Engineering Applications of Artificial Intelligence 79 (2019) 58–66

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Identification of DNA–protein binding sites by bootstrap multiple convolutional neural networks on sequence information☆

Yongqing Zhang [a,b], Shaojie Qiao [c,*], Shengjie Ji [a], Nan Han [d], Dingxiang Liu [c], Jiliu Zhou [a]

[a] School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China
[b] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
[c] School of Cybersecurity, Chengdu University of Information Technology, Chengdu 610225, China
[d] School of Management, Chengdu University of Information Technology, Chengdu 610103, China

ARTICLE INFO

Keywords:
DNA–protein binding sites
Bootstrap method
Convolutional neural networks
ADASYN sampling

ABSTRACT

Identification of DNA–protein binding sites in protein sequence plays an essential role in a wide variety of biological processes. In particular, there are huge volumes of protein sequences accumulated in the post-genomic era. In this study, we propose a new prediction approach appropriate for imbalanced DNA–protein binding sites data. Specifically, motivated by the imbalanced problem of the distribution of DNA–protein binding and non-binding sites, we employ the Adaptive Synthetic Sampling (ADASYN) approach to over-sample the positive data and Bootstrap strategy to under-sample the negative data to balance the number of the binding and non-binding samples. Furthermore, we employ the three types of features: the position specific scoring matrix, one-hot encoding and predicted solvent accessibility, to encode the sequence-based feature of each protein residue. In addition, we design an ensemble convolutional neural network classifier to handle the imbalance problem between binding and non-binding sites in protein sequence. Extensive experiments were conducted on the real DNA–protein binding sites dataset, PDNA-543, PDNA-224 and PDNA-316, in order to validate the effectiveness of our method on predicting the binding sites by ten-fold cross-validation metric. The experimental results demonstrate that our method achieves a high prediction performance and outperforms the state-of-the-art sequence-based DNA–protein binding sites predictors in terms of the Sensitivity, Specificity, Accuracy, Precision and Mathew's Correlation Coefficient (MCC). Our method can obtain the MCC values of 0.63, 0.48 and 0.67 on PDNA-543, PDNA-224 and PDNA-316 datasets, respectively. Compared with the state-of-the art prediction models, the MCC values for our method are increased by at least 0.24, 0.13 and 0.23 on PDNA-543, PDNA-224 and PDNA-316 datasets, respectively.

1. Introduction

Our work is motivated by the imbalanced problem of the distribution of DNA–protein binding and non-binding sites. Specifically, the number of DNA–protein binding sites (minority class) is significantly fewer than that of the non-binding sites (majority class). In addition, the performance of the minority class can be greatly underestimated. Therefore, we propose the ensemble learning approach by combining multiple convolutional neural networks to utilize the abundant number of non-binding sites.

There has been a growing interest in the identification of DNA–protein binding sites in protein sequence which play crucial roles in vital biological process (Si et al., 2015; Wong et al., 2016; Zhang et al., 2016; Qiao et al., 2018a,b), including gene regulatory, transcription, epigenome, splicing, DNA replication and DNA repair. Hence, accurately identification the protein–DNA binding sites are very importance to understand the mechanism between them. Meanwhile, identification of DNA–protein binding sites can also contribute to understand the pathogenesis of diseases. Many experimental techniques have been proposed to confirm the interactions between DNA and proteins, such as protein binding microarray (PBM) (Hume et al., 2014), ChIP-seq (Steube et al., 2017) and protein microarray assays (Cretich et al., 2014). However, these experimental methods are time-consuming and very expensive. Moreover with the huge amount of protein sequence data available, there is an urgent need to develop computational methods of identifying DNA–protein binding sites from protein sequences.

Currently, many computational methods (Qiao et al., 2015a,b) have been proposed to identify DNA–protein binding sites in protein sequence. During the past decade, a number of machine learning algorithms (Qiao et al., 2018c) have been used to predict DNA-binding

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.engappai.2019.01.003.

* Corresponding author.
E-mail address: sjqiao@cuit.edu.cn (S. Qiao).

https://doi.org/10.1016/j.engappai.2019.01.003
Received 25 July 2018; Received in revised form 1 December 2018; Accepted 4 January 2019
Available online 8 January 2019
0952-1976/

		Position-Specific Score Matrix						One-hot Encoding						Predicted Solvent Accessibility		
		A	R	N	⋯	Y	V									
1	E	−1	0	0	⋯	−2	−2	1	0	⋯	0	0	0	0.00	0.09	0.90
2	T	0	−1	0	⋯	−2	0	0	0	⋯	0	0	0	0.17	0.78	0.03
3	D	−2	−2	1	⋯	−3	−3	0	0	⋯	0	1	0	0.00	0.12	0.88
4	C	−1	−4	−4	⋯	−3	−2	0	0	⋯	0	0	0	0.50	0.36	0.13
5	E	−1	0	0	⋯	−2	−2	1	0	⋯	0	0	0	0.00	0.09	0.90
6	S	0	3	0	⋯	−3	−3	0	0	⋯	0	0	0	0.04	0.25	0.69
7	E	−1	0	0	⋯	−2	−2	1	0	⋯	0	0	0	0.00	0.09	0.90
8	T	0	−1	0	⋯	−2	0	0	0	⋯	0	0	0	0.17	0.78	0.03
9	D	−2	−2	1	⋯	−3	−3	0	0	⋯	0	1	0	0.00	0.12	0.88
10	C	−1	−4	−4	⋯	−3	−2	0	0	⋯	0	0	0	0.50	0.36	0.13
11	E	−1	0	0	⋯	−2	−2	1	0	⋯	0	0	0	0.00	0.09	0.90
12	S	0	3	0	⋯	−3	−3	0	0	⋯	0	0	0	0.04	0.25	0.69

Fig. 1. Example of features among PSSM, One-hot encoding and PSA.

sites from protein sequences, including BindN (Wang and Brown, 2006), BindN+ (Wang et al., 2010), ProteDNA (Chu et al., 2009), DP_Bind (Hwang et al., 2007), MetaDBSite (Si et al., 2011), DNABind (Li et al., 2014), TargetDNA (Hu et al., 2016), EL_PSSM_RT (Zhou et al., 2017) and MLAB (Shen et al., 2017). For example, Wang et al. (2010) proposed a prediction model, referred as BindN+, by integrated PSSM and three physiochemical properties. Hu et al. (2016) proposed a sequenced-based predictor by integrating under-sampling with an appropriate boosting ensemble algorithm. Zhou et al. (2017) proposed and described a novel PSSM encoding method, combining physicochemical features and ensemble learning to predict DNA–protein binding sites. In MLAB (Shen et al., 2017), a multi-scale local average blocks algorithm was proposed by ensemble weighted sparse representation and random under-sampling with evolutionary information and predicted solvent accessibility features from protein primary sequences. These sequence-based predictors only utilize protein sequence information and recognize DNA-binding sites with one or more machine learning algorithms. Despite the promising results of these methods, there remains room for further improvements in accurately predicting DNA-binding sites form protein sequence.

In view of the issues mentioned above, we use position specific scoring matrix (PSSM), One-hot encoding and predicted solvent accessibility (PSA) features to encode residues. In addition, because the number of the DNA–protein binding site (minority class) being significantly lower than that of non-binding sites (majority class), the performance of binding sites identification will be underestimated. However, most of the previous research cannot take advantages of the abundant number of non-binding sites for the identification. In this study, we propose an ensemble classifier by combining multiple convolutional neural networks to utilize the abundant number of non-binding sites. Extensive experiments were conducted on real DNA–protein binding sites dataset including the PDNA-543, PDNA-224 and PDNA-316 data, to validate the effectiveness of our method on predicting the binding sites by ten-fold cross-validation metric. The experimental results demonstrate that our method achieves a high prediction performance and outperforms many existing sequence-based DNA–protein binding sites predictors in terms of the Sensitivity, Specificity, Accuracy, Precision and Mathew's Correlation Coefficient (MCC).

The original contributions of the proposed model are three-fold: (1) we introduce a new feature representation approach by combining position specific scoring matrix, one-hot encoding and predicted solvent accessibility features; (2) we apply Adaptive Synthetic Sampling to oversample the minority class and Bootstrap strategy for majority class to deal with the imbalance problem. In addition, the ensemble technique was used to improve the final performance; (3) the experimental results demonstrate that the proposed approach performs better in identification of DNA–protein binding sites for handling the imbalanced data.

2. Methodology

In this section, we provide the details of the methods and materials of this paper. Fig. 1 provides a system frame of our proposed method. For the training phase, PSSM, one-hot encoding and PSA features are concatenated together and then used as training feature set to train the classifier. ADASYN sampling is used to oversampling the positive class (binding sites). We use CNN and ensemble classifier as the classification algorithm and trained model is stored for the testing phase. Testing phase is the same as the training phase, but the labels for the test dataset are not provided to the classifier.

2.1. Feature representation

From the point of view of machine learning, the prediction of DNA binding sites in proteins is a traditional binary classification problem. Thus, training a machine learning based prediction model on how to encode DNA binding sites in proteins with discriminative feature is one of the most crucial steps. Various effective sequence-based feature, such as position specific scoring matrix(PSSM) (Kelley et al., 2015), predicted secondary structure (Drozdetskiy et al., 2015) and physicochemical properties (Wong et al., 2015), have been explored for predicting protein–DNA binding residues. In this study, we employed PSSM feature and sequence feature for predicting DNA binding sites in proteins.

2.1.1. Position specific scoring matrix

PSSM, being a very important type of evolutionary features, is obtain by running the PSI-BLAST (Kelley et al., 2015) program to search against the SwissProt database (Khoury et al., 2011) through three iteration, with 10^{-3} as the E-value cutoff for multiple sequence alignment. In PSSM, there are 40 scores for each sequence position and each score means the conservation degree of a specific residue type on that position. Before feeding into a prediction engine, all the scores in PSSM need to be scaled between 0 and 1 using the following equation.

$$NPSSM(i,j) = \frac{1}{1+e^{-PSSM(i,j)}} \tag{1}$$

For every data instance, all the scaled scores in the PSSM are used as its evolution features. Since the biological function of binding sites is often influenced by its neighboring sites, a binding site data is defined as a window of length w with the target binding site positioned in the middle and $(w-1)/2$ neighboring sites on either side.

Sequence features include local amino acid composition, predicted second structure and predicted solvent accessible area. Each protein sequence is converted into a $20 \times L$ one-hot encoding binary matrix (L is the probe length) and the intensity values are normalized.

2.1.2. Predicted solvent accessibility

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 (Heffernan et al., 2017). Ahmad et al. (2010) demonstrated the important role of solvent accessibility to amino acid residues in predicting protein–DNA binding. Therefore, we used solvent accessibility information to predict protein–DNA binding residues. We get the predicted solvent accessibility (PSA) characteristics of each residue by feeding the corresponding sequence to the standalone SANN program, which can be download at http://lee.kias.re.kr/~newton/sann/. For each protein sequence, SANN precisely predicts its PSA matrix (L rows and 3 columns), which includes the probabilities of three solvent accessibility classes (i.e., buried (B), intermediate (I), and exposed (E)) of each residue.

In order to improve the identification of DNA–protein binding sites, we combine the PSSM, One-hot encoding and PSA features together. For PSSM, the conservation scores correspond to twenty amino acid position in PSI-BLAST. For One-hot encoding, there are twenty dimensions for each residue with twenty columns of features. For PSA, there are three solvent accessibility classes of each residue.

Fig. 1 is an illustrative example of the features of PSSM, One-hot encoding and PSA. By Fig. 1, for example, the window size is set to five. The input for every prediction is the feature score as given in the row corresponding to this target binding site, the totaling number of input values is $5 \times 43 = 215$.

2.2. ADASYN sampling method

Adaptive Synthetic Sampling (ADASYN) sampling method uses a systematic method to adaptively create different amounts of synthetic data according to their distributions (Krawczyk, 2016). The difference between SMOTE is, instead of using a uniform distribution for data generation, ADASYN applies a density distribution as a criterion to automatically decide the number of synthetic samples which need to be generated for each minority example. The density distribution criterion is defined as the normalized amount of majority cased within the k-nearest neighbor of each minority case. This is achieved as follows: Firstly, calculate the number of synthetic data examples that need to be generated for the entire minority examples:

$$G = (|M_{maj}| - |M_{min}|) \times \beta \tag{2}$$

where M_{maj} is the number of majority classes, M_{min} is the number of minority classes and $\beta \in [0, 1]$ is a parameter used to specify the desired balanced level after the synthetic data generation process. Secondly, for each case $x_i \in M_{min}$, find the k-nearest neighbors according to the Euclidean distance and calculate the ratio Γ_i defined as:

$$\Gamma = \frac{\Delta_i / K}{Z}, i = 1, \ldots, |M_{min}| \tag{3}$$

where Δ_i is the number of cases in the K-nearest neighbors of x_i that belong to M_{maj}, and Z is normalization constant so that Γ_i is a distribution function. Then, determine the number of synthetic data samples that need to be generated for each $x_i \in M_{min}$:

$$g_i = \Gamma_i \times G \tag{4}$$

Finally, for each $x_i \in M_{min}$, generate g_i synthetic data samples according to below formula:

$$x_{new} = x_i + (\hat{x}_i - x_i) \times \delta \tag{5}$$

where $\hat{x}_i$ is one of the k-nearest neighbors for x_i and $\hat{x}_i \in M_{min}$ and $\delta \in [0, 1]$ is a random number.

Table 1 summarizes the commonly used symbols in this paper.

2.3. Convolutional Neural Network (CNN)

Most modern deep learning models are based on an artificial neural network (Huang, 1999; Huang and Du, 2008). Recently, deep learning techniques have proved highly effective in a number of diverse tasks compared with other machine learning methods, and have been widely applied to the field of bioinformatics (Min et al., 2017), such as, gene expression regulation (Chen et al., 2016), protein structure prediction (Spencer et al., 2015), protein classification (Asgari and Mofrad, 2015), etc.

CNN is a well-known deep learning architecture that has been extensively applied to image recognition (He et al., 2016), speech recognition (Qian et al., 2016), natural language processing (Goldberg, 2016), bioinformatics (Zhou and Troyanskaya, 2015; Alipanahi et al., 2015) and other artificial intelligence research fields. CNN perform adaptive feature extraction to map input data to informative representations during training. The basic components of a CNN include convolutional, pooling and fully connected layers. The convolution operation is the engine of the CNN. The convolutional layer aims to extract and represent the local information of raw features by several feature maps and kernels. The pooling layer aims to compress the resolution of the feature maps to achieve spatial invariance. After several convolution and pooling operations, there are one or more fully connected layers to perform high level reasoning. The output of the last fully connected layer is fed back to an output layer. For a classifier or regression task, softmax regression is commonly used. CNNs generalize well by encoding spatial invariance during training.

Table 1
Symbols and descriptions.

Symbol	Description
PSSM(i,j)	The Position Specific Scoring Matrix value between protein residue i and j
NPSSM(i,j)	The normalization of PSSM(i,j) value
L	The length of protein sequence
M_{min}	The number of the minority class
M_{max}	The number of the majority class
G	The total number of synthetic data be generated for minority class
β	A factor used to specify the desired balanced level (see Eq. (2))
δ	A factor used to generate the new minority data (see Eq. (5))
K	The number of K-nearest neighbors
Δ_i	The number of cases in K-nearest neighbors of x_i
Z	The normalization constant (See Eq. (3))
Γ	The Distribution function
g_i	The number of synthetic data for each minority example x_i

2.4. Back-propagation (BP) algorithm

The back-propagation (BP) algorithm (Hameed et al., 2016) is a common method for training a neural network. The basic idea behind BP is to minimize the overall output error gradually during the learning process. Whereas the training set are estimated iteratively through the input layer to predict the correct output. The BP process is segmented into two stages: forward and backward process. In the forward process, the input signal is propagated from the input layer to the output layer. During the forward process, the weight value and offset value of the network are maintained constant and the status of each layer of neuron will only exert an effect on that of next layer of neuron. In the backward process, weights on the connections between all layers will be updated to minimize the error between target and predict result until finding the optimum weights with minimum error function.

2.5. Ensemble classifier and Bootstrap strategy

In this study, the basic idea of the propose method is to integrate the ensemble learning algorithm into the sampling technique, which will allow us to reduce the imbalance ratio of samples (Qiu et al., 2018; Zhang et al., 2012), therefore, make the learning task more tractable. In this method, the majority class is under-sampling with the Bootstrap strategy (Khosravi et al., 2015). By performing random sampling on m subsets from the majority category of non-binding cases, we can make all negative subsets have the same or similar size as the minority class of binding cases. In addition, the minority class performs over-sampling by ADASYN technique. After the phase of sampling, each negative

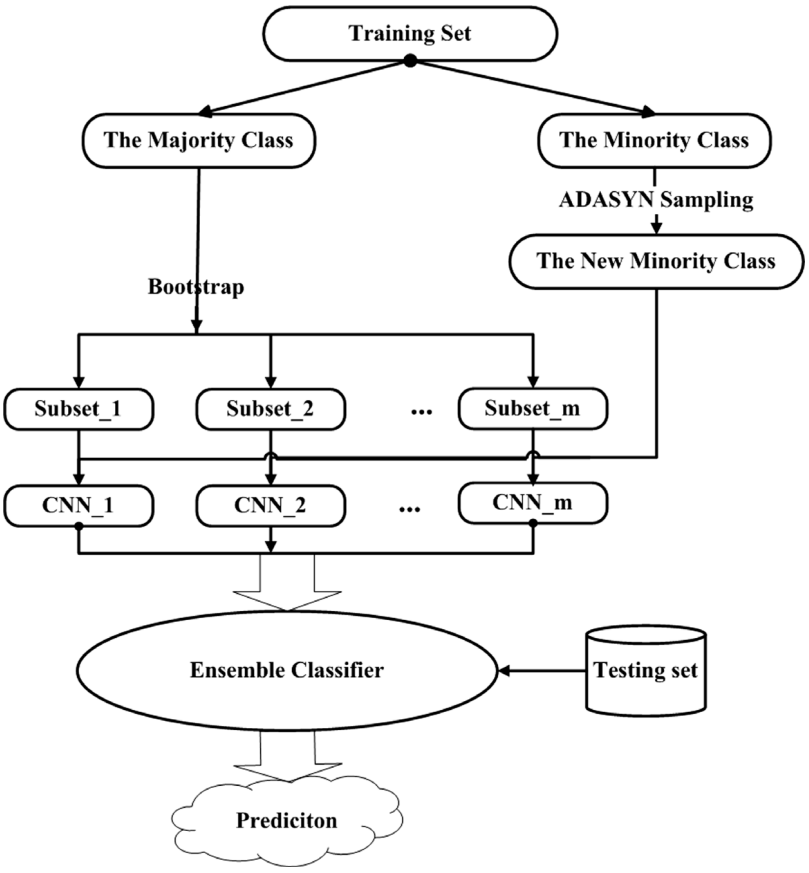


Fig. 2. The working mechanism of the ensemble classifier.

subset will combine with the set of positive cases and generate m new training sets. Then, we train the multiple DNA-binding sites predictor with CNNs as a base classifier. Finally, the result is obtained by applying the majority voting strategy to ensemble the base classifiers.

In the imbalanced problem, the minority class is more important than the majority class, so, in this study, the vote will be assigned to the minority class when the label of base classifiers is equally divided. For example, the numbers of classifier are 10 in this study, if the classified labels are both five, we will view this sample to be the minority class. The working mechanism of the ensemble classifier is shown in Fig. 2.

2.6. Model parameters and training procedure

Our proposed training classifier consists of four convolutional layers. The input layer is a $w \times 43$, where w is the window size for each binding site. Specifically, the size of PSSM feature is $w \times 20$, the size of One-hot encoding is $w \times 20$ and the size of PSA is $w \times 3$. The first layer of our network is a convolutional layer, which uses 32 filters of size 9×1 over the input. The second layer is a max-pooling layer whose size is 7×1. The third layer is a fully connected layer with 32 neurons. Since it is fully connected, each of its 32 neurons is connected to all of the neurons in the max-pooling layer. The final output layer consists of two neurons corresponding to these two classification results.

In this study, our method is trained using the standard back-propagation algorithm and mini-batch gradient descent with the Ada-grad variation (Duchi et al., 2011). Dropout (Srivastava et al., 2014) and early stopping strategies are used for regularization and model selection. The detailed introduction of parameter specifications are given in Section 3. All models used in this study share a genetic SGD forward and backward training procedure. In each iteration, the whole training data are divided into batches and process one batch at one time. Each batch contains a list of sentences which is determined by the parameter of batch size. For most of machine learning methods, the algorithms perform well when the batch size is specified between 30 and 200 (Alipanahi et al., 2015). By empirical studies, we use fifty batch sizes for all experiments in this study. All the initializations of weights and bias were set to the default in Keras. In this manuscript, the value of 100 is specified for epoch, which means we run the whole dataset (including all batches) for 100 times. Code for all experiments was written in the Python library Theano. All experiments were run on a machine with 24 Xeon processor, and 256 GB of memory and 1 Nvidia Tesla K80c GPU.

2.7. Benchmark datasets

In this study, the three DNA–protein binding sites datasets, PDNA-543 (Hu et al., 2016), PDNA-224 (Li et al., 2013) and PDNA-316 (Si et al., 2011), are used to evaluate the performance of the proposed method. A summary of these datasets are shown as follows:

PDNA-543 consists of 7186 protein sequences, which are related to the PDB (Protein Data Bank) and collected before October 10, 2014. After removing redundant sequences by wielding CD-hit (Qiao et al., 2018b) software, there are 543 non-redundant protein sequences and there is not more than 30% similarity between two sequences. Specifically, there are 9549 binding sites and 134 995 non-binding sites while the imbalanced ratio is about 14.2.

PDNA-224 comprise of 224 non-redundant protein chains containing 57,348 amino acids. There are 3778 binding sites and 53 570 non-binding sites while the imbalanced ratio is about 14.2.

PDNA-316 is constructed by Si et al., which has 316 DNA-binding protein chains and 5609 binding sites and 67,109 non-binding sites while the imbalanced ratio is about 12.

These three datasets are used for different purpose in different experiments, so we train the each dataset individually. Table 2 summarizes the detailed compositions of PDNA-543, PDNA-224 and PDNA-316.

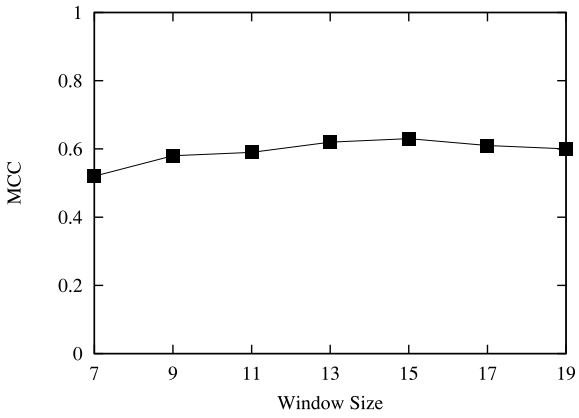


Fig. 3. The performance variation curves of MCC versus window size under our method.

Table 2
Composition of dataset. (numP represent the number of positive sample; numN represent the number of negative samples; Ratio = numN/numP).

Dataset	No. of sequences	numP	numN	Ratio
PDNA-543	543	9549	134,995	14.2
PDNA-224	224	3778	53,570	14.2
PDNA-316	316	5609	67,109	12.0

3. Experimental results and analysis

In this section, we test on three DNA–protein binding sites datasets to evaluate the performance of our proposed method, including PDNA-543, PDNA-224 and PDNA-316. Firstly, we present a detailed explanation of the measurement used for assessing the performance of the deep architecture and consider various aspects of the deep learning and their effects on predicting performance. Then, we analyze the performance of binding site representation, such as PSSM, one-hot encoding, PSA and combine them together. Lastly, we compare the performance of our method with some outstanding methods on the above datasets.

3.1. Evaluation metrics

In this study, five evaluation indexes routinely used in this field, Sensitivity (Sen), Specificity (Spe), Accuracy (Acc), Precision (Pre) and the Mathew's Correlation Coefficient (MCC) are utilized to evaluate predictive ability. They are calculated according to the following formulate:

$$Sen = \frac{TP}{TP + FN} \tag{6}$$

$$Spe = \frac{TN}{TN + FP} \tag{7}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{8}$$

$$Pre = \frac{TP}{TP + FP} \tag{9}$$

$$MCC = \frac{TP \bullet TN - FN \bullet FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{10}$$

where TP is the number of true positives; TN is the number of true negatives; FP is the number of false positives; FN is the number of false negatives; P is the number of positives and N is the number of negatives.

3.2. Window size selection in PSSM

In this section, we discuss how to choose the window size on the training dataset PDNA-543 over a ten-fold cross-validation. We evaluate the MCC on the training dataset by gradually varying the value from 7 to 19. Fig. 3 shows the performance variation curves of MCC under different size of window in PSSM. As shown in Fig. 3, we find that the value of MCC continues to increase with the window size and reaches to the peak value when window size equals 15.

In this study, we have the convention that 15 vicinal amino acid residues as a window ($w = 15$), where the window specifically indicates the target residue and 7 neighbors on either side of the target residue itself. This causes the dimensionality of the PSSM feature to be $15 \times 20 = 300$, $15 \times 20 = 300$ dimensional feature vector for one-hot encoding and the dimensionality of the PSA feature is $15 \times 3 = 45$.

Table 3
The performance on PDNA-543 for various features by ten-fold cross-validation.

Feature	Sen(%)	Spe(%)	Acc(%)	Pre(%)	MCC
PSSM	72.31	78.16	76.89	31.02	0.35
One-hot encoding	77.75	89.97	88.45	51.78	0.57
PSSM+PSA	76.27	85.85	84.73	41.68	0.49
One-hot encoding+PSA	78.55	91.51	89.95	55.93	0.61
Combine three features	78.77	92.36	90.77	57.76	0.63

3.3. Performance of different features on PDNA-543

To analyze the performance of PSSM, one-hot encoding and PSA features, we evaluate these features by our method on PDNA-543 dataset. The result is shown in Fig. 4 and the detail results are given in Table 3. As mentioned above, Sen, Spe, Acc, Pre and MCC are the main metrics.

It can be observed that the PSSM+PSA features achieve 76.27% for Sen, 85.85% for Spe, 84.73% for Acc, 41.68% for Pre and 0.49 for MCC, which outperform PSSM features by 3.96%, 7.69%, 7.84%, 10.66% and 0.14, respectively. The One-hot encoding+PSA features achieve 78.55% for Sen, 91.51% for Spe, 89.95% for Acc, 55.93% for Pre and 0.61 for MCC, which outperform One-hot encoding features by 0.8%, 1.54%, 1.5%, 4.15% and 0.04, respectively. Because of additional solvent accessibility information, the PSSM+PSA and One-hot encoding+PSA are all higher than single PSSM and one-hot encoding features.

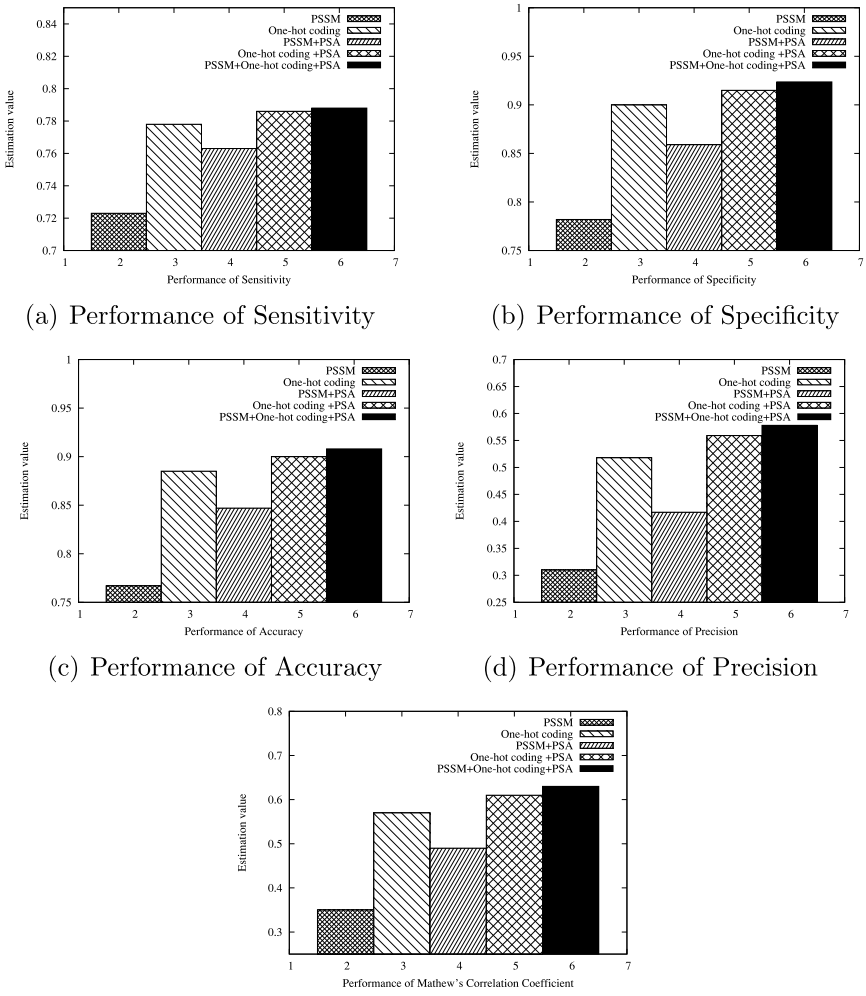
Obviously, the combination approach of PSSM, one-hot encoding and PSA gets better performance the other features. Furthermore, it achieves 78.77% for Sen, 92.36% for Spe, 90.77% for Acc, 57.76% for Pre and 0.63 for MCC. The MCC of PSSM + One-hot encoding + PSA is higher than PSSM (0.28), One-hot coding (0.06), PSSM + PSA (0.14) and One-hot encoding + PSA (0.02), respectively. In Fig. 4, we can see that the fusion feature of PSSM, One-hot encoding and PSA has the best performance than the other features in the PDNA-543 dataset, which indicates that these features are complementary for each other and the one-hot encoding feature is the important method for effectively predicting protein–DNA binding.

In addition, we also compare the ANOVA for Spe and ACC, Pre and MCC under five different features, PSSM, One-hot encoding, PSSM + PSA, One-hot encoding + PSA and PSSM + One-hot encoding + PSA. Specifically, the null hypothesis in ANOVA is always that there is no difference in the mean values. The alternative hypothesis is always that the mean values are not all equal.

The results are given in Table 4. According to Table 4, we can see that the P-value of Spe and ACC, Pre and MCC is 0.708 and 0.474, respectively. Since the P-value of Spec and ACC, Pre and MCC all bigger than 0.05, so we accept the null hypothesis, which means Spe and ACC is no difference in means. For the metrics of Pre and MCC, there is no difference in the mean values.

3.4. Comparison of ensemble classifier with base classifier

In this section, we compare the performance of the ensemble classifier and base classifier. The result is shown in Table 5, where the



(a) Performance of Sensitivity

(b) Performance of Specificity

(c) Performance of Accuracy

(d) Performance of Precision

(e) Performance of Mathew's Correlation Coefficient

Fig. 4. The performance on PDNA-543 for various features by ten-fold crossvalidation. (a), (b), (c), (d) and (e) are the performance of Sensitivity, Specificity, Accuracy, Precision and MCC, respectively.

performance is PSSM + One-hot encoding + PSA features. Table 5 shows that compare to the CNN classifier, the ensemble classifier achieves significant performance improvement on both PDNA-543 and PDNA-316. More specifically, on the PDNA-543, the increase to the CNN is 1.77% on Sen, 15.58% on Spe, 13.97% on Acc and 0.25 on MCC. For PDNA-316 dataset, the increase to the CNN classifier is 3.83% on Sen, 15.92% on Spe, 14.28% on ACC, 0.25 on MCC. It validates the important of ensemble classifier to deal with the imbalanced data for identification of DNA–protein binding sites.

3.5. Comparison with existing predictors on PDNA-543

In this section, we demonstrate the effectiveness of the proposed method by comparing it with other commonly-used predictors of DNA binding sites in proteins, including TargetDNA (Hu et al., 2016)and EC-RUS(WSRC) (FPR5%) (Shen et al., 2017), the results of which are shown in Fig. 5 and Table 6.

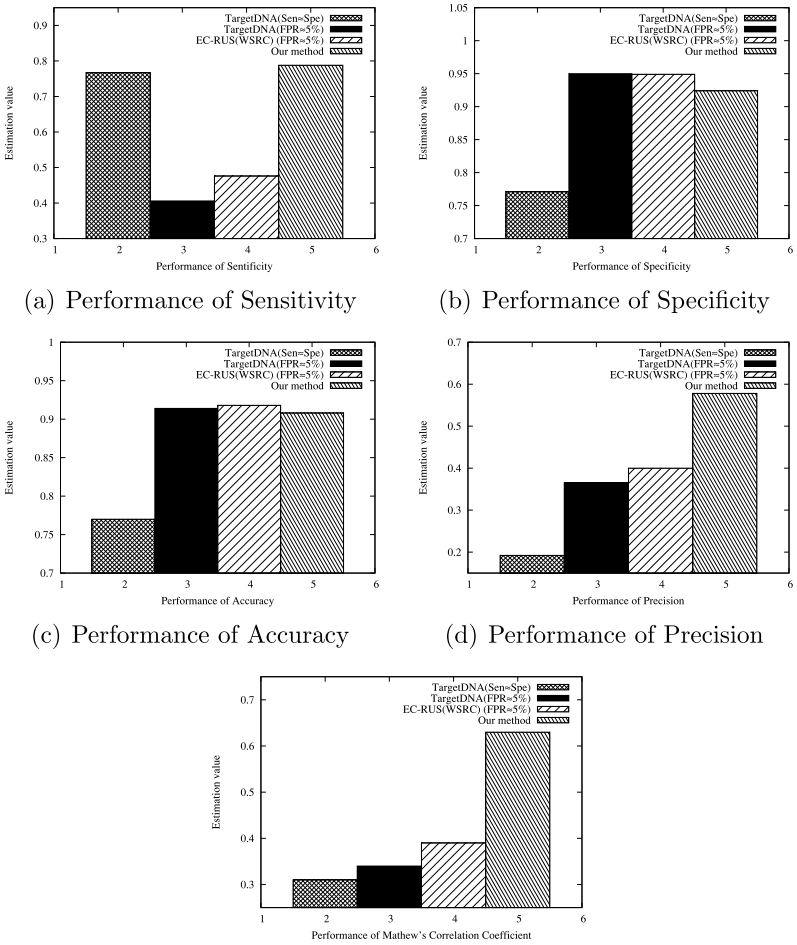
The results shown in Table 6 and Fig. 5 clearly demonstrate that our method outperforms the other predictors in terms of MCC, which is an overall index for evaluation the quality of binary prediction. It is noted that TargetDNA (FPR5%) and EC-RUS(WSRC) (FPR5%) gained the higher specificity and accuracy but comes with the lower sensitivity than our method (38.17% and 31.15%). The sensitivity is an important measure of imbalanced data, which is of the most interest for many researchers. The low sensitivity value of TargetDNA (FPR5%) and EC-RUS(WSRC) (FPR5%) indicates that this method is most likely to mistake identification DNA–protein binding sites.

Table 4
ANOVA analysis of PDNA-543 for various features by ten-fold cross-validation. SS is the sum of square; MS is the mean sum of square; df is the degree of freedom.

	Source	SS	df	MS	F-ratio	*P*-value
Specificity and accuracy	Groups	0.0005	1	0.0005	0.15	**0.708**
	Error	0.0265	8	0.0033		
	Total	0.027	9			
Precision and MCC	Groups	0.0072	1	0.0072	0.56	**0.474**
	Error	0.102	8	0.0128		
	Total	0.1092	9			

Table 5
Comparison of ensemble classifier with base classifier on PDNA-543 and PDNA-316.

	Evaluation method	Sen(%)	Spe(%)	Acc(%)	MCC
PDNA-543	Base-classifier	77.00	76.78	76.80	0.38
	Ensemble-classifier	78.77	92.36	90.77	0.63
PDNA-316	Base-classifier	78.91	76.42	76.76	0.42
	Ensemble-classifier	82.74	92.34	91.04	0.67



(a) Performance of Sensitivity

(b) Performance of Specificity

(c) Performance of Accuracy

(d) Performance of Precision

(e) Performance of Mathew's Correlation Coefficient

Fig. 5. The performance comparisons between our method and other predictors on PDNA-543. (a), (b), (c), (d) and (e) are the performance of Sensitivity, Specificity, Accuracy, Precision and MCC, respectively.

Table 6
The performance comparisons between our method and other predictors on PDNA-543 over ten-fold cross-validation.

Evaluation methods	Sen(%)	Spe(%)	Acc(%)	Pre(%)	MCC
TargetDNA(Sen ≈ Spe)	76.98	77.05	77.04	19.18	0.31
TargetDNA(FPR ≈ 5%)	40.60	95.00	91.40	36.47	0.34
EC-RUS(FPR ≈ 5%)	47.62	94.92	91.80	39.91	0.39
Our method	78.77	92.36	90.77	57.76	0.63

Table 7
The performance comparisons between our method and PreDNA on PDNA-224 over ten-fold cross-validation.

Evaluation methods	Sen(%)	Spe(%)	Acc(%)	MCC
PreDNA	76.1	82.2	81.8	0.35
Our method	76.8	84.5	83.5	0.48

3.6. The predicted results on PDNA-224

To further evaluate the performance of our method in predicting DNA-binding sites, we compared our method with PreDNA (Li et al., 2013) predictors using the PDNA-224 test sets. The results are shown in Table 7. From Table 7, we observe that our method consistently outperforms the PreDNA method in terms of four evaluation indexes, especially for MCC value. The Sen, Spe, Acc and MCC of our method are 76.8%, 84.5%, 83.5% and 0.48, respectively, which are improvements of approximately 0.7%, 2.3%, 1.7% and 0.13, respectively, over the PreDNA.

3.7. Predicted results on the PDNA-316

In order to highlight the advantage of our method, we also test on the PDNA-316 dataset, which is described by Si et al. (2011). We compare the prediction performance of our proposed method with other previous research including BindN (Wang and Brown, 2006), BindN + (Wang et al., 2010), EC-RUS(WSRC) (Shen et al., 2017), DISIS (Si et al., 2011), MetaDBSite (Si et al., 2011), DP_Bind (Hwang et al., 2007), DNABindR (Si et al., 2011) and TargetDNA (Hu et al., 2016), the results of which are shown in Table 8.

Table 8 shows that our method achieves satisfactory results, such as 82.74% on Sen, 92.34% on Sep, 91.03% on ACC, 0.67 on MCC, outperforming BindN, EC-RUS(WSRC) (SenSpe), BindN-RF, MetaDBSite, DP-Bind, DNABindR and TargetDNA(SenSpe) predictor for all of four evaluation criterion. When compared with TargetDNA(FPR5%), EC-RUS(WSRC) (FPR5%) and DISIS, these predictors have higher Precision value than our method. However, our method perform better than TargetDNA(FPR5%), EC-RUS(WSRC) (FPR5%) and DISIS by 39.72%, 31.66%, 63.74% and 0.29, 0.23, 0.42 on Sen and MCC, respectively. The results on PDNA-316 dataset indicate that the effect use of protein sequence information and ensemble classifiers.

Table 8
Comparison of the predicting performance between other predictors on PDNA-316.

Predictor	Sen(%)	Spe(%)	Acc(%)	MCC
BindN	54.00	80.00	78.00	0.21
EC-RUS(Sen ≈ Spe%)	80.67	78.18	78.37	0.36
EC-RUS(FPR ≈ 5%)	51.08	94.99	91.61	0.44
DISIS	19.00	98.00	92.00	0.25
BindN-RF	67.00	83.00	82.00	0.32
MetaDBSite	77.00	77.00	77.00	0.32
DP-Bind	69.00	79.00	78.00	0.29
DNABindR	66.00	74.00	73.00	0.23
TargetDNA(Sen ≈ Spe)	77.96	78.03	78.02	0.34
TargetDNA(FPR ≈ 5%)	43.02	95.00	90.99	0.38
Our method	82.47	92.34	91.03	0.67

4. Conclusion

In this paper, we have designed and implemented a novel sequence-based predictor of DNA–protein binding sites. It is trained on the PDNA-543, PDNA-224 and PDNA-316 with an ADASYN, undersampling, CNN classifier, and the bootstrap classifier ensemble strategy. When different sequence features combined, the prediction performance is further improved. This indicates that the PSSM, one-hot encoding and PSA are complementary for prediction. The comparison of ensemble classifier with the CNN classifier indicates that ensemble learning is indeed useful for DNA–protein binding sites. Our method achieves MCC of 0.63, 0.48 and 0.67 on PDNA-543, PDNA-224 and PDNA-316 datasets, respectively. Compared with the state-of-the art prediction models, MCC for our method is increased by at least 0.24, 0.13 and 0.23 on PDNA-543, PDNA-224 and PDNA-316 datasets, respectively. Our method has obtained a desired performance on three datasets. In the future, we will consider some other physicochemical features to construct the model and try to explain the biology meaning of CNN filters.

Abbreviations

ADASYN: Adaptive Synthetic Sampling; Acc: Accuracy; BP: Back-propagation algorithm; CNN: Convolutional Neural Network; FN: the number of false negative; FP: the number of false positive; MCC: Mathews Correlation Coefficient; PSA: Predicted Solvent Accessibility; Pre: Precision; PSSM: Position Specific Scoring Matrix; Sen: Sensitivity; Spe: Specificity; TN: the number of true negatives; TP: the number of true positive.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants Nos. 61702058, 61772091, 61802035, the China Postdoctoral Science Foundation Funded Project under Grant No. 2017M612948, the Scientific Research Foundation for Advanced Talents of Chengdu University of Information Technology, China under Grant Nos. KYTZ201717, KYTZ201715, KYTZ201750, the Scientific Research Foundation for Young Academic Leaders of Chengdu University of Information Technology, China under Grant Nos. J201701, J201706, the Sichuan Science and Technology Program, China under Grant No. 2018JY0448, the National Natural Science Foundation of Guangxi, China under Grant No. 2018GXNSFDA138005, the Scientific Research Foundation for Education Department of Sichuan Province, China under Grant No. 18ZA0098, the Innovative Research Team Construction Plan in Universities of Sichuan Province, China under Grant No. 18TD0027, and Guangdong Key Laboratory Project, China under Grant No. 2017B030314073.

Conflict of interest

There is no conflict of interest.

References

Ahmad, S., Gromiha, M.M., Sarai, A., 2010. Real value prediction of solvent accessibility from amino acid sequence. Proteins Struct. Funct. Bioinform. 50 (4), 629–635.
Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J., 2015. Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. Nature Biotechnol. 33 (8), 831–839.
Asgari, E., Mofrad, M.R.K., 2015. Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One 10 (11), 1–15.
Chen, Y., Li, Y., Narayan, R., Subramanian, A., Xie, X., 2016. Gene expression inference with deep learning. Bioinformatics 32 (12), 1–8.
Chu, W.Y., Huang, Y.F., Huang, C.C., Cheng, Y.S., Huang, C.K., Oyang, Y.J., 2009. Protedna: a sequence-based predictor of sequence-specific dna-binding residues in transcription factors. Nucleic Acids Res. 37 (Web Server issue), W396.
Cretich, M., Damin, F., Chiari, M., 2014. Protein microarray technology: how far off is routine diagnostics? Analyst 139 (3), 528–542.
Drozdetskiy, A., Cole, C., Procter, J., Barton, G.J., 2015. Jpred4: a protein secondary structure prediction server. Nucl. Acids Res. 43 (W1), W389–W394.
Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12 (7), 257–269.
Goldberg, Y., 2016. A primer on neural network models for natural language processing. J. Artificial Intelligence Res. 57, 345–420.
Hameed, A.A., Karlik, B., Salman, M.S., 2016. Back-propagation algorithm with variable adaptive momentum. Knowl.-Based Syst. 114, 79–87.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
Heffernan, R., Yang, Y., Paliwal, K., Zhou, Y., 2017. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics 33 (18), 2842–2849.
Hu, J., Li, Y., Zhang, M., Yang, X., Shen, H.B., Yu, D.J., 2016. Predicting protein-dna binding residues by weightedly combining sequence-based features and boosting multiple svms. IEEE/ACM Trans. Comput. Biol. Bioinform. PP (99), 1389–1398.
Huang, D.-S., 1999. Radial basis probabilistic neural networks: Model and application. Int. J. Pattern Recogn. Artif. Intell. 13 (07), 1083–1101.
Huang, D.S., Du, J.X., 2008. A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks. IEEE Trans. Neural Netw. 19 (12), 2099.
Hume, M.A., Barrera, L.A., Gisselbrecht, S.S., Bulyk, M.L., 2014. Uniprobe, update 2015: new tools and content for the online database of protein-binding microarray data on protein–dna interactions. Nucl. Acids Res. 43 (D1), D117–D122.
Hwang, S., Gou, Z., Kuznetsov, I.B., 2007. Dp-bind: a web server for sequence-based prediction of dna-binding residues in dna-binding proteins. Bioinformatics 23 (5), 634–636.
Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., Sternberg, M.J., 2015. The phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. 10 (6), 845–858.
Khosravi, A., Nahavandi, S., Srinivasan, D., Khosravi, R., 2015. Constructing optimal prediction intervals by using neural networks and bootstrap method. IEEE Trans. Neural Netw. Learn. Syst. 26 (8), 1810–1815.
Khoury, G.A., Baliban, R.C., Floudas, C.A., 2011. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. Sci. Rep. 1, 90.
Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Prog. Artif. Intell. 5 (4), 221–232.
Li, B.Q., Feng, K.Y., Ding, J., Cai, Y.D., 2014. Predicting dna-binding sites of proteins based on sequential and 3d structural information. Mol. Genet. Genomics 289 (3), 489–499.
Li, T., Li, Q.Z., Liu, S., Fan, G.L., Zuo, Y.C., Peng, Y., 2013. Predna: accurate prediction of dna-binding sites in proteins by integrating sequence and geometric structure information. Bioinformatics 29 (6), 678–685.
Min, S., Lee, B., Yoon, S., 2017. Deep learning in bioinformatics. Brief. Bioinform. 18 (5), 851–869.
Qian, Y., Bi, M., Tan, T., Yu, K., 2016. Very deep convolutional neural networks for noise robust speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (12), 2263–2276.
Qiao, S., Han, N., Gao, Y., Li, R.-H., Huang, J., Guo, J., Gutierrez, L.A., Wu, X., 2018a. A fast parallel community discovery model on complex networks through approximate optimization. IEEE Trans. Knowl. Data Eng. 30 (9), 1638–1651.
Qiao, S., Han, N., Wang, J., Li, R.H., Gutierrez, L.A., Wu, X., 2018b. Predicting long-term trajectories of connected vehicles via the prefix-projection technique. IEEE Trans. Intell. Transp. Syst. 2018 (7), 2305–2315.
Qiao, S., Han, N., Zhou, J., Li, R.H., Jin, C., Gutierrez, L.A., 2018c. Socialmix: A familiarity-based and preference-aware location suggestion approach. Eng. Appl. Artif. Intell. 68, 192–204.
Qiao, S., Han, N., Zhu, W., Gutierrez, L.A., 2015a. Traplan: An effective three-in-one trajectory-prediction model in transportation networks. IEEE Trans. Intell. Transp. Syst. 16 (3), 1188–1198.
Qiao, S., Shen, D., Wang, X., Han, N., Zhu, W., 2015b. A self-adaptive parameter selection trajectory prediction approach via hidden markov models. IEEE Trans. Intell. Transp. Syst. 16 (1), 284–296.

Qiu, W.-R., Sun, B.-Q., Xiao, X., Xu, Z.-C., Jia, J.-H., Chou, K.-C., 2018. iKcr-pseens: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics 110 (5), 239–246.

Shen, C., Ding, Y., Tang, J., Song, J., Guo, F., 2017. Identification of dna-protein binding sites through multi-scale local average blocks on sequence information. Molecules 22 (12), 2079.

Si, J., Zhang, Z., Lin, B., Schroeder, M., Huang, B., 2011. Metadbsite: a meta approach to improve protein dna-binding sites prediction. BMC Syst. Biol. 5 (S1), S7.

Si, J., Zhao, R., Wu, R., 2015. An overview of the prediction of protein dna-binding sites. Int. J. Mol. Sci. 16 (3), 5194–5215.

Spencer, M., Eickholt, J., Cheng, J., 2015. A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Trans. Comput. Biol. Bioinform. 12 (1), 103–112.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Steube, A., Schenk, T., Tretyakov, A., Saluz, H.P., 2017. High-intensity uv laser chip-seq for the study of protein-dna interactions in living cells. Nat. Commun. 8 (1), 1303.

Wang, L., Brown, S.J., 2006. Bindn: a web-based tool for efficient prediction of dna and rna binding sites in amino acid sequences. Nucleic Acids Res. 34 (Web Server issue), W243.

Wang, L., Huang, C., Yang, M.Q., Yang, J.Y., 2010. Bindn+ for accurate prediction of dna and rna-binding residues from protein sequence features. BMC Syst. Biol. 4 (S1), S3.

Wong, K.C., Li, Y., Peng, C., Moses, A.M., Zhang, Z., 2015. Computational learning on specificity-determining residue-nucleotide interactions. Nucleic Acids Res. 43 (21), 10180–10189.

Wong, K.C., Li, Y., Peng, C., Wong, H.S., 2016. A comparison study for dna motif modeling on protein binding microarray. IEEE/ACM Trans. Comput. Biol. Bioinform. 13 (2), 261–271.

Zhang, Y., Cao, X., Zhong, S., 2016. Genemo: a search engine for web-based functional genomic data. Nucl. Acids Res. 44 (W1), W122–W127.

Zhang, Y., Zhang, D., Mi, G., Ma, D., Li, G., Guo, Y., Li, M., Zhu, M., 2012. Using ensemble methods to deal with imbalanced data in predicting protein–protein interactions. Comput. Biol. Chem. 36, 36–41.

Zhou, J., Lu, Q., Xu, R., He, Y., Wang, H., 2017. El_pssm_rt: Dna-binding residue prediction by integrating ensemble learning with pssm relation transformation. BMC Bioinformatics 18 (1), 379.

Zhou, J., Troyanskaya, O.G., 2015. Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods 12 (10), 931–934.