An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles

Beisi Xu[a,b,c,*], **Yuedong Yang**[b,c,*], **Haojun Liang**[a], and **Yaoqi Zhou**[b,c,*]

[a]Department of Polymer Science and Engineering, University of Science and Technology of China, Hefei, Anhui 230026, China

[b]Indiana University School of Informatics, Indiana University-Purdue University, Indianapolis, Indiana 46202, USA

[c] Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA

Abstract

How to make an accurate representation of protein-DNA interactions by an energy function is a long-standing unsolved problem in structural biology. Here, we modified a statistical potential based on the distance-scaled, finite ideal-gas reference state (DFIRE) so that it is optimized for protein-DNA interactions. The changes include a volume-fraction correction to account for unmixable atom types in proteins and DNA in addition to the usage of a low-count correction, residue/base-specific atom types, and a shorter cutoff distance for protein-DNA interactions. The new statistical energy functions are tested in threading and docking decoy discriminations and prediction of protein-DNA binding affinities and transcription-factor binding profiles. Results indicate that new proposed energy functions are among the best in existing energy functions for protein-DNA interactions. The new energy functions are available as a web-server called DDNA 2.0 at http://sparks.informatics.iupui.edu. The server version was trained by the entire 212 protein-DNA complexes.

INTRODUCTION

Precise control of the timing and amount of genes expressed is the basis for the existence of different cell types arranged in a complex structural pattern in a multicelluar organism despite having the identical genome of the organism. The regulation of gene expression is accomplished by specific binding between *cis*-regulatory regions of the genome and proteins such as transcription factors. Such specific binding is made possible by specific interactions between DNA and proteins.

Interaction between DNA and proteins could be described by various types of energy functions. Existing energy functions for protein-DNA interactions can be separated into direct and indirect readout components. Indirect readout refers to binding specificity caused by minimizing the energy penalty of DNA deformation upon protein binding[1⁻7] while direct readout involves specific binding due to specific interactions between proteins and DNA.

This paper focuses on searching for the specific energy function responsible for direct readout. Existing energy functions for protein-DNA binding can be classified as molecular-

mechanics-based2,8⁻12 and knowledge-based13⁻15. A molecular-mechanics-based energy function is approximated by physical interaction terms including bonded and nonbonded interactions whose parameters and weights are derived from experimental results and quantum/theoretical calculations of small 16⁻18 or macro-molecules 9,19. A knowledge-based energy function 13⁻15, on the other hand, is derived from statistical analysis of known protein-DNA structures, similar to knowledge-based potentials for proteins 20.

Different knowledge-based energy functions differ on how a reference state is defined. A reference state is a state when interactions are turned off. For example, Kono and Sarai 15,21 proposed a residue/base-level, three-dimensional grid potential based on a statistically averaged reference state proposed by Sippl 22. Zhang et al 14 employed a distance-scaled, finite ideal-gas (DFIRE) reference state23⁻25 for deriving protein-DNA interactions. Liu et al 13 developed a multi-body residue-base potential with an optimized, distance-dependent reference state. Robertson and Varani 26 applied a conditional probability formalism due to Samudrala and Moult 27. Donald et al. 11 applied several approximations including Quasi-chemical approximation 28 and generalized topological Go approximation 29.

The purpose of this paper is to develop a knowledge-based protein-DNA energy function based on a finite ideal-gas reference state. Our initial application of the DFIRE state to protein-DNA complexes was based on the idea that protein and DNA molecules share common atom types (only 19 atom types employed for both) 30. That is, the complexes are treated as a single mixable system and the original physical foundation of the DFIRE state (a state of ideal-gas mixture in a finite sphere) remains reasonable. However, if the atom types of proteins and those of DNA are different, the two types of atoms will locate at physically separated locations. A direct application of the DFIRE state to protein-DNA interactions 11,26 is no longer suitable.

In this paper, we introduce a volume-fraction correction to account for unmixable nature of protein and DNA atom types. In addition, we employ low-count corrections and a reduced interaction-distance cutoff to the finite ideal-gas reference state for protein-DNA interactions. The new proposed energy functions are tested in protein-DNA threading, docking decoy discrimination, binding affinity prediction, and prediction of transcription-factor binding profiles. To avoid overtraining, we employed separate training structural databases for different testing benchmarks.

MATERIALS AND METHODS

Residue/Base-Specific Atom Types

As the first application of the DFIRE energy function for protein-DNA interactions (referred as DDNA here and hereafter) 24, proposed statistical energy functions will be derived from the known structures of protein-DNA complexes. Unlike DDNA, the proposed energy functions treat atoms in proteins and those in DNA as completely different atom types (i.e. no overlapping atom types). More specifically, we employed residue or base-specific atom types as in Robertson and Varani 26. In other words, every protein and nucleic-acid heavy-atom type is considered in a residue/base-specific manner (e.g. C_α in alanine is a different atom type from that in leucine and C1' in adenine is a different atom type from that in guanine). All non-protein, non-DNA atom types were not employed. There are a total of 167 atom types for proteins and 82 atom types for DNA. For example, 82 atom types for DNA are resulted from 21, 19, 22, and 20 atoms in bases A, C, G, and T, respectively.

The Original DFIRE Energy Function

The following equation was employed to obtain the DFIRE-based, statistical atom-atom potential of mean force $\bar{u}^{DFIRE}$ between atom types i and j that are distance r apart: 23

$$
\bar{u}^{DFIRE}(i, j, r) = \begin{cases} -RT\ln \dfrac{N_{obs}(i,j,r)}{\left[\frac{r^{\alpha}\Delta r}{r_{cut}^{\alpha}\Delta r_{cut}}\right] N_{obs}(i,j,r_{cut})}, & r<r_{cut} \\ 0, & r \geq r_{cut} \end{cases}
$$
(1)

where R is the gas constant, $T=300K$, $\alpha=1.61$, $N_{obs}(i,j,r)$ is the number of *ij* pairs within the spherical shell at distance *r* observed in a given structure database, $r_{cut}=14.5\text{Å}$, and $\Delta r(\Delta r_{cut})$ is the bin width at $r(r_{cut})$ ($\Delta r=2\text{Å}$ for $r<2\text{Å}$, 0.5Å for $2\text{Å}<r<8\text{Å}$, and 1Å for $8\text{Å}<r<15\text{Å}$). The value of α was determined by the best fit of r^{α} to the actual distance-dependent number of ideal-gas points in finite protein-size spheres. We shall label the outcome of this equation as the DFIRE energy function for residue/base specific atom types. This equation was used to generate DDNA with 19 atom types for both proteins and DNA 14. It should be emphasized that choosing T=300K is arbitrary and RT is a scaling coefficient that does not have any effect on the results presented here because we are interested in the relative rather than the absolute energy value. Moreover, all knowledge-based energy function assumes that various protein structures belong to different snapshots of the same thermodynamic ensemble.

Distance-scaling and cutoff

When the DFIRE energy function was applied to protein-protein 31 and protein-DNA 14 interactions, it was applied only to interfacial residues 32 or atoms 14. Such a limit to interfacial atoms or residues indicates that it will be beneficial to limit the interaction range of the DFIRE for binding interactions. Similarly, Robertson and Varani 26 found that a shorter distance (10Å) cutoff leads to a more discriminative energy function for selecting native complex structures from protein-DNA docking decoys. Here, we will employ a 10Å cutoff without distance scaling. That is,

$$
\bar{u}^{FIRE}(i, j, r) = \begin{cases} -RT\ln \dfrac{P(i,j,r)}{P^{ref}(r)}, & r<r_{cut} \\ 0, & r \geq r_{cut} \end{cases}
$$
(2)

where $P(i,j,r) = N_{obs}(i,j,r)/\Sigma_r N_{obs}(i,j,r)$, $P^{ref}(r) = r^{\alpha}\Delta r/\Sigma_r r^{\alpha}\Delta r$, $r_{cut}=10\text{Å}$, and Δr is the bin width at r ($\Delta r=3\text{Å}$ for the first bin and 1Å for the next seven bins). We shall call this energy function as the FIRE energy function because distance scaling is no longer employed in this equation. The value of α was unchanged from 1.61 because this parameter was obtained physically by the best fit of r^{α} to the actual distance-dependent number of ideal-gas points in finite protein-size spheres.

Dirichlet Pseudocounts for low-count correction

Using residue/base-specific atom types will encounter low counts in same distance bins because of the small size of the existing database for protein-DNA complexes. Here, we adopt the low-count correction according to Bayesian statistics 26. In this method, number of atomic pairs in a given distance bin is corrected by a background distribution as followed:

$$
N_{obs}^{c}(i, j, r) = N_{obs}(i, j, r) + N_0 \sum_{i \in P, j \in D} N_{obs}(i, j, r) \Big/ \sum_{r', i \in P, j \in D} N_{obs}(i, j, r').
$$

where $N_0=75$ 26 and the summation over i,j is only over atomic pairs between a protein (P) and a DNA (D). This pseudocount correction leads to the energy function called cFIRE given by the equation:

$$\underline{u}^{cFIRE}(i, j, r) = \begin{cases} -RT\ln\frac{P^c(i,j,r)}{P^{ref}(r)}, & r<r_{cut} \\ 0, & r \geq r_{cut} \end{cases}$$

(3)

where $P^c(i, j, r) = N^c_{obs}(i, j, r) / \Sigma_r N^c_{obs}(i, j, r)$.

One could further introduce a low counter correction as Sippl first introduced for deriving a distance-dependent potential with a small number of protein structures22. We found that it did not make a statistically significant additional improvement in testing. Thus, we will not introduce the correction here.

Volume-Fraction Correction

The above equation was derived with a reference state of uniformly distributed points within finite-sized spheres. That is, all atom types mix with each other well. However, residue/base-specific atom types do not mix with each other and they are located in either DNA or proteins. As a result, it becomes necessary to replace the volume element ($4\pi r^2 \Delta r$ for an infinite ideal-gas mixture or $4\pi r^\alpha \Delta r$ for the DFIRE approximation) by the fraction of the volume element occupied by protein-DNA atomic pairs. This volume-fraction correction leads to an equation for vcFIRE between protein and DNA atoms given by

$$\underline{u}^{vcFIRE}(i, j, r) = \begin{cases} -RT\ln\frac{P^c(i,j,r)}{P_V^{ref}(r)}, & r<r_{cut} \\ 0, & r \geq r_{cut} \end{cases}$$

(4)

where $P_V^{ref}(r) = (r^\alpha \Delta r) f_V(r) / \Sigma_r [(r^\alpha \Delta r) f_V(r)]$ and the molar fraction of protein-DNA interaction pairs $f_V(r) = N^{PD}_{obs}(r) / N_{obs}(r)$ with the number of all atomic pairs $N_{obs}(r) = \Sigma_{i,j} N_{obs}(i, j, r)$ and the number of atomic pairs between a protein and a DNA $N^{PD}_{obs}(r) = \Sigma_{i \in P, j \in D} N_{obs}(i, j, r)$.

Database of Protein-DNA complexes

$N_{obs}(i,j,r)$ is obtained from a structural database of non-redundant high-resolution protein-DNA complexes (X-Ray, resolution<3.0Å). The database is built on protein-DNA complexes collected from the PDB database and culled by the PISCES server 33 at http://dunbrack.fccc.edu/PISCES.php with maximum sequence identity of 35% by PDB entry. The database contains 212 protein-DNA complexes. To avoid overtraining, we used different training sets of protein-DNA complexes for different tests because different test sets are made of different protein-DNA complexes. In each test, we removed those training protein-DNA complexes whose protein sequences have more than 35% sequence identity (blastp with an expectation value of 0.000134) with the proteins in the test set.

Testing proposed energy functions

The free energy for the formation of a protein-DNA complex, ΔG, is approximated as follows.

$$\Delta G = \sum_{i,j} \bar{u}(i, j, r)$$

where the summation is between all atomic pairs between a protein and a DNA. $\bar{u}(i,j,r)$ can be from either DFIRE [Eq. (1)], FIRE [Eq. (2)], cFIRE [Eq. (3)], or vcFIRE [Eq. (4)]. This equation assumes a rigid-body docking during the formation of protein-DNA complexes and neglects the contributions from DNA deformation and from possible binding-induced change of the protein conformation. That is, intra-protein and intra-DNA interactions are assumed to be unchanged during the binding.

Test 1: DNA threading decoys—The protein-DNA threading benchmark is made of 51 complexes collected by Kono and Sarai 15. For each protein-DNA complex, we generate 50,000 evenly distributed random DNA sequences. That is, each base has a probability of 0.25. The DNA structure of a random sequence is constructed by fixing the phosphate-deoxyribose backbone and overlapping the new base pair with the position of the native base pair. In this test, we employ a training database of 166 complexes after removing 46 complexes in the dataset of 212 complexes that have higher sequence identity than 35% with the 51 testing complexes (See Table 1).

The ability of an energy function to discriminate a native DNA sequence from randomly generated DNA sequences is measured by *Z-Score* with *Z-Score=($\Delta G_{\mathrm{native}}-\Delta G_{\mathrm{ave}}$)/S* and ΔG_{ave} and *S* are the average and standard deviation of the free energy values of threading decoy complexes, respectively. To ensure the accuracy of obtained *Z-score* values, we calculated the average of 200 *Z-Score* values by generating 199 additional sets of 50000 decoys per protein-DNA complexes. Each set was generated with different random numbers. We report the average and standard deviations of the *Z-Score* values.

Test 2: Docking Decoy discrimination—We obtained near-native docking decoy sets of 45 protein-DNA complexes from Robertson and Varani26. There are 2000 lowest-RMSD decoys for each complex generated by FTDock and near-native structures generated from restraints around native complex structures. For this test, a non-homologous training dataset of 167 complexes is employed (removing 45 complexes in 212 training complexes with sequence identity higher than 35% with these 45 test complexes, see Table 1). Similar to the DNA threading test above, the ability of an energy function to discriminate a native conformation from decoy conformations is measured by Z-Score.

Test 3: Recovering native base pairs—For a given protein-DNA complex, each base pair is replaced by three other possible base pairs. The total free energy between the native base pair and the protein is compared to the energy values between the three other base pairs and the protein. If the native base pair has the lowest energy, that native base pair is successfully recovered by the energy function. We measure the success rate for recovering native base pairs by calculating the fraction of recovered native base pairs in total number of base pairs in the DNA sequence. This success rate is averaged over the number of protein-DNA complexes. (Here, we have assumed that the contribution from the intra-DNA interaction is negligible with the assumption of rigid-body docking.) A ten-fold cross validation is performed for this test based on a randomly selected 200 complexes in the dataset of 212 complexes. We randomly divide the 200 complexes into 10 parts ("folds"). Each fold has 20 complexes. In each test, nine folds are used for training and the remaining fold is for testing. This test is repeated 10 times to cover every fold.

Test 4:Binding free energy prediction—We employed a binding database (ΔG) due to Donald et al. 11, which is a modified version of Zhang et al14. This database contains 30 protein-DNA complexes. For this test, we use a training dataset of 185 protein-DNA complexes after removing 17 complexes from the dataset of 212 complexes. These 185 protein-DNA complexes have less than 35% sequence identity with the 30 testing complexes (see Table 1).

Test 5:Mutation-induced change in binding free energy prediction—For mutation-induced change (ΔΔG) in binding free energy, we also approximated it as the energy difference between mutant and wild type (ΔΔG=ΔG^{mutant}- ΔG^{wild}). The ΔΔG dataset is from Morozov et al9 and modified by Donald et al11. This database contains 189 mutants of 10 protein-DNA complexes. For this test, we also remove 10 sequence-homologous protein-DNA complexes from the training set. That is, the training set for this test contains 202 complexes (See Table 1).

Test 6:Prediction of Position-Specific Weight Matrix (PWM)—Our approximate protein-DNA interaction for the binding free energies allows the decomposition of the predicted binding free energies into the contributions by each individual base. That is,

$$\Delta G = \sum_i \Delta G_\alpha^i$$

where ΔG_α^i is the binding free energy of a base α (A, C, G, or T) at position i. In our proposed energy functions, ΔG_α^i is independent of all other bases. We can calculate position-specific weight matrix (PWM) of a given base α at a given position i by using the Boltzmann formula:

$$p_\alpha^i = \frac{\exp\left(-\beta \Delta G_\alpha^i\right)}{\sum\limits_{\gamma=1}^{4} \exp\left(-\beta \Delta G_\gamma^i\right)}$$

where γ represents different bases,$\beta=1/RT$ is the inverse of temperature and employed as a fitting parameter. The significance of PWM prediction is evaluated by $\Psi-$ test. $\Psi-$ test9 is a generalization of well-know $\times^2$ – test9:

$$\psi(p,q) = \frac{1}{L}\left[\sum_{i=1}^{L} \sum_{\gamma=1}^{4} q_\gamma^i \ln \frac{q_\gamma^i}{p_\gamma^i}\right]$$

where p_γ^i is the predicted probability of base γ at position i, q_γ^i is the experimental frequency and L is the length of base pairs. To avoid zero denominators, both p and q distributions are smoothed by adding 0.05 to all PWM entries and re-normalizing to avoid zero probabilities at denominator. Morozov et al9 also evaluated $\Psi(p_{random},q)$by comparing randomly predicted p_{random} matrix against the experimental matrix q. Each random weight matrix was calculated by sampling four numbers in (0, 1) interval and normalized. An average of 10000 $\Psi(p_{random},q)$ was obtained. The difference between $<\Psi(p_{random},q)>$ and $\Psi(p,q)$ measures the successfulness of the predicted PWM. We use the database of 19 complexes with experimental PWM values collected by Morozov et al 9. We have removed 1ihf from their original 20-complex set because of the mismatch between the PWM and the DNA bases in the 1ihf complex structure. Homologous protein sequences to these 19 complexes are

excluded from our training set. That is, 194 complexes are used for training our energy functions in this particular test (see Table 1).

RESULTS

Test 1: Sequence-Decoy Discrimination

In Table 2, we compare the average Z-scores given by different variants of DFIRE energy functions along with the results given by Gromiha et al 35. Each average Z-score is an average of 200 Z-scores generated by random 50000 sequence decoys. A more negative Z-score indicates a larger normalized gap between the energy of a native complex structure and the average energy of sequence decoys. The standard deviations of the Z-score values for all 51 protein-DNA complexes are between 0 and 0.03. Thus, the results are stable. Table 2 shows that reducing the range of interaction from DFIRE to FIRE makes a significant improvement in mean Z-scores from -0.5 to -2.2. Addition of volume correction (vFIRE) makes no significant change. A low-count correction based on Dirichlet pseudocount (cFIRE) further improves the Z-score to -2.8 from -2.2 (P value less than 0.0001 according to the paired T-test, GraphPad Software: http://www.graphpad.com/quickcalcs/ttest1.cfm), while no significant change is observed for the introduction of further volume correction (vcFIRE) in this test (P value=0.17). The number of positive Z-score values (where the average energy of sequence decoys is lower than the energy of native DNA sequence) is reduced from 3 in FIRE, 2 in vFIRE, 1 in cFIRE, to 0 in vcFIRE. For majority of protein-DNA complexes, the Z-score values given by vcFIRE are lower than that given by FIRE or by DFIRE. There are a few exceptions. For example, Z-score for 1dp7 is -3.65 by FIRE, -3.61 by vFIRE, -3.47 by cFIRE and -3.00 by vcFIRE. In this case, all correction terms failed to improve Z-score. This is somewhat expected because proposed corrections are approximations and unlikely to improve Z-score in every case.

Nevertheless, the average Z-score values (-2.2 to -2.86) given by various FIRE energy functions are significantly lower than the two methods proposed by Gromiha et al35 (-1.7 and -1.8, respectively). As a comparison, we also applied DDNA 14 to this threading set. DDNA's Z-scores are close to DFIRE's, in average.

Test 2: Docking-Decoy Discrimination

The second test measures the ability of the proposed energy functions to recognize the native complex structures from near-native docking decoys made by Robertson and Varani. Table 3 compares Z-Score values given by four variants of DFIRE-based energy functions (DDNA, FIRE, cFIRE, and vcFIRE) along with the results based on the Robertson and Varani energy function 26 trained by the same 167 complexes. The average Z-Score changes from -2.22 (FIRE), -2.14 (vFIRE), -2.02 (cFIRE) to -2.80 (vcFIRE) as the low-count and volume-fraction corrections are added to the residue/base-specific FIRE energy function. In this test, a combination of volume-fraction and low-count corrections makes a significant improvement over FIRE while individual correction term makes a small but negative impact on Z-score [from -2.22 to -2.02 (cFIRE) or to -2.14 (vFIRE)]. This indicates that individual correction term may not be always beneficial because small databases affect both correction terms. The average Z-score given by vcFIRE (-2.80) is also lower than that (-2.06) given by the Robertson and Varani energy function 26.

A more challenging test is the ability to identify near-native complexes by various energy functions (i.e. predicting the best structure from available decoys). Table 4 compares the lowest rmsd structure in top five decoys ranked by various DFIRE energy functions, along with the best possible decoy structure in the decoy set. The median of the best rmsd values in top five for the 45 protein-DNA complexes is 0.51Å by FIRE, 0.55 by vFIRE, 0.50Å by

cFIRE, and 0.46Å by vcFIRE. The latter is close to the best possible median value of 0.44Å. The Robertson and Varani energy function 26 yields a median value of 0.50Å, the same as cFIRE and higher than vcFIRE. If we define a failure in prediction as the best rmsd value in top five predictions is greater than 2Å, there are 3 by FIRE, 10 by vFIRE, 0 by cFIRE, 0 by vcFIRE and 1 by the Robertson and Varani energy function26 . This indicates that volume fraction correction without low count correction significantly reduces the ability of the energy function for locating near-native structures.

Test 3: Recovering Native Base Pairs

Table 5 reports 10-fold cross-validated average success rates for recovering native base pairs of 200 protein-DNA complexes (See methods). All four tested methods yield essentially the same success rate of 40%. This success rate is substantially higher than 25% success rate by random selection and 31% by DDNA.

Tests 4: Binding free energy

Figure 1 compares theoretically predicted binding affinities with experimentally measured ones for 30 protein-DNA complexes (See methods). The correlations between theoretical results and experimental data are all significant. The correlation coefficients are 0.84 by DDNA, 0.79 by FIRE, 0.55 by vFIRE (not shown), 0.85 by cFIRE, and 0.72 by vcFIRE. It is not clear how to interpret the variations observed in correlation coefficients by different approximations. More studies are certainly needed when a large database of protein-DNA binding affinities with corresponding complex structures becomes available.

Tests 5: Mutation-induced change in stability

Table 6 compares the correlation coefficients between theoretically predicted changes and experimentally measured changes in stability due to mutation. For a majority of protein-DNA complexes, there is no significant correlation. In fact, the overall correlation coefficients for all 189 mutants are nearly zero for DDNA, FIRE, vFIRE, cFIRE, and vcFIRE. This highlights the challenge for $\Delta\Delta G$ prediction 11,26.

Test 6: Prediction of Position-Specific Weight Matrix (PWM)

Table 7 compares the average values of Ψ-test of 19 complexes given by various DFIRE energy functions. A smaller value indicates a better agreement with experimental results. There is a statistically significant improvement from DDNA (0.46), FIRE (0.36), vFIRE (0.36) to cFIRE (0.33) or vcFIRE (0.33) (e.g. P-value is 0.035 between FIRE and vcFIRE). All these values are significantly lower than 0.71, the average random value of Ψ-test. As shown in the table, the overall accuracy of FIRE-based energy functions is similar to that of the dynamic model proposed by Morozov et al. 9 but is not as accurate as their static model. Higher accuracy of the static model is likely due to its direct training by experimental $\Delta\Delta G$ values (More in discussion).

Fig. 2 shows the most successful PWM prediction by the variants of FIRE-based energy functions for the phage lambda repressor protein (lambdaR). For example, vcFIRE's prediction yields 13/17 of the base pairs with the highest weight the same as the experiment results.

DISCUSSION

In this paper, we have developed statistical energy functions based on finite, ideal-gas reference (FIRE) state for protein-DNA interactions. The new proposed methods further extend the statistical energy function based on the distance-scaled FIRE (DFIRE) state that was originally developed for proteins 23¯25 and applied to protein-DNA interactions

(DDNA) 14. Significant improvements over DDNA by FIRE-based energy functions are observed for threading and docking decoy discriminations, recovery of native base pairs, and prediction of binding profiles. These improvements are due to a combination of following factors: a reduction of interaction range from 15Å to 10Å, an employment of residue/base-specific atom types, a low-count correction, and volume-fraction correction. We further show that low-count correction alone (cFIRE) is found useful for DNA threading and PWM prediction but not for docking while the volume correction is most effective only if it is combined with the low count correction. The first three factors were also used to improve the accuracy of the RAPDF statistical potential for protein-DNA interactions 26.

It is of interest to compare the performance of FIRE-based energy functions with other statistical energy functions. For DNA threading decoys (Test 1), the Z-score values given by various FIRE energy functions are significantly lower than the two methods proposed by Gromiha et al35. For the docking decoy set (Test 2), Robertson and Varani's energy function is less discriminative than vcFIRE with a higher average Z-score of -2.06 26. In Tests 4 and 5, we found that FIRE-based energy functions can make a reasonable prediction of ΔG but not $\Delta\Delta G$. This result puts FIRE-based energy functions close to the performance of physical-based energy functions tested by Donald et al. 11. (For example, a simple Lennard-Jones energy function gives a correlation of 0.76 for ΔG and 0.23 for $\Delta\Delta G$.) No knowledge-based energies tested by Donald et al. 11 give accurate prediction of either ΔG or $\Delta\Delta G$. Finally, position specific weight matrices (PWM) predicted by FIRE-based energy functions are less accurate than the static model and comparably accurate to the dynamic model given by Rosetta 9 as demonstrated in Table 7. Compared to FIRE-based energy functions, the Rosetta energy function contains many physical and knowledge-based energy terms whose relative weights were optimized by using experimental ΔG and $\Delta\Delta G$ data in the static model or native recovery of native protein amino acid side chains in the dynamic model. Because the protein-DNA complexes in the database for ΔG and $\Delta\Delta G$ overlap with the complexes in the database for PWM test, the static model may have been over-trained for the PWM test. Here, we make an effort to avoid over training by employing separate training sets for different test sets.

It has been shown that a model with reduced atom types is less accurate than a model with residue/base specific atom types (for example, in Robertson et. al's work26). We also tested a version of FIRE with unmixable 12 atom types for proteins and 11 atom types for DNA (same as 19 atom types in DDNA except that atom types for proteins and DNA do not mix with each other). This version of FIRE leads to an average Z-score of -1.66 for the threading decoy set, a significant reduction (P-value of 0.0009 for paired t-test) from -2.23 for FIRE based on residue/base-specific atom types. This confirms the utility of residue/base-specific atom types.

This work represents an optimized version of the finite ideal gas reference state for protein-DNA interactions. Initial tests of the proposed FIRE-based energy functions indicate that they are among the best in existing energy functions for protein-DNA interactions. This is encouraging because there is room for further improvement. Examples are incorporation of the effect of DNA conformational changes and orientation-dependence of the protein-DNA interaction. Recently, we have developed a dipolar DFIRE (dDFIRE) energy function for proteins36. In this energy function, each polar atom is treated as a dipole with a direction and the orientation dependence of polar interactions is extracted from protein structures. This approach takes into account the hydrogen-bonding interaction via the physical dipole-dipole interaction and the possible orientation-dependent interactions between polar and nonpolar atoms and between polar atoms that are non-hydrogen-bonded. The development of a corresponding dipolar vcFIRE is in progress.

Acknowledgments

REFERENCES

1. Olson, WK.; Gorin, AA.; Lu, XJ.; Hock, LM.; Zhurkin, VB. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes; Proceedings of the National Academy of Sciences of the United States of America; 1998. p. 11163-11168.

2. Endres RG, Schulthess TC, Wingreen NS. Toward an atomistic model for predicting transcription-factor binding sites. Proteins-Structure Function and Bioinformatics. 2004; 57(2):262–268.

3. Endres RG, Wingreen NS. Weight matrices for protein-DNA binding sites from a single co-crystal structure. Phys Rev E Stat Nonlin Soft Matter Phys. 2006; 73(6):061921. [PubMed: 16906878]

4. Paillard G, Deremble C, Lavery R. Looking into DNA recognition: zinc finger binding specificity. Nucleic Acids Research. 2004; 32(22):6673–6682. [PubMed: 15613596]

5. Arauzo-Bravo MJ, Fujii S, Kono H, Ahmad S, Sarai A. Sequence-dependent conformational energy of DNA derived from molecular dynamics simulations: Toward understanding the indirect readout mechanism in protein-DNA recognition. Journal of the American Chemical Society. 2005; 127(46): 16074–16089. [PubMed: 16287294]

6. Aeling KA, Opel ML, Steffen NR, Tretyachenko-Ladokhina V, Hatfield GW, Lathrop RH, Senear DF. Indirect recognition in sequence-specific DNA binding by Escherichia coli integration host factor - The role of DNA deformation energy. Journal of Biological Chemistry. 2006; 281(51): 39236–39248. [PubMed: 17035240]

7. Becker NB, Wolff L, Everaers R. Indirect readout: detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials. Nucleic Acids Research. 2006; 34(19):5638–5649. [PubMed: 17038333]

8. Paillard G, Lavery R. Analyzing protein-DNA recognition mechanisms. Structure. 2004; 12(1):113–122. [PubMed: 14725771]

9. Morozov AV, Havranek JJ, Baker D, Siggia ED. Protein-DNA binding specificity predictions with structural models. Nucleic Acids Research. 2005; 33(18):5781–5798. [PubMed: 16246914]

10. Huang N, MacKerell AD. Specificity in protein-DNA interactions: Energetic recognition by the (cytosine-C5)-methyltransferase from HhaI. Journal of Molecular Biology. 2005; 345(2):265–274. [PubMed: 15571720]

11. Donald JE, Chen WW, Shakhnovich EI. Energetics of protein-DNA interactions. Nucleic Acids Research. 2007; 35(4):1039–1047. [PubMed: 17259221]

12. Siggers TW, Honig B. Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. Nucleic Acids Research. 2007; 35(4):1085–1097. [PubMed: 17264128]

13. Liu Z, Mao F, Guo J-t, Yan B, Wang P, Qu Y, Xu Y. Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential. Nucl Acids Res. 2005; 33(2):546–558. [PubMed: 15673715]

14. Zhang C, Liu S, Zhu QQ, Zhou YQ. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. Journal of Medicinal Chemistry. 2005; 48(7):2325–2335. [PubMed: 15801826]

15. Kono H, Sarai A. Structure-based prediction of DNA target sites by regulatory proteins. Proteins-Structure Function and Genetics. 1999; 35(1):114–131.

16. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. Journal of Computational Chemistry. 1983; 4(2):187–217.

17. Cheatham TE, Young MA. Molecular dynamics simulation of nucleic acids: Successes, limitations, and promise. Biopolymers. 2000; 56(4):232–256. [PubMed: 11754338]

18. Ponder JW, Case DA. Force fields for protein simulations. Protein Simulations. 2003; 66:27–+.

19. Havranek JJ, Duarte CM, Baker D. A simple physical model for the prediction and design of protein-DNA interactions. Journal of Molecular Biology. 2004; 344(1):59–70. [PubMed: 15504402]

20. Skolnick J. In quest of an empirical potential for protein structure prediction. Current Opinion in Structural Biology. 2006; 16(2):166–171. [PubMed: 16524716]

21. Selvaraj S, Kono H, Sarai A. Specificity of protein-DNA recognition revealed by structure-based potentials: Symmetric/asymmetric and cognate/non-cognate binding. Journal of Molecular Biology. 2002; 322(5):907–915. [PubMed: 12367517]

22. Sippl MJ. Calculation of Conformational Ensembles from Potentials of Mean Force - an Approach to the Knowledge-Based Prediction of Local Structures in Globular-Proteins. Journal of Molecular Biology. 1990; 213(4):859–883. [PubMed: 2359125]

23. Zhou HY, Zhou YQ. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science. 2002; 11(11):2714–2726. [PubMed: 12381853]

24. Zhou HY, Zhou YQ. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction (vol 11, pg 2714, 2002). Protein Science. 2003; 12(9):2121–2121.

25. Zhou Y, Zhou HY, Zhang C, Liu S. What is a desirable statistical energy function for proteins and how can it be obtained? Cell Biochemistry and Biophysics. 2006; 46(2):165–174. [PubMed: 17012757]

26. Robertson TA, Varani G. An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure. Proteins-Structure Function and Bioinformatics. 2007; 66(2):359–374.

27. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of Molecular Biology. 1998; 275(5):895–916. [PubMed: 9480776]

28. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins-Structure Function and Genetics. 2001; 44(3):223–232.

29. Kussell, E.; Shimada, J.; Shakhnovich, EI. A structure-based method for derivation of all-atom potentials for protein folding; Proceedings of the National Academy of Sciences of the United States of America; 2002. p. 5343-5348.

30. Zhang C, Liu S, Zhou HY, Zhou YQ. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Science. 2004; 13(2):400–411. [PubMed: 14739325]

31. Liu S, Zhang C, Zhou HY, Zhou YQ. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. Proteins-Structure Function and Bioinformatics. 2004; 56(1):93–101.

32. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein-protein interactions. Biophys J. 2003; 84(3):1895–1901. [PubMed: 12609891]

33. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics. 2003; 19(12):1589–1591. [PubMed: 12912846]

34. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997; 25(17):3389–3402. [PubMed: 9254694]

35. Gromiha MM, Siebers JG, Selvaraj S, Kono H, Sarai A. Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. Journal of Molecular Biology. 2004; 337(2):285–294. [PubMed: 15003447]

36. Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins: Structure, Function, and Bioinformatics. 2008; 72(2):793–803.
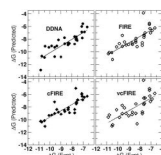
Figure 1.
Experimentally measured binding affinity (-log(Kd) unit) as compared to theoretically predicted values by DDNA (Filled circles), FIRE (Open circles), cFIRE (Filled Diamonds), and vcFIRE (Open Diamonds). The correlation coefficients between respective theoretical predictions and experimental measurements are 0.84, 0.79, 0.85 and 0.72, respectively. Theoretical results are scaled and shifted so that all results can be viewed in a single diagram.
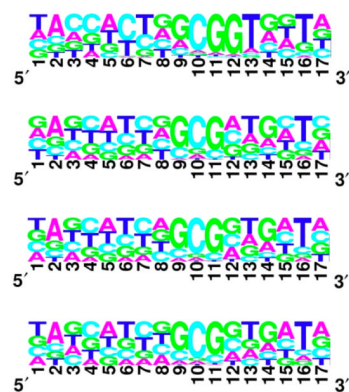
Figure 2.
PWM prediction of phage lambda repressor protein (lambdaR, PDB id:1lmb) given by experiment, FIRE, cFIRE, and vcFIRE (from top to bottom), respectively. vFIRE is not shown because it is similar to that of cFIRE.

Table 1

A list of separate training sets that are made for different test sets. The maximum allowed sequence identity between a protein in a training set and a protein in the corresponding test set is 35%.

#	Name	Training Set[a]	Test Set[b]
1	Threading Decoys	166	51 (200[c])
2	Docking Decoys	167	45 (2000[d])
3	Native base-pair recovery	180	20 (10 fold[e])
4	Binding affinity	185	30
5	Mutation-induced binding affinity	202	10 (189[f])
6	Binding profiles	194	19

[a]The number of protein-DNA complexes in the training set.

[b]The number of protein-DNA complexes in the testing set.

[c]200 sets of 50000 DNA sequence decoys are generated for each protein-DNA complex.

[d]2000 decoys per complex.

[e]10 fold cross validation for success rate in recovering the native base pairs.

[f]10 protein-DNA complexes with a total of 189 mutants.

Table 2

The Z-scores given by various methods for random sequence decoys (DNA threading) of 51 complexes. A lower Z-score given by a method indicates a stronger bias toward native DNA sequence.

PDB ID[a]	G1[b]	G2[c]	DDNA[d]	DFIRE[e]	FIRE[f]	vFIRE[g]	cFIRE[h]	vcFIRE[i]
1a02	−3.4	−1.8	−2.27	1.48	−3.21	−3.29	−3.19	−3.29
1a74	−1.6	0.7	1.50	−1.40	−3.86	−3.94	−4.07	−4.17
1b3t	−1.4	−2.1	−1.15	−0.68	−0.79	−0.61	−2.78	−2.38
1bhm	−2.9	−1.3	−0.05	−0.32	−1.08	−1.08	−3.13	−3.26
1bl0	−2.7	−2.5	−2.23	−0.45	−2.46	−2.42	−3.24	−3.25
1cdw	−2.2	−0.6	1.64	1.59	0.83	0.87	0.04	−0.02
1cjg	−1.1	−1.4	−2.58	−0.39	−0.28	−0.33	−0.50	−0.81
1cma	−0.2	−1.6	1.02	−1.19	−1.63	−1.72	−1.59	−1.59
1e66	−1.8	−1.7	−3.22	−1.60	−1.96	−1.88	−3.18	−3.12
1dp7	−0.8	−0.7	0.76	−1.92	−3.65	−3.51	−3.47	−3.00
1ecr	−1.1	−1.1	0.53	0.43	−1.22	−1.19	−1.66	−1.58
1fjl	−2.7	−1	2.59	1.15	−2.10	−1.98	−2.76	−2.63
1gat	−0.4	−1.7	1.73	1.80	−1.04	−0.98	−1.48	−1.27
1gdt	−2	−1.7	−0.04	0.95	−2.25	−1.99	−3.98	−3.75
1glu	−1.1	−1.1	1.72	−0.12	−0.66	−0.70	−1.80	−1.95
1hcq	−1.7	−2.5	−0.85	−2.48	−3.14	−2.97	−4.35	−4.09
1hcr	−1.8	0.4	−0.25	0.94	−1.09	−0.89	−2.47	−2.43
1hdd	−1.1	−1.8	0.95	1.20	−0.73	−0.81	−1.29	−1.57
1hlo	0.1	−1.6	0.29	−2.26	−3.60	−3.67	−3.70	−3.95
1hry	−0.2	−0.9	0.23	−0.78	−1.06	−0.99	−1.42	−1.33
1if1	−0.4	−1.7	−1.62	0.70	0.03	0.18	−2.00	−1.96
1ign	0	−2.2	−0.23	−3.16	−4.47	−4.41	−5.33	−5.32
1ihf	−1.2	−2.3	1.79	0.65	−1.48	−1.60	−1.59	−1.81
1j59(1ber)	−2	−0.8	−2.33	−0.37	−3.55	−3.56	−3.87	−3.79
1lmb	−2.9	−4.3	−1.48	−1.93	−3.49	−3.72	−3.73	−4.25
1mdy	−0.7	−2.5	2.81	−1.13	−2.95	−2.90	−2.95	−2.83
1mey	−3.6	−2.2	−1.52	−2.45	−4.85	−4.74	−5.12	−4.92
1mhd	−1.9	−1.9	0.56	−0.46	−2.34	−2.43	−2.61	−2.74

PDB ID[a]	G1[b]	G2[c]	DDNA[d]	DFIRE[e]	FIRE[f]	vFIRE[g]	cFIRE[h]	vcFIRE[i]
1mmm	−4.4	−3	0.20	−0.10	−3.77	−3.89	−3.86	−4.04
1mse	−0.4	−2	−0.69	−0.25	−1.76	−1.84	−2.06	−2.13
1oct	−1.6	−2.1	−0.37	0.96	−2.77	−2.52	−3.09	−2.85
1par	0.6	−1.7	−0.96	1.24	−0.67	−1.02	−1.99	−2.42
1pdn	−2	−2.5	−1.06	−0.57	−0.44	−0.44	−1.86	−1.92
1per	−2.5	−1.1	0.20	0.77	−0.59	−0.90	−1.45	−1.92
1pue	−1.1	−2.7	−1.27	−0.81	−1.50	−1.73	−1.86	−2.21
1rep	−2	−3.2	−2.20	−1.25	−3.04	−3.05	−3.11	−3.01
1rv5	−2.3	−0.3	0.11	1.01	−1.87	−1.95	−1.69	−1.67
1srs	−3	−2.4	0.67	−1.15	−3.13	−3.45	−2.97	−3.62
1svc	−2.6	−2.2	−1.68	−2.75	−3.80	−3.82	−4.20	−4.27
1tc3	−1.7	−2.5	−0.24	−2.20	−2.51	−2.51	−2.28	−2.29
1tf3	−3.2	−2.3	−1.19	−2.28	−2.26	−2.21	−3.61	−3.56
1tro	−1.3	−3.1	−0.19	0.08	−3.43	−3.40	−4.16	−4.05
1tsr	−1.1	−1.2	−2.38	−0.43	−1.75	−1.57	−2.90	−2.68
1ubd	−1.3	−2.1	−0.12	−1.08	−3.46	−3.49	−3.85	−4.00
1xbr	−2	−2.4	−2.76	−0.18	−1.88	−1.86	−2.39	−2.40
1yrn	−4.4	−2.9	−0.05	−0.13	−2.98	−2.76	−4.03	−3.78
1ysa	−3	−2.1	0.14	−0.68	−3.66	−3.92	−3.65	−4.01
2bop	−0.9	−1.7	−2.16	−3.24	−2.75	−2.67	−3.28	−3.12
2drp	−1.2	−2.3	1.40	0.14	−3.75	−3.91	−4.37	−4.75
3cro	−2	0.3	−1.52	1.74	0.17	−0.23	−0.00	−0.57
6cro	0	−2.3	−3.86	−2.01	−3.80	−3.79	−3.78	−3.79
Mean	−1.7	−1.8	−0.43	−0.50	−2.23	−2.24	−2.82	−2.86
s[j]	1.1	0.9	1.52	1.34	1.34	1.33	1.21	1.17

[a] Protein databank identification code.

[b] Distant-dependent statistical potentials for the specific base-amino acid interactions derived from the spatial distributions of Cα atoms of amino acid residues around a base. Consider intermolecular interactions only 35.

[c] Consider intramolecular only15,21.

[d] DDNA [DFIRE with 19-atomic types acted on interfacial atoms only, 35].

[e] DFIRE with residue/base-specific atom types and 15Å cutoff.

[f] FIRE with residue/base-specific atom types and 10Å cutoff without distance scaling.

[g] Volume-fraction correction added to FIRE (vFIRE).

[h] Low-count correction with Dirichlet pseudo counts added to FIRE (cFIRE).

[i] Volume-fraction correction added to cFIRE (vcFIRE).

[j] Standard Derivation of Z-Score values in 51 complexes.

Table 3

Z-score values between the native complex structure and near-native docking decoys of 45 protein-DNA complexes given by various energy functions.

PDB[a]	Γ(°)[b]	DDNA[c]	FIRE[d]	vFIRE[e]	cFIRE[f]	vcFIRE[g]	RV[h]
1qna	35.7	−1.21	−1.65	−1.07	−1.64	−1.79	−1.57
1d02	13.62	−1.47	−2.02	−1.42	−1.89	−2.63	−1.95
1eon	13.41	−1.66	−2.41	−2.49	−1.99	−3.09	−1.98
1ckq	12.29	−1.02	−1.51	−1.55	−1.22	−1.94	−1.14
1dmu	9.12	−1.55	−2.92	−1.36	−2.05	−4.16	−2.06
1qpz	8.53	−2.20	−2.71	−2.06	−2.52	−3.48	−2.55
1au7	8.48	−1.52	−2.28	−3.59	−1.91	−2.55	−1.96
1je8	8.15	−1.85	−2.23	−2.66	−2.04	−2.91	−2.04
2cgp	7.84	−0.97	−1.51	−2.15	−1.36	−1.99	−1.42
1b3t	7.74	−1.38	−2.47	−0.74	−1.95	−2.99	−1.94
1tc3	7.3	−1.56	−1.44	−1.27	−1.78	−2.67	−1.56
1g9z	7.17	−2.63	−4.19	−2.50	−3.31	−5.45	−3.29
1zme	6.84	−2.13	−1.63	−0.40	−2.20	−2.38	−2.26
1a73	6.56	−1.85	−2.45	−2.67	−2.23	−3.41	−2.30
1jko	6.55	−1.77	−2.39	−2.46	−2.29	−3.12	−2.16
1bdt	6.41	−1.77	−1.91	−1.01	−1.82	−3.19	−1.88
2bop	6.28	−1.68	−2.51	−2.88	−2.04	−2.97	−2.13
1a1i	6.21	−1.44	−2.13	−2.71	−1.92	−2.49	−1.98
1bc8	6.1	−1.50	−2.43	−2.27	−2.16	−2.67	−2.10
1pdn	6.04	−1.45	−1.42	−1.31	−2.13	−2.47	−2.17
1skn	5.96	−1.23	−2.14	−2.83	−1.96	−2.60	−2.06
1mjo	5.94	−2.09	−2.18	−2.21	−2.13	−2.55	−2.16
1bl0	5.88	−0.96	−2.23	−1.76	−1.47	−1.92	−1.40
2dgc	5.75	−1.46	−2.07	−2.49	−1.97	−2.36	−2.06
3pvi	5.71	−1.65	−1.97	−2.47	−1.83	−2.34	−1.86
2hdd	5.61	−2.37	−2.80	−2.98	−2.64	−3.13	−2.70
1ign	5.19	−1.74	−2.59	−2.35	−2.24	−3.44	−2.30
1qpi	5.09	−2.12	−3.40	−3.50	−2.99	−3.67	−3.07

PDB[a]	$\Gamma(\circ)$[b]	DDNA[c]	FIRE[d]	vFIRE[e]	cFIRE[f]	vcFIRE[g]	RV[h]
1a3q	5.08	−1.46	−2.23	−2.43	−1.87	−2.49	−1.91
1dfm	5.05	−1.23	−1.96	−1.87	−1.52	−2.60	−1.51
1lq1	5.04	−1.94	−2.45	−2.97	−2.31	−3.26	−2.38
1tro	5.02	−1.43	−2.37	−2.67	−2.03	−2.78	−2.05
1fjl	4.95	−1.36	−1.77	−2.22	−1.55	−2.12	−1.58
1h8a_a	4.82	−1.29	−2.03	−2.58	−1.92	−2.35	−2.00
1h8a_b	4.82	−1.02	−2.04	−2.41	−1.58	−2.18	−1.59
1f4k	4.8	−1.16	−2.04	−1.88	−2.02	−2.58	−2.10
6pax	4.73	−1.21	−1.42	−1.08	−1.79	−2.74	−1.96
1hlv	4.53	−1.77	−2.29	−2.34	−2.16	−3.17	−2.23
1mnn	4.46	−1.59	−2.52	−2.63	−2.33	−3.40	−2.49
1dsz	4.38	−1.12	−1.71	−1.55	−1.69	−2.38	−1.82
1hwt	4.13	−1.77	−1.98	−1.09	−2.29	−1.96	−2.40
1per	4.09	−1.44	−2.32	−2.51	−1.99	−2.70	−2.08
1l3l	4.02	−1.76	−2.59	−3.08	−2.37	−3.10	−2.42
3hts	3.87	−0.95	−2.52	−2.41	−1.98	−3.03	−2.05
3bam	3.77	−1.66	−2.26	−1.34	−1.98	−2.86	−1.99
Mean		−1.56	−2.22	−2.14	−2.02	−2.80	−2.06
Standard Deviation		0.38	0.51	0.73	0.38	0.65	0.40

[a]Protein databank identification code

[b]The degree of overall DNA deformation26

[c]DDNA [DFIRE with 19-atomic types acted on interfacial atoms only, 35].

[d]FIRE with residue/base-specific atom types and 10Å cutoff without distance scaling.

[e]Volume-fraction correction added to FIRE(vFIRE).

[f]Low-count correction with Dirichlet pseudo counts added to FIRE (cFIRE).

[g]Volume-fraction correction added to cFIRE (vcFIRE).

[h]Robertson and Varani26 energy function that was trained by the same 167 complexes as FIRE-based energy functions.

Table 4

The lowest rmsd value in top five complexes selected by various energy functions, compared to the lowest possible rmsd value in the decoy sets.

PDB ID[a]	Γ(°)[b]	DDNA[c]	FIRE[d]	vFIRE[e]	cFIRE[f]	vcFIRE[g]	RV[h]	Lowest[i]
1qna	35.70	1.43	1.24	6.38	1.05	0.54	1.05	0.49
1d02	13.62	1.23	0.69	1.44	0.43	0.43	0.43	0.43
1eon	13.41	0.73	0.42	0.42	0.42	0.73	0.42	0.42
1ckq	12.29	0.82	0.61	0.61	0.61	0.41	0.61	0.39
1dmu	9.12	0.25	0.25	25.66	0.25	0.25	0.25	0.25
1qpz	8.53	0.72	1.10	0.20	0.20	0.20	0.20	0.20
1au7	8.48	1.18	0.50	0.50	0.50	0.50	0.50	0.50
1je8	8.15	0.41	0.35	0.41	0.35	0.41	0.35	0.35
2cgp	7.84	1.28	0.45	0.58	0.31	0.45	0.31	0.31
1b3t	7.74	0.90	0.62	25.07	0.88	0.88	0.88	0.51
1tc3	7.30	1.54	12.65	12.71	1.00	1.00	2.38	0.59
1g9z	7.17	1.16	0.41	0.41	0.41	0.49	0.49	0.41
1zme	6.84	0.72	0.96	20.80	0.90	0.90	0.90	0.50
1a73	6.56	0.80	0.45	0.45	0.45	0.45	0.45	0.45
1jko	6.55	0.92	0.66	0.66	0.66	1.06	1.27	0.38
1bdt	6.41	1.01	0.90	4.24	0.46	0.46	0.54	0.46
2bop	6.28	1.10	0.50	0.50	0.58	0.58	0.58	0.50
1a1i	6.21	1.74	0.17	0.17	0.23	0.72	0.17	0.17
1bc8	6.10	1.20	0.33	0.33	0.87	0.33	0.87	0.33
1pdn	6.04	0.80	4.63	4.63	1.01	1.18	1.01	0.54
1skn	5.96	1.23	0.50	0.62	0.50	0.50	0.50	0.50
1mjo	5.94	0.78	0.37	0.37	0.37	0.37	0.37	0.37
1bl0	5.88	1.13	0.47	0.47	0.35	0.35	0.35	0.35
2dgc	5.75	2.01	0.06	0.06	0.06	0.06	0.06	0.06
3pvi	5.71	0.78	0.63	0.58	0.61	0.61	0.61	0.58
2hdd	5.61	1.24	0.80	0.55	0.77	0.77	0.52	0.41
1ign	5.19	1.04	0.66	0.51	0.51	0.51	0.51	0.51
1qpi	5.09	1.13	0.59	0.59	0.59	0.60	0.59	0.59

PDB ID[a]	Γ(°)[b]	DDNA[c]	FIRE[d]	vFIRE[e]	cFIRE[f]	vcFIRE[g]	RV[h]	Lowest[i]
1a3q	5.08	0.87	0.42	0.42	0.42	0.87	0.42	0.42
1dfm	5.05	1.40	0.40	0.40	0.40	0.63	0.40	0.40
1lq1	5.04	0.89	0.42	0.42	0.42	0.42	0.42	0.42
1tro	5.02	1.13	0.55	0.63	0.55	0.55	0.55	0.55
1fjl	4.95	0.51	0.51	0.46	0.50	0.50	0.50	0.46
1h8a_b	4.82	2.71	0.75	0.11	0.80	1.20	0.80	0.53
1h8a_a	4.82	3.51	0.11	0.75	0.11	0.11	0.11	0.11
1f4k	4.80	1.94	0.45	0.45	0.45	0.53	0.45	0.44
6pax	4.73	0.97	2.13	8.00	0.63	0.63	0.63	0.44
1hlv	4.53	0.86	0.34	0.34	0.34	0.34	0.34	0.34
1mnn	4.46	1.14	0.73	0.73	0.83	0.71	0.74	0.56
1dsz	4.38	0.72	0.40	0.40	0.40	0.24	0.40	0.24
1hwt	4.13	0.89	1.64	7.35	1.12	0.62	1.09	0.55
1per	4.09	0.69	0.54	0.54	0.54	0.74	0.54	0.54
1l3l	4.02	1.61	0.67	0.67	0.64	0.64	0.64	0.64
3hts	3.87	1.74	0.69	0.69	0.69	0.69	0.69	0.69
3bam	3.77	1.03	0.47	24.45	0.44	0.89	0.44	0.42
Median		1.04	0.51	0.55	0.50	0.46	0.50	0.44

[a]Protein databank identification code

[b]The degree of overall DNA deformation26.

[c]DDNA [DFIRE with 19-atomic types acted on interfacial atoms only, 35].

[d]FIRE with residue/base-specific atom types and 10Å cutoff without distance scaling.

[e]Volume-fraction correction added to FIRE(vFIRE).

[f]Low-count correction with Dirichlet pseudo counts added to FIRE (cFIRE)

[g]Volume-fraction correction added to cFIRE (vcFIRE).

[h]Robertson and Varani26 energy function that was trained by the same 167 complexes as FIRE-based energy functions.

[i]The lowest rmsd structure in decoys.

Table 5

The ten-fold-cross-validated success rates and their standard deviations for recovering native base pairs of 200 protein-DNA complexes by various energy functions.

	Random	DDNA[a]	FIRE[b]	vFIRE[c]	cFIRE[d]	vcFIRE[e]
Success Rate	25%	30.8±2.5%	40.1±4.9%	41.1±5.0%	40.1±3.3%	40.5±3.8%

[a]DDNA [DFIRE with 19-atomic types acted on interfacial atoms only, 35. There is no change of training sets for different folds here].

[b]FIRE with residue/base-specific atom types and 10Å cutoff without distance scaling.

[c]Volume-fraction correction added to FIRE (vFIRE).

[d]Low-count correction with Dirichlet pseudo counts added to FIRE (cFIRE)

[e]Volume-fraction correction added to cFIRE (vcFIRE).

Table 6

The correlation coefficient between theoretically predicted and experimental measured ΔΔG (189 mutants) values given by various energy functions.

PDB ID	# of Mutants	DDNA[a]	FIRE[b]	vFIRE[c]	cFIRE[d]	vcFIRE[e]
1aay	15	−0.04	−0.08	0.01	0.02	0.21
1ckq	13	0.11	0.21	0.21	0.20	0.17
1ecr	20	0.28	0.02	−0.00	0.17	0.14
1efa	5	−0.63	0.81	0.78	0.64	0.43
1hcq	7	0.39	0.66	0.65	0.66	0.61
1jk1	6	−0.44	−0.32	−0.14	−0.51	−0.47
1lmb	51	0.21	0.32	0.35	0.38	0.44
1run	15	0.19	0.59	0.60	0.63	0.72
1tro	9	−0.44	−0.00	−0.09	0.02	−0.06
6cro	48	0.15	0.39	0.45	0.36	0.44
All	189	0.04	0.05	0.05	0.08	0.09

[a] DDNA [DFIRE with 19-atomic types acted on interfacial atoms only 35].

[b] FIRE with residue/base-specific atom types and 10Å cutoff without distance scaling.

[c] Volume-fraction correction added to FIRE (vFIRE).

[d] Low-count correction with Dirichlet pseudo counts added to FIRE (cFIRE)

[e] Volume-fraction correction added to cFIRE (vcFIRE).

Table 7

Accuracy of PWM prediction based on ψ-test values for 19 complexes by various methods.

PDB ID	Random[a]	Static[b]	Dynamics[b]	DDNA[c]	FIRE[d]	vFIRE[e]	cFIRE[f]	vcFIRE[g]
1aay	0.95	0.19	0.35	0.67	0.35	0.35	0.34	0.36
1yui	0.95	0.26	-	0.63	0.49	0.49	0.53	0.56
1ysa	0.91	0.31	0.39	0.6	0.46	0.46	0.44	0.38
1b8i	0.87	0.34	0.36	0.29	0.40	0.40	0.39	0.40
1fjl	0.83	0.32	0.51	0.47	0.42	0.42	0.36	0.38
2puc	0.81	0.26	0.63	0.45	0.43	0.43	0.42	0.42
1ym	0.73	0.26	0.36	0.2	0.33	0.33	0.28	0.30
1r0o	0.72	0.25	0.38	0.47	0.37	0.37	0.33	0.37
1tro	0.71	0.31	0.39	0.42	0.42	0.42	0.42	0.43
1j1v	0.7	0.22	0.36	0.27	0.32	0.32	0.31	0.29
2drp	0.69	0.24	0.23	0.46	0.47	0.55	0.49	0.48
1mj2	0.69	0.38	0.33	0.26	0.55	0.47	0.55	0.55
1mnn	0.68	0.12	0.20	0.46	0.25	0.25	0.22	0.22
1gxp	0.68	0.28	0.41	0.49	0.48	0.48	0.43	0.44
1gcc	0.57	0.12	-	0.64	0.41	0.41	0.18	0.26
1mse	0.55	0.24	-	0.66	0.21	0.21	0.10	0.09
1run	0.51	0.17	0.38	0.55	0.23	0.24	0.23	0.19
1lmb	0.47	0.09	0.14	0.52	0.09	0.09	0.09	0.08
6cro	0.47	0.10	0.21	0.26	0.10	0.10	0.09	0.10
Means	0.71	0.24	0.35[g]	0.46	0.36	0.36	0.33	0.33(0.34[h])
S[i]	0.15	0.08	0.12[g]	0.14	0.13	0.13	0.14	0.14(0.13[h])

[a]Random-test value.

[b]Static and dynamic models from Ref. 9.

[c]DDNA [DFIRE with 19-atomic types acted on interfacial atoms only, 35].

[d]FIRE with residue/base-specific atom types and 10Å cutoff without distance scaling.

[e]Volume-fraction correction added to FIRE (vFIRE).

[f] Low-count correction with Dirichlet pseudo counts added to FIRE (cFIRE)

[g] Volume-fraction correction added to cFIRE (vcFIRE).

[h] Average over 16 protein-DNA complexes (excluding 1yui, 1gxp and 1gcc).

[i] Standard Derivation of ψ-test values.