Protein-DNA Binding Residue Prediction via Bagging Strategy and Sequence-Based Cube-Format Feature

Jun Hu, Yan-Song Bai, Lin-Lin Zheng, Ning-Xin Jia, Dong-Jun Yu, and Gui-Jun Zhang

Abstract—Protein-DNA interactions play an important role in diverse biological processes. Accurately identifying protein-DNA binding residues is a critical but challenging task for protein function annotations and drug design. Although wet-lab experimental methods are the most accurate way to identify protein-DNA binding residues, they are time consuming and labor intensive. There is an urgent need to develop computational methods to rapidly and accurately predict protein-DNA binding residues. In this study, we propose a novel sequence-based method, named PredDBR, for predicting DNA-binding residues. In PredDBR, for each query protein, its position-specific frequency matrix (PSFM), predicted secondary structure (PSS), and predicted probabilities of ligand-binding residues (PPLBR) are first generated as three feature sources. Secondly, for each feature source, the sliding window technique is employed to extract the matrix-format feature of each residue. Then, we design two strategies, i.e., square root (SR) and average (AVE), to separately transform PSFM-based and two predicted feature source-based, i.e., PSS-based and PPLBR-based, matrix-format features of each residue into three corresponding cube-format features. Finally, after serially combining the three cube-format features, the ensemble classifier is generated via applying bagging strategy to multiple base classifiers built by the framework of 2D convolutional neural network. The computational experimental results demonstrate that the proposed PredDBR achieves an average overall accuracy of 93.7% and a Mathew's correlation coefficient of 0.405 on two independent validation datasets and outperforms several state-of-the-art sequenced-based protein-DNA binding residue predictors. The PredDBR web-server is available at https://jun-csbio.github.io/PredDBR/.

Index Terms—Protein-DNA binding residue, convolutional neural network, sequence-based features, bagging strategy

1 Introduction

Interactions between proteins and DNAs play a crucial role in a wide variety of biological processes, such as, DNA replication, recombination, repair, gene transcription and expression [1], [2], [3]. Hence, the accurate prediction of protein-DNA binding residues contributes to elaborate the interaction mechanism of them, and facilitate our understanding of these biological processes. Traditionally, protein-DNA binding residues can be identified by experimental techniques, such as electrophoretic mobility shift assays (EMSAs) [4], [5], Fast ChIP [6], and X-ray crystallography [7]. However, these techniques are time-consuming and laborious. With the rapid advance of protein sequencing technology, a large amount of unannotated protein-DNA complexes is sequenced and deposited. Therefore, there is an urgent need to develop computational methods that can rapidly and reliably identify DNA-binding residues from protein sequences.

During the past decades, many computation-based methods have been proposed to predict protein-DNA binding residues. Generally speaking, these existing methods are roughly divided into three categories: sequenced-based methods [8], [9], [10], structure-based methods [11], [12], and hybrid methods [13], [14], [15] that both sequence information and structural information are used. In the early stage, structure-based methods, such as, PreDs [16], DNABINDPROT [17], DISPLAR [18], DBD-Hunter [19], and DR_bind [20], dominated in the field of predicting protein-DNA binding residues. Due to most of the existing structure-based methods only extracted the available information from the three-dimension (3D) structures, that is the sequence information is ignored, their prediction performance is limited. To overcome this issue, hybrid methods attempt to integrate sequence-based/sequence-driven features and 3D structure-based features to further improve prediction accuracy. Such as, Igor B. Kuznetsov *et al.* [14] incorporated multiple sequence-based features and low-resolution structure information to generate a DNA-binding residue prediction model using Support Vector Machine (SVM) [21] algorithm; Li *et al.* [13] employed not only sequenced-based/sequenced-driven features, i.e., position specific scoring matrix (PSSM), residual disorder, protein secondary structure (PSS), protein solvent accessibility (PSA), but also five 3D structural features to identify protein-DNA binding sites. Although the structure-based and hybrid

- *Jun Hu, Yan-Song Bai, Lin-Lin Zheng, Ning-Xin Jia, and Gui-Jun Zhang are with the College of Information Engineering, Zhejiang University of Technology, Hangzhou, Zhejiang 310023, China. E-mail: {hujunum, 2111903289, 2111903253, zgj}@zjut.edu.cn, jianingxin97@163.com.*
- *Dong-Jun Yu is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China. E-mail: njyudj@njust.edu.cn.*

Manuscript received 14 January 2021; revised 22 September 2021; accepted 24 October 2021. Date of publication 29 October 2021; date of current version 8 December 2022.
This work was supported in part by the National Natural Science Foundation of China under Grants 61902352, 62072243, 61772273, and 61773346, in part by the Natural Science Foundation of Zhejiang under Grants LY21F020025 and LZ20F030002, in part by the Natural Science Foundation of Jiangsu under Grant BK20201304, and in part by the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant RF-A2020012.
(Corresponding authors: Dong-Jun Yu and Gui-Jun Zhang.)
Digital Object Identifier no. 10.1109/TCBB.2021.3123828

methods could achieve a good prediction performance of DNA-binding residues, all of them depend on the protein 3D structure information and could not be directly employed to predict these proteins whose 3D structures are not available. Luckily, the 3D structure of proteins could be predicted via many existing computational methods, such as I-TASSER [22], ROSETTA [23], and AlphaFold [24]; however, there are still lots of proteins whose 3D structures cannot achieve high-quality structures prediction via these methods. It is urgent to develop the methods which only utilize protein sequence information.

In recent years, many sequence-based methods have been proposed to identify protein-DNA binding residues, most of which utilize the machine learning algorithms to complete this prediction task, such as, DP-Bind [9], BindN+ [25], DNABR [26], enDNA-Prot [27], MetaDBSite [28], DNABind [29], MLAB [30], ProteDNA [31], TargetDNA [32], EL_PSSM-RT [33], iProDNA-CapsNet [34], CNNsite [35], and DNAPred [36]. For example, DP-Bind [9] extracted evolutionary information from protein sequences and employed three machine learning algorithms, i.e., SVM, kernel logistic regression, and penalized logistic regression, to predict protein-DNA binding residues. EL_PSSM-RT [33] proposed a novel position-specific scoring matrix (PSSM) encoding method and combined the algorithms of SVM and random forest [37] to locate the protein-DNA binding residues. MLAB [30] designed a novel method, called Multi-scale Local Average Blocks, to predict the DNA-binding residues of proteins, which utilizes the ensemble weighted sparse representation algorithm to dig out the available information from the evolutionary information and PSA. CNNsite [35] designed a convolutional neural network based on multiple sequence features, e.g., PSSM, PSS, and PSA, to locate the protein-DNA binding residues. DNAPred [36], designed by our group, combined the PSSM, PSS, PSA, and amino acid frequency difference between binding and non-binding residues to generate the feature vector of each residue, which is inputted to the prediction model built by a two-stage imbalanced learning algorithm to improve the prediction accuracy. These existing methods have obtained some promising results, but effective information of protein sequences is not extracted adequately. There remains room to further improve the sequence-based prediction accuracy in identifying protein-DNA binding residues.

In this study, we propose a new sequence-based method, called PredDBR, to further improve the performance of protein-DNA binding residue prediction. Specifically, we first extract three feature sources, i.e., position specific frequency matrix (PSFM), predicted secondary structure (PSS), and predicted probabilities of ligand-binding residues (PPLBR), from protein sequences. Secondly, for each feature source, the sliding window technique is employed to extract the matrix-format feature of each residue. Then, to extract more effective information, we design two strategies, i.e., square root (SR) and average (AVE), to separately transform PSFM-based and two predicted feature source-based, i.e., PSS-based and PPLBR-based, matrix-format features of each residue into three corresponding cube-format features. Finally, after serially combining the three cube-format features, we obtain the ensemble classifier via applying bagging strategy to multiple base classifiers built by the framework of 2D convolutional neural network. Experimental results show that the proposed PredDBR outperforms other existing state-of-the-art predictors.

TABLE 1
Statistical Summary of PDNA-543, PDNA-335, PDNA-316, PDNA-52, and PDNA-41

Dataset	No. of Sequence	P_{pos} [a]	P_{neg} [b]	Ratio [c]
PDNA-543	543	9549	134995	14.14
PDNA-335	335	6461	71320	11.04
PDNA-316	316	5609	67109	11.96
PDNA-52	52	973	16225	16.68
PDNA-41	41	734	14021	19.10

[a] P_{pos} *represents the number of positive samples.*

[b] P_{neg} *represents the number of negative samples.*

[c] *Ratio* $= P_{neg}/P_{pos.}$

2 Materials and Methods

2.1 Benchmark Datasets

To evaluate the performance of the proposed PredDBR, we employ five protein-DNA binding residue datasets, i.e., PDNA-543 [32], PDNA-335 [38], and PDNA-316 [28], PDNA-52 [38], and PDNA-41 [32]. Here, PDNA-543, PDNA-335, and PDNA-316 are used for ten-fold cross-validation tests. PDNA-41 and PDNA-52 are utilized for independent validation tests.

PDNA-543 and PDNA-41 are constructed during our previous work [32]. PDNA-543 consists of 543 DNA-binding protein sequences, which are released into the Protein Data Bank (PDB) [39] before October 10, 2014. PDNA-41 consists of 41 DNA-binding protein sequences, which are released into the PDB after October 10, 2014. The identity between any two sequences selected from the union set of PDNA-543 and PDNA-41 is no more than 30%. PDNA-335 and PDNA-52 are collected in the previous work [38]. PDNA-335 consists of 335 DNA-binding protein sequences, which are released into the PDB before 10 March 2010, from BioLip [40]. PDNA-52 consists of 52 DNA-binding protein sequences, which are released into the PDB after 10 March 2010, from BioLip. There is no sequence in PDNA-335 that has more than 40% pairwise identity to the sequences in PDNA-52. PDNA-316, which is constructed by Si *et al.* [28], consists of 316 DNA-binding protein sequences. The identity of any pairwise of sequences in PDNA-316 is no more than 30%.

The detailed statistical summary of the five datasets used in this study is demonstrated in Table 1. The five datasets could be easily downloaded at https://jun-csbio.github.io/PredDBR/ freely for academic use. The names of proteins in the five data sets are listed in Supporting Text S1.

2.2 Feature Sources

To effectively predict DNA-binding residues from protein sequence information, three feature sources, i.e., position-specific frequency matrix (PSFM), predicted secondary structure (PSS), and predicted probabilities of ligand-binding residues (PPLBR), are employed to encode the feature representation of each residue.

2.2.1 Position-Specific Frequency Matrix

The protein evolutionary information has been demonstrated to be effective in predicting DNA-binding residues [38]. In this study, to obtain the profile that expresses the evolutionary information of each protein, the HHblits [41] software is first employed to thread the corresponding protein sequence against the Uniclust30 sequence database [42] through three iterations with 0.001 as the *E*-value cutoff to generate multiple sequence alignments (MSA). A profile, named as position-specific frequency matrix (PSFM), is then computed from the MSA.

Let the size of MSA is $M \times L$, where M is the number of protein sequences in MSA and L is the length of query protein. There are 21 element types in MSA numbered from 1 to 21, including 20 natural amino acid types and the gap type. Then, the *i*th row and *j*th column value in the corresponding PSFM, i.e., $\mathrm{PSFM}_{i,j}$, could be calculated as:

$$\mathrm{PSFM}_{i,j} = \frac{\delta_{\mathrm{MSA}}(i,j)}{\mathrm{M}} \tag{1}$$

where $\delta_{\mathrm{MSA}}(i,j)$ represents the total number of the *j*th element type in the *i*th column in MSA, $i = 1, 2, \cdots, L$, and $j = 1, 2, \cdots, 21$.

2.2.2 Predicted Secondary Structure

The information of predicted secondary structure has been demonstrated to be available for DNA-binding residues prediction [38]. In this study, we also employ the PSIPRED software [43] to generate the predicted secondary structure information. Concretely, given a protein sequence with L residues, the PSIPRED [43] software will output an $L \times 3$ probability matrix, which includes the probabilities of three secondary structure classes (i.e., coil (C), helix (H), and strand (E)) of each residue.

2.2.3 Predicted Probabilities of Ligand-Binding Residues

Generally, the information of ligand-binding residues has a positive effect on the prediction of DNA-binding residues. Hence, in this study, the predicted probabilities of ligand-binding residues (PPLBR) are employed to be one important feature source. In order to generate the PPLBR of each protein, we have designed a sequence-based ligand-binding residue predictor method I-LBR [44], which is available at https://jun-csbio.github.io/I-LBR for academic use. The standalone package of I-LBR could also be downloaded at https://github.com/jun-csbio/I-LBR. For each query protein sequence with L residues, I-LBR with the default settings is used to predict its PPLBR, whose size is $L \times 1$. Concretely, the *i*th element in PPLBR represents the ligand-binding probability value of the *i*th residue in the query protein.

2.3 Cube-Format Feature Extraction

It has been found that one target residue is influenced by its context residues in a protein sequence [33]. In order to extract more detailed information about the relationship between one target residue and its context residues, the feature information of the target residue is represented as cube-format for containing more meaningful information. In this study, we design two different strategies, i.e., square root (SR) and average (AVE), to generate the cube-format feature of each target residue from the three above-described feature sources. The detailed descriptions of SR and AVE are described as follows:

2.3.1 Strategy of SR

In this strategy of SR, for each feature source with size of $L \times K$, a sliding window with size of W is first employed to generate the corresponding matrix-format feature of each target residue in one protein with L residues. Here, the sliding window is centered at this target residue and the size of matrix-format feature of each residue should be $W \times K$. Then, based on the matrix-format feature (F^M), each element ($F^{C_{SR}}_{k,i,j}$) of the corresponding cube-format feature ($F^{C_{SR}}$) could be easily generated by the SR strategy as follows:

$$F^{C_{SR}}_{k,i,j} = \sqrt{F^M_{i,k} \times F^M_{j,k}} \tag{2}$$

where $F^M_{i,k}$ represents the *i*th row and *k*th column element value in F^M, $k = 1, 2, \cdots, K$, $i = 1, 2, \cdots, W$, and $j = 1, 2, \cdots, W$. Hence, the size of $F^{C_{SR}}$ is $K \times W \times W$.

2.3.2 Strategy of AVE

Similar to the strategy of SR, for each feature source with size of $L\times K$, the matrix-format feature (F^M) of each target residue in each protein is first generated via using the sliding window with size of W in the AVE strategy. Then, based on the F^M, each element ($F^{C_{AVE}}_{k,i,j}$) of the corresponding cube-format feature ($F^{C_{AVE}}$) could be calculated by the AVE strategy as follows:

$$F^{C_{AVE}}_{k,i,j} = \frac{F^M_{i,k} + F^M_{j,k}}{2} \tag{3}$$

where $F^M_{i,k}$ represents the *i*th row and *k*th column element value in F^M, $k = 1, 2, \cdots, K$, $i = 1, 2, \cdots, W$, and $j = 1, 2, \cdots, W$. Hence, the size of $F^{C_{AVE}}$ is $K \times W \times W$.

In this study, we have empirically tested the performance of the two strategies (SR and AVE) on three different feature sources, i.e., PSFM, PSS, and PPLBR (see details in Section 3.1). It is found that the SR strategy is more suitable for generating the cube-format feature on the PSFM feature source than the AVE strategy. On the contrary, the AVE strategy could achieve a higher performance than the SR strategy on the feature sources of PSS and PPLBR.

2.4 Framework of 2D Convolutional Neural Network

In order to dig out more detailed information from the above cube-format feature, in this study, a new framework of 2D convolutional neural network (2D-CNN) is designed to train one classifier of protein-DNA binding residue prediction. Fig. 1 demonstrates the 2D-CNN framework.

As shown in Fig. 1, there are four modules in the new 2D-CNN framework. The first three modules are convolutional modules and the last module is a fully connectional module (FCM). For each convolutional module (ConvM), there is one convolutional (Conv) layer, one batch normalized (BN) layer, and one rectified linear unit (ReLU) layer. The Conv layer of the first ConvM employs 50 filters of kernel size 3 ×

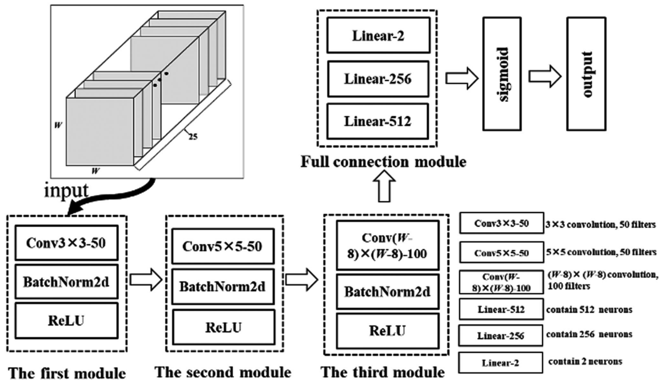
W
W
25
input
Linear-2
Linear-256
Linear-512
Full connection module
sigmoid
output
Conv3×3-50
BatchNorm2d
ReLU
Conv5×5-50
BatchNorm2d
ReLU
Conv(W-8)×(W-8)-100
BatchNorm2d
ReLU
The first module
The second module
The third module
Conv3×3-50 3×3 convolution, 50 filters
Conv5×5-50 5×5 convolution, 50 filters
Conv(W-8)×(W-8)-100 (W-8)× (W-8) convolution, 100 filters
Linear-512 contain 512 neurons
Linear-256 contain 256 neurons
Linear-2 contain 2 neurons

Fig. 1. Framework of 2D convolutional neural network.

3. The Conv layer of the second ConvM utilizes 50 filters of kernel size 5×5. The Conv layer of the third ConvM uses 100 filters whose kernel sizes depends on the size of the inputted cube-format feature, so as to the output element number is 900. Concretely, for the inputted cube-format feature with size of K$\times W \times W$, the size of each filter in the Conv layer of the third ConvM is $(W-8) \times (W-8)$.

The FCM contains three linear layers and one sigmoid activation layer. The first linear layer (LL) has 512 neurons connected to all 900 elements output by the third ConvM. The second LL has 256 neurons. The third LL includes two neurons. The sigmoid activation layer is finally employed to output two probability values of belonging to the classes of DNA-binding and non-DNA-binding residues.

2.5 Bagging-Based Ensemble Learning Scheme

The problem of DNA-binding residue prediction is a typical imbalanced learning problem [45]. By revisiting Table 1, it is easy to find that the imbalance rate is larger than 10. The imbalanced learning problem inherent in the prediction of DNA-binding residues has a potential negative effect on the final prediction performance. Hence, to enhance the performance of DNA-binding residue prediction, it is an imperative task to alleviate the negative influence of imbalanced learning problem.

The scheme of ensemble learning has been demonstrated to be an effective method for reducing the above negative influence and enhancing the performance of DNA-binding residue prediction [10], [30], [33], [36]. Hence, in this study, we also employ the ensemble learning scheme to enhance the prediction performance via integrating several base classifiers learned by the above 2D-CNN framework.

The bagging strategy [46] is an easy and effective method to tackle the imbalanced data problem. To make full use of the positive samples, we modify the classical bagging strategy to learn an ensemble model. As Fig. 2 shown, in the modified bagging strategy, there are the following steps: (1) separating positive samples (i.e., DNA-binding residues) and negative samples (i.e., non-DNA-binding residues) from original training sets to comprise the positive sample set (*PS*) and the negative sample set (*NS*); (2) randomly selecting $T \cdot |PS|$ samples from *NS* to construct a negative sample subset (NS_{sub}), where *T* means the imbalance degree value and $|PS|$ means the sample number in *PS*; (3) using step 2 to generate *N* different negative sample subsets; (4) combing each NS_{sub} and *PS* to generate one new training subset (*TrS*) and generating *N* new *TrS*s; (5) using the above-described 2D-CNN framework to learn one base classifier on each *TrS* and learning *N* base classifiers; and (6) applying the mean ensemble strategy to decide the final probability values of belonging to the positive and negative classes each target residue. Finally, the target residue is labeled as this class corresponding to the higher probability value. Note that, due to the positive samples are paid more attentions, the sample is labeled as a positive one, when its probability values of two classes are equal.

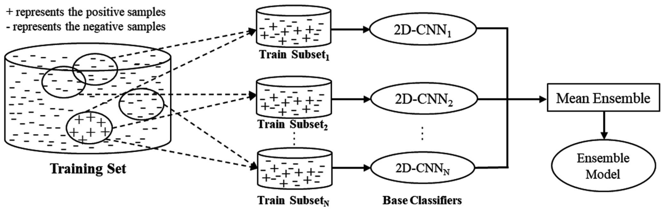

Fig. 2. The schematic diagram of the modified bagging strategy.

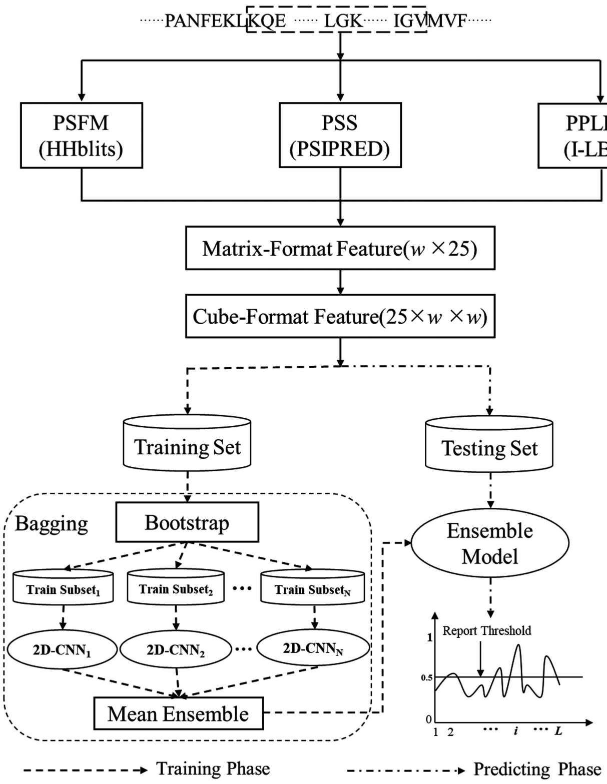

Fig. 3. Architecture of PredDBR.

2.6 Architecture of PredDBR

Fig. 3 illustrates the architecture of the proposed PredDBR. For each query protein sequence, PredDBR first employs the HHblits [41], PSIPRED [43], and I-LBR [44] to generate three features sources, i.e., PSFM, PSS, and PPLBR. Secondly, the strategies of SR and AVE described in Section 2.3 are employed to generate the PSFM-based cube-format feature and the PSS-based and PPLBR-based cube-format features of each target residue, respectively. The final cube-format feature of each target residue is easily obtained via serially combining the first dimension of the PSFM-based, PSS-based, and PPLBR-based cube-format features. In the training phase, after generating the cube-format feature of all target residues in the training dataset, we can gain the corresponding imbalanced training sample set. The bagging-based ensemble learning scheme described in Section 2.5 is used to learn the final prediction model. In the prediction phase, for each protein to be predicted, the ensemble prediction model could be adopted to give the probability output for each target residue belonging to the class of DNA-binding residues. The web-server of PredDBR is freely available at https://jun-csbio.github.io/PredDBR/ for academic use.

2.7 Evaluation Indexes

In this study, to evaluate the performance of PredDBR, we use six common evaluation indexes, i.e., Sensitivity (*Sen*), Specificity (*Spe*), Precision (*Pre*), F_1-score (F_1), Accuracy (*Acc*), and the Mathew's correlation coefficient (*MCC*). The five metrics can be easily calculated according to the following formula.

$$Sen = \frac{TP}{TP + FN} \times 100 \tag{4}$$

$$Spe = \frac{TN}{TN + FP} \times 100 \tag{5}$$

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \times 100 \tag{6}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \times 100 \tag{7}$$

$$Pre = \frac{TP}{TP + FP} \tag{8}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}, \tag{9}$$

where *TP* is the number of true positive samples; *TN* is the number of true negative samples; *FP* is the number of false positive samples; *FN* is the number of false negative samples. The *Pre* and *MCC* values range from 0 to 1 and other three metrics range from 0 to 100. The higher values of these five measures indicate the better performance of DNA-binding residues prediction. By default, the values of the above threshold-dependent evaluation indexes are calculated under the threshold of 0.5.

3 Results and Discussions

3.1 Performance Comparison Between Two Different Strategy Based Cube-Format Features

In this section, we compare the efficacy of the cube-format features generated by two different strategies, i.e., SR and AVE, on three different feature sources, i.e., PSFM, PSS, and PPLBR. For the sake of description, the cube-format features generated by using the SR and AVE on the feature sources of PSFM, PSS, and PPLBR are abbreviated to SR-PSFM and AVE-PSFM, SR-PSS and AVE-PSS, and SR-PPLBR and AVE-PPLBR, respectively. Each cube-format feature is evaluated by a ten-fold cross-validation tests on PDNA-543. Here, the sliding window size *W* is set to 15, which is used in the previous study [47]. In each training phase of the cross-validation tests, the parameters of *T* and *N* in the ensemble learning scheme are both set to 1. Table 2 summarizes the results of the SR-based and AVE-based cube-format features over ten-fold cross-validation tests on PDNA-543.

TABLE 2
Performance Comparisons of Different Cube-Format Features on PDNA-543 Over Ten-Fold Cross-Validation Tests Under $W = 15$, $T = 1$, and $N = 1$

Feature	*Sen*	*Spe*	*Acc*	*Pre*	F_1	*MCC*
SR-PSFM	67.21	82.89	81.90	0.209	0.319	0.303
AVE-PSFM	68.52	77.98	77.29	0.199	0.308	0.279
SR-PSS	49.20	69.54	68.26	0.097	0.162	0.097
AVE-PSS	63.54	60.12	60.33	0.096	0.166	0.116
SR-PPLBR	53.27	80.53	79.01	0.139	0.220	0.189
AVE-PPLBR	50.52	82.43	80.65	0.145	0.225	0.192

By visiting Table 2, it is easy to find that SR-PSFM outperforms AVE-PSFM concerning the *Spe*, *Acc*, *Pre*, F_1, and *MCC* evaluation indexes. Concretely, the *Spe*, *Acc*, *Pre*, and *MCC* values of SR-PSFM are 82.89, 81.90, 0.209, 0.319, and 0.303, which are 6.30%, 5.96%, 1.00%, 3.45%, and 8.60% higher than that of AVE-PSFM, respectively, although SR-PSFM has a lower *Sen* (67.21). That is, the SR strategy is more suitable than the AVE strategy for extracting the cube-format feature on the PSFM feature source. The *Sen*, F_1, and *MCC* values of AVE-PSS are 63.54, 0.166, and 0.116, which are 29.15%, 2.41%, and 19.59% higher than that of SR-PSS, respectively, although AVE-PSS achieves a slightly lower *Spe* (60.12), *Acc* (60.33), and *Pre* (0.096) values. The *Spe*, *Acc*, *Pre*, F_1, and *MCC* values of AVE-PPLBR are 2.36%, 2.08%, 4.32%, 2.22%, and 1.59% higher than that of SR-PPLBR, respectively. These comparisons have demonstrated that the AVE strategy is a better choice than the SR strategy to generate the cube-format feature on the PSS and PPLBR feature sources. Hence, in all the subsequent experiments, we use the features of SR-PSFM, AVE-PSS, and AVE-PPLBR to encode the discriminative information of each target residue.

3.2 Performance Comparison Between Matrix-Format and Cube-Format Features

In order to evaluate the performance of the cube-format feature, we compare it to the matrix-format feature over ten-fold cross-validation tests on PDNA-543 under the parameters of *T* and *N* are both set to 1. Concretely, in this section, we employ the SR-PSFM, AVE-PSS, and AVE-PPLBR features described in Section 3.1 to generate the final cube-format feature (denoted as CFF) via serially combining their first dimensions. The corresponding final matrix-format feature (abbreviated to MFF) is extracted via two steps: (1) using the sliding window technique with size 15 to generate the PSFM-based, PSS-based, and PPLBR-based matrix-format features; (2) serially combining the three matrix-format features. Table 3 lists the performance of MFF and CFF on the training dataset PDNA-543. Note that, due to the 2D-CNN framework described in Section 2.4 is not suitable to directly learn the prediction model on MFF, the results of MFF are obtained by using a newly designed 1D convolutional neural network framework (see detail in Supporting Text S2), which is similar to the framework of the 2D-CNN.

From Table 3, it is easy to observe that CFF outperforms MFF in terms of all the six evaluation indexes, i.e., *Sen*, *Spe*, *Acc*, *Pre*, F_1, and *MCC*. Concretely, the *Sen*, *Spe*, *Acc*, *Pre*, F_1,

TABLE 3
Performance Comparison Between Matrix-Format and Cube-Format Features on PDNA-543 Over Ten-Fold Cross-Validation Under $W = 15$, $T = 1$, and $N = 1$

Feature	*Sen*	*Spe*	*Acc*	*Pre*	F_1	*MCC*
MFF*	71.31	73.90	74.15	0.188	0.296	0.290
CFF#	71.32	81.11	80.37	0.235	0.353	0.329

**MFF means matrix-format feature.*
#CFF means cube-format feature.

and *MCC* values of CFF are 71.32, 81.11, 80.37, 0.235, 0.353, and 0.329, which are 0.01%, 9.75%, 8.39%, 16.15%, 4.70%, and 13.45% higher than that of MFF. The comparison result demonstrates that the cube-format feature could dig out more available information than MFF for improving the performance of protein-DNA binding residue prediction.

3.3 Choosing the Value of the Sliding Window Size

As described in the above section, CFF is a good choice to encode each target residue. However, the size of CFF depends on the sliding window size (*W*). Although $W = 15$ has been demonstrated to be useful one for extracting the matrix-format feature in the previous study [47], it is not clear whether it is an optimal choose to extract the cube-format feature. In this section, we try to empirically choose an appropriate value for *W* to generate CFF. Concretely, we evaluate the performance variations of *Pre* and *MCC* on the training dataset PDNA-543 over ten-fold cross-validation tests by gradually varying the value of *W* from 9 to 21 with a step size of 2. Again, the parameters of *T* and *N* in the ensemble learning scheme are both set to 1.

Fig. 4 shows the performance of *MCC* and *Pre* of different sliding window sizes. By visiting Fig. 4, it is easily found that $W = 17$ outperforms $W = 15$, which is the second-best one concerning both *Pre* and *MCC*. The *Pre* and *MCC* values of $W = 17$ are 0.244 and 0.331, which are higher than that of $W = 15$ ($Pre = 0.235$ and $MCC = 0.329$). In addition, Fig. 4 clearly demonstrates that the overall trend exhibited by the values of *MCC* tends to increase with *W* when $W < 17$. When $W > 17$, the values of *MCC* tend to decrease. The similar trend could be also found on the values of *Pre*. These results demonstrate that $W = 17$ is a suitable choice for extracting CFF in this study. Hence, in all the subsequent experiments, the value of *W* is set to 17.

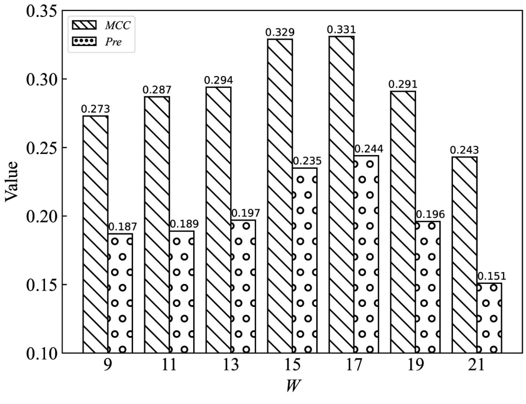

Fig. 4. Comparison of *MCC* and *Pre* of different sliding window sizes on PDNA-543 over ten-fold cross-validation tests under $T = 1$ and $N = 1$.

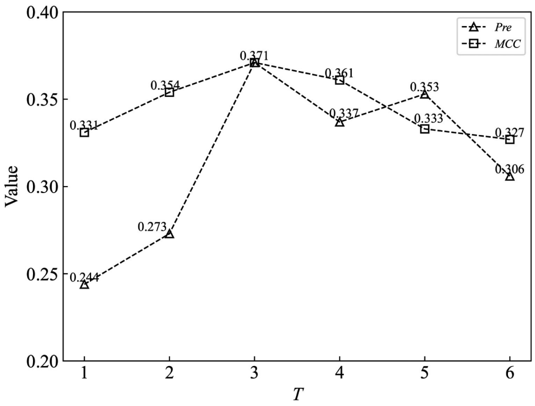

Fig. 5. The performance of *MCC* and *Pre* versus *T* values.

3.4 Selecting the Imbalance Degree Parameter

In this section, we will tune the imbalance degree parameter *T* to an appropriate value, which is used in the ensemble learning scheme. Specifically, we evaluate the *Pre* and *MCC* performance variations of a classifier learned by the ensemble learning scheme on the training dataset PDNA-543 over ten-fold cross-validation tests via gradually varying the value of *T* from 1 to 6 with a step size of 1. Here, the parameter of *N*, i.e., the number of the base classifiers, is set to 1.

Fig. 5 shows the performance variation curves of *Pre* and *MCC* versus *T*. From Fig. 5, it is clearly observed that the overall trend exhibited by the values of *MCC* tends to increase with *T* when $T \leq 3$. When $3 < T \leq 6$, the values of *MCC* tend to decrease. It can be expected that the performance of *MCC* will further deteriorate with the increase in *T*, when $T > 6$, because the severity of imbalance will become increasingly serious with the increase in *T*. In addition, it is easy to find the similar trend concerning the *Pre* values. Thus, in all subsequent experiments, the value of *T* is set to 3 to train each base classifier in the ensemble learning scheme.

3.5 Selecting the Number of the Base Classifiers

To select a suitable number of the base classifier (*N*) in the ensemble learning scheme, in this section, we evaluate the *Pre* and *MCC* performance variations on the training dataset PDNA-543 over ten-fold cross-validation tests by gradually varying the value of *N* from 1 to 16 with a step size of 1. Fig. 6 plots the performance variation curves of *Pre* and *MCC* versus *N*.

By visiting Fig. 6, it is easy to find that $N = 8$ is the best choice in this study. Concretely, the *Pre* and *MCC* values of $N = 8$ are 0.450 and 0.413, which are 3.93% and 3.25% higher than that of the second-best choice, i.e., $N = 15$, respectively. Hence, the value of *N* is finally set to 8 in this study.

3.6 Comparisons With Other Protein-DNA Binding Residue Prediction Methods

In this section, to evaluate the performance of the proposed PredDBR, we compare it with the existing DNA-binding residue methods, including BindN [8], DP-Bind [9], BindN-rf [10], BindN+ [25], DNABind [29], TargetDNA [32], iProDNA-CapsNet [34], DBS-PRED [48], DNABindR [49], DISIS [50], MetaDBSite [28], DNAPred [36], TargetS [38], and EC-RUS [51].

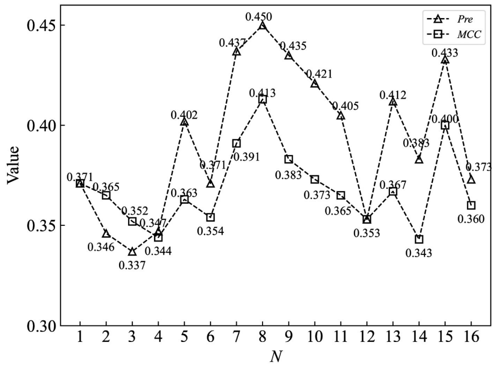

Fig. 6. The performance of *MCC* and *Pre* versus *N* values.

3.6.1 Performance Comparison on PDNA-543

Table 4 shows the performance comparisons of PredDBR and two control methods, i.e., TargetDNA [32] and DNAPred [36], which are two of the most recently released methods, on PDNA-543 over ten-fold cross-validation tests. For the purpose of fair comparison, the performance of PredDBR is evaluated under two different thresholds, as do in TargetDNA and DNAPred. One is the threshold that makes *Sen* ≈ *Spe*, and the other is the threshold that makes *Spe* ≈ 95%.

From Table 4, it can be observed that PredDBR outperforms other two predictors, i.e., TargetDNA and DNAPred, concerning the *Pre* and *MCC* evaluation indexes under both of the considered thresholds. Concretely, under the threshold that makes *Sen* ≈ *Spe*, the *Pre*, F_1, and *MCC* values of PredDBR are 0.233, 0.358, and 0.338, which are 21.35%, 14.25%, and 11.18% higher than that of TargetDNA, respectively. Under the threshold that makes *Spe* ≈ 95%, the *Pre*, F_1, and *MCC* values of PredDBR are 0.451, 0.458, and 0.409, which are 23.56%, 16.16%, and 20.65% higher than that of TargetDNA, respectively. Compared to DNAPred, PredDBR improves the value of *MCC* by 6.29% and 9.65% under the thresholds that make *Sen* ≈ *Spe* and *Spe* ≈ 95%, respectively. The performance of PredDBR is also evaluated under the threshold of 0.5 on PDNA-543 and the values of *Sen*, *Spe*, *Acc*, *Pre*, F_1, and *MCC* of PredDBR are 45.35, 95.50, 91.43, 0.471, 0.462, and 0.415, respectively.

TABLE 4
Performance Comparisons Between PredDBR, TargetDNA, and DNAPred on PDNA-543 Over Ten-Fold Cross-Validation Tests

Method	*Sen*	*Spe*	*Acc*	*Pre*	F_1	*MCC*
TargetDNA$_{(Sen \approx Spe)}$ [a]	76.98	77.05	77.04	0.192	0.307	0.304
DNAPred$_{(Sen \approx Spe)}$ [b]	77.10	78.50	78.40	-	-	0.318
PredDBR$_{(Sen \approx Spe)}$	77.64	77.41	77.43	0.233	0.358	0.338
TargetDNA$_{(Spe \approx 95\%)}$ [a]	40.60	95.00	91.40	0.365	0.384	0.339
DNAPred$_{(Spe \approx 95\%)}$ [b]	44.90	95.00	91.70	-	-	0.373
PredDBR$_{(Spe \approx 95\%)}$	46.49	95.00	91.06	0.451	0.458	0.409
PredDBR	45.35	95.50	91.43	0.471	0.462	0.415

[a] *Results excerpted from TargetDNA [32];* [b] *Results excerpted from DNAPred [36]; "*Sen ≈ Spe*" means the threshold that makes* Sen ≈ Spe; *"*Spe ≈ *95%" means the threshold that makes* Spe ≈ *95%. '-' means the value is not given.*

TABLE 5
Performance Comparisons Between PredDBR and the Existing Methods on PDNA-316 Over Ten-Fold Cross-Validation Tests

Predictor	*Sen*	*Spe*	*Acc*	*Pre*	*MCC*
DBS-PRED [a]	53.00	76.00	75.00	-	0.170
BindN [a]	54.00	80.00	78.00	-	0.210
DNABindR [a]	66.00	74.00	73.00	-	0.230
DISIS [a]	19.00	98.00	92.00	-	0.250
DP-Bind [a]	69.00	79.00	78.00	-	0.290
BindN-rf [a]	67.00	83.00	82.00	-	0.320
MetaDBSite [a]	77.00	77.00	77.00	-	0.320
TargetDNA$_{(Sen \approx Spe)}$ [a]	77.96	78.03	78.02	-	0.339
TargetDNA$_{(Spe \approx 95\%)}$ [a]	43.02	95.00	90.99	-	0.375
DNAPred$_{(Sen \approx Spe)}$ [b]	80.00	79.90	79.90	-	0.370
DNAPred$_{(Spe \approx 95\%)}$ [b]	52.10	95.10	91.80	-	0.452
PredDBR$_{(Sen \approx Spe)}$	81.54	80.71	80.78	0.274	0.398
PredDBR$_{(Spe \approx 95\%)}$	56.06	95.34	92.12	0.519	0.497
PredDBR	53.08	95.82	92.30	0.532	0.489

[a] *Results excerpted from TargetDNA [32];* [b] *Results excerpted from DNAPred [36]. "*Sen ≈ Spe*" means the threshold that makes* Sen ≈ Spe; *"*Spe ≈ *95%" means the threshold that makes* Spe ≈ *95%. '-' means the value is not given.*

3.6.2 Performance Comparison on PDNA-316

In this section, the performance of PredDBR is evaluated on PDNA-316 over ten-fold cross-validation tests, comparing with other state-of-the-art DNA-binding residue prediction methods, i.e., DBS-PRED [48], BindN [8], DNABindR [49], DISIS [50], DP-Bind [9], BindN-rf [10], MetaDBSite [28], TargetDNA [32], and DNAPred [36]. The detail results are shown in Table 5.

By visiting Table 5, it is easy to find that PredDBR enjoys the better performance than other nine predictors in terms of *MCC*. Compared to the second-best method DNAPred, under the threshold that makes *Sen* ≈ *Spe*, the values of *Sen*, *Spe*, *Acc*, and *MCC* of PredDBR are improved by 1.93%, 1.01%, 1.10%, and 7.56%, respectively; under the threshold that makes *Spe* ≈ 95%, the values of *Sen*, *Spe*, *Acc*, and *MCC* of PredDBR are 56.06, 95.34, 92.12, and 0.497, which are 7.60%, 0.25%, 0.35%, and 9.96% higher than that of DNAPred. Furthermore, under the threshold of 0.5, the *Sen*, *Spe*, *Acc*, *Pre*, and *MCC* values of PredDBR are 53.08, 95.82, 92.30, 0.532, and 0.489, respectively. In addition, the *Sen*, *Spe*, *Acc*, and *MCC* values of PredDBR, which are evaluated under the threshold that makes *Sen* ≈ *Spe*, are higher than the values evaluated for DBS-PRED, BindN, DNABindR, DP-Bind, MetaDBSite, and TargetDNA$_{(Sen \approx Spe)}$. Taking MetaDBSite as an example, PredDBR$_{(Sen \approx Spe)}$ achieves the improvements of 5.90%, 4.82%, 4.91%, and 24.38% on *Sen*, *Spe*, *Acc*, and *MCC*, respectively. It is also easy to notice that DISIS obtains the highest *Spe* value (98.00) but the lowest value of *Sen* (19.00). That is, DISIS predicts too many false negatives.

3.6.3 Performance Comparison on PDNA-335

Table 6 demonstrates the performance of PredDBR, TargetS [38], EC-RUS [51], and DNAPred [36] on PDNA-335 over five-fold cross-validation tests. The *Sen*, *Spe*, *Acc*, *Pre*, and *MCC* values of PredDBR are 42.59, 95.34, 90.96, 0.453, and

TABLE 6
Performance Comparisons Between EC-RUS, TargetS, DNAPred, and PredDBR on PDNA-335 Over Five-Fold Cross-Validation

Predictor	*Sen*	*Spe*	*Acc*	*Pre*	*MCC*
EC-RUS[a]	48.70	95.10	92.60	-	0.378
TargetS[b]	41.70	94.50	89.90	-	0.362
DNAPred[c]	54.30	91.70	88.60	-	0.390
PredDBR	42.59	95.34	90.96	0.453	0.390

[a]*Results excerpted from EC-RUS;* [b]*Results excerpted from TargetS;* [c]*Results excerpted from DNAPred. '-' means the value is not given.*

0.390, respectively. The *MCC* value of PredDBR is equal to that of DNAPred, 7.73% higher than that of TargetS, and 3.17% higher than that of EC-RUS.

3.6.4 Performance Comparison on PDNA-52

To evaluate the generation ability of the proposed PredDBR, its performance is calculated on the independent validation dataset PDNA-52, comparing to five control methods, i.e., DNABR [26], MetaDBSite [28], TargetS [38], DNAPred [36], and COACH [52]. COACH is a general-purpose protein-ligand binding residues predictor. To facilitate the calculation of evaluation indexes and performance comparison, the protein-ligand binding residues predicted by COACH are all regarded as DNA-binding residues. Note that, in this section, the prediction model of PredDBR is trained on the training dataset PDNA-335. Table 7 lists the detail results.

By visiting Table 7, we can find that PredDBR gains the highest values of *Acc* (93.46) and *MCC* (0.451). Compared to COACH on the independent dataset PDNA-52, which obtains a higher *Sen* value (59.91), the improvements of 2.49%, 2.04%, 22.03%, and 6.89% are achieved by PredDBR on *Spe*, *Acc*, *Pre*, and *MCC*, respectively. Compared to DNAPred, which is the third-best method, the improvements of 4.00%, 1.00%, 1.04%, and 11.40% are achieved by PredDBR on the evaluation indexes of *Sen*, *Spe*, *Acc*, and *MCC*, respectively. Compared to TargetS, PredDBR obtains the improvements of 30.39%, 0.17%, and 19.63% on *Sen*, *Acc*, and *MCC*, respectively, although PredDBR has a lower *Spe* value (95.83). Compared to MetaDBSite, PredDBR achieves higher *Spe*, *Acc*, and *MCC* values and a lower *Sen* value. Compared to DNABR, PredDBR gains the improvements of 32.31%, 9.77%, 10.47%, and 143.78% on *Sen*, *Spe*, *Acc*, and *MCC*, respectively. In addition, PredDBR achieves a *Pre* value of 0.454.

TABLE 7
Performance Comparisons of PredDBR, MetaDBSite, TargetS, and DNAPred on the Independent Validation Dataset PDNA-52

Predictor	*Sen*	*Spe*	*Acc*	*Pre*	*MCC*
DNABR[a]	40.70	87.30	84.60	-	0.185
MetaDBSite[a]	58.00	76.40	75.20	-	0.192
TargetS[a]	41.30	96.50	93.30	-	0.377
DNAPred[b]	51.80	94.90	92.50	-	0.405
COACH[c]	59.91	93.45	91.55	0.354	0.420
PredDBR	53.85	95.83	93.46	0.454	0.451

[a]*Results excerpted from TargetS [38];* [b]*Results excerpted from DNAPred [36]. '-' means the value is not given;* [c]*results are computed by COACH program.*

TABLE 8
Performance Comparisons Between PredDBR and Other Existing Predictors on PNDA-41 Under Independent Validation

Predictor	*Sen*	*Spe*	*Acc*	*Pre*	*MCC*
BindN [a]	45.64	80.90	79.15	0.111	0.143
ProteDNA [a]	4.77	99.84	95.11	0.603	0.160
BindN+ $_{(Spe \approx 95\%)}$ [a]	24.11	95.11	91.58	0.205	0.178
BindN+ $_{(Spe \approx 85\%)}$ [a]	50.81	85.41	83.69	0.154	0.213
MetaDBSite [a]	34.20	93.35	90.41	0.212	0.221
DP-Bind [a]	61.72	82.43	81.40	0.155	0.241
DNABind [a]	70.16	80.28	79.78	0.157	0.264
TargetDNA$_{(Sen \approx Spe)}$ [a]	60.22	85.79	84.52	0.182	0.269
TargetDNA$_{(Spe \approx 95\%)}$ [a]	45.50	93.27	90.89	0.261	0.300
iProDNA-CapsNet$_{(Sen \approx Spe)}$ [b]	75.36	75.34	75.34	0.135	0.245
iProDNA-CapsNet$_{(Spe \approx 95\%)}$ [b]	42.17	94.93	92.38	0.298	0.315
DNAPred$_{(Sen \approx Spe)}$ [c]	76.10	76.70	76.10	-	0.260
DNAPred$_{(Spe \approx 95\%)}$ [c]	44.70	94.90	92.40	-	0.337
COACH [d]	46.19	95.10	92.67	0.330	0.352
PredDBR$_{(Sen \approx Spe)}$	76.43	75.80	75.83	0.131	0.264
PredDBR$_{(Spe \approx 95\%)}$	43.05	95.77	93.14	0.348	0.351
PredDBR	39.10	96.79	93.93	0.389	0.359

[a]*Results excerpted from TargetDNA [32];* [b]*Results excerpted from iProDNA-CapsNet [34];* [c]*Results excerpted from DNAPred [36];* [d]*results are computed by COACH program.* "Sen ≈ Spe" *means the threshold that makes* Sen ≈ Spe *on PDNA-543;* "Spe ≈ 95%" *means the threshold that makes* Spe ≈ 95% *on PDNA-543. '-' means the value is not given.*

3.6.5 Performance Comparison on PDNA-41

To further evaluate the generation ability of PredDBR, its performance is also evaluated on the independent validation dataset PDNA-41, comparing to ten control methods, i.e., BindN [8], ProteDNA [31], BindN+ [25], MetaDBSite [28], DP-Bind [9], DNABind [29], TargetDNA [32], iProDNA-CapsNet [34], DNAPred [36], and COACH [52]. Note that, in this section, the prediction model of PredDBR is trained on the training dataset PDNA-543. The detail results are demonstrated in Table 8.

From Table 8, it is easily found that the values of 39.10, 96.79, 93.93, 0.389, and 0.359 are obtained by PredDBR on the evaluation indexes of *Sen*, *Spe*, *Acc*, *Pre*, and *MCC* under the threshold of 0.5, respectively. Although COACH obtains a higher *Sen* value (46.19) on the independent dataset PDNA-41, PredDBR achieves the improvements of 1.75%, 1.34%, 15.17% and 1.95% on *Spe*, *Acc*, *Pre* and *MCC*, respectively. Compared to the third-best method, i.e., DNAPred, under Spe ≈ 95%, where the corresponding threshold is calculated on the training dataset PDNA-543, the *Spe*, *Acc*, and *MCC* values of PredDBR are 95.77, 93.14, and 0.351, which are 0.92%, 0.80%, and 4.15% higher than that of DNAPred, respectively, although PredDBR achieves a lower *Sen* value (43.05); under *Sen* ≈ . *Spe*, PredDBR also achieves a comparable performance with a slightly higher *MCC* (0.264). Compared to iProDNA-CapsNet, under *Spe* ≈ 95%, the improvements of 2.09%, 0.88%, 0.82%, 16.78%, and 11.43% are achieved by PredDBR on the evaluation indexes of *Sen*, *Spe*, *Acc*, *Pre*, and *MCC*, respectively; under *Sen* ≈ *Spe*, the *MCC* value of PredDBR (0.264) is also higher than that of iProDNA-CapsNet (0.245). Compared to TargetDNA, under *Spe* ≈ 95%, PredDBR obtains

2.68%, 2.48%, 33.33%, and 17.00% increases in *Spe*, *Acc*, *Pre*, and *MCC*, respectively, although PredDBR achieves a lower *Sen* value; under *Sen* $\approx$ *Spe*, PredDBR gains a comparable performance. Compared to DNABind, PredDBR achieves the higher *Pre* (0.389) and *MCC* (0.359) values. In addition, it is also easy to find that PredDBR outperforms BindN, ProteDNA, BindN+, MetaDBSite, and DP-Bind. Taking DP-Bind as an example, the *Pre* and *MCC* values of PredDBR (*Pre* $=$ 0.389 and *MCC* $=$ 0.359) increased by 150.97% and 48.96%, respectively.

4 Conclusion

In this study, we have developed and implemented a new method, called PredDBR, to predict protein-DNA binding residues from protein sequence information. Benchmarked results have demonstrated the efficacy of the proposed PredDBR via comparing to several state-of-the-art sequence-based methods on five protein-DNA binding residue datasets. The superior performance of PredDBR is mainly attributed to the strong capabilities of 2D-CNN and bagging strategy, which can effectively dig out the available information embedded in the cube-format feature and deal with the imbalanced learning problem, respectively. The web-server is freely available at https://jun-csbio.github.io/PredDBR/ for academic use.

In the future work, to further improve the DNA-binding residue prediction performance of PredDBR, three directions are considered: (1) extracting more discriminative feature sources from the results of research problems similar to protein-DNA binding residue prediction, such as transcription factor binding site prediction [53], [54], [55], association mapping from DNA methylation to disease [56], protein-protein interaction prediction [57], [58], protein crystallization prediction [59], protein-vitamin binding residues prediction [60], and sequence-based protein structure/function analysis [61]; (2) developing more useful strategies to extract more discriminative cube-format feature information; (3) using the deep learning algorithms [62], [63], [64], [65] to train the prediction model on a newly collected big dataset. In addition, RNA is structurally similar to DNA. Learning from the existing RNA-protein or DNA-protein binding prediction method [66], [67], we will extend the prediction model of PredDBR to predict protein-RNA binding residues. Except RNA, although the structures of ATP and Ca^{2+} are different from that of DNA, we will also employ the transfer learning algorithms to predict the binding residues of them.

References

[1] 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 . Bioinformat .*, vol. 4, no. 1, pp. 117–125, Fourth Quarter 2007.

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

[3] K. Wong, Y. Li, C. Peng, and H. Wong, "A comparison study for DNA motif modeling on protein binding microarray," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 13, no. 2, pp. 261–271, Mar. 2016.

[4] 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.

[5] S. Jones, P. van Heyningen, H. M. Berman, and J. M. Thornton, "Protein-DNA interactions: A structural analysis," *J. Mol. Biol.*, vol. 287, no. 5, pp. 877–896, Apr. 1999.

[6] Y. Mandelgutfreund and H. Margalit, "Quantitative parameters for amino acid-base interaction: Implications for prediction of protein-DNA binding sites," *Nucleic Acids Res.*, vol. 26, no. 10, pp. 2306–2312, 1998.

[7] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, "CATH – a hierarchic classification of protein domain structures," *Structure*, vol. 5, no. 8, pp. 1093–1109, 1997.

[8] 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, pp. 243–248, 2006.

[9] 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," *Bioinformatics*, vol. 23, no. 5, pp. 634–636, 2007.

[10] L. Wang, M. Q. Yang, and J. Y. Yang, "Prediction of DNA-binding residues from protein sequence information using random forests," *BMC Genomic.*, vol. 10, no. 1, pp. 1–9, 2009.

[11] S. Jones, J. 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.

[12] Y. Tsuchiya, K. Kinoshita, and H. Nakamura, "Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces," *Proteins Struct. Function Bioinformat.*, vol. 55, no. 4, pp. 885–894, 2004.

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

[14] I. B. Kuznetsov, Z. Gou, R. Li, and S. Hwang, "Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins," *Proteins Struct. Function Bioinformat.*, vol. 64, no. 1, pp. 19–27, 2006.

[15] L. Tao, Q. Z. Li, L. Shuai, G. L. Fan, Y. C. Zuo, and P. Yong, "PreDNA: Accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information," *Bioinformatics*, vol. 29, no. 6, pp. 678–685, 2013.

[16] Y. Tsuchiya, K. Kinoshita, and H. Nakamura, "PreDs: A server for predicting dsDNA-binding site on protein molecular surfaces," *Bioinformatics*, vol. 21, no. 8, pp. 1721–1723, 2005.

[17] P. Ozbek, S. Soner, B. Erman, and T. Haliloglu, "DNABINDPROT: Fluctuation-based predictor of DNA-binding residues within a network of interacting residues," *Nucleic Acids Res.*, vol. 38, pp. 417–423, 2010.

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

[19] M. Gao, and J. Skolnick, "DBD-Hunter: A knowledge-based method for the prediction of DNA–protein interactions," *Nucleic Acids Res.*, vol. 36, no. 12, pp. 3978–3992, 2008.

[20] C. Y. Chi, J. D. Wright, and L. Carmay, "DR_bind: A web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry," *Nucleic Acids Res.*, vol. 40, no. 1, pp. 249–256, 2012.

[21] C. Cortes, and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[22] J. Yang and Y. Zhang, "I-TASSER server: New development for protein structure and function predictions," *Nucleic Acids Res.*, vol. 43, no. 1, pp. 174–181, 2015.

[23] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, and D. Baker, "Protein structure prediction using ROSETTA," *Methods Enzymol.*, vol. 383, pp. 66–93, 2004.

[24] A. W. Senior *et al.*, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, Jan. 2020.

[25] L. Wang, C. Huang, M. Q. and J. Y. Yang "BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features," *BMC Syst. Biol.*, vol. 4, no. 1, 2010, Art. no. S3.

[26] X. Ma, J. Guo, H. Liu, J. Xie, and X. Sun, "Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 9, no. 6, pp. 1766–1775, Nov./Dec. 2012.

[27] R. Xu *et al.*, "enDNA-Prot: Identification of DNA-binding proteins by applying ensemble learning," *BioMed. Res. Int.*, vol. 2014, pp. 294279–294279, 2014.

[28] 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. 1, pp. 1–7, 2011.

[29] R. Liu and J. Hu, "DNABind: A hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning- and template-based approaches," *Proteins Struct. Function Bioinformat.*, vol. 81, no. 11, pp. 1885–1899, 2013.

[30] C. Shen, Y. Ding, J. Tang, J. Song, and F. Guo, "Identification of DNA–protein binding sites through multi-scale local average blocks on sequence information," *Molecules*, vol. 22, no. 12, 2017, Art. no. 2079.

[31] W. Chu, Y. Huang, C. Huang, Y. Cheng, C. Huang, and Y. Oyang, "ProteDNA: A sequence-based predictor of sequence-specific DNA-binding residues in transcription factors," *Nucleic Acids Res.*, vol. 37, pp. 396–401, 2009.

[32] J. Hu, Y. Li, M. Zhang, X. Yang, H. Shen, and D. Yu, "Predicting Protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 14, no. 6, pp. 1389–1398, 2017.

[33] J. Zhou, Q. Lu, R. Xu, Y. He, and H. Wang, "EL_PSSM-RT: DNA-binding residue prediction by integrating ensemble learning with PSSM relation transformation," *BMC Bioinformat.*, vol. 18, no. 1, 2017, Art. no. 379.

[34] B. P. Nguyen, Q. H. Nguyen, G. N. Doan-Ngoc, T. H. Nguyen-Vo, and S. Rahardja, "iProDNA-CapsNet: Identifying protein-DNA binding residues using capsule neural networks," *BMC Bioinformat.*, vol. 20, no. Suppl 23, 2019, Art. no. 634.

[35] J. Zhou, Q. Lu, R. Xu, L. Gui, and H. Wang, "CNNsite: Prediction of DNA-binding residues in proteins using convolutional neural network with sequence features," in *Proc. IEEE Int. Conf. Bioinformat . Biomed .*, 2016, pp. 78–85,.

[36] Y. Zhu, J. Hu, X. Song, and D. Yu, "DNAPred: Accurate identification of dna-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines," *J. Chem. Inf. Model.*, vol. 59, no. 6, pp. 3057–3071, 2019.

[37] Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[38] D. Yu, J. Hu, J. Yang, H. Shen, J. Tang, and J. Yang, "Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering," *IEEE/ACM Trans. Computat. Biol. Bioinformat.*, vol. 10, no. 4, pp. 994–1008, 2013.

[39] P. W. Rose *et al.*, "The RCSB protein data bank: Views of structural biology for basic and applied research and education," *Nuclc Acids Res.*, vol. 43, no. D1, pp. 345–356, 2015.

[40] J. Yang, A. Roy, and Y. Zhang, "BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions," *Nucleic Acids Res.*, vol. 41, pp. 1096–1103, Jan. 2013.

[41] M. Remmert, A. Biegert, A. Hauser, and J. Söding, "HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nat. Methods*, vol. 9, no. 2, pp. 173–175, 2012.

[42] M. Milot, von den D. Lars, G. Clovis, M. J. Martin, S. Johannes, and S. Martin, "Uniclust databases of clustered and deeply annotated protein sequences and alignments," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D170–D176, 2017.

[43] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Molecul. Biol.*, vol. 292, no. 2, pp. 195–202, 1999.

[44] J. Huab, L. Raoa, X. Fana, and G. Zhang, "Identification of ligand-binding residues using protein sequence profile alignment and query-specific support vector machine model," *Anal. Biochem.*, vol. 604, 2020, Art. no. 113799.

[45] H. He, and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl . Data Eng .*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[46] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[47] Y. Zhang, S. Qiao, S. Ji, N. Han, D. Liu, and J. Zhou, "Identification of DNA–protein binding sites by bootstrap multiple convolutional neural networks on sequence information," *Eng. Appl. Artif. Intell.*, vol. 79, pp. 58–66, 2019.

[48] 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," *Bioinformatics*, vol. 20, no. 4, pp. 477–486, Mar. 2004.

[49] 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 Bioinformat.*, vol. 7, 2006, Art. no. 262.

[50] Y. Ofran, V. Mysore, and B. Rost, "Prediction of DNA-binding residues from sequence," *Bioinformatics*, vol. 23, no. 13, pp. i347–i353, 2007.

[51] Y. Ding, J. Tang, and F. Guo, "Identification of protein–ligand binding sites by sequence information and ensemble classifier," *J. Chem. Inf. Model.*, vol. 57, no. 12, pp. 3149–3161, 2017.

[52] Yang, J., A. Roy, and Y. Zhang, "Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment," *Bioinformatics*, vol. 29, no. 20, pp. 2588–2595, 2013.

[53] Q. Zhang, L. Zhu, W. Bao, and D. S. Huang, "Weakly-supervised convolutional neural network architecture for predicting protein-DNA binding," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 17, no. 2, pp. 679–689, Mar./Apr. 2020.

[54] Q. Zhang, L. Zhu, and D. S. Huang, "High-Order convolutional neural network architecture for predicting DNA-protein binding sites," *IEEE/ACM Trans. Computat. Biol. Bioinformat.*, vol. 16, no. 4, pp. 1184–1192, Jul. /Aug. 2019.

[55] Q. Zhang, Z. Shen, and D. S. Huang, "Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network," *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 8484.

[56] L. Yuan and D. S. Huang, "A network-guided association mapping approach from DNA methylation to disease," *Scientific Rep.*, vol. 9, no. 1, 2019, Art. no. 5601.

[57] J. F. Xia, X. M. Zhao, and D. S. Huang, "Predicting protein-protein interactions from protein sequences using meta predictor," *Amino Acids*, vol. 39, no. 5, pp. 1595–1599, Nov. 2010.

[58] Z. S. Wei, J. Y. Yang, H. B. Shen, and D. J. Yu, "A cascade random forests algorithm for predicting protein-protein interaction sites," *IEEE Trans . Nanobiosci.*, vol. 14, no. 7, pp. 746–760, Oct. 2015.

[59] J. Hu, K. Han, Y. Li, J. Y. Yang, H. B. Shen, and D. J. Yu, "TargetCrys: Protein crystallization prediction by fusing multi-view features with two-layered SVM," *Amino Acids*, vol. 48, no. 11, pp. 2533–2547, Nov. 2016.

[60] D. J. Yu, J. Hu, H. Yan, X. B. Yang, J. Y. Yang, and H. B. Shen, "Enhancing protein-vitamin binding residues prediction by multiple heterogeneous subspace SVMs ensemble," *BMC Bioinformat.*, vol. 15, no. 1, Sep. 2014, Art. no. 297.

[61] S. P. Deng, and D. S. Huang, "SFAPS: An r package for structure/function analysis of protein sequences based on informational spectrum method," *Methods*, vol. 69, no. 3, pp. 207–12, Oct. 2014.

[62] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, 2019.

[63] J. M. Stokes *et al.*, "A deep learning approach to antibiotic discovery," *Cell*, vol. 180, no. 4, pp. 688–702, 2020.

[64] E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen, "Deep learning for cellular image analysis," *Nat. Methods*, vol. 16, no. 12, pp. 1233–1246, 2019.

[65] Q. Zhang, Z. Shen, and D. S. Huang, "Predicting in-vitro transcription factor binding sites using DNA sequence + shape," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 18, no. 2, pp. 667–676, Mar./Apr. 2021.

[66] Z. Shen, S. P. Deng, and D. S. Huang, "Capsule network for predicting rna-protein binding preferences using hybrid feature," *IEEE/ACM Trans. Computat. Biol. Bioinformat.*, vol. 17, no. 5, pp. 1483–1492, Sep./Oct. 2020.

[67] Z. Shen, Q. Zhang, K. Han, and D. S. Huang, "A deep learning model for rna-protein binding preference prediction based on hierarchical LSTM and attention network," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, early access, Jul. 7, 2020, doi: 10.1109/TCBB.2020.3007544.

Jun Hu received the BS degree in computer science from Anhui Normal University in 2011. From 2011 to 2018, he was a PhD student with the School of Computer Science and Engineering, Nanjing University of Science and Technology and a member of Pattern Recognition and Bioinformatics Group, led by professor Dong-Jun Yu. From 2016 to 2017, he was a visiting student with the University of Michigan, Ann Arbor, MI, USA He is currently a teacher with the College of Information Engineering, Zhejiang University of Technology. His research interests include pattern recognition, data mining, and bioinformatics.

Yan-Song Bai received the BS degree in automation from the Hubei University of Arts and Science, China, in 2019. He is currently a postgraduate student with the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China, led by Jun Hu. His research interests include pattern recognition, data mining, and bioinformatics.

Lin-Lin Zheng received the BS degree in electrical engineering and automation from Huainan Normal University, China, in 2019. She is currently a postgraduate student with the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. Her research interests include pattern recognition, data mining, and bioinformatics.

Ning-Xin Jia received the BS degree in automation from Qingdao University of Science and Technology, China, in 2018. She is currently a postgraduate student with the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China, led by Jun Hu. Her research interests include pattern recognition, data mining, and bioinformatics.

Dong-Jun Yu received the PhD degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technologyin 2003. In 2008, was an academic visitor with the Department of Computer, University of York, U.K. He also visited the Department of Computational Medicine, University of Michigan, Ann Arbor, in 2016. He is currently a full professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. He has authored or coauthored more than 50 scientific papers in pattern recognition and bioinformatics. His research interests include pattern recognition, machine learning, and bioinformatics. He is a senior member of China Computer Federation and a senior member of China Association of Artificial Intelligence.

Gui-Jun Zhang received the PhD degree in control theory and control engineering from Shanghai Jiaotong University, Shanghai, China, in 2004. He is currently a professor with the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His current research interests include intelligent information processing, optimization theory and algorithm design, and bioinformatics.