Computational Biology and Chemistry 83 (2019) 107088

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

DM-RPIs: Predicting ncRNA-protein interactions using stacked ensembling strategy

Shuping Cheng, Lu Zhang, Jianjun Tan*, Weikang Gong, Chunhua Li, Xiaoyi Zhang

College of Life Science and Bioengineering, Beijing University of Technology, Intelligent Physiological Measurement and Clinical Translation, Beijing International Base for Scientific and Technological Cooperation, Beijing, 100124, China

ARTICLE INFO

Keywords:
ncRNA-protein interactions
Deep Stacking Auto-encoders Networks (DSANs)
Support Vector Machine (SVM)
Random Forest (RF)
Convolution Neural Network (CNN)
Stacked integrate

ABSTRACT

ncRNA-protein interactions (ncRPIs) play an important role in a number of cellular processes, such as post-transcriptional modification, transcriptional regulation, disease progression and development. Since experimental methods are expensive and time-consuming to identify the ncRPIs, we proposed a computational method, Deep Mining ncRNA-Protein Interactions (DM-RPIs), for identifying the ncRPIs. In order to descending dimension and excavating hidden information from k-mer frequency of RNA and protein sequences, using the Deep Stacking Auto-encoders Networks (DSANs) model refined the raw data. Three common machine learning algorithms, Support Vector Machine (SVM), Random Forest (RF), and Convolution Neural Network (CNN), were separately trained as individual predictors and then the three individual predictors were integrated together using stacked ensembling strategy. Based on the RPI2241 dataset, DM-RPI obtains an accuracy of 0.851, precision of 0.852, sensitivity of 0.873, specificity of 0.826, and MCC of 0.701, which is promising and pioneering for the prediction of ncRPIs.

1. Introduction

Non-coding RNA (ncRNA) plays an important role in many biological processes, especially when ncRNA bind with a variety of proteins, such as post-transcriptional modification, transcriptional regulation, protein synthesis, human disease and so on. Although the interactions between ncRNA and protein in the regulation of gene expression is important, but only a few number of ncRNA-protein interactions (ncRPIs) have been studied. Some experimental methods have been developed into analyze ncRPIs, for example, HITS-CLIP, (Weyn-Vanhentenryck et al., 2014), PAR-CLIP (Friedersdorf and Keene, 2014), RIPiT-Seq (Guramrit et al., 2014) and RNAcompete-S (Cook et al., 2017). However, these experimental methods are time-consuming, expensive and labor-intensive. It is necessary to develop computational methods to predicting ncRPIs.

Several computational methods have been proposed to predict ncRPIs, which can be divided into semi-supervised and supervised methods the former trained on a combination of labeled and unlabeled data and the latter trained on datasets that include labels. Some researchers developed semi-supervised methods for predicting RNA-protein interactions. For example, Liu et al. proposed LPI-NRLMF method (Liu et al., 2017) for predicting lncRNA-protein interactions by neighborhood regularized logistic matrix factorization in 2017. Zhao et al. proposed IRWNRLPI method (Zhao et al., 2018a, 2018b), integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interactions prediction in 2018. In the same year, Zhao et al. proposed LPI-BNPRA method (Zhao et al., 2018a, 2018b) using the bipartite network projection recommended algorithm to identify lncRNA-protein interactions. In the same year, Chen et al. proposed BNPMDA method (Chen et al., 2018). The above methods extraced lncRNA and protein sequence similarity matrixes, used semi-supervised algorithm to predict lncRNA-protein interactions and all of them obtained high predictive accuracy. They performed well only for predicting interactive pairs but performed weakly for predicting non-interactive pairs. In the same way, supervised methods are essential for predicting ncRPIs. In 2011, Pancaldi et al. trained support vector machines (SVM) and random forest (RF) models to predict mRNA-protein interactions in yeast, extracting secondary structure and physical property of protein and RNA as features, this approach performed well for RNA binding proteins with known targets (Vera Pancaldi, 2011). In the same year, Bellucci et al. proposed CatRAPID method to predict mRNA-protein interactions, which extracted features from the secondary structure of protein and RNA, hydrogen bond and Van Edward force, to calculate the binding preference of polypeptide and nucleotide chain (Bellucci et al., 2011). In 2013, Lu et al. developed the method of lncPro, which extracted the biological properties of RNA and protein

* Corresponding author.
E-mail address: tanjianjun@bjut.edu.cn (J. Tan).

https://doi.org/10.1016/j.compbiolchem.2019.107088
Received 28 January 2019; Received in revised form 30 June 2019; Accepted 2 July 2019
Available online 06 July 2019
1476-9271/

(Lu et al., 2013), such as the secondary structure of RNA and protein, the hydrogen bond between RNA and protein, and so on. In 2015, Suresh et al. trained a SVM classifier called RPI-Pred basing on sequence and 3D structure of RNA and protein (Suresh et al., 2015). In 2016, Pan et al. provided IP-Miner method using the Deep Stacking Auto-encoders Networks (DSANs) and stacked ensembling strategy to predict ncRPIs (Pan et al., 2016). In 2018, Hu et al. proposed HLPI-Ensemble method (Hu et al., 2018) for identifying lncRNA-protein interactions in human only, which integrated three common machine learning algorithms, SVM, RF and Extreme Gradient Boosting (XGB), which performed well for predicting lncRNA-protein interactions in human. In the same year, Chen et al. Proposed BNPMDA method (Chen et al., 2018) for MiRNA-Disease Association prediction (BNPMDA) model based on the rating-integrated bipartite network recommendation and the know miRNA-disease associations.

Nowadays several benchmark datasets have been constructed, some of them are small-scale datasets, such as RPI369, RPI488. SVM and RF performed better on those small-scale datasets. But some of them are larger-scale datasets, such as RPI2241, RPI13254. Neural network performed better on these larger-scale datasets. And with the development of deep learning, all kinds of neural networks (such as a convolutional neural network (CNN)) are widely used in many areas, for example, speech recognition (Achanta and Gangashetty, 2017), image processing (Pham et al., 2018), etc.. Thus, it is necessary and meaningful to integrate SVM, RF and CNN together to identify ncRPIs, by integrating the classifier can perform well not only on small-scale datasets but also performs well on large-scale datasets. In the study, we developed DM-RPIs (Deep Mining ncRPIs) for predicting ncRPIs. An ensembling classifier was trained to predict ncRPIs by sequence information. Firstly, DSANs was trained to preprocess raw data. Then, three individual classifiers, SVM, RF, and CNN, were separately trained to identify ncRPIs. The performance of three individual classifiers were comparable, the predictive accuracy increased about 15% than without DSANs to preprocess raw data on RPI369 and RPI2241. Finally, the three individual classifiers were integrated using stacked strategy and tested using 5-fold cross validation on RPI369, RPI488, RPI1807, RPI2241 and RPI13254, respectively. Overall, DM-RPIs is superior to the three individual predictors in predicting ncRPIs.

2. Materials and methods

2.1. Source of the datasets

We collected five datasets from the published papers, including RPI2241, RPI369 (Muppirala et al., 2011), RPI1807 (Suresh et al., 2015), RPI13254 (Pancaldi and Bähler, 2011) and RPI488 (Pan et al., 2016) as Table 1 shown. RPI2241 is generated by extracting 943 protein-RNA complexes from The Protein-RNA Interface Database (PRIDB) (Lewis et al., 2011), the RNA include rRNA, ncRNA, mRNA and so on. Interactions are generated by using distance threshold (8 Å) on the dataset. RPI369 is a subset of RPI2241, which removes all RNA-protein interactions that contain ribosomal protein or ribosomal RNA. Above two datasets were published including positive pairs only, the negative pairs were generated at random. The RPI1807 was constructed by Suresh (Suresh et al., 2015), whose RNA-protein interactions were extracted from the Nucleic Acid Database (NDB) (Narayanan et al., 2014) and the protein-RNA interface database (PRIDB) (Lewis et al., 2011) using an 3.4 Å distance cut-off. RPI13254 is a large-scale non-structure-based experimental dataset, which including 13,254 positive pairs and 5172 negative pairs. We randomly selected 5172 positive pairs, which were balanced with the negative pairs. RPI488 is a structure-based lncRNA-protein interactions dataset, and the interactive pairs were selected with 5 Å cut-off.

Table 1
The more information about 5 datasets.

Datasets	Interactions	Of RNAs	Of proteins	Positive + negative	Cut-off (Å)
RPI2241	2241	443	952	2241	8
RPI369	369	332	338	369	8
RPI1807	1807	1078	1807	1807 + 1436	3.4
RPI13254	13,254	4500	42	13,254 + 5172	–
RPI488	243	25	247	243 + 245	5

2.2. Conjoint 3-mer residues for protein and 4-mer nucleic acids for RNA

20 amino acids are divided into 7 groups according to their dipole moments and volume of their side chain: [A, G, V], [I, L, F, P], [Y, M, T, S], [H, N, Q, W], [R, K], [D, E] and [C] (Pandey et al., 2018). The protein chain is represented by conjoint triad features (CTF), where each feature represents normalized frequency of 3-mer in the 7-letter representation of the protein sequence, resulting in 343 ($7 \times 7 \times 7$) dimensional feature vectors. Similarly, each RNA chain is represented by normalized frequency of 4-mer sequence fragment. Thus, each RNA chain is quantified by 256 ($4 \times 4 \times 4 \times 4$) dimensional feature vector. After the above steps, we got a vector of 599 (343 + 256) dimensions to represent each interaction.

$$f_i = \frac{N_i}{\sum_{i=1}^{n} N_i} \quad (1)$$

In the Formula (1), n is 343 or 256 standing for the dimension of a vector, N_i stands for the amount of the i-th feature.

2.3. Data preprocess

The raw features which were obtained from above were inputted into the DSANs model (Fig. 1), so that we can obtain low dimensional features and remove noise among data. A DSANs model was trained including 3 hidden layers and fine tuning, the number of neurons for 3 hidden layers in DSANs model is 256,128, and 64, respectively. The protein and RNA raw sequence features are inputted the DSANs model. At last we obtained the vectors of 128 (64 + 64) dimensions for each pair via the model, features were lower dimensions (128) than before (599). It was unsupervised in 3 hidden layers, the labels needed not to participate. In the fine tuning process, it is supervised to update parameters by comparing the really labels with the predicted results. As this process as shown in Fig. 1, we could obtain low dimensions and representative features which can improve the performance of the method for predicting ncRPIs.

2.4. The brief introduction of SVM, RF and CNN

SVM is a popular supervised machine learning method, and it is applied widely for many binary classification problems. The RBF was selected as the kernel function. The SVM classifier was trained to efficiently predict ncRPIs with C = 32 and $\gamma = 0.125$. RF is a simple, easy to implement and little computational storage algorithm. It shows well performance in many classified tasks and be used more and more widely. In the predictor, we selected n_estimators, max_depth, min_samples_split, min_samples_leaf and max_features with 100, 13, 40, 6 and 13, respectively, by grid search. CNN is a kind of widely used neural network in image processing. It is usually including convolutional layer, pooling layer, batchnormalization layer and so on. In the study, a 21 layers CNN classifier is constructed, including 7 BatchNormalization layers, 5 Conv1D layers, 3 Dropout layers, 3 Dense layers, 2 MaxPooling layers, 1 GlobalAveragePooling layer, we chosen 'Adam' function as optimizer, the learning rate was 0.0005. The last dense layer includes 2 neurons, 'softmax' function is the activation function of the layer, so that the CNN classifier performs two classified task by the layer.

We used stacked ensembling strategy to integrate three individual

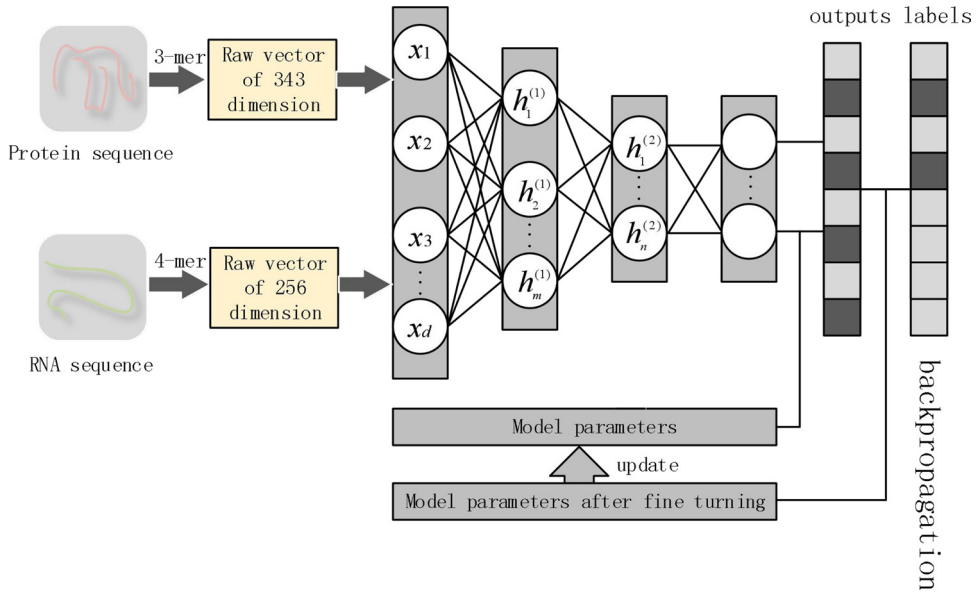

Fig. 1. The Deep Stacking Auto-encoders Networks and fine tuning.

predictors, SVM, RF and CNN. Compared to general averaging ensembling strategy, the stacked ensembling strategy performed better (Pan et al., 2016) for predicting ncRPIs, which is denoted as

$$P_{\mathbf{w}}(y = \pm 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-y\mathbf{w}^{\mathrm{T}}\mathbf{x})} \tag{2}$$

where $\mathbf{x}$ is the vectors of outputted probabilities from three single classifiers, y is the corresponding label of every interactive pair, and $\mathbf{w}$ is the weight vectors of the three individual classifiers.

2.5. Performance evaluation

We trained three individual classifiers, SVM, RF and CNN, and one ensembling classifier, DM-RPIs. These classifiers were calculated by 5-fold cross-validation from six measures, accuracy, precision, sensitivity, specificity, F-measure and MCC. In the following formulas, the *TP* is the number of true positives, meaning the positive interactions predicted as positive interactions. *FP* is the number of false positives, meaning the negative interactions predicted as positive interactions. *TN* is the number of true negatives, meaning the negative interactions predicted as negative interactions. *FN* is the number of false negatives, meaning the positive interactions predicted as negative interactions. In addition, we made the Receiver Operating Characteristic (ROC) curve on RPI1807, RPI2241 and RPI13254, and calculated the area under the ROC curve as the AUC value.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$F - Measure = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{7}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN) + (TN + FP)(TN + FN)}} \tag{8}$$

3. Result

3.1. Comparison between three individual classifiers and the ensembling classifier–DM-RPIs

On RPI369, the results of six evaluation indicators of three individual classifiers (SVM, RF, CNN) and one ensembling classifier, DM-RPIs are showed in Fig. 2. Three individual classifiers perform differently in different measures. In these three individual models, the RF classifier achieves highest performance with the accuracy of 0.760, precision of 0.772, specificity of 0.788 and MCC of 0.540, respectively. The CNN classifier achieves highest performance with the sensitivity of 0.859 and F-measure of 0.770. It obtains the highest score with sensitivity of 0.859, which increases by about 10% over SVM of 0.750 and RF of 0.738. The CNN classifier performs worst with the minimum specificity of 0.632, which descends by about 10% over SVM of 0.763 and RF of 0.788. It indicates that the CNN classifier can predict positive interactions effectively, but its performance is fairly worse for predicting negative interactions. On the whole, it implies that three individual classifiers have low correlation on predicting ncRPIs, which is relatively promising to integrate them together (Pan et al., 2016). By further comparing, the ensembling classifier DM-RPIs performs a little better than the three individual classifiers with accuracy of 0.791, F-measure of 0.797 and MCC of 0.582, respectively. Especially the MCC value which is improved by 10% approximately over three individual classifiers. It indicates that the ensembling classifier could identity ncRPIs effectively comparing to an individual classifier. It is fairly promising to integrate different individual classifiers together.

As is shown in Fig. 3, on RPI488, the RF classifier performs best with accuracy of 0.853, precision of 0.879, specificity of 0.902, F-measure of 0.851 and MCC of 0.726, respectively. Only the sensitivity is lower than the CNN classifier of 0.952. The SVM classifier achieves with accuracy of 0.851, precision of 0.846, specificity of 0.852, F-measure of 0.849 and MCC of 0.710, respectively, which are a little worse than RF classifier, besides the sensitivity of 0.848 is a little higher than RF of 0.818. The predictive results of CNN classifier are far below than other classifiers with accuracy of 0.561, precision of 0.520, specificity of 0.201, F-measure of 0.675 and MCC of 0.248, respectively, only sensitivity of 0.952 is above than other classifiers by 10% approximately. The reason may be that the RPI488 is relatively small-scale including only 243 interactions, it is too small to fit an effective deep learning network model. DM-RPIs achieves better results with accuracy of 0.851,

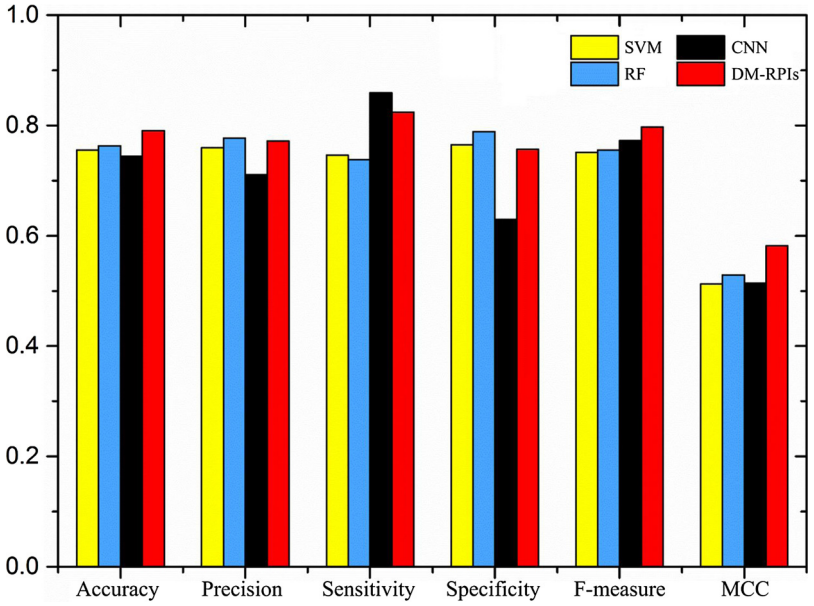

Fig. 2. Results comparison on three individual classifiers and DM-RPIs on RPI369.

precision of 0.848, sensitivity of 0.849, specificity of 0.853, F-measure of 0.849 and MCC of 0.711, respectively. On the whole, the SVM classifier, the RF classifier and DM-RPIs can identify ncRPIs effectively, and achieve relatively substantially on RPI488.

ROC curve of RPI1807 is shown in Fig. 4, RF and DM-RPIs achieve the best AUC of 0.99, both SVM and CNN achieve the AUC of 0.98. no matter the three individual predictors SVM, RF and CNN, or ensemble predictor DM-RPIs, all of them obtain ideal AUC results, which is close to 1.

We further tested DM-RPIs on other two large-scale datasets, which are RPI2241 and RPI13254. RPI2241 includes 2241 interactive pairs and 2241 non-interactive pairs, which is structure-based dataset. On RPI2241, DM-RPIs performs best with AUC of 0.92 (Fig. 5), which exceeds other individual classifiers in identifying ncRPIs. In addition, we downloaded the RPI13254, which is a large-scale non-structure-based experimental dataset. It covers 13,254 positive interactions and 5172 negative interactions, but we only selected 5172 positive interactions

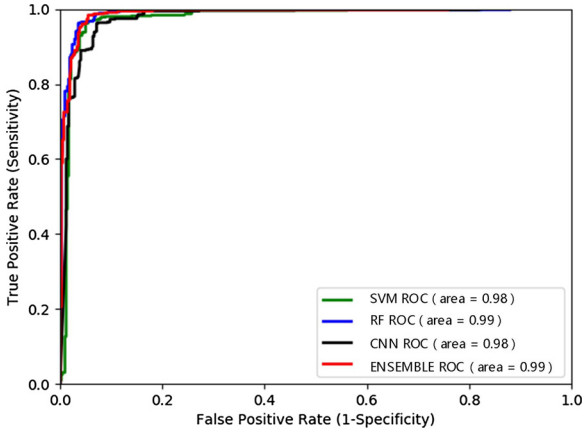

Fig. 4. The ROC curve of three individual classifiers and DM-RPIs on RPI1807.

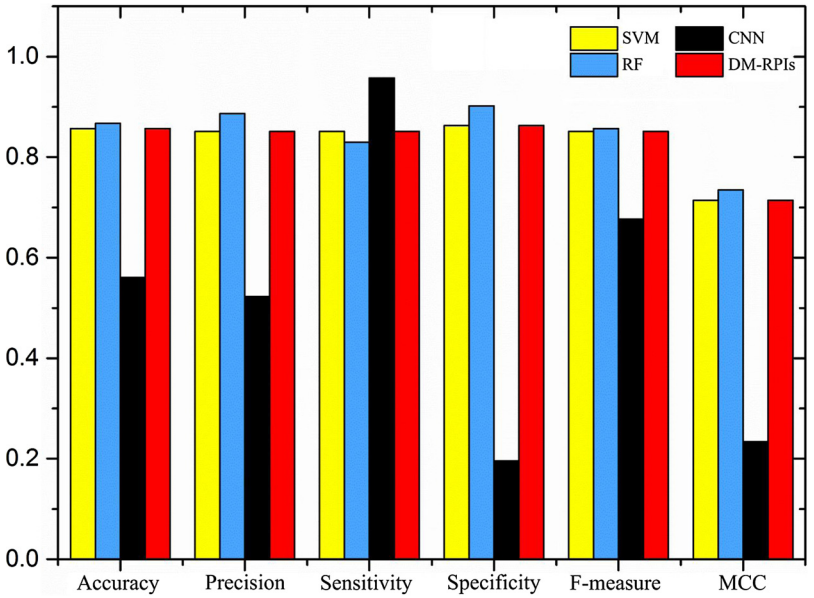

Fig. 3. Results comparison on three individual classifiers and DM-RPIs on RPI488.

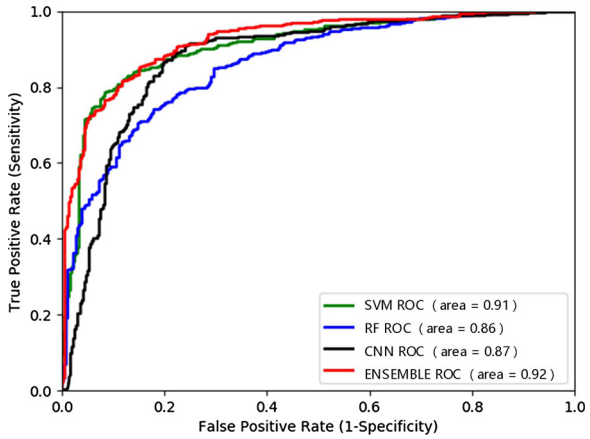

Fig. 5. The ROC curve of three individual classifiers and DM-RPIs on RPI2241.

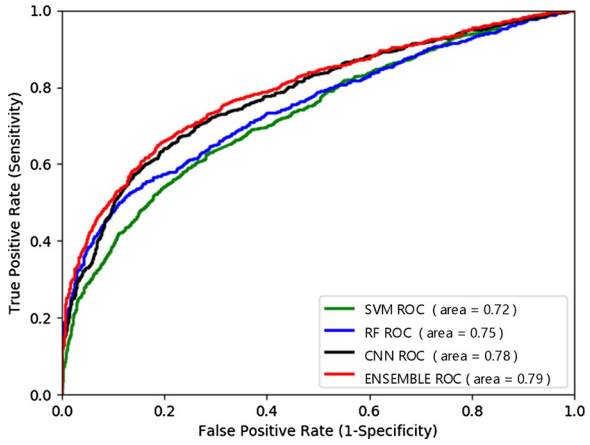

Fig. 6. The ROC curve of three individual classifiers and DM-RPIs on RPI13254.

from 13,254 positive interactions in order to balance the negative interactions. On the non-structure-based dataset RPI13254, DM-RPIs still achieves a little better than other individual classifiers with AUC of 0.79 (Fig. 6), which indicates that it is fairly promising to combine individual predictors together.

On the whole, individual classifiers perform differently on different datasets. On RPI1807, RF achieves the best AUC of 0.99 (Fig. 4). On RPI2241, the SVM get the best AUC of 0.91 (Fig. 5). However, on RPI13254, CNN get the best AUC of 0.78 (Fig. 6). On the three datasets, the three individual classifiers get the best results respectively, and none of the three individual classifiers can surpass other two on all datasets. But on all of three datasets, DM-RPIs gets the best AUC of 0.99, 0.92, 0.79, respectively (Figs. 4–6), surpassing all of three individual classifiers. The reason is that stacked ensembling strategy can improve the predictive performance (Pan et al., 2016), which is demonstrated by the performance of DM-RPIs on predicting ncRPIs.

3.2. Comparison with previous methods for predicting ncRPIs

Because the best prediction ability of DM-RPIs, we compared DM-RPIs with previous methods, which are based on sequence information. RPISeq-RF and RPISeq-SVM was both from the paper of Muppirala (Muppirala et al., 2011), but RPISeq-SVM performed worse than RPISeq-RF on both RPI369 and RPI2241.Thus we only compared the performance of RPISeq-RF with DM-RPIs in the work. The negative interactions of RPISeq-RF and lncPro methods have not been published in their papers, their papers only published the positive interactions, we could not compare the results from original papers. In the Table 2, the results of RPISeq-RF and lncPro are from the paper of Pan (Pan et al., 2016), whose positive interactions and negative interactions are same as ours.

Table 2
Performance comparison with previous sequence-based methods.

Datasets	Methods	Accuracy	Precision	Sensitivity	Specificity	MCC
RPI1807	DM-RPIs	0.967	0.968	0.975	0.957	0.933
	RPISeq-RF	0.973	0.960	0.968	0.984	0.946
	lncPro	0.969	0.960	0.965	0.984	0.938
	IP-Miner	**0.986**	**0.978**	**0.982**	**0.993**	**0.972**
RPI2241	DM-RPIs	**0.851**	**0.852**	**0.873**	**0.826**	**0.701**
	RPISeq-RF	0.646	0.663	0.652	0.630	0.293
	lncPro	0.654	0.669	0.659	0.640	0.310
	IP-Miner	0.824	0.836	0.833	0.812	0.650
RPI369	DM-RPIs	**0.791**	**0.772**	**0.824**	0.757	**0.582**
	RPISeq-RF	0.704	0.707	0.705	0.702	0.409
	lncPro	0.704	0.713	0.708	0.696	0.409
	IP-Miner	0.752	0.713	0.735	**0.791**	0.507

The bold values are the best results of four kinds of methods on every data. For example,the first bold value, 0.986 represent that IP-Miner obtained the highest accuracy on RPI1807.

As shown in Table 2, on RPI1807 DM-RPIs obtains high score more than 0.900 no matter which measures, but DM-RPIs performs slightly worse than RPISeq-RF and IP-Miner. IP-Miner performs best with each measure more than 0.950, which are fairly perfect results on identifying ncRPIs. On the whole, on RPI1807 all the four methods achieve high performance with AUC more than 0.900. The mainly reason is that RPI1807 is constructed strictly, all methods perform well almost on RPI1807 dataset. The reason is that RPI1807 dataset was constructed fairly strictly by Suresh (Suresh et al., 2015), who set up the threshold (3.4 Å) to distinguish the positive interactions and negative interactions, the threshold is the distance between of two atoms, one from protein and other from RNA. However, the RPI488 dataset was constructed by 5 Å, RPI369 and RPI2241 are 8 Å. Moreover, negative interactions of RPI369 and RPI2241 were generated by randomly, however, the negative interactions of RPI1807 were determinate by setting the threshold greater than 3.4 Å. Lastly, in order to reduce the bias of sequence homology, the cut-off of redundant sequences set up the threshold 30%, RPI369 and RPI2241 are also 30%, but RPI488 set up 90%. This is why RPI1807 perform well in different measures regardless of classifiers. On RPI2241, DM-RPIs obtains the best results in each measure with accuracy of 0.851, precision of 0.852, sensitivity of 0.873, specificity of 0.826 and MCC of 0.701, respectively. Compared to RPISeq-RF and lncPro, the performance of DM-RPIs significantly increases in each measure, RPISeq-RF and lncPro are individual classifiers and without features preprocessing, DM-RPIs is ensembling classifier and using DSANs to preprocess raw features. It indicates that it is fairly effectively to adopt ensembling strategy and DSANs. The results of DM-RPIs increase about 3% over IP-Miner on accuracy and precision, which is also ensembling classifier and applies the DSANs to preprocess raw features. It indicates that DM-RPIs is more effective than IP-Miner to identify ncRPIs. On RPI369, DM-RPIs performs better than previous methods with accuracy of 0.791, precision of 0.772, sensitivity of 0.824 and MCC of 0.582, respectively, besides the specificity (0.757) is worse than IP-Miner (0.791). In conclusion, it indicates that DM-RPIs is fairly effectively to identify ncRPIs. It is fairly promising to apply ensembling strategy in DM-RPIs for predicting ncRPIs.

4. Discussion

A new classifier, DM-RPIs, was proposed to predict the ncRPIs using only sequence information. It integrated three individual classifiers, SVM, RF and CNN. In the preprocessing stage, the DSANs was applied to process the raw features, which can mine the hidden information from the features effectively, reduce the dimension of the features and remove the noise which perhaps interferes the predictive results among

the data (Lyons et al., 2015). In DM-RPIs, we used the stacked ensembling strategy, that the outputs predictive probabilities of the three individual classifiers were served as training data for the ensembling classifier. We adopted the logistic regression to further improve the classifier performance. The performance of a single predictive classifier is limited, by using ensembling strategy with multiple single classifiers can improve the overall performance of the model. DM-RPIs performed better than previous methods on RPI369 and RPI2241, and it also obtained ideal results on other large-scale datasets, such as RPI13254. On the whole it is a pioneering and effective classifier which integrates SVM, RF and CNN models together.

The k-mer frequency indicates sequence-binding preference of RNA-protein interactions, it has been demonstrated that the higher frequency this k-mer sequence exists in a sequence, the higher probability it is a binding motif (Jungkamp et al., 2011). For secondary structures of RNA and protein, experimental data is lacking and theoretical predictions is not accurate enough and existing some limits (Liao et al., 2010). For example, the method of RNAfold limits the length of the sequence on 7500 nucleic acids (Langdon et al., 2018). The method of mfold only predicts one RNA secondary structure for one time, which can't batch process in the engineering (Wiese and Hendriks, 2006). For the 3D-structures of RNA and protein, it is flexibility, unstable and uneasy to obtain (Bressanelli et al., 2000).Thus it is difficult to obtain the 3D-features accurately for training a model. For example, Mariusz et al. presents a novel method of the fully automated prediction of RNA 3D-structures from a user-defined secondary structure (Mariusz et al., 2012), but the method is limited on 500 nucleic acids only, this is unsuitable in the study whose most of RNAs are more than 500 nucleic acids. Moreover, the method needs to obtain secondary structures of RNA, which is unable to obtain accurately enough as mentioned above. We extracted sequence features only to train the model. Compared to structure features, the sequence features is easy to obtain and accurate enough for RNA and protein. Moreover, the k-mer frequency carrys out abundant the sequence-binding motifs information, thus we trained the model basing on sequence information in the work.

Because deep learning model is able to learn complex and statistical information from large-scale dataset, DSANs can automatically extract hidden relationship between RNA and protein. At the start of DM-RPIs classifier, the dimensions of raw features were reduced from 599 to 128, not only it reduces the computational cost for training the classifier, but also it can obtain representative features from raw data in the process. The CNN can predict the ncRPIs effectively, we sets 3 as the length of convolution kernel in the work, the features extracted by Convolution layers can pay more attention to the local sequence, it is reasonable in the work because ncRNA-protein interactions is local binding site, instead of the combination of the whole sequence (Sahiner et al., 1996). The study of Pan (Pan et al., 2018) also demonstrated the effect of CNN on identifying ncRPIs.

Different classifiers have different performances on different datasets. In the field some of the datasets are large-scale, such as RPI13254. And some are little-scale, such as RPI488, RPI369. For little-scale datasets, some traditional machine learning, for example, SVM and RF are good at carrying out classification task, and for large-scale datasets, a multilayered neural network model can be trained to carry out classification task more effectively (Ginneken, 2017). Usually only large-scale and reasonably distributed dataset can train a robust and stable deep learning network to identify ncRPIs. If a dataset is little-scale, it perhaps trains an unsound and under-fitting network. Or if a dataset is not reasonably distributed and only includes a kind of or several kinds of RNA or protein. For example, the RPI488 only includes long non-coding RNA and the RPI369 contains only non-ribosomal complexes. The classifier which is trained by these datasets can only predict the kind of interactions accurately, but it is worse effective for predicting other kinds of interactions. A kind of organism exists all kinds of RNA including ncRNA, miRNA, tRNA and so on, if we want to construct the interaction network for a kind of organism, we must train a classifier that can identify all kinds of interactions. In the work we integrated the SVM, RF and CNN classifiers training a robust and comprehensive classifier. It has been demonstrated that the classifier is promising to identify the ncRPIs.

The research is meaningful and widely used, for example, to investigate the RNA moonlighting and the target protein localization diversity based on the RNA-protein interaction data. In 2018, Cheng et al. developed MoonFinder method to identify of moonlighting lncRNAs (Cheng and Leung, 2018a, 2018b), MoonFinder is a statistical method identifying moonlighting lncRNAs without a priori knowledge through the integration of protein interactome, RNA-protein interactions and functional annotation of proteins. In the same year, Cheng et al. proposed ncTALENT (non-coding RNA target localization coefficient) method to quantify the target localization diversity of ncRNAs based on the ncRNA-protein interaction and protein subcellular localization data (Cheng and Leung, 2018a, 2018b).

Although DM-RPIs achieved better performance in different measures, there are still something need to be done in the field. The negative interactions were generated randomly on some datasets, it is unreasonable but necessary to balance the positive interactions for training an effective classifier. Researchers should also construct the negative interactions not only the positive interactions. It is very useful and necessary to develop computational methods to predict ncRPIs. Large-scale dataset is also still necessary to train a more stable and robust model. And the predictive accuracy remains to be improved for identifying ncRPIs.

5. Conclusions

We have proposed DM-RPIs method to identify ncRPIs using RNA and protein sequence information. The DSANs model was trained to preprocess the raw features, which can descend features dimension and mine hidden information from raw data. Three classifiers, SVM, RF and CNN, were separately trained for predicting ncRPIs. At last, we adopted the stacked assembling strategy to integrate the three individual classifiers to improve the predicting performance. DM-RPIs outperforms the other methods based on four datasets, RPI369, RPI488, RPI2241 and RPI1807.

Acknowledgment

The authors thank the Chinese Natural Science Foundation Project (No. 21173014) for financial support.

References

Achanta, S., Gangashetty, S.V., 2017. Deep elman recurrent neural networks for statistical parametric speech synthesis. Speech Commun. 93, 31–42.
Bellucci, M., Agostini, F., Masin, M., Tartaglia, G.G., 2011. Predicting protein associations with long noncoding RNAs. Nat. Methods 8 (6), 444–445.
Bressanelli, S., Tomei, L., Roussel, A., Incitti, I., Vitale, R.L., Mathieu, M., Rey, F.A., 2000. The RNA-dependent RNA polymerase of hepatitis C virus: 3D structure and implications for viral replication. Acta Crystallogr. 56, s76 (Suppl).
Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z., Liu, H., 2018. BNPMDA: bipartite network projection for MiRNA-Disease association prediction. Bioinformatics 34 (18), 3178–3186.
Cheng, L., Leung, K.S., 2018a. Quantification of non-coding RNA target localization diversity and its application in cancers. J. Mol. Cell Biol. 10 (2), 130–138.
Cheng, L., Leung, K.S., 2018b. Identification and characterization of moonlighting long non-coding RNAs based on RNA and protein interactome. Bioinformatics 34 (20), 3519–3528.
Cook, K.B., Vembu, S., Ha, K.C.H., Hong, Z., Laverty, K.U., Hughes, T.R., Morris, Q.D., 2017. RNAcompete-S: combined RNA sequence/structure preferences for RNA binding proteins derived from a single-step in vitro selection. Methods 126, 18–28.
Friedersdorf, M.B., Keene, J.D., 2014. Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. Genome Biol. 15 (1) (2014-01-07), 15(1), R2.
Ginneken, B.V., 2017. Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. Radiol. Phys. Technol. 10 (1), 23–32.
Guramrit, S., Ricci, E.P., Moore, M.J., 2014. RIPiT-Seq: a high-throughput approach for footprinting RNA:protein complexes. Methods 65 (3), 320–332.

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., Liu, H., 2018. HLPI-Ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy. RNA Biol. 15 (6), 797–806.
Jungkamp, A.C., Stoeckius, M., Mecenas, D., Grün, D., Mastrobuoni, G., Kempa, S., Rajewsky, N., 2011. In vivo and transcriptome-wide identification of RNA binding protein target sites. Mol. Cell 44 (5), 828–840.
Langdon, W.B., Petke, J., Lorenz, R., 2018. Evolving better RNAfold structure prediction. Paper Presented at the European Conference on Genetic Programming. pp. 220–236.
Lewis, B.A., Walia, R.R., Terribilini, M., Ferguson, J., Zheng, C., Honavar, V., Dobbs, D., 2011. PRIDB: a Protein-RNA interface database. Nucleic Acids Res. 39 (Database issue):D277.
Liao, B., Luo, J., Li, R., Zhu, W., 2010. RNA secondary structure 2D graphical representation without degeneracy. Int. J. Quantum Chem. 106 (8), 1749–1755.
Liu, H., Ren, G., Hu, H., Zhang, L., Ai, H., Zhang, W., Zhao, Q., 2017. LPI-NRLMF: lncRNA-protein interaction prediction by neighborhood regularized logistic matrix factorization. Oncotarget 8 (61), 103975–103984.
Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X., Li, T., 2013. Computational prediction of associations between long non-coding RNAs and proteins. BMC Genomics 14 (1), 651.
Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., Sattar, A., Yang, Y., 2015. Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. J. Comput. Chem. 35 (28), 2040–2046.
Mariusz, P., Marta, S., Maciej, A., Purzycka, K.J., Piotr, L., Natalia, B., Adamiak, R.W., 2012. Automated 3D structure composition for large RNAs. Nucleic Acids Res. 40 (14), e112.
Muppirala, U.K., Honavar, V.G., Drena, D., 2011. Predicting RNA-Protein interactions using only sequence information. BMC Bioinformatics 12 (1), 489.
Narayanan, B.C., Westbrook, J., Ghosh, S., Petrov, A.I., Sweeney, B., Zirbel, C.L., Berman, H.M., 2014. The Nucleic Acid Database: new features and capabilities. Nucleic Acids Res. 42 (Database issue):114-122.
Pan, Fan., Yan, J., Shen, H.B., 2016. IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. BMC Genomics 17 (1), 582.
Pan, X., Rijnbeek, P., Yan, J., Shen, H.B., 2018. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genomics 19 (1), 511.
Pancaldi, V., Bähler, J., 2011. In silico characterization and prediction of global protein-mRNA interactions in yeast. Nucleic Acids Res. 39 (14), 5826–5836.
Pandey, C., Sandeep, R., Priyam, A., Mahapatra, S., Sahu, S.S., 2018. Predicting protein-RNA interaction using sequence derived features and machine learning approach. Int. J. Data Min. Bioinform. 19 (3), 270.
Pham, H.H., Khoudour, L., Crouzil, A., Zegers, P., Velastin, S.A., 2018. Exploiting deep residual networks for human action recognition from skeletal data. Comput. Vis. Image Underst. 170, 51–66.
Sahiner, B., Chan, H.P., Petrick, N., Wei, D., Helvie, M.A., Adler, D.D., Goodsitt, M.M., 1996. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans. Med. Imaging 15 (5), 598–610.
Suresh, V., Liu, L., Adjeroh, D., Zhou, X., 2015. RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. Nucleic Acids Res. 43 (3), 1370–1379.
Vera Pancaldi, J.B., 2011. In silico characterization and prediction of global protein–mRNA interactions in yeast. Nucleic Acids Res. 39 (14), 5826–5836.
Weyn-Vanhentenryck, S., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., Zhang, M., 2014. HITS-CLIP and integrative modeling define the rbfox splicing-regulatory network linked to brain development and autism. Cell Rep. 6 (6), 1139.
Wiese, K.C., Hendriks, A., 2006. Comparison of P-RnaPredict and mfold—algorithms for RNA secondary structure prediction. Bioinformatics 22 (8), 934–942.
Zhao, Q., Yu, H., Ming, Z., Hu, H., Ren, G., Liu, H., 2018a. The bipartite network projection-recommended algorithm for predicting long non-coding RNA-Protein interactions. Mol. Ther. Nucleic Acids 13, 464–471.
Zhao, Q., Zhang, Y., Hu, H., Ren, G., Zhang, W., Liu, H., 2018b. IRWNRLPI: integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-Protein interaction prediction. Front. Genet. 9, 239.