RNA Secondary Structure Representation Network for RNA-proteins Binding Prediction

Ziyi Liu,[1] **Fulin Luo,**[2*] **Bo Du**[1†]

[1] National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science, and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, China
[2] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, China
ziyiliu@whu.edu.cn, luoflyn@163.com, dubo@whu.edu.cn

Abstract

RNA-binding proteins (RBPs) play a significant part in several biological processes in the living cell, such as gene regulation and mRNA localization. Several deep learning methods, especially the model based on convolutional neural network (CNN), have been used to predict the binding sites. However, previous methods fail to represent RNA secondary structure features. The traditional deep learning methods generally transform the RNA secondary structure to a regular matrix that cannot reveal the topological structure information of RNA. To effectively extract the structure features of RNA, we propose an RNA secondary structure representation network (RNASSR-Net) based on graph convolutional neural network (GCN) and convolution neural network (CNN) for RBP binding prediction. RNASSR-Net constructs the graph model derived from the RNA secondary structure to learn the topological properties of RNA. Then, it obtains the spatial importance of each base in RNA with CNN to guide the representation of the RNA secondary structure. Finally, RNASSR-Net combines the structure and sequence features to predict the binding sites. Experimental results demonstrate the proposed method outperforms a few state-of-the-art methods on the benchmark datasets and gets a higher improvement on the small-size data. Besides, the proposed RNASSR-Net is also used to detect the accurate motifs compared with the experimentally verified motifs, which reveals the binding region location and RNA structure interpretation for some biological guidance in the future.

Introduction

RNAs and proteins are important components in life, which are involved in biochemical reaction in living cells(Ule and Rinn 2014). RNA-binding proteins are the proteins that binds to the RNAs to regulate the gene expression and control RNA processes and translation within the cells(Van Nostrand et al. 2016). The gene regulation comprises a huge number of co- and post-transcriptional gene expression in living organisms, including polyadenylation, RNA splicing, modification, capping, localization, translation and turnover(Ray et al. 2013). The dysfunctions of certain RNA-binding proteins (RBPs) may cause some serious diseases, such as neurodegenerative disorders, cancers and cardiovascular diseases(Musunuru 2003; Lukong et al. 2008).

RBPs also have a great influence on the viral RNA transcription and replication. The outbreak of coronavirus disease 2019 (COVID-19) is a disease caused by a novel coronavirus. The coronavirus(CoV) N proteins inside the coronavirus construct helical ribonucleoproteins during the packaging of the viral RNA genome, regulating viral RNA synthesis in replication/transcription, and modulating infected cell metabolism(Nelson, Stohlman, and Tahara 2000; Stohlman et al. 1988; Cong et al. 2020). Therefore, the study of RBPs benefits to the understanding of gene regulation and the treatment of some genetic diseases and infectious diseases.

To investigate how the RBPs affect the RNA processing, the RNA substrates that each RBP interacts with have been analyzed(Ule et al. 2003). The binding sites of these RNA substrates are highly related to the function of RBPs. Both RNA sequences and RNA structures determine the RBPs binding intensities (Buckanovich and Darnell 1997; Hackermüller et al. 2005). The RNA sequences reflect the base distribution information and the RNA structures reflect the topological information. In previous works, the RNA sequences are widely used to get the RNA binding prediction(Zhang et al. 2016; Kazan et al. 2010a). However, the missing of RNA structures won't guarantee the model to get the optimal prediction. For example, hairpin loop containing "UGGC" has been shown to bind Vts1p-SAM with high affinity(Aviv et al. 2006). Thus, the topological information extracted by RNA structures also contributes to the RNA binding prediction.

To analyze the binding of RNA, biologist uses the RBPs to obtain the binding information of RNA by biological experiments. However, the experimental way is very time- and material- consuming. To reduce the cost, high-throughput technologies have been widely used in genome-wise study of RNA-protein interactions, such as the cross-linking immunoprecipitation coupled with high- throughput sequencing (CLIP-seq)(Anders et al. 2011; Ferre, Colantoni, and Helmer-Citterich 2015) and RNAcompete(Ray et al. 2013). These high-throughput technologies provide a considerable amount of available data, which make it possible for traditional machine learning and deep learning methods to train an accurate model for binding prediction.

[*] Corresponding author.

[†] Corresponding author.

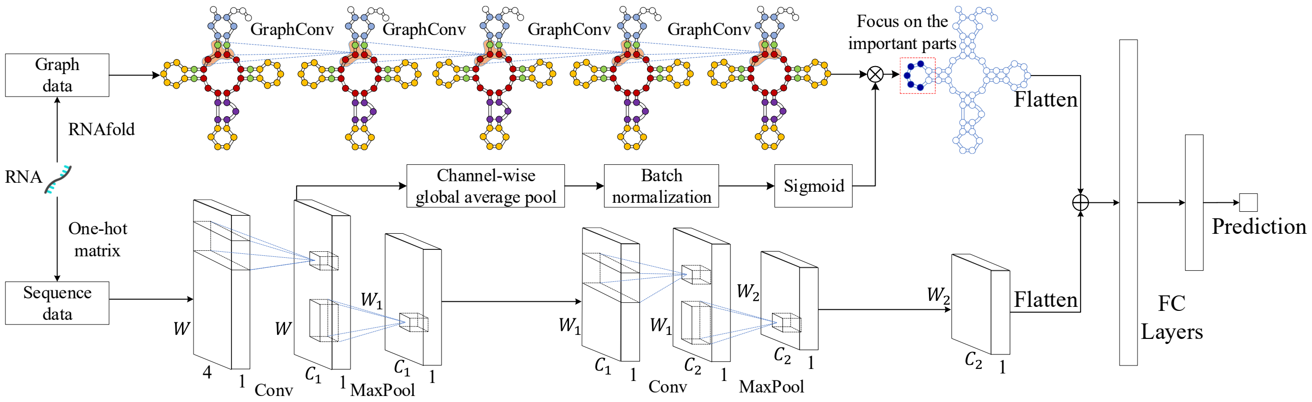


Figure 1: Network architecture of RNASSR-Net. The RNA sequences and structures are processed by GCN and CNN respectively. Different colored nodes in the RNA graph represent nucleotides of different structural categories. For the nucleotide base importance, GCN focuses on the some sub-structures that are critical to the binding classification.

As we known, the RNA data are considered as sequences, which neglects the topological structures of RNA. The structure data, which benefits for the research of RBPs binding can be directly obtained by some computational software (Gruber et al. 2008). For RNA secondary structure, it is represented as simple strings with dot and bracket. These dot-bracket strings make it very hard for models, such as convolutional neural network(CNN) and long short-term memories(LSTMs), to learn the binding pattern. Some approaches transform the dot-bracket strings to the one-hot matrix according to nucleotide categories of RNA secondary structure. However, the transformation will cause irreversible information loss of RNA structure. Therefore, it is very urgent to develop a method for the analysis of RNA secondary structure.

Therefore, we propose a RNA secondary structure representation network (RNASSR-Net) based on graph convolutional neural network (GCN) and convolution neural network (CNN) for RBPs binding prediction. RNASSR-Net simultaneously uses the sequences and the structure data to learn the topological and binding properties of RNA, as shown in Figure 1. Our source code is available at https://github.com/ziniBRC/RNASSR-Net. The main contributions of this paper include the following:

- The proposed RNASSR-Net model uses GCN to represent the secondary structures of RNA and CNN to learn the sequence features of RNA. This method simultaneously considers the topological and sequence properties of RNA to achieve the complementarity of different information and improve the prediction accuracy of RNA binding.
- We use CNN to learn the base weights which reveal the spatial importance of RNA. The weights can guide GCN to learn the secondary structures of RNA. The base weights corresponds to the node in the graph of RNA structure. Under the similar receptive field, we share the base weights to make GCN focus on some important nucleotide bases.
- We fuse the structure and sequence features to predict the RBP binding. The proposed method outperforms other baselines on the benchmark dataset. We detect accurate motifs that are consistent with the experimentally verified motifs. Besides, we analyze the response value of each nucleotide base to extract the importance region and structure mutation information.

Related Work

A lot of traditional machine learning methods have been proposed to address the RNA binding prediction problems. Traditional methods focus on how to extract the important features manually. These machine learning algorithms and computational tools analyze different aspects of features to predict RBP binding sites and generate the motifs(Stražar et al. 2016). BioBayesNet is the first tool that uses the structural features to perform the target recognition problem for the transcription factor binding sites(Pudimat, Schukat-Talamazzini, and Backofen 2005). RNAContext is a motif binding method to achieve the relative binding preferences of RBPs for RNA sequences and structures(Kazan et al. 2010b). RCK developed a new model, which is from the extension of RNAcontext, to obtain the k-mer sequence and structure preferences of RNA-binding proteins(Orenstein, Wang, and Berger 2016). GraphProt extracts a very large number of features from sequence and structure information by graph encoding and uses the Support Vector Machine (SVM) family to predict the RNA binding sites(Maticzka et al. 2014).

For the complexity of RNA sequences, it is hard to extract important latent features through traditional machine learning methods. To address this problem, deep learning has been introduced into the fields of bioinformatics and computational biology, such as protein location(Almagro Armenteros et al. 2017), protein structure prediction(Heffernan et al. 2015), RNA prediction(Alipanahi et al. 2015) and chemoinformatics(Lusci, Pollastri, and Baldi 2013). Recently, several deep learning methods have been developed to analyze RNA binding prediction and detect candidate motifs automatically. DeepBind(Alipanahi et al. 2015) firstly introduced the CNN model to predict RNA binding sites and structures. In DeepBind, Position weight matrices(PWM), which are composed of one layer of convolution and pooling followed by a fully connected layer, are embedded into the CNN model to show the probability distribution of binding sites. Deepnet-rbp(Zhang et al. 2016) uses deep belief network (DBN) to predict RBP binding sites from sequence, secondary and tertiary structural features. iDeepE(Pan and Shen 2018) integrates the multi-channel local and global sequence information for predicting RBP binding sites and motifs. Another similar work iDeepS trains two individual CNNs and a LSTM to get the binding prediction. Although these methods are effective and accurate, they mainly focus on the RNA sequence features using CNN model. The features of RNA structures are converted to position weight matrix according to the categories of RNA structures, which are processed by CNN models. The topological information of RNA structure isn't fully used because the link information of RNA bases is missing after the conversion.

In recent years, graph neural network(GNN) has been developed to learn the topological information of data, which is widely used in the fields of social sciences(Kipf and Welling 2017; Monti et al. 2019), knowledge graphs(Schlichtkrull et al. 2018; Chami et al. 2020), chemistry(Duvenaud et al. 2015; Gilmer et al. 2017). In order to process irregular data format, recursive neural network is used to build the graph neural networks(Gori, Monfardini, and Scarselli 2005; Scarselli et al. 2008), which aims to deal with a general class of graphs. In the case of fixed-size graphs, a series of convolutional neural networks based on the spectral representation of the graphs have been applied on the node classification and graph classification. Specifically, Kipf & Welling(Kipf and Welling 2017) proposed a simplified spectral neural network using 1-hop filters to address overfitting problem and minimize the number of operations. As we known, few graph neural networks have been applied in the analysis of RNA structures. Recently, RNA-protein interactions network (RPI-Net) based GNN was used to learn and exploit a graph representation of RNA molecules(Yan, Hamilton, and Blanchette 2020).

Methodology

Sequence Coding

RNA is composed of four kinds of nucleotides, which can be distinguished by different bases of nucleotides. The bases that are denoted as 'A','C','G','U' respectively represent different kinds of nucleotides in our model. The lengths of RNA sequences in the datasets are different. However, the input features of convolutional neural networks must have the same fixed size. We pad RNA sequences to a fixed window size using an identifier 'N'. Given an RNA sequence s as $\{s_1, s_2, s_3, s_n\}$ with n nucleotides ($s_i \in$ 'A','C','G','U','N'), we encode the RNA sequence data as one-hot matrix:

$$M_{i,:} = \begin{cases} [0.25 \quad 0.25 \quad 0.25 \quad 0.25] & \text{if } s_{i-m+1} \text{ is `N'} \\ [1.00 \quad 0.00 \quad 0.00 \quad 0.00] & \text{if } s_{i-m+1} \text{ is `A'} \\ [0.00 \quad 1.00 \quad 0.00 \quad 0.00] & \text{if } s_{i-m+1} \text{ is `C'} \\ [0.00 \quad 0.00 \quad 1.00 \quad 0.00] & \text{if } s_{i-m+1} \text{ is `G'} \\ [0.00 \quad 0.00 \quad 0.00 \quad 1.00] & \text{if } s_{i-m+1} \text{ is `U'} \end{cases} \tag{1}$$

where *i* is the location of nucleotides in the RNA sequences. In our study, we suppose that 4 different nucleotides obey an average distribution at the start and end of the sequences. Therefore, we use [0.25, 0.25, 0.25, 0.25] for the padded nucleotides and 'N' in the one-hot matrix. 'N' denotes the unknown base in RNA sequences.

Graph Construction

The one-hot matrix reflects the composition of nucleotide sequences. However, the structures of RNA take an important part in RNA function. One-hot matrix of RNA sequences can't reflect the nucleotide structure information. To better extract the intrinsic properties of RNA, the RNA structures should be considered in the model. In this paper, we use RNAfold to abstract certain structural details. RNAfold can sample all possible structures and retain highly probable candidates. In several previous methods, the secondary structures are used in the model. The nucleotides in the secondary structure can be classified to six categories, which are denoted as stems (S), multiloops (M), hairpins (H), internal loops (I), bulges (B) and external regions (E). Same as the RNA sequences, the nucleotides of these six categories can be transferred to one-hot matrix in the previous methods. Nevertheless, the one-hot matrix neglects the structure relationships of RNA, which may lose some structure information. Therefore, we use the graph model to represent the structural information of RNA.

According to the RNAfold abstracted results, some structures may have the link between two nucleotides *i* and *j* , but others may not. In this approach, we denote the probability of a secondary structure *s* associated with sequence *x* as $p(s|x)$. The construction of graph is also used in (Yan, Hamilton, and Blanchette 2020). The probability $p(s|x)$ is defined as:

$$p(s|x) = \frac{1}{z} e^{-\beta E(s,x)} \tag{2}$$

where *Z* is a normalization constant and *E(s, x)* is the free energy of *x* under structure *s*. The base-pairing probability for nucleotides *i* and *j* is defined as:

$$p([i,j]|x) = \sum_{[i,j] \in s} p(s \mid x) \tag{3}$$

After running RNAplfold, we obtain a probabilistic adjacency matrix $A_{n \times n}$, where $A_{i,j} = p([i,j]|x)$. For graphs constructed by different RNA sequences, the node numbers of graphs are also different. We pad the graphs using the empty nodes, which aren't linked to other nodes. The node numbers of the padded graphs are the same as the length of the padded RNA sequences.

Network Architecture

For RNA data, the structures and sequences can be denoted as graph data and one-hot matrix. Because the previous CNN-based methods aren't used for the graph data, we propose a novel network based on GNN and CNN to learn the structure and sequence features of RNA. This network integrates the node information in graph and the spatial information in one-hot matrix for RNA binding prediction. In the network, the graph data and the one-hot matrix are inputted into GNN and CNN, respectively. GNN and CNN both have two convolution layers and two pool layers. For graph data and one-hot matrix, one nucleotide corresponds to one node in the graph and a location in the one-hot matrix. Thus, the nodes in the RNA graph and the nucleotide bases in the RNA sequence should have shared some similar information. Under the similar receptive field, the spatial importance extracted from CNN is passed to GCN to guide the training process of GCN. Finally, the features extracted by GCN and CNN are flattened, which are concatenated with each other. Then, the features are inputted into fully connected layers to get the RNA binding prediction.

Convolutional Neural Network

The convolutional neural network(CNN) can help extract the non-linear intrinsic features of inputs, including convolution, max-pool, and fully connected layers. Any convolutional neural network contains at least one convolution operation. And the pointwise product between input one-hot matrix and filters are outputted after convolution operation. The output of convolution operation $x_{i,k}$ is the score of filter k aligned to the nucleotide i in the padded sequence s. In this study, filters are stored in the matrix F, where the element $F^{(l)}_{k,j,c}$ is the coefficient of k^{th} filter at the position j and channel c in the l^{th} layer. The output of convolution operation $x^{(l)}_{i,k}$ can be defined as:

$$x^{(l)}_{i,k} = \Sigma_{j=1}^{m}\Sigma_{c=1}^{C^{(l-1)}} x^{(l)}_{i-\frac{m}{2}+j,c}F^{(l)}_{k,j,c} \tag{4}$$

where $x^{(l)}$ is the output of the l^{th} layer, i is the index of features, $F^{(l)}_{k,:}$ is the coefficients of the filter k in the l^{th} layer, m is the size of the filters, $C^{(l)}$ is the channel number of the input in the l^{th} layer. In the convolution operation of the first layer, the input of convolution operation is the one-hot matrix M and the channel of M is the base of nucleotide. So the first convolution filters are called as motifs detectors, which can learn the pattern of base distribution.

The activation function of convolution operation is rectified linear unit (ReLU). The positive scores that are greater than 0 are passed to next layer, while the negative scores are assigned to 0. The ReLU activation operation is defined as:

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{otherwise} \end{cases} \tag{5}$$

In CNN, convolution operation is followed by max-pool or average-pool layers. After the convolution layers, the size of the output feature map remains almost unchanged. Pool layers can reduce the dimensionality and preserve some important features by selecting the maximum or average values inside a certain window. Max-pool layers can be denoted as:

$$\text{Maxpool}_F(X): x^{(l)}_{i,k} = \max\left(x^{(l-1)}_{i-\frac{p}{2}+r,k}\right) \quad r \in [0, p-1] \tag{6}$$

where X is the output after convolution and ReLU layer, p is the length of pool kernel, $x^{(l)}_{i,k}$ is the output of max-pooling $x^{(l-1)}_{i,k}$, which will be inputted into next convolutional layer or fully connected layer.

Graph Neural Network

Graph neural network(GNN) is an end-to-end network that learns hierarchical graph representation and graph node embedding. In this paper, we use GNN to learn the adjacent node information and node embedding features to analyze the RNA structure. For previous methods, the one-hot matrix of RNA structures is constructed and inputted into convolutional neural networks to get prediction. However, the process of converting RNA structures into one-hot matrix will lose some important information. Besides, one-hot matrixes can only analyze the node features and the adjacent node relationship is often missed. While GNN can directly process the structure data. The adjacent matrix is inputted into the GNN model to extract the adjacent node information.

Given a graph $G = (V, E, A)$, where V is a finite set of $|V| = n$ nodes, E is a set of edges and $A \in \mathbb{R}^{n\times n}$ is an adjacency matrix encoding the connection weight between two nodes. For comprehension, we consider the graph convolution following general "message-passing" architecture:

$$H^{(l)} = Gh\left(A, H^{(l-1)}; \theta^{(l)}\right) \tag{7}$$

where $H^{(l)} \in R^{n\times d}$ are the node embeddings after l steps of graph convolution operation, $Gh(\cdot)$ is the graph convolution operation which is known as the message propagation function, $H^{(l-1)}$ is the output of last convolution operation, $\theta^{(l)}$ is the trainable parameters. The initial $H^{(0)}$ is the node features of RNA secondary structures.

Many implementations of message propagation function have been proposed to get the output of the graph convolution operation. A popular methods is the graph convolution network, which is implemented by linear transformations and ReLU non-linearities:

$$H^{(l)} = \text{ReLU}\left(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X^{(l-1)}W^{(l)}\right) \tag{8}$$

where $\tilde{A} = A + I_N$, $\widetilde{D}_{ij} = \Sigma_j \tilde{A}_{ij}$ and $W^{(l)}$ is a trainable matrix. $\widetilde{D}$ is the degree matrix of $\tilde{A}$. $\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}$ is a renormalization trick which is introduced to alleviate the numerical instability and exploding/vanishing gradient problem.

Although GCN has been applied into many research fields, GCN doesn't results in very good prediction performance. In this paper, we introduce a new method to help GCN perform better by the features of CNN. In graph learning representation of RNA structure, each base in the RNA sequence corresponds to a node in graph. Under the similar receptive field, the base importance weight learned by CNN can guide the training process of GCN . For one graph convolutional layer, the node integrates the information of the neighbor nodes, which could be understood as the 3*1 filters in CNN. For regular data, 4 graph convolutional layers have the similar receptive field compared to 9*1 filters in CNN. Thus, we extract the base importance weights from the first CNN layer to guide the training process of GCN. The base importance weights are defined as:

$$\text{weight } = \operatorname{sigmoid}\left(BN\left(\frac{1}{\mathrm{C}}\sum_{k=1}^{C} x_{i,k}^{(1)}\right)\right) \tag{9}$$

where $x_{i,k}^{(1)}$ is the output of the first CNN layer, BN is the operation of batch normalization. This importance weight is multiplied by the graph features obtained by GCN, which is calculated as follows:

$$H^{(4)} = H^{(4)} \otimes \text{ weight} \tag{10}$$

where $H^{(4)}$ is the ouptuts of the 4^{th} GCN layer, $\otimes$ is the Hadamard product.

Loss Function

In our work, the outputs of our model are used to predict the binding labels of the RNA data. To obtain the optimal parameters, we use the binary cross-entropy loss function:

$$L(\theta) = -\sum_{i=1}^{N} y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right)\log\left(1 - \hat{y}_i\right) \tag{11}$$

where y_i is the true label, $\hat{y}_i$ is the predicted labels of the trained model.

Motifs Detection

In our study, we will adopt MEME (Bailey et al. 2009) to accomplish the visualization of our results. In this study, we investigate the convolutional outputs of CNN and GCN in our model. The convolutional filters can be considered as the “motifs detector”. If the filters have the same base distribution of motifs, the outputs of these filters reflect the binding intensities. We convert the outputs of the filters into position weight matrices(PWM) using the same strategy in DeepBind and Basset (Alipanahi et al. 2015). We calculate the average value of the outputs of the first convolution in CNN and the outputs of the last graph convolution in GCN. If the average output value at position *i* is larger than 0.75 maximum value across this set of sequence, k-mer sequence $\left\{s_{i-\frac{k_c}{2}}, s_{i-\frac{k_e}{2}+1}, \cdots, s_{i+\frac{k_e}{2}}\right\}$ will be selected to calculate the probability of each base at corresponding location of motifs.

Experiments

Baseline

There are a lot of computational methods that have been developed to predict the binding sites and motif structure from sequence datasets. In this study, we will compare our method with four other state-of-the-art methods, and the performance is measured by the area under the receiver operating characteristic curve (AUROC).

- GraphProt(Maticzka et al. 2014): It used a graph kernel that can represent a sequence feature and the potential secondary structure of potential motifs. These graph kernel features are fed into support vector machine (SVM) to classify bound sites from unbound sites.
- Deepnet-rbp(Zhang et al. 2016): It constructs the features by fusing the RNA primary sequence, secondary and tertiary structural features together and inputs the feature into a DBN to predict RBP binding sites.
- iDeepE(Pan and Shen 2018): On the one hand, it uses two-layer local multi-channel CNNs to convolve the multiple subsequences of RNA sequence in parallel to get the local outputs. On the other hand, it uses two-layer global 1-channel for padded RNA sequences to get the global outputs. Lastly, it takes the average of local and global outputs as the final output.
- RPI-Net(Yan, Hamilton, and Blanchette 2020): A gated GNN based model is proposed to learning the graph representation of the RNA secondary structures. It uses a LSTM that treats the node embedding as a hidden state and the messages coming to each node as input.

Datasets

We compare the performance of our method with other baselines on the benchmark RBP binding dataset in GraphProt originally created by (Maticzka et al. 2014). This dataset consists of 24 sets of HITS-CLIP-, PAR-CLIP- and iCLIP-derived binding sites, where 23 sets were derived from doRiNA (Anders et al. 2011) and PTB HITS-CLIP binding sites was taken from (Xue et al. 2009). For this dataset, the region identified in the CLIP-seq experiment are flanked by 150 nucleotides on both ends. According to the positive sites, the ”viewpoint” region in each RNA positive sequence is identified by experiment. The positive sites are subsequence anchored at the peak center, which are derived from CLIP-seq processed in doRiNA. The negative RNA samples are generated by choosing at random viewpoint-sized portions of human transcripts, where there is no supportive evidence of existing binding sites.

According to the CLIP-Seq protocol, including PAR-CLIP (Hafner et al. 2010) and HITS-CLIP (Licatalosi et al. 2008), the RNA regions that contain RBP binding sites are separated from the original transcript using particular RNA cleavage enzymes such as RNase T1. The borders of the RNA regions have the same base distribution because of the same cleavage enzymes, which may misguide the model to learn the information of the cleavage enzymes. To avoid the borders misleading problem, dataset is debiased through randomizing the borders of the RNA regions using the four kinds of nucleotides.

In this paper, we still focus on the original binding dataset. However, RPI-Net doesn’t provide the results on the original dataset. Thus, we conduct some extended experiments on the debiased dataset to compare RNASSR-Net with RPI-Net.

Parameters Setting

To investigate the performance of our model, we randomly select 90% of origin training set from RBP-24 as the training

RBP	GraphProt	Deepnet-rbp	iDeepE	Ours
ALKBH5	68.0%	71.4%	75.8%	**77.1%**
C17ORF85	80.0%	82.0%	83.0%	**88.9%**
C22ORF28	75.1%	79.2%	83.7%	**86.5%**
CAPRIN1	85.5%	83.4%	89.3%	**92.4%**
Ago2	76.5%	80.9%	88.4%	**89.0%**
ELAVL1H	95.5%	96.6%	97.9%	**98.3%**
SFRS1	89.8%	93.1%	94.6%	**95.3%**
HNRNPC	95.2%	96.2%	97.6%	**98.0%**
TDP43	87.4%	87.6%	94.5%	**95.2%**
TIA1	86.1%	89.1%	93.7%	**94.8%**
TIAL1	83.3%	87.0%	93.4%	**94.6%**
Ago1-4	89.5%	88.1%	91.5%	**93.7%**
ELAVL1B	93.5%	96.1%	97.1%	**98.0%**
ELAVL1A	95.9%	96.6%	96.4%	**97.7%**
EWSR1	93.5%	96.6%	96.9%	**97.0%**
FUS	96.8%	98.0%	98.5%	**98.6%**
ELAVL1C	99.1%	**99.4%**	98.8%	99.1%
IGF2BP1-3	88.9%	87.9%	94.7%	**97.0%**
MOV10	86.3%	85.4%	91.6%	**94.0%**
PUM2	95.4%	97.1%	96.7%	**97.9%**
QKI	95.7%	**98.3%**	97.0%	98.1%
TAF15	97.0%	98.3%	97.6%	**98.4%**
PTB	93.7%	**98.3%**	94.4%	94.9%
ZC3H7B	82.0%	79.6%	90.7%	**91.7%**
Mean	88.7%	90.2%	93.1%	**94.4%**

Table 1: Performance of our method and other baseline methods across 24 experiments on original dataset. The bold font indicates the best AUC among compared methods.

set and the remaining 10% as the validation set. We train our model for a maximum of 100 epochs using Adam (Kingma and Ba 2015). For each epoch, we set the size of batched training data as 128. For the learning rate parameters, we set the initial learning rate as 0.001 and the learning rate reduce factor as 0.5. The learning rate will decrease by learning rate reduce factor if the validation loss does not decrease for 10 consecutive epochs. When the learning rate is equal to 1e-5, the training procedure will stop. We save the model when the model gets the highest AUC on the validation data. Small change of the other parameters did not change the results much. We set the weight decay and dropout as 0.01 and 0.25 respectively. For baseline models, we set the parameters same as their original papers.

We run our experiments on a Ubuntu server with NVIDIA GTX 2080Ti GPU with memory 12 GB. The initial weights and bias use default setting in PyTorch. We construct the graph using DGL, which is known as an open source framework for graph neural networks.

Binding Prediction Results on the Original Dataset

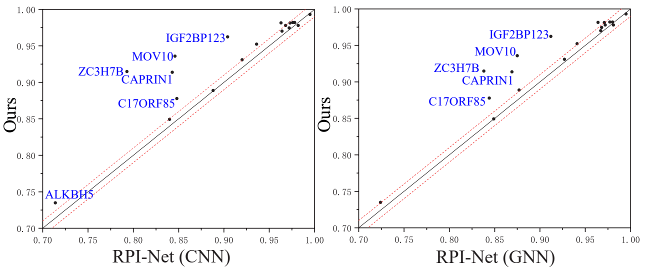
We compare our method with other baseline methods on the 24 proteins sub-datasets. The detailed results are listed in the Table 1. In general, our method performs better than other baseline methods on almost all the proteins, which proves that our method can perform better in the task of predicting RNA binding sites.

As shown in Table 1, our method yields the best mean AUC 94.4% on the 24 sub-datasets, which is better than GraphProt by 5.7%, deepnet-rbp by 4.2%, ideepE by 1.3%. GraphProt and deepenet-rbp are sequence-structure-profile methods, whereas iDeepE and our method are PWM-profile models. iDeepE and our method can better learn the potential motifs from the convolutional kernel in the neural network automatically. iDeepE only extracts the RNA features from sequence data, while our method integrates the RNA structure features using GCN. The introduction of RNA structure improves the binding prediction a lot. In addition, GraphProt performs worse than the other approaches. Support vector machine (SVM), which is a traditional machine learning method, is used in GraphProt to predict the occurrence of the binding sites. Although GraphProt uses the graph kernel to represent the RNA structure, the graph kernel cannot adequately represent structural information for prediction. From the results, deep learning methods can show obvious advantages over traditional machine learning methods. Deepnet-rbp extracted the RNA secondary structure and tertiary structure features using DBNs, However, the higher dimension is, the less accuracy of RNA structure is. Deepnet-rbp codes the RNA structure data, which means information missing. Our method selects the original RNA structure denoted as graph data for prediction to extract valuable information as much as possible. Our method yields the best AUC on 21 proteins among the 4 methods. For another three protein datasets, deepnet-rbp obtains the best AUC than the other methods, while our method has very similar AUCs as deepnet-rbp.

Besides our method significantly improves the performance on the small-size sub-datasets. Our method increases ALKBH5 by 1.3%, C17ORF85 by 5.9%, C22ORF28 by 2.8% and CAPRIN1 by 3.4% compared to the best baseline AUC.

Binding Prediction Results on the Debiased Dataset

For the debiased dataset, the border of the "viewpoint" region in the positive samples are randomized. From the previous research, the performance of the CNN based methods will decrease because of the lost of enzyme cleave information. To make sure that our method is able to learn



(a) Ours vs RPI-Net(CNN) **(b) Ours vs RPI-Net(GNN)**

Figure 2: Comparison of AUC between our method and RPI-Net on the debiased dataset. The red dotted lines correspond to a 1% difference.

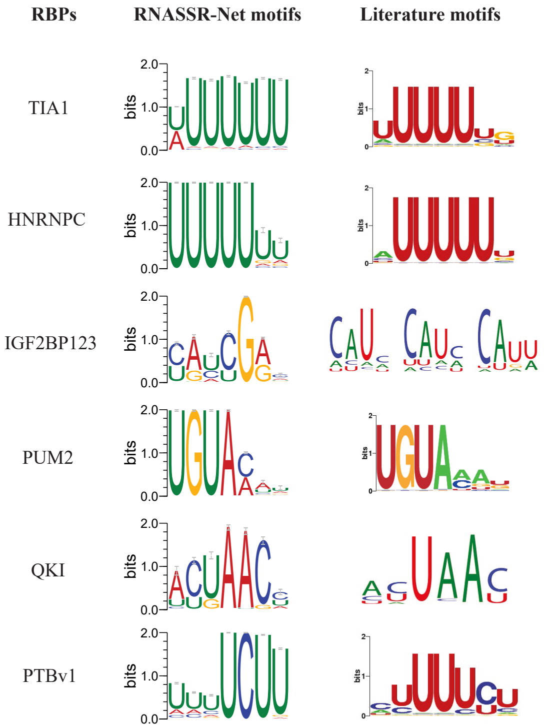


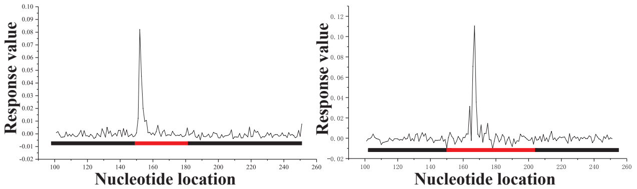
Figure 3: Samples of detected motifs compared to the literature motifs.

the prediction pattern without the enzyme cleave information, we conduct some comparative experiments on the 19 debiased sub-datasets. Specifically, our method outperforms other baseline methods on the debiased dataset. Our method yields the best mean AUC 93.6% on the 19 debiased sub-datasets, which is better than GraphProt by 7.3%, deepnetrbp by 3.9%, ideepE by 5.8% and RPI-Net by 1.9%.

RPI-Net gets better performance than other baseline methods on the debiased datasets. Thus, we show the comparison between our method and RPI-Net in Figure 2. Our method outperforms the RPI-Net generally, because the integration features of the secondary structure and sequence can improve the performance of RNA binding prediction. Our mehtod on several RBPs(ALKBH5, C17ORF85, CAPRIN1, SFRS1, ELAVL1 and ZC3H7B) achieves the AUCs exceeding 1% compared with RPI-Net, which indicates the importance of secondary structure features.

Identified Motifs

As the other deep learning methods for RBPs, our method can automatically identify the binding sequence motifs from the learned parameters of the first convolution filters in the model. After several experiments, we get the candidate motifs for RBPs in RBP-24 dataset. We sample some identified motifs to show the motifs detection ability of RNASSR-Net in the Figure 3. With the knowledge of the current CISBP-RNA and some other research, we calculate the frequency of candidate motifs in the sequences and show the known motifs and the detected motifs of our method using TOMTOM.



(a) CNN response result
(b) GCN response result

Figure 4: Nucleotide base response curves of CNN and GCN. The sequence under the curve corresponds to the sequence category and structure category of bases. Red regions indicate the “viewpoint” region.

However, there are many proteins that don’t have corresponding verified motifs in current CISBP-RNA. So we conduct the enrichment analysis to show the most possible motifs for these RBPs. We compare these detected motifs with some data from literature. FUS, TAF15 and EWSR1 show preference for AU-rich sites (Hoell et al. 2011).

Response Analysis of RNA

We novelly extract the response value of each nucleotide in the RNA from the features of CNN and GCN in Figure 4. The peak areas are the predicted “viewpoint” in the sequence, which contributes to the binding prediction. The sequence under the curve corresponds to the sequence category and structure category of bases. Specially, the structure in the GCN result that has the minimum free energy is predicted by RNAfold. From Figure 4, the peak areas are consistent with the real “viewpoint” region, which shows that RNASSR-Net has a good ability of identifying motifs regions.

The nucleotide base response curves of GCN can reflect the structure mutation. In Figure 4(b), the left border of the “viewpoint” region is the transition of ‘S’ to ‘H’, where exists a trough of the curve. In the peak area, a trough exists between the first and the second high peak, which corresponds to the “SSBBSSS” in the structure. The mutation of ‘B’ results in fluctuation between two peaks.

Conclusion

In this study, we develop a novel deep learning model based on CNN and GCN to predict the RBP binding sites from the data derived from CLIP-seq. To guide the GCN training process, we introduce the base weights extracted by CNN to the nodes of RNA graph. We conduct several experiments on the original dataset and the debiased dataset. From the experiment results, our method yields better performance than the other state-of-the-art methods. The integration of GCN and CNN improves deep learning models on both the original dataset and the debiased dataset. Besides, our method can detect accurate binding motifs automatically. We novelly analyze the response curve of CNN and GCN to interpret the focused region and structure mutation of RNA for binding prediction.

Acknowledgments

This work was supported in part by National Key Research and Development Program of China under Grant: 2018AAA0101100, the National Natural Science Foundation of China under Grants 61822113, 41871243, 61801336, 62041105 and 62071340, the Natural Science Foundation of Hubei Province under Grants 2018CFA050, the National Key R & D Program of China under Grant 2018YFA0605500, the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170, and the China Postdoctoral Science Foundation under Grant 2019M662717 and 2020T130480.

The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Alipanahi, B.; Delong, A.; Weirauch, M. T.; and Frey, B. J. 2015. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology* 33(8): 831–838.

Almagro Armenteros, J. J.; Sønderby, C. K.; Sønderby, S. K.; Nielsen, H.; and Winther, O. 2017. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 33(21): 3387–3395.

Anders, G.; Mackowiak, S. D.; Jens, M.; Maaskola, J.; Kuntzagk, A.; Rajewsky, N.; Landthaler, M.; and Dieterich, C. 2011. doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic acids research* 40(D1): D180–D186.

Aviv, T.; Lin, Z.; Ben-Ari, G.; Smibert, C. A.; and Sicheri, F. 2006. Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1p. *Nature structural & molecular biology* 13(2): 168–176.

Bailey, T. L.; Boden, M.; Buske, F. A.; Frith, M.; Grant, C. E.; Clementi, L.; Ren, J.; Li, W. W.; and Noble, W. S. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* 37(suppl_2): W202–W208.

Buckanovich, R. J.; and Darnell, R. B. 1997. The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. *Molecular and cellular biology* 17(6): 3194–3201.

Chami, I.; Wolf, A.; Juan, D.; Sala, F.; Ravi, S.; and Ré, C. 2020. Low-Dimensional Hyperbolic Knowledge Graph Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 6901–6914.

Cong, Y.; Ulasli, M.; Schepers, H.; Mauthe, M.; V'kovski, P.; Kriegenburg, F.; Thiel, V.; de Haan, C. A.; and Reggiori, F. 2020. Nucleocapsid protein recruitment to replication-transcription complexes plays a crucial role in coronaviral life cycle. *Journal of virology* 94(4).

Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, 2224–2232.

Ferre, F.; Colantoni, A.; and Helmer-Citterich, M. 2015. Revealing protein–lncRNA interaction. *Briefings in bioinformatics* 17(1): 106–116.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 1263–1272.

Gori, M.; Monfardini, G.; and Scarselli, F. 2005. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, 729–734. IEEE.

Gruber, A. R.; Lorenz, R.; Bernhart, S. H.; Neuböck, R.; and Hofacker, I. L. 2008. The vienna RNA websuite. *Nucleic acids research* 36(suppl_2): W70–W74.

Hackermüller, J.; Meisner, N.-C.; Auer, M.; Jaritz, M.; and Stadler, P. F. 2005. The effect of RNA secondary structures on RNA-ligand binding and the modifier RNA mechanism: a quantitative model. *Gene* 345(1): 3–12.

Hafner, M.; Landthaler, M.; Burger, L.; Khorshid, M.; Hausser, J.; Berninger, P.; Rothballer, A.; Ascano Jr, M.; Jungkamp, A.-C.; Munschauer, M.; et al. 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141(1): 129–141.

Heffernan, R.; Paliwal, K.; Lyons, J.; Dehzangi, A.; Sharma, A.; Wang, J.; Sattar, A.; Yang, Y.; and Zhou, Y. 2015. Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Scientific reports* 5(1): 1–11.

Hoell, J. I.; Larsson, E.; Runge, S.; Nusbaum, J. D.; Duggimpudi, S.; Farazi, T. A.; Hafner, M.; Borkhardt, A.; Sander, C.; and Tuschl, T. 2011. RNA targets of wild-type and mutant FET family proteins. *Nature structural & molecular biology* 18(12): 1428–1431.

Kazan, H.; Ray, D.; Chan, E. T.; Hughes, T. R.; and Morris, Q. 2010a. RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins. *PLOS Computational Biology* 6(7): 1–10.

Kazan, H.; Ray, D.; Chan, E. T.; Hughes, T. R.; and Morris, Q. 2010b. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol* 6(7): e1000832.

Kingma, D. P.; and Ba, J. L. 2015. Adam: A method for stochastic gradient descent. In *ICLR: International Conference on Learning Representations*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Licatalosi, D. D.; Mele, A.; Fak, J. J.; Ule, J.; Kayikci, M.; Chi, S. W.; Clark, T. A.; Schweitzer, A. C.; Blume, J. E.; Wang, X.; et al. 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456(7221): 464–469.

Lukong, K. E.; Chang, K.-w.; Khandjian, E. W.; and Richard, S. 2008. RNA-binding proteins in human genetic disease. *Trends in Genetics* 24(8): 416–425.

Lusci, A.; Pollastri, G.; and Baldi, P. 2013. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of chemical information and modeling* 53(7): 1563–1575.

Maticzka, D.; Lange, S. J.; Costa, F.; and Backofen, R. 2014. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome biology* 15(1): 1–18.

Monti, F.; Frasca, F.; Eynard, D.; Mannion, D.; and Bronstein, M. M. 2019. Fake News Detection on Social Media using Geometric Deep Learning. *CoRR* abs/1902.06673.

Musunuru, K. 2003. Cell-specific RNA-binding proteins in human disease. *Trends in cardiovascular medicine* 13(5): 188–195.

Nelson, G. W.; Stohlman, S. A.; and Tahara, S. M. 2000. High affinity interaction between nucleocapsid protein and leader/intergenic sequence of mouse hepatitis virus RNA. *Journal of General Virology* 81(1): 181–188.

Orenstein, Y.; Wang, Y.; and Berger, B. 2016. RCK: accurate and efficient inference of sequence-and structure-based protein–RNA binding models from RNAcompete data. *Bioinformatics* 32(12): i351–i359.

Pan, X.; and Shen, H.-B. 2018. Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics* 34(20): 3427–3436.

Pudimat, R.; Schukat-Talamazzini, E.-G.; and Backofen, R. 2005. A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics* 21(14): 3082–3088.

Ray, D.; Kazan, H.; Cook, K. B.; Weirauch, M. T.; Najafabadi, H. S.; Li, X.; Gueroussov, S.; Albu, M.; Zheng, H.; Yang, A.; et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499(7457): 172–177.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE Transactions on Neural Networks* 20(1): 61–80.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, 593–607. Springer.

Stohlman, S.; Baric, R.; Nelson, G.; Soe, L.; Welter, L.; and Deans, R. 1988. Specific interaction between coronavirus leader RNA and nucleocapsid protein. *Journal of virology* 62(11): 4288–4295.

Stražar, M.; Žitnik, M.; Zupan, B.; Ule, J.; and Curk, T. 2016. Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics* 32(10): 1527–1535.

Ule, J.; Jensen, K. B.; Ruggiu, M.; Mele, A.; Ule, A.; and Darnell, R. B. 2003. CLIP identifies Nova-regulated RNA networks in the brain. *Science* 302(5648): 1212–1215.

Ule, J.; and Rinn, John, L. 2014. 'Oming in on RNA-protein interactions. *Genome Biology* .

Van Nostrand, E. L.; Pratt, G. A.; Shishkin, A. A.; Gelboin-Burkhart, C.; Fang, M. Y.; Sundararaman, B.; Blue, S. M.; Nguyen, T. B.; Surka, C.; Elkins, K.; et al. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature methods* 13(6): 508–514.

Xue, Y.; Zhou, Y.; Wu, T.; Zhu, T.; Ji, X.; Kwon, Y.-S.; Zhang, C.; Yeo, G.; Black, D. L.; Sun, H.; et al. 2009. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Molecular cell* 36(6): 996–1006.

Yan, Z.; Hamilton, W. L.; and Blanchette, M. 2020. Graph neural representational learning of RNA secondary structures for predicting RNA-protein interactions. *Bioinformatics* 36: i276–i284.

Zhang, S.; Zhou, J.; Hu, H.; Gong, H.; Chen, L.; Cheng, C.; and Zeng, J. 2016. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic acids research* 44(4): e32–e32.