Briefings in Bioinformatics, 22(5), 2021, 1–12

https://doi.org/10.1093/bib/bbaa397
Problem Solving Protocol

NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning

Jun Zhang, Qingcai Chen and Bin Liu

Corresponding author: Bin Liu, Harbin Institute of Technology, HIT Campus Shenzhen University Town, Xili, Shenzhen 518055, China, and School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China. Tel.: (+86) 010-68911310; E-mail: bliu@bliulab.net

Abstract

The interactions between proteins and nucleic acid sequences play many important roles in gene expression and some cellular activities. Accurate prediction of the nucleic acid binding residues in proteins will facilitate the research of the protein functions, gene expression, drug design, etc. In this regard, several computational methods have been proposed to predict the nucleic acid binding residues in proteins. However, these methods cannot satisfactorily measure the global interactions among the residues along protein. Furthermore, these methods are suffering cross-prediction problem, new strategies should be explored to solve this problem. In this study, a new computational method called NCBRPred was proposed to predict the nucleic acid binding residues based on the multilabel sequence labeling model. NCBRPred used the bidirectional Gated Recurrent Units (BiGRUs) to capture the global interactions among the residues, and treats this task as a multilabel learning task. Experimental results on three widely used benchmark datasets and an independent dataset showed that NCBRPred achieved higher predictive results with lower cross-prediction, outperforming 10 existing state-of-the-art predictors. The web-server and a stand-alone package of NCBRPred are freely available at http://bliulab.net/NCBRPred. It is anticipated that NCBRPred will become a very useful tool for identifying nucleic acid binding residues.

Key words: nucleic acid binding residue prediction; cross-prediction problem; multilabel learning; sequence labeling model

Introduction

The interactions between proteins and nucleic acids (DNA/RNA) play many crucial roles in biological processes, such as transcription control, translation, DNA replication, posttranscriptional gene regulation [1–3]. Accurate identification of the nucleic acid binding residues in proteins is important for studying and characterizing the interactions between proteins and nucleic acids [4, 5].

The nucleic acid binding residues in proteins were mainly detected by wet-lab experimental methods, such as X-ray crystallography and nuclear magnetic resonance. However, these methods are relatively expensive and slow, not suitable for whole-genome scale analysis [6–8]. Therefore, only a small number of complexes of proteins and nucleic acids have been resolved and deposited in Protein Data Bank (PDB) [9]. Many nucleic acid binding proteins remain to be discovered and learned. In this regard, several efforts have been made to develop computational methods to detect the nucleic acid binding residues in proteins based on the protein sequence or structure information. These methods are faster with lower cost [8, 10, 11] comparing with the wet-lab experimental methods. For examples, DP-bind [12] predicts DNA-binding residues in proteins based on Position-Specific Scoring Matrix (PSSM) and three machine learning classifiers, including Support

Jun Zhang is currently pursuing the PhD degree in computer science and technology with Harbin Institute of Technology, Shenzhen, China. His research interests include bioinformatics, natural language processing and machine learning.
Qingcai Chen, PhD, is a professor at the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. His research interests include machine learning, pattern recognition, speech signal processing and natural language processing.
Bin Liu, PhD, is a professor at the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China, and the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. His expertise is in bioinformatics, natural language processing and machine learning.
Submitted: 1 October 2020; **Received (in revised form):** 5 November 2020

Vector Machine (SVM), kernel logistic regression and penalized logistic regression; DBS–PSSM [13] employs neural network to extract evolutionary information of protein sequences from the PSSMs to predict DNA-binding residues; RNABindR [14, 15] identifies and shows RNA-binding residues based on Naive Bayes classifier and protein structure in PDB; Pprint [16] predicts RNA-binding residues by combining evolutionary information features and SVM algorithm. Some other predictors with different classification algorithms or features have also been proposed for improving DNA-binding residue prediction [17, 18] and RNA-binding residue prediction [19, 20].

All these computational methods have greatly promoted the development of this field. However, they can only detect the DNA-binding residue or RNA-binding residue, and are suffering from cross-prediction problem (a DNA-binding residue predictor accurately detects the DNA-binding residues, but also cross-predicts many RNA-binding residues as DNA-binding residues, and vice versa) [5]. The reason is that DNA-binding residues and RNA-binding residues share some similar characteristics [5, 21]. If a DNA-binding residue predictor is only trained with DNA-binding proteins and does not consider the RNA-binding proteins, it can accurately predict the DNA-binding residues but also prefers to identify the RNA-binding residues as DNA-binding residues. For similar reason, an RNA-binding residue predictor trained only with RNA-binding proteins tends to predict the DNA-binding residues as RNA-binding residues. Although some methods can predict both DNA-binding residues and RNA-binding residues, such as BindN [22], NAPS [23], BindN+ [24], SVMnuc [10] and NucBind [10, 25], unfortunately, these methods still treat DNA-binding residue prediction and RNA-binding residue prediction as two separate tasks during their training processes. As a result, they are still suffering from cross-prediction problem.

To solve the cross-prediction problem, the predictor DRNApred using a two-layer strategy has been proposed to predict both DNA-binding residues and RNA-binding residues [8]. However, its prediction accuracy is significantly lower than that of SVMnuc and NucBind. Besides, most of existing methods treat DNA- or RNA-binding residue recognition as a classification task, the unit of training and test is residue. The protein sequences were segmented into fixed length subsequences, and each residue was represented by several neighbors before and after it. Although this classification method can achieve the purpose of identifying nucleic acid binding residues, it ignored the long-distance dependence (contextual information) among residues in a protein, resulting in limited prediction performance.

Therefore, the following two aspects should be revisited in this very important field: (i) cross-prediction problem. New frameworks should be explored to solve the cross-prediction problem and improve prediction accuracy by considering both the DNA and RNA-binding residues as a whole in their training and test processes; (ii) global and long dependencies among residues. The binding residues are typically located on the protein surface in clusters. However, the neighbor residues in protein structures could be far away from each other in their primary sequences. Therefore, the global and long-distance dependencies among residues in proteins should be considered when constructing the prediction models.

In this study, we introduced a new method called NCBRPred to predict both DNA-binding residues and RNA-binding residues in proteins based on the multilabel sequence labeling model (MSLM). Compared with the existing methods, NCBRPred has the following two advantages: (i) it treats the identification of DNA-binding residues and RNA-binding residues as a multilabel learning task by using both the DNA-binding proteins and RNA-binding proteins to train the model so as to reduce the cross-prediction rate; (ii) the sequence labeling model that can measure the global and long-distance dependencies among residues and capture the sequential characteristics of the nucleic acid binding residues were employed so as to improve prediction performance.

Methods

Datasets

Three benchmark datasets were used to train and evaluate different methods, including YK17 [8], YFK16-3.5 [5] and YFK16-5 [5]. YFK16-3.5 and YFK16-5 are two widely used benchmark datasets constructed by Yan *et al*. [5]. YK17 is an extension of YFK16-3.5 by adding new nucleic acid binding proteins [8]. Each of these three benchmark datasets contains two subsets, including a training set and an independent test set. The sequence similarity between any protein in the training set and any protein in the test set is less than 30%. The dataset MW15 [21] was also used to evaluate different methods as an independent dataset. In order to avoid overestimating the performance of our method, the proteins sharing more than 25% sequence similarity with any protein in MW15 were removed from the benchmark dataset YFK16-5 by using BLASTClust [26], and the proposed model was trained with the refined YFK16-5 to predict the proteins in MW15. The statistical information of these four datasets is summarized in Table 1.

Protein representation

Nucleic acid binding residues are conservative during evolution process, and nucleic acids usually interact with proteins on the protein surface [5]. Therefore, two kinds of evolutionary profiles (PSSM and Hidden Markov Model (HMM) profile), the predicted secondary structure (SS), and the predicted solvent accessibility (SA) were used to represent the proteins in this study. The performance of different feature combinations on the performance of the proposed method will be discussed in the 'Results and Discussion' section.

The PSSMs were generated by using PSI-BLAST [27] with parameters of '-num_iterations 3 -evalue 0.001' to search against the nonredundant database NRDB90 [28]. Each element e in the PSSM profile was normalized by:

$$p_{i,j} = \frac{e_{i,j} - u_i}{s_i} \tag{1}$$

where i and j indicate the row index and the column index of element in PSSM profile, respectively, u_i and s_i are the mean and the standard deviation of each row in PSSM, respectively, which can be calculated by following equations:

$$u_i = \frac{1}{n}\sum_{j=1}^{n} e_{i,j} \tag{2}$$

$$s_i = \sqrt[2]{\frac{1}{n}\sum_{j=1}^{n}\left(e_{i,j} - u_i\right)^2} \tag{3}$$

where n is the number of standard amino acids. The dimension of PSSM-based features is $L \times 20$ and L is the length of the protein.

The HMM profiles were generated by using HHblits [29] to search against the database uniprot20_2016_02 with default

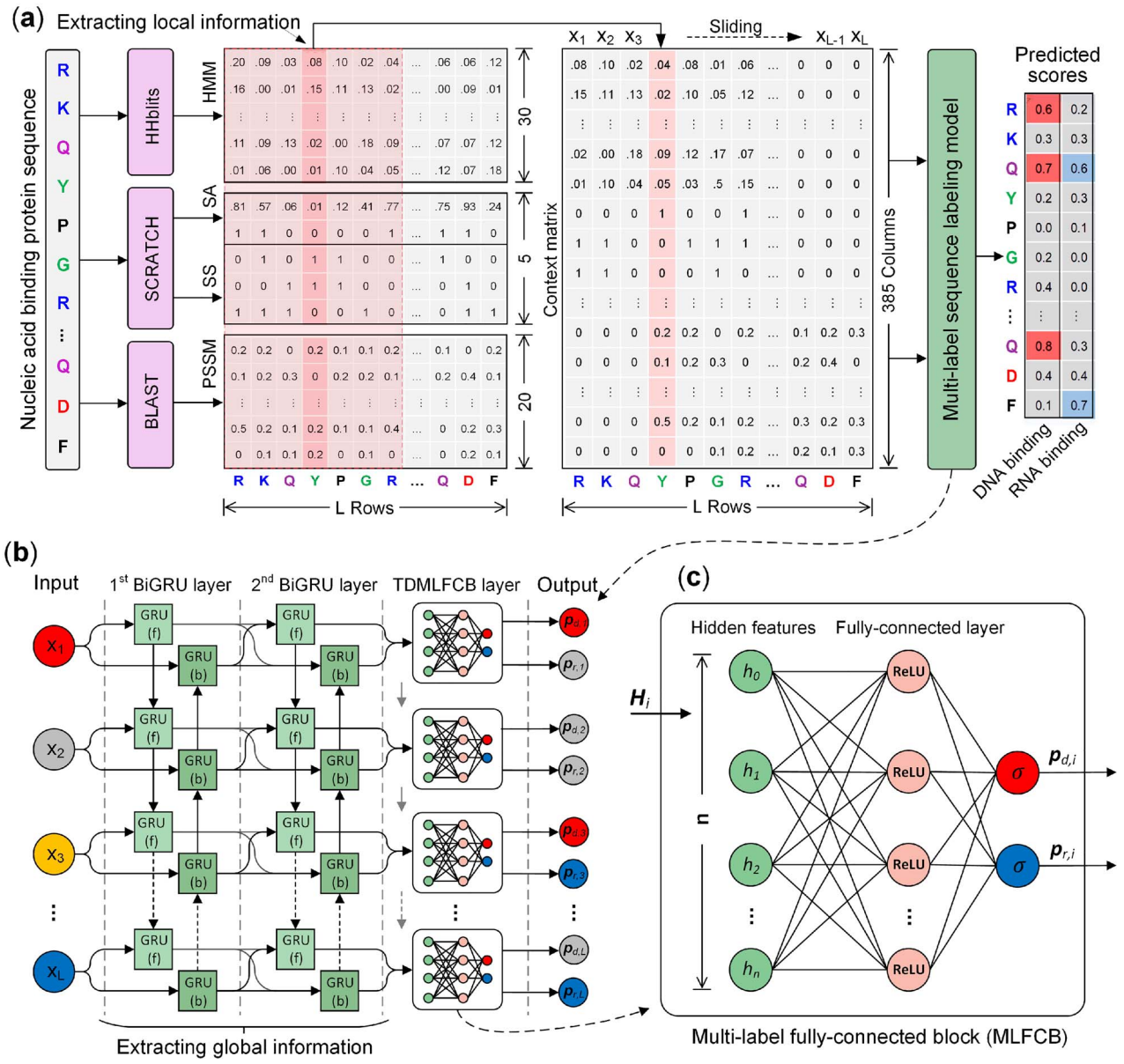

Figure 1. The framework and architecture of NCBRPred. (**A**) The overall framework of NCBRPred. Both DNA-binding proteins and RNA-binding proteins are fed into NCBRPred for training and test. The sliding window strategy was used to capture the local dependencies among residues in a protein. (**B**) The network architecture of MSLM. It contains three layers, including two BiGRU layers and a TDMLFCB layer. The two BiGRU layers measure the correlations among residues along the protein in a global fashion so as to capture the long and short distance dependencies among residues. The TDMLFCB layer predicts DNA-binding residues and RNA-binding residues based on the learned hidden features by the former two BiGRU layers. The red, blue, orange and gray circles in the input layer represents DNA-binding residue, RNA-binding residue, DNA and RNA-binding residue, and non-DNA/RNA-binding residue, respectively. (**C**) The network architecture of MLFCB. It integrates the predictive results for binding residues via the multilabel learning strategy trained with both DNA and RNA-binding residues, leading to lower cross-prediction rate.

parameters. Each element e in the HMM profile was normalized by:

$$f_{i,j} = \begin{cases} 2.0^{-0.001 \times e_{i,j}} & \text{if } e_{i,j} \neq * \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

where * represents an infinite integer in HMM profile; i and j indicate the row index and the column index of element in HMM profile, respectively. The HMM profile contains L rows and 30 columns. L is the length of the query protein sequence. The first 20 columns indicate the frequencies of the 20 standard amino acids in the corresponding positions, and the last 10 columns contain the seven transition frequencies and three local diversities in the corresponding position. For more information of HMM profile, please refer to [30]. Therefore, the dimension of HMM-based features is $L \times 30$.

The SS and SA were predicted by SSpro and ACCpro [31], respectively. The predicted SS provides the information of three class SSs, including helix, strand and other, whose dimension is $L \times 3$. Two predicted SA profiles are the binary SA and the score of SA. The element in the profile of SA score was normalized by $s \times 0.01$, where s represents the SA score. The dimensions of the combination of two predicted SA profiles are $L \times 2$.

By merging these four different features, the protein was represented as a matrix with dimension of $L \times 55$. Then each protein

Table 1. The statistical information of the four datasets used in this study

Dataset	Training set				Test set			
	Protein chains	Residues			Protein chains	Residues		
		D[a]	R[b]	N[c]		D[a]	R[b]	N[c]
YK17 [8]	488	7764	4684	90 594	82	955	807	17 119
YFK16-3.5 [5]	467	6932	4647	88 508	64	875	409	13 679
YFK16–5 [5]	469	10 848	6941	82 811	65	1452	648	12 927
MW15 [21]	NA	NA	NA	NA	46	760	368	9447

[a]DNA-binding residues.
[b]RNA-binding residues.
[c]Residues neither bind to nucleic acid residue nor are disordered residue.

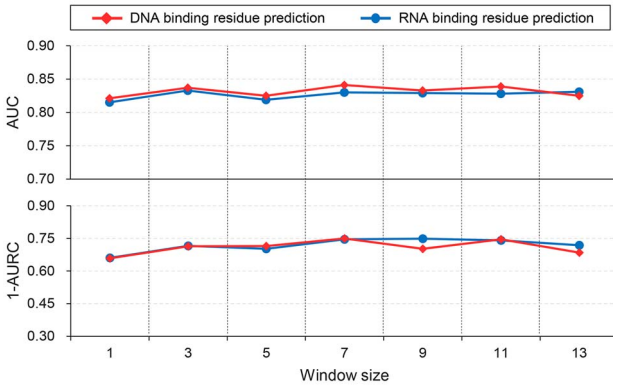

Figure 2. The performance of NCBRPred with different window sizes on the training set of YK17 via 5-fold cross-validation. For the clearer results, we used 1-AURC instead of AURC to show the trend of the cross-prediction of NCBRPred under different window sizes.

was converted into a context matrix by using a sliding window approach (see Figure 1A) with a window size of 7 optimized by 5-fold cross-validation on the training dataset of YK17. Therefore, the final dimension of the context matrix is $L\times385$.

Architecture of NCBRPred

Sequence labeling models, especially for the models based on recurrent neural networks, are able to measure the global and long-distance dependencies among residues in protein, which usually were ignored by classification models. They have been proven to be more effective than classification models in the identifications of special sites or regions in proteins [32–34]. Inspired by these studies, we applied recurrent neural networks to nucleic acid binding residue identification. In this study, two kinds of sequence labeling models were tested, including models based on Long Short-Term Memory (LSTM) [35] and models based on Gated Recurrent Unit (GRU) [36]. GRU-based models showed better performance for identifying nucleic acid binding residues than LSTM-based models, and GRU-based models are more efficient than LSTM-based models during training and test processes. This is because GRU has a simpler network structure with fewer parameters compared with LSTM, which is more suitable for the task with fewer samples. The performance of different models will be discussed in the 'Results and Discussion' section.

In order to reduce the complexity of the prediction model and the risk of overfitting, we used the GRU to construct the MSLM. It contains three layers (see Figure 1B), including two bidirectional GRU layers (BiGRU) and a Time-Distributed MultiLabel Fully-Connected Block (TDMLFCB) layer. The two BiGRU layers measure the correlations among residues along the protein in a global fashion so as to capture the long and short distance dependencies among residues [37]. The output dimension of the memory cell in the 1st BiGRU layer is 32, and it is 40 in the 2nd BiGRU layer, which were selected through grid search by considering both predictive performance and calculating cost. The TDMLFCB layer predicts DNA-binding residues and RNA-binding residues based on the learned hidden features by the former two BiGRU layers. The Multi-Label Fully-Connected Block (MLFCB) in the 3rd layer contains two sublayers (see Figure 1C). The 1st sublayer consists of 80 neurons using rectified linear unit (ReLU) [38] as activation functions to learn potential patterns or associations among hidden features, and takes both DNA-binding resides and RNA-binding residues into consideration so as to reduce the cross-prediction rate. The 2nd sublayer contains two neurons with sigmoid function to identify DNA-binding residues and RNA-binding residues, respectively. The dropout strategy was employed between the two sublayers to further avoid the overfitting problem. The dropout rate was 0.5. The MLFCB shares its weights at each time step. The lightweight structure of MSLM designed in current work makes it a multi-layer neural network rather than a deep neural network, making it suitable for analyzing the current data.

In this study, we used Keras and TensorFlow [39] to construct the proposed model. To process proteins with various lengths, the masking technique was employed by NCBRPred. For sequences shorter than the longest protein in a dataset, their corresponding context matrices were padded with zero values. The padded parts were masked during subsequent operations. Because the lack of coordinate information for the residues in disordered regions, the disordered residues were not annotated binding when constructing datasets [5, 8]. To utilize the complete sequence information of a protein, the proposed model processes the whole protein sequence including disordered residues. But the disordered residues were not considered when calculating the loss during the training process. The loss was calculated by:

$$loss = \frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{N} l_{i,j} \tag{5}$$

$$l_{i,j} = \begin{cases} y_{i,j}\times\log\left(p_{i,j}\right) + \left(1-y_{i,j}\right)\times\log\left(1-p_{i,j}\right) & y_{i,j}\geq 0 \\ 0 & y_{i,j} < 0 \end{cases} \tag{6}$$

where N represents the total number of residues in the training data; $y_{i,j}$ is the true label for the i-th label of the j-th residue; $p_{i,j}$ is the predicted score for the i-th label of the j-th residue. For the disorder residues in training datasets, the true labels were set as -1.

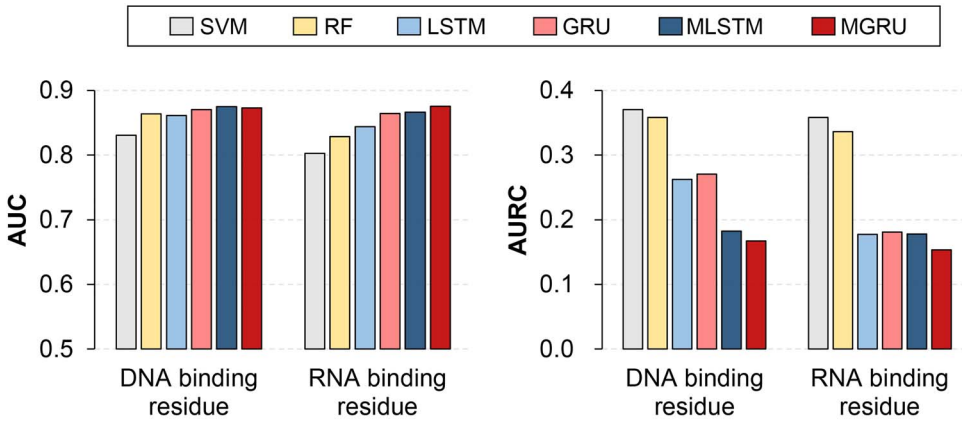

Figure 3. Comparison of six models for detecting DNA-binding residues and RNA-binding residues on training set of YK17 via 5-fold cross-validation. The MLSTM and MGRU are multilabel sequence labeling models. The LSTM and GRU are similar as MLSTM and MGRU except that they are not multilabel classifiers but binary classifiers. For SVM, the radial basis function (rbf) was employed as the kernel with optimal parameters of 'C = 3, *gamma* = 0.001'. For RF, the optimal number of estimators was 250.

Table 2. Performance of NCBRPred based on different features and their combinations on the training set of YK17

Features	DNA-binding residue prediction		RNA-binding residue prediction	
	AUC	AURC	AUC	AURC
One-hot	0.55	0.56	0.62	0.43
PSSM	0.82	0.25	0.81	0.30
HMM	0.82	0.31	0.79	0.39
PSSM + SA	0.82	0.28	0.81	0.32
PSSM + SS	0.83	0.25	0.82	0.25
PSSM + SS + SA	0.83	0.25	0.82	0.29
HMM + SS + SA	0.83	0.33	0.80	0.38
PSSM + SS + SA + HMM	**0.85**	**0.22**	**0.84**	**0.24**
PSSM +SS + SA + HMM + One-hot	0.84	0.22	0.82	0.27

Note: The best results are highlighted in bold type.

Performance evaluation

In this study, two evaluation metrics were used to evaluate the performance of different methods, including the area under the ROC curve (AUC) and the area under the cross-prediction rate–true positive rate (CPR–TPR) curve (AURC). The CPR–TPR plot uses the TPR on the *x*-axis and the CPR on the *y*-axis. CPR is the ratio of DNA-binding residues incorrectly predicted as RNA-binding residues or the ratio of RNA-binding residues incorrectly predicted as DNA-binding residues [5, 8]. TPR is defined as the fraction of DNA-binding residues correctly predicted as DNA-binding residues or the fraction of RNA-binding residues correctly predicted as RNA-binding residues. The CPR–TPR curve shows the dynamic change of CPR under different TPRs of a predictor. Under the same TPR, the lower CPR is, the better the predictive performance is. The AUC reflects the overall predictive performance of a predictor. The higher the AUC is, the better the predictive performance is. The AURC reflects the cross-prediction problem of a predictor. The lower the AURC is, the better predictive performance is.

Results and discussion

The predictive performance of NCBRPred based on different features and their combinations

We explored the impact of different features and their combinations on the performance of NCBRPred by using 5-fold cross-validation on the training dataset of YK17. The corresponding features and results were listed in Table 2. From it we can see that the NCBRPred based on the combination of PSSM, SS, SA and HMM achieved the best performance, and its performance decreased when adding the one-hot encoding as an additional feature. These results are not surprising because these four features (PSSM, SS, SA and HMM) describe the protein sequences in different aspects, and they are complementary. The information of one-hot encoding is limited compared with PSSM or HHM (refer to Table 2). As a result, performance improvement cannot be observed when adding the one-hot encoding as an extra feature.

Based on the above analysis, the final NCBRPred predictor was constructed based on four features, including PSSM, SS, SA and HMM. These features were also proven useful for predicting nucleic acid binding residues in some previous studies [8, 10, 40, 41].

Impact of the window sizes on the predictive performance of NCBRPred

As introduced in the section of protein representation, a sliding window approach was employed to generate the context matrix, which incorporates the features of the neighbor residues along the protein sequences [42]. In this section, the impact of the window size on prediction performance was investigated by using 5-fold cross-validation on the training dataset of YK17. Experimental results were shown in Figure 2. For the better visualization, we used 1-AURC instead of AURC to show the trend of the cross-prediction results of NCBRPred with different window sizes. We can see from Figure 2 that the proposed method

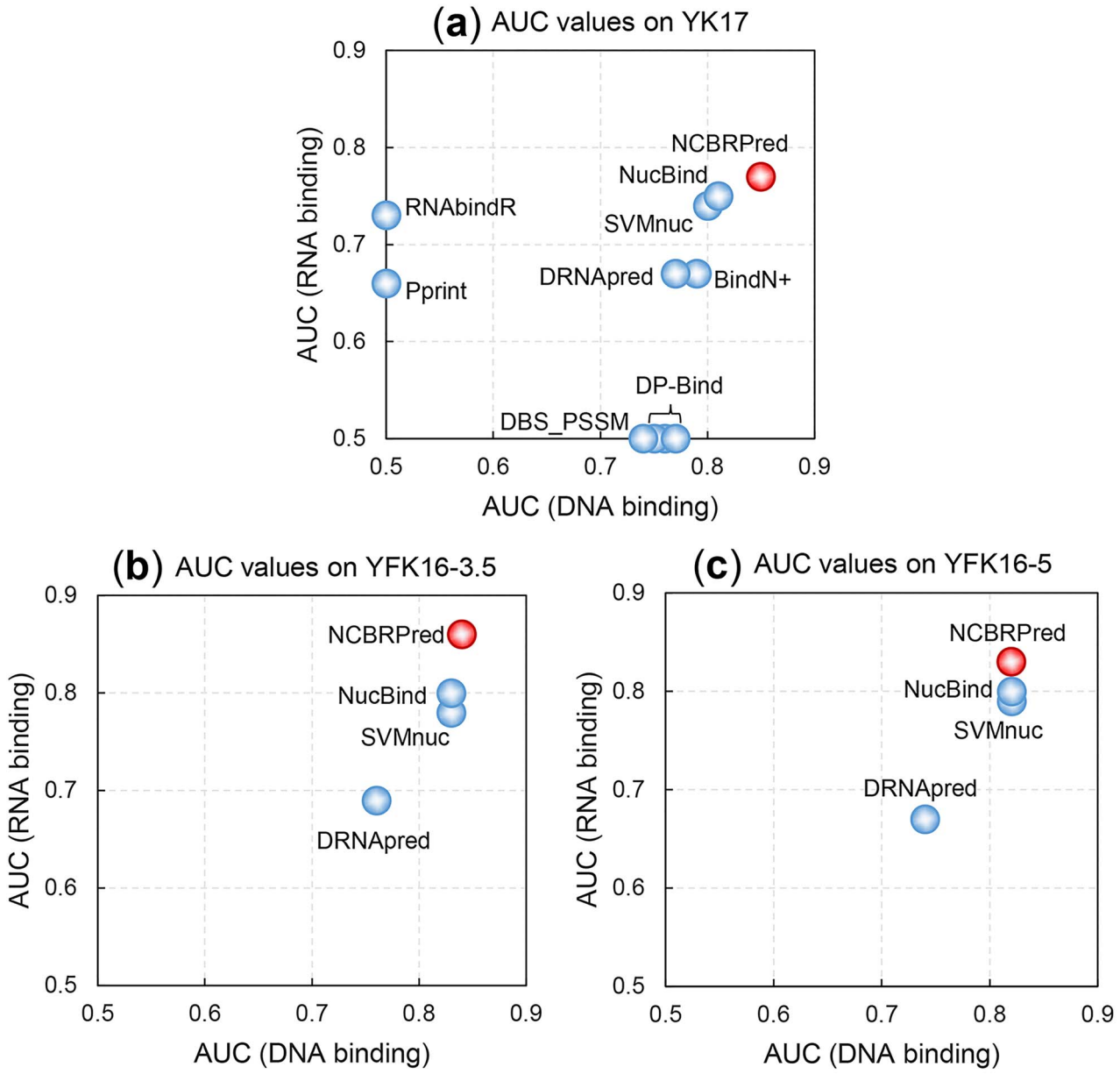

Figure 4. Comparison of different predictors for detecting DNA-binding residues and RNA-binding residues on three datasets. (**A**) The results of different methods on the test sets of YK17. (**B**) The results of different methods on the test sets of YFK16-3.5. (**C**) The results of different methods on the test sets of YFK16-5.

showed similar performance in terms of both AUC and 1-AURC. The slight difference would be caused by random initialization of the weights of neural networks. In general, NCBRPred achieved relatively stable performance with different window sizes, and it achieved the best performance with window size of 7. The reason is that NCBRPred is based on the MSLM, which is a sequence labeling framework to model the protein sequence in a global fashion. Therefore, it is not sensitive with the window size. Therefore, the window size was set as 7 considering both the predictive performance and computational cost.

Comparison of different models

In order to investigate whether the sequence labeling models are more suitable for nucleic acid binding residue identification than the classification models, and whether the multilabel learning strategy can reduce the cross-prediction rate or not, we compared different machine learning models on the training set of YK17 via 5-fold cross-validation, including classification model based on SVM, classification model based on Random Forest (RF), and the sequence labeling model based on LSTM, sequence labeling model based on GRU, multilabel sequence labeling model based on LSTM (MLSTM), and multilabel sequence labeling model based on GRU (MGRU). The sequence labeling model is similar as the MSLM except that it is not a multilabel model but a binary model. The predictive results of these six predictors were shown in Figure 3, from which we can see the followings: (i) the four sequence labeling models outperformed the two classification models (SVM and RF), indicating that the correlations among residues along

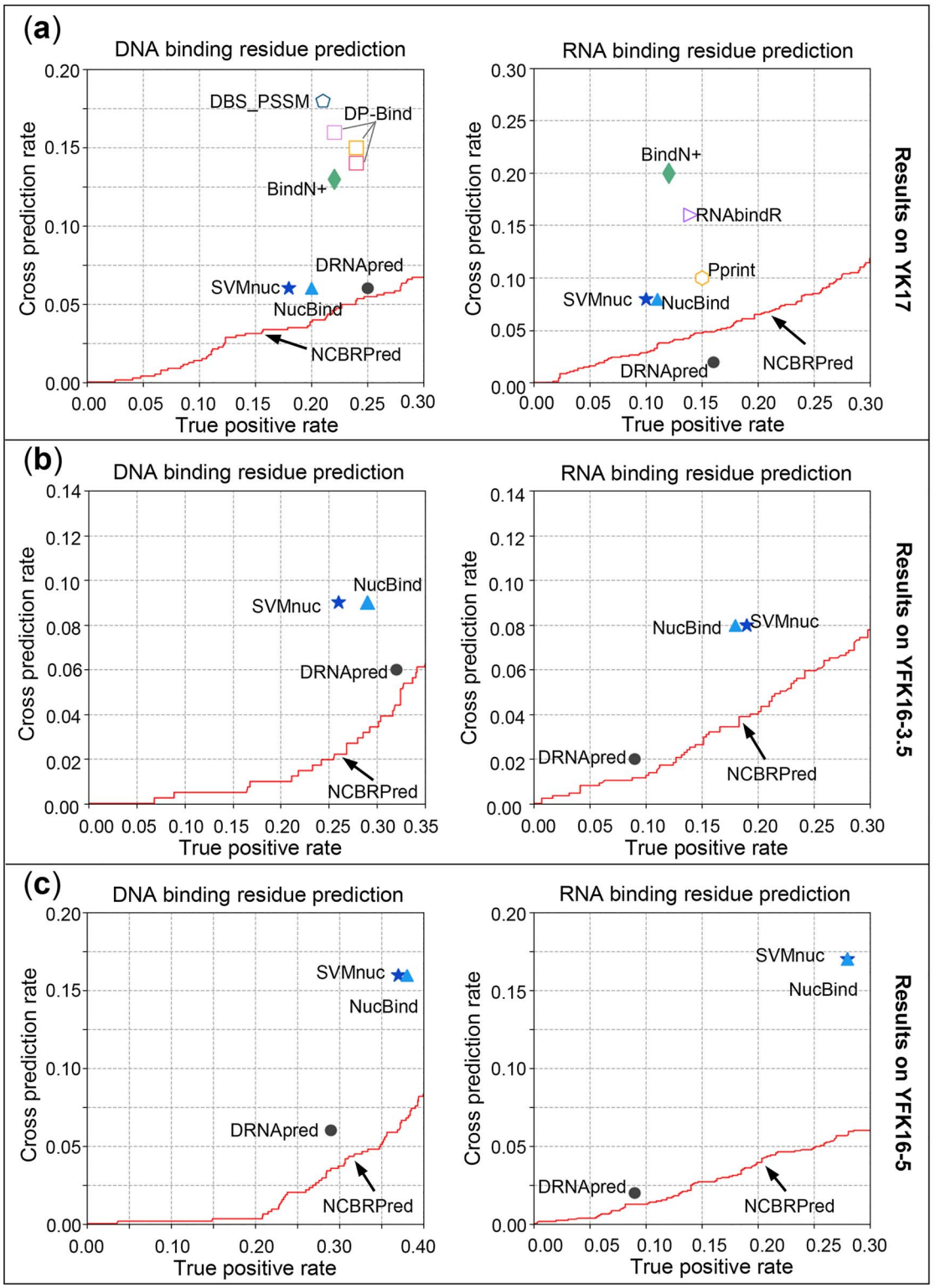

Figure 5. The CPR–TPR curves of NCBRPred and comparison with other competing methods. (**A**) DNA-binding residue prediction and RNA-binding residue prediction on the test set of YK17. (**B**) DNA-binding residue prediction and RNA-binding residue prediction on the test set of YFK16-3.5. (**C**) DNA-binding residue prediction and RNA-binding residue prediction on the test set of YFK16-5.

the protein sequences measured by the sequence labeling models are useful for reflecting the patterns of nucleic acid binding residues (clusters in the 3D structures of proteins); (ii) the two MSLMs outperformed the two sequence labeling models, especially in terms of AURC. These results are not surprising because the MSLM employed the multilabel framework to use the information of both the RNA-binding residues and DNA-binding residues, and therefore, it avoids the

Table 3. The experimental results of different methods on the independent dataset MW15

Type	Method	AUC	AULC	AURC	AURLC
DNA-binding residue	DRNApred[a]	0.72	0.010	0.48	0.136
	SVMnuc[b]	0.83	0.026	0.45	0.086
	NucBind[b]	**0.83**	0.026	0.40	0.086
	NCBRPred	0.81	**0.027**	**0.31**	**0.048**
RNA-binding residue	DRNApred[a]	0.47	0.001	0.50	0.118
	SVMnuc[b]	0.79	0.005	0.41	0.071
	NucBind[b]	0.79	0.005	0.34	0.041
	NCBRPred	**0.80**	**0.006**	**0.20**	**0.022**

Note: The best results are highlighted in bold type.
[a]The results were calculated by using the web-server of DRNApred.
[b]The AUC and AULC were reported in [10], the AURC and AULRC were calculated by the web-server of NucBind.

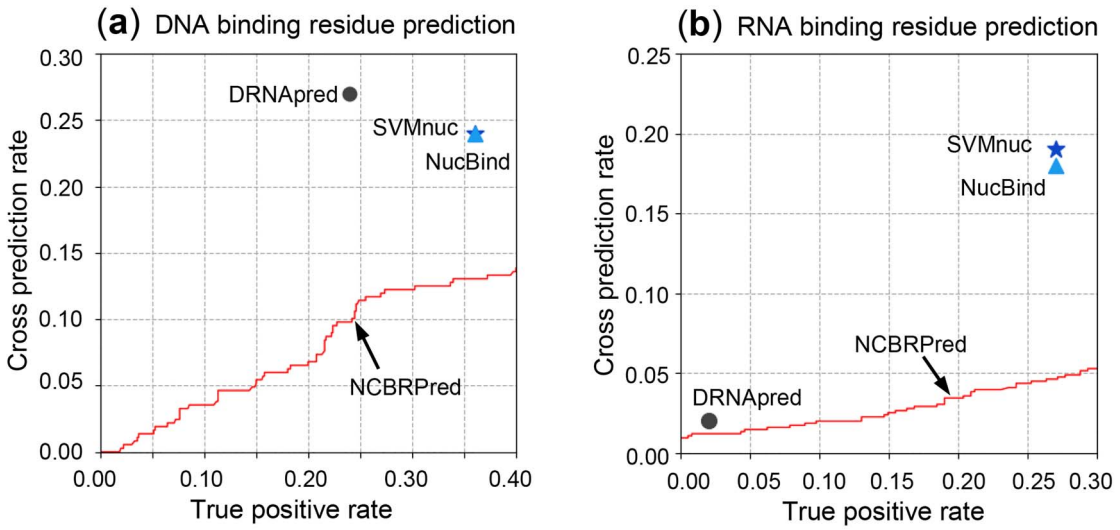

Figure 6. The comparison of NCBRPred and other three competing methods on MW15.

cross-prediction problem; (iii) in general, GRU-based models showed better prediction performance than LSTM-based models. This indicates that GRU-based models are more suitable for the identification of nucleic acid binding residues than LSTM-based models. In addition, we also analyzed the training efficiency of the MLSTM and MGRU models. The MGRU model was 30 s faster than the MLSTM model for each training epoch evaluated on the same data (training set of YK17). This is because GRU has a simpler network structure with fewer parameters compared with LSTM.

Considering the predictive performance and computational cost, the GRU-based MSLM was finally used to identify nucleic acid binding residues in this study.

Performance comparison of various computational methods

The performance of NCBRPred was evaluated on the three widely used benchmark datasets, including YK17 [8], YFK16-3.5 [5] and YFK16-5 [5], and its performance was compared with 10 state-of-the-art methods, including DP-Bind(klr) [12], DP-Bind(svm) [12], DP-Bind(plr) [12], DBS_PSSM [13], Pprint [16], RNAbindR [14, 15], BindN+ [24], DRNApred [8], SVMnuc [10] and NucBind [10].

The AUCs of different methods on the three datasets were shown in Figure 4, from which we can see that NCBRPred outperformed the other 10 methods by incorporating both DNA-binding residue and RNA-binding residue into the predictive model.

In order to evaluate the cross-prediction of different methods, we plotted the CPR–TPR curve of NCBRPred, and compared it with other methods as shown in Figure 5. We can see that NCBRPred achieved the 2nd-best performance for predicting RNA-binding residues on test set of YK17 (see the right panel of Figure 5A), and it outperforms all the other competing methods on other five subsets. Interestingly, among these 11 methods, five methods (BindN+, DRNApred, SVMnuc, NucBind and NCBRPred) can predict both the DNA-binding residues and RNA-binding residues, but only DRNApred and NCBRPred were trained by using both DNA-binding resides and RNA-binding residues in training processes. As a result, these two methods outperformed the other nine methods in terms of cross-prediction. This indicates that considering both the DNA-binding residues and RNA-binding residues during the training process is able to decrease the cross-prediction.

From the results of above three datasets, we can find that the NCBRPred predictor is the only method achieves more accurate prediction results with lower cross-prediction. All the results indicate that the proposed MSLM is effective for nuclei acid binding residue prediction.

Performance of various methods on the independent dataset MW15

In order to further validate its performance, NCBRPred was evaluated on the independent dataset MW15. The results of

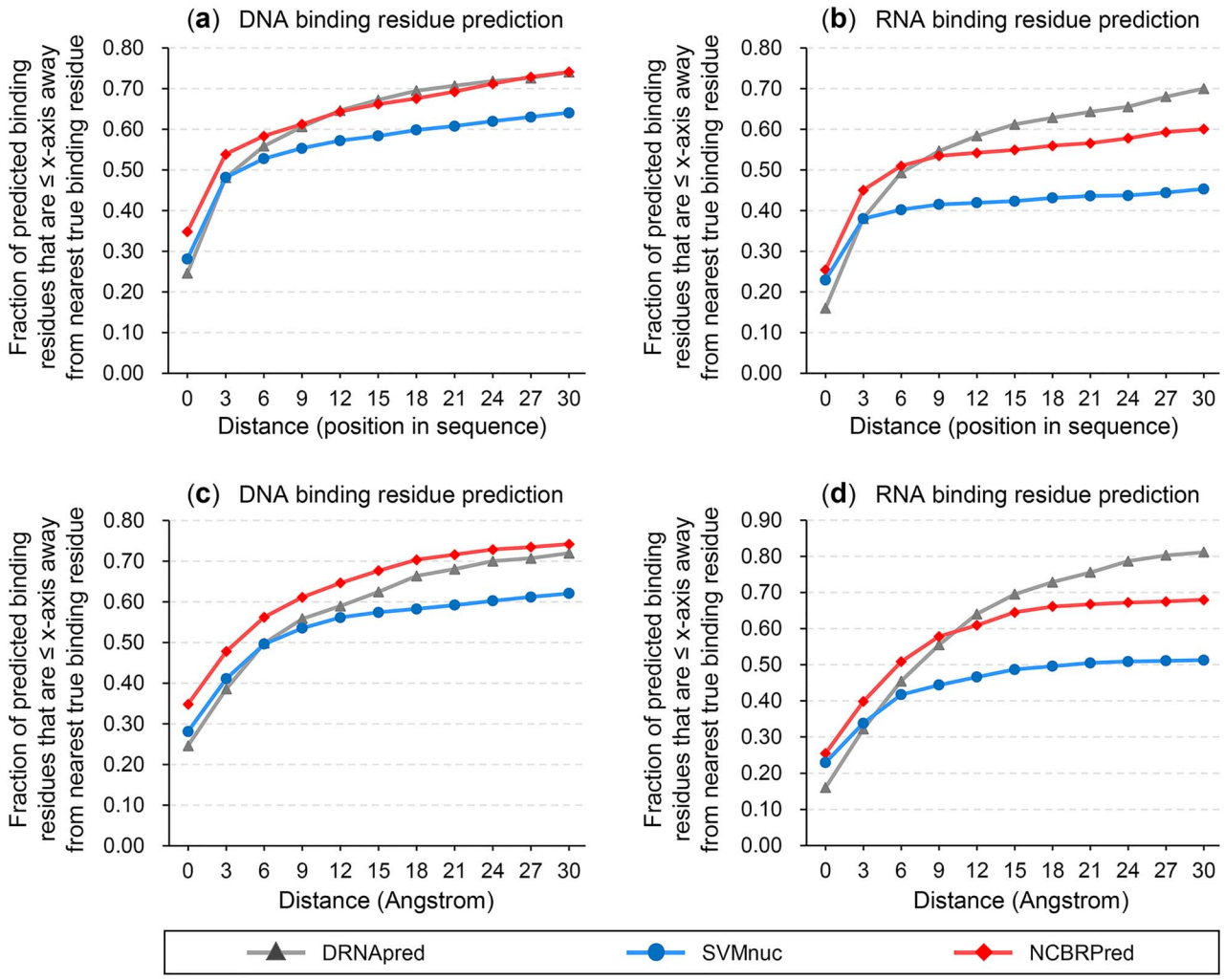

Figure 7. Analysis of the fraction of predicted nucleic acid binding residues shorter than a certain distance from the true nucleic acid binding residues. (**A**) The results of different methods for DNA-binding residue prediction in terms of the position distance in protein sequences. (**B**) The results of different methods for RNA-binding residue prediction in terms of the position distance in protein sequences. (**C**) The results of different methods for DNA-binding residue prediction in terms of the space distance in the protein structures. (**D**) The results of different methods for RNA-binding residue prediction in terms of the space distance in the protein structures. The fraction of predicted binding residues shorter than a certain distance from the true binding residues was defined as the count of predicted binding residues shorter than a certain distance from the true binding residues divided by the total number of the predicted binding residues. The data for curves of DRNApred and SVMnuc were obtained from their corresponding web-servers.

NCBRPred along with the other three best predictors (DRNApred, SVMnuc and NucBind) were shown in Table 3 and Figure 6.

The results show that NCBRPred outperformed the other three competing methods in terms of AURC, and it has lower CPR than other methods under the same TPR. As we can see the AUC value of NCBRPred is slightly lower than those of SVMnuc and NucBind for predicting DNA-binding residues. The reason is that MW15 is an imbalanced dataset and AUC usually overestimates a prediction method on an imbalanced dataset. The AULC and AURLC are the variants of AUC and AURC [8], which are more suitable for evaluating a method on the imbalanced dataset. In order to more objectively evaluate different methods, the AULCs and AURLCs of four methods were calculated on the independent dataset MW15, as listed in Table 3. The proposed method achieved the best AULC and AURLC among four competing methods. This demonstrates that NCBRPred outperforms three existing methods, and is a useful method for solving cross-prediction problem and predicting both DNA-binding residues and RNA-binding residues.

Analysis of the predicted nucleic acid binding residues

We also analyzed and compared the nucleic acid binding residues predicted by the top three predictors on the YK17 dataset, including NCBRPred, SVMnuc and DRNApred. Following the previous study [8], two kinds of distances were considered, including the number of amino acids between two residues in the protein sequences (position distance), and the Euclidian distance between two residues in the protein structures (space distance). For each predictor, we calculated the fraction of predicted nucleic acid binding residues shorter than a certain distance from the true nucleic acid binding residues. The fraction was defined as the count of predicted binding residues shorter than a certain distance from the true binding residues divided by the total number of the predicted binding residues. The results of these three predictors under different distance thresholds were shown in Figure 7. The results showed that NCBRPred outperformed the other two competing methods with distance shorter than 9, whereas DRNApred showed higher

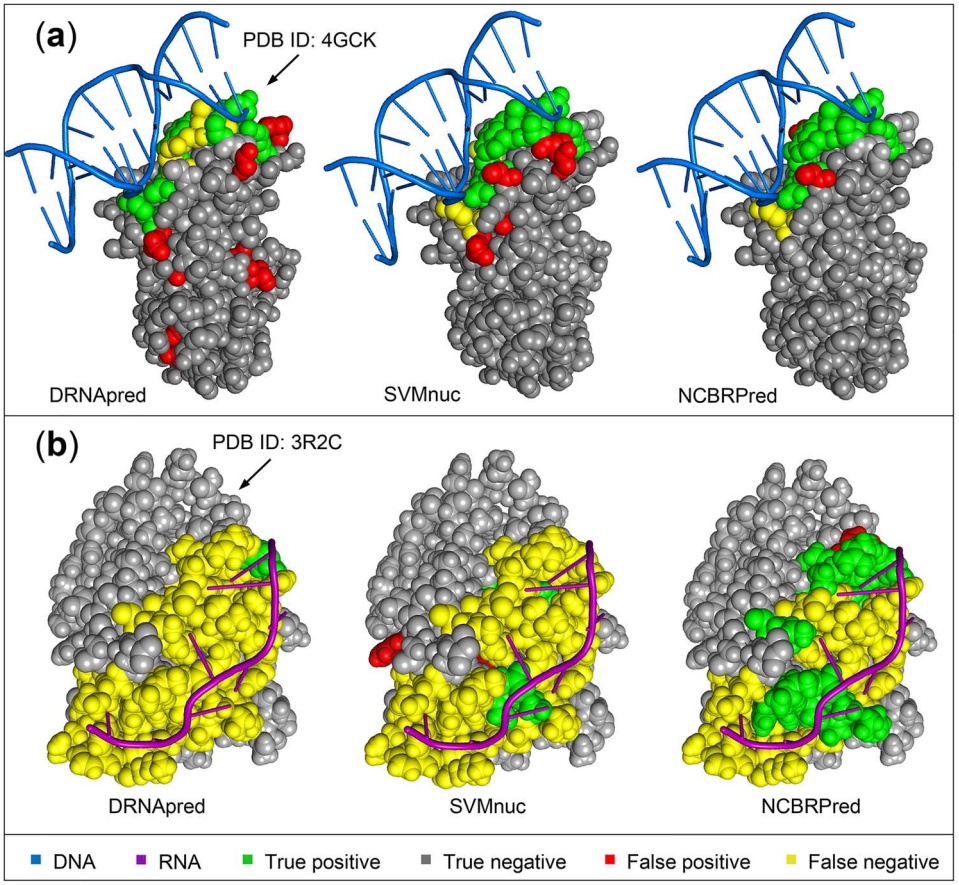

Figure 8. Visualization of the case study. (**A**) The prediction results of different methods in the DNA-binding protein 4GCK (PDB ID). (**B**) The prediction results of different methods in the RNA-binding protein 3R2C (PDB ID).

fraction for larger distances. These results indicated that most of the nucleic acid binding residues predicted by NCBRPred are near the true binding residues. In contrast, most of the nucleic acid binding residues predicted by the other two competing methods are far away from the true binding residues. The reason is that the sequence labeling model employed by NCBRPred is able to capture the global characteristics of residues along the protein (the nucleic acid binding residues prefer to occur in clusters in protein 3D structures). As a result, the false positive samples far away from the true nucleic acid binding residues obviously reduce. This point will be further discussed in the next section.

Predictive results visualization

The nucleic acid binding residues in a DNA-binding protein (PDB ID: 4GCK [43]) and an RNA-binding protein (PDB ID: 3R2C [44]) predicted by the top three predictors were visualized in Figure 8, including NCBRPred, SVMnuc and DRNApred. Both the two proteins were selected from the test set of YK17. From the figure we can see the followings: (i) the nucleic acid binding residues occur in clusters in the 3D structures of the proteins so as to interact with DNA sequence or RNA sequence; (ii) NCBRPred identified fewer false positives and false negatives than those predicted by SVMnuc and DRNApred; (iii) the false positives predicted by SVMnuc and DRNApred are far away from the true nucleic acid binding residues. In contrast, the false positives predicted by NCBRPred are obviously closer to the true positive samples, which is fully consistent with the results shown in the prior section.

Conclusion

In this study, we proposed a new computational predictor NCBRPred to predict both DNA-binding residues and RNA-binding residues in proteins by using multilabel learning framework and sequence labeling model. The NCBRPred predictor achieved the state-of-the-art predictive performance and overcame the cross-prediction problem suffered by many existing methods.

Different from traditional predictor, in this study, the nucleic acid binding residue prediction was treated as a multilabel sequence labeling task, and the NCBRPred is the 1st method to use multilabel learning framework to consider both DNA-binding residues and RNA-binding residues simultaneously during training process. As a result, the proposed predictor is able to reduce the cross-prediction between DNA-binding residues and RNA-binding residues. The sequence labeling model can measure the global dependencies among residues along a protein, ignored by traditional classification model. NCBRPred employed the sequence labeling model based on GRU with few parameters and simple structures, making it suitable for small-scale data analysis. Besides, we refined the traditional binary cross-entropy loss function to filter out of the disordered residues when calculating loss during the training process so as to utilize the precise sequence information of a protein.

Experimental results on three benchmark datasets and an independent dataset demonstrated the feasibility and effectiveness of the proposed method for identifying nucleic acid binding residues. The corresponding web-server and stand-alone package of NCBRPred are freely available at http://bliulab.net/NCBRPred. Besides, the new framework used in NCBRPred would also be applied to solve other tasks in bioinformatics, such as the prediction of the functional sites in protein intrinsic disorder region, the identification of the small ligand binding residues in proteins, the detection of DNA mutations, and etc. It is anticipated that NCBRPred will become a useful tool for identifying nucleic acid binding residues.

Key Points

- Interactions of proteins and nucleic acids are playing various crucial roles in cellular activities. Accurate identification of nucleic acid binding residues in proteins is significant for characterizing the interactions between proteins and nucleic acids.
- Most of the existing nucleic acid binding residue predictors are suffering from cross-prediction problem. As a result, some DNA-binding residues and RNA-binding residues were incorrectly predicted. Therefore, new computational predictors should be investigated so as to solve the cross-prediction problem and improve the predictive performance.
- A new computational method called NCBRPred was proposed to predict the nucleic acid binding residues in proteins, which considers both DNA-binding residues and RNA-binding residues, and the global distance dependencies among residues by using the multilabel sequence labeling model.
- Experimental results on three widely used benchmark datasets and an independent dataset showed that NCBRPred outperformed 10 existing state-of-the-art predictors.

Acknowledgments

The authors are very much indebted to the three anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of this paper.

Funding

National Key R&D Program of China (No. 2018AAA0100100), the National Natural Science Foundation of China (No. 61672184, 61822306, 61732012, and 61861146002), and Beijing Natural Science Foundation (No. JQ19019).

Conflict of interest

The authors declare that they have no competing interests.

Availability

The NCBRPred is freely available at http://bliulab.net/NCBRPred as a web-server and a stand-alone package.

References

1. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat Rev Genet* 2014;**15**:829–45.
2. Vaquerizas JM, Kummerfeld SK, Teichmann SA, *et al*. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 2009;**10**:252–63.
3. Jolma A, Yan J, Whitington T, *et al*. DNA-binding specificities of human transcription factors. *Cell* 2013;**152**:327–39.
4. Zhang J, Ma Z, Kurgan L. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief Bioinform* 2017;**20**:1250–68.
5. Yan J, Friedrich S, Kurgan L. A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief Bioinform* 2016;**17**:88–105.
6. Liu B. BioSeq-analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief Bioinform* 2019;**20**:1280–94.
7. Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res* 2019;**47**:e127.
8. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res* 2017;**45**:84.
9. Berman HM, Westbrook JD, Feng Z, *et al*. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–42.
10. Su H, Liu M, Sun S, *et al*. Improving the prediction of protein–nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. *Bioinformatics* 2019;**35**:930–6.
11. Qu K, Wei L, Zou Q. A review of DNA-binding proteins prediction methods. *Current Bioinformatics* 2019;**14**:246–54.
12. Hwang S, Gou Z, Kuznetsov IB. DP-bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics* 2007;**23**:634–6.
13. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics* 2005;**6**:33.
14. Terribilini M, Lee J, Yan C, *et al*. Prediction of RNA binding sites in proteins from amino acid sequence. *RNA* 2006;**12**:1450–62.
15. Terribilini M, Sander JD, Lee J, *et al*. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res* 2007;**35**:578–84.
16. Kumar M, Gromiha MM, Raghava GPS. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins* 2008;**71**:189–94.
17. Chu W, Huang Y, Huang C, *et al*. ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors. *Nucleic Acids Res* 2009;**37**:396–401.
18. Chen YC, Wright JD, Lim C. DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry. *Nucleic Acids Res* 2012;**40**:249–56.
19. Liu Z, Wu L, Wang Y, *et al*. Prediction of protein–RNA binding sites by a random forest method with combined features. *Bioinformatics* 2010;**26**:1616–22.
20. Yoichi M, Spriggs RV, Haruki N, *et al*. PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences. *Nucleic Acids Res* 2010;**38**:W412–6.
21. Miao Z, Westhof E. A large-scale assessment of nucleic acids binding site prediction programs. *PLoS Comput Biol* 2015;**11**:e1004639.

22. Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res* 2006;**34**:243–8.
23. Carson MB, Langlois R, Lu H. NAPS: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Res* 2010;**38**:431–5.
24. Wang L, Huang C, Yang MQ, *et al*. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst Biol* 2010;**4**:1–9.
25. Wu Q, Peng Z, Zhang Y, *et al*. COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Res* 2018;**46**:W438–42.
26. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
27. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
28. Holm L, Sander C. Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* 1998;**14**:423–9.
29. Remmert M, Biegert A, Hauser A, *et al*. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2012;**9**:173–5.
30. Steinegger M, Meier M, Mirdita M, *et al*. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 2019. doi: 10.1186/s12859-019-3019-7.
31. Magnan CN, Baldi P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* 2014;**30**:2592–7.
32. Hanson J, Yang Y, Paliwal K, *et al*. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* 2016;**33**:685–92.
33. Liu Y, Wang X, Liu B. RFPR-IDP: reduce the false positive rates for intrinsically disordered protein and region prediction by incorporating both fully ordered proteins and disordered proteins. *Brief Bioinform* 2020. doi: 10.1093/bib/bbaa018.
34. Pan X, Rijnbeek P, Yan J, *et al*. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics* 2018;**19**:511.
35. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735–80.
36. Cho K, Van Merrienboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *In: Proc. Conference on Empirical Methods in Natural Language Processing*. 2014:1724–34.
37. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform* 2016;**18**:851.
38. Nair V, Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines. *In: Proc. International conference on machine learning*. 2010:807–14.
39. Abadi M, Agarwal A, Barham P, *et al*. TensorFlow: large-scale machine learning on heterogeneous distributed systems. *In: 12th USENIX Conference on Operating Systems Design and Implementation*. 2016:265–283.
40. Peng Z, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res* 2015;**43**:e121.
41. Peng Z, Wang C, Uversky VN, *et al*. Prediction of disordered RNA, DNA, and protein binding regions using DisoRDPbind. *Methods Mol Biol* 2017;**1484**:187–203.
42. Liu Y, Wang X, Liu B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief Bioinform* 2019;**20**:330–46.
43. Tonthat NK, Milam SL, Chinnam N, *et al*. SlmA forms a higher-order structure on DNA that inhibits cytokinetic Z-ring formation over the nucleoid. *Proc Natl Acad Sci USA* 2013;**110**:10586–91.
44. Stagno JR, Altieri AS, Bubunenko M, *et al*. Structural basis for RNA recognition by NusB and NusE in the initiation of transcription antitermination. *Nucleic Acids Res* 2011;**39**:7803–15.