Computational and Structural Biotechnology Journal 18 (2020) 20–26

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Learning distributed representations of RNA and protein sequences and its application for predicting lncRNA-protein interactions

Hai-Cheng Yi [a,b], Zhu-Hong You [a,*], Li Cheng [a], Xi Zhou [a], Tong-Hai Jiang [a], Xiao Li [a], Yan-Bin Wang [a]

[a] *The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China*
[b] *University of Chinese Academy of Sciences, Beijing 100049, China*

ARTICLE INFO

Article history:
Received 11 April 2019
Received in revised form 19 October 2019
Accepted 15 November 2019
Available online 30 November 2019

Keywords:
Distribution representation
Natural language processing
Word2vec
RNA-protein interaction

ABSTRACT

The long noncoding RNAs (lncRNAs) are ubiquitous in organisms and play crucial role in a variety of biological processes and complex diseases. Emerging evidences suggest that lncRNAs interact with corresponding proteins to perform their regulatory functions. Therefore, identifying interacting lncRNA-protein pairs is the first step in understanding the function and mechanism of lncRNA. Since it is time-consuming and expensive to determine lncRNA-protein interactions by high-throughput experiments, more robust and accurate computational methods need to be developed. In this study, we developed a new sequence distributed representation learning based method for potential lncRNA-Protein Interactions Prediction, named LPI-Pred, which is inspired by the similarity between natural language and biological sequences. More specifically, lncRNA and protein sequences were divided into *k*-mer segmentation, which can be regard as "word" in natural language processing. Then, we trained out the RNA2vec and Pro2vec model using word2vec and *human* genome-wide lncRNA and protein sequences to mine distribution representation of RNA and protein. Then, the dimension of complex features is reduced by using feature selection based on Gini information impurity measure. Finally, these discriminative features are used to train a Random Forest classifier to predict lncRNA-protein interactions. Five-fold cross-validation was adopted to evaluate the performance of LPI-Pred on three benchmark datasets, including RPI369, RPI488 and RPI2241. The results demonstrate that LPI-Pred can be a useful tool to provide reliable guidance for biological research.

1. Introduction

The emerging recognition of RNA is that any transcripts, regardless of protein coding potential, can have intrinsic functions [1]. One kind of this transcripts that are no less than 200 nucleotides, known as long non-coding RNA (lncRNA). Existing studies demonstrate that only less than 2% of the human genome can be translated into proteins, whereas more than 80% of it has biochemical functions [2,3]. Furthermore, more than 70% of ncRNA are long ncRNA [4], which means there is massive of precious information lncRNAs contained awaiting our effective mining. The lncRNA often act through functions by binding to partner proteins, and play critical roles in gene regulation, splicing, translation, chromatin modification and poly-adenylation [5–8]. Moreover, emerging evidences have revealed that various complex diseases have strong correlation with lncRNAs, such as Alzheimer [9], lung cancer [10] and cardiovascular diseases [11]. Therefore, the basis for understanding the functions of lncRNA is to identify lncRNA-protein interactions. It's inefficient to examine a large number of under-researched lncRNAs and proteins though wet experiments.

Due to the time-consuming and laborious of high throughput experiments, such as CLIP-seq, RIP-seq and fRIP-seq [12], several computational lncRNA-protein interaction prediction methods have been put forward in recent years, which can be used as guide tools for biological experiments. These methods can be divided into two categories. The first kind of methods mainly based on sequence information, structural information, evolutionary knowledge or physicochemical properties to exploit discriminative features of lncRNA and protein. For instance, Muppirala et al. proposed RPISeq, which adopted k-mer composition to encode RNA and protein sequences and trained support vector machine (SVM) and Random Forest (RF) model to identify interactions [13]. Suresh et al. used sequence information and structure information to build a SVM predictor to predict novel protein-RNA interactions, named PRI-Pred [14]. Bellucci et al. developed catPA-

[*] Corresponding author.
E-mail address: zhuhongyou@ms.xjb.ac.cn (Z.-H. You).

https://doi.org/10.1016/j.csbj.2019.11.004
2001-0370/© 2019 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

PID by using the physicochemical properties of nucleotide and polypeptide, include secondary structure, Van der Waals propensities and hydrogen bonding, to evaluate the interaction propensities, and they further applied this model to predicted protein interactions in the Xist network [15,16]. Lu et al. scored RNA-protein pair by using matrix multiplication and Fisher's linear discriminant. More recently, Yi et al. presented a deep learning framework RPI-SAN, using stacked autoencoder to extract high-level hidden feature from sequence, then they trained RF classifier and ensemble strategy to robustly and accurately predict ncRNA-protein interactions [17]. These methods suggested that the sequence carried enough information for prediction tasks.

There is another category of methods in this domain, which considered the known interactions between lncRNA and protein. Yun et al. considered the relatedness of heterogeneous objects path-constrained, introduced a method using HeteSim measure to compute the relatedness score, called PLPIHS [18]. Zhang et al. using graph regularized nonnegative matrix factorization to discover unknown interacted pairs based on the hypothesis that similar lncRNAs (proteins) have similar corresponding proteins (lncRNAs) [19]. Shen et al. proposed LPI-KTASLP to identify lncRNA-protein interactions with kernel target alignment and semi-supervised link prediction model using multivariate information [20]. Zhang et al. combined multiple sequence-based features and lncRNA-lncRNA similarities and protein-protein similarities, which is calculated by using RNA sequences and protein sequences and known lncRNA-protein interactions [21]. But these kind methods have limitations when predicting new samples, especially those never appeared in the similarity matrices.

This paper aims to develop a new sequence distributed representation learning based method for novel **l**ncRNA-**P**rotein **I**nteractions **Pred**iction, named **LPI-Pred**, which is inspired by the similarity between biological sequences and natural languages [22]. More specifically, lncRNA and protein sequences were divided into *k*-mer segmentation, which can be regard as "word" in natural language processing. Furthermore, we trained the RNA2vec and Pro2vec model using skip-gram word embedding model and *Human* genome-wide lncRNA and protein sequences for lncRNA and protein, respectively. The aforementioned train sequences data are provided by the GENCODE project (release v29) [23]. And then, we measured the importance of features via Gini information impurity, and select top-50 feature as final discriminative features. Finally, these features are used to train RF predictor. We evaluated our model on three benchmark datasets under five-fold cross-validation, including RNA-protein interaction datasets, RPI369 and RPI1807, and lncRNA-protein interaction dataset, RPI488, using six widely used evaluation indicators in machine learning field. And we compared our model with other state-of-the-art models such as RPISeq [13], lncPro [24], and RPI-SAN [17]. The rigorous experimental results prove the validity and reliability of our method.

2. Materials and methodology

2.1. Datasets exploration

In practice, three benchmark datasets, including RPI369 [13], RPI1807 [14] and RPI488 [25] were selected to execute our evaluation. The first two are RNA-protein interactions datasets, while the third is lncRNA-protein interactions dataset. The RPI369 dataset is a non-redundant data set, which is generated from RPIDB [26], and only have non-ribosomal complexes (e.g., mRNA, miRNA, tRNA). The dataset RPI369 contains 332 RNA sequences, 338 protein sequences and 369 positive interaction pairs. In the same work, the authors also constructed another dataset RPI2241, which is larger than RPI369 but is strongly biased to ribosomal RNA-protein interactions. That's why we're not inclined to adopt it. The RPI1807 also is a non-redundant data set of RNA-protein interactions complexes, generated by parsing the RPIDB and Nucleic Acid Database (NDB) [24]. There are 1078 RNA sequences and 1807 protein sequences in RPI1807, consisting 1807 pairs positive samples and 1436 pairs negative samples. The RPI488 is a lncRNA-protein interactions dataset, contains 245 negative lncRNA-protein pairs, 243 interacted lncRNA-protein pairs. The number of lncRNA and protein in this dataset are 25 and 247, respectively. The details of these three benchmark datasets are listed in Table 1 as below:

2.2. k-mer segmentation

In this section, we will introduce the feature representation scheme used in this study, which is aims to fully exploit the hidden high-level feature from the sequence information. For a given lncRNA or protein sequence, *k*-mer composition is used to spilt them into subsequences, which can be considered as "word" in the fellow step. Scan a sequence from beginning to the end, one nucleic acid once time. For a given sequence of length L, we will obtain $L-k+1$ *k*-mers, and the count of possible *k*-mer are 4^k for RNA (A, C, G, U) and 20^k for protein (Ala, Gly, Val, Ile, Leu, Phe, Pro, Tyr, Met, Thr, Ser, His, Asn, Gln, Tpr, Arg, Lys, Asp, Glu, Cys), different from common usage, we do not use the 7-letter reduce alphabet, which reduced 20 amino acids into 7 groups based on their similarity of dipole moments and side chain volume. We set the k to 4 for lncRNA and set k to 3 for protein, which are two commonly accepted empirical parameters [13,17,25,27]. The process of splitting nucleic acids sequence and amino acids sequences into *k*-mers shown in Fig. 1.

2.3. Distribution representation of lncRNA and protein sequences

And then, we using the genome-wide human lncRNA and protein sequences to train a word embedding model, named RNA2vec and pro2vec, respectively. The training data provided by the GENCODE project and their goal of this project is to identify and classify all gene features in the human and mouse genomes with high accuracy based on biological evidence, and to release these annotations [23,28]. We use the skip-gram [29,30] word representation model to learn distribution representation of RNA and protein sequences. In nature, the model is a neural network with projection layer for learning word representation. The structure of skip-gram is shown in Fig. 2 below.

For a given sequence $(w_1, w_2, \ldots, w_{l-k+1})$, the goal of training model is to maximize the mean log probability:

$$\max \frac{1}{N}\sum_{n=1}^{N}\sum_{-c\le m\le c, m\ne 0} logP(w_{n+m}|w_n) \tag{1}$$

c stands for the distance to the central word; the log probability distribution can be defined as follow:

$$\log P(w_o|w_i) = log\frac{e^{v'^{T}_{w_o} v_{w_i}}}{\sum_{w=1}^{W} e^{v'^{T}_{w} v_{w_i}}} \tag{2}$$

where the v_w and v'_w are the input and output vector of word w, respectively. W is the size of training lncRNA or protein training lexicon.

In natural language processing, the word embedding model has achieved great success [31,32], it has also made progress in computational biology [33–35]. In this work, we regard each k-mer as a word and a sequence as a sentence, and then learning the distribution representation by using skip-gram word2vec model. The procedure for training RNA2vec and pro2vec is shown as Fig. 3.

Table 1
The details of two RNA-protein interactions datasets RPI369 and RPI1807 and lncRNA-protein interactions dataset RPI488.

Datasets	# of RNAs	# of proteins	Positive samples	Negative samples	References
RPI369	332	338	369	369	[13]
RPI1807	1078	1807	1807	1807	[14]
RPI488	25	247	243	245	[25]

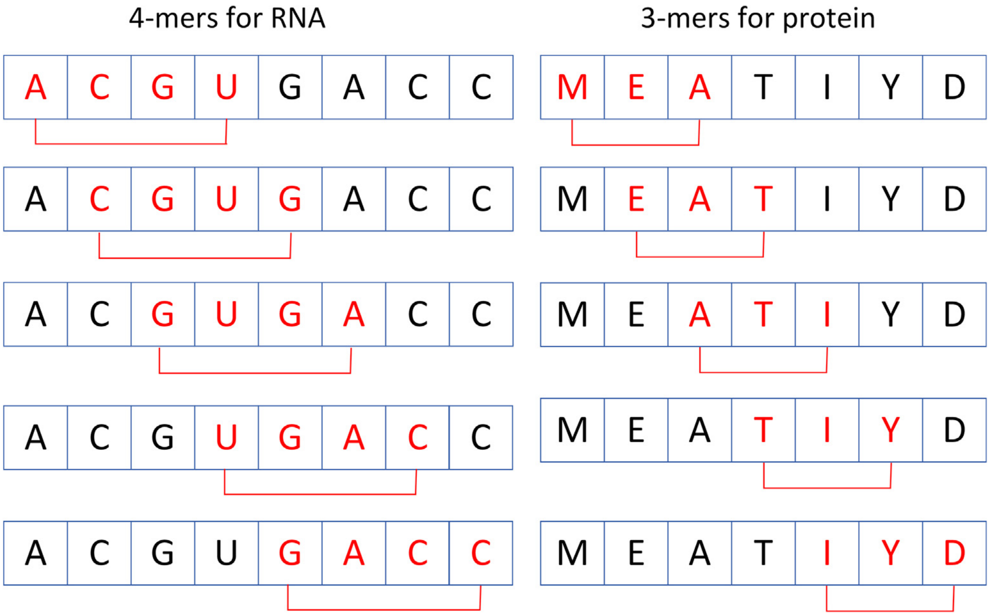

Fig. 1. Procedure of splitting RNA nucleotides and protein amino acids sequences into smaller *k*-mers.

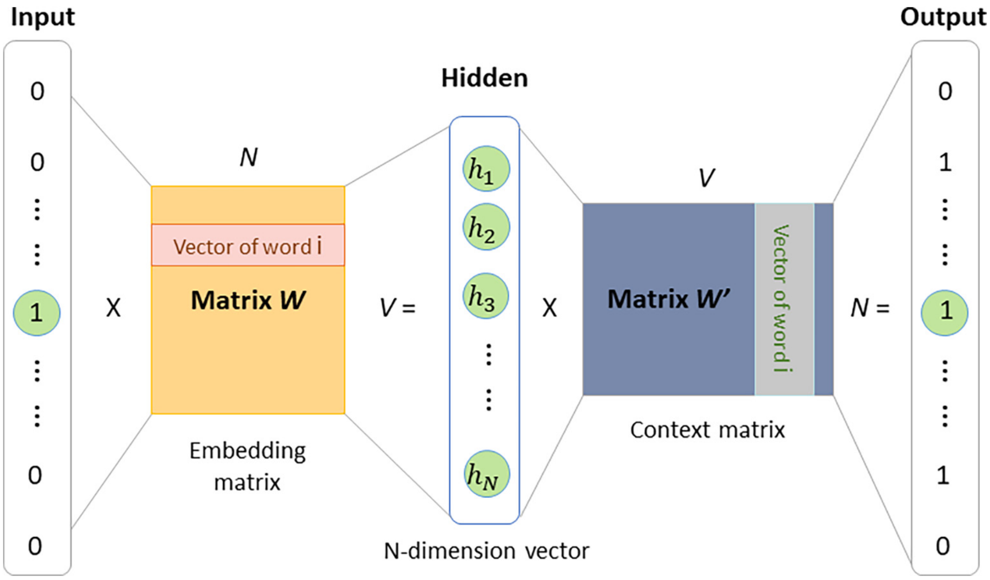

Fig. 2. The skip-gram word embedding model. Lnc2vec and pro2vec model were trained by using this model and genome-wide human lncRNA and protein sequences. Skip-gram is trained by predicting words surrounding the central word, after training, the weights matrix *W* of the hidden layer is obtained, that is word vectors.

The parameters of the model are *min_count* = 1, *size* = 300, *window* = 5, *iter* = 10, *batch_words* = 100. Where the *size* represents the dimensions of output word vector, and *window* stands for maximum distance between the current and predicted word within a sentence, *iter* is the count of iterations (epochs) over the corpus, *batch_words* is the target size (in words) for batches of examples passed to worker threads. When the *min_count* (means minimum word frequency) is set too high, the model only counts high-frequency words, which is not conducive to learning discriminative word vectors from sequence representation. Other parameters are default. Inspired by the additivity of word embedding [30], we represented a given sequence by summing all its *k*-mer word embed-

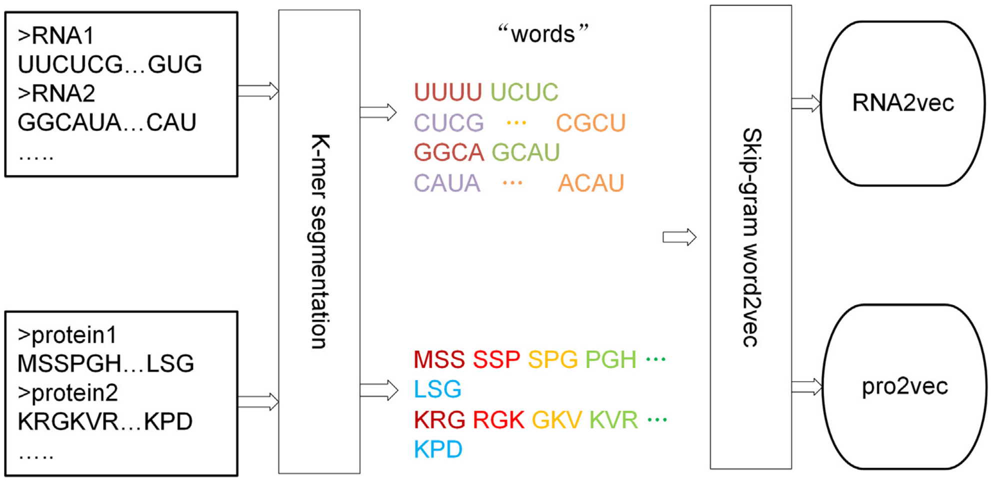

Fig. 3. . The procedure for training RNA2vec and pro2vec. The corpus of RNA and protein sequences obtained from GENCODE project. And the model implemented by word2vec.

dings. Here, we obtained the word embedding feature as base feature. The procedure for training RNA2vec and pro2vec is shown as Fig. 3.

2.4. Gini information impurity-based feature selection

A data set often has hundreds of previous features. How to choose the features that have the greatest impact on the results, so as to reduce the number of features when building the model. There are many such methods, for instance, principal component analysis, Lasso [36,37], mRMR [38] and so on. However, here we are going to introduce the use of Random Forest to feature screening based on Gini information impurity.

Assuming that there are m features $f_1, f_2, f_{\dots}, f_m$, we can calculate the Variable Importance Measures (*VIM*) by the Gini index $VIM_i^{(Gini)}$ for each feature f_i, that is, the average change of node splitting impurity in all RF decision trees by f_i feature. The Gini index (*GI*) can be defined as:

$$GI_i = \sum_{k=1}^{|K|} \sum\nolimits_{k \neq k'} p_{ik} p_{ik'} = 1 - \sum_{k=1}^{|K|} p_{ik}^2 \tag{3}$$

where the K means there are k categories, and p_{ik} indicates the proportion of categories k in i_{th} node. The *VIM* of feature f_i in j_{th} node can be computed from the variation of *GI* before and after branching of j_{th} node:

$$VIM_{ij}^{(Gini)} = GI_i - GI_r - GI_l \tag{4}$$

Among them, GI_r and GI_l respectively represent the *GI* of the right and left nodes after branching. Suppose there are N decision trees, so:

$$VIM_i^{(Gini)} = \sum\nolimits_{n=1}^{N} VIM_{ij}^{(Gini)} \tag{5}$$

Finally, all the obtained importance scores can be normalized by:

$$VIM_i = \frac{VIM_i}{\sum_{j=1}^{c} VIM_j} \tag{6}$$

Here, we selected the most important top-50 features as final feature.

2.5. Training an LPI-Pred model

The selected top-50 feature would be used to train an LPI-Pred model for predicting potential lncRNA-protein interactions on test data set. In summary, the procedure for training an LPI-Pred is shown in Fig. 4:

- Using *human* genome-wide lncRNA and protein sequences as corpus, segment them into k-mers as the words.
- Using word2vec model to train out RNA2vec and pro2vec for lncRNA and protein sequence distribution representation.
- Obtaining the word embedding of the protein and ncRNA sequences in the benchmark RNA-protein interaction datasets.
- Select top-50 features based on feature importance to train Random Forest predictor.

2.6. Performance evaluation metrics

In this study, we proposed a novel lncRNA-protein interactions prediction model LPI-Pred, based on sequence distributed representation learning and Gini information impurity measure. The common metrics and five-fold cross-validation are used to evaluate the performance of LPI-Pred. Divided all data into five equal sub-set. For each training, one-fold set data is taken as test data, the rest four-fold are taken as training data. Take the mean performance metrics of five training as final performance. There is no overlap between train data and test data, and this is unbiased comparison. The metrics used in performance evaluation including accuracy (Acc), Sensitivity (Sens), Specificity (Spec), Precision (Pre) and Matthews Correlation Coefficient (MCC). Certainly, and the area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curve are also adopted to evaluate the performance. These metrics can be defined as:

$$\text{Acc} = \frac{TN + TP}{TN + TP + FN + FP} \tag{7}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{8}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

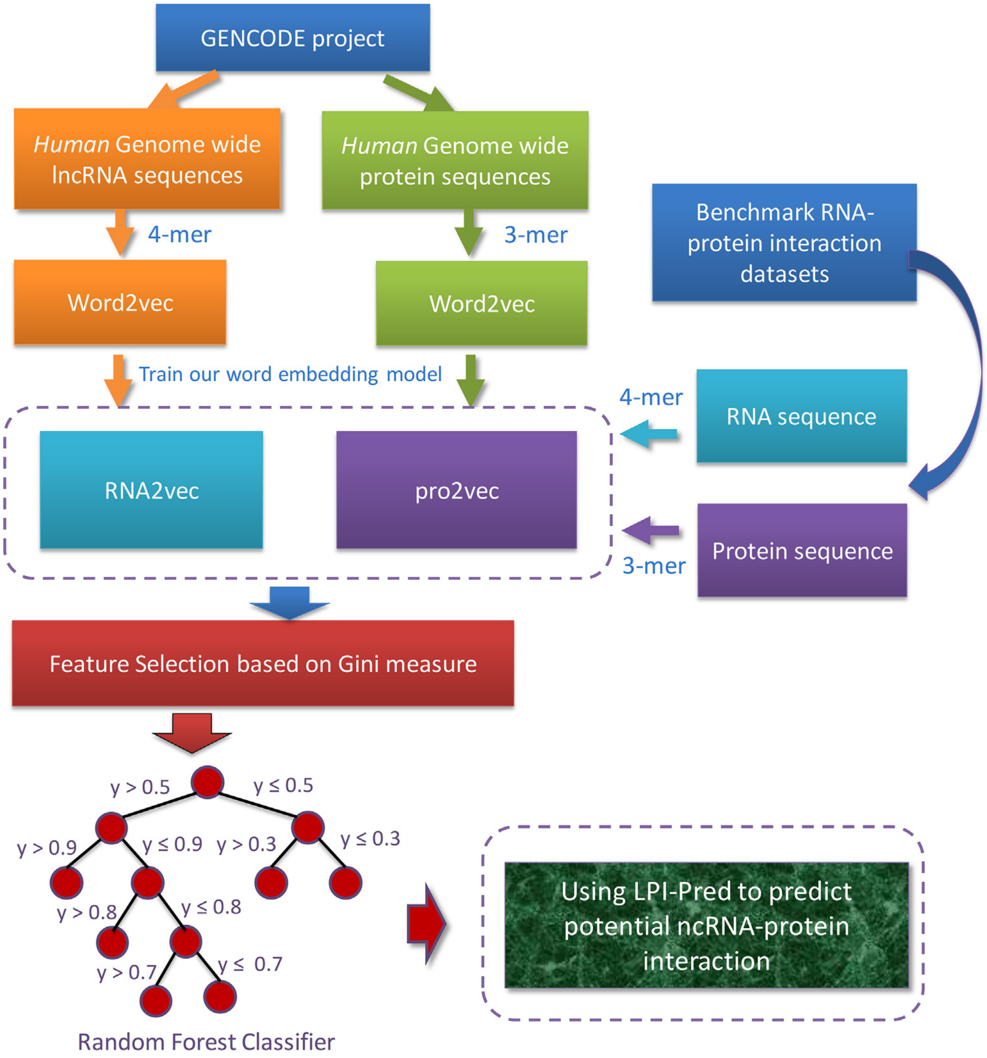
GENCODE project
Human Genome wide lncRNA sequences
Human Genome wide protein sequences
4-mer
3-mer
Word2vec
Word2vec
Benchmark RNA-protein interaction datasets
Train our word embedding model
4-mer
RNA sequence
RNA2vec
pro2vec
Protein sequence
3-mer
Feature Selection based on Gini measure
y > 0.5
y ≤ 0.5
y > 0.9
y ≤ 0.9
y > 0.3
y ≤ 0.3
y > 0.8
y ≤ 0.8
y > 0.7
y ≤ 0.7
Random Forest Classifier
Using LPI-Pred to predict potential ncRNA-protein interaction

Fig. 4. The workflow of LPI-Pred. The word embedding model RNA2vec and pro2vec are trained to obtain the sequence information of RNA and protein, and these features after feature selection are used to train Random Forest predictor.

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{11}$$

where *TN*, *TP* indicates the correctly predicted negative samples and positive samples number, *FN*, *FP* represents the false wrongly predicted negative and positive samples number.

3. Results and discussion

In this study, we proposed a novel lncRNA-protein interactions prediction model LPI-Pred, based on sequence distributed representation learning and Gini information impurity measure. In this section, we designed the following experiments to verify the performance of the model. First, we compared the effects of different sequence coding schemes on lncRNA-protein interaction dataset, and the effect of feature selection. Second, we did a performance comparison with different individual predictors. And then, we verify LPI-Pred's ability to predict lncRNA-protein interactions and compared with other state-of-the-art methods. Final, we apply our model to lncRNA-protein interactions network construction.

3.1. Comparison between different sequences encoding strategies

We applied a new RNA and protein sequences encoding method in this work, using skip-gram distribution representation model. In order to verify the effectiveness of this sequence numerical coding scheme, we first compare it with the widely used *k*-mer frequency on three benchmark datasets. The comparison results are shown in Table 2.

In all three gold standard datasets, the selected word embedding feature, obtained though RNA2vec and pro2vec model, have improved performance compared to *k*-mer method. This can prove that distribution representation word vector is effectiveness for biological sequences encoding, for RNA and protein. It can achieve and even exceed the performance of *k*-mer, which is very widely used in biological sequence representation. The comparison between LPI-Pred (using RNA2vec and pro2vec with feature selection) and LPI-Pred without feature selection demonstrate the necessity of feature selection.

3.2. Comparison with individual predictors

To verify the effect of RF classifier separately, we compared RF and other machine learning modals including SVM (with RBF kernel), Logistic Regression (LR), under same set of features and the same experimental conditions. These models were trained with default parameters. The results are shown in Table 3:

Several Random Forest-based methods have achieved remarkable performance on many issues in the field of computational biology. We trained LPI-Pred based on random forest classifier.

Table 2
Comparing the five-fold cross-validation performance of *k*-mer and word embedding with and without feature selection on three gold standard datasets.

Datasets	feature	Acc (%)	Sens (%)	Spec (%)	Pre (%)	MCC (%)
RPI369	*k*-mer	68.71	67.29	70.30	69.88	37.74
	embedding without feature selection	71.97	70.27	**73.76**	**73.19**	44.24
	embedding with feature selection	**73.06**	**75.32**	71.14	72.64	**46.67**
RPI488	*k*-mer	89.29	**83.17**	95.17	94.33	79.09
	embedding without feature selection	87.64	83.17	91.93	90.82	75.52
	embedding with feature selection	**89.92**	82.75	**96.72**	**96.32**	**80.59**
RPI1807	*k*-mer	96.88	**98.44**	94.96	96.04	93.72
	embedding without feature selection	96.73	97.90	95.28	96.28	93.37
	embedding with feature selection	**97.10**	97.89	**96.14**	**96.91**	**94.13**

The boldface indicates this measure performance is the best among the compared sequence feature encoding.

Table 3
Comparing the five-fold cross-validation performance of LPI-Pred and other machine learning classifiers on three gold standard datasets.

Datasets	Methods	Acc (%)	Sens (%)	Spec (%)	Pre (%)	MCC (%)
RPI369	SVM	65.17	66.20	64.34	65.48	30.61
	LR	58.37	44.06	**73.12**	62.51	18.05
	LPI-Pred	**73.06**	**75.32**	71.14	**72.64**	**46.67**
RPI488	SVM	88.68	81.97	95.17	94.26	77.95
	LR	88.68	81.97	95.17	94.26	77.95
	LPI-Pred	**89.92**	**82.75**	**96.72**	**96.32**	**80.59**
RPI1807	SVM	92.35	94.11	90.17	92.29	84.52
	LR	87.26	90.17	83.56	87.39	74.17
	LPI-Pred	**97.10**	**97.89**	**96.14**	**96.91**	**94.13**

The boldface indicates this measure performance is the best among the compared methods for individual dataset.

As shown in the comparison results in the above table, LPI-Pred outperformed all other classifiers using same feature set and under same experimental conditions.

3.3. Evaluation of LPI-Pred's capability to predict lncRNA-protein interactions

Furthermore, we compared our model with other state-of-the-art methods including RPISeq [13], lncPro [24], and RPI-SAN [17] to evaluate the predictive ability to lncRNA-protein interactions of LPI-Pred. The RPISeq and lncPro use only sequence information, which is similar to LPI-Pred. More recently, the RPI-SAN use deep learning model, based on sequence information and evolutionary information to predict novel ncRNA-protein interactions. We follow same performance evaluation measurements. The comparison details are shown as below Table 4.

On dataset RPI369, LPI-Pred performs better than RPISeq and lncPro on all measurements, with accuracy of 73.06%, sensitivity of 75.32%, specificity of 71.14%, precision of 72.64%, MCC of 46.67% and AUC of 0.802. For dataset RPI1807, LPI-Pred is not best on all 6 indicators, but it still has an accuracy of up to 97.1%, and perform better on sensitivity and precision. Essentially, the RPI488 is the full lncRNA-protein interactions dataset. As the results shown, the accuracy, sensitivity, specificity, precision, MCC and AUC of LPI-Pred are 89.92%, 82.75%, 96.72%, 96.32%, 80.59% and 0.911. It has the best performance on accuracy, specificity, precision and MCC compared with all existing methods. Overall, the evaluation between LPI-Pred and other methods on three benchmark datasets can prove the high robustness and accuracy of LPI-Pred. It suggests that the word embedding can provide hidden high-level feature of sequence and the feature selection can further enhance the expressiveness of features and reduce the complexity of model training.

4. Conclusion

The lncRNA-protein interactions play numerous roles in life activities, cellular function and disease. The first step in studying its function and mechanism is to identify interacting lncRNA-protein pairs. In this study, we present a novel lncRNA-protein interaction prediction model named LPI-Pred. First, we trained distribution representation model, RNA2vec and pro2vec, by using skip-gram word embedding model and *human* genome-wide lncRNA and protein sequences. Then, we convert the lncRNA and

Table 4
Comparing five-fold cross-validation performance of LPI-Pred and other state-of-the-art methods on three gold standard datasets.

Datasets	Methods	Acc (%)	Sens (%)	Spec (%)	Pre (%)	MCC (%)	AUC
RPI369	RPISeq	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
	LPI-Pred	**73.06**	**75.32**	**71.14**	**72.64**	**46.67**	**0.802**
RPI1807	RPISeq	**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**
	LPI-Pred	97.10	**97.89**	96.14	**96.91**	94.13	0.994
RPI488	RPISeq	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**
	LPI-Pred	**89.92**	82.75	**96.72**	**96.32**	**80.59**	0.911

The boldface indicates this measure performance is the best among the compared methods for individual dataset.

protein sequence into word vector using the model trained above. The Gini impurity-based feature selection is used to obtain discriminative features. Then we training LPI-Pred to predict lncRNA-protein interactions. We compared the performance of different feature representations and predictors, and we also compared LPI-pred with other state-of-the-art methods. The rigorous evaluation experimental results show the effectiveness and robustness of our model.

Inspired by the similarity between biological sequences and natural language sentences, we divided sequence into *k*-mers, which can be considered as "words" in biological language. The experimental proved this feature extraction scheme works well. However, rethinking of the procedure of RNA2vec and pro2vec, we recognize that *k*-mer may not be the best way to sequence word segmentation. More bio-semantic sequence segmentation should be explored in the future.

Author contributions

H-C. Y and Z-H. Y conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript; L. C, X. Z, T-H. J and X. L designed, performed and analyzed experiments and wrote the manuscript; All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China, under Grant 61572506, in part by the National Outstanding Youth Science Fund Project of National Natural Science Foundation of China, under Grant 61722212, in part by the Pioneer Hundred Talents Program of Chinese Academy of Sciences.

References

[1] Han S, Du W, Xu Y, Zhang Y, Li Y, Liang Y, Ma Q, Wang C: LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. 2018.

[2] Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature 2012;489:101.

[3] Pennisi E. ENCODE project writes eulogy for junk DNA. Science 2012;337(6099):1159–61.

[4] Yang Q, Zhang S, Liu H, Wu J, Xu E, Peng B, et al. Oncogenic role of long noncoding RNA AF118081 in anti-benzo[a]pyrene-trans-7,8-dihydrodiol-9,10-epoxide-transformed 16HBE cells. Toxicol Lett 2014;229(3):430–9.

[5] Tsai M-C, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, et al. Long noncoding RNA as modular scaffold of histone modification complexes. Science 2010;329(5992):689–93.

[6] Nie L, Wu H-J, Hsu J-M, Chang S-S, Labaff AM, Li C-W, et al. Long non-coding RNAs: versatile master regulators of gene expression and crucial players in cancer. Am J Transl Res 2012;4(2):127–50.

[7] Zeng X, Lin W, Guo M, Zou Q. A comprehensive overview and evaluation of circular RNA detection tools. PLoS Comput Biol 2017;13(6):e1005420.

[8] Wang Kevin C, Chang Howard Y. Molecular mechanisms of long noncoding RNAs. Mol Cell 2011;43(6):904–14.

[9] Ng S-Y, Lin L, Soh BS, Stanton LW. Long noncoding RNAs in development and disease of the central nervous system. Trends Genet 2013;29(8):461–8.

[10] Shi X, Sun M, Liu H, Yao Y, Kong R, Chen F, et al. A critical role for the long non-coding RNA GAS5 in proliferation and apoptosis in non-small-cell lung cancer. Mol Carcinog 2015;54(S1):E1–E12.

[11] Congrains A, Kamide K, Oguro R, Yasuda O, Miyata K, Yamamoto E, et al. Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B. Atherosclerosis 2012;220(2):449–55.

[12] Colantoni A, Ferrè F, Helmer-Citterich M. Revealing protein–lncRNA interaction. Briefings Bioinf 2015;17(1):106–16.

[13] Muppirala UK, Honavar VG, Dobbs D. Predicting RNA-protein interactions using only sequence information. BMC Bioinf 2011;12(1):489.

[14] 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.

[15] Bellucci M, Agostini F, Masin M, Tartaglia GG. Predicting protein associations with long noncoding RNAs. Nat Methods 2011;8:444.

[16] Agostini F, Cirillo D, Bolognesi B, Tartaglia GG. X-inactivation: quantitative predictions of protein interactions in the Xist network. Nucleic Acids Res 2012;41(1):e31.

[17] 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.

[18] Xiao Y, Zhang J, Deng L. Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks. Sci Rep 2017;7(1):3664.

[19] Zhang T, Wang M, Xi J, Li A. LPGNMF: predicting long non-coding RNA and protein interaction using graph regularized nonnegative matrix factorization. IEEE/ACM Trans Comput Biol Bioinf 2018:1.

[20] Shen C, Ding Y, Tang J, Jiang L, Guo F. LPI-KTASLP: Prediction of LncRNA-Protein Interaction by Semi-Supervised Link Learning With Multivariate Information. IEEE Access 2019;7:13486–96.

[21] Zhang W, Yue X, Guifeng T, Wu W, Huang F, Zhang X: SFPEL-LPI: sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions, 14; 2018.

[22] Wang Y, You Z-H, Yang S, Li X, Jiang T-H, Zhou X. A high efficient biological language model for predicting protein-protein interactions. Cells 2019;8(2):122.

[23] Frankish A, Bignell A, Berry A, Yates A, Parker A, Schmitt BM, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 2018;47(D1):D766–73.

[24] Lu Q, Ren S, Lu M, Zhang Y, Zhu D, Zhang X, et al. Computational prediction of associations between long non-coding RNAs and proteins. BMC Genomics 2013;14(1):651.

[25] 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;17:582.

[26] Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, Honavar V, et al. PRIDB: a protein–RNA interface database. Nucleic Acids Res 2010;39(suppl_1):D277–82.

[27] Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, et al. Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci 2007;104(11):4337–41.

[28] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for the ENCODE project. Genome Res 2012;22.

[29] Le Q, Mikolov T. Distributed representations of sentences and documents. In: International conference on machine learning: 2014. 1188–1196.

[30] Gittens A, Achlioptas D, Mahoney MW: Skip-gram-zipf+ uniform= vector additivity. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers): 2017. 69–76.

[31] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 2013.:3111–9.

[32] Mikolov T, Chen K, Corrado G, Dean J: Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781 2013.

[33] Asgari E, Mofrad MR. Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS ONE 2015;10(11):e0141287.

[34] Pan X, Shen H-B. Learning distributed representations of RNA sequences and its application for predicting RNA-protein binding sites with a convolutional neural network. Neurocomputing 2018;305:51–8.

[35] Wang L, You Z-H, Chen X, Li Y-M, Dong Y-N, Li L-P, et al. LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities. PLoS Comput Biol 2019;15(3):e1006865.

[36] Kim S-J, Koh K, Lustig M, Boyd S, Gorinevsky D. An interior-point method for large-scale $\ell_1 $-regularized least squares. IEEE J Sel Top Signal Process 2007;1(4):606–17.

[37] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1.

[38] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;8:1226–38.