Molecular Therapy
Nucleic Acids
Original Article

The Bipartite Network Projection-Recommended Algorithm for Predicting Long Non-coding RNA-Protein Interactions

Qi Zhao,[1] Haifan Yu,[1] Zhong Ming,[2,3] Huan Hu,[4] Guofei Ren,[5] and Hongsheng Liu[4,6,7]

[1]School of Mathematics, Liaoning University, Shenyang 110036, China; [2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China; [3]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China; [4]School of Life Science, Liaoning University, Shenyang 110036, China; [5]School of Information, Liaoning University, Shenyang 110036, China; [6]Research Center for Computer Simulating and Information Processing of Bio-Macromolecules of Liaoning Province, Shenyang 110036, China; [7]Engineering Laboratory for Molecular Simulation and Designing of Drug Molecules of Liaoning, Shenyang 110036, China

With the development of science and biotechnology, many evidences show that ncRNAs play an important role in the development of important biological processes, especially in chromatin modification, cell differentiation and proliferation, RNA progressing, human diseases, etc. Moreover, lncRNAs account for the majority of ncRNAs, and the functions of lncRNAs are expressed by the related RNA-binding proteins. It is well known that the experimental verification of lncRNA-protein relationships is a waste of time and expensive. So many time-saving and inexpensive computational methods are proposed to uncover potential lncRNA-protein interactions. In this work, we propose a novel computational method to predict the potential lncRNA-protein interactions with the bipartite network projection recommended algorithm (LPI-BNPRA). Our approach is a semi-supervised method based on the lncRNA similarity matrix, protein similarity matrix, and lncRNA-protein interaction matrix. Compared with three previous methods under the leave-one-out cross-validation, our model has a more high-confidence result with the AUC value of 0.8754 and the AUPR value of 0.6283. We also do case studies by the *Mus musculus* dataset to further reflect the reliability of our approach. This suggests that LPI-BNPRA will be a reliable computational method to uncover lncRNA-protein interactions in biomedical research.

INTRODUCTION

With the development of biological research, people gradually realize the importance of non-coding RNAs (ncRNAs) over the years. ncRNAs are classified into sncRNAs (short non-coding RNAs) with a length of less than 200 nucleotides and lncRNAs (long non-coding RNAs) with a length of more than 200 nt in molecular biology. Moreover, lncRNAs occupy a large proportion of ncRNAs. In recent years, more and more ncRNAs are confirmed to be related to many important biological processes,[1] especially lncRNAs, which are confirmed to play a critical role in the chromatin modification,[2] cell differentiation and proliferation,[3] RNA progressing,[4] cellular apoptosis,[5] and human diseases.[6] For example, Wang et al.[7] in 2017 found that the lncRNA ACOD1 can respond to viral infection and rapidly upregulate expression and further enhance the enzyme activity of metabolic enzyme GOT2, and it will increase the influence ability of virus. In another study, Yang et al.[8] revealed that lncRNAs can regulate the Warburg effect of tumor cells in order to promote tumor growth. lncRNAs do not work alone; they need to combine with the corresponding proteins to play their roles in biological process and complete the delivery of information. So, it is important to identify the lncRNA-binding protein interactions[9] if we want to understand the molecular mechanism underlying the functions of lncRNAs.[10–15]

An increasing number of proteins related to RNAs are confirmed[16] with the development of biotechnology includes high-throughput sequencing technology (microarray, RNA-seq, CLIP-seq, etc.), RNAi, radioimmunoprecipitation (RIP), and so on. However, our understanding of lncRNA-protein relationships is limited.[17] Further, using experimental methods to predict lncRNA-protein associations is costly and time consuming.[18] Therefore, developing time-saving and inexpensive computational methods becomes more and more necessary. During recent years, many computational methods are widely used in the field of bioinformatics, for instance, lncRNA-disease, genome-cancer, and drug-target interaction predictions.[1,19–27] But there are only a few computational methods that can be used to infer lncRNA-protein correlations. For instance, Bellucci et al.[28] proposed a method named catRAPID in 2011, which was based on the physicochemical properties, including secondary structure, hydrogen bonding, and van der Waals propensities, to forecast lncRNA-protein connections by encoding lncRNA-protein pairs as a feature vector. Then, in the same year, Muppirala et al.[29] developed RPI-seq based on the sequences of lncRNAs and proteins, which trained RF (random forest)[30] and SVM (support vector machine)[31] classifiers

Received 23 May 2018; accepted 25 September 2018;
https://doi.org/10.1016/j.omtn.2018.09.020.

Correspondence: Hongsheng Liu, School of Life Science, Liaoning University, Shenyang 110036, China.
E-mail: liuhongsheng@lnu.edu.cn

to indicate RNA-protein interactions. Next, a method was presented by Wang et al.[32] based on naive bias and extended naive bias classifiers based the same dataset and feature with RPI-seq. Thereafter, Lu et al.[33] used hydrogen-bond, Van der Waals propensities and RSS (six types of RNA secondary structures) to develop a method named IncPro to infer lncRNA-protein relationships. In 2015, Suresh et al.[34] proposed RPI-Pred, which extracted lncRNA sequences, protein sequences and high-order 3D structural proteins features to develop methods based on SVM. Later on, lncRNA-protein bipartite network inference (LPBNI) was developed by Ge et al.,[35] which was based on a bipartite network to predict lncRNA-protein associations that only used known lncRNA-protein interactions. lncRNA-protein interactions prediction using Eigenvalue transformation-based semi-supervised link prediction (LPI-ETSLP) was presented by Hu et al.[36] in 2017, which is a semi-supervised method based on eigenvalue transformation with the known lncRNA-protein interaction matrix, the lncRNA and protein similarity matrix to uncover the relationship between lncRNA and protein. In the same year, Liu et al.[37] also explored a semi-supervised model named LPI-neighborhood regularized logistic matrix factorization (NRLMF) under the same dataset as LPI-ETSLP to predict unknown lncRNA-protein interactions by neighborhood-regularized logistic matrix factorization.

There are also some limitations of those traditional machine-learning methods mentioned above. First, those methods are based on the confirmed RNA-protein relationships instead of ncRNA-protein relationships, which would make bias. Second, they are used to infer lncRNA-protein correlations based on the NPInter[38] database, so that it may bias the forecast because NPInter contains ncRNAs gene-protein interaction sequences. Third, it is difficult to choose the features of lncRNAs and proteins in lncRNA-protein relationships by the machine-learning-methods predictions. Fourth, there are no negative samples to forecast potential lncRNA-protein interactions, which may lead to some biases as well.

LPI-ETSLP and LPI-NRLMF can overcome the above-mentioned drawbacks of traditional machine-learning models, but they also have some shortcomings. The model construction of LPI-ETSLP and LPI-NRLMF is very dependent on the dataset and will bias the result. In addition, LPI-ETSLP and LPI-NRLMF use theoretical parameters, and these parameters may not apply to new data. They also have a complex calculation process and take a lot of time.

Compared with LPI-ETSLP and LPI-NRLMF, our study aims to explore a novel semi-supervised method based on a bipartite network-projection-recommended algorithm (LPI-BNPRA) to infer new lncRNA-protein associations and solve problems mentioned above as well. In our model, the lncRNA similarity matrix, protein similarity matrix, and known lncRNA-protein interaction matrix are also working for forecasting unknown lncRNA-protein correlations with a bipartite network. The calculation process of our work is not complex and time saving. The Smith-Waterman algorithm is utilized to calculate the similarity score between lncRNA and lncRNA (or protein and protein) in this new method. We present a bias-ratings idea under the degree of lncRNAs or protein sequence similarity and increase the accuracy of the prediction. We also obtain an optimal threshold according to previous work of Shi et al.,[39] which is not based on the theoretical parameters and performs stably for the prediction results when we repeat experiments many times. In addition, we use leave-one-out cross-validation (LOOCV) to verify the dependability of LPI-BNPRA and get a higher AUC (the area under the ROC curve)[40] value of 0.8754 compared with other methods. Apart from this, we compute other index values of our model as well, such as area under precision recall (AUPR), precision (PRE), accuracy (ACC), etc. What's more, we do case studies under the *Mus musculus* dataset and get high-confidence lncRNA-protein interaction predictions with the predictive score ranks compared with other methods. These comparisons show LPI-BNPRA dependability of predicting new lncRNA-protein relationships.

RESULTS AND DISCUSSION

Comparison with Other Methods Based on NPInter V2.0

In this subsection, we compare the predictive ability of LPI-BNPRA with RWR,[41,42] LPBNI,[35] and RPI-seq based on the same dataset to illustrate the advantage of our method. RPI-seq is a classic machine-learning method based on the development of RF and SVM. However, RWR and LPBNI are semi-supervised learning algorithms, which depend on the similarity matrix of lncRNAs and proteins. Our method provides an idea on how to forecast lncRNA-protein interactions by the integrating protein similarity matrix, lncRNA similarity matrix, and lncRNA-protein interaction matrix. Thus, we compare results of these approaches by LOOCV under the same dataset, and the comparison results are listed in Figure 1 and Table 1.

As shown in Figure 1, LPI-BNPRA has an AUC value of 0.8754, while the AUCs of RWR, LPBNI, RPI-seq-RF, and RPI-seq-SVM are 0.8323, 0.8568, 0.3949, and 0.3987. Particularly, the AUCs of RWR and LPBNI are a little lower than that of LPI-BNPRA. This is because RWR does the prediction by using protein and lncRNA sequence information under a multiple network and LPBNI only takes average strategy under a bipartite network with confirmed lncRNA-protein interactions. LPI-BNPRA proposes bias ratings regarded as initial resources under the agglomerative hierarchical clustering, where bias ratings produce transfer rates and makes our model more exact than those two methods. It is clear that RPI-seq has a lower AUC value than that of the other three approaches and shows RPI-seq is not reliable compared with LPI-BNPRA. Because RPI-seq is a classical machine-learning method, which needs training features and negative and positive samples to do the prediction. There are many differences of biological function between RNAs and lncRNAs; it is also one of the reasons that makes RPI-seq ineffective, because RPI-seq uses RNA-protein interactions instead of lncRNA-protein interactions. Compared with RPI-seq, LPI-BNPRA is based on the lncRNA-protein relationships confirmed by experimental verifications, which makes it possible to avoid the problems of negative samples. We examine the quality of computational methods by other index values, for instance, AUPR, PRE, sensitivity (SEN), F1 score, and so on. It is illustrated in Table 1 that LPI-BNPRA has higher

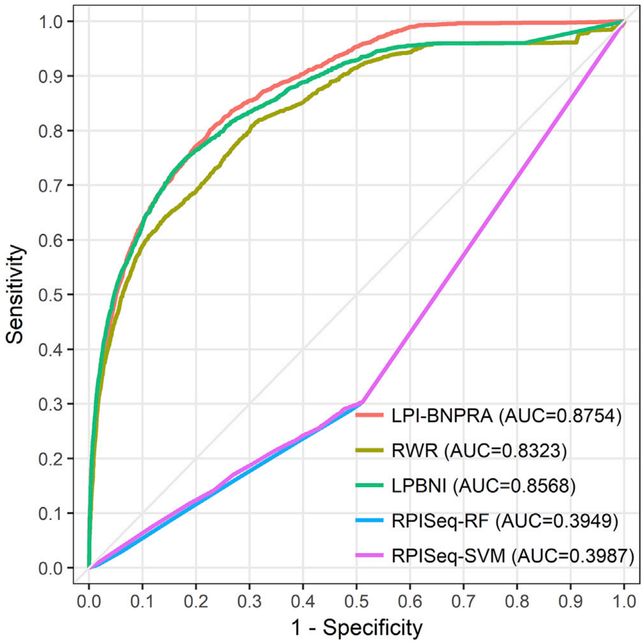

Figure 1. The ROC Curves of LPI-BNPRA, RWR, LPBNI, RPI-Seq-RF, and RPI-Seq-SVM

The ROC curves of LPI-BNPRA, RWR, LPBNI, RPI-seq-RF, and RPI-seq-SVM are plotted in red, brown, green, blue, and purple. The light gray line is the ROC curve of the relationship between LPI-BNPRA and the randomized lncRNA-protein pairs.

values than RWR and LPBNI apart from AUC value; for example, the AUPR values of LPI-BNPRA, RWR, and LPBNI are 0.6283, 0.2893, and 0.3306. The AUPR value of our method is much higher than those of RWR and LPBNI. The PRE value is 0.6540, which is 43.71%, 43.22% higher than RWR and LPBNI. Although our approach has a lower ACC value than that of RWR and LPBNI, it is more reliable to assess unstable dataset with F1 score than using the ACC value. From Table 1, the F1 score of our model is 0.5564 and obviously higher than that of other three models, further suggesting that our model achieves a reliable prediction. Due to the limited resources of the lncRNA-protein interactions database and because we only extract a few available datasets, the predictive results are also limited. Therefore, with the lncRNA-protein connections increasingly verified by experiments, the database is continually expanding, and the accuracy and efficiency of LPI-BNPRA will be correspondingly improved.

Case Studies

From the above studies, using known human lncRNA-protein interactions to forecast potential lncRNA-protein relationships can obtain a reliable result based on LPI-BNPRA. Therefore, in order to better assess the predictive ability of this method, we apply it to test its performance based on the *Mus musculus* dataset. Mice are physiologically similar to other mammals and humans, and there is also a high genetic similarity between mice and humans. Mice also have a strong ability to reproduce and grow up. Experimenters often do experiments using the mouse as the experimental animal, and there are a lot of mouse genome-sequencing experiments. Therefore, we use the *Mus musculus* dataset to test our model performance.

Table 1. Comparison of LPI-BNPRA with RWR, LPBNI, RPI-seq-RF, and RPI-seq-SVM Models

Methods	AUC	AUPR	PRE	SEN	ACC	F1 Score
LPI-BNPRA	0.8754	0.6283	0.6540	0.4841	0.8799	0.5564
RWR	0.8332	0.2893	0.3681	0.3538	0.9536	0.3603
LPBNI	0.8586	0.3306	0.3713	0.3713	0.9581	0.3868
RPI-seq-RF	0.3949	0.0631	0.0983	0.0983	0.4626	0.1481
RPI-seq-SVM	0.3987	0.0698	0.1003	0.1003	0.4823	0.1493

We do case studies under the *Mus musculus* dataset extracted from NPInter v3.0 and list the top 10 inferred new lncRNA-protein connections in Table 2; these new connections are checked in the *Mus musculus* dataset finally. Embryonic stem cells are highly undifferentiated cells, which can splinter and proliferate cells to produce various organs and then generate an organism. Therefore, embryonic stem cells play a key role in the biological processes. Guttman et al.[43] proposed that many lncRNA gene-protein pairs are important in the circuitry of the mouse to control mouse embryonic stem cell state, for example, NONMMUG030867-A2AC19, NONMMUG078379-Q8CHK4, NONMMUG002214-A2AC19, etc. We find these interaction predictions forecasted by our method, which are rank advanced from Table 2.

Table 2 also lists the rankings of the top 10 lncRNA-protein predicted interactions of other models. It is evident that the rankings of our model are more advanced than those of other models, and some of the top 10 relationship-pair prediction rankings predicted by other three methods are not very high—for instance, RPI-seq had lower rankings. Many new interactions found by our model may be neglected in other models. Therefore, LPI-BNPRA achieves a high-confidence performance compared with other methods as well by these important lncRNA gene-protein interactions.

Conclusions

lncRNAs have large quantities, large molecular weight, and poor stability *in vitro*, and they are also difficult to crystallize. The characteristics of lncRNA make it difficult to study its structure, and only a few studies have reported its structure at present.[44] Although the role of lncRNAs is undoubted in the regulation of gene expression, only a few lncRNAs have been studied for their function and mechanism of action.[45] Since lncRNAs play a regulatory role in the coordination of protein molecules, the identification of protein molecules that bind specific lncRNAs has become the main study to reveal the function and mechanism of lncRNAs.[46] It is well known that lncRNAs are widely involved in biological processes such as DNA methylation, protein modification, and chromosome remodeling *in vivo*. lncRNAs can directly interact with transcription factors, functional RNA molecules, and chromatin remodeling modifiers as well. In addition,

Table 2. Top 10 Novel Predicted lncRNA-Protein Interactions Based on LPI-BNPRA and Their Ranks Based on Other Methods

lncRNA	Protein	Confirmed?	LPI-BNPRA	RWR	LPBNI	RPI-seqFR	RPI-seqSVM
NONMMUG030867	A2AC19	confirmed	1	17	16	3	157
NONMMUG078379	Q8CHK4	confirmed	2	26	34	124	131
NONMMUG002214	A2AC19	confirmed	3	62	71	98	147
NONMMUG009968	Q8CHK4	confirmed	4	118	101	171	107
NONMMUG022640	Q13185	confirmed	5	136	131	64	156
NONMMUG045923	Q8CHK4	confirmed	6	116	120	122	141
NONMMUG013483	Q9NQR1	confirmed	7	9	42	38	122
NONMMUG035346	O09106	confirmed	8	122	128	129	132
NONMMUG009968	O09106	confirmed	9	127	127	91	79
NONMMUG030867	Q9NQR1	confirmed	10	80	87	102	112

lncRNAs regulate the target gene expression at the transcriptional level and post-transcriptional level.[47,48] Some lncRNAs are also involved in the regulation of target genes as precursors for certain functional sncRNAs, such as small interfering RNAs (siRNAs), microRNAs, and piwiRNAs.[1,49–52] Particularly, the functions of lncRNAs are expressed by the related lncRNA-binding proteins.[10,53,54] Therefore, predicting potential lncRNA-protein interactions is helpful to study the function of lncRNAs. But experimental methods to forecast lncRNA-protein relationships are costly and time-consuming. Therefore, some computational methods are developed to infer unknown lncRNA-protein connections. In this paper, we propose a method named LPI-BNPRA to predict lncRNA-protein associations by the known lncRNA-protein interactions matrix, the lncRNA similarity matrix, and the protein similarity matrix. Further, the performance of LPI-BNPRA is assessed by LOOCV; it also has a higher AUC score with 0.8754 than that of RWR, LPBNI, and RPI-seq. We also calculate many other reliable index values, which show the perfect performance of our method as well. Besides, in order to further check the effectiveness of LPI-BNPRA, we apply it to infer lncRNA-protein correlations on the *Mus musculus* dataset extracted from NPInter v3.0; the results show our method also has high-accuracy predictability based on the other species datasets. With the development of biotechnology, more lncRNA-protein interactions will be confirmed, and the accuracy of the prediction of LPI-BNPRA will increase as well. In summary, we find that LPI-BNPRA is a reliable computational method to predict unknown lncRNA-protein relationships in the future.

MATERIALS AND METHODS

Datasets

We got ncRNA-protein interactions downloaded from NPInter v2.0[55] in our approach, which is a database includes conformed interactions between ncRNAs and other biomolecules (proteins, RNAs, and genomic DNAs). We chose lncRNA-protein interactions involved in human lncRNAs, and lncRNA sequences were downloaded from NONCODE.[56,57] So, we obtained 141,353 lncRNA sequences according to NONCODE 4.0. Then we removed unreliable lncRNA sequences, which were not human lncRNAs, and obtained a dataset containing 4,158 high-confidence lncRNA-protein associations generated by 990 lncRNAs and 27 proteins.[58,59] Especially when we selected reliable datasets, the accuracy of our method was improved by removing lncRNAs only related to one protein and proteins only related to one lncRNA.

lncRNA-lncRNA Similarity Matrix

The Smith-Waterman algorithm is a sequence algorithm, and we used it to calculate the similarity scores of every lncRNA-lncRNA pair. LSM stands for the similarity matrix of lncRNAs, in which the entity $LSM(l_i, l_j)$ reflects the similarity score between lncRNA l_i and lncRNA l_j. $LSM(l_i, l_j)$ was scored according to the following formula:

$$LSM\left(l_i, l_j\right) = \frac{sw\left(l_i, l_j\right)}{\max\left(sw\left(l_i, l_i\right), sw\left(l_j, l_j\right)\right)}, \qquad \text{(Equation 1)}$$

where $sw(l_i, l_j)$ is the sequence similarity score of lncRNA l_i and lncRNA l_j, which is calculated by the Smith-Waterman algorithm.

Protein-Protein Similarity Matrix

Similar to lncRNA similarity scores, we used the Smith-Waterman algorithm to calculate protein-protein similarity scores as well. The protein similarity matrix is marked as PSM, where the entity $PSM(p_i, p_j)$ is the similarity score of protein p_i and protein p_j. $PSM(p_i, p_j)$ is defined as

$$PSM\left(p_i, p_j\right) = \frac{sw\left(p_i, p_j\right)}{\max\left(sw\left(p_i, p_i\right), sw\left(p_j, p_j\right)\right)}, \qquad \text{(Equation 2)}$$

where $sw(p_i, p_j)$ shows the sequence similarity score between protein p_i and protein p_j calculated by Smith-Waterman algorithm.

lncRNA-Protein Interaction Matrix

We computed the interaction scores between lncRNAs and proteins according to the sequence similarity matrixes. Therefore, the adjacency matrix Y appeared to describe the lncRNA-protein interactions, in which entity $Y(l_i, p_j)$ is 1 if lncRNA l_i is confirmed to be related to the protein p_j, otherwise 0.

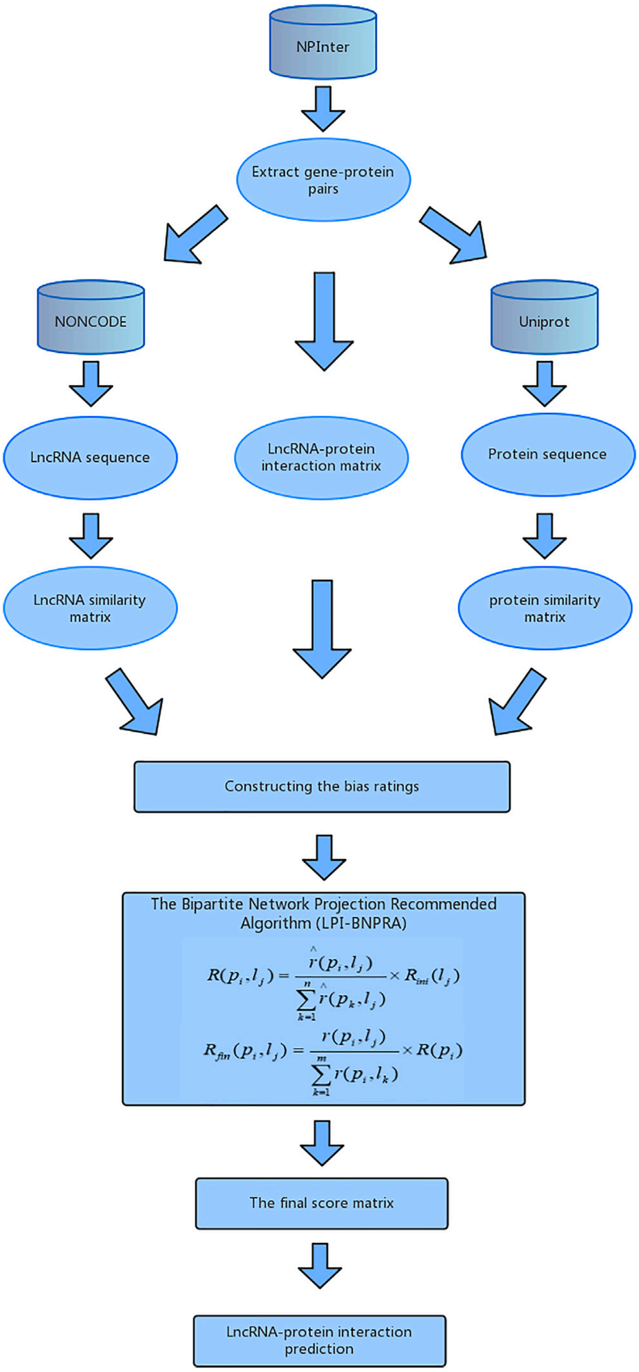
NPInter
Extract gene-protein pairs
NONCODE
Uniprot
LncRNA sequence
LncRNA-protein interaction matrix
Protein sequence
LncRNA similarity matrix
protein similarity matrix
Constructing the bias ratings
The Bipartite Network Projection Recommended Algorithm (LPI-BNPRA)
$R(p_i,l_j)=\frac{\hat{r}(p_i,l_j)}{\sum_{k=1}^{n}\hat{r}(p_k,l_j)}\times R_{ini}(l_j)$
$R_{fin}(p_i,l_j)=\frac{r(p_i,l_j)}{\sum_{k=1}^{m}r(p_i,l_k)}\times R(p_i)$
The final score matrix
LncRNA-protein interaction prediction

Figure 2. The Workflow Chart of LPI-BNPRA

LPI-BNPRA

Figure 2 reflects the flow chart of LPI-BNPRA. First, we obtained the extracted gene-protein pairs from NPInter v2.0 database. Then, we removed unreliable lncRNA and protein sequences. Next, we calculated the lncRNA-lncRNA similarity score and protein-protein similarity score and constructed the bias ratings. Finally, we integrated the lncRNA similarity matrix, protein similarity matrix, and lncRNA-protein interaction matrix with LPI-BNPRA to compute lncRNA-protein interaction scores and infer potential lncRNA-protein connections.

We got the bias ratings of every lncRNA for proteins based on the agglomerative hierarchical clustering. For instance, a given lncRNA l_i tends to be relevant to many proteins that have the similar sequence information. Thus, we can construct bias ratings of this given lncRNA l_i to proteins.

The agglomerative hierarchical clustering takes a bottom-up strategy, which assumes each protein (or each lncRNA) is a single cluster in the first step, and then combines these single clusters based on the linkage criterion of minimum variance method. Where $LD(l_i, l_j)$ is the distance between lncRNA l_i and lncRNA l_j, and $PD(p_i, p_j)$ represents the distance between protein p_i and protein p_j, $LD(l_i, l_j)$ and $PD(p_i, p_j)$ are denoted as

$$LD(l_i, l_j) = 1 - LSM(l_i, l_j) \qquad \text{(Equation 3)}$$

$$PD(p_i, p_j) = 1 - PSM(p_i, p_j). \qquad \text{(Equation 4)}$$

In the same manner, many reliable clusters with an appropriate threshold are gained after cutting the hierarchical clustering tree. The threshold performs stably for the prediction results when we repeat experiments many times based on LOOCV. Therefore, we counted the bias rating of protein p_i to lncRNA l_j as the following equation

$$r(p_i, l_j) = \frac{n_{cr}}{T(p_i)}, \qquad \text{(Equation 5)}$$

where n_{cr} represents the number of the lncRNA l_j in the cluster cr, which includes the lncRNA l_j; let $T(p_i)$ be the number of all the lncRNAs related to protein p_i.

Different proteins have different interaction score ranges, according to the known lncRNA-protein interaction matrix that contains confirmed interactions of lncRNA-protein pairs, and we obtained a bias-rating range of each protein to related lncRNAs by using the agglomerative hierarchical clustering. Therefore, we get the original bias rating $\hat{r}(p_i, l_j)$ as follows:

$$\hat{r}(p_i, l_j) = \frac{r(p_i, l_j)}{\bar{r}(p_i)} \qquad \text{(Equation 6)}$$

$$\bar{r}(p_i) = \frac{\sum_{j=1}^{m} r(p_i, l_j)}{T(p_i)}. \qquad \text{(Equation 7)}$$

The original bias ratings demonstrate the difference of biases from every protein to different lncRNAs, which reflect the differences of lncRNA-protein interactions. Then, the initial resource from lncRNAs to related proteins is defined as

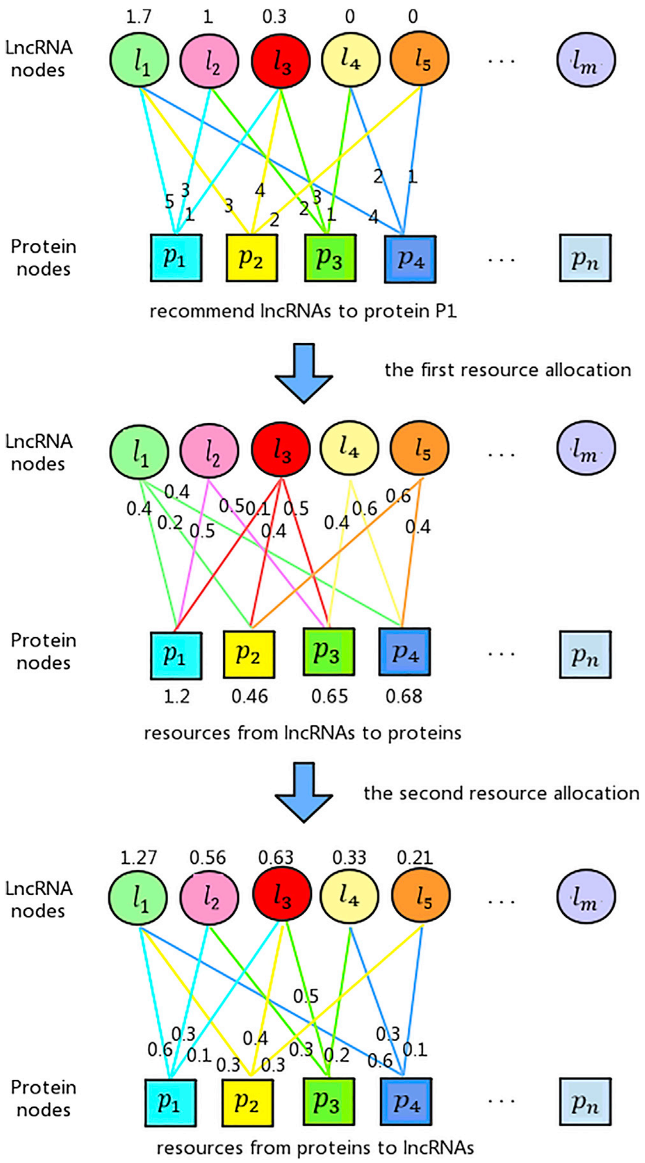

Figure 3. The Basic Idea of LPI-BNPRA

$$R_{ini}(l_j) = \hat{r}(p_i, l_j). \qquad \text{(Equation 8)}$$

As shown in the Figure 3, there are two resource allocations in our model under the bipartite network. In the first resource allocation, we allocated the initial resource from lncRNAs to proteins. Resources of proteins are allocated to lncRNAs again by the principle similar to the first allocation in the second resource allocation. So, we need to calculate the resource of protein $R(p_i) = \sum_{j=1}^{m} R(p_i, l_j)$ from lncRNA l_j in the first resource allocation.

$$R(p_i, l_j) = \frac{\hat{r}(p_i, l_j)}{\sum_{k=1}^{n} \hat{r}(p_k, l_j)} \times R_{ini}(l_j). \qquad \text{(Equation 9)}$$

The resource of protein p_i is the sum of the allocated initial resource from all lncRNAs correlated with protein p_i:

$$R(p_i) = \sum_{j=1}^{m} R(p_i, l_j). \qquad \text{(Equation 10)}$$

Similar to the first resource allocation, we assigned a transfer weight to every link in the second allocation (see Figure 3). Next, we obtained the final resource from protein p_i to lncRNA l_j:

$$R_{fin}(p_i, l_j) = \frac{r(p_i, l_j)}{\sum_{k=1}^{m} r(p_i, l_k)} \times R(p_i). \qquad \text{(Equation 11)}$$

Then, the final resources of lncRNA l_j are allocated from all the proteins connected with the lncRNA l_j. The formula is presented as follows:

$$R_{fin}(l_j) = \sum_{i=1}^{n} R_{fin}(p_i, l_j). \qquad \text{(Equation 12)}$$

Finally, we gained a bias-rating score matrix containing the bias ratings from proteins to lncRNAs. Therefore, we can recommend the corresponding lncRNAs to a given protein with the final resource scores in a descending order.

For example, the ratings of protein p_1 for lncRNAs l_1, l_2, and l_3 confirmed have relation with p_1 are 5, 3, and 1, respectively, and the median is 3. The rest of the numbers on the edge of each link are the ratings of proteins for its associated lncRNAs. Therefore, the initial resources from p_1 to lncRNAs l_1 to l_5 are 1.7, 1, 0.3, 0, and 0 according to the Equations 6 and 7. In the first resource allocation, resources are distributed from lncRNAs to proteins, and protein p_1 gets 1.2 in total by Equations 9 and 10. Then, in the second resource allocation, resources are transferred back to lncRNAs. In the first and second resource allocations, the number on the edge of each link is the resource allocation rate from every lncRNA to its related protein. lncRNA l_4 receives 0.33 in total in the light of Equations 11 and 12. Similarly, lncRNA l_5 obtains 0.21 in total. Above all, we can predict that lncRNA l_4 may be more correlative with protein p_1 than lncRNA l_5 based on the final resource scores of lncRNA l_4 and l_5.

Performance Evaluation

LOOCV was applied for checking the effectiveness of LPI-BNPRA in this step. In the cross-validation process, the initial dataset was divided into K sub-samples, from which an individual sub-sample was taken as the data for the validation model. K = 4,158 in this work, and every lncRNA-protein interaction was regarded as one sub-sample; the other 4,157 samples were used for training. The cross-validation was repeated 4,158 times until each sub-sample was verified once, and an estimated value was obtained by averaging the 4,158 results. We selected a confirmed lncRNA-protein interaction as a testing sample in turn and looked at other confirmed lncRNA-protein correlations as training samples in our model. We plotted the ROC curve (receiver operating characteristic curve) to describe the performance of our approach. And AUC was also utilized to evaluate the method reliability; AUC = 1 demonstrates the perfect performance and 0.5 reflects the random performance. We also computed ACC, PRE, SEN, and F1 to test the ability of LPI-BNPRA. These index values are denoted as

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \quad \text{(Equation 13)}$$

$$REC = \frac{TP}{P} \quad \text{(Equation 14)}$$

$$PRE = \frac{TP}{TP + FP} \quad \text{(Equation 15)}$$

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} = 2 \times \frac{PRE \times REC}{PRE + REC}. \quad \text{(Equation 16)}$$

TP is true positive, TN indicates true negative, FP shows false positive, FN reflects false negative. The system errors and statistical biases are marked as ACC. PRE (the precision is regarded as the number of positive prediction as well) is the example of the related retrieval fraction, and REC (the recall is regarded as the number of sensitivity) defines the example of the related retrieval of search fraction based on the relevant understandings and measures. The F1 score (also named F degree of measure or F fraction) is more typical than ACC compared with different classifiers based on the class imbalance databases, which is because of taking the test accuracy and the score of calculation into consideration. F1 is accuracy weighted average and recall, and our model can reach an optimistic number if F1 = 1, otherwise F1 = 0 reveals the worst number of our method.

AUTHOR CONTRIBUTIONS

Q.Z. and H.L. conceived the project, developed the prediction method, designed and implemented the experiments, analyzed the results, and wrote the paper. H.Y. implemented the experiments, analyzed the results, and wrote the paper. Z.M., H.H., and G.R. analyzed the results and revised the paper. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors have no conflicts of interest.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (no 31570160 and 61772531), the Innovation Team Project of Education Department of Liaoning Province under grant no. LT2015011, the Doctor Startup Foundation from Liaoning Province under grant no. 20170520217, the Important Scientific and Technical Achievements Transformation Project under grant no. Z17-5-078, the Large-Scale Equipment Shared Services Project under grant no. F15165400, and the Applied Basic Research Project under grant no. F16205151.

REFERENCES

1. Chen, X., Yan, C.C., Zhang, X., and You, Z.H. (2017). Long non-coding RNAs and complex diseases: from experimental results to computational models. Brief. Bioinform. *18*, 558–576.

2. Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

3. Wapinski, O., and Chang, H.Y. (2011). Long noncoding RNAs and human disease. Trends Cell Biol. *21*, 354–361.

4. Wilusz, J.E., Sunwoo, H., and Spector, D.L. (2009). Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. *23*, 1494–1504.

5. Yu, F., Zheng, J., Mao, Y., Dong, P., Li, G., Lu, Z., Guo, C., Liu, Z., and Fan, X. (2015). Long non-coding RNA APTR promotes the activation of hepatic stellate cells and the progression of liver fibrosis. Biochem. Biophys. Res. Commun. *463*, 679–685.

6. Tripathi, R., Patel, S., Kumari, V., Chakraborty, P., and Varadwaj, P. (2016). DeepLNC, a long non-coding RNA prediction tool using deep neural network. Netw. Model. Anal. Health Inform. Bioinform. *5*, 21.

7. Wang, P., Xu, J., Wang, Y., and Cao, X. (2017). An interferon-independent lncRNA promotes viral replication by modulating cellular metabolism. Science *358*, 1051–1055.

8. Yang, F., Zhang, H., Mei, Y., and Wu, M. (2014). Reciprocal regulation of HIF-1α and lincRNA-p21 modulates the Warburg effect. Mol. Cell *53*, 88–100.

9. Prasanth, K.V., and Spector, D.L. (2007). Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. Genes Dev. *21*, 11–42.

10. Khalil, A.M., and Rinn, J.L. (2011). RNA-protein interactions in human health and disease. Semin. Cell Dev. Biol. *22*, 359–365.

11. Li, Z., and Nagy, P.D. (2011). Diverse roles of host RNA binding proteins in RNA virus replication. RNA Biol. *8*, 305–315.

12. Zhang, L., Ai, H., Chen, W., Yin, Z., Hu, H., Zhu, J., Zhao, J., Zhao, Q., and Liu, H. (2017). CarcinoPred-EL: Novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods. Sci. Rep. *7*, 2118.

13. Huang, Y.A., You, Z.H., Chen, X., Chan, K., and Luo, X. (2016). Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding. BMC Bioinformatics *17*, 184.

14. Li, J.Q., You, Z.H., Li, X., Ming, Z., and Chen, X. (2017). PSPEL: In Silico Prediction of Self-Interacting Proteins from Amino Acids Sequences Using Ensemble Learning. IEEE/ACM Trans. Comput. Biol. Bioinformatics *14*, 1165–1172.

15. Chen, X., Sun, Y.-Z., Liu, H., Zhang, L., Li, J.-Q., and Meng, J. (2017). RNA methylation and diseases: experimental results, databases, Web servers and computational models. Brief. Bioinform. Published online November 18, 2017. https://doi.org/10.1093/bib/bbx142.

16. Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172–177.

17. Kishore, S., Luber, S., and Zavolan, M. (2010). Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. Brief. Funct. Genomics *9*, 391–404.

18. Xia, T., Xiao, B.X., and Guo, J.M. (2013). [Acting mechanisms and research methods of long noncoding RNAs]. Yi Chuan *35*, 269–280.

19. Zhao, Q., Yao, C., Tang, J., and Liu, L. (2016). Study of spatial signal transduction in bistable switches. Front. Phys. *11*, 110501.

20. Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018). The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. Neurocomputing *273*, 526–534.

21. Chen, X., and Yan, G.Y. (2013). Novel human lncRNA-disease association inference based on lncRNA expression profiles. Bioinformatics *29*, 2617–2624.

22. Chen, X., Yan, C.C., Zhang, X., Zhang, X., Dai, F., Yin, J., and Zhang, Y. (2016). Drug-target interaction prediction: databases, web servers and computational models. Brief. Bioinform. *17*, 696–712.

23. Chen, X., Huang, Y.A., You, Z.H., Yan, G.Y., and Wang, X.S. (2017). A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics *33*, 733–739.

24. Wang, E., Zou, J., Zaman, N., Beitel, L.K., Trifiro, M., and Paliouras, M. (2013). Cancer systems biology in the genome sequencing era: part 1, dissecting and modeling of tumor clones and their networks. Semin. Cancer Biol. *23*, 279–285.

25. Wang, E., Zou, J., Zaman, N., Beitel, L.K., Trifiro, M., and Paliouras, M. (2013). Cancer systems biology in the genome sequencing era: part 2, evolutionary dynamics of tumor clonal networks and drug resistance. Semin. Cancer Biol. *23*, 286–292.

26. Wang, E., Zaman, N., Mcgee, S., Milanese, J.S., Masoudi-Nejad, A., and O'Connor-McCourt, M. (2015). Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. Semin. Cancer Biol. *30*, 4–12.

27. Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.H., and Liu, H. (2018). BNPMDA: Bipartite Network Projection for MiRNA-Disease Association prediction. Bioinformatics *34*, 3178–3186.

28. Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G.G. (2011). Predicting protein associations with long noncoding RNAs. Nat. Methods *8*, 444–445.

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

30. Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News: The Newsletter of the R Project. *2/3*, 18–28.

31. Hearst, M.A. (1998). Support Vector Machines. IEEE Intell. Syst. *13*, 18–28.

32. Wang, Y., Chen, X., Liu, Z.-P., Huang, Q., Wang, Y., Xu, D., Zhang, X.-S., Chen, R., and Chen, L. (2013). De novo prediction of RNA-protein interactions from sequence information. Mol Biosyst. *9*, 133–42.

33. Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X., and Li, T. (2013). Computational prediction of associations between long non-coding RNAs and proteins. BMC Genomics *14*, 651.

34. 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*, 1370–1379.

35. Ge, M., Li, A., and Wang, M. (2016). A Bipartite Network-based Method for Prediction of Long Non-coding RNA-protein Interactions. Genomics Proteomics Bioinformatics *14*, 62–71.

36. Hu, H., Zhu, C., Ai, H., Zhang, L., Zhao, J., Zhao, Q., and Liu, H. (2017). LPI-ETSLP: lncRNA-protein interaction prediction using eigenvalue transformation-based semi-supervised link prediction. Mol. Biosyst. *13*, 1781–1787.

37. Liu, H., Ren, G., Hu, H., Zhang, L., Ai, H., Zhang, W., and Zhao, Q. (2017). LPI-NRLMF: lncRNA-protein interaction prediction by neighborhood regularized logistic matrix factorization. Oncotarget *8*, 103975–103984.

38. Hao, Y., Wu, W., Li, H., Yuan, J., Luo, J., Zhao, Y., and Chen, R. (2016). NPInter v3.0: an upgraded database of noncoding RNA-associated interactions. Database (Oxford) *2016*, baw057.

39. Shi, J.Y., Yiu, S.M., Li, Y., Leung, H.C., and Chin, F.Y. (2015). Predicting drug-target interaction for new drugs using enhanced similarity measures and super-target clustering. Methods *83*, 98–104.

40. Lobo, J.M., Jiménez-Valverde, A., and Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. *17*, 145–151.

41. Li, A., Ge, M., Zhang, Y., Peng, C., and Wang, M. (2015). Predicting Long Noncoding RNA and Protein Interactions Using Heterogeneous Network Model. BioMed Res. Int. *2015*, 671950.

42. Gan, M. (2014). Walking on a user similarity network towards personalized recommendations. PLoS ONE *9*, e114662.

43. Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.

44. Wang, T., Qu, L., and Li, Y. (2015). Structures and Functions of Long Non-coding RNAs and Its Roles in Diseases. Zhongguo Sheng Wu Hua Xue Yu Fen Zi Sheng Wu Xue Bao *31*, 659–666.

45. Li, L., and Song, X. (2014). [In vivo functions of long non-coding RNAs]. Yi Chuan *36*, 228–236.

46. Lan, Y., and Song, X. (2014). Interaction of long non-coding RNA and protein. Chem. Life *34*, 473–478.

47. Wierzbicki, A.T. (2012). The role of long non-coding RNA in transcriptional gene silencing. Curr. Opin. Plant Biol. *15*, 517–522.

48. Bai, Y., Dai, X., Harrison, A.P., and Chen, M. (2015). RNA regulatory networks in animals and plants: a long noncoding RNA perspective. Brief. Funct. Genomics *14*, 91–101.

49. Chen, X., and Huang, L. (2017). LRSSLMDA: Laplacian Regularized Sparse Subspace Learning for MiRNA-Disease Association prediction. PLoS Comput. Biol. *13*, e1005912.

50. You, Z.H., Huang, Z.A., Zhu, Z., Yan, G.Y., Li, Z.W., Wen, Z., and Chen, X. (2017). PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. PLoS Comput. Biol. *13*, e1005455.

51. Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., and Yan, G. (2016). NLLSS: Predicting Synergistic Drug Combinations Based on Semi-supervised Learning. PLoS Comput. Biol. *12*, e1004975.

52. Chen, X., Huang, L., Xie, D., and Zhao, Q. (2018). EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction. Cell Death Dis. *9*, 3.

53. Da Sacco, L., Baldassarre, A., and Masotti, A. (2012). Bioinformatics tools and novel challenges in long non-coding RNAs (lncRNAs) functional analysis. Int. J. Mol. Sci. *13*, 97–114.

54. Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., and Liu, H. (2018). HLPI-Ensemble: Prediction of human lncRNA-protein interactions based on ensemble strategy. RNA Biol. *15*, 797–806.

55. Yuan, J., Wu, W., Xie, C., Zhao, G., Zhao, Y., and Chen, R. (2014). NPInter v2.0: an updated database of ncRNA interactions. Nucleic Acids Res. *42*, D104–D108.

56. Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R., and Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. *42*, D98–D103.

57. Zhao, Y., Yuan, J., and Chen, R. (2016). NONCODEv4: Annotation of Noncoding RNAs with Emphasis on Long Noncoding RNAs. Methods Mol. Biol. *1402*, 243–254.

58. Consortium, U.P.; UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204–D212.

59. Pundir, S., Martin, M.J., and O'Donovan, C.; UniProt Consortium (2016). UniProt Tools. Curr. Protoc. Bioinformatics *53*, 1–15.