Yi et al. BMC Bioinformatics (2020) 21:60
https://doi.org/10.1186/s12859-020-3406-0

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

RPI-SE: a stacking ensemble learning framework for ncRNA-protein interactions prediction using sequence information

Hai-Cheng Yi[1,2], Zhu-Hong You[1,2*], Mei-Neng Wang[3], Zhen-Hao Guo[1], Yan-Bin Wang[1] and Ji-Ren Zhou[1]

Abstract

Background: The interactions between non-coding RNAs (ncRNA) and proteins play an essential role in many biological processes. Several high-throughput experimental methods have been applied to detect ncRNA-protein interactions. However, these methods are time-consuming and expensive. Accurate and efficient computational methods can assist and accelerate the study of ncRNA-protein interactions.

Results: In this work, we develop a stacking ensemble computational framework, RPI-SE, for effectively predicting ncRNA-protein interactions. More specifically, to fully exploit protein and RNA sequence feature, Position Weight Matrix combined with Legendre Moments is applied to obtain protein evolutionary information. Meanwhile, *k*-mer sparse matrix is employed to extract efficient feature of ncRNA sequences. Finally, an ensemble learning framework integrated different types of base classifier is developed to predict ncRNA-protein interactions using these discriminative features. The accuracy and robustness of RPI-SE was evaluated on three benchmark data sets under five-fold cross-validation and compared with other state-of-the-art methods.

Conclusions: The results demonstrate that RPI-SE is competent for ncRNA-protein interactions prediction task with high accuracy and robustness. It's anticipated that this work can provide a computational prediction tool to advance ncRNA-protein interactions related biomedical research.

Keywords: Sequence analysis, RNA-protein interaction, ncRNA, Ensemble learning, Position weight matrix, Legendre moments

Background

Protein is the main bearer of cellular activities. However, only a small fraction of the *Human* genome (about 2%) contains protein-coding genes [1]. The remaining 98% of the genes are mainly responsible for regulation, that is, they are involved in controlling when and where genes are expressed and activated [2]. This part of the huge genome produces RNA molecules that vary in size, structure, and function. They are called non-coding RNAs (ncRNA) [3]. Different types of non-coding RNA interact with proteins in different ways. NcRNA can be divided into several categories, which are widely present in most cells and are vital in life activities. And there are some ncRNAs that play a role in specific species. Highly conserved ncRNAs are considered molecular fossils and functional redundancy in the RNA world, and they have been found to act as structural or regulatory molecules involved in the complex flow of life information from DNA to proteins [4].

NcRNA-Protein interactions (ncRPIs) play an essential role in many biological functions. Many ncRNAs play a regulatory role in DNA replication, translation, RNA splicing, and gene expression (such as trans-acting and cis-acting), genome defense and so on [5–7]. Meanwhile, a variety of diseases can be caused by mutations or imbalances in the composition of ncRNAs in the body, such as cancer [8], Prader-Wills syndrome [9], autism [10], Alzheimer's disease [11], cartilage-hair hypoplasia [12], hearing loss [13]. Because the role of ncRNAs usually depends on binding to specific proteins, identifying

* Correspondence: zhuhongyou@ms.xjb.ac.cn
[1]Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
Full list of author information is available at the end of the article

the protein molecules that bind to specific ncRNAs is the key to studying the function and mechanism of ncRPIs. Thanks to the Human Genome Project, research in the life sciences has entered the era of post-genomics. The application of various advanced high-throughput experimental methods has generated and accumulated huge amounts of data that are in urgent need of analysis. There is already a gap between the known ncRNAs and their interactions.

High-throughput methods are valuable but time-consuming and expensive. In recent years, there have been extensive research on computational prediction of proteins-RNAs interactions (RPIs) [14–18]. Pancaldi et al. applied both Random Forest (RF) and Support Vector Machine (SVM) model for RPIs prediction, using more than 100 different functional and physical features, such as genomic context, structure or localization, experimental translation and so on [19]. Muppirala et al. introduced a mothed named RPISeq, which also used RF and SVM classifiers based on primary sequences information [20]. In 2013, Lu et al. trained different types Fisher linear discriminant model using the information of hydrogen bonding propensities, the secondary structure and Van der Waals of long ncRNAs and proteins [14]. Suresh et al. presented RPI-Pred, a computational approach based on SVM to predict RPIs by using both high-order structure information [21]. Recently, Cirillo et al. introduced Global Score for protein-RNA interactions prediction. The main contribution of this method is to integrate the local characteristics of protein and RNA structures into the overall binding tendency, and calibrate it based on high-throughput data [22]. Pan et al. put forward a model combined stacking autoencoder with random forest classifiers named IPMiner, archived great prediction performance of ncRPIs [23]. As can be seen, both efficient feature extraction and machine learning model are important to achieve great predictive performance in this domain.

In our previous work, we presented a deep learning stacked autoencoder network based framework to predict ncRNA-protein interactions, named RPI-SAN. The main contribution of RPI-SAN is the application of deep stacked autoencoder to obtain efficient hidden representation of RNA and protein sequence information [18, 24]. Deep learning shows excellent ability with large-scale data support in many fields, however, ncRPIs data sets generally don't have large scales, thus it's not very suitable or urgent need for deep learning methods. Previous research confirmed that in ncRPIs prediction task, tree-based model and SVM model can work well, and sequences contain enough information for predicting ncRPIs [25, 26]. Traditional machine learning techniques have the potential to be explored for accuracy and interpretability in small sample learning tasks, especially ncRNA-protein interactions prediction task.

To this end, we propose a stacking ensemble based computational model, RPI-SE, by integrating Gradient Boosting Decision Tree (GBDT, implemented by XGBoost) [27], SVM [28, 29] and Extremely randomized Trees [30] (ExtraTree) algorithms to predict ncRNA-protein interactions. Specifically, *k*-mer sparse matrix is used to exploit the sequence information of RNA, which retains not only the nucleic acid components, but also the sequence order information [18, 31, 32]. Meanwhile, the Legendre Moments (LMs) descriptor is applied to convert the information contained in a the Position Weight Matrix (PWM) [33, 34] in a feature vector, which can retain the evolutionary information contained in amino acid sequences corresponding to physicochemical properties. And the Singular Value Decomposition (SVD) [35] is further applied to reduce the dimension of vectors. Then, these evolutionary features are used to train three base predictors include GBDT, SVM and ExtraTree. Finally, stacking ensemble is adopted to integrate these base predictors. To thoroughly verify the performance, the RPI-SE is evaluated on three benchmark data sets under five-fold cross-validation, including RPI369 [20], RPI488 [23] and RPI1807 [21], and compared with other methods, including RPISeq-RF [20], RPI-Pred [21], lncPro [14], IPMiner [23] and RPI-SAN [18]. The experimental results demonstrate that RPI-SE is competent for ncRPIs prediction task, obtained predictive performance with high accuracy and robustness. The workflow of the proposed method is shown in Fig. 1.

Results

In this work, we proposed a stacking ensemble based computational model to predict ncRNA-protein interactions, called RPI-SE, which integrated XGBoost, SVM and ExtraTree algorithms and using high efficiency features. Above all, we evaluated RPI-SE's predictive performance of RNA-protein interactions on benchmark data sets. Moreover, we compare RPI-SE with other computational methods on different data sets, including RPI488, RPI369 and RPI1807. Furthermore, the performance of different integration strategies has also been analyzed. The evaluation indicators used in the assessment include accuracy (Acc), true negative rate (TNR), true positive rate (TPR), positive predictive value (PPV), Matthews Correlation Coefficient (MCC) and the Area Under the

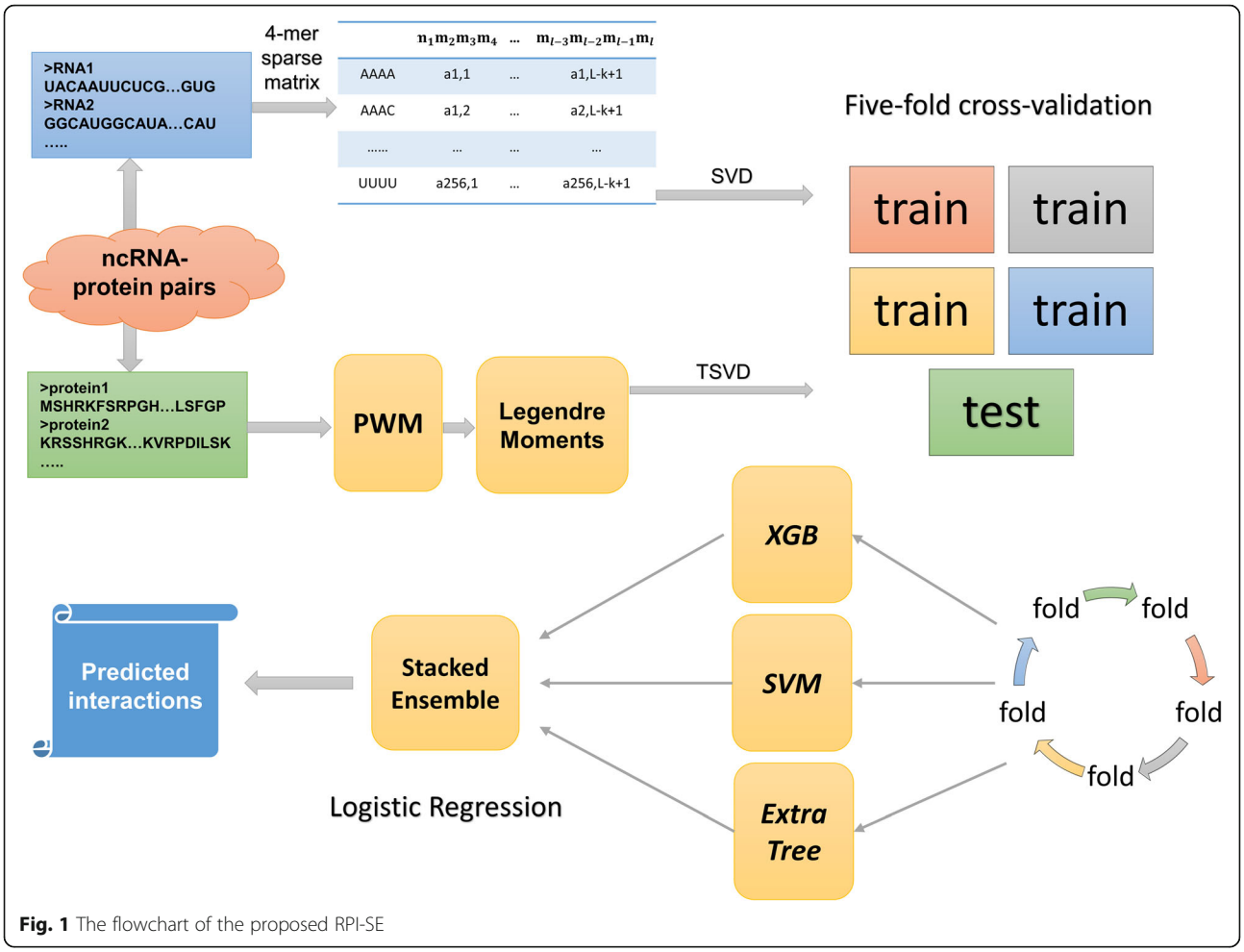

Fig. 1 The flowchart of the proposed RPI-SE

Receiver Operating Characteristic curve (AUC) are also adopted to evaluate the performance of RPI-SE.

Evaluate RPI-SE's performance of RNA-protein interactions prediction

To evaluate RPI-SE's ability of predicting RNA-protein interactions, the RPI-SE is carried out on RPI369 data set under five-fold cross-validation. The Table 1 shows the results of five-fold cross-validation of RPI-SE on the RPI369 data set. Meanwhile, a comparison of the results of individual base classifiers and stack integration is shown in Table 2. Certainly, the same experiments were performed on RPI488 and RPI1807 data sets, and their results are reported in Additional file 1.

Under five-fold cross-validation, RPI-SE performs much better than compared methods on RPI369 data set. From Table 2, RPI-SE performs an accuracy of 88.44%, a TPR of 83.69%, a TNR of 95.87%, a PPV of 80.85%, an MCC of 77.73% and as shown in Fig. 2, RPI-SE performed an AUC 0.924. It's the best of the four comparison predictors. XGBoost achieves an accuracy of 84.54%, a TPR of 81.45%, a TNR of 90.08%, a PPV of 78.87% and MCC of 69.51%. It is the best performing base classifier. The accuracy, TPR, TNR, PPV and MCC

Table 1 The five-fold cross-validation performance on RPI369 data set

Fold set	Acc (%)	TPR (%)	TNR (%)	PPV (%)	MCC (%)
1	90.28	86.42	95.89	84.51	81.03
2	88.19	82.56	97.26	78.87	77.61
3	88.19	84.15	94.52	81.69	76.95
4	87.41	81.40	97.22	77.46	76.27
5	88.11	83.95	94.44	81.69	76.81
Average	88.44 ± 1.08	83.69 ± 1.88	95.87 ± 1.38	80.85 ± 2.75	77.73 ± 1.90

Table 2 Performance of individual predictors and RPI-SE on RPI369 data set

Predictors	Acc(%)	TPR(%)	TNR(%)	PPV(%)	MCC(%)
XGBoost	84.54	81.45	90.08	78.87	69.51
SVM	75.63	72.50	83.49	67.61	51.86
ExtraTree	68.66	67.65	72.74	64.51	37.57
RPI-SE	**88.44**	**83.69**	**95.87**	**80.85**	**77.73**

The boldface indicates this measure performance is the best among the compared methods

of kernel SVM are 75.3, 72.50, 83.49, 67.61 and 51.86% and those of ExtraTree are 68.66, 67.65, 72.74, 64.51% and only 37.57%. The experimental results demonstrate our model is suitable for RNA-protein interaction prediction.

Comparison between different integrated learning strategies

To demonstrate the performance improvement of integration strategies, we compared stacked ensemble with base predictors and general averaged ensemble strategies on RPI369. Stacked ensemble is implemented by a Logistic Regression function. Logistic regression automatically learns respective weights for the three base predictors, including XGBoost, SVM and ExtraTree. As Fig. 3 shows, stacked ensemble archived an AUC of 0.925, better than averaged ensemble method and three base classifiers. Experimental results prove that the stacked integration strategy improves the performance of the prediction framework and is more powerful and flexible than the averaged integration strategy.

Compared with other state-of-the-art methods

We further compared RPI-SE with other computational methods under same conditions. The contrast methods include IPMiner [23], lncPro [14], RPISeq-RF [20], and RPI-SAN.

As Table 3 shows, on RPI369 data set, RPI-SE is obviously better than other methods, with an accuracy of 88.44%, a TPR of 83.69%, a TNR of 95.87%, a PPV of 80.85%, an MCC of 77.73% and AUC of 0.924 (shown in Fig. 2). RPI-SE increased the accuracy, TPR, TNR, PPV, MCC, and AUC by more than 13.2, 10, 16.7, 9.5, 27 and 15%, respectively. For RPI488 data set, RPI-SE also obtained acceptable performance (as AUC shown in Fig. 4), with an accuracy of 89.3%, better than other comparison methods but only closed to RPI-SAN. As shown in Table 3 and Fig. 5, on the data set RPI1807, the results of all the methods are close, with an accuracy rate of over 96%. RPI-SE attains a high accuracy of 96.86%.

Discussion

RPI-SE is composed of three basic predictors, XGBoost classifier, SVM classifier with RBF kernel, and ExtraTree classifier. Different classifiers have different adaptability to the data. XGBoost has advantages in accuracy and TPR, while SVM has advantages in stability. At the same time, basic classifiers have their own disadvantages. It is necessary to integrate them for best performance. The degree to which the stacking strategy improves the final prediction performance is different. When the difference between the classifiers is greater, stacking integration is more effective. The RPI488 and RPI1807 data sets have stronger correlations, so the base predictors have more consistent output on these two

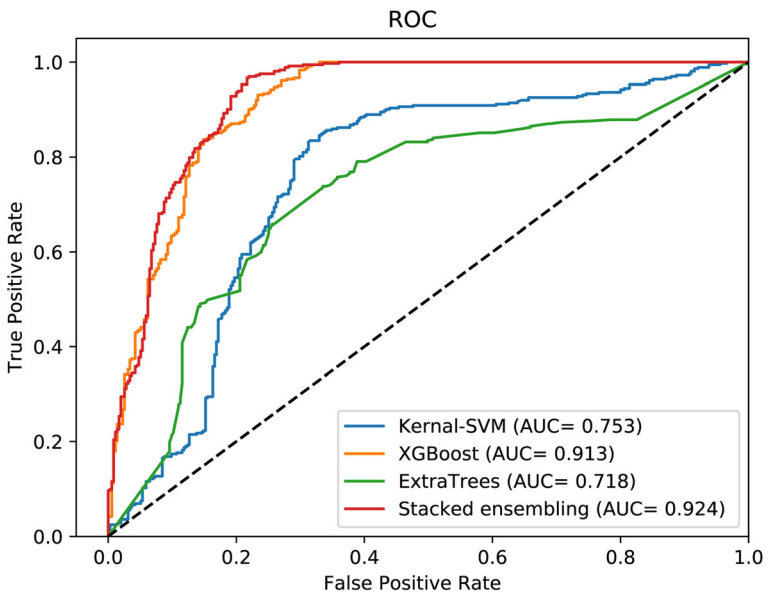

Fig. 2 The performance of RPI-SE and contrast methods on RPI369

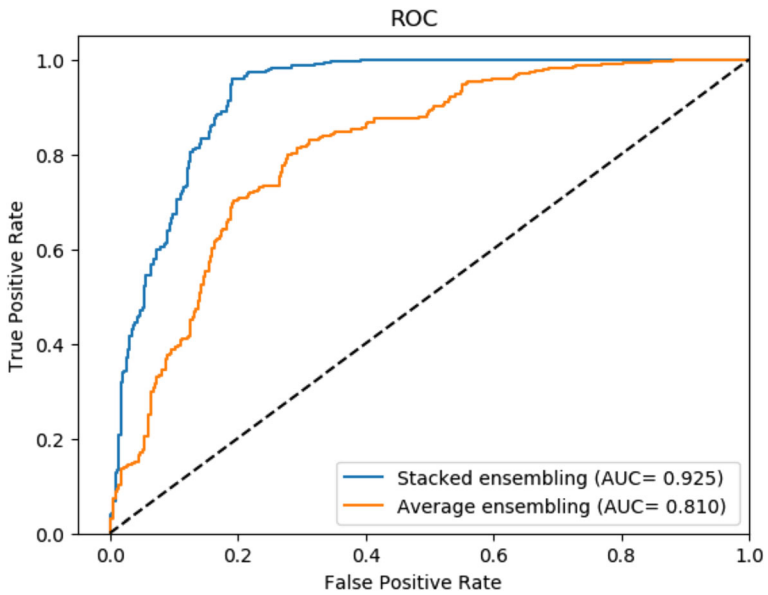

Fig. 3 Comparison of different integration strategies

data sets, and the stacking ensemble improves the performance of the prediction framework less on these two data sets. RPI-SE uses PWM to convert a protein sequence into a probabilistic description, which requires that the sequence length be greater than 50. Therefore, sequences less than 50 in length were removed. The performance of machine learning models is highly dependent on the parameter set, while the model parameters of RPI-SE are only adjusted on the RPI369 data set, which makes it not perform optimally on the other two data sets. It uses only simple machine learning models and integration strategies to achieve results that are close to or better than the most advanced models. These results proved it is an acceptable methodological innovation in terms of simplicity and efficiency.

Table 3 Compared RPI-SE with other computational methods on RPI369, RPI488 and RPI1807 data sets

Data sets	Methods	Acc(%)	TPR(%)	TNR(%)	PPV(%)	MCC(%)	AUC
RPI369	IPMiner	75.2	73.5	79.1	71.3	50.7	0.773
	RPISeq-RF	70.4	70.5	70.2	70.7	40.9	0.767
	lncPro	70.4	70.8	69.6	71.3	40.9	0.740
	RPI-SAN	74.9	74.1	78.7	71.7	50.4	0.778
	RPI-SE	88.44	83.69	95.87	80.85	77.73	0.924
RPI488	IPMiner	89.1	93.9	83.1	94.5	78.4	0.914
	RPISeq-RF	88.0	92.6	82.2	93.2	76.2	0.903
	lncPro	87.0	90.0	82.7	91.0	74.0	0.901
	RPI-SAN	89.7	94.3	83.7	95.2	79.3	0.920
	RPI-SE	89.30	94.49	83.48	95.15	79.31	0.904
RPI1807	IPMiner	98.6	98.2	99.3	97.8	97.2	0.998
	RPISeq-RF	97.3	96.8	98.4	96.0	94.6	0.996
	lncPro	96.9	96.5	98.1	95.5	93.8	0.994
	RPI-SAN	96.1	93.6	99.9	91.4	92.4	0.999
	RPI-SE	96.86	96.71	97.69	95.83	93.65	0.994

Conclusion

In this research, we put forward a stacking ensemble computational method, RPI-SE, integrated three individual models, including XGBoost, SVM and ExtraTree, to predict ncRNA-protein interactions using sequence information. PWM and *k*-mer sparse matrix were employed to fully mine efficient features from protein and RNA sequences. The presented method gained a great performance on benchmark data sets. Experimental results prove that the proposed method can accurately and efficiently predict potential ncRNA-protein interactions. RPI-SE uses only simple machine learning models and ensemble learning strategy, and has the advantages of simplicity and interpretability. Meanwhile, RPI-SE has better performance on small data sets, which is in line with the limited scale of the ncRNA-protein interaction data. Although deep learning has been widely adopted in many fields, there is still plenty valuable work worth doing. As a general machine learning model, RPI-SE can perform ncRPI predictions more conveniently and rapidly than complex deep learning models, which can provide useful guidance for ncRPI related biomedical research.

Methods

Data sets

Three benchmark data sets from the previous research, including RPI369, RPI488 and RPI1807, are used to evaluate

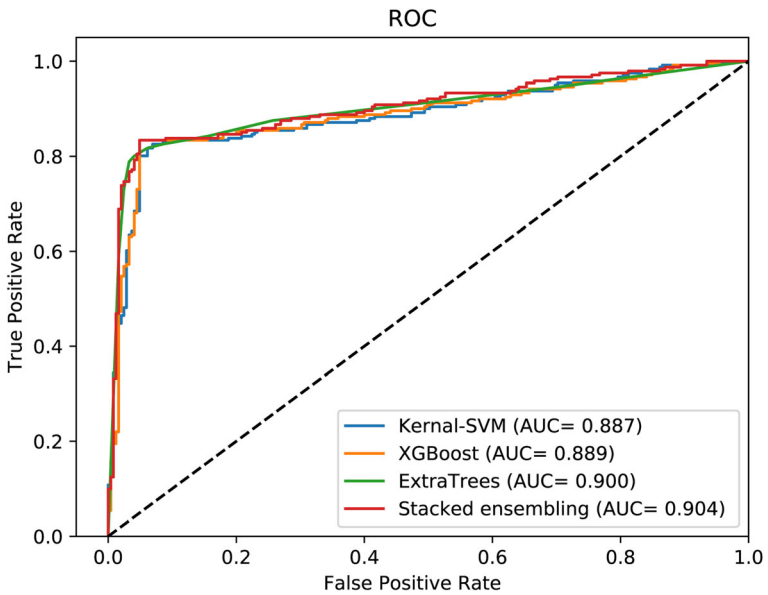

Fig. 4 The performance of RPI-SE and contrast methods on data set RPI488

the performance of RPI-SE. The RPI369 is a non-redundant data set without ribosomal proteins or ribosomal RNAs, from PRIDB [36], which is a comprehensive database calculated from the Protein Data Bank (PDB) [37] of protein-RNA complexes. It includes a total of 332 RNA chains and 338 protein chains, and 369 positive interactive pairs. The RPI488 is a non-redundant lncRNA-protein interactions data set, has 245 negative samples and 243 positive samples [38–40]. The RPI1807 data set includes 1078 RNA and 1807 protein chains. And the number of positive and negative samples is 1807 and 1436, contains 493 RNA and 1436 protein chains. The details of the data sets used in this work are shown in Table 4.

The ncRNA and protein sequences representation

To fully explore the evolutionary features of ncRNA and protein sequences, *k*-mer sparse matrix and position weight matrix are used to represent RNA and protein sequences, respectively. RNA sequence were represented by the *k*-mer sparse matrix [31]. From beginning to end,

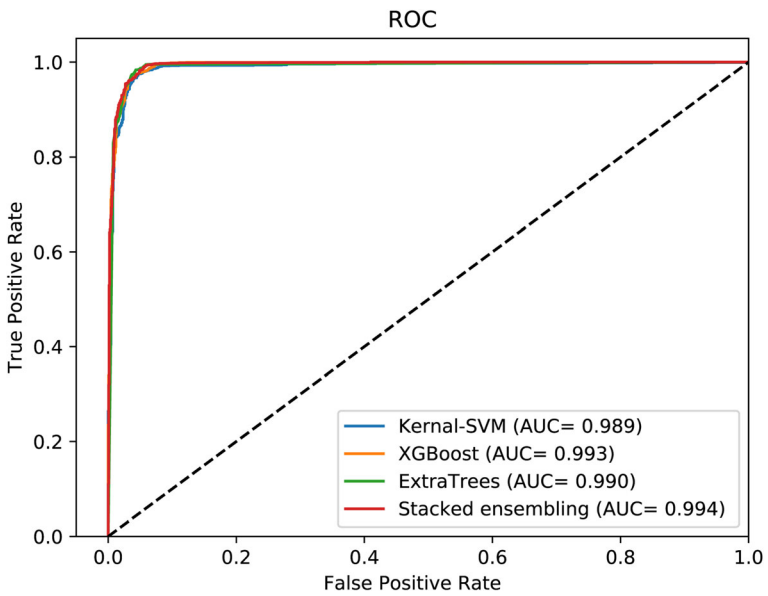

Fig. 5 The performance of RPI-SE and contrast methods on data set RPI1807

Table 4 The details of the RNA-protein interaction data sets

Data set	Interaction pairs	# of proteins	# of RNAs
RPI369	369	338	332
RPI488	243	25	247
RPI1807	1807	1807	1078

it scans each RNA sequence (A, C, G, U) with a k nucleotides window, move one nucleotide at a time. Suppose a RNA sequence with length of L, there are 4^k different possible k-mers and $L - k + 1$ steps.

As shown in Table 5, the dimension of the corresponding k-mer sparse matrix M is $4^k \times (L - k + 1)$. When $m_j m_{j+1} m_{j+2} m_{j+3}$ are same to the i_{th} k-mer among 4^k different k-mers, set the element a_{ij} to 1.

The k-mer sparse matrix M can be defined as follows and the k is set to 4 for RNA sequence.

$$\mathrm{M} = \left(\mathrm{a_{ij}}\right)_{4^k} \times (\mathrm{L}-\mathrm{k}+1) \tag{1}$$

$$a_{ij} = \begin{cases} 1, if\ \mathrm{m_j m_{j+1} m_{j+2} m_{j+3}} = k\text{–}mer(i) \\ 0, else \end{cases} \tag{2}$$

Moreover, the SVD is adopted to reduce M into a 1×256 feature vector.

In consideration of the different structures between RNA and protein sequences have, we employed a more biological method for protein sequences to contain biological evolution information, the position weight matrix (PWM), which is a widely used representation of motifs in biological sequences, to convert it. A PWM has one row for each symbol of the alphabet and 20 rows for amino acids in protein sequences. The PWM of a protein sequence with length of l can defined as follow:

$$PWM = \begin{bmatrix} w_{1,1}, w_{1,2}, \ldots, w_{1,20} \\ w_{2,1}, w_{2,2}, \ldots, w_{2,20} \\ \vdots\vdots\vdots\vdots \\ w_{l,1}, w_{l,2}, \ldots, w_{l,20} \end{bmatrix} \tag{3}$$

In practice, both the Position-Specific Iterated BLAST (PSI-BLAST) tool and against database *SwissProt* can be freely downloaded from http://blast.ncbi.nlm.nih.gov/Blast.cgi. And we set err-value to 0.001, set the value of iteration to 3.

Table 5 *K*-mer sparse matrix representation of RNA sequence

	$R_1R_2R_3R_4$	$R_2R_3R_4R_5$	…	$R_{L-3}R_{L-2}R_{L-1}R_L$
AAAA	a_{11}	a_{12}	…	$a_{1,L-k+1}$
AAAC	a_{21}	a_{22}	…	$a_{2,L-k+1}$
AACA	a_{31}	a_{32}	…	$a_{3,L-k+1}$
…	…	…	…	…
UUUU	$a_{256,1}$	$a_{256,2}$	…	$a_{256,L-k+1}$

Then we extracted Legendre Moment (LMs) [41] feature vectors from the PWM of protein sequence. LMs can exploit eigenvectors of a matrix without losing information, in which the Legendre polynomial is adopted as the kernel function. It is a type of class orthogonal moment, which is widely used in image analysis and pattern recognition.

The 2-D Legendre moments of order (m, n), with image intensity function $f(x; y)$, are defined as:

$$L_{mn} = \mu_{mn} \int_{-1}^{1} \int_{-1}^{1} V_m(x) V_n(x) f(x,y) dx dy \tag{4}$$

where m, $n = 0, 1, 2\ldots$, $\mu_{mn} = (2m + 1)(2n + 1)/4$, and the m_{th} order LMs is given by:

$$V_m(x) = \frac{1}{2^m m!} \frac{d_m}{d_{x^m}} \left(x^2 - 1\right)^m \tag{5}$$

which has the following orthogonality, where ϑ_{mn} represents the Kronecker function.:

$$\int_{-1}^{1} \mathrm{V_m(x) V_n(x)} = \frac{2}{2\mathrm{m}+1} \vartheta_{mn} \tag{6}$$

Hence, a matrix of $\mathrm{R} \times \mathrm{S}$ elements with function $f(i, j)$ can be indicated in discrete form as follow:

$$L_{mn} = \mu_{mn} \sum_{i=1}^{R} \sum_{j=1}^{S} h_{mn}(x,y) f(x,y) \tag{7}$$

For the Legendre polynomials,

$$\int V_m(x) dx = {V_{m+1}(x) - V_{m-1}(x)} \Big/ {2m+1}, x \in [-1, 1] \tag{8}$$

So, according to the above formula, the accuracy expression can be defined as follows.

$$L_{mn} = \mu_{mn} \sum_{i=0}^{R-1} \sum_{j=0}^{S-1} \frac{\Delta(m,x)}{2m+1} \times \frac{\Delta(n,y)}{2n+1} \tag{9}$$

$$\Delta(\mathrm{p,t}) = V_{p+1}\left(t + \frac{\Delta t}{2}\right) - V_{p-1}\left(t + \frac{\Delta t}{2}\right) - V_{p+1}\left(t - \frac{\Delta t}{2}\right) + V_{p-1}\left(t - \frac{\Delta t}{2}\right) \tag{10}$$

Therefore, a PWM of a target protein sequence will be converted into a 1×676 feature vector by using LMs. The truncated SVD was further employed to reduce the influence of noise and retain the principal features. Truncated SVD is very similar to principal component analysis (PCA), but differs in that it works on sample matrices directly instead of their covariance matrices. Contrary to PCA, this estimator does not center the data before computing the singular value decomposition. This means it can work with sparse matrices efficiently. When truncated SVD is applied to term-document matrices, it is known as Latent Semantic Analysis [42].

The feature vectors of the protein will be reduced to 500 dimensions. Finally, each pair of ncRNA-protein contains 1 × 756 conjoined feature vector.

To-be-integrated machine learning classifier

Three kinds of machine learning classifiers are used as to-be-integrated base classifiers, including GBDT [27], SVM [28, 29] and ExtraTree [30].

XGBoost is a scalable end-to-end tree boosting model implementation, which is a great sparsity-aware approach for sparse data and weighted quantile sketch for approximate tree learning. Traditional GBDT only uses first-order derivative information when optimizing. XGBoost performs second-order Taylor expansion for cost function, and uses first and second derivatives. It adds a regularization term in the cost function to control the complexity of the model. A regular term contains the number of leaf nodes of a tree, and the sum of squares of the L2 modules of score on each leaf node. From the Bias-variance tradeoff point of view, the regular term reduces the variance of the model, making the learning model simpler and preventing over fitting, which is also a characteristic of its superior to the traditional GBDT. After iteration, XGBoost multiplied the weight of the leaf node, mainly to weaken the impact of each tree, and let the behind have a larger learning space. XGBoost draws on the practice of random forest and supports column sampling, which not only reduces over-fitting but also reduces calculations. A parallel approximate histogram algorithm is also proposed to generate candidate segmentation points efficiently. XGBoost's objective function can be defined as follows:

$$\mathrm{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{11}$$

$$\Omega(f_t) = \gamma \mathrm{T} + \frac{\lambda}{2} \sum_{j=1}^{T} w_j^2 \tag{12}$$

Here, l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i. The regular term controls the complexity of the model, including the number of leaf nodes T and the l_2 modulus square of the leaf score.

SVM constructs a hyperplane or a series of hyperplanes in a high-dimensional or infinite-dimensional space that can be used for classification, regression, or other tasks. Intuitively, by using a hyperplane to achieve a good segmentation, it is possible to maximize the distance between the closest training data points (function margins) in any class. This is usually due to a larger margin. The advantages of support vector machines are: It's very efficient in high dimensional space. Even if the data dimension is larger than the sample size, it is still valid. The subset of training sets is used in support vectors, so it is also efficient in memory utilization. The disadvantages of support vector machines include: If the number of features is much larger than the number of samples, it is necessary to avoid overfitting when selecting kernel functions.

Suppose the labeled training data $[(x_i, y_i), i = 1, 2, 3\ldots, n, y_i = (-1, 1), x_i \in \mathrm{R}]$. and the separating hyperplane is: $(w(x) + b) = 0$. In the linear separable situation, the SVM maximized the margin by minimizing $\|w\|^2/2$ subject to looking for the separating hyperplane as following constraint:

$$y_i(w_{x_i} + \mathrm{b}) \geq 1, \forall x_i \tag{13}$$

In the linear non-separable situation, we can find the optimal separating hyperplane by introducing slack variables: ξ_i, $i = 1, 2\ldots, n$ and user-adjustable parameter C, then minimizing:

$$\|w\|^2/2 + C\sum_{i=1}^{n} \xi_i, \xi_i \geq 0, \forall x_i \tag{14}$$

$$y_i(w_{x_i} + \mathrm{b}) \geq 1 - \xi_i, \xi_i \geq 0, \forall x_i \tag{15}$$

Radial Basis Function (RBF) kernel is adopted in this experiment, which can be defined as:

$$f(\mathrm{x}) = e^{-\gamma \left\| x - x' \right\|^2} \tag{16}$$

Extremely randomized trees essentially consist of randomizing strongly both attribute and cut-point choice while splitting a tree node. It builds totally randomized trees whose structures are independent of the output values of the learning sample. The strength of the randomization can be tuned to problem specifics by the appropriate choice of a parameter. Randomness in the computation of segmentation points is further enhanced. In a random forest, a random subset of the candidate features is used in a random forest. Unlike a threshold for finding the most regional diversity, the threshold here is randomly generated for each candidate feature and selects the best one of these randomly generated thresholds as a segmentation rule. This method usually reduces the variance of one-point model, while the cost slightly increases the deviation.

Implementation of stacking ensemble integration strategy

The Logistic Regression (LR) is used as the merge layer to integrate three base classifiers' output, which can learn the integration weight w for each base classifier. The predicted probability value outputs of individual

classifiers be the level 0 layer, while successive logistic regression was the level 1. The definition of LR is:

$$P_w(\pm 1|p) = \frac{1}{1 + e^{-w^T p(\pm 1|p)}} \quad (17)$$

where the p is the level 0 classifiers' probability outputs and it will degenerate to average strategy when the weight for each individual classifier of logistic regression is judged as the same.

Performance evaluation indicators

The evaluation of the experiments in this work was performed under five-fold cross-validation. In each validation, all data randomly divides into five equal subsets, four-fold data are used for training, and the rest one-fold is used for testing. There is no overlap between train data and test data. The average performances of five-fold are taken as the final validation performance. The evaluation indicators used in the experiments can be defined as:

$$\mathrm{Acc} = \frac{TN + TP}{TN + TP + FN + FP} \quad (18)$$

$$\mathrm{TPR} = \frac{TP}{TP + FN} \quad (19)$$

$$\mathrm{TNR} = \frac{TN}{TN + FP} \quad (20)$$

$$\mathrm{PPV} = \frac{TP}{TP + FP} \quad (21)$$

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (22)$$

where TN, TP, FN, and FP indicates the number of true negative, true positive, false negative and false positive samples.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3406-0.

Additional file 1: Table S1. The 5-fold cross-validation details on RPI488 dataset. **Table S2.** Performance of individual predictors and RPI-SE on RPI488 dataset. **Table S3.** The 5-fold cross-validation details on RPI1807 dataset. **Table S4.** Performance of individual predictors and RPI-SE on RPI1807 dataset.

Abbreviations

Acc: Accuracy; AUC: Area Under the receiver operating characteristic Curve; ExtraTree: Extremely randomized Trees; GBDT: Gradient Boosting Decision Tree; LMs: Legendre Moments; LR: Logistic Regression; MCC: Matthews Correlation Coefficient; ncRNA: non-coding RNA; ncRPI: non-coding RNA-protein interaction; PCA: Principal Component Analysis; PDB: Protein Data Bank; PPV: Positive Predictive Value; PSI-BLAST: Position-Specific Iterated BLAST; PWM: Position Weight Matrix; RF: Random Forest; RPI: RNA-protein interaction; SVD: Singular Value Decomposition; SVM: Support Vector Machine; TNR: True Negative Rate; TPR: True Positive Rate

Acknowledgements

The authors would like to thank all the editors and anonymous reviewers for their constructive advices.

Authors' contributions

HCY and ZHY conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript; MNW, ZHG, YBW, and JRZ designed, performed and analyzed experiments and wrote the manuscript; The author(s) read and approved the final manuscript.

Funding

This work is supported by the National Natural Science Foundation of China under Grants 61722212, 61873212, 61861146002, and 61732012. The funders have no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript.

Availability of data and materials

The datasets generated and/or analysed during this study are available under open licenses in the data repository, https://github.com/haichengyi/RPI-SE.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China. [2]University of Chinese Academy of Sciences, Beijing 100049, China. [3]School of Mathematics and Computer Science, Yichun University, Yichun 336000, China.

Received: 18 July 2019 Accepted: 11 February 2020
Published online: 18 February 2020

References

1. Taft RJ, Pheasant M, Mattick JS. The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays. 2007;29(3):288–99.
2. Esteller M. Non-coding RNAs in human disease. Nat Rev Genet. 2011;12(12):861.
3. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H. starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. 2013;42(D1):D92–7.
4. Poole AM, Jeffares DC, Penny D. The path from the RNA world. J Mol Evol. 1998;46(1):1–17.
5. Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet. 2016;17(1):47.
6. Slack FJ, Chinnaiyan AM. The role of non-coding RNAs in oncology. Cell. 2019;179(5):1033–55.
7. Wang L, You Z-H, Huang D-S, Zhou F. Combining high speed ELM learning with a deep convolutional neural network feature encoding for predicting protein-RNA interactions. IEEE/ACM Trans Comput Biol Bioinform. 2018:1.
8. Shahrouki P, Larsson E. The non-coding oncogene: a case of missing DNA evidence? Front Genet. 2012;3:170.
9. Sahoo T, del Gaudio D, German JR, Shinawi M, Peters SU, Person RE, Garnica A, Cheung SW, Beaudet AL. Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. Nat Genet. 2008;40(6):719–21.
10. Cook EH Jr, Scherer SW. Copy-number variations associated with neuropsychiatric conditions. Nature. 2008;455(7215):919–23.
11. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, Finch CE, Laurent GS, Kenny PJ, Wahlestedt C. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase expression. Nat Med. 2008;14(7):723–30.
12. Ridanpää M, van Eenennaam H, Pelin K, Chadwick R, Johnson C, Yuan B, vanVenrooij W, Pruijn G, Salmela R, Rockas S, et al. Mutations in the RNA

component of RNase MRP cause a pleiotropic human disease, Cartilage-Hair Hypoplasia. Cell. 2001;104(2):195–203.
13. Lewis MA, Quint E, Glazier AM, Fuchs H, De Angelis MH, Langford C, van Dongen S, Abreu-Goodger C, Piipari M, Redshaw N, et al. An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. Nat Genet. 2009;41:614.
14. Lu Q, Ren S, Lu M, Zhang Y, Zhu D, Zhang X, Li T. Computational prediction of associations between long non-coding RNAs and proteins. BMC Genomics. 2013;14(1):651.
15. Bellucci M, Agostini F, Masin M, Tartaglia GG. Predicting protein associations with long noncoding RNAs. Nat Methods. 2011;8(6):444–5.
16. Agostini F, Zanzoni A, Klus P, Marchese D, Cirillo D, Tartaglia GG. cat RAPID omics: a web server for large-scale prediction of protein–RNA interactions. Bioinformatics. 2013;29(22):2928–30.
17. Livi CM, Blanzieri E. Protein-specific prediction of mRNA binding using RNA sequences, binding motifs and predicted secondary structures. BMC Bioinformatics. 2014;15(1):123.
18. Yi H-C, You Z-H, Huang D-S, Li X, Jiang T-H, Li L-P. A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information. Mol Ther Nucleic Acids. 2018;11:337–44.
19. Pancaldi V, Bähler J. In silico characterization and prediction of global protein–mRNA interactions in yeast. Nucleic Acids Res. 2011;39(14):5826–36.
20. Muppirala UK, Honavar VG, Dobbs D. Predicting RNA-protein interactions using only sequence information. Bmc Bioinformatics. 2011;12(1):489.
21. Suresh V, Liu L, Adjeroh D, Zhou X. RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. Nucleic Acids Res. 2015;43(3):1370–9.
22. Cirillo D, Blanco M, Armaos A, Buness A, Avner P, Guttman M, Cerase A, Tartaglia GG. Quantitative predictions of protein interactions with long noncoding RNAs. Nat Methods. 2016;14(1):5.
23. Pan X, Fan YX, Yan J, Shen HB. IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. BMC Genomics. 2016;17(1):582.
24. Wang L, You Z-H, Chen X, Xia S-X, Liu F, Yan X, Zhou Y. Computational methods for the prediction of drug-target interactions from drug fingerprints and protein sequences by stacked auto-encoder deep neural network. In: International Symposium on Bioinformatics Research and Applications. Cham: Springer; 2017. p. 46–58.
25. Yi H-C, You Z-H, Cheng L, Zhou X, Jiang T-H, Li X, Wang Y-B. Learning distributed representations of RNA and protein sequences and its application for predicting lncRNA-protein interactions. Comput Struct Biotechnol J. 2020;18:20–6.
26. Yi H-C, You Z-H, Zhou X, Cheng L, Li X, Jiang T-H, Chen Z-H. ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. Mol Ther Nucleic Acids. 2019;17:1–9.
27. Chen T, Guestrin C. XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16; 2016. p. 785–94.
28. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
29. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol. 2011;2(3):1–27.
30. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63(1):3–42.
31. You Z-H, Zhou M, Luo X, Li S. Highly efficient framework for predicting interactions between proteins. IEEE Trans Cybern. 2016;47(3):731–43.
32. Yi H-C, You Z-H, Guo Z-H. Construction and Analysis of Molecular Association Network by Combining Behavior Representation and Node Attributes. Front Genet. 2019;10:1106.
33. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics. 2005;6(1):33.
34. cheol Jeong J, Lin X, Chen X-W. On position-specific scoring matrix for protein function prediction. IEEE/ACM Trans Comput Biol Bioinformatics. 2011;8(2):308–15.
35. De Lathauwer L, De Moor B, Vandewalle J. A multilinear singular value decomposition. SIAM J Matrix Anal Appl. 2000;21(4):1253–78.
36. Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, Honavar V, Dobbs D. PRIDB: a protein–RNA interface database. Nucleic Acids Res. 2010;39(suppl_1):D277–82.
37. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235.
38. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM. Computational methods for prediction of protein–RNA interactions. J Struct Biol. 2012;179(3):261.
39. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26(5):680–2.
40. Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, Honavar V, Dobbs D. PRIDB: a protein-RNA interface database. Nucleic Acids Res. 2011;39(Database issue):D277.
41. Zhang H, Shu H, Coatrieux G, Zhu J, Wu QM, Zhang Y, Zhu H, Luo L. Affine Legendre moment invariants for image watermarking robust to geometric distortions. IEEE Trans Image Process. 2011;20(8):2189–99.
42. Deerwester S. Indexing by latent semantic analysis. J Am Soc Inf Sci. 1990;41(6):391–407.

Publisher's Note