Computational and Structural Biotechnology Journal 20 (2022) 2020–2028

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

DeepMC-iNABP: Deep learning for multiclass identification and classification of nucleic acid-binding proteins

Feifei Cui [a,b,c,1], Shuang Li [d,1], Zilong Zhang [a,b,c], Miaomiao Sui [e], Chen Cao [f], Abd El-Latif Hesham [g], Quan Zou [b,c,*]

[a] School of Computer Science and Technology, Hainan University, Haikou 570228, China
[b] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China
[c] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324000, China
[d] Beidahuang Industry Group General Hospital, Harbin 150001, China
[e] Graduate School Agricultural and Life Science, The University of Tokyo, Tokyo 1138657, Japan
[f] School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 211166, China
[g] Genetics Department, Faculty of Agriculture, Beni-Suef University, Beni-Suef 62511, Egypt

ARTICLE INFO

Article history:
Received 8 February 2022
Received in revised form 6 April 2022
Accepted 20 April 2022
Available online 26 April 2022

Keywords:
Deep learning
Multiclass classification
DNA-binding protein
RNA-binding protein
Nucleic acid-binding protein

ABSTRACT

Nucleic acid-binding proteins (NABPs), including DNA-binding proteins (DBPs) and RNA-binding proteins (RBPs), play vital roles in gene expression. Accurate identification of these proteins is crucial. However, there are two existing challenges: one is the problem of ignoring DNA- and RNA-binding proteins (DRBPs), and the other is a cross-predicting problem referring to DBP predictors predicting DBPs as RBPs, and vice versa. In this study, we proposed a computational predictor, called DeepMC-iNABP, with the goal of solving these difficulties by utilizing a multiclass classification strategy and deep learning approaches. DBPs, RBPs, DRBPs and non-NABPs as separate classes of data were used for training the DeepMC-iNABP model. The results on test data collected in this study and two independent test datasets showed that DeepMC-iNABP has a strong advantage in identifying the DRBPs and has the ability to alleviate the cross-prediction problem to a certain extent. The web-server of DeepMC-iNABP is freely available at http://www.deepmc-inabp.net/. The datasets used in this research can also be downloaded from the website.

1. Introduction

Interactions between proteins and nucleic acids (general terms of DNA and RNA) define and regulate diverse cellular functions, such as splicing, translation, posttranscriptional modifications, protein synthesis, DNA transcription and replication [1–7]. They are also closely related to some diseases such as cancer [8–12] and involved in virus infection [13,14]. Therefore, it is important to accurately identify the proteins that bind to nucleic acids (DNA or RNA).

Nucleic acid-binding proteins (NABPs) are traditionally divided into proteins that have the ability to bind DNA or RNA, namely, DNA-binding proteins (DBPs) and RNA-binding proteins (RBPs). However, there are some proteins that can bind to both DNA and RNA (DRBPs), and these types of protein also play an important role in gene expression [15–17]. Therefore, accurate identification of DBPs, RBPs and DRBPs is vitally important.

In recent years, with the rapid development of high-throughput sequencing technologies and the reduction of sequencing costs, a large amount of bioinformatics data has been accumulated [18,19]. Machine learning and deep learning techniques have been widely applied for the analysis of large amounts of biological data [20–25]. Among them, protein sequences have shown exponential growth, which makes deep learning models for the prediction of nucleic acid-binding proteins from primary sequences strongly feasible [26–28].

Existing computational models treat DBP identification and RBP identification as two independent tasks, and most of them are binary-class prediction tasks. Because DBPs and RBPs have high similarities, the existing computational models generally have cross-predicting problems. Cross-prediction refers to the DBP predictor identifying true RBPs as DBPs, whereas the RBP predictor

* Corresponding author at: Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China.
E-mail address: zouquan@nclab.net (Q. Zou).
[1] Equally contributed.

https://doi.org/10.1016/j.csbj.2022.04.029

predicts true DBPs as RBPs. Moreover, they usually ignore the existence of DRBPs. That is because in these independent binary-class classification models, only DBPs or RBPs are used to build predictors of DNA binding or RNA binding, respectively. Thus, the predictors identified either "DBPs and non-DNA binding proteins (non-DBPs)", or "RBPs and non-RNA binding proteins (non-RBPs)".

Recent studies [2,29] on the prediction of nucleic acid-binding residues show that although the existing DNA-binding or RNA-binding residue predictors have strong predictive performance, they cannot discriminate DNA- and RNA-binding residues. Because independent predictors are built using exclusive data (either DNA-binding or RNA-binding residues), they are unable to learn to separate DNA-binding from RNA-binding residues. William et al. revealed through experiments that DRBPs are an important part of cellular proteins and play important cellular roles [17].

To solve the above problems, in this study, we proposed a new computational predictor called DeepMC-iNABP by applying a multiclass learning strategy to identify NABPs based on deep learning. Four categories, "DBP", "RBP", "DRBP" and "non-nucleic acid binding protein (non-NABP)", were defined and used for learning the multiclass classification model. By applying the multiclass classification model based on a convolutional neural network and a recurrent neural network, the DeepMC-iNABP predictor predicts a protein as each of these classes. The DeepMC-iNABP predictor overcomes the problem of cross-predicting because both DBPs and RBPs are used for training the multiclass learning model. DeepMC-iNABP can not only identify nuclei acid-binding protein but also detect DBPs and RBPs, which will help with the functional annotation of proteins. Moreover, because DRBPs as separate class data are used for training, DRBPs can be identified by the DeepMC-iNABP predictor.

2. Material and methods

2.1. Data resources

In this study, an initial dataset containing 15,654 DBP chains, 15,009 RBP chains,1,218 DRBP chains and 146,900 non-NABP chains was first collected from the UnitProt Knowledgebase (UniProtKB) [30]. DBP chains were collected by searching reviewed data with the Gene Ontology (GO) annotation keywords "DNA binding" and "NOT RNA binding", whereas RBP chains were collected by searching reviewed data with the GO annotation keywords "RNA binding" and "NOT DNA binding". DRBP chains were obtained by searching reviewed data with the GO annotation keywords "DNA binding" and "RNA binding". Non-NABP chains were collected by searching the reviewed data with the keyword "NOT nucleic acid binding". Then, we deleted duplicate chains and filtered the chains with lengths of less than 40 or more than 1,000. The redundancy between each dataset was reduced by the BLASTP program [31] with the bit-score set as greater than 50.0. After data preprocessing, 12,471 DBP chains, 12,068 RBP chains,1,218 DRBP chains and 143,650 non-NABP chains were remained. Then, we randomly selected sequences data to form the final datasets containing 12,000 DBP chains, 12,000 RBP chains,1,200 DRBP chains and 12,000 non-NABP chains. We first split the final dataset into two (training and testing) at a ratio of 10:1. After this, we kept aside the test set and randomly chose 90% of our training data to be the actual training set and the remaining 10% to be the validation set. The statistical information of the training data, validation data and test data we collected is listed in Table 1.

In this study, we also used two independent test datasets (Table 2) to evaluate the performance of the DeepMC-iNABP predictor, including TEST474 [32] and DRBP206 [32]. Due to very little information found in the prior studies on the identification of DRBPs, TEST474 and DRBP206 are rare independent datasets containing DRBP data. A total of 175 DBP chains, 68 RBP chains, 8 DRBP chains and 233 non-NABP chains collected from the Swiss-Prot database to construct the TEST474 dataset. The DRBP206 dataset contains only 103 DRBP chains, and 103 non-NABP chains were also collected from the Swiss-Prot database.

2.2. Multiclass learning strategy

RBP identification and DBP identification are usually considered two separate prediction tasks in most previous studies, resulting in cross-prediction problems. Although a few studies [33,34] have focused on the cross-prediction problem for the identification of nucleic acid-binding protein, they obtained low recognition precision, possibly due to the limitations of datasets or sequence feature representations, such as small training datasets and manual selection of sequence features. Here, we treat the identification of NABPs as a multiclass learning task. Multiclass classification refers to the prediction of more than two classes; in fact, there are four protein categories containing DBPs, RBPs, DRBPs and non-NABPs. In multiclass classification, we first design the classifier model, then train the model using training data and validation data, and finally predict the collected test data or independent test data into multiple class labels.

In this study, we applied the "one-vs.-all" multiclass classification technique. In one-vs.-all classification, the number of class labels present in the dataset and the number of generated binary classifiers must be the same [35,36]. Thus, we create four classifiers here for four respective classes. For each data instance, our model will output one variable containing four different values (score) to show the probabilities of the input protein predicted as the corresponding class. The output value with the largest score will be taken as the class predicted by the model. Therefore, a protein can be identified as one of four categories of proteins (DBP, RBP, DRBP or non-NABP).

In this study, we modelled a multiclass classification problem using deep neural networks, including convolutional neural networks (CNNs) and long short-term memory networks (LSTMs). It should be noted that we applied one hot encoding to reshape the categorical variable (label variable) for the data instance to implement this deep multiclass classification model.

2.3. Network architecture of DeepMC-iNABP

The fundamental architecture of the DeepMC-iNABP model is shown in Fig. 1. The architecture outlines several parts: (1) deep representation learning for protein sequences, (2) the combination of the two neural networks (CNN and LSTM), and (3) a multiclass fully connected network. First, one-hot encoding and deep representation learning with neural networks were utilized to convert the protein sequence into a fixed-dimension matrix of vector embeddings. CNNs have the advantage of selecting good feature, and LSTMs have a good ability to learn sequential data, which is of great importance in protein structure and function prediction [37–43]. In this study, CNN and LSTM were combined for protein sequence feature learning [44,45]. Moreover, a one-vs.-all multiclass fully connected network was employed to address the cross-prediction problem, in which each of the output nodes represents a different class. More specifically, the network of the DeepMC-iNABP model contains 15 layers, including an input layer, an embedding layer, 3 convolution layers, 3 pooling layers, 4 dropout layers, a long short -term layer, a dense layer, and an output layer.

Table 1
Data collected in this study for training, validation and testing.

Classes	Train data	Validation data	Test data	In total
DBP chains	9,720	1,080	1,200	12,000
RBP chains	9,720	1,080	1,200	12,000
DRBP chains	972	108	120	1,200
Non-NABP chains	9,720	1,080	1,200	12,000
In total	30,132	3,348	3,720	37,200

Abbreviations: DBP, DNA-binding protein; RBP, RNA-binding protein; DRBP, DNA- and RNA- binding protein; non-NABP, non-nucleic acid-binding protein.

Table 2
Independent test datasets.

Test datasets	DBPs	RBPs	DRBPs	non-NABP	In total
TEST474	175	68	8	233	474
DRBP206	103	0	0	103	206

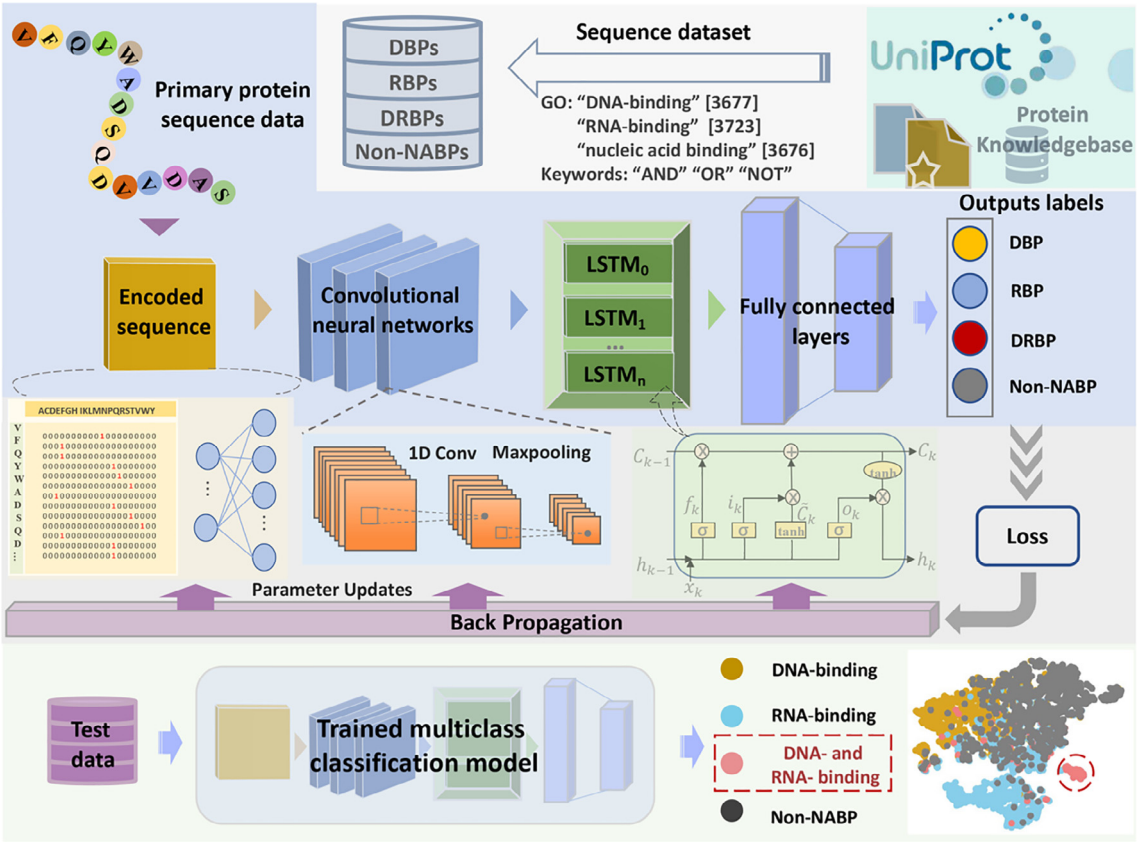

Fig. 1. Fundamental architecture of the DeepMC-iNABP model.

2.4. Sequence representation learning

In protein-related tasks, deep learning approaches can directly learn rich data representations directly from primary sequences instead of manually extracting feature information for protein sequence numeric representation, as in traditional machine learning methods [46–48]. Deep representation learning in an end-to-end model [49–51] can be relatively easy to implement because the intermediate layers, including sequence embedding layers and other neural network layers for prediction between inputs and outputs, are trained as a whole part (i.e., treated as a "black box"). Here, we utilized the widely used end-to-end representation scheme, a one-hot encoding-based deep representation of protein sequences.

2.5. Combination of CNNs and LSTMs

In this study, we combine CNNs and LSTMs to conduct a multiclass classification model. Multiple convolutional filters slide over the embedded sequence matrix to produce a new feature map in convolutional layers. In the max-pooling layer, the maximum value is calculated as a feature corresponding feature to the specific filter. Then, the outputs of the max-pooling layer are inputted into the cell of the LSTM layer to learn the long-term dependencies of motifs (features of sequences). The output vectors of the LSTM cells are concatenated and become inputs to the dense layer. A softmax activation function is applied to generate the final outputs containing 4 output values (one for each class).

2.6. Loss function

The loss function assesses the distance between the predicted output and the actual output to evaluate the performance of the neural network during training. For multiclass classification tasks, the loss function of the neural network usually chooses categorical cross entropy, since one example can be considered to belong to a specific category with probability 1, and to other categories with probability 0. The categorical cross-entropy loss function calculates the loss of an example by computing the following sum:

$$LOSS = -\sum_{i=1}^{N} y_i \cdot log\widehat{y}_i$$

where $\widehat{y}_i$ is the i-th value in the model output, y_i is the corresponding target value (true value), and N is the number of values in the model output (total number of classes).

2.7. Deep neural network parameters

As the length of input sequence was set to 1,000 (processing of padding or truncating), the number of nodes in the input layers of DeepMC-iNABP model was set as the number of the length of each sequence. As DeepMC-iNABP was designed to implement multiclass classification, the number of nodes in output layer was set just the same as the number of categories, i.e., 4. In embedding layer, the output dimensionality was set to 100 which showed the best performance after comparing the results of serval situations (64, 100, 110, 120). In DeepMC-iNABP, there are three convolutional layers, and the number of filters in each convolutional layer was set to 128, 64, 32; the number of kernels was set to 10, 5, 3, respectively. Max pooling is added to CNNs following individual convolutional layers. A dropout layer was designed after each convolutional part which helps to prevent overfitting, as the - dropout layer randomly sets input units to 0 with a frequency of rate at each step during training time. The dropout rate was set to 0.2. The output dimension of the LSTM layer was set to 70. All layers used rectified linear unit (ReLU) activation function except the output layer. The activation function of the output layer used softmax function. For the optimizer, we chose Adam, and the learning rate was set to 0.001. We chose categorical crossentropy loss function, and epoch was set to 100.

2.8. Performance evaluation

To evaluate the performance of the prediction, the assessment measurements used herein included accuracy, recall, precision and F1 score. These are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1score = \frac{2 \times TP}{2 \times TP + FP + FN}$$

where TP, TN, FN, and FP are the numbers of true positives, true negatives, false negatives, and false positives, respectively. Among these measures, recall indicates the accuracy of predicting positive samples, precision is the ratio of correctly predicted positive observations to total predicted positive observations, F1 score is the weighted average of precision and recall, and accuracy as the most intuitive performance measure is simply a ratio of correctly predicted observations to total observations.

For all measures, the maximum value is 1, and the minimum value is 0. A value of 0 indicates the worst performance, which means that the predicted observations differ greatly from the actual observations, while a value of 1 indicates the best performance, which means that the predicted observations are very close to the actual observations.

Moreover, an AU-ROC curve (area under the receiver operating characteristics curve) was used to visualize the performance of the multiclass classification problem in this study. The ROC curve is a graph showing the performance of a classification model at all classification thresholds, which can evaluate classifier output quality. ROC curves typically feature true-positive rates on the Y axis, and false-positive rates on the X axis, which means that a larger area under the curve (AUC) is usually better.

3. Results and discussion

3.1. Performance of the DeepMC-iNABP predictor on the test data and independent dataset

We evaluated the performance of DeepMC-iNABP on the test data we collected and two dependent datasets (TEST474 and DRBP206 datasets). The corresponding results are shown in Table 3 and Fig. 2.

The confusion matrix shown in Fig. 2A shows the results of multiclass classification on the test data collected in this study. It shows that 1024 of 1200 total sequences of non-NABPs (Class 0), 848 of 1200 total sequences of DBPs (Class 1), 983 of 1200 total sequences of RBPs (Class 2) and 75 of 120 total sequences of DRBP (Class 3) were correctly identified. Most multiclass data in the test data were correctly classified. From the confusion matrix (Fig. 2A), 64 and 112 sequences of the non-NABP class were identified as DBPs and RBPs, respectively; 241 sequences of the DBP class and 182 sequences of the RBP class were identified as non-NABPs. Furthermore, 109 sequences of the DBP class were identified as RBPs, while 31 sequences of the RBP class were identified as DBPs. However, few non-NABPs, DBPs and RBPs were identified as DRBPs. Moreover, the evaluation metrics, including precision, recall, F1 score and accuracy, were calculated, as shown in Table 3, to assess the performance of the DeepMC-iNABP predictor on the test data we collected. These results indicate that the evaluation values of DBP, RBP, DRBP and non-NABP are good, especially the classification of DRBP. Fig. 2B, 3C and 3D show the ROC curves and AUCs of DeepMC-iNABP on test data collected in this study, TEST474 and DRBP206, respectively.

The confusion matrix (Fig. 2A), ROC-AUC curves (Fig. 2B) and evaluations (Table 3) indicate that the DeepMC-iNABP predictor achieves good results on the collected test data. Meanwhile, the ROC curves and AUCs in Fig. 2C and 3D also show that the DeepMC-iNABP model preforms well on the two independent test datasets.

3.2. Comparison with other methods on the independent dataset TEST474

For objective evaluation, we compared our model with other nucleic acid-binding protein classification methods on the independent test dataset TEST474. The predicted results are presented in Table 4 and Fig. 3.

To date, there are almost no multiclass classifiers for NABPs. The iDRBP_MMC predictor [32] is actually composed of two binary classifiers (DBP predictor and RBP predictor), not a multiclass classifier. However, the combination of two classifiers can achieve the

Table 3
Performance evaluation of test data collected in this study.

	Precision	Recall	F1-score
DBP class	0.889	0.707	0.787
RBP class	0.804	0.819	0.812
DRBP class	0.926	0. 625	0.746
Non-NABP class	0.700	0.853	0.769
Average value	0.835	0.725	0.759

Abbreviations: DBP, DNA-binding protein; RBP, RNA-binding protein; DRBP, DNA- and RNA- binding protein; non-NABP, non-nucleic acid-binding protein.

classification of non-NABP, DBP, RBP and DRBP. The DeepDRBP-2L predictor utilizes a two-level framework, in which the first level detects NABP or not and the second level further identifies the predicted NABP as DBP, RBP or DRBP [34]. In short, the DeepDRBP-2L predictor consists of a binary classifier (the first level) and a multiclass classifier (the second level). Because of the two-level strategy, it can finally achieve the classification of non-NABP, DBP, RBP and DRBP. Thus, we compared with these two existing nucleic acid-binding protein predictors. The evaluation metrics and predicted results have shown in Table 4, Fig. 3, Fig. 4A and 4C.

DeepMC-iNABP predicted 5 proteins as DRBPs, 4 of which are true DRBPs (Fig. 3C). The precision, F1 score and recall of DeepMC-iNABP are 0.80, 0.615 and 0.50, respectively. While iDRBP_MMC and DeepDRBP-2L predicted 3 and 23 proteins as DRBPs, respectively, but only one of them was a true DRBP (Fig. 3A and 4B). The precision, F1 score and recall of DRBPs in DeepMC-iNABP are far beyond those of iDRBP_MMC and DeepDRBP-2L. For the identification of DBPs and RBPs, DeepMC-iNABP predicted 168 proteins as DBPs, 146 of which were true DBPs, 6 of which were RBPs; 56 proteins were RBPs, of which 40 were true RBPs, and 6 were true DBPs. Whereas DeepDRBP-2L identified 7 proteins of true RBPs as DBPs, and 6 proteins of true DBPs as RBPs; iDRBP_MMC predicted 2 true RBPs as DBPs, and 4 true DBPs as RBPs.

According to the results in this study, DeepMC-iNABP alleviates the cross-prediction problem to a certain extent. These results are likely to be related to the multiclass classification strategy adopted in this study, in which DBPs and RBPs as data of different classes were used for training the classification model. However, although a deep neural network-based model is able to automatically learn features from primary protein sequences, because similarity sequences exist between DBPs and RBPs, it may affect sequence representation learning, which may affect the identification of DBPs and RBPs. Further studies that take feature learning of similar sequences for deep neural networks into account will need to be undertaken.

For the prediction of all proteins, including DBPs, RBPs, DRBPs and non-NABPs, the recall, precision and F1 score of DeepMC-iNABP were much better than those of the DeepDRBP-2L predictor, as shown in Table 4. Compared with the iDRBP_MMC predictor, our model outperforms it at the identification of DRBPs, although it is slightly closer to it at predicting the DBPs, RBPs and non-NABPs. The results show that the identification of DRBPs is a strong point

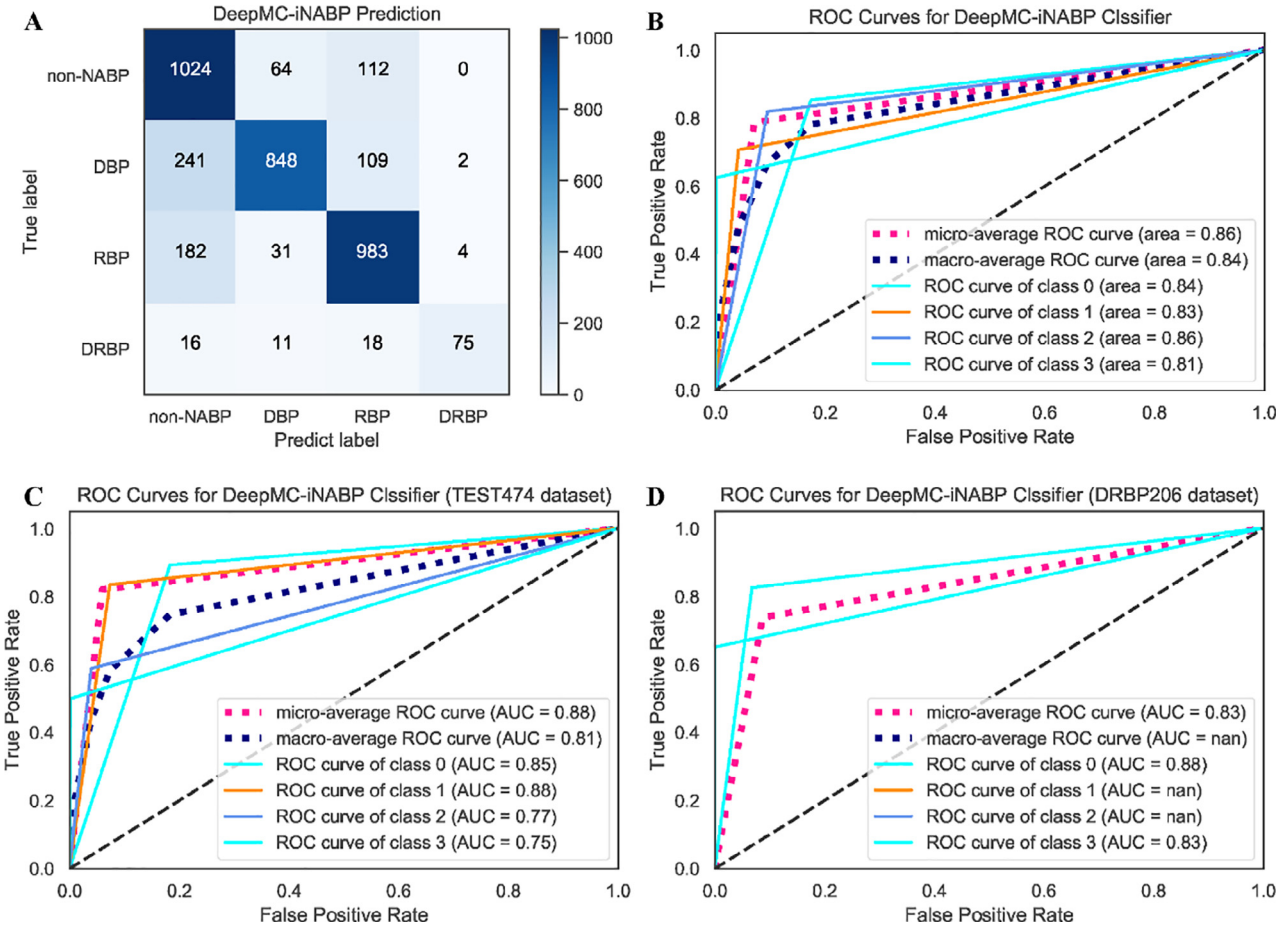

Fig. 2. Performance of the DeepMC-iNABP model on the test dataset and independent test datasets. **A and B.** Confusion matrix and ROC-AUC curves of our model on the test dataset collected in this study, **C and D.** ROC-AUC curves of our model on independent test datasets (TEST474 and DRBP206). Class 0–3 in ROC-AUC curves refer to non-NABPs, DBPs, RBPs and DRBPs, respectively.

Table 4
Comparison of DeepMC-iNABP and existing models on the independent dataset TEST474.

Model	DNA-binding			RNA-binding			DNA- and RNA-binding			non-NABP		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
DeepDRBP-2L[a]	0.817	0.877	0.846	0.456	0.620	0.525	0.125	0.043	0.065	0.888	0.832	0.859
iDRBP_MMC [b]	0.869	0.950	0.907	0.706	0.727	0.716	0.125	0.333	0.182	0.933	0.849	0.889
DeepMC-iNABP	0.834	0.869	0.851	0.588	0.714	0.645	**0.500**	**0.800**	**0.615**	0.892	0.812	0.850

[a] The results were obtained using the webserver of DeepDRBP-2L [34].
[b] The results were obtained using the webserver of iDRBP_MMC [32].

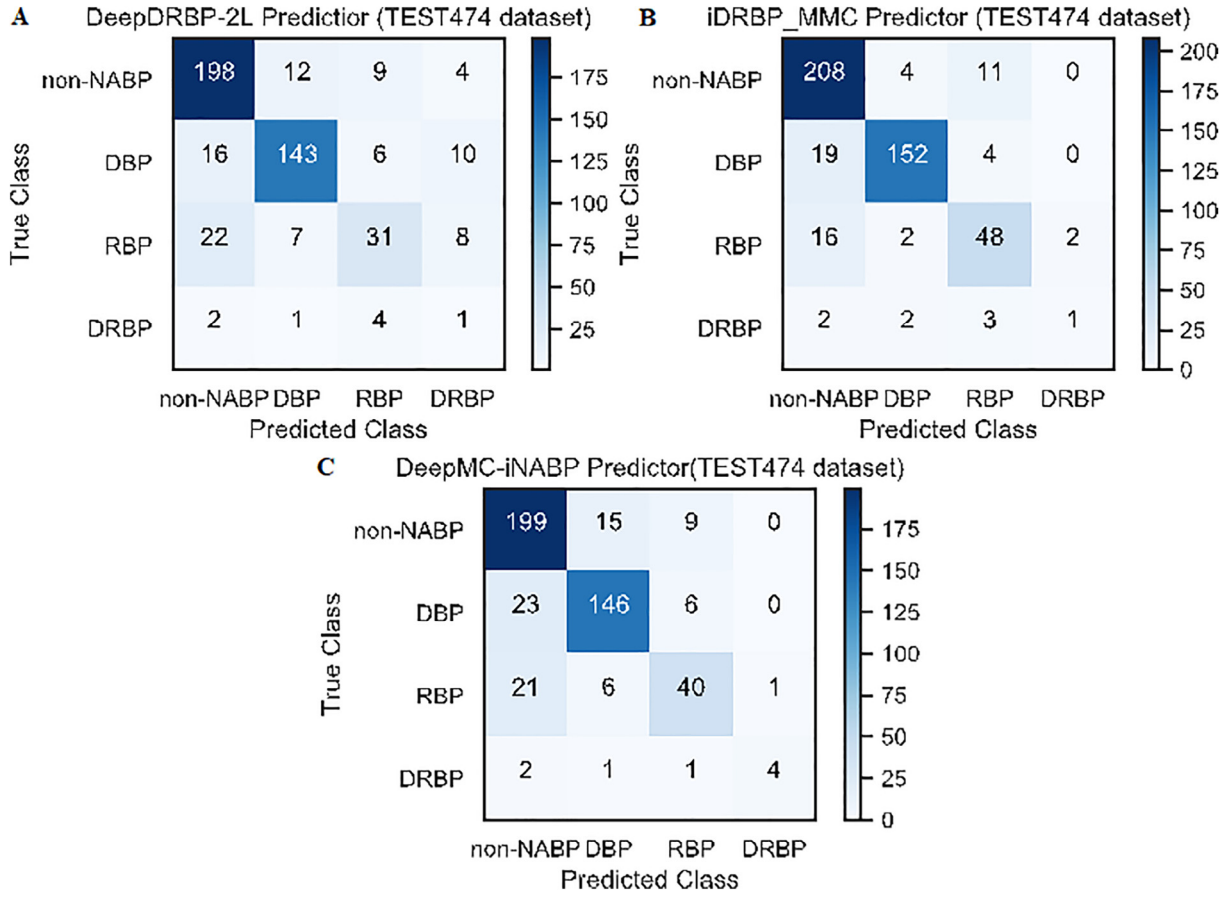

Fig. 3. Confusion matrix of DeepMC-iNABP and existing models on the independent dataset TEST474.

of the DeepMC-iNABP predictor, rather than ignoring the existence of DRBPs like other methods.

3.3. Performance for identifying DNA- and RNA-binding proteins

For the purpose of comprehensively evaluating the ability of the DeepMC-iNABP predictor to identify DRBPs, we utilized another independent dataset called DRBP206 which contains only DRBPs and non-NABPs. Fig. 4B presents the performance of DeepMC-iNABP and the existing predictors on the DRBP206 dataset. According to the comparison, the accuracy, recall and F1 score of DeepMC-iNABP on the DRBP206 test dataset were quite higher than those of DeepDRBP-2L and iDRBP_MMC. The precision values of DeepMC-iNABP were slightly better or almost equal to those of iDRBP_MMC but were even better than those of DeepDRBP-2L. Taken together, these results tested on the DRBP206 dataset suggest that DeepMC-iNABP does have a strong advantage in identifying DRBPs. Moreover, feature visualization of represented sequence data also indicates that DeepMC-iNABP model classified the four types of protein sequences well, especially the DRBPs as shown in Fig. 5.

In the current study, comparing our model with some NABP predictors showed that DeepMC-iNABP can successfully identify DRBPs, while few predictors, such as DeepDRBP-2L and iDRBP_MMC introduced above, may have the ability to recognize DRBPs but perform very poorly. A possible explanation for this might be that DeepDRBP-2L or iDRBP_MMC applied binary classification or multilabel classification, whereas the DeepMC-iNABP predictor employed a multiclass classification scheme, and DRBP data as a separate class were used to train the DeepMC-iNABP model.

4. Conclusions

There are two challenges in the prediction of NABP: one is the problem of ignoring DRBPs, and the other is the cross-predicting problem. Very little was found in the prior studies. In this study, we focused on these two problems, and proposed a NABP predictor,

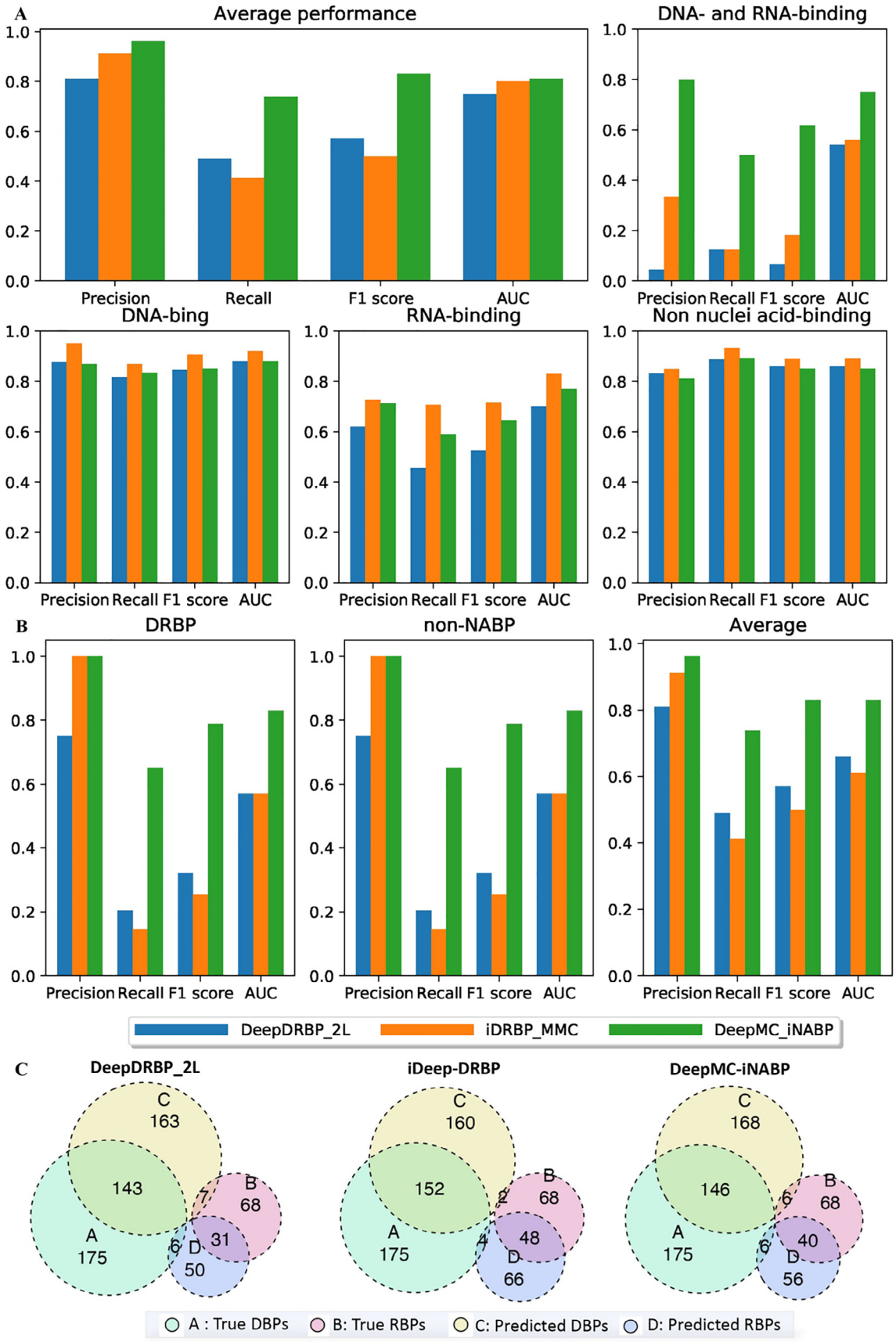

Fig. 4. Comparison of DeepMC-iNABP and existing models on the independent dataset. **A** and **C.** Independent dataset TEST474, **B.** Independent dataset DRBP206.

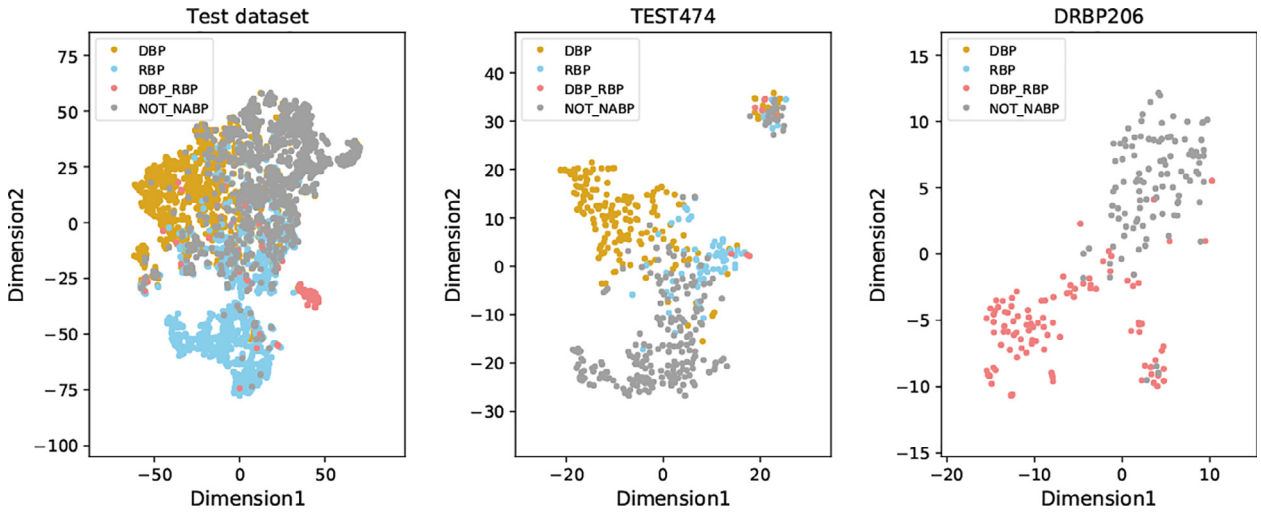

Fig. 5. Feature visualization of DeepMC-iNABP by t-SNE for dimension reduction. **A.** Feature representation of test dataset collected in this study, **B.** Feature representation of independent dataset TEST474, **C.** Feature representation of independent dataset DRBP206.

called DeepMC-iNABP, with the goal of solving these difficulties by utilizing a multiclass classification strategy and deep learning approaches. The classic deep learning model, the architecture of which contains one-hot encoding-based sequence representation and the combination of deep neural networks (CNN and LSTM), was constructed for identifying the NABPs. DBPs, RBPs, DRBPs and non-NABPs were used as separate classes of data for training the DeepMC-iNABP model. The results on several test datasets showed that DeepMC-iNABP has a strong advantage in identifying DRBPs and alleviates the cross-prediction problem to a certain extent. Moreover, the web server of DeepMC-iNABP (https://www.deepmc-inabp.net/) was provided, which will be useful for researchers in the field of protein-nucleic acid interactions.

5. Code availability

The web server of DeepMC-iNABP, data resource and codes are freely available from https://www.deepmc-inabp.net/.

Funding

The work was supported by the National Natural Science Foundation of China (No. 61922020, No. 62102064 and No.62101100) and the Special Science Foundation of Quzhou (2021D004).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Yin Y, Morgunova E, Jolma A, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science 2017;356.
[2] Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res 2017;45:e84.
[3] Dou L, Yang F, Xu L, et al. A comprehensive review of the imbalance classification of protein post-translational modifications. Brief Bioinform 2021.
[4] Xin H, Deng K, Fu M. Post-transcriptional gene regulation by RNA-binding proteins in vascular endothelial dysfunction. Sci China Life Sci 2014;57:836–44.
[5] Gyebi GA, Ogunyemi OM, Ibrahim IM, et al. Dual targeting of cytokine storm and viral replication in COVID-19 by plant-derived steroidal pregnanes: An in silico perspective. Comput Biol Med 2021;134:104406.
[6] Hu Y, Qiu S, Cheng L. Integration of multiple-omics data to analyze the population-specific differences for coronary artery disease. Comput Math Methods Med 2021;2021:7036592.
[7] Qiu S, Hu Y, Cheng L. BIN1 rs744373 located in enhancers of brain tissues upregulates BIN1 mRNA expression, thereby leading to Alzheimer's disease. Alzheimers Dement 2022.
[8] Sebestyen E, Singh B, Minana B, et al. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res 2016;26:732–44.
[9] van Kouwenhove M, Kedde M, Agami R. MicroRNA regulation by RNA-binding proteins and its implications for cancer. Nat Rev Cancer 2011;11:644–56.
[10] Ao C, Yu L, Zou Q. Prediction of bio-sequence modifications and the associations with diseases. Brief Funct Genomics 2021;20:1–18.
[11] Jiao S, Zou Q, Guo H, et al. iTTCA-RF: a random forest predictor for tumor T cell antigens. J Transl Med 2021;19:449.
[12] Zhu T, Dai Q, He P-A. Identification of Potential Immune-related Biomarkers in Gastrointestinal Cancers. Curr Bioinform 2021;16:1203–13.
[13] Schmidt N, Lareau CA, Keshishian H, et al. The SARS-CoV-2 RNA-protein interactome in infected human cells. Nat Microbiol 2021;6:339–53.
[14] Tabaie A, Orenstein EW, Nemati S, et al. Predicting presumed serious infection among hospitalized children on central venous lines with machine learning. Comput Biol Med 2021;132:104289.
[15] Pan H, Jin M, Ghadiyaram A, et al. Cohesin SA1 and SA2 are RNA binding proteins that localize to RNA containing regions on DNA. Nucleic Acids Res 2020;48:5639–55.
[16] Peled S, Leiderman O, Charar R, et al. De-novo protein function prediction using DNA binding and RNA binding proteins as a test case. Nat Commun 2016;7:13424.
[17] Hudson WH, Ortlund EA. The structure, function and evolution of proteins that bind DNA and RNA. Nat Rev Mol Cell Biol 2014;15:749–60.
[18] Zhang Z, Cui F, Wang C, et al. Goals and approaches for each processing step for single-cell RNA sequencing data. Brief Bioinform 2021;22. bbaa314.
[19] Zhang Z, Cui F, Lin C, et al. Critical downstream analysis steps for single-cell RNA sequencing data. Briefings Bioinf 2021.
[20] Cao C, Kwok D, Edie S, et al. kTWAS: integrating kernel machine with transcriptome-wide association studies improves statistical power and reveals novel genes. Briefings Bioinf 2021;22. bbaa270.
[21] Zhang S, Su Q, Chen Q. Application of machine learning in animal disease analysis and prediction. Curr Bioinform 2021;16:972–82.
[22] Tohka J, van Gils M. Evaluation of machine learning algorithms for health and wellness applications: A tutorial. Comput Biol Med 2021;132:104324.
[23] Ramzan Z, Hassan MA, Asif HMS, et al. A machine learning-based self-risk assessment technique for cervical cancer. Curr Bioinform 2021;16:315–32.
[24] Hu Y, Sun JY, Zhang Y, et al. rs1990622 variant associates with Alzheimer's disease and regulates TMEM106B expression in human brain tissues. BMC Med 2021;19:11.
[25] Hu Y, Zhang H, Liu B, et al. rs34331204 regulates TSPAN13 expression and contributes to Alzheimer's disease with sex differences. Brain 2020;143:e95.
[26] 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;10:1–13.
[27] Cui F, Zhang Z, Cao C, et al. Protein–DNA/RNA interactions: Machine intelligence tools and approaches in the era of artificial intelligence and big data. Proteomics 2022. 2100197.
[28] Da L-T, Pardo-Avila F, Xu L, et al. Bridge helix bending promotes RNA polymerase II backtracking through a critical and conserved threonine residue. Nat Commun 2016;7:1–10.

[29] Yan J, Friedrich S, Kurgan L. A comprehensive comparative review of sequence-based predictors of DNA-and RNA-binding residues. Briefings Bioinf 2016;17:88–105.
[30] Boutet E, Lieberherr D, Tognolli M, et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. Plant Bioinform Springer 2016:23–54.
[31] Ye J, McGinnis S, Madden TL. BLAST: improvements for better sequence analysis. Nucleic Acids Res 2006;34:W6–9.
[32] Zhang J, Chen Q, Liu B. iDRBP_MMC: identifying DNA-binding proteins and RNA-binding proteins based on multi-label learning model and motif-based convolutional neural network. J Mol Biol 2020;432:5860–75.
[33] Xu L, Jiang S, Wu J, et al. An in silico approach to identification, categorization and prediction of nucleic acid binding proteins. Brief Bioinform 2021;22.
[34] Zhang J, Chen Q, Liu B. DeepDRBP-2L: A new genome annotation predictor for identifying DNA-binding proteins and RNA-binding proteins using convolutional neural network and long short-term memory. IEEE/ACM Trans Comput Biol Bioinform 2021;18:1451–63.
[35] Rifkin R, Klautau A. In defense of one-vs-all classification. J Machine Learn Res 2004;5:101–41.
[36] Galar M, Fernández A, Barrenechea E, et al. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. Pattern Recogn 2011;44:1761–76.
[37] Alayba AM, Palade V, England M, et al. A combined CNN and LSTM model for arabic sentiment analysis. In: International cross-domain conference for machine learning and knowledge extraction. Springer; 2018. p. 179–91.
[38] Alipanahi B, Delong A, Weirauch MT, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 2015;33:831–8.
[39] Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res 2016;44:e107.
[40] Hou J, Adhikari B, Cheng J. DeepSF: deep convolutional neural network for mapping protein sequences to folds. Bioinformatics 2018;34:1295–303.
[41] Chen J, Zou Q, Li J. DeepM6ASeq-EL: prediction of human N6-methyladenosine (m6A) sites with LSTM and ensemble learning. Front Comput Sci 2022;16:162302.
[42] Sharma AK, Srivastava R. Protein secondary structure prediction using character bi-gram embedding and Bi-LSTM. Curr Bioinform 2021;16:333–8.
[43] Rafiei A, Rezaee A, Hajati F, et al. SSP: Early prediction of sepsis using fully connected LSTM-CNN model. Comput Biol Med 2021;128:104110.
[44] Neeraj, Singhal V, Mathew J, et al. Detection of alcoholism using EEG signals and a CNN-LSTM-ATTN network. Comput Biol Med 2021;138:104940.
[45] Dastider AG, Sadik F, Fattah SA. An integrated autoencoder-based hybrid CNN-LSTM model for COVID-19 severity prediction from lung ultrasound. Comput Biol Med 2021;132:104296.
[46] Ao C, Zhou W, Gao L, et al. Prediction of antioxidant proteins using hybrid feature representation method and random forest. Genomics 2020;112:4666–74.
[47] Dou L, Li X, Zhang L, et al. iGlu_AdaBoost: identification of lysine glutarylation using the Adaboost classifier. J Proteome Res 2020;20:191–201.
[48] Liu B. BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. Brief Bioinform 2019;20:1280–94.
[49] Cui F, Zhang Z, Zou Q. Sequence representation approaches for sequence-based protein prediction tasks that use deep learning. Brief Funct Gen 2021;20:61–73.
[50] Naseer S, Hussain W, Khan YD, et al. NPalmitoylDeep-pseaac: A predictor of N-palmitoylation sites in proteins using deep representations of proteins and PseAAC via modified 5-steps rule. Curr Bioinform 2021;16:294–305.
[51] Edera AA, Small I, Milone DH, et al. Deepred-Mt: Deep representation learning for predicting C-to-U RNA editing in plant mitochondria. Comput Biol Med 2021;136:104682.