OXFORD

Briefings in Bioinformatics, 2022, **23(4)**, 1–12

https://doi.org/10.1093/bib/bbac236

Advance access publication date: 17 June 2022

Problem Solving Protocol

iDRNA-ITF: identifying DNA- and RNA-binding residues in proteins based on induction and transfer framework

Ning Wang[†], Ke Yan[†], Jun Zhang and Bin Liu

Corresponding author. B. Liu, Beijing Institute of Technology, No. 5, South Zhongguancun Street, Haidian District, Beijing, 100081, China. Tel.: +86-010-68911310. E-mail: bliu@bliulab.net

[†]Ning Wang and Ke Yan contributed equally to this work.

Abstract

Protein-DNA and protein-RNA interactions are involved in many biological activities. In the post-genome era, accurate identification of DNA- and RNA-binding residues in protein sequences is of great significance for studying protein functions and promoting new drug design and development. Therefore, some sequence-based computational methods have been proposed for identifying DNA- and RNA-binding residues. However, they failed to fully utilize the functional properties of residues, leading to limited prediction performance. In this paper, a sequence-based method iDRNA-ITF was proposed to incorporate the functional properties in residue representation by using an induction and transfer framework. The properties of nucleic acid-binding residues were induced by the nucleic acid-binding residue feature extraction network, and then transferred into the feature integration modules of the DNA-binding residue prediction network and the RNA-binding residue prediction network for the final prediction. Experimental results on four test sets demonstrate that iDRNA-ITF achieves the state-of-the-art performance, outperforming the other existing sequence-based methods. The webserver of iDRNA-ITF is freely available at http://bliulab.net/iDRNA-ITF.

Keywords: induction and transfer framework, DNA- and RNA-binding residue identification, nucleic acid-binding residue identification, convolutional attention neural network

Introduction

Proteins and nucleic acid (DNA or RNA) interactions are involved in many biological processes, such as regulation of gene expression, signal transduction, post-transcriptional modification and regulation, etc. [1–4]. Therefore, accurate identification of DNA- and RNA-binding residues is essential in designing novel drugs and studying protein and nucleic acid interaction mechanisms [5]. Many wet-lab experimental methods were employed to detect DNA- and RNA-binding residues in proteins, such as nuclear magnetic resonance spectroscopy and X-ray [6]. However, they are time-consuming and relatively expensive. In this regard, it is very important to develop accurate and low-cost computational methods for large-scale screening of DNA- and RNA-binding residues in proteins [2, 7–9].

The computational methods include sequence- and structure-based methods. The sequence-based methods identify the functional sites from the protein sequences, such as TargetS [10], RNABindRPlus [11], TargetDNA [5], DNAPred [12], DRNAPred [1], SVMnuc [13], NCBR-Pred [6], etc. They are based on various sequence-derived features, including evolutionary information, physicochemical properties and predicted secondary structures (SS). iDeepMV [14] utilizes several multi-view features, including amino acid sequence view and dipeptide component view. PreRBP-TL [15] predicts the specific RNA-binding proteins by using the evolutionary information. The structure-based methods identify the functional sites from the three-dimensional structure of a protein. Structure-based methods additionally use the spatial features extracted from the structures to identify the DNA- and RNA-binding residues compared with the sequence-based methods, such as aaRNA [16], NucleicNet [17], DNABind [18], GraphBind [2], etc. The performance of the structure-based method is generally better than that of the sequence-based method. This is because the structures can provide spatial features that are more closely related to functions. With the development of sequencing technology, more and more protein sequences with unknown functions and structures should be analyzed, the applicability of the structure-based method is limited. Although some protein structure prediction tools, such as Modeller [19], trRosseta [20] and AlphaFold [21] can predict the three-dimensional structures based on the protein sequences,

Ning Wang is a master student at the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. Her research interests include bioinformatics, natural language processing and machine learning.

Ke Yan is a postdoctoral researcher at the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. His expertise is in bioinformatics and machine learning.

Jun Zhang is a doctoral candidate at the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Guangdong, China. His research interests include bioinformatics, natural language processing and machine learning.

Bin Liu is a professor at the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. His expertise is in bioinformatics, natural language processing and machine learning.

Received: March 15, 2022. **Revised:** May 6, 2022. **Accepted:** May 20, 2022

the differences between the predicted structures and the real structures still exist. The structure-based methods often failed to accurately predict the DNA- and RNA-binding residues in proteins without known structures. For example, GraphBind [2] is the state-of-the-art structure-based method for the task of identifying DNA- and RNA-binding residues in proteins with known structures, but its performance decreased obviously when it is only based on the predicted protein structures.

Currently, sequence-based methods usually use feature extraction tools to obtain sequence-derived features, such as the amino acid information, etc. The extraction process of these features fails to consider the functional properties of the residues. However, rich residue representation is critical for improving model performance. Therefore, we designed induction and transfer framework to incorporate the functional properties in residue representation to identify DNA- and RNA-binding residues. In the induction phase, we introduced the nucleic acid-binding function properties of residues, related to both DNA- and RNA-binding residues. Nucleic acid-binding residue training signal was used to guide the network to learn nucleic acid-binding residue features. In the transfer stage, the inductive features and the features learned by the other modules were fused, and the DNA-binding residue training signal and the RNA-binding residue training signal were respectively used to guide the learning to obtain task-transfer features and make the final prediction. The ablation experiments showed that the different modules in the induction and transfer framework are complementary. The experimental results showed that iDRNA-ITF achieves the best performance among the sequence-based methods, and outperforms the structure-based methods when identifying DNA- and RNA-binding residues in proteins without known structures.

Materials and methods

Datasets

The datasets used in this study were constructed by Graphbind [2], including DNA-573_Train, RNA-495_Train, DNA-129_Test and RNA-117_Test. All the proteins were collected from the BioLiP database [22]. The sequence similarity of protein chains in each training set is <30%, and the sequence similarity between the protein chains in the test set and the protein chains in the training set is <30% [2]. For better training models and evaluation methods for the performance of predicting DNA- and RNA-binding residues, we recombine these datasets because each dataset only contains a single type of binding proteins. The DNA-573_Train and RNA-495_Train were combined to construct a hybrid benchmark set DRNA-1068_Bench. The benchmark set was split into DRNA-962_Train (90%) and DRNA-106_Valid (10%), which were used for training the model and optimizing the parameters, respectively. The hybrid test DRNA-246_Test was constructed by combining DNA-129_Test and RNA-117_Test, it was used to evaluate the cross-prediction problem. The statistical information is listed in Table 1.

Empirical feature

In this study, a variety of feature extraction tools were used to extract the sequence-derived features, including evolutionary information, physicochemical properties and predicted SS.

We used Position-Specific Scoring Matrix (PSSM) [23], Position-Specific Frequency Matrix (PSFM) and Hidden Markov Model (HHM) to represent the evolution information of proteins. For a given protein sequence **P** with L amino acids, we used PSI-BLAST [24] with default parameters to search against the nrdb90 database [25] to obtain the PSSM profiles and PSFM profiles with the size of $L \times 20$. The sigmoid function [2] was used to normalize each element x in PSSM to the range [0,1]:

$$\bar{x} = \frac{1}{1 + e^{-x}} \tag{1}$$

The HHblits [26] with default parameters was employed to search against the uniclust30 database [27] to obtain the HHM profiles with the size of $L \times 30$. Each element e in HHM was normalized to the range of [0,1] by the following rule [6]:

$$\bar{e} = \begin{cases} 0.0, & \text{if } e \text{ is} * \\ 2.0^{-e \times 0.001}, & \text{otherwise} \end{cases} \tag{2}$$

The predicted structural features of amino acids were generated by SPIDER2 [28], including 8-dimensional SS, 2-dimensional CN and 4-dimensional HSE. In addition, seven physicochemical properties (SEVEN; [29]) including steric parameter, polarizability, volume, hydrophobicity, isoelectric point, helix probability and sheet probability were obtained. The element s in predicted SS and SEVEN were normalized to the range of [0,1] by the min–max scaling:

$$\bar{s} = \frac{s - s_{\min}}{s_{\max} - s_{\min}} \tag{3}$$

where $s_{\min}$ is the minimum values of each column in these features, $s_{\max}$ is the maximum values of each column in these features. Finally, for the jth residue, empirical features $h_j^{\mathrm{Emp}} \in \mathbb{R}^{1\times 91}$ were obtained by concatenating PSSM, PSFM, HHM, SS, CN, HSE and SEVEN.

Induction and transfer framework

A site in the sequences may have different molecular functions, which we called residues with different functional properties. Nucleic acid-binding function is associated with both DNA-binding function and RNA-binding function, and they have similar characteristics. Motivated by transfer learning, we designed an induction and transfer framework (see Figure 1) to extract and transfer

Table 1. The statistical information of the datasets

Dataset	DBRs[a]	RBRs[b]	Non-NABRs[c]	DBPs[d]	RBPs[e]
DNA-573_Train	14 479	0	145 404	573	0
RNA-495_Train	0	14 609	122 290	0	495
DNA-129_Test	2240	0	35 275	129	0
RNA-117_Test	0	2031	35 314	0	117
DRNA-1068_Bench	14 479	14 609	267 694	573	495
DRNA-962_Train	13 256	13 252	233 997	503	446
DRNA-106_Valid	1223	1357	33 697	50	49
DRNA-246_Test	2240	2031	70 589	129	117

[a] DBRs represent the DNA-binding residues. [b] RBRs represent the RNA-binding residues. [c] Non-NABRs represent the non-nucleic acid-binding residues. [d] DBPs represent the DNA-binding proteins. [e] RBPs represent the RNA-binding proteins.

nucleic acid-binding functional features to identify DNA-binding residues and RNA-binding residues. In the induction phase, the task of identifying nucleic acid-binding residues was used to induce the nucleic acid-binding features, which are related to both DNA- and RNA-binding residues. In the transfer phase, the inductive features and the features learned by the other modules were fused, and the DBR training signal and the RBR training signal were respectively used to guide the learning to obtain task-transfer features and make the final prediction. To develop the networks for the three residue recognition tasks, we used the 'one-versus-all' method [30] to transform the DRNA-962_Train and DRNA-106_Valid into three binary training and validation sets of the task network ($\mathbb{S}_{\mathrm{NAB}}^{\mathrm{Train}}$ and $\mathbb{S}_{\mathrm{NAB}}^{\mathrm{Valid}}$, $\mathbb{S}_{\mathrm{DB}}^{\mathrm{Train}}$ and $\mathbb{S}_{\mathrm{DB}}^{\mathrm{Valid}}$, $\mathbb{S}_{\mathrm{RB}}^{\mathrm{Train}}$ and $\mathbb{S}_{\mathrm{RB}}^{\mathrm{Valid}}$). For the nucleic acid-binding residue network, both the DBRs and RBRs were treated as positive samples. For the DNA-binding residue network and RNA-binding residue network, the DBRs and RBRs were respectively treated as positive samples. Through the induction of the common characteristics and transferring the common characteristics to downstream tasks with specific signals, the final prediction models can obtain better generalization performance.

Architecture of neural networks

The general network structures of the three tasks were shown in Figure 2A. We used feature extraction tools to extract residue-level empirical features from the protein sequences in the dataset, and then specific local features are learned through the convolutional attention module according to the training signals of different tasks, and the feature integration module is employed to concatenate empirical features and specific local features. The integrated features are input to the BiGRU module to extract sequence information, and then make classification by the fully connected layer. The networks of different tasks have differences in training signal, network parameters and feature integration module.

Convolutional attention module

If the protein sequence is treated as a text and the amino acid is treated as a character, the nucleic acid-binding residues can be considered as the named entities in the text. Therefore, the idea of named entity recognition (NER) and the other sequence labeling tasks in the natural language process can be applied to identify DNA- and RNA-binding residues in proteins. CAN-NER [31] was proposed in the task of Chinese named entity recognition. It mainly proposed a convolutional attention module to solve the problem of inaccurate Chinese word segmentation. It uses local attention to capture the relations of the central character and each context token in the window. Considering that DNA- and RNA-binding residue recognition task also suffers from the problem of no clear word-level segmentation between each residue, we learn from the convolutional attention module in CAN-NER to extract the local features of residues. The convolutional attention module was shown in Figure 2B.

For each window in the CNN, whose size is k. In each window, local attention is applied to capture the relationship between the central residue and each context residue in the window, and then the sum-pooling strategy is used to extract the local features. For the jth residue, the local attention output h_m is calculated as follows [31]:

$$h_m = \alpha_m h_m^{\mathrm{Emp}} \tag{4}$$

where $m \in \left\{j - \frac{k-1}{2}, \ldots, j + \frac{k-1}{2}\right\}$, α_m is the attention distribution, which is calculated as [31]:

$$\alpha_m = \frac{\exp\left(s\left(h_j^{\mathrm{Emp}}, h_m^{\mathrm{Emp}}\right)\right)}{\sum_{n \in \left\{j - \frac{k-1}{2}, \ldots, j + \frac{k+1}{2}\right\}} \exp\left(s\left(h_j^{\mathrm{Emp}}, h_n^{\mathrm{Emp}}\right)\right)} \tag{5}$$

$s\left(h_j^{\mathrm{Emp}}, h_m^{\mathrm{Emp}}\right)$ is the attention scoring function, which is calculated as [31]:

$$s\left(h_j^{\mathrm{Emp}}, h_m^{\mathrm{Emp}}\right) = v^{\mathrm{T}} \tanh\left(W h_j^{\mathrm{Emp}} + U h_m^{\mathrm{Emp}}\right) \tag{6}$$

where v, W **and** U are the learnable parameters. The CNN layer contains h kernels on a context window of k residues [31],

$$h_j^c = \sum_k \left[W^C * h_{j - \frac{k-1}{2} : j + \frac{k-1}{2}} + b^c\right] \tag{7}$$

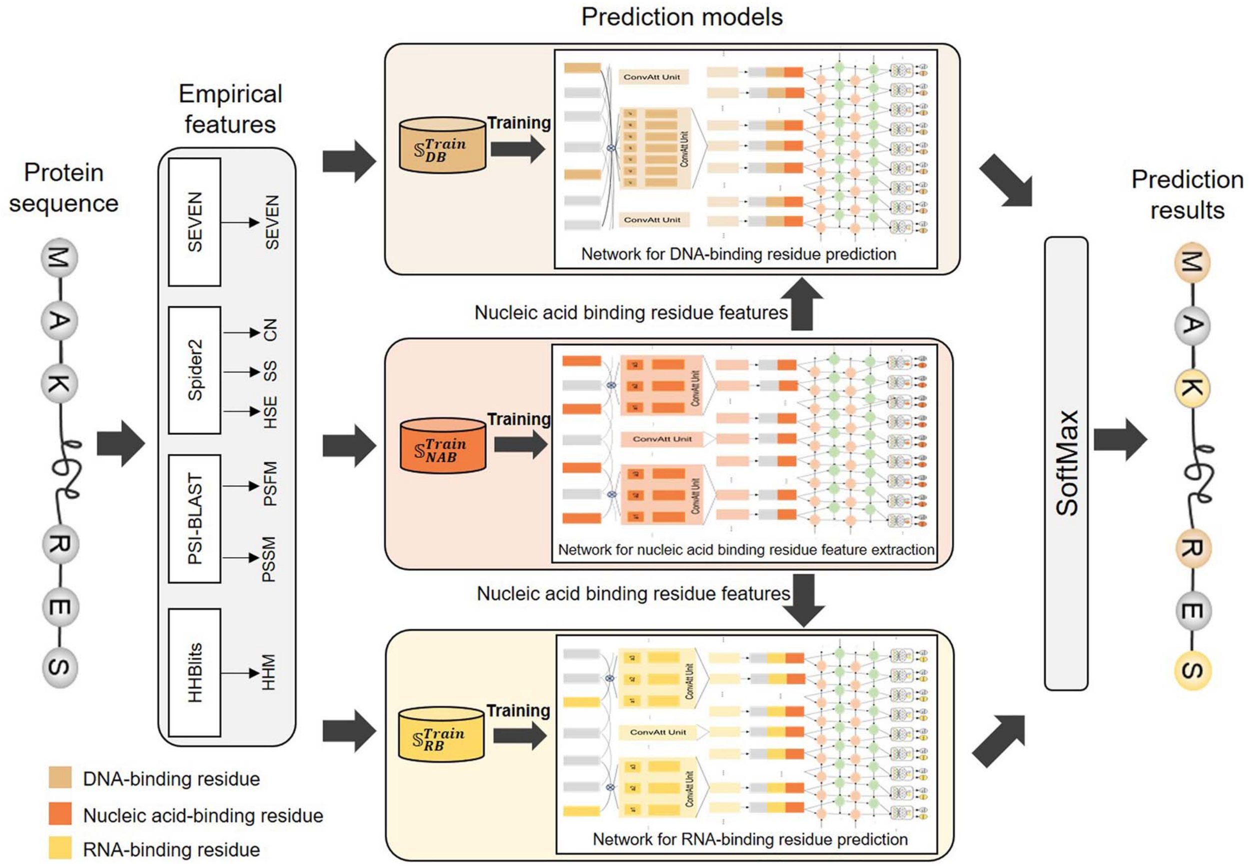
Prediction models
Empirical features
Protein sequence
M
A
K
R
E
S
SEVEN
SEVEN
Spider2
CN
SS
HSE
PSI-BLAST
PSFM
PSSM
HHBlits
HHM
$\mathbb{S}^{Train}_{DB}$
Training
ConvAtt Unit
Network for DNA-binding residue prediction
Nucleic acid binding residue features
$\mathbb{S}^{Train}_{NAB}$
Training
Network for nucleic acid binding residue feature extraction
Nucleic acid binding residue features
$\mathbb{S}^{Train}_{RB}$
Training
Network for RNA-binding residue prediction
SoftMax
Prediction results
M
A
K
R
E
S
DNA-binding residue
Nucleic acid-binding residue
RNA-binding residue

Figure 1. Induction and transfer framework of iDRNA-ITF. The nucleic acid-binding residue network was trained with $\mathbb{S}^{\text{Train}}_{\text{NAB}}$ to summarize the features of nucleic acid-binding residues from empirical features. Then, the features of nucleic acid-binding residues were input into the DNA-binding residue network and the RNA-binding residue network respectively, and the features were transferred to a new feature space according to more specific task signals. The DNA-binding residue network was trained with $\mathbb{S}^{\text{Train}}_{\text{DB}}$, the RNA-binding residue network was trained with $\mathbb{S}^{\text{Train}}_{\text{RB}}$. Finally, the prediction scores of the DNA-binding residue network and RNA-binding residue network were respectively passed through SoftMax to obtain the final prediction results.

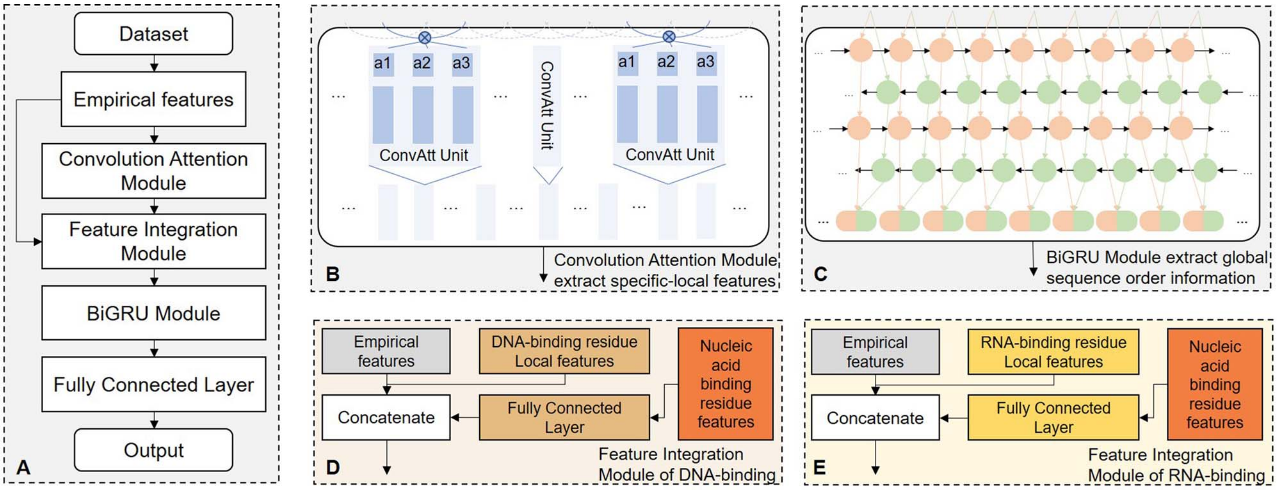
Dataset
Empirical features
Convolution Attention Module
Feature Integration Module
BiGRU Module
Fully Connected Layer
Output
A
a1
a2
a3
ConvAtt Unit
ConvAtt Unit
a1
a2
a3
ConvAtt Unit
Convolution Attention Module extract specific-local features
B
BiGRU Module extract global sequence order information
C
Empirical features
DNA-binding residue Local features
Nucleic acid binding residue features
Concatenate
Fully Connected Layer
Feature Integration Module of DNA-binding
D
Empirical features
RNA-binding residue Local features
Nucleic acid binding residue features
Concatenate
Fully Connected Layer
Feature Integration Module of RNA-binding
E

Figure 2. The network architecture of iDRNA-ITF. (**A**) The general network structures of DBRs identification task, RBRs identification task and NABRs identification task. (**B**) The convolutional attention module extracts specific-local features according to the different training signals. (**C**) The BiGRU module captures the long- and short-distance dependence among residues along with the protein. (**D**) The feature integration module of DNA-binding concatenates empirical features, DNA-binding residue local features and the task-transfer features obtained by $\text{trans}_{\text{DB}}\left(h_j^{\text{BiGRU-NAB}}\right)$. (**E**) The feature integration module of RNA-binding concatenates empirical features, RNA-binding residue local features and the task-transfer features obtained by $\text{trans}_{\text{RB}}\left(h_j^{\text{BiGRU-NAB}}\right)$.

where W^C and b^c are learnable parameters, the operation $*$ represents the element-wise product and $h_{j-\frac{k-1}{2}:j+\frac{k-1}{2}}$ means a concatenation of the hidden states $h_{j-\frac{k-1}{2}}, \ldots, h_{j+\frac{k-1}{2}}$. Finally, we perform the sum-pooling to get the local features of each residue. For the *j*th residue, the local features extracted by the convolutional attention module of the three tasks are $h_j^{\mathrm{ConvAtt-NAB}} \in \mathbb{R}^{1\times 110}$, $h_j^{\mathrm{ConvAtt-DB}} \in \mathbb{R}^{1\times 110}$ and $h_j^{\mathrm{ConvAtt-RB}} \in \mathbb{R}^{1\times 50}$, respectively.

Feature integration module

The feature integration module of the nucleic acid-binding residue network is used to concatenate empirical features and local-specific features, the output of this module is defined as follows:

$$f_j^{\mathrm{NAB}} = \left[h_j^{\mathrm{Emp}}; h_j^{\mathrm{BiGRU-NAB}}\right] \tag{8}$$

The feature integration modules of the DNA-binding residue network and the RNA-binding residue network were shown in Figure 2D and E, respectively. The feature integration modules in these two networks transfer the features of nucleic acid-binding residues through the fully connected layer, and then we concatenate the empirical features, local-specific features and task-transfer features. The output of the DNA-binding feature integration modules is defined as follows:

$$f_j^{DB} = \left[h_j^{\mathrm{Emp}}; h_j^{\mathrm{ConvAtt-DB}}; \mathbf{trans}_{\mathrm{DB}}\left(h_j^{\mathrm{BiGRU-NAB}}\right)\right] \tag{9}$$

where $\mathbf{trans}\_DB(h_j^{\mathrm{BiGRU-NAB}})$ is the task-transfer function and means DNA-binding residue identification as a training signal to perform task-transfer on $h_j^{\mathrm{BiGRU-NAB}}$ to extract task-transfer features. It can be represented by:

$$\mathbf{trans}_{\mathrm{DB}}\left(h_j^{\mathrm{BiGRU-NAB}}\right) = W_{\mathrm{DB}} h_j^{\mathrm{BiGRU-NAB}} + b_{\mathrm{DB}} \tag{10}$$

where W_{DB} and b_{DB} are learnable parameters.

The output of the RNA-binding feature integration modules is defined as follows:

$$f_j^{RB} = \left[h_j^{\mathrm{Emp}}; h_j^{\mathrm{ConvAtt-RB}}; trans_{RB}\left(h_j^{\mathrm{BiGRU-NAB}}\right)\right] \tag{11}$$

where $\mathbf{trans}_{\mathrm{RB}}(h_j^{\mathrm{BiGRU-NAB}})$ is the task-transfer function and means DNA-binding residue identification as a training signal to perform task-transfer on $h_j^{\mathrm{BiGRU-NAB}}$ to extract task-transfer features. It can be represented by:

$$\mathbf{trans}_{\mathrm{RB}}\left(h_j^{\mathrm{BiGRU-NAB}}\right) = W_{\mathrm{RB}} h_j^{\mathrm{BiGRU-NAB}} + b_{\mathrm{RB}} \tag{12}$$

where W_{RB} and b_{RB} are learnable parameters.

Finally, the outputs ($f_j^{\mathrm{NAB}} \in \mathbb{R}^{1\times 201}$, $f_j^{\mathrm{DB}} \in \mathbb{R}^{1\times 251}$ and $f_j^{\mathrm{RB}} \in \mathbb{R}^{1\times 251}$) of the feature integration modules of the three tasks are respectively input into their BiGRU modules for sequence feature learning.

BiGRU module and output layer

The BiGRU module was shown in Figure 2C, which contains two layers of BiGRU. This module is used to capture long- and short-distance dependencies between residues along with the protein. The BiGRU is defined as:

$$h_j^{\mathrm{BiGRU}} = \mathbf{BiGRU}\left(h_{j-1}^{\mathrm{BiGRU}}, f_j; W_1, W_2\right) \tag{13}$$

where f_j is the output of the feature integration module, h_{j-1}^{BiGRU} is the previous hidden state for the BiGRU layer, W_1 and W_2 are learnable parameters. Finally, we get the sequence information of the three tasks, denoted as $h_j^{\mathrm{BiGRU-NAB}} \in \mathbb{R}^{1\times 160}$, $h_j^{\mathrm{BiGRU-DB}} \in \mathbb{R}^{1\times 160}$, $h_j^{\mathrm{BiGRU-RB}} \in \mathbb{R}^{1\times 160}$, respectively.

For nucleic acid-binding residue networks, $h_j^{\mathrm{BiGRU-NAB}}$ is input into the feature integration module of the DNA-binding residue network and RNA-binding residue network. Feature transfer is performed through more specific task training signals. For DNA-binding residue network and RNA-binding residue network, $h_j^{\mathrm{BiGRU-DB}}$ and $h_j^{\mathrm{BiGRU-RB}}$ are respectively input into the fully connected layer for classification to calculate the final prediction results.

Implementation and training

In this study, we used Pytorch (https://pytorch.org/) to implement iDRNA-ITF. The Adam optimization algorithm [32, 33] was used to optimize parameters during the training process. To avoid network overfitting, we used Dropout algorithm [34, 35] during the training process. We used the weighted cross-entropy loss function to measure loss and solve the problem of imbalance in the datasets. The early stopping strategy was employed to control the training process based on the model's performance on the validation set. The hyperparameters were optimized based on the grid search strategy according to the maximum AUC (see section 'Performance evaluation metrics'). The grid search parameter range and the detailed parameters of the three networks of iDRNA-ITF were listed in Tables S1–S4, (see supplementary Data available online).

Performance evaluation metrics

The iDRNA-ITF and the other compared methods were evaluated by the following five evaluation metrics [8, 36–40], including the area under the ROC curve (AUC), Matthews correlation coefficient (MCC), F1-score (F1), recall (REC) and precision (PRE), which can be calculated by:

$$\mathrm{AUC} = \text{The area under the ROC curve} \tag{14}$$

$$\mathrm{MCC} = \frac{TP \times TN - FN \times FP}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \tag{15}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{REC} \times \mathrm{PRE}}{\mathrm{REC} + \mathrm{PRE}} \tag{16}$$

$$\mathrm{REC} = \frac{TP}{TP+FN} \tag{17}$$

Table 2. The influence of different inputs of the feature integration module in DNA-binding residue network on the performance of iDRNA-ITF

Model	Feature extraction tool module[a]	DNA-binding convolutional attention module[b]	Nucleic acid-binding module[c]	AUC	1-AURC	MCC	F1	REC	PRE
A_DB	√			0.857	0.707	0.268	0.273	0.440	0.198
B_DB		√		0.850	0.665	0.262	0.261	0.471	0.180
C_DB			√	0.868	0.641	0.279	0.272	0.511	0.185
D_DB	√	√		0.866	0.683	0.279	0.279	0.479	0.197
E_DB	√		√	0.878	0.666	0.279	0.258	0.584	0.165
F_DB		√	√	0.881	0.711	0.302	0.302	0.492	0.218
iDRNA-ITF_DB[d]	√	√	√	0.886	0.722	0.320	0.329	0.454	0.258

[a] Feature extraction tool module provides empirical features. [b] DNA-binding convolutional attention module provides DNA-binding residue local features. [c] Nucleic acid-binding module provides task-transfer features transferred from nucleic acid-binding residue features. [d] iDRNA-ITF_DB represents DNA-binding residue network in iDRNA-ITF.

Table 3. The influence of different inputs of the feature integration module in RNA-binding residue network on the performance of iDRNA-ITF

Model	Feature extraction tool module[a]	RNA-binding convolutional attention module[b]	Nucleic acid-binding Module[c]	AUC	1-AURC	MCC	F1	REC	PRE
A_RB	√			0.711	0.417	0.124	0.131	0.469	0.076
B_RB		√		0.741	0.430	0.111	0.136	0.124	0.152
C_RB			√	0.772	0.356	0.176	0.176	0.488	0.107
D_RB	√	√		0.750	0.590	0.158	0.187	0.253	0.148
E_RB	√		√	0.779	0.362	0.191	0.191	0.482	0.119
F_RB		√	√	0.785	0.597	0.193	0.221	0.249	0.198
iDRNA-ITF_RB[d]	√	√	√	0.799	0.642	0.208	0.233	0.306	0.188

[a] Feature extraction tool module provides empirical features. [b] RNA-binding convolutional attention module provides RNA-binding residue local features. [c] Nucleic acid-binding module provides task-transfer features transferred from nucleic acid-binding residue features. [d] iDRNA-ITF_RB represents RNA-binding residue network in iDRNA-ITF.

$$\mathrm{PRE} = \frac{TP}{TP + FP} \tag{18}$$

where TP is true positive, TN is true negative, FP is false positive, FN is false negative. In addition, the 1 − the area under the CPR-TPR curve (1 − AURC) [1] was used to evaluate the cross-prediction problem in the hybrid dataset.

Results and discussion

The induction and transfer framework can improve predictive performance and weaken the cross-prediction problem

In order to analyze the importance of different components in the proposed method, and explore the interaction among different tasks and modules, we conducted an ablation experiment on DRNA-106_Valid set. Different components were processed by the feature integration module, which includes three parts: empirical features, local-specific features and task-transfer features. Tables 2 and 3 listed the influence of different parts on the final prediction results.

Experiments A_DB-C_DB and A_RB-C_RB evaluate the impact of using only one module feature on prediction performance. A_DB and A_RB show that the empirical features used in traditional methods can only achieve the most basic predictive performance. Compared with empirical features, the local-specific features extracted by the convolutional attention module improve the performance of RNA-binding residue prediction (see B_RB). This indicates that RNA-binding residues have more conservative local patterns. Compared with A_DB and B_DB (A_RB and B_RB), C_DB (C_RB) achieved the best prediction performance, which shows the characteristics of nucleic acid-binding residue network induction are more comprehensive and generalized, and they can extract effective features to achieve better performance in downstream tasks by task-transfer. Especially for the task of RNA-binding residue recognition, the AUC of C_RB is 6 points higher than that of A_RB. Because nucleic acid-binding residue features describe both RNA-binding residue characteristics and DNA-binding residue characteristics, task-transfer had limitations in distinguishing between two types of residues, leading to the lowest 1-AURC.

Experiments D_DB-F_DB and D_RB-F_RB evaluate the contributions of the pairwise concatenation features for different modules. It can be seen that the prediction performance of the models based on pairwise concatenation is better than models based on a single module. This shows that the features of different modules are complementary. When combining local-specific features and task-transfer features, the prediction results are greatly improved (see F_DB, F_RB), because task information and generalization information are considered in the features. iDRNA-ITF_DB and iDRNA-ITF_RB used the features of the three modules in the feature

Table 4. Performance comparison of different methods on two test datasets

Dataset	Method	AUC	MCC	F1	Rec	Pre
DNA-129_Test	TargetS[a]	N/A	0.262	0.291	0.239	0.370
	TargetDNA[a]	0.825	0.291	0.335	0.417	0.280
	DNAPred[a]	0.845	0.332	0.373	0.396	0.353
	COACH-D[a]	0.761	0.302	0.341	0.324	0.360
	NucBind[a]	0.797	0.309	0.346	0.323	0.373
	SVMnuc[a]	0.812	0.304	0.341	0.316	0.371
	NCBRPred[b]	0.823	0.313	0.348	0.312	0.392
	Graphbind[a]	0.816	0.320	0.362	0.439	0.310
	Graphbind*[c]	0.855	0.292	0.295	0.567	0.200
	iDRNA-ITF	0.883	0.401	0.438	0.500	0.389
RNA-117_Test	RNABindRPlus[a]	0.717	0.202	0.248	0.273	0.227
	COACH-D[a]	0.663	0.195	0.235	0.221	0.252
	NucBind[a]	0.715	0.189	0.233	0.231	0.235
	SVMnuc[a]	0.729	0.192	0.235	0.231	0.240
	NCBRPred[b]	0.667	0.172	0.187	0.135	0.300
	Graphbind[a]	0.718	0.168	0.218	0.303	0.171
	Graphbind*[c]	0.738	0.164	0.177	0.439	0.111
	iDRNA-ITF	0.760	0.236	0.281	0.349	0.235

[a]Results were reported in [2]. [b]Results were calculated by using the webserver of NCBRPred [6]. [c]Results were calculated by using the source code of GraphBind [2] based on the proteins whose structures were predicted by MODELLER [19].

integration module to get the best prediction performance, indicating that the induction and transfer framework can improve predictive performance and weaken the cross-prediction problem.

Comparison of different methods on test datasets

iDRNA-ITF and the other sequence-based methods were evaluated and compared on two test datasets (see Table 4). Furthermore, the structure-based method GraphBind [2] used the predicted protein structures of the two test sets to predict binding residues, and its performance was also compared with iDRNA-ITF.

Compared with the sequence-based methods, the AUC and MCC of iDRNA-ITF were increased by 0.038–0.122 and 0.069–0.139 on the DNA-129_Test dataset, the AUC and MCC of iDRNA-ITF were improved by 0.031–0.097 and 0.034–0.072 on the RNA-117_Test dataset. The reason is that the introduced nucleic acid-binding residue prediction task can provide additional sequence information for identifying DNA- and RNA-binding residues. Compared with GraphBind, iDRNA-ITF performed better in both test datasets, indicating that iDRNA-ITF is even better than the structure-based method for predicting the proteins without known structures.

Comparison of different methods on the hybrid test dataset

iDRNA-ITF and the other methods that can predict both DBRs and RBRs were evaluated and compared on the hybrid test dataset (see Table 5 and Figure 3). The results showed that iDRNA-ITF achieves the best results in terms of all the four evaluation metrics (AUC, 1-AURC, MCC and F1) when identifying DBRs and RBRs. This is because the feature integration module uses three complementary features to represent residues, taking into account not only the specificity characteristics of the residues but also the generalization characteristics of the residues.

Analysis of the predicted DNA- and RNA-binding residues

Residues located closer to the binding residues in the protein sequences are more likely to bind with DNA/RNA [1, 6]. Therefore, we analyzed the false positives predicted by the three best predictors by calculating the fraction of predicted DBRs and RBRs within the range of true DBRs and RBRs. The fractions of the three predictors at different ranges were depicted in Figure 4, which shows that iDRNA-ITF outperforms the compared predictors when the distance range is <2 for predicting DBRs, and it outperforms the compared predictors when the distance range is <6 for predicting RBRs. Although the other two predictors showed higher scores at longer distances, this was because iDRNA-ITF predicted fewer false positives and the convolutional attention module in the inductive transfer framework learned local features, making the predicted false positives closer to true binding residues.

Predictive result visualization

The DNA-binding protein chain 5k7z_A was selected from DNA-129_Test, the RNA-binding protein chain 5z9x_A was selected from RNA-117_Test. Pymol (https://pymol.org/2/) was used to visualize the prediction results of the top three predictors for these two protein chains (see Figure 5). We can see the following: (i) compared with GraphBind* and SVMNuc, iDRNA-ITF predicted the most true positive samples and the fewest false positive samples, indicating that iDRNA-ITF learns the differences between positive samples and negative samples benefited from the multi-level features provided by the induction and transfer framework. (ii) GraphBind* tends to predict residues as binding residues, indicating

Table 5. Performance comparison of different methods on the hybrid test dataset

Class	Method	AUC	1-AURC	MCC	F1	Rec	Pre
DNA-binding residue	SVMnuc[a]	0.824	0.623	0.227	0.247	0.340	0.195
	NucBind[a]	0.812	0.624	0.223	0.245	0.326	0.196
	NCBRPred[b]	0.819	0.648	0.231	0.254	0.312	0.214
	Graphbind*[c]	0.851	0.636	0.252	0.238	0.567	0.150
	iDRNA-ITF	0.881	0.711	0.324	0.331	0.500	0.247
RNA-binding residue	SVMnuc[a]	0.743	0.490	0.148	0.172	0.235	0.136
	NucBind[a]	0.712	0.490	0.149	0.172	0.238	0.135
	NCBRPred[b]	0.695	0.631	0.155	0.168	0.135	0.220
	Graphbind*[c]	0.739	0.418	0.143	0.141	0.439	0.084
	iDRNA-ITF	0.790	0.643	0.203	0.217	0.349	0.158

[a] Results were calculated by using the webserver of NucBind [13]. [b] Results were calculated by using the YK17 trained model accessed from in the webserver of NCBRPred [6]. [c] Results were calculated by using the source code of GraphBind [2] based on the proteins whose structures were predicted by MODELLER [19].

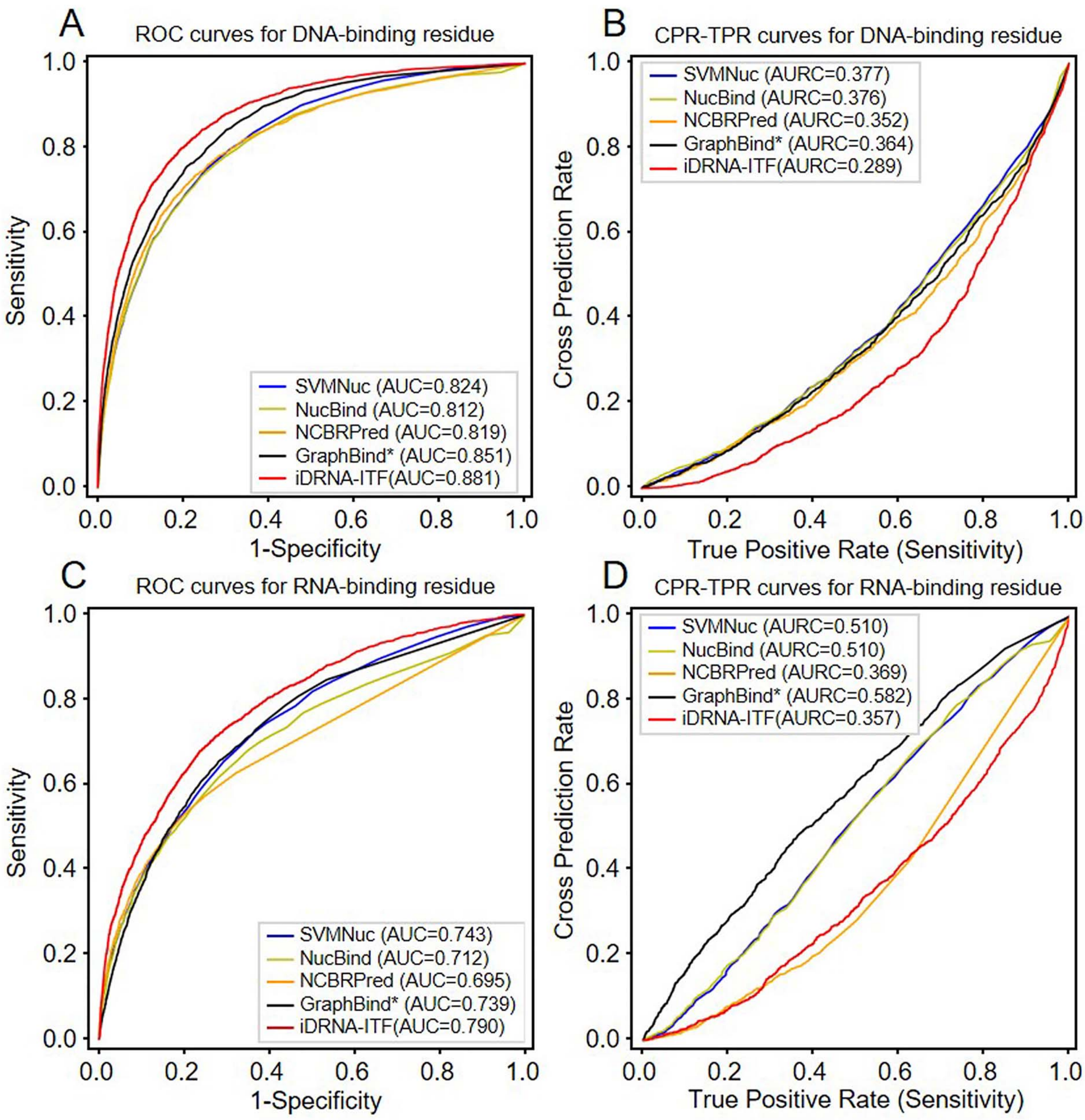

Figure 3. The performance of different methods on DRNA-246_Test. (**A**) ROC curves of different methods for DNA-binding residue. (**B**) CPR-TPR curves of different methods for DNA-binding residue. (**C**) ROC curves of different methods for RNA-binding residue. (**D**) CPR-TPR curves of different methods for RNA-binding residue.

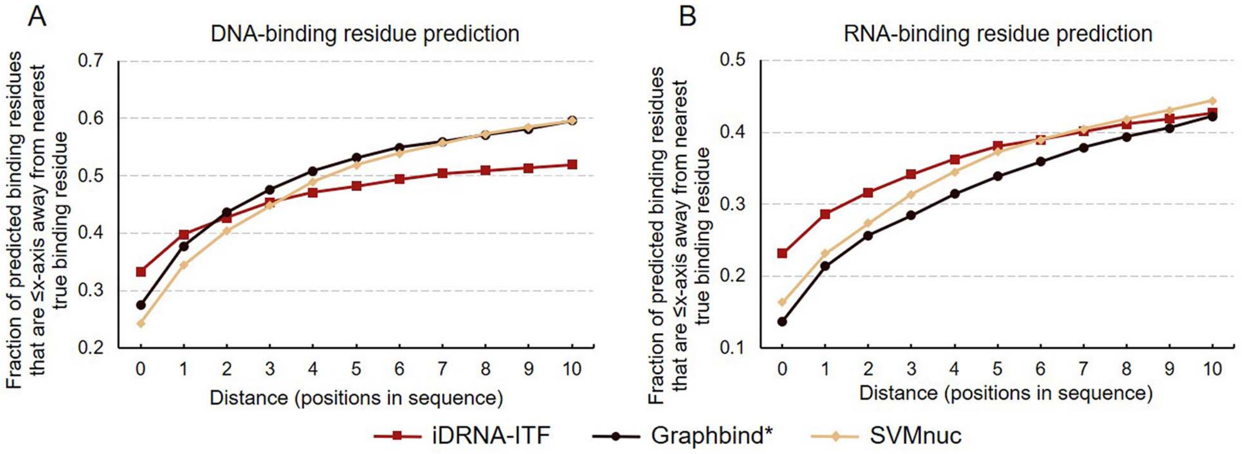

Figure 4. Analyze the predicted binding residues within a certain range of true binding residues. (**A**) Analysis of DNA-binding residue positions predicted by different methods in protein sequences. (**B**) Analysis of RNA-binding residue positions predicted by different methods in protein sequences.

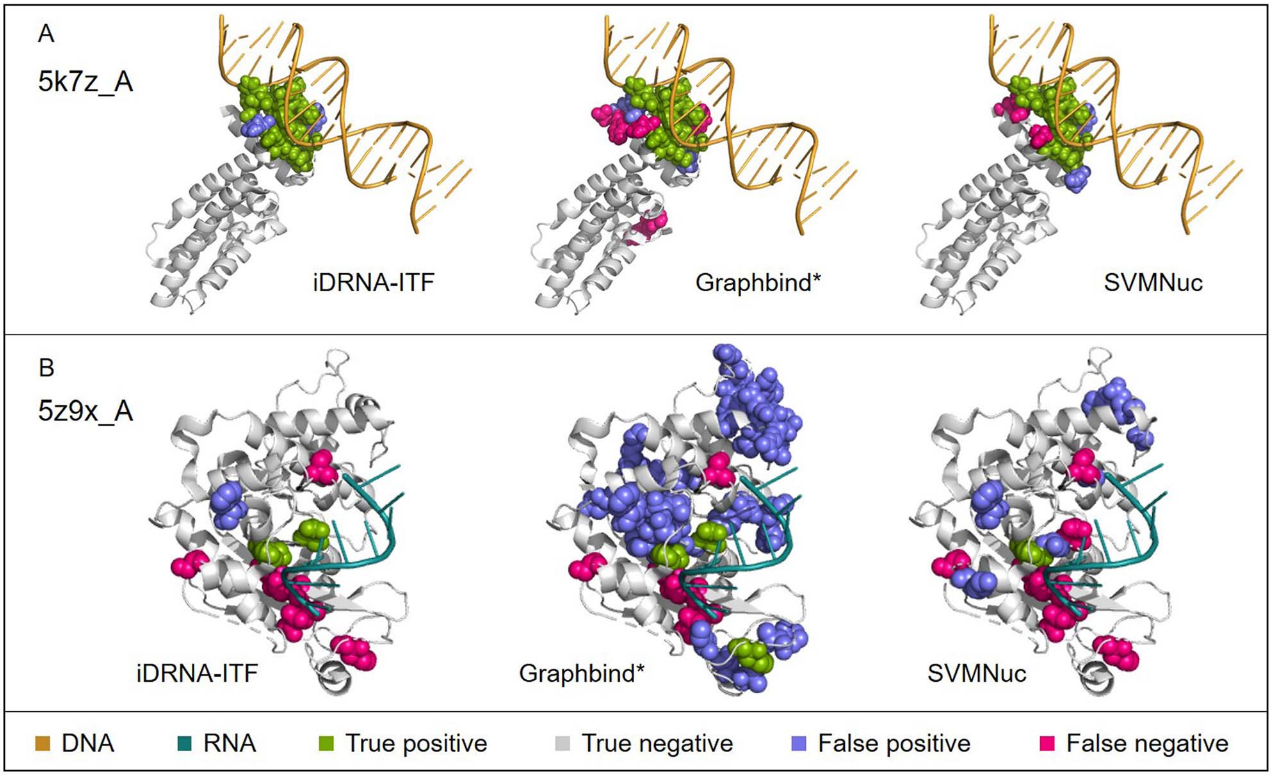

Figure 5. Visualization of prediction results by different methods. The results for identifying the DNA/RNA-binding residues in protein chain 5k7z_A and 5z9x_A predicted by iDRNA-ITF, GraphBind*, SVMNuc are shown in (**A**) and (**B**), respectively.

that the judgment conditions for binding residues were looser, leading to more false positives.

Comparison of different methods on the MW15 dataset

The MW15 dataset [41] is a commonly used test set for evaluating the performance of the methods for identifying both DBRs and RBRs. There are 46 protein sequences in this dataset with 760 DBRs, 368 RBRs and 9447 non-NABRs. In addition, we removed proteins that share over 25% sequence similarity with any protein in MW15 from the training dataset, and then retrained the model to predict the protein sequences in MW15. The performance of different methods on the MW15 test dataset was shown in Table 6 and Figure 6, from which we can see that iDRNA-ITF outperforms the other competing methods for identifying both DBRs and RBRs. These results further demonstrate that iDRNA-ITF has good generalization ability and can achieve stable performance on different test sets.

Table 6. Performance comparison of different methods on the MW15 dataset

Class	Method	AUC	1-AURC	MCC	F1	Rec	Pre
DNA-binding residue	DRNApred[a]	0.725	0.521	0.164	0.226	0.236	0.217
	SVMnuc[b]	0.808	0.546	0.343	0.380	0.332	0.444
	NucBind[b]	0.819	0.598	0.368	0.403	0.354	0.468
	NCBRPred[d]	0.810	0.695	0.407	0.450	0.450	0.450
	iDRNA-ITF	0.870	0.743	0.460	0.500	0.549	0.460
RNA-binding residue	DRNApred[a]	0.467	0.499	0.006	0.027	0.019	0.044
	SVMnuc[b]	0.777	0.595	0.190	0.220	0.255	0.193
	NucBind[b]	0.780	0.660	0.293	0.317	0.386	0.269
	NCBRPred[d]	0.799	0.800	0.236	0.263	0.264	0.262
	iDRNA-ITF	0.839	0.760	0.270	0.280	0.478	0.198

[a]Results were calculated by using the webserver of DRNApred [1]. [b]Results were calculated by using the webserver of NucBind [13]. [c]Results were calculated by using the YK16-5 trained model accessed from the webserver of NCBRPred [6].

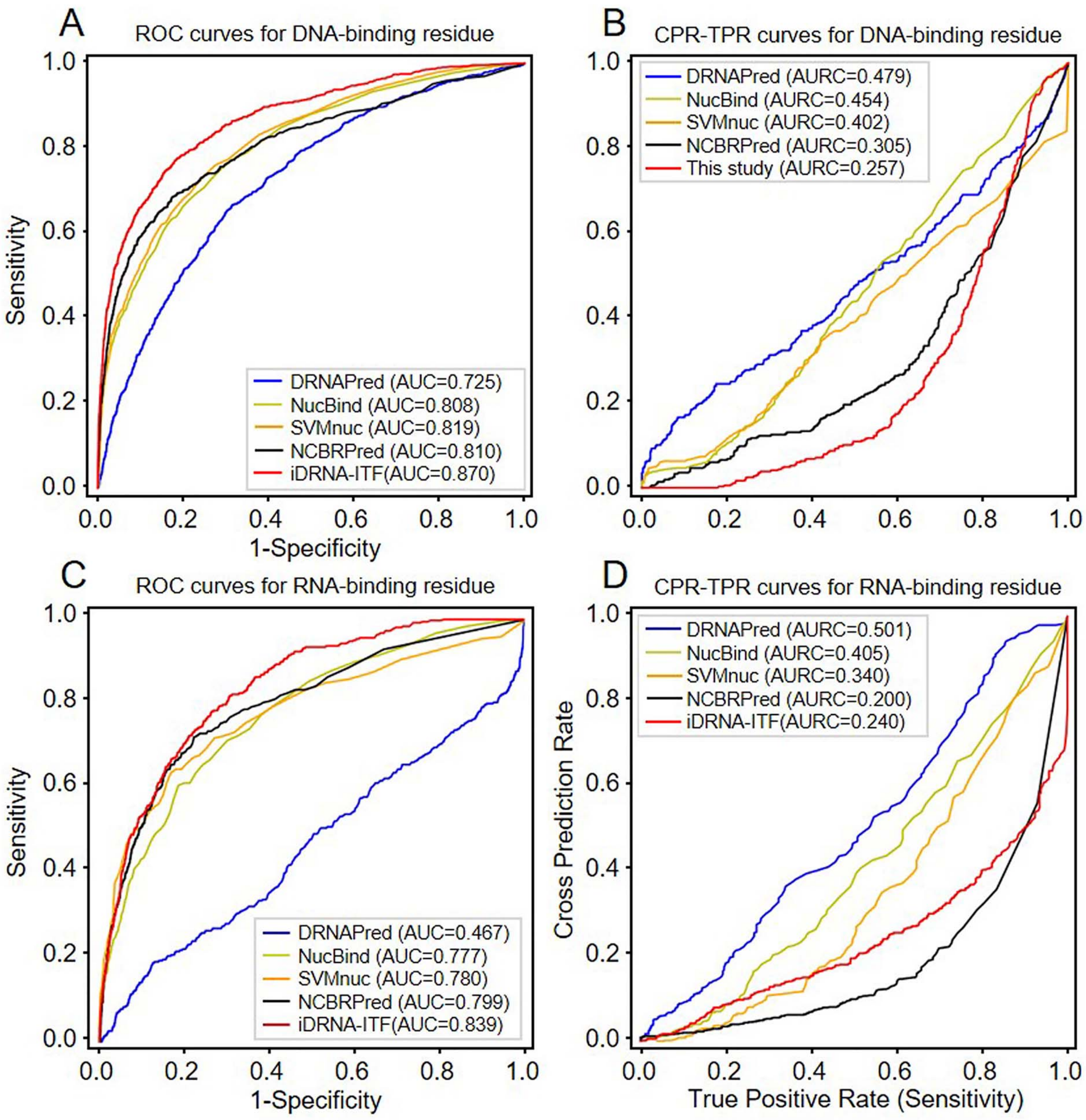

Figure 6. The performance of different methods on the MW15 dataset. (**A**) ROC curves of different methods for DNA-binding residue. (**B**) CPR-TPR curves of different methods for DNA-binding residue. (**C**) ROC curves of different methods for RNA-binding residue. (**D**) CPR-TPR curves of different methods for RNA-binding residue.

Discussion

DNA- and RNA-binding residues in proteins are essential for gene expression, signal transduction, etc. In this regard, several predictors were proposed to predict DBR and RBR in proteins. However, those methods fail to use the functional properties of the residues to assist in the identification of DNA- and RNA-binding residues. The nucleic acid-binding function is associated with both DNA-binding function and RNA-binding function, and nucleic acid-binding features can provide more generalized features for DNA-binding residue recognition tasks and RNA-binding residue recognition tasks. Therefore, we use the idea of transfer learning to design an induction and transfer framework to induct nucleic acid-binding features, and perform task transfer. The experimental results show that the features obtained by induction and transfer framework can provide effective information and are complementary to the empirical features. In addition, the proposed method also has some limitations. For example, iDRNA-ITF fails to distinguish between DBR and RBR from some specific species. In the future, we will construct a more suitable sequence analysis model to predict the species specific DNA- and RNA-binding residues.

Conclusion

Feature extraction of protein sequences is critical for constructing the sequence-based methods, traditional methods do not fully utilize the functional properties of residues in feature extraction. In this study, a sequence-based method iDRNA-ITF was proposed to incorporate the functional properties in residue representation by using an induction and transfer framework. The characteristics of the method can be summarized as follows: (i) The independent networks for the DBRs identification task and the RBRs identification task were designed respectively, and learn specific parameters during the training process; (ii) The nucleic acid-binding features of the residues were summarized and input into the DNA-binding residue network and the RNA-binding residue network for improving prediction performance and (iii) The feature integration module in two networks concatenates three complementary features to represent residue, which ensures the balance between prediction accuracy and cross-prediction problems. It can be anticipated that the induction and transfer framework has potential applications in many fiels, such as functional sites in protein disordered regions prediction [42], therapeutic peptide recognition [43, 44], etc.

Key Points

- In this study, we proposed a sequence-based method called iDRNA-ITF to identify DNA- and RNA-binding residues in proteins.
- iDRNA-ITF summarizes the nucleic acid-binding residue features and inputs them into the DNA-binding residue recognition network and the RNA-binding residue recognition network for feature transfer.
- A feature integration module was designed in the network structure, concatenating three types of features as the input of the downstream network. This enhances the spread of features and reduces overfitting problems.
- Experimental results on the four independent datasets showed that iDRNA-ITF outperforms the other competing sequence-based methods. The web server of iDRNA-ITF is accessible at http://bliulab.net/iDRBP-ITF.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

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

Funding

This work was supported by the Beijing Natural Science Foundation (no. JQ19019), National Natural Science Foundation of China (no. 62102030).

References

1. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res* 2017;**45**(10):e84.
2. Xia Y, Xia CQ, Pan X, *et al.* GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic Acids Res* 2021;**49**(9):e51.
3. Liu Z-P. Predicting lncRNA-protein Interactions by machine learning methods: a review. *Curr Bioinform* 2020;**15**(8):831–40.
4. Ao CY, Yu L, Zou Q. Prediction of bio-sequence modifications and the associations with diseases. *Brief Funct Genomics* 2021;**20**(1):1–18.
5. Hu J, Li Y, Zhang M, *et al.* Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**(6):1389–98.
6. Zhang J, Chen Q, Liu B. NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning. *Brief Bioinform* 2021;**22**(5). https://doi.org/10.1093/bib/bbaa397.
7. Wang X, Wang S, Fu H, *et al.* DeepFusion-RBP: using deep learning to fuse multiple features to identify RNA-binding protein sequences. *Curr Bioinform* 2021;**16**(8):1089–100.
8. Zou Y, Wu H, Guo X, *et al.* MK-FSVM-SVDD: a multiple kernel-based fuzzy SVM model for predicting DNA-binding proteins via support vector data description. *Curr Bioinform* 2021;**16**(2): 274–83.

9. Niu M, Wu J, Zou Q, *et al.* rBPDL: predicting RNA-binding proteins using deep learning. *IEEE J Biomed Health Inform* 2021;**25**(9): 3668–76.
10. Yu DJ, Hu J, Yang J, *et al.* Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM Trans Comput Biol Bioinform* 2013;**10**(4):994–1008.
11. Walia RR, Xue LC, Wilkins K, *et al.* RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. *PLoS One* 2014;**9**(5): e97725.
12. Zhu YH, Hu J, Song XN, *et al.* DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines. *J Chem Inf Model* 2019;**59**(6):3057–71.
13. 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**(6):930–6.
14. Yang H, Deng Z, Pan X, *et al.* RNA-binding protein recognition based on multi-view deep feature and multi-label learning. *Brief Bioinform* 2021;**22**(3). https://doi.org/10.1093/bib/bbaa174.
15. Zhang J, Yan K, Chen Q, *et al.* PreRBP-TL: prediction of species-specific RNA-binding proteins based on transfer learning. *Bioinformatics* 2022;**38**(8):2135–43. https://doi.org/10.1093/bioinformatics/btac106.
16. Li S, Yamashita K, Amada KM, *et al.* Quantifying sequence and structural features of protein-RNA interactions. *Nucleic Acids Res* 2014;**42**(15):10086–98.
17. Lam JH, Li Y, Zhu L, *et al.* A deep learning framework to predict binding preference of RNA constituents on protein surface. *Nat Commun* 2019;**10**(1):4941.
18. Liu R, Hu J. DNABind: a hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning- and template-based approaches. *Proteins* 2013;**81**(11): 1885–99.
19. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 1993;**234**(3):779–815.
20. Yang J, Anishchenko I, Park H, *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A* 2020;**117**(3):1496–503.
21. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**(7873): 583–9.
22. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res* 2013;**41**(Database issue):D1096–103.
23. Li Q, Yu J, Yan Y, *et al.* PsePSSM-based prediction for the protein-ATP binding sites. *Curr Bioinform* 2021;**16**(4):576–82.
24. 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**(17):3389–402.
25. Holm L, Sander C. Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* 1998;**14**(5): 423–9.
26. Remmert M, Biegert A, Hauser A, *et al.* HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2011;**9**(2):173–5.
27. Mirdita M, von den Driesch L, Galiez C, *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res* 2017;**45**(D1):D170–6.
28. Yang Y, Heffernan R, Paliwal K, *et al.* SPIDER2: a package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. *Methods Mol Biol* 2017;**1484**:55–63.
29. Meiler J, Müller M, Zeidler A. Schm? Schke F: generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *J Mol Model* 2001;**7**(9): 360–9.
30. Rifkin RM, Klautau A. In defense of one-vs-all classification. *J Mach Learn Res* 2004;**5**:101–41.
31. Zhu Y, Wang G, Karlsson BF. CAN-NER: convolutional attention network for chinese named entity recognition. 2019.
32. Kingma D, Ba J. Adam: a method for stochastic optimization. *Comput Sci* 2014.
33. Zafar S, Nazir M, Sabah A, *et al.* Securing bio-cyber interface for the internet of bio-nano things using particle swarm optimization and artificial neural networks based parameter profiling. *Comput Biol Med* 2021;**136**:104707.
34. Srivastava N, Hinton G, Krizhevsky A, *et al.* Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**(1):1929–58.
35. Garifullin A, Lensu L, Uusitalo H. Deep Bayesian baseline for segmenting diabetic retinopathy lesions: advances and challenges. *Comput Biol Med* 2021;**136**:104725.
36. Majumder R, Das CK, Banerjee I, *et al.* Screening of the Prime bioactive compounds from Aloe vera as potential anti-proliferative agents targeting DNA. *Comput Biol Med* 2022;**141**: 105052–2.
37. Zeb A, Ali SS, Azad AK, *et al.* Genome-wide screening of vaccine targets prioritization and reverse vaccinology aided design of peptides vaccine to enforce humoral immune response against Campylobacter jejuni. *Comput Biol Med* 2021;**133**:104412.
38. Chauhan A, Avti P, Shekhar N, *et al.* Structural and conformational analysis of SARS CoV 2 N-CTD revealing monomeric and dimeric active sites during the RNA-binding and stabilization: insights towards potential inhibitors for N-CTD. *Comput Biol Med* 2021;**134**:104495.
39. Niu M, Zou Q, Lin C. CRBPDL: identification of circRNA-RBP interaction sites using an ensemble neural network approach. *PLoS Comput Biol* 2022;**18**(1):e1009798.
40. Li H, Pang Y, Liu B. BioSeq-BLM: a platform for analyzing DNA, RNA, and protein sequences based on biological language models. *Nucleic Acids Res* 2021;**49**(22):e129.
41. Miao Z, Westhof E. A Large-scale assessment of nucleic acids binding site prediction programs. *PLoS Comput Biol* 2015;**11**(12):e1004639.
42. Tang YJ, Pang YH, Liu B. DeepIDP-2L: protein intrinsically disordered region prediction by combining convolutional attention network and hierarchical attention network. *Bioinformatics* 2022;**38**(5):1252–60.
43. Yan K, Lv H, Wen J, *et al.* TP-MV: therapeutic peptides prediction by multi-view learning. *Curr Bioinform* 2022;**17**(2):174–83.
44. Yan K, Lv H, Guo Y, *et al.* TPpred-ATMV: therapeutic peptide prediction by adaptive multi-view tensor learning model. *Bioinformatics* 2022;**38**(10):2712–8. https://doi.org/10.1093/bioinformatics/btac200.