Research Article

Proteins: Structure, Function and Bioinformatics
DOI 10.1002/prot.24330

DNABind: A hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning and template-based approaches

Rong Liu[1, 2] and Jianjun Hu[1*]

[1]Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina, 29208, USA

[2]Center for Bioinformatics, College of Life Science and Technology, Huazhong Agricultural University, Wuhan, 430070, P.R. China

[*] Corresponding Author

Email: jianjunh@cse.sc.edu

Tel: 803-777-7304

Fax: 803-777-3767

Short Title: DNA-binding residue prediction

Keywords: Protein-DNA interaction, DNA-binding residue, machine learning, template, structural analysis, conformational change

Received: Jan 15, 2013; Revised: May 02, 2013; Accepted: May 12, 2013

Abstract

Accurate prediction of DNA-binding residues has become a problem of increasing importance in structural bioinformatics. Here we presented DNABind, a novel hybrid algorithm for identifying these crucial residues by exploiting the complementarity between machine learning and template-based methods. Our machine learning-based method was based on the probabilistic combination of a structure-based and a sequence-based predictor, both of which were implemented using Support Vector Machines algorithms. The former included our well-designed structural features, such as solvent accessibility, local geometry, topological features, and relative positions, which can effectively quantify the difference between DNA-binding and non-binding residues. The latter combined evolutionary conservation features with three other sequence attributes. Our template-based method depended on structural alignment and utilized the template structure from known protein-DNA complexes to infer DNA-binding residues. We showed that the template method had excellent performance when reliable templates were found for the query proteins, but tended to be strongly influenced by the template quality as well as the conformational changes upon DNA binding. In contrast, the machine learning approach yielded better performance when high quality templates were not available (about 1/3 cases in our dataset) or the query protein was subject to intensive transformation changes upon DNA binding. Our extensive experiments indicated that the hybrid approach can distinctly improve the performance of the individual methods for both bound and unbound structures. DNABind also significantly outperformed the-state-of-art algorithms by around 10% in terms of Matthews's correlation coefficient. The proposed methodology could also have wide application in various protein functional site annotations. DNABind is freely available at http://mleg.cse.sc.edu/DNABind/.

Introduction

Protein-DNA interactions play critical roles in a wide range of fundamental cellular processes, including replication, transcription, recombination, and repair.[1] Despite intensive studies on the interactions between proteins and DNA over the past decades, the mechanism of protein-DNA binding and recognition remains poorly understood. Identification of the DNA-binding regions of a protein is the first step toward uncovering the nature of this interaction mode. To date, different experimental techniques, such as mutagenesis and binding assays,[2,3] have been commonly applied to this problem, but experimental determination of binding residues is a time-consuming and labor-intensive process. Hence, it is highly desirable that computational methods could guide or assist experimental approaches for systematically identifying DNA-binding sites on a large scale.

With the efforts of structural genomics projects, an increasing number of protein-DNA complex structures have been solved and deposited into the structural databases such as PDB, which provides the possibility to conduct structural and functional analysis on the interfaces between proteins and DNA with computational methods. Such studies have greatly enhanced our knowledge of DNA-binding regions. For instance, the residues located at the binding interfaces prefer to be positively charged and polar residues.[4] To facilitate contacting with DNA, the binding residues are always well exposed to solvents.[5] Compared to non-binding regions, DNA-binding residues generally have higher electrostatics potential.[6] Additionally, binding residues are more conserved than non-binding residues,[7] and putative hotspots tend to occur as clusters of conserved residues.[8] Besides the aforementioned well-known features that can be used to characterize DNA-binding residues, other structural properties such as packing density,[8,9] surface curvature,[10] B-factor,[9] residue fluctuations,[11] and hydrogen bond donors,[12] have also been successfully exploited. These structure-based analyses can provide important clues to discriminate DNA-binding residues from the protein surface.

In principle, there are two classes of computational algorithms for DNA-binding residue prediction. One is feature-based approaches,[9,13-16] which extract effective features to describe the microenvironment of the target residue and feed them into machine learning model. The other is template-based approaches,[17,18] which utilize structural alignment to retrieve the best template of the query protein from known protein-DNA complexes. Recently, several structure-based algorithms with different combinations of sequence and structural attributes have been developed, where the prediction models were implemented with machine learning techniques. Kuznetsov et al.[13] first proposed a support vector machine (SVM) predictor by combining position specific scoring matrix (PSSM) with low-resolution structural information, such as spatial neighbors, missing residues, solvent accessibility, and secondary structure. Tjong and Zhou[14] developed a prediction method called DISPLAR, which was composed of consensus neural networks using PSSM and solvent accessibility as the input features. Xiong et al.[9] integrated SVM with four features including sequence profile, solvent accessibility, packing density, and pK_a value to recognize DNA-binding residues. Dey et al.[12] also introduced a SVM-based method, but utilizing a combination of sequence conservation, spatial clustering, hydrogen bond donors, and residue propensity to identify binding sites at the patch level. Despite the promising results of these methods, accurate prediction of

DNA-binding residues is still an unsolved problem. It is therefore necessary to find novel structural features that could be more efficient in detecting these crucial residues.

In addition to structure-based predictions, a large number of sequence-based machine learning approaches have been developed for DNA-binding residue prediction during the last decade.[15,16,19-21] Generally, with the exception of evolutionary conservation, sequence features such as amino acid identity[16] and residue physicochemical properties[15] are difficult to distinctly reflect the difference between binding and non-binding regions compared to their structural counterparts. As the result, the prediction performance of sequence-based predictors was usually not as good as that of the predictors incorporating structural features. Nevertheless, sequence-based methods have a unique advantage: they do not need the protein structure to be determined. Moreover, if we attempt to combine structure and sequence-based approaches, the latter might provide additional information to achieve better overall prediction performance.

As an alternative to the commonly used machine learning approaches, the template-based predictors have been successfully applied to recognizing various protein functional sites.[22-25] Currently, there are two template-based methods related to DNA-binding residue prediction, of which the initial motivation is predicting DNA-binding function of new proteins. Gao and Skolnick[17] developed a knowledge-based method, DBD-Hunter, which combined structural alignment and the evaluation of statistical potential for identifying DNA-binding proteins and associated binding residues. Zhao et al.[18] proposed a similar procedure with an improved statistical energy function for improving the prediction accuracy. In these two studies, the binding residue prediction will be conducted only if the target protein was considered as a DNA-binding protein. Hence, the prediction performance at the residue level was just evaluated on the set of target proteins having a highly similar template. Despite the good results reported in the literature, the extent to which the template-based methods can be reliably used to identify binding residues in the comprehensive structures is still not clear. They could not be applicable when the target proteins cannot find a good template. Additionally, to the best of our knowledge, there is no study that conducted a systematic comparison between template and machine learning-based approaches. And it remains to investigate the feasibility and the best way to combine these two strategies for improving DNA-binding residue prediction.

In this paper, we propose a hybrid method for computational identification of DNA-binding residues by exploiting the complementary relationship between machine learning and template-based methods. We first conducted a comprehensive analysis of various structural features of binding regions, including solvent accessibility, local geometry, topological features, and relative positions. These features can effectively quantify the difference between binding and non-binding residues. They were then coupled with evolutionary conservation to construct a structure-based predictor using the SVM algorithm. Meanwhile, we constructed a pure sequence-based predictor on the basis of evolutionary conservation in combination with three other sequence features. Outputs of these two feature-based predictors were further combined using a linear function, which produced a slightly better performance than the individual predictors. Alternatively, we presented a template-based method for binding residue recognition, in which structural alignment was used to detect the best reference structure from a template library. Based on a representative dataset including both bound and unbound structures, we systemically demonstrated the advantages and limitations of the template and machine learning methods. Finally, we designed and evaluated an integrative predictor, called DNABind, by exploiting the complementarity between these two strategies, which distinctly outperformed the existing methods. Although our current study focused on DNA-binding residue prediction, the proposed approach can be easily extended to other types of protein functional site annotations.

Materials and Methods

Datasets

Training and test sets

In this study, we used the dataset collected by Xiong et al.,[9] which was composed of 206 DNA-binding protein chains with pairwise sequence identity less than 25%. These chains were split into two datasets: DS123 for model training and HOLO83 for independent testing. Additionally, APO83 that contains corresponding unbound structures of the chains in HOLO83 was also used to evaluate our method. Following the steps of Tjong and Zhou,[14] we defined surface residues for each chain. A residue was considered as a surface residue if its exposed surface area is larger than 10% of its nominal maximum area.[26] Further, a surface residue was defined as a binding residue if the distance between its any heavy atom and any heavy atom of the DNA molecules is less than 4.5Å. According to this definition, DS123 contains 2912 binding residues and 16016 non-binding residues, and HOLO83/APO83 includes 2038/1871 binding residues and 12200/12187 non-binding residues respectively.

Template library

A template library of 312 protein-DNA complex structures was collected in this study. We retrieved all the protein-DNA complex structures solved by X-ray crystallography with a resolution better than 3Å from the PDB database (August 2012 release). The complex structures that include only one DNA chain were eliminated. All the protein chains from the remaining 1779 complexes were further analyzed. We obtained 3567 chains with more than 40 amino acids long and at least five binding residues within 4.5Å of the DNA molecules. These chains were clustered to remove redundancy using the BLASTCLUST program[27] with identity threshold of 35% and length coverage threshold of 60%. As a result, 352 clusters were retained and the longest chain in each cluster was selected as a representative. By manually checking with the PYMOL package,[28] we considered the 312 protein chains interacting with the double-stranded DNA composed of at least six base pairs as the final templates. The list of the template chains is provided in Table S1.

Feature generation

To construct machine learning-based predictors, each residue in the query protein was characterized by a group of structural and sequence descriptors, which included evolutionary conservation, solvent accessibility, local geometry, topological features, relative position, statistical potential, predicted structural features, and amino acids indices. More details about these descriptors are given below.

Position specific scoring matrix

Position specific scoring matrix (PSSM) is widely used to reflect the evolutionary conservation of each residue in a protein sequence. The PSI-BLAST program was used to generate the PSSM of each query protein with parameters j = 3 and e = 0.001. The search was performed against the non-redundant database from NCBI.

Relative solvent accessibility

Accessible surface area is the exposed region of a molecule that is accessible to solvents. The DSSP program[29] was used to calculate the exposed surface area of each residue in the monomer structure. The ratio of the exposed surface area to the nominal maximum area was considered as the relative solvent accessibility.

Depth index and protrusion index

Depth index and protrusion index are important measures used to describe the geometric shape of a protein, which reflect the local concavity and convexity of the protein surface respectively. Their definitions are given in Supporting Information. The PSAIA software[30] with default parameters was utilized to generate the depth and protrusion-related features of each residue, including the average and standard deviation of all atom values, the average and standard deviation of all side-chain atom values, and the minimal and maximal atom values.

Topological features

Each protein structure can be considered as a residue interaction network, where each vertex represents a residue in the protein and each edge denotes that there is a physical contact between a residue pair. A contact was identified if the residue pair contains at least a pair of heavy atoms with a distance less than 5Å. Based on the network, four well-established measures, including degree, closeness, betweenness, and clustering coefficient, were extracted to describe the topological characteristics of each residue in a protein structure. The definitions of these measures are provided in Supporting Information. The resulting features should be converted into Z-score as following:

$$Z(i) = \frac{X(i) - \bar{X}}{\sigma}$$

where X(i) is the raw value of residue i for a given feature, $\bar{X}$ and σ are the average and standard deviation over all the residues in a protein structure respectively.

Distance to the centroid of the protein surface

Distance to the centroid is a global feature and reflects the relative positions of the exposed residues in the protein structure. Briefly, each surface residue was represented by its Cα atom and the geometric center of the protein surface was the average of the Cα coordinates of all the surface residues. Accordingly, the Euclidean distance between each surface residue and the centroid can be calculated. The distance of each residue should be normalized as following:

$$N(i) = \frac{D(i) - \min(i)}{\max(i) - \min(i)}$$

where D(i) is the raw distance of residue i, max(i) and min(i) are the longest and shortest distance to the centroid respectively.

Statistical pair potential

The existing studies proposed that the frequencies of observed interactions between DNA bases and protein residues follow a Boltzmann distribution.[31] Based on this assumption, Gao and Skolnick[17] derived a knowledge-based statistical pair potential by analyzing a representative set of 179 protein-DNA complex structures. Here we used this statistical potential to represent the probabilities of a given residue type interacting with fourteen types of DNA functional groups.

Predicted structural features

Sequence-derived structural features have been commonly used to predict protein binding residues. In our study, we utilized the predicted secondary structure and solvent accessibility generated by the SPINE program[32] to describe the putative structural properties of each residue as only sequence information was provided. Normalization should be conducted on predicted solvent accessibility as done on the real solvent accessibility.

Amino acid indices

An amino acid index utilizes different numerical values to represent physiochemical and biological properties of 20 residue types. Atchley et al.[33] conducted multivariate statistical analyses on 494 amino acid indices and achieved five highly interpretable patterns of amino acid variability. These patterns represent polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge, respectively. We used the reduced amino acid indices to reflect the respective attributes of each residue in a protein sequence.

Machine learning-based prediction protocol

Here a machine learning-based algorithm was proposed for the prediction of DNA-binding residues, where the Support Vector Machines (SVMs) were used to build the prediction models. We developed a structure-based and a sequence-based predictor respectively. The input of the structure-based predictor is a spatial window of 11 residues containing the target residue and its nearest surface residues. The features used in this predictor included PSSM, relative solvent accessibility, depth index and protrusion index, topological features, distance to the centroid, and statistical potentials. On the other hand, the input of the sequence-based classifier is a linear window of 11 consecutive residues centered on the target residue. The features used in this predictor contained PSSM, predicted structural features, amino acid indices, and statistical potentials. These two predictors were implemented using the LIBSVM package[34] with the radial basis function as the kernel. The optimal parameters C and γ were 10/10 and 0.1/0.05 respectively. It should be pointed out that except relative solvent accessibility, normalized centroid distance, and predicted structural attributes, the remaining features were scaled into the range from 0 to 1 using the standard logistic function. Using the probability estimation of LIBSVM, we got a probability score of the target residue to be a binding residue by each predictor. To utilize the complementarity between these two predictors, we combined their outputs as following:

$$\text{MLscore} = \alpha \times \text{STRscore} + (1-\alpha) \times \text{SEQscore}$$

where MLscore is the probability score produced by our machine learning-based protocol, STRscore and SEQscore are the probability scores generated by the structure and sequence-based predictors respectively, and the weighting factor α is set at 0.6. The optimal probability cutoffs of these three predictors are 0.57, 0.57 and 0.56 respectively.

Template-based prediction protocol

We also developed a template-based approach to identify the residues involved in DNA binding. For a query protein, any template with more than 35% sequence identity to the target chain was removed from the template library to exclude highly similar structures that may lead to biased results. The structure of the query protein was then scanned against the remaining template structures using the structural alignment program TM-align.[35] The templates were ranked according to their TM-scores. The protein-DNA complex containing the top-ranked template was selected and the query protein was superimposed onto the template structure. The rotation matrix generated by TM-align was used to achieve the superimposition. According to the putative complex structure between the target protein and the DNA, a residue was considered to be located at the binding interface if any heavy atom of this residue is within 4.5Å of the DNA molecules. Therefore, we defined that the predicted binding residue has a probability score of 1 to be a real binding residue and the non-binding residue has a probability score of 0.

DNABind: a combination of machine learning and template-based protocols

The proposed DNABind method is a hybrid prediction algorithm on the basis of combining machine learning and template-based approaches. Given a query protein, if a template can be found with a TM-score larger than a given threshold, we integrated the prediction results generated by the machine learning and template approaches using the weighted sum method similar to the construction of MLscore. Otherwise, we only considered the prediction results produced by the machine learning method. The scoring functions are defined as following:

$$\begin{cases} \text{Cscore} = \beta \times \text{MLscore} + (1-\beta) \times \text{TLscore} & \textit{if} \quad \text{TM-score} \geq \textit{cutoff} \\ \text{Cscore} = \text{MLscore} & \textit{Otherwise} \end{cases}$$

where Cscore is the probability score of the target residue produced by the integrative protocol, MLscore and TLscore are the probability scores generated by the machine learning and template-based protocols respectively, and both the TM-score cutoff and the weighting factor β are set at 0.6. According to the integrative prediction, if the Cscore of a target residue is not less than 0.47, we considered it as a binding residue; otherwise, it was categorized as a non-binding residue.

Training and testing

To estimate the effectiveness of our method, we first tested the machine learning-based algorithm on the DS123 dataset using 5-fold cross-validation. This dataset was randomly split into five partitions, four of which were used for training and the remaining one for testing. Especially, we used all the binding residues and an equal number of randomly extracted non-binding residues as the training set in each fold. The final result was the average performance of the five partitions. The optimal parameters of SVM-based predictors and the weighting factor α were determined during this stage. Furthermore, datasets HOLO83 and APO83 were considered as independent datasets to test our machine learning and template-based methods. Since the ultimate aim of our proposed prediction algorithm should be identifying the potential binding residues in the unbound structures, HOLO83 can be used to optimize the TM-score threshold and the weighting factor β in the integrative protocol. The resulting parameters were used to evaluate our performance on the APO83 dataset.

Performance measures

The prediction performance of our method was evaluated by recall, precision, F1-score and Matthews's correlation coefficient (MCC). These measures are defined as following:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{F1-score} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall}+\text{Precision}}$$

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

Here TP, TN, FP and FN are true positive, true negative, false positive, and false negative, respectively. Additionally, the receiver operating characteristic (ROC) curve was drawn by changing the probability score cutoff. The area under the ROC curve (AUC) was also reported to assess the overall performance.

Results and Discussion

Statistical analysis of structural features

In this study, we utilized a variety of structural features to describe DNA-binding residues. It is interesting to quantify the differences of these attributes between the residues involved in DNA-binding regions and those within non-binding regions. The distributions and mean values of the samples in binding and non-binding groups from DS123 were shown in Figure 1 and Table S2 respectively, where Kolmogorov-Smirnov test and t-test were combined to evaluate the statistical significance.

We first focused on the three local geometric properties including relative solvent accessibility, depth index, and protrusion index. From Figure 1(A), we can see that DNA-binding residues are more exposed on the protein surface compared to non-binding residues, which has been widely accepted in the existing studies.[9,12] To our knowledge, however, there are few reports about the comparison of depth index and protrusion index. As show in Figure 1(B)-(C), it is interesting to find that the binding residues tend to have lower depth values but higher protrusion values, suggesting that they are usually located at the convex regions of the protein surface. This might be attributed to the need of the geometric complementarity between the concave surface of DNA grooves and the protein binding regions.

Another way to characterize functional residues is to utilize topological features defined on the residue interaction network. It was found that protein functional sites such as active sites and ligand-binding residues,[36,37] typically have higher centrality values. Here we presented a systematic comparison of the network-based features in Figure 1(D)-(G). It was shown that, similar to above functional residues, the closeness and betweenness values of DNA-binding residues are also generally higher than those of the remaining protein surface. For the degree measure, conversely, the binding group has a slightly lower mean value, although the difference is not significant (Table S2). To some extent, this feature is similar to the packing density feature used by Xiong et al,[9] which also represents the local connectivity of a target residue. They showed that the average packing density of binding residues was moderately smaller than that of non-binding residues. In addition, we found that the clustering coefficients of DNA-binding regions were generally lower. This is probably due to the fact that relatively relaxed neighborhood of binding residues could provide a certain degree of flexibility for the binding of DNA molecules.

Motivated by Ben-Shimon and Eisenstein's study,[38] we proposed the centroid distance feature for characterizing DNA-binding residues. As shown in Figure 1(H), the distribution of this feature for the binding group is clearly different from that of the non-binding group, and the discrepancy of this feature is most apparent among those of all the structural attributes. Concretely, the binding residues are much closer to the geometric center of the protein surface than other exposed residues. Additionally, we observed that there is a strong negative correlation between the closeness values of residues and their distance to the centroid (Pearson's correlation coefficient = -0.827). The above analysis indicated that this new feature could be used to identify DNA-binding residues.

Since the proposed method will be finally used to predict binding residues in the unbound structures, it is necessary to check whether these structural features could characterize the binding regions of the unbound structures as well as those of the bound structures. In Figure S1-S2 and Table S3, as expected, similar phenomenon was observed for each attribute when the same statistical analysis was conducted on the unbound dataset APO83 and the bound dataset HOLO83. These results strongly suggested that the structural features introduced here are not sensitive to the conformational changes triggered by DNA binding.

Performance of machine learning-based approach tested on DS123 dataset

In this section, the machine learning-based approach which integrated the structure and sequence-based predictors was evaluated on the DS123 dataset using 5-fold cross-validation. In order to estimate the improvements introduced by our features, we considered the PSSM-based predictors as the baseline models and constructed several other predictors using different combinations of the PSSM feature with increasing number of other features.

From the Table I, we can see that when only the PSSM feature was used to build the prediction models, the predictors constructed with both spatial and linear input windows produced similar performance, with a F1-score of about 0.47 and MCC of about 0.36. This confirmed that evolutionary conservation is very important for predicting DNA-binding residues. Next, by incorporating different structural features into the baseline model with a step-by-step approach, we observed gradually increased performance of the structure-based predictors. In particular, we achieved additional gains in the prediction when the centroid distance feature was incorporated, even though it is highly related to the closeness feature. When all the structural features plus statistical potential were used, the performance reached the peak with a F1-score of 0.515 and MCC of 0.421. Meanwhile, compared to the baseline model, the AUC value of the final structure-based predictor was significantly increased from 0.803 to 0.830. The results suggested that these structural features coupled with evolutionary conservation can provide largely complementary information for DNA-binding residue prediction. Alternatively, if only sequence information was used, the performance can be improved moderately by integrating PSSM with three other sequence features. The final sequence-based predictor gave a promising result with a F1-score of 0.479, MCC of 0.376, and AUC of 0.814. It is therefore evident that both structure and sequence-derived features can contribute to the effective prediction of DNA-binding residues.

Additionally, we investigated the possibility of combining the best structure and sequence-based predictors for recognizing the binding residues. By testing different values of the weighting factor, we achieved the highest AUC value as $\alpha = 0.6$ (Figure S3). As shown in Table I and Figure 2, the integrative machine learning-based predictor showed enhanced performance with a F1-score of 0.524, MCC of 0.432, and AUC of 0.845, respectively. This indicated that the individual predictors exploited different aspects of the binding signals, and the complementarity between them resulted in the improved performance. According to these results, we selected the integrative prediction model as the machine learning-based protocol for DNA-binding residue prediction, which was to be hybridized with the template-based protocol.

Performance of machine learning-based approach tested on independent datasets: HOLO83 and APO83

To further evaluate the machine learning-based method, we tested several representative predictors from the previous section on HOLO83 and APO83, which were considered as independent datasets. As shown in Table II, the performances of two baseline models on HOLO83 were not as good as the result of 5-fold cross-validation on dataset DS123, but were still acceptable. Compared to using evolutionary conservation alone, we found that incorporation of the structural features remarkably improved the prediction performance with 3% increase in the AUC score. Adding additional sequence attributes also marginally raised the performance. Similar phenomena were also observed on the unbound dataset APO83. More importantly, the results showed that the performances of our predictors on the unbound structures, especially the structure-based predictors, were just slightly lower than those on the bound structures, which further indicated that the features used in our study are tolerable to the conformational changes upon DNA binding. Finally, we applied the integrative machine learning predictor to the independent datasets. We observed improvements in performance for the protein structures both in the absence and presence of DNA. The F1-score and MCC values were 0.483 and 0.396 for APO83, and 0.502 and 0.411 for HOLO83, respectively. The ROC curves in Figure 3 also clearly illustrated that the integrative strategy is indeed effective to enhance the prediction on the unbound structures as well as that on the bound structures.

Performance of template-based approach tested on HOLO83 and APO83

Besides the machine learning-based method, we also applied the template-based protocol to datasets HOLO83 and APO83. For each query structure, we picked out the best template with sequence identity less than 35% from the template library. As shown in Figure 4, the TM-score distributions of the top-ranked templates for HOLO83 and APO83 were largely similar, but the bound structures tended to easily identify more closely resembled templates. For instance, 13 holo queries retrieved a template with a TM-score better than 0.85, whereas only 8 apo queries detected such highly similar templates. Similar phenomenon was also reported by Gao and Skolnick.[17] Additionally, we found that there were 63 bound-unbound structure pairs sharing the same template, indicating that most of the corresponding structures in HOLO83/APO83 would experience small conformational changes before and after DNA binding.

After the best template was retrieved for the query protein, we replaced the template structure with the query structure to get the predicted protein-DNA complex structure. The putative binding residues were determined in terms of the distance between amino acids and nucleotides. From Table III, we can see that the template-based approach achieved a F1-score and MCC of 0.525 and 0.452 on the bound dataset HOLO83, and 0.417 and 0.336 on the unbound dataset APO83, respectively. Obviously, compared with the performance on the bound structures, there was a substantial deterioration in the performance on the unbound structures, suggesting that the template method is very sensitive to conformational changes. The reason will be discussed in the next section. Alternatively, we proposed another method to predict binding residues by directly aligning the target residues of the query structure with the known binding residues of the template. The results are also reported in Table III. It showed that the distance-derived prediction method clearly outperformed the alignment-derived counterpart on HOLO83, but they achieved comparable performance on APO83. Due to the fact that the former tended to produce more positive predictions, it yielded higher recall but lower precision compared to the latter. Based on this performance comparison, the distance-derived prediction was chosen in our template-based protocol.

Complementarity between machine learning and template-based methods

Our hybrid DNABind algorithm exploited the complementarity between machine learning and template-based methods to improve DNA-binding residue prediction. In this section, we provided some evidence for this motivation from two perspectives that are tightly related to the performance: one is the quality of template selection, and the other is the effect of conformational changes.

To evaluate the dependence of prediction performance on the templates, we divided the TM-score ranging from 0.4 to 1.0 into 5 partitions and deposited the query structures from HOLO83/APO83 into corresponding partitions by the TM-scores of their best templates. As shown in Table IV, when the TM-score was larger than 0.6, the template-based method can give respectable performance on each partition. Especially, the prediction performance on the bound structures of the highest two partitions was excellent. In contrast, when the TM-score was below 0.6, the performance of the template method decreased drastically. These results showed that the performance of our template-based approach strongly depended on the quality of the template. On the other hand, the performance discrepancy of the machine learning method on these two different subsets (separated by the TM-score of 0.6) was relatively smaller. Even on the partition with TM-score less than 0.5, the machine learning method achieved an acceptable result with a F1-score and MCC of around 0.39 and 0.29 for both bound and unbound structures, which were remarkably higher than the corresponding values of about 0.21 and 0.09 achieved by the template-based approach. It is thus clear that when no high quality template can be found for the query proteins, the machine learning-based approach depending on effective features to describe the microenvironment of the target residue is an alternative way to distinguish its binding function.

Another factor that affects the performance of the template-based method is the conformational changes due to DNA binding. To verify this, we first calculated the transformation changes in terms of the TM-score by aligning each bound-unbound structure pair in the HOLO83 and APO83 datasets. We then divided the 83 structure pairs into three groups with *small* (TM-score ≥ 0.95), *medium* (0.85 ≤ TM-score < 0.95), and *large* (TM-score

< 0.85) conformational changes based on the TM-scores. Next, we analyzed how the prediction performance of the template method on each group is related to the number of identical templates that were retrieved by the query protein pairs and the average TM-scores of the templates for the bound and unbound groups.

Table V showed the performance of the template method on different groups of bound and unbound structures with varying conformational changes. For the *small* transformation group with 45 structure pairs, there was no significant difference between the average TM-score of the retrieved templates for the proteins in the bound group (0.651±0.124) and that of the unbound group (0.651±0.126), suggesting that all the protein pairs in this group retrieved identical (34 pairs) or highly similar templates. Accordingly, the performance on the unbound structures was comparable to that on the bound structures with a MCC score of 0.363 against 0.382. For the protein pairs in the *medium* transformation group, 92% (24/26) of them retrieved the identical templates, and the average TM-score of the bound group (0.681±0.115) was slightly higher than that of the unbound group (0.665±0.102). In terms of the MCC measure, however, the performance on the unbound group was about 14% lower than that on the bound group. The reason might be that the conformational changes before and after DNA binding mainly happened in the local binding regions, which led to more incorrect predictions for the unbound structures despite a reasonable degree of global similarity with the template structure. As expected, in the *large* transformation group, there was only 42% (5/12) overlap between the templates for the protein pairs in the bound and unbound groups, and the average TM-score of the bound group (0.703±0.144) was obviously higher than that of the unbound group (0.623±0.085). This suggested that it is difficult for the monomer structures with large conformational changes to find ideal reference structures from existing templates derived from the complex structures. It is thus not surprising that the performance on the unbound structures in the *large* group was distinctly inferior to the performances on the *small* and *medium* groups. Furthermore, the performance discrepancy between unbound and bound structures became more remarkable as the conformational change is large, with a MCC score of 0.203 for unbound structures in comparison to 0.597 for bound structures.

Compared to the template-based approach, the machine learning-based method had much higher tolerance as regard to the conformational changes. For example, as given in Table V (results in brackets), we observed that the performance of this predictor on the unbound structures in the *large* transformation group was even slightly better than the result achieved on the *small* transformation group. Moreover, the performance differences for the bound and unbound structures were significantly smaller for all the three transformation groups. It is therefore expected that for unbound structures, the hybrid predictor DNABind, which takes advantage of the information provided by machine learning-based approach, should demonstrate better performance than the template-based method alone.

Performance of hybrid algorithm DNABind tested on HOLO83 and APO83

Ultimately, the purpose of binding residue prediction is to identify potential binding regions in protein structures without DNA molecules. Here, we utilized the bound dataset HOLO83 to optimize the TM-score cutoff and the weighting factor β in the integrative predictor DNABind. Previously, we showed that the performance of the template-based method decreased rapidly as the TM-score of the template was less than 0.6. Consequently, we expected that the optimal TM-score cutoff should be around 0.6. To achieve the highest overall performance, we conducted a grid search by testing TM-score threshold from 0.55 to 0.65 and β from 0 to 1, respectively. In Table S4, the AUC score peaked when TM-score = 0.6 and β = 0. 6. These two optimal parameters were used to validate the unbound dataset APO83.

We then conducted a systematic comparison of the machine learning-based, template-based, and hybrid predictors in Table VI. It was shown that for the overall performance on the whole datasets, the template method yielded a better result than the machine learning method on HOLO83 with a MCC of 0.452 compared to 0.411, while the machine learning method won on APO83 with a MCC of 0.396 compared to 0.336. In general, the template-based predictor tended to have higher precision, because the predicted binding residues are usually clustered and close to the real binding regions once a reliable template was retrieved, whereas some of the positive predictions generated by machine learning-based method might be scattered on the protein surface, leading to relatively high false positive rate. Additionally, it is clear that the integrative algorithm, DNABind, remarkably outperformed both individual protocols. In particular, as shown in Figure 3, compared to machine learning-based method, the AUC measures of DNABind were distinctly increased from 0.839 to 0.885 on HOLO83, and from 0.837 to 0.861 on APO83. DNABind also achieved the better MCC measures of 0.513 and 0.442 on HOLO83 and APO83 respectively in comparison to 0.453 and 0.336 provided by the template-based method. We further found that the performance difference of DNABind on bound and unbound structures became much smaller, suggesting that it had effectively alleviated the effect from conformational changes.

To further illustrate the complementary relationship of two individual methods, we separated HOLO83/APO83 into two groups based on the TM-score cutoff of 0.6. For the group with TM-score above the cutoff, the performance comparison with the template-based method showed that incorporating machine learning-based prediction into DNABind significantly improved its performance on unbound structures (APO83) while with moderate improvement on bound structures (HOLO83). As discussed earlier, although there was a reasonable global alignment between the apo query and the resulting template, some binding residues might not be correctly inferred from the template due to the local conformational changes happening in binding regions. In this case, the machine learning-based method, whose predictions mainly depended on characterizing the local context of target residues, can provide additional information to correct these mistakes. It is also worth mentioning that since the machine learning and integrative protocols used different probability score cutoffs (0.57 and 0.47), there was a difference in the results of them on the group with TM-score below 0.6. Taken together, the complementarity between the individual prediction strategies is indeed helpful in improving DNA-binding residue prediction.

Case studies

To further show the complementarity between machine learning and template approaches for DNA-binding residue prediction, we selected two DNA-binding proteins from the HOLO83 and APO83 datasets for visualizing their prediction results. The first example is the complex structure composed of the transcriptional regulator CprK (3E6C:C) and its associated DNA.[39] From Figure 5, we observed that when the machine learning predictor was used to identify the binding residues of CprK, we got 10 true positives and 7 false positives, with a F1-score of 0.606 and MCC of 0.560. For the template method, an ideal template (2CGP:A) was identified with a TM-score of 0.73, although the sequence identity between the query and template proteins was only 0.21. Based on the predicted complex structure, we obtained 7 true positives without any false positive. The F1-score and MCC of template-based method were 0.609 and 0.641. When the hybrid method was applied to the prediction, the number of true positives was the same as that produced by the machine learning approach, whereas there was only one false positive. And the F1-score and MCC were increased to 0.741 and 0.733 respectively, clearly demonstrating the advantage of DNABind for the bound structure.

We also applied these three approaches to an apo example, EcoRV endonuclease (1RVE:A),[40] which experienced a medium conformational change before and after DNA binding (the TM-score between the unbound and bound structures is 0.94). As represented in Figure 6, although the TM-score of the best template (2WIW:B) was 0.603, which was just equal to the cutoff used in DNAbind, we still correctly recognized 11 binding residues but missed the remaining 15 residues. Compared with the template-based method, the machine learning-based predictor output more positive samples, including 21 true positives and 19 false positives. The F1-score and MCC values of the individual methods were 0.537/0.636 and 0.505/0.579 respectively. Through the combination of the individual predictors, we yielded an improved result with a F1-score and MCC of 0.704 and 0.652, which produced 19 true positives, 9 false positives, and 7 false negatives, respectively. The results suggested that the complementarity between template and machine learning-based strategies is useful in enhancing the prediction of unbound structures as well as that of bound structures.

Comparison with other prediction algorithms

We compared our algorithms with two recent DNA-binding residue prediction algorithms, including Xiong et al.'s method and DISPLAR,[9,14] which also combined machine learning with structural information. To make a fair comparison, we re-implemented Xiong et al.'s method and conducted 5-fold cross-validation on the DS123 dataset. Furthermore, we evaluated these two methods using the HOLO83 and APO83 datasets.

The results in Table VII showed that when Xiong et al.'s method was tested on DS123, it yielded a F1-score and MCC of 0.483 and 0.378, which was marginally better than the performance of the baseline machine learning predictor using only the PSSM feature (Table I). Similar observation was also reported in their study. Clearly, our machine learning-based predictor, $\text{DNABind}_{\text{ML}}$, outperformed their predictor by about 5% in the corresponding measures. For the independent datasets (HOLO83 and APO83), $\text{DNABind}_{\text{ML}}$ was still obviously superior to Xiong et al.'s method for both bound and unbound structures. Compared to DISPLAR, $\text{DNABind}_{\text{ML}}$ achieved slightly better performance on HOLO83, with a F1-score of 0.502 (compared to 0.479) and MCC of 0.411 (compared to 0.396), respectively. However, for the unbound structures, the performance of DISPLAR degraded dramatically and was approximately 6% lower than that of our predictor, suggesting that DISPLAR was affected by the conformational changes upon DNA binding. Furthermore, our hybrid machine learning and template-based algorithm, DNABind, outperformed both Xiong et al.'s method and DISPLAR by more than 10% in terms of the F1-score and MCC measure.

In addition, the PDNA62 dataset collected by Ahmad et al.[5] was used as an independent dataset to compare DNABind with Xiong et al.'s method and DISPLAR. To train the $\text{DNABind}_{\text{ML}}$ model, we only used the chains in DS123 sharing less than 25% sequence identity with any chain in PDNA62. As given in Table VII, the performance of $\text{DNABind}_{\text{ML}}$ on PDNA62 is much better than that on HOLO83/APO83, with a F1-score of 0.62 and MCC of 0.5. When the template-based prediction was integrated, the corresponding measures were raised by about 5% and 7%, respectively. More importantly, the performance of DNABind was consistently superior to that of the other two structure-based algorithms. Besides, we also compared DNABind with other existing structure and sequence-based methods, such as DR_Bind,[41] BindN+,[42] NAPS,[43] and DNABINDPROT.[44] The detailed procedures and results are provided in Supporting Information (Table S5), which further demonstrated the robustness and advantage of our proposed algorithm. The success of our method is probably due to two reasons. On the one hand, compared to the structural features used in the other two methods, our comprehensive feature set more effectively captured the nature of DNA-binding residues. On the other hand, the template-based prediction method can yield excellent performance when a reasonable template can be found. It is thus complementary to the machine learning-based algorithm which has better performance for the query proteins without good templates.

In this study, we did not directly compare our approach with existing pure template-based methods,[17,18] since the same structural alignment program TM-align were used in these methods and DNABind and similar performance was expected. Zhao et al.[18] showed that there was a slight difference in the performances of DNA-binding residue prediction when they used different energy functions to re-rank the templates. This suggested that the major power of template-based predictors comes from the structural alignment procedure, although the evaluation of statistical potential is beneficial for discriminating the binding functions of proteins. It is thus reasonable to believe that compared to existing template methods, our template-based approach could achieve comparable performance and our hybrid algorithm DNABind has better predictive capability than all these template methods.

While we are working on DNABind, a new structural alignment program SPalign was developed by Yang et al.,[45] which showed improvement in predicting DNA-binding proteins over TMalign. We thus used SPalign to find the best templates for the query proteins in the HOLO83 and APO83 datasets. As shown in Table S6, SPalign has slightly improved the template-based binding residue prediction by about 2% in the MCC measure, which is agreement with the observation reported in Yang et al.'s paper. However, when SPalign algorithm was integrated with the machine learning method, the combined model did not remarkably outperform DNABind. Even so, the results clearly demonstrated that different structural alignment algorithms could be incorporated into our proposed method for improving DNA-binding residue prediction.

Conclusion

In this study, we proposed DNABind, the first hybrid algorithm for predicting DNA-binding residues, based on combining machine learning and template-based prediction strategies. Our study showed that there exists a dichotomy in the DNA-binding residue prediction problem. When good templates (such as the TM-score larger than 0.6) can be found for a query protein, the template-based algorithm achieved much higher performance than the machine learning-based predictor. Instead, when no good templates are available, the machine learning-based predictor was obviously superior to the template-based counterpart. We also demonstrated the effectiveness of our well-designed structural features, which can well quantify the difference between binding and non-binding residues and thus contributed to remarkable progress in the performance of feature-based approaches. Additionally, our systematic comparison of these two strategies showed that the performance of template-based methods was easily affected by the conformational changes upon DNA binding as well as the template quality, whereas the machine learning-based methods provided relatively stable performance. Utilizing the complementarity between these two approaches, our hybrid DNABind algorithm not only demonstrated clearly better predictive power than either method on its own, but also significantly outperformed the state-of-the-art algorithms. Since machine learning and template-based methods are widely available for other protein functional site annotations, our hybrid approach can be also extended to solve these problems.

Acknowledgements

This work was supported by the National Science Foundation Career Award (Grant BIO-DBI-0845381).

References

1. Gangloff S, Soustelle C, Fabre F. Homologous recombination is responsible for cell death in the absence of the Sgs1 and Srs2 helicases. Nat Genet 2000;25(2):192-194.
2. Rao H, Stillman B. The origin recognition complex interacts with a bipartite DNA binding site within yeast replicators. Proc Natl Acad Sci U S A 1995;92(6):2224-2228.

3. Tropel D, van der Meer JR. Characterization of HbpR binding by site-directed mutagenesis of its DNA-binding site and by deletion of the effector domain. FEBS J 2005;272(7):1756-1766.
4. Jones S, van Heyningen P, Berman HM, Thornton JM. Protein-DNA interactions: A structural analysis. J Mol Biol 1999;287(5):877-896.
5. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 2004;20(4):477-486.
6. Jones S, Shanahan HP, Berman HM, Thornton JM. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 2003;31(24):7189-7198.
7. Luscombe NM, Thornton JM. Protein-DNA interactions: Amino acid conservation and the effects of mutations on binding specificity. J Mol Biol 2002;320(5):991-1009.
8. Ahmad S, Keskin O, Sarai A, Nussinov R. Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. Nucleic Acids Res 2008;36(18):5922-5932.
9. Xiong Y, Liu J, Wei DQ. An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins 2011;79(2):509-517.
10. Tsuchiya Y, Kinoshita K, Nakamura H. Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. Proteins 2004;55(4):885-894.
11. Ozbek P, Soner S, Erman B, Haliloglu T. DNABINDPROT: fluctuation-based predictor of DNA-binding residues within a network of interacting residues. Nucleic Acids Res 2010;38:W417-W423.
12. Dey S, Pal A, Guharoy M, Sonavane S, Chakrabarti P. Characterization and prediction of the binding site in DNA-binding proteins: improvement of accuracy by combining residue composition, evolutionary conservation and structural parameters. Nucleic Acids Res 2012;40(15):7150-7161.
13. Kuznetsov IB, Gou ZK, Li R, Hwang SW. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins 2006;64(1):19-27.
14. Tjong H, Zhou HX. DISPLAR: an accurate method for predicting DNA-binding sites on

protein surfaces. Nucleic Acids Res 2007;35(5):1465-1477.

15. Wu JS, Liu HD, Duan XY, Ding Y, Wu HT, Bai YF, Sun X. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Bioinformatics 2009;25(1):30-35.
16. Yan C, Terribilini M, Wu F, Jernigan RL, Dobbs D, Honavar V. Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinformatics 2006;7:262.
17. Gao M, Skolnick J. DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. Nucleic Acids Res 2008;36(12):3978-3992.
18. Zhao H, Yang Y, Zhou Y. Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. Bioinformatics 2010;26(15):1857-1863.
19. Hwang S, Gou Z, Kuznetsov IB. DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics 2007;23(5):634-636.
20. Ofran Y, Mysore V, Rost B. Prediction of DNA-binding residues from sequence. Bioinformatics 2007;23(13):I347-I353.
21. Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 2006;34(Web Server issue):W243-248.
22. Brylinski M, Skolnick J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci U S A 2008;105(1):129-134.
23. Roy A, Zhang Y. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. Structure 2012;20(6):987-997.
24. Schmidt T, Haas J, Gallo Cassarino T, Schwede T. Assessment of ligand-binding residue predictions in CASP9. Proteins 2011;79 Suppl 10:126-136.
25. Zhang QC, Petrey D, Norel R, Honig BH. Protein interface conservation across structure space. Proc Natl Acad Sci U S A 2010;107(24):10896-10901.
26. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20(3):216-226.
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic

Acids Res 1997;25(17):3389-3402.

28. The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.
29. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577-2637.
30. Mihel J, Sikic M, Tomic S, Jeren B, Vlahovicek K. PSAIA - protein structure and interaction analyzer. BMC Struct Biol 2008;8:21.
31. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein-protein interactions. Biophys J 2003;84(3):1895-1901.
32. Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J Comput Chem 2012;33(3):259-267.
33. Atchley WR, Zhao J, Fernandes AD, Druke T. Solving the protein sequence metric problem. Proc Natl Acad Sci U S A 2005;102(18):6395-6400.
34. Fan RE, Chen PH, Lin CJ. Working set selection using second order information for training SVM. J Mach Learn Res 2005;6:1889-1918.
35. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33(7):2302-2309.
36. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S. Network analysis of protein structures identifies functional residues. J Mol Biol 2004;344(4):1135-1146.
37. Liu R, Hu J. Computational prediction of heme-binding residues by exploiting residue interaction network. PLoS One 2011;6(10):e25560.
38. Ben-Shimon A, Eisenstein M. Looking at enzymes from the inside out: The proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. J Mol Biol 2005;351(2):309-326.
39. Levy C, Pike K, Heyes DJ, Joyce MG, Gabor K, Smidt H, van der Oost J, Leys D. Molecular basis of halorespiration control by CprK, a CRP-FNR type transcriptional regulator. Mol Microbiol 2008;70(1):151-167.
40. Winkler FK, Banner DW, Oefner C, Tsernoglou D, Brown RS, Heathman SP, Bryan RK, Martin PD, Petratos K, Wilson KS. The crystal structure of EcoRV endonuclease and of its

complexes with cognate and non-cognate DNA fragments. EMBO J 1993;12(5):1781-1795.

41. Chen YC, Wright JD, Lim C. DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry. Nucleic Acids Res 2012;40(Web Server issue):W249-256.
42. Wang L, Huang C, Yang MQ, Yang JY. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst Biol 2010;4 Suppl 1:S3.
43. Carson MB, Langlois R, Lu H. NAPS: a residue-level nucleic acid-binding prediction server. Nucleic Acids Res 2010;38(Web Server issue):W431-435.
44. Ozbek P, Soner S, Erman B, Haliloglu T. DNABINDPROT: fluctuation-based predictor of DNA-binding residues within a network of interacting residues. Nucleic Acids Res 2010;38(Web Server issue):W417-423.
45. Yang Y, Zhan J, Zhao H, Zhou Y. A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction. Proteins 2012;80(8):2080-2088.

Figure Legends

Figure 1. Comparison of structural characteristics between DNA-binding and non-binding residues. (A) Relative solvent accessibility, (B) Depth index, (C) Protrusion index, (D) Degree, (E) Closeness, (F) Betweenness, (G) Clustering coefficient, (H) Distance to the centroid. The distributions were obtained by dividing all the residues from DS123 into different bins according to their attribute values and calculating the percentages of binding and non-binding residues in each bin.

Figure 2. The ROC curves of representative predictors tested on DS123. $PSSM_Q$ and $PSSM_T$ represented the baseline predictors using sequence and structural windows as inputs. SEQ and STR denoted the best sequence and structure-based predictors. STR+SEQ denoted the combination of sequence and structure-based predictors. All the predictors were evaluated using 5-fold cross-validation on DS123.

Figure 3. The ROC curves of representative predictors tested on HOLO83 and APO83. (A) HOLO83, (B) APO83. With the exception of DNABind, the predictors are the same as those in Figure 2. All the predictors were evaluated by independent test on HOLO83/APO83.

Figure 4. The TM-score distributions of the top-ranked templates for HOLO83 and APO83. The TM-score of each top-ranked template was achieved using TM-align to align the query structure with our template library.

Figure 5. Prediction results of transcriptional regulator CprK in the bound state. (A) Machine learning-based predictor, (B) Template-based predictor, (C) DNABind. The following color scheme is used: query protein in purple, template protein in grey, true positives in red, false positives in yellow, false negatives in green. In addition, we superimposed the query structure onto the template with the rotation matrix. The TM-score and sequence identity between query and template proteins were 0.73 and 0.21, respectively.

Figure 6. Prediction results of EcoRV endonuclease in the unbound state. (A) Machine learning-based predictor, (B) Template-based predictor, (C) DNABind. The color scheme is the same as that of Figure 5. The TM-score and sequence identity between query and template proteins were 0.603 and 0.24, respectively.

Table I. Performance of machine learning-based protocol on DS123

Model[a]	Type[b]	Recall	Precision	F1	MCC	AUC
PSSM	Q	0.656	0.364	0.466	0.358	0.803
PSSM+PS	Q	0.678	0.359	0.466	0.360	0.807
PSSM+PS+AA	Q	0.674	0.370	0.473	0.369	0.807
PSSM+PS+AA+SP	Q	0.678	0.373	0.479	0.376	0.814
PSSM	T	0.656	0.376	0.474	0.368	0.803
PSSM+SA	T	0.660	0.381	0.480	0.375	0.806
PSSM+SA+DP	T	0.667	0.387	0.486	0.383	0.811
PSSM+SA+DP+TP	T	0.689	0.404	0.505	0.408	0.821
PSSM+SA+DP+TP+DC	T	0.699	0.405	0.510	0.414	0.828
PSSM+SA+DP+TP+DC+SP	T	0.697	0.413	0.515	0.421	0.830
STR+SEQ	C	0.697	0.424	0.524	0.432	0.845

[a] PS: predicted structural features, AA: amino acids indexes, SP: statistical potential, SA: solvent accessibility, DP: depth index and protrusion index, TP: topological features, DC: distance to the centroid, STR: the best structure-based predictor, SEQ: the best sequence-based predictor.

[b] Q, T, and C denote the sequence-based, structure-based and combined predictors, respectively.

Table II. Performance of machine learning-based protocol on HOLO83/APO83

Dataset	Model[a]	Recall	Precision	F1	MCC	AUC
HOLO83	$PSSM_Q$	0.585	0.347	0.435	0.328	0.788
	SEQ	0.601	0.348	0.441	0.335	0.794
	$PSSM_T$	0.592	0.371	0.456	0.354	0.799
	STR	0.590	0.420	0.491	0.397	0.828
	STR+SEQ	0.590	0.437	0.502	0.411	0.839
APO83	$PSSM_Q$	0.587	0.325	0.419	0.318	0.788
	SEQ	0.603	0.327	0.424	0.326	0.795
	$PSSM_T$	0.575	0.335	0.423	0.323	0.791
	STR	0.569	0.394	0.465	0.374	0.822
	STR+SEQ	0.582	0.414	0.483	0.396	0.837

[a] Q and T denote the sequence and structure-based input windows, and SEQ and STR denote the best sequence and structure-based predictors, respectively.

Table III. Performance of template-based protocol on HOLO83/APO83

Method	Dataset	Recall	Precision	F1	MCC
Distance	HOLO83	0.496	0.558	0.525	0.452
	APO83	0.388	0.450	0.417	0.336
Alignment	HOLO83	0.351	0.613	0.447	0.401
	APO83	0.305	0.560	0.395	0.351

Table IV. Performance of individual protocols on different TM-score ranges

Method	TM-score	Recall	Precision	F1	MCC
Template	0.8-1.0	0.619[a] (0.464)[b]	0.809 (0.710)	0.701 (0.561)	0.668 (0.526)
	0.7-0.8	0.633 (0.537)	0.721 (0.614)	0.674 (0.573)	0.628 (0.522)
	0.6-0.7	0.529 (0.405)	0.503 (0.485)	0.516 (0.441)	0.438 (0.364)
	0.5-0.6	0.321 (0.255)	0.451 (0.295)	0.375 (0.274)	0.275 (0.147)
	0.4-0.5	0.221 (0.220)	0.207 (0.199)	0.214 (0.209)	0.082 (0.088)
Machine learning	0.8-1.0	0.666 (0.677)	0.496 (0.530)	0.568 (0.595)	0.493 (0.532)
	0.7-0.8	0.593 (0.617)	0.387 (0.349)	0.468 (0.446)	0.374 (0.369)
	0.6-0.7	0.635 (0.564)	0.440 (0.444)	0.519 (0.497)	0.439 (0.409)
	0.5-0.6	0.520 (0.556)	0.517 (0.480)	0.519 (0.515)	0.418 (0.416)
	0.4-0.5	0.472 (0.500)	0.354 (0.319)	0.405 (0.389)	0.297 (0.289)

[a] The performance was tested on HOLO83.

[b] The performance was tested on APO83.

Table V. Performance of individual protocols on different conformational change groups

Group	NO. of chain pairs	Structure	Ave. of TM-scores[c]	Recall	Precision	F1	MCC
Small	45[a] (34)[b]	Holo	0.651±0.124	0.456[d] (0.561)[e]	0.463 (0.406)	0.459 (0.471)	0.382 (0.387)
		Apo	0.651±0.126	0.418 (0.550)	0.454 (0.383)	0.436 (0.452)	0.363 (0.370)
Medium	26 (24)	Holo	0.681±0.115	0.522 (0.645)	0.636 (0.491)	0.573 (0.558)	0.499 (0.458)
		Apo	0.665±0.102	0.421 (0.642)	0.484 (0.465)	0.451 (0.540)	0.357 (0.445)
Large	12 (5)	Holo	0.703±0.144	0.560 (0.573)	0.755 (0.436)	0.643 (0.495)	0.597 (0.390)
		Apo	0.623±0.085	0.242 (0.565)	0.359 (0.416)	0.289 (0.479)	0.203 (0.384)

[a] The number of protein pairs in each group.

[b] The number of protein pairs sharing the same template in each group.

[c] The p-values of small, medium, and large groups provided by paired-test were 0.89, 0.02, and 0.09, respectively.

[d] The performance provided by template-based protocol.

[e] The performance provided by machine learning-based protocol.

Table VI. Performance comparison of different protocols on HOLO83/APO83

Group	NO. of Chains	Method	Recall	Precision	F1	MCC
TM-score ≥ cutoff	57[a] (55)[b]	Template	0.593[c] (0.465)[d]	0.659 (0.572)	0.624 (0.513)	0.569 (0.453)
		Machine Learning	0.630 (0.606)	0.437 (0.417)	0.516 (0.494)	0.432 (0.414)
		DNABind	0.684 (0.581)	0.610 (0.561)	0.645 (0.571)	0.585 (0.507)
TM-score < cutoff	26 (28)	Template	0.279 (0.241)	0.325 (0.252)	0.301 (0.247)	0.182 (0.122)
		Machine Learning	0.500 (0.534)	0.439 (0.406)	0.467 (0.462)	0.361 (0.360)
		DNABind	0.669 (0.689)	0.381 (0.342)	0.486 (0.457)	0.380 (0.359)
All	83 (83)	Template	0.496 (0.388)	0.558 (0.450)	0.525 (0.417)	0.452 (0.336)
		Machine Learning	0.590 (0.582)	0.437 (0.414)	0.502 (0.483)	0.411 (0.396)
		DNABind	0.679 (0.618)	0.515 (0.451)	0.586 (0.521)	0.512 (0.442)

[a] The number of chains in HOLO83.

[b] The number of chains in APO83.

[c] The performance was tested on HOLO83.

[d] The performance was tested on APO83.

Table VII. Performance comparison between our approach and existing methods

Dataset	Method	Recall	Precision	F1	MCC	AUC
DS123	Xiong et al.	0.625	0.398	0.483	0.378	0.806
	$\text{DNABind}_{\text{ML}}$[a]	0.698	0.424	0.524	0.432	0.845
HOLO83	Xiong et al.	0.586	0.378	0.459	0.358	0.800
	DISPLAR	0.461	0.499	0.479	0.396	n/a[b]
	$\text{DNABind}_{\text{ML}}$	0.590	0.437	0.502	0.411	0.839
	DNABind	0.679	0.515	0.586	0.512	0.885
APO83	Xiong et al.	0.562	0.349	0.431	0.332	0.794
	DISPLAR	0.420	0.423	0.421	0.333	n/a
	$\text{DNABind}_{\text{ML}}$	0.582	0.414	0.483	0.396	0.837
	DNABind	0.618	0.451	0.521	0.442	0.861
PDNA62	Xiong et al.	0.753	0.462	0.573	0.434	0.819
	DISPLAR	0.684	0.570	0.620	0.501	n/a
	$\text{DNABind}_{\text{ML}}$	0.784	0.512	0.620	0.500	0.856
	DNABind	0.820	0.563	0.667	0.566	0.896

[a] ML denotes the machine learning-based protocol in DNABind.

[b] Since DISPLAR only output predicted label for each residue, we cannot calculate its AUC measure.

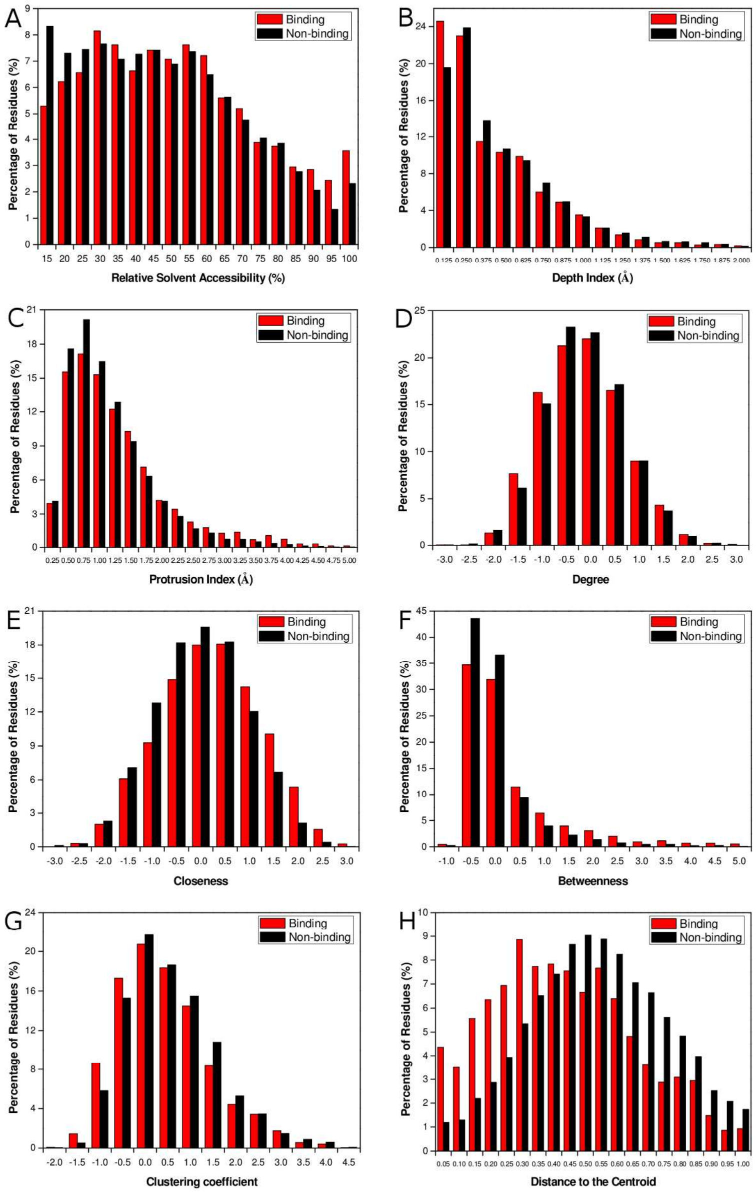

Figure 1. Comparison of structural characteristics between DNA-binding and non-binding residues. (A) Relative solvent accessibility, (B) Depth index, (C) Protrusion index, (D) Degree, (E) Closeness, (F) Betweenness, (G) Clustering coefficient, (H) Distance to the centroid. The distributions were obtained by dividing all the residues from DS123 into different bins according to their attribute values and calculating the percentages of binding and non-binding residues in each bin.
154x241mm (300 x 300 DPI)

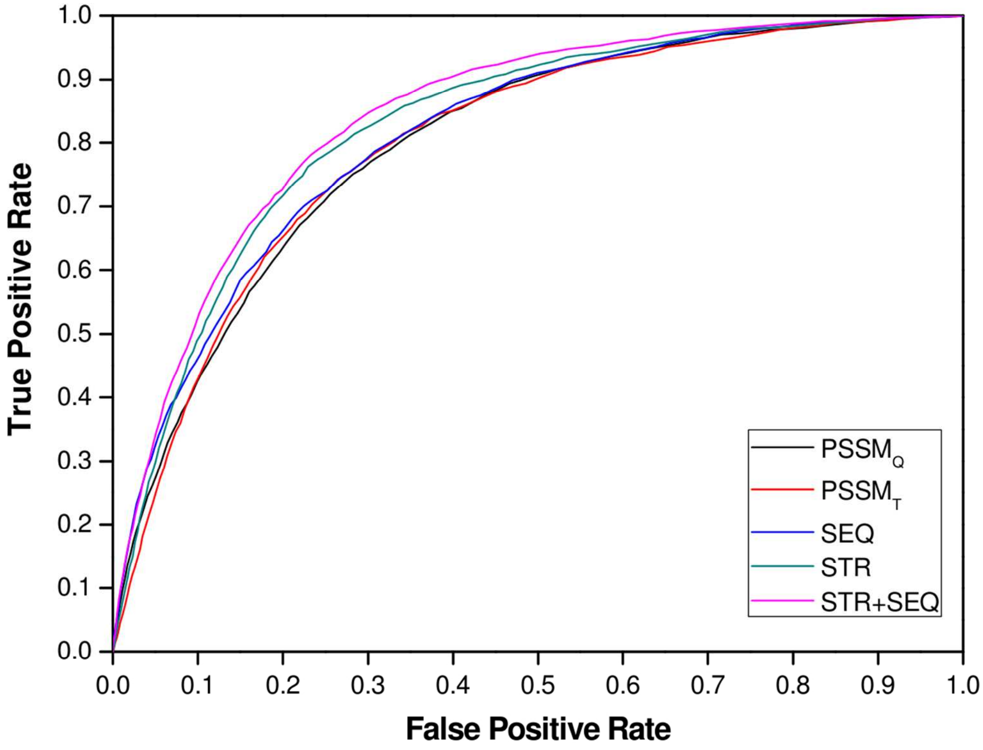

Figure 2. The ROC curves of representative predictors tested on DS123. PSSMQ and PSSMT represented the baseline predictors using sequence and structural windows as inputs. SEQ and STR denoted the best sequence and structure-based predictors. STR+SEQ denoted the combination of sequence and structure-based predictors. All the predictors were evaluated using 5-fold cross-validation on DS123.
114x86mm (300 x 300 DPI)

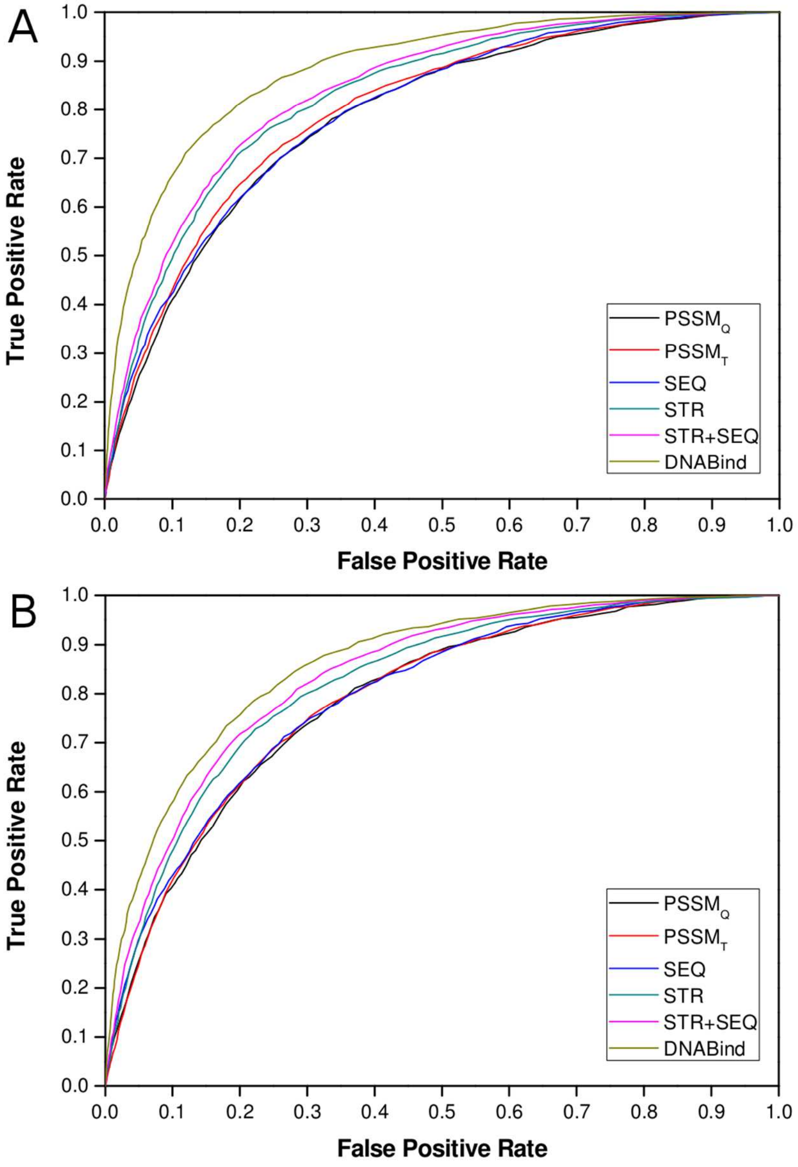

Figure 3. The ROC curves of representative predictors tested on HOLO83 and APO83. (A) HOLO83, (B) APO83. With the exception of DNABind, the predictors are the same as those in Figure 2. All the predictors were evaluated by independent test on HOLO83/APO83.
93x135mm (300 x 300 DPI)

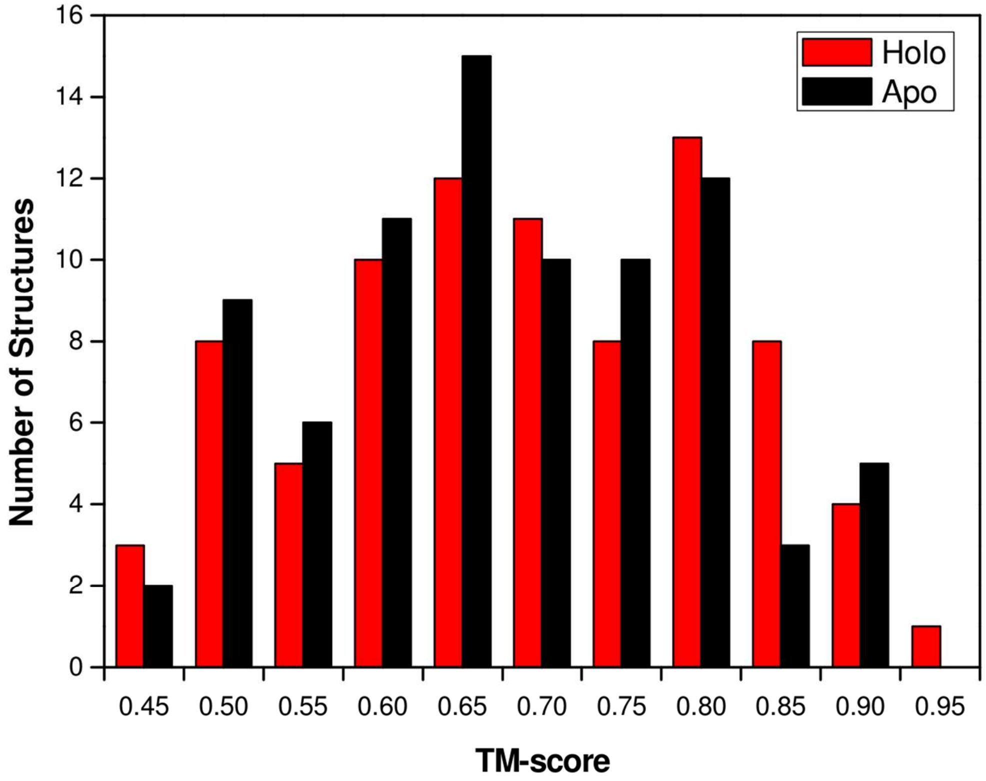

Figure 4. The TM-score distributions of the top-ranked templates for HOLO83 and APO83. The TM-score of each top-ranked template was achieved using TM-align to align the query structure with our template library.
112x88mm (300 x 300 DPI)

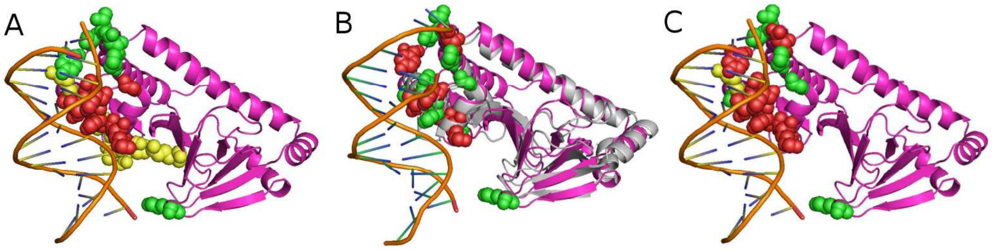

Figure 5. Prediction results of transcriptional regulator CprK in the bound state. (A) Machine learning-based predictor, (B) Template-based predictor, (C) DNABind. The following color scheme is used: query protein in purple, template protein in grey, true positives in red, false positives in yellow, false negatives in green. In addition, we superimposed the query structure onto the template with the rotation matrix. The TM-score and sequence identity between query and template proteins were 0.73 and 0.21, respectively.
174x43mm (300 x 300 DPI)

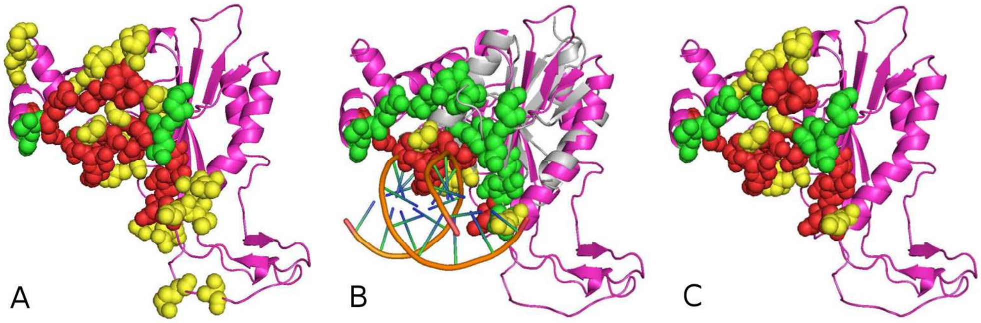

Figure 6. Prediction results of EcoRV endonuclease in the unbound state. (A) Machine learning-based predictor, (B) Template-based predictor, (C) DNABind. The color scheme is the same as that of Figure 5. The TM-score and sequence identity between query and template proteins were 0.603 and 0.24, respectively.
173x56mm (300 x 300 DPI)