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

Huiying Zhao¹,†, Yuedong Yang¹,† and Yaoqi Zhou²,∗

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

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 prediction On-line Tools (SPOT) package available at http://sparksinformatics.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 DNA binding function by introducing a new volume-fraction correction to a statistical energy function based 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, 60bp 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 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 SG179, 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 (>5%) 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):

\[ f_{|ij|}^{\text{DFIRE}}(r) = \begin{cases} -RT \ln \frac{N_{\text{DB}}(i,j,r)}{N_{\text{DB}}(i,j,r_{\text{cut}})} , & r < r_{\text{cut}} , \\ 0 , & r \geq r_{\text{cut}} . \end{cases} \]  

where \( R \) is the gas constant, \( T = 300 \text{ K} \), \( \sigma = 1.61 \), \( N_{\text{DB}}(i,j,r) \) is the number of \( ij \) pairs within the spherical shell at distance \( r \) observed in a given structure.
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)

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

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 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-modeling based 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 DDNA2.

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 structures but weaker for highly similar structures between template and query. This has led Gao and Skolnick (2008) to develop TM-score-dependent energy-score thresholds lead to DDNA3O (open black circles) compared to optimized DBD-Hunter (open green triangle).

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.

In addition, we employed a Matthews correlation coefficient (MCC) which is defined as

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

where TP, TN, FP and FN refer to true positives, true negatives, false positives and false negatives, respectively.
Table 1. Optimized TM-score-dependent energy thresholds based on DB179 and NB3797 (DDNA3O):

<table>
<thead>
<tr>
<th>TM-score range</th>
<th>Energy threshold</th>
<th>ΔTPa</th>
<th>TPb</th>
<th>ΔFPc</th>
<th>FPd</th>
<th>Max MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.74–1.00</td>
<td>–9.87</td>
<td>53</td>
<td>55</td>
<td>3</td>
<td>3</td>
<td>0.52</td>
</tr>
<tr>
<td>0.62–0.74</td>
<td>–13.95</td>
<td>52</td>
<td>105</td>
<td>4</td>
<td>7</td>
<td>0.73</td>
</tr>
<tr>
<td>0.58–0.62</td>
<td>–16.50</td>
<td>3</td>
<td>108</td>
<td>1</td>
<td>8</td>
<td>0.74</td>
</tr>
<tr>
<td>0.55–0.58</td>
<td>–18.64</td>
<td>4</td>
<td>112</td>
<td>0</td>
<td>8</td>
<td>0.76</td>
</tr>
<tr>
<td>0.52–0.55</td>
<td>–29.10</td>
<td>2</td>
<td>114</td>
<td>0</td>
<td>8</td>
<td>0.76</td>
</tr>
</tbody>
</table>

*Number of true positives predicted in each TM-Score bin.
*Number of false positives predicted in each TM-Score bin.
*Number of false positives predicted in each TM-Score bin.
*Number 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 11/3797 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 62 by DDNA3O (out of 104) for the APO sets, and 61 by DDNA3 and 60% to 51%). 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).

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).

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 (1trx..). 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, 1b5F5 and 1exgA) 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. DDNA3O (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

<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>1nfaA</td>
<td>1i02N</td>
<td>1hjC</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>1ukIC</td>
<td>1am9A</td>
<td>1lwB</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>1rxtA</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>1es8A</td>
<td>1dfmA</td>
<td>2bamB</td>
<td>88</td>
<td>0.68</td>
<td>−30.68</td>
<td>14.1</td>
<td>0.64</td>
<td>0.89</td>
</tr>
<tr>
<td>1ij6A</td>
<td>1ef4A</td>
<td>1rzA</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>1li1A</td>
<td>1gt0D</td>
<td>1cktA</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>1iq39A</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>1sxpA</td>
<td>1y6fA</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>1mi7R</td>
<td>1trrA</td>
<td>1gdtA</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>2audA</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>

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

- Targets from APO set.
- Targets from HOLO set.
- Template.
- Sequence identity between APO and HOLO target calculated by bl2seq in blast2.2.
- TM-score between HOLO target and template protein.
- TM-score between APO target and template protein.
- Energy value between template–target complex.
- TM-score between HOLO target and APO target.
- Template used for HOLO is unable to be used for APO because of TM-score between HOLO target and APO target.

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 0.55 and energy threshold of 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 DDNA3U 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 (DDNA3A 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 databases. The resulting DDNA3U 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 (DDNA3A and DDNA3O) trained by DB179 to structural genomics targets.

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

To further examine the possibility of overfitting in DDNA3U, we perform a 10-fold cross-validation tests on the DB250/NB3797 databases. The resulting DDNA3U 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 (DDNA3A and DDNA3O) trained by DB179 to structural genomics targets.

Fig. 3. (a) Structural comparison between APO target 1f43A and HOLO target 1le8A. Red: fragment of binding domain of 1f43A. Orange: fragment of binding domain of 1le8A. Green: template DNA of 2bamB. (b) Structural comparison between APO target 1f43A (red) and HOLO target 1le8A (green). Orange: template DNA of 1rzrA.
As shown in Table 3, application of DDNA3 leads to 32 DNA binding proteins by DBD-Hunter, DDNA3 and DDNA3O. We found that introducing an atom-type-dependent volume fraction energy function plays a significant role in prediction. There are four putative DNA binding proteins (1ug2A, 1y9bA, 2cqxA and 2fb1A) by DDNA3 overlap with those by DBD-Hunter highlights that the improvement in confidence of prediction when a consensus correction and DNA mutation in the DFIRE statistical energy function for predicting DNA binding proteins is accomplished by developing a new statistical energy function. The overlap between predicted proteins by DDNA3O and DBD-Hunter is only 19 proteins, 15 (79%) of which are 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 DNA binding proteins are from the annotations in the NCBI database.

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 DBD179/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 overlap between proteins by DBD-Hunter and DDNA3O is 0.69, thus our results are reasonably robust with different training.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Prediction</th>
<th>Putative</th>
<th>Other function</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDNA3</td>
<td>32</td>
<td>19</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>DDNA3O</td>
<td>27</td>
<td>19</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>DBD-Hunter</td>
<td>37</td>
<td>18</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>Overlap*</td>
<td>19</td>
<td>15</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

*Overlap between DBD-Hunter and DDNA3O.

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 DNA binding proteins 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 (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 DNA binding proteins 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 (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 DNA binding proteins 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 (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 DNA binding proteins 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 (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 DNA binding proteins 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 (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 DNA binding proteins 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 (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 DNA binding proteins 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 (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 DNA binding proteins 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 (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 DNA binding proteins 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/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


