frontiers in Genetics

TECHNOLOGY REPORT
published: 08 October 2018
doi: 10.3389/fgene.2018.00458

Accurate Prediction of ncRNA-Protein Interactions From the Integration of Sequence and Evolutionary Information

Zhao-Hui Zhan[1], *Zhu-Hong You*[2]*, *Li-Ping Li*[2], *Yong Zhou*[1] *and Hai-Cheng Yi*[2]

[1] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China, [2] Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Ürümqi, China

Non-coding RNA (ncRNA) plays a crucial role in numerous biological processes including gene expression and post-transcriptional gene regulation. The biological function of ncRNA is mostly realized by binding with related proteins. Therefore, an accurate understanding of interactions between ncRNA and protein has a significant impact on current biological research. The major challenge at this stage is the waste of a great deal of redundant time and resource consumed on classification in traditional interaction pattern prediction methods. Fortunately, an efficient classifier named LightGBM can solve this difficulty of long time consumption. In this study, we employed LightGBM as the integrated classifier and proposed a novel computational model for predicting ncRNA and protein interactions. More specifically, the pseudo-Zernike Moments and singular value decomposition algorithm are employed to extract the discriminative features from protein and ncRNA sequences. On four widely used datasets RPI369, RPI488, RPI1807, and RPI2241, we evaluated the performance of LGBM and obtained an superior performance with AUC of 0.799, 0.914, 0.989, and 0.762, respectively. The experimental results of 10-fold cross-validation shown that the proposed method performs much better than existing methods in predicting ncRNA-protein interaction patterns, which could be used as a useful tool in proteomics research.

Keywords: ncRNA-protein interactions, PSSM, LightGBM, Pseudo-Zernike moments, k-mers

OPEN ACCESS

Edited by:
Quan Zou,
Tianjin University, China

Reviewed by:
Kang Wei,
The Chinese University of Hong Kong, Hong Kong
Dongya Jia,
Rice University, United States

***Correspondence:**
Zhu-Hong You
zhuhongyou@ms.xjb.ac.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *07 August 2018*
Accepted: *19 September 2018*
Published: *08 October 2018*

Citation:
Zhan Z-H, You Z-H, Li L-P, Zhou Y and Yi H-C (2018) Accurate Prediction of ncRNA-Protein Interactions From the Integration of Sequence and Evolutionary Information. Front. Genet. 9:458. doi: 10.3389/fgene.2018.00458

INTRODUCTION

Non-coding RNAs (ncRNAs) are regarded as the "dark matter" in the genome because of their inability in coding proteins. These years, a variety of ncRNA has been discovered by researchers which plays an indispensable role in most processes of vital movements in the field of biology including amino acids transporting, RNA modification and so on (Pan et al., 2016). According to recent research on ncRNA, ncRNA has been proved to be inextricably associated with human diseases and even cancer. For instance, Tian Y et al. have demonstrated that the role of ncRNA in diabetes is emerging significantly since ncRNA is involved in the modulation of 0205 cell mass, insulin synthesis, secretion and signaling (Tian et al., 2018). However, compared to those ncRNAs with known functions in vital processes occurring in living organisms, there is still a large part of ncRNAs whose functions are not yet clear. In order to gain insight into the function of ncRNA, it is essential to determine whether these ncRNAs interact with other proteins which

subserve the comprehension of the mechanism behind biological processes involving RNA-Binding proteins(RBPs) (Li and Nagy, 2011). Although reliable models in predicting ncRNA and protein were composed by a large number of experimental analyses such as RBPs (Pan et al., 2017), RPI-Bind (Luo et al., 2017), RNA Compete-S (Cook et al., 2017), there is still a limited number of structural features available in the protein data bank(PDB) about RNA-protein complexes causing these experiments were time-consuming and resource-consuming (Berman et al., 2000) Therefore, researchers focused their attention on predicting interactions between ncRNA and protein only based on sequences which was regarded as a reliable computational approach since the sequences carried enough information required for prediction (Suresh et al., 2015). This sequence-based method can be used to identify potential ncRNA and protein partners in the absence of their structural information during the experiment (Muppirala et al., 2011).

Machine learning provides researchers one of the most cost-effective ways to construct predictive models in an experimental environment where validated training data is available (Muppirala et al., 2011). In Mohammad et al.'s article, they collected motif information and repetitive patterns extracted from validated interactions between RNA and protein with the combination of sequence composition as descriptors to build a RPI prediction model called rpiCOOL by using a random forest classifier (Akbaripour-Elahabad et al., 2016). The random forest classifier is an ensemble of decision trees of which each tree is constructed through training a subset of features that are sampled from the input feature sets randomly (Akbaripour-Elahabad et al., 2016). And in Wang Ying et al.'s article, they proposed a new ncRNA-protein interaction model extended Bayesian classifier which selected valid features by reducing likelihood scores and allowed transparent feature integration during prediction (Wang et al., 2013). After feature extraction, the extracted features were sent to Bayesian classifier for training. Bayesian classifier is one of the most basic statistical classification methods which principle is to calculate the posterior probability of an object by using Bayesian formula, and select the class with the maximum posterior probability as the class to which the object belongs (Cheng et al., 2017). Hai-cheng Yi et al. proposed a computational RPI-SAN model by using the deep-learning stacked auto-encoder network to mine the hidden high-level features from RNA and protein sequences and fed them into a Random forest classifier to predict ncRNA binding proteins (Yi et al., 2018). They further employed Stacked assembling to improve the accuracy of the proposed method (Long et al., 2017; Patel et al., 2017). Including random forests and Bayesian classifiers, these classifiers are traditional classical machine learning classifiers which effectiveness have verified by a large-scale number of experiments (Liu et al., 2016; Wang et al., 2016; Luo and Liu, 2017). However, these traditional classifiers still have much room for improvement in classification performance and time consumption.

In recent years, an improved gradient boosting decision tree classifier named LightGBM has been proposed. LightGBM is a histogram-based decision tree algorithm, which divides continuous feature values into discontinuous feature blocks, and transforms these feature blocks into feature histograms during training (Shi et al., 2018). This LightGBM classifier algorithm had been used to speed up the decision tree building process on GPUs (Graphics Processing Units) and improved its scalability in the article of Huan Zhang et al. (Zhang et al., 2017). In their paper, a large number of experimental data shown that the training speed in constructing decision trees of LightGBM classification algorithm was much faster than general decision tree algorithms with the same classification accuracy (Mitchell and Frank, 2017).

In the field of biology, the discovery of ncRNAs has far exceeded the speed of research on their functions in ncRNA and protein interactions. Therefore, it is urgent to study an efficient prediction tool in the field of ncRNA-protein interactions which is less-time consuming and resource saving. Hence, we applied this efficient LightGBM classifier to large-scale ncRNA and protein interaction prediction and proposed a new machine learning model using sequence-based information named LGBM in this context. More specifically, each sequence of ncRNA is converted into a k-mers sparse matrix and the feature vectors of ncRNA are extracted from the resulting k-mers sparse matrices using the singular value decomposition (SVD). For proteins, based on the evolutionary point mutation model of protein sequences, we converted each protein sequence into a position-specific scoring matrix (PSSM) where the position information and frequency information were contained. Afterwards, each protein sequence was characterized by the feature vector obtained from a transform processing by using the pseudo-Zernike moment (PZM) algorithm. After extracting features of ncRNA and protein, we fed these reprehensive features into LightGBM classifier for classifying learning and predicting interactions between ncRNA and protein. In order to evaluate the predictive performance of the machine learning model, we used a 10-fold cross-validation to reduce overfitting. During the experiment, we employed four benchmark datasets to evaluate the performance of our model which was RPI369, RPI488, RPI1807, and RPI2241, respectively, and compared the prediction results of our model with other advanced models at the present stage. Experimental results indicated that our model LGBM performed well on four datasets above.

METHODS

Protein Feature Extraction

In this section, we selected the PZM feature extraction algorithm to extract sequence-based protein feature vectors using PSSM (Maali et al., 2016; Kheirkhah et al., 2017). The PSSM algorithm first integrates the biological evolution information to predict distantly related proteins, and has achieved good performances in protein binding sites and disordered region prediction (Yi et al., 2018). Let P be a PSSM matrix as the representative of an arbitrary protein. A matrix P consists of r rows and 20 columns with the explanation that r means the length of the primary sequence of an arbitrary protein while 20 means the quantity of amino acids (Sharma et al., 2013). Based on this, a PSSM matrix is

represented as follows:

$$P = \begin{bmatrix} p_{1,1} & \cdots & p_{1,20} \\ \vdots & \ddots & \vdots \\ p_{r,1} & \cdots & p_{r,20} \end{bmatrix} \quad (1)$$

Where $p_{i,j}$ in i_{th} row j_{th} column denotes the relative probability of j_{th} amino acid at the i_{th} position of the same protein sequence with which PSSM matrix comes from (Hayat and Khan, 2011). In experiments, the position-specific iterated BLAST (PSI-BLAST) tool was used to transform original protein sequences into PSSM matrices with the parameter *err-value* set to be 0.001.

Then we extracted PZM feature vectors from the resulting PSSM matrices above. The PZM is a statistical feature extraction algorithm that is computationally efficient for using global information to extract features (Haddadnia, 2001). Pseudo-Zernike polynomials are orthogonal sets of complex-valued polynomials defined as follows (Haddadnia et al., 2003):

$$V_{\alpha\beta}(x,y) = R_{\alpha\beta}(\rho)\ \exp\left(j\beta \tan^{-1}\left(\frac{y}{x}\right)\right) \quad (2)$$

Where $x^2 + y^2 \leq 1$, $\alpha \geq 0$, $|\beta| \leq \alpha$ and $\rho = \sqrt{x^2 + y^2}$ is the length of the vector from the origin to the pixel (x, y). And the radial polynomials $R_{\alpha\beta}$ are defined as:

$$R_{\alpha\beta}(x,y) = \sum_{t=0}^{\alpha-|\beta|} Z_{\alpha,|\beta|,t}\left(x^2+y^2\right)^{\frac{\alpha-t}{2}} \quad (3)$$

Where

$$Z_{\alpha,|\beta|,t} = (-1)^t \frac{2\alpha+1-t}{t!\,(\alpha-|\beta|-t)!\,(\alpha-|\beta|-t+1)!} \quad (4)$$

And $R_{\alpha,-\beta}(\rho) = R_{\alpha,\beta}(\rho)$ Therefore, the Zernike moments of order α with repetition β for a continuous image function $f(x,y)$ that vanishes outside the unit circle are as follows (Kim and Lee, 2003):

$$M_{\alpha\beta} = \frac{\alpha+1}{\pi} \iint_{x^2+y^2\leq 1} f(x,y) V^*_{\alpha\beta}(\rho,\theta)\, dxdy \quad (5)$$

Pseudo-Zernike polynomials are orthogonal and satisfy the following equation:

$$\iint_{x^2+y^2\leq 1} \left[V^*_{\alpha\beta}(x,y)\right] \times V_{mn}(x,y)\, dxdy = \frac{\pi}{\alpha+1}\delta_{\alpha m}\delta_{\beta n} \quad (6)$$

With

$$\delta_{ab} = f(x) = \begin{cases} 1, & a = b \\ 0, & otherwise \end{cases} \quad (7)$$

Hence, based on the derivation of the above formulas, the feature vectors of protein sequences can be represented as follows (Wang Y. et al., 2017):

$$\overrightarrow{F} = \left[|M_{11}|, |M_{22}|, \cdots, \left|M_{\alpha\beta}\right|\right]^T \quad (8)$$

ncRNA Feature Extraction

As for ncRNA, we used the SVD algorithm to extract feature vectors from the k-mers sparse matrix represented ncRNA sequences. In the k-mers sparse matrix construction algorithm, we traversed each complete ncRNA sequence (A, C, G, U) stepping one nucleotide at a time, which is considered characteristic of each nucleotide (Yi et al., 2018). After that, the frequency of the combined triplet feature based on 4 nucleotide letters was extracted for each RNA sequence and obtained 4^k dimensional features (You et al., 2016). Each characteristic value is the normalized frequency of 4-mers nucleotides in the ncRNA sequences, which is AAAA, AAAC … TTTT (Pan et al., 2016). Therefore, we obtained matrices including frequency information, location information and more hidden information represented the ncRNA sequences (Yi et al., 2018).

Furthermore, we used SVD algorithm to decompose k-mers sparse matrix. The Q represent the original k-mers sparse matrix from above and there is singular value decomposition as follows:

$$Q = U\Sigma V \quad (9)$$

Where the elements of diagonal in Σ represent the singular value of Q. It obtained the most information from original matrix Q. Consequently, We reconstruct a 1×4^k dimensional vector from Q shows as follows:

$$\overrightarrow{F} = U \quad (10)$$

LightGBM Algorithm

After obtaining potential features of ncRNA and protein calculated from above feature representation approaches, we fed these high-level features into LightGBM classifier to train the prediction scheme for predicting RPIs.

The traditional gradient boosting decision tree (GBDT) algorithm is a widely used machine learning algorithm which ensemble decision trees in an integrated learning model (Ke et al., 2017). This GBDT algorithm learns the decision trees by fitting the negative gradients (Friedman, 2001). In the process of learning decision trees, the most time-consuming and labor-consuming step is to find the best split points (Appel et al., 2013). The traditional GBDT algorithm uses the histogram-based algorithm to store continuous eigenvalues into discrete regions which are used to construct feature histograms during training instead of selecting the best split points (Li et al., 2007). However, with the increase of data volume, the workload of scanning all the data instances to estimate the information gain of all possible split points is increasing which costs time-consuming a lot (Chen and Guestrin, 2016). In order to address the limitation of this problem, an improved algorithm based on GBDT named LightGBM was proposed which improving the accuracy of classification in proposing two new novel techniques called Gradient Based One-side Sampling (GOSS) and Exclusive Feature Bundling (EFB) (Ke et al., 2017).

Through the GOSS algorithm, the problem that no native sample weights in GBDT avoiding hurting the accuracy of the learned model was solved by discarding those data instances with small gradients. Firstly, training instances were sort by their

gradients from high to low in order. Second, select top $p \times 100\%$ instances with high gradients and sample q percent data instances in the remaining subsets randomly. Let $A \cup B$ represents their collection. Hence, the estimated variance gain $\tilde{V}_s(b)$ of splitting feature s at point b over the subset $A \cup B$ can be define as follows (Ke et al., 2017):

$$\tilde{V}_s(b) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-p}{q}\sum_{x_i \in B_l} g_i\right)^2}{n_l^s(b)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-p}{q}\sum_{x_i \in B_r} g_i\right)^2}{n_r^s(b)}\right) \quad (11)$$

Where $A_l = \{x_i \in A; x_{ij} \leq b\}$, $A_r = \{x_i \in A; x_{ij} > b\}$, $B_l = \{x_i \in B; x_{ij} \leq b\}$ and $B_r = \{x_i \in B; x_{ij} > b\}$.

On the second step, the EFB algorithm was used to effectively reduce the number of features by bundling exclusive features into a single feature avoiding hurting the accuracy. By adopting the EFB algorithm, building the same feature histograms from the resulting feature bundles above were available as those from individual features (Meng et al., 2016). Therefore, the complexity of histogram building was reduced from $O(\#data \times \#feature)$ to $O(\#data \times \#bundle)$ since $\#bundle \ll \#feature$. First, we used NP-hard to partition features into a smallest number of exclusive bundles just as the graph coloring problem (Zuev, 2015). Second, offsets were added to the original values of feature vectors to merging the features in the same bundle and ensured that the values of the original values can be identified from the resulting feature bundles.

Evaluation Criteria

In this study, we used a 10 - fold cross-validation method to avoid overfitting and guarantee the accuracy of our algorithm of our model which divided the datasets into 10 equal parts randomly. During each training test, one part was taken as the testing dataset, while the remaining nine parts were the training datasets[1]. Therefore, a total of 10 experiments were conducted. To evaluate the performance of our model LGBM, we followed several widely used evaluation criteria including accuracy, sensitivity, specificity, precision, and Matthews Correlation Coefficient(MCC) as follows (Liu and Chen, 2012):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (13)$$

$$Specificity = \frac{TN}{TN + FP} \quad (14)$$

$$Precision = \frac{TP}{TP + FP} \quad (15)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (16)$$

[1] *K-Fold Cross Validation*. Classification.

TABLE 1 | The specific composition of four required datasets.

Datasets	Positive pairs	Negative pairs	The number of proteins	The number of ncRNAs
RPI369	369	369	338	332
RPI488	243	245	247	25
RPI1807	1,807	1,436	1,807	1,078
RPI2241	2,241	943	2,043	842

where *TP*, *FP*, *TN*, and *FN* are respectively interpreted as the number of true positive, false positive, true negative and false negative. The Receiver Operating Characteristic(ROC) curve can be represented as the threshold between *SP* and *SN*, which *x*-ray depicts false positive rate (FPR) while *y*-ray depicts true positive rate (TPR) (Huang et al., 2015). Meanwhile, the AUC is regarded as the area of the graphical under the ROC curve.

Datasets

To verify the robust and effectiveness of our model LGBM, we selected four ncRNA and protein interactions datasets including RPI369, RPI488, RPI1807, and RPI2241. Among them, the dataset RPI369 and RPI2241 were selected from the databases PRIDB which is a database of ncRNA-protein interfaces calculated from their complexes in the protein data bank (Berman et al., 2000; Wang et al., 2013). RPI2241 is a positive sample set consisting of 2,241 pairs of experimentally verified ncRNA-protein pairs including 2,043 protein chains and 842 ncRNA chains. RPI369 is a subpart of RPI2241 with 369 pairs including 338 protein chains and 332 ncRNA chains which excludes all ncRNA-protein interaction pairs that interact with ribosomal proteins or ribosomal ncRNA in various organisms (Muppirala et al., 2011). For dataset RPI369 and RPI2241, an approximately negative sample dataset was constructed with twice number of pairs by pairing ncRNA and protein sequences after removing the pairs in the positive sample dataset randomly (Wang et al., 2013). RPI488 is a non-redundant lncRPI dataset based on structural complexes which consists of 488 lncRNA-protein pairs, including 245 non-interacting pairs and 243 interacting pairs from shen et.al. (Pan et al., 2016). And RPI488 is smaller than other datasets since there are fewer lncRNA-protein complexes in PDB where ncRNA-protein complexes are destroyed from downstream (Ying et al., 2010). The dataset RPI1807 consists of 1807 positive ncRNA-protein pairs including 1078 ncRNA chains and 1807 protein chains and 1436 negative pairs with 493 ncRNA chains and 1436 Protein chains. It is established by parsing a nucleic acid database (NAD) that provides RNA protein complex data and protein RNA interface data (Yi et al., 2018). The specific composition of these four datasets are described in **Table 1**.

Experimental Results

In this study, we proposed a machine learning classification model based on improved gradient boosting decision tree to predict interactions between ncRNA and protein named LGBM which used PSSM and PZM algorithms to extract protein feature

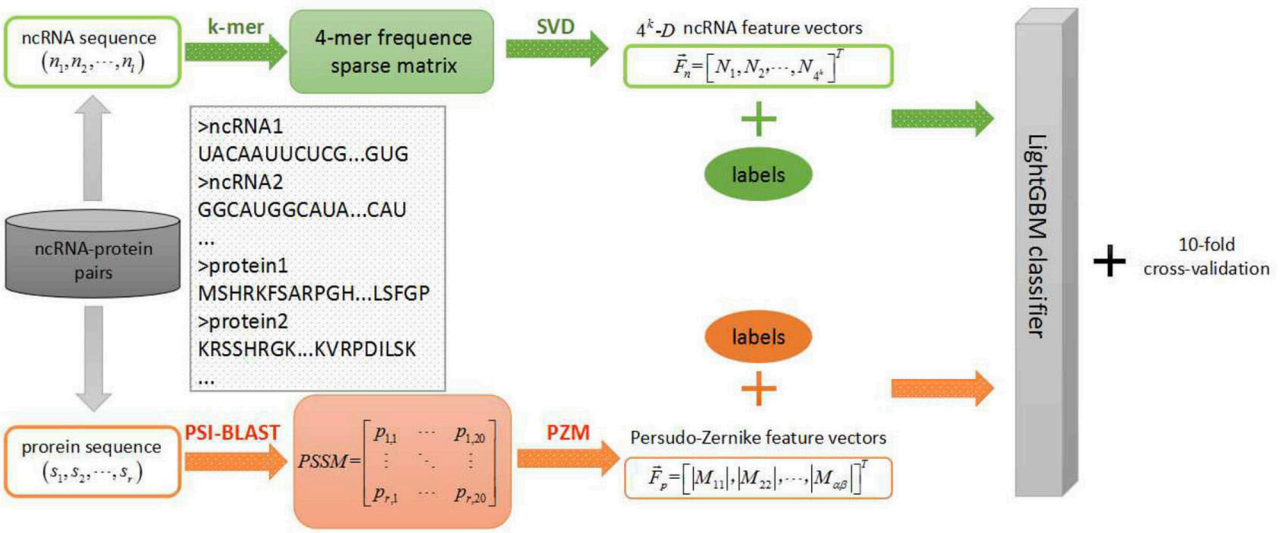

FIGURE 1 | Step-wise work flow for the purposed LGBM machine learning model.

TABLE 2 | Ten-fold cross-validation results on dataset RPI369.

Testing set	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	MCC (%)
1	75.00	73.17	81.08	68.57	50.12
2	70.83	73.53	67.57	74.29	41.90
3	76.39	76.32	78.28	74.29	52.73
4	76.39	75.68	77.78	75.00	52.80
5	76.39	71.11	88.89	63.89	54.51
6	69.44	65.91	80.56	58.33	39.89
7	73.61	72.97	75.00	72.22	47.24
8	75.00	72.50	80.56	69.44	50.31
9	74.65	71.43	83.33	65.71	49.89
10	70.42	69.23	75.00	65.71	40.91
Average	73.81	72.18	68.75	78.81	48.03

TABLE 3 | Ten-fold cross-validation results on dataset RPI488.

Testing set	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	MCC (%)
1	91.84	100.0	83.33	100.0	84.76
2	87.76	94.00	77.27	96.30	75.91
3	87.76	88.89	88.89	86.36	75.25
4	95.92	100.0	92.00	100.0	92.15
5	75.51	75.00	60.00	86.21	48.43
6	91.84	91.30	91.30	92.31	83.61
7	93.75	100.0	86.36	100.0	87.99
8	87.50	96.30	83.87	94.12	75.19
9	91.60	95.24	86.96	96.00	83.54
10	91.67	91.67	91.67	91.67	83.83
Average	89.52	93.28	94.30	84.17	79.02

vectors and combined k-mers matrices and SVD algorithms to extract RNA feature vectors. The specific steps of the machine learning model are shown in the **Figure 1**. To verify the performance of the proposed model LGBM, we evaluated the prediction ability of LGBM on datasets RPI369 and RPI488 and had a comparison with the prediction performance of other classifiers under the same feature extraction condition firstly. In addition, we also evaluated the predictive performance of datasets RPI1807 and RPI2241 and compared the prediction results of these two datasets with those of other proposed models in earlier papers.

Prediction Ability of LGBM

In this section, we validated our machine learning model LGBM with 10-fold cross-validation on datasets RPI369 and RPI488 to predicting ncRNA-protein interactions. The 10-fold cross-validation contributed LGBM to avoid over-fitting and had a better performance. As a result, the summary of experimental prediction results under 10-fold cross-validation are shown in **Tables 2**, **3**.

As shown in **Tables 2**, **3**, when LGBM machine learning model was used to predict interactions between ncRNA and protein on dataset RPI369, the mean performance of accuracy, precision, sensitivity, specificity and MCC were 73.81, 72.18, 68.75, 78.81, and 48.03%, respectively. While for dataset RPI488, the mean performance of accuracy, precision, sensitivity, specificity and MCC highly achieved 89.52, 93.28, 94.30, 84.17, and 79.02%, respectively. At the meantime, in 10-fold cross-validation, the accuracy of one validation was even as high as 95.92% while there were other five validations achieved the accuracy of 90%.

TABLE 4 | Performance evaluation on different classifiers.

Dataset	Classifier	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	MCC (%)
RPI369	LGBM	**73.81**	**72.18**	68.75	**78.81**	**48.03**
	SVM	71.60	71.70	**70.74**	72.51	43.62
	GBDT	71.74	71.79	**70.74**	72.79	43.90
RPI488	LGBM	**89.52**	**93.28**	**94.30**	**84.17**	**79.02**
	SVM	86.22	88.62	89.86	82.27	72.44
	GBDT	86.01	88.54	89.86	81.81	72.04

The bold value indicates this measure performance is the best among the compared methods.

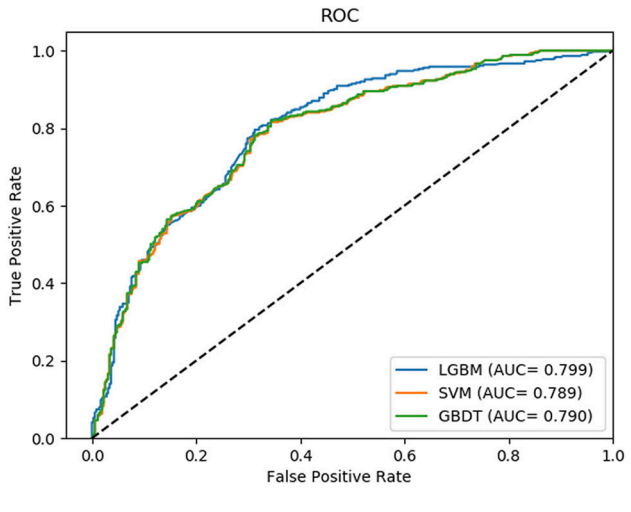

FIGURE 2 | The ROC curve of dataset RPI369 on three classifiers.

The prediction accuracy of LGBM on datasets RPI369 and RPI488 illustrated the feasibility of predicting ncRNA and protein interactions only based on their sequence information. In fact, the protein and ncRNA feature extraction methods can extract more in-depth information hidden in sequences including location, frequency and interaction information into PSSM matrices and k-mers matrices (You et al., 2016). In addition, selecting PZM algorithm to extract feature vectors makes better use of the properties of PZM (Khotanzad and Hong, 1990).

Comparison Between Different Classifiers

In this comparison module, we compared the prediction performance of LightGBM classifier, SVM classifier and traditional gradient boosting decision tree classifier in datasets RPI369 and RPI488 sharing the same feature extraction condition. As a result, the summary of experimental prediction results under 10-fold cross-validation is shown in **Table 4** and the corresponding trade-off between false positive rate and true positive rate shown in the receiver operating characteristic (ROC) curve in **Figures 2**, **3**.

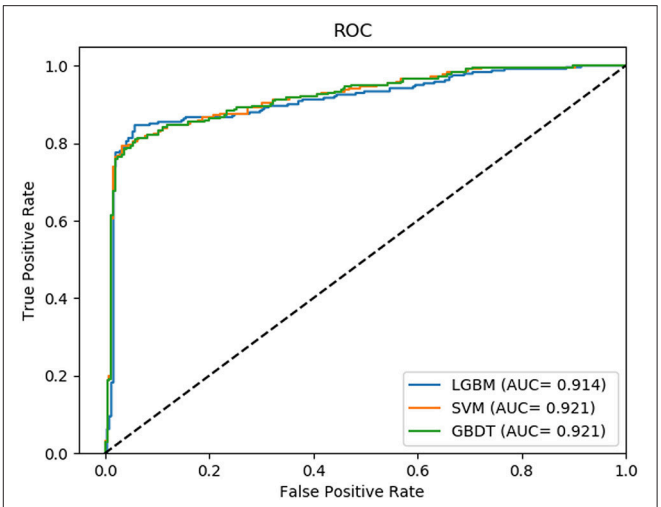

FIGURE 3 | The ROC curve of dataset RPI488 on three classifiers.

As can be seen from **Table 4**, the LightGBM classifier achieved an accuracy of 73.81% in predicting interactions between ncRNA and protein in dataset RPI369, which was higher than 71.60% of SVM classifier and 71.74% of traditional GBDT classifier. And as for precision, sensitivity and MCC except specificity, the LightGBM classifier also had a better performance with exact percent of 72.18, 78.81, and 48.03% respectively, while 71.70, 72.51, and 43.62% under SVM classifier and 71.79, 72.79, and 43.90% under traditional GBDT classifier. For dataset RPI488, whether accuracy, precision and sensitivity or specificity and MCC, LightGBM classifier performed better than the other two classifiers with the exact results of 89.52, 93.28, 84.17, 94.30, and 79.02%, respectively. That is to say, under the evaluation of each evaluation criterion, our LightGBM classifier had a better classification performance than SVM and traditional GBDT classifiers which proved the feasibility and effectiveness of choosing LightGBM classifier to process sequence information in our model LGBM.

The comparison results shown the feasibility and effectiveness of selecting LightGBM as classifier in our model (Zhu et al., 2017). In fact, LightGBM, as an improved gradient boosting decision tree, processing the advantages of reducing the number of features and gaining enough information gain through smaller datasets by EFB and GOSS, is superior to other classifiers in terms of computational speed and memory consumption (Wang et al., 2017).

Comparison With Other Existing Methods

In this section, we compared the prediction performance combined with 10-fold cross-validation of LGBM model at datasets RPI488, RPI1807, and RPI2241 with RPI-Pred, RPISeq-RF, and Inc-Pro. RPI-Pred is a SVM-based ncRNA-protein interactions prediction model proposed by Suresh et al. which based on sequence and structure information (Suresh et al., 2015). The accuracy of the RPI-Pred model on dataset RPI1807 is 93.00%. RPISeq-RF is a random forest classifier-based model

TABLE 5 | Comparison between LGBM and other methods in RPI488, RPI1807, and RPI2241.

Dataset	Method	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	MCC (%)
RPI488	LGBM	**89.52**	**93.28**	**94.30**	**84.17**	**79.02**
	RPISeq-RF	88.00	93.20	92.60	82.20	76.20
	IncPro	87.00	91.00	90.00	82.70	74.00
RPI1807	LGBM	96.42	**96.21**	95.20	97.40	92.76
	RPI-Pred	93.00	94.00	95.00	N/A	N/A
	IncPro	**96.90**	95.50	**96.50**	**98.10**	**93.80**
RPI2241	LGBM	**68.86**	**72.76**	**76.38**	61.50	**38.33**
	RPISeq-RF	63.96	65.37	64.83	62.59	27.98
	IncPro	65.40	66.90	65.90	**64.00**	31.00

The bold value indicates this measure performance is the best among the compared methods.

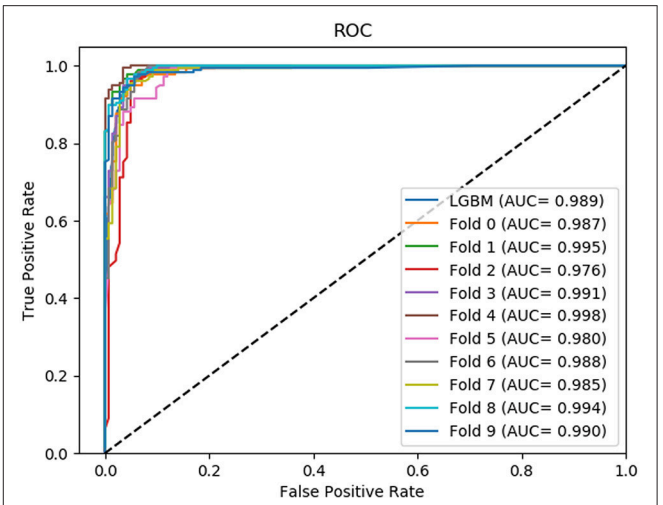

FIGURE 5 | The ROC curve of dataset RPI1807 on 10-fold cross- validation.

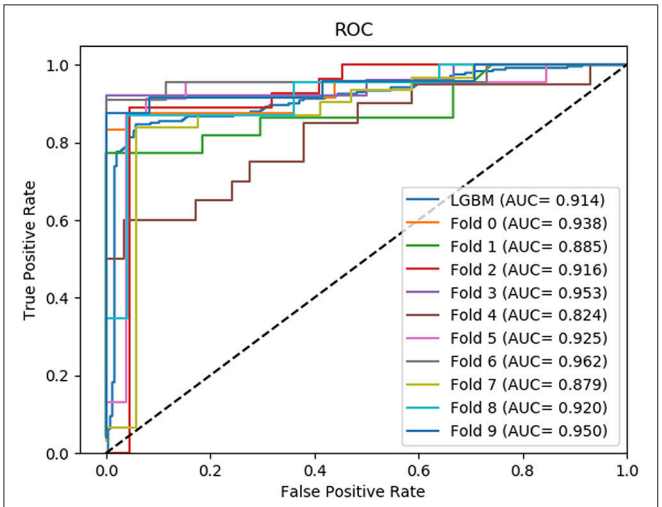

FIGURE 4 | The ROC curve of dataset RPI488 on 10-fold cross- validation.

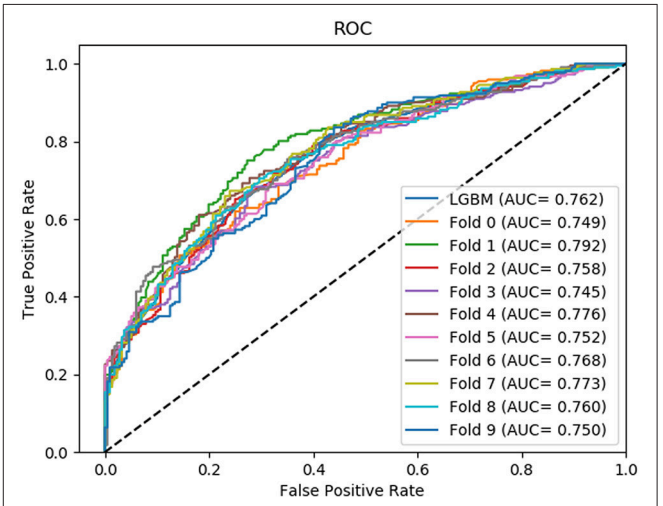

FIGURE 6 | The ROC curve of dataset RPI2241 on 10-fold cross- validation.

proposed by Usha K Muppirala et al. which extracts feature vectors from ncRNA and protein sequence information only (Muppirala et al., 2011). And the accuracy of the RPISeq-RF model on datasets RPI488, RPI1807, and RPI2241 are 88.00, 97.30, and 63.96%, respectively. IncPro is a model proposed by Lu et al. which encodes lncRNA and protein sequences as digital vectors and scores each lncRNA-protein pair using matrix multiplication (Lu et al., 2013). Based on this IncPro model, the accuracy of datasets RPI488, RPI1807, and RPI2241 achieves 87.00, 96.90, and 65.40%, respectively. The summary comparative results of the experiments are shown in **Table 5.** And the 10-fold cross-validation ROC curve for our model LGBM at RPI488, RPI1807, and RPI2241 are shown in **Figures 4–6**.

As shown in **Table 5**, our machine learning model LGBM achieved an experimental prediction accuracy of 89.52 %, higher than 88.00% of RPISeq-RF and 87.00% of IncPro on dataset RPI488. At the meantime, LGBM also had a better performance in other evaluation criterions including precision, sensitivity, specificity and MCC of 93.28, 94.30, 84.17, and 79.02%, respectively. While the performance of RPISeq-RF were 88.00, 93.20, 92.60, 82.20, 76.20% and IncPro were 87.00, 91.00, 90.00, 82.70, and 74.00%. On dataset RPI2241, except specificity, our model had a better performance of 68.86, 72.76, 76.38% on accuracy, precision and sensitivity. While on dataset RPI1807, although our experimental prediction performance was not as good as IncPro, it was still as high as 96.42, 96.21, 95.20, 97.40, and 97.26%, which is not much lower than 96.90, 95.50, 96.50, 98.10, and 93.80% of IncPro on accuracy, precision, sensitivity, specificity and MCC respectively. PRI-Pred performed slightly worse which was 93.00, 94.00, and 95.00% on accuracy, precision and sensitivity.

By comparing the prediction results, we are able to see that our prediction model LGBM has a better performance on datasets RPI488 and RPI2241, however, on dataset RPI1807, the prediction accuracy is worse than IncPro, while the accuracy is still more than 96%. In general, our model LGBM is effective and robust in predicting interactions between ncRNA and protein.

CONCLUSION

In this study, we proposed an efficient prediction model LGBM using sequence and evolutionary information to predict interactions between ncRNA and protein. In order to obtain evolutionary information from protein sequences, the Zernike Moment algorithm is used to extract feature vectors of proteins from PSSM. Meanwhile, the SVD was used to extract features from k-mers sparse matrix of ncRNA, in which both the location and frequency information is preserved. On this basis, we fed the high-level feature vectors into the LightGBM classifier to predict the interaction between ncRNA and protein. To verify the accuracy and robustness of our model, 10-fold cross validation was used. Experimental results on datasets RPI369, RPI488, RPI1807 and RPI2241 demonstrated the robustness and effectiveness of our model. Therefore, the proposed LGBM model is feasible, reliable and full of generalization ability to predict ncRNA-protein interaction. Our research can be a useful tool to further biological research.

AUTHOR CONTRIBUTIONS

Z-HZ, Z-HY, and YZ conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript. L-PL and H-CY designed, performed and analyzed experiments and wrote the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation of China, under Grants 61373086, 61572506. The authors would like to thank all the editors and reviewers for their constructive advices.

REFERENCES

Akbaripour-Elahabad, M., Zahiri, J., Rafeh, R., Eslami, M., and Azari, M. (2016). rpiCOOL: a tool for In Silico RNA–protein interaction detection using random forest. *J. Theor. Biol.* 402, 1–8. doi: 10.1016/j.jtbi.2016.04.025

Appel, R., Fuchs, T., and Perona, P. (2013). "Quickly boosting decision trees, pruning underachieving features early," in *International Conference on International Conference on Machine Learning: 2013* (Atlanta, GA), III-594.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank, 1999–. *Int. Tables Crystallograp.* 67, 675–684. doi: 10.1107/97809553602060000722

Chen, T., and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. *arXiv:1603.02754*. 2016, 785–94.

Cheng, Z., Kai, H., Yang, W., Hui, L., Guan, J., and Zhou, S. (2017). Selecting high-quality negative samples for effectively predicting protein-RNA interactions. *BMC Syst. Biol.* 11(Suppl. 2):9. doi: 10.1186/s12918-017-0390-8

Cook, K. B., Vembu, S., Kch, H., Zheng, H., Laverty, K. U., Hughes, T. R., et al. (2017). RNAcompete-S: combined RNA sequence/structure preferences for RNA binding proteins derived from a single-step *in vitro* selection. *Methods* doi: 10.1016/j.ymeth.2017.06.024

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Haddadnia, J. (2001). "Neural network based face recognition with moments invariant,"in *IEEE Intconfimage Processing Thessaloniki*. Greece 2001.

Haddadnia, J., Ahmadi, M., and Faez, K. (2003). An efficient feature extraction method with pseudo-zernike moment in RBF neural network-based human face recognition system. *EURASIP J. Adv. Signal Process* 2003:267692. doi: 10.1155/S1110865703305128

Hayat, M., and Khan, A. (2011). Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. *J. Theor. Biol.* 271:10. doi: 10.1016/j.jtbi.2010.11.017

Huang, Y. A., You, Z. H., Xin, G., Leon, W., and Wang, L. (2015). Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence. *Biomed Res. Int.* 2015:902198. doi: 10.1155/2015/902198

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "LightGBM: a highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30 (NIPS 2017),* 3146–3154.

Kheirkhah, A., Mohd Daud, S., and Kamardin, K. (2017). Enhancing efficiency of protein functional prediction through association network using greedy weighting method. *Curr. Bioinform.* 12, 275–284. doi: 10.2174/1574893611666161118142028

Khotanzad, A., and Hong, Y. H. (1990). Invariant image recognition by zernike moments. *IEEE Trans. Pattern Anal. Mach. Intell.* 12:489–497. doi: 10.1109/34.55109

Kim, H. S., and Lee, H. K. (2003). Invariant image watermark using Zernike moments. *Circ. Syst. Video Technol. IEEE Transact.* (2003)13, 766–775. doi: 10.1109/TCSVT.2003.815955

Li, P., Burges, C. J. C., and Wu, Q. (2007). "McRank: learning to rank using multiple classification and gradient boosting," in *International Conference on Neural Information Processing Systems: 2007* (Vancouver, BC), 897–904.

Li, Z., and Nagy, P. D. (2011). Diverse roles of host RNA-binding proteins in RNA virus replication. *RNA Biol.* 8, 305–315. doi: 10.4161/rna.8.2.15391

Liu, X., Zou, Q., Wu, Y., Li, D., and Zeng, J. (2016). An empirical study of features fusion techniques for protein-protein interaction prediction. *Curr. Bioinform.* 11, 4–12. doi: 10.2174/1574893611666151119221435

Liu, Z. P., and Chen, L. (2012). Proteome-wide prediction of protein-protein interactions from high-throughput data. *Protein Cell* 3, 508–520. doi: 10.1007/s13238-012-2945-1

Long, H., Wang, M., and Fu, H. (2017). Deep convolutional neural networks for predicting hydroxyproline in proteins. *Curr. Bioinform.* 12, 233–238. doi: 10.2174/1574893612666170221152848

Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X., et al. (2013). Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics* 14:651. doi: 10.1186/1471-2164-14-651

Luo, J., and Liu, C. (2017). An effective method for identifying functional modules in dynamic PPI networks. *Curr. Bioinform.* 12, 66–79. doi: 10.2174/1574893611666160831113726

Luo, J., Liu, L., Venkateswaran, S., Song, Q., and Zhou, X. (2017). RPI-Bind, a structure-based method for accurate identification of RNA-protein binding sites. *Sci. Rep.* 7:614. doi: 10.1038/s41598-017-00795-4

Maali, A. A., Mahdavi, M., and Gheshlaghi, R. (2016). Suitability of sequence-based feature vector for classification algorithm improves accuracy of human protein-protein interaction prediction: a red blood cell case study. *Curr. Bioinform.* 11, 291–300. doi: 10.2174/1574893610666151026215233

Meng, Q., Ke, G., Wang, T., Chen, W., Ye, Q., Ma, Z. M., et al. (2016). A communication-efficient parallel algorithm for decision tree. arXiv:1611.01276

Mitchell, R., and Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing. *PeerJ Comput. Sci.* 3:e127. doi: 10.7717/peerj-cs.127

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

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

Pan, X., Rijnbeek, P., Yan, J., and Shen, H. B. (2017). Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics* 19:511. doi: 10.1186/s12864-018-4889-1

Patel, S., Tripathi, R., Kumari, V., and Varadwaj, P. (2017). DeepInteract: deep neural network based protein-protein interaction prediction tool. *Curr. Bioinform.* 12, 551–557. doi: 10.2174/1574893611666160815150746

Sharma, A., Lyons, J., Dehzangi, A., and Paliwal, K. K. (2013). A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. *J. Theor. Biol.* 320:41. doi: 10.1016/j.jtbi.2012.12.008

Shi, Y., Li, J., and Li, Z. (2018). Gradient boosting with piece-wise linear regression trees. arXiv:1802.05640.

Suresh, V., Liu, L., Adjeroh, D., and Zhou, X. (2015). RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. *Nucleic Acids Res.* 43:1370. doi: 10.1093/nar/gkv020

Tian, Y., Xu, J., Du, X., and Fu, X. (2018). The interplay between noncoding RNAs and insulin in diabetes. *Cancer Lett.* 419, 53–63. doi: 10.1016/j.canlet.2018.01.038

Wang, D., Zhang, Y., and Zhao, Y. (2017). LightGBM: an effective miRNA classification method in breast cancer patients. *ICCBB 2017* (Newark, NJ), 1–7. doi: 10.1145/3155077.3155079

Wang, F., Song, B., Zhao, X., Miao, Y., Li, D., Zhou, N., et al. (2016). Prediction and analysis of the protein-protein interaction networks for chickens, cattle, dogs, horses and rabbits. *Curr. Bioinform.* 11, 131–142. doi: 10.2174/1574893611666151203221255

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

Wang, Y., You, Z., Li, X., Chen, X., Jiang, T., and Zhang, J. (2017). PCVMZM: using the probabilistic classification vector machines model combined with a zernike moments descriptor to predict protein–protein interactions from protein sequences. *Int. J. Mol. Sci.* 18:1029. doi: 10.3390/ijms18051029

Yi, H. C., You, Z. H., Huang, D. S., Li, X., Jiang, T. H., and Li, L. P. (2018). A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information. *Mol. Ther. Nucleic Acids* 11, 337–344. doi: 10.1016/j.omtn.2018.03.001

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

You, Z. H., Zhou, M., Luo, X., and Li, S. (2016). Highly efficient framework for predicting interactions between proteins. *IEEE Transactions on Cybernetics* 47, 1–13. doi: 10.1109/TCYB.2016.2524994.

Zhang, H., Si, S., and Hsieh, C. J. (2017). GPU-acceleration for large-scale tree boosting. arXiv:1706.08359. Available Online at: https://arxiv.org/abs/1706.08359

Zhu, J., Shan, Y., Mao, J. C., Yu, D., Rahmanian, H., and Zhang, Y. (2017). Deep embedding forest: forest-based serving with deep embedding features. arXiv:1703.05291.

Zuev, Y. A. (2015). A graph coloring problem. *Mathematical Notes* 97, 965–7. doi: 10.1134/S0001434615050338

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.