Bioinformatics, 32, 2016, i306–i313
doi: 10.1093/bioinformatics/btw264
ISMB 2016

Structure-based prediction of transcription factor binding specificity using an integrative energy function

Alvin Farrel, Jonathan Murphy, and Jun-tao Guo*

Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

*To whom correspondence should be addressed

Abstract

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 article, we present a structure-based method for computational prediction of TFBSs using a novel, integrative energy (IE) function. The new energy function combines a multibody (MB) knowledge-based potential and two 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. We applied the new energy function to the TFBS prediction using a non-redundant dataset that consists of TFs from 12 different families. Our results show that the new IE function improves the prediction accuracy over the knowledge-based, statistical potentials, especially for homeodomain TFs, the second largest TF family in mammals.

Contact: jguo4@uncc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transcription factors (TFs) regulate gene expression by interacting with specific DNA sequences called transcription factor binding sites (TFBSs) (Lemon and Tjian, 2000; Levine and Tjian, 2003). Identification of TFBSs on a genomic scale, a crucial step in genomic annotation and in deciphering transcription regulatory networks, remains a great challenge in post-genomic bioinformatics. With the rapid increase of available genomic data, effective sequence-based methods for TFBS predictions have been developed (Stormo, 2000). However, one issue of sequence-based methods is the high number of false-positive results, especially when the binding signal is weak or the TF's DNA-binding site is significantly different from the consensus sequence.

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 biophysical interactions. Whereas the 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 TFs bind at these locations and sequences. Moreover, understanding the mechanisms and effects of mutations on gene expression and diseases can guide rational design of therapeutic agents. Although research on protein–DNA recognition began in the 1970s (Seeman *et al.*, 1976), structure-based methods for prediction of TFBSs were not developed until years ago when the high-resolution protein structures became available in the Protein Data Bank (PDB) (Berman *et al.*, 2000).

One of the major issues in structure-based TFBS prediction is the scoring function for evaluating binding affinity or binding energy between proteins and DNA. There are two major types of energy functions for studying protein–DNA interactions, the physics-based molecular mechanics force fields and the knowledge-based statistical potentials. Physics-based energy functions consist of physicochemical interactions including electrostatic interactions, van der Waals (VDW) forces, solvation energy and others (Liu and Bradley, 2012). 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). Besides the general terms such as VDW and electrostatic interactions that include hydrogen bonds, π-cation and π–π interactions have also been studied in protein–DNA recognition. It was previously thought that these interactions have a primary role in establishing the stability of the protein–DNA complexes, but new data suggest that these interactions may have a bigger role in protein–DNA recognition (Baker and Grant, 2007; Luscombe *et al.*, 2001). Although the physics-based energies can accurately describe protein–DNA interactions, they are computationally expensive. In addition, protein–DNA complexes are intrinsically dynamic. A protein–DNA complex structure used for TFBS prediction only represents a snapshot of the structure with a large number of possible conformations. Different conformations may result in different TFBS predictions because physics-based energy functions are sensitive to conformational changes.

Knowledge-based potentials are derived from statistical analysis of a set of known, non-redundant protein–DNA complexes. They are often preferred because they are relatively simple, less computationally expensive and less sensitive to conformational changes while producing results comparable with 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). Recent residue-level potentials have proven to work well in protein–DNA interaction studies (Liu *et al.*, 2005; Takeda *et al.*, 2013). However, 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. Though the hydrogen bonds are implicitly captured in knowledge-based potentials, they are 'averaged' with the non-specific interactions. The other is caused 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 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 (Corona and Guo, 2016). We hypothesize that π interactions between aromatic residues and DNA bases contribute to TF–DNA binding specificity. However, 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 MB 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 TFs. The results show that TFBS prediction using our new IE function improves accuracy when compared with other residue-level and atomic-level knowledge-based potentials.

2 Methods

IE function

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

$$\mathrm{E_{Total}} = \mathrm{W_{MB}E_{MB}} + \mathrm{W_{HB}E_{HB}} + \mathrm{W_{\pi}E_{\pi}} \quad (1)$$

where E_{Total} is the total energy, E_{MB}, E_{HB}, and E_{π} represent the normalized MB energy, hydrogen bond energy and π interaction energy, respectively, and W_{MB}, W_{HB} and W_{π} are weights for each term. As 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 owing to its important contribution to protein–DNA binding specificity (Luscombe *et al.*, 2001). The weight for π interaction is half the weight of the MB 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.

The knowledge-based, MB statistical potential

We have previously developed two residue-level knowledge-based potentials, a MB potential and an orientation potential, for assessing protein–DNA interactions in TFBS prediction and protein–DNA docking (Liu *et al.*, 2005; Takeda *et al.*, 2013). The MB potential uses structural environment for accurate assessment, whereas the orientation potential uses both distance and angle information to better capture hydrogen bond information implicitly. As we propose an explicit hydrogen bond term in our new IE function to capture the key hydrogen bond interactions, we chose the MB 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 MB potential for TF–DNA docking (Liu *et al.*, 2005; Takeda *et al.*, 2013), the MB potential predicts TF–DNA binding motifs better than the orientation potential possibly because of the capture of interaction context, as structure-based prediction of TFBSs and protein–DNA docking are two different computational problems (data not shown). The MB 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).

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. (1997; Thorpe *et al.*, 2001).

$$E_{HB} = V_0\left\{5\left(\frac{d_0}{d}\right)^{12} - 6\left(\frac{d_0}{d}\right)^{10}\right\}F(\theta, \phi, \varphi) \quad (2)$$

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), to calculate the hydrogen bond energy between amino acids and nucleotides in the protein–DNA complexes (Abecasis *et al.*, 2010).

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, $\pi - \pi$

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 (Fig. 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 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 1) were determined using MarvinSketch 6.1.4, a software package from Chemaxon (Marvin6.1.4, 2013).

Mecozzi *et al*. calculated the binding energies of benzene as well as other aromatic compounds of biological and medicinal interest (Mecozzi *et al*., 1996). 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).

The electrostatic potential was then calculated using:

$$\Delta E_{ac} = \frac{k_e N_A q_a q_c}{\varepsilon r} \tag{3}$$

where ΔE_{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 electrostatic potential is then converted from joules/mol to kcal/mol using the conversion factor of 2.39×10^{-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 (4)).

$$E_\pi = \sum_a^{N_a} \Delta E_{ac} \tag{4}$$

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 and ΔE_{ac} is the energy between an atom a on the base and the electron cloud c.

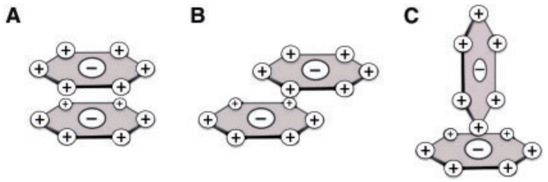

Fig. 1. Geometries of π Interactions between aromatic structures. (**A**) Parallel stacked geometry, the least energetically favorable geometry. (**B**) Parallel displaced geometry, the most energetically favorable geometry. (**C**) T-shaped or edge-to-face geometry, more energetically favorable than the parallel stacked geometry but less favorable than the parallel displaced geometry

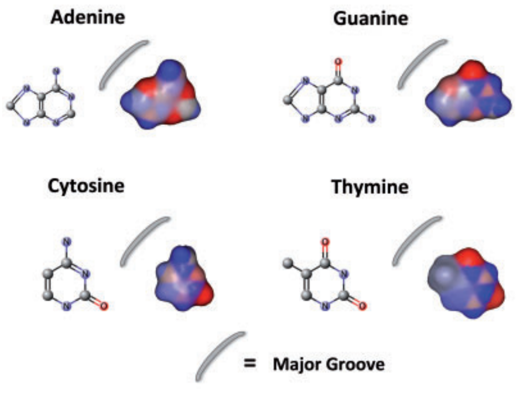

Fig. 2. Electronic landscape of the bases. Charge distributions of the four bases in the major groove. The blue regions represent partial positive charges, whereas the red regions represent partial negative charges. The gray regions are neutral. MarvinSketch 6.1.4, a software package from Chemaxon (Marvin6.1.4), was used to generate the electronic landscape and calculate the charges on the atoms

Prediction algorithm

The flowchart for the structure-based TFBS prediction is shown in Figure 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*., 2004). 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 bp, 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, 65 536 sequences, using 3DNA (Lu and Olson, 2003). The three energy terms were then calculated for each of the 65 536 TF–DNA complex structures. The score for each of the three terms, MB energy, hydrogen bond energy and π interaction energy, was normalized using Equation (5):

$$E_N = \frac{E - E_{max}}{E_{min} - E_{max}} \tag{5}$$

where E_N is the normalized energy, E is the energy for a specific complex with a sequence and E_{max} and E_{min} are the maximum and

Table 1. Quantified charges on nucleotide major groove atoms (blue and red regions on the electronic landscapes illustrated in Fig. 2)

Atom	A	C	G	T
N/O	0.34(N6)	0.34(N4)	−0.44(O6)	−0.478(O4)
C5	−0.015	0.066	0.007	0.087
C6	–	0.085	–	0.096
N7	−0.21	–	−0.215	–
C8	0.115	–	0.107	–

Table 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

minimum energies in the set of 65 536 TF–DNA complexes, respectively. The total energy is then calculated using Equation (1). The distribution of IE 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 (Fig. 3).

PWM prediction and validation

PWMs are generated using the selected sequences from the distribution of IE scores. First, a 4 × 8 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 JASPAR PWMs (Mathelier *et al.*, 2014). We used three quantitative measures to score the similarity between the predicted and the reference PWMs: Chi-square test, averaged Kullback–Leibler (AKL) divergence (Wu *et al.*, 2001; Xu and Su, 2010) and Euclidean distance (Blaisdell, 1986; Xu and Su, 2010). 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 reference PWMs. These columns represent aligned base positions in the binding motif. A predicted column is considered to be a correct prediction if the IC-weighted PCC between the corresponding predicted and reference columns is at least 0.25 (Persikov and Singh, 2014).

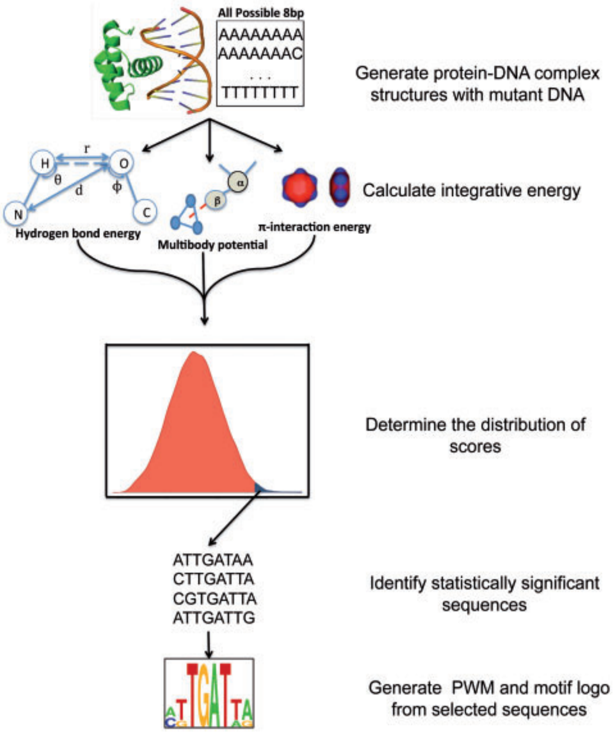
All Possible 8bp
AAAAAAAA
AAAAAAAC
. . .
TTTTTTTT
Generate protein-DNA complex structures with mutant DNA
H
r
O
θ
d
ϕ
N
C
α
β
Calculate integrative energy
Hydrogen bond energy
Multibody potential
π-interaction energy
Determine the distribution of scores
ATTGATAA
CTTGATTA
CGTGATTA
ATTGATTG
Identify statistically significant sequences
Generate PWM and motif logo from selected sequences

Fig. 3. Flowchart for structure-based TFBS prediction

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 PDB with corresponding JASPAR PWMs. These structures were solved by X-ray crystallography with a resolution <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 from 12 TF families: helix loop helix, zinc fingers, homeodomains, leucine zippers, signal transducer and activator of transcription 1 (STAT1), fork head, ETS family, high mobility group (HMG), NFAT, SMAD, P53 DNA binding domain and runt domains. The PDB chains in the dataset are 1AM9:A, 1BC8:C, 1BF5:A, 1DSZ:A, 1GU4:A, 1H8A:C, 1H9D:A, 1JNM:A, 1LLM:C, 1NKP:A, 1NLW:A, 1NLW:B, 1OZJ:A, 1P7h:M, 1PUF:A, 1PUF:B, 1T2K:A, 2A07:A, 2AC0:A, 2DRP:A, 2HDD:A, 2QL2:A, 2QL2:B, 2UZK:A, 2YPA:A, 3F27:A, 4F6M:A, 4HN5:A and 4IQR:A.

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 TFs (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.*, 1994). 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 <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 IE function to see if we can accurately predict different binding sites for highly similar proteins.

3 Results

We applied the new IE function to the prediction of TFBSs using the non-redundant dataset of 29 TF–DNA complex structures and compared the prediction with MB potential and DDNA3, a knowledge-based atomic-level protein–DNA interaction potential (Zhang *et al.*, 2005). Five examples of the predicted TF-binding motifs and the corresponding JASPAR motifs are shown in Figure 4A (all 29 predicted motifs are available in Supplementary Fig. S1). We also applied three different quantitative methods, Chi-square test, AKL divergence and Euclidean distance, to compare the prediction accuracy as described in Section 2. The lower the AKL divergence value, the more is the similarity between the predicted PWMs and JASPAR PWMs. Figure 4B 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 (data not shown). As shown in Figure 4, 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 where 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, whereas 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 between IE and DDNA3, respectively, suggesting that the improvements are statistically significant.

Zinc fingers and homeodomains represent the two largest and extensively studied TF families. In our non-redundant dataset, we found six zinc finger chains (Fig. 5) and three homeodomains (Fig. 6). 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, whereas the other three have no significant differences (1DSZ:A, 4HN5:A and 4IQR:A, Figs 4 and 5).

Homeodomains are the second largest TF family (Tupler *et al.*, 2001). Each homeodomain recognizes a variation of the typical TAAT core binding site. There were three homeodomains in the non-redundant dataset. Figure 6 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 4B. In all three cases, predictions using the IE consistently outperform both MB and DDNA3 potentials.

Homeodomain is a well-studied, highly conserved structural domain for DNA binding. As 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

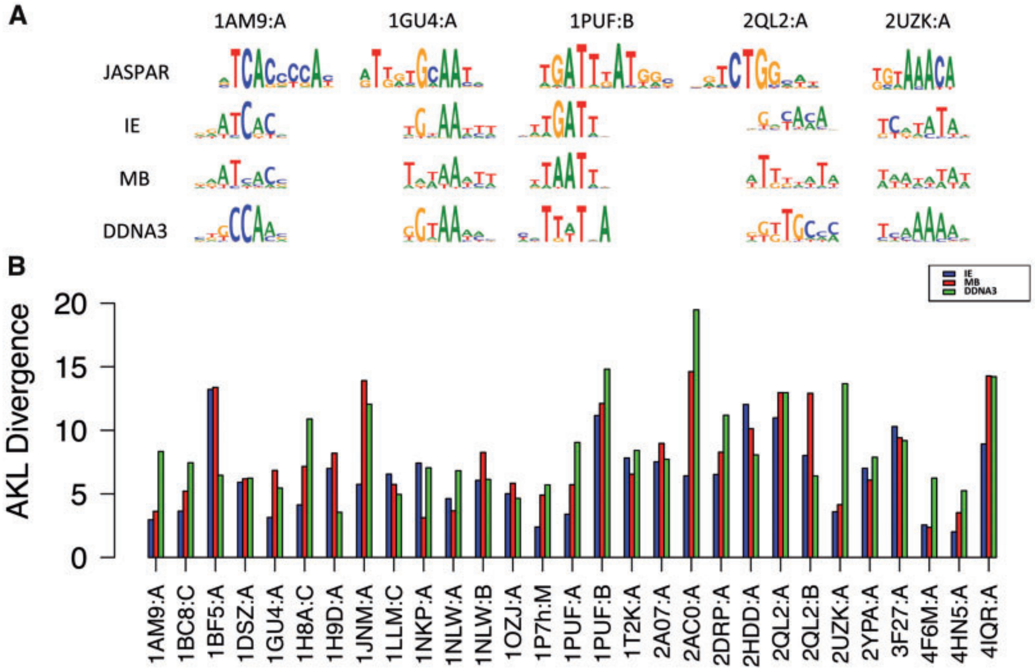

Fig. 4. Comparison of IE prediction accuracy with MB and DDNA3 energies. (**A**) Five examples of the non-redundant dataset; (**B**) AKL divergence of the predicted PWMs with JASPAR PWMs using the integrative function (IE: blue), multibody potential (MB: red) and DDNA3 (green)

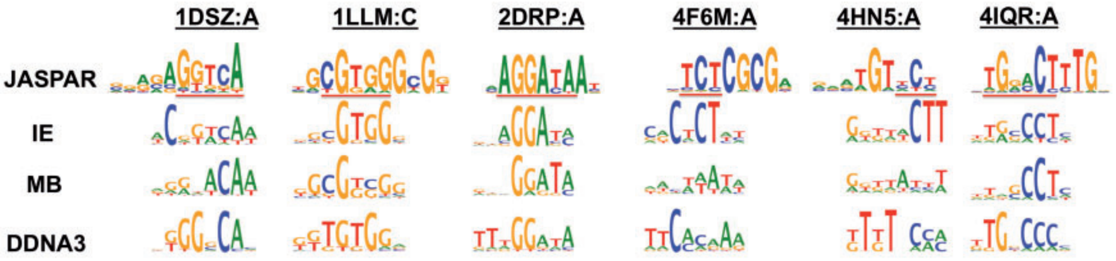

Fig. 5. Comparison of zinc finger binding site predictions. Red lines under the JASPAR logos indicate the DNA sequences involved in binding to the TF-chain/domain

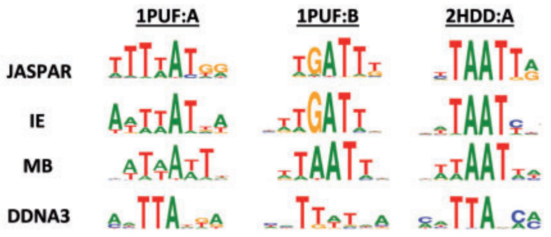

Fig. 6. Binding site prediction of three homeodomains in the non-redundant dataset

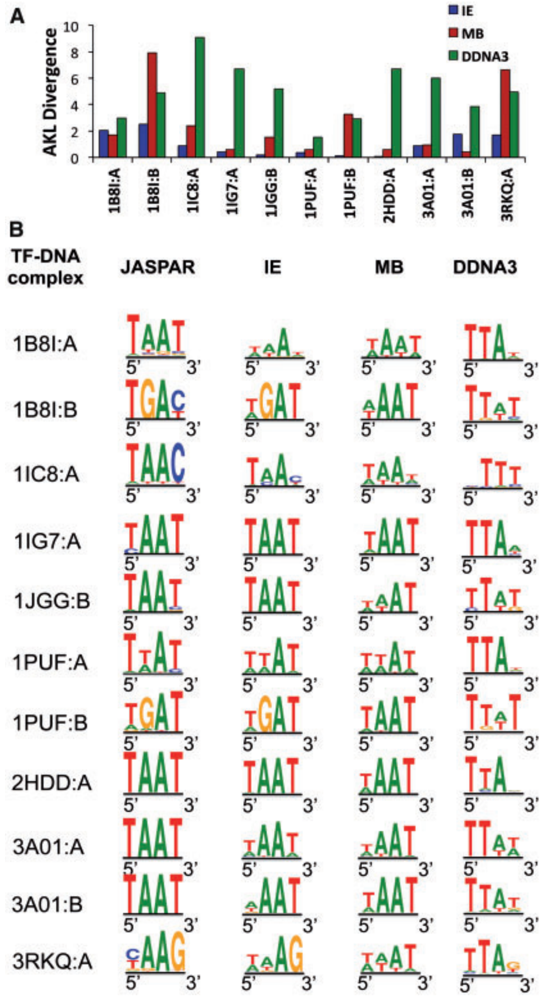

Fig. 7. Prediction of homeodomain binding sites. (**A**) Quantitative comparison between the predicted binding motifs and JASPAR motifs of the homeodomain dataset using the IE (blue), MB potential (red) and DDNA3 (green) using AKL divergence. (**B**) Comparison of the predicted binding motifs

described in Section 2. Figure 7 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 IE function can not only accurately predict the binding sites of homeodomains with low sequence identity (Fig. 6), it can also accurately predict the binding sites of homeodomains with high sequence similarity but with different binding sites (Fig. 7).

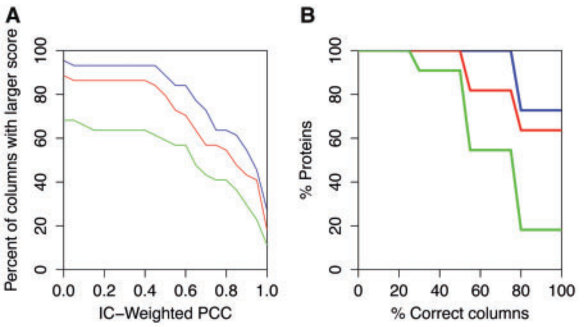

Fig. 8. Performance comparison of the IE (blue), MB (red) and DDNA3 (green) based on IC-weighted PCC. (**A**) Distribution of IC-weighted PCC. For each threshold of IC-weighted PCC score (*x*-axis), the fraction of predicted columns that achieves a score that high or more when compared with their corresponding 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

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 8 shows that approximately 93% of the core base positions (44 columns) are correctly predicted by the IE 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.

4 Discussion

We report here improved accuracy of structure-based TF binding site prediction using an IE function. The IE function consists of the MB potential (Liu *et al.*, 2005), and two atomic terms: hydrogen bond energy and π interaction energy. The MB energy is a residue-level knowledge-based protein-DNA interaction potential derived from the mean force theory. Even though this MB 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.*, 2005, 2008), 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 MB potential improves the TFBS prediction accuracy of 1B8I:B and 1IC8:A in the homeodomain dataset (Fig. 9A), as it captures the hydrogen bonds formed between arginine 258 and lysine 273, respectively, and the guanine of the conserved G:C base pair (Fig. 9B and C).

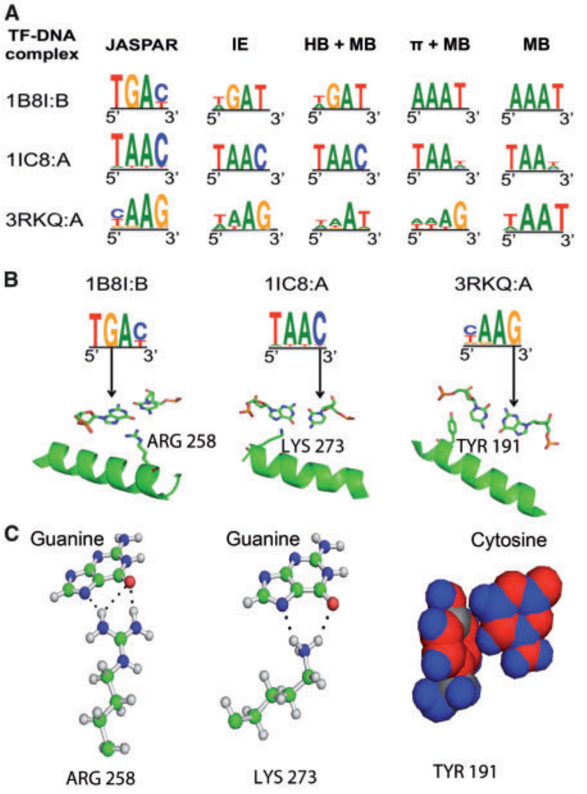

Fig. 9. Contribution of energy terms to prediction accuracy. (**A**)The hydrogen bond energy term improves the prediction accuracy of 1B8I:B and 1IC8:A when compared with the MB energy. The increased prediction accuracy of 3RKQ:A has a major contribution from the π-interaction energy term. (**B**) Physical 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 gray spheres represent partial positive, partial negative and neutral charged atoms, respectively.

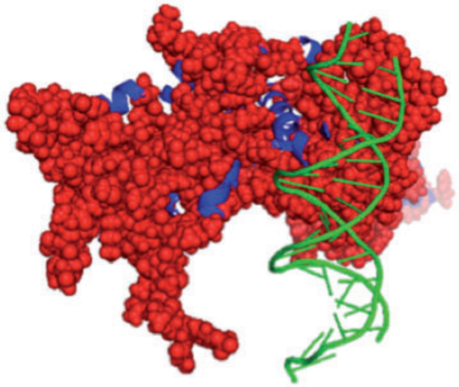

Fig. 10. Complex structure STAT–1/DNA complex (1BF5:A). The interaction involves many coils (red spheres) with DNA (green).

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 (Fig. 2). However, these interactions are masked owing 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 (Fig. 9B), improving the TFBS prediction accuracy. This suggests that the partial positively charged atoms (large blue spheres in Fig. 9C) of cytosine interact electrostatically with the partial negatively charged atoms (large red spheres in Fig. 9C) in the aromatic ring of tyrosine 191, which may contribute to TF–DNA binding specificity.

The IE function shows an overall improvement in TFBS prediction over other knowledge-based potentials. However, in several cases in the multi-family dataset, the IE function does not perform as well as the MB and DDNA3 potentials (Fig. 4). 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 (Fig. 10). As discussed in the introduction, both hydrogen 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.

5 Conclusion

We developed a novel IE function that consists of three components, a knowledge-based MB potential, a hydrogen bond energy function and an electrostatic potential for π interaction energy. We applied the new IE function to the prediction of TFBSs. The results show an overall improvement in binding site prediction, and there is a significant improvement in predicting binding sites of homeodomains when compared with the MB and DDNA3 potentials. The improved accuracy using the integrative function demonstrates the importance of considering hydrogen bonds and π−interactions explicitly in structure-based TFBS predictions, as they are not accurately captured by the knowledge-based potentials.

Funding

This work has been supported by the National Institutes of Health [R15GM110618 to J.G]; and National Science Foundation [DBI1356459 to J.G].

Conflict of Interest: none declared.

References

UniProt. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*,**36**, D190–D195.

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

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

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

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

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Blaisdell,B.E. (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Natl. Acad. Sci. USA*, **83**, 5155–5159.

Corona,R.I. and Guo,J-T. (2016) Statistical analysis of structural determinants for protein-DNA binding specificity. *Proteins: Structure, Function, and Bioinformatics*, in press.

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

Dahiyat,B.I. *et al.* (1997) Automated design of the surface positions of protein helices. *Protein Sci.*, **6**, 1333–1337.

Donald,J.E. *et al.* (2007) Energetics of protein-DNA interactions. *Nucleic Acids Res.*, **35**, 1039–1047.

Gehring,W.J. *et al.* (1994) Homeodomain proteins. *Ann. Rev. Biochem.*, **63**, 487–526.

Gromiha,M.M. *et al.* (2004) Structural analysis of cation-pi interactions in DNA binding proteins. *Int. J. Biol. Macromol.*, **34**, 203–211.

Havranek,J.J. *et al.* (2004) A simple physical model for the prediction and design of protein-DNA interactions. *J. Mol. Biol.*, **344**, 59–70.

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

Lemon,B. and Tjian,R. (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.*, **14**, 2551–2569.

Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.

Liu,L.A. and Bradley,P. (2012) Atomistic modeling of protein-DNA interaction specificity: progress and applications. *Curr. Opin. Struct. Biol.*, **22**, 397–405.

Liu,Z. *et al.* (2005) Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential. *Nucleic Acids Res.*, **33**, 546–558.

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

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

Luscombe,N.M. *et al.* (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.

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

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

Mathelier,A. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.

McGaughey,G.B. *et al.* (1998) pi-Stacking interactions. Alive and well in proteins. *J. Biol. Chem.*, **273**, 15458–15463.

Mecozzi,S. *et al.* (1996) Cation-pi interactions in aromatics of biological and medicinal interest: electrostatic potential surfaces as a useful qualitative guide. *Proc. Natl. Acad. Sci. USA*, **93**, 10566–10571.

Morozov,A.V. *et al.* (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, **33**, 5781–5798.

Persikov,A.V. and Singh, M. (2014) De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Res.*, **42**, 97–108.

Pettersen,E.F. *et al.* (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.

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

Schneider,T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Seeman,N.C. *et al.* (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA*, **73**, 804–808.

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

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

Takeda,T. *et al.* (2013) A knowledge-based orientation potential for transcription factor-DNA docking. *Bioinformatics*, **29**, 322–330.

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

Tupler,R. *et al.* (2001) Expressing the human genome. *Nature*, **409**, 832–833.

Wilson,K.A. *et al.* (2014) 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 Res.*, **42**, 6726–6741.

Wilson,K.A. and Wetmore,S.D. (2015)A survey of DNA–protein π–interactions: a comparison of natural occurrences and structures, and computationally predicted structures and strengths. In: Scheiner, S. (eds), *Noncovalent Forces, Challenges and Advances in Computational Chemistry and Physics*, Vol. **19**. Springer International Publishing, Switzerland.

Wintjens,R. *et al.* (2000) Contribution of cation-pi interactions to the stability of protein-DNA complexes. *J. Mol. Biol.*, **302**, 395–410.

Wu,T.J. *et al.* (2001) Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics*, **57**, 441–448.

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

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