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 83% 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: yzhou@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 (Szalayi 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
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, 60 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-DNAB179 pairs and 92 HOLO-DNB179 pairs have sequence identity >35%. These pairs are excluded from target-template pairs testing. An additional test set of 1697 pairs (the SG2167 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 database, 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 536 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):

$$u_{DFIRE}^{i,j}(r) = \begin{cases} -RT \ln \left( \frac{N_{obs}(i,j,r)}{N_{pairs}(i,j,r)} \right), & r < r_{cut}, \\ 0, & r \geq r_{cut}, \end{cases}$$

where $R$ is the gas constant, $T = 300$ K, $a = 1.61$. $N_{obs}(i,j,r)$ is the number of $ij$ pairs within the spherical shell at distance $r$ observed in a given structure database, $r_{cut}$ is the cutoff distance, $\Delta r_{cut}$ is the bin width at $r_{cut}$. The value of $a(1.61)$ was determined by the best fit of $r^3$ 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 $N_{obs}(i,j,r)$.

$$N_{obs}(i,j,r) = N_{obs}(i,j,r) + \beta$$

We further updated the release on November 2009 and obtained 2235 chains. This updated dataset of DNA binding proteins is divided into protein chains and clustered by the CD-HIT (Li and Godzik, 2006). With any chain contained in DB179 and with each other, we obtained 35% sequence identity cutoff, a resolution of 3 Å or better, a minimum length of 40 residues for proteins, 60 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-DNAB179 pairs and 92 HOLO-DNB179 pairs have sequence identity >35%. These pairs are excluded from target-template pairs testing. An additional test set of 1697 pairs (the SG2167 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 database, 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 536 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).

2.3 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 XS/DNA 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 (Chuahm 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)

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}
\]

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 \(\beta\), TM-score and binding affinity thresholds to achieve the highest MCC values. Optimization is performed by a grid-based search. The grids for \(\beta\) 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 \(\beta\) 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 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.

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 for 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).

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).
This small conformational change apparently does not prohibit the
binding of DNA to the DNA-binding domain because the sensitivity of 59% for the HOLO dataset decreases to 48% for the APO dataset. There are also three targets identified as DNA binding proteins correctly in the APO set but not in the HOLO set. All three (1llzA, 1hs5A 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 1. Optimized TM-score-dependent energy thresholds based on DB71 and NB3797 (DDNA3O)

<table>
<thead>
<tr>
<th>TM-score range</th>
<th>Energy threshold</th>
<th>ΔTP</th>
<th>TP</th>
<th>ΔFP</th>
<th>FP</th>
<th>Max MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.74–1.00</td>
<td>−9.87</td>
<td>53</td>
<td>3</td>
<td>3</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>0.62–0.74</td>
<td>−13.95</td>
<td>52</td>
<td>4</td>
<td>7</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>0.58–0.62</td>
<td>−16.50</td>
<td>3</td>
<td>8</td>
<td>0.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.55–0.58</td>
<td>−18.64</td>
<td>4</td>
<td>8</td>
<td>0.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.52–0.55</td>
<td>−29.10</td>
<td>2</td>
<td>8</td>
<td>0.76</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3Number of false positives predicted in each TM-Score bin.
4Number of false positives predicted in each TM-Score bin.
5Number of false positives predicted in each TM-Score bin.
6Number of false positives predicted in each TM-Score bin.

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 11.3/3797 to 8.3/3797).

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 60% to 51%). The performance of DBD-Hunter (47% for the APO and 55% for the HOLO sets) is somewhat in between DDNA2 and DDNA3O.

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 1h43A 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). 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 1jg9A 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 (1mx..). 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 (1lsaA, 1hs5A and 1esgA) are just outside of arbitrary boundaries generated by optimization. This highlights the empirical nature of the proposed approach.

### Figure 2. (a) Structural comparison between APO target protein 1mjkA (green) and targeted template 1ea4A (red). The TM-score between them is 0.79 and the interaction 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.

The small conformational change apparently does not prohibit the binding of DNA to the DNA-binding domain because the sensitivity of 59% for the HOLO dataset decreases to 48% for the APO dataset. There are also three targets identified as DNA binding proteins correctly in the APO set but not in the HOLO set. All three (1lsaA, 1hs5A and 1esgA) are just outside of arbitrary boundaries generated by optimization. This highlights the empirical nature of the proposed approach.

![Figure 2](image-url)
Table 2. Targets are predicted as DNA binding on HOLO set but not on APO set

<table>
<thead>
<tr>
<th>APO</th>
<th>HOLO</th>
<th>TMP</th>
<th>Seqid</th>
<th>HOLO-TMP</th>
<th>HOLOEN</th>
<th>APOEN</th>
<th>AP_TMP</th>
<th>HOLO-APO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1faA</td>
<td>1f02N</td>
<td>1jbgC</td>
<td>82</td>
<td>0.67</td>
<td>−25.70</td>
<td>−1.1</td>
<td>0.53</td>
<td>0.64</td>
</tr>
<tr>
<td>1i6C</td>
<td>1am9A</td>
<td>1iwbB</td>
<td>70</td>
<td>0.82</td>
<td>−24.99</td>
<td>−6.5</td>
<td>0.84</td>
<td>0.86</td>
</tr>
<tr>
<td>1r7x</td>
<td>1by4A</td>
<td>1kb4A</td>
<td>83</td>
<td>0.90</td>
<td>−29.57</td>
<td>−20.5</td>
<td>0.81</td>
<td>0.80</td>
</tr>
<tr>
<td>1e6A</td>
<td>1d1mA</td>
<td>2bamB</td>
<td>88</td>
<td>0.66</td>
<td>−30.68</td>
<td>14.1</td>
<td>0.64</td>
<td>0.89</td>
</tr>
<tr>
<td>1jyA</td>
<td>1f66A</td>
<td>1twA</td>
<td>100</td>
<td>0.90</td>
<td>−12.97</td>
<td>−1.6</td>
<td>0.89</td>
<td>0.96</td>
</tr>
<tr>
<td>1i1A</td>
<td>1g60D</td>
<td>1k4A</td>
<td>52</td>
<td>0.78</td>
<td>−26.68</td>
<td>−9.5</td>
<td>0.73</td>
<td>0.74</td>
</tr>
<tr>
<td>1ev7A</td>
<td>1iawA</td>
<td>1cf7A</td>
<td>97</td>
<td>0.55</td>
<td>−23.51</td>
<td>−20.0</td>
<td>0.53</td>
<td>0.82</td>
</tr>
<tr>
<td>1q39A</td>
<td>1k3wA</td>
<td>2f5pA</td>
<td>90</td>
<td>0.82</td>
<td>−20.67</td>
<td>−18.4</td>
<td>0.48</td>
<td>0.55</td>
</tr>
<tr>
<td>1f43A</td>
<td>1le8A</td>
<td>1fjlA</td>
<td>100</td>
<td>0.88</td>
<td>−19.47</td>
<td>−7.5</td>
<td>0.58</td>
<td>0.64</td>
</tr>
<tr>
<td>1bgt</td>
<td>1xspA</td>
<td>1y6A</td>
<td>93</td>
<td>0.75</td>
<td>−19.17</td>
<td>−2.0</td>
<td>0.78</td>
<td>0.98</td>
</tr>
<tr>
<td>1m7R</td>
<td>1tmA</td>
<td>1gkA</td>
<td>89</td>
<td>0.68</td>
<td>−21.58</td>
<td>−15.0</td>
<td>0.38</td>
<td>0.52</td>
</tr>
<tr>
<td>2anl</td>
<td>1tx3A</td>
<td>4rveB</td>
<td>96</td>
<td>0.56</td>
<td>−24.53</td>
<td>−20.2</td>
<td>0.54</td>
<td>0.95</td>
</tr>
</tbody>
</table>

* Targets from APO set.
* Targets from HOLO set.
* Template.
* Sequence identity between APO and HOLO target calculated by blast2.2.
* TM-score between HOLO target and template protein.
* Energy value between template–target complex.
* TM-score between APO target and template protein.
* Template used for HOLO is unable to be used for APO because of TM-score between HOLO target and APO target.

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 $0.70 \pm 0.02$ 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 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.

One can employ TM-score-dependent energy thresholds to the updated DB250/NB3797 databases. The resulting DDNA3UO 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.

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. This is, at a lower false positive rate, a larger template database in fact decreases sensitivity and precision.

Indeed, if 249 templates are permitted to use, the average MCC 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.

Fig. 3. (a) Structural comparison between APO target 1f43A and HOLO target 1e6A. Red: fragment of binding domain of 1e6A. Orange: template DNA of 2bamB. (b) Structural comparison between APO target 1jyA (red) and HOLO target 1f66A (green). Orange: template DNA of 1rzrA.
As shown in Table 3, application of DDNA3 leads to 32 DNA structural genomics targets (SG1697) predicated as DNA binding proteins. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database.

The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database.

The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database.

The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database.

The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database. The overlap between predicted proteins by DDNA3O and 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 functions. All the putative functions are from the annotations in the NCBI database.
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/Holol) and structural databases from 179 to 250 proteins leads to significant improvement prediction even for the APO structures (protein structures in the complex structures allow prediction of DNA binding residues. High sensitivity from training (DB179) to test (apo/Holol, DB1 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 GM085003).

Conflict of Interest: none declared.

REFERENCES


