frontiers in Genetics

METHODS
published: 20 November 2019
doi: 10.3389/fgene.2019.01148

Projection-Based Neighborhood Non-Negative Matrix Factorization for lncRNA-Protein Interaction Prediction

Yingjun Ma[1,2], *Tingting He*[2,3] *and Xingpeng Jiang*[2,3]*

[1] School of Mathematics & Statistics, Central China Normal University, Wuhan, China, [2] Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, Wuhan, China, [3] School of Computer, Central China Normal University, Wuhan, China

OPEN ACCESS

Edited by:
Wen Zhang,
Huazhong Agricultural University, China

Reviewed by:
Yang Yang,
Shanghai Jiao Tong University, China
Lingling Jin,
Thompson Rivers University, Canada
Le Ou-Yang,
Shenzhen University, China

***Correspondence:**
Xingpeng Jiang
xpjiang@mail.ccnu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *08 August 2019*
Accepted: *21 October 2019*
Published: *20 November 2019*

Citation:
Ma Y, He T and Jiang X (2019) Projection-Based Neighborhood Non-Negative Matrix Factorization for lncRNA-Protein Interaction Prediction. Front. Genet. 10:1148. doi: 10.3389/fgene.2019.01148

Many long ncRNAs (lncRNA) make their effort by interacting with the corresponding RNA-binding proteins, and identifying the interactions between lncRNAs and proteins is important to understand the functions of lncRNA. Compared with the time-consuming and laborious experimental methods, more and more computational models are proposed to predict lncRNA-protein interactions. However, few models can effectively utilize the biological network topology of lncRNA (protein) and combine its sequence structure features, and most models cannot effectively predict new proteins (lncRNA) that do not interact with any lncRNA (proteins). In this study, we proposed a projection-based neighborhood non-negative matrix decomposition model (PMKDN) to predict potential lncRNA-protein interactions by integrating multiple biological features of lncRNAs (proteins). First, according to lncRNA (protein) sequences and lncRNA expression profile data, we extracted multiple features of lncRNA (protein). Second, based on protein GO ontology annotation, lncRNA sequences, lncRNA(protein) feature information, and modified lncRNA-protein interaction network, we calculated multiple similarities of lncRNA (protein), and fused them to obtain a more accurate lncRNA(protein) similarity network. Finally, combining the similarity and various feature information of lncRNA (protein), as well as the modified interaction network, we proposed a projection-based neighborhood non-negative matrix decomposition algorithm to predict the potential lncRNA-protein interactions. On two benchmark datasets, PMKDN showed better performance than other state-of-the-art methods for the prediction of new lncRNA-protein interactions, new lncRNAs, and new proteins. Case study further indicates that PMKDN can be used as an effective tool for lncRNA-protein interaction prediction.

Keywords: lncRNA-protein interaction, feature projection, neighborhood completion, graph non-negative matrix factorization, kernel neighborhood similarity

INTRODUCTION

RNA represents the direct output of genomic encoded genetic information, and a large part of the regulatory capacity of cells focuses on the synthesis, processing, transportation, modification, and translation of RNA. With the continuous improvement of RNA analysis, cell type isolation, and culture technology, people's understanding of many biological functions of RNA is also getting

higher and higher (DjebaliDavis and Merkel et al., 2012). Studies have shown that up to 85% of human genes are transcribed, but the proportion of RNA transcriptional codons encoding proteins is extremely low, suggesting that most RNA transcripts are non-coding (Fang and Fullwood, 2016). A large part of human genes plays their functions through non-coding RNA (ncRNA) (Mattick, 2005). Transcriptional ncRNA has similar chromosome modification functions to protein-coding genes. In multiple sites of human genome, the deletion of ncRNA will lead to the decline of the specificity of adjacent protein-coding genes (Ulf Andersson ørom et al., 2010). Long non-coding RNA (lncRNA) is an important type of ncRNA, which has more than 200 nucleotide transcripts and no obvious protein coding function (Volders et al., 2013). With the development of biological information, people are becoming more and more aware of the important role of lncRNA in various biological processes; lncRNA is involved in the regulation of gene expression and function of multiple networks, affects the formation of the kernel structure domain and whole chromosome state of transcription, and participates in the interaction of two different chromosomal regions through direct mechanisms regulating the chromosome structure (Batista and Chang, 2013). In addition, a growing number of studies have shown that mutations and disorders of lncRNA are associated with different human diseases. The primary structure, secondary structure, expression level of lncRNA, and changes in its homologous binding protein can lead to a variety of diseases ranging from neuropathy to cancer (Wapinski and Chang, 2011). Currently, more and more lncRNA have been discovered, but their functions and mechanisms are still poorly understood. In general, almost all lncRNA functions are expressed through the interaction with the corresponding RNA-binding proteins, and their functions and mechanisms depend on their interaction with various protein complexes in cells (Khalil and Rinn, 2011). Therefore, it is important to determine the potential interactions between lncRNAs and proteins to study the functions of lncRNA. It is expensive and time-consuming to detect large-scale lncRNA-protein interactions by experimental means, so a large number of computational models are proposed based on existing experimental data (Suresh et al., 2015).

Based on the physicochemical properties of peptide chains and nucleotide chains, Bellucci et al. (2011) proposed catRAPID in 2011, which combined secondary structure, hydrogen bonding, and van der Waals to predict the interactions between lncRNAs and proteins. Subsequently, Lu et al. (2013) proposed the lncPro model, which used the secondary structure, hydrogen bonds, van der Waals, and other features to encode nucleotide and amino acid sequences into feature vector, and calculated the interaction scores between lncRNAs and proteins by Fisher's linear discriminant method. Suresh et al. (2015) proposed the RPI-Pred to predict the interactions between lncRNAs and proteins, which combined the secondary structural feature of RNA sequences with the three-dimensional structural feature of proteins and used support vector machine (SVM) model for prediction. Xiao et al. (2017) proposed a PLPIHS model, which constructed a heterogeneous model by using lncRNA-lncRNA similarity network, lncRNA-protein interaction network, and protein-protein interaction network, and then established a SVM classifier to predict lncRNA-protein interaction by HeteSim score. Subsequently, Deng et al. (2018) improved on PLPIHS and proposed a PLIPCOM model, which simultaneously obtained the low-dimensional features of lncRNA (protein) by restarted random walk and singular value decomposition on heterogeneous networks, and then used the gradient asymptotic tree algorithm to predict by combining the HeteSim score and low-dimensional features. Both algorithms achieved high AUC values, but they used the known lncRNA-protein interaction information to construct heterogeneous network, which also led to the reuse of the known interactions. Recently, Hu et al. (2018) proposed an ensemble strategy to predict potential lncRNA-protein interactions (HLPI-Ensemble), which used the strategy of random pairing to generate negative samples of lncRNA-protein interactions, and integrated support vector machine (SVM), random forest (RF), and extreme gradient enhancement (XGB) three mainstream machine learning algorithms to predict interaction scores. This ensemble learning strategy can not only improve the prediction performance of the model, but can also prevent the over-fitting of the model to some extent. Pan and Shen. (2017) used hybrid convolutional neural network and deep belief network to predict RNA-protein binding sites on RNAs, which used multimodal deep learning to fuse shared features of different sources of data, and found the explainable binding motifs. The above supervised learning method has achieved certain effects in predicting lncRNA-protein interactions, but there are still some problems. First, the key to supervised learning is to construct as balanced as possible positive and negative samples, but at present, most databases only provide lncRNA-protein interaction information, while the construction of negative samples is still a problem. Second, lncRNA-protein interaction prediction problem is a serious unbalanced classification problem, and the known interaction accounts for less than 1% of the total lncRNA-protein pairs, while many supervisory models often choose the same number of positive and negative samples as training set and test set, which artificially reduces the prediction range of the model to some extent. Finally, both lncRNA and protein exist in a whole biological network, and the rational use of lncRNA (protein) network topology can greatly improve the predictive performance of the model.

Recently, many network-based models have been proposed for predicting lncRNA-protein interactions. Li et al. (2015) proposed a heterogeneous network model to predict lncRNA-protein interactions, which constructed a lncRNA similarity network using lncRNA expression profiles and protein similarity network using weighted protein-protein interactions (PPIs), then combined with known lncRNA-protein interaction network uses the restart random walk model to make predictions. Ge et al. (2016) proposed a binary network inference algorithm (LPBNI) using only the known lncRNA-protein interactions to infer potential lncRNA-associated proteins. Zheng et al. (2017) predicted potential lncRNA-protein interactions by fusing multiple network information. Specifically, based on protein sequence, protein domain, protein GO term and STRING dataset, the method constructed four protein similarity networks, respectively, and integrated with similarity network fusion algorithm (SNF), and then used random walk algorithm to

calculate the score. Recently, Zhang et al. (2018a) proposed a linear neighborhood propagation algorithm (LPLNP) to predict the potential lncRNA-protein interactions. Specifically, based on various feature extracted, LPLNP calculated the linear neighborhood similarity of the corresponding lncRNA (protein), and used the label propagation algorithm to calculate the interaction scores, and finally the linear combination of all prediction scores as the final result. Subsequently, Zhang et al. (2018b) proposed a sequence-based feature projection ensemble learning algorithm (SFPEL-LPI). Specifically, based on lncRNA sequences, protein sequences, and known lncRNA-protein interactions, SFPEL-LPI extracted a variety of lncRNA (protein) features and similarity information, and uses feature projection ensemble learning framework to predict lncRNA-protein interaction scores. Compared to LPLNP, SFPEL-LPI has fewer parameters and higher precision and can predict new lncRNAs and new proteins. Most network-based models build similarity networks by mining lncRNA (protein) related information and use their network topological structure and known lncRNA-protein interaction information for prediction and have the advantage of not requiring negative sample construction. In addition, this type of method is also global; based on the prediction results, we can get the prediction ranking of all unknown interaction pairs, which is more convenient for us to study the higher-ranking unknown interaction. However, in addition to SFPEL-LPI, other network-based methods only focus on the construction of similarity networks and ignore important feature information. Although SFPEL-LPI makes use of both feature information and similarity information, it separates the lncRNA network and protein network for prediction, which also limits the improvement of model performance.

Based on this, this study proposes a projection-based neighborhood non-negative matrix factorization (PMDKN) to predict potential lncRNA-protein interactions in heterogeneous omics data, which is also applicable to the prediction of new lncRNAs and new proteins. First, based on the lncRNA sequences, lncRNA expression profile, and protein sequences, we extracted a variety of features of lncRNA and protein. Second, based on multiple features of lncRNA and protein, lncRNA sequences, gene ontology annotation of the protein and the modified lncRNA-protein interaction network, we calculated multiple similarities of lncRNA and protein and fused to obtain more accurate lncRNA (protein) similarity network. Finally, PMDKN uses these features and fused similarity network to predict lncRNA-protein interaction scores. The results indicate that PMDKN exhibits higher predictive performance than other state-of-the-art methods for the prediction of lncRNA-protein interactions, new lncRNAs, and new proteins. Case study further demonstrates that PMDKN can be an effective tool for lncRNA-protein interaction.

MATERIALS AND METHODS

Dataset

The noncoding RNAs and protein related biomacromolecules interaction database (Npinter) (Wu et al., 2006) provides a large number of experimentally verified interactions between non-coding RNA and other biomolecules. So far, Npinter has been updated to version 3.0, which includes more lnRNA-protein interactions than the previous version (Hao et al., 2016). In order to evaluate the predictive performance of the algorithm, we performed cross-experiment using the interactive data provided in Npinter v2.0 (Yuan et al., 2013) as the benchmark dataset and used Npinter v3.0 to test the final prediction ability of the model. Li et al. (2015) extracted interactions from Npinter v2.0 by limiting the organization to 'Homo sapiens' and ncRNA to 'NONCODE' and processed 4,870 interactions between 1,113 lncRNAs and 96 proteins. On this basis, Zhang et al. (2018a) deleted lncRNAs and proteins with no sequence information and only one interaction, resulting in 4,158 interactions between 990 lncRNAs and 27 proteins. Meanwhile, various features and similarity information were extracted based on the sequence data of lncRNAs and proteins. In order to facilitate the experimental comparison, we used the dataset provided by Zhang et al. (2018a) as the benchmark DATASET 1 for verification.

In benchmark DATASET 1, all lncRNAs (proteins) interact with at least two proteins (lncRNAs), and the number of lncRNA-protein interactions is relatively dense. To investigate the predictive ability of the algorithm for sparse interactions, lncRNAs without sequence information were deleted from the data provided by Li et al., and a total of 4,679 interactions between 1,068 lncRNAs and 90 proteins were finally obtained. Meanwhile, sequence information of lncRNA and expression profile information of lncRNA in 24 human tissues and cells were extracted from the integrated knowledge database of non-coding RNAs database (NONCODE) (Liu, 2004; Xie et al., 2013; Fang et al., 2018), and sequence information of protein and gene ontology annotation of protein were extracted from the protein-protein interaction networks dataset (STRING 9.1) (Franceschini et al., 2012). Based on the relevant information of lncRNA and proteins, multiple features and similarities of lncRNA (proteins) were calculated to construct benchmark DATASET 2.

Features for lncRNAs and Proteins

Let $\mathcal{L} = \{l_1, l_2, \cdots, l_{N_l}\}$ and $\mathcal{P} = \{p_1, p_2, \cdots, p_{N_p}\}$ represent the set of N_l lncRNAs and N_p proteins obtained, respectively. In this section, we introduce the three features of lncRNA, the two features of the protein, and the similarity of lncRNA and the similarity of protein.

Features of lncRNA

We extracted three features of lncRNA, namely expression profile feature and two sequence-based features: pseudo-k-tuple nucleotide composition (PseKNC) (Chen et al., 2014) and parallel related pseudo dinucleotide composition (PCPseDNC) (Guo et al., 2014). For lncRNA, k-mer (nucleotide sequence of length k) is generally used to describe the short-term ordered information of the sequences, while the overall or long-term information of the sequences is described by the physicochemical properties of nucleotides. PseKNC and PCPseDNC describe the lncRNA by integrating the short-term and long-term features of the sequences (Chen et al., 2014). We calculated the PseKNC and PCPseDNC of lncRNA using python "repDNA" package (Liu et al., 2015).

Features of Protein

The hydrophilicity and hydrophobicity of proteins play an important role in protein folding, environmental and molecular interactions, and catalytic effects. Combining the frequency of regularization of 20 amino acids in the protein sequence and the distribution pattern of hydrophilicity and hydrophobicity along the protein chain, we calculated the characteristics of the two proteins, which are the amphiphilic pseudo amino acid composition (APseAAC) (Chou, 2001; Chou, 2005) and the combined triad descriptor (CTriad). Among them, Ctriad was proposed by Shen et al. (2007) to predict protein-protein interactions. First, in order to reduce the size of the feature space, 20 amino acids were grouped into 7 classes according to the dipole and volume of the side chains. Second, using the classes of amino acids to distinguish any conjoint triad (combination of any three consecutive amino acids) and counting the frequency. $f(v_i)$ i=1,2,.⋯,7^3 of the occurrence of the conjoint triad in the amino acid sequence, where v_i represents the i-th conjoint triad. Finally, normalizing $f(v_i)$, we could get the conjoint triplet descriptor feature CTriad(P)=[q_1,q_2,⋯,q_{343}] of protein P as follows:

$$q_i = \frac{f(v_i) - \min\left\{f(v_i)\right\}_{i=1}^{343}}{\max\left\{f(v_i)\right\}_{i=1}^{343} - \min\left\{f(v_i)\right\}_{i=1}^{343}}$$

Where, $\min\left\{f(v_i)\right\}_{i=1}^{343}$ and $\max\left\{f(v_i)\right\}_{i=1}^{343}$ represent the minimum and maximum frequencies of all conjoint triads, respectively. It should be noted that in order to prevent the over-fitting problem caused by the lncRNA (protein) feature due to the high dimension, we use the PCA for dimensionality reduction on the high-dimensional features.

Similarities for LncRNAs and Proteins

In this section, we introduce the lncRNA-lncRNA similarity and the protein-protein similarity.

lncRNA-lncRNA Sequence Similarity

Kirk et al. (2018) found that lncRNAs with related functions, although lacking linear homology, often have a similar k-tuple spectrum, which is related to lncRNA binding protein and its subcellular localization. Song et al. (2014) introduced a variety of alignment-free genome and metagenome comparison methods based on word frequency and proved that d_2^* has a stronger statistical ability to measure sequence correlation. Therefore, d_2^* was used in this study to calculate the sequence similarity between lncRNAs. For any two lncRNA sequences L_1 and L_2 with m and n nucleotides, respectively, the dissimilarity $d_2^*(L_1, L_2)$ is as follows:

$$d_2^*(L_1, L_2) = \frac{1}{2}\left(1 - \frac{D_2^*(L_1, L_2)}{\sqrt{\sum_{w\in A^k} X_w^2/(\bar{m}p_w^X)}\sqrt{\sum_{w\in A^k} Y_w^2/(\bar{n}p_w^Y)}}\right)$$

Where $D_2^*(L_1, L_2)$ represents the D_2^* statistic of L_1 and L_2, and p_w^X and p_w^Y respectively represent the probability of k-tuple w occurring in L_1 and L_2 of lncRNA under the background model. $\bar{m} = m - k$, $\bar{n} = n - k$, $X_w = X_w - \bar{m}p_w^X$, $Y_w = Y_w - \bar{n}p_w^Y$, where X_w and Y_w represent the frequencies at which the k-tuples in the sequences L_1 and L_2 occur, respectively. Further, the similarity of L_1 and L_2 is $(1 - d_2^*(L_1, L_2))$. We used the program provided by Ahlgren et al. (2016) to calculate the d_2^* similarity of lncRNA.

Protein-Protein Semantic Similarity

The semantic comparison of gene ontology annotations provides a quantitative method for calculating the semantic similarity of gene products (Yu et al., 2010). There are currently two classic methods for computing the semantic similarity of GO annotation items: information-based methods (Jiang and Conrath, 1997; Lin, 1998; Resnik, 1999) and graph-based (Wang et al., 2007) methods, respectively. In this study, the graph-based method was first used to calculate the semantic similarity of GO items, and then the semantic similarity of proteins was calculated according to the association between protein and GO items. Specifically, any GO item A could be expressed as DAG(A)=(A,T_A,E_A), where T_A represents the set containing item A and all its ancestor items in the GO diagram, and E_A represents the set connecting all edges of GO item in DAG(A). Then, for any two GO annotation items A and B, their semantic similarity could be defined as:

$$S(A, B) = \frac{\sum_{t\in T_A \cap T_B} S_A(t) + S_B(t)}{SV(A) + SV(B)}$$

Where, $S_A(t)$ and $S_B(t)$ represent the S-value of GO item t related to item A and item B respectively, and $SV(A) = \sum_{t\in T_A} S_A(t)$ represents the semantic value of GO item A. At this point, according to the correlation between protein and GO term, we can get the semantic similarity of protein. We use R package "protr" to obtain semantic similarity of proteins; more details are shown in (Xiao et al., 2015).

Kernel Neighborhood Similarity

In Section "Features for lncRNAs and Proteins", we obtained three features of lncRNA and two features of protein, and the known lncRNA-protein interaction network also contains important lncRNA (protein) feature information. Based on these feature vectors, there are many methods for calculating similarities, such as Gaussian, linear neighborhood similarity (Zhang et al., 2018a) (LNS), and so on. Here, we adopt kernel neighborhood similarity (KSNS) (Ma et al., 2018a; Ma et al., 2018b), which not only considers the neighbor and non-neighbor similarity of samples hierarchically, but also explores nonlinear relations, which was well applied to a variety of biological problems. It should be noted that the currently known lncRNA-protein interaction matrix is incomplete. Therefore, in order to reduce the error caused by information loss, we first use the Weighted K nearest neighbor profiles (WKNNP) (Xiao

et al., 2018) to complete the known interaction matrix, and then calculated the KSNS of lncRNA(protein) interaction profile.

Based on the above steps, we obtained a total of 5 similarities of lncRNAs and 4 similarities of proteins, which reflected the similarity relationship of lncRNAs (proteins) from different perspectives. Due to the limitations of data and the selection of computational methods, these similarity networks may contain noise. Hence, we adopted a clusDCA proposed by Wang et al. (2015) for similarity network fusion, which can not only eliminate network noise and effectively capture network topology, but also have high computational efficiency in large-scale networks. The general procedure for predicting lncRNA-protein interaction using PMDKN is shown in **Figure 1**.

Prediction of lncRNA-Protein Interaction

Based on various features of lncRNA (protein) and the integrated lncRNA (protein) similarity network, we proposed projection-based neighborhood non-negative matrix factorization (PMDKN) to predict potential lncRNA-protein interactions. $\{FL_i\}_{i=1}^{N_1}$ represents the N_1 feature matrices of lncRNA, $\{FP_i\}_{i=1}^{N_2}$ represents the N_2 feature matrices of protein, similarity matrix of lncRNA and protein are SL and SP respectively, A represents known lncRNA-protein interaction matrix, and $\overline{A}$ represents lncRNA-protein interaction matrix completed by WKNNP.

First, we mapped lncRNA and protein to the common non-negative space R^d, that is, any lncRNA l_i and protein p_j can be represented by non-negative latent vectors $u_i \in R^{d\times 1}$ and

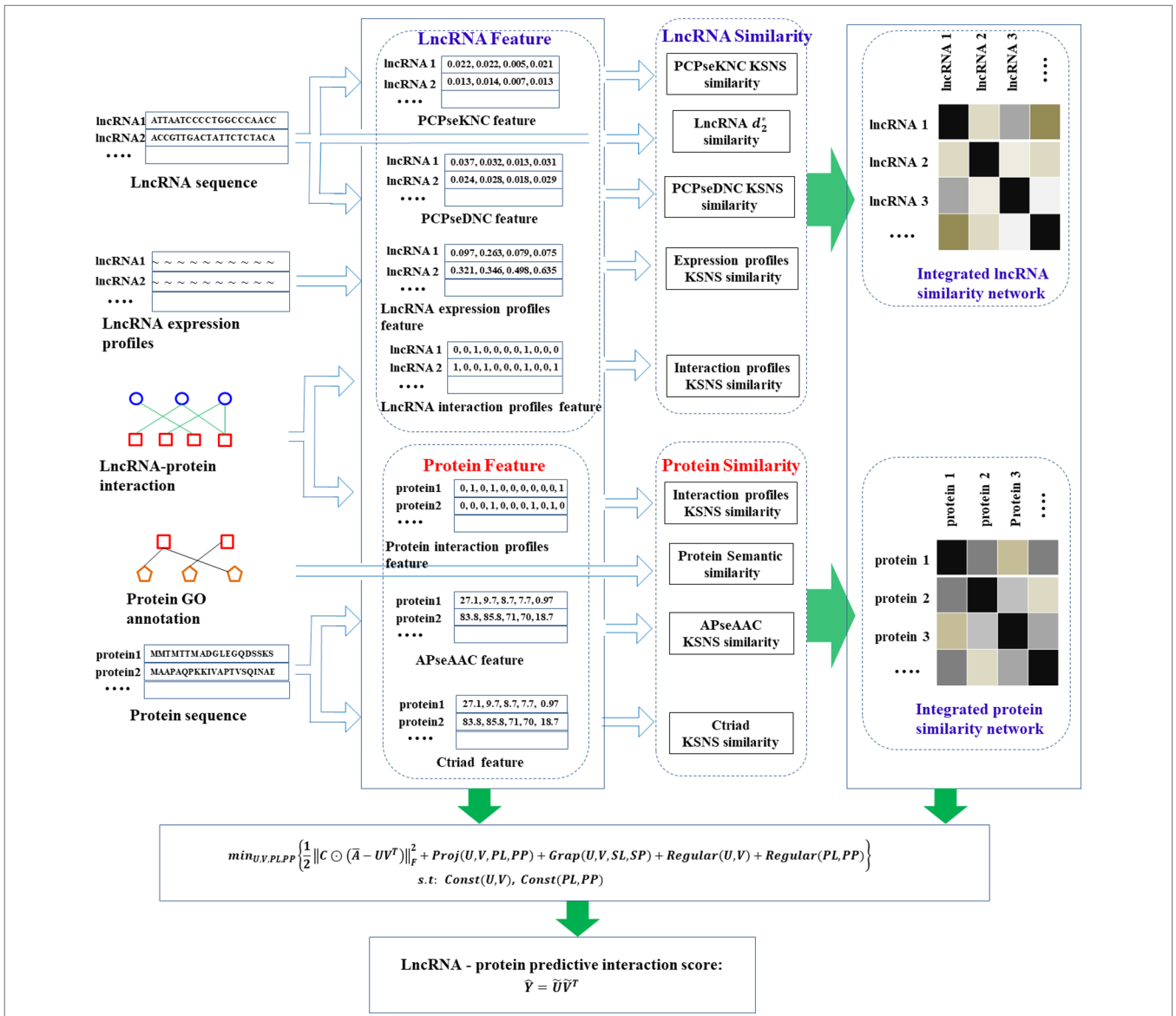

FIGURE 1 | Flow chart of lncRNA-protein interaction prediction by PMDKN algorithm. As shown in the figure, we first calculated three features of lncRNAs and two features of proteins, and then calculated five similarities of lncRNAs and four similarities of proteins according to lncRNA sequence, protein GO annotation and their features.

$v_j \in R^{d\times 1}$. For simplicity, we further denote the latent vectors of all lncRNAs and all proteins by $\mathrm{U}=\left(u_1,\cdots,u_{N_l}\right)^T \in R^{N_l\times d}$ and $\mathrm{V}=\left(v_1,\cdots,v_{N_h}\right)\in R^{d\times N_p}$, then, the product of the U and V can be used to approximate the modified interaction matrix $\bar{A}$. Since the observed interactions have been verified by experiments and have higher reliability than the unknown interactions, the observed lncRNA-protein interactions are assigned a higher level of importance and can be obtained as follows:

$$\boldsymbol{min}_{U,V}\left\{\frac{1}{2}\|\boldsymbol{C}\odot\left(\bar{\boldsymbol{A}}-\boldsymbol{U}\boldsymbol{V}^T\right)\|_F^2+\frac{\boldsymbol{\gamma}}{2}\left(\|\boldsymbol{U}\|_F^2+\|\boldsymbol{V}\|_F^2\right)\right\}$$

$$\textbf{s.t.}\ \ \mathbf{U}\geq 0,\ \mathbf{V}\geq 0 \tag{1}$$

where C is the importance level distribution matrix, that is, if there is interaction between the lncRNA l_i and the protein p_j, $C_{i,j}=\delta$, otherwise, $C_{i,j}=1$, where $\delta>1$ is an important level parameter. $\|\cdot\|_F$ denotes the F-norm and γ denotes the regularization parameter of latent vectors.

In addition, in order to integrate different types of lncRNA features, we project all lncRNA features onto the non-negative space R^d, and required the difference between it and U to be as small as possible, so as to obtain:

$$\min_{PL_i}\left\{\sum_{i=1}^{N_1}\boldsymbol{\alpha}_i^{\boldsymbol{\eta}}\|\boldsymbol{FL}_i\boldsymbol{PL}_i^T-\boldsymbol{U}\|_F^2+\frac{\mu}{2}\sum_{i=1}^{N_1}\sum_{k=1}^{r}\|\boldsymbol{PL}_i(\boldsymbol{k},:)\|_1^2\right\}$$

$$\textbf{s.t.}\ \boldsymbol{PL}_{\mathbf{i}}\geq 0 \tag{2}$$

where $FL_i\in R^{N_l\times d_{li}}$ represents the i-th feature matrix of lncRNA, d_{li} represents the dimension of the feature, and $PL_i\in R^{r\times d_{li}}$ represents the corresponding projection matrix. In order to facilitate calculation and convenient interpretation, PL_i is required to be non-negative. The Weight vector $\alpha=\left(\alpha_1,\alpha_2,\cdots,\alpha_{N_1}\right)$ controls the effect of different feature projections on U. The projection index parameter $\eta>1$ is the index of α, indicating that all features contribute to the generation of U. μ is the regularization parameter of projection matrix, and $P(k,:)$ is the k-th row of the matrix P. $||PL_i(k,:)||$ represents the ℓ_1 -norm of the vector $PL_i(k,:)$ (ie, $\|P(k,:)\|_1=\sum_j|P(k,j)|$), ensuring that the projection vector $PL_i(k,:)$ is as sparse as possible, and $\sum_{k=1}^{r}\|PL_i(k,:)\|_1^2$ is equivalent to the square of the $\ell_{1,2}$ -norm of the matrix PL_i. Therefore, equation (2) can be expressed as follows:

$$\min_{PL_i}\left\{\sum_{i=1}^{N_1}\boldsymbol{\alpha}_i^{\boldsymbol{\eta}}\|\boldsymbol{FL}_i\boldsymbol{PL}_i^T-\boldsymbol{U}\|_F^2+\frac{\mu}{2}\sum_{i=1}^{N_1}\|\boldsymbol{PL}_i\|_{1,2}^2\right\}$$

$$\textbf{s.t.}\ \boldsymbol{PL}_{\mathbf{i}}\geq 0 \tag{3}$$

Similarly, for proteins, we have:

$$\min_{PP_j}\left\{\sum_{j=1}^{N_2}\boldsymbol{\beta}_j^{\boldsymbol{\eta}}\|\boldsymbol{FP}_j\boldsymbol{PP}_j^T-\boldsymbol{V}\|_F^2+\frac{\mu}{2}\sum_{j=1}^{N_2}\|\boldsymbol{PP}_j\|_{1,2}^2\right\}$$

$$\textbf{s.t.}\ \boldsymbol{PP}_j\geq 0 \tag{4}$$

where $FP_j\in R^{N_p\times d_{pj}}$ represents the j-th feature matrix of the protein, and non-negative matrix $PP_j\in R^{r\times d_{pi}}$ represents the corresponding projection matrix. The weight vector $\beta=\left(\beta_1,\beta_2,\cdots,\beta_{N_2}\right)$ controls the effect of feature projection on V.

It is generally believed that lncRNAs with higher similarity are more likely to interact with the same protein, but due to the incomplete data set, the similarity network of lncRNAs (proteins) obtained may contain noise. In order to eliminate the influence of non-neighborhood noise and improve the prediction accuracy, we only consider strong neighborhood similarity relationship of the samples. Therefore, lncRNA neighborhood similarity matrix (SL) was constructed as follows:

$$\overline{\boldsymbol{SL}}_{\boldsymbol{i},\boldsymbol{j}}=\begin{cases}\boldsymbol{SL}_{\boldsymbol{i},\boldsymbol{j}} & \textit{if } \boldsymbol{l_j}\in\boldsymbol{N}\left(\boldsymbol{l_i}\right)\textit{ or }\boldsymbol{l_i}\in\boldsymbol{N}\left(\boldsymbol{l_j}\right)\\ \mathbf{0} & \textit{otherwise}\end{cases} \tag{5}$$

Among them, $\overline{SL}_{i,j}$ represents the local similarity of lncRNA l_i and l_j, and $N(l_i)$ represents the K neighbor sets closest to lncRNA l_i. In order to adaptively select the number of neighbors according to the sample size, we make $K=0.3\times N_l$, $\lfloor\cdot\rfloor$ indicates rounding up. It is known from equation (5) that $\overline{\mathrm{SL}}$ is a symmetric matrix. According to lncRNAs with higher similarity, their features are as close as possible; we have:

$$\frac{\boldsymbol{\lambda}}{\mathbf{2}}\sum_{\boldsymbol{i}}\sum_{\boldsymbol{j}}\overline{\boldsymbol{SL}}_{\boldsymbol{i},\boldsymbol{j}}\|\boldsymbol{u_i}-\boldsymbol{u_j}\|_F^2=\boldsymbol{\lambda tr}\left(\boldsymbol{U}^T\left(\boldsymbol{D}_l-\overline{\boldsymbol{SL}}\right)\boldsymbol{U}\right)=\lambda tr\left(\boldsymbol{U}^T\boldsymbol{LP}_l\boldsymbol{U}\right) \tag{6}$$

Where $tr(\cdot)$ represents the trace of the matrix, λ is the neighborhood Laplacian regularization parameter, and $LP_l=DL-SL$ is the Laplacian matrix of the lncRNA. The diagonal matrix $D_l=diag\left(dL_1,dL_2,\cdots,dL_{N_l}\right)$, whose diagonal elements are $dL_i=\sum_k\overline{SL}_{i,k}$, respectively. Similarly, we can calculate the neighborhood similarity matrix $\overline{SP}$ of the protein as follows:

$$\overline{\boldsymbol{SP}}_{\boldsymbol{i},\boldsymbol{j}}=\begin{cases}\boldsymbol{SP}_{\boldsymbol{i},\boldsymbol{j}} & \textit{if }\boldsymbol{p_j}\in\boldsymbol{N}\left(\boldsymbol{p_i}\right)\textit{ or }\boldsymbol{p_i}\in\boldsymbol{N}\left(\boldsymbol{p_j}\right)\\ \mathbf{0} & \textit{otherwise}\end{cases} \tag{7}$$

Furthermore, the objective function can be obtained as follows:

$$\frac{\boldsymbol{\lambda}}{\mathbf{2}}\sum_{\boldsymbol{i}}\sum_{\boldsymbol{j}}\overline{\boldsymbol{SP}}_{\boldsymbol{i},\boldsymbol{j}}\left\|\boldsymbol{p_i}-\boldsymbol{p_j}\right\|_F^2=\boldsymbol{\lambda tr}\left(\boldsymbol{V}^T\left(\boldsymbol{D}_p-\overline{\boldsymbol{SP}}\right)\boldsymbol{V}\right)=\boldsymbol{\lambda tr}\left(\boldsymbol{V}^T\boldsymbol{LP}_p\boldsymbol{V}\right) \tag{8}$$

where $LP_p=D_p-\overline{SP}$ is the Laplacian matrix of the protein. The diagonal matrix $D_p=diag\left(dP_1,dP_2,\cdots,dP_{N_p}\right)$, whose diagonal

elements are $dP_i = \sum_k \overline{SP}_{i,k}$, respectively. Combined with the above formulas, the objective function of PMKDN algorithm can be obtained as follows:

$$\begin{aligned}
&\min_{U,V,GU_i,GV_j,\alpha_i,\beta_j} \left\{ \frac{1}{2} \| C \odot (\overline{A} - UV^T) \|_F^2 \right. \\
&= \frac{1}{2} \sum_{i=1}^{N_1} \alpha_i^\eta \| FL_i PL_i^T - U \|_F^2 \\
&+ \frac{1}{2} \sum_{j=1}^{N_2} \beta_j^\eta \| FP_j PP_j^T - V \|_F^2 \\
&+ \frac{\lambda}{2} \left(tr(U^T LP_l U) + tr(V^T LP_p V) \right) \\
&+ \frac{\mu}{2} \left(\sum_{i=1}^{N_1} \| PL_i \|_{1,2}^2 + \sum_{j=1}^{N_2} \| PP_j \|_{1,2}^2 \right) \\
&\left. + \frac{\gamma}{2} (\| U \|_F^2 + \| V \|_F^2) \right\}
\end{aligned}$$

$$\begin{aligned}
&\text{st.} \quad U \geq 0, V \geq 0, \; PL_i \geq 0, PP_j \geq 0, \sum_{i=1}^{N_1} \alpha_i = 1, \\
&\sum_{j=1}^{N_2} \beta_j = 1, \alpha_i \geq 0, \; \beta_j \geq 0
\end{aligned} \tag{9}$$

We use the two-step method to solve (9). First, by fixing α_i, β_j, and using the Lagrangin multiplier and the KKT condition, we can get the iterative formula of U, V, PL_i and PP_j as follows:

$$U = U \odot \frac{\left(C \odot \overline{A}\right) V + \sum_{i=1}^{N_1} \alpha_i^\eta FL_i PL_i^T + \lambda_1 \overline{SL} U}{\left(C \odot UV^T\right) V + \sum_{i=1}^{N_1} \alpha_i^\eta U + \lambda_1 D_l \; U + \gamma_1 U} \tag{10}$$

$$V = V \odot \frac{\left(C \odot \overline{A}^T\right) U + \sum_{i=1}^{N_1} \beta_j^\eta FP_j PP_j^T + \lambda_2 \overline{SP} V}{\left(C \odot VU^T\right) U + \sum_{j=1}^{N_2} \beta_j^\eta V + \lambda_2 D_p V + \gamma_2 V} \tag{11}$$

$$PL_i = PL_i \odot \frac{\alpha_i^\eta U^T FL_i}{\alpha_i^\eta PL_i FL_i^T FL_i + \mu_1 PL_i ee^T} \tag{12}$$

$$PP_j = PP_j \odot \frac{\beta_j^\eta V^T FP_j}{\beta_j^\eta PP_j FP_j^T FP_j + \mu_2 PP_j ee^T} \tag{13}$$

Then, fix U, V, PL_i and PP_j, and let $a_i = \| FL_i PL_i^T - U \|_F^2 \geq 0$, $b_j = \| FP_j PP_j^T - V \|_F^2 \geq 0$, C_1 represents the terms unrelated to α_i and β_j (3.8). We can get the objective function about α_i and β_j as follows:

$$\min_{\alpha,\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N_u} a_i \alpha_i^\eta + \frac{1}{2} \sum_{j=1}^{N_v} b_j \beta_j^\eta + C_1 \right\}$$

$$\sum_{i=1}^{N_u} \alpha_i = 1, \sum_{j=1}^{N_v} \beta_j = 1, \alpha_i \geq 0, \; \beta_j \geq 0$$

Using the Lagrangian multiplier, the iterative formula for α_i and β_j can be obtained as follows:

$$\alpha_i = \frac{\left(\frac{1}{a_i}\right)^{\frac{1}{\eta-1}}}{\sum_{i=1}^{N_1} \left(\frac{1}{a_i}\right)^{\frac{1}{\eta-1}}} = \frac{\left(\frac{1}{\| FL_i PL_i^T - U \|_F^2}\right)^{\frac{1}{\eta-1}}}{\sum_{i=1}^{N_1} \left(\frac{1}{\| FL_i PL_i^T - U \|_F^2}\right)^{\frac{1}{\eta-1}}} \tag{14}$$

$$\beta_j = \frac{\left(\frac{1}{b_j}\right)^{\frac{1}{\eta-1}}}{\sum_{j=1}^{N_2} \left(\frac{1}{b_j}\right)^{\frac{1}{\eta-1}}} = \frac{\left(\frac{1}{FP_j PP_j^T - V_F^2}\right)^{\frac{1}{\eta-1}}}{\sum_{j=1}^{N_2} \left(\frac{1}{FP_j PP_j^T - V_F^2}\right)^{\frac{1}{\eta-1}}} \tag{15}$$

According to (14) and (15), α_i and β_j always satisfy non-negative constraints. In formula (9), U and V are obtained based on the decomposition of the known lncRNA-protein interaction matrix. In order to reduce the prediction error of the new lncRNA (lncRNA without any protein interaction information) and the new protein, we utilized the method proposed by Liu et al. (2016), that is, the lncRNA(protein) was modified by using the neighborhoodlatent vectors. Let $\tilde{u}_i$ the modified latent vector of lncRNA l_i, which can be calculated as follows:

$$\tilde{u}_i = \begin{cases} u_i & if \quad \sum_{j=1}^{N_p} A_{i,j} > 0 \\ \frac{1}{QL_i} \sum_{s \in N^+(l_i)} SL(i,s) u_s & otherwise \end{cases} \tag{16}$$

where, the first item $\sum_{j=1}^{N_p} A_{i,j} > 0$ indicates that the latent vector of lnRNA with protein interaction remain unchanged. The second term refers to the modification of latent vector of lncRNAs without protein interaction, where $N^+(l_i)$ refers to the set composed of K lncRNAs with the highest similarity to l_i among lncRNA sets with protein interaction. In order to make the number of neighbors automatically adapt to the size of samples, we set $K = \max\left(5, \lfloor 0.1 \times N_l \rfloor\right)$, where $Q_i = \sum_{s \in N^+(d_i)} SL(i,s)$ represents the normalized term. Similarly, we modified the latent vector of proteins as follows:

$$\tilde{v}_j = \begin{cases} v_j & if \quad \sum_{i=1}^{N_l} A_{i,j} > 0 \\ \frac{1}{QP_j} \sum_{t \in N^+(l_j)} SP(j,t) v_t & otherwise \end{cases} \tag{17}$$

By using the modified latent vector $\tilde{U} = \left[\tilde{u}_1, \tilde{u}_2, \cdots, \tilde{u}_{N_l}\right]$ of lncRNA and the modified latent vector $\tilde{V} = \left[\tilde{v}_1, \tilde{v}_2, \cdots, \tilde{v}_{N_p}\right]$ of protein, we can obtain the final lncRNA-protein interaction score $\bar{Y} = \tilde{U}\tilde{V}^T$.

Algorithm

In the process of model derivation, we assume that the features of lncRNA and protein are non-negative, so the original features need to be normalized before algorithm calculation. Let $\hat{F} \in R^{N \times M}$ represent the original feature matrix of lncRNA (protein), where $\hat{F}_{i,j}$ represents the j-th dimension of the i-th sample, then the normalized feature matrix F is as follows:

$$\boldsymbol{F}_{i,j} = \frac{\hat{\boldsymbol{F}}_{i,j} - \min\left(\hat{\boldsymbol{F}}_{\cdot j}\right)}{\max(\hat{\boldsymbol{F}}_{\cdot j}) - \min\left(\hat{\boldsymbol{F}}_{\cdot j}\right)} \tag{18}$$

Where, min (F_j) and max (F_j) represent the minimum and maximum of the j-th dimension, respectively. Algorithm 1 summarizes the general process of solving lncRNA-protein interaction prediction by KDMPN.

ALGORITHM 1 | KDMPN

Input: Known lncRNA-protein interaction matrix A; Modified lncRNA-protein interaction matrix $\bar{A}$; Importance level parameter δ; LncRNA original feature matrix $\left\{\widehat{FL}_i\right\}_{i=1}^{N_1}$; Protein initial feature matrix $\left\{\widehat{FP}_j\right\}_{j=1}^{N_2}$; LncRNA similarity matrix SL; Protein similarity matrix SP; Potential subspace regularization parameter r; Projection index parameter η>1; Projection matrix regularization parameter μ; Neighborhood Laplacian regularization parameter λ; Potential subspace regularization parameter γ.

Output: LncRNA latent vector $\tilde{U}$; Protein latent vector $\tilde{V}$; Predictive interaction matrix $\bar{Y}$; LncRNA feature projection matrix $\left\{PL_i\right\}_{i=1}^{N_1}$; Protein feature projection matrix $\left\{PP_j\right\}_{j=1}^{N_2}$; LncRNA projection parameter $\left\{\alpha_i\right\}_{i=1}^{N_1}$; Protein projection parameter $\left\{\beta_j\right\}_{j=1}^{N_2}$.

Initialize:

1 The importance level distribution matrix $(C)_{N_l \times N_p}$ is calculated from δ and A, andthe normalized lncRNA feature matrix $\left\{FL_i\right\}_{i=1}^{N_1}$ and the protein projection matrix $\left\{FP_j\right\}_{j=1}^{N_2}$ are obtained by using Equation (18) for $\left\{\widehat{FL}_i\right\}_{i=1}^{N_1}$ and $\left\{\widehat{FP}_j\right\}_{i=1}^{N_2}$. Based on SL and SP, the neighborhood similarity matrices $\overline{SL}$ and $\overline{SP}$ of lncRNA and protein were obtained using equations (5) and (7), respectively.

Initialize $U, V \,\|\, \left\{PL_i\right\}_{i=1}^{N_1}$ and $\left\{PP_j\right\}_{j=1}^{N_2}$ using the random number of the [0, 1] interval.

2 for $i \leftarrow 1,2,\cdots,N_1$ do
  Fix PL_i and U, calculate α_i according to formula (14).
 end for

for $i \leftarrow 1,2,\cdots,N_2$ do
  Fix PP_j and V, calculate β_j according to formula (15).
 end for

repeat

3 Fix $\left\{\alpha_i\right\}_{i=1}^{N_1}$ and $\left\{PL_i\right\}_{i=1}^{N_1}$, Update U according to formula (10).

4 Fix $\left\{\beta_j\right\}_{j=1}^{N_2}$ and $\left\{PP_j\right\}_{j=1}^{N_2}$, Update V according to formula (11).

5 for $i \leftarrow 1,2,\cdots,N_1$ do
  Fix $\left\{\alpha_i\right\}_{i=1}^{N_1}$ and U, Update PL_i according to formula (12).
  Fix PL_i and U, Update α_i according to formula (14).
 end for

6 for $j \leftarrow 1,2,\cdots,N_2$ do
  Fix $\left\{\beta_j\right\}_{j=1}^{N_2}$ and V, Update PP_j according to formula (13).
  Fix PP_j and V, Update β_j according to formula (15).
 end for

until Converges

7 $\tilde{U}$ was obtained by completing the subspace feature U of lncRNA according to formula (16).

8 $\tilde{V}$ was obtained by completing the subspace feature V of protein according to formula (17).

9 $\bar{Y} = \tilde{U}\tilde{V}^T$

RESULTS AND DISCUSSION

Experimental Settings

According to previous studies, the performance of the interactive prediction method was evaluated by the 5-fold cross validation (CV), and the area under ROC curve (AUC), area under Precision-Recall curve (AUPR), and F1 value (F1) were used as evaluation indexes. Since the known lncRNA-protein interactions were much less than the unknown lncRNA-protein interactions, AUPR was usually used as the most important evaluation index to punish false positives (Zhang et al., 2018a; Zhang et al., 2018b).

In addition, in order to eliminate the influence of random partition on the results in the crossover experiment, we selected the method of Liu et al. (2016), set 5 random seeds for CV, and took the mean value of the cross experiment results under all random seeds as the final prediction result. Specifically, the lncRNA-protein interaction matrix $A \in R^{N_l \times N_p}$ has N_l rows for lncRNAs and N_p columns for proteins. In order to investigate the prediction ability for lncRNA-protein interactions, new lncRNAs and new proteins, we performed CV under three different settings, as follows:

1. CV_a: CV on known lncRNA-protein interaction pairs. Specifically, we randomly divided the known lncRNA-protein interactions into 5 equal parts. Take turns to select one and all the unknown interactions to form the test set and the remaining four and all the unknown interactions to form the training set (that is, change the 1 corresponding to the test set in A into 0 as the training set).
2. CV_l: CV on lncRNAs. Specifically, all lncRNAs are randomly divided into five equal parts; one is selected as a test set in turn, and the remaining four are training sets (that is, all the rows corresponding to the test set in A were changed to zeros).
3. CV_p: CV on proteins. Specifically, all proteins are randomly divided into five equal parts; one is selected as a test set in turn, and the remaining four are training sets (that is, all the columns corresponding to the test set in A were changed to zeros).

It should be noted that with regard to CV_a, we selected all zeros in A as the test set. For example, for DATA2, the test set of each crossover experiment contains 4,870/5 = 947 known interactions and 97,658 unknown interactions (that is, the ratio

of positive and negative examples is approximately 1:100). This selection method ensures that all the unknown interactions can be included in each crossover experiment, which expands the search range and is also in line with the actual situation.

Parameter Setting

The PMDKN algorithm have six parameters, namely the projection index parameter η, the projection regularization parameter μ, the latent vector regularization parameter γ, the neighborhood Laplacian regularization parameter λ, the potential subspace dimension d, and the known interaction important level parameter δ. Among them, μ and γ control the influence of feature projection, γ controls subspace feature contribution, λ describes the effect of similarity network, and δ controls the importance level of observed interaction. In order to study the effect of parameters on the prediction results, we calculated all the parameter combinations. Specifically, η was selected from {2,3,4,5,6}, μ was selected from $\{10^{-3},10^{-2},10^{-1},10^{0},10^{1}\}$, γ was selected from $\{10^{-3},10^{-2},10^{-1},10^{0},10^{1}\}$, and λ was selected from $\{2^{-2},2^{-1},2^{0}2^{1},2^{2},2^{3}\}$; according to the previous research (Zheng et al., 2013, Liu et al., 2016, Xiao et al., 2018), for methods based on matrix decomposition, the potential subspace dimension d = 100, δ was selected from {1,2,⋯, 6}.

It should be noted that unlike DATASET 1, DATASET 2 contained more lncRNAs and proteins, and the initially constructed lncRNA (protein) similarity network did not utilize any known interaction information and therefore has higher predictive value. In addition, since CV_a, CV_l, and CV_p are considered the predictive power of the algorithm for new interactions, new lncRNAs, and new proteins, respectively, we believe that the three experimental setups are equally important for algorithm evaluation. Therefore, based on DATASET 2, for the combination of different parameters, the average evaluation index of the algorithm under the three experimental settings is the final evaluation standard. We take AUPR as the evaluation index, and the influence of the analysis parameters on the prediction results was shown in **Figure 2**.

As shown in Figure 2, the optimal parameters obtained by the PMDKN algorithm are η = 5, μ = 100, λ = 1, γ = 1, δ = 2, and the average optimal AUPR value under the three experimental settings is 0.4735. Specifically, we first analyze the influence of the projection parameters η and μ. Fixed λ = 1, γ = 1, δ = 2, and calculate the AUPR value of the model under all possible combinations of η and μ. As shown in (A) of **Figure 2**, as η becomes larger, the AUPR value of the model increases, but the overall AUPR value of the model fluctuates a little. Then, we fixed η = 5, μ = 100, γ = 1, δ = 2, and analyzed the influence of the change of λ on the AUPR value. As shown in (B) of **Figure 2**, when λ increases, the AUPR value of the model first becomes larger and then decreases, and when λ = 1, the AUPR value is the largest. Similarly, as shown in (C) in **Figure 2**, when γ < 1, the change of AUPR was relatively flat; when γ > 1, the AUPR value decreased sharply with the increase of gamma. In

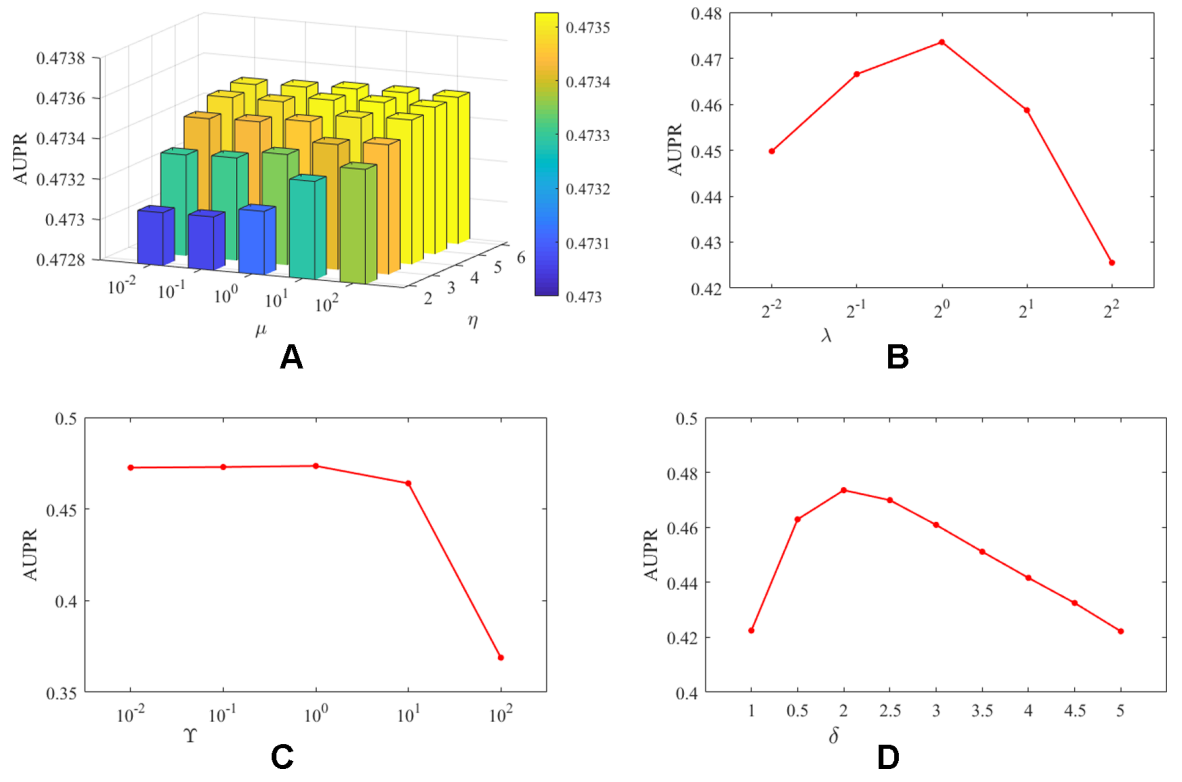

FIGURE 2 | The influence of parameters on the AUPR value of PMDKN. Among them, **(A)** represents the influence of the projection parameters μ and η on the AUPR value. **(B)** shows the effect of the neighborhood Laplacian regularization parameter λ. **(C)** shows the effect of the feature regularization parameter γ. **(D)** indicates the effect of observing the important level parameter δ.

(D), $\delta = 1$ indicates that the known interactions and the unknown interactions are equally important, and the corresponding AUPR value of the model is only 0.42; however, when $\delta = 2$, the model has the maximum AUPR value, which further emphasized that the setting of δ is necessary to improve the performance of the model.

Based on the above discussion, in the following study, we select $\eta = 5$, $\mu = 100$, $\lambda = 1$, $\gamma = 1$, $d = 100$, and $\delta = 2$ as parameters of PMDKN.

Comparison With State-of-the-Art Prediction Methods

In order to evaluate the predictive ability of PMDKN algorithm equitably, we conducted 5-fold cross validation on DATASET 1 and DATASET 2, and compared them with the following methods: SFPEL-LPI (Zhang et al., 2018b), LPLNP (Zhang et al., 2018a), LPBNI (Ge et al., 2016), and LKSNF (Ma et al., 2018b). Since DATASET 1 itself was the benchmark dataset for SFPEL-LPI, LPLNP, and LKSNF, we do not need to re-extract the features. For DATASET 2, we calculated the PCPseDNC and SCPseAAC of lncRNA according to the requirements of SFPEL-LPI, and calculated the PCPseAAC and SCPseAAC of the protein. Since SWSS similarity leads to the reuse of known interaction information, only the Smith Waterman similarity of lncRNA (protein) were calculated. For LPLNP and LKSNF, we calculated the sequence feature and expression profile feature of lncRNA and the CTD of the protein according to their requirements. While LPBNI only uses known lncRNA-protein interactions for prediction, we did not need to extract additional information. According to previous studies, LPLNP, LPBNI, and LKSNF only predicted the unknown interaction of lncRNA-protein, while SFPEL-LPI not only predicted unknown lncRNA-protein interactions, but also predicted new lncRNA and new protein. Therefore, based on DATASET 1 and DATASET 2, we perform CV_a on all models, and CV_l and CV_p on SFPEL-LPI. We performed the crossover experiment using the experimental setup in Section "Experimental Settings" and used the mean of the five-fold crossover experimental results of the five random seeds as the evaluation index of the algorithm, and the parameters of these models were selected using the recommended parameters.

Table 1 shows the comparison of predictive performance of PMDKN and other state-of-the-art methods for new lncRNA-protein interaction prediction. It can be seen that, no matter in DATASET 1 or DATASET 2, the AUPR, AUC, and F1 values of PMDKN are higher than other models. Specifically, on DATASET 1, as for the most important evaluation index AUPR, PMDKN can reach 0.4959, which increases by 50.46%, 8.37%, 4.31%, and 6.07%, respectively, compared with LPBNI's 0.3296, LPLNP's 0.4576, LKSNF's 0.4754, and SFPEL-LPI's 0.4675. Regarding the commonly used evaluation index AUC, PMDKN can reach 0.9223, which is higher than 0.8546 of LPBNI, 0.9095 of LPLNP, 0.9150 of LKSNF, and 0.9201 of SFPEL-LPI. The F1 value of PMDKN can reach 0.4814, which is 24.04%, 6.50%, 4% and 3.37%, respectively, compared with 0.3881 for LPBNI, 0.4520 for LPLNP, 0.4629 for LKSNF, and 0.4657 for SFPEL-LPI. In DATASET 2, the AUPR of PMDKN could reach 0.4808, which improved by 40.67%, 2.45%, 6.18%, and 14.07%, respectively, compared with 0.3418 of LPBNI, 0.4693 of LPLNP, 0.4528 of LKSNF, and 0.4215 of SFPEL-LPI. The AUC value of PMDKN can reach 0.9732, higher than 0.9340 of LPBNI, 0.9700 of LPLNP, 0.9710 of LKSNF, and 0.9728 of SFPEL-LPI. The F1 value of PMDKN can reach 0.4761, which is 19.71%, 3.37%, 2.67%, and 7.04%, respectively, compared with 0.3977 for LPBNI, 0.4606 for LPLNP, 0.4637 for LKSNF, and 0.4448 for SFPEL-LPI. These demonstrate that the PMDKN algorithm of this paper has good predictive power for unknown lncRNA-protein interactions.

The prediction of new lncRNAs and new proteins are also the important criterion for evaluating the performance of the method. Among the four comparison algorithms above, only SFPEL-LPI performs the prediction of new lncRNA and new protein. Therefore, we only compare the prediction performance of SFPEL-LPI and PMDKN on CV_l and CV_p. As shown in **Table 2**, except for the F1 value of PMDKN on DATASET 2, which is 0.4864, slightly lower than the 0.4892 of SFPEL-LPI, PMDKN was better than SFPEL-LPI for other evaluation indicators, especially for the prediction of new proteins (CVp). Specifically, on DATASET 1, the AUPR values of PMDKN for CV_l and CV_p can reach 0.6301 and 0.4918, which is 30.92% and 49.71%, respectively, relative to SFPEL-LPI of 0.4813 and 0.3285. The AUC values of the PMDKN algorithm for CV_l and CV_p can reach 0.8907 and 0.7843, which are 7.52% and 17.66% higher than the 0.8284 and 0.6666 of SFPEL-LPI, respectively. The F1 value of the PMDKN algorithm for CV_l and CV_p can reach 0.6081 and 0.5251, which is 23.32% and 38.95% higher than the 0.4931 and 0.3779 of SFPEL-LPI, respectively. Similarly, in DATASET 2, the AUPR value and AUC value of CV_l of PMDKN were higher than SFPEL-LPI, especially for CV_p, the AUPR value, AUC value, and F1 value of PMDKN could reach 0.4604, 0.9019, and 0.4818, respectively, improving 281.13%, 37.78%, and 148.35% compared with the 0.1208, 0.6546, and 0.1940 of SFPEL-LPI.

Comparative Analysis of Model Stability

Due to technical limitations, some noises may be hidden in the known lncRNA-protein interactions, such as lack of interaction information, unreal interaction information and so on. In order to test the dependence of the prediction performance of the model on the known interactions, according to the method of Zhang et al. (2018b), we randomly deleted some of the known interactions to represent the missing

TABLE 1 | Comparison of predicted performance of new lncRNA-protein interactions based on DATASET1 and DATASET2.

DATA	Method	AUPR	AUC	F1 value
DATASET 1	LPBNI	0.3296	0.8546	0.3881
	LPLNP	0.4576	0.9095	0.4520
	LKSNF	0.4754	0.9150	0.4629
	SFPEL-LPI	0.4675	0.9201	0.4657
	PMDKN	**0.4959**	**0.9223**	**0.4814**
DATASET 2	LPBNI	0.3418	0.9340	0.3977
	LPLNP	0.4693	0.9700	0.4606
	LKSNF	0.4528	0.9710	0.4637
	SFPEL-LPI	0.4215	0.9728	0.4448
	PMDKN	**0.4808**	**0.9732**	**0.4761**

In the above table, the best results under the current metric are shown in bold on each data set.

TABLE 2 | Comparison of predicted performance of new lncRNAs and new proteins based on DATASET1 and DATASET2.

DATA	Method	CV_l			CV_p		
		AUPR	AUC	F1 value	AUPR	AUC	F1 value
DATASET 1	SFPEL-LPI	0.4813	0.8284	0.4931	0.3285	0.6666	0.3779
	PMDKN	**0.6301**	**0.8907**	**0.6081**	**0.4918**	**0.7843**	**0.5251**
DATASET 2	SFPEL-LPI	0.4756	0.9446	**0.4892**	0.1208	0.6546	0.1940
	PMDKN	**0.4794**	**0.9465**	0.4864	**0.4604**	**0.9019**	**0.4818**

In the above table, the best results under the current metric are shown in bold on each data set.

information and randomly added the nonexistent interactions to represent the false interactions, and then studied the change of prediction performance of the model. Since only a few interactions have been detected at present, it indicates that there are still a large number of interactions that have not been discovered. Therefore, we deleted 20% of the known lncRNA-protein interactions and added 5% of the interactions that actually do not exist as noise. At this point, the test set of the model becomes 20% known interactions and all unknown interactions. As shown in **Figure 3**, on the disturbance dataset of DATASET 1, the AUC values of LPBNI, LKSNF, LPLNP, SFPEL-LPI, and PMDKN are 0.8417, 0.8980, 0.8856, 0.9077, and 0.9116, respectively, and the AUPR values are 0.2776, 0.2564, 0.2431, 0.2596, and 0.3392. On the perturbed data set of DATASET 2, the AUC values of LPBNI, LKSNF, LPLNP, SFPEL-LPI, and PMDKN were 0.9297, 0.9707, 0.9646, 0.9687, and 0.9714, respectively, and the AUPR values were 0.2969, 0.2662, 0.2526, 0.2415, and 0.4081, respectively. Comparing the results of **Table 1**, it can be seen that the introduction of partial noise in the perturbed dataset leads to a decrease in the AUPR and AUC values of all prediction models, but PMDKN still achieves satisfactory results and outperforms LPBNI, LKSNF, LPLNP, and SFPEL-LPI.

Case Study

LncRNA-protein interactions in DATASET 1 and DATASET 2 used in this paper were extracted from Npinter2.0, and the current version of Npinter has been updated to Npinter v3.0 (Hao et al., 2016). Compared with version 2.0(Yuan et al., 2013), Npinter v3.0 contains more lncRNAs, more proteins, and more interactive information. To test the predictive ability of new proteins, we extracted 95 new proteins that

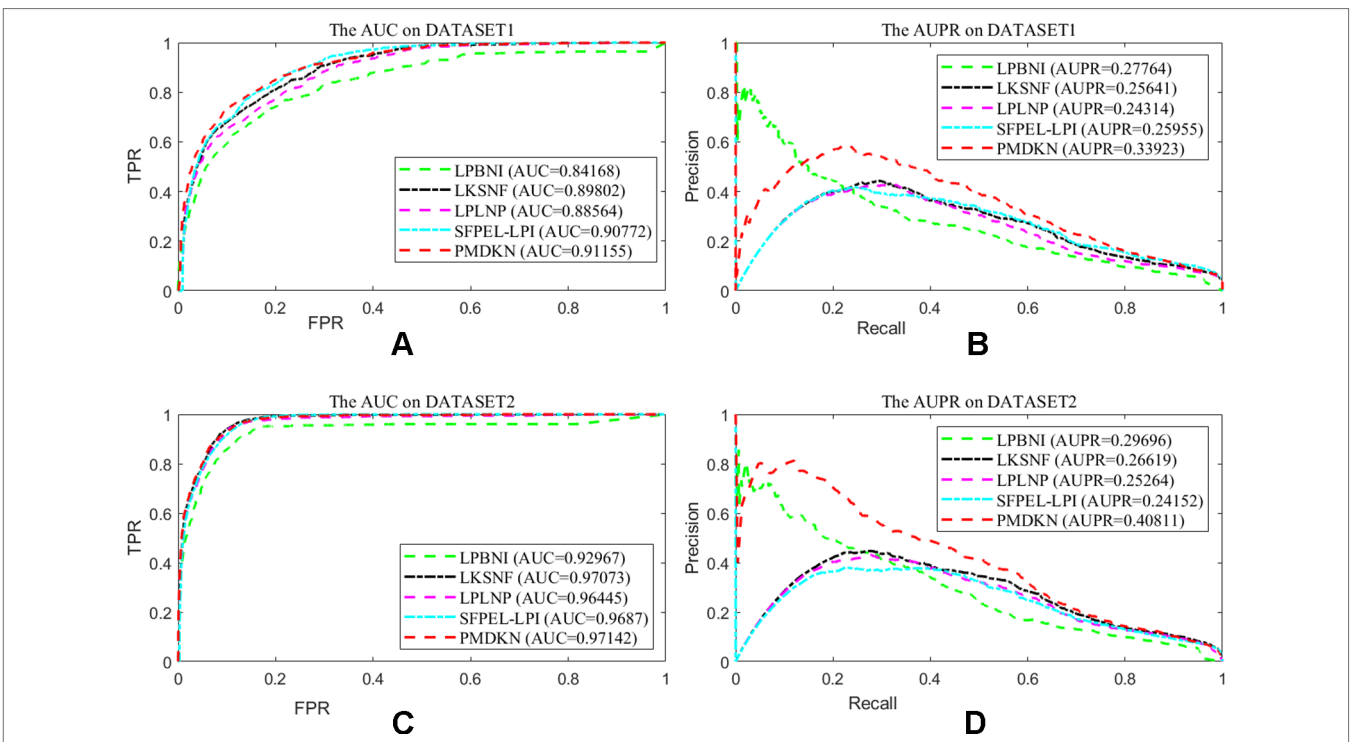

FIGURE 3 | Prediction performance of the model on disturbed data set. Among them, **(A)** shows the ROC curve and AUC value of the five methods after DATASET1 adds noise. **(B)** shows the P-R curve and AUPR values of the five methods after DATASET1 is added with noise. **(C)** shows the ROC curve and AUC value of the five methods after DATASET2 adds noise. **(D)** indicates the P-R curve and AUPR value of the five methods after DATASET2 is added with noise.

did not exist in DATASET 2 from Npinter v3.0, extracted the amino acid sequence and gene ontology annotation of these new proteins, and combined with DATASET2 information to predict the interactions between these new proteins and lncRNAs. For the prediction score of each new protein, we calculated its AUPR and AUC values, and calculated the hit rate of the top 10, 20, 50, and 100 candidate lncRNAs (Nourania et al., 2016). For the new protein p_i, the hit rate hit(p_i) can be expressed as follows:

$$\mathrm{hit}\left(p_i\right)=\frac{\left|\left|Cand\left(p_i\right)\cap Test\left(p_i\right)\right|\right|}{\left|\left|Test\left(p_i\right)\right|\right|}$$

Among them, *Cand* (p_i) represents the candidate lncRNA set of protein p_i, and in this paper *Cand* (p_i) represents the *top*-10, *top*-20, *top*-50, *top*-100 candidate lncRNAs sorted according to the predicted score, respectively. *Test* (p_i) represents the set of lncRNAs for all interactions of protein p_i in Npinter v3.0. $|\cdot|$ indicates the number of elements. As SFPEL-LPI can predict new proteins and new lncRNAs, the predicted results of SFPEL-LPI and PMDKN were compared. The predicted scores, actual labels and evaluation indicators of 95 new proteins are shown in **Supplementary Table 1**. The average AUPR value, the average AUC value, the average hit rate of the *top*-10, *top*-20, *top*-50, and *top*-100 predicted by SFPEL-LPI and PMDKN for 95 proteins are shown in **Figure 4**.

As shown in **Figure 4**, for the prediction of new proteins, PDMKN not only has higher AUPR and AUC values than SFPEL-LPI, but also the top 10, 20, 50, 100 hit ratios of candidate lncRNAs are much higher than SFPEL-LPI. Specifically, the average AUPR and AUC values for PMDKN were 0.204 and 0.839, respectively, which were 20.66% and 8.49% higher than 0.169 and 0.773 for SFPEL-LPI, respectively. The hit rates of candidate lncRNAs in the *top*-10, *top*-20, *top*-50 and *top*-100 reached 42.8%, 47.1%, 52.1%, 57.2%, and increased by 266.32%, 264.37%, 125.75%, and 80.68%, respectively, compared with the 11.7%, 12.9%, 23.1%, and 31.7% of SFPEL-LPI, which further demonstrated that PMDKN had strong predictive ability.

DISCUSSION

In this study, we proposed a new lncRNA-protein interaction prediction model, which not only can predict the unknown interactions between lncRNAs and proteins, but also has strong prediction ability for new lncRNAs and new proteins. To fairly evaluate the predictive performance of the model, we performed three 5-fold cross-validation on the two benchmark datasets, namely, CV_a for the new lncRNA-protein interactions, CV_l for the new lncRNAs, and CV_p for the new proteins. The results show that, on DATASET 1, the AUPR values of PMDKN under the three experimental settings could reach 0.4959 (on CV_a), 0.6301 (on CV_l), and 0.4918(on CV_p) respectively; on DATASET 2, the AUPR values of PMDKN under the three experimental settings can reach 0.4808 (on CV_a), 0.4794 (on CV_l), and 0.4604 (on CV_p) respectively, higher than other state-of-the-art methods. In the case study, 95 new proteins were predicted, and the results

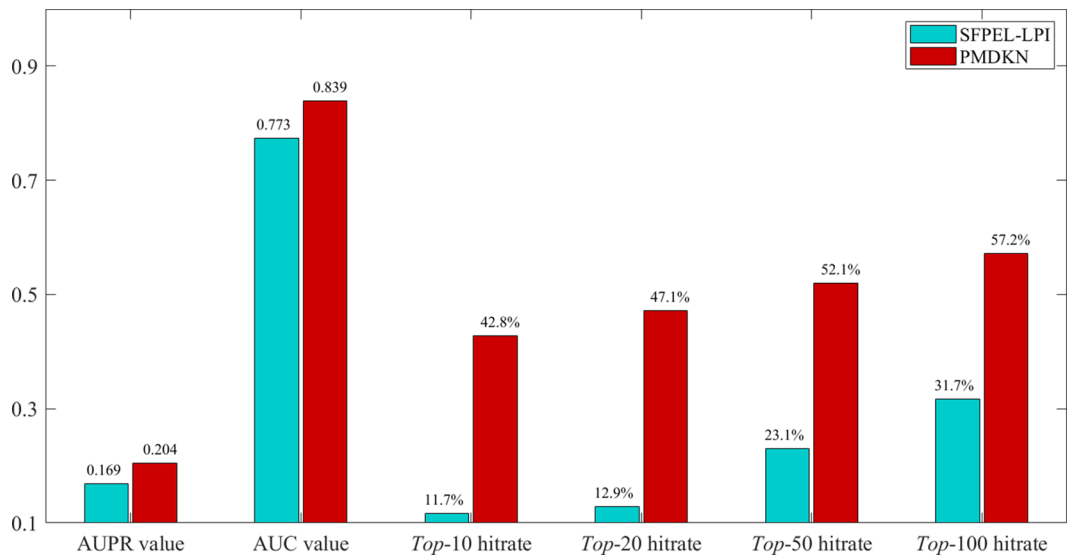

FIGURE 4 | Comparison of SFPEL-LPI and PMDKN prediction results for new proteins. The AUPR and AUC values in the figure represent the average AUPR and average AUC values predicted by PMDKN for 79 proteins, respectively. Top-10 hitrate, Top-20 hitrate, Top-50 hitrate, and Top-100 hitrate represent the mean hit rates of the first 10, 20, 50, and 100 candidate lncRNAs, respectively.

showed that for the top-10 candidate lncRNAs, the hit rate of PMDKN algorithm could reach 42.8%, much higher than other method. Therefore, PMDKN can be used as an effective tool for lncRNA-protein interaction prediction.

The good performance of PMDKN may have the following reasons: First, feature extraction and network construction. We extract multiple features to describe lncRNA and protein in all directions and integrate multiple infomation to construct a more accurate lncRNA (protein) similarity network, effectively avoiding the over-fitting problem that may be caused by the information deviation of a single data source. Second, the use of neighborhood information. We modified the initial lncRNA-protein interaction network to overcome the network sparsity problem, and used the adaptive neighborhood completion strategy to eliminate the errors caused by the lack of information in the latent vectors of new lncRNAs (new protein), so as to ensure the predictive ability of new proteins and new lncRNAs. Finally, the construction of the ensemble predictive model. We combine the multiple sequence features of lncRNA (protein) and the integrated similarity networks to construct the predictive model, which distinguishes positive and negative observations by setting important levels and establishes the relationship between features and potential vectors through the projection of the features, so as to improve the accuracy of model prediction.

DATA AVAILABILITY STATEMENT

The source code and datasets used in the paper can be found in the **Supplementary Files**.

AUTHOR CONTRIBUTIONS

YM and XJ designed the projection-based neighborhood non-negative matrix factorization for lncRNA-protein interaction prediction. YM and XJ designed the experiment and wrote the manuscript. TH and XJ supervised and helped conceive the study. All authors read and approved the final manuscript.

FUNDING

This research is supported by National Key Research and Development Program of China (2017YFC0909502) and the National Natural Science Foundation of China (61532008 and 61872157).

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01148/full#supplementary-material

S1 TABLE | Prediction results of 95 new proteins by SFPEL-LPI and PMDKN.

S1 FILE | Eps format for all pictures in the manuscript.

S2 FILE | The code and data of PMDKN.

REFERENCES

Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A., and Sun, F. (2016). Alignment-free d ∗ 2 oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* 45, 39–53. doi: 10.1093/nar/gkw1002

Batista, P. J., and Chang, H. Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. *Cell* 152, 1298–1307. doi: 10.1016/j.cell.2013.02.012

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs. *Nat. Methods* 8, 444–445. doi: 10.1038/nmeth.1611

Chen, W., Lei, T., Jin, D., Lin, H., and Chou, K. (2014). PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.* 456, 53–60. doi: 10.1016/j.ab.2014.04.001

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins-Struct. Funct. And Bioinf.* 43, 246–255. doi: 10.1002/prot.1035

Chou, K. C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinf.* 21, 10–19. doi: 10.1093/bioinformatics/bth466

Deng, L., Wang, J., Xiao, Y., Wang, Z., and Liu, H. (2018). Accurate prediction of protein-lncRNA interactions by diffusion and HeteSim features across heterogeneous network. *BMC Bioinf.* 19, 370. doi: 10.1186/s12859-018-2390-0

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. (2012). Landscape of transcription in human cells. *Nat.* 489, 101–108. doi: 10.1038/nature11233

Fang, Y., and Fullwood, M. J. (2016). Roles, functions, and mechanisms of long non-coding RNAs in cancer. *Genomics Proteomics Bioinf.* 14, 42–54. doi: 10.1016/j.gpb.2015.09.006

Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., et al. (2018). NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* 46, D308–D314. doi: 10.1093/nar/gkx1107

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., et al. (2012). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–D815. doi: 10.1093/nar/gks1094

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

Guo, S. H., Deng, E. Z., Xu, L. Q., Ding, H., Lin, H., Chen, W., et al. (2014). iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinf.* 30, 1522–1529. doi: 10.1093/bioinformatics/btu083

Hao, Y., Wu, W., Li, H., Yuan, J., Luo, J., Zhao, Y., et al. (2016). NPInter v3.0: an upgraded database of noncoding RNA-associated interactions. *Database*. doi: 10.1093/database/baw057

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

Jiang, J. J., and Conrath, D. W.Jiang, J. J., and Conrath, D. W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: In Tenth International Conference on Research on Computational Linguistics (ROCLING X). (Ed.)

Khalil, A. M., and Rinn, J. L. (2011). RNA–protein interactions in human health and disease. *Semin. In Cell Dev. Biol.* 22, 359–365. doi: 10.1016/j.semcdb.2011.02.016

Kirk, J. M., Kim, S. O., Inoue, K., Smola, M. J., Lee, D. M., Schertzer, M. D., et al. (2018). Functional classification of long non-coding RNAs by k-mer content. *Nat. Genet.* 50, 1474–1482. doi: 10.1038/s41588-018-0207-8

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, 1–11. doi: 10.1155/2015/671950

Lin, D. (1998). In Proceedings of the Fifteenth International Conference on Machine Learning in *An Information-Theoretic Definition of Similarity*. lcml. 296–304.

Liu, B., Liu, F., Fang, L., Wang, X., and Chou, K. C. (2015). repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinf.* 31, 1307–1309. doi: 10.1093/bioinformatics/btu820

Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X. (2016). Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction. *PloS Comput. Biol.* 12, e1004760. doi: 10.1371/journal.pcbi.1004760

Liu, C. (2004). NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.* 33, D112–D115. doi: 10.1093/nar/gki041

Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X., et al. (2013). Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics* 14, 651. doi: 10.1186/1471-2164-14-651

Ma, Y., Ge, L., Ma, Y., Jiang, X., He, T., and Hu, X. (2018a). "2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)," in *Kernel Soft-neighborhood Network Fusion for MiRNA-Disease Interaction Prediction*. IEEE, 197–200. doi: 10.1109/BIBM.2018.8621122

Ma, Y., Yu, L., He, T., Hu, X., and Jiang, X. (2018b). 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), in *Prediction of Long Non-coding RNA-protein Interaction through Kernel Soft-neighborhood Similarity*. IEEE, 193–196. doi: 10.1109/BIBM.2018.8621460

Mattick, J. S. (2005). The functional genomics of noncoding RNA. *Science* 309, 1527–1528. doi: 10.1126/science.1117806

Nourania, E., Khunjush, F., and Durmuş, S. (2016). Computational prediction of virus-human protein-protein interactions using embedding kernelized heterogeneous data. *Mol. Biosyst.* 12, 1976–1986. doi: 10.1039/C6MB00065G

Pan, X., and Shen, H. (2017). RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC Bioinf.* 18, 136. doi: 10.1186/s12859-017-1561-8

Resnik, P. (1999). Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language *J. Artif. Intell. Res.* 11, 95–130. doi: 10.1186/s12859-017-1561-8

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., et al. (2007). Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. U. S. A* 104, 4337–4341. doi: 10.1073/pnas.0607879104

Song, K., Ren, J., Reinert, G., Deng, M., Waterman, M. S., and Sun, F. (2014). New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings Bioinf.* 15, 343–353. doi: 10.1093/bib/bbt067

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. doi: 10.1093/nar/gkv020

Ulf Andersson ørom, T. D. M. B., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al. (2010). Long Noncoding RNAs with Enhancer-like Function in Human Cells. *Cell* 143, 46–58. doi: 10.1016/j.cell.2010.09.001

Volders, P., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., et al. (2013). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.* 41, D246–D251. doi: 10.1093/nar/gks915

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C. (2007). A new method to measure the semantic similarity of GO terms. *Bioinf.* 23, 1274–1281. doi: 10.1093/bioinformatics/btm087

Wang, S., Cho, H., Zhai, C., Berger, B., and Peng, J. (2015). Exploiting ontology graph for predicting sparsely annotated gene function. *Bioinf.* 31, i357–i364. doi: 10.1093/bioinformatics/btv260

Wapinski, O., and Chang, H. Y. (2011). Long noncoding RNAs and human disease. *Trends In Cell Biol.* 21, 354–361. doi: 10.1016/j.tcb.2011.04.001

Wu, T., Wang, J., Liu, C., Zhang, Y., Shi, B., Zhu, X., et al. (2006). NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic Acids Res.* 34, D150–D152. doi: 10.1093/nar/gkj025

Xiao, N., Cao, D., Zhu, M., and Xu, Q. (2015). protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinf.* 31, 1857–1859. doi: 10.1093/bioinformatics/btv042

Xiao, Y., Zhang, J., and Deng, L. (2017). Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks. *Sci. Rep.* 7, 3664. doi: 10.1038/s41598-017-03986-1

Xiao, Q., Luo, J., Liang, C., Cai, J., and Ding, P. (2018). A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinf.* 34, 239–248. doi: 10.1093/bioinformatics/btx545

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., et al. (2013). NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res.* 42, D98–D103. doi: 10.1093/nar/gkt1222

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinf.* 26, 976–978. doi: 10.1093/bioinformatics/btq064

Yuan, J., Wu, W., Xie, C., Zhao, G., Zhao, Y., and Chen, R. (2013). NPInter v2.0: an updated database of ncRNA interactions. *Nucleic Acids Res.* 42, D104–D108. doi: 10.1093/nar/gkt1057

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.neucom.2017.07.065

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

Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S. (2013). "KDD '13 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining" in *Collaborative Matrix Factorization with Multiple Similarities for Predicting Drug-Target Interactions*. ACM, 1025–1033. doi: 10.1145/2487575.2487670

Zheng, X., Wang, Y., Tian, K., Zhou, J., Guan, J., Luo, L., et al. (2017). Fusing multiple protein-protein similarity networks to effectively predict lncRNA-protein interactions. *BMC Bioinf.* 18, 420. doi: 10.1186/s12859-017-1819-1

Conflict of Interest: The 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.