BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 3 2013, pages 322–330
doi:10.1093/bioinformatics/bts699

Structural bioinformatics

Advance Access publication December 5, 2012

A knowledge-based orientation potential for transcription factor-DNA docking

Takako Takeda, Rosario I. Corona and Jun-tao Guo*

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

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Computational modeling of protein–DNA complexes remains a challenging problem in structural bioinformatics. One of the key factors for a successful protein–DNA docking is a potential function that can accurately discriminate the near-native structures from decoy complexes and at the same time make conformational sampling more efficient. Here, we developed a novel orientation-dependent, knowledge-based, residue-level potential for improving transcription factor (TF)-DNA docking.

Results: We demonstrated the performance of this new potential in TF–DNA binding affinity prediction, discrimination of native protein–DNA complex from decoy structures, and most importantly in rigid TF–DNA docking. The rigid TF–DNA docking with the new orientation potential, on a benchmark of 38 complexes, successfully predicts 42% of the cases with root mean square deviations lower than 1 Å and 55% of the cases with root mean square deviations lower than 3 Å. The results suggest that docking with this new orientation-dependent, coarse-grained statistical potential can achieve high-docking accuracy and can serve as a crucial first step in multi-stage flexible protein–DNA docking.

Availability and implementation: The new potential is available at http://bioinfozen.uncc.edu/Protein_DNA_orientation_potential.tar.

Contact: jguo4@uncc.edu

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

Received on August 13, 2012; revised on November 25, 2012; accepted on December 1, 2012

1 INTRODUCTION

Protein–DNA interactions play crucial roles in many biological processes. Transcription factors (TFs), which bind to specific *cis*-regulatory elements on DNA and regulate gene expression, represent one of the largest groups of proteins in many genomes (Babu *et al.*, 2004; Janga and Collado-Vides, 2007). How TFs recognize and bind specifically to their DNA target sequences, despite of decades of efforts, is still not fully understood. Benefited from the technical advances in experimental structure determination, high-resolution structures of TF–DNA complexes have provided a glimpse of TF–DNA interaction in each complex. A collection of such views may offer valuable insights into the molecular mechanism of TF–DNA recognition and the evolution of gene regulatory networks (Luscombe and Thornton, 2002). A TF–DNA complex model is also a starting point for structure-based TF-binding site prediction, which has received much research attention recently (Angarica *et al.*, 2008; Kaplan *et al.*, 2005; Liu *et al.*, 2008; Siggers and Honig, 2007; Xu *et al.*, 2009).

Despite technical advances in experimental structure determination, only a small percentage of TF–DNA complex structures have been solved and deposited in Protein Data Bank (PDB) (Berman *et al.*, 2000). Computational docking between a protein and DNA, on the other hand, has been considered as a cost-efficient alternative to the usually time-consuming experimental methods. Macromolecule docking relies on some sorts of energy functions for building complex models (Pande, 2011). There are two major types of the potentials for studying protein–DNA interactions: physics-based (Donald *et al.*, 2007; Endres *et al.*, 2004) and knowledge-based (Liu *et al.*, 2005; Robertson and Varani, 2007; Xu *et al.*, 2009; Zhang *et al.*, 2005). For example, van Dijk *et al.* applied a physics-based potential, originally developed for protein–protein docking, to protein–DNA docking (van Dijk and Bonvin, 2010; van Dijk *et al.*, 2006). Knowledge-based potentials, derived from experimental structures, are considered more attractive and have more practical value in structural bioinformatics studies owing to their relative simplicity (Miyazawa and Jernigan, 1985; Pande, 2011; Sippl, 1990; Sippl, 1995; Zhou and Zhou, 2002). These knowledge-based potentials generally vary in their resolutions, from residue-level to atom-level and in their distance scales, from distance-independent to distance-dependent (Gao and Skolnick, 2008; Kono and Sarai, 1999; Liu *et al.*, 2005; Luscombe *et al.*, 2001; Robertson and Varani, 2007; Xu *et al.*, 2009; Zhang *et al.*, 2005; Zhao *et al.*, 2010).

Although high-resolution, atomic-level potentials can provide the details needed to accurately discriminate near-native structures from decoys, coarse-grained potentials can have a smooth and less-rugged energy landscape, making it less likely to get trapped in local minima during conformational search (Ayton *et al.*, 2007; Bradley *et al.*, 2005; Flores *et al.*, 2012; Kim and Hummer, 2008; Poulain *et al.*, 2008). Another advantage of the coarse-grained potentials at residue-nucleotide level is their capability in addressing the dynamic nature of macromolecules, as they are less sensitive to small conformational changes (Bradley *et al.*, 2005; Gopal *et al.*, 2010; Vreven *et al.*, 2011). To take advantage of both the coarse-level and the atomic-level potentials, multi-scale approaches are often adopted, in which near-native models are constructed first with a coarse-level potential followed by refinement with high-resolution potentials (Chen and Xu, 2006; Murphy *et al.*, 2003; Vreven *et al.*, 2011).

*To whom correspondence should be addressed.

In this study, we focus on the development of a novel knowledge-based, residue-level potential for accurate docking between TFs and their DNA target sequences. Previous studies have revealed different interaction 'modes' between TFs and other major types of DNA-binding proteins, such as restriction enzymes and non-specific DNA-binding proteins (Ashworth and Baker, 2009; Contreras-Moreira *et al.*, 2010; Kim *et al.*, 2011). In addition, the negatively charged residues, aspartate and glutamate, are overrepresented in restriction enzymes compared with other types of DNA-binding proteins (Pingoud *et al.*, 2005). As the major goal of this study is to develop a protein–DNA interaction potential for assessing TF–DNA binding affinity in TF–DNA docking, non-TF protein–DNA complexes including restriction enzymes and non-specific DNA-binding proteins are not included in the dataset for potential development.

Liu *et al.* have previously developed a knowledge-based, residue-level potential based on statistical analysis of known TF–DNA complex structures (Liu *et al.*, 2005). The potential uses DNA tri-nucleotides, called triplets, as an interaction unit to study the interactions between TF and DNA molecules. The triplets could be real nucleotides with explicit positions (native nucleotides) or pseudo-nucleotide placeholders that do not make any structural or energy contribution toward potential calculation. The triplet representation has the advantage of covering both the preference of individual bases and the local environment around the nucleotides. It has been shown that this multi-body potential performs well in assessing TF–DNA interactions and in protein–DNA docking (Liu *et al.*, 2005; Liu *et al.*, 2008).

The binding specificity between a DNA and a protein is generally contributed by hydrogen bonds. It has been shown that two-thirds of the hydrogen bonds between amino acids and bases lead to specific complex interactions (Angarica *et al.*, 2008). Kono and Sarai have previously studied the radial and angular distributions for hydrogen bonds and found that the angular distributions of protein atoms around potential hydrogen-bond forming atoms of bases have different patterns (Kono and Sarai, 1999). The strength of a hydrogen bond is usually defined by bond length(s) and angle(s) between a donor and an acceptor (Baker and Hubbard, 1984; Frishman and Argos, 1995; Wade and Goodford, 1993). Therefore, adding angular information to a statistical potential can be expected to improve the accuracy in assessing TF–DNA binding affinity and specificity.

In this article, we present an orientation-dependent, knowledge-based potential derived from a non-redundant set of TF–DNA complex structures through converting the observed frequencies of base-residue pairs with respect to both distances and angles to a potential based on Boltzmann's principle. The performance of the new potential was assessed through binding affinity prediction and TF–DNA rigid docking. The results show much better accuracy in TF–DNA binding affinity prediction and rigid TF–DNA docking with the new orientation potential when compared with the residue potential without angle information.

2 METHODS

2.1 Datasets

A non-redundant dataset of 160 TF–DNA complex structures was first generated from PDB for deriving the new orientation potential (Supplementary Table S1) (Berman *et al.*, 2000). These complex structures were solved by X-ray crystallography with resolutions 3.5 Å or better and R-factors of at most 0.3. The TFs in the complexes have 40–1000 amino acids. Redundant TF chains are removed with a sequence identity cutoff of 55%. To eliminate any potential bias when constructing a non-redundant dataset in structural bioinformatics, it would be ideal to use a lower sequence identity cutoff. However, when only a limited number of complex structures are available, there are drawbacks for selecting representatives with a lower cutoff, as discussed in previous studies. First, statistical analysis based on a small dataset would suffer the low-count problem, particularly in cases with a large number of combinations of cases (Luscombe *et al.*, 2001). Second, homologous TFs can bind to different DNA sequences, and the binding patterns may be unique to the specific TF–DNA complex. Inclusion of these entries may maximize the diversity of protein–DNA interactions (Kim and Guo, 2009; Luscombe *et al.*, 2001; Prabakaran *et al.*, 2006). Therefore, there is a trade-off between the degree of redundancy and the statistical significance of the potentials. Taking both into account, we settled on a cutoff of 55%.

For TF–DNA docking evaluations, we used our previously developed rigid TF–DNA docking benchmark (Kim *et al.*, 2011). This benchmark contains 38 non-redundant cases that are classified into two groups in terms of expected docking difficulty. Each case in the benchmark is a TF–DNA binding unit defined as an entity of a DNA double-helix and one or more TF-chains that interact with each other with at least three residue–residue contacts based on a heavy-atom distance cutoff of 4.5 Å. The TFs in the 38 complexes have <35% sequence identity and do not have overlap with the 160 complexes for potential development based on the selection criteria.

2.2 Development of the orientation potential

For the development of our new orientation potential, we applied a similar statistical approach used in developing the multi-body potential (Liu *et al.*, 2005). Owing to the limited number of non-redundant TF–DNA complexes, we only consider distance and angle information while dropping the multi-body term. Figure 1 illustrates the angle used for the new potential. The angle φ represents the angle between two vectors. One vector is defined based on the DNA bases, which is either from N9 to N1 for adenine and guanine or from N1 to C4 for cytosine and thymine (Fig. 1A and B). The other vector is a projection of the residue sidechain vector (from the C_β atom to the sidechain centroid) onto the base plane (Fig. 1C). A pseudo C_β position is calculated for glycine as described previously (Liu *et al.*, 2005). For glycine and alanine, C_α–C_β vector is used instead, as there are no heavy atoms beyond the C_β position of the sidechain. The angles are grouped into three bins ($-60^\circ \le \varphi < 0^\circ$, $0^\circ \le \varphi < 60^\circ$, and $-120^\circ \le \varphi < -60^\circ$ and $60^\circ \le \varphi < 120^\circ$). To further reduce the total number of possible combinations, some residues are grouped together based on both the physicochemical properties of amino acids and the low raw count in known TF–DNA complexes. In this study, ALA, ILE, LEU, PRO and VAL are combined into one group, whereas SER and THR belong to another group.

The distance r between a residue and a base is represented by the distance between the centroid of the residue sidechain and the centroid of the base. The bin width is set at 1 Å with a distance cutoff of 15 Å, meaning there is no interaction between a residue and a base if they are separated by >15 Å. The correction of the observed fractions of interactions, owing to incomplete training set, is first carried out by introducing two exponential parameters (α and β) for the fraction of interacting residues $W_{residue}$ and the fraction of interacting bases W_{base}, respectively (Equation 1). In Equation 1, $N_{ij}(r, \varphi)$ is the corrected number of interactions between protein residue i and DNA base j, and $N^0_{ij}(r,\varphi)$ is the initial number of interactions between protein residue i and DNA base j. Low count effects are removed by an offset parameter σ. A cutoff on the number of interactions N_{cutoff} is set to 1.

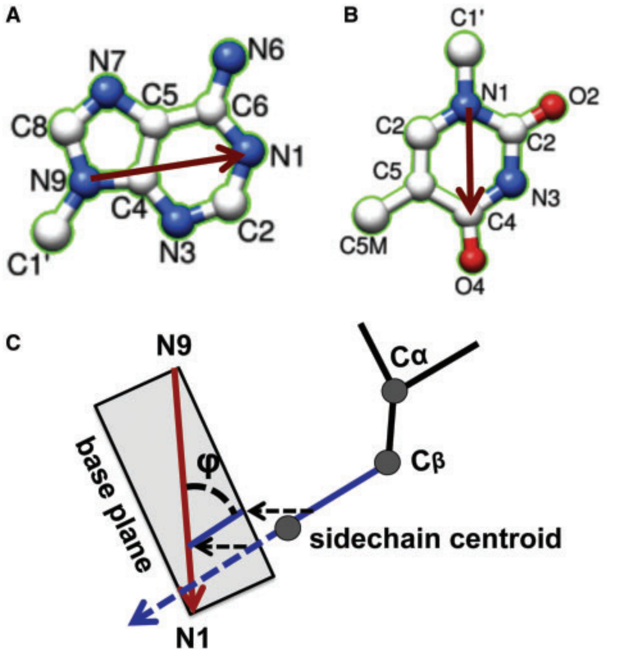


Fig. 1. Schematic representation of vectors for defining the angle φ. Vectors for the DNA bases [(**A**) purines; (**B**) pyrimidines] are shown in red arrows. The definition of angle φ between residue sidechain and DNA base is shown in (**C**)

$$N_{ij}(r,\varphi)=\begin{cases}\dfrac{N^0_{ij}(r,\varphi)}{W^{\alpha}_{residue}\cdot W^{\beta}_{base}\cdot F(r)} & \text{for}\quad N^0_{ij}(r,\varphi)>N_{cutoff}\\ \dfrac{N^0_{ij}(r,\varphi)}{W^{\alpha}_{residue}\cdot W^{\beta}_{base}\cdot F(r)}+\sigma & \text{for}\quad N^0_{ij}(r,\varphi)\le N_{cutoff}\end{cases}\qquad(1)$$

The two exponential parameters (α, β) and the offset parameter σ are optimized using a Z-score optimization approach (see detailed description for Equations 6 and 7). $F(r)$ is a distance normalization function for removing the distance effect. We used $F(r)=0.535*(r\verb|^|3.345)$ as described previously (Liu *et al.*, 2005). After the correction of the observed interaction frequency with those parameters, we convert the frequency to the potential by Boltzmann law (Equations 2–4).

$$p_{ij}(r,\varphi)=\frac{N_{ij}(r,\varphi)}{\sum_{i,j}N_{ij}(r,\varphi)}=\frac{e^{-E_{ij}(r,\varphi)/kT}}{\sum_{i,j}e^{-E_{ij}(r,\varphi)/kT}}=\frac{e^{-E_{ij}(r,\varphi)/kT}}{Z}\qquad(2)$$

Equation 2 describes the relationship between the observed probability $p_{ij}(r,\ \varphi)$ and the statistical thermodynamic interaction energy $E_{ij}(r,\ \varphi)$ of an interaction between residue i and base j with a distance r and an angle φ, where T is temperature, k is Boltzmann constant and Z is the partition function. The uniform density reference is used for each distance-angle bin (Equation 3), where $\overline{p_{ij}(r,\varphi)}$ is the mean probability of the interactions between residues and bases with a distance r and an angle φ. Energy $E^0_{ij}(r,\ \varphi)$ for the final potential is shown in Equation 4, where $\overline{E(r,\varphi)}$ is the reference mean energy.

$$\sum\nolimits_{i,j}p_{ij}(r,\varphi)=1,\quad \overline{p(r,\varphi)}=\frac{\sum_{i,j}p_{ij}(r,\varphi)}{\sum_{i,j}1}=\frac{1}{15\times 4}=\frac{1}{60}\qquad(3)$$

$$\begin{aligned}E^0_{ij}(r,\varphi)&=E_{ij}(r,\varphi)-\overline{E(r,\varphi)}\\&=-kT\big(\ln(p_{ij}(r,\varphi)\cdot Z)-\ln\overline{p(r,\varphi)}\cdot Z)\big)\\&=-kT\cdot\ln\left(\frac{p_{ij}(r,\varphi)}{\overline{p(r,\varphi)}}\right)=-kT\cdot\ln\big(60\times p_{ij}(r,\varphi)\big)\end{aligned}\qquad(4)$$

The interaction energy E for a protein–DNA complex is the sum of the energies $E^0_{ij}(r,\ \varphi)$ of all residue-base interactions (Equation 5). In Equation 6, Z_t, the critical Z-score, is the gap between the native energy $E_{native\text{-}t}$ of complex t, and the average energy of the decoys $<E_t>$ and $\delta(E_t)$ is the standard deviation of the decoy energies. An average Z-score for M complexes is computed as in Equation 7. Z-score optimization is performed to derive α, β and σ by a Monte Carlo simulated annealing approach. The cooling rate is set at 0.998 with a convergence of 10^{-6}. The goal is to minimize the average Z-score by changing the parameter values of α, β and σ using the native and decoy complexes of the 160 TF–DNA complexes. In this work, parameters α, β and σ are 0.644, 0.787 and 0.440, respectively.

$$E=\sum\nolimits_i\sum\nolimits_j\sum\nolimits_r\sum\nolimits_\varphi E^0_{ij}(r,\varphi)\qquad r\le 15\text{Å}\qquad(5)$$

$$Z_t=\frac{E_{native-t}-\langle E_t\rangle}{\delta(E_t)}\qquad(6)$$

$$Z=\ln\left(\frac{\sum_t e^{Z_t}}{M}\right)\qquad(7)$$

2.3 Assessment of the orientation potential

2.3.1 Binding affinity We compared the predicted binding affinity with the experimental binding-free energies of 25 protein–DNA complexes (PDBID: 1aay, 1apl, 1az0, 1azp, 1bc7, 1bhm, 1bp7, 1ca5, 1cdw, 1cma, 1cw0, 1ecr, 1glu, 1hcr, 1ipp, 1lmb, 1nfk, 1oct, 1par, 1qrv, 1tro, 1run, 1tsr, 1ysa, 1ytf). These complexes were selected from the dataset with 30 complexes used by Xu *et al.* (2009). The five complexes that appeared in our 160-complex dataset for potential development are not considered.

2.3.2 Discrimination of native structures from decoys Z-scores were calculated to test how well the new potential can discriminate the native protein–DNA complex from docking decoys. The 2000 lowest root mean square deviation (RMSD) docking decoys for each of the 27 DNA–protein complexes (PDBID: 1a1i, 1a73, 1au7, 1bc8, 1ckq, 1d02, 1dfm, 1dmu, 1eon, 1f4k, 1g9z, 1h8a_a, 1h8a_b, 1hlv, 1jko, 1l3l, 1mjo, 1mnn, 1pdn, 1qna, 1tc3, 1tro, 1zme, 2hdd, 3bam, 3pvi, 6pax) were selected from the 45 cases used by Robertson and Varani (Robertson and Varani, 2007) after removing the 18 complexes that are in our dataset for potential development to avoid potential biases. Z-scores were calculated as described in Equation 6. The RMSD was computed between the backbone heavy atoms of the native DNA and the docked DNA structure after fixing the protein positions.

2.3.3 Rigid TF–DNA docking The performance of the orientation potential in rigid TF–DNA docking was assessed with our previously developed protein–DNA docking program (Liu *et al.*, 2008). The program uses a Monte Carlo simulated annealing approach to search for a docked TF–DNA conformation with the optimal interaction energy. The energy function consists of binding affinity and van der Waals packing energy. The primary role of adding the packing energy to the docking is to guide the docking process without affecting the final docked structures (its contribution to the final energy approaches zero as the random walk progresses). The movements include rotations with a step size of 2° and translations with a step size of 0.1 Å. The simulation stops when it converges or it reaches a total number of 1.5 million steps. Two hundred independent Monte Carlo simulations were carried out for each TF–DNA complex. Protein and DNA are harmonically constrained with a cutoff of 14 Å between the protein pocket and the centroids of DNA. We tested the performance of the new orientation potential on our previously developed TF–DNA rigid docking benchmark (Kim *et al.*, 2011). To evaluate the docking accuracy, we compared the docked DNA conformations with the corresponding DNA structures in the native TF–DNA complexes by fixing the protein positions and calculated the RMSDs between the predicted and the native complex using DNA backbone heavy atoms.

3 RESULTS AND DISCUSSION

3.1 Binding affinity and decoy discrimination

We first tested how well the new potential can predict the binding affinity. Figure 2 shows the correlations between experimental binding affinity and the predicted binding affinity using either the orientation potential (Fig. 2A) or the multi-body potential, a knowledge-based residue level potential developed by Liu *et al.* (Fig. 2B) (Liu *et al.*, 2005). The correlation coefficient between the predicted binding affinity with orientation potential and the experimental energy is 0.57 (P-value = 0.01), whereas the multi-body potential has a correlation coefficient of 0.41 (P-value = 0.02). The performance of the orientation potential is on par with vFIRE, an all-atom, knowledge-based DFIRE function that includes volume fraction (correlation coefficients 0.57 for the orientation potential versus 0.55 for vFIRE) (Xu *et al.*, 2009).

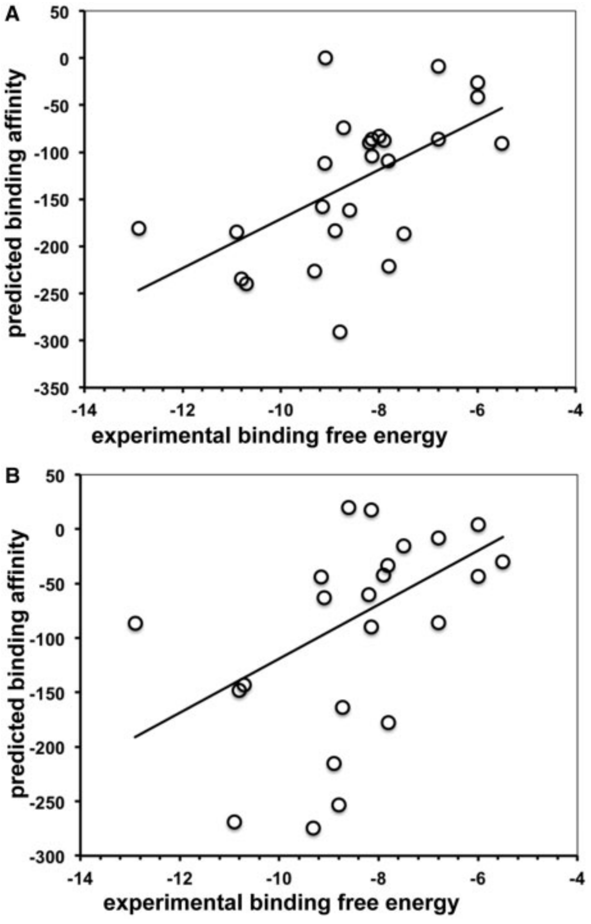


Fig. 2. Correlation analysis between the predicted binding affinity and the experimental binding affinity [*-log*(*Kd*) *unit*]. (**A**) Predicted binding affinity with the orientation potential; (**B**) Predicted binding affinity with the multi-body potential

Our main goal of this study is to develop a better and efficient potential for improving TF–DNA docking performance. Thus, it is important to assess the potential's capability for discriminating the native or near native structures from decoy structures. We tested it on a docking decoy set by Robertson and Varani (see Methods) and demonstrated the discriminative power based on Z-scores: Z-score = $(E_{native} - E_{avg})/S$, where E_{native} is the predicted binding affinity for the native complex, and E_{avg} and S are the average and standard deviation of the binding affinities of decoy complexes, respectively. A higher Z-score, especially a positive Z-score, would suggest a lower discriminative power. Overall, the orientation potential outperforms the multi-body potential based on Z-score comparisons (Fig. 3). All except for one case (1ckq) with orientation potential have negative Z-scores, whereas there are eight complexes (1d02, 1hlv, 1l3l, 1tc3, 1tro, 1zme, 3bam and 3pvi) having positive Z-scores calculated with the multi-body potential. We should point out that the binding affinities calculated from the orientation and multi-body potentials are at a similar scale, otherwise the direct comparison of Z-scores would be less meaningful.

In addition to Z-score comparison, another way to test the performance improvement of the new potential is to assess the relationship between the binding affinity of a decoy and the structural distance between this decoy structure and the native complex in terms of RMSD. Two such examples, 1a1i and 1l3l, are shown in Figure 4 (A–C for 1ali; D–F for 1l3l), in

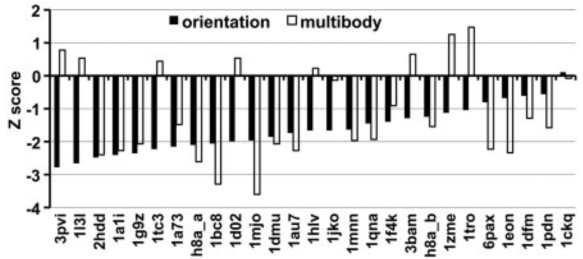


Fig. 3. Comparison of Z-scores between the orientation potential (filled) and the multi-body potential (open). The entries are arranged from left to right based on sorted Z-scores of orientation potential

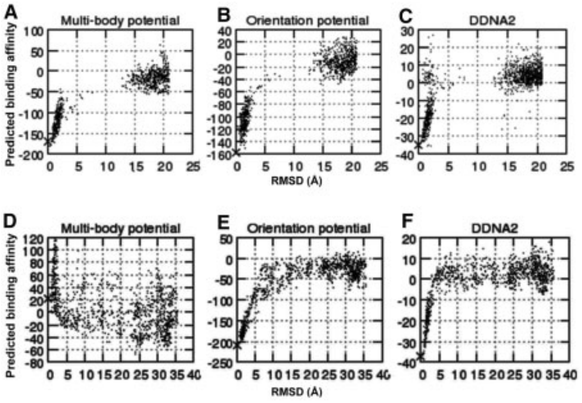


Fig. 4. Scatter plots of RMSDs versus the predicted binding affinities for 1a1i (**A–C**) and 1l3l (**D–F**). (A and D) multi-body potential; (B and E) orientation potential; (C and F) DDNA2 potential

which we compared the binding affinities predicted by the orientation potential, multi-body potential and DDNA2. DDNA2 is a program for predicting protein–DNA binding affinity using a knowledge-based, atomic-level potential (Xu *et al.*, 2009). Consistent with the *Z*-score data, the orientation potential clearly performs better than the multi-body potential and achieves a similar performance to DDNA2. Similar results were observed in other cases (Supplementary Fig. S1). To quantitatively compare the performance of the orientation potential with the multi-body potential and DDNA2, we counted the number of 'false positive' decoys in each of the 27 cases. A decoy is considered as a 'false positive' if its binding affinity is lower than the binding affinity of the native structure, and its RMSD with respect to the native structure is >5 Å. The number of false positives using the orientation potential is much smaller than that using the multi-body potential. The performance of the orientation potential is close to that of DDNA2, which is an atomic-level potential (Supplementary Fig. S2).

Table 1. Rigid docking results using orientation potential on rigid docking benchmark set

Class	PDBID	Conformation with the lowest energy		Conformation with the smallest RMSD	
		RMSD (Å)	$E_{docking}$	RMSD (Å)	$E_{docking}$
Easy	1aay	0.96	−210.26	0.94	−208.08
	1an2	0.62	−158.29	0.6	−156.72
	1jj4	0.81	−147.16	0.66	−142.28
	1jt0	24.2	−96.53	5.07	−87.86
	1lmb	0.32	−127.49	0.32	−127.49
	1qn4	0.37	−302.72	0.16	−301.16
	1qpi	24.36	−102.29	12.26	−72.29
	1sax	0.39	−356.55	0.36	−353.64
	1tro	0.7	−206.91	0.21	−203.89
	1z9c	1.93	−148.96	1.84	−146.41
	1zs4	0.17	−193.43	0.14	−188.99
	2ac0	0.39	−225.61	0.38	−223.84
	2cgp	0.56	−255.63	0.52	−252.22
	2e1c	0.28	−220.63	0.19	−217.09
	2it0	6.3	−135.85	6.08	−126.05
	2or1	1.62	−186.78	0.39	−177.79
	2yvh	16.81	−109.73	13.5	−104.7
	3clc	1.37	−149.86	1.37	−149.86
	3dnv	0.79	−171.79	0.74	−170.66
	3e6c	0.31	−200.28	0.12	−197.33
	3gz6	1.35	−241.93	0.58	−236.42
Hard	1b01	0.6	−142.22	0.56	−141.23
	1by4	27.6	−74.98	9.69	−50.49
	1cma	5.09	−134.55	1.62	−106.67
	1gxp	36.71	−98.82	1.08	−91.65
	1h8a	16.01	−98.06	15.83	−97.96
	1hjc	2.64	−102.36	2.43	−100.08
	1r8d	6.89	−106.5	5.79	−91.28
	1rio	60.41	−58.85	23.86	38.33
	1xpx	19.69	−88.91	2.21	−58.17
	1zme	20.29	−117.21	3.07	−88.43
	2bnw	8.45	−136.83	2.25	−117.88
	2c6y	0.49	−157.92	0.41	−156.61
	2fio	28.67	−140.85	18.66	−93.61
	2irf	0.39	−135.64	0.3	−132.51
	2rbf	6.09	−117.52	4.4	−108.04
	2zhg	28.91	−170.1	0.5	−138.28
	3hdd	4.32	−135.93	3.88	−132.91

It is worth mentioning that the decoy set by Robertson and Varani was not designed for testing only TFs (Robertson and Varani, 2007). In addition to TFs, the 27 complexes also contain other DNA-binding proteins, including nine restriction endonucleases and five others. Although the orientation potential was developed using a set of TFs, no obvious differences were observed between TFs and non-TF complexes. This is not surprising, as majority of these non-TF proteins are specific DNA-binding proteins. The key difference though is the high occurrences of aspartate and glutamate involved in metal coordination and catalytic activities in these enzymes compared with TFs (data not shown) (Pingoud *et al.*, 2005).

3.2 Assessment of the orientation potential in rigid TF–DNA docking prediction

TF–DNA rigid docking was carried out with either the orientation or the multi-body potential using our previously developed Monte Carlo-based protein–DNA docking program (Liu *et al.*, 2008). The TF–DNA docking benchmark with 21 easy and 17 hard targets was used for performance comparison (Kim *et al.*, 2011). The docking results based on 200 independent Monte Carlo simulations for each case are shown in Table 1 (for orientation potential) and Supplementary Table S2 (for multi-body potential). Each table shows the docking energy and RMSD for both the conformation with the lowest energy and the conformation with the smallest RMSD.

Figure 5 shows that the docking accuracy with the orientation potential is significantly better than that with the multi-body

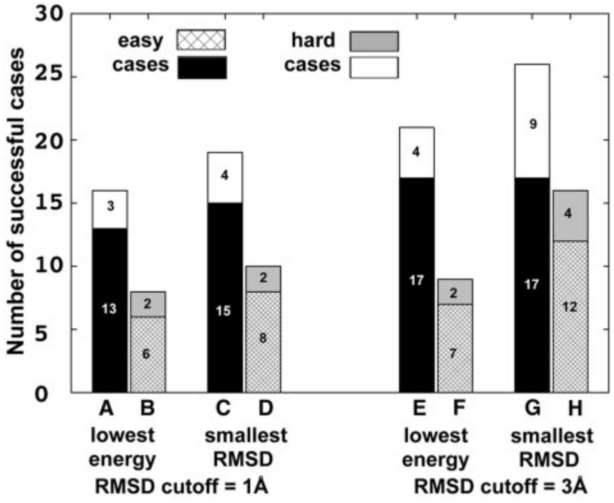


Fig. 5. Docking results on the rigid TF–DNA docking benchmark. The docking performance is evaluated based on either the lowest docking energy or the smallest RMSD using either the orientation potential (**A**, **C**, **E**, **G**) or the multi-body potential (**B**, **D**, **F**, **H**). The filled and the patterned columns represent the easy cases, whereas the open and the gray boxes represent the hard docking targets. Two different RMSD cutoffs are used to tally the successful cases, 1 Å (A, B, C, D) or 3 Å (E, F, G, H)

potential. Docking simulations with the orientation potential reconstructed 16 (42%, 13 easy and 3 hard targets) TF–DNA complexes with RMSDs of ≤1 Å (Fig. 5, column A), whereas only eight (21%, six easy and two hard targets) complexes with RMSDs of ≤1 Å were reconstructed with the multi-body potential (Fig. 5, column B). When the evaluation was done using the smallest RMSD, we found that 19 of 38 complexes (50%, 15 easy and 4 hard targets) have at least one docked conformation with RMSD <1 Å using the orientation potential (Fig. 5, column C); only 10 (26%, eight easy and two hard targets) have at least one docked conformation with RMSD <1 Å for docking with multi-body potential (Fig. 5, column D). We found similar performance improvement when using 3 Å as the RMSD cutoff. Docking with the orientation potential predicted 21 (55%, 17 easy and 4 hard cases) based on the lowest energy and produced 26 (68%, 17 easy and 9 hard cases) targets with docked structures having <3 Å RMSD (Fig. 5E and G), whereas the multi-body potential docking correctly predicted nine (24%, seven easy and two hard cases) and produced 16 (42%, 12 easy and 4 hard cases) targets with docked structures having <3 Å RMSD (Fig. 5F and

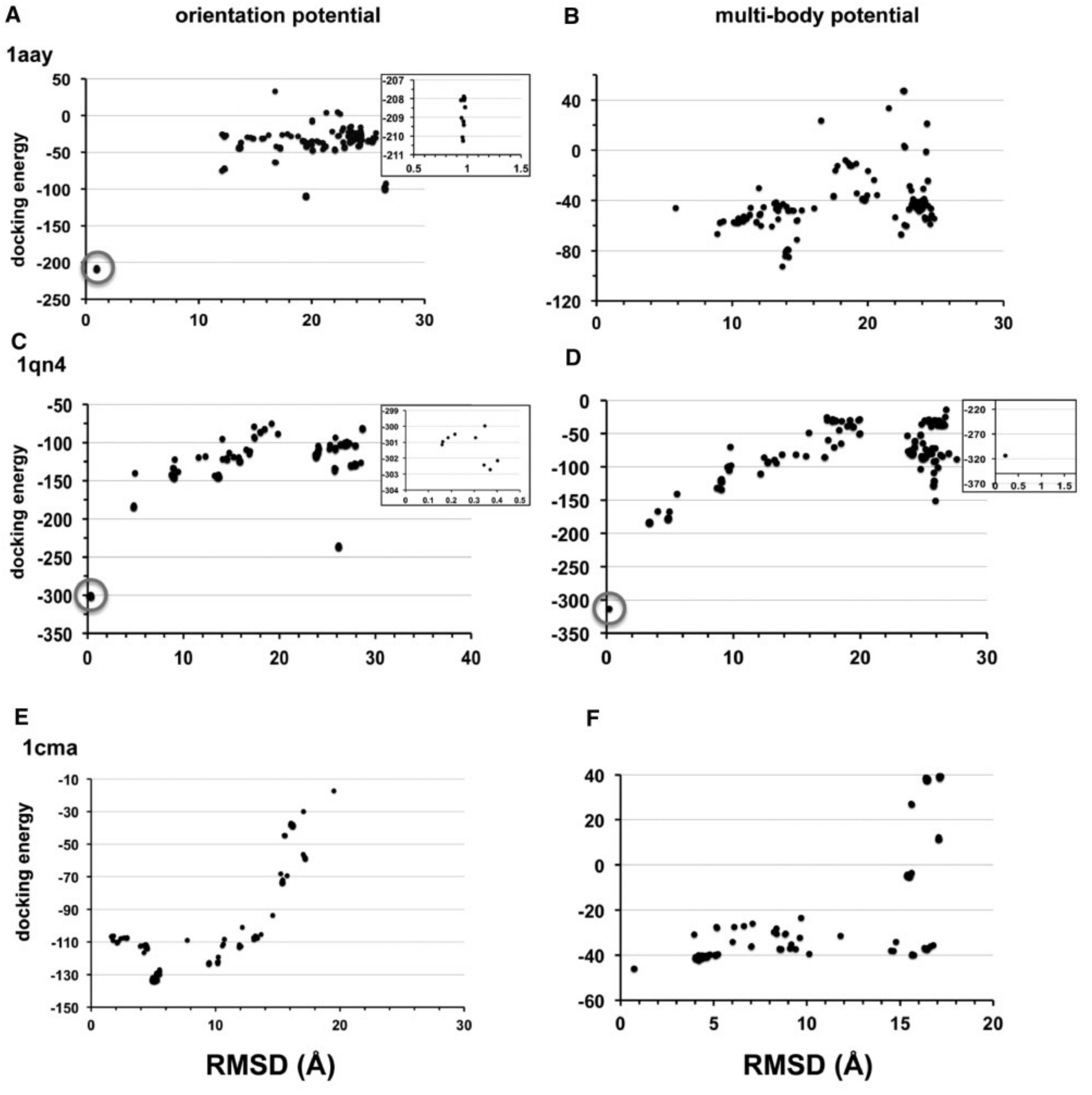


Fig. 6. Scatter plots of docking energy against RMSD of the 200 predicted docking conformations for each target. Three docking targets, 1aay, 1qn4 and 1cma, are shown either with the orientation potential (**A**, **C**, **E**) or the multi-body potential (**B**, **D**, **G**). The insets are the enlarged portions for docked structures in the gray circles

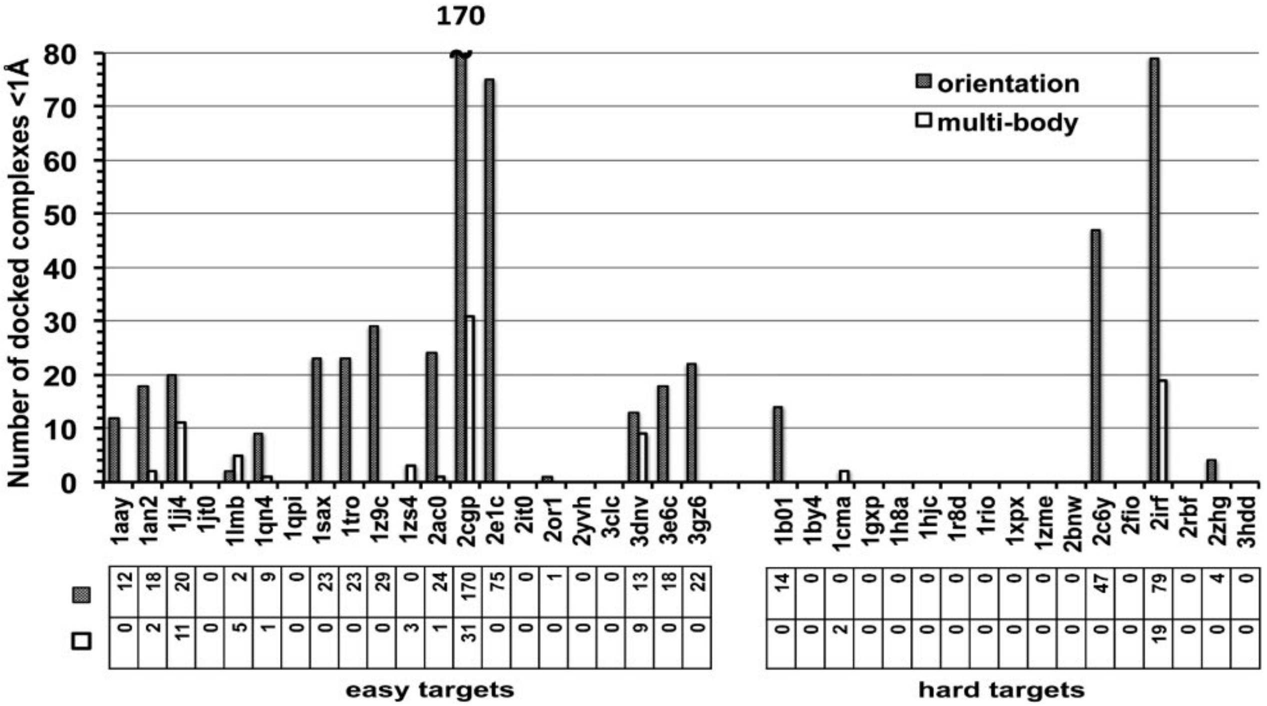
H). The results also support the classification of docking difficulty based on the interaction strength between protein and DNA (Kim *et al.*, 2011). Both potentials made much more correct prediction for the easy cases than for the hard cases (Fig. 5, Table 1 and Supplementary Table S2).

Docking with the orientation potential correctly reconstructed eight more targets with $\leq$1 Å RMSD based on the lowest docking energy (Table 1, Supplementary Table S2 and Fig. 5). Target 1aay represents one of these eight cases (Fig. 6A and B). Docking with the orientation potential produced 12 near-native structures with RMSDs of $\leq$1 Å (inset plot in Fig. 6A), whereas docking with the multi-body potential failed to produce any near-native structures (Fig. 6B).

Owing to the statistical nature of the Monte Carlo docking algorithm, another way to assess the docking performance is to check the 'easiness' (or difficulty) of finding a 'hit' in docking simulations (Wu *et al.*, 2012). For example, if only one of the 200 docked TF–DNA conformations is a correct structure, it is highly possible that a second docking experiment for this target with 200 independent simulations fails to reconstruct a native or near-native structure. On the other hand, it would be easier to make consistent and correct predictions if more near-native structures are reconstructed from 200 independent Monte Carlo docking simulations. One such example is shown in Figure 6 (C and D). Docking simulations for 1qn4 resulted in correct predictions with both the orientation and the multi-body potentials based on the lowest docking energy. The numbers of near-native conformations, however, are dramatically different. There are nine for the orientation potential and only one for the multi-body potential (Fig. 6C and D).

The orientation potential correctly predicted all the cases that were reconstructed using the multi-body potential with 1cma as the only exception (Fig. 6E and F). The energy-RMSD scatterplot for all the docking targets are shown in Supplementary Figure S3. Figure 7 demonstrates major improvement in producing more near-native structures (<1 Å) when using the orientation potential for rigid docking than those with the multi-body potential. The number of 'hits' based on 3 Å cutoff is shown in Supplementary Figure S4.

Owing to the relatively small size of the benchmark set, we further performed the rigid docking experiments on a larger set with 66 TF–DNA complex structures. To construct a larger dataset, a higher sequence identity cutoff is necessary owing to the limited number of TF–DNA complex structures in PDB. These 66 complexes have <70% protein sequence identity with resolutions of 3.5 Å or better. We observed a similar performance improvement to the benchmark set. Docking with the orientation potential resulted a significantly better docking accuracy in each category (~20% improvement) when compared with the multi-body potential (Supplementary Figure S5, Supplementary Tables S3 and S4). We also noticed that the docking accuracy of this large dataset is better than the benchmark. The difference can be a result of two contributing factors. One is that the redundancy level of the large set (70% sequence identity cutoff) is higher than that in the benchmark (<35% sequence identity). The other is that the large set has more entries with larger protein–DNA interaction interface (Supplementary Fig. S6) as we showed earlier that a complex with larger protein–DNA interaction interface is relatively easier to predict than the ones with smaller interaction interface (Kim *et al.*, 2011).



	1aay	1an2	1jj4	1jt0	1lmb	1qn4	1qpi	1sax	1tro	1z9c	1zs4	2ac0	2cgp	2e1c	2it0	2or1	2yvh	3clc	3dnv	3e6c	3gz6
▨	12	18	20	0	2	9	0	23	23	29	0	24	170	75	0	1	0	0	13	18	22
□	0	2	11	0	5	1	0	0	0	0	3	1	31	0	0	0	0	0	9	0	0

	1b01	1by4	1cma	1gxp	1h8a	1hjc	1r8d	1rio	1xpx	1zme	2bnw	2c6y	2fio	2irf	2rbf	2zhg	3hdd
▨	14	0	0	0	0	0	0	0	0	0	0	47	0	79	0	4	0
□	0	0	2	0	0	0	0	0	0	0	0	0	0	19	0	0	0

Fig. 7. The number of successfully docked structures on the rigid docking benchmark. The total number of independent docking simulations for each target is 200, and the RMSD cutoff is set at 1 Å

Taken together, we observed a much better docking performance for the orientation potential in reconstructing complexes in terms of finding the native or near-native conformations (Fig. 5, Supplementary Figs S5 and S6). Docking with the orientation potential also produced more near-native structures than the multi-body potential when using the same docking procedure (Fig. 5, Supplementary Figs S5 and S6). In addition, the computation time using the new orientation potential is only slightly more than that with the multi-body potential, 51 ± 20.6 min versus 42 ± 18.2 min based on the 66 docking targets (Supplementary Fig. S7).

4 CONCLUSION

We have developed an orientation-dependent, knowledge-based residue-level potential and assessed its performance with a variety of tests—binding affinity prediction, decoy discrimination and rigid TF–DNA docking. The new potential has a much better protein–DNA binding affinity prediction capability than our previously developed multi-body residue-level potential. Our results also show that the performance of this residue-level, orientation potential is close to some of the atomic-level potentials, such as vFIRE, though it is less accurate than cFIRE or vcFIRE (Xu *et al.*, 2009). This is not surprising, as the atom-level potentials have more detailed information on the interaction. However, the main purpose of this new, coarse-grained potential is for improving TF–DNA docking predictions. A coarse-grained potential has an advantage in addressing the dynamic nature of macromolecules, as it is less sensitive to small conformational changes.

Although the use of atomic details offers the accuracy in scoring, the rugged energy landscape and the time-consuming energy calculations at atomistic level can get a docking simulation trapped in the local minima (Ayton *et al.*, 2007; Bradley *et al.*, 2005; Flores *et al.*, 2012; Kim and Hummer, 2008; Poulain *et al.*, 2008), making a thorough sampling of the conformational space nearly impossible. The common strategy in many protein folding or docking studies is to apply a multi-scale approach by exploring the conformational space first at the residue-level and then refining the structure(s) at the atomistic level (Chen and Xu, 2006; Murphy *et al.*, 2003; Vreven *et al.*, 2011).

By introducing an angle term, we were able to achieve much better prediction accuracy when compared with the multi-body potential in all the tests. In our current procedure, we adopted the distance correction function in the multi-body potential without normalizing the angle term (Liu *et al.*, 2005). As the distance and the angle are related, one potential future improvement of this orientation potential is to develop a methodology for normalizing a distance and angle function $F(r,\varphi)$.

This novel orientation potential could be useful in development of new docking algorithms, especially in flexible TF–DNA docking in which the starting structures are in unbound state and undergo conformational change on binding, as our residue-level potential is less sensitive to conformational changes (Chen and Xu, 2006; Murphy *et al.*, 2003; Vreven *et al.*, 2011). We will develop a strategy for applying the new potential to flexible protein–DNA docking, a much more challenging problem in structural bioinformatics.

ACKNOWLEDGEMENT

The authors thank Dr Jiancheng Jiang for discussions and comments on the manuscript.

Funding: This work was supported by the National Science Foundation (DBI0844749 to J.G.).

Conflict of Interest: none declared.

REFERENCES

Angarica,V.E. *et al.* (2008) Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC Bioinformatics*, **9**, 436.

Ashworth,J. and Baker,D. (2009) Assessment of the optimization of affinity and specificity at protein-DNA interfaces. *Nucleic Acids Res.*, **37**, e73.

Ayton,G.S. *et al.* (2007) Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr. Opin. Struct. Biol.*, **17**, 192–198.

Babu,M.M. *et al.* (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, **14**, 283–291.

Baker,E.N. and Hubbard,R.E. (1984) Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.*, **44**, 97–179.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bradley,P. *et al.* (2005) Toward high-resolution de novo structure prediction for small proteins. *Science*, **309**, 1868–1871.

Chen,Z. and Xu,Y. (2006) Structure prediction of helical transmembrane proteins at two length scales. *J. Bioinform. Comput. Biol.*, **4**, 317–333.

Contreras-Moreira,B. *et al.* (2010) Comparison of DNA binding across protein superfamilies. *Proteins*, **78**, 52–62.

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

Endres,R.G. *et al.* (2004) Toward an atomistic model for predicting transcription-factor binding sites. *Proteins*, **57**, 262–268.

Flores,S.C. *et al.* (2012) Multiscale modeling of macromolecular biosystems. *Brief. Bioinform.*, **13**, 395–405.

Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.

Gao,M. and Skolnick,J. (2008) DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. *Nucleic Acids Res.*, **36**, 3978–3992.

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

Janga,S.C. and Collado-Vides,J. (2007) Structure and evolution of gene regulatory networks in microbial genomes. *Res. Microbiol.*, **158**, 787–794.

Kaplan,T. *et al.* (2005) Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.*, **1**, e1.

Kim,R. and Guo,J.T. (2009) PDA: an automatic and comprehensive analysis program for protein-DNA complex structures. *BMC Genom.*, **10** (**Suppl. 1**), S13.

Kim,Y.C. and Hummer,G. (2008) Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. *J. Mol. Biol.*, **375**, 1416–1433.

Kim,R. *et al.* (2011) Benchmarks for flexible and rigid transcription factor-DNA docking. *BMC Struct. Biol.*, **11**, 45.

Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.

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.

Luscombe,N.M. and Thornton,J.M. (2002) Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.*, **320**, 991–1009.

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.

Miyazawa,S. and Jernigan,R.L. (1985) Estimation of effective interresidue contact energies from protein crystal-structures—quasi-chemical approximation. *Macromolecules*, **18**, 534–552.

Murphy,J. *et al.* (2003) Combination of scoring functions improves discrimination in protein-protein docking. *Proteins*, **53**, 840–854.

Pande,V.S. (2011) (Compressed) sensing and sensibility. *Proc. Natl Acad. Sci. USA*, **108**, 14713–14714.

Pingoud,A. *et al.* (2005) Type II restriction endonucleases: structure and mechanism. *Cell. Mol. Life Sci.*, **62**, 685–707.

Poulain,P. *et al.* (2008) Insights on protein-DNA recognition by coarse grain modelling. *J. Comput. Chem.*, **29**, 2582–2592.

Prabakaran,P. *et al.* (2006) Classification of protein-DNA complexes based on structural descriptors. *Structure*, **14**, 1355–1367.

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.

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.

Sippl,M.J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, **213**, 859–883.

Sippl,M.J. (1995) Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, **5**, 229–235.

van Dijk,M. and Bonvin,A.M. (2010) Pushing the limits of what is achievable in protein-DNA docking: benchmarking HADDOCK's performance. *Nucleic Acids Res.*, **38**, 5634–5647.

van Dijk,M. *et al.* (2006) Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Res.*, **34**, 3317–3325.

Vreven,T. *et al.* (2011) Integrating atom-based and residue-based scoring functions for protein-protein docking. *Protein Sci.*, **20**, 1576–1586.

Wade,R.C. and Goodford,P.J. (1993) Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds. *J. Med. Chem.*, **36**, 148–156.

Wu,J. *et al.* (2012) High performance transcription factor-DNA docking with GPU computing. *Proteome Sci.*, **10** (**Suppl. 1**), S17.

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

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.

Zhao,H. *et al.* (2010) Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. *Bioinformatics*, **26**, 1857–1863.

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