Predicting Protein-DNA Binding Residues by Weightedly Combining Sequence-Based Features and Boosting Multiple SVMs

Jun Hu, Yang Li, Ming Zhang, Xibei Yang, Hong-Bin Shen, and Dong-Jun Yu

Abstract—Protein-DNA interactions are ubiquitous in a wide variety of biological processes. Correctly locating DNA-binding residues solely from protein sequences is an important but challenging task for protein function annotations and drug discovery, especially in the post-genomic era where large volumes of protein sequences have quickly accumulated. In this study, we report a new predictor, named TargetDNA, for targeting protein-DNA binding residues from primary sequences. TargetDNA uses a protein's evolutionary information and its predicted solvent accessibility as two base features and employs a centered linear kernel alignment algorithm to learn the weights for weightedly combining the two features. Based on the weightedly combined feature, multiple initial predictors with SVM as classifiers are trained by applying a random under-sampling technique to the original dataset, the purpose of which is to cope with the severe imbalance phenomenon that exists between the number of DNA-binding and non-binding residues. The final ensembled predictor is obtained by boosting the multiple initially trained predictors. Experimental simulation results demonstrate that the proposed TargetDNA achieves a high prediction performance and outperforms many existing sequence-based protein-DNA binding residue predictors. The TargetDNA web server and datasets are freely available at http://csbio.njust.edu.cn/bioinf/TargetDNA/ for academic use.

Index Terms—Protein-DNA binding residues, kernel alignment, feature weighting, classifier ensemble, imbalance learning

1 Introduction

Interactions between proteins and DNA are indispensable for biological activities and play important roles in a wide variety of biological processes [1], [2], [3], such as DNA replication, transcription, splicing, and repair. Hence, accurately locating the protein-DNA binding residues is of great importance for both analyzing protein function and designing novel drugs [4]. Much effort has been made to uncover the intrinsic mechanism of protein-DNA interactions [5], [6], and a number of high-throughput experimental technologies have been developed to confirm the interactions between DNA and proteins, such as protein binding microarray (PBM) [7], ChIP-Seq [8], and protein microarray assays [9]. However, the identification of protein-DNA binding residues via experimental technologies is often cost-intensive and time-consuming. Due to the importance of protein-DNA interactions and the difficulty in experimentally identifying DNA-binding residues, together with the fact that a huge number of unannotated protein sequences have quickly accumulated, the development of computational methods for the fast prediction of protein-DNA binding residues solely from sequences has become a hot topic in bioinformatics [1], [5], [10].

In the last decade, a series of computational methods have emerged for predicting DNA-binding residues, which have been well characterized by Si et al. [1] and Miao et al. [11]. These existing methods can be grouped into the following three main categories according to the base features used: sequence-based methods [10], [12], structure-based methods [13], [14], and hybrid methods [15] that utilize both the sequence and structural information.

It is undeniable that the prediction accuracies of structure-based and hybrid methods often outperform those of sequence-based methods [15], likely because structure-based features are more effective than sequence-based features at expressing the differences between DNA-binding and non-binding residues [15]. Many structure-based features, such as the B-factor, surface curvature and depth index (DPX), have been successfully exploited to characterize DNA-binding residues [15]. However, the applicability of structure-based and hybrid methods is limited in the common scenario where only the sequence of a given protein target is known and no corresponding 3D structure is available. Although several homology modeling tools, such as MODELLER [16]

- *J. Hu and Y. Li are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei Road, Nanjing, 210094, China. E-mail: junh_cs@126.com, balllee2011@sina.com.*
- *M. Zhang is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei Road, Nanjing, 210094, China, and the School of Computer Science and Engineering, Jiangsu University of Science and Technology, 2 Huancheng Road, Zhenjiang, 212003, China. E-mail: zhangming@just.edu.cn.*
- *X. Yang is with the School of Computer Science and Engineering, Jiangsu University of Science and Technology, 2 Huancheng Road, Zhenjiang, 212003, China. E-mail: yangxibei@hotmail.com.*
- *H.B. Shen is with the Department of Automation, Shanghai Jiao Tong University, and the Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China. E-mail: hbshen@sjtu.edu.cn.*
- *D.-J. Yu is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei Road, Nanjing, 210094, China, and the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing, 210094, P.R. China. E-mail: njyudj@njust.edu.cn.*

Manuscript received 14 Mar. 2016; revised 24 Aug. 2016; accepted 7 Oct. 2016. Date of publication 11 Oct. 2016; date of current version 6 Dec. 2017. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TCBB.2016.2616469

and I-TASSER [17], have been developed and demonstrated as feasible tools for modeling 3D structure from a given protein sequence, discrepancies between the predicted structure and the actual structure still exist, particularly for proteins that do not fit a structural template [18]. Furthermore, with ever-evolving gene-sequencing technologies, the gap between protein sequences and structures continues to widen. Therefore, sequence-based computational methods for predicting DNA-binding residues are more practical, economic, and in urgent need.

Compared to structure-based methods, sequence-based methods can quickly predict DNA-binding residues without using protein structure information. During the past decade, a number of machine-learning algorithms have been used to predict DNA-binding residues from protein sequences, and a series of sequence-based predictors have been developed, including BindN [10], DP-Bind [12], BindN+ [19], MetaDBsite [6], and DNABR [20], among others. These sequence-based predictors often utilize only protein sequence information and recognize DNA-binding residues with one or more machine-learning algorithms, such as support vector machine (SVM) [21] or random forest (RF) [22]. For example, in BindN [10], the prediction models are constructed by SVM with three sequence features, including the pK_a value of the side chain, the hydrophobicity index, and the molecular mass of an amino acid. In DP-Bind [12], three machine-learning algorithms, including SVM, kernel logistic regression, and penalized logistic regression, are integrated to predict DNA-binding residues based on the profile of evolutionary conservation of a query protein sequence in the form of a position-specific scoring matrix (PSSM) [23]. Wong et al. [24] proposed and described a computational approach, which takes into account both protein sequence and DNA information, for learning the specificity-determining residue-nucleotide interactions of different known DNA-binding domain families. In addition, Wong et al. [25] developed a HMM-based approach using belief propagations (named kmerHMM), which accepts and pre-processes PBM raw data into median-binding intensities of individual k-mers to identify DNA motifs. Despite the promising results of these methods, there remains room for further improvements in accurately predicting DNA-binding residues from protein sequences.

Another important issue that warrants careful consideration for developing machine-learning-based predictors of protein-DNA binding residues is the severe intrinsic class imbalance: the number of DNA-binding residues (minority class) is significantly fewer than that of non-binding residues (majority class). Sample rescaling is the most straightforward strategy for dealing with the issue of class imbalance [26], [27]. In this strategy, over-sampling and under-sampling are the two most commonly used implementations. As demonstrated in previous work [26], [27], [28], over-sampling will obtain an enlarged training dataset and thus will inevitably increase the training and predicting time. In addition, over-sampling may also lead to a potential over-fitting problem. On the other hand, under-sampling can obtain a more compact training dataset but comes with the risk of losing data. In view of this, in this study, we address the class imbalance by integrating under-sampling with an appropriate boosting ensemble algorithm. More specifically, we trained multiple different classifiers on balanced datasets obtained by applying random under-sampling (RUS); then, these trained classifiers are ensembled with a boosting procedure.

TABLE 1
Composition of the Training and Independent Validation Datasets

Dataset	No. of Sequences	numP [a]	numN [b]	Ratio [c]
PDNA-543	543	9,549	134,995	14.137
PDNA-TEST	41	734	14,021	19.102

[a] *numP represents the number of positive samples.*
[b] *numN represents the number of negative samples.*
[c] *Ratio = numN / numP.*

In view of the issues mentioned above, we propose a sequence-based predictor, named "TargetDNA", for the computational identification of DNA-binding residues. First, we employed the protein evolutionary information and the predicted solvent accessibility, which are determined solely from protein sequences, as two base features (refer to Section 2.2 for details). Next, to further quantify the difference between DNA-binding and non-binding residues, we utilized a centered linear kernel target alignment algorithm to learn the weights for weightedly combining the two features. Then, based on the weightedly combined feature, we trained multiple DNA-binding residue predictors with SVM as a base classifier by applying a RUS technique on the original imbalanced dataset. Finally, we obtained the ensembled predictor by using a boosting ensemble algorithm. We also created an online web server of TargetDNA, which is freely accessible for academic use at http://csbio.njust.edu.cn/bioinf/TargetDNA/.

2 Methods

2.1 Benchmark Datasets

We constructed a dataset of 7,186 DNA-binding protein chains, which had clear target annotations in the Protein Data Bank (PDB) [29] before October 10, 2015. After removing the redundant sequences using CD-hit software [30], a total of 584 non-redundant protein sequences were obtained such that no two sequences had more than 30 percent identity. Then, we divided the non-redundant sequences into two parts, the training dataset (PDNA-543) and the independent test dataset (PDNA-TEST). PDNA-543 consists of 543 protein sequences, which were all released into the PDB before October 10, 2014. PDNA-TEST includes 41 protein chains, which were all released into the PDB after October 10, 2014. More specifically, there are 9,549 DNA-binding residues (i.e., positive samples) and 134,995 non-binding residues (i.e., negative samples) in PDNA-543. PDNA-TEST consists of 734 positive samples and 14,021 negative samples. Table 1 summarizes the detailed compositions of PDNA-543 and PDNA-TEST.

2.2 Feature Representation

From the point of view of machine learning, the prediction of protein-DNA binding residues is a traditional binary classification problem. Thus, training a machine-learning-based prediction model on how to encode protein-DNA binding residues with discriminative features is one of the most crucial steps. Various effective sequence-based features, such as PSSM [12], predicted secondary structure [5],

predicted solvent accessibility [5], [14], and physicochemical properties [24], have been explored for predicting protein-DNA binding residues. In this study, we only employed two typical features for predicting protein-DNA binding residues, as follows:

2.2.1 Position Specific Scoring Matrix

PSSM profiles have been demonstrated to be an effective feature for expressing residue conservations and have been applied to many bioinformatics problems, such as the prediction of protein function [31], protein secondary structure [32], and protein-nucleotide binding residues [33], [34]. In this study, we employed the PSSM feature for predicting DNA-binding residues. The PSSM profile of a sequence is generated by using the PSI-BLAST [23] to search against the Swiss-Prot database [35] through three iterations, with 10^{-3} as the *E*-value cutoff for multiple sequence alignment. The standard logistic function was used to rescale the score of each element, denoted as x, in a PSSM profile in the interval (0,1):

$$f(x) = \frac{1}{1+e^{-x}}, \tag{1}$$

After obtaining the rescaled PSSM, the sliding window technique was utilized to extract the PSSM feature of each residue [36]. We used the sliding-window technique because we hypothesize that the DNA-binding ability of a residue depends on its own PSSM scores as well as the PSSM scores of its neighboring residues. In this study, we evaluated different window sizes (from 1 to 25, with a step size of 1) on the training dataset PDNA-543 over a ten-fold cross-validation and found that 9 is the best choice. This causes the dimensionality of the PSSM feature to be $9 \times 20 = 180$.

2.2.2 Predicted Solvent Accessibility

The concept of solvent accessibility has been widely studies since it was first introduced by Lee and Richards [37] because surface residues can be delineated by solvent accessibility information and are directly involved in interactions with other biological molecules [38]. Solvent accessibility is particularly significant in that it is closely related to the spatial arrangement and the packing of residues during the process of protein folding [38]. Moreover, there is an inseparable relationship between solvent accessibility and protein-ligand interactions, suggesting that solvent accessibility information can be used to determine protein functions. Ahmad et al. [39] demonstrated the important role of solvent accessibility to amino acid residues in predicting protein-DNA binding. Therefore, in this study, we also employed solvent accessibility information for predicting protein-DNA binding residues. We obtained the predicted solvent accessibility (PSA) characteristics of each residue by feeding the corresponding sequence to the standalone SANN program [38], which can be downloaded at http://lee.kias.re.kr/~newton/sann/. For each protein sequence, SANN precisely predicts its PSA matrix (L rows and 3 columns, where L is the length of the protein sequence), which includes the probabilities of three solvent accessibility classes (i.e., buried (B), intermediate (I), and exposed (E)) of each residue. In this study, the sliding window of size 9 was employed to construct the PSA feature of each residue. Accordingly, the dimensionality of the PSA feature is $9 \times 3 = 27$.

2.3 Learning the Weights for Combining the Two Features

It is believed that the PSSM and the PSA features potentially contain complementary discriminative information for predicting protein-DNA binding residues because these two features are extracted from different views (the evolutionary view and the physicochemical view). The most straightforward and convenient method is to serially combine the two features and obtain a super feature (i.e., PSSM+PSA) to use for training a prediction model. However, this simple combination method neglects the relative importance of the two features and thus is not guaranteed to obtain an optimal discriminative capability. Hence, developing an effective method to measure the relative importance of the two features for predicting DNA-binding residues would be especially useful. In this section, inspired by centered kernel target alignment [40], we developed a centered linear kernel target alignment (CLKTA) algorithm to learn the weights of the two features. Then, a more discriminative feature could be obtained by weightedly combining the two features.

Let B and C be two feature spaces (e.g., PSSM and PSA in this study) defined in the training sample space Ω. The dimensionalities of B and C are n and m, respectively.

For a given sample $r \in \Omega$, its corresponding feature vectors are $\mathbf{b} \in B$ and $\mathbf{c} \in C$. Our goal is to learn two optimal weights, $w_b \geq 0$ and $w_c \geq 0$, for the B and C feature spaces, respectively, such that the super feature $\mathbf{z} = (w_b\mathbf{b}^T, w_c\mathbf{c}^T)^T$ in the feature space Z has a better discriminative capability. Here, we utilize CLKTA for achieving this goal and define CLKTA as follows:

Definition 1 (CLKTA). Let $\mathbf{L} \in R^{N\times N}$ be the kernel matrix derived from the linear kernel $k_l(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^T\mathbf{z}_j$ based on a sample feature set $\mathbf{S} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ with the corresponding label vector $\mathbf{y} \in \{-1, 1\}^N$, and let $\mathbf{y}\mathbf{y}^T$ be the ideal kernel matrix for the target. The centered alignment between the linear kernel and the target on $\mathbf{S}$ is defined as

$$\mathrm{CA}\left(\mathbf{L}, \mathbf{y}\mathbf{y}^T\right) = \frac{<\mathbf{U}_N\mathbf{L}\mathbf{U}_N, \mathbf{y}\mathbf{y}^T>_F}{||\mathbf{U}_N\mathbf{L}\mathbf{U}_N||_F||\mathbf{y}\mathbf{y}^T||_F} = \frac{<\mathbf{U}_N\mathbf{L}\mathbf{U}_N, \mathbf{y}\mathbf{y}^T>_F}{N||\mathbf{U}_N\mathbf{L}\mathbf{U}_N||_F}, \tag{2}$$

where $\mathbf{U}_N = \mathbf{I}_N - (1/N)\mathbf{1}_N\mathbf{1}_N^T$ is the centering matrix, $<\cdot,\cdot>_F$ denotes the Frobenius inner product and $||\cdot||_F$ represents the Frobenius norm, which are defined as follows:

$$\forall \mathbf{X}, \mathbf{Y} \in R^{N\times N}, \quad <\mathbf{X}, \mathbf{Y}>_F = Trace(\mathbf{X}^T\mathbf{Y}), \quad ||\mathbf{X}||_F = \sqrt{<\mathbf{X}, \mathbf{X}>_F}, \tag{3}$$

Let $\mathbf{L}_b$, $\mathbf{L}_c$, and $\mathbf{L}$ be the base linear kernel matrices built from the B, C, and Z feature spaces, respectively, for training sample space Ω with the corresponding label vector $\mathbf{y} \in \{-1, 1\}^N$ (N denotes the number of samples in the training dataset). Then, we argue that the equation $\mathbf{L}_{ij} = w_b^2(\mathbf{L}_b)_{ij} + w_c^2(\mathbf{L}_c)_{ij}$ holds. We prove this argument as follows:

$$\begin{aligned}\mathbf{L}_{ij} &= k_l(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^T\mathbf{z}_j \\ &= \left(w_b\mathbf{b}_i^T, w_c\mathbf{c}_i^T\right)\left(w_b\mathbf{b}_j w_c\mathbf{c}_j\right) = w_b^2\mathbf{b}_i^T\mathbf{b}_j + w_c^2\mathbf{c}_i^T\mathbf{c}_j \\ &= w_b^2 k_l(\mathbf{b}_i, \mathbf{b}_j) + w_c^2 k_l(\mathbf{c}_i, \mathbf{c}_j) = w_b^2(\mathbf{L}_b)_{ij} + w_c^2(\mathbf{L}_c)_{ij}\end{aligned} \tag{4}$$

where $\mathbf{z}_i$, $\mathbf{b}_i$, and $\mathbf{c}_i$ are three feature vectors of the ith sample from Z, B, and C feature spaces, respectively. According to Eq. (4), we easily conclude that $\mathbf{L} = w_b^2\mathbf{L}_b + w_c^2\mathbf{L}_c$.

Inspired by centered-alignment-based kernel learning [40], we maximize the centered alignment between $\mathbf{L}$ and $\mathbf{y}\mathbf{y}^T$ to obtain the optimal weights w_b and w_c in CLKTA. It can be formulated as the following optimization problem:

$$\max_{w_b,w_c\geq 0} \mathrm{CA}(\mathbf{L}, \mathbf{y}\mathbf{y}^T) = \max_{w_b,w_c\geq 0} \frac{< \mathbf{U}_N\mathbf{L}\mathbf{U}_N, \mathbf{y}\mathbf{y}^T >_F}{||\mathbf{U}_N\mathbf{L}\mathbf{U}_N||_F}, \quad (5)$$

Let $\mathbf{w} = (w_b^2, w_c^2)^T$, $\mathbf{a}$ be the vector $(< \mathbf{U}_N\mathbf{L}_b\mathbf{U}_N, \mathbf{y}\mathbf{y}^T >_F, < \mathbf{U}_N\mathbf{L}_c\mathbf{U}_N, \mathbf{y}\mathbf{y}^T >_F)^T$, and $\mathbf{M}$ be the matrix as follows:

$$\mathbf{M} = \begin{pmatrix} < \mathbf{U}_N\mathbf{L}_b\mathbf{U}_N, \mathbf{U}_N\mathbf{L}_b\mathbf{U}_N >_F & < \mathbf{U}_N\mathbf{L}_b\mathbf{U}_N, \mathbf{U}_N\mathbf{L}_c\mathbf{U}_N >_F \\ < \mathbf{U}_N\mathbf{L}_c\mathbf{U}_N, \mathbf{U}_N\mathbf{L}_b\mathbf{U}_N >_F & < \mathbf{U}_N\mathbf{L}_c\mathbf{U}_N, \mathbf{U}_N\mathbf{L}_c\mathbf{U}_N >_F \end{pmatrix}, \quad (6)$$

Then, Eq. (5) can be rewritten as

$$\max_{w_b,w_c\geq 0} \frac{\mathbf{w}^T\mathbf{a}}{\sqrt{\mathbf{w}^T\mathbf{M}\mathbf{w}}}, \quad (7)$$

The solution to Eq. (7) can be obtained by solving a quadratic programming (QP) problem, as follows [40]:

$$\min_{w_b,w_c\geq 0} \frac{1}{2}\mathbf{w}^T\mathbf{M}\mathbf{w} - \mathbf{w}^T\mathbf{a}, \quad (8)$$

After obtaining the optimal weights, w_b and w_c, the final super feature $\mathbf{z} = (w_b\mathbf{b}^T, w_c\mathbf{c}^T)^T$, which contains more discriminative ability to represent the sample, can be used to train a high-performance prediction model.

It is worth noting that we can easily extend CLKTA to learn the weights of >2 feature spaces. Because we employed the linear kernel, according to Eq. (9), the final kernel matrix ($\mathbf{L}$), which corresponds to the final feature, can be represented as the weighted sum of all single-feature-based kernel matrixes ($\{\mathbf{L}_m\}_{m=1}^M$), i.e., $\mathbf{L} = \sum_{m=1}^M w_m^2\mathbf{L}_m$.

$$\mathbf{L}_{ij} = w_1^2(\mathbf{L}_1)_{ij} + \cdots + w_M^2(\mathbf{L}_M)_{ij} = \sum_{m=1}^M w_m^2(\mathbf{L}_m)_{ij}, \quad (9)$$

where $w_m \geq 0$ represents the weight of the mth feature space and $M \geq 2$ is the feature space number. Then, we let $\mathbf{w} = (w_1^2, \ldots, w_M^2)^T$, $\mathbf{a} = (< \mathbf{U}_N\mathbf{L}_1\mathbf{U}_N, \mathbf{y}\mathbf{y}^T >_F, \ldots, < \mathbf{U}_N\mathbf{L}_M\mathbf{U}_N, \mathbf{y}\mathbf{y}^T >_F)^T$, and $\mathbf{M} = \{\mathbf{M}_{ij}\}_{i=1,j=1}^{i=M,j=M}$, where $\mathbf{M}_{ij} = < \mathbf{U}_N\mathbf{L}_i\mathbf{U}_N, \mathbf{U}_N\mathbf{L}_j\mathbf{U}_N >_F$, and we can use the same solution (8) to solve the Eq. (5) for obtaining the optimal weights $\{w_m\}_{m=1}^M$.

2.4 Boosting Multiple SVMs Trained with Random Under-Sampling

The prediction of protein-DNA binding residues is a standard imbalanced-learning problem. By revisiting Table 1, we see that the ratio between the number of non-binding residues and that of binding residues is larger than 14. As a severe imbalance phenomenon exists between the majority class and the minority class, traditional statistical machine-learning algorithms will be biased toward the majority class [28].

In this study, we employed the RUS technique to facilitate subsequent statistical machine learning methods. RUS can effectively change the sample distributions of different classes to obtain a newly balanced training dataset. Furthermore, it can decrease the size of the training dataset, consequently accelerating the training and prediction processes [26]. However, RUS comes with the disadvantage of potentially losing useful information [26]. Thus, the prediction accuracy of the final prediction model may be decreased.

To circumvent this problem, we utilized an ensemble method by boosting multiple classifiers trained with RUS, termed B-RUS. More specifically, B-RUS first trains T base classifiers, denoted as $\{f_t\}_{t=1}^T$, with RUS on the original imbalanced training dataset; then, these trained base classifiers are ensembled by applying the boosting framework described in [41] to obtain the final classifier, denoted as $F(\mathbf{z})$, as follows:

$$F(\mathbf{z}) = \sum_{t=1}^T \beta_t f_t(\mathbf{z}), \quad (10)$$

where β_t $(1 \leq t \leq T)$ is the weight of the tth base classifier calculated by

$$\beta_t = \frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t}, \quad (11)$$

where ε_t is the weighted classification error of the tth base classifier defined as follows:

$$\varepsilon_t = \frac{\sum_{i=1}^N w(i)\cdot|f_t(\mathbf{z}_i)|\cdot U(-y_i f_t(\mathbf{z}_i))}{\sum_{i=1}^N w(i)\cdot|f_t(\mathbf{z}_i)|}, \quad (12)$$

where $U(x)$ is a function that equals 1 when $x > 0$ and 0 otherwise, $w(i)$ is the weight of the ith sample in the training dataset, N is the size of training dataset, and $f_t(\mathbf{z}_i)$ is the real-valued classification output of the tth base classifier on the input $\mathbf{z}_i$ [41].

Algorithm 1 summarizes the main steps of the proposed B-RUS, which is a customized implementation of the boosting framework described in [41] obtained by replacing the multiple-kernel base classifiers with RUS base classifiers. The purpose of using multiple-kernel base classifiers in [41] is to more effectively fuse multiple features, while that of using RUS base classifiers in B-RUS is to better cope with the class imbalance problem.

As described in Algorithm 1, we first initialize equal weights for all training samples. During each iteration k, the weight of the base classifier that owns the smallest weighted classification error is calculated. If the classifier weight is > 0, the base classifier is selected; otherwise, the procedure terminates because the performance of the currently best non-selected base classifier is worse than guessing, and we actively set the weight of each non-selected base classifier to be 0 (i.e., the non-selected base classifiers do not work for the final decision classifier). Then, in the next iteration, we assign a larger weight to each training sample, which is incorrectly classified by the newly selected based classifier. Finally, the boosting algorithm outputs a strong classifier consisting of a number of single weighted base classifiers.

TABLE 2
Performance Comparison Between PSSM, PSA, and CLKTA Features on PDNA-543 over a Ten-Fold Cross-Validation Test with a Single SVM Classifier Under $Sen \approx Spe$

Feature Type	*Sen* (%)	*Spe* (%)	*Acc* (%)	*Pre* (%)	*MCC*	AUC
PSSM	74.73	75.15	75.12	17.54	0.276	0.824
PSSM+PSA	75.71	75.75	75.75	18.09	0.285	0.835
CLKTA	76.69	76.64	76.42	18.69	0.297	0.839

TABLE 3
Performance Comparison Between PSSM, PSA, and CLKTA Features on PDNA-543 over a Ten-Fold Cross-Validation Test with a Single SVM Classifier Under $FPR \approx 5\%$

Feature Type	*Sen* (%)	*Spe* (%)	*Acc* (%)	*Pre* (%)	*MCC*	AUC
PSSM	33.79	94.91	90.88	31.98	0.280	0.824
PSSM +PSA	36.19	95.00	91.11	33.86	0.302	0.835
CLKTA	39.21	95.00	91.31	35.67	0.327	0.839

In this study, the SVM [42] is utilized to train base classifiers. We use LIBSVM [21], which is freely available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm/, to implement the SVM function. Here, a radial basis function is chosen as the kernel function. The two most important parameters, i.e., the kernel width parameter σ and the regularization parameter γ, are optimized over a five-fold cross-validation using a grid search strategy in the LIBSVM tool.

Algorithm 1. Boosting multiple classifiers trained with RUS (B-RUS)

Input: $\{(\mathbf{z}_i, y_i)\}_{i=1}^{N}$: training dataset;
T: the number of base classifiers.
Output: The final ensembled classifier $F(\mathbf{z})$
1 Training T base classifiers with RUS on original training dataset:
$S_f = \{f_t\}_{t=1}^{T}$
2 Initialize the sample weights and the classifier weights as follows:
$w_1(i) = 1/N, 1 \le i \le N$
$\beta_t = 0, 1 \le t \le T$
3 FOR $k = 1$ to T
1) For each non-selected f_t, compute weighted classification error ε_t using Eq. (12);
2) Set the weighted classification errors of all selected f_t to be $+\infty$;
3) Obtain the index t^* of the smallest weighted classification error $t^* = \arg\min_{f_t \in S_f} \varepsilon_t$;
4) Compute weight $\beta_{t^*} = \frac{1}{2}\log\frac{1-\varepsilon_{t^*}}{\varepsilon_{t^*}}$ for f_{t^*};
5) IF $\beta_{t^*} < 0$
6) Break;
7) ELSE
8) The base classifier f_{t^*} is selected;
9) END IF
10) Update the weights of samples:
$w_{k+1}(i) = \frac{w_k(i)}{L_k} e^{-\beta_{t^*} y_i f_{t^*}(\mathbf{z}_i)}$
where L_k is a normalization factor and $f_{t^*}(\mathbf{z}_i)$ is the prediction output (1 or -1) of the t^*th base classifier on the ith sample.
4 END FOR
Return The final ensembled classifier $F(\mathbf{z}) = \sum_{t=1}^{T} \beta_t f_t(\mathbf{z})$

2.5 Assessing Predictive Ability

In this study, five evaluation indexes routinely used in this field, i.e., Sensitivity (*Sen*), Specificity (*Spe*), Accuracy (*Acc*), Precision (*Pre*), and the Mathew's Correlation Coefficient (*MCC*) are utilized to evaluate predictive ability, as follows:

$$Sen = \frac{TP}{TP + FN}, \tag{13}$$

$$Spe = \frac{TN}{TN + FP}, \tag{14}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP}, \tag{15}$$

$$Pre = \frac{TP}{TP + FP}, \tag{16}$$

$$MCC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP + FN) \cdot (TP + FP) \cdot (TN + FN) \cdot (TN + FP)}}, \tag{17}$$

where *TN*, *TP*, *FN*, and *FP* are abbreviations for true negatives, true positives, false negatives, and false positives, respectively.

However, these five indexes are threshold-dependent. Hence, the method chosen for reporting these evaluation indexes is critical for making a fair comparison between different predictors. In this study, we use two strategies for selecting thresholds: 1) we select the threshold that makes $Sen \approx Spe$ (i.e., the balanced evaluation described in [43]), and 2) we select the threshold that makes $FPR \approx 5\%$ ($FPR = 1 - Spe$). Furthermore, the area under the receiver operating characteristic (ROC) curve (termed AUC), which is threshold-independent and increases in direct proportion to the overall prediction performance, is used to assess the overall predictive abilities.

3 Experimental Results and Analysis

3.1 Performance Comparisons between Different Features

In this section, the discriminative performances of the three sequence-based features, including PSSM, PSSM+PSA (simple serial combination), and CLKTA (the weighted combination described in Section 2.3), will be investigated. Each feature was evaluated by performing ten-fold cross-validation on the training dataset PDNA-543 with a single SVM classifier. In each training phase of cross-validation, we first use RUS to make the number of the majority samples equal to that of the minority samples, and we then utilize LIBSVM [21] to train a single SVM model under the balanced sample distribution. Tables 2 and 3 summarize the discriminative performance comparison between the three features on PDNA-543 over a ten-fold cross-validation test with a single SVM classifier under $Sen \approx Spe$ and $FPR \approx 5\%$, respectively.

From Tables 2 and 3, we observe that the PSSM+PSA feature consistently outperforms the PSSM feature in terms of all six evaluation indexes. Using Table 3 as an example, the *Sen*, *Pre*, *MCC*, and AUC of the PSSM+PSA feature are 36.19,

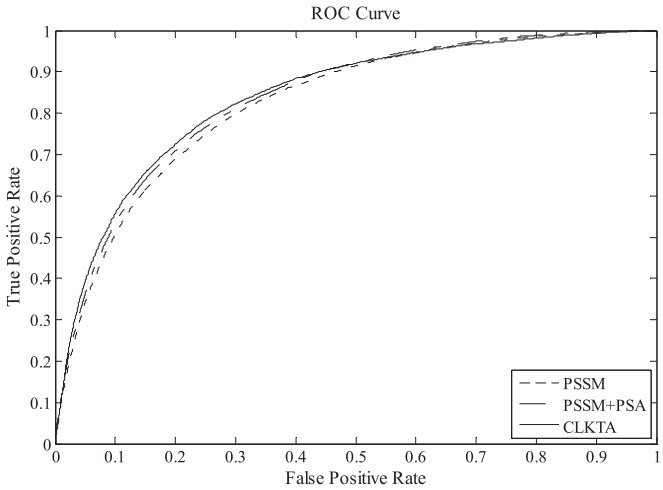


Fig. 1. ROC curves for PSSM, PSSM+PSA, and CLKTA features on PDNA-543 over ten-fold cross-validation.

33.86 percent, 0.302, and 0.835, respectively, which are improvements of approximately 2.40, 1.88, 2.20, and 1.10 percent, respectively, over the PSSM feature. As for the remaining two evaluation indexes, the PSSM+PSA feature also slightly outperforms the PSSM feature.

As for the CLKTA and PSSM+PSA features, Tables 2 and 3 show that the discriminative performance of the CLKTA feature is consistently better than that of the PSSM+PSA feature. In Table 2, the six evaluation index values of the CLKTA feature are all slightly higher than those of the PSSM+PSA feature. Table 3 shows that the *Sen*, *Pre*, and *MCC* of the CLKTA feature are 39.21, 35.67 percent, and 0.327, respectively, which represent improvements of approximately 3.02, 1.81, and 2.5 percent, respectively, over the PSSM+PSA feature. Furthermore, to directly show the overall prediction performance of the three features, Fig. 1 illustrates the ROC curves of the three features on PDNA-543 over a ten-fold cross-validation test.

From the comparison results between the three features listed in Tables 2, 3, and Fig. 1, we empirically demonstrate that the three features are highly useful, and the CLKTA feature is the best method for effectively predicting protein-DNA binding residues.

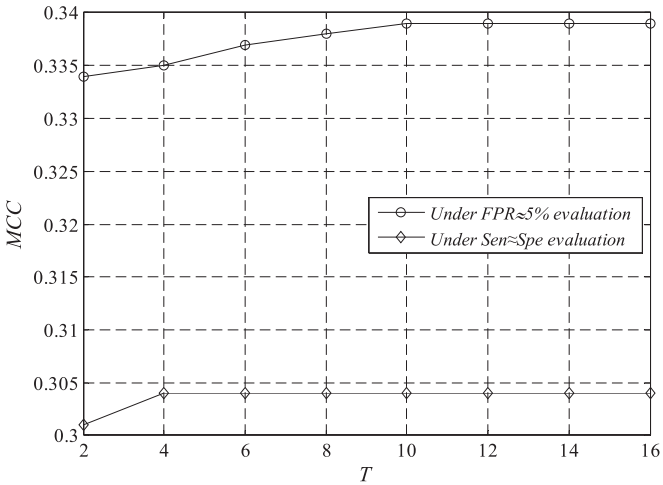


Fig. 2. The performance variation curves of *MCC* versus *T* under $Sen \approx Spe$ and $FPR \approx 5\%$.

TABLE 4
The Detailed *MCC* Values Under $Sen \approx Spe$ and $FPR \approx 5\%$ for Different Values of *T*

T	2	4	6	8	10	12	14	16
MCC under $Sen \approx Spe$	0.301	0.304	0.304	0.304	0.304	0.304	0.304	0.304
MCC under $FPR \approx 5\%$	0.334	0.335	0.337	0.338	0.339	0.339	0.339	0.339

3.2 Selecting the Number of Base Classifiers in B-RUS

In this section, we attempt to empirically demonstrate how to choose the number of base classifiers (*T*) for training an ensembled classifier with B-RUS. Under the two threshold selection strategies (i.e., $Sen \approx Spe$ and $FPR \approx 5\%$), we evaluated the *MCC* performance variations of an ensembled classifier on the training dataset (i.e., PDNA-543) over a ten-fold cross-validation by gradually varying the value of *T* from 2 to 16, with a step size of 2.

Fig. 2 plots the performance variation curves of *MCC* versus *T* under the two threshold selection strategies. Table 4 summarizes the detailed values of *MCC* under each threshold selection strategy for different values of *T*.

From Fig. 2 and Table 4, we find that the value of *MCC* first increases with increasing *T* and then converges. Under $Sen \approx Spe$, the first maximum *MCC* value is achieved when $T = 4$. Continually increasing the number of base classifiers will not further enhance the performance of the ensembled classifier. Under $FPR \approx 5\%$, the first maximum *MCC* value is achieved when $T = 10$, and no improvement can be observed with larger values of *T*. Hence, in all the subsequent experiments, we set $T = 10$ to train an ensembled classifier with B-RUS.

To demonstrate the efficacy of B-RUS, Table 5 and Fig. 3 list the performance comparisons between with B-RUS and without B-RUS on the PDNA-543 dataset over a ten-fold cross-validation test under both $Sen \approx Spe$ and $FPR \approx 5\%$. As shown in Table 5, values obtained with B-RUS are consistently better than those obtained without B-RUS under both evaluation methods and in terms of all six evaluation indexes. The results shown in Table 5 and Fig. 3 indicate that the prediction performance is indeed improved after applying B-RUS.

3.3 Comparisons with Existing Predictors of Protein-DNA Binding Residues

In this section, we demonstrate the efficacy of the proposed method, TargetDNA, by comparing it with other common predictors of protein-DNA binding residues, including BindN [10], BindN+ [19], ProteDNA [44], DP-Bind [12],

TABLE 5
Performance Comparisons Between with and Without B-RUS on PDNA-543 over Ten-Fold Cross-Validation Under $Sen \approx Spe$ and $FPR \approx 5\%$

Evaluation methods	with/without B-RUS	*Sen* (%)	*Spe* (%)	*Acc* (%)	*Pre* (%)	*MCC*	AUC
$Sen \approx Spe$	without	76.69	76.64	76.42	18.69	0.297	0.839
	with	76.98	77.05	77.04	19.18	0.304	0.845
$FPR \approx 5\%$	without	39.21	95.00	91.31	35.67	0.327	0.839
	with	40.60	95.00	91.40	36.47	0.339	0.845

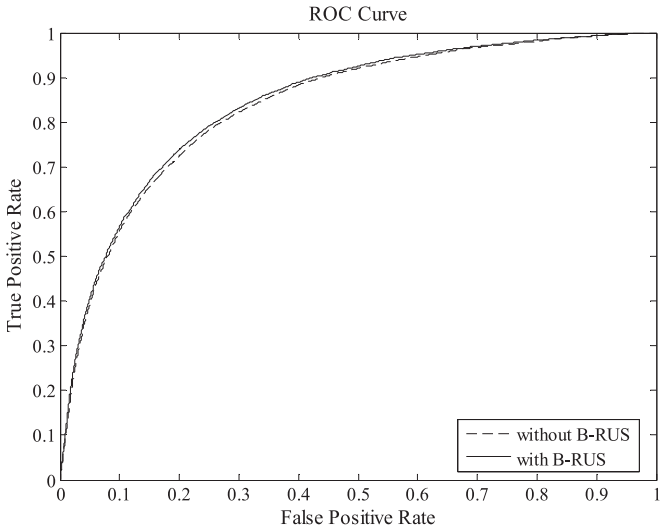


Fig. 3. ROC curves of with and without B-RUS predictors on PDNA-543 over a ten-fold cross-validation test.

MetaDBSite [6], and DNABind [15], by performing independent validation tests on PDNA-TEST, the results of which are listed in Table 6.

By observing Table 6, we see that TargetDNA achieves satisfactory results with the best and second-best *MCC* values of 0.300 and 0.269 under $FPR \approx 5\%$ and $Sen \approx Spe$, respectively. Compared with BindN+, which is an updated version of BindN, TargetDNA achieves an improvement of 12.2 percent on *MCC* under $FPR \approx 5\%$. Additionally, TargetDNA under $Sen \approx Spe$ consistently outperforms BindN+ under $Spe \approx 85\%$ (default setting of BindN+) for all of the five evaluation indexes. As for MetaDBSite, a meta approach for predicting sequence-specific protein-DNA binding residues, the proposed TargetDNA also achieves improvements of 11.30, 4.91, 0.48, and 7.9 percent on *Sen*, *Pre*, *Acc*, and *MCC*, respectively, under $FPR \approx 5\%$. We note that there are four base predictors (or web servers) of MetaDBSite, i.e., BindN-RF [45], DBS-PRED [5], DISIS [46], and DNABindR [47], which cannot normally be used to predict protein-DNA binding residues. We have not failed to notice

TABLE 6
Performance Comparisons Between the Proposed TargetDNA and Other Predictors of Protein-DNA Binding Residues on PDNA-TEST

Predictor	*Sen* (%)	*Spe* (%)	*Acc* (%)	*Pre* (%)	*MCC*
BindN [a]	45.64	80.90	79.15	11.12	0.143
ProteDNA [b]	4.77	99.84	95.11	60.30	0.160
BindN+ ($FPR \approx 5\%$) [c]	24.11	95.11	91.58	20.51	0.178
BindN+ ($Spe \approx 85\%$) [c]	50.81	85.41	83.69	15.42	0.213
MetaDBSite [a]	34.20	93.35	90.41	21.22	0.221
DP-Bind [a]	61.72	82.43	81.40	15.53	0.241
DNABind [d]	70.16	80.28	79.78	15.70	0.264
TargetDNA ($Sen \approx Spe$)	60.22	85.79	84.52	18.16	0.269
TargetDNA ($FPR \approx 5\%$)	45.50	93.27	90.89	26.13	0.300

[a] *Results computed using the MetaDBSite server at http://projects.biotec.tu-dresden.de/metadbsite/.*
[b] *Results computed using the ProteDNA server at http://serv.csbb.ntu.edu.tw/ProteDNA/service.php.*
[c] *Results computed using the BindN+ server at http://bioinfo.ggc.org/bindn+/.*
[d] *Results computed using the DNABind server at http://mleg.cse.sc.edu/DNABind/.*

TABLE 7
Performance Comparison Between the Proposed TargetDNA and Other Predictors of Protein-DNA Binding Residues on PDNA-316 [6] over a Ten-Fold Cross-Validation Test

Predictor	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*
DBS-PRED *	53.00	76.00	75.00	0.170
BindN *	54.00	80.00	78.00	0.210
DNABindR *	66.00	74.00	73.00	0.230
DISIS *	19.00	98.00	92.00	0.250
DP-Bind *	69.00	79.00	78.00	0.290
BindN-rf *	67.00	83.00	82.00	0.320
MetaDBSite *	77.00	77.00	77.00	0.320
TargetDNA ($Sen \approx Spe$)	77.96	78.03	78.02	0.339
TargetDNA ($FPR \approx 5\%$)	43.02	95.00	90.99	0.375

* *Data excerpted from [6].*

that ProteDNA achieves the highest *Pre* value (60.30 percent). However, the corresponding *Sen* value is the lowest (4.77 percent), denoting too many false negatives are incurred during prediction with this method. On the other hand, DNABind [15], which uses both protein sequence information and structural information, achieves the best performance on *Sen* (70.16 percent) but a much lower *Spe* value, implying too many false positives are incurred during prediction with this method.

3.4 Performance Comparison on a Dataset Used with Other Predictors

In addition to the ten-fold cross-validation test and independent test performed above, to fairly evaluate the prediction performance of the proposed TargetDNA, we also compared it with other protein-DNA predictors using the same datasets employed by the compared predictors. In this section, the PDNA-316 dataset, which was constructed by Si et al. [6], is employed to further demonstrate the efficacy of TargetDNA. The details of PDNA-316 can be found in [6].

We compare TargetDNA with DP-Bind [12], DNABindR [47], DISIS [46], DBS-PRED [5], BindN-rf [45], BindN [10], and MetaDBSite [6] on PDNA-316 over a ten-fold cross-validation as performed in [6], and the comparison results are listed in Table 7. The results shown in Table 7 clearly demonstrate that TargetDNA outperforms the other predictors in terms of the *MCC*, which is an overall index for evaluating the quality of binary prediction. Compared with the second-best performer MetaDBSite [6], the *Sen*, *Spe*, *Acc*, and *MCC* values of TargetDNA under $Sen \approx Spe$ evaluation method are 77.96, 78.03, 78.02 percent, and 0.339, respectively, which are improvements of approximately 0.96, 1.03, 1.02, and 1.9 percent over MetaDBSite [6], respectively. To further demonstrate the performance of TargetDNA on the PDNA-316 dataset, we also calculated the prediction results of TargetDNA under the $FPR \approx 5\%$ evaluation method, and the highest *MCC* value 0.375 is obtained with a satisfactory sensitivity value of 43.02 percent. It is noted that DISIS [46] gained the highest specificity and accuracy but comes with the lowest sensitivity (19.00 percent). However, sensitivity is a measure of DNA-binding residue prediction, which is of the most interest for many researchers [6]. The low sensitivity value of DISIS indicates that this method is most likely to mistakenly predict DNA-binding residues as non-binding residues, leading to a low *MCC*.

4 Conclusions

In this study, we have designed and implemented a new sequence-based predictor of protein-DNA binding residues, named TargetDNA. TargetDNA is trained on the DNA-binding protein dataset collected from the most recently released PDB [48] with a CLKTA method, RUS technique, SVM, and the boosting classifier ensemble strategy. Experimental results with a training dataset and an independent validation dataset have demonstrated the efficacy of the proposed TargetDNA. The superior performances of TargetDNA are due to several reasons, including an appropriate benchmark dataset, more discriminative feature design, and careful construction of the prediction model. Currently, the TargetDNA prediction server has already been made available online and can predict DNA-binding residues for each query protein sequence.

We note that the current CLKTA method has two potential disadvantages. First, CLKTA may lack the ability to remove noisy features or information. Second, CLKTA lacks a non-linear learning ability, which means that CLKTA cannot effectively learn the dataset features with non-linear distributions. In our future work, to use CLKTA to process the prediction tasks with noisy information, we will employ effective denoising methods, such as sparse learning [49] and deep learning [50], to remove the noise contained in the original features before applying CLKTA. We will also update CLKTA with a non-linear kernel trick to cope with the non-linear learning issue.

Another point of concern is the relatively long computation time of TargetDNA (approximately 450 s for a sequence of 300 residues). This long computation time stems from the fact that TargetDNA has to perform PSI-BLAST [23], SANN [38] and LIBSVM [21] to extract features and predict protein-DNA binding residues. In the future, we will attempt to accelerate the computation speed by using several servers to concurrently perform these computations.

Molecules binding motifs mining is a long-term challenge for understanding their functions. The failure of forming correct interactions between some critical molecules has been revealed as one of the important causes for diseases like cancer [51]. The TargetDNA model developed in this study is specifically for identifying the protein-DNA binding residues, and in the future work, we will further investigate the applicability of our model to other types of molecules binding residues prediction problems, e.g., ATP-protein bindings [52], RNA-protein bindings [53], and the sequence specificities of DNA- and RNA-binding proteins prediction [54]. The current model is also expected to be applied on topics of cancer research regarding protein-DNA binding residues mining [55], genome-scale sequence analysis [56], and human single-nucleotide polymorphisms (SNPs) predictions [57].

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61373062, 61671288, and 61572242), the Natural Science Foundation of Jiangsu (No. BK20141403), the Fundamental Research Funds for the Central Universities (No. 30916011327), "The Six Top Talents" of Jiangsu Province (No. 2013-XXRJ-022), Science and Technology Commission of Shanghai Municipality (No. 16JC1404300), China Scholarship Council (No. 201606840087), National Key Research and Development Program: Key Projects of International Scientific and Technological Innovation Cooperation between Governments (No. S2016G9070), and the Open Project Program of Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education (No. JYB201603). Dong-Jun Yu is the corresponding author for this paper.

References

[1] J. N. Si, R. Zhao, and R. L. Wu, "An overview of the prediction of protein DNA-binding sites," *Int. J. Mol. Sci.*, vol. 16, no. 3, pp. 5194–5215, 2015.

[2] K. A. Aeling, N. R. Steffen, M. Johnson, G. W. Hatfield, R. H. Lathrop, and D. F. Senear, "DNA deformation energy as an indirect recognition mechanism in protein-DNA interactions," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 4, no. 1, pp. 117–125, Jan.-Mar. 2007.

[3] K. C. Wong, Y. Li, C. Peng, H. S. Wong, "A comparison study for DNA motif modeling on protein binding microarray," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 2, pp. 1–1, Mar./Apr. 2016.

[4] P. Schmidtke, and X. Barril, "Understanding and predicting druggability. A high-throughput method for detection of drug binding sites," *J. Medicinal Chemistry*, vol. 53, no. 15, pp. 5858–5867, 2010.

[5] S. Ahmad, M. M. Gromiha, and A. Sarai, "Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information," *Bioinf.*, vol. 20, no. 4, pp. 477–486, 2004.

[6] J. Si, Z. Zhang, B. Lin, M. Schroeder, and B. Huang, "MetaDBSite: A meta approach to improve protein DNA-binding sites prediction," *BMC Syst. Biol.*, vol. 5, no. Suppl 1, 2011, Art. no. S7.

[7] M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. 3rd Estep, and M. L. Bulyk, "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities," *Nature Biotechnol.*, vol. 24, no. 11, pp. 1429–1435, Nov. 2006.

[8] A. Valouev, et al., "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data," *Nature Methods*, vol. 5, no. 9, pp. 829–834, Sep. 2008.

[9] S. W. Ho, G. Jona, C. T. L. Chen, M. Johnston, and M. Snyder, "Linking DNA-binding proteins to their recognition sequences by using protein microarrays," *Proc. Nat. Academy Sci. United States Amer.*, vol. 103, no. 26, pp. 9940–9945, Jun. 27, 2006.

[10] L. Wang and S. J. Brown, "BindN: A web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences," *Nucleic Acids Res.*, vol. 34, no. Suppl 2, pp. W243–W248, 2006.

[11] Z. Miao and E. Westhof, "A Large-scale assessment of nucleic acids binding site prediction programs," *PLoS Comput. Biol.*, vol. 11, no. 12, Dec. 2015.

[12] S. Hwang, Z. Gou, and I. B. Kuznetsov, "DP-Bind: A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins," *Bioinf.*, vol. 23, no. 5, pp. 634–636, 2007.

[13] S. Jones, J. A. Barker, I. Nobeli, and J. M. Thornton, "Using structural motif templates to identify proteins with DNA binding function," *Nucleic Acids Res.*, vol. 31, no. 11, pp. 2811–2823, 2003.

[14] H. Tjong and H.-X. Zhou, "DISPLAR: An accurate method for predicting DNA-binding sites on protein surfaces," *Nucleic Acids Res.*, vol. 35, no. 5, pp. 1465–1477, 2007.

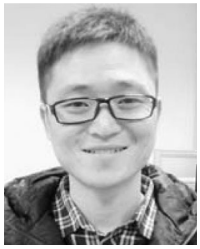
[15] B.-Q. Li, K.-Y. Feng, J. Ding, and Y. D. Cai, "Predicting DNA-binding sites of proteins based on sequential and 3D structural information," *Mol. Genetics Genomics*, vol. 289, no. 3, pp. 489–499, 2014.

[16] N. Eswar, B. Webb, and M. A. Marti-Renom, et al., "Comparative protein structure modeling using MODELLER," *Current Protocols Bioinf./Editoral Board*, pp. 5.6.1–5.6.30, 2006, doi:10.1002/0471250953.bi0506s15.

[17] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC Bioinf.*, vol. 9, no. 1, 2008, Art. no. 40.

[18] C. Kauffman and G. Karypis, "Computational tools for protein-DNA interactions," *Wiley Interdisciplinary Rev.-Data Mining Knowl. Discovery*, vol. 2, no. 1, pp. 14–28, Jan./Feb., 2012.

[19] L. Wang, C. Huang, M. Q. Yang, and J. Y. Yang, "BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features," *BMC Syst. Biol.*, vol. 4, no. Suppl 1, 2010, Art. no. S3.
[20] X. Ma, J. Guo, H.-D. Liu, J. M. Xie, and X. Sun, "Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 6, pp. 1766–1775, Nov./Dec. 2012.
[21] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.
[22] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[23] A. A. Schaffer, et al., "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements," *Nucleic Acids Res.*, vol. 29, no. 14, pp. 2994–3005, Jul. 15, 2001.
[24] K. C. Wong, Y. Li, C. B. Peng, A. M. Moses, and Z. Zhang, "Computational learning on specificity-determining residue-nucleotide interactions," *Nucleic Acids Res.*, vol. 43, no. 21, pp. 10180–10189, Dec. 2, 2015.
[25] K. C. Wong, T. M. Chan, C. B. Peng, Y. Li, and Z. Zhang, "DNA motif elucidation using belief propagation," *Nucleic Acids Res.*, vol. 41, no. 16, Sep. 2013, Art. no. e153.
[26] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[27] J. Hu, Y. Li, W. X. Yan, J.-Y. Yanga, H.-B. Shenb, and D.-J. Yua, "KNN-based dynamic query-driven sample rescaling strategy for class imbalance learning," *Neurocomputing*, vol. 191, pp. 363–373, May 26, 2016.
[28] D. J. Yu, J. Hu, Z. M. Tang, H.-B. Shen, J. Yang, and J.-Y. Yang, "Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling," *Neurocomputing*, vol. 104, pp. 180–190, Mar. 2013.
[29] P. W. Rose, et al., "The RCSB protein data bank: views of structural biology for basic and applied research and education," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D345–D356, 2015.
[30] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinf.*, vol. 22, no. 13, pp. 1658–1659, 2006.
[31] J. C. Jeong, X. Lin, and X.-W. Chen, "On position-specific scoring matrix for protein function prediction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 2, pp. 308–315, Mar./Apr. 2011.
[32] M. H. Zangooei and S. Jalili, "Protein secondary structure prediction using DWKF based on SVR-NSGAII," *Neurocomputing*, vol. 94, pp. 87–101, 2012.
[33] D. J. Yu, J. Hu, J. Yang, H. B. Shen, J. Tang, and J. Y. Yang, "Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 4, pp. 994–1008, Jul./Aug. 2013.
[34] K. Chen, M. J. Mizianty, and L. Kurgan, "Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors," *Bioinf.*, vol. 28, no. 3, pp. 331–41, Feb. 1, 2012.
[35] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 45–48, 2000.
[36] J. S. Chauhan, N. K. Mishra, and G. P. Raghava, "Identification of ATP binding residues of a protein from its primary sequence," *BMC Bioinf.*, vol. 10, 2009, Art. no. 434.
[37] B. Lee and F. M. Richards, "The interpretation of protein structures: estimation of static accessibility," *J. Mol. Biol.*, vol. 55, no. 3, pp. 379–400, 1971.
[38] K. Joo, S. J. Lee, and J. Lee, "Sann: Solvent accessibility prediction of proteins by nearest neighbor method," *Proteins-Structure Function Bioinf.*, vol. 80, no. 7, pp. 1791–1797, 2012.
[39] S. Ahmad, M. M. Gromiha, and A. Sarai, "Real value prediction of solvent accessibility from amino acid sequence," *Proteins-Structure Function Genetics*, vol. 50, no. 4, pp. 629–635, Mar. 1, 2003.
[40] Y. Lu, L. Wang, J. Lu, J. Yang, and C. Shen, "Multiple kernel clustering based on centered kernel alignment," *Pattern Recognit.*, vol. 47, no. 11, pp. 3656–3664, 2014.
[41] F. Yang, H. Lu, and M.-H. Yang, "Robust visual tracking via multiple kernel boosting with affinity constraints," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 2, pp. 242–254, Feb. 2014.
[42] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley-Interscience, 1998.
[43] D. J. Yu, et al., "TargetATPsite: A template-free method for ATP-binding sites prediction with residue evolution image sparse representation and classifier ensemble," *J. Comput. Chemistry*, vol. 34, no. 11, pp. 974–985, Apr. 2013.
[44] W.-Y. Chu, Y.-F. Huang, C.-C. Huang, Y. S. Cheng, C. K. Huang, and Y. J. Oyang, "ProteDNA: A sequence-based predictor of sequence-specific DNA-binding residues in transcription factors," *Nucleic Acids Res.*, pp. W396–W401, 2009.
[45] L. Wang, M. Q. Yang, and J. Y. Yang, "Prediction of DNA-binding residues from protein sequence information using random forests," *BMC Genomics*, vol. 10, no. Suppl 1, 2009, Art. no. S1.
[46] Y. Ofran, V. Mysore, and B. Rost, "Prediction of DNA-binding residues from sequence," *Bioinf.*, vol. 23, no. 13, pp. I347–I353, Jul. 1, 2007.
[47] C. Yan, M. Terribilini, F. Wu, R. L. Jernigan, D. Dobbs, and V. Honavar, "Predicting DNA-binding sites of proteins from amino acid sequence," *BMC Bioinf.*, vol. 7, no. 1, 2006, Art. no. 262.
[48] P. W. Rose, et al., "The RCSB protein data bank: redesigned web site and web services," *Nucleic Acids Res.*, vol. 39, pp. D392–D401, Jan. 2011.
[49] W. Zhang, L. Zhang, Z. Jin, R. Jin, D. Cai, X. Li, R. Liang, and X. He, "Sparse learning with stochastic composite optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. 99, pp. 1–1, 2016, doi: 10.1109/TPAMI.2016.2578323.
[50] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2014.
[51] Y. Y. Xu, F. Yang, Y. Zhang, and H. B. Shen, "An image-based multi-label human protein subcellular localization predictor (iLocator) reveals protein mislocalizations in cancer tissues," *Bioinf.*, vol. 29, no. 16, pp. 2032–2040, 2013.
[52] K. Chen, M. J. Mizianty, and L. Kurgan, "ATPsite: sequence-based prediction of ATP-binding residues," *Proteome Sci.*, vol. 9, Suppl 1, no. Suppl 1, pp. 295–297, 2011.
[53] H. Ren and Y. Shen, "RNA-binding residues prediction using structural features," *BMC Bioinf.*, vol. 16, no. 1, pp. 1–10, 2015.
[54] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nature Biotechnol.*, vol. 33, no. 8, pp. 831–838, 2015.
[55] K. C. Wong, "A novel approach to predict core residues on cancer-related DNA-binding domains," *Cancer Inf.*, vol. 15, no. Suppl 2, pp. 1–7, 2016.
[56] K. C. Wong, Y. Li, C. B. Peng, and Z. Zhang, "SignalSpider: Probabilistic pattern discovery on multiple normalized ChIP-Seq signal profiles," *Bioinf.*, vol. 31, no. 1, pp. 17–24, 2015.
[57] K. C. Wong and Z. L. Zhang, "SNPdryad: Predicting deleterious non-synonymous human SNPs using only orthologous protein sequences," *Bioinf.*, vol. 30, no. 8, pp. 1112–1119, 2014.



Jun Hu received the BS degree in computer science from Anhui Normal University, China, in 2011. Currently, he is working towards the PhD degree in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include bioinformatics, data mining, and pattern recognition.

Yang Li received the BS degree in computer science from the Nanjing University of Science and Technology, China, in 2014. He is currently working toward the PhD degree in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include bioinformatics, data mining, and pattern recognition.

Ming Zhang received the BS and MS degrees in computer science from the Jiangsu University of Science and Technology in 2002 and 2005, respectively, and the PhD degree in pattern recognition and machine intelligence from the Nanjing University of Science and Technology, in 2013. He is currently an associate professor in the School of Computer Science and Engineering at the Jiangsu University of Science and Technology. His current research interests include granular computing, pattern recognition, and bioinformatics.

Xibei Yang received the BS degree from Xuzhou Normal University (XZNU), Xuzhou, China, in 2002, the MS degree from the Jiangsu University of Science and Technology (JUST), Zhenjiang, China, in 2006, and the PhD degree from the Nanjing University of Science and Technology (NJUST), Nanjing, China, in 2010, both in computer applications. He is currently an associate professor with JUST, and he is also a postdoctoral researcher with NJUST. He has published more than 100 articles in international journals and international conferences. His research interests include granular computing, rough set, and decision making.

Hong-Bin Shen received the PhD degree from Shanghai Jiaotong University, China, in 2007. He was a postdoctoral research fellow of Harvard Medical School from 2007 to 2008. Currently, he is a professor in the Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University. His research interests include data mining, pattern recognition, and bioinformatics. He has published more than 60 papers and constructed 20 bioinformatics servers in these areas and he serves the editorial members of several international journals.

Dong-Jun Yu received the BS degree in computer science and the MS degree in artificial intelligence from the Jiangsu University of Science and Technology, in 1997 and 2000, respectively, and the PhD degree in pattern analysis and machine intelligence from the Nanjing University of Science and Technology, in 2003. In 2008, he acted as an academic visitor with the University of York, UK. He is currently a professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His current interests include pattern recognition, data mining, and bioinformatics.