Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 19(5): 4749–4764.
DOI: 10.3934/mbe.2022222
Received: 27 November 2021
Revised: 28 February 2022
Accepted: 02 March 2022
Published: 11 March 2022

Research article

RLF-LPI: An ensemble learning framework using sequence information for predicting lncRNA-protein interaction based on AE-ResLSTM and fuzzy decision

Jinmiao Song[1,2], Shengwei Tian[3,4,5,*], Long Yu[1], Qimeng Yang[1], Qiguo Dai[2], Yuanxu Wang[2], Weidong Wu[6,*] and Xiaodong Duan[2]

[1] Department of Information Science and Engineering, Xinjiang University, Urumqi 830008, China

[2] Key Laboratory of Big Data Applied Technology, State Ethnic Affairs Commission, Dalian Minzu University, Dalian 116600, China

[3] Department of Software, Xinjiang University, Urumqi 830008, China

[4] Key Laboratory of Signal and Information Processing, Xinjiang University, Urumqi 830008, China

[5] Key Laboratory of Software Engineering Technology, Xinjiang University, Urumqi 830008, China

[6] Center for Science Education, People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi 830001, China

* **Correspondence:** Email: tianshengwei@163.com, xjwudong@126.com.

Abstract: Long non-coding RNAs (lncRNAs) play a regulatory role in many biological cells, and the recognition of lncRNA-protein interactions is helpful to reveal the functional mechanism of lncRNAs. Identification of lncRNA-protein interaction by biological techniques is costly and time-consuming. Here, an ensemble learning framework, RLF-LPI is proposed, to predict lncRNA-protein interactions. The RLF-LPI of the residual LSTM autoencoder module with fusion attention mechanism can extract the potential representation of features and capture the dependencies between sequences and structures by k-mer method. Finally, the relationship between lncRNA and protein is learned through the method of fuzzy decision. The experimental results show that the ACC of RLF-LPI is 0.912 on ATH948 dataset and 0.921 on ZEA22133 dataset. Thus, it is demonstrated that our proposed method performed better in predicting lncRNA-protein interaction than other methods.

Keywords: deep learning; lncRNA-protein interaction; fuzzy decision; extra trees; attention mechanism

1. Introduction

With the in-depth understanding of non-coding RNAs (ncRNAs), it is discovered that ncRNAs play an important part in many living activities such as cell cycle regulation, epigenetic regulation and cell differentiation, which makes it one of the research hotspots in the field of genetics [1], particularly lncRNA, with a length of more than 200nt. Studies have shown that the types and quantities of lncRNAs are far from the same with those of proteins. Only a few of the biological functions of lncRNAs have been revealed. Abnormal expressions of lncRNAs are closely related to various complex diseases, such as hepatocellular carcinoma, liver cancer, breast cancer, Alzheimer's disease, etc. [2–5]. Identification of lncRNA biological functions is helpful to improve the deep cognition of living activities. In recent years, exploring the interactions between lncRNAs and proteins is one of the main ways to infer the lncRNAs functions and to do further studies on lncRNAs. LncRNA-protein interactions (LPIs) play a vital part in protein synthesis, viral replication and transcriptional regulation. Thus, it is of significance to study lncRNA-protein interactions [6, 7]. At present, relevant studies are mainly being divided into experiments and computational methods, while traditional experimental methods can only conduct experimental studies on a pair of lncRNA and protein each time, which is time-consuming and costly [8]. The computational methods are being used to reveal the potential lncRNAs and proteins interactions increasingly [9, 10]. The computational methods of lncRNA-protein interaction recognition can be divided into two works: network-based methods and machine learning-based methods. The network-based LPIs prediction methods construct a heterogeneous lncRNA-protein network, which can capture the hidden feature information of the topology containing lncRNAs in the related biological heterogeneous network. Lu et al. [11] produced feature vector coding by sequences of lncRNAs and proteins, and scored each RNA-protein pair by matrix multiplication, which is further used to measure RNA-protein interactions. Zhang et al. [12] calculated the linear neighborhood similarity in the feature space and transferred it to the interaction space. Then, LPIs were predicted by a label propagation progress. Zhao et al. [13] searched for the potential interactions between lncRNAs and proteins through random walk and neighborhood regularized logistic matrix factorization. Zhu et al. [14] proposed a lncRNA-protein bipartite network based on ant colony clustering (ACCBN) to predict LPIs. Zhang et al. [15] calculated the similarity of lncRNAs and proteins by integrating lncRNA expression information and gene ontology information, and predicted all lncRNA-protein interactions by using graph regularized nonnegative matrix factorization framework. Zhang et al. [16] integrated protein semantic similarity, lncRNA functional similarity, known human LPIs and Gaussian interaction profile kernel similarity to predict LPIs using a heterogeneous graph and depth-first search algorithm. The network-based methods deliver LPI labels in heterogeneous graph and path propagation effectively, however, it is of less capability to predict interaction performance for isolated proteins or lncRNAs.

Machine learning-based methods combine the feature of sequences, structures and physicochemical properties to construct a classifier, forming an interactive or non-interactive classification model. Muppriala et al. [17] used k-mer features as SVM and random forest input to predict the associations of RNA and protein, which achieved high prediction accuracy in prediction. Wang et al. [18] proposed an extended naïve-Bayes-classifier method to extract the sequence features of proteins and ncRNAs. Then, based on likelihood ratio score, the method selected effective features and reduced data dimensions to predict LPIs more accurately. Pan et al. [19] proposed IPMiner to extract the sequence features of proteins and ncRNAs and used stack autoencoder and fine tuning to preprocess

data, and SDA-RF, SDA-FT-RF and IP-Miner methods were used to predict LPIs, respectively. The final results showed that IPMiner had the best performance and obtained high prediction accuracy in verification set. Peng et al. [20] developed an ensemble framework (LPI-EnEDT) with extra tree classifier and decision tree classifier to implement imbalanced LPIs data classification by integrating multiple biological features of lncRNAs and proteins. Deep learning-based methods are widely used in biological applications. Compared with other sequence-based methods, deep learning can learn the sequence features of RNAs and proteins automatically and discover specific interactions between these sequences. Peng et al. [21] proposed a hierarchical deep learning framework RPITER to predict RNA-protein interactions and input the combination of primary sequence information and structure information into convolutional neural network (CNN) and stacked auto-encoder (SAE), which could achieve good results. Wekesa et al. [22] proposed a graph representation learning method (GPLPI) to predict plant lncRNA-protein interactions from sequence and structural information. Then, a long and short-term memory encoder-decoder network with multiple self-attention was proposed to extract advanced features, and the majority voting mechanism was used to improve the prediction model to achieve classification performance [22, 23]. In addition, some deep learning-based methods introduce noise artificially to reduce overfitting, which further improve the generalization ability and robustness.

In this study, an ensemble learning model for predicting LPIs using sequence information and structure information, RLF-LPI, is proposed, which can determine whether there is an interaction between specific lncRNA-protein pair. This experiment extracts the sequence and structural features of lncRNA-protein pairs, and inputs the features into the residual LSTM autoencoder module of the fusion attention mechanism in different ways to extract the potential representation of the original feature information, which is then used as the input of the base classification. Finally, the ensemble learning module based on fuzzy decision is used to improve the prediction performance. The performance of RLF-LPI on lncRNA-protein datasets and other ncRNA-protein datasets are evaluated to compare with other existing methods. The results show that RLF-LP has higher predictive performance, better generalization ability and robustness.

2. Materials and methods

2.1. Datasets

Datasets of Arabidopsis thaliana ATH948 and Zea mays ZEA22133 are collected from published papers [24]. In this experiment, CD-HIT [25] algorithm is used to retain effective sequences with less than 10% sequence similarity between lncRNAs and proteins. During non-interaction pair construction, the same number of positive and negative sample sequences are generated by randomly pairing with lncRNA, and eliminating the existing interaction pair sequences. The ATH948 dataset includes 948 interaction pairs and 948 non-interaction pairs, consisting of 35 proteins and 109 lncRNAs. In addition, ZEA22133 dataset is obtained, including 22,133 interaction pairs and 22,133 non-interaction pairs, consisting of 42 proteins and 1704 lncRNAs. Futhermore, the experiment collects multiple ncRNA-protein datasets from other studies, such as RPI1807, RPI369, RPI2241 [24] and RPI488 [19], to verify the generalization ability of RLF-LPI, covering a wide range of species, including humans, animals and plants. The specific datasets for this experiment are shown in Table 1. For RNAs, the RNAfold program in the ViennaRNA package [29] is used to calculate the secondary structure information of RNA through the minimum free energy. For proteins, the SOPMA algorithm [30] is used to

predict the three-state structure of proteins, including α -helix, β -fold and helix.

Table 1. Experimental dataset.

Dataset	lncRNA	Protein	Interaction pair	Non-interaction pair
ATH948	109	35	948	948
ZEA22133	1704	42	22,133	22,133
RPI2241	842	2043	2241	2241
RPI369	332	338	369	369
RPI488	25	247	243	245
RPI1807	1078	1807	1807	1436

2.2. Methods

The sequence length of lncRNAs and proteins vary greatly in the dataset, during the process of extracting experimental features. Simple one-hot coding only encodes a single nucleotide or amino acid, which cannot make fully use of the long-term dependence between adjacent nucleotides or amino acids. Therefore, it is of great importance to select an appropriate sequence coding method. In this section, k-mer method is used to obtain sequence and structural features and to eliminate the above deficiencies. Firstly, in terms of sequence information, the corresponding frequency of nucleotide (AUGC) is calculated in the lncRNA sequence to fully extract features. Then, the 340-dimensional feature vector is obtained by taking the combination features of k= 1,2,3 and 4. For proteins, 20 kinds of amino acids need to be divided into seven categories according to the physical and chemical properties of amino acids, including: {Val, Gly, Ala}, {Phe, Pro, Leu, Ile}, {Ser, Tyr, Met, Thr}, {His, Asn, Tpr, Gln}, {Arg, Lys}, {Glu, Asp} and {Cys}. The coding frequency of each amino acid sequence is calculated based on the seven categories of amino acids, and the 399-dimensional feature vector is obtained by taking the combination features of k= 1,2 and 3. Finally, local features of different levels are extracted to enhance the information representation of global features by k-mer.For sequence structure information, protein and lncRNA secondary structure can enhance the expression of spatial information and sequence information, making the result more accurate. K-mer method can fully extract secondary structure features and process data. Protein and RNA secondary structure frequencies of 1-mer, 2-mer and 3-mer were extracted to supplement the sequence coding, therefore, the 39-dimensional protein structure features and the 399-dimensional lncRNA structure features were obtained.

2.3. Model design

In this paper, an ensemble learning framework based on RLF-LPI is proposed to predict lncRNA-protein interactions. Firstly, the effects of sequence and structural features on classification performance are taken into account. The sequence information of lncRNAs and proteins is extracted respectively and combined to obtain 739-dimension vector. Similarly, 438-dimension structure feature vector is also obtained. Secondly, the complementary relationship between sequence features and structural features is fully considered to obtain 1177 dimension combine-feature vector. The sequence feature representation, structural feature representation and combined feature representation are obtained by inputting three types of features into AE-ResLSTM. Sequence feature representation and structure feature representation are fused to obtain enhanced feature representation. As shown in Figure 1, the

original combine-feature, combine-feature depth representation and depth enhancement feature representation are input into the base learner respectively to improve feature representation ability, and three scores are obtained. Finally, the scores are taken as the fuzzy decision inputs, and the greedy algorithm is used to optimize the parameters in the fuzzy decision, so that the given lncRNA-protein pair can be predicted more efficiently. In AE-ResLSTM module, the deep encode-decode structure is used to solve the long-term dependence between sequences and to learn potential feature information through LSTM. Meanwhile, residual structure and attention mechanism are used in the decoder to calculate feature weights.

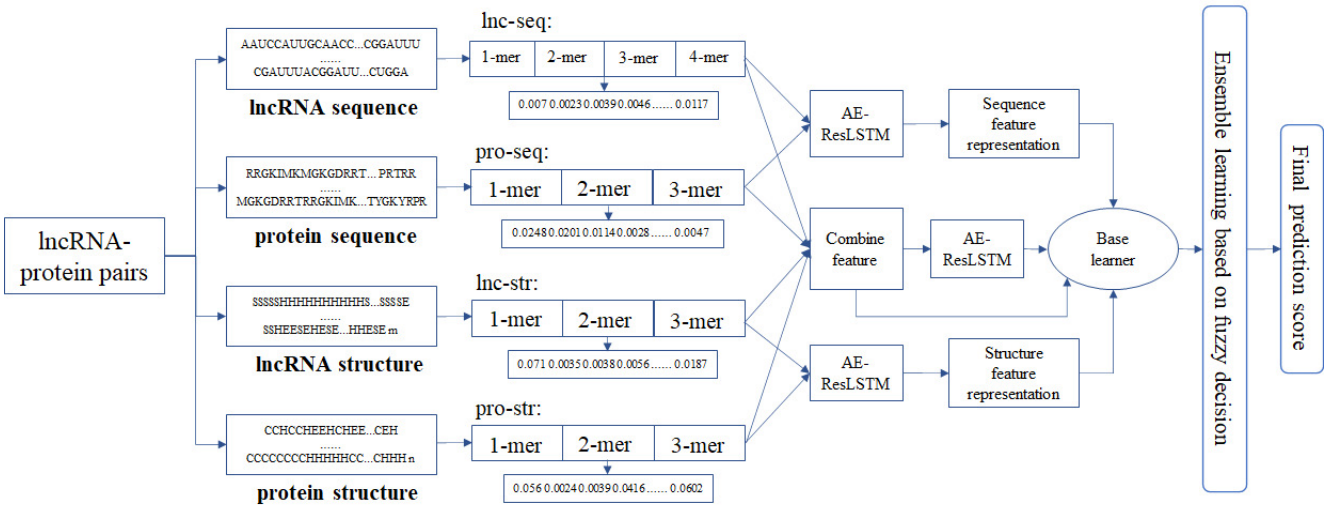

Figure 1. The framework of the RLF-LPI. The sequence information features and structural information features are extracted by k-mer, and the high-level abstract representation is learned by using the AE-ResLSTM module. Finally, the ensemble learning based on fuzzy decision is used to obtain the prediction result.

2.3.1. AE-ResLSTM with attention module

Autoencoder is a data compression algorithm, belonging to unsupervised learning, which can learn the implied features of input data and reconstruct the original input data [26]. In this paper, the structure of AE-ResLSTM network is proposed, which is divided into an encoder and a decoder, as shown in Figure 2, to depict the details of the architecture. The whole feature information is learned by encoder and represented by fixed-length vectors. Those fixed-length vectors, as the input to the decoder, are interpreted as feature representations.

The first part of the module is the encoder. LSTM can process sequence structure and learn the long-term dependence between RNA/protein sequence structure modules. Encoder structure can learn the potential representation of given data. The potential feature information of sequence structure data can be learned by deep learning model mentioned above. The second part is a stacked residual LSTM decoder, which is composed of multiple residual LSTM network layers and attention mechanism modules. Stacked LSTM layers can enrich the expression of the model, but it may also lead to gradient explosion or gradient disappearance. Here, skip connections between encode-decode layers are introduced, which fuse the encoder information to the decoder at the corresponding level, to improve

the sensitivity of optimization gradient and the learning process of the nonlinear neural layer, to prevent possible gradient problems in the model, to enhance effective information, and to enable the back propagation to obtain better feature expression. Meanwhile, the information of the residual module outputs to the internal attention layer to calculate the weights of each feature, to weaken the redundant information and to improve the expression ability of the model.

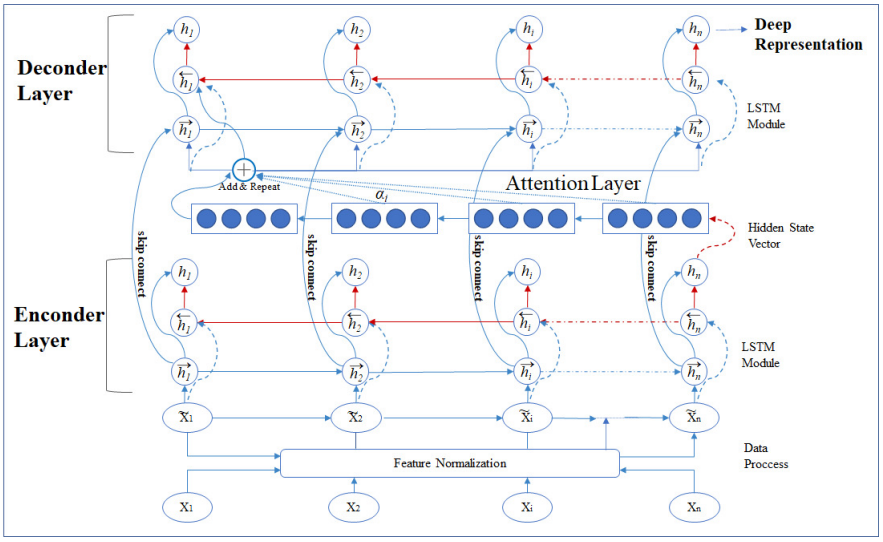

Figure 2. The structure of AE-ResLSTM module.

Equations (2.1)–(2.6) represent the calculation of LSTM units. Each unit consists of three types of gates, including forget, input and output gates. Where, H_{t-1} represents the previous state, and x_t represents the current input.

$$h_t = LSTM(v_t) = o_t \odot tanh(C_t) \tag{2.1}$$

$$f_t = \sigma(W_f(H_{t-1} + x_t) + b_f) \tag{2.2}$$

$$l_t = \sigma(W_l(H_{t-1} + x_t) + b_l) \tag{2.3}$$

$$\widetilde{C}_t = tanh(W_C(H_{t-1} + x_t) + b_C) \tag{2.4}$$

$$C_t = f_t * C_{t-1} + l_t * \widetilde{C}_t \tag{2.5}$$

$$o_t = \sigma(W_o(H_{t-1} + x_t) + b_o) \tag{2.6}$$

The symbols $\odot$ represent multiplication calculations, W represents different weight matrices, b represents bias value, and σ represents sigmoid function. Equation (2.7) represents residual calculation. LSTM is combined with a residual network and outputted to base classifier.

$$H_t = ResLSTM(v_t) = [LSTM(v_t) + v_t] \tag{2.7}$$

The attention layer assigns different weights to each output layer of a decoder. Attention was first proposed by Yang et al. [27], who proposed a hierarchical attention network for document classification. The performance of the model in this paper is improved by adding an attention mechanism to calculate the weights of each feature to enhance the key features. The attention mechanism is described by the following Eqs (2.7)–(2.10):

$$m_i = tanh(W_w h_i + b_w) \tag{2.8}$$

$$\alpha = \frac{exp(m_i^T m_w)}{\sum_i exp(m_i^T m_w)} \tag{2.9}$$

$$z = \sum_i \alpha_i h_i \tag{2.10}$$

where m_i is the hidden representation of the i_{th} feature and h_i comes from the residual output of ResLSTM decoder. The significance of this feature is measured by the similarity between m_i and the context vector m_w. Multiply all the eigenvectors by their corresponding weights and sum to the final output vector z. W_w, b_w, and m_x are initialized randomly, and they are being learned during training.

2.3.2. Extra trees classifier

Extra Trees (Extremely randomized trees, ET) were proposed by Pierre Geurts et al. in 2006. ET algorithm is similar to Random Forest (RF), which is composed of decision trees. However, RF uses Bagging model, and ET uses all of the samples (features were chosen randomly). To some extent, the generalization ability of ET is higher than that of RF, and ET is different from the RF in a random subset of the optimal bifurcation properties. Therefore, Extra Trees Classifier model is used to be the base learner here.

2.3.3. Fuzzy decision integration and fusion

The prediction of lncRNA-protein interactions is a binary classification problem. The residual LSTM autoencoder module with attention mechanism is used to obtain the sequence potential information, and the Extra Trees Classifier base classifier is used to predict whether there is an interaction between lncRNA and protein. In this paper, the method in reference [28] is introduced. Three Extra Trees Classifier based classification models are used, and three Extra Trees outputs are taken as decisions, respectively. The greedy algorithm is used to judge whether fuzzy rules are satisfied, and then the final classification results are obtained. In this experiment, the base classification model uses Extra Trees Classifier, which has different data training methods, so the final decision selection is complicated. Using the greedy algorithm index as the criterion for the result of the base classifier model, the formula is as follows:

$$G = abs(2cp - 1), cp \in [0, 1] \tag{2.11}$$

The G is the greed index, 'abs' is the absolute value function, and cp is the probability of confounding. It can be seen from Eq (2.11) that G and cp are in direct proportion, which shows that the Extra Trees Classifier can determine whether there is an interaction in the unlabeled sample, if $cp \geq 0.5$, it means there is an interaction, otherwise there is no interaction. The larger the greed index is, the higher the probability of confounding is, and the output result of the base classifier model is optimal.

2.4. Implementation of predictors

The experimental code is implemented in the Keras 2.1.6 environment and is written based on python3.6.9. The ExtraTreesClassifier module is implemented by sklearn in Python. The specific experimental system is configured with 2.81GHz CPU, 6GB GPU and 8GB memory under the Windows10 operating system. The Adam optimizer is used for fast convergence of the module, and the Mean Square Error (MSE) loss function is used. Furthermore, the early stopping strategy [31] and

dropout [32] are used to avoid overfitting in the training process, setting dropout to 0.5, batchsize to 256, and epoch to 100.

3. Results and discussion

3.1. Performance evaluation metrics

For the experimental model, six indicators are evaluated, including accuracy (ACC), precision (Pre), sensitivity (Sn), specificity (Sp), Matthews Correlation Coefficient (MCC) and Area Under the Curve (AUC).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

$$Pre = \frac{TP}{TP + FP} \tag{3.2}$$

$$Sn = \frac{TP}{TP + FN} \tag{3.3}$$

$$Sp = \frac{TN}{TN + FP} \tag{3.4}$$

$$MCC = \frac{TP \times TN - TP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.5}$$

where TP, FP, TN, FN represent true positive, false positive, true negative and false negative.

3.2. Results

RLF-LPI performance is evaluated using two datasets. Table 2 shows the results of the five-fold cross-validation. On the ATH948 dataset, the ACC of the method is 91.2%, the precision is 95.6%, and the AUC is 95.3%. On the ZEA22133 dataset, the ACC of the method is 92.1%, the precision is 91.9%, and the AUC is 98.0%. According to the result analysis, the accuracy of ZEA22133 dataset is not as good as ATH948. One possible reason is that the amount of ZEA22133 data is too large. Overall, those information used in our model produced good results for prediction.

Table 2. Performance of dataset on RLF-LPI (%).

Dataset	ACC	Pre	Sn	Sp	MCC	AUC
ATH948	91.2	95.6	86.4	96.0	71.8	95.3
ZEA22133	92.1	91.9	92.5	91.8	78	98.0

3.3. Comparison with other methods

On the experimental dataset of this paper, RLF-LPI is compared with five LPI prediction models based on other sequence methods, including LPI-DL [33], PLRPI [24], IPMiner [19], RPISeq [17] and lncPro [11]. As shown in Table 3, RLF-LPI achieves the best performance on the ATH948 and ZEA22133 datasets. On the ZEA22133 dataset, the accuracy is 1.4% higher and the AUC is 1.0% higher than that of LPI-DL. Compared to other LPI prediction methods, RLF-LPI performs better on other indicators and has certain advantages compared with IPMiner, RPISeq-RF and lncPro. This

indicates that the performance of our proposed RLF-LPI model in predicting LPIs reaches the expected standard, and can effectively extract high-level representations of features, making it better to be used in predicting models.

Table 3. Performance in dataset compared with other methods (%).

Dataset	Method	ACC	Pre	Sn	Sp	MCC	AUC
ATH948	RLF-LPI	**91.2**	**95.6**	86.4	**96.0**	71.8	**95.3**
	LPI-DL	88.1	90.4	—	—	77.7	94.9
	PLRPI	90.4	92.8	**87.6**	93.2	**81.1**	—
	IPMiner	88.2	89.2	86.9	89.5	76.5	94.1
	RPISeq-RF	75.6	76.2	75.2	73.0	79.4	90.2
	IncPro	75.4	76.9	75.4	74.7	71.5	89.2
ZEA22133	RLF-LPI	**92.1**	91.9	**92.5**	91.8	78.0	**98.0**
	LPI-DL	90.7	91.5	—	—	**81.5**	97.0
	PLRPI	82.6	**99.9**	67.5	**99.6**	69.6	—
	IPMiner	68.7	69.6	66.5	70.9	37.5	84.6
	RPISeq-RF	65.4	64.1	62.5	70.3	35.9	81.4
	IncPro	60.3	61.3	60.8	69.6	30.9	80.8

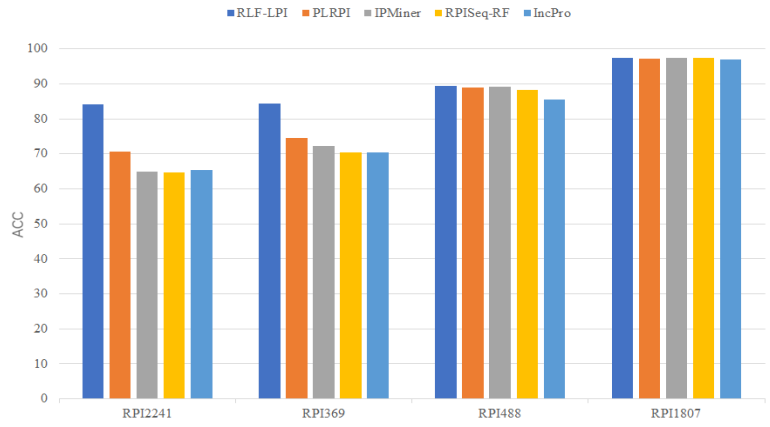

Figure 3. Comparison of Accuracy in different methods of the open dataset.

As shown in Table 3, RLF-LPI outperforms other models on the ATH948 and ZEA22133 datasets. On one hand, in order to enhance the input of the model, the structural features of lncRNAs and proteins are used as supplementary information of sequence features, which enable the model to have stronger expression ability. Meanwhile, different k-mer combinations are also adopted to enrich the feature information. On the other hand, the model uses the LSTM residual network structure with attention mechanism, which further enhances the feature representation ability. In addition, extra trees are used as the base learner to improve performance, which can provide additional randomness, inhibit over-fitting, and accelerate training speed. Finally, fuzzy decision is introduced to select the optimal object by greedy algorithm and optimize the prediction performance of the model.

3.4. Testing the robustness of RLF-LPI

We compared this with other published papers on the interactions between ncRNAs and proteins based on other methods, with the aim of testing the robustness of RLF-LPI. As shown in Table 4 and Figure 3, it can be found that RLF-LPI has higher performance on datasets RPI2241, RPI369 and RPI488. On RPI1807 dataset, RLF-LPT, IPMiner and RPISeq-RF achieve similar accuracy and are all superior to lncPro, which show that the proposed method has strong robustness.

Table 4. Performance in dataset compared with other methods (%).

Dataset	Method	ACC	Pre	Sn	Sp	MCC	AUC
RPI2241	RLF-LPI	84.2	87.4	80.0	**88.4**	54.9	**92.4**
	PLRPI	70.7	72.9	65.9	75.5	41.7	—
	IPMiner	**86.1**	**88.2**	**87.7**	84.1	**72.4**	90.6
	RPISeq-RF	85.0	86.3	86.1	83.8	70.7	69.0
	IncPro	61.6	66.9	52.9	69.5	31.0	72.2
RPI369	RLF-LPI	**84.4**	**81.9**	**88.2**	**80.6**	**61.0**	**90.5**
	PLRPI	74.5	73.3	77.2	71.8	49.2	—
	IPMiner	70.3	72.4	72.3	72.3	42.8	77.3
	RPISeq-RF	69.4	70.7	70.5	70.2	40.6	76.7
	IncPro	50.4	71.3	70.8	69.6	40.9	74.0
RPI488	RLF-LPI	**89.3**	**94.5**	83.5	**95.1**	66.4	91.4
	PLRPI	89.0	93.9	83.3	94.6	78.5	—
	IPMiner	89.1	93.5	**84.0**	94.4	**78.8**	91.4
	RPISeq-RF	88.3	93.5	82.8	83.6	77.2	88.3
	IncPro	85.6	94.1	77.6	94.0	72.5	**92.9**
RPI1807	RLF-LPI	**97.3**	96.8	**98.4**	95.9	92.7	98.7
	PLRPI	97.2	**97.2**	98.2	96.5	**94.3**	—
	IPMiner	96.8	95.5	96.5	96.5	93.5	**99.8**
	RPISeq-RF	97.0	96.2	97.0	**97.6**	93.8	99.6
	IncPro	56.9	55.5	56.5	58.1	43.8	99.4

RLF-LPI shows better performance on the public datasets because of the use of the AE-ResLSTM module. Since the previous autoencoder learn automatically, the error will continue to decrease. If the amount of data is relatively small, the gradient will disappear. Using an encode-decode structure and LSTM can fully learn features data of potential information. The stacked residual LSTM with attention mechanism is used in the decoder, where the depth of the residual network by increasing fairly can improve accuracy. Its internal residual block uses the skip connects, alleviating the problem of gradient disappearance caused by increasing depth in deep neural networks, and better feature expression can be obtained. The above proves that RLF-LPI is well adapted to the problem of lncRNA-protein interactions.The ROC curves of the dataset in this paper are shown in the following Figure 4.

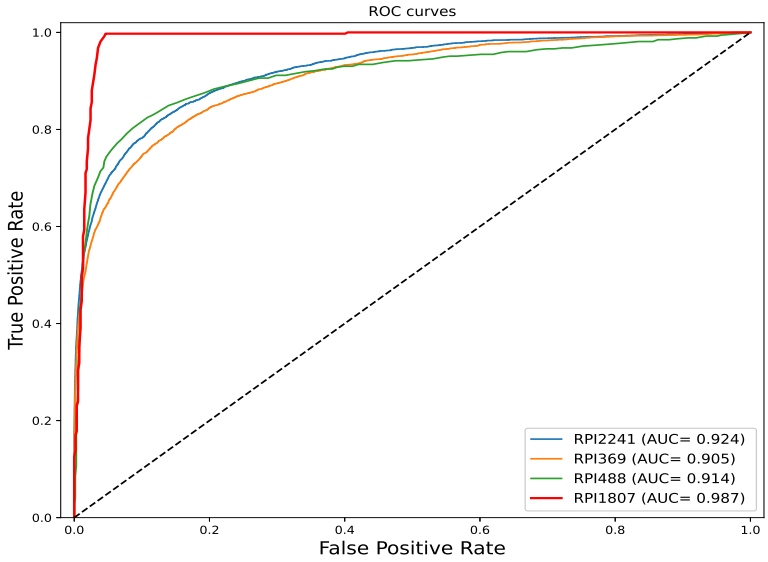

Figure 4. ROC curves for different datasets.

3.5. Comparing with different base classifiers

Five classical machine learning algorithms are tested on ZEA22133 dataset, including Extra Trees Classifier(ETC), Logistic regression (LR), Random Forest (RF), Gradient Boosting Decison Tree (GBDT) and Decision Tree (DT). As shown in Table 5 and Figure 5, RLF-LPI with ETC is significantly superior to other methods in all evaluation indicators. The values in the table represent the average values (%) under the five-fold cross-validation, and the bold values represent the best values in the dataset obtained by different methods. The average accuracy of this model is 3.2% higher and the AUC is 1.7% higher than that of other methods, which show that our approach is effective.

Table 5. Comparison of ZEA22133 on different base classifiers (%).

	ACC	Pre	Sn	Sp	MCC	AUC
Our	**92.1**	**91.9**	**92.5**	**91.8**	**78.0**	**98.0**
LR	88.9	89.1	88.8	89.1	69.2	96.3
RF	83.3	81.4	86.6	80.1	58.0	90.4
GBDT	88.5	87.9	89.2	87.7	68.7	96.0
DT	85.7	85.6	85.9	85.5	61.4	85.7

3.6. Ablation experiments

As shown in Table 6, RLF-LPI has a significant effect on predicting lncRNA-protein interactions. Among them, AE-ResLSTM network structure plays an obvious role in the whole experiment process. The network structure mainly consists of three parts, including the attention mechanisms and residual network structures in the LSTM encode-decode structure, as well as fuzzy decision. A series of ablation experiments are conducted to study the performance of partial network structures on the overall model

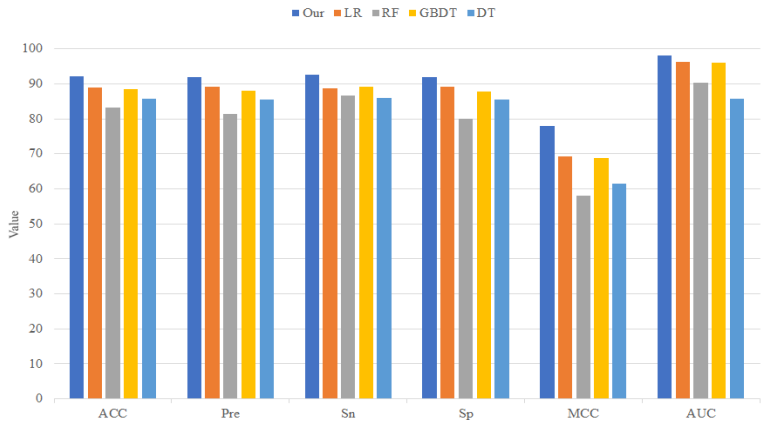

Figure 5. Performance indicators under different base classifiers.

and explore the generalization ability. Table 6 shows the ablation experiment results of the model under the five-fold cross-validation. It can be seen that ACC and Pre are 3.8% and 3.1% lower than our model without the residual structures, and can not reach the performance of RLF-LPI after eliminating attention and fuzzy decision. The experimental results show that residual structures can increase the depth of the model and improve the performance, and the expression ability of the model can be effectively enhanced by using attention mechanisms and fuzzy decision.

Table 6. Performance of dataset on RLF-LPI (%).

Attention	Residual	Fuzzy Decision	ACC	Pre	MCC
✓	✓	✓	**92.1**	**91.9**	**78.0**
×	✓	✓	91.2	90.7	77.8
✓	×	✓	88.3	88.8	67.2
✓	✓	×	91.1	90.2	75.0

4. Discussion and conclusions

The reason for the success of RLF-LPI may come from the following factors. First, taking sequence and structural coding parts as input, RLF-LPI fully exploits the dependencies on nucleotides and amino acids in lncRNAs and proteins, which enhance the expression ability of features. Second, feature extraction methods with different k-mer combinations are applied to lncRNAs and proteins, and the comprehensive prediction results were generated by AE-ResLSTM. The deep feature information of lncRNA-protein pairs is effectively captured by the combination of residual structure and LSTM. Finally, the fuzzy decision can better divide the decision boundary, effectively reduce the prediction error, and the overall strategy improves the performance of the overall model. As what shown in Table 7, our model has good performance on most datasets, but the AUC values are slightly lower than other models on individual datasets. For example, our AUC value on the NPInter v2.0 dataset is slightly lower than that with IPMiner method, mainly due to the disadvantage of parallel processing of the LSTM used in the encoder network. Compared with some state-of-the-art networks, although the gradient problem in shorter sequences can be solved to a certain extent, it is still very difficult

for longer sequences. As in the calculation process, LSTM combined with residual structure is used, which increase the parameters and deepen the network, the computational amount is increased and it takes longer. Our AUC values in the PRI1488 and RPI2241 datasets are slightly lower than that of RPI-SAN. This is because the RPI-SAN model extracts pseudo-Zernike moment (PZM) features using protein position-specific matrices, resulting in a more robust effect.

Table 7. Comparison of multiple datasets on different methods (%).

Method	Inputs	Classifier	Database	Performance(AUC)
IPMiner [19]	Sequences	Stacked Autoencoder + RF	NPInter v2.0 / PRI1488 / RPI2241	**99.5**/91.4/90.6
HLPI-Ensemble [10]	Sequences	SVM + RF + XGB	NPInter v2.0	96.0
RPI-SAN [34]	Sequences + Conservation	Stacked Autoencoder + RF	RPI1488/ RPI2241	**92.0**/**96.2**
BGFE [35]	Sequences	Stacked Autoencoder + RF	RPI1488/ RPI2241	89.8/94.7
RPI-SE [36]	Sequences	XGBoost + SVM + Extra Tree	RPI1488	90.4
LPI-BNPRA [37]	Sequences	Bipartite Network	NPInter V2.0	87.5
RLF-LPI	Sequences + Structures	AE-ResLSTM + Extra Tree	NPInter v2.0 / PRI1488 / RPI2241	98.1/91.4/92.4

In conclusion, compared with previous models, RLF-LPI has good robustness in predicting lncRNA-protein interactions. Of course, RLF-LPI has some limitations that need to be improved in the future. For example, known lncRNA-protein interaction data is still insufficient. Limited by the influence of datasets, there are still great challenges in the acquisition of negative samples and the selection of parameters. More lncRNA-protein interaction data can further improve the performance of RLF-LPI. All in all, as deep learning continues to develop, the problems raised will be solved. We look forward to future applications of RLF-LPI in more complex biological networks and to playing an active role in other biomolecular prediction methods.

Data availability

Datasets and code are available for download at https://github.com/JinmiaoS/LPI.

Acknowledgments

This research is supported by Scientific Research Fund Project of the Education Department of Liaoning Province (No.LJKZ0028), Dalian Young Science and Technology Star Project (No.2020RQ059) and the National Natural Science Foundation of China (No.61701073). We thank Jun Meng and her colleagues for providing very valuable help and guidance for our research.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. D. Guan, W. Zhang, G. H. Liu, J. C. Belmonte, Switching cell fate, ncRNAs coming to play, *Cell Death Dis.*, **4** (2013), e464. https://doi.org/10.1038/cddis.2012.196
2. J. J. Quinn, H. Y. Chang, Unique features of long non-coding RNA biogenesis and function, *Nat. Rev. Genet.*, **17** (2016), 47–62. https://doi.org/10.1038/nrg.2015.10
3. K. Panzitt, M. M. O. Tschernatsch, C. Guelly, T. Moustafa, M. Stradner, H. M. Strohmaier, et al., Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA, *Gastroenterology*, **132** (2007), 330–342. https://doi.org/10.1053/j.gastro.2006.08.026
4. J. Wang, X. Liu, H. Wu, P. Ni, Z. Gu, Y. Qiao, et al., CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer, *Nucleic Acids Res.*, **38** (2010), 5366–5383. https://doi.org/10.1093/nar/gkq285
5. A. C. Kaushik, A. Mehmood, X. Wang, D. Q. Wei, X. Dai, Globally ncrnas expression profiling of tnbc and screening of functional lncrna, *Front. Bioeng. Biotechnol.*, **8** (2021), 1480. https://doi.org/10.3389/fbioe.2020.523127
6. X. Pan, P. Rijnbeek, J. Yan, H. B. Shen, Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks, *BMC Genomics*, **19** (2018). https://doi.org/10.1186/s12864-018-4889-1
7. D. Adjeroh, M. Allaga, J. Tan, J. Lin, Y. Jiang, A. Abbasi, et al., Feature-based and string-based models for predicting RNA-protein interaction, *Molecules*, **23** (2018), 697. https://doi.org/10.3390/molecules23030697
8. S. W. Zhang, X. N. Fan, Computational methods for predicting ncRNA-protein interactions, *Med. Chem.*, **13** (2017), 515–525. https://doi.org/10.2174/1573406413666170510102405
9. L. Peng, F. Liu, J. Yang, X. Liu, Y. Meng, X. Deng, et al., Probing lncRNA–protein interactions: data repositories, models, and algorithms, *Front. Genet.*, (2020), 1346. https://doi.org/10.3389/fgene.2019.01346
10. H. Hu, L. Zhang, H. Ai, H. Zhang, Y. Fan, Q. Zhao, H Liu, et al., HLPI-Ensemble: Prediction of human lncRNA-protein interactions based on ensemble strategy, *RNA Biol.*, **15** (2018), 797–806. https://doi.org/10.1080/15476286.2018.1457935
11. Q. Lu, S. Ren, M. Lu, Y. Zhang, D. Zhu, X. Zhang, et al., Computational prediction of associations between long non-coding RNAs and proteins, *BMC Genomics*, **14** (2013). https://doi.org/10.1186/1471-2164-14-651
12. W. Zhang, Q. Qu, Y. Zhang, W. Wang, The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions, *Neurocomputing*, **273** (2018), 526–534. https://doi.org/10.1016/j.neucom.2017.07.065

13. Q. Zhao, Y. Zhang, H. Hu, G. Ren, W. Zhang, H. Liu, IRWNRLPI: integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction, *Front. Genet.*, **9** (2018), 239. https://doi.org/10.3389/fgene.2018.00239

14. R. Zhu, G. Li, J. X. Liu, L. Y. Dai, Y. Guo, ACCBN: Ant-Colony-clustering-based bipartite network method for predicting long non-coding RNA-protein interactions, *BMC Bioinf.*, **20** (2019). https://doi.org/10.1186/s12859-018-2586-3

15. T. Zhang, M. Wang, J. Xi, A. Li, LPGNMF: predicting long non-coding RNA and protein interaction using graph regularized nonnegative matrix factorization, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **17** (2018), 189–197. https://doi.org/10.1109/TCBB.2018.2861009

16. H. Zhang, Z. Ming, C. Fan, Q. Zhao, H. Liu, A path-based computational model for long non-coding RNA-protein interaction prediction, *Genomics*, **112** (2020), 1754–1760. https://doi.org/10.1016/j.ygeno.2019.09.018

17. U. K. Muppirala, V. G. Honavar, D. Dobbs, Predicting RNA-protein interactions using only sequence information, *BMC Bioinf.*, **12** (2011). https://doi.org/10.1186/1471-2105-12-489

18. Y. Wang, X. Chen, Z. P. Liu, Q. Huang, Y. Wang, D. Xu, et al., De novo prediction of RNA-protein interactions from sequence information, *Mol. Biosyst.*, **9** (2013), 133–142. https://doi.org/10.1039/C2MB25292A

19. X. Pan, Y. X. Fan, J. Yan, H. B. Shen, IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction, *BMC Genomics*, **17** (2016), 582. https://doi.org/10.1186/s12864-016-2931-8

20. L. Peng, R. Yuan, L. Shen, P. Gao, L. Zhou, LPI-EnEDT: an ensemble framework with extra tree and decision tree classifiers for imbalanced lncRNA-protein interaction data classification, *Biodata Min.*, **14** (2021), 50. https://orcid.org/0000-0002-2321-3901

21. C. Peng, S. Han, H. Zhang, Y. Li, RPITER: a hierarchical deep learning framework for ncRNA–protein interaction prediction, *Int. J. Mol. Sci.*, **20** (2019), 1070. https://doi.org/10.3390/ijms20051070

22. J. S. Wekesa, J. Meng, Y. Luan, A deep learning model for plant lncRNA-protein interaction prediction with graph attention, *Mol. Genet. Genomics*, **295** (2020), 1091–1102. https://doi.org/10.1007/s00438-020-01682-w

23. J. S. Wekesa, J. Meng, Y. Luan, Multi-feature fusion for deep learning to predict plant lncRNA-protein interaction, *Genomics*, **112** (2020), 2928–2936. https://doi.org/10.1016/j.ygeno.2020.05.005

24. H. Zhou, Y. Luan, J. S. Wekesa, J. Meng, Prediction of plant lncRNA-protein interactions using sequence information based on deep learning, in *International Conference on Intelligent Computing*, (2019), 358–368. https://doi.org/10.1007/978-3-030-26766-7_33

25. Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, *Bioinformatics*, **26** (2010), 680–682. https://doi.org/10.1093/bioinformatics/btq003

26. I. Goodfellow, Y. Bengio, A. Courville, Regularization for deep learning, *Deep learn.*, (2016), 216–261.

27. Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, (2016), 1480–1489. https://doi.org/10.18653/v1/N16-1174

28. Q. Kang, J. Meng, J. Cui, Y. Luan, M. Chen, PmliPred: a method based on hybrid model and fuzzy decision for plant miRNA–lncRNA interaction prediction, *Bioinformatics*, **36** (2020), 2986–2992. https://doi.org/10.1093/bioinformatics/btaa074

29. R. Lorenz, S. H. Bernhart, C. H. Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, et al., ViennaRNA Package 2.0, *Algorithms Mol. Biol.*, **6** (2011). https://doi.org/10.1186/1748-7188-6-26

30. C. Geourjon, G. Deleage, SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments, *Bioinformatics*, **11** (1995), 681–684. https://doi.org/10.1093/bioinformatics/11.6.681

31. G. Montavon, G. Orr, K. R. Müller, *Neural Networks: Tricks of the Trade*, springer, 2012.

32. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.*, **15** (2014), 1929–1958.

33. J. S. Wekesa, Y. Luan, J. Meng, LPI-DL: A recurrent deep learning model for plant lncRNA-protein interaction and function prediction with feature optimization, in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (2020), 499–502. https://doi.org/10.1109/BIBM49941.2020.9313431

34. H. C. Yi, Z. H. You, D. S. Huang, X. Li, T. H. Jiang, L. P. Li, A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information, *Mol. Ther.-Nucleic Acids*, **11** (2019), 337–344. https://doi.org/10.1016/j.omtn.2018.03.001

35. Z. H. Zhan, L. N. Jia, Y. Zhou, L. P. Li, H. C. Yi, BGFE: a deep learning model for ncRNA-protein interaction predictions based on improved sequence information, *Int. J. Mol. Sci.*, **20** (2019), 978. https://doi.org/10.3390/ijms20040978

36. H. C. Yi, Z. H. You, M. N. Wang, Z. H. Guo, Y. B. Wang, J. R. Zhou, RPI-SE: a stacking ensemble learning framework for ncRNA-protein interactions prediction using sequence information, *BMC Bioinf.*, **21** (2020), 60. https://doi.org/10.1186/s12859-020-3406-0

37. Q. Zhao, H. Yu, Z. Ming, H. Hu, G. Ren, H. Liu, The bipartite network projection-recommended algorithm for predicting long non-coding RNA-protein interactions, *Mol. Ther.-Nucleic Acids*, **13** (2018), 464–471. https://doi.org/10.1016/j.omtn.2018.09.020

AIMS Press