MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **85** (2021) 555-574

ISSN 0340 - 6253

LncRNA-Protein Interaction Prediction Based on Regularized Nonnegative Matrix Factorization and Sequence Information

Da Xu[1], Hanxiao Xu[1], Yusen Zhang[1,*], Wei Chen[1], Rui Gao[2]

[1]*School of Mathematics and Statistics, Shandong University, Weihai 264209, China*

[2]*School of Control Science and Engineering, Shandong University, Jinan 250061, China*

(Received May 8, 2020)

Abstract

lncRNA affects the expression of nearby protein-coding genes and interfaces with related RNA binding proteins to exert functions. It is necessary to develop new computational models, which can reduce the cost and time of the biological experiments and select the most promising lncRNA-protein pairs for experimental validation. In this work, we propose a novel model called LPI-RNMF to identify the lncRNA-protein interaction (LPI) by using a new regularized nonnegative matrix factorization (RNMF) algorithm. First, LPI-RNMF extracts integrated lncRNA and protein similarity matrixes by sequences-based normalized Smith-Waterman score and known lncRNA-protein association matrix-based Gaussian interaction profile kernel, respectively. Then, a new regularized nonnegative matrix factorization algorithm is proposed and utilized to predict potential interactions. We conduct 5-fold cross-validation experiments on the benchmark data set, the AUC value is 0.9102 and AUPR value is 0.7245. In addition, leave-one-out cross-validation (LOOCV) is implemented and the AUC value is 0.9210. The comparison results are significantly higher than other methods mentioned. Moreover, case studies and implementing a test on a novel data set also demonstrate the stable performance of our method. These experimental results suggest that LPI-RNMF is a useful tool in predicting unknown lncRNA-protein interactions.

* Corresponding author: zhangys@sdu.edu.cn

1 Introduction

Non-coding RNA (ncRNA) was considered as transcriptional noise until within the development of biological research, more and more evidence showed up ncRNA is important functional expression in the genome [1,2]. Long ncRNAs (lncRNAs) (> 200 bp) account for the largest proportion in the human transcriptome and have gained wide attention in recent decades [3–5].

Accumulating evidence has shown that lncRNA plays a key role in various biological processes, including epigenetic regulation, gene transcriptional regulation, chromatin modification, protein transport, trafficking, cell differentiation and cellular apoptosis, and so on [6–10]. Furthermore, lncRNAs remain important in the regulations of many human complex diseases such as various types of cancer [1,11–13]. There are some databases of lncRNA associated with diseases such as the LncRNADisease [14] and the Lnc2Cancer [15].

Generally, lncRNA affects the expression of nearby protein-coding genes and interfaces with related RNA binding proteins to exert functions [16–19]. It is an important way to reveal functions and enrich the annotations of lncRNA by predicting the lncRNA-protein interactions (LPIs). Now, more and more lncRNAs have been discovered but the functionality of most lncRNAs remains unknown [20,21]. Computational approaches could reduce significantly the cost and time of the biological experiments, we can select the most promising lncRNA-protein pairs for experimental verification [22,23]. At present, various computational models have been widely proposed to solve the biological problems such as miRNA-disease [23,24], drug-target [25,26], lncRNA-protein [27,28], microbe-disease associations [29] and protein-protein interaction predictions [30]. But there are only a few computational approaches for predicting LPIs.

Overall, there are two types of computational methods for predicting LPIs: machine learning-based models and network-based models [18,19,27,31]. Machine learning-based models construct supervised classification prediction models by fusing the features of sequence, structure, and physicochemical property. The work is usually formulated as the binary classification. For example, in 2011, Bellucci et al. [32] designed catRAPID method based on the secondary structure, hydrogen and van der Waals of sequences. Muppirala et al. [33]

presented RPISeq method, in the same year, which adopted Random Forest (RF) and Support Vector Machine (SVM) classifiers using only sequence information. Two years later Lu et al. [34] proposed lncPro method by using matrix multiplication and encoding sequence information into numeric vector. Later, Wang et al. [35] presented an extended naive-Bayes-classifier. In 2015, Suresh et al. [36] presented a model (RPI-Pred) using sequence and structural information, which was based on the SVM classifier. In 2019, Wekesa et al. [37] proposed PLRPIM method by combining shallow machine learning and deep learning methods for plant LPIs.

Nowadays, the network-based method has been a widely used tool for predicting potential LPIs. It is primarily based on the fusion of known interactions and heterogeneous data to construct network. Usually, it is considered as a semi-supervised task and more suitable for predicting unobserved LPIs. For example, in 2015, Yang et al. [38] adopted the Hetesim algorithm to evaluate relevance between proteins and lncRNAs. Li et al. [39] presented the LPIHN method using a random walk with restart in the heterogeneous network, in the same year. Ge et al. [40] presented LPBNI method based on the lncRNA-protein bipartite network. In 2017, in order to boost the predictable performance, Zheng et al. [41] fused multiple protein similarity networks. In 2018, Zhang et al. [19] designed a sequence-based feature projection ensemble learning method for predicting. Hu et al. [42] designed an eigenvalue transformation-based prediction model. They also proposed an integration model using the neighborhood regularized logistic matrix factorization and the random walk algorithm for predicting [43]. Recently, Shen et al. proposed LPI-KTASLP model using multivariate information fusion [44].

However, there are some computational models have limitations: (1) Machine learning-based methods are affected by the imbalance of negative and positive samples, and rely heavily on the negative samples that are difficult to obtain. Moreover, they are usually unable to retain the topological information of the known LPI bilayer network. (2) Some network-based models utilize various lncRNA and protein information for multivariate information fusion, but these methods cannot work if some kinds of information are unavailable for some lncRNAs or proteins [19]. Beyond that, LPI-ETSLP and IRWNRLPI used theoretical parameters that may not apply to new data [21]. (3) The constructions of some methods are based on multivariate information fusion or multiple algorithms, which may be time-consuming.

The matrix factorization method is a useful tool and has been widely used in the recommended system [45–47]. To boost the predictable performance, we transform the identification problem of LPIs into a recommended task and propose a regularized nonnegative matrix factorization algorithm called LPI-RNMF to predict the potential lncRNA-protein associations. LPI-RNMF integrates the lncRNA similarity matrix, protein similarity matrix, and known lncRNA-protein association matrix. It uses a semi-supervised learning strategy to discover unknown associations and does not need negative samples. In the calculation process, it uses less prior feature information and model parameters, thus saving time and getting robust performance. The experimental results show that LPI-RNMF achieves superior performance compared with the state-of-the-art methods and is a promising method in the predicting of unknown LPIs.

2 Methods

2.1 Method overview

To infer the potential lncRNA-protein associations, we propose a new model called LPI-RNMF. The model consists of two steps (Fig. 1). In the first step, we got the integrated lncRNA similarity matrix S_l and protein similarity matrix S_p by fully exploiting sequence similarity and Gaussian interaction profile (GIP) kernel similarity, respectively. Then, we propose a novel regularized nonnegative matrix factorization algorithm to discover the potential lncRNA-protein associations.

2.2 Similarity measures

Let $P = \{P_1, P_2, \dots, P_m\}$ and $L = \{L_1, L_2, \dots L_n\}$ denote the sets of proteins and lncRNAs, where m and n denote the numbers of proteins and lncRNAs, respectively. To facilitate the description of the method and lncRNA-protein associations, we introduced the matrix $A \in R^{n \times m}$ as an adjacency matrix of lncRNA-protein associations, where $A_{ij} = 1$ if lncRNA l_i has a known association with protein p_j, otherwise $A_{ij} = 0$. Set $A(l_i) = (l_{i1}, l_{i2}, \dots, l_{im})$ denotes the ith row of the adjacency matrix A, which represents the interaction profile of the lncRNA l_i. And $A(p_j) = (p_{1j}, p_{2j}, \dots, p_{nj})$ denotes the jth column of the adjacency matrix

A, which represents the interaction profile of the protein p_j.

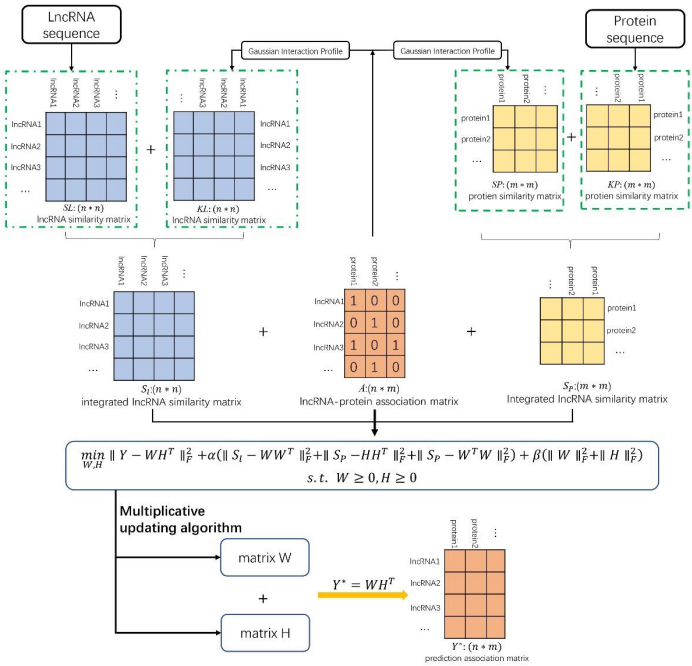

Figure 1. The overall workflow of LPI-RNMF

2.2.1 lncRNA similarity measures

The basic assumption that similar lncRNAs (proteins) share similar interaction or noninteraction pattern with proteins (lncRNAs) is widely used in the related studies [18,27]. The GIP kernel similarity has been widely applied to bioinformatics and related fields [29,31]. In this work, for a given lncRNA l_i, we first extract the $A(l_i)$ to represent the interaction profile of lncRNA l_i from the training adjacency matrix A. Subsequently, we use the following equation to calculate the GIP kernel similarity between the lncRNA l_i and l_q:

$$KL(l_i, l_q) = exp(-\sigma_l ||A(l_i) - A(l_q)||^2) \quad (1)$$

$$\sigma_l = \sigma_l' / \left(\frac{1}{n} \sum_{k=1}^{n} ||A(l_k)||^2 \right) \quad (2)$$

where σ_l represents the normalized interaction profile kernel bandwidth, σ_l' is a parameter

and set $\sigma_l' = 1$ in the experiments. Matrix KL denotes the lncRNA GIP kernel similarity matrix. Practically, we will recalculate the GIP kernel similarity based on the training samples every time during the LOOCV or 5-fold cross-validation.

Next, we calculate the normalized Smith-Waterman (SW) score [48,49] for measuring the sequence similarities between lncRNA pairs. The sequence similarity between lncRNA l_i and l_q can be calculated as follows:

$$SL(l_i, l_q) = \frac{SW\left(S_{l_i}, S_{l_q}\right)}{\sqrt{SW(S_{l_i}, S_{l_i})SW\left(S_{l_q}, S_{l_q}\right)}} \tag{3}$$

where $SW\left(S_{l_i}, S_{l_q}\right)$ denotes the normalized SW score of lncRNA l_i and l_q. S_{l_i} and S_{l_q} represent the sequence information of lncRNA l_i and l_q, respectively. Matrix SL denotes the lncRNA sequence similarity matrix.

2.2.2 Protein similarity measures

We calculate the similarity of proteins in the same way as the lncRNA. The GIP kernel similarity of protein p_j and p_t can be calculated as follows:

$$KP(p_j, p_t) = exp(-\sigma_p||A(p_j) - A(p_t)||^2) \tag{4}$$

$$\sigma_p = \sigma_p' / \left(\frac{1}{m}\sum_{k=1}^{m}||A(p_k)||^2\right) \tag{5}$$

where σ_p represents the normalized interaction profile kernel bandwidth, σ_p' is a parameter and set $\sigma_p' = 1$ in the experiment. Matrix KP denotes the protein GIP similarity matrix.

In the same manner, the sequence similarity between protein p_j and p_t can be measured by the normalized SW score as follows:

$$SP(p_j, p_t) = \frac{SW\left(S_{p_j}, S_{p_t}\right)}{\sqrt{SW\left(S_{p_j}, S_{p_j}\right)SW(S_{p_t}, S_{p_t})}} \tag{6}$$

where $SW\left(S_{p_j}, S_{p_t}\right)$ denotes the normalized SW score of protein p_j and p_t. S_{p_j} and S_{p_t} represent the sequence information of protein p_j and p_t, respectively. Matrix SP denotes the protein sequence similarity matrix.

2.2.3 Integrating similarity

Note that the GIP kernel similarity is measured by the association information and interaction profile, providing an extensible prediction framework by complementing sequence similarity. Enlightened by some studies [29,31], we obtain integrated lncRNA similarity matrix S_l and an integrated protein similarity matrix S_p as follows:

$$S_l(l_i, l_q) = \frac{KL(l_i, l_q) + SL(l_i, l_q)}{2} \tag{7}$$

$$S_p(p_j, p_t) = \frac{KP(p_j, p_t) + SP(p_j, p_t)}{2} \tag{8}$$

Specifically, the integrated similarity matrix is calculated as a mean matrix of the sequence similarity matrix and GIP kernel similarity matrix.

2.3 Regularized nonnegative matrix factorization algorithm

2.3.1 Standard NMF

Nonnegative matrix factorization (NMF) is a popular technique for multivariate analysis of nonnegative data and widely used for feature learning and data faithful representation in the field of bioinformatics and computer vision [50,51]. The goal of NMF is to seek a decomposition of one nonnegative matrix Y and obtain two nonnegative matrices whose product will be the best approximation of Y. The problem of lncRNA-protein association prediction can be transformed into the NMF problem:

$$\min_{W,H} \| Y - WH^T \|_F^2 \tag{9}$$

$$s.t.\ \ W \geq 0, H \geq 0$$

where $||\cdot||_F$ is Frobenius norm of the matrix, Y is the lncRNA-protein adjacency matrix, $Y \geq 0$ and $Y \epsilon R^{n \times m}$. W and H are nonnegative matrices and $W \epsilon R^{n \times k}$, $H \epsilon R^{m \times k}$. Lee et al. [52] proposed an updating algorithm to solve this optimization problem.

2.3.2 RNMF

Collaborative Matrix Factorization (CMF) has been applied in the field of bioinformatics [53,54]. According to the basic hypothesis, proteins with similar functions will tend to be involved in similar lncRNAs and vice versa, which is consistent with the observations of biological experiments [19,27]. So, similar proteins will be more likely to share similar

interaction or noninteraction pattern with lncRNAs. Inspired by the former research [55], to integrate more effective information, the regularization terms are incorporated into the LPI-RNMF framework to guide the nonnegative matrix factorization process and measure the low-dimensional representation of lncRNA-protein adjacency matrix. A novel objective function was designed and the optimization problem can be formularized as follows:

$$\min_{W,H} \| Y - WH^T \|_F^2 + \alpha\left(\| S_l - WW^T \|_F^2 + \| S_p - HH^T \|_F^2 + \| S_p - W^T W \|_F^2\right)$$
$$s.t.\ \ W \geq 0, H \geq 0 \tag{10}$$

where S_l and S_p are integrated similarity matrixes defined above, α is the regularization coefficient for adjusting the contribution of lncRNA and protein similarity information. In the experiment, k was set to the number of proteins on account of regularization terms. In the objective function, the first part is a standard NMF for searching the latent low-dimensional information matrices. In the second part, the effect of the first two regularization items is to minimize the squared error between S_l (S_p) and WW^T (HH^T), and the last proposed regularization item is based on the basic hypothesis that similar proteins will be more likely to share similar interaction or noninteraction pattern with lncRNAs to integrate more effective information.

To adjust the smoothness of W and H and prevent over-fitting, the Tikhonov (L_2) regularization terms [45,56] are incorporated into the LPI-RNMF framework for lncRNA-protein association prediction:

$$\min_{W,H} \| Y - WH^T \|_F^2 + \alpha\left(\| S_l - WW^T \|_F^2 + \| S_p - HH^T \|_F^2 + \| S_p - W^T W \|_F^2\right)$$
$$+ \beta(\| W \|_F^2 + \| H \|_F^2)$$
$$s.t.\ \ W \geq 0, H \geq 0 \tag{11}$$

where β is also the regularization coefficient used to adjust the Tikhonov regularization term. In the objective function, the third part is used to prevent over-fitting. To simplify the complex problem and promote the robust performance of the model, we introduced the same parameter for the same regularization term.

2.3.3 Optimization

In this work, the Lagrange multipliers method is applied to the optimization problem Eq. (11). First, according to the Frobenius norm and trace property of matrix, two kinds of regularization terms can be transformed into:

$$\begin{aligned} R_1 =& \| S_l - WW^T \|_F^2 + \| S_p - HH^T \|_F^2 + \| S_p - W^T W \|_F^2 \\ =& Tr\left(S_l {S_l}^T\right) - 2Tr(S_l W W^T) + Tr(WW^T WW^T) \\ &+ Tr\left(S_p {S_p}^T\right) - 2Tr\left(S_p H H^T\right) + Tr(HH^T HH^T) \\ &+ Tr\left(S_p {S_p}^T\right) - 2Tr\left(S_p W^T W\right) + Tr(W^T WW^T W) \end{aligned} \tag{12}$$

$$R_2 = \| W \|_F^2 + \| H \|_F^2 = Tr(WW^T) + Tr(HH^T) \tag{13}$$

where $Tr(\cdot)$ denotes the trace of matrix.

We set $\Phi = [\varphi_{ik}]$ and $\Psi = [\psi_{jk}]$ are the Lagrange multipliers of the constrains $w_{ik} \geq 0$ and $h_{jk} \geq 0$, respectively. Then, we define the Lagrange function L_f as follows:

$$\begin{aligned} L_f =& Tr(YY^T) - 2Tr(YHW^T) + Tr(WH^T HW^T) + \alpha R_1 + \beta R_2 \\ &+ Tr(\Phi W^T) + Tr(\Psi H^T) \end{aligned} \tag{14}$$

After using the Karush-Kuhn-Tucker (KKT) conditions [57], we can obtain the following updating rules:

$$w_{ik} = w_{ik} \frac{\left(YH + 2\alpha\left(S_l W + W S_p\right)\right)_{ik}}{(WH^T H + \beta W + 4\alpha WW^T W)_{ik}} \tag{15}$$

$$h_{jk} = h_{jk} \frac{\left(Y^T W + 2\alpha S_p H\right)_{jk}}{(HW^T W + \beta H + 2\alpha HH^T H)_{jk}} \tag{16}$$

The matrices W and H are updated alternately based on the Eq. (15) and Eq. (16) until convergence. Eventually, we will obtain the predicted lncRNA-protein adjacency matrix $Y^* = WH^T$ and prioritize the protein-related lncRNAs (lncRNA-related proteins) based on the predictive values in the matrix Y^*. Y_{ij}^* is the score measuring how likely lncRNA l_i is associated with protein p_j.

3 Results

3.1 Datasets

To further evaluate the performance of LPI-RNMF, the summary of two publicly lncRNA-protein data sets used in the experiment is tabulated in Table 1. The first data set was a widely-used data set and collected by Zhang et al. [18]. It is used as a benchmark data set. The LPI data were downloaded from the NPInter v2.0 [58] database which is one of the most comprehensive databases of interactions between other biomolecules and ncRNAs. In the experiment, the lncRNA sequences were downloaded from the NONCODE [59] database which is an integrated knowledge database about ncRNAs. In addition, the protein sequences were downloaded from the SUPERFAMILY 2.0 [60] database which stores millions of protein sequences. The second data set was collected by Zheng et al. [41], it is used as a novel data set in the experiment. We implement a test on the novel data set to demonstrate the reliable and effective prediction performance of the method proposed by us.

Table 1. Summary of two data sets in the experiment.

Data sets	lncRNAs	Proteins	Associations
Benchmark data set	990	27	4158
Novel data set	1050	84	4467

3.2 Evaluation metrics

To assess the effectiveness of LPI-RNMF and other computational models, in the same experimental conditions, we implemented leave-one-out cross-validation (LOOCV) and 5-fold cross-validation (5-CV) schemes. In the 5-CV framework, the experiment was repeated for 20 times for every model since random sample division may cause the potential experimental bias. Sensitivity and 1-specificity were obtained and plotted the receiver-operating characteristics (ROC) curves to intuitively assess the performance. Moreover, we also calculated the areas under the ROC curve (AUC) to evaluate the performance of different methods. The AUC value of 0.5 means random performance and the AUC value of 1 indicates perfect prediction. The ratio between the known lncRNA-protein associations and the unobserved ones has a serious imbalance in the data set. The precision-recall (PR) curve [61] is more suitable than the ROC

curve when different classes are imbalanced. Therefore, we also used the PR curve and AUPR to measure the performance. In the LOOCV framework, the corresponding AUC values were also calculated to intuitively assess the performance.

Beyond that, we also used several widely used evaluation metrics, including accuracy (ACC), sensitivity (SEN), precision (PRE), and $F_1\ score$, expressed as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

$$SEN = \frac{TP}{TP + FN} \tag{18}$$

$$PRE = \frac{TP}{TP + FP} \tag{19}$$

$$F_1\ score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{20}$$

where TP, FP, TN, and FN represent the number of true positives, false positives, true negatives, and false negatives, respectively. In the lncRNA-protein association data set, the unknown lncRNA-protein associations are considered negative samples, while the known associations are called positive samples.

Average performances of all evaluation metrics were obtained during the experiment. The performance of the model will be better along with the value of the metric becomes larger.

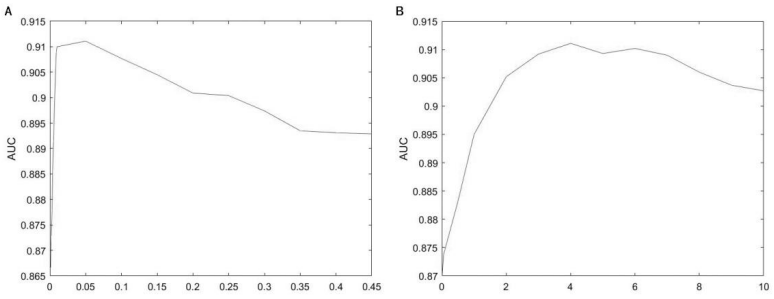

Figure 2. (A) The relationship between AUC value of the LPI-RNMF and parameter α. (B) The relationship between AUC value of the LPI-RNMF and parameter β.

3.3 Model setting

The values of α and β are important parameters of LPI-RNMF, and their different scale values will influence the prediction performance. To explore the properties of the proposed method, we applied the grid search method and some comparison experiments to find the best

model parameters. In order to explore the influences of α and β, we plotted the figures as shown in Fig. 2. From the figure, the prediction performance of LPI-RNMF is greatly enhanced when α increases from 0 to 0.05, indicating the second part of objective function provides very effective information (see Fig. 2A), and the performance keeps a decreasing trend as the value of α increases from 0.05 to 0.45. The trend of β is similar with the parameter α (see Fig. 2B). Finally, the parameter values of α and β were set 0.05 and 4 for the model to obtain optimal and stable prediction performance.

The proposed optimization function is convergence which is demonstrated through the experiments. Fig. 3A shows the convergence curve on the benchmark data set. From the figure, we can see that the objective function value decreases as the iterations.

3.4 Comparison with benchmark prediction methods

To evaluate the performance of LPI-RNMF, we compared it with five benchmark computational methods, including LPI-FKLKRR [48], LPBNI [40], LPIHN [39], collaborative filtering (CF) and random walk with restart (RWR) [18]. In this paper, we reproduced these computational models under the same experimental conditions on the same benchmark data set, and conducted 5-CV framework for comparison. The comparison results have been shown in Fig. 3B, Fig. 3C and Fig. 4, and all details are presented in Table 2. Fig. 3B is the ROC curves of different methods under 5-CV. From the figure, the AUC values of other comparative methods are the following: LPI-FKLKRR (AUC: 0.9045), RWR (AUC: 0.8309), CF (AUC: 0.7683), LPBNI (AUC: 0.8564) and LPIHN (AUC: 0.8442). Our method achieves average AUC value (0.9102), which is higher than other methods under the same condition. Fig. 3C is the PR curves of different methods under 5-fold cross-validation. We can observe that our method obtains the highest average AUPR value (0.7245), indicating that our method is more reliable. In addition, we also obtain other values of the evaluation metrics, including ACC, SEN, PRE and F_1 score, which are presented in Table 2. From Fig. 4, we can observe that LPI-RNMF is superior to other methods under the same experimental conditions. In general, these can suggest that LPI-RNMF is a promising tool in predicting unknown LPIs.

In addition, we reproduced these computational methods on the benchmark data set under the LOOCV framework. The corresponding AUC values and ROC curves were calculated in a

similar way with 5-fold cross-validation. As shown in Fig. 3D, the AUC value of LPI-RNMF is 0.9210, which is significantly better than that of RWR (0.8577), CF (0.7927), LPBNI (0.8823) and LPIHN (0.8800). To sum up, the experimental results demonstrate the prediction performance of LPI-RNMF is reliable and effective.

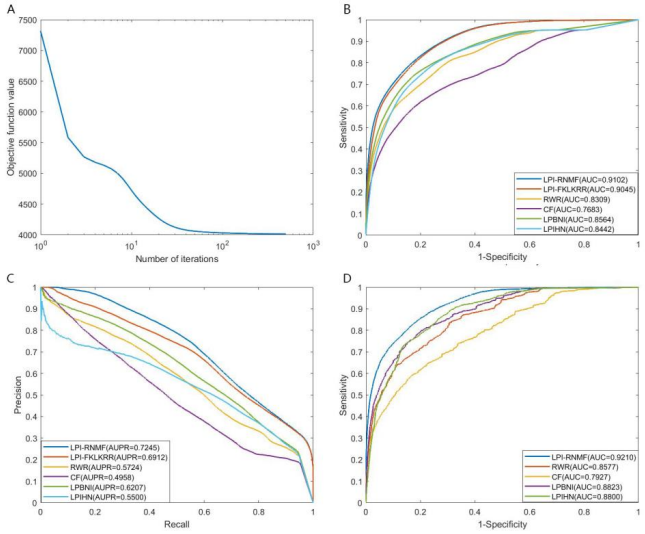

Figure 3. (A) Convergence behavior of objection function on the benchmark data set. (B) The ROC curves of different methods for LPI prediction under 5-fold cross-validation. (C) The PR curves of different methods for LPI prediction under 5-fold cross-validation. (D)The ROC curves of different methods for LPI prediction under LOOCV.

Table 2. The comparison results of different methods via 20 runs of 5-CV on benchmark dataset.

	AUC	AUPR	SEN	PRE	ACC	F_1 score
LPI-RNMF	**0.9102**	**0.7245**	0.6259	**0.6754**	**0.8943**	**0.6482**
LPI-FKLKRR	0.9045	0.6912	0.6174	0.6544	0.8893	0.6344
RWR	0.8309	0.5724	0.5641	0.5505	0.8597	0.5557
CF	0.7683	0.4958	0.4928	0.4870	0.8376	0.4860
LPBNI	0.8564	0.6207	0.6033	0.5791	0.8683	0.5877
LPIHN	0.8442	0.5500	**0.6528**	0.5002	0.8436	0.5649

3.5 Case studies

The purpose of the case study is to verify the power of the model for predicting new protein without any known related lncRNAs. We masked all relationships between all lncRNAs and the same protein. The model was trained with the rest of the known associations and tested on

masked associations (validation data sets). Specifically, in the protein case studies, we converted all 1 to 0 corresponding to the same protein in the lncRNA-protein adjacency matrix A and sorted all lncRNA samples (positive and unknown samples) that are related to this protein based on the model. Then, we chose the top ranked predictions to further verify the performance. The model will be efficient if the top predictions have more positive samples.

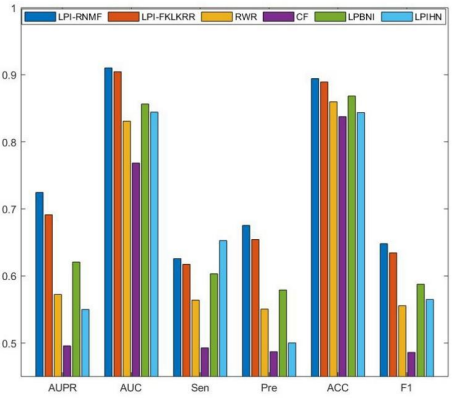

Figure 4. Performance of different methods on benchmark data set via 20 runs of 5-CV.

In the protein case studies, the top 10 prediction associations of two protein (ENSP00000401371 and ENSP00000381031) were extrapolated by LPI-RNMF as shown in Table 3. Our method achieved 9/10 and 10/10 successful prediction performance, respectively. The previous works LPI-FKLKRR [48] and LPI-KTASLP [44] also used the protein ENSP00000401371 to make case studies and achieved 6/10 and 5/10 successful prediction performance, respectively. In the lncRNA case studies, since the number of the protein is small, the top 5 associations of two lncRNAs (NONHSAT031708 and NONHSAT007429) were extrapolated by the LPI-RNMF as shown in Table 4, respectively. All predictions of these two lncRNAs are right. The previous work LPI-FKLKRR also used the lncRNA NONHSAT031708 to make a case study and achieved 3/5 successful prediction performance. If we have a new protein, the results demonstrate that our method can predict the possibility of interaction between this new protein and 990 lncRNAs used in the experiment.

3.6 Evaluation on novel data set

A novel data set is further tested to demonstrate the stable performance of our algorithm. The summary information of the novel data set can be found in Table 1. The 5-fold cross-validation was implemented on the novel data set, and the comparison results have been shown in Table 5. From the table, we can observe that LPI-RNMF achieves the average AUPR value of 0.7459 and the average AUC value of 0.9739. The performance is higher than LPI-KTASLP and LPI-FKLKRR methods. The results indicate that better robustness performance can be obtained by our method on the new data set.

Table 3. Top 10 associations of protein ENSP00000401371 and ENSP00000381031

lncRNA ID	Protein ID	Rank	Confirm?
NONHSAT021830	ENSP00000401371	1	Confirmed
NONHSAT137541	ENSP00000401371	2	Confirmed
NONHSAT135796	ENSP00000401371	3	Confirmed
NONHSAT041921	ENSP00000401371	4	Confirmed
NONHSAT009703	ENSP00000401371	5	Confirmed
NONHSAT138142	ENSP00000401371	6	Confirmed
NONHSAT027070	ENSP00000401371	7	Confirmed
NONHSAT104639	ENSP00000401371	8	——
NONHSAT104991	ENSP00000401371	9	Confirmed
NONHSAT011652	ENSP00000401371	10	Confirmed
NONHSAT021830	ENSP00000381031	1	Confirmed
NONHSAT137541	ENSP00000381031	2	Confirmed
NONHSAT135796	ENSP00000381031	3	Confirmed
NONHSAT009703	ENSP00000381031	4	Confirmed
NONHSAT041921	ENSP00000381031	5	Confirmed
NONHSAT138142	ENSP00000381031	6	Confirmed
NONHSAT027070	ENSP00000381031	7	Confirmed
NONHSAT104991	ENSP00000381031	8	Confirmed
NONHSAT011652	ENSP00000381031	9	Confirmed
NONHSAT001511	ENSP00000381031	10	Confirmed

Table 4. Top 5 associations of lncRNA NONHSAT031708 and NONHSAT007429

lncRNA ID	Protein ID	Rank	Confirm?
NONHSAT031708	ENSP00000385269	1	Confirmed
NONHSAT031708	ENSP00000258729	2	Confirmed
NONHSAT031708	ENSP00000240185	3	Confirmed
NONHSAT031708	ENSP00000371634	4	Confirmed
NONHSAT031708	ENSP00000290341	5	Confirmed
NONHSAT007429	ENSP00000385269	1	Confirmed
NONHSAT007429	ENSP00000258729	2	Confirmed
NONHSAT007429	ENSP00000371634	3	Confirmed
NONHSAT007429	ENSP00000240185	4	Confirmed
NONHSAT007429	ENSP00000290341	5	Confirmed

Table 5. Performance of different methods via 5-fold cross-validation on the novel dataset.

Methods	AUC	AUPR
LPI-RNMF	**0.9739**	**0.7459**
LPI-KTASLP	0.9152	0.7173
LPI-FKLKRR	0.9669	0.7062

4 Conclusion and discussion

In this paper, a novel model called LPI-RNMF is proposed to reveal the potential LPIs based on the regularized nonnegative matrix factorization algorithm. In the 5-fold cross-validation and LOOCV evaluation framework, the AUC values of LPI-RNMF are 0.9102 and 0.9210, respectively, which are better than other computational methods compared in this paper. Furthermore, case studies and implementing a test on a novel data set also demonstrate that LPI-RNMF has a robust and stable performance. The results confirm that the proposed method is promising in predicting unknown LPIs.

Several factors leading to LPI-RNMF prediction performance are summarized as follows. Firstly, LPI-RNMF discover unknown associations using a semi-supervised learning strategy and does not require negative samples. Secondly, the calculation process of LPI-RNMF uses less prior information and model parameters, thus saving time and promoting the robust performance. Thirdly, it is based on the regularized nonnegative matrix factorization, not only integrating GIP kernel similarity and sequence similarity but also mining conveniently the topological structure information. Of course, LPI-RNMF also has some limitations, such as combining reasonably various kinds of data sources will be a challenge for the computational model if some supplementary data is introduced.

Acknowledgments: We thank the reviewers for their time reading the paper and constructive comments. This work was supported by the National Natural Science Foundation of China under Grant (No. 61877064, U1806202 and 61533011).

References

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

[2] V. Mohanty, Y. Gökmen-Polar, S. Badve, S. C. Janga, Role of lncRNAs in health and disease-size and shape matter, *Brief. Funct. Genomics*. **14** (2015) 115–129.

[3] P. Volders, K. Helsens, X. Wang, L. Martens, K. Gevaert, J. Vandesompele, P. Mestdagh, LNCipedia: a database for annotated human lncRNA transcript sequences and structures, *Nucleic Acids Res*. **41** (2013) 246–251.

[4] J. R. Prensner, A. M. Chinnaiyan, The emergence of lncRNAs in cancer biology, *Cancer Discov*. **1** (2011) 391–407.

[5] M. Hajjari, S. J. Mowla, M. A. Faghihi, Editorial: Molecular function and regulation of non-coding RNAs in multifactorial diseases, *Front. Genet*. **7** (2016) 2015–2016.

[6] M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, E. S. Lander, Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals, *Nature* **458** (2009) 223–227.

[7] C. P. Ponting, P. L. Oliver, W. Reik, Evolution and functions of long noncoding RNAs, *Cell*. **136** (2009) 629–641.

[8] K. C. Wang, H. Y. Chang, Molecular mechanisms of long noncoding RNAs, *Mol. Cell*. **43** (2012) 904–914.

[9] M. Guttman, J. L. Rinn, Modular regulatory principles of large non-coding RNAs, *Nature*. **482** (2012) 339–346.

[10] S. Geisler, J. Coller, RNA in unexpected places: Long non-coding RNA functions in diverse cellular contexts, *Nat. Rev. Mol. Cell Biol*. **14** (2013)699–712.

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

[12] L. W. Harries, Long non-coding RNAs and human disease, *Biochem. Soc. Trans*. **40** (2012) 902–906.

[13] X. Chen, G. Y. Yan, Novel human lncRNA-disease association inference based on lncRNA expression profiles, *Bioinf.* **29** (2013) 2617–2624.

[14] G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan, Q. Cui, LncRNADisease: a database for long-non-coding RNA-associated diseases, *Nucleic Acids Res*. **41** (2013) 983–986.

[15] S. Ning, J. Zhang, P. Wang, H. Zhi, J. Wang, Y. Liu, Y. Gao, M. Guo, M. Yue, L. Wang, X. Li, Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers, *Nucleic Acids Res*. **44** (2016) 980–985.

[16] T. R. Mercer, M. E. Dinger, J. S. Mattick, Long non-coding RNAs: Insights into functions, *Nat. Rev. Genet*. **10** (2009) 155–159.

[17] J. König, K. Zarnack, N. M. Luscombe, J. Ule, Protein–RNA interactions: New genomic technologies and perspectives, *Nat. Rev. Genet*. **13** (2012) 77–83.

[18] W. Zhang, Q. Qu, Y. Zhang, W. Wang, The linear neighborhood propagation method for predicting long non-coding RNA-protein interactions, *Neurocomputing* **273** (2017) 526–534.

[19] W. Zhang, X. Yue, G. Tang, W. Wu, F. Huang, X. Zhang, SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting lncRNA-protein interactions, *PLoS Comput. Biol*. **14** (2018) 1–21.

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

[21] Q. Zhao, H. Yu, Z. Ming, H. Hu, G. Ren, H. Liu, The bipartite network projection-recommended algorithm for predicting long non-coding RNA-protein interactions, *Mol. Ther. Nucleic Acids*. **13** (2018) 464–471.

[22] T. Zhang, M. Wang, J. Xi, A. Li, LPGNMF: Predicting long non-coding RNA and protein interaction using graph regularized nonnegative matrix factorization, *IEEE/ACM Trans. Comput. Biol. Bioinf*. **17** (2018) 189–197.

[23] X. Chen, D. Xie, Q. Zhao, Z. H. You, MicroRNAs and complex diseases: From experimental results to computational models, *Brief. Bioinf*. **20** (2019) 515–539.

[24] X. Chen, L. Wang, J. Qu, N. N. Guan, J. Q. Li, Predicting miRNA-disease association based on inductive matrix completion, *Bioinf*. **34** (2018) 4256–4265.

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

[26] Z. Wu, W. Li, G. Liu, Y. Tang, Network-based methods for prediction of drug-target interactions, *Front. Pharm*. **9** (2018) 1–14.

[27] H. Zhang, Y. Liang, S. Han, C. Peng, Y. Li, Long noncoding RNA and protein interactions: From experimental results to computational models based on network methods, *Int. J. Mol. Sci*. **20** (2019) 1284.

[28] Y. Xiao, J. Zhang, L. Deng, Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks, *Sci. Rep*. **7** (2017) 1–12.

[29] X. Chen, Y. A. Huang, Z. H. You, G. Y. Yan, X. S. Wang, A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases, *Bioinf*. **33** (2016) 733–739.

[30] D. Xu, H. X. Xu, Y. S. Zhang, W. Chen, R. Gao, Protein-protein interactions prediction based on graph energy and protein sequence information, *Molecules* **25** (2020) 1841.

[31] G. B. Xie, C. Wu, Y. Sun, Z. Fan, J. Liu, LPI-IBNRA: Long non-coding RNA-protein interaction prediction based on improved bipartite network recommender algorithm, *Front. Genet*. **10** (2019) 1–10.

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

[33] U. K. Muppirala, V. G. Honavar, D. Dobbs, Predicting RNA-protein interactions using only sequence information, *BMC Bioinf.* **12** (2011) 489.

[34] Q. Lu, S. Ren, M. Lu, Y. Zhang, D. Zhu, X. Zhang, T. Li, Computational prediction of associations between long non-coding RNAs and proteins, *BMC Genomics* **14** (2013) 1.

[35] Y. Wang, X. Chen, Z. P. Liu, Q. Huang, Y. Wang, D. Xu, X. S. Zhang, R. Chen, L. Chen, De novo prediction of RNA-protein interactions from sequence information, *Mol. Biosyst*. **9** (2013) 133–142.

[36] V. Suresh, L. Liu, D. Adjeroh, X. Zhou, RPI-Pred: Predicting ncRNA-protein interaction using sequence and structural information, *Nucleic Acids Res*. **43** (2015) 1370–1379.

[37] J. S. Wekesa, Y. Luan, M. Chen, J. Meng, A hybrid prediction method for plant lncRNA-protein interaction, *Cells* **8** (2019) 521.

[38] J. Yang, A. Li, M. Ge, M. Wang, Prediction of interactions between lncRNA and protein by using relevance search in a heterogeneous lncRNA-protein network, *Chinese Control Conf. CCC*. (2015) 8540–8544.

[39] A. Li, M. Ge, Y. Zhang, C. Peng, M. Wang, Predicting long noncoding RNA and protein interactions using heterogeneous network model, *Biomed Res. Int*. **2015** (2015) 671950.

[40] M. Ge, A. Li, M. Wang, A bipartite network-based method for prediction of long non-coding RNA-protein interactions, *Genomics Proteomics Bioinf.* **14** (2016) 62–71.

[41] X. Zheng, Y. Wang, K. Tian, J. Zhou, J. Guan, L. Luo, S. Zhou, Fusing multiple protein-protein similarity networks to effectively predict lncRNA-protein interactions, *BMC Bioinf.* **18** (2017) 420.

[42] H. Hu, C. Zhu, H. Ai, L. Zhang, J. Zhao, Q. Zhao, H. Liu, LPI-ETSLP: LncRNA-protein interactions prediction using eigenvalue transformation-based semi-supervised link prediction, *Mol. Biosyst*. **3** (2017) 10715–10722.

[43] Q. Zhao, Y. Zhang, H. Hu, G. Ren, W. Zhang, H. Liu, IRWNRLPI: Integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction, *Front. Genet*. **9** (2018) 1–12.

[44] C. Shen, Y. Ding, J. Tang, L. Jiang, F. Guo, LPI-KTASLP: Prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information, *IEEE Access*. **7** (2019) 13486–13496.

[45] Q. Xiao, J. Luo, C. Liang, J. Cai, P. Ding, A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations, *Bioinf.* **34** (2018) 239–248.

[46] A. Hernando, J. Bobadilla, F. Ortega, A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model, *Knowledge-Based Syst*. **97** (2016) 188–202.

[47] X. Luo, M. C. Zhou, S. Li, Z. You, Y. Xia, Q. Zhu, A nonnegative latent factor model for large-scale sparse matrices in recommender systems via alternating direction method, *IEEE Trans. Neural Networks Learn. Syst*. **27** (2016) 579–592.

[48] C. Shen, Y. Ding, J. Tang, F. Guo, Multivariate information fusion with fast kernel learning to kernel ridge regression in predicting lncRNA-protein interactions, *Front. Genet*. **10** (2019) 1–12.

[49] T. F. Smith, M. S. Waterman, Identification of common molecular subsequences, *J. Mol. Biol*. **147** (1981) 195–197.

[50] S. Choi, Algorithms for orthogonal nonnegative matrix factorization, *2008 IEEE Int. Jt. Conf. Neural Networks*. (2008) 1828–1832.

[51] X. Li, G. Cui, Y. Dong, Graph regularized non-negative low-rank matrix factorization for image clustering, *IEEE Trans. Cybern*. **47** (2016) 3840-3853.

[52] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature*. **401** (1999) 788–791.

[53] X. Zheng, H. Ding, H. Mamitsuka, S. Zhu, Collaborative matrix factorization with multiple similarities for predicting drug-target interactions, *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. data Min*. (2013) 1025–1033.

[54] Z. Shen, Y. Zhang, K. Han, A. K. Nandi, B. Honig, D. Huang, MiRNA-disease association prediction with collaborative matrix factorization, *Complexity*. **2017** (2017) 1-9.

[55] P. Xuan, T. Shen, X. Wang, T. Zhang, W. Zhang, Inferring disease-associated microRNAs in heterogeneous networks with node attributes, *IEEE-ACM Trans. Comput. Biol. Bioinf*. (2019) 1-1.

[56] N. Guan, D. Tao, Z. Luo, B. Yuan, Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent, *IEEE Trans. Image Process*. **20** (2011) 2030–2048.

[57] F. Facchinei, C. Kanzow, S. Sagratella, Solving quasi-variational inequalities via their KKT conditions, *Math. Program*. **144** (2014) 369–412.

[58] J. Yuan, W. Wu, C. Xie, G. Zhao, Y. Zhao, R. Chen, NPInter v2.0: an updated database of ncRNA interactions, *Nucleic Acids Res.* **42** (2014) 104–108.

[59] C. Xie, J. Yuan, H. Li, M. Li, G. Zhao, D. Bu, W. Zhu, W. Wu, R. Chen, Y. Zhao, NONCODEv4: Exploring the world of long non-coding RNA genes, *Nucleic Acids Res*. **42** (2014) 98–103.

[60] P. Prasad, J. Stahlhacke, M. E. Oates, B. Smithers, J. Gough, The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver, *Nucleic Acids Res*. **47** (2019) 490–494.

[61] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, *PLoS One* **10** (2015) 1–21.