frontiers in Genetics

ORIGINAL RESEARCH
published: 16 July 2021
doi: 10.3389/fgene.2021.690096

Predicting lncRNA–Protein Interaction With Weighted Graph-Regularized Matrix Factorization

Xibo Sun[1], Leiming Cheng[2], Jinyang Liu[3,4], Cuinan Xie[3,4], Jiasheng Yang[5] and Fu Li[6]**

[1] Yidu Central Hospital of Weifang, Weifang, China, [2] Huaibei Kuanggong Zong Yiyuan, Huaibei, China, [3] Geneis Beijing Co., Ltd., Beijing, China, [4] Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China, [5] Academician Workstation, Changsha Medical University, Changsha, China, [6] Department of Thoracic Surgery, The Second Affiliated Hospital of Hainan Medical University, Haikou, China

OPEN ACCESS

Edited by:
Lihong Peng,
Hunan University of Technology, China

Reviewed by:
Guanghui Li,
East China Jiaotong University, China
JunLin Xu,
Hunan University, China

***Correspondence:**
Jiasheng Yang
jsyang.mcc@gmail.com
Fu Li
lifu_3251@163.com

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: *02 April 2021*
Accepted: *21 May 2021*
Published: *16 July 2021*

Citation:
Sun X, Cheng L, Liu J, Xie C, Yang J and Li F (2021) Predicting lncRNA–Protein Interaction With Weighted Graph-Regularized Matrix Factorization. Front. Genet. 12:690096. doi: 10.3389/fgene.2021.690096

Long non-coding RNAs (lncRNAs) are widely concerned because of their close associations with many key biological activities. Though precise functions of most lncRNAs are unknown, research works show that lncRNAs usually exert biological function by interacting with the corresponding proteins. The experimental validation of interactions between lncRNAs and proteins is costly and time-consuming. In this study, we developed a weighted graph-regularized matrix factorization (LPI-WGRMF) method to find unobserved lncRNA–protein interactions (LPIs) based on lncRNA similarity matrix, protein similarity matrix, and known LPIs. We compared our proposed LPI-WGRMF method with five classical LPI prediction methods, that is, LPBNI, LPI-IBNRA, LPIHN, RWR, and collaborative filtering (CF). The results demonstrate that the LPI-WGRMF method can produce high-accuracy performance, obtaining an AUC score of 0.9012 and AUPR of 0.7324. The case study showed that SFPQ, SNHG3, and PRPF31 may associate with Q9NUL5, Q9NUL5, and Q9UKV8 with the highest linking probabilities and need to further experimental validation.

Keywords: lncRNA–protein interaction, weighted graph-regularized matrix factorization, lncRNA similarity, protein similarity, SFPQ, SNHG3, PRPF31

INTRODUCTION

Long non-coding RNAs (lncRNAs) are closely associated with many key biological processes, for example, immune response, embryonic stem cell pluripotency, and cell cycle regulation (Chen et al., 2016; Agirre et al., 2019; Gil and Ulitsky, 2020). lncRNAs regulate cellular activities to achieve their biological function through interactions with proteins (Chen and Yan, 2013; Zhang et al., 2018b). Therefore, finding potential lncRNA–protein interactions (LPIs) is important to uncover lncRNA-related biological activities. Wet experiments found a few LPIs; however, experimental methods are costly and time-consuming. Thus, computational methods are developed to identify possible associations between lncRNAs and proteins (Bester et al., 2018; Chen et al., 2018).

LPI prediction methods can be roughly classified into two groups: network-based methods and machine learning-based methods. Network-based LPI identification methods integrated various biological data and network propagation methods (Peng et al., 2019). Li et al. (2015) used random walk with restart on the constructed lncRNA–protein heterogeneous network to find LPI candidates. Zhang et al. (2018a) developed a linear neighborhood propagation method to score for lncRNA–protein pairs. Ge et al. (2016), Zhao et al. (2018a), and Xie et al. (2019) applied bipartite network projection recommended methods to compute the association probabilities between lncRNAs and proteins.

Machine learning-based methods mainly contain matrix factorization-based LPI prediction methods and ensemble learning-based LPI prediction methods. Matrix factorization methods have been widely applied to various association prediction areas (Peng et al., 2020). Liu et al. (2017), Zhang T. et al. (2018), Zhao et al. (2018a), and Shen et al. (2019) used matrix factorization methods to predict possible LPIs. Hu et al. (2018) and Zhang et al. (2018b) utilized ensemble techniques and generated ensemble learning frameworks to discover potential LPIs based on the constructed benchmark datasets. Computational methods effectively revealed the possible associations between lncRNAs and proteins. However, the performance obtained by the above methods is limited and can be further improved.

In this study, we first integrated lncRNA similarity, protein similarity, known LPIs. We then developed a novel LPI prediction method based on weighted graph-regularized matrix factorization (LPI-WGRMF). LPI-WGRMF was compared with five state-of-the-art LPI methods [LPBNI, LPI-IBNRA, LPIHN, RWR, and collaborative filtering (CF)] to measure the performance of the proposed LPI-WGRMF method. LPI-WGRMF obtained the AUC value of 0.9057 and the AUPR value of 0.7324. The results showed that LPI-WGRMF is a useful tool for identifying LPIs. Case study analysis suggests that there are possibly joint links between SFPQ and Q9NUL5, SNHG3 and Q9NUL5, and PRPF31 and Q9UKV8.

MATERIALS AND METHODS

In this manuscript, we developed an LPI prediction model, LPI-WGRMF. The method can be summarized to three steps. First, experimentally validated LPIs from the NPInter 2.0 database were collected. Second, lncRNA similarity matrix and protein similarity matrix are computed based on the assumption that lncRNAs tend to associate with similar proteins and vice versa. Finally, lncRNA similarity, protein similarity, and LPI matrix were integrated to the weight graph-regularized matrix factorization model for computing the association scores for each lncRNA–protein pair.

Materials

LPI Data

We obtained experimentally validated LPI dataset, which was provided by Zhang et al. (2018a). The dataset contains 4158 LPIs between 990 lncRNAs and 27 proteins after preprocessing. The LPI matrix between n lncRNAs and m proteins was denoted as $Y_{n\times m}$.

lncRNA Similarity Matrix

The sequence and expression information of lncRNAs can be downloaded from the NONCODE database. We computed lncRNA similarity matrix by integrating the sequence similarity, expression similarity, and interaction similarity to the similarity kernel fusion technique.

Sequence statistical similarity

Each lncRNA was described a 20-dimensional vector based on the methods provided by Zhang et al. (2018b). Based on the assumption that each vector can be denoted by their k-nearest neighbors, linear neighborhood similarity between two lncRNAs l_i and l_j can be computed and denoted as $s_{l,0}\left(i,j\right)$.

Expression similarity

Suppose that the expression profile of the i^{th} lncRNA can be represented as e_i and thus the expression similarity between two lncRNAs l_i and l_j can be defined as:

$$s_{l,1}\left(i,j\right)=\begin{cases}\frac{1}{2}\left(1+\rho_{i,j}\right) & i\neq j\\ 0 & i=j\end{cases} \tag{1}$$

where $\rho_{i,j}$ is the Pearson's correlation coefficient between two expression profiles e_i and e_j and is defined as:

$$\rho_{i,j}=\frac{cov(e_i,e_j)}{\sigma(e_i)\sigma(e_j)} \tag{2}$$

where $cov()$ denotes the covariance and σ denotes the standard deviation.

Interaction profile similarity

Suppose that the interaction profile of the i^{th} lncRNA can be represented as the i^{th} row $Y_{i.}$ Of the LPI matrix Y, the interaction profile similarity between two lncRNAs l_i and l_j can be defined as:

$$s_{l,2}\left(i,j\right)=\exp\left(-\frac{1}{\gamma_l}||Y_{i.}-Y_{j.}||^2\right) \tag{3}$$

where

$$\gamma_l=\frac{1}{n}\sum_{i\ 1}^{n}||Y_{i.}||^2 \tag{4}$$

where $||\cdot||$ denotes the 2-norm of a matrix.

Protein Similarity Matrix

Sequence alignment similarity

The sequences of proteins were downloaded from the SUPERFAMILY database. The alignment score of the u^{th} protein against the v^{th} protein can be computed by Blast and be denoted as $b_{u,v}$. The sequence similarity between two proteins p_u and p_v can be defined as:

$$s_{p,0}\left(u,v\right)=\begin{cases}\frac{b_{u,v}}{b_{u,u}} & u\neq v\\ 0 & u=v\end{cases} \tag{5}$$

Sequence statistical feature similarity

Each protein can be represented as a 504-dimensional vector based on the method provided by Zhou et al. (2020). Linear neighborhood similarity between two proteins p_u and p_v can be computed and denoted as $s_{p,1}$.

Interaction profile similarity

Suppose that the interaction profile of the u^{th} protein can be represented as the u^{th} column $Y_{.u}$ of the LPI matrix Y, the interaction profile similarity between two proteins p_u and p_v can be defined as:

$$s_{p,2}(u,v) = \exp\left(-\frac{1}{\gamma_l}||Y_{.u} - Y_{.v}||^2\right) \tag{6}$$

where

$$\gamma_l = \frac{1}{n}\sum_{u=1}^{m}||Y_{.u}||^2 \tag{7}$$

Similarity Kernel Fusion

In the above sections, three lncRNA similarity measurements and three protein similarity measurements were proposed. The similarity kernel fusion method provided by Zhou et al. (2020) was applied to integrate this similarity information to compute a more comprehensive similarity.

First, the three lncRNA similarities were normalized as follows:

$$\theta_{l,q}(i,j) = \frac{s_{l,q}(i,j)}{\sum_{t=1}^{n} s_{l,q}(t,j)}, \ (q = 0, 1, 2) \tag{8}$$

The normalized similarity matrix was denoted as:

$$\Theta_{l,q} = \{\theta_{l,q}(i,j)\}_{n\times n} \tag{9}$$

Second, for an lncRNA l_i and $s_{l,q}$, the k most similar lncRNAs were collected as a set $N_{l,q}(i,k)$ and $s_{l,q}$ can be normalized in constraint based on the neighborhood information:

$$\varphi_{l,q}(i,j) = \frac{s_{l,q}(i,j)\, I_{l,q,k}(i,j)}{\sum_{t=1}^{n} s_{l,q}(i,t)\, I_{l,q,k}(i,t)} \tag{10}$$

where

$$I_{l,q,k}(i,j) = \begin{cases} 1 & l_j \in N_{l,q}(u,k) \\ 0 & l_j \notin N_{l,q}(u,k) \end{cases} \tag{11}$$

The neighborhood constrained normalized matrix was denoted as:

$$\phi_{l,q} = \{\varphi_{l,q}(i,j)\}_{n\times n} \tag{12}$$

The above three normalized matrices were integrated based on the following iterative process:

$$\Theta_{l,q}(\lambda+1) = \frac{1}{2}\alpha\left(\phi_{l,q}\sum_{r\neq q}\Theta_{l,r}(\lambda)\phi_{l,q}^{T}\right) + \frac{1}{2}(1-\alpha)\sum_{r\neq q}\Theta_{l,r}(0) \tag{13}$$

where α was a weight parameter with $0 \quad \alpha \quad 1$, T was the transpose of the matrix, λ represented the iterative parameter, and $\Theta_{l,r}(0) \quad \Theta_{l,r}$.

We computed the integrated similarity matrix after z rounds of iteration:

$$\Theta_l = \frac{1}{3}\left(\Theta_{l,0}(z) + \Theta_{l,1}(z) + \Theta_{l,2}(z)\right) \tag{14}$$

By considering data noise, we defined the following indicator function based on the k most similar lncRNAs for each lncRNA:

$$w_{l,k} = \begin{cases} 1 & I_{l,0,k}(i,j) = I_{l,1,k}(i,j) = I_{l,2,k}(i,j) = 1 \\ 0 & I_{l,0,k}(i,j) = I_{l,1,k}(i,j) = I_{l,2,k}(i,j) = 0 \\ 0.5 & otherwise \end{cases} \tag{15}$$

The final lncRNA similarity matrix can be denoted as follows:

$$S_{l,k} = \{\vartheta_l(i,j)\, w_{l,k}(i,j)\}_{n\times n} \tag{16}$$

where $\vartheta_l(i,j)$ is the $(i,j)^{th}$ element in the matrix Θ_l.

Nearest Neighbor Information

Based on the graph regularization theory, similar lncRNAs should tend to interact with similar proteins and vice versa in an LPI network, and thus we first observe the nearest neighbor information for lncRNAs and proteins. Given the lncRNA similarity matrix S^l, we represented a p-nearest neighbor graph N as

$$N_{ij} = \begin{cases} 1 & j \in N_p(i) \ \& \ i \in N_p(j) \\ 0 & j \notin N_p(i) \ \& \ i \notin N_p(j) \\ 0.5 & otherwise \end{cases} \tag{17}$$

where $N_p(i)$ denotes the set of p nearest neighbors of lncRNA l_i. N is applied to increase the sparsify of the lncRNA similarity matrix S^l as

$$\forall i,j \quad \hat{S}_{ij}^{l} = N_{ij}S_{ij}^{l} \tag{18}$$

Thus, the sparse similarity matrix of lncRNAs can be computed. Similarity, the sparse similarity matrix of protein can be done.

Low-Rank Approximation

Based on low-rank approximation idea, the LPI matrix $Y \in \mathbb{R}^{n=m}$ can be decomposed into two low-rank latent feature matrices $A \in \mathbb{R}^{n=k}$ (for lncRNAs) and $B \in \mathbb{R}^{m=k}$ (for proteins) by minimizing the following low-rank approximation objective:

$$\min_{A,B}||Y - AB^T||_F^2 \tag{19}$$

where $||\cdot||_F$ denotes the Fronbenius norm and k is the rank of matrices A and B, that is, the number of features in A and B.

We decomposed $Y \in \mathbb{R}^{n=m}$ into $U \in \mathbb{R}^{n=k}$, $S_k \in \mathbb{R}^{k=k}$, and V$\in \mathbb{R}^{m=k}$ so that US_kV^T is the closest k-rank approximation to Y where U and V are matrices with orthonormal columns, S_k is a diagonal matrix, and $k_{max} = \min(n,m)$. Thus, the feature matrices A and B can be represented as $A = US_k^{1/2}$ and $B = VS_k^{1/2}$.

Graph-Regularized Matrix Factorization

To boost generalization ability and prevent overfitting, we minimize the following GRMF's objective function by adding Tikhonov and graph regularization terms to the above low-rank approximation:

$$\min_{A,B} ||Y - AB^T||_F^2 + \lambda_f(||A||_F^2| + |B||_F^2) + \lambda_l \sum_{i,r=1}^{n} \hat{S}_{ij}^l ||a_i - a_r||^2 + \lambda_p \sum_{j,q=1}^{n} \hat{S}_{ij}^p ||b_j - b_q||^2 \quad (20)$$

where λ_f, λ_l, and λ_p are positive parameters, a_i and b_j are the i^{th} and j^{th} rows of A and B, respectively, and n and m are the numbers of lncRNAs and proteins, respectively. The first term is used to make the model approximate the matrix Y. The second term (Tikhonov regularization) minimizes the norms of A and B. The third and final terms are lncRNA graph regularization and protein graph regularization, respectively. The two terms are applied to minimize the distance between feature vectors of two neighboring lncRNAs or proteins. Based on graph regularization, the above model can be redescribed as

$$\min_{A,B} ||Y - AB^T||_F^2 + \lambda_f(||A||_F^2| + |B||_F^2) + \lambda_l Tr(A^T \mathcal{L}_l A) + \lambda_p Tr(B^T \mathcal{L}_p B) \quad (21)$$

where $Tr(\cdot)$ denotes the trace of matrix, $\mathcal{L}_l = D^l - \hat{S}^l$ and $\mathcal{L}_p = D^p - \hat{S}^p$ represent the graph Laplacian terms for $\hat{S}^l$ and $\hat{S}^p$, respectively, and D^l and D^p are diagonal matrices where $D_{ii}^l = \sum_r \hat{S}_{ir}^l$ and $D_{jj}^t = \sum_q \hat{S}_{jq}^p$.

To improve LPI prediction performance, we normalize graph Laplacians $\mathcal{L}_l$ and $\mathcal{L}_p$ by $\widetilde{\mathcal{L}}_l = (D^l)^{-1/2}\widetilde{\mathcal{L}}_l(D^l)^{-1/2}$ and $\mathcal{L}_p = D^p - \hat{S}^p$. Equation (4) can be rewritten as

$$\min_{A,B} ||Y - AB^T||_F^2 + \lambda_f(||A||_F^2| + |B||_F^2) + \lambda_l Tr(A^T \widetilde{\mathcal{L}}_l A) + \lambda_p Tr(B^T \widetilde{\mathcal{L}}_p B) \quad (22)$$

Weighted Graph-Regularized Matrix Factorization

To prevent unknown lncRNA–protein pairs from affecting the performance of singular value decomposition produced by Y, we add a weight matrix W into the objective function as follows:

$$\min_{A,B} ||W \odot (Y - AB^T)||_F^2 + \lambda_f(||A||_F^2| + |B||_F^2) + \lambda_l Tr(A^T \widetilde{\mathcal{L}}_l A) + \lambda_p Tr(B^T \widetilde{\mathcal{L}}_p B) \quad (23)$$

Based on the alternating least square method provided by Ezzat et al. (2016), we can solve the model (6). Let $\frac{\partial L}{\partial a_i} = 0$ and $\frac{\partial L}{\partial b_j} = 0$, run alternatingly the following two update rules until convergence:

$$\forall i = 1, 2, ...n,$$

$$a_i = \left(\sum_{j=1}^{m} W_{ij} Y_{ij} b_j - \lambda_l (\tilde{L}_l)_{i*} A\right)\left(\sum_{j=1}^{m} W_{ij} b_j^T b_j \lambda_f I_k\right)^{-1} \quad (24)$$

$$\forall j = 1, 2, ...m,$$

$$b_i = \left(\sum_{i=1}^{n} W_{ij} Y_{ij} a_i - \lambda_p (\tilde{L}_p)_{j*} B\right)\left(\sum_{j=1}^{n} W_{ij} a_i^T a_i \lambda_f I_k\right)^{-1} \quad (25)$$

where $(\tilde{L}_l)_{i*}$ and $(\tilde{L}_p)_{j*}$ are the i^{th} and j^{th} rows vectors of $\widetilde{\mathcal{L}}_l$ and $\widetilde{\mathcal{L}}_p$, respectively.

We can obtain A and B based on Eqs 7 and 8. Finally, the interaction probability between the i^{th} lncRNA and the j^{th} protein can be computed by

$$Y = AB^T \quad (26)$$

RESULTS

Experimental Settings

We conducted three different fivefold cross validation on the training dataset to set LPI-WGRMF's parameters, that is, k (the rank of matrices A and B), p (the number of nearest neighbors), λ_l, λ_d, *and* λ_t. We set the parameters as $k \in \{50, 100\}$, $p \in \{1, 2, 3, 4, 5, 6, 7\}$, $\lambda_f \in \{2^{-2}, 2^{-1}, 2^0, 2^1\}$, $\lambda_l \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, *and* $\lambda_p \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. And we used grid search and found that the best parameter combination is $k = 50$, $p = 7$, $\lambda_f = 0.5$, $\lambda_l = 0.3$, *and* $\lambda_p = 0.005$.

Evaluation Metrics

Precision, recall, f1 score, accuracy, AUC, and AUPR are widely applied to measure the performance of machine learning methods on association prediction. In this study, we used the six measurements to evaluate the performance of our proposed LPI-WGRMF. AUC is the area under the receiver operating characteristics curve. AUPR is the area under precision–recall curve. The other four criteria are defined as follows:

$$Precision = \frac{TP}{TP + FP} \quad (27)$$

$$Recall = \frac{TP}{TP + FN} \quad (28)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (29)$$

$$\text{f1 } score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (30)$$

where TP and FP denote the predicted true and false number of positive LPIs, respectively, and TN and FN denote the predicted true and false number of negative LPIs, respectively. The experiments were conducted 20 times. The average precision, recall, accuracy, AUC, and AUPR values for 20 times of experiments were computed as the final performance.

Performance Comparison of LPI-WGRMF and Other Methods

To measure the performance of our proposed LPI-WGRMF method, we compared LPI-WGRMF and five state-of-the-art methods, that is, LPBNI, LPI-IBNRA, LPIHN, RWR, and CF. LPBNI is a bipartite network inference method; LPIHN is a heterogeneous network inference method based on random walk with restart. The two models obtained better prediction performance in the area of LPI identification and are state-of-the-art LPI prediction methods. The experiments were conducted 20 times under fivefold cross validation. The results are shown in **Table 1**. The best performance in each column (measurement metric) is denoted in bold in **Table 1**.

Higher precision, recall, accuracy, and AUC denote better performance. From **Table 1**, we can find that LPI-WGRMF significantly outperformed other five methods in terms of precision, recall, and AUC. Precision computed by LPI-WGRMF was better 59.27, 45.32, 55.74, 61.17, and 67.44% than LPBNI, LPI-IBNRA, LPIHN, RWR, and CF, respectively. Recall computed by LPI-WGRMF was better 36.83, 34.83, 56.19, 44.91, and 53.86%, respectively. F1-score computed by LPI-WGRMF was better 36.83, 30.37, 56.19, 44.91, and 53.86%, respectively. AUC of LPI-WGRMF was higher 5.39, 3.74, 6.69, 10.19, and 15.14%, respectively. AUPR of LPI-WGRMF was higher 54.92, 40.59, 68.61, 61.40, and 67.82%, respectively.

Although accuracy computed by LPI-WGRMF was lower than LPBNI, LPI-WGRMF obtained better precision, recall, and AUC. More importantly, AUC and AUPR are more representative measurement metrics compared with other three evaluation metrics. Thus, AUC and AUPR can be more effectively applied to evaluate the performance of LPI prediction models. LPI-WGRMF is a powerful tool for LPI identification because of its better precision, recall, AUC, and AUPR. **Figures 1**, **2** demonstrate the AUC and AUPR values obtained by the six LPI prediction methods. The results show that LPI-WGRMF obtained the best AUC value, thereby demonstrating LPI-WGRMF's powerful LPI prediction capability.

Case Study

We further conducted four case studies after confirming the performance of LPI-WGRMF. The lncRNAs in the four cases are Splicing Factor Proline and Glutamine Rich (SFPQ), FOrkhead boX protein D2-Adjacent Opposite Strand RNA 1 (FOXD2-AS1), Small Nucleolar RNA Host Gene 3 (SNHG3), and Pre-mRNA-Processing Factor 31 (PRPF31), respectively. We predicted possible LPIs based on lncRNA similarities, protein similarities, known LPIs, and LPI-WGRMF. **Table 2** lists the predicted top five proteins associated with the above four lncRNAs.

SFPQ is a multifunctional nuclear protein participating in a few cellular activities including RNA transport, apoptosis, and DNA repair. SFPQ is densely associated with several diseases including renal cell carcinoma, Xp11-associated tumor, and dyslexia. More importantly, the expression levels of SFPQ impact on the sensitivity of ovarian cancer cells to PT-induced death (Gao et al., 2019; Pellarin et al., 2020). **Table 2** shows that SFPQ has joint connection with Q9NUL5 (ranked as 2). More importantly, the association between SFPQ and Q9NUL5 is ranked as 1 in all other five LPI identification methods. The fact suggests that SFPQ is possibly to link with Q9NUL5.

FOXD2-AS1 is an RNA gene and is abnormally expressed in a variety of malignant tumors. FOXD2-AS1 has close associations with many diseases, for example, nasopharyngeal carcinoma, esophageal cancer, bladder cancer, multiple pterygium syndrome, escobar variant, and ulcerative colitis (Bao et al., 2018; Chen et al., 2018; Su et al., 2018; Huang et al., 2020; Liu et al., 2020). FOXD2-AS1 was predicted to be closely linking with O00425, Q9NZI8, Q9Y6M1, and Q9NUL5, which was ranked as 1, 2, 3, and 4. All these connections were ranked in the top five associations among other five LPI prediction models. Therefore, FOXD2-AS1 is associated with O00425, Q9NZI8, Q9Y6M1, and Q9NUL5.

SNHG3 is a newly found lncRNA and was discovered as a biomarker of malignant cancers, for example, ovarian cancer, hepatocellular carcinoma, colorectal cancer, lung cancer, and

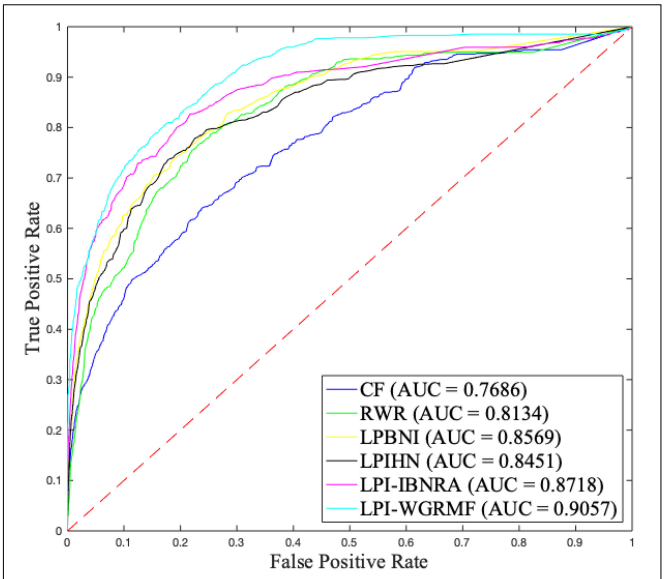

FIGURE 1 | The AUC values of six LPI prediction methods.

TABLE 1 | The performance of five LPI prediction methods.

Methods	Precision	Recall	Accuracy	F1-score	AUC	AUPR
LPBNI	0.3794	0.4037	0.9573	0.3876	0.8569	0.3302
LPI-IBNRA	0.5093	0.4165	**0.9641**	0.4521	0.8718	0.4351
LPIHN	0.4122	0.2800	0.9412	0.3324	0.8451	0.2299
RWR	0.3617	0.3521	0.9531	0.3543	0.8134	0.2827
CF	0.3033	0.2949	0.9488	0.2965	0.7686	0.2357
LPI-WGRMF	**0.9314**	**0.6391**	0.8906	**0.6493**	**0.9057**	**0.7324**

The best performance in each column (measurement metric) is denoted in bold.

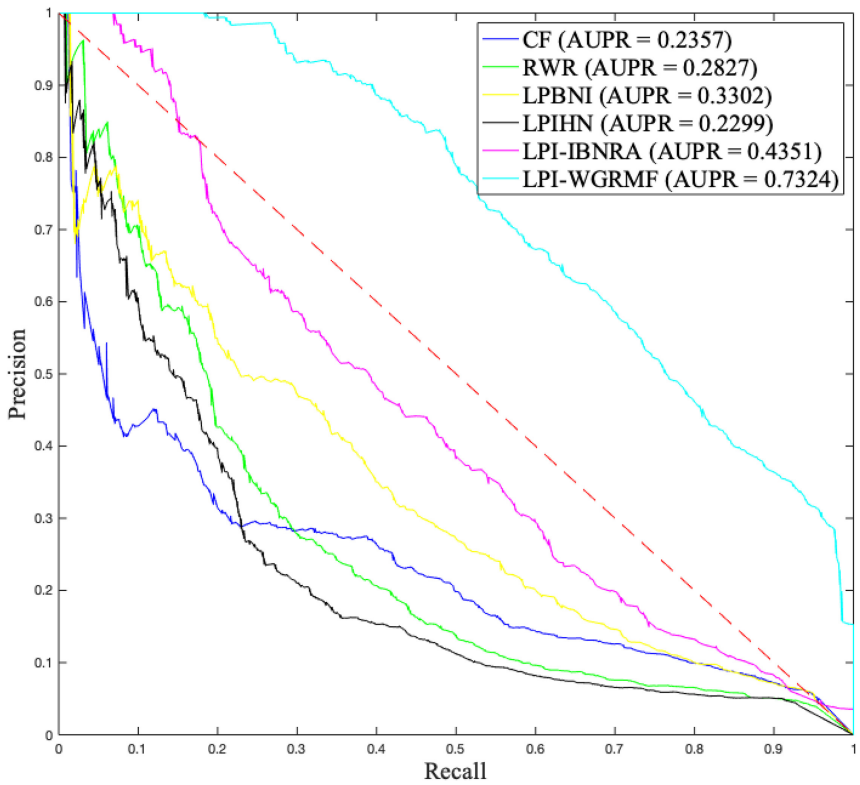

FIGURE 2 | The AUPR values of six LPI prediction methods.

TABLE 2 | The top five proteins associated with the four lncRNAs.

lncRNAs	Proteins	Confirmed	LPI-WGRMF	LPBNI	LPIIBNRA	LPIHN	RWR	CF
MTND2P28	Q9NUL5	NO	1	1	4	2	7	2
	O00425	YES	2	2	2	1	1	1
	P26599	YES	3	8	10	11	4	11
	Q07955	YES	4	16	17	18	5	15
	Q9Y6M1	YES	5	3	1	3	2	3
RPI001_1001892	Q9NUL5	YES	1	1	1	1	1	1
	Q07955	YES	2	9	13	15	8	13
	P35637	YES	3	5	5	5	4	5
	P26599	YES	4	15	17	16	9	16
	Q9NZI8	YES	5	4	4	3	5	3
RPI001_1002045	Q9NUL5	YES	1	1	1	1	1	1
	P35637	YES	2	4	2	5	4	5
	Q01844	YES	3	6	6	6	6	6
	P31483	YES	4	9	10	8	7	9
	Q9Y6M1	YES	5	3	4	3	3	3
RP11-169K16.7	Q9UKV8	YES	1	1	1	1	2	1
	Q9H9G7	YES	2	2	4	2	1	7
	Q9UL18	YES	3	7	3	4	4	10
	Q9HCK5	YES	4	6	2	3	3	9
	Q9NUL5	YES	5	5	5	6	5	2

glioma (Zhang et al., 2016; Huang et al., 2017; Lu et al., 2019; Liu and Tao, 2020). The results from case study analyses showed that SNHG3 tends to link with Q9NUL5 (ranked as 1) and has highest association scores with the protein in LPNI, BPIHN, and CF. Thus, SNHG3 may be possibly linked with Q9NUL5.

PRPF31 is one retinitis pigmentosa-causing gene. Its genetic variants have joint connections with variation in response to metformin in patients with type 2 diabetes (Kiser et al., 2019). In our predicted results, PRPF31 was found to be densely associated with Q9UKV8 (ranked as 1). More importantly, the association between PRPF31 and Q9UKV8 was identified to be ranked as 1, 1, 2, and 1 in LPBNI, LPIHN, RWR, and CF, respectively. PRPF31 obtained the highest association score with Q9UKV8 in five models.

DISCUSSION AND CONCLUSION

In this manuscript, we developed a novel method LPI-WGRMF for identifying possible LPIs, based on lncRNA similarity, protein similarity, known LPIs, and weighted graph regularization-based matrix factorization. We first integrated the similarity information and known LPIs as the initial resource. We then proposed a weighted graph-regularized matrix factorization model to compute the association scores for lncRNA–protein pairs.

LPI-WGRMF was compared with five classical LPI methods, that is, LPBNI, LPI-IBNRA, LPIHN, RWR, and CF. Cross-validation experiments were conducted for 20 times. The results showed the powerful performance of LPI-WGRMF. We conducted four case study analyses after confirming the LPI-WGRMF's accuracy. The results suggest that there are possibly close associations between SFPQ and Q9NUL5, SNHG3 and Q9NUL5, and PRPF31 and Q9UKV8 and need to further experimental validation.

In the future, other sources of LPI-related data may be used to improve the prediction performance, for example, using multiple kernels and designing a multiple kernel learning-based algorithm to effectively integrate the abundant lncRNA and protein information.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

FL and JY conceived, designed, and managed the study. XS and LC designed the LPI-WGRMF method, ran LPI-WGRMF, and wrote the original manuscript. JL and CX revised the original draft. XS, JL, and CX discussed the proposed method and gave further research. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We would like to thank all authors of the cited references.

REFERENCES

Agirre, X., Meydan, C., Jiang, Y., Garate, L., Doane, A. S., Li, Z., et al. (2019). Long non-coding RNAs discriminate the stages and gene regulatory states of human humoral immune response. *Nat. Commun.* 10:821.

Bester, A. C., Lee, J. D., Chavez, A., Lee, Y.-R., Nachmani, D., Vora, S., et al. (2018). An integrated genome-wide crispra approach to functionalize lncrnas in drug resistance. *Cell* 173, 649–664. doi: 10.1016/j.cell.2018.03.052

Bao, J., Zhou, C., Zhang, J., Mo, J., Ye, Q., He, J., et al. (2018). Upregulation of the long noncoding RNA FOXD2-AS1 predicts poor prognosis in esophageal squamous cell carcinoma. *Cancer Biomark.* 21, 527–533. doi: 10.3233/CBM-170260

Chen, X., Sun, Y.-Z., Guan, N.-N., Qu, J., Huang, Z.-A., Zhu, Z.-X., et al. (2018). Computational models for lncrna function prediction and functional similarity calculation. *Brief. Funct. Genom.* 18, 58–82. doi: 10.1093/bfgp/ely031

Chen, X., Yan, C. C., Zhang, X., and You, Z.-H. (2016). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 18, 558–576. doi: 10.1093/bib/bbw060

Chen, X., and Yan, G. Y. (2013). Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

Ezzat, A., Zhao, P., Wu, M., Li, X. L., and Kwoh, C. K. (2016). Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 646–656. doi: 10.1109/TCBB.2016.2530062

Gao, Z., Chen, M., Tian, X., Chen, L., Chen, L., Zheng, X., et al. (2019). A novel human lncRNA SANT1 cis-regulates the expression of SLC47A2 by altering SFPQ/E2F1/HDAC1 binding to the promoter region in renal cell carcinoma. *RNA Biol.* 16, 940–949. doi: 10.1080/15476286.2019.1602436

Ge, M., Li, A., and Wang, M. (2016). A bipartite network-based method for prediction of long non-coding rna-protein interactions. *Genomics Proteomics Bioinform.* 14, 62–71. doi: 10.1016/j.gpb.2016.01.004

Gil, N., and Ulitsky, I. (2020). Regulation of gene expression by cis-acting long non-coding RNAs. *Nat. Rev. Genet.* 21, 102–117. doi: 10.1038/s41576-019-0184-5

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). Hlpi- ensemble: prediction of human lncrna-protein interactions based on ensemble strategy. *RNA Biol.* 15, 797–806. doi: 10.1080/15476286.2018.1457935

Huang, W., Tian, Y., Dong, S., Cha, Y., Li, J., Guo, X., et al. (2017). The long non-coding RNA SNHG3 functions as a competing endogenous RNA to promote malignant development of colorectal cancer. *Oncol. Rep.* 38, 1402–1410. doi: 10.3892/or.2017.5837

Huang, Y., Yuan, K., Tang, M., Yue, J. M., Bao, L. J., Wu, S., et al. (2020). Melatonin inhibiting the survival of human gastric cancer cells under ER stress involving autophagy and Ras-Raf-MAPK signalling. *J. Cell. Mol. Med.* 2020, 1480–1492. doi: 10.1111/jcmm.16237

Kiser, K., Webb-Jones, K. D., Bowne, S. J., Sullivan, L. S., Daiger, S. P., and Birch, D. G. (2019). Time course of disease progression of PRPF31-mediated retinitis pigmentosa. *Am. J. Ophthalmol.* 200, 76–84.

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. doi: 10.1155/2015/671950

Liu, H., Ren, G., Chen, H., Liu, Q., Yang, Y., Zhao, Q., et al. (2020). Predicting lncRNA–miRNA interactions based on logistic matrix factorization with neighborhood regularized. *Knowl. Based Syst.* 191:105261. doi: 10.1016/j.knosys.2019.105261

Liu, H., Ren, G., Hu, H., Zhang, L., Ai, H., Zhang, W., et al. (2017). Lpi-nrlmf: lncrna-protein interaction prediction by neighborhood regularized logistic matrix factorization. *Oncotarget* 8:103975. doi: 10.18632/oncotarget.21934

Liu, Z., and Tao, H. (2020). Small nucleolar RNA host gene 3 facilitates cell proliferation and migration in oral squamous cell carcinoma via targeting nuclear transcription factor Y subunit gamma. *J. Cell. Biochem.* 121, 2150–2158.

Lu, W., Yu, J., Shi, F., Zhang, J., Huang, R., Yin, S., et al. (2019). The long non-coding RNA Snhg3 is essential for mouse embryonic stem cell self-renewal and pluripotency. *Stem Cell Res. Ther.* 10:157. doi: 10.1002/jcb.29421

Pellarin, I., Dall'Acqua, A., Gambelli, A., Pellizzari, I., D'Andrea, S., Sonego, M., et al. (2020). Splicing factor proline-and glutamine-rich (SFPQ) protein regulates platinum response in ovarian cancer-modulating SRSF2 activity. *Oncogene* 39, 4390–4403. doi: 10.1038/s41388-020-1292-6

Peng, L., Liu, F., Yang, J., Liu, X., Meng, Y., Deng, X., et al. (2019). Probing lncRNA-protein interactions: data repositories, models, and algorithms. *Front. Genet.* 10:1346. doi: 10.3389/fgene.2019.01346

Peng, L., Shen, L., Liao, L., Liu, G., and Zhou, L. (2020). RNMFMDA: a microbe-disease association identification method based on reliable negative sample selection and logistic matrix factorization with neighborhood regularization. *Front. Microbiol.* 11:592430. doi: 10.3389/fmicb.2020.592430

Su, F., He, W., Chen, C., Liu, M., Liu, H., Xue, F., et al. (2018). The long non-coding RNA *FOXD2-AS1* promotes bladder cancer progression and recurrence through a positive feedback loop with Akt and E2F1. *Cell Death Dis.* 9, 1–17. doi: 10.1038/s41419-018-0275-9

Shen, C., Ding, Y., Tang, J., Jiang, L., and Guo, F. (2019). Lpi-ktaslp: prediction of lncrna-protein interaction by semi-supervised link learning with multivariate information. *IEEE Access* 7, 13486–13496. doi: 10.1109/ACCESS.2019.2894225

Xie, G., Wu, C., Sun, Y., Fan, Z., and Liu, J. (2019). Lpi-ibnra: Long non-coding rna- protein interaction prediction based on improved bipartite network recommender algorithm. *Front. Genet.* 10:343. doi: 10.3389/fgene.2019.00343

Zhang, T., Cao, C., Wu, D., and Liu, L. (2016). *SNHG3* correlates with malignant status and poor prognosis in hepatocellular carcinoma. *Tumor Biol.* 37, 2379–2385. doi: 10.1007/s13277-015-4052-4

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

Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018a). The linear neighborhood propagation method for predicting long non-coding rna-protein interactions. *Neurocomputing* 273, 526–534. doi: 10.1016/j.jpdc.2017.08.009

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018b). Sfpel-lpi: Sequence-based feature projection ensemble learning for predicting lncrna-protein interactions. *PLoS Comput. Biol.* 14:e1006616. doi: 10.1371/journal.pcbi.1006616

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

Zhao, Q., Zhang, Y., Hu, H., Ren, G., Zhang, W., and Liu, H. (2018b). Irwnrlpi: integrating random walk and neighborhood regularized logistic matrix factorization for lncrna-protein interaction prediction. *Front. Genet.* 9:239. doi: 10.3389/fgene.2018.00239

Zhou, Y. K., Hu, J., Shen, Z. A., Zhang, W. Y., and Du, P. F. (2020). LPI-SKF: predicting lncRNA-protein interactions using similarity kernel fusions. *Front. Genet.* 11:615144. doi: 10.3389/fgene.2020.615144

Conflict of Interest: JL and CX were employed by the company Geneis Beijing Co., Ltd.

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