ALGORITHMS FOR STRUCTURE BASED-PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES

by

Alvin Lemuel Farrel

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2016

Approved by:

_____
Dr. Jun-tao Guo

_____
Dr. Anthony Fodor

_____
Dr. Donald Jacobs

_____
Dr. Xinghua Shi

_____
Dr. Bao-Hua Song

ABSTRACT

ALVIN LEMUEL FARREL. Algorithms for structure based-prediction of transcription factor binding sites. (Under the direction of DR. JUN-TAO GUO)

Transcription factors (TFs) regulate gene expression through binding to specific target DNA sites. Accurate annotation of transcription factor binding sites (TFBSs) at genome scale represents an essential step toward our understanding of gene regulation networks. In this dissertation, we present a structure-based method for computational prediction of TFBSs using a novel, integrative energy (IE) function and an efficient pentamer algorithm. The integrative energy function combines a multibody (MB) knowledge-based potential and atomic energy terms (hydrogen bond and π-interaction) that might not be accurately captured by the knowledge-based potential owing to the mean force nature and low count problem. A pentamer algorithm is developed to address the computational complexity issue due to the exponential increase of the number of DNA sequences for longer binding sites that need to be evaluated. Test results show that the new energy function improves the prediction accuracy over the knowledge-based, statistical potentials based on a non-redundant dataset that consists of TF-DNA complexes from 12 different families. The pentamer algorithm improves TFBS prediction accuracy while greatly reducing the time complexity for long binding sites.

DEDICATION

For my father, Andrew Joel Farrel, who gave me unfailing support and encouragement in all my academic goals; and my mother, Genevieve Estelle Farrel, who was my first teacher and taught me that most learning happens outside of the classroom.

ACKNOWLDGEMENTS

First and foremost I would like to thank my advisor, Dr. Jun-tao Guo for his invaluable guidance, excellent teaching mentorship, and unceasing patience over the years. Sincere gratitude to my committee members, Dr. Anthony Fodor, Dr. Donald Jacobs, Dr. Xinghua Shi, Dr. Bao-Hua Song, and Dr. Dennis Livesay for their valuable support and suggestions. I would like to thank all the members of the Guo Lab, past and present, for opportunity to share ideas and work in such a friendly environment. Finally I would like to thank the GAANN Computing Scholars program at UNCC, FASEB MARC program, and the generous financial support from my advisor, Dr. Jun-tao Guo during my Ph.D. studentship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

TF	Transcription factor
TFBS	Transcription factor binding site
ChIP	Chromatin ImmunoPrecipitation
SELEX	Systematic Evolution of Ligands by Exponential enrichment
PBM	Protein Binding Mircoarrays
PWM	Position Weight Matrix
EM	Expectation-Maximization
DNA	Deoxyribonucleic acid
PDB	Protein data bank
VDW	van der Waals
DBP	DNA-binding protein
IE	Integrative energy
MB	Multibody
AKL	Averaged Kullback-Leibler
PFM	Position frequency matrix
IC	Information content
PCC	Pearson correlation coefficient
Bp	Base pair
Exd	Extradenticle
Ubx	Ultrabithorax

CHAPTER 1: INTRODUCTION

1.1 Background

Regulation of gene expression is critical for proper cellular function. Discovering gene regulatory networks embedded in the genome and fully understanding the mechanism of sequence-specific protein-DNA interactions remains a key challenge in post-genomic bioinformatics. Transcription factors (TFs) regulate gene expression by interacting with specific DNA sequences called transcription factor binding sites (TFBSs) and identification of TFBSs on a genomic scale represents a crucial step in deciphering transcription regulatory networks and in genomic annotation (Lemon and Tjian, 2000; Levine and Tjian, 2003). Knowledge of protein-DNA interactions at the structural-level can provide insights into the mechanisms of gene regulation. In addition, understanding the mechanisms of protein-DNA interactions can also help engineer novel TF specificity, design new therapeutic drugs, and elucidate pathologies of genetic disorders with altered gene expressions. Mutations in TFs can be deleterious and lead to diseases (Alibes, et al., 2010; D'Elia, et al., 2001; Muller and Vousden, 2013) or they can be advantageous evolutionary adaptations (Luscombe and Thornton, 2002). Changes in binding affinities of TFs to specific DNA sequences can also affect how the transcription factors interact with other regulatory proteins, leading to phenotypic consequences potentially causing evolutionary changes (Borneman, et al., 2007; Schmidt, et al., 2010; Wray, 2007).

1.2 Experimental Identification of Transcription Factor Binding Sites

DNase I footprinting and gel-mobility shift assay represent two traditional experimental methods for determining binding sites of TFs. However, these methods are time-consuming and unsuitable for large-scale studies (Bulyk, 2003). High-throughput experimental methods such as Systematic Evolution of Ligands by Exponential enrichment (SELEX)(Ellington and Szostak, 1990; Oliphant, et al., 1989; Tuerk and Gold, 1990), Chromatin ImmunoPrecipitation (ChIP)-based technologies such as ChIP-chip (Ren, et al., 2000) and ChIP-seq (Johnson, et al., 2007), and Protein Binding Mircoarrays (PBM) (Bulyk, et al., 1999), are more efficient methods for determining TFBSs in large scale studies. SELEX and PBM are *in vitro* experimental methods. In a recent comparative study, SELEX and PBM derived TFBSs were in agreement for most transcription factors (Orenstein and Shamir, 2014). These *in vitro* methods directly measure TF-DNA specificity, represented by a Position Weight Matrix (PWM) (Stormo and Zhao, 2010). However, it's not a true representation of DNA binding within a cellular environment, which may include cofactors, epigenetic factors, and other regulatory machinery. The ChIP-based technologies are *in vivo* high-throughput methods where TFBSs are identified by microarrays or parallel DNA sequencing technologies after antibodies are used to isolate TFs bound to their binding sequences. The DNA is sequenced, aligned, and used to generate a binding motif represented by a PWM (Boeva, et al., 2010; Georgiev, et al., 2010; Guo, et al., 2012; Hu, et al., 2010; Johnson, et al., 2007; Kulakovskiy, et al., 2010; Ren, et al., 2000). While these experimental techniques for determining TFBSs are fairly accurate, they require time and resources. An accurate

computational method for determining TFBSs of native and mutated TFs can complement the experimental methods and save time and resources.

1.3 Computational methods for prediction of transcription factor binding sites

With the rapidly increasing genomic data becoming available, very effective sequence-based methods for TFBSs predictions have been developed (Stormo, 2000). A number of algorithms use promoter sequences for TFBS prediction including Expectation-Maximization (EM) (Lawrence and Reilly, 1990) and Gibbs sampling (Lawrence, et al., 1993). Some algorithms now include phylogenetic footprinting or orthologous sequences into the conventional prediction methods (Bulyk, 2003). Sequence-based methods have also been combined with ChIP-based methods for the identification of TFBSs (Furey, 2012). One issue of sequence-based methods is that they tend to generate high number of false positives. This can occur when the binding signal is weak or the TF's DNA-binding site is significantly different from the consensus sequence. In addition, some TFs also bind to multiple distinct sequence motifs, adding more complications to TF binding predictions (Badis, et al., 2009; Dowell, 2010; Friedman and O'Brian, 2003).

Structure-based prediction methods, on the other hand, focus on protein-DNA interactions rather than sequence conservation. Therefore, they are not constrained by sequence information. These prediction methods mimic real binding and recognition events because specific binding between a TF and its binding sites in the cell relies on their biophysical interactions. Sequence-based methods and experimental technologies can identify the genome binding site locations and binding site sequences. Structure-based methods can also explain why and how these proteins bind at these locations and

sequences because they provide insight into the mechanisms of TF-DNA interactions. Understanding these mechanisms, how mutations affect these mechanisms, and the downstream effects on gene expression can contribute to our understanding of diseases and lead to rational design of therapeutic agents.

Although research on protein-DNA recognition began in the 1970s (Seeman, et al., 1976), structure based methods weren't developed until years ago when the high-resolution protein-DNA complex structures became available in the protein data bank (PDB) (Berman, et al., 2000). The basic workflow of structure-based prediction of TFBSs starts with a TF-DNA complex structure. A scoring function is used to calculate the interaction energy between the TF and every permutation of the DNA sequence in the structure. The energy scores and their corresponding DNA sequences are then used to generate a binding motif (Figure 1.1) (Liu and Bradley, 2012). The binding motif can be generated by aligning the top scoring sequences to generate a PWM (Stormo and Zhao, 2010) or using innovative alignment-independent statistical approaches to determine a representative PWM of the binding motif (Newburger and Bulyk, 2009).

One key component in structure-based TFBS prediction is the scoring function for evaluating binding affinity or binding energy between proteins and DNA. While there is no simple recognition code or pairing between amino acids and DNA bases, it has been found that some amino acids has preferred pairings with some DNA bases (Matthews, 1988; Pabo and Nekludova, 2000). There are two general types of energy functions in studying protein-DNA interactions: the physics-based molecular mechanics force fields and the knowledge-based statistical potentials. Both the physics-based energy functions

and the knowledge-based protein-DNA interaction potentials have their distinct advantages as well as limitations.

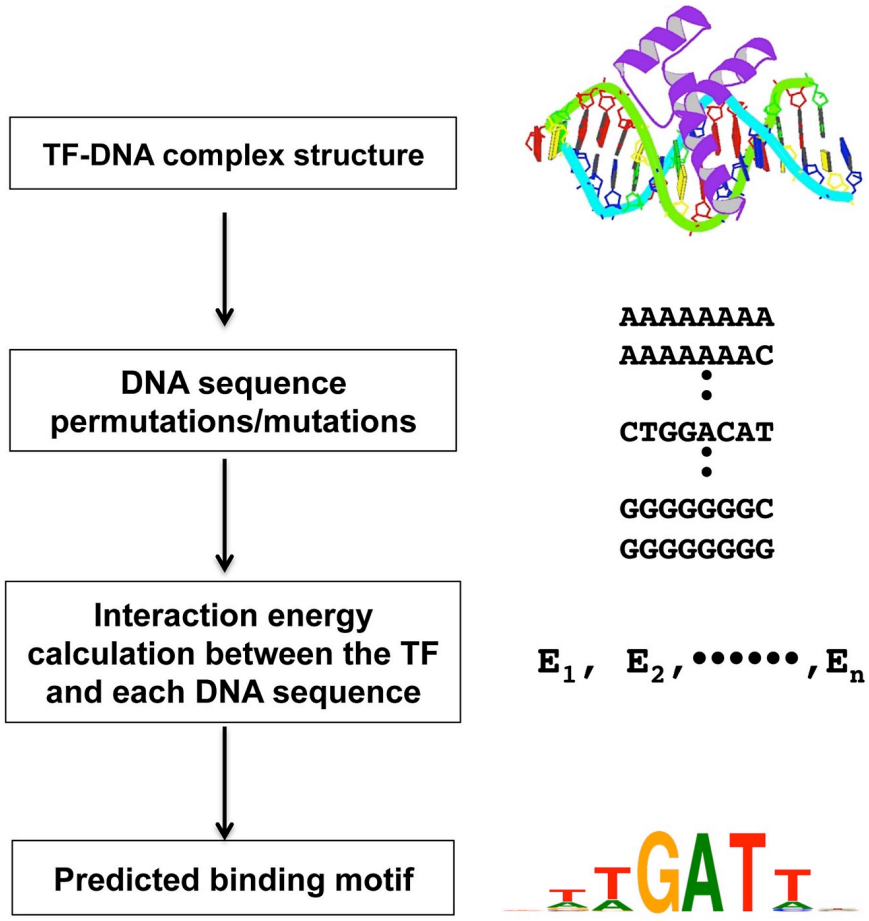

Figure 1.1: Flowchart of structure-based transcription factor binding site prediction.

1.4 Physics-based Energy Functions for Protein-DNA Interactions

Physics based energy functions consist of physicochemical interactions including electrostatic interactions, van der Waals (VDW) forces, solvation energy, and others (Liu and Bradley, 2012). Some methods use both experimental and theoretical data from small molecules for parameter training such as AMBER and CHARMM (Donald, et al., 2007; Kollman, et al., 2000; MacKerell and Banavali, 2000). Other methods use experimental

results from macromolecules to derive parameters, such as ROSETTA (Havranek, et al., 2004). These physics-based potentials rely on approximations and often assume fixed charges. They have been applied to protein-DNA interaction studies with some success (Alibes, et al., 2010; Havranek, et al., 2004; Morozov, et al., 2005; Siggers and Honig, 2007). Other approaches use quantum mechanical calculations to extend the Schrodinger equation for molecular modeling of complex systems, however, these approaches are very computationally intensive (Donald, et al., 2007).

Besides the general terms such as VDW and electrostatic interactions, which include hydrogen bonds, π-cation and π-π interactions have been studied in protein-DNA recognition. It was previously thought that these interactions have a primary role of establishing the stability of the protein-DNA complexes but new data suggests that these interactions may have a bigger role in protein-DNA recognition (Baker and Grant, 2007; Corona and Guo, 2016; Luscombe, et al., 2001). Understanding the intricate mechanisms of these molecular interactions may help improve the accuracy of structure-based TFBS prediction.

The impact of water mediated hydrogen bonds and their importance in protein-DNA recognition prediction is highly debated (Jiang, et al., 2005; Li and Bradley, 2013; Li, et al., 2011; Schneider, et al., 1992; Tucker-Kellogg, et al., 1997; van Dijk, et al., 2013). The effects of water on the energetics of TF-DNA complex can be used at different details. Some methods neglect the interactions completely while including an implicit solvation potential (Morozov, et al., 2005). On the other end of the spectrum, the water molecules are included as part of the complex structure when the molecular mechanics energy functions are applied (Beierlein, et al., 2011; Liu and Bader, 2009;

Seeliger, et al., 2011). The including of explicit water molecules typically improves predictions but the improvement tends to be modest (Li and Bradley, 2013; van Dijk, et al., 2013).

Protein-DNA complexes are intrinsically dynamic. A protein-DNA structure in the PDB represents either a snapshot of one of the many possible conformations, or a derived averaged structure. Protein-DNA complexes change their conformations due to protein backbone flexibility or at a much more detailed level due to amino acid side chain flexibility. Different conformations may result in different TFBS predictions because physics-based predictions depend on energies of specific distance dependent interactions in a crystal structure's conformation. Flexibility of the amino acid side chains in the protein-DNA interface, where the distance and conformation dependent details matter most, need to be addressed in a manner that is not too computationally expensive but can improve the accuracy of physics-based predictions.

1.5 Knowledge-based Protein-DNA Interaction Potentials

Knowledge-based potentials are based on statistical analysis of a set of known, non-redundant protein-DNA complexes. The potentials are generally derived from the mean force theory and are often preferred because they are relatively simple and less computationally expensive while producing comparable predictions to physics-based predictions. Knowledge-based potentials vary in resolution from residue-based (Aloy, et al., 1998; Liu, et al., 2005; Mandel-Gutfreund and Margalit, 1998; Takeda, et al., 2013) to atom-based potentials (Donald, et al., 2007; Robertson and Varani, 2007; Zhang, et al., 2005). They also vary in their distance scales from distance independent (Aloy, et al., 1998; Mandel-Gutfreund and Margalit, 1998) to distance dependent (Liu, et al., 2005;

Robertson and Varani, 2007; Takeda, et al., 2013; Zhang, et al., 2005). All knowledge-based energy functions are calculated using a log ratio of the observed frequencies over the expected frequencies.

$$e(i,j,r) = -RT \ln\left[\frac{N(i,j,r)_{obs}}{N(i,j,r)_{\exp}}\right] \tag{1.1}$$

where R is the gas constant, T is the temperature, $N(i,j,r)_{obs}$ and $N(i,j,r)_{exp}$ 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 knowledge-based potentials can produce relatively good predictions, the mean force nature affects accuracy in capturing the hydrogen bond interaction as they are affected by distance as well as the angles of the atoms involved in the hydrogen bond potential (Robertson and Varani, 2007). This is important because about two-thirds of the hydrogen bonds between amino acids and bases lead to specific complex interactions (Luscombe, et al., 2001). Carefully choosing a bins size can help this issue and improve prediction accuracy (Burghardt, et al., 2002). One bin is typically used for combining distances less than 0.3 nm, which add noise in describing the energies. Alternatively, finer bins require many data points to avoid the low count problem, however, there aren't sufficient non-redundant high-resolution protein-DNA complexes available in the PDB for this purpose.

High-resolution all-atom based potentials can provide detailed atomic positions for a more accurate calculation of the energies present in a protein-DNA complex. However, these high-resolution potentials are very sensitive to protein backbone, side chains, and docking conformations, which could be a problem due to the dynamic nature

of macromolecules (Bradley, et al., 2005; Gopal, et al., 2010; Vreven, et al., 2011). Therefore, residue level potentials are considered advantageous because they are not as sensitive to slight changes in the protein-DNA complex conformations. Also, in protein-DNA docking studies, residue level potentials produce less rugged energy landscapes, which make it less likely for complexes to get stuck in local energy minima during a conformational search (Ayton, et al., 2007; Flores, et al., 2012; Liu, et al., 2008; Poulain, et al., 2008; Takeda, et al., 2013; Wu, et al., 2012).

Both atomic level and residue-level energy potentials have been applied to predict TFBSs. Xu *et al*. developed an energy function that uses structure-based templates for DNA binding sites, which lead to increased accuracy over their previous all-atom based potential vcFIRE (Xu, et al., 2013; Xu, et al., 2009). Two of the residue level knowledge-based potentials with comparable results to all-atom based potentials are a multibody energy function by Liu *et al*. (Liu, et al., 2005) and an orientation-dependent potential by Takeda *et al*. (Takeda, et al., 2013). The multibody potential uses tri-nucleotides, called triplets, as an interaction unit to score interactions between the TF's amino acids and the DNA bases. The multibody potential considers the environment of the TF-DNA interactions and captures the essential physical interactions between the TF and the DNA including hydrogen bond interactions within short distances and van der Waals interactions. The orientation-dependent potential introduces an angle term to represent the angle between two vectors from the bases and the amino acids' sidechains to compensate for the angle term lost in capturing hydrogen bond energy in other distant-dependent knowledge based potentials. The predictions of the orientation potential are

close to those of some atomic-level energy potentials such as vFIRE (Takeda, et al., 2013).

Atomic resolution statistical potentials are sometimes used to predict protein-DNA binding specificity because it is thought that residue level potentials do not always have sufficient resolution to make accurate predictions, however they are well suited for protein-DNA docking studies (Joyce, et al., 2015). Recent residue-level potentials have proven to work just as well as atomic-level predictions (Takeda, et al., 2013). Furthermore, the atomic potentials accuracy depends heavily on the conformation of the complex and the amino acid side chains. Overall, statistical potentials require much less computational power than physics-based potentials. While neither the atomic-level nor the residue-level method is ideal in determining TF recognition sequences, their advantages can be combined for a better description of TF-DNA interactions.

1.6 Evaluating Binding Site Prediction Performance.

The quality of a transcription factor binding site prediction can be evaluated by comparing it to an experimentally derived TFBS. These experimentally annotated TFBSs, determined using methods discussed in 1.2, are typically stored in two databases, JASPAR (Mathelier, et al., 2016) and UniPROBE (Newburger and Bulyk, 2009). UniPROBE only contains TFBSs generated by PBM, while JASPAR includes TFBSs identified by various *in vitro* and *in vivo* methods including PBM, SELEX, and ChIP-based methods. The PWMs representing TF-DNA specificity in these databases can be directly compared to the PWMs generated by TFBS prediction algorithms using various statistical measures such as Pearson correlation coefficient, average log-likelihood ratio, Pearson chi-squared test, Fisher-Irwin exact test, Kullback-Leibler divergence, Euclidean

distance, and the Sandelin –Wasserman similarity function (Gupta, et al., 2007). These statistical methods can be built on to better evaluate the similarity between binding motifs. For example, asymmetrical methods, such as Kullback-Leibler divergence, can be symmetrized by averaging the two Kullback-Leibler distances (Seghouane and Amari, 2007). Also, information content from the PWMs being evaluated can be used to weight the importance of the various matrix columns, representing base positions in the binding motif, by conservation when performing Pearson correlation coefficient. (Persikov and Singh, 2014).

1.7 Dissertation Goals

There are two major issues in structure-based prediction of transcription factor binding sites: an accurate scoring function to assess the protein-DNA interactions and the availability of TF-DNA complex models. Since simple base recognition codes for TF-DNA recognition do not exist (Benos, et al., 2002), many interaction models, including biophysical and statistical approaches, have been developed for studying specific protein-DNA recognitions (Benos, et al., 2002; Desjarlais and Berg, 1992; Harr, et al., 1983; Luscombe, et al., 2001; Mandel-Gutfreund and Margalit, 1998; Mandel-Gutfreund, et al., 1995; Mulligan, et al., 1984; Staden, 1984; Suzuki, et al., 1995; Suzuki and Yagi, 1994; von Hippel and Berg, 1989). Each model has its advantages and limitations in TFBS prediction (Benos, et al.). While physics-based energy is more accurate in describing protein-DNA interactions, it is very computationally expensive. Moreover, a physics-based energy may not fully capture the essence of protein-DNA recognition. Knowledge based statistical potentials include residue-level and atom-level energy functions (Liu, et al., 2005; Takeda, et al., 2013; Xu, et al., 2009). Though they are relatively simple when

compared to the physics-based energy functions, they can have comparable prediction performance to physics-based potentials.

Another issue with structure-based transcription factor binding site prediction lies in its requirement of TF-DNA complex models. Due to technical limitations, the number of TF-DNA complex structures in the PDB is rather small when compared to the number of transcription factors in genomes of all three domains in the tree of life. This raises the question, assuming we have a near-perfect scoring function for assessing the TF-DNA interactions, how can we expand application of structure-based approaches for TFBS prediction? Protein structure modeling is a cost effective method to complement the experimental approaches, especially homology modeling techniques, which can offer relatively high accuracy structural models.

In this dissertation project I have addressed two questions. Can we develop novel energy functions and algorithms for efficient and accurate prediction of TF binding sites? Can we expand structure-based TFBS prediction models to cases without known TF-DNA complex structures in the PDB? There are three specific aims in this dissertation research: 1) develop an integrative energy function for structure-based prediction of transcription factor binding specificity; 2) develop an efficient algorithm to improve the structure-based TFBS prediction with longer binding sites, especially for binding sites of TF dimers or tetramers; and 3) develop a homology modeling-based approach to expand the application of the above methods to transcription factors without known TF-DNA complex structures.

CHAPTER 2: STRUCTURE-BASED PREDICTION OF TRANSCRIPTION FACTOR BINDING SPECIFICTY USING AN INTEGRATIVE ENERGY FUNCTION

2.1 Introduction

As described in Chapter 1, structure-based TFBS prediction methods focus on physical protein-DNA interactions by mimicking real binding and recognition events as specific binding between a TF and its binding sites in the cell relies on biophysical interactions. One of the key issues in structure-based TFBS prediction is accurate assessment of the binding affinity or binding energy between proteins and DNA. Of the two major types of energy functions, the physics-based energies can accurately describe protein-DNA interactions but are computationally expensive, while the knowledge-based potentials are computationally efficient with reasonable accuracy.

Knowledge-based potentials, derived from statistical analysis of known TF-DNA complex structures, are simple to use. However, these statistical potentials may be limited by two factors. One is the mean force nature of the knowledge-based potentials. For example, amino acids arginine and lysine can contribute to both specific interactions with DNA through hydrogen bonding and non-specific interactions through electrostatic interaction with the DNA backbone. They form hydrogen bonds in the major groove when highly specific DNA-binding proteins (DBPs) interact with DNA. Contrastingly, they form hydrogen bonds predominantly in the minor grooves when non-specific DBPs interact with DNA (Corona and Guo, 2016). Though the hydrogen bonds important for DNA sequence recognition are implicitly captured in knowledge-based potentials, they

are “averaged” with the non-specific interactions. The accuracy of the knowledge-based potentials is also affected by the low count problem. More recent studies have suggested that π-interactions between aromatic amino acids and DNA bases are more prevalent than previously thought, though very little is known about their critical role in specific protein-DNA binding (Wilson, et al., 2014; Wilson and Wetmore, 2015). Through comparative analysis, we recently found that tyrosine and histidine are enriched in interacting with DNA bases in highly specific DNA-binding proteins. We hypothesize that π-interactions between aromatic residues and DNA bases contribute to TF-DNA binding specificity. These interactions may not be accurately captured in knowledge-based potentials, as the number of aromatic residues that are involved in protein-DNA interactions is relatively low.

Here we propose a novel, integrative energy (IE) function that combines a knowledge-based multibody potential with hydrogen bond and π-interaction information for prediction of TFBSs and apply it to the binding site prediction of non-redundant datasets of transcription factors. The results show that TFBS prediction using our new integrative energy function improves accuracy when compared to other residue-level and atomic-level knowledge-based potential.

2.2 Methods

2.2.1 Integrative Energy Function

The integrative energy function consists of a knowledge-based multibody (MB) potential (Liu, et al., 2005; Takeda, et al., 2013) and two physics-based terms, hydrogen bond energy and electrostatic potentials from π interactions:

$$E_{Total} = W_{MB}E_{MB} + W_{HB}E_{HB} + W_{\pi}E_{\pi} \qquad (2.1)$$

where E_{Total} is the total energy, E_{MB}, E_{HB}, and E_{π} represent the normalized multibody energy, hydrogen bond energy, and π interaction energy respectively, and W_{MB}, W_{HB}, and W_{π} are weights for each term. Since there are only a limited number of non-redundant TF-DNA complexes with known TFBSs, we were unable to use training methods to get an optimal set of weights. We used 1, 1, and 0.5 for W_{MB}, W_{HB}, and W_{π} respectively in this study. The hydrogen bond energy has equal weight to the knowledge-based potential due to its important contribution to protein-DNA binding specificity (Luscombe, et al., 2001). The weight for π-interaction is half the weight of the multibody and hydrogen bond terms because it is less abundant and its role in specific protein-DNA interaction is not as well defined as the hydrogen bonds.

2.2.2 Knowledge-based, Multibody Statistical Potential

We have previously developed two residue-level knowledge-based potentials, a multibody potential and an orientation potential, for assessing protein-DNA interactions in transcription factor binding site prediction and protein-DNA docking (Liu, et al., 2005; Takeda, et al., 2013). The multibody potential utilizes structural environment for accurate assessment while the orientation potential uses both distance and angle information to better capture hydrogen bond information implicitly. Since we propose an explicit hydrogen bond term in our new integrative energy function to capture the key hydrogen bond interactions, we chose the multibody potential over the orientation potential to minimize the overlap between the hydrogen bond energy and the orientation potential while taking the structural environment into consideration. In addition, we found that even though the orientation potential performs better than the multibody potential for TF-DNA docking (Liu, et al., 2005; Takeda, et al., 2013), the multibody potential predicts

TF-DNA binding motifs better than the orientation potential possibly due to the capture of interaction context as structure-based prediction of TFBSs and protein-DNA docking are two different computational problems (data not shown). The multibody potential uses the distance between an amino acid's β-carbon and the geometric center of a nucleotide triplet. The position of a nucleotide is represented by the N_1 atom in pyrimidines or the N_9 atom in purines (Liu, et al., 2005; Takeda, et al., 2013).

2.2.3 Hydrogen Bond Energy

The hydrogen bond energy is calculated using the model described by Thorpe *et al.* (Eq. 2), which was adapted from Dahiyat *et al*. (Dahiyat, et al., 1997; Thorpe, et al., 2001).

$$E_{HB} = V_0 \left\{ 5 \left(\frac{d_0}{d}\right)^{12} - 6 \left(\frac{d_0}{d}\right)^{10} \right\} F(\theta, \phi, \varphi) \qquad (2.2)$$

where d_0 (2.8 Å) and V_0 (8 *kcal/mol*) are the hydrogen-bond equilibrium distance and well-depth respectively, and d is the distance between the donor and the acceptor. The angle function, F, varies depending on the hybridization state of the acceptor and donor atoms (Dahiyat, et al., 1997; Thorpe, et al., 2001). We used FIRST (Jacobs, et al., 2001), which implements Equation 2.2, to calculate the hydrogen bond energy between amino acids and nucleotides in the protein-DNA complexes (Abecasis, et al.).

2.2.4 π-interaction Energy

π-interactions typically exist between aromatic compounds and cations, partially charged atoms, or other aromatic compounds. These interactions consist of VDW forces and electrostatic interactions (Gromiha, et al., 2004; Luscombe, et al., 2001; McGaughey, et al., 1998; Wintjens, et al., 2000). In aromatic compounds, π-π interactions occur when the partially positive charges on the edges of an aromatic molecule interact with the

negatively charged electron cloud of another aromatic compound. These interactions can be in a parallel stacked, parallel displaced, or edge to face conformation (Figure 2.1). It appears that the VDW forces do not have a major impact on DNA-binding specificity of TFs, but they assist greatly in protein-DNA complex stability (2008; Gromiha, et al., 2004; Wintjens, et al., 2000). However, the electrostatic charges on the edges of the bases, especially in the major groove, are different in the four DNA bases. Figure 2.2 shows the electronic landscape of the atoms on each base at the resonant state assuming a physiological pH. The partially charged edges of the bases exposed in the major groove (Table 2.1) were determined using MarvinSketch 6.1.4, a software package from Chemaxon (Marvin6.1.4).

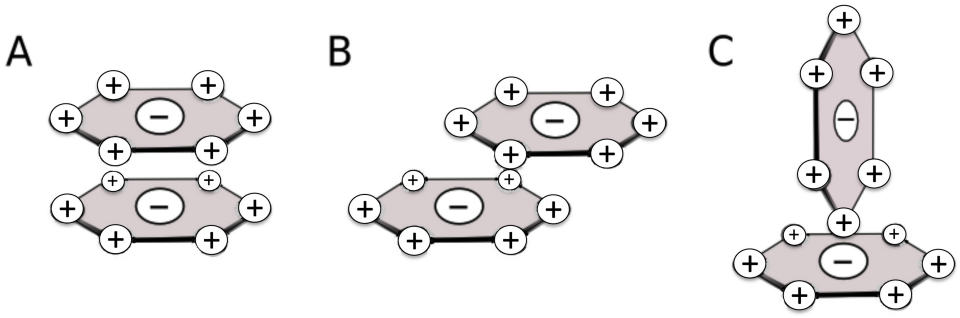

Figure 2.1: Geometries of π-Interactions between aromatic structures. A: Parallel stacked geometry, the least energetically favorable geometry. B: Parallel displaced geometry, the most energetically favorable geometry. C: T-shaped or edge to face geometry, more energetically favorable than the parallel stacked geometry but less favorable than the parallel displaced geometry.

Mecozzi *et al*. calculated the binding energies of benzene as well as other aromatic compounds of biological and medicinal interest (Mecozzi, et al.). Based on the relationships between the binding energy of benzene and the binding energy of the side chains of the aromatic compounds, we estimated the charges on the electron clouds of the aromatic residues (Table 2.2).

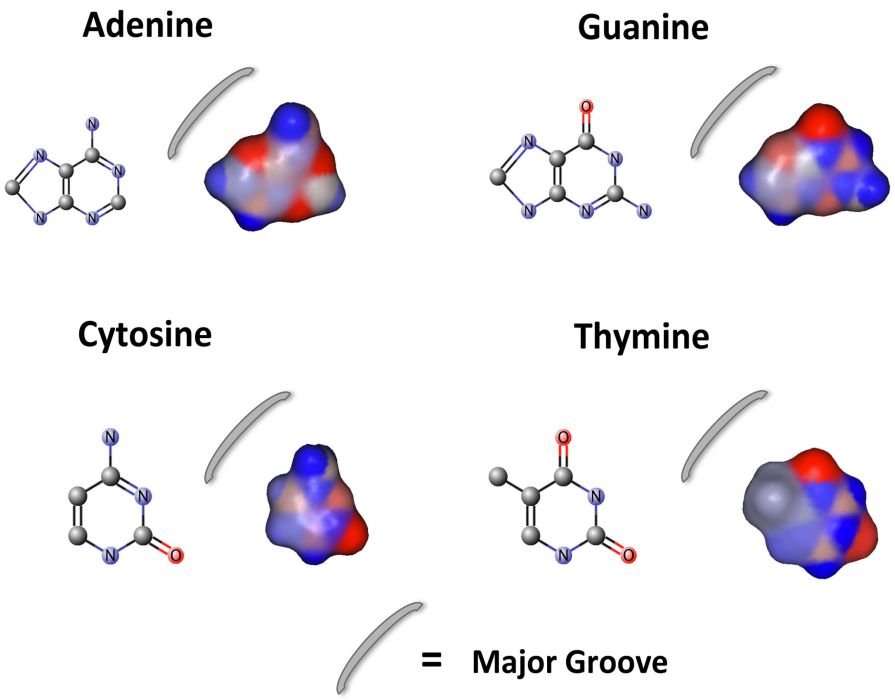

Figure 2.2: Electronic landscape of the bases. Charge distributions of the four bases in the major groove. The blue regions represent partial positive charges while the red regions represent partial negative charges. The grey regions are neutral. MarvinSketch 6.1.4, a software package from Chemaxon (Marvin6.1.4), was used to generate the electronic landscape and calculate the charges on the atoms.

Table 2.1: Quantified charges on nucleotide major groove atoms (blue and red regions on the electronic landscapes illustrated in Figure 2.2).

Atom	A	C	G	T
N4/O4	0.34	0.34	-0.44	-0.478
C5	-0.015	0.066	0.00	0.087
C6	-	0.085	-	0.096
N7	-0.21	-	-0.21	-
C8	0.115	-	0.10	-

Table 2.2: Estimated electron cloud charges of aromatic amino acids.

Molecule	Electron cloud charge
Benzene	-0.372
Tyrosine	-0.369
Phenylalanine	-0.372
Tryptophan	-0.447

The electrostatic potential was then calculated using:

$$E_{ac} = \frac{k_e N_A q_a q_c}{\varepsilon r} \quad (2.3)$$

where E_{ac} is the energy between an atom a on the base and the electron cloud c on the aromatic amino acid, k_e is Coulomb's constant, N_A is Avogadro's number, q_a and q_c are the charges of the atom and the electron cloud respectively, ε is the dielectric constant and r is the distance between the point charges (meters). The charges, q_a and q_c, are determined by multiplying the partial charge values by the charge of an electron, 1.6 x 10^{-19}. The electrostatic potential is then converted from joules/mol to kcal/mol using the conversion factor of 2.39 x 10^{-4}. The electrostatic potential is then converted from joules/mol to kcal/mol using the conversion factor of 2.39 x 10^{-1}. The electrostatic potential of each atom on the base with direct access to the electron cloud on the amino acid is summed together to calculate the total π interaction energy between the amino acid and base (Equation 2.4).

$$E_{\pi} = \sum_{a}^{N_a} E_{ac} \quad (2.4)$$

where E_{π} is the total π-π interaction energy between the base and the amino acid, N_a is the number of atoms of the base that have an unblocked pathway to the electron cloud on the aromatic residue, E_{ac} is the energy between an atom a on the base and the electron cloud c.

2.2.5 Prediction Algorithm

The flowchart for structure-based TFBS prediction is shown in Figure 2.3. It begins with a TF-DNA complex structure consisting of a single TF-chain/domain interacting with a duplex DNA. Hydrogen atoms were added to the complex structure, which are needed for hydrogen bond calculations, using UCSF Chimera 1.8 (Pettersen, et al.). The addition of hydrogen atoms may introduce steric clashes, which was addressed by energy minimization using Chimera with the following parameters: 100 steepest descent steps with a step size of 0.02, 100 conjugate gradient steps with a step size of 0.02, and an update interval of 10. A total of 8 base pairs, which include residues contacting bases and flanking bases, were used for the energy calculation. A residue-base contact is defined if the atom distance between the residue side chain and the base is within 3.9 Å. The native DNA sequence in the TF-DNA complex was mutated to generate all possible combinations of the 8 bases, 65536 sequences, using 3DNA (Lu and Olson). The three energy terms were then calculated for each of the 65536 TF-DNA complex structures. The score for each of the three terms, multibody energy, hydrogen bond energy, and π-interaction energy, was normalized using equation 2.5:

$$\mathrm{E_N} = \frac{\mathrm{E\text{-}E_{max}}}{\mathrm{E_{min}\text{-}E_{max}}} \qquad (2.5)$$

where E_N is the normalized energy, E is the energy for a specific complex with a sequence, $\mathrm{E_{max}}$ and $\mathrm{E_{min}}$ are the maximum and minimum energies in the set of 65536 TF-DNA complexes respectively. The total energy is then calculated using Equation 1. The distribution of integrative energy scores is generated using R and a significance level α is used to select the statistically significant sequences. In this study, we used α of 0.01 divided by the number of contacted DNA bases to normalize the number of expected

sequences. The rationale of using adjusted α is that for a fixed number of DNA binding sequences, if more bases are involved in TF-DNA interaction and are conserved, the expected number of binding sequences should be smaller. The sequences with energy scores in the adjusted α region were then selected to generate a position weight matrix (PWM) and motif logo (Figure 2.3).

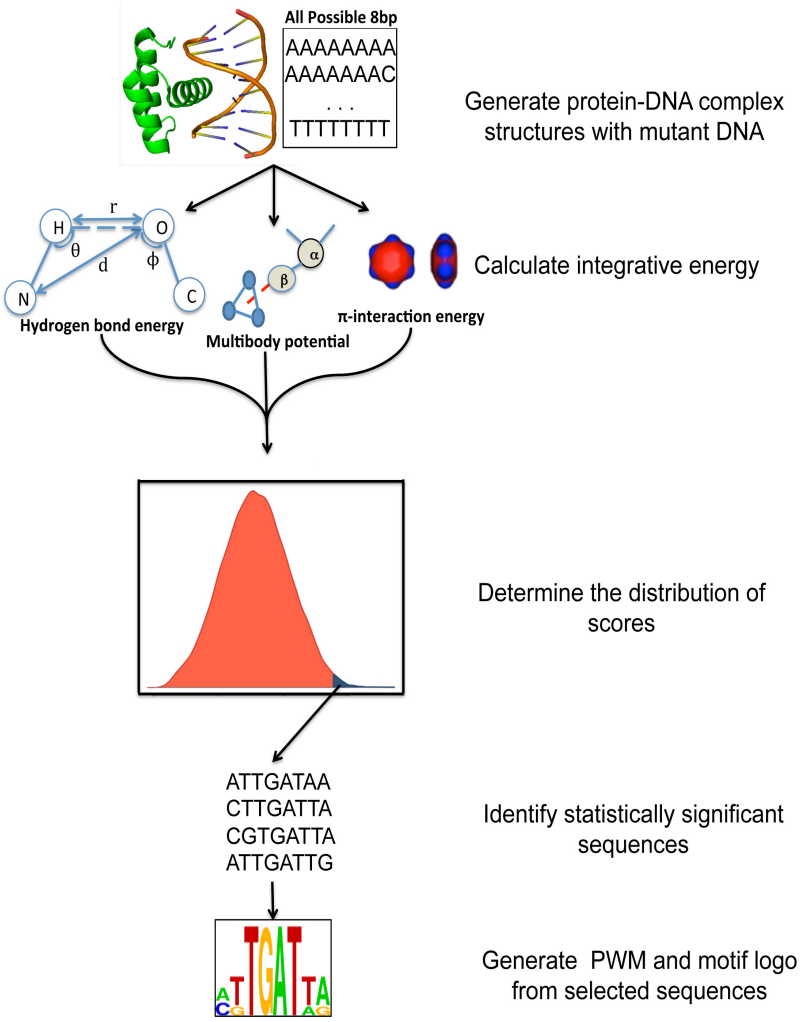

Figure 2.3: Flowchart for structure-based TFBS prediction.

2.2.6 Binding Motif Prediction and Validation

PWMs are generated using the selected sequences from the distribution of integrative energy scores. First, a 4x8 position frequency matrix (PFM) is generated using these sequences. The PFM is then converted to a PWM and subsequently converted to a motif logo using the method described by Schnieder and Stephens (Crooks, et al., 2004; Schneider and Stephens, 1990). The predicted PWMs were compared with their corresponding annotated JASPAR PWMs (Mathelier, et al., 2016). We used averaged Kullback-Leibler divergence (AKL) (Equation 2.6) to quantitatively measure the similarity between the predicted and reference TF-binging site PWMs(Wu, et al., 2001; Xu and Su, 2010).

$$D_{AKL} = \sum_{i} \sum_{B \in \{A,C,T,G\}} \frac{\left(P_{iB}\ log\frac{P_{iB}}{Q_{iB}} + Q_{iB}\ log\frac{Q_{iB}}{P_{iB}}\right)}{2} \qquad (2.6)$$

where D_{AKL} is the AKL divergence, i represents the corresponding columns of the base positions being compared in the predicted and reference matrices, B represents the four bases A,C,G, and T, P_{iB} and Q_{iB} represent the frequency of a particular base B in corresponding columns *i,* in the predicted and reference matrices respectively.

We also used a method called Information content (IC)-weighted Pearson correlation coefficient (PCC) (Persikov and Singh, 2014), developed recently by Persikov and Singh, to measure the similarity of corresponding columns from the predicted and experimentally derived PWMs. IC-weighted PCC is a method to measure the similarity of the corresponding columns from the predicted and reference PWMs representing the same base positions in the binding motif (Persikov and Singh, 2014). The information content is calculated using equation 2.7:

$$IC(m) = 2 + \sum_{B \in \{A,C,G,T\}} m_B \, log \;\; m_B \qquad (2.7)$$

where the $IC(m)$ is the the information content function for column *m* in a PWM, and *B* represents the DNA bases frequencies in that PWM column. The IC-weighted PCC is then calculated using equation 2.8:

$$PCC_{m,n}^{IC} = \frac{\sum_{b \in \{A,C,G,T\}} (m_b - \overline{m})(n_b - \overline{n})}{\sqrt{\sum_{b \in \{A,C,G,T\}} (m_b - \overline{m})^2 \bullet \sum_{b \in \{A,C,G,T\}} (n_b - \overline{n})^2}} \times \frac{IC(m)}{2} \qquad (2.8)$$

where $PC_{m,n}^{IC}$ is the IC-weighted PCC between the reference column m, and the predicted column, n. m_b and n_b are the frequencies of the DNA bases, b, found in the rows of the corresponding reference and predicted PWM columns respectively. $\overline{m}$ and $\overline{n}$ are the mean frequencies in the reference and predicted columns respectively. A predicted column is considered a correct prediction when the IC-weighted PCC between the corresponding predicted and reference columns is at least 0.25 (Persikov and Singh, 2014). The advantage of using IC-weighted PCC measure is that it takes into consideration the conservation of a base-position in the reference binding motif (information content) and how well it matches the predicted binding motif (Pearson's correlation coefficient).

2.2.7 Datasets

The first dataset is a non-redundant set of TF chain-DNA complexes. It was generated using all the high quality crystal structures of TF-DNA complexes in the Protein Data Bank (PDB) (Berman, et al., 2000) with corresponding JASPAR PWMs. These structures were solved by X-ray crystallography with a resolution less than 3Å and R-factors ≤ 0.3. All structures with a sequence identity of 35% or greater were first

grouped together. The TF-DNA complex structure with a corresponding JASPAR PWM and the highest resolution in a group was chosen as the group's representative. This dataset has 29 non-redundant TF chain-DNA complexes found in Table 2.3.

Table 2.3: Non-redundant dataset of 29 TF chain-DNA structures.

PDB ID	Chain	Protein Family Annotation
1AM9	A	HLH, Helix Loop Helix DNA-binding domain
1BC8	C	ets domain
1BF5	A	STAT,STAT DNA-binding domain, SH2 domain
1DSZ	A	Nuclear Receptor
1GU4	A	Leucine Zipper Domain
1H8A	C	Myb/SANT domain
1H9D	A	RUNT domain
1JNM	A	Leucine Zipper Domain
1NKP	A	HLH, Helix Loop Helix DNA-binding domain
1NLW	A	HLH, Helix Loop Helix DNA-binding domain
1NLW	B	HLH, Helix Loop Helix DNA-binding domain
1OZJ	A	SMAD MH1 domain
1P7H	M	Rel/Dorsal transcription factors, DNA-binding domain
1PUF	A	Homeodomain
1PUF	B	Homeodomain
1T2K	A	Interferon regulatory factor
2A07	F	Forkhead DNA-binding domain
2AC0	A	p53 DNA-binding domain-like
2DRP	A	Classic zinc finger, C2H2
2QL2	A	HLH, helix-loop-helix DNA-binding domain (CATH)
2QL2	B	HLH, helix-loop-helix DNA-binding domain (CATH)
2UZK	A	Fork head domain (PFAM)
3F27	D	HMG (high mobility group) box (PFAM)
3HDD	A	Homeodomain
2YPA	A	Helix-loop-helix DNA-binding domain (PFAM)
2YPA	B	Helix-loop-helix DNA-binding domain (PFAM)
4F6M	A	Kaiso zinc finger DNA binding domain - Transcriptional regulator Kaiso (Gene annotation)
4HN5	A	Erythroid Transcription Factor GATA-1 (CATH)
4IQR	A	Zinc finger, C4 type (PFAM)

Family classifications in Table 2.3 are primarily based on SCOP. CATH classifications are used when SCOP domain classifications are unavailable. If both SCOP and CATH classifications are unavailable, PFAM classifications are adopted.

We also generated a second non-redundant set for special case studies. Homeodomain proteins are involved in regulation of many cellular processes in mammals and represent the second largest family of transcription factors (Tupler, et al., 2001). There are a large number of experimentally determined PWMs for homeodomains and a relatively large number of homeodomain-DNA complex structures in the PDB. A homeodomain is a three α-helical DNA binding domain that binds to both the major groove and minor groove of the target DNA sequences (Gehring, et al.). To generate this dataset, we combined both the protein sequence similarity and binding site similarity. The homeodomain dataset consists of TF chain-DNA complexes with a corresponding JASPAR PWM. Each pair of the homeodomains in the dataset has less than 55% protein sequence similarity and different annotated binding sites in JASPAR (based on the IC-weighted PCC criteria of 0.25 or larger for the matching positions). This dataset includes: 1B8I:A, 1B8I:B, 1IC8:A, 1IG7:A, 1JGG:B, 1PUF:A, 1PUF:B, 3RKQ:A, 2HDD:A, 3A01:A, and 3A01:B. One exception is that we included both 1B8I:B and 1PUF:B because they have different binding sites even though they share 82% sequence identity. This is to test the capability of the new integrative energy function to see if we can accurately predict very different binding sites for highly similar proteins.

2.3 Results

We applied the new integrative energy function to the prediction of TFBSs using the non-redundant dataset of 29 TF-DNA complex structures and compared the

prediction with multibody potential and DDNA3, a knowledge-based atomic-level protein-DNA interaction potential (Zhang, et al., 2005). The predicted TF-binding motifs and the corresponding JASPAR motifs are shown in Figure 2.4. We also applied three different quantitative methods, Chi-square test, averaged Kullback-Leibler divergence and Euclidean distance, to compare the prediction accuracy as described in the Methods. The lower the AKL divergence value, the more similar between the predicted PWMs and JASPAR PWMs. Figure 2.5 shows the results based on AKL divergence to demonstrate the similarity between the predicted PWMs and the reference JASPAR PWMs. Results from the other two methods are consistent with the AKL divergence results. As shown in Figure 2.4 and 2.5, IE outperforms both MB and DDNA3 or at least one of them in the majority of the cases, for example, 1AM9:A and 1PUF:B. There are three cases that IE performs worse than MB and/or DDNA3, such as 1BF5:A and 2UZK:A. In several cases, the prediction accuracies are similar among all three energy functions, for example, 1DSZ:A.

To check if the overall improvements are statistically significant, we performed Wilcoxon signed rank test to compare the predictions between IE and MB as well as between IE and DDNA3 based on the predicted similarity to JASPAR PWMs. The null hypothesis is that prediction accuracy of the IE method is equal or worse than the MB (or DDNA3) method while the alternative hypothesis is that the prediction accuracy of the IE method is better than MB and DDNA3. The *p*-values for the three comparison metrics, Chi-square, AKL divergence and Euclidian distance are 0.003, 0.003, and 0.048 between IE and MB predictions, and 0.003, 0.005, and 0.025 tween IE and DDNA3 respectively, suggesting that the improvements are statistically significant.

Zinc fingers and homeodomains represent the two largest and extensively studied transcription factor families. In our non-redundant dataset, we found six zinc finger chains (Figure 2.6) and three homeodomains (Figure 2.7). Zinc fingers usually function as a dimer or multimers. A single zinc finger domain typically contains three to four conserved recognition bases (Persikov and Singh, 2014). Three of the six zinc finger cases (1LLM:C, 2DRP:A and 4F6M:A) show better binding site prediction using the IE function while the other three have no significant differences (1DSZ:A, 4HN5:A, and 4IQR:A, Figures 2.4, 2.5 and 2.6).

Each homeodomain recognizes a variation of the typical TAAT core binding site. There were three homeodomains in the non-redundant dataset. Figure 2.7 shows the predicted binding motifs and significant improvement in prediction accuracy when using the IE function over the MB and DDNA3 statistical potentials. The quantitative improvement is shown in Figure 2.5. In all three cases, predictions using the integrative energy consistently outperform both MB and DDNA3 potentials. Since we have a relatively large number of high quality homeodomain-DNA complex structures in the PDB and a large number of experimentally derived homeodomain binding motifs, we generated a larger dataset of homeodomains by combining the protein sequence similarity and binding site similarity as described in the Methods section. Figure 2.8 shows the predicted binding motifs using the IE (blue), MB (red), and DDNA3 (green) energy functions and their accuracy when compared with the JASPAR motifs. The data demonstrate that our new integrative energy function can also accurately predict the binding sites of homeodomains with high sequence similarity but with different binding sites (Figure 2.8).

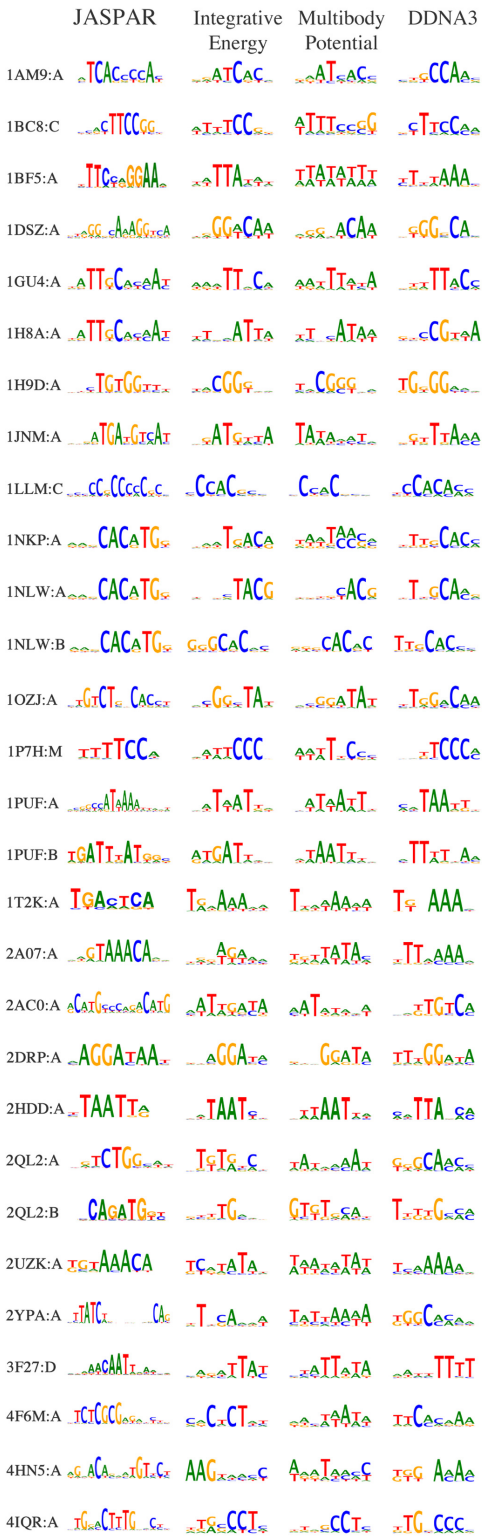

Figure 2.4; Comparison of JASPAR motifs with the predicted motifs using the IE, multibody and DDNA3.

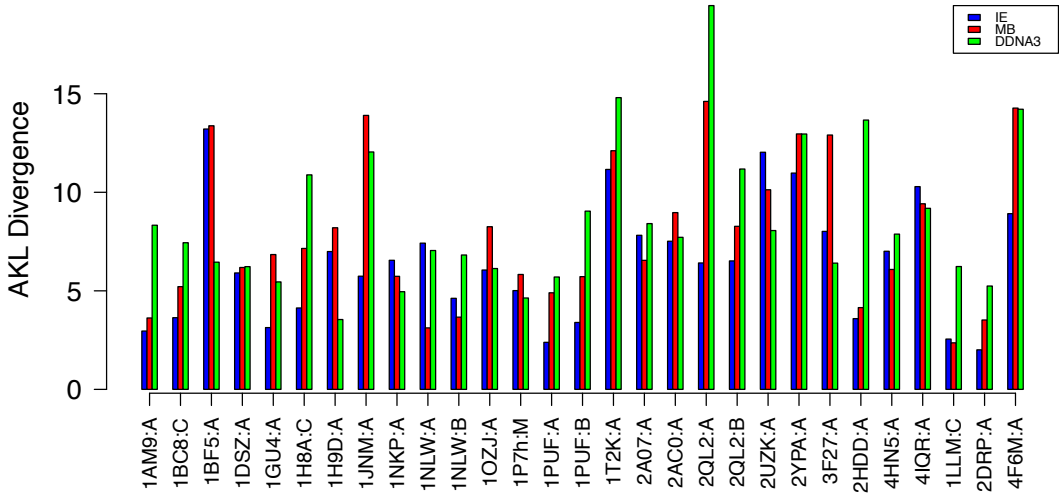

Figure. 2.5: Comparison of integrative energy prediction accuracy with multibody and DDNA3 potentials. AKL divergence of the predicted PWMs with JASPAR PWMs using the integrative function (IE: blue), multibody potential (MB: red), and DDNA3 (green).

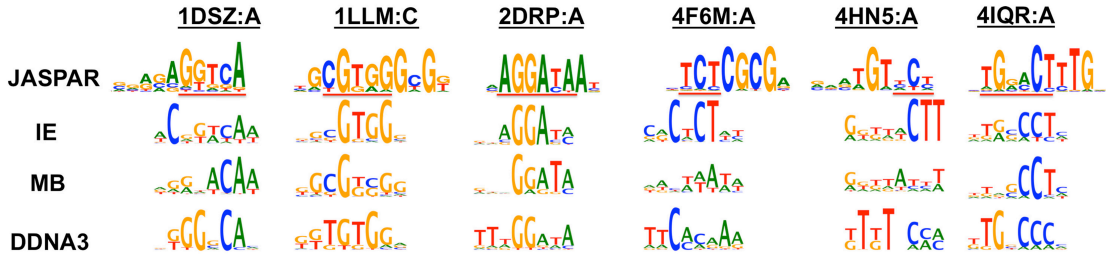

Figure 2.6: Comparison of zinc finger binding site predictions. Red lines under the JASPAR logos indicate the DNA sequences involved in binding to the TF-chain/domain.

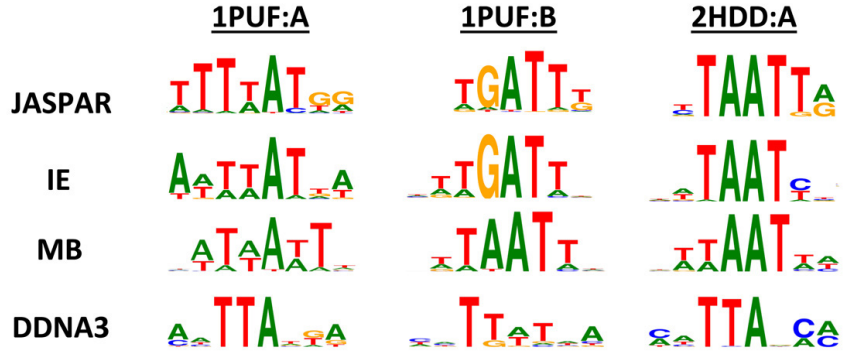

Figure 2.7: Binding site prediction of three homeodomains in the non-redundant dataset.

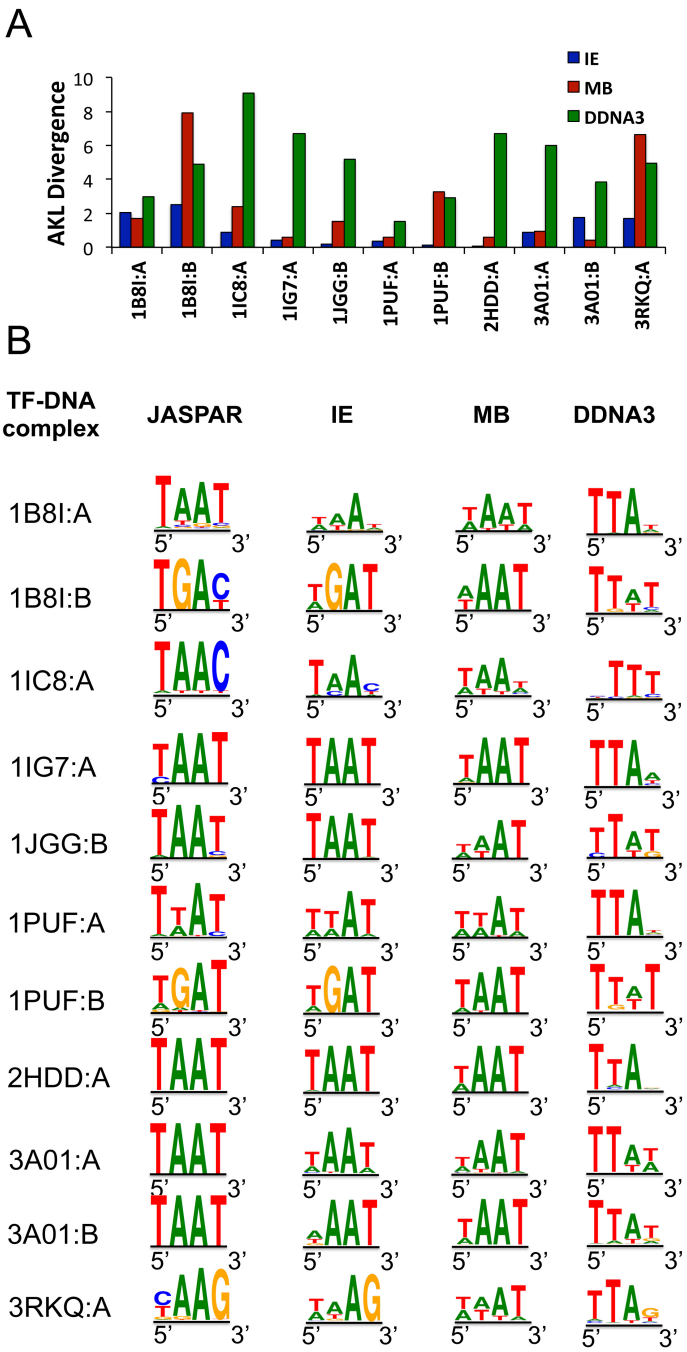

Figure 2.8: Prediction of homeodomain binding sites. (A) Quantitative comparison between the predicted binding motifs and JASPAR motifs of the homeodomain dataset using the integrative energy (blue), multibody potential (red), and DDNA3 (green) using Averaged Kullback-Leibler divergence. (B) Comparison of the predicted binding motifs

We also used a recently developed IC-weighted PCC method to calculate the correctly predicted core-binding positions (PWM columns) in the homeodomain dataset. Persikov and Singh suggested that a reference column is correctly predicted if the IC-weighted PCC between the corresponding predicted and reference columns is at least 0.25 (Persikov and Singh, 2014). Figure 2.9 shows that approximately 93% of the core base positions (44 columns) are correctly predicted by the integrative energy function, 86% by the MB potential, and 63% by the DDNA3 potential. The columns predicted by the IE function have a higher correlation to their corresponding JASPAR columns than the MB and DDNA3 energy functions.

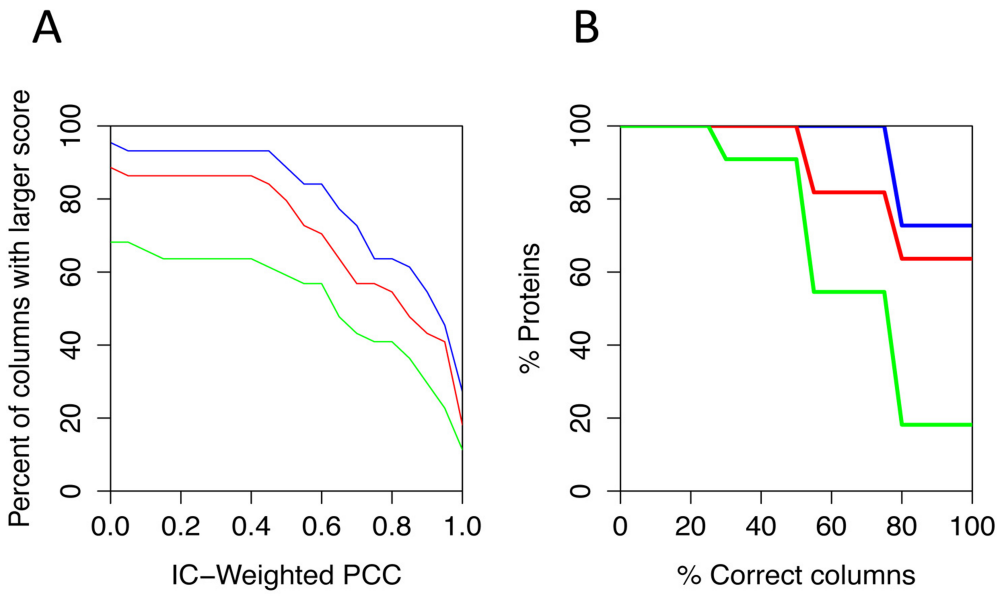

Figure 2.9: Performance comparison of integrative energy (blue), multibody (red), and DDNA3 (green) based on IC-weighted PCC. (**A**) Distribution of IC-weighted PCC. For each threshold of IC-weighted PCC score (*x*-axis), the fraction of predicted columns that achieves a score that high or more when compared to their corresponding JASPAR PWMs. (**B**) Percent of correctly predicted positions in the core 4mer PWMs. The percent of proteins with correct columns (percentage) using an IC-weighted PCC threshold of 0.25.

2.4 Discussion

We report here improved accuracy of structure-based TF binding site prediction using an integrative energy function. The integrative energy function consists of the multibody potential (Liu, et al., 2005), and two atomic terms: hydrogen bond energy and π-interaction energy. The multibody energy is a residue-level knowledge-based protein-DNA interaction potential derived from the mean force theory. Even though this multibody potential implicitly captures biophysical interactions including hydrogen bonds and π-interactions and showed its predictive power in both TF binding site prediction and protein-DNA docking studies (Liu, et al., 2008; Liu, et al., 2005), the mean force nature and the typical low count problem limit its ability to accurately capture the key hydrogen bond and π-interactions. For example, arginine has the ability to form bidentate hydrogen bonds, which allows it to bind specifically to guanine because guanine has two hydrogen acceptors present in the major groove of DNA. Bidentate hydrogen bonds are considered key contributors to protein-DNA binding specificity (Luscombe, et al., 2001; Seeman, et al., 1976). In the case of arginine and lysine, both can contribute to specific (through simple and complex hydrogen bonding) and non-specific (through electrostatic interactions) interactions; however, knowledge-based potentials cannot differentiate these two types of interactions. Therefore, adding explicit hydrogen bond terms can improve the accuracy of TFBS prediction by distinguishing hydrogen bonds that contribute to specificity from other interaction energies. We found that adding the explicit hydrogen bond term to the multibody potential improves the TFBS prediction accuracy of 1B8I:B and 1IC8:A in the homeodomain dataset (Figure

2.10A) as it captures the hydrogen bonds formed between arginine 258 and lysine 273 respectively and the guanine of the conserved G:C base pair (Figures 2.10B and 2.10C).

Aromatic residues can interact with DNA through π-interactions (Baker and Grant, 2007; Wilson, et al., 2014). T-shaped π-interaction with a base having partial positive charges in the major groove can contribute to binding specificity because of the variations of the electronic landscape of the bases in the major groove (Figure 2.2). However, these interactions are masked due to the low count problem and the mean force nature in knowledge-based potentials. Adding an explicit π-interaction term increases the accuracy of TFBS prediction. For example, the explicit π-interaction term captures the π-interaction formed between tyrosine 191 and the cytosine in the conserved G:C pair in 3RKQ:A (Figure 2.10B), improving the TFBS prediction accuracy. This suggests that the partial positively charged atoms (large blue spheres in Figure 2.10C) of cytosine interact electrostatically with the partial negatively charged atoms (large red spheres in Figure 2.10C) in the aromatic ring of tyrosine 191, which may contribute to TF-DNA binding specificity.

The integrative energy function shows an overall improvement in TFBS prediction over other knowledge-based potentials. However, in several cases in the multi-family dataset, the integrative energy function does not perform as well as the multibody and DDNA3 potentials (Figure 2.4 and 2.5). We investigated the complex structures and performed rigidity tests using FIRST (Jacobs, et al., 2001) and found that in those cases, the amino acids that interact with the DNA were from flexible regions or loops. For example, in the STAT1-DNA complex (1BF5:A), the residues involved in interacting with DNA are on the loops (Figure 2.11). As discussed in the introduction, both hydrogen

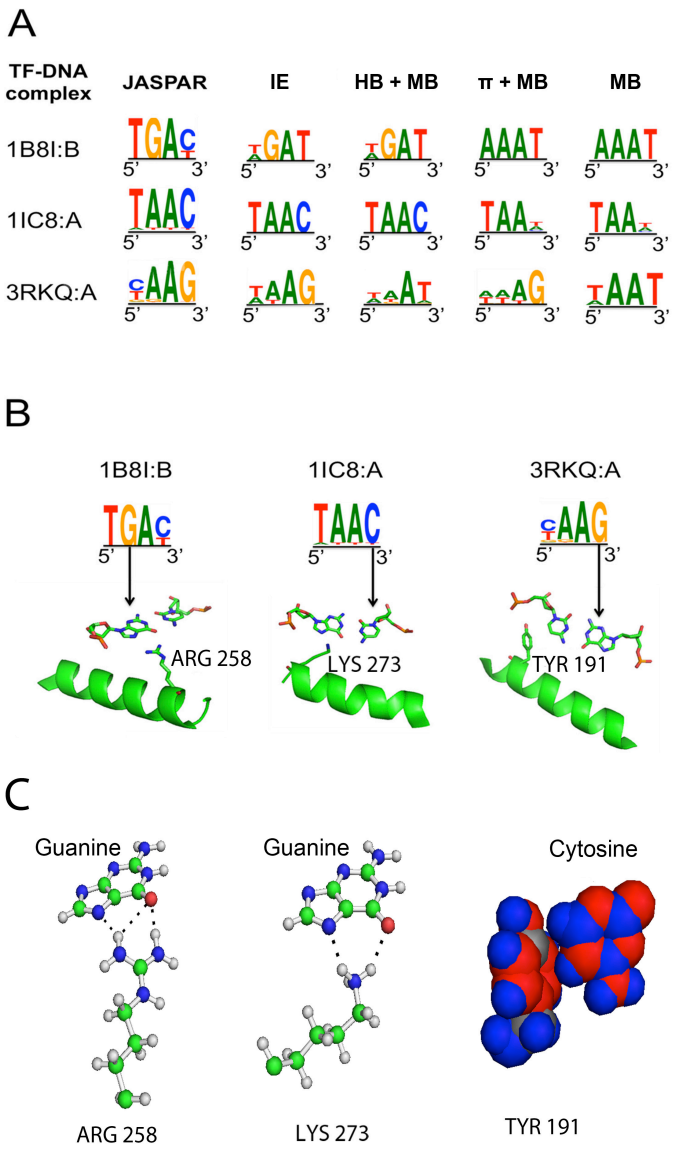

Figure 2.10: Contribution of energy terms to prediction accuracy. (A)The hydrogen bond energy term improves the prediction accuracy of 1B8I:B and 1IC8:A when compared to the multibody energy. The increased prediction accuracy of 3RKQ:A has a major contribution from the π-interaction energy term. (B) Physical interactions involving hydrogen bonds from arginine (1B8I:B), lysine (1IC8:A), and a π-interaction involving tyrosine (3RKQ:A) with the conserved G:C base pairs. (C) All-atom rendering of residue-base interactions showing the hydrogen bonds (black dotted lines) between Arg258 and guanine in 1B8I:B, between Lys273 and guanine in 1IC8:A where the green, blue, red, and white atoms represent carbon, nitrogen, oxygen, and hydrogen respectively. Tyrosine 191 is involved in π-interaction with cytosine where the blue, red, and grey spheres represent partial positive, partial negative, and neutral charged atoms respectively.

bonds and π-interactions are high-resolution functions that are sensitive to conformational changes. For complex structures with highly flexible regions for DNA contacts, there is a large variation of interaction energies for different conformations of the complex and the structure used for prediction is just a snapshot of multiple possible conformations. In addition, if a TF structure is not in an ideal docked conformation and the amino acids do not have favorable torsion angles to achieve favorable bidentate hydrogen bonds with the DNA, then the sensitive physical energies may not help the prediction, which is the case in 1NLW:A and 2UZK:A. Future work will need to incorporate the flexibility information into the prediction process.

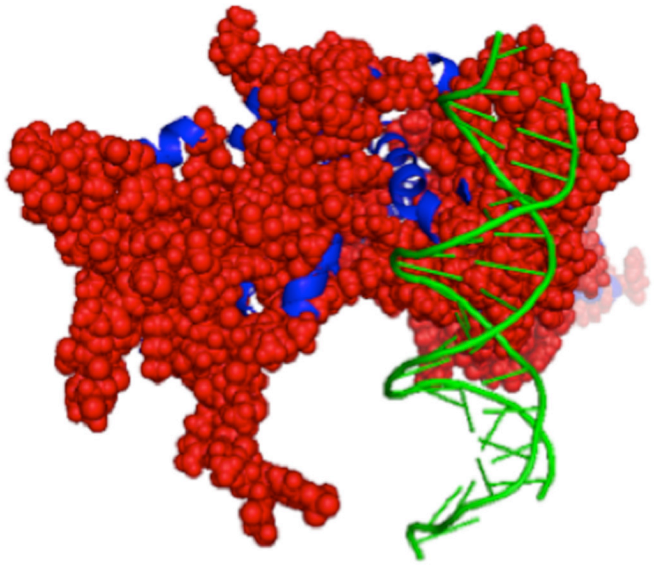

Figure 2.11: Complex structure STAT-1/DNA complex (1BF5:A). The interaction involves many coils (red spheres) with DNA (green).

2.5 Conclusion

We developed a novel integrative energy function that consists of three components, a knowledge-based multibody potential, a hydrogen bond energy function,

and an electrostatic potential for π-interaction energy. We applied the new integrative energy function to the prediction of transcription factor binding sites. The results show an overall improvement in binding site prediction and there is a significant improvement in predicting binding sites of homeodomains when compared to the multibody and DDNA3 potentials. The improved accuracy using the integrative function demonstrates the importance of considering hydrogen bonds and π-interactions explicitly in structure-based transcription factor binding site predictions, as they are not accurately captured by the knowledge-based potentials.

CHAPTER 3. A PENTAMER ALGORITHM FOR IMPROVING STRUCTURE-BASED TRANSCRIPTION FACTOR BINDING SITE PREDICTION

3.1. Introduction

In chapter 2, we developed a structure-based transcription factor binding site prediction algorithm using an integrative energy (IE) function that consists of three terms, a residue-level knowledge-based multibody potential (Liu, et al., 2005), an explicit hydrogen bond energy (Dahiyat and Mayo, 1997; Thorpe, et al., 2001), and an electrostatic potential for π-interaction energy (Farrel, et al., 2016). The new IE scoring function improves TFBS prediction accuracy over both residue-based and atomic-based knowledge-based potentials. However, the algorithm cannot scale well for predicting longer TF binding sites, especially for binding sites from homo/hetero TF dimers or tetramers. The previous algorithm first generates TF-DNA complexes consisting of a TF and every possible permutation of its target-sequence, an octamer (8 base pairs), which typically covers the full length of a single TF-domain binding site. The IE function is then applied to each TF-DNA complex to calculate their binding energy and subsequently predict their binding sites. The total number of TF-DNA complex energy calculations is 4^L, where L is the length of the binding motif. For example, in our previous full-length (octamer) approach, a binding motif of length eight base pairs requires evaluating a total of $4^8 = 65{,}536$ TF-DNA complexes. As the size of the binding sites increases, the time complexity increases exponentially. Here we propose a new approach, called the pentamer algorithm, for more efficient and accurate TFBS prediction. We also modified

our IE function to simplify the calculation of the hydrogen bond energy and π-interaction energy proposed in our previous method. Our results show that the new approach improves the prediction speed and accuracy, especially in the cases of TF dimers.

3.2 Methods

3.2.1 Modified Integrative Energy Function

The integrative energy function combines a residue level, knowledge-based multibody potential (Liu, et al., 2005), with a physics-based electrostatic energy potential (Equation 3.1). In our previous study, in addition to the multibody potential, we added a hydrogen bond term (Dahiyat and Mayo, 1997; Thorpe, et al., 2001) and a π-interaction term (Farrel, et al., 2016). The FIRST program is used to calculate the hydrogen bond energy (Farrel, et al., 2016; Jacobs, et al., 2001), which makes the calculation less efficient to use. Since both types of interactions are fundamentally electrostatic interactions, in this study, we combine the hydrogen bond energy and π-interaction into one electrostatic energy term to reduce the complexity of the calculation. The modified integrative energy function is:

$$E_{IE} = W_{MB}E_{MB} + W_{EE}E_{EE} \tag{3.1}$$

where E_{IE} is the integrative energy score, W_{MB} and E_{MB} are the weight and normalized energy score of the multibody potential respectively, and W_{EE} and E_{EE} are the weight and normalized energy score of the electrostatic energy respectively. Each energy term is normalized using the Min-Max normalization method as described in Chapter 2 (Farrel, et al., 2016). Since there are only a limited number of non-redundant TF-DNA complexes with known TFBSs, which is not enough to have a separate training set to optimize the weights, we use a weight of 1 for both W_{MB} and W_{EE}. Partial charges of the atoms

involved in electrostatic interactions between the TF and the DNA were determined by MarvinSketch, (Marvin6.1.4.) (Table 3.1):

Table 3.1: Charges of atoms used in electrostatic potential calculation.

Amino Acid	Atom	Charge fraction	Amino Acid	Atom	Charge fraction
SER	HG	0.21	PHE	CZ	-0.372
	HB2	0.056		CG	-0.372
	HB3	0.056		HD1	0.062
THR	HG1	0.21		HD2	0.062
	HB	0.059		HE1	0.062
ASP	OD1	-0.482		HE2	0.062
	OD2	-0.482		HZ	0.062
GLU	OE1	-0.482	TRP	CG	-0.685
	OE2	-0.482		CH2	-0.685
ASN	OD1	-0.466		HD1	0.104
	D21	0.157		HE1	0.252
	D22	0.157		HZ2	0.054
GLN	OE1	-0.466	DG	N3	-0.206
	E21	0.157		H21	0.152
	E22	0.157		H22	0.152
ARG	HE	0.140		H1	0.174
	H11	0.268		O6	-0.441
	H12	0.268		N7	-0.215
	H21	0.292		H8	0.107
	H22	0.292	DA	N3	-0.239
LYS	HZ1	0.255		N1	-0.241
	HZ2	0.255		H2	0.089
	HZ3	0.255		H61	0.157
TYR	CZ	-0.249		H62	0.157
	CG	-0.249		N7	-0.21
	HD1	0.062		H8	0.115
	HD2	0.062	DC	O2	-0.468
	HE1	0.065		N3	-0.22
	HE2	0.065		H41	0.157
	HH	0.218		H42	0.157
CYS	HG	0.102		H5	0.066
HIS	ND1	-0.255		H6	0.085
	HD2	0.068	DT	O2	-0.373
	HE1	0.107		H3	0.194
	HE2	0.217		O4	-0.478
				H6	0.096

The electrostatic potential is calculated with a variation of Coulomb's law (Equation 3.2):

$$E_{ab} = \frac{k_e N_A q_a q_b}{\varepsilon d} \tag{3.2}$$

where E_{ab} is the electrostatic energy between an atom a of the amino acid and an atom b of the base, k_e is Coulomb's constant, N_A is Avogadro's number, q_a and q_b are the charges of the amino acid atom and base atom respectively, ε is the dielectric constant and d is the distance between the point charges (meters). The charges, q_a and q_b, are determined by multiplying the partial charge values (Table 3.1) by the charge of an electron, 1.6 x 10^{-19}. The electrostatic potential is then converted from joules/mol to kcal/mol using the conversion factor of 2.39 x 10^{-4}. The electrostatic potential of each atom is added together to calculate the total electrostatic energy between the TF and the DNA sequence (Equation 3.3).

$$E_{EE} = \sum_{N_{ab}} E_{ab} \tag{3.3}$$

where E_{EE} is the total electrostatic energy between the TF and the DNA, N_{ab} is the number of amino acid-base atom interactions, E_{ab} is the electrostatic energy between an amino acid atom a and the base atom b. The interaction distance for atoms involved in a possible hydrogen bond, the hydrogen atoms and hydrogen bond acceptor atoms, was 1.5Å <= d <= 2.9Å (Dahiyat and Mayo, 1997; Thorpe, et al., 2001). REDUCE, a program for adding and removing hydrogen atoms to PDB structure files, is used to add hydrogen atoms to the TF-DNA complexes (Word, et al., 1999). The cutoff distance for atoms involved in a possible π-interaction between an aromatic amino acid and a base was at 4.5 Å based on previous studies (Gallivan and Dougherty, 1999). The sum of the charges found in the electron cloud of aromatic residues, are used as the charge for the

electrostatic energy calculation to account for the delocalization of electrons in π-systems and their involvement in π-π interaction (Farrel, et al., 2016; Michael Gromiha, et al., 2004) (McGaughey, et al., 1998; Wintjens, et al., 2000).

3.2.2 Pentamer Algorithm

The pentamer algorithm applies the IE function to TF-pentamer DNA complexes derived from a single TF-DNA complex structure. The first step is to determine the binding sequence for the transcription factor. It can be determined by prior knowledge or automatically detected based on the TF-DNA complex structure. The TF-DNA complex is checked for the first and the last bases that are in contact with the TF using an atom distance cutoff of 5Å. Though the non-interacting, flanking bases are less conserved, recent studies have shown that these flanking bases contribute to DNA binding specificity by affecting DNA shape and stability (SantaLucia, et al., 1996) (Afek, et al., 2014; Barrera, et al., 2016; Gordan, et al., 2013; Slattery, et al., 2014; Zhou, et al., 2015). Therefore, we add two base pairs on each side of the binding sequence of length *n* (Figure 3.1A). For example, a DNA binding sequence of 5 base pairs (bps) becomes a 9 bp sequence after adding two flanking bases on each side (Figure 3.1B). A TF-DNA complex was first energy minimized using UCSF Chimera 1.8 as described in Chapter 2 (Farrel, et al., 2016; Pettersen, et al., 2004). The DNA sequence from the previous step is then split into a series of overlapping 5bp DNA sequences using each contacted bp as the center of a pentamer. Five TF-pentamer DNA complex structures are generated for energy calculation (Figure 3.1B). The DNA sequence in each TF-pentamer is mutated to every possible permutation of the four bases using 3DNA (Lu and Olson, 2008), which results in 4^5 or 1024 TF-pentamer complex structures for each original TF-pentamer

(Figure 3.1B). In total, there are $n*1024$ TF-pentamer complex structures to be evaluated, where n is the number of base pairs of the initial binding sequence. The binding energy for each TF-pentamer DNA complex is then calculated (Equation 3.1).

To predict the TF binding sites from these pentamer interaction energies, we applied two different methods, the tiling array algorithm and position weight matrix (PWM) stacking algorithm. In the tiling array algorithm, the IE score of a binding sequence is the sum of the interaction energy of overlapping pentamer sequences (Figure 3.1C). The statistically significant scores from the binding sequence IE score distribution of all the permutations are determined and their corresponding DNA sequences are used to generate the binding motif as we did previously (Chapter 2)(Farrel, et al., 2016). In this study, the critical value for statistical significance in the tiling array algorithm was 0.01 normalized by the length of the predicted motif. For the PWM stacking algorithm, the binding sequence is broken up into pentamer subsequences. The IE score of each permutation of each pentamer sequence is calculated. For a given pentamer representing 5 contiguous bases on the binding motif, a PWM representing the statistically significant pentamer sequences is determined from the distribution of the IE scores of all the possible sequence permutations of that pentamer. Each position (column) in each pentamer PWM represents a specific position (column) in the binding motif PWM. All of the corresponding cells representing the frequency of a particular nucleotide in a specific position are added together to generate a PFM of the binding motif (Figure 3.1D). The PFM is then converted to a PWM and converted to a motif logo using the method described by Schnieder and Stephens(Crooks, et al., 2004; Schneider and Stephens, 1990).

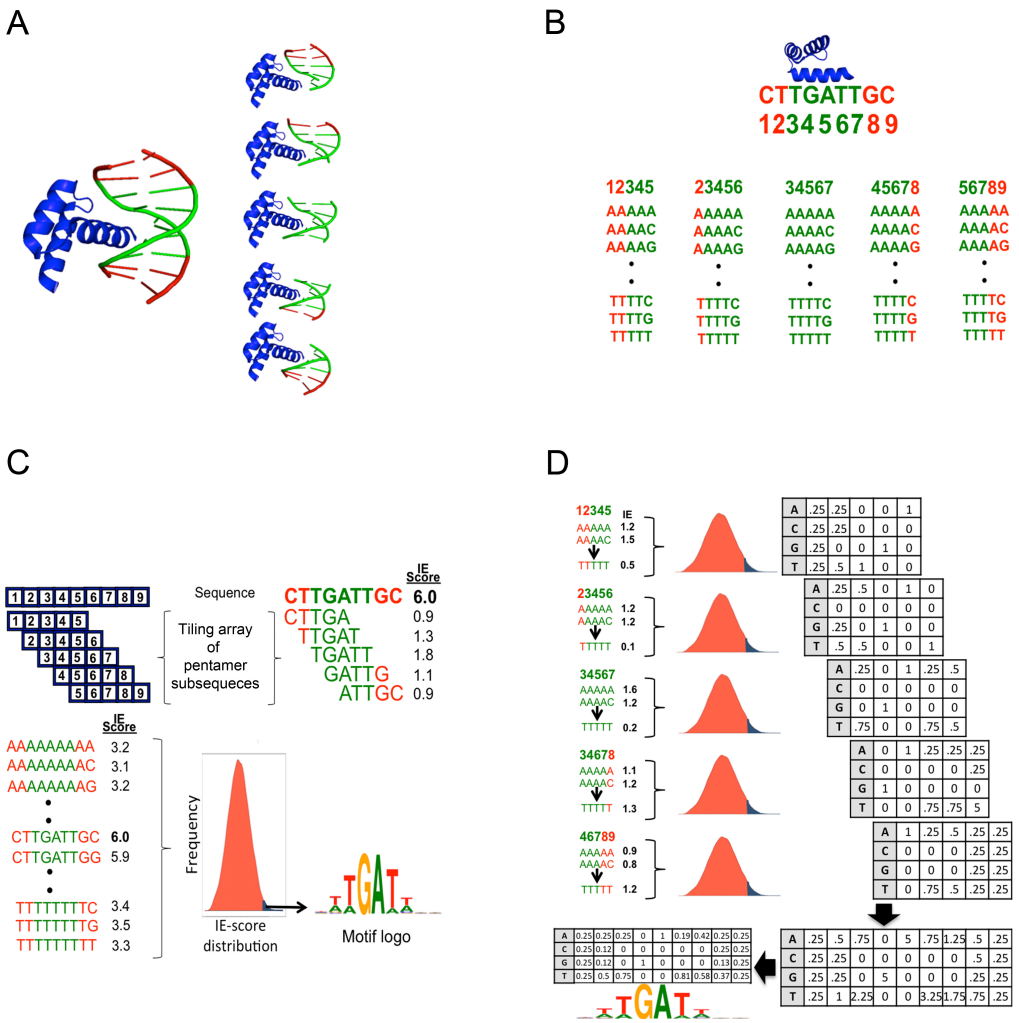

Figure 3.1: The pentamer algorithm. (A) The original TF-DNA structure is broken into TF-DNA pentamer structures. The green bases are TF-DNA contacts and the red bases are the 2 flanking bases on each side. (B) The DNA sequence is broken into overlapping pentamer subsequences. (C) The tiling array algorithm adds the IE scores of the TF-pentamer structures to determine a full length sequence IE score. The motif is predicted based on the sequences with statistically significant IE scores among all possible sequences. (D) The PWM stacking algorithm generates a distribution of IE scores based on the sequence permutations for each pentamer of the original sequence. The PWM positions corresponding to the same position in the original structure are added together to form a PFM representing the TF's TFBS, which is then converted to a PWM for the binding motif.

3.2.3 Dataset

The modified energy function and pentamer algorithm was tested on a non-redundant shown in Table 3.2. This dataset was generated in a similar way as described in Chapter 2. In addition, each TF chain structure was checked manually for similarity between the bound DNA and the corresponding JASPAR PWM. Uniprot (Wu, et al., 2006) and JASPAR were used to determine TF PDB structures with annotated binding PWMs by cross-referencing the Uniprot IDs used in both databases. However, not all DNA-binding domains in a multi-domain TF are represented in JASPAR. Furthermore, in selecting a representative for TF chains with high sequence identity, TF chains with more contacts in the major groove have higher priority. The dataset contains 27 non-redundant TF chain-DNA complexes (Table 3.2). The same dataset of homeodomains in Chapter 2 was also used for testing the new algorithm. For dimer binding site prediction, eight TF-DNA complex structures were used: 1AM9, 1GU4, 1JNM, 1NKP, 1NLW, 1OZJ, 2QL2, and 2YPA. Similar to chapter 2, domain family classifications are primarily based on SCOP. CATH classifications are used when SCOP domain classifications are unavailable. PFAM classifications are used if both SCOP and CATH do not have annotations.

3.2.4 Performance Evaluation

The predicted PWMs were compared with their corresponding JASPAR PWMs using the same methods reported in Chapter 2 (Mathelier, et al., 2016). Averaged Kullback-Leibler divergence (AKL) (Equation 2.6) was used to quantitatively measure the similarity between the predicted and reference TF-binding site PWMs (Wu, et al., 2001; Xu and Su, 2010). IC-weighted PCC was used to determine the number of

correctly predicted columns in the aligned predicted and reference PWMs (Equations 2.7 and 2.8).

Table 3.2: Non-redundant dataset of 27 TF chain-DNA structures.

PDB ID	Chain	Protein Family Annotation
1AM9	A	HLH, Helix Loop Helix DNA-binding domain
1BC8	C	ets domain
1BF5	A	STAT,STAT DNA-binding domain, SH2 domain
1DSZ	A	Nuclear Receptor
1GU4	A	Leucine Zipper Domain
1H9D	A	RUNT domain
1JNM	A	Leucine Zipper Domain
1NKP	A	HLH, Helix Loop Helix DNA-binding domain
1NKP	B	HLH, Helix Loop Helix DNA-binding domain
1NLW	A	HLH, Helix Loop Helix DNA-binding domain
1OZJ	A	SMAD MH1 domain
1P7H	L	Rel/Dorsal transcription factors, DNA-binding domain
1PUF	A	Homeodomain
1PUF	B	Homeodomain
2A07	F	Forkhead DNA-binding domain
2AC0	A	p53 DNA-binding domain-like
2DRP	A	Classic zinc finger, C2H2
2QL2	A	HLH, helix-loop-helix DNA-binding domain (CATH)
2QL2	B	HLH, helix-loop-helix DNA-binding domain (CATH)
2UZK	A	Fork head domain (PFAM)
3F27	D	HMG (high mobility group) box (PFAM)
3HDD	A	Homeodomain
4IQR	A	Zinc finger, C4 type (PFAM)
2YPA	A	Helix-loop-helix DNA-binding domain (PFAM)
2YPA	B	Helix-loop-helix DNA-binding domain (PFAM)
4F6M	A	Kaiso zinc finger DNA binding domain - Transcriptional regulator Kaiso (Gene annotation)
4HN5	A	Erythroid Transcription Factor GATA-1 (CATH)

3.3 Results

The modified and original integrative energy function functions were applied to the multi-family non-redundant dataset of 27 TF-DNA complex structures using the original full-length algorithm (Farrel, et al., 2016) to evaluate the performance of the modified IE function. The old and modified IE functions have overall similar performances when the predicted binding motifs are compared to their corresponding experimentally annotated binding motifs found in JASPAR (Figure 3.2). A Wilcoxon Signed Rank test was performed to evaluate the null hypothesis that the modified and original IE functions had equal AKL divergences. The p-value of this test was 0.67 inferring that the performance of the two IE functions are similar.

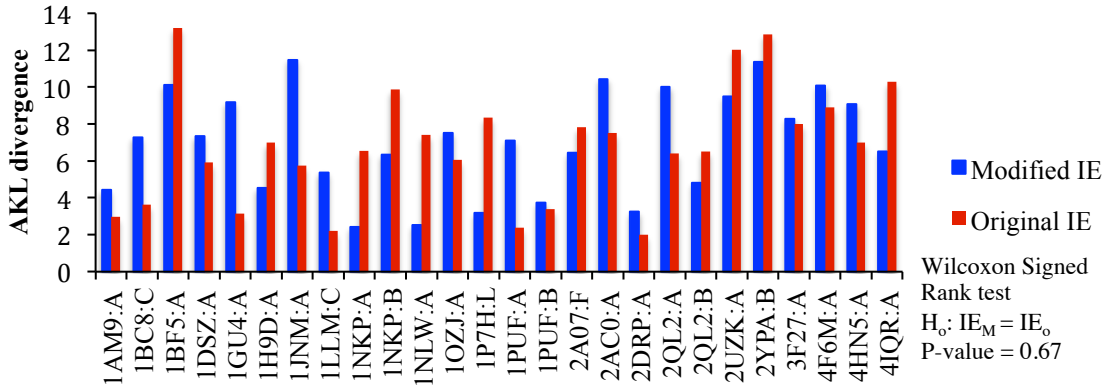

Figure 3.2: Comparison of TFBS prediction accuracy using the original full-length algorithm with the modified integrative energy function and original integrative function.

The pentamer algorithm with the modified energy function was tested on the multi-family non-redundant dataset of 27 TF-DNA complex structures. The PWMs of the predicted TF binding sites and their corresponding JASPAR binding sites were compared using IC-weighted PCC to determine the number of matching columns. We compared the performance of the new pentamer algorithm with our previous full-length algorithm

(Chapter 2), which showed overall improved performance of the new algorithm (Figures 3.3 and 3.4). Not only does the pentamer algorithm run much faster due to the greatly reduced total number of interaction energy calculations, in majority of the cases, the pentamer method produced better or comparable results than the original full-length method in terms of the number of correctly predicted columns based the IC-weighted PCC (Figure 3.3A). For example, more columns are predicted correctly using either the tiling array or the PWM stacking pentamer algorithm for 1GU4:A and 1P7H:L, which also reflected in the binding motifs (Figure 3.3A and Figure 3.4). We performed statistical analysis using Wilcoxon Signed Rank test with an alternative hypothesis that the pentamer algorithm generated a greater number of correctly predicted base positions than the full-length algorithm. The *p*-values are 0.016 and 0.0051 for the tiling array and PWM stacking algorithms respectively, suggesting that performance increase is significant. There are no apparent performance differences between the two pentamer methods, tiling array and PWM stacking (Figure 3.3).

In some cases, both the pentamer and the full-length methods have similar results, such as 1AM9:A and 1PUF:B (Figure 3.3A and Figure 3.4). However, since the cutoff for correctly predicted columns is set at 0.25 as proposed by Persikov and Singh (Persikov and Singh, 2014), there is a large range of IC-weighted PCC values for the correctly predicted columns, from 0.25 to 1. A closer look at distributions of the IC-weighted PCC values revealed that even though the correctly predicted columns are comparable between the pentamer and full-length algorithms in some cases, such as 1NKP:B, 1NLW:A, and 2DRP:A, the IC-weighted PCC values are better (closer to 1) from the pentamer prediction than the full-length prediction (Figure 3.3B). The

distribution of the IC-weighted PCC of all 27 cases shows a similar trend (Figure 3.3C). More data points are close to the perfect IC-weighted PCC score in the pentamer algorithms than the full-length method. Among the 27 cases, there are only 4 cases (1BC8:C, 1NLW:A, 2QL2:A, and 3F27:D) that the full-length method has a slightly better prediction (Figure 3.3A, D and Figure 3.4).

We also tested the new pentamer algorithm on a non-redundant dataset of homeodomains as described in Chapter 2 (Farrel, et al., 2016). Since we know the core binding sites of homeodomains, it is easier to distinguish the core binding sites from the flanking regions of reference binding motifs in JASPAR. The pentamer and full-length algorithms have comparable performance in predicting homeodomain binding sites (Figures 3.5A and B). The pentamer algorithm predicts better for 1IC8:A, 1PUF:A, and 3RKQ:A while full-length algorithm produces better predictions for 1JGG:B, 2HDD:A, and 3A01:B (Figure 3.5A and B). The remaining 5 cases have similar accuracy. As for the core binding sites, we used AKL divergence to compare the prediction accuracy with the reference binding core motifs in JASPAR (underlined in Figure 3.5A). A smaller AKL divergence value indicates a better match. While statistical test (Wilcoxon Signed Rank test) shows no significant AKL divergence difference between the pentamer algorithms, tiling array or PWM stacking, and the full-length algorithm for the 4-base core site prediction with *p*-values of 0.22 and 0.056 respectively, the predictions of core binding sites for 1JGG:B and 3A01:A are much better using the full-length method (Figure 3.5C).

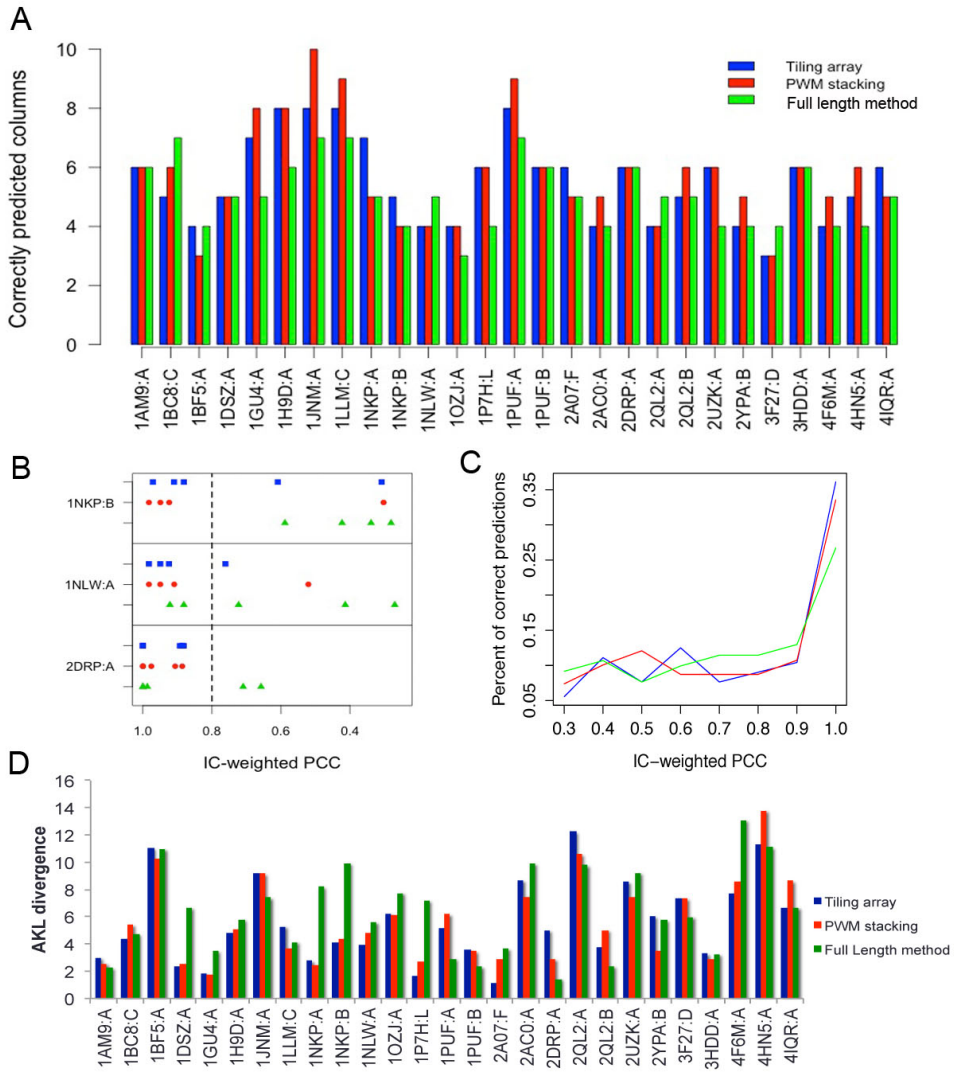

Figure 3.3: Comparison of TFBS prediction between pentamer and full-length algorithms. (A) Comparison of the number of correctly predicted columns (based on the IC-weighted PCC scores) by the tiling array (blue), PWM stacking (red), and full-length (green) algorithms. (B) Distributions of IC-weighted PCC values of correctly predicted columns by tiling array (blue squares), PWM stacking (red circles), and full-length (green triangles) algorithms. (C) Distributions of IC-weighted PCC scores predicted correctly in 27 cases by the tiling array (blue), PWM stacking (red), and full-length (green) algorithms in the multi-family dataset. (D) AKL divergence between JASPAR annotated binding motifs and motifs predicted by the tiling array (blue), PWM stacking (red), and full-length (green) algorithms.

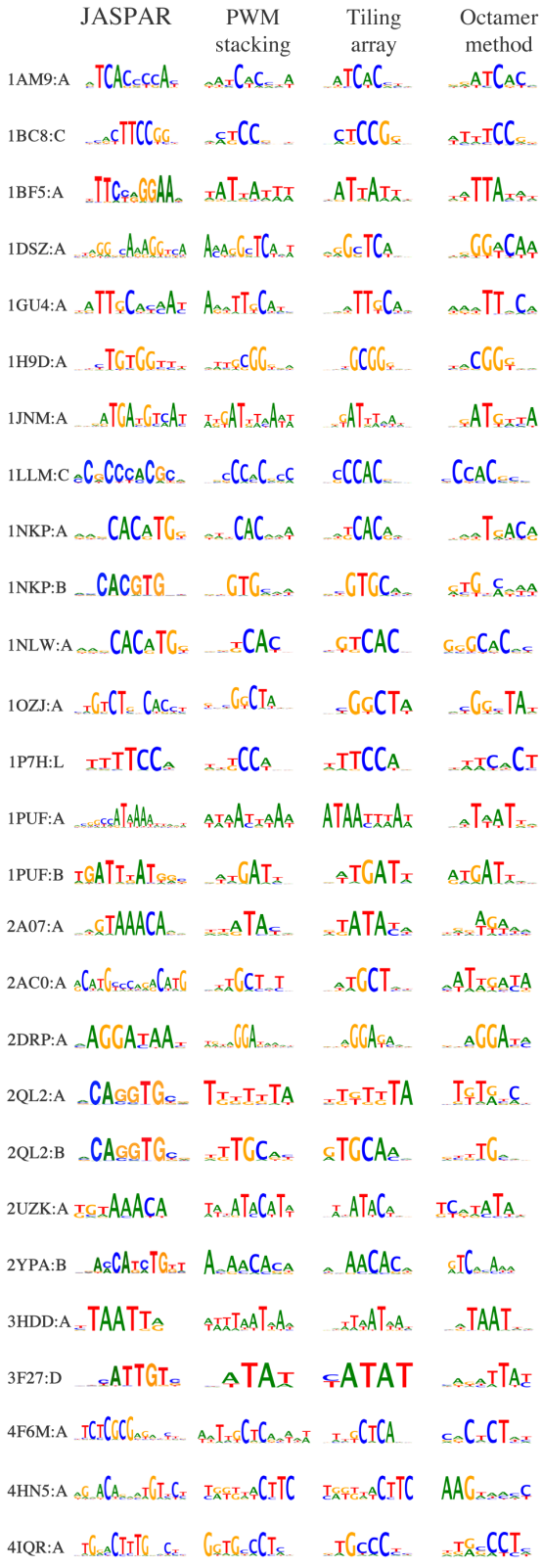

Figure 3.4: Comparison of the reference binding motif logos in JASPAR with the motif logos predicted by the tiling array, and PWM stacking, and full-length algorithms.

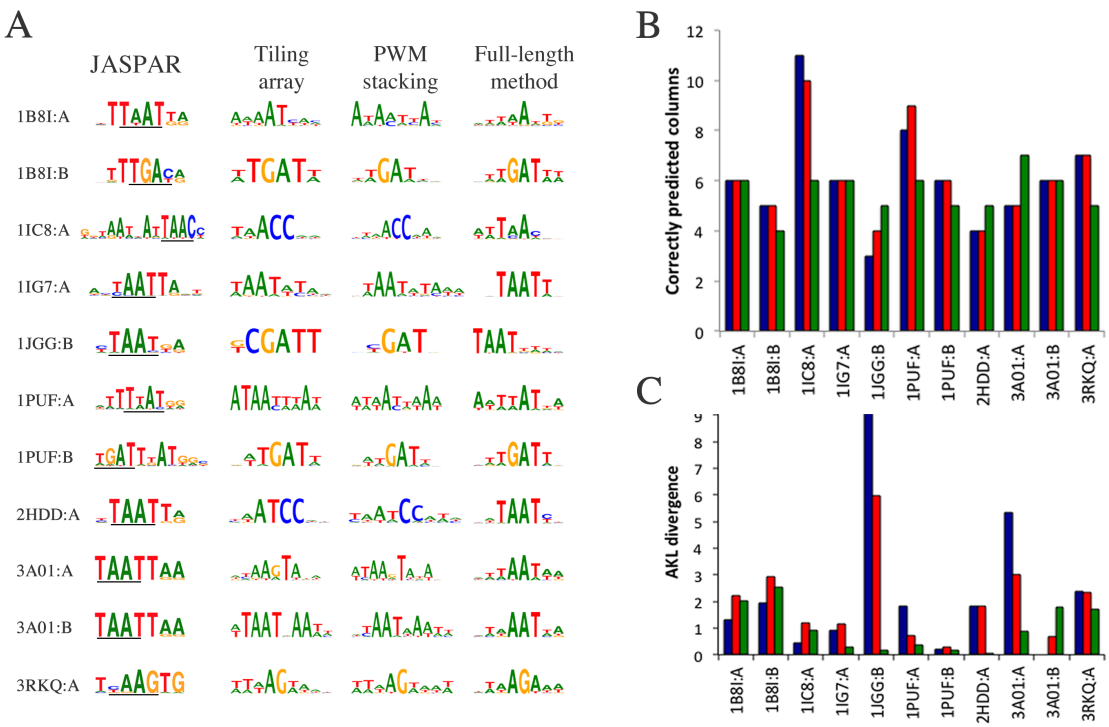

Figure 3.5: Comparison of reference and predicted homeodomain binding sites. (A) Comparison of homeodomain binding motifs predicted by the tiling array, PWM stacking, and full-length algorithms. The core binding positions are underlined. (B) Comparison of the number of correctly predicted columns based on IC-weighted PCC and (C) comparison of the similarity between the 4-base core-binding region based on AKL divergence between the JASPAR annotated binding motifs and the predicted binding motifs by the tiling array (blue), PWM stacking (red), and full-length (green) algorithms.

A close examination of 1JGG:B and 3A01:A revealed that both proteins have long N-terminal tails interacting with the minor groove, especially in 3A01:A (Figure 3.6). Homeodomains consist of 3 alpha helices and an N-terminal arm with the carboxyl-terminal recognition helix binding in the major groove of DNA (Gehring, et al., 1994; Noyes, et al., 2008). Typically arginine and lysine residues in the N-terminal tail within 6 residues of a homeodomain's first alpha helix interact in the minor groove for stability and may contribute to specificity (Figure 3.6) (Christensen, et al., 2012; Noyes, et al., 2008). The N-terminal tails of homeodomains are relatively flexible and interact with the

minor groove. The complex structure is just a snapshot of the dynamic N-terminal and DNA minor groove interaction. In addition, most of the minor groove interactions are non-specific due to the relatively non-discriminative surface of the minor groove (Corona and Guo, 2016). Since the pentamer algorithm captures relative local interactions, the "noise" from these non-specific interactions may get amplified when compared with the full-length algorithm.

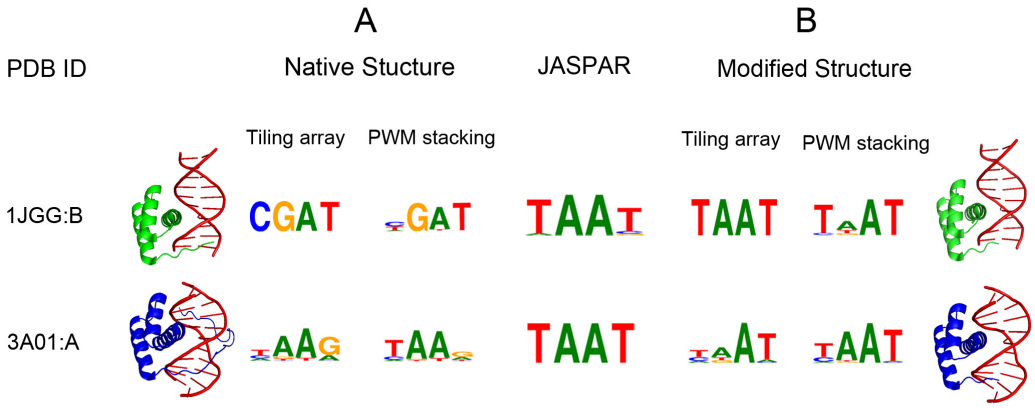

Figure 3.6: Effect of homeodomain N-terminal tails on core prediction accuracy using the pentamer algorithm. The original complex structures and the predicted binding sites are shown on the left and the structures after removing the N-terminal tails and their binding site predictions are shown on the right.

To test if the N-terminal tails affect the prediction accuracy, we redid the pentamer prediction after removing most of the N-terminal tails and only keeping 6 residues upstream of the first alpha helix. Removing the N-terminal residues greatly improves the binding site prediction (Figure 3.6). The AKL divergences between the predicted and JASPAR core binding motifs are reduced from 9.30 and 5.98 to 0.19 and 0.82 for the tiling array and PWM stacking algorithms respectively for 1JGG:B, and from 5.36 and 3.02 to 1.47 and 0.53 for the tiling array and PWM stacking algorithms respectively for 3A01:A when using the structures without the N-terminal tails. These

results are consistent with our previous observation that more contacts from dynamic coils in a crystal structure may affect the integrative energy function due to the sensitivity of the physics-based electrostatic energy (Chapter 2) (Farrel, et al., 2016), which affects the binding energy more for the pentamer algorithm than the full-length algorithm since the "noise" can be masked better in the full-length method.

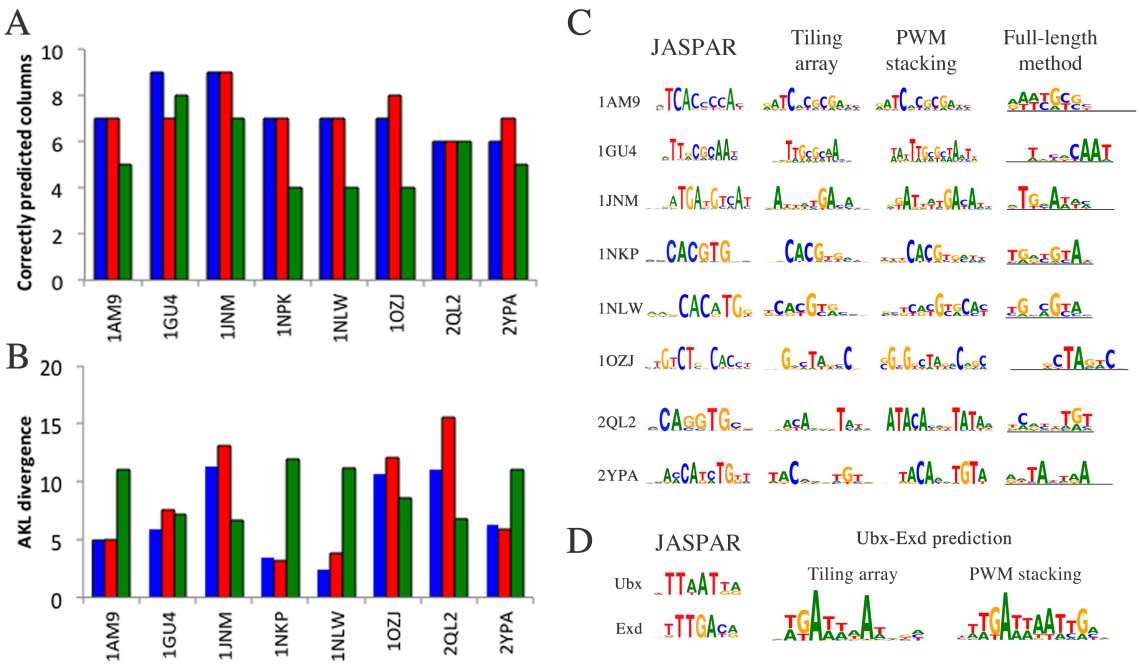

Figure 3.7: Comparison of TF dimer binding site predictions between pentamer and full length algorithms. (A) Comparison of the number of correctly predicted columns in TF dimers by the tiling array (blue), PWM stacking (red), and full-length (green) algorithms. (B) AKL divergence between dimer's JASPAR binding motifs and their corresponding binding predictions by the tiling array (blue), PWM stacking (red), and full-length (green) algorithms. (C) Comparison of TFBSs of TF dimers predicted by the tiling array, PWM stacking, and full-length algorithms. (D) Multi-domain TF prediction of the Ubx-Exd TFBS, which does not have an annotated binding site for the hetero-dimer in JASPAR.

The pentamer algorithm performs calculations at a linear time complexity for subsections of the binding site allowing the prediction of longer binding motifs, for example, multi-domain TFBSs. We compiled eight TFs from the multi-family dataset

that appear to work as homo or hetero dimers (See 3.2 Methods). While the pentamer method significantly reduces the time for energy calculation, except for 1GU4 and 2QL2, all other six dimer cases have improved TFBS predictions when compared to predictions using all possible sequences of the binding lengths (Figure 3.7). We also made predictions for Hox proteins Extradenticle (Exd) and Ultrabithorax (Ubx). Ubx and Exd form a dimer to regulate gene expression (Crocker, et al., 2015; Passner, et al., 1999). Though both Ubx and Exd have separate annotations of PWMs in JASPAR (Figure 3.7D), there are no JASPAR binding motif for the Ubx-Exd dimer. However, binding sites of Ubx-Exd dimer have been reported in structural studies (Foos, et al., 2015; Passner, et al., 1999). The predicted Ubx-Exd dimer binding sites are consistent with the published data. Furthermore, the *Drosophila* limb-promoting gene *Distalless* regulatory element, which is in part regulated by Ubx-Exd interactions (Gebelein, et al., 2002; Merabet, et al., 2007), also has a similar binding site to the our predicted binding motif using the pentamer algorithm.

As for the two pentamer prediction algorithms, tiling array and PWM stacking, there are no apparent differences. While both pentamer algorithms reduce the time complexity by lowering the total number of TF-DNA complexes for energy calculation, calculating the final binding motif using the tiling array algorithm requires more computation than the PWM stacking approach as it requires calculating scores for all possible sequence permutations of the binding site (Figure 3.1C). For TF dimer binding sites that have longer spacing (>= 4bps) between two monomer binding sites, it is more efficient to calculate each of the binding sites separately using the pentamer algorithms and then combine the two to form one binding motif.

3.4. Discussion

In this chapter, we developed an efficient pentamer algorithm and a simplified energy function that combines the hydrogen bond term and the π-interaction term into one electrostatic energy term. While the previous method described in Chapter 2 improves TF binding site prediction over the knowledge-based multibody and DDNA3 potentials, it requires evaluation of a total of n^L (where L is the length of the binding sequence) TF-DNA complexes. As the length of the binding motif increases, the number of energy calculations increases exponentially. In Chapter 2, we used a fixed sequence length (8bps or octamer), which typically covers the full-length of a binding site for a single TF-domain, for TFBS prediction of single TF domain-DNA complexes. Each TFBS prediction requires energy calculations for each of the 65536 (4^8) possible permutations of the TFBS sequence for the full-length algorithm, which can be calculated in reasonable time (Farrel, et al., 2016). However, for longer binding sequences, it becomes impractical even with large computer clusters.

There are many examples where we need to evaluate longer binding sequences. For example, we need to consider flanking sequences for binding site predictions as it has been demonstrated that flanking sequences contribute to binding specificity even though these they are not conserved (Gordan, et al., 2013). Secondly, some sites are regulated by the binding of TF dimers or tetramers, which are usually longer than an octamer. Finally, in homology model based TF binding site prediction, it would be nice to consider multiple homology models to increase the conformation coverage, which demands more energy calculations.

Our new algorithm with the modified IE energy improves both the speed and accuracy, especially for longer binding sites. The increase of prediction speed is obvious since we only need to calculate energies of 1024 (4^5) TF-pentamer complexes times the number of splits (Figure 3.1). The overall improvement of accuracy may lie in the fact that long range interactions from the coarse multibody function may introduce noise to the original full-length algorithm. In the pentamer algorithm the noise level is reduced. Therefore, for short binding site predictions that do not command more computations, an full-length approach can be applied. For longer binding site predictions, a pentamer method offers better accuracy and is more time-efficient. There is no apparent difference between the two methods for deriving the binding motifs based on the energies of all TF-pentamer complexes. The tiling array approach uses more computing time than the PWM stacking approach with an advantage of providing the actual binding sequences, while the PWM stacking approach only produces the binding motifs using a statistical approach.

3.5 Conclusion

We developed a pentamer algorithm with a modified energy function that speeds up the TFBS prediction by reducing the number energy calculations and improves TFBS prediction accuracy. Two algorithms, tiling array and PWM stacking, have been used to combine the TF-pentamer results for binding motif prediction with comparable performance in terms of TFBS prediction accuracy. The PWM stacking algorithm is relatively faster than the tiling array approach, as it does not need to calculate the individual sequence binding energy. Our results also show that the longer the binding sequences, the higher the increase in speed and accuracy can be achieved, for example, the binding site prediction of multi-domain TFs.

CHAPTER 4. PREDICTION OF HOMEODOMAIN BINDING SPECIFCITY USING HOMOLOGY MODELS

4.1 Introduction

As described earlier, structure-based prediction of TFBSs requires known TF-DNA complex structures. Experimental determination of high-resolution TF-DNA complex structures remains a difficult task and most TFs do not have a resolved TF-DNA complex structure in the PDB. As a result, the number of known TF sequences vastly outnumbers the number of TF-DNA structures in the PDB. The unavailability of TF-DNA complex structures limits the application of structure-based TFBS prediction. Computationally predicting TF-DNA complexes can address this issue but remains a challenge in computational structural biology.

In general, there are two computational methods for generating TF-DNA complex models. One is homology modeling using existing homologous TF-DNA complexes as templates. The other approach is to generate TF-DNA models through computational docking studies (Dominguez, et al., 2003; Roberts, et al., 2013; Takeda, et al., 2013; van Zundert, et al., 2015). While computational docking is a promising approach, protein-DNA docking has lagged behind other docking studies, such as protein-protein and protein-ligand docking. Currently, TF-DNA docking accuracy is not good enough for routine applications of TFBS predictions, especially using methods with atomic terms that are sensitive to atomic distances. On the other hand, homology modeling is quite

mature and can offer very good structural models (Morozov, et al., 2005; Schueler-Furman, et al., 2005).

Homology modeling is a common method used to predict structure of proteins using known protein structures with high sequence identity as templates. Baker and Sali showed that homology modeling can be used to predict structures up to 1 Å root mean square (RMS) error if the template has a high sequence identity (>50%), and mid range sequence identity (30-50%) up 1.5 Å RMS error (Baker and Sali, 2001). However, homology modeling of TF-DNA complexes for the application of structure-based TFBS prediction adds another level of complexity as some energy functions are sensitive to atomic terms and rely on the amino acid sidechains and nucleotides to be in a specific range of conformations to produce an accurate TFBS prediction. In this chapter we present a method of generating multiple TF-DNA homology models, and use these models to predict the TFBSs of homeodomains.

4.2. Method

4.2.1 Dataset

The dataset of homeodomains described in Chapter 2 and Chapter 3 were used to compare how well homology TF-DNA models can be applied to predict TF binding sites. In a recent study by Barrera *et al*. (Barrera, et al., 2016), transcription factors with point mutations were investigated for changes in binding specificity and binding affinity. HOXD13, a homeodomain associated with developmental regulatory systems has mutations that can cause synpolydactyly. Barrera *et al.* experimentally studied the change in binding specificity and affinity of five mutated variants of HOXD13, I322L, Q325K, Q325R, R306Q, and S316C, using protein binding microarrays (Newburger and Bulyk,

2009). We used homology models to investigate the changes of binding affinity and binding specificity of the five mutations.

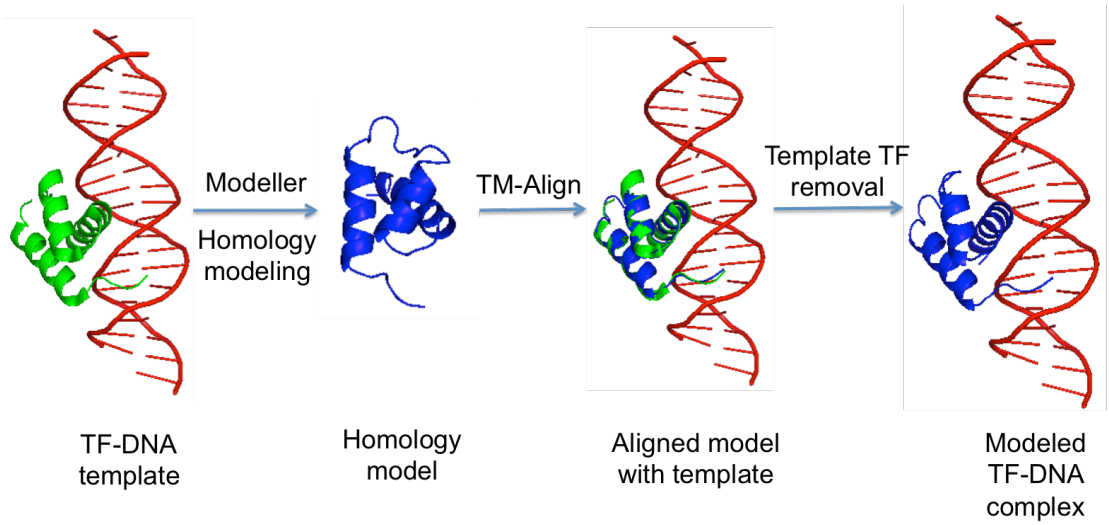

Figure 4.1: Homology modeling of TF-DNA complexes. MODELLER is used to generate homology models of a transcription factor. TM-Align is used to perform a structural alignment of the homology model onto the template TF-DNA complex. The template TF is then removed from the aligned structure, producing in a homology model of a TF-DNA complex.

4.2.2 Homology Modeling

The homology models of transcription factors are generated using MODELLER (Eswar, et al., 2006). Available homeodomain-DNA complex structures from the PDB are used as templates. The five TF-DNA structures with the highest sequence identity and sequence coverage of the query sequence are chosen as templates. Ten models are generated for each TF template and each model is evaluated using PROCHECK (Laskowski, et al., 1993). The five models with the least number of residues in the disallowed region of the Ramachandran plot and the least number of bad contacts are selected for each template, resulting in twenty-five total models. TF-DNA complex structures are then generated using TM-align (Zhang and Skolnick, 2005) to structurally

align a modeled TF on the template TF-DNA complex (Figure 4.1). The template TF is then removed, leaving the TF model with the template DNA. Each of the twenty-five TF models was paired to each of the 5 DNA templates producing 125 TF-DNA models (Figure 4.2).

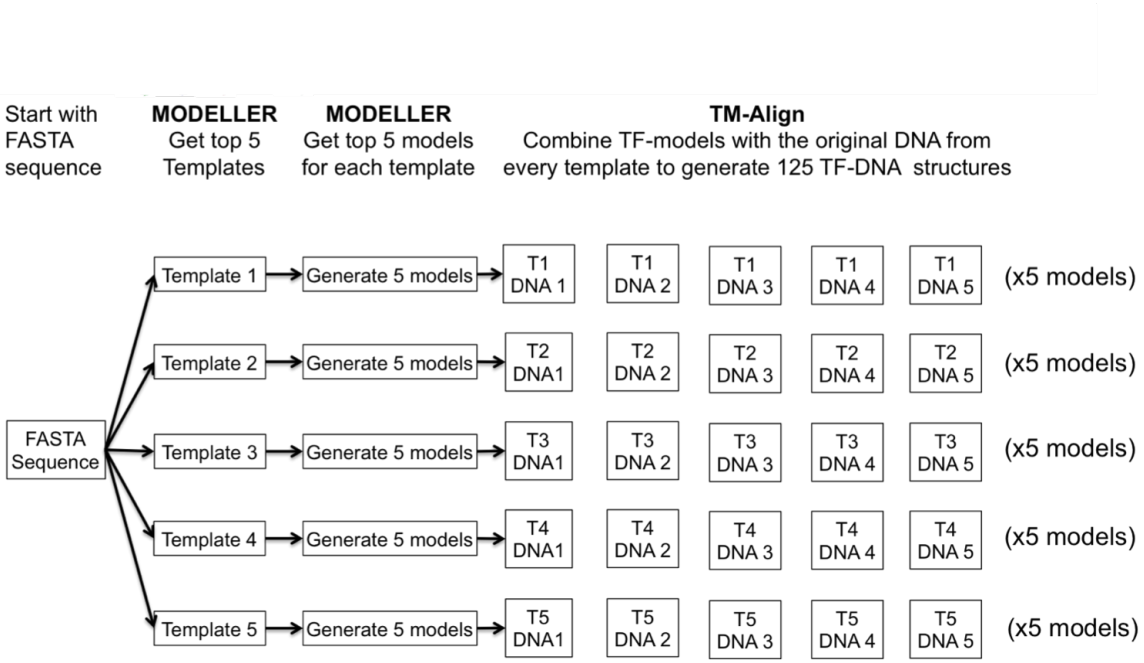

Figure 4.2: Generating 125 TF-DNA homology models.

4.2.3 TF-DNA Complex Model Selection and TF Binding Site Prediction

van der Waals (VDW) energy and TF-DNA contacts are used to assess the quality of the 125 TF-DNA complex models. TF-DNA models with low van der Waals energy are preferred because they have minimal steric clashes. Also, complexes having more amino acids within interacting distance of the major groove of the DNA are preferred as these interactions are important for the energy calculation required for the TFBS prediction. The VDW energy is estimated using our in-house protein DNA docking energy function (Liu, et al., 2008; Liu, et al., 2005). The number of TF-DNA contacts is the number of unique TF-DNA heavy atom pairs within 3.9Å of each other. A

quality score is determined simply by subtracting the number of TF-DNA contacts from the VDW energy. The lower quality scores represent better structures. The energy of the complex models with the lowest quality scores is minimized using GROMACS (Van Der Spoel, et al., 2005) to add hydrogen atoms to the structures and optimize them for the energy calculations. GROMACS is also used to filter out TF-DNA models with other structural issues not captured in the previous evaluation steps. The five structures with the lowest quality scores that underwent successful energy minimization were then revaluated for VDW energy and amino acid-base contacts as the structure has changed after energy minimization.

TFBS prediction is performed on the 3 lowest scoring TF-DNA models using the tiling array and PWM stacking pentamer algorithms and the modified integrative energy functions (Chapter 3). In Chapter 3, we observed that TF-DNA structures with accurate TFBS predictions had similar binding motifs generated by the tiling array and PWM stacking algorithms. Therefore, we used the similarity between the binding motifs generated by the tiling array algorithm and the PWM stacking algorithm as a confidence measure for the TFBS predictions of the TF-DNA models. For this reason, the TFBS predictions of the 3 lowest scoring TF-DNA models are ranked based on the similarity of the PWMs generated by the two pentamer algorithms. To compare the prediction accuracy of homology TF-DNA models, the predicted PWMS using TF-DNA models were compared to their corresponding JASPAR binding motifs and the predicted motifs using the TF-DNA structures from the PDB. We used averaged Kullback-Liebler divergence (Wu, et al., 2001; Xu and Su, 2010) to compare the overall similarity between

the predicted and reference binding sites, and IC-weighted PCC (Persikov and Singh, 2014) to compare the number of correctly predicted columns.

4.2.4 Prediction of binding sites of HOXD13 variants using homology models

To test if homology modeling can be used to predict the change in binding specificity of HOXD13 with point mutations, HOXD13 and HOXD13 mutant models were generated using the homology modeling techniques described above. The predicted binding sites of the variants were compared to their experimentally annotated binding site and the binding site of the wildtype to determine if this method can predict a change in specificity between wildtype and mutants.

4.3 Results and Discussion

The homeodomain binding site predictions using homology models were comparable to the binding site predictions using the native PDB structures (Figure 4.3). The quantitative analysis also reveals that Model 1, the best scoring model, has a lower AKL divergence in seven of the eleven homeodomains. Furthermore, Model 1 has equal or greater correctly predicted binding site positions in eight of the eleven homeodomains. TFBS Predictions using Models 2 and 3 were comparable to the predictions using the native PDB structure. Further developments to the TF-DNA model selection scoring function and increasing the initial number of models generated may improve the probability of selecting near-native TF-DNA models that produce accurate TFBS predictions.

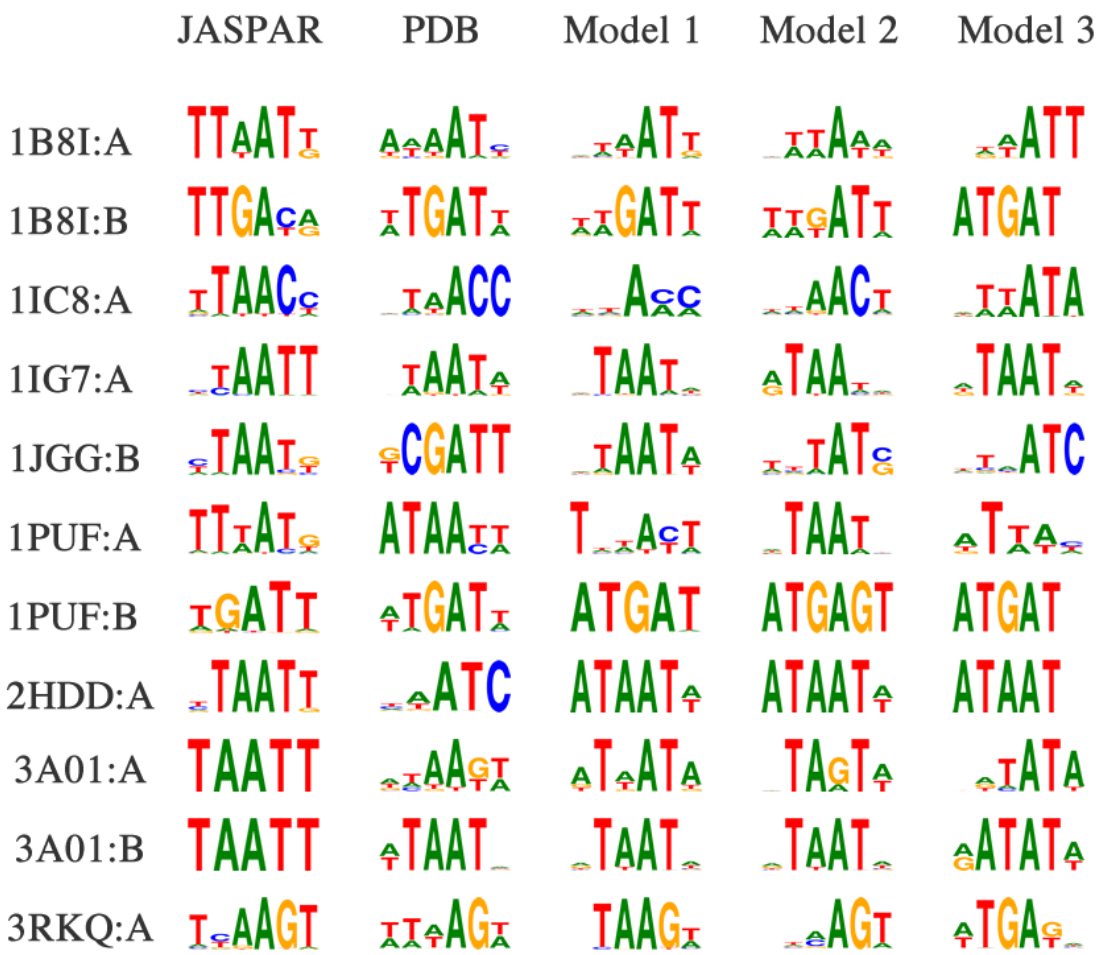

Figure 4.3: Comparison of TF binding site prediction using the native TF-DNA complex structures (PDB) and the top 3 homology models.

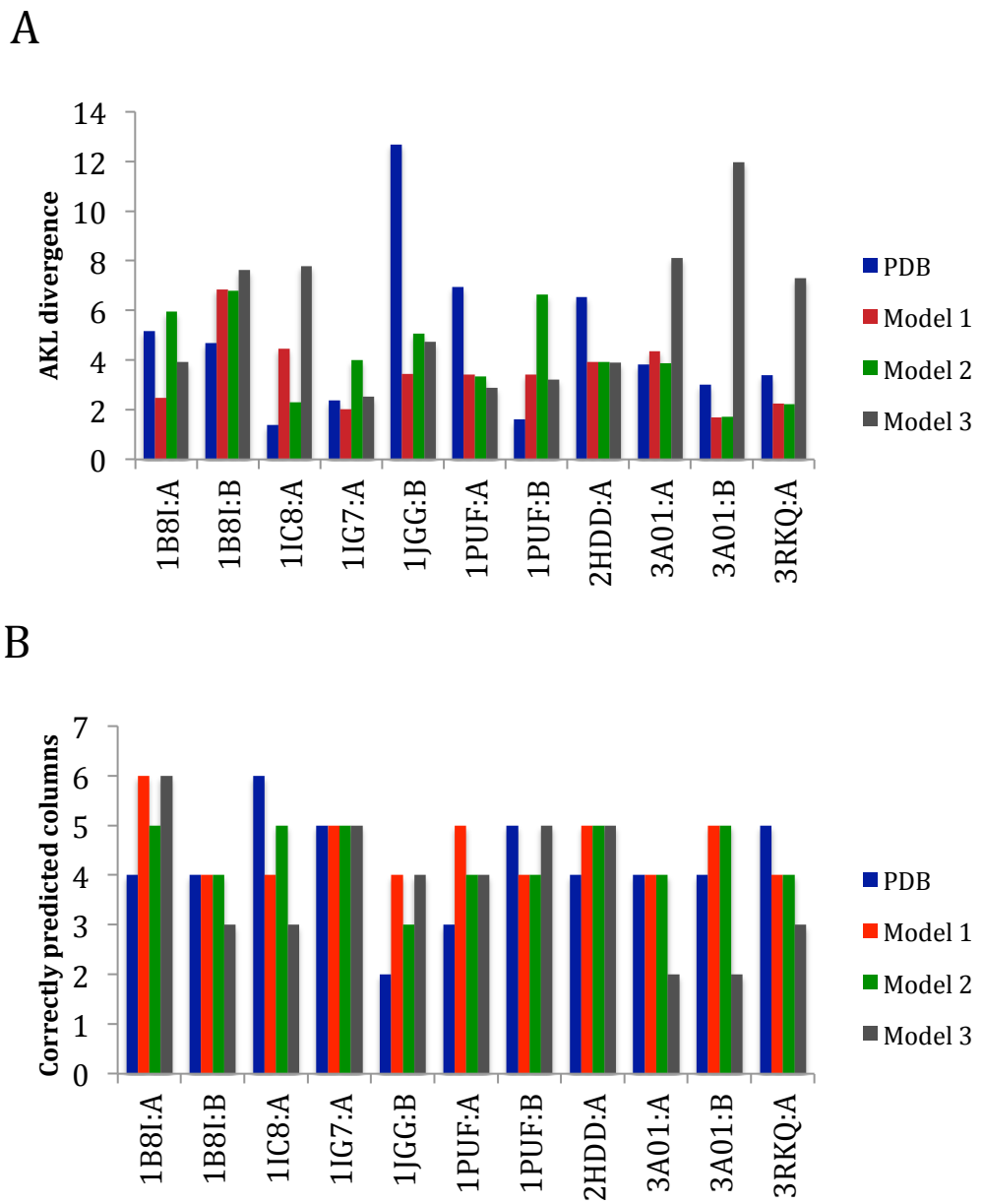

Figure 4.4: Quantitative comparison of homeodomain binding site predictions using the native complex structures from PDB and top 3 homology models. The binding site predictions were compared with their corresponding JASPAR PWMs using (A) AKL divergence and (B) IC-weighted PCC for the number of correctly predicted columns

Barrera *et al*. reported that the I322L and Q325K variants of HOXD13 had decreased binding affinity and changed binding specificity, while the Q325R and R306W variants only had a change in binding specificity. While R306W is not on the recognition helix (Figure 4.5), this variant still affects the binding specificity. The S316C variant does not affect the binding specificity or the affinity (Barrera, et al., 2016). Mutant HOXD13-DNA complex models were generated as shown in Figure 4.1 and Figure 4.2. TFBS predictions were carried out using the modified integrative energy function and the pentamer algorithm, and compared with the binding sites determined experimentally by PBM in UniPROBE (Figure 4.6). The binding site predictions of the variants on the recognition helix (Q325K, Q325R, and I322L) were in agreement with the UniPROBE annotations when compared to the TFBS predictions using this simple modeling approach.

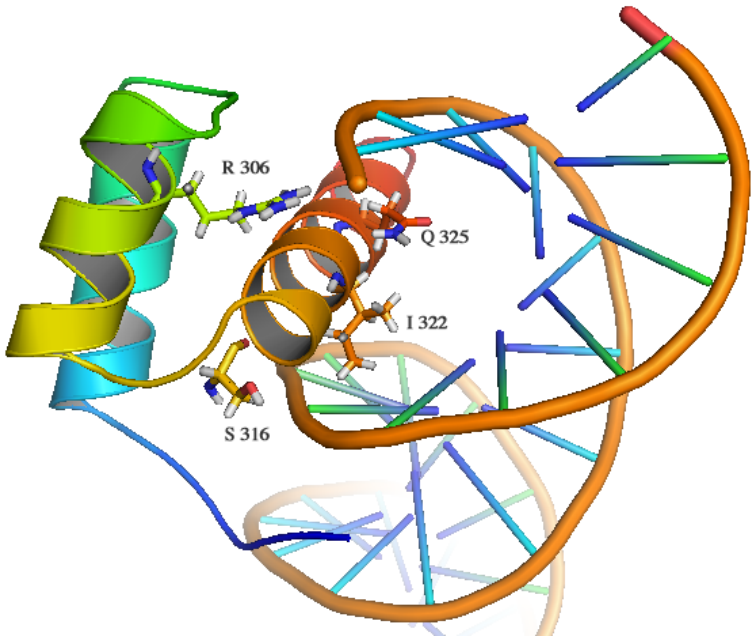

Figure 4.5: TF-DNA model of HOXD13 showing the amino acid positions of the variants' mutations.

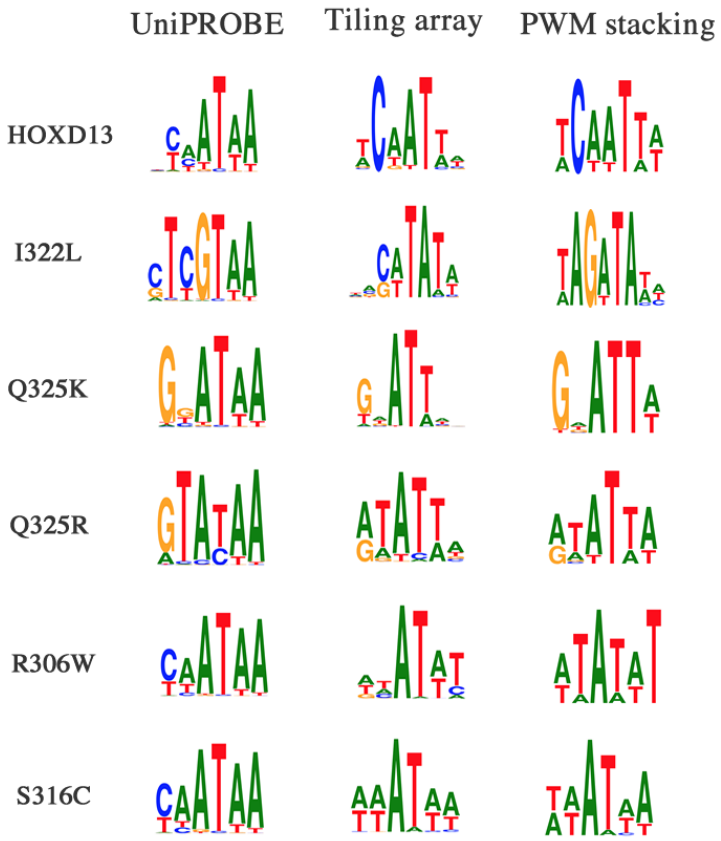

Figure 4.6: TFBS prediction of variants using models generated with MODELLER.

The TFBS predictions of the variants were compared to the wildtype binding sites and their corresponding binding sites from UniPROBE using IC-weighted PCC (Figure 4.7). All four models of the variants that were reported to have a change in specificity showed less matching columns than the wildtype model when the predicted binding motifs were compared to the wildtype binding motif from UniPROBE (Figure 4.7A). The S316C variant and wildtype TFBS predictions have similar matching columns to the wildtype binding motif from UniPROBE. This is in agreement with the observation that this S316C variant does not change binding specificity. Interestingly, the two variants that have lower affinity and specificity have relatively low matching predicted motif columns with both the wildtype and their corresponding binding motif from UniPROBE.

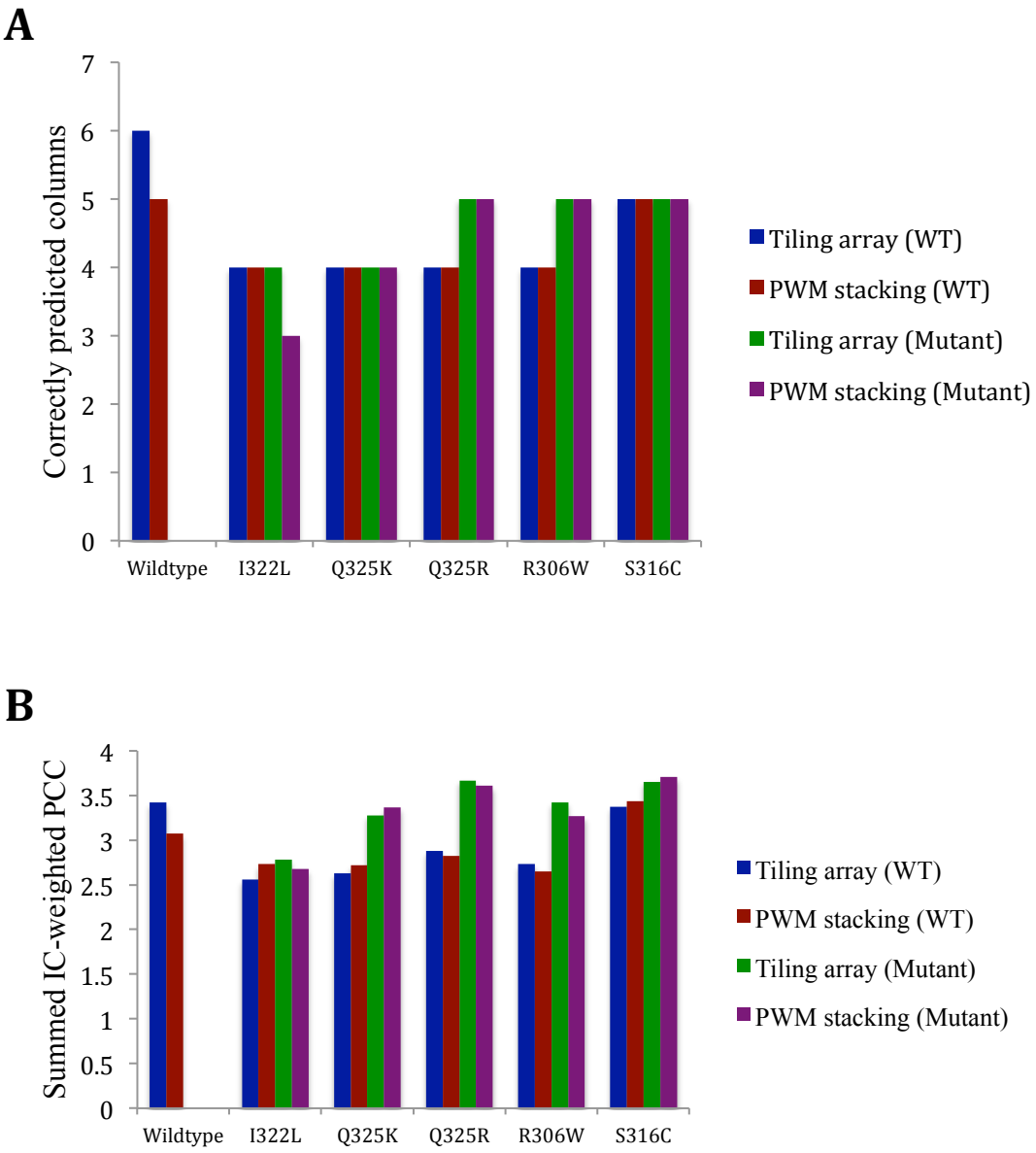

Figure 4.7: Comparison of variant predicted binding sites with UniPROBE binding sites. (A) IC-weighted PCC for the number of correctly predicted columns when compared to the wildtype (blue and red) and their corresponding (green and purple) UniPROBE binding sites. (B) Summed IC-weighted PCC scores of correctly predicted columns.

We discussed in chapter 3 that a correctly predicted column using IC-weighted PCC can range between 1 and 0.25. When we sum the IC-weighted PCCs of the matching columns, we observed that the predicted binding motif of the Q325K variant is more similar to the variant binding motif from UniPROBE than the wildtype. On the other hand, the S316C variant's summed IC-weighted PCC values from the comparison with the wildtype binding site are similar to values of the wildtype model. These results show the potential of this method to study the effects of mutations in transcription factors

4.4. Conclusion

We developed a method to generate TF-DNA homology models and used them for TFBS predictions. This method was tested on a small dataset of homeodomains. The structure-based binding site predictions of the modeled structures were comparable to the binding site predictions of the corresponding experimentally derived crystal structures in the PDB. Furthermore, a preliminary study of HOXD13 and mutants showed that homology models could be used to predict the changes in specificity of homeodomains with point mutations. Future work is needed to improve the methods of selecting TF-DNA models for the application of the energy functions. Furthermore, this method needs to be tested on other TF families.

CHAPTER 5. CONCLUSIONS AND FUTURE DIRECTIONS

In this dissertation, we developed new energy functions and efficient algorithms for improving structure-based transcription factor binding site predictions. We demonstrated that physics-based and knowledge-based potentials could be combined to improve TFBS predictions by maximizing their strengths while compensating for each other's limitations. The importance of π-interactions and hydrogen bonds in TF-DNA recognition and binding specificity is clearly demonstrated in this study. The method is extended to prediction of longer binding sites with a novel pentamer algorithm. It increases the prediction efficiency as well as prediction accuracy. Using homeodomain family as an example, we also showed that homology TF-DNA models can be used to predict TF binding sites with good accuracy

One of the key steps in structure-based prediction of TFBS is the choice of energy functions. Knowledge-based methods are efficient but sometimes generate inaccurate predictions due to mean force potentials and the low count problem. Physics-based methods are typically time consuming and sometimes generate inaccurate predictions because we do not fully understand all the force fields involved in molecular interactions. We developed an integrative energy function that combines a knowledge-based multibody potential with physics-based atomic potentials, specifically, hydrogen bond interactions and π-interactions. Both hydrogen bond interactions and π-interactions contribute to TF-DNA binding specificity. Incorporation of these interactions explicitly

into energy function improves the binding site prediction accuracy of multiple transcription factor families, especially for homeodomains family. It also revealed the importance of π-interactions between aromatic amino acids and bases in transcription factor binding specificity, a finding that has been previously overlooked. Overall, developing the integrative energy function improved the accuracy of structure-based predictions and our understanding of protein-DNA interactions.

Structured-based predictions require an energy function to be applied to protein-DNA complexes with all sequence permutations. As the length of the binding site increases, the number of permutations increases exponentially and thus requires an exponential increase of energy calculations. We addressed this problem by implementing a novel pentamer algorithm that breaks a DNA structure into a series of pentamers. This algorithm requires a linear increase of energy calculations as the binding site length increases instead of an exponential increase of energy calculations required by traditional methods. The pentamer algorithm increases efficiency of structure-based methods without compromising the prediction accuracy. In fact, it improves prediction accuracy for longer binding sites. This new algorithm expands the application of structure-based binding-site prediction to multi-domain transcription factors with longer binding sequences.

Finally, we developed a workflow to apply structure-based prediction methods to transcription factors without known TF-DNA experimental structures. We generated TF-DNA homology models and applied the integrative energy function and pentamer algorithm for binding site predictions. This preliminary study showed potential for future expansion of applying structure based methods to homology models.

Future work needs to be done to expand the application of structured-base prediction of TFBSs using homology models of TFs to other TF families. Furthermore, these approaches can go beyond transcription factor binding site prediction and be applied to the study of the effects of mutations on binding specificity changes and their implications and possible roles in diseases, such as cancer.

REFERENCES

The universal protein resource (UniProt). *Nucleic acids research* 2008;36(Database issue):D190-195.

Abecasis, G.R.*, et al.* A map of human genome variation from population-scale sequencing. *Nature* 2010;467(7319):1061-1073.

Afek, A.*, et al.* Protein-DNA binding in the absence of specific base-pair recognition. *Proc Natl Acad Sci U S A* 2014;111(48):17140-17145.

Alibes, A.*, et al.* Using protein design algorithms to understand the molecular basis of disease caused by protein-DNA interactions: the Pax6 example. *Nucleic acids research* 2010;38(21):7422-7431.

Aloy, P.*, et al.* Modelling repressor proteins docking to DNA. *Proteins* 1998;33(4):535-549.

Ayton, G.S., Noid, W.G. and Voth, G.A. Multiscale modeling of biomolecular systems: in serial and in parallel. *Current opinion in structural biology* 2007;17(2):192-198.

Badis, G.*, et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* 2009;324(5935):1720-1723.

Baker, C.M. and Grant, G.H. Role of aromatic amino acids in protein-nucleic acid recognition. *Biopolymers* 2007;85(5-6):456-470.

Baker, D. and Sali, A. Protein structure prediction and structural genomics. *Science* 2001;294(5540):93-96.

Barrera, L.A.*, et al.* Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science* 2016;351(6280):1450-1454.

Beierlein, F.R., Kneale, G.G. and Clark, T. Predicting the effects of basepair mutations in DNA-protein complexes by thermodynamic integration. *Biophysical journal* 2011;101(5):1130-1138.

Benos, P.V., Lapedes, A.S. and Stormo, G.D. Is there a code for protein-DNA recognition? Probab(ilistical)ly. *BioEssays : news and reviews in molecular, cellular and developmental biology* 2002;24(5):466-475.

Berman, H.M.*, et al.* The Protein Data Bank. *Nucleic acids research* 2000;28(1):235-242.

Boeva, V.*, et al.* De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic acids research* 2010;38(11):e126.

Borneman, A.R., *et al.* Divergence of transcription factor binding sites across related yeast species. *Science* 2007;317(5839):815-819.

Bradley, P., *et al.* Free modeling with Rosetta in CASP6. *Proteins* 2005;61 Suppl 7:128-134.

Bulyk, M.L. Computational prediction of transcription-factor binding site locations. *Genome biology* 2003;5(1):201.

Bulyk, M.L., *et al.* Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nature biotechnology* 1999;17(6):573-577.

Burghardt, T.P., *et al.* Cation-pi interaction in a folded polypeptide. *Biopolymers* 2002;63(4):261-272.

Christensen, R.G., *et al.* Recognition models to predict DNA-binding specificities of homeodomain proteins. *Bioinformatics* 2012;28(12):i84-89.

Corona, R.I. and Guo, J.T. Statistical analysis of structural determinants for protein-DNA-binding specificity. *Proteins* 2016;84(8):1147-1161.

Crocker, J., *et al.* Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* 2015;160(1-2):191-203.

Crooks, G.E., *et al.* WebLogo: a sequence logo generator. *Genome Res* 2004;14(6):1188-1190.

D'Elia, A.V., *et al.* Missense mutations of human homeoboxes: A review. *Human mutation* 2001;18(5):361-374.

Dahiyat, B.I., Gordon, D.B. and Mayo, S.L. Automated design of the surface positions of protein helices. *Protein science : a publication of the Protein Society* 1997;6(6):1333-1337.

Dahiyat, B.I. and Mayo, S.L. De novo protein design: fully automated sequence selection. *Science* 1997;278(5335):82-87.

Desjarlais, J.R. and Berg, J.M. Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proceedings of the National Academy of Sciences of the United States of America* 1992;89(16):7345-7349.

Dominguez, C., Boelens, R. and Bonvin, A. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* 2003;125(7):1731-1737.

Donald, J.E., Chen, W.W. and Shakhnovich, E.I. Energetics of protein-DNA interactions. *Nucleic acids research* 2007;35(4):1039-1047.

Dowell, R.D. Transcription factor binding variation in the evolution of gene regulation. *Trends in genetics : TIG* 2010;26(11):468-475.

Ellington, A.D. and Szostak, J.W. In vitro selection of RNA molecules that bind specific ligands. *Nature* 1990;346(6287):818-822.

Eswar, N.*, et al.* Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* 2006;Chapter 5:Unit 5 6.

Farrel, A., Murphy, J. and Guo, J.T. Structure-based prediction of transcription factor binding specificity using an integrative energy function. *Bioinformatics* 2016;32(12):i306-i313.

Flores, S.C.*, et al.* Multiscale modeling of macromolecular biosystems. *Briefings in bioinformatics* 2012;13(4):395-405.

Foos, N.*, et al.* A flexible extension of the Drosophila ultrabithorax homeodomain defines a novel Hox/PBC interaction mode. *Structure* 2015;23(2):270-279.

Friedman, Y.E. and O'Brian, M.R. A novel DNA-binding site for the ferric uptake regulator (Fur) protein from Bradyrhizobium japonicum. *J Biol Chem* 2003;278(40):38395-38401.

Furey, T.S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature reviews. Genetics* 2012;13(12):840-852.

Gallivan, J.P. and Dougherty, D.A. Cation-pi interactions in structural biology. *Proceedings of the National Academy of Sciences of the United States of America* 1999;96(17):9459-9464.

Gebelein, B.*, et al.* Specificity of Distalless repression and limb primordia development by abdominal Hox proteins. *Developmental cell* 2002;3(4):487-498.

Gehring, W.J., Affolter, M. and Burglin, T. Homeodomain proteins. *Annual review of biochemistry* 1994;63:487-526.

Georgiev, S.*, et al.* Evidence-ranked motif identification. *Genome biology* 2010;11(2):R19.

Gopal, S.M.*, et al.* PRIMO/PRIMONA: a coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. *Proteins* 2010;78(5):1266-1281.

Gordan, R.*, et al.* Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell reports* 2013;3(4):1093-1104.

Gromiha, M.M., Santhosh, C. and Ahmad, S. Structural analysis of cation-pi interactions in DNA binding proteins. *International journal of biological macromolecules* 2004;34(3):203-211.

Guo, Y., Mahony, S. and Gifford, D.K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS computational biology* 2012;8(8):e1002638.

Gupta, S.*, et al.* Quantifying similarity between motifs. *Genome biology* 2007;8(2):R24.

Harr, R., Haggstrom, M. and Gustafsson, P. Search algorithm for pattern match analysis of nucleic acid sequences. *Nucleic acids research* 1983;11(9):2943-2957.

Havranek, J.J., Duarte, C.M. and Baker, D. A simple physical model for the prediction and design of protein-DNA interactions. *Journal of molecular biology* 2004;344(1):59-70.

Hu, M.*, et al.* On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic acids research* 2010;38(7):2154-2167.

Jacobs, D.J.*, et al.* Protein flexibility predictions using graph theory. *Proteins* 2001;44(2):150-165.

Jiang, L.*, et al.* A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins* 2005;58(4):893-904.

Johnson, D.S.*, et al.* Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007;316(5830):1497-1502.

Joyce, A.P.*, et al.* Structure-based modeling of protein: DNA specificity. *Briefings in functional genomics* 2015;14(1):39-49.

Kollman, P.A.*, et al.* Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts of Chemical Research* 2000;33(12):889-897.

Kulakovskiy, I.V.*, et al.* Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics* 2010;26(20):2622-2623.

Laskowski, R.A.*, et al.* PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst* 1993;26:283-291.

Lawrence, C.E.*, et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993;262:208-214.

Lawrence, C.E. and Reilly, A.A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* 1990;7(1):41-51.

Lemon, B. and Tjian, R. Orchestrated response: a symphony of transcription factors for gene control. *Genes & development* 2000;14(20):2551-2569.

Levine, M. and Tjian, R. Transcription regulation and animal diversity. *Nature* 2003;424(6945):147-151.

Li, S. and Bradley, P. Probing the role of interfacial waters in protein-DNA recognition using a hybrid implicit/explicit solvation model. *Proteins* 2013;81(8):1318-1329.

Li, Y.*, et al.* Modeling of the water network at protein-RNA interfaces. *Journal of chemical information and modeling* 2011;51(6):1347-1352.

Liu, L.A. and Bader, J.S. Structure-based ab initio prediction of transcription factor-binding sites. *Methods in molecular biology* 2009;541:23-41.

Liu, L.A. and Bradley, P. Atomistic modeling of protein-DNA interaction specificity: progress and applications. *Current opinion in structural biology* 2012;22(4):397-405.

Liu, Z.*, et al.* Structure-based prediction of transcription factor binding sites using a protein-DNA docking approach. *Proteins* 2008;72(4):1114-1124.

Liu, Z.*, et al.* Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential. *Nucleic acids research* 2005;33(2):546-558.

Lu, X.J. and Olson, W.K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic acids research* 2003;31(17):5108-5121.

Lu, X.J. and Olson, W.K. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature protocols* 2008;3(7):1213-1227.

Luscombe, N.M., Laskowski, R.A. and Thornton, J.M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic acids research* 2001;29(13):2860-2874.

Luscombe, N.M. and Thornton, J.M. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *Journal of molecular biology* 2002;320(5):991-1009.

MacKerell, A.D. and Banavali, N.K. All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. *Journal of computational chemistry* 2000;21(2):105-120.

Mandel-Gutfreund, Y. and Margalit, H. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic acids research* 1998;26(10):2306-2312.

Mandel-Gutfreund, Y., Schueler, O. and Margalit, H. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *Journal of molecular biology* 1995;253(2):370-382.

Marvin6.1.4. 2013. ChemAxon (http://www.chemaxon.com)

Mathelier, A.*, et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic acids research* 2016;44(D1):D110-115.

Matthews, B.W. Protein-DNA interaction. No code for recognition. *Nature* 1988;335(6188):294-295.

McGaughey, G.B., Gagne, M. and Rappe, A.K. pi-Stacking interactions. Alive and well in proteins. *J Biol Chem* 1998;273(25):15458-15463.

Mecozzi, S., West, A.P., Jr. and Dougherty, D.A. Cation-pi interactions in aromatics of biological and medicinal interest: electrostatic potential surfaces as a useful qualitative guide. *Proceedings of the National Academy of Sciences of the United States of America* 1996;93(20):10566-10571.

Merabet, S.*, et al.* A unique Extradenticle recruitment mode in the Drosophila Hox protein Ultrabithorax. *Proceedings of the National Academy of Sciences of the United States of America* 2007;104(43):16946-16951.

Michael Gromiha, M.*, et al.* Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *Journal of molecular biology* 2004;337(2):285-294.

Morozov, A.V.*, et al.* Protein-DNA binding specificity predictions with structural models. *Nucleic acids research* 2005;33(18):5781-5798.

Muller, P.A. and Vousden, K.H. p53 mutations in cancer. *Nature cell biology* 2013;15(1):2-8.

Mulligan, M.E.*, et al.* Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity. *Nucleic acids research* 1984;12(1 Pt 2):789-800.

Newburger, D.E. and Bulyk, M.L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic acids research* 2009;37(Database issue):D77-82.

Noyes, M.B.*, et al.* Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* 2008;133(7):1277-1289.

Oliphant, A.R., Brandl, C.J. and Struhl, K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Molecular and cellular biology* 1989;9(7):2944-2949.

Orenstein, Y. and Shamir, R. A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic acids research* 2014;42(8):e63.

Pabo, C.O. and Nekludova, L. Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *Journal of molecular biology* 2000;301(3):597-624.

Passner, J.M.*, et al.* Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. *Nature* 1999;397(6721):714-719.

Persikov, A.V. and Singh, M. De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic acids research* 2014;42(1):97-108.

Pettersen, E.F.*, et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry* 2004;25(13):1605-1612.

Poulain, P.*, et al.* Insights on protein-DNA recognition by coarse grain modelling. *Journal of computational chemistry* 2008;29(15):2582-2592.

Ren, B.*, et al.* Genome-wide location and function of DNA binding proteins. *Science (New York, N.Y* 2000;290(5500):2306-2309.

Roberts, V.A.*, et al.* Predicting protein-DNA interactions by full search computational docking. *Proteins* 2013;81(12):2106-2118.

Robertson, T.A. and Varani, G. An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure. *Proteins* 2007;66(2):359-374.

SantaLucia, J., Jr., Allawi, H.T. and Seneviratne, P.A. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* 1996;35(11):3555-3562.

Schmidt, D.*, et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 2010;328(5981):1036-1040.

Schneider, B., Cohen, D. and Berman, H.M. Hydration of DNA bases: analysis of crystallographic data. *Biopolymers* 1992;32(7):725-750.

Schneider, T.D. and Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic acids research* 1990;18(20):6097-6100.

Schueler-Furman, O.*, et al.* Progress in modeling of protein structures and interactions. *Science* 2005;310(5748):638-642.

Seeliger, D.*, et al.* Towards computational specificity screening of DNA-binding proteins. *Nucleic acids research* 2011;39(19):8281-8290.

Seeman, N.C., Rosenberg, J.M. and Rich, A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proceedings of the National Academy of Sciences of the United States of America* 1976;73(3):804-808.

Seghouane, A.K. and Amari, S. The AIC criterion and symmetrizing the Kullback-Leibler divergence. *IEEE transactions on neural networks* 2007;18(1):97-106.

Siggers, T.W. and Honig, B. Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic acids research* 2007;35(4):1085-1097.

Slattery, M.*, et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci* 2014;39(9):381-399.

Staden, R. Computer methods to locate signals in nucleic acid sequences. *Nucleic acids research* 1984;12(1 Pt 2):505-519.

Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)* 2000;16(1):16-23.

Stormo, G.D. and Zhao, Y. Determining the specificity of protein-DNA interactions. *Nature reviews. Genetics* 2010;11(11):751-760.

Suzuki, M.*, et al.* DNA recognition code of transcription factors. *Protein engineering* 1995;8(4):319-328.

Suzuki, M. and Yagi, N. DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proceedings of the National Academy of Sciences of the United States of America* 1994;91(26):12357-12361.

Takeda, T., Corona, R.I. and Guo, J.T. A knowledge-based orientation potential for transcription factor-DNA docking. *Bioinformatics* 2013;29(3):322-330.

Thorpe, M.F.*, et al.* Protein flexibility and dynamics using constraint theory. *J Mol Graph Model* 2001;19(1):60-69.

Thorpe, M.F.*, et al.* Protein flexibility and dynamics using constraint theory. *Journal of Molecular Graphics & Modelling* 2001;19:60-69.

Tucker-Kellogg, L.*, et al.* Engrailed (Gln50-->Lys) homeodomain-DNA complex at 1.9 A resolution: structural basis for enhanced affinity and altered specificity. *Structure* 1997;5(8):1047-1054.

Tuerk, C. and Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 1990;249(4968):505-510.

Tupler, R., Perini, G. and Green, M.R. Expressing the human genome. *Nature* 2001;409(6822):832-833.

Van Der Spoel, D.*, et al.* GROMACS: fast, flexible, and free. *Journal of computational chemistry* 2005;26(16):1701-1718.

van Dijk, M.*, et al.* Solvated protein-DNA docking using HADDOCK. *Journal of biomolecular NMR* 2013;56(1):51-63.

van Zundert, G.C.*, et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of molecular biology* 2015.

von Hippel, P.H. and Berg, O.G. Facilitated target location in biological systems. *The Journal of Biological Chemistry* 1989;264:675-678.

Vreven, T., Hwang, H. and Weng, Z. Integrating atom-based and residue-based scoring functions for protein-protein docking. *Protein science : a publication of the Protein Society* 2011;20(9):1576-1586.

Wilson, K.A., Kellie, J.L. and Wetmore, S.D. 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* 2014;42(10):6726-6741.

Wilson, K.A. and Wetmore, S.D. A Survey of DNA–Protein π–Interactions: A Comparison of Natural Occurrences and Structures, and Computationally Predicted Structures and Strengths. In: Scheiner, S., editor, *Noncovalent Forces, Challenges and Advances in Computational Chemistry and Physics 19*. Springer International Publishing Switzerland 2015; 2015.

Wintjens, R.*, et al.* Contribution of cation-pi interactions to the stability of protein-DNA complexes. *Journal of molecular biology* 2000;302(2):395-410.

Word, J.M.*, et al.* Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of molecular biology* 1999;285(4):1735-1747.

Wray, G.A. The evolutionary significance of cis-regulatory mutations. *Nature reviews. Genetics* 2007;8(3):206-216.

Wu, C.H.*, et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic acids research* 2006;34(Database issue):D187-191.

Wu, J.*, et al.* High performance transcription factor-DNA docking with GPU computing. *Proteome science* 2012;10 Suppl 1:S17.

Wu, T.J., Hsieh, Y.C. and Li, L.A. Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics* 2001;57(2):441-448.

Xu, B.*, et al.* A structural-based strategy for recognition of transcription factor binding sites. *PloS one* 2013;8(1):e52460.

Xu, B.*, et al.* An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. *Proteins* 2009;76(3):718-730.

Xu, M. and Su, Z. A novel alignment-free method for comparing transcription factor binding site motifs. *PloS one* 2010;5(1):e8797.

Zhang, C.*, et al.* A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J Med Chem* 2005;48(7):2325-2335.

Zhang, Y. and Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* 2005;33(7):2302-2309.

Zhou, T.*, et al.* Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc Natl Acad Sci U S A* 2015;112(15):4654-4659.