Methods 181-182 (2020) 15–23

Contents lists available at ScienceDirect

Methods

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

econvRBP: Improved ensemble convolutional neural networks for RNA binding protein prediction directly from sequence

Yuze Zhao, Xiuquan Du*

School of Computer Science and Technology, Anhui University, Hefei, Anhui, China

ARTICLE INFO

Keywords:
Ensemble deep learning
RNA binding proteins
Convolutional neural network
Residual block

ABSTRACT

RNA binding proteins (RBPs) determine RNA process from synthesis to decay, which play a key role in RNA transport, translation and degradation. Therefore, exploring RBPs' function from the amino acid sequence using computational methods has become one of the momentous topics in genome annotation. However, there still have some challenges: (1) shallow feature: Although the sequence determines structure is self-evident, it is difficult to analyze the essential features from simple sequence. (2) Poorly understand: feature-based prediction methods mainly emphasize feature extraction, while in-depth understanding of protein mysteries limits the application of feature engineering. (3) Feature fusion: multi-feature fusion is often used, but the features are not well integrated. In view of these challenges, we propose a novel ensemble convolutional neural network (econvRBP) to predict RBPs. In order to capture the local and global features of RNA binding proteins simultaneously, first of all, One Hot and Conjoint Triad encoding methods are used to transform amino acid sequence into local and global features, respectively. After that the local and global features are combined for further high-level feature extraction using convolutional neural networks. Some experiments are constructed to evaluate our method with 10-fold cross validation and the results show that it has achieved the best performance among all the predictors so far. We correctly predicted 99% of 2875 RBPs and 99% of 6782 non-RBPs with accuracy of 0.99. In addition, the datasets provided by RBPPred are also used to validate our models with an accuracy of 0.87. These results indicate that the econvRBP is the most excellent method at present, and will provide reliable guidance for the detection of RBPs. econvRBP is available at http://47.100.203.218:3389/home.html/.

1. Introduction

RNA binding proteins (RBPs) are a general term for a class of proteins that bind to mRNA or non-coding RNA. These RBPs are involved in the regulation and metabolism of RNA and participate in all aspects of RNA processing, and control the life cycle of RNA from synthesis to degradation [10]. Without RBPs, it is not an exaggeration to say, RNA can't do anything. As we know, RBPs account for 5–10% of eukaryotic proteomes and are important in post-transcriptional biological processes such as gene regulation, alternative splicing and translation [1]. From neurological disorders to cancer, RBPs participate in many human diseases [20]. For example, TDP-43 is an RBP that can cause amyotrophic lateral sclerosis when it is mutated [22]. And the expression of Sam68 will lead to further proliferation of prostate cancer cells and survival of cytotoxic substances [3]. Therefore, the recognition of RBPs and the understanding of their regulatory mechanisms are crucial to solve physiology and disease.

Unfortunately, RBPs prediction is far from enough so far. RBPs in all kinds of species cannot be fully detected. It has been an important challenge in the field of genomic annotation. In recent years, some high-throughput experimental techniques have been developed to detect RBPs. The proteome-wide identification of RBPs, especially the RNA interactome capture (RIC) and its modifications have become powerful tools for RBPs identification. For instance, Kwon et al. successfully detected 555 mRNA binding proteins in mouse embryonic stem cells by combining UV cross-linking of RBPs to RNA in living cells [16]. Although RIC has been successfully applied to detect RBPs, experimental variability and technical noise limit its utility. Perez-Perri et al. proposed an improved method, enhanced RIC (eRIC), which significantly improved specificity and increased signal-to-noise ratio by using a locked nucleic acid (LNA)-modified capture probe, and optimized washing conditions, so as to enhance the detection capability of

This document is the results of the research project funded by the Anhui Provincial Natural Science Foundation under Grant 1708085QF143.

* Corresponding author.
E-mail address: dxqllp@163.com (X. Du).

https://doi.org/10.1016/j.ymeth.2019.09.008
Received 21 May 2019; Received in revised form 21 August 2019; Accepted 5 September 2019
Available online 09 September 2019

RBPs [25]. Determining the function of a large number of proteins experimentally, however, is a nearly impossible task [23]. This gives birth to a series of computational methods based on machine learning algorithms, such as Random Forest (RF) and Support Vector Machine (SVM). BindN+ proposed a method based on homologous sequence alignment [30]. It effectively improved the evolution information descriptor and experimental results showed that the improved descriptor better reflects the evolution information of proteins, and yielded MCC of 0.440. Zhao et al. were inspired by the previous successful prediction of RNA-binding domains and RNA-binding sites (SPOT-stru) [9], they developed SPOT-seq with structure and sequence. This method broke the traditional computational method based on sequence homology or evolutionary information between characterized and un-characterized proteins, and yielded a MCC of 0.61 [33]. Kumar et al. took the lead in proposing PPRINT to predict the protein binding residues [14]. If the percentage of protein binding residues exceeds a certain threshold, it will be directly determined as RBPs; otherwise, the amino acid sequence will be encoded using PSSM-400 and classified by SVM [15]. This hybrid approach yielded a MCC of 0.62. RNABindRPlus combined homologous sequence alignment method HomPRIP and machine learning method RNABindRPlus to study RNA-binding residues [29]. HomPRIP and RNABindRPlus achieved an MCC of 0.83 and 0.37 on RB111. Ma et al. put forward a new method to predict RBPs directly from the amino acid sequence, PRBP. They combined evolutionary information with six physicochemical properties including the pKa value of the amino group, the pKa value of the carboxyl group, the molecular mass, the electron-ion interaction potential (EIIP), the number of lone pairs, and the Wiener index, and then used the RF classifier to identify RBPs with the MCC of 0.66 [21]. Shazman et al. studied the structure and electrostatic properties of proteins and classified them using SVM, called NAbind algorithm [26,24]. And then they developed the web server called BindUP based on NAbind algorithm. In order to further improve the performance, Zhang et al. proposed SVM-based method, called RBPPred. In their method, amino acid was encoded according to hydrophobicity, polarity, normalized van der Waals volume, polarizability, predicted secondary structure, predicted solvent accessibility, charge and polarity of side chain, Position Specific Scoring Matrix (PSSM) profile. Then the combined features were sent to the SVM for prediction and yielded a MCC of 0.808 [32]. However, all of the above methods have obvious problems: traditional machine learning algorithms often require complex feature engineering, which usually include the following steps: (1) perform a deep exploratory data analysis on the dataset. (2) Do a simple dimensionality reduction process. (3) Feature selection. The emergence of deep learning provides a new way to predict RBPs. Feature engineering is not necessary when using deep learning. Deep learning forms a more abstract high-level representation by combining low-level features to discover distributed feature representations of data [17]. It has beaten machine learning in many areas such as natural language processing [28], computer vision [12,13], and drug discovery [11]. In addition, deep learning techniques have been applied in many aspects of bioinformatics [2,31] and proven to be a powerful tool. For example, Deep-RBPPred introduced deep learning technology into the prediction of RBPs for the first time [34]. Compared with RBPPred, this method only adopt six properties including hydrophobicity, polarity, normalized van der Waals volume, polarizability, charge and polarity of side chain. These physicochemical properties are fed into CNN to train its weights. However, in the 160 dimensional feature vector of Deep-RBPPred, 84 dimensional is its physicochemical properties, 64 dimensional is its global properties, 12 dimensional zero-padding. Hence, its physicochemical properties are preferred and the global properties are ignored. And its features are just simple stacked. Furthermore, Du et al. fused the protein features from multi-view and used deep belief networks (DBN) to predict RBPs, DeepMVF-RBP, and yielded a MCC of 0.818 [6]. Taking Deep-RBPPred and DeepMVF-RBP as examples, these methods still need to extract complicate features. In fact, it is not necessary to extract many physicochemical properties mentioned above, only to carry out composition-transition-distribution (C-T-D) transformation [7]. To avoid complicated feature engineering and feature stacking, we propose ensemble deep learning method. Ensemble deep learning can make a comprehensive measurement of the features extracted by deep learning, and result that the model more generalizable. Therefore, we only need to provide comprehensive and basic features to our model. All remains are to focus on the design of model structure, which will undoubtedly make the prediction faster and more effective.

In this paper, we present RBPs prediction method, namely econvRBP. This method combines local features with global features and adopts different CNN structures for each feature. Different encoding methods and structures give the model different perspectives. Model analyzes the amino acid sequence, and finally output the prediction probability. Compared with the previous methods, econvRBP has the following advantages:

1. econvRBP adopts ensemble deep learning to integrate weak models to build strong predictors. Therefore, it is more robust and better performance than a single model.
2. econvRBP only uses a simple Conjoint Triad encoding to express its global features, without additional manual feature extraction steps.
3. Taking into account the global-local features of the amino acid sequence, the two kinds of feature complement each other when the individual feature is not good. According to different features, CNNs with different structures are designed for feature extraction, so that CNN can analyze the input data from different perspectives
4. A friendly web server is provided for biological researchers, and has guiding significance for experimental methods. econvRBP are available at http://47.100.203.218:3389/home.html/.

2. Method

The model is shown in Fig. 1. Protein sequences are separately encoded by One Hot method and Conjoint Triad method. The One Hot encoding mainly focuses on the local features of the amino acid sequence. First of all, the sequence is analyzed by a convolution kernel, then feature's dimension is reduced by pooling, which facilitates a more global analysis of the next convolution kernel. Finally, the dimension is further reduced by pooling. The Conjoint Triad encoding mainly focuses on the electrostatic and global properties of amino acid sequence. We used 3-layers with $3 \times 3 \times 3$ convolution kernel to extract the features, replacing a single $7 \times 7 \times 7$ convolution kernel. Without changing receptive field, econvRBP can capture more nonlinear features and reduce number of parameters. We also establish a direct mapping from input to output for 3-layers convolution, called the residual block. At last, majority vote method is used to integrate two models and output the finally probability. The detail of the model will be introduced below.

2.1. Sequence encoding

Feature engineering is an old-fashioned problem in RBPs prediction. However, it is no longer a decisive step in deep learning. Therefore, in this paper, we prefer to call sequence encoding instead of feature extraction. We mainly focus on the integrated convolutional neural network, and sequence encoding only serves this model. Hence, we use some simple encoding methods to obtain shallow features, and the real correlation features are analyzed by CNN. The powerful ensemble convolutional neural networks learn the essential and effective features of RBPs from simple features.

The dipoles and volumes of the amino acid side chain have an impact on the electrostatic (including hydrogen bonding) and hydrophobic interactions, which in turn determine the RBPs. Hence, we divided 20 amino acids into 7 categories based on dipole and volume. Based on the density-functional theory method B3LYP/6-31G* and

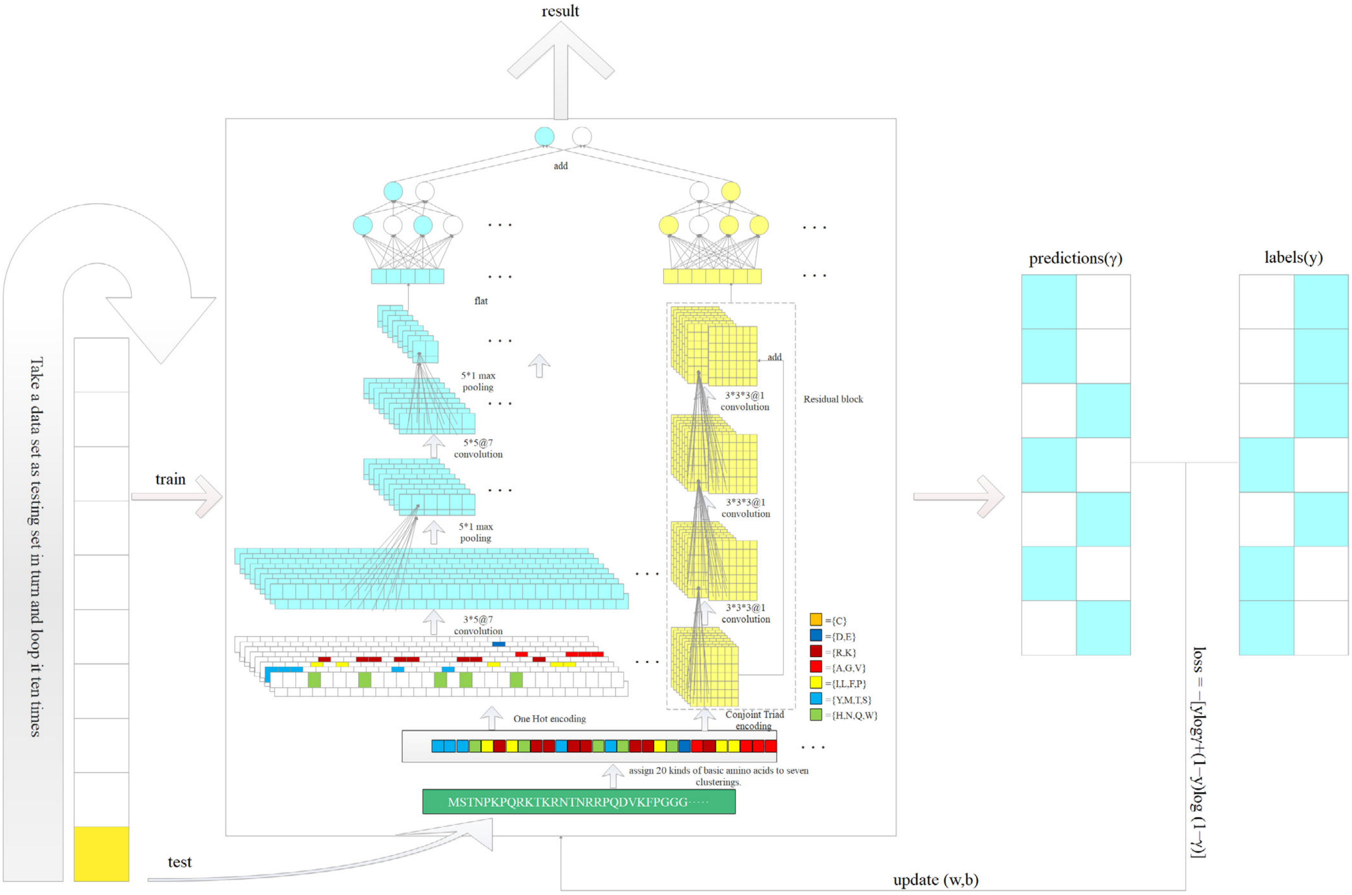

Fig. 1. The flowchart of econvRBP. As we can see, different encoding methods are sent to different network architectures. On the right side of the model, outlined by a dotted line, is the residual block.

Table 1
Classification of amino acids.

Class No.	*Dipolescale*[a]	*Volumescale*[b]	Class
1	−	−	Ala, Gly, Val
2	−	+	Ile, Leu, Phe, Pro
3	+	+	Tyr, Met, Thr, Ser
4	+ +	+	His, Asn, Gln, Tpr
5	+ + +	+	Arg, Lys
6	+′ +′ +′	+	Asp, Glu
7	+[c]	+	Cys

[a] Dipole scale (Debye): −, dipole < 1.0; +, 1.0 < Dipole < 2.0; + +, 2.0 < Dipole < 3.0; + + +, Dipole > 3.0;

[b] Volume Scale ($Å^3$): −, Volume < 50; +, Volume > 50;

[c] Cys is separated from class 3 because of its ability to form disulfide bonds.

molecular modeling approach, these two parameters were calculated [27]. The twenty essential amino acids are divided into the following seven clusters (Table 1):[A, G, V], [I, L, F, P], [Y, M, T, S], [H, N, Q, W], [R, K], [D, E], [C].

2.1.1. One hot encoding

In order to reveal the local features of the amino acid sequence, we carry out the One Hot encoding. One Hot encoding is also called a bit effective encoding. The method uses K bits to encode K states, each state has its own bit, and at any time, only one bit is effective. Given an amino acid sequence $P = (P_1, P_2, \cdots, P_n)$ with length of n and seven clusters $C = (\text{cluster}_1, \text{cluster}_2, \cdots, \text{cluster}_7)$, One Hot encoding converts an amino acid sequence into a matrix M of $n \times 7$ dimensions, i.e.:

$$M_{ij} = \begin{cases} 1, & P_i \in \text{cluster}_j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

For CNN, the dimensions of the input matrix are fixed, but the length after the One Hot encoding depends on the length of the sequence itself. Therefore, this paper uses the fixed-length amino acid sequence for One Hot encoding as an input to CNN. We work out the average length of all samples, and then cut off the amino acid sequences from the start location whose length is greater than the average. We will add zero if the sequence length less than the average. As thus, we can get the One Hot encoding matrixes of the same length.

In fact, in order to capture the local features of protein sequence, we break the sequence ($length = n$) into multiple subsequences with window size of W (Fig. 2). Meanwhile, there is overlap of length S between two adjacent subsequences. And the number of subsequences is $(n - W)/(W - S)$. Through this method, each subsequence will contain local information of protein sequence. In the experiment, we use the category of each amino acid as the input channel, the length and quantity of the subsequence as height and weight of the feature map, respectively.

2.1.2. Conjoint Triad encoding

The Conjoint Triad encoding was originally used in the prediction of protein-protein interactions proposed by Shen et al. [27]. The same method is also used herein to encode the amino acid sequences. Fig. 3 shows the detail encoding process. This encoding method takes into account the position of an amino acid and its two adjacent amino acids, and then three amino acids as the whole. Thus, each of three amino acids belongs to one of the seven clusters mentioned above, with a total of $7 \times 7 \times 7 = 343$. We count all occurrences of triples and then perform

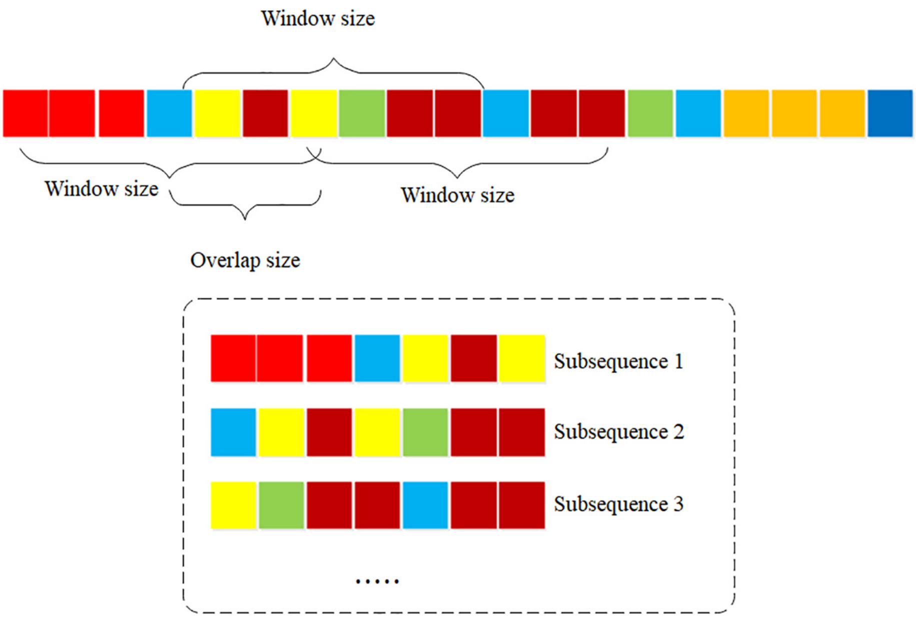

Fig. 2. Flowchart to extract local information. The sliding window slides over the sequence with the size of *w*, and finally we break the sequence into subsequence bag, where the length of each subsequence is length *w* and (n − W)/(W − S) subsequences in it.

a Min-Max normalization operation. For example, assuming that protein sequence has been mapped to a 7 × 7 × 7 dimensional feature vector $V = (V_1, V_2, \cdots, V_{343})$, after Min-Max normalization we can get the frequency vector $F = (F_1, F_2, \cdots, F_{343})$, which are defined by formula (2):

$$F_i = (V_i - \min\{V_1, V_2, \cdots, V_{343}\})/(\max\{V_1, V_2, \cdots, V_{343}\} - \min\{V_1, V_2, \cdots, V_{343}\}) \tag{2}$$

Thus, we mapped an amino acid sequence with arbitrary length into a 7 × 7 × 7 dimensional feature vector.

2.2. Ensemble convolutional neural network

2.2.1. Convolutional neural networks

CNN, which have been widely used in fields such as image and natural language processing, were raised in 1998 [18]. CNN consists of convolutional layers, pooling layers and fully connected layers. A convolutional layer, which may be composed of multiple convolution kernels, is used to extract high dimensional non-liner features. And the size of the convolution kernel determines the “field of view” of a convolutional layer, namely the receptive field. It is precisely because the multi-layer convolution kernel has different “field of view”. Because CNN is mainly a network for images, we need to make some modifications in our research. Just like the previous encoding method (One Hot and Conjoint Triad), before a protein sequence is sent to CNN, we transform it into a matrix by encoding. This matrix takes into account both the global characteristics of the protein sequence (Conjoint Triad) and the global characteristics of the protein sequence (One Hot). In fact, after we get a matrix through encoding; it is actually equivalent to an image. We now apply CNN on the prediction of RBPs, and we also use a different “field of view” to analyze the amino acid sequence. However,

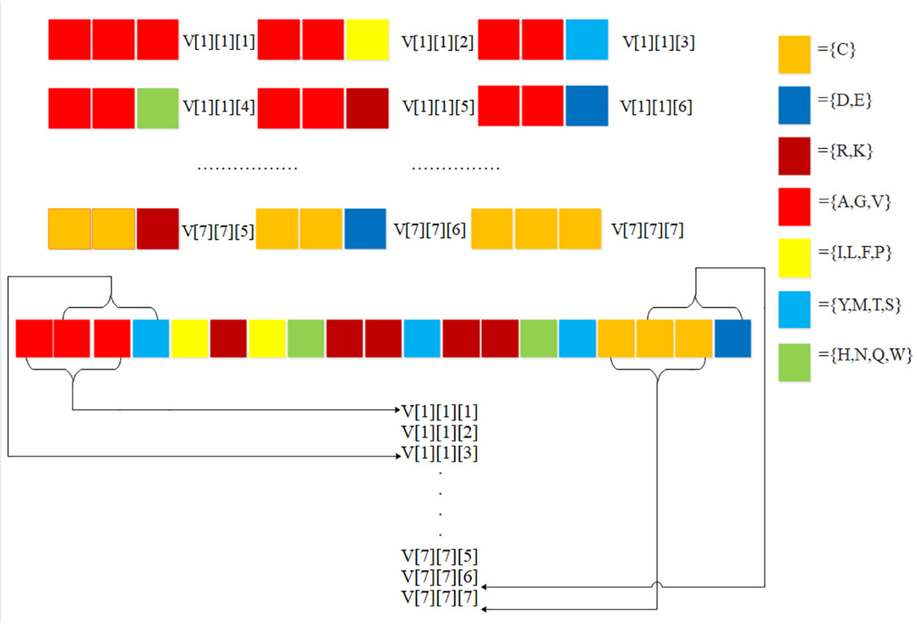

Fig. 3. Detailed process of Conjoint Triad encoding. We counted the frequency of all Conjoint Triads across the sequence.

the features directly from matrix do not contain entire effective information; we also need to "tell" CNN some high-level local and global information. In this way we analyzed sequence from different perspectives and different receptive fields. Based on different encoding methods, we respectively used two dimensional convolution structure and three dimensional convolution structure. In two dimensional convolutions, supposed that a sequence has broken into n subsequences with length w of each subsequence. We took 7 of the $n \times w \times 7$ One Hot encoding matrix as 7 channels, and the $n \times w$ matrix was carried out with two dimensional convolution. Formally, the features before and after convolution are denoted as P^l and P^{l+1}. The convolution kernel scans input features in the receptive field regularly. We call it Two Dimensional Local Convolution (2DLC), which can be formulated by (3):

$$P^{l+1}(x, y) = (P^l \otimes\ w^l)(x, y) = \sum_{i=1}^{nk} \sum_{j=1}^{L} \sum_{k=1}^{W} [P_i^l(s_0 x + j,\ s_1 y + k) w_i^l(x, y)] \tag{3}$$

where, (x, y) are pixels of the feature map, w is the parameter of the convolution kernel, nk is the number of the feature channels, L and W are the length and width of the input feature matrix, s_0 and s_1 are sliding stride of the convolution kernel on the feature map to the right and down, respectively.

In 2DLC, the first layer adopts 3 × 5 × 7@7 (motif detector) convolution to extract features (called motif in biology), and uses the maximum pooling of 5 × 1 to reduce the dimensions of features. The second layer adopts 5 × 5 × 7@7 convolution is adopted to extract higher-dimensional features, and the maximum pooling of 5 × 1 is adopted to further reduce dimensions.

In three dimensional convolution, we transformed the 7 × 7 × 7 dimensional feature vector into a feature cubic matrix of 7 × 7 × 7 according to the encoding characteristics of the Conjoint Triad encoding, and then three dimensional convolution was applied to the feature, which is called the Three Dimensional Global Convolution (3DGC):

$$\begin{aligned} P^{l+1}(x, y, z) &= (P^l \otimes\ w^l)(x, y, z) \\ &= \sum_{i=1}^{nk} \sum_{j=1}^{L} \sum_{k=1}^{W} \sum_{m=1}^{D} [P_i^l(s_0 x + j,\ s_1 y + k,\ s_2 z + m) w_i^l \\ &\quad (x, y, z)] \end{aligned} \tag{4}$$

where, (x, y, z), w, nk, L, w, s_0 and s_1 have the same meaning as above, D is the depth of the cubic matrix, and s_2 is the backward sliding stride of the convolution kernel.

In 3DGC, we use 3 × 3 × 3 × 1@1 convolution of 3 layers to extract global features.

In practical applications, we employ multi-layer convolution and then the fully connected is used for prediction. The purpose of multi-layer convolution is that the features learned by single layer convolution are local. The more layers, the more global features are learned. Different CNN structures are applied for different encoding methods, and the outputs of different structures are combined. For 2DLC, we use a two-layer convolution for feature extraction. For 3DGC, three-layer convolution is applied for feature extraction.

2.2.2. *Residual block applied on 3DGC*

The learning ability of model continues to increase as the deeper of layers. In practice, it is counterproductive. One of the reasons is that the deeper of layer leads to the problem of gradient vanishing. The regularization initialization and batch normalization can solve this problem. However, with the further deeper of the layer, the accuracy of training set will continue to decline, which is called degradation problem [8]. The model degradation problem is due to the fact that with the deepening of the neural network layer, the low dimensional feature has been sufficiently fitted by shallow layers, while the deeper layers simply become an identity mapping without learning the high dimensional features. The network degradation problem indicates that the deep model is not easy to train. So we introduce the residual block, which allows the original input information to be transmitted directly to the later layers. Assuming that the input of a deep neural network is x and the expected output is $H(x)$, the target of the deep neural network is $H(x)$ before the introduction of the residual block. After introducing the residual block, the expected output of the neural network is $H(x) = F(x) + x$, so the learning target of the deep neural network at this layer turns into $F(x) = H(x) - x$. $H(x)$ and $F(x)$ are original system function and new system function, respectively. In the case of an identity map, $H(x) = x$, $F(x) = 0$. Obviously, learning new system function $F(x)$ is much simpler than original system function $H(x)$. We observed that after the introduction of the residual block the difficulty of fitting is greatly reduced. The direct mapping from input to output makes the model have stronger learning ability and faster convergence speed.

2.2.3. *Ensemble deep learning*

Many studies have demonstrated that ensemble learning is often superior to the individual classifier, which enhances not only the performance of the classification, but also the confidence of the results. Therefore, in this study, the deep learning-based ensemble method is used to further improve the performance. Ensemble learning uses a series of learners to learn and a certain rule (majority vote) is used to integrate individual learning result to achieve better outcomes than a single learner. In general, multiple learners in ensemble learning are homogeneous "weak learners". Common machine-learning-based ensemble methods mainly include: Boosting, Bagging, Random Forest, etc. Due to the neural network model is nonlinear and has high variance, which can be frustrating when preparing a final model to make predictions. A successful way to reduce the variance of neural network models is to train multiple models rather than individual models and to combine the predictions of these models. This is called ensemble deep learning, which not only reduces the variance of the prediction, but also produces predictions that are better than any single model. Simple features turn into abstract and high dimensional complex features after feature extraction by 2DLC and 3DGC, then the fully connected layer serves as the classifier and outputs probability. Finally, we average the probabilities of the two models to get the result by majority vote.

2.3. *Performance evaluation*

Our model is trained on tensorflow 1.9.0 (https://github.com/tensorflow/tensorflow), which supports GPU-accelerated calculation. The number of iterations is set to be 2000, the batch size is 966, and the probability of dropout is 0.5 during the train and we turn the dropout layer off in test. The learning rate is set to be 0.0001 using Adam optimizer. The loss function uses the cross-entropy function.10-fold cross validation is employed on our model. We randomly divide the datasets into ten parts, took out nine subsets for training and the remaining one is for verification, and took turns to perform ten times, ensure that each subsample can be tested once. Finally we averaged all these test results to get the final result. The final results are shown in Table 2. We calculated accuracy (ACC), precision (PRE), Matthews correlation coefficient (MCC), sensitivity (SN), and specificity (SP). They are respectively defined as follows:

$$\mathrm{ACC} = (TP + TN)/(TP + TN + FP + FN) \tag{5}$$

$$\mathrm{PRE} = TP/(TP + FP) \tag{6}$$

$$\begin{aligned} &\mathrm{MCC} \\ &\quad = (TP{*}TN - FP{*}FN)/\sqrt{(TP + FN){*}(TP + FP){*}(TN + FP){*}(TN + FN)} \end{aligned} \tag{7}$$

$$\mathrm{SN} = TP/(TP + FN) \tag{8}$$

Table 2
The comparison of results from different models. We can see that the performance of econvRBP is the best in all models.

Feature group	ACC	PRE	SN	SP	AUC	MCC
econvRBP	**0.998**	**0.993**	**0.999**	**0.997**	**0.999**	**0.995**
2DLO	0.954	0.846	0.999	0.939	0.969	0.890
3DGO	0.991	0.972	0.998	0.988	0.985	0.979
econv-20	0.937	0.790	0.998	0.918	0.943	0.851
econvWR	0.969	0.900	0.994	0.959	0.992	0.925

The significance of bold values denote the best result.

$$SP = TN/(TN + FP) \tag{9}$$

We also used Conjoint Triad encoding to compare the performance of traditional machine learning algorithms such as SVM and RF, and made their ROC curve and calculated the Area Under Curve (AUC) (Fig. 5).

3. Results

3.1. Datasets

Although the prediction methods of RBPs are various, unfortunately, there are still no unified public datasets so far. Cai et al. proposed a scientific method to obtain protein data. We followed the same method as Cai et al. [4] and Ma et al. [21] to obtain RBPs (Fig. 4). We use the keyword “RNA-binding” to search Uniprot database and downloaded 60048 protein sequences [5], as a rough positive sample. Then the rough positive datasets is removed ambiguous proteins, such as sequences with the length more than 6000 or less than 50. Besides, sequences containing illegal amino acids were deleted (‘X’ and ‘Z’). Finally, 59660 positive samples are obtained. For comparison, the list of keyword “tDNA binding — DNA binding — core protein” with logical “or” is used to search from Uniport with a logical “or”, we obtain a total of 48,528 rough negative samples and perform the same preliminary screening. Then the positive and negative samples are mixed, and the CD-HIT program [19]is used to remove the homologous sequences with identity cutoff ⩾25%. In this way, a total of 9657 mixed samples are obtained, we call it Mix9657.

In order to correctly separate positive and negative samples, all of the reviewed proteins that “GO: Molecular Function” section contains “RNA-binding protein” are downloaded. These reviewed proteins functions are all annotated as RNA-binding, we call it ALLRBP. We only select positive samples that both belong to 59660 samples and exist in ALLRBP as positive samples, with the rest as negative samples. This verified separation yield 2875 positive samples and 6782 negative samples named BP2875 and nBP6782, respectively.

At the same time, we download protein seqeuence provided by RBPPred for verification and comparison. For the training set of RBPPred, we call it Comparison train (Cmptrain). Similarly we call its test set as Comparison test (Cmptest).

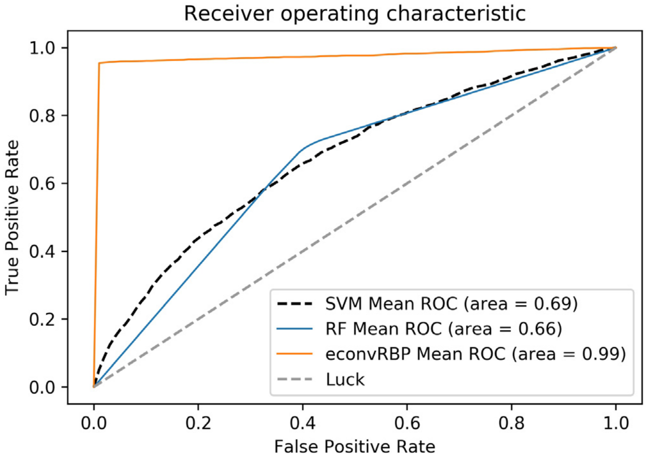

Fig. 5. ROC curves for SVM, RF and econvRBP. Where the parameter of RF is set to n_estimators = 460, min_samples_split = 2 and $(C, \gamma) = (110, 0.1)$ of SVM using grid search strategy.

3.2. Performance comparison of different network structures

Different models were designed to compare the influences of different structures and features on the original model. We separately designed one model called Two Dimensional Local Only (2DLO) which removed 3DGC from the origin model, and the other removed 2DLC, called Three Dimensional Global Only (3DGO). To verify the effect of residual block on the model, we also analyzed the model removing residual block, called econvWR. We compared the impact of using 20 kinds of amino acids on the model as well. It is worth noting that the use of 20 kinds of amino acids will expand the number of channels to 20 for One Hot matrix, and expand the Conjoint Triad cubic matrix to $20 \times 20 \times 20 = 8000$. If the original design were followed, neurons in fully connected will be excessively preferred to its global features. Therefore, we added max pooling with $2 \times 2 \times 2$ filter and $2 \times 2 \times 2$ step in the first two layers of convolution in order to reduce dimension of features, and balance global-local features, we call that econv-20. The results of different structures are shown in Table 2. From the results Table 2, we made the following two assumptions: (1) it is necessary to establish a mapping from input to output. Although deep CNNs can extract high dimensional and abstract features, these complex features make it more difficult for the next layer to learn. The residual block combines the features of the deep CNNs output with low dimensional, simple features that are easier to understand for the next layer. The experimental process also validated it: during the training process, the model with the residual block will converge faster and predict better. (2) It is a good choice to classify 20 kinds of amino acids into seven categories. The dipoles and volumes can reflect the properties of RBPs, so they can be an essential feature of RBPs prediction. In this way, the noise is reduced and the overfitting is prevented. It is obvious from the

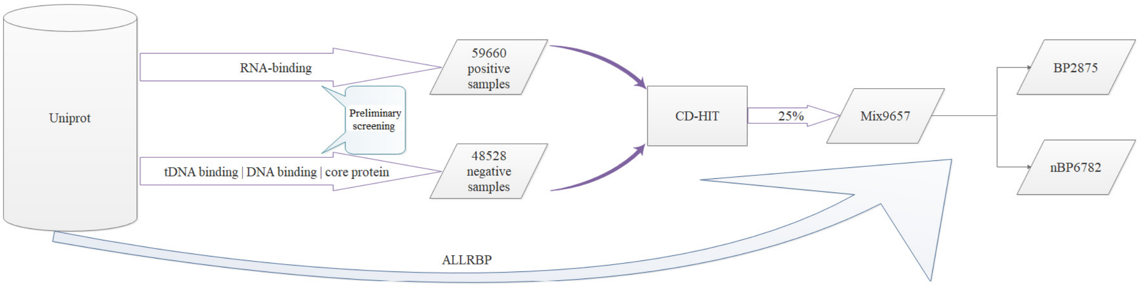

Fig. 4. The flowchart of sample acquisition. The keywords “RNA binding” and “tDNA binding — DNA binding — core protein” are used to retrieve positive and negative samples, respectively. Then the mixed data was used together with CD-HIT program to remove homologous sequences at a threshold of 25%. Finally, the intersection of ALLRBP and Mix9657 is taken as the final positive sample BP2875 and the rest as the final negative sample nBP6782.

Table 3
Performance comparison with traditional learning algorithm (SVM and RF) using Conjoint Triad encoding. One Hot encoding is difficult to use in machine learning and is not listed here.

Learning algorithm	ACC	PRE	SN	SP	AUC	MCC
SVM	**0.713**	0.632	0.104	**0.971**	**0.686**	0.155
RF	0.654	**0.673**	**0.598**	0.710	0.662	**0.310**

The significance of bold values denote the best result.

comparison among econvRBP, 2DLO and 3DGO that simply using the local or global sequence purely cannot make the right judgment. However, after combining with the global or local features, the classifier improves powerful performance.

3.3. Comparison with conventional machine learning algorithm

Traditional machine learning algorithms have been frequently used in the field of bioinformatics, such as SVM, RF, etc. We used the grid search strategy to find the optimal super parameters of SVM and RF. Researchers have proposed a large number of complex methods in feature engineering and have achieved relatively good results. However, because of its own characteristics, CNN does not require a lot of time and efforts for feature engineering. We can make good prediction by designing model structures and adjusting parameters. Therefore, this study compares the performance of the deep learning method with the machine learning method (Table 3) using the same encoding. It can be seen that CNN is far better than the machine learning using the same Conjoint Triad encoding as the input feature, and shows that the CNN is a powerful tool.

3.4. Comparison with other existing methods

3.4.1. Compared with existing machine learning methods

Mix9657 is an obviously unbalanced dataset. There have 6782 negative samples in Mix9657, accounting for about 70%. Therefore, Mix9657 causes the classifier to have additional preferences for negative samples. From the result of cross validation, the classifier misjudges some positive samples as negative samples, which is caused by the imbalance. Similarly, it also will cause great trouble to the other models. Although there are many prediction methods of RBPs, few RBPs prediction web server are available online. Kumar et al. provided a web server called RNApred for RBPs prediction [15]. RNApred provides three alternative methods: Amino acid composition, PSSM, and Hybrid, the latter two cannot produce results in tolerable time. Here, we just picked the Amino acid composition method. And the SVM threshold is constantly adjusted for the best performance. From the results of RNApred (Table 4), we can see that a large number of errors are mainly concentrated in negative samples.

3.4.2. Compared with existing deep learning methods

In order to further evaluate the validity of the model, we reproduced the same CNN-based method Deep-RBPPred proposed by Zheng et al. and applied this model on MIX9657 (Table 5). Unfortunately, due to the different dataset, gradient explosion problem occurred in the training process, so we appropriately reduced the learning rate and increased the training steps compared with the original model. Deep-RBPPred was trained 18,000 steps with 0.000001 learning rate and got the final result. We believed that the stacking of features as machine learning is no longer suitable for deep neural networks represented by CNN. The convolution kernel regularly scanned on the feature map. When the kernel scanned to the junction of different features, the convolution results will mislead the next layer that the adjacent features have strong correlation (actually they are just being stacked together).

Table 4
Comparison of confusion matrix between RNApred and econvRBP on Mix9657. For RNApred, we adjusted the SVM threshold to -0.2, so that it has a preference for negative samples to deal with the unbalanced dataset. It can be seen from the results that the imbalance of the sample will lead to a significant increase in False Positive. Both RNApred and econvRBP show different degrees of misjudgment on positive samples.

Method: RNApred		Prediction	
		Positive	Negative
Real	True	2410	465
	False	1770	5012
Method: econvRBP		Prediction	
		Positive	Negative
Real	True	2859	16
	False	1	6781

Table 5
Comparison of confusion matrix between Deep-RBPPred and econvRBP on Mix9657. Deep-RBPPred and econvRBP are both CNN-based models. The feature encoding of Deep-RBPPred is more complicated than econvRBP, considering more physical and chemical properties of proteins. However, econvRBP considers the essential properties of protein sequence to construct model according to objective laws, thus achieving better results.

Method: Deep-RBPPred		Prediction	
		Positive	Negative
Real	True	2269	606
	False	127	6655
Method: econvRBP		Prediction	
		Positive	Negative
Real	True	2859	16
	False	1	6781

3.5. Performance comparison among other datasets

We want to further test whether econvRBP could accurately identify unlearned RBPs in a large number of negative samples. The data provided by RBPPred is downloaded for verification. For Cmptrain, BP2875 are mixed and the homologous sequence are removed with a threshold of 25% and fed to econvRBP for verification. For the independent testing set, we fed it to econvRBP directly with nothing changes. One thing we have to mention is that econvRBP uses fixed-length One Hot encoding. Therefore, the shape of input conflicts with the data provided by RBPPred. For this case, we find that the average length of the Cmptrain is 283, and the average length of the Cmptest is 347, far below the average length of 549 we trained, we had to pad zero at the end of Cmptrain and Cmptest, causing data to mix in a lot of noise. Numerous noises will make wrong distinguish. Even so, econvRBP accurately predict 3661 of the 4,250 amino acid sequences in Cmptrain, with an accuracy rate of 86%. In Cmptest, 5322 of the 6105 sequences were accurately predicted, and the accuracy rate reached 87%. This proves that econvRBP has strong robustness and is able to deal with a large amount of noise.

3.6. Web Server for econvRBP

With the development of high-throughput computation methods, more and more web servers have been applied in bioinformatics. More practical and friendly servers are the order of the day. Therefore, we also provide an econvRBP-based web server here. The specific steps are

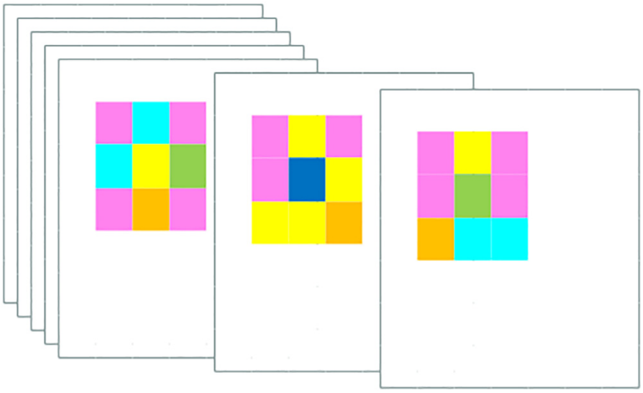

Fig. 6. Convolution kernel training results. The darker the color, the greater the weight. We found that 3DGC automatically learns that the position in the central of the convolution kernel is most relevant to other locations, so its weight is the largest. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

as follows:

1. Open the urlhttp://47.100.203.218:3389/home.html/.
2. Click the 'Submission' button on the left to submit your jobs.
3. Protein sequences in the 'fasta' format will be prepared and submitted.
4. Click the 'submit' button and wait for the final result. The output consists of the sequence ID and the result.

4. Discussion

As mentioned above, we divided 20 kinds of amino acids into seven categories according to dipoles and volumes. Using conjoint triad encoding, an amino acid sequence is encoded as a 7 × 7 × 7 cube. The index of the cube represents the frequency of occurrence in three groups. The convolution kernel of 3 × 3 × 3 is scanned regularly on the feature map of 7 × 7 × 7. Another reason we use a 3 × 3 × 3 size convolution kernel is that all features in the receptive field are correlated. In the receptive field, the convolution kernel and the feature map are multiplied and summed. The weight of the convolution kernel represents the preference for the feature map in the receptive field. We extract a convolution kernel from the trained 3DGC, as shown in Fig. 6, this is a process of convolution kernel scanning on feature graph. Color represents the area where the convolution kernel is located, and white represents other unrelated areas in the feature map. The darker the color, the greater the weight in the convolution kernel. In this case, the index of convolution kernel in feature map is [2:4][2:4][1:3] (the index starts at 1). The index [3][3][2] does not only mean that the frequency of conjoint triad, and is also the center of the kernel. Other locations in the receptive field are related to the center element. For example, the index [3][2][2] means the transition from group 3 to group 2 of middle amino acid in the conjoint triad. Obviously, the most important position in the convolution kernel is the middle, which has the strongest correlation with other positions. Therefore, the color in the middle position in Fig. 5 is the darkest and the weight is the largest. This inspires us that convolutional neural network can be well applied to bioinformatics to analyze the proximity information of sequences. By acquiring protein sequence encoding, we can use CNN to process RNA binding protein just like image. The filter of convolutional neural network can scan a motif with biological significance, i.e. feature subsequences, which are very important to RNA binding protein. From the whole experiment, we can see that our method is very effective.

5. Conclusion

We compared the effects of different network structures on RBPs prediction. The local features are separated from the global features are not satisfactory due to the loss of some global information and fall into local optimum during prediction; the global features that are separated from the local features lose important local information. Some of the details in the experiment are thought-provoking: In the early stages of training, econvRBP, 2DLO and econvWR both simply predicts that all samples are negative samples, and after a few steps, econvRBP first jumps out of the local optimum, and continues to converge to the global optimum. The combination of global information and local information can promote each other. The global information and residual block provide a relatively gentle gradient for the amino acid sequence, so that the model can jump out of the local minimum value in the training process. The classification of 20 amino acids into 7 categories is also the key point of our model. First, the classification into 7 categories can reduce the data dimension, effectively reduce the data quantity and increase batch size on training. Secondly, the expression of useless features can be reduced. All the sequence information of 20 amino acids is often not necessary. Although econvRBP can be used to effectively predict the RNA binding protein, further features, including those related to alternative splicing, have not yet been explored. In addition, we will further explore the correlation between the high-level features learned by CNN and alternative splicing in the future work, which will further link our work with practical biological problems.

References

[1] Castello Alfredo, Horos Rastislav, Strein Claudia, Fischer Bernd, Eichelbaum Katrin, Lars M Steinmetz, Krijgsveld Jeroen, Matthias W Hentze, System-wide identification of rna-binding proteins by interactome capture, Nat. Protoc. 8 (2013) 491–500.

[2] B. Alipanahi, A. Delong, M.T. Weirauch, B.J. Frey, Predicting the sequence specificities of dna-and rna-binding proteins by deep learning, Nature Biotechnol. 33 (2015) 831.

[3] R. Busà, M.P. Paronetto, D. Farini, E. Pierantozzi, F. Botti, D.F. Angelini, F. Attisani, G. Vespasiani, C. Sette, The rna-binding protein sam68 contributes to proliferation and survival of human prostate cancer cells, Oncogene 26 (2007) 4372–4382.

[4] Y.D. Cai, S.L. Lin, Support vector machines for predicting rrna-, rna-, and dna-binding proteins from amino acid sequence, Biochimica et Biophysica Acta (BBA)-Proteins Proteomics 1648 (2003) 127–133.

[5] Consortium, U., Uniprot: a hub for protein information, Nucleic Acids Res. 43 (2014) D204–D212.

[6] X. Du, Y. Diao, Y. Yao, H. Zhu, Y. Yan, Y. Zhang, Deepmvf-rbp: Deep multi-view fusion representation learning for rna-binding proteins prediction, 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 65–68.

[7] I. Dubchak, I. Muchnik, S.R. Holbrook, S.H. Kim, Prediction of protein folding class using global description of amino acid sequence, Proc. Nat. Acad. Sci. 92 (1995) 8700–8704.

[8] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[9] Z. Huiying, Y. Yuedong, Z. Yaoqi, Structure-based prediction of rna-binding domains and rna-binding sites and application to structural genomics targets, Nucleic Acids Res. 39 (2011) 3017–3025.

[10] M. Ibba, D. Söll, Protein-rna molecular recognition, Nature 381 (1996) 656–656.

[11] Ma. Junshui, Robert P Sheridan, Liaw Andy, George E Dahl, Svetnik Vladimir, Deep neural nets as a method for quantitative structure-activity relationships, J. Chem. Inf. Model. 55 (2015) 263–274.

[12] K. Kang, W. Ouyang, H. Li, X. Wang, Object detection from video tubelets with convolutional neural networks, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 817–825.

[13] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Information Processing Syst. (2012) 1097–1105.

[14] M. Kumar, Gromiha, G. Mmraghava, Prediction of rna binding sites in a protein using svm and pssm profile, Proteins-structure Function Bioinformatics 71 (2010) 189–194.

[15] M. Kumar, M.M. Gromiha, G.P. Raghava, Svm based prediction of rna-binding proteins using binding residues and evolutionary information, J. Mol. Recognit. 24 (2015) 303–313.

[16] S.C. Kwon, H. Yi, K. Eichelbaum, S. Föhr, B. Fischer, K.T. You, A. Castello, J. Krijgsveld, M.W. Hentze, V.N. Kim, The rna-binding protein repertoire of embryonic stem cells, Nature Struct. Mol. Biol. 20 (2013) 1122.

[17] Y. LeCun, Y. Bengio, G. Hinton, Deep Learning Nature 521 (2015) 436.

[18] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al., Gradient-based learning applied to

document recognition, Proc. IEEE 86 (1998) 2278–2324.

[19] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.

[20] K.E. Lukong, K.W. Chang, E.W. Khandjian, S. Richard, Rna-binding proteins in human genetic disease, Trends Genet. 24 (2008) 416–425.

[21] X. Ma, J. Guo, K. Xiao, X. Sun, Prbp: Prediction of rna-binding proteins using a random forest algorithm combined with an rna-binding residue predictor, IEEE/ACM Trans. Comput. Biol. Bioinf. 12 (2015) 1385–1393.

[22] Polymenidou Magdalini, Lagier Tourenne Clotilde, Kasey R Hutt, Stephanie C Huelga, Moran Jacqueline, Tiffany Y Liang, Ling Shuo-Chien, Sun Eveline, Wancewicz Edward, Mazur Curt, Long pre-mrna depletion and rna missplicing contribute to neuronal vulnerability from loss of tdp-43, Nat. Neurosci. 14 (2011) 459–468.

[23] D. Marchese, N.S. de Groot, N. Lorenzo Gotor, C.M. Livi, G.G. Tartaglia, Advances in the characterization of rna-binding proteins, Wiley Interdisciplinary Rev.: RNA 7 (2016) 793–810.

[24] I. Paz, E. Kligun, B. Bengad, Y. Mandel-Gutfreund, Bindup: a web server for non-homology-based prediction of dna and rna binding proteins, Nucleic Acids Res. 44 (2016) W568–W574.

[25] J.I. Perez-Perri, B. Rogell, T. Schwarzl, F. Stein, Y. Zhou, M. Rettel, A. Brosig, M.W. Hentze, Discovery of rna-binding proteins and characterization of their dynamic responses by enhanced rna interactome capture, Nature Commun. 9 (2018).

[26] S. Shazman, Y. Mandel-Gutfreund, Classifying rna-binding proteins based on electrostatic properties, Plos Comput. Biol. 4 (2008) e1000146 .

[27] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, Predicting protein-protein interactions based only on sequences information, Proc. Nat. Acad. Sci. 104 (2007) 4337–4341.

[28] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, Adv. Neural Inform. Processing Syst. (2014) 3104–3112.

[29] R.R. Walia, L.C. Xue, K. Wilkins, Y. El-Manzalawy, D. Dobbs, V. Honavar, 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 9 (2014) e97725 .

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

[31] B. Yang, F. Liu, C. Ren, Z. Ouyang, Z. Xie, X. Bo, W. Shu, Biren: predicting enhancers with a deep-learning-based model using the dna sequence alone, Bioinformatics 33 (2017) 1930–1936.

[32] X. Zhang, S. Liu, Rbppred: predicting rna-binding proteins from sequence using svm, Bioinformatics 33 (2016) 854–862.

[33] H. Zhao, Y. Yang, Y. Zhou, Highly accurate and high-resolution function prediction of rna binding proteins by fold recognition and binding affinity prediction, Rna Biology 8 (2011) 988–996.

[34] J. Zheng, X. Zhang, X. Zhao, X. Tong, X. Hong, J. Xie, S. Liu, Deep-rbppred: Predicting rna binding proteins in the proteome scale based on deep learning, Sci. Rep. 8 (2018) 15264.