BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 15 2010, pages 1857–1863
doi:10.1093/bioinformatics/btq295

Structural bioinformatics

Advance Access publication June 4, 2010

Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function

Huiying Zhao[1,†], Yuedong Yang[1,†] and Yaoqi Zhou[2,*]

[1]School of Informatics, Indiana University Purdue University, Indianapolis, IN 46202 and [2]Center for computational Biology and Bioinformatics, Indiana University School of Medicine, 719 Indiana Ave Ste 319, Walker Plaza Building, Indianapolis, IN 46202, USA

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Template-based prediction of DNA binding proteins requires not only structural similarity between target and template structures but also prediction of binding affinity between the target and DNA to ensure binding. Here, we propose to predict protein–DNA binding affinity by introducing a new volume-fraction correction to a statistical energy function based on a distance-scaled, finite, ideal-gas reference (DFIRE) state.

Results: We showed that this energy function together with the structural alignment program TM-align achieves the Matthews correlation coefficient (MCC) of 0.76 with an accuracy of 98%, a precision of 93% and a sensitivity of 64%, for predicting DNA binding proteins in a benchmark of 179 DNA binding proteins and 3797 non-binding proteins. The MCC value is substantially higher than the best MCC value of 0.69 given by previous methods. Application of this method to 2235 structural genomics targets uncovered 37 as DNA binding proteins, 27 (73%) of which are putatively DNA binding and only 1 protein whose annotated functions do not contain DNA binding, while the remaining proteins have unknown function. The method provides a highly accurate and sensitive technique for structure-based prediction of DNA binding proteins.

Availability: The method is implemented as a part of the Structure-based function-Prediction On-line Tools (SPOT) package available at http://sparks.informatics.iupui.edu/spot

Contact: yqzhou@iupui.edu

Received on March 16, 2010; revised on May 13, 2010; accepted on June 2, 2010

1 INTRODUCTION

DNA binding proteins are proteins that make specific binding to either single or double-stranded DNA. They play an essential role in transcription regulation, replication, packaging, repair and rearrangement. With completion of many genome projects and many more in progress, more and more proteins are discovered with unknown function (Jaroszewski *et al.*, 2009). The structures for some of those function-unknown proteins are solved because of structural genomics projects (Burley, 2000). Functional annotations of these proteins are particularly challenging because the goal of structural genomics is to cover the sequence space of proteins so that homology modeling becomes a reliable tool for structure prediction of any proteins and, thus, many targets in structural genomics have low sequence identity to the proteins with known function. Therefore, it is necessary to develop computational tools that utilize not only sequence but also structural information for function prediction (Ahmad *et al.*, 2004; Gao and Skolnick, 2008; Lee *et al.*, 2007a; Punta and Ofran, 2008; Sadowski and Jones, 2009; Watson *et al.*, 2005).

Many methods have been developed for structure-based prediction of DNA binding proteins. These include function prediction through homology and structural comparisons (Bhardwaj *et al.*, 2005; Ferrer-Costa *et al.*, 2005a, b; Lee *et al.*, 2007b; Pazos and Sternberg, 2004; Shanahan *et al.*, 2004). Others explore sequence and structural features of DNA binding and non-binding proteins with sophisticated machine learning methods such as neural network (Ahmad *et al.*, 2004; Mahony *et al.*, 2006; Stawiski *et al.*, 2003; Tjong and Zhou, 2007), logistic regression (Lee *et al.*, 2006) and support vector machines (Bhardwaj *et al.*, 2005; Cai and Lin, 2003; Kumar *et al.*, 2008; Langlois *et al.*, 2007; Tjong and Zhou, 2007).

Recently, Gao and Skolnick (2008) proposed a new two-step approach, called DBD-Hunter (Gao and Skolnick, 2008), for structure-based prediction of DNA binding proteins. In DBD-Hunter, the structure of a target protein is first structurally aligned to known protein–DNA complexes and the aligned complex structures are used to build the complex structures between DNA and the target protein. The predicted complex structures are, then, employed for judging DNA binding or not by structural similarity scores (TM-Score) and predicted protein–DNA binding affinities. TM-align (Zhang *et al.*, 2005) and a contact-based statistical energy function are employed in the first and second steps of DBD-Hunter, respectively. DBD-Hunter is found to substantially improve over the methods based on sequence comparison only (PSI-BLAST), structural alignment only (TM-align) and a logistic regression technique (Szilagyi and Skolnick, 2006).

In this study, we investigate if one can further improve the prediction of DNA binding proteins by employing a different statistical energy function for predicting binding affinity. Our knowledge-based energy function is distance-dependent and built on a distance-scaled, finite, ideal gas reference (DFIRE) state originally

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

developed for proteins (Yang and Zhou, 2008a, b; Zhou and Zhou, 2002) and extended to protein–DNA interactions (Xu *et al.*, 2009; Zhang *et al.*, 2005). Here, we introduce a new volume-fraction correction for the DFIRE energy function in extracting protein–DNA statistical energy function from protein–DNA complex structures. This volume fraction correction term, unlike previously introduced one (Xu *et al.*, 2009), is atom-type dependent to better account for the fact that protein and DNA atom types are unmixable and occupy in physically separated volumes. In addition to introduction of a new energy function, we further optimize protein–DNA binding affinity by performing DNA mutation. These two techniques lead to a highly accurate and sensitive tool for structure-based prediction of DNA binding proteins.

2 METHODS

2.1 Datasets

We employed the datasets compiled by Gao and Skolnick (2008). One positive and one negative datasets for training are 179 DNA binding proteins (DB179) and 3797 non-DNA binding proteins (NB3797), respectively. These structures were obtained based on 35% sequence identity cutoff, a resolution of 3 Å or better, a minimum length of 40 residues for proteins, 6 bp for DNA and 5 residues interacting with DNA (within 4.5 Å of the DNA molecule). As in Gao and Skolnick (2008), we use significantly larger number of non-DNA binding proteins in order to reduce false positive rate because DNA binding proteins are only a small fraction of all proteins. APO and HOLO testing datasets are made up of 104 DNA binding proteins whose structures are determined in the absence and presence of DNA, respectively. A maximum of 35% sequence identity was also employed in selecting these 104 proteins. For APO/HOLO datasets, 93 APO–DB179 pairs and 92 HOLO–DB179 pairs have sequence identity >35%. These pairs are excluded from target–template pairs during testing. An additional test set of 1697 proteins (the SG1697 set) was compiled from structural genome targets with a sequence identity cutoff at 90% by Gao and Skolnick (2008) from the January 2008 PDB release. We further updated the release on November 2009 and obtained 2235 chains (the SG2235 set). This was done by queried ‘structural genomic’ words in the PDB databank, resulting in 2447 PDB entries. These PDB entries were divided into protein chains and clustered by the CD-HIT (Li and Godzik, 2006). For the clusters that contain a protein chain in SG1679, we chose the protein chain as the representation. For other clusters, we randomly chose one protein chain. There are 538 additional proteins and a total of 2235 protein chains.

To provide an additional test set and examine the effect of a larger database of DNA binding proteins, we have also updated DNA binding proteins from DB179 to DB250. This updated dataset of DNA binding proteins is selected from PDB released on December 2009 based on the same criteria that produced DB179. After removing the chains with high sequence identity (>35%) with any chain contained in DB179 and with each other, we obtained 71 additional protein–DNA complexes. This leads to an additional test dataset DB71 and an expanded training set DB250 (DB179 + DB71).

2.2 Knowledge-based energy function

We employ a knowledge-based energy function to predict the binding affinity of a protein–DNA complex. We have developed a knowledge-based energy function for proteins based on the DFIRE that satisfies the following equation (Zhou and Zhou, 2002):

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

where R is the gas constant, $T=300$ K, $\alpha=1.61$, $N_{\mathrm{obs}}(i,j,r)$ is the number of *ij* pairs within the spherical shell at distance *r* observed in a given structure database, r_{cut} is the cutoff distance, Δr_{cut} is the bin width at r_{cut}. The value of $\alpha(1.61)$ was determined by the best fit of r^{α} to the actual distance-dependent number of ideal-gas points in finite protein-size spheres.

Equation (1) for proteins was initially applied to protein–DNA interactions unmodified with 19 atom types for both proteins and DNA (DDNA; Zhang *et al.*, 2005). In DDNA2 (Xu *et al.*, 2009), a low count correction is made to $N_{\mathrm{obs}}(i,j,r)$

$$N_{\mathrm{obs}}^{\mathrm{lc}}(i,j,r)=N_{\mathrm{obs}}(i,j,r)+\frac{75\sum_{i,j}N_{ij}^{\mathrm{Protein\text{–}DNA}}(r)}{\sum_{i,j,r}N_{ij}^{\mathrm{Protein\text{–}DNA}}(r)}\tag{2}$$

In addition, we employed residue/base-specific atom types with a distance-dependent volume-fraction correction defined as $f^{v}(r)=\frac{\sum_{i,j}N_{ij}^{\mathrm{Protein\text{–}DNA}}(r)}{\sum_{i,j}N_{ij}^{\mathrm{All}}(r)}$. This volume fraction correction was made to take into account the fact that DNA and protein atoms with residue/base-specific atom types do not mix with each other. However, we found that DDNA2 is unable to go beyond existing techniques for predicting DNA binding proteins. To further improve DDNA2, we introduce atom-type-dependent volume fractions: $f_i^{v}(r)=\frac{\sum_{j}N_{ij}^{\mathrm{Protein\text{–}DNA}}(r)}{\sum_{j}N_{ij}^{\mathrm{All}}(r)}$. Our final equation for the statistical energy function is

$$\bar{u}_{i,j}^{\mathrm{DDNA3}}(r)=\begin{cases}-\eta\ln\dfrac{N_{\mathrm{obs}}(i,j,r)}{\left(\frac{f_i^{v}(r)f_j^{v}(r)}{f_i^{v}(r_{\mathrm{cut}})f_j^{v}(r_{\mathrm{cut}})}\right)^{\beta}\frac{r^{\alpha}\Delta r}{r_{\mathrm{cut}}^{\alpha}\Delta r_{\mathrm{cut}}}N_{obs}^{lc}(i,j,r_{\mathrm{cut}})}, & r<r_{\mathrm{cut}},\\ 0, & r\geq r_{\mathrm{cut}},\end{cases}\tag{3}$$

where we have introduced a parameter β. Physically, β should be around 1/2 so that volume fraction is counted once. We will employ it as an adjustable parameter here for the same reason that makes $\alpha < 2$: proteins are finite in size. As in DDNA2, we will use residue/base-specific atom types (167 atom types for proteins and 82 for DNA) and $r_{\mathrm{cut}}=15$ Å, $\Delta r=0.5$ Å. We also set the factor η arbitrarily to 0.01 to control the magnitude of the energy score. For convenience, we shall label the volume-fraction corrected DFIRE as DDNA3.

2.3 Training of the method for predicting DNA binding proteins

DB179 is used to generate the DDNA3 statistical energy function [Equation (3)]. To avoid overfitting, we employed the leave-one-out scheme to train DDNA3 statistical energy function. A target protein is chosen from DB179/NB3797. The TM-align program is employed to make a structural alignment between this target protein with a protein in DB179 (except itself if it is in DB179). If the alignment score (TM-score) is greater than a threshold, the proposed complex structure between the target protein and DNA is obtained by replacing the template protein from its protein–DNA complex structure. The binding affinity between DNA and the target protein is evaluated by the DDNA3 energy function [Equation (3)]. Instead of using template DNA sequences, we perform exhaustive mutations of DNA base pairs to search for the highest binding affinity. DNA bases are paired by X3DNA software package (Lu and Olson, 2003). Unlike mutations in proteins, DNA mutation is relatively easy because the dihedral angles of bases are unchanged. The conformations of mutated bases are built using default bond length, bond angle and dihedral angle parameters as defined in AMBER98 force field (Cheatham *et al.*, 1999). A DNA base, if it does not have a corresponding pairing base, is not mutated. If the highest binding affinity is greater than an optimized threshold, the target protein is considered as a DNA binding protein. The method described above has two important differences from DBD-hunter: the use of our distance-dependent energy function and the search for the strongest binding DNA fragment.

2.4 Evaluation of the method for predicting DNA binding proteins

The measures of the method performance are: sensitivity [SN = TP/(TP + FN)]; specificity [SP = TN/(TN + FP)]; accuracy [AC = (TP + TN)/(TP + FN + TN + FP)]; and precision [PR = TP/(TP + FP)]. In addition, we employed a Matthews correlation coefficient (MCC)

$$\mathrm{MCC}=\frac{\mathrm{TP}*\mathrm{TN}-\mathrm{FP}*\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \qquad (4)$$

Here, TP, TN, FP and FN refer to true positives, true negatives, false positives and false negatives, respectively.

3 RESULTS

3.1 Training based on DB179/NB3797 (DDNA3)

We have optimized volume-fraction exponent β, TM-score and binding affinity thresholds to achieve the highest MCC values. Optimization is performed by a grid-based search. The grids for β and TM-score are 0.02 and 0.01, respectively. For the binding affinity threshold, the lowest energy of each aligned complex under different TM-score thresholds is calculated and these energy values are considered sequentially as the energy threshold. We found that the highest MCC is 0.73 for $\beta = 0.4$, the structural similarity threshold of 0.60 and the energy threshold of −11.6. The corresponding accuracy, precision and sensitivity are 98%, 91% and 60%, respectively. The effect of a knowledge-based energy function can be revealed by replacing DDNA3 with DDNA2. The optimized MCC value (structural similarity threshold of 0.53 and energy threshold of −4.2) is 0.61. (Note, there is no β parameter in DDNA2.) The corresponding accuracy, precision and sensitivity are 97%, 85% and 55%, respectively. It is clear that the reference state of a statistical energy function has a significant impact on the performance in predicting DNA binding proteins. The largest improvement is 6% improvement in precision, the fraction of correct prediction in all prediction. The overall performance of DDNA3 significantly improves over that of DBD-Hunter, which has an MCC of 0.64, 98% accuracy, 84% precision and 55% sensitivity, respectively.

Figure 1 shows sensitivity as a function of false positive rate. Our results were obtained by fixing structural similarity threshold and varying the energy threshold. It is clear that DDNA3 yields a substantially higher sensitivity than either DDNA2 or DBD-Hunter for a given false positive rate.

The predicted binding complexes can be employed to examine predicted DNA binding residues. An amino acid residue is considered as a DNA binding residue, if any heavy atom of that residue is <4.5 Å away from any heavy atom of a DNA base. Predicted binding residues from template-based modeling can be compared to actual binding residues. For the training set (179 DB and 3797 NB proteins), there are 108 predicted DB proteins with 11 false positives. For these 108 predicted complexes, specificity, accuracy, precision, sensitivity and MCC of predicting DNA binding residues are 94%, 89%, 74%, 68% and 0.64, respectively. For a comparison, DDNA2 has predicted 99 DB proteins and the corresponding performance in predicting DNA binding residues are 93%, 88%, 75%, 67% and 0.63, respectively. These performances are similar to a specificity of 93%, an accuracy of 90%, a precision of 71% and a sensitivity of 72% achieved by DBD-hunter. Similar performance in predicting DNA binding residues is due to the same structural alignment (TM-align) method used in the first step by three methods. The slight difference in binding residue prediction is caused by two reasons: the change in the number of predicted DNA binding proteins and possibly different templates recognized by different methods.

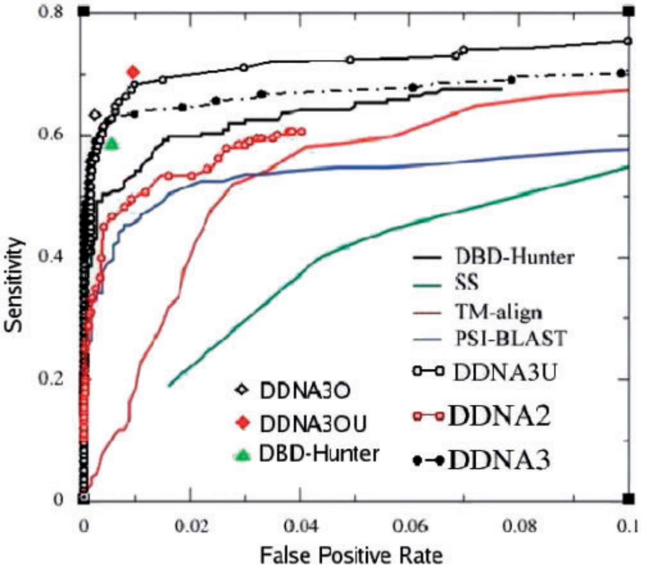

Fig. 1. Sensitivity versus false positive rate, given by DDNA3 (filled black circles) and DDNA2 (open red circles) reveals the importance of an appropriate reference state for method performance in predicting DNA binding proteins. The results of other methods are adapted from Gao and Skolnick (2008). DDNA3U (open black circles) is the sensitivity versus false positive rate given by DDNA3 based on updated DB250 dataset. TM-score-dependent energy-score thresholds lead to DDNA3O (open diamond) and DDNA3OU (red filled diamond), compared to optimized DBD-Hunter (open green triangle).

3.2 TM-score-dependent energy threshold (DDNA3O)

Obviously, one threshold for energy and one for structural similarity (TM-score) are too simple to capture the complex relation between structure and binding affinity. For example, one expects that the binding-energy requirement should be stronger for less similar structures but weaker for highly similar structures between template and query. This has led Gao and Skolnick (2008) to develop TM-score-dependent energy thresholds (nine energy thresholds for nine TM-score bins ranging from 0.40 to 1.0 to maximize MCC value in each bin), and they finally set a minimum TM-score cutoff at 0.55 for achieving the maximum MCC value. If we followed their method, we found the same minimum TM-score cutoff at 0.55 for the maximum MCC value of 0.76. We revise this method slightly here. Instead of optimizing the MCC for each TM-score bin, we search for the energy threshold for a given TM-score bin by optimizing the MCC value for the TM-score bin plus all other bins with higher TM-scores. The results are shown in Table 1. This revised method leads to the same maximum MCC value of 0.76 but with a minimum TM-score cutoff at 0.52, slightly increased sensitivity (two additional true positives) without increase of false positives. To distinguish this further optimized method, we labeled it as DDNA3O. DDNA3O yields a sensitivity of 64% and specificity of 99.8%. By comparison, the corresponding optimized

Table 1. Optimized TM-score-dependent energy thresholds based on DB179 and NB3797 (DDNA3O)

TM-score range	Energy threshold	ΔTP[a]	TP[b]	ΔFP[c]	FP[d]	Max MCC
0.74–1.00	−9.87	53	53	3	3	0.52
0.62–0.74	−13.95	52	105	4	7	0.73
0.58–0.62	−16.50	3	108	1	8	0.74
0.55–0.58	−18.64	4	112	0	8	0.76
0.52–0.55	−29.10	2	114	0	8	0.76

[a] Number of true positives predicted in each TM-Score bin.

[b] Number of true positives predicted in each TM-Score bin plus bins with higher TM scores.

[c] Number of false positives predicted in each TM-Score bin.

[d] Number of false positives predicted in each TM-Score bin plus bins with higher TM scores.

DBD-Hunter with the same dataset has a MCC value of 0.69 with the corresponding sensitivity of 58% and specificity of 99.5%, while the DDNA3 has a MCC value of 0.73 with sensitivity of 60% and specificity of 99.7%. Thus, the most significant improvement from DDNA3 to DDNA3O is significant increase in sensitivity (from 60% to 64%) also with slight reduction in rate of false positives (from 11/3797 to 8/3797).

There are 114 complexes predicted as DNA binding proteins by DDNA3O. For these 114 complexes, predicted DNA binding residues are compared to native complexes. The specificity, accuracy, precision, sensitivity and MCC are 95%, 90%, 77%, 69% and 0.67, respectively.

3.3 Test on the APO104/HOLO104 datasets

The methods trained above (DDNA3 and DDNA3O) are applied to predict DNA binding proteins of APO104/HOLO104 datasets. The numbers of positive prediction are 50 by DDNA3 and 53 by DDNA3O (out of 104) for the APO sets, and 61 by DDNA3 and 62 by DDNA3O (out of 104) for the HOLO sets, respectively. That is, using monomer structures, rather than the complex structures, leads to a reduction of 11% in sensitivity (from 59% for the HOLO to 48% for the APO set) by DDNA3 and 9% by DDNA3O (from 60% to 51%). The corresponding sensitivity values for DDNA2 are 43.3% (45/104) and 53.8% (56/104) for the APO and HOLO sets, respectively. The performance of DBD-Hunter (47% for the APO and 55% for the HOLO sets) is somewhat in between DDNA2 and DDNA3. The test confirms an increase in sensitivity by DDNA3O over by DDNA3 for the APO set, in particular.

A more detailed analysis on predictions made by DDNA3O shows that there is an overlap of 50 predictions between the APO and HOLO sets. Figure 2 shows one example of the test on target proteins 1mjkA (contained in APO104) and 1mjmA (contained in HOLO104). 1mjkA and 1mjmA are the structures of the same methionine repressor protein in the absence and presence of DNA fragment, respectively. There is a small conformational change before and after DNA binding (TM-score between the two is 0.93). This small conformational change apparently does not prohibit the successful match to the same template protein 1ea4A with strong binding affinity.

On the other hand, there are 12 correctly predicted HOLO targets but incorrectly predicted APO targets as shown in Table 2. The difference is caused by significant local conformational change in binding regions (high TM-align score but low binding affinity). An example (1le8A in HOLO and corresponding 1f43A in APO) is shown in Figure 3a, where significant change in binding regions (from red in APO to green in HOLO) leads to incorrect prediction despite insignificant structural change in non-binding regions of the protein. In another more extreme case (Fig. 3b), disordered region in APO structure (1jyfA) changes to ordered binding domain in HOLO structure (1efaA).

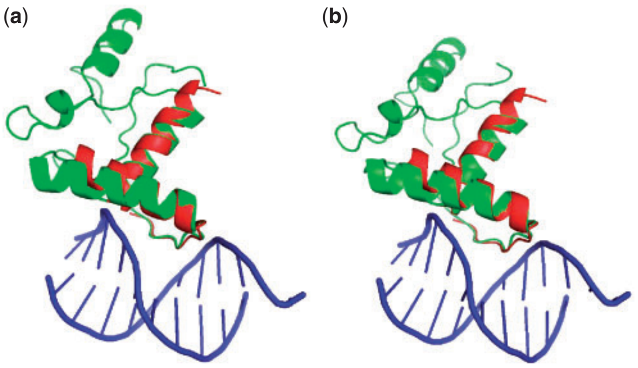

Fig. 2. (**a**) Structural comparison between APO target protein 1mjkA (green) and template protein 1ea4A (red). The TM-score between them is 0.79 and the interaction energy between 1mjkA and template DNA is −20.9. (**b**) Structural comparison between HOLO target protein 1mjmA (green) and template protein (1ea4A). The TM-score between them is 0.76 and the interaction energy between 1mjmA and template DNA is −20.6.

Another cause of incorrect prediction in APO and correct prediction in HOLO is large overall structural change. The large overall structural changes lead to poor structural alignment to templates so that their TM-scores are lower than the threshold. For example, despite 90% sequence identity, TM-score between 1q39A in APO and 1k3w in HOLO structures is only 0.55 and leads to the poor alignment of APO structure to template (best is 0.48 in TM-score). We also discovered a technical reason for an APO target (1rxr_). We are unable to use the template employed for the corresponding HOLO target because the sequence identity between the template and its respective APO target is slightly higher than 35%.

There are also three targets identified as DNA binding proteins correctly in the APO set but not in the HOLO set. All three (1llzA, 1bf5A and 1esgA) are just outside of arbitrary boundaries generated by optimization. This highlights the empirical nature of the proposed approach.

3.4 Test on the DB71 dataset

The additional 71 proteins contained in the updated protein/DNA complex structural dataset (DB71) offer a challenging test set. DDNA3 (DDNA3O) predicts 34 (39) out of 71 proteins as DNA binding proteins. Thus, the sensitivity is 34/71 (48%) by DDNA3 and 55% by DDNA3O. DDNA3O continues to make significant improvement in sensitivity over DDNA3. This 55% sensitivity is 5% lower than the sensitivity of 60% for the HOLO dataset but is higher than the sensitivity of 51% for the APO dataset. This suggests that >50% new complex structures are recognizable by DDNA3O

Table 2. Targets are predicted as DNA binding on HOLO set but not on APO set

APO[a]	HOLO[b]	TMP[c]	Seqid[d]	HOLO-TMP[e]	HOLOEN[f]	APOEN[f]	AP_TMP[g]	HOLO-APO[h]
1nfa_	1a02N	1hjbC	82	0.67	−25.70	−1.1	0.53	0.64
1uklC	1am9A	1nlwB	70	0.82	−24.99	−6.5	0.84	0.86
1rxr_[i]	1by4A	1kb4A	83	0.90	−29.57	−20.5	0.81	0.80
1es8A	1dfmA	2bamB	88	0.68	−30.68	14.1	0.64	0.89
1jyfA	1efaA	1rzrA	100	0.90	−12.97	−1.6	0.89	0.96
1i11A	1gt0D	1cktA	52	0.78	−26.68	−9.5	0.73	0.74
1ev7A	1iawA	1cf7A	97	0.55	−23.51	−20.0	0.53	0.82
1q39A	1k3wA	2f5pA	90	0.82	−20.67	−18.4	0.48	0.55
1f43A	1le8A	1fjlA	100	0.88	−19.47	−7.5	0.58	0.64
1bgt_	1sxpA	1y6fA	93	0.75	−19.17	−2.0	0.78	0.98
1mi7R	1trrA	1gdtA	89	0.68	−21.58	−15.0	0.38	0.52
2audA	1tx3A	4rveB	96	0.56	−24.53	−20.2	0.54	0.95

[a] Targets from APO set.
[b] Targets from HOLO set.
[c] Template.
[d] Sequence identity between APO and HOLO target calculated by bl2seq in blast2.2.
[e] TM-score between HOLO target and template protein.
[f] Energy value between template–target complex.
[g] TM-score between APO target and template protein.
[h] TM-score between HOLO target and APO target.
[i] Template used for HOLO is unable to be used for APO because of >35% sequence ID.

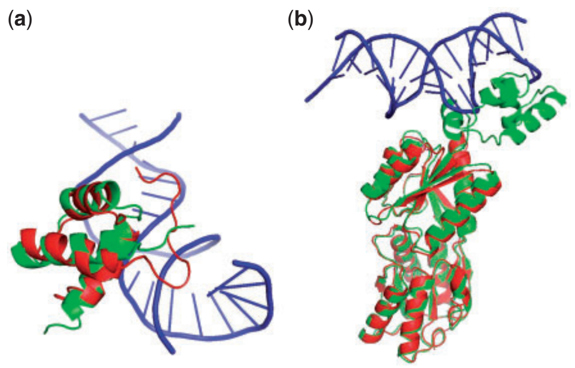

Fig. 3. (**a**) Structural comparison between APO target 1f43A and HOLO target 1le8A. Red: fragment of binding domain of 1f43A. Green: fragment of binding domain of 1le8A. Orange: template DNA of 2bamB. (**b**) Structural comparison between APO target 1jyfA (red) and HOLO target 1efaA (green). Orange: template DNA of 1rzrA.

with DB179 as templates for protein–DNA complexes for all the sets tested (APO, HOLO and DB71).

3.5 The effect of a larger, updated dataset of DNA binding proteins (DDNA3U)

To examine the effect of a larger dataset of DNA binding proteins, we use DB250 and NB3797 as the training set. We found that for this larger, updated dataset, the highest MCC is 0.75 with the same or similar values for three parameters ($\beta = 0.4$, TM-score threshold of 0.55 and energy threshold of −13.7) as DDNA3. This result highlights the stability of trained parameters with a 40% increase in DNA binding proteins. The corresponding accuracy, precision and sensitivity are 97%, 87% and 67%, respectively. In particular, 45 out of 71 additional proteins outside DB179 are recognized as DNA binding by DB250-trained DDNA3 (DDNA3U), compared to 34 by DB179-trained DDNA3. However, the sensitivity increases at the cost of more false positives (26, more than doubled from 11 for DB179-trained DDNA3).

Application of this newly trained method to APO104 and HOLO104 sets leads to 52 (50%) and 64 (62%) predicted DNA binding proteins, respectively. That is, a 40% expansion of DNA binding proteins (from 179 to 250) leads to about 3% improvement in sensitivity. However, as Figure 1 indicates, newly trained DDNA3 (labeled as DDNA3U) yields higher sensitivity only when false positive rate >0.005. That is, at a lower false positive rate, a larger template database in fact decreases sensitivity and precision.

One can employ TM-score-dependent energy thresholds to the updated DB250/NB3797 databases. The resulting DDNA3UO further increases the number of true positives from 167 to 176 but the number of false positives also increases from 26 to 34. Since we are interested in predicting DNA binding proteins with very low false positive rate (<0.005), we will employ the methods (DDNA3 and DDNA3O) trained by DB179 to structural genomics targets.

To further examine the possibility of overfitting in DDNA3U, we perform a 10-fold cross-validation tests on the DB250/NB3797. That is, all the binding and non-binding sets are randomly divided into 10 folds. Each time, one fold is chosen as the test set while the other nine folds are employed for all training including the statistics of potential energy function, the structure templates for protein–DNA binding, and re-training of the threshold parameters. The test is repeated for 10 times. The method performance is analyzed by 1000 times of bootstrap resampling (Angarica *et al.*, 2008). We found that the average MCC value is 0.70 ± 0.02 with the accuracy of 97%, the precision of 88% and the sensitivity of 58%, respectively. It is clear that the only significant change from the leave-one-out results is the reduction of sensitivity from 67% to 58%. This is likely caused by the reduced number of templates in the 10-fold cross-validation. Indeed, if 249 templates are permitted to use, the average MCC

Table 3. Structural Genomics targets (SG1697) predicated as DNA binding proteins by DBD-Hunter, DDNA3 and DDNA3O

Method	Prediction	Putative	Other function	Unknown
DDNA3	32	19	3	10
DDNA3O	27	19	1	7
DBD-Hunter	37	18	3	16
Overlap[a]	19	15	0	4

[a] Overlap between DBD-Hunter and DDNA3O.

value is 0.72 ± 0.02. Thus, our results are reasonably robust with different training.

3.6 Application to structural genomics targets

As shown in Table 3, application of DDNA3 leads to 32 DNA binding proteins from SG1697. Among them, 19 out of 32 proteins (59%) are putative DNA binding proteins, 3 out of 32 proteins (10%) are annotated to having other functions, while others (31%) have unknown function. DDNA3O decreases the prediction of DNA binding proteins from 32 to 27 without change on the number of putative DNA binding proteins (19) but reduces the number of proteins with other annotated function from 3 to 1 and with unknown functions from 10 to 7. This result further confirms the improvement of DDNA3O over DDNA3. By comparison, DBD-Hunter predicts 37 DNA binding proteins. Among the 37 proteins, there are 18 (48.6%) putative DNA binding proteins, 3 (8.1%) with other putative functions and 16 (43.2%) with unknown function. All the putative functions are from the annotations in the NCBI database.

The overlap between predicted proteins by DDNA3O and DBD-Hunter is only 19 proteins, 15 (79%) of which are putative DNA binding proteins. The large fraction of putative DNA binding proteins in overlapped predictions highlights significant improvement in confidence of prediction when a consensus prediction is made. Meanwhile, only 70% (19/27) proteins predicted by DDNA3O overlap with those by DBD-Hunter highlights that the energy function plays a significant role in prediction. There are four putative DNA binding proteins (1ug2A, 1y9bA, 2cqxA and 2fb1A) predicted by DDNA3O but missed by DBD-Hunter. Similarly, there are three putative DNA binding proteins (2hytA, 2iaiA and 2od5A) predicted by DBD-Hunter but missed by DDNA3O. The complete list of predicted DNA binding proteins is shown in Table 4. Table 4 includes 10 additional predicted proteins from SG2235, 8 of which are putative DNA binding proteins. That is, 73% (27/37) of predicted proteins from SG2235 are putative DNA binding proteins. This result confirms the prediction quality of the proposed DDNA3O technique.

Table 4. Targets are predicted as DNA binding proteins by DDNA3O from SG1697 and SG2235 with function annotated in NCBI database

Target	Template	TM-score	Energy	Putative function
2keyA[a]	1p7dB	0.58	−22.19	DB
2khvA[a]	1p7dB	0.72	−30.06	DB
2kobA[a]	1p7dB	0.75	−26.52	DB[b]
3cecA[a]	3croL	0.75	−21.67	DB
3edpA[a]	1sfuA	0.74	−13.42	DB
3frwF[a]	1trrG	0.77	−23.04	DB
3ic7A[a]	1cf7A	0.61	−17.48	DB
3ikbA[a]	4sknE	0.62	−16.54	DB
3iuvA[a]	1jt0A	0.77	−14.97	UK[c]
3ke2A[a]	1gdtA	0.58	−18.58	UK
1iuyA	1f4kB	0.61	−19.25	NB[d]
1s7oA	1gdtA	0.67	−14.37	DB
1sfxA	1u8rJ	0.72	−24.89	DB
1ug2A	1fjlA	0.58	−17.92	DB
1wi9A	1repC	0.62	−17.50	UK
1x58A	1w0tA	0.87	−24.86	DB
1y9bA	1ea4A	0.67	−22.76	DB
1z7uA	1u8rJ	0.66	−14.75	DB
1zelA	1cgpA	0.56	−20.67	UK
2cqxA	1akhA	0.69	−17.87	DB
2da4A	1akhA	0.74	−27.67	DB
2e1oA	1akhA	0.87	−18.37	DB
2eshA	1f4kB	0.67	−17.10	DB
2esnA	1u8rJ	0.62	−21.74	DB
2ethA	1u8rJ	0.71	−20.94	DB
2f2eA	1u8rJ	0.71	−14.07	DB
2fb1A	2as5F	0.62	−14.47	DB
2fyxA	2a6oB	0.78	−18.83	DB
2g7uA	1u8rJ	0.70	−15.83	DB
2jn6A	1gdtA	0.70	−17.11	DB
2jtvA	2ex5A	0.61	−21.07	UK
2nx4A	1jt0A	0.76	−16.34	DB
2qvoA	1z9cF	0.80	−10.19	UK
3b73A	1z9cF	0.68	−23.89	UK
3bddA	1u8rJ	0.76	−21.56	DB
3bhwA	1fokA	0.58	−19.04	UK
3bz6A	1u8rJ	0.73	−17.02	UK

[a] Targets in SG2235.

[b] Targets are annotated as protein which has putative functions related with DNA binding in PDB.

[c] It is unknown whether a target has putative functions related with DNA binding.

[d] Non-binding to DNA according to GO annotation.

4 DISCUSSION

We have developed a highly accurate method (DDNA3O) to predict DNA binding proteins. This is accomplished by developing a new statistical energy function for predicting DNA binding proteins. We found that introducing an atom-type-dependent volume fraction correction and DNA mutation in the DFIRE statistical energy function leads to a significant improvement in the performance in predicting DNA binding proteins (MCC = 0.76 for DB179/NB3797 by DDNA3O). This is a significant improvement from MCC of 0.69 given by optimized DBD-Hunter. Application of DDNA3O to structural genome targets confirms the accuracy of the proposed method with 73% potentially correct prediction of DNA binding proteins (annotated as putative DNA binding), 3% potentially false positives (function annotated but not DNA binding) and the rest unknown.

For DDNA3, the effect of DNA mutation is small for improving the MCC value of the training set (from 0.72 to 0.73) but is significant for improving the sensitivity from 46/104 (44%) to 50/104 (48%) of the APO test set. We further find that the mutation leads to no significant improvement in sequence identity between template DNA sequence and wild-type DNA sequence. The sequence identities to wild-type DNA sequences before and after mutation are both close to the random value of 25%. One possible reason is the absence of structural refinement for protein during mutation. This result also suggests that DDNA3 is not yet specific enough to identify binding DNA bases.

In principle, exhaustive mutations of DNA base pairs can lead to significant increase in computing time for a long DNA segment. However, because our energy function does not consider base–base interaction by assuming a rigid DNA structure before and after binding, the computing requirement for the exhaustive mutations of DNA base pairs is only four times more than that without base mutations.

One potential concern is insufficient statistics due to the small number of complex structures for deriving the DDNA3 energy function. We have addressed this question by employing the leave-one-out (for both DB179 and DB250 sets) and 10-fold cross-validation (for the DB250 set) techniques. The consistency between different training and test sets provides the confidence about the energy functions obtained.

Another concern is potential overfitting due to five threshold parameters in DDNA3O because of the small number of true positives for each TM-score bins (Table 1). This concern is reduced somewhat as the energy threshold mostly satisfies the expectation that less similar structures (low TM-scores) requires stricter energy thresholds. Moreover, there is a consistent improvement in sensitivity from training (DB179) to test (APO/HOLO104, DB71 and structural genomics targets). This consistency makes the improvement statistically significant. However, one certainly cannot completely rule out overfitting. More studies as larger dataset becomes available are certainly needed.

One advantage of the proposed structure-based prediction method is the prediction of protein–DNA complex structures. The predicted complex structures allow prediction of DNA binding residues. High specificity and accuracy (>90%) are achieved for binding residue prediction even for the APO structures (protein structures in the absence of DNA).

The success of DDNA3O is limited by the availability of protein–DNA complexes as templates. A 40% expansion of template databases from 179 to 250 proteins leads to significant improvement in sensitivity if false positive rate >0.005 (Fig. 1) but also slightly decreases sensitivity if false positive rate <0.005. Thus, there is a clear need to further improve the energy function that discriminates binding from non-binding proteins. The rigid-body approximation employed here likely has limited the performance of DDNA3O. Introducing flexibility to DNA and proteins to DDNA3 is in progress.

Funding: National Institutes of Health (grant R01 GM 085003).

Conflict of Interest: none declared.

REFERENCES

Ahmad,S. *et al.* (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.

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

Bhardwaj,N. *et al.* (2005) Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Res.*, **33**, 6486–6493.

Burley,S.K. (2000) An overview of structural genomics. *Nat. Struct. Biol.,* **7**, 932–934.

Cai,Y.-d. and Lin,S.L. (2003) Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim. Biophys. Acta*, **1648**, 127–133.

Cheatham,T.E. *et al.* (1999) A modified version of the cornell *et al.* force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.*, **16**, 845–862.

Ferrer-Costa,C. *et al.* (2005a) PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics*, **21**, 3176–3178.

Ferrer-Costa,C. *et al.* (2005b) HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif. *Bioinformatics*, **21**, 3679–3680.

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.

Jaroszewski,L. *et al.* (2009) Exploration of uncharted regions of the protein universe. *PLoS Biol.*, **7**, e1000205.

Kumar,M. *et al.* (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins*, **71**, 189–194.

Langlois,R.E. *et al.* (2007) Learning to translate sequence and structure to function: identifying DNA binding and membrane binding proteins. *Ann. Biomed. Eng.*, **35**, 1043–1052.

Lee,D. *et al.* (2007a) Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.*, **8**, 995–1005.

Lee,D. *et al.* (2007b) Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.*, **8**, 995–1005.

Lee,H. *et al.* (2006) Diffusion kernel-based logistic regression models for protein function prediction. *Omics*, **10**, 40–55.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

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

Mahony,S. *et al.* (2006) Self-organizing neural networks to support the discovery of DNA-binding motifs. *Neural Netw.*, **19**, 950–962.

Pazos,F. and Sternberg,M.J.E. (2004) Automated prediction of protein function and detection of functional sites from structure. *Proc. Natl Acad. Sci. USA*, **101**, 14754–14759.

Punta,M. and Ofran,Y. (2008) The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS Comput. Biol.*, **4**, e1000160.

Sadowski,M.I. and Jones,D.T. (2009) The sequence-structure relationship and protein function prediction. *Curr. Opin. Struct. Biol.*, **19**, 357–362.

Shanahan,H.P. *et al.* (2004) Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Res.*, **32**, 4732–4741.

Stawiski,E.W. *et al.* (2003) Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol.*, **326**, 1065–1079.

Szilagyi,A. and Skolnick,J. (2006) Efficient prediction of nucleic acid binding function from low-resolution protein structures. *J. Mol. Biol.,* **358**, 922–933.

Tjong,H. and Zhou,H.-X. (2007) DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res.*, **35**, 1465–1477.

Watson,J.D. *et al.* (2005) Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.*, **15**, 275–284.

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.

Yang,Y. and Zhou,Y. (2008a) Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins*, **72**, 793–803.

Yang,Y. and Zhou,Y. (2008b) Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Sci.*, **17**, 1212–1219.

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.

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.