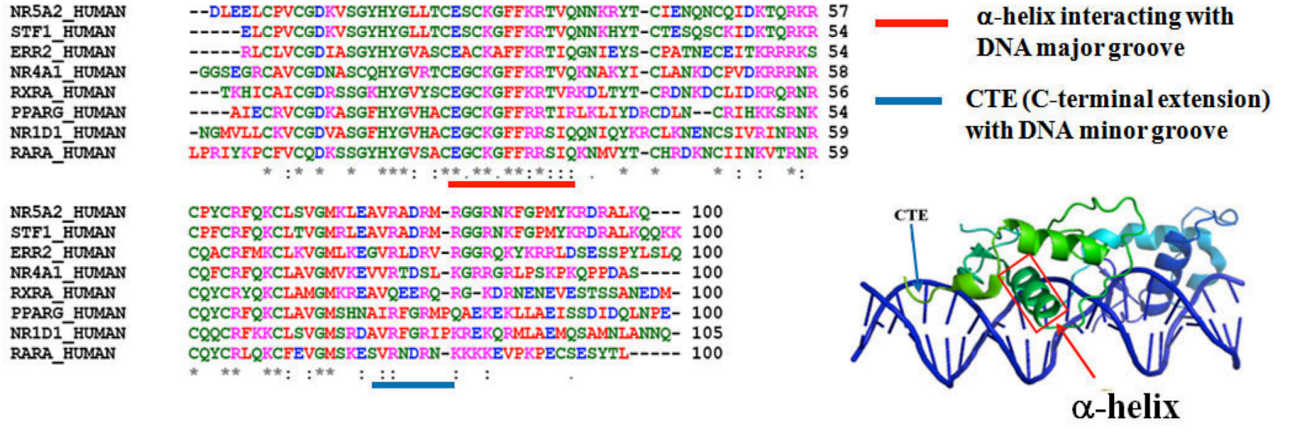

Fig. 1 Sequence alignment of eight nuclear receptors using ClustalW. The red bar represents the positions of conserved DNA binding residues that fold as an α-helix while the blue bar shows the less conserved CTE region. The protein-DNA complex structure is Rev-ERB and its response element (PDB id: 1GA5, image was created with Pymol).

with more than 95% sequence identity) binds to at least two different DNA binding sequences. For each of the 10 sets identified, the structures of the transcription factor were first superimposed and the structural differences of DNA molecules in terms of backbone Root Mean Square Deviation (RMSD) were then calculated (Fig. 2). We found that the DNA structures of different binding sequences of a transcription factor are generally conserved well with RMSDs less than 0.2 nm, suggesting that the interaction modes are conserved between the transcription factor and its specific binding sequences regardless of the variability at certain positions. For example, only 10 of the 18 positions of the six purine repressor binding sites are conserved while the other positions show different degrees of variations (Fig. 2a). Yet, the DNA structures between each pair of the complexes show very little divergence, with a range from 0.02 to 0.12 nm RMSDs. The only set with relatively higher DNA RMSDs (from 0.25 to 0.35 nm) is from three different CAP (Catabolite gene Activator Protein)-DNA complexes (Fig. 2b), in which the DNA structures are bent and the structural differences between DNAs result mainly from the different DNA terminal structures as shown by the arrow in Fig. 2b. If the flanking sequences, which are not in contact with CAP, are not taken into account, the RMSDs are less than 0.2 nm (red squares). In practice especially when atomic energy functions are employed (discussed in detail later), it is necessary to consider the conformational changes when certain bases are "mutated". This can be achieved through energy minimization or other conformational optimization methods[40].

One of the key issues in structure-based TF-binding site prediction is the scoring function for assessing the binding energy or binding affinity between protein and DNA. Earlier efforts that aimed to find simple recognition rules between particular amino acids and specific bases turned out to be futile, as it became obvious that there were no straightforward codes for protein-DNA recognition, although some preferred pairings of amino acids and bases were observed[41,42]. There are two major types of energy functions, the physics-based Molecular Mechanics (MM) force fields and knowledge-based statistical potentials derived from known complex structures, for structure-based TF-binding site prediction. Below we discuss in detail about these two types of energy functions.

2.1 Physics-based energy functions for protein-DNA interaction

Physics-based energy potentials consist of physicochemical components including van der Waals (VDW) forces, electrostatic interactions, solvation energy, and others[40]. There are two types of potentials, grouped based on the approaches for parameter fitting. One uses both experimental and theoretical data from small molecules for parameter training, such as AMBER and CHARMM[43-45]. The other type

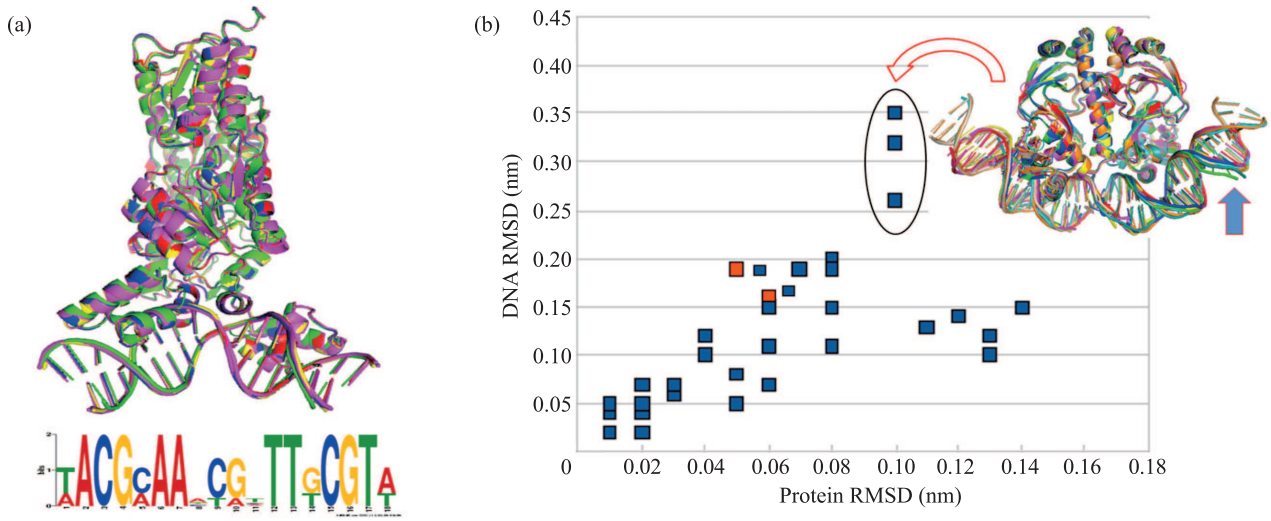

Fig. 2 Transcription factor binding to multiple TF-biding sequences. (a) Superimposed purine repressor-DNA complexes (1bdh, 1jh9, 1qp0, 1qp4, 1qqa, 1zay) and the sequence logo from 6 DNA binding sequences. (b) Relationship between the protein RMSD and DNA RMSD of the aligned CAP-DNA complexes.

utilizes experimental results from macromolecules to derive the parameters. ROSETTA is an example of the second type[46]. This group of physics-based potentials relies on approximations, which tend to abstract away electrons and assume that charge is a fixed property of atoms. A number of studies have applied physics-based energy for studying protein-DNA interactions with some success[36,39,46,47]. There are also extensions of the Schrodinger equation for molecular modeling of more complex systems, which form a class of molecular modeling using Quantum Mechanical (QM) calculation. However, QM modeling is extremely computationally intensive[43].

Besides the general energy terms, such as VDW forces and electrostatic interactions, a special type of interaction, cation-π and π-π interactions, has been investigated in protein-DNA complexes[48-50]. π interactions occur when the negatively charged electron cloud on π systems, generally formed on aromatic compounds, interacts with partially positively charged atoms or cations[51]. Aromatic molecules can undergo aromatic stacking, which is primarily caused by π interactions. Previous studies suggest that the predominant energetic contributions to π interactions are VDW forces and electrostatic forces[51-53]. It is thought that the geometries of interactions between two aromatic structures are governed by electrostatic interactions while the major energy contribution come from VDW forces[52]. Typically π-π interactions between two aromatic compounds can be described as either parallel stacked, T-shaped (or edge to face), or parallel displaced (Fig. 3)[51,52]. Aromatic interactions in the parallel displaced geometry are the strongest among the three types. The electron clouds from both aromatic molecules are involved in attractive electrostatic interactions with the other compound's partially positively charged edges. T-shaped geometries have the second strongest interaction because the partial positive charges from one system's edge have an attractive electrostatic interaction with the electron cloud of the other aromatic molecule. The parallel geometry is the least favorable because it has repulsive forces between the stacked electron clouds of both systems and the stacked partial positive charged edges (Fig. 3)[51,52].

There have been several models for computing or estimating π-π and π-cation interactions[51]. More recent methods use *ab initio* calculations and molecular mechanics to compute these energies[51,52,54]. Some

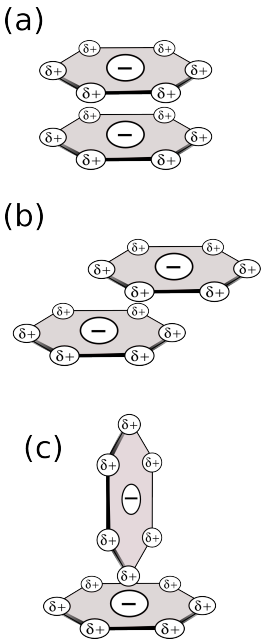

Fig. 3 Geometries of interactions between two aromatic structures. (a), Parallel stacked geometry, the least energetically favorable geometry; (b), Parallel displaced geometry, the most energetically favorable geometry; (c) T-shaped (or edge to face) geometry, more energetically favorable than the parallel stacked geometry but less energetically favorable than the parallel displaced geometry

methods include a simple sum of the VDW and electrostatic energy while others use more sophisticated computational methods that include more detailed molecular mechanics and statistical methods for optimization and correction[49,55]. In proteins, π interactions exist between aromatic amino acids[48,52,56,57]. In DNA, all the bases are aromatic and are involved in base stacking which is essential for DNA stability. Base stacking occurs by π interactions between the staggered bases resulting in displaced parallel π-π interactions. Cation-π interactions between cations and aromatic molecules have been studied to assess their contribution to the structural stability of protein-DNA complexes[49,50,54]. In addition, the π interactions of DNA bases leave the atoms in the major and minor groove of DNA partially charged and thus capable of forming π-cation interactions or π-π interactions with proteins. The landscape of atoms with different charges creates opportunities for specificity in protein-DNA interactions. Bases with partial positively charged atoms in the major groove can form favorable interactions with aromatic residues in the T-shaped (or edge to face) geometry and acidic residues. Bases

with partial negatively charged atoms in the major groove can form interactions with positively charged residues. The impact of these electrostatic interactions on the protein-DNA specificity is thought to be less than hydrogen bonds but may have a bigger impact than VDW forces which play a large role in strength of interaction rather than specificity (unpublished data)[58,59].

2.2 Knowledge-based protein-DNA interaction potentials

Knowledge-based potentials, generally derived from the mean-force theory, are considered more attractive in that they are relatively simple yet with comparable prediction performance to physics-based potentials[60]. These potentials are based on statistical analysis of a set of known, non-redundant protein-DNA complexes. They generally vary in their resolution levels, from residue-based[61-64] to atom-based potentials[43,60,65] and in their distance scales, from distance-independent[61,62] to distance-dependent[60,63-65]. Nevertheless, all knowledge-based potentials are essentially calculated based on the log ratio of the observed frequency over the expected frequency (Eq. (1))

$$e(i,j,r) = -RT \ln \left[\frac{N(i,j,r)_{\text{obs}}}{N(i,j,r)_{\text{exp}}} \right] \tag{1}$$

where R is the gas constant, T is the temperature, $N(i,j,r)_{\text{obs}}$ and $N(i,j,r)_{\text{exp}}$ represent the observed and the expected number between residues (for residue-based) or atoms (for atomic-based) i and j separated by a distance r. While counting $N(i,j,r)_{\text{obs}}$ is easy, $N(i,j,r)_{\text{exp}}$ can be different in different methods[66]. For example, Lu and Skolnick[67] used a simple quasichemical formulation,

$$N(i,j,r)_{\text{exp}} = N(r)\chi_i\chi_j \tag{2}$$

where $N(r)$ is the number of occurrences of an interaction separated by a distance r. χ_i and χ_j represent the mole fraction of residue or atom types i and j. Zhang et al.[60] proposed a more complex, distance-scaled, finite ideal-gas (DFIRE) reference state for protein-DNA interactions:

$$\bar{\mu}(i,j,r) = \begin{cases} -RT \ln \dfrac{N_{\text{obs}}(i,j,r)}{\left(\dfrac{r}{r_{\text{cut}}}\right)^{\alpha} \left(\dfrac{\Delta r}{\Delta r_{\text{cut}}}\right) N_{\text{obs}}(i,j,r)}, & r < r_{\text{cut}}; \\ 0, & r \geqslant r_{\text{cut}} \end{cases} \tag{3}$$

The exponent α is 1.61 on the basis of a state of uniformly distributed points in finite spheres[66]. Further improvements have been made to improve the DFIRE potential by a volume-fraction correction to account for unmixable atom types in proteins and DNA, and a shorter cutoff distance for protein-DNA interactions[68,69]. These potential have been applied to the prediction of TF-binding sites using either the TF-DNA complex structures[68] or unbounded TF structures[70].

In distance-dependent potentials, carefully chosen bin sizes can improve prediction accuracy[65]. Robertson and Varani[65] noted that this is especially problematic for predicting energy from hydrogen bonds as they are affected not only by distance, but also by angle among the participating atoms. It has been shown that two-thirds of the hydrogen bonds between amino acids and bases lead to specific complex interactions[58]. Generally one bin is used for aggregating distances shorter than 0.3 nm, which affects the resolutions for accurately describing the energies. On the other hand, finer bins require large amount of data points to avoid the low count problem. Unfortunately, there are not sufficient non-redundant protein-DNA complexes available for this purpose at this time.

High-resolution, atomic-level potentials can provide the details needed to accurately describe protein-DNA interactions. However, one advantage of residue-level potentials is their capability in addressing the dynamic nature of macromolecules, as they are less sensitive to small conformational changes[71-73]. Another advantage of residue-level potentials lies in their application in protein-DNA docking, in which coarse-grained potentials can have a smooth and less-rugged energy landscape, making it less likely to get trapped in local minima during conformational search[64,74-78]. In 1999, Kono and Sarai[35] developed a simple residue level potential and applied it to TF target site prediction. Two other knowledge-based residue-level potentials were later developed for evaluation of TF-DNA binding affinities as well as for protein-DNA docking predictions[63,64]. The first one is a multi-body potential, which uses DNA tri-nucleotides, called triplets, as an interaction unit to study the interactions between TF and DNA molecules. The triplets could be real nucleotides with explicit positions (native nucleotides) or pseudo-nucleotide placeholders that do not make any structural or energy contribution toward

potential calculation. This multi-body potential, which considers the environment of protein-DNA interactions, can capture the essential physical interactions between protein and DNA, as it shows specific strong hydrogen-bond contributions at short distances, as well as van der Waals repulsion and dispersion attraction[63]. This potential was applied to the prediction of the binding affinity for zinc-finger protein-DNA complexes and showed a high correlation between the predicted binding affinities and the experimental relative binding free energy[63]. Another residue-level potential is an orientation-dependent potential that introduces an angle term, ϕ, to represent the angle between two vectors from the base and the amino acid sidechain respectively[64]. It has been demonstrated that the performance of this orientation-based potential is close to some of the atomic-level potentials, such as vFIRE in binding affinity prediction[64].

2.3 Latest developments

AlQuraishi and McAdams[79] recently proposed a *de novo* statistical potential. While their potential is based on the same input data as the knowledge-based potentials, they do not impose a priori, a specific mathematical formula relating the structural data to the binding energy. They formulated the problem as a compressed sensing problem and used techniques from signal processing. Specifically, a structure is reduced to a vector of energies, each of a different possible type. These microscopic measurements reflect the mesoscopic signal of binding affinity. It has been demonstrated that this *de novo* potential performs well when compared with the knowledge-based potentials[79]. AlQuraishi and McAdams[80] also investigated the statistical potentials by introducing three approaches. The first is the incorporation of binding affinity data into the model. They also varied the parameters to realize the fact that important *in vivo* microenvironments, such as solvent accessibility, differ at different points of the interface. The third is an ensemble-based meta-parameter fitting. Traditionally, meta-parameters are fit by finding a single set of parameters for a given set of data. In their approach, they assign each case a set of the top fitting parameters. Grinter and Zou[81] recently proposed another method to deal with cases that might fail using knowledge-based potentials. They developed a Bayesian method that will predict when a given knowledge-based potential would perform poorly and switch to physics-based potential for better prediction.

3 Conclusions

In this survey we discussed the structure-based approaches for TF binding site prediction. The two key components are a protein-DNA complex model and a scoring function for assessing the protein-DNA interactions. The complex model can be an experimentally determined structure or computationally modeled structure. In evaluating all the possible combinations of the DNA sequences, how to efficiently handle the conformational changes caused by base "mutation", especially DNA backbone flexibility, is a difficult task. In terms of assessing the protein-DNA interaction energy, both physics-based and knowledge-based energy functions have advantages and disadvantages. A combination of both types of potentials may help improve prediction accuracy.

Acknowledgements

We thank Dr. RyangGuk Kim for his help with one of the figures. This work was supported by the National Science Foundation #DBI-0844749 and #DBI-1356459 to JTG.

References

[1] B. Lemon and R. Tjian, Orchestrated response: A symphony of transcription factors for gene control, *Genes Dev*., vol. 14, pp. 2551-2569, 2000.

[2] M. Levine and R. Tjian, Transcription regulation and animal diversity, *Nature*, vol. 424, pp. 147-151, 2003.

[3] M. L. Bulyk, Computational prediction of transcription-factor binding site locations, *Genome Biology*, vol. 5, p. 201, 2003.

[4] A. R. Oliphant, C. J. Brandl, and K. Struhl, Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: Analysis of yeast GCN4 protein, *Mol. Cell Biol*., vol. 9, pp. 2944-2949, 1989.

[5] A. D. Ellington and J. W. Szostak, *In vitro* selection of RNA molecules that bind specific ligands, *Nature*, vol. 346, pp. 818-822, 1990.

[6] C. Tuerk and L. Gold, Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase, *Science*, vol. 249, pp. 505-510, 1990.

[7] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, et al., Genome-wide location and function of DNA binding proteins, *Science*, vol. 290, pp. 2306-2309, 2000.

[8] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, Genome-wide mapping of *in vivo* protein-DNA interactions, *Science*, vol. 316, pp. 1497-1502, 2007.

[9] M. L. Bulyk, E. Gentalen, D. J. Lockhart, and G. M. Church, Quantifying DNA-protein interactions by double-stranded DNA arrays, *Nat. Biotechnol*., vol. 17, pp. 573-577, 1999.

[10] G. D. Stormo and Y. Zhao, Determining the specificity of protein-DNA interactions, *Nat. Rev. Genet*., vol. 11, pp. 751-760, 2010.

[11] M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. 3rd Estep, and M. L. Bulyk, Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities, *Nat. Biotechnol*., vol. 24, pp. 1429-1435, 2006.

[12] Y. Orenstein and R. Shamir, A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data, *Nucleic Acids Research*, vol. 42, no. 8, p.e63, 2014.

[13] T. S. Furey, ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions, *Nat. Rev. Genet.*, vol. 13, pp. 840-852, 2012.

[14] E. G. Wilbanks and M. T. Facciotti, Evaluation of algorithm performance in ChIP-seq peak detection, *PLoS One*, vol. 5, p. e11471, 2010.

[15] S. Georgiev, A. P. Boyle, K. Jayasurya, X. Ding, S. Mukherjee, and U. Ohler, Evidence-ranked motif identification, *Genome Biology*, vol. 11, p. R19, 2010.

[16] Y. Guo, S. Mahony, and D. K. Gifford, High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints, *PLoS Computational Biology*, vol. 8, p. e1002638, 2012.

[17] V. Boeva, D. Surdez, N. Guillon, F. Tirode, A. P. Fejes, O. Delattre, and E. Barillot, De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis, *Nucleic Acids Research*, vol. 38, p. e126, 2010.

[18] I. V. Kulakovskiy, V. A. Boeva, A. V. Favorov, and V. J. Makeev, Deep and wide digging for binding motifs in ChIP-Seq data, *Bioinformatics*, vol. 26, pp. 2622-2623, 2010.

[19] M. Hu, J. Yu, J. M. Taylor, A. M. Chinnaiyan, and Z. S. Qin, On the detection and refinement of transcription factor binding sites using ChIP-Seq data, *Nucleic Acids Research*, vol. 38, pp. 2154-2167, 2010.

[20] P. J. Park, ChIP-seq: Advantages and challenges of a maturing technology, *Nat. Rev. Genet*., vol. 10, pp. 669-680, 2009.

[21] G. D. Stormo, DNA binding sites: Representation and discovery, *Bioinformatics*, vol. 16, pp. 16-23, 2000.

[22] C. E. Lawrence and A. A. Reilly, An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences, *Proteins*, vol. 7, pp. 41-51, 1990.

[23] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment, *Science*, vol. 262, pp. 208-214, 1993.

[24] Y. E. Friedman and M. R. O'Brian, A novel DNA binding site for the ferric uptake regulator (Fur) protein from Bradyrhizobium japonicum, *J. Biol. Chem*., vol. 278, pp. 38395-38401, 2003.

[25] R. D. Dowell, Transcription factor binding variation in the evolution of gene regulation, *Trends Genet*., vol. 26, pp. 468-475, 2010.

[26] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, et al., Diversity and complexity in DNA recognition by transcription factors, *Science*, vol. 324, pp. 1720-1723, 2009.

[27] S. Tuupanen, M. Turunen, R. Lehtonen, O. Hallikas, S. Vanharanta, T. Kivioja, M. Bjorklund, G. Wei, J. Yan, I. Niittymaki, et al., The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling, *Nature Genetics*, vol. 41, pp. 885-890, 2009.

[28] N. C. Seeman, J. M. Rosenberg, and A. Rich, Sequence-specific recognition of double helical nucleic acids by proteins, *Proc. Natl. Acad. Sci. USA*, vol. 73, pp. 804-808, 1976.

[29] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, The Protein Data Bank, *Nucleic Acids Res*., vol. 28, pp. 235-242, 2000.

[30] S. Khorasanizadeh and F. Rastinejad, Nuclear receptor interactions on DNA-response elements, *Trends Biochem. Sci*., vol. 26, pp. 384-390, 2001.

[31] D. L. Bain, A. F. Heneghan, K. D. Connaghan-Jones, and M. T. Miura, Nuclear receptor structure: Implications for function, *Annu. Rev. Physiol*., vol. 69, pp. 201-220, 2007.

[32] M. L. Sierk, Q. Zhao, and F. Rastinejad, DNA deformability as a recognition feature in the reverb response element, *Biochemistry*, vol. 40, pp. 12833-12843, 2001.

[33] L. Yin, N. Wu, J. C. Curtin, M. Qatanani, N. R. Szwergold, R. A. Reid, G. M. Waitt, D. J. Parks, K. H. Pearce, G. B. Wisely, et al., Rev-erbalpha, a heme sensor that coordinates metabolic and circadian pathways, *Science*, vol. 318, pp. 1786-1789, 2007.

[34] R. G. Endres, T. C. Schulthess, and N. S. Wingreen, Toward an atomistic model for predicting transcription factor binding sites, *Proteins*, vol. 57, pp. 262-268, 2004.

[35] H. Kono and A. Sarai, Structure-based prediction of DNA target sites by regulatory proteins, *Proteins*, vol. 35, pp. 114-131, 1999.

[36] A. V. Morozov, J. J. Havranek, D. Baker, and E. D. Siggia, Protein-DNA binding specificity predictions with structural models, *Nucleic Acids Research*, vol. 33, pp. 5781-5798, 2005.

[37] C. W. Garvie and C. Wolberger, Recognition of specific DNA sequences, *Molecular Cell*, vol. 8, pp. 937-946, 2001.

[38] T. Kaplan, N. Friedman, and H. Margalit, *Ab initio* prediction of transcription factor targets using structural knowledge, *PLoS Computational Biology*, vol. 1, p. e1, 2005.

[39] T. W. Siggers and B. Honig, Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry, *Nucleic Acids Res.*, vol. 35, pp. 1085-1097, 2007.

[40] L. A. Liu and P. Bradley, Atomistic modeling of protein-DNA interaction specificity: Progress and applications, *Current Opinion in Structural Biology*, vol. 22, pp. 397-405, 2012.

[41] B. W. Matthews, Protein-DNA interaction. No code for recognition, *Nature*, vol. 335, pp. 294-295, 1988.

[42] C. O. Pabo and L. Nekludova, Geometric analysis and comparison of protein-DNA interfaces: Why is there no simple code for recognition? *J. Mol. Biol*., vol. 301, pp. 597-624, 2000.

[43] J. E. Donald, W. W. Chen, and E. I. Shakhnovich, Energetics of protein-DNA interactions, *Nucleic Acids Res*., vol. 35, pp. 1039-1047, 2007.

[44] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. H. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, et al., Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models, *Accounts of Chemical Research*, vol. 33, pp. 889-897, 2000.

[45] A. D. MacKerell and N. K. Banavali, All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution, *Journal of Computational Chemistry*, vol. 21, pp. 105-120, 2000.

[46] J. J. Havranek, C. M. Duarte, and D. Baker, A simple physical model for the prediction and design of protein-DNA interactions, *Journal of Molecular Biology*, vol. 344, pp. 59-70, 2004.

[47] A. Alibes, A. D. Nadra, F. De Masi, M. L. Bulyk, L. Serrano, and F. Stricher, Using protein design algorithms to understand the molecular basis of disease caused by protein-DNA interactions: The Pax6 example, *Nucleic Acids Research*, vol. 38, pp. 7422-7431, 2010.

[48] J. P. Gallivan and D. A. Dougherty, Cation-pi interactions in structural biology, *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 9459-9464, 1999.

[49] R. Wintjens, J. Lievin, M. Rooman, and E. Buisine, Contribution of cation-pi interactions to the stability of protein-DNA complexes, *Journal of Molecular Biology*, vol. 302, pp. 395-410, 2000.

[50] K. A. Wilson, J. L. Kellie, and S. D. Wetmore, DNA protein pi-interactions in nature: Abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar, *Nucleic Acids Research*, vol. 42, no. 10, pp. 6726-6741, 2014.

[51] C. Hunter and J. Sanders, The nature of pi-pi interactions, *Journal of the American Chemical Society*, vol. 1121, pp. 5525-5534, 1990.

[52] G. B. McGaughey, M. Gagne, and A. K. Rappe, pi-Stacking interactions. Alive and well in proteins, *J. Biol. Chem*., vol. 273, pp. 15458-15463, 1998.

[53] N. Allinger and J. H. Lii, Benzene, aromatic rings, van der Waals molecules, and crystals of aromatic molecules in molecular mechanics (MM3), *Journal of Computational Chemistry*, vol. 8, pp. 1146-1153, 1987.

[54] M. M. Gromiha, C. Santhosh, and M. Suwa, Influence of cation-pi interactions in protein-DNA complexes, *Polymer*, vol. 45, p. 633, 2004.

[55] P. Mignon, S. Loverix, J. Steyaert, and P. Geerlings, Influence of the pi-pi interaction on the hydrogen bonding capacity of stacked DNA/RNA bases, *Nucleic Acids Research*, vol. 33, pp. 1779-1789, 2005.

[56] Z. Shi, C. A. Olson, and N. R. Kallenbach, Cationpi interaction in model alpha-helical peptides, *Journal of the American Chemical Society*, vol. 124, pp. 3284-3291, 2002.

[57] T. P. Burghardt, N. Juranic, S. Macura, and K. Ajtai, Cation-pi interaction in a folded polypeptide, *Biopolymers*, vol. 63, pp. 261-272, 2002.

[58] N. M. Luscombe, R. A. Laskowski, and J. M. Thornton, Amino acid-base interactions: A three-dimensional analysis of protein-DNA interactions at an atomic level, *Nucleic Acids Res*., vol. 29, pp. 2860-2874, 2001.

[59] C. M. Baker and G. H. Grant, Role of aromatic amino acids in protein-nucleic acid recognition, *Biopolymers*, vol. 85, pp. 456-470, 2007.

[60] C. Zhang, S. Liu, Q. Zhu, and Y. Zhou, A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes, *J. Med. Chem*., vol. 48, pp. 2325-2335, 2005.

[61] Y. Mandel-Gutfreund and H. Margalit, Quantitative parameters for amino acid-base interaction: Implications for prediction of protein-DNA binding sites, *Nucleic Acids Research*, vol. 26, pp. 2306-2312, 1998.

[62] P. Aloy, G. Moont, H. A. Gabb, E. Querol, F. X. Aviles, and M. J. Sternberg, Modelling repressor proteins docking to DNA, *Proteins*, vol. 33, pp. 535-549, 1998.

[63] Z. Liu, F. Mao, J. T. Guo, B. Yan, P. Wang, Y. Qu, and Y. Xu, Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential, *Nucleic Acids Res*., vol. 33, pp. 546-558, 2005.

[64] T. Takeda, R. I. Corona, and J. T. Guo, A knowledge-based orientation potential for transcription factor-DNA docking, *Bioinformatics*, vol. 29, no. 3, pp. 322-330, 2013.

[65] T. A. Robertson and G. Varani, An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure, *Proteins*, vol. 66, pp. 359-374, 2007.

[66] H. Zhou and Y. Zhou, Distance-scaled, finite ideal gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction, *Protein Sci*., vol. 11, pp. 2714-2726, 2002.

[67] H. Lu and J. Skolnick, A distance-dependent atomic knowledge-based potential for improved protein structure selection, *Proteins*, vol. 44, pp. 223-232, 2001.

[68] B. Xu, Y. Yang, H. Liang, and Y. Zhou, An allatom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles, *Proteins*, vol. 76, pp. 718-730, 2009.

[69] H. Zhao, Y. Yang, and Y. Zhou, Structure based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function, *Bioinformatics*, vol. 26, pp. 1857-1863, 2010.

[70] C. Y. Chen, T. Y. Chien, C. K. Lin, C. W. Lin, Y. Z. Weng, and D. T. Chang, Predicting target DNA sequences of DNA-binding proteins based on unbound structure, *PLoS One*, vol. 7, p. e30446, 2012.

[71] P. Bradley, L. Malmstrom, B. Qian, J. Schonbrun, D. Chivian, D. E. Kim, J. Meiler, K. M. Misura, and D. Baker, Free modeling with Rosetta in CASP6, *Proteins*, vol. 61, no. Suppl 7, pp. 128-134, 2005.
[72] S. M. Gopal, S. Mukherjee, Y. M. Cheng, and M. Feig, PRIMO/PRIMONA: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy, *Proteins*, vol. 78, no. 1266-1281, 2010.
[73] T. Vreven, H. Hwang, and Z. Weng, Integrating atom-based and residue-based scoring functions for protein-protein docking, *Protein Sci*., vol. 20, pp. 1576-1586, 2011.
[74] G. S. Ayton, W. G. Noid, and G. A. Voth, Multiscale modeling of biomolecular systems: In serial and in parallel, *Current Opinion in Structural Biology*, vol. 17, pp. 192-198, 2007.
[75] P. Poulain, A. Saladin, B. Hartmann, and C. Prevost, Insights on protein-DNA recognition by coarse grain modeling, *Journal of Computational Chemistry*, vol. 29, pp. 2582-2592, 2008.
[76] S. C. Flores, J. Bernauer, S. Shin, R. Zhou, and X. Huang, Multiscale modeling of macromolecular biosystems, *Briefings in Bioinformatics*, vol. 13, pp. 395-405, 2012.
[77] Z. Liu, J. T. Guo, T. Li, and Y. Xu, Structure based prediction of transcription factor binding sites using a protein-DNA docking approach, *Proteins*, vol. 72, pp. 1114-1124, 2008.
[78] J. Wu, B. Hong, T. Takeda, and J. T. Guo, High performance transcription factor-DNA docking with GPU computing, *Proteome Sci.*, vol. 10, no. Suppl 1, p. S17, 2012.
[79] M. AlQuraishi and H. H. McAdams, Direct inference of protein-DNA interactions using compressed sensing methods, *Proc. Natl. Acad. Sci. USA*, vol. 108, pp. 14819-14824, 2011.
[80] M. AlQuraishi and H. H. McAdams, Three enhancements to the inference of statistical protein-DNA potentials, *Proteins*, vol. 81, pp. 426-442, 2013.
[81] S. Z. Grinter and X. Zou, A Bayesian statistical approach of improving knowledge-based scoring functions for protein-ligand interactions, *Journal of Computational Chemistry*, vol. 35, pp. 932-943, 2014.

Jun-tao Guo received his PhD degree in molecular and cellular biochemistry and a master's degree in computer science from the University of Kentucky in 2001 and 2002, respectively. He is currently an associate professor in the Department of Bioinformatics and Genomics at the University of North Carolina at Charlotte. His research interests are in the broad area of structural bioinformatics.

Shane Lofgren received his BS degree in economics, Magna Cum Laude, from the University of Oregon in 2011. After a brief career in finance and education, he transitioned into the bioinformatics field as a research assistant in the lab of Dr. Jun-tao Guo at the University of North Carolina at Charlotte, working on improving prediction of transcription factor binding sites. He is currently working in the lab of Julien Sage at the Stanford University Medical School.

Alvin Farrel received his MS degree in biochemistry in 2009 from Loma Linda University, and his BS degrees in both biology and computing from Andrews University in 2006. He is currently a PhD student in the Department of Bioinformatics and Genomics at the University of North Carolina at Charlotte working in the lab of Dr. Jun-tao Guo. His research interests include studying the mechanisms of protein-DNA interactions and the effects of genetic mutations using structural bioinformatics and computational biophysics.