GILPI: Graphlet Interaction - based lncRNA-Protein Interaction Prediction

Hong-Yi Zhang[1*], Yan Zhou[2]

[*1]Big data statistics specialization, Guizhou University of Finance and Economics, China-GuiYang
[2]Specialization in subject teaching (language), Guizhou Normal University, China-GuiYang
*Corresponding Author

Received: 01 July 2024/ Revised: 08 July 2024/ Accepted: 15 July 2024/ Published: 31-07-2024

Abstract— *Identification of lncRNA-protein interactions is important for understanding the biological functions and molecular mechanisms of lncRNAs. In this study, we proposed a computational model for predicting lncRNA-protein interactions based on Graphlet interactions to find potential LPIs (GILPI). First, five LPI datasets were collected. Second, vector features of lncRNAs and proteins were extracted from the sequence data by pyfeat and BioTriangle, respectively. Third, these features were subjected to Pearson's correlation coefficient to calculate the similarity between lncRNAs and the similarity between proteins. Fourth, the Jaccard similarity between lncRNAs and proteins was calculated based on the LPI network, and then the corresponding Pearson similarity and Jaccard similarity were taken as the average value of the final lncRNA-lncRNA similarity and protein-protein similarity to construct the network. Finally, lncRNA-protein classification prediction was performed on both networks. Comparing GILPI with five state-of-the-art LPI prediction methods through 5-fold cross-validation, the results show that the GILPI prediction model has strong LPI classification performance. The case studies show that there may be interactions between NONHSAT021830 and Q9H9S0, n385685 and Q07955, and NONHSAT098243 and P25490.The novelty of GILPI is that it integrates the two similarities to construct a network, and then utilizes Graphlet interactions on the network to directly and indirectly link the features to mine out potential features, thus greatly improving the performance of the model.*

Keywords— Graphlet interaction, Jaccard similarity, Pearson similarity, lncRNA-protein interaction.

I. INTRODUCTION

1.1 Motivation:

Long non-coding RNAs (lncRNAs) are transcripts composed of more than 200 nucleotides but lack coding capabilities [1]. lncRNAs play key roles in biological processes such as gene expression regulation, epigenetic regulation, and cell differentiation [2].

For example, HOXA-AS2 and SNHG12 in lncRNAs have been identified as potential therapeutic targets and biomarkers for human cancers [3]. DLEU1 is closely related to colorectal cancer through activation of KPNA3, the expression of HOTAIR is elevated in lung cancer, and ZFAS1 is closely related to the chemosensitivity of cervical cancer cells [4]. In summary, more and more experiments have confirmed that lncRNAs are tumor-related biomolecules. However, to date, the relationship between lncRNAs and known tumor suppressor entities remains largely elusive. There is evidence that lncRNAs exert their biological functions through binding to RNA-binding proteins. Therefore, identifying potential lncRNA-protein interactions (LPIs) contributes to understanding many important biological processes and the treatment of various complex diseases.

1.2 Related Work:

Identifying lncRNA-protein interactions (LPIs) generally adopts two methods: experimental methods and computational methods. In experimental methods, biologists initially detect lncRNA-protein interactions through bioexperiments, such as RNA pulldown [5], RNA binding protein immunoprecipitation (RIP) [6], etc. However, this method is time-consuming and wasteful of resources. Gradually, people explore potential LPIs with computational methods, mainly divided into machine learning-based methods and network-based methods.

Machine learning-based methods mainly describe lncRNA-protein pairs by selecting features of lncRNAs and proteins, and use the extracted features as input to train a supervised learning model to identify potential LPIs. Liu et al.[7], Zhang et al.[8], Ma et al.[9] explored the neighborhood regularized logistic matrix decomposition method, graph regularized nonnegative matrix factorization model, and projection-based neighborhood nonnegative matrix factorization method (PMKDN), respectively.

Network-based methods usually construct some associated networks of lncRNAs or proteins, and then design a network algorithm to calculate the probability or score of interaction between lncRNAs and proteins. Zhao et al.[10] and Ge et al.[11] designed two recommendation algorithms based on bipartite networks to score each lncRNA-protein pair. Jia[12] et al. proposed a multifeature fusion method based on linear neighborhood propagation to calculate the linear neighborhood similarity of feature space and predict the results through label propagation.

Computational methods can effectively discover many potential relationships between lncRNAs and proteins. However, most machine learning-based LPI prediction methods are measured on a single dataset, which may lead to prediction bias. Secondly, cross-validation is performed in the case of lncRNA-protein pairs, ignoring the performance under other cross-validations. Network-based methods cannot find possible potential associated proteins or lncRNAs for a single lncRNA or protein.

1.3 Research Contributions

In this paper, we developed a network-based LPI prediction model, GILPI, to predict the interaction relationships between lncRNAs and proteins. The GILPI model integrates the bioinformatics of lncRNAs and proteins, Pearson similarity, Jaccard similarity, and Graphlet interactions into a unified prediction framework to identify potential LPIs. The main contributions of this work are as follows:

1) It reasonably integrates a variety of biological characteristics of lncRNAs and proteins, including 13 types for lncRNAs and 14 types for proteins, enabling a more effective description of lncRNA-protein pairs.
2) It creates networks of lncRNAs and proteins composed of Pearson similarity and Jaccard similarity, and utilizes Graphlet interactions on these networks to classify and predict unknown lncRNA-protein pairs.
3) By leveraging the direct and indirect connections of Graphlet interactions, it deeply mines the hidden features between lncRNA-protein pairs, thereby enhancing the predictive performance of GILPI.

II. MATERIALS AND METHODS

2.1 Data Preparation:

2.1.1 Dataset Acquisition:

In this paper, we have compiled five datasets related to LPI. Datasets 1, 2, and 3 contain human LPI data, while Datasets 4 and 5 contain plant LPI data. Dataset 1 is provided by Li et al. [13]. After removing lncRNAs and proteins with unknown sequence information from NPInter[14], NONCODE[15], and UniProt[16], we obtained 3,479 known associations from 935 lncRNAs and 59 proteins. Dataset 2 was constructed by Zheng et al. [17]. After similar preprocessing to Dataset 1, we filtered out 3,265 known associations from 885 lncRNAs and 84 proteins. Dataset 3 was constructed by Zhang et al. [18]. and contains 4,158 interactions from 990 lncRNAs and 27 proteins. Datasets 4 and 5 are from Arabidopsis thaliana and maize, respectively. The former contains 948 interactions from 109 lncRNAs and 35 proteins, while the latter contains 22,133 associations from 1,704 lncRNAs and 42 proteins. The sequence data was extracted from the PlncRNADB database [19], and the interaction data was obtained from http://bis.zju.edu.cn/PlncRNADB/. The five data details are shown in Table 1:

TABLE 1
LPI DATA

Dataset	lncRNAs	Protein	LPIs
Data1	935	59	3479
Data2	885	84	3265
Data3	990	27	4158
Data4	109	35	948
Data5	1704	42	22133

We represent the LPI network as a matrix Y, where elements contain:

$$y(i,j) = \begin{cases} 1, & \text{If lncRNA interacts with protein} \\ 0, & \text{Other} \end{cases} \quad (1)$$

2.1.2 Feature of lncRNAs:

After obtaining the sequence information for the five datasets, we selected 13 features to describe lncRNAs, which are as follows: zCurve, gcContent, atgcRatio, cumulativeSkew, pseudoKNC, monoMonoKGap, mono-DiKGap, monoTriKGap, diMo-noKGap, diDiKGap, diTriK-Gap, triMonoKGap, and tri-DiKGap. The corresponding features were extracted using the Pyfeat [20] Python tool, resulting in a 14,892-dimensional vector.

2.1.3 Feature of Proteins

To describe the biological information of proteins, we selected 14 features, which are as follows: amino acid composition, dipeptide composition, tri-peptide composition, CTD composition, CTD transition, CTD distribution, M-B autocorrelation, Moran autocorrelation, Geary autocorrelation, conjoint triad features, quasi-sequence order descriptors, sequence order coupling number, pseudo amino acid composition 1, and pseudo amino acid composi-tion 2. Features generated by BioTriangle [21] can effectively distinguish the captured amino acid information. In this study, we utilized the BioTriangle software to extract protein features, resulting in a 10,029-dimensional vector.

2.2 Overview of GILPI:

In this study, we created a framework for the LPI prediction model GILPI that integrates Pearson similarity, Jaccard similarity, and Graphlet interactions to classify unknown lncRNA-protein pairs. The following figure describes the GILPI framework.

In Fig. 1, the lncRNA-lncRNA Pearson similarity network, protein-protein Pearson similarity network, lncRNA-lncRNA Jaccard similarity network, protein-protein Jaccard similarity network were obtained after putting the lncRNA vectors and the protein vectors through the Pearson similarity and Jaccard similarity calculations. Then the lncRNA-lncRNA similarity network was constructed by adding the lncRNA-lncRNA Pearson similarity and lncRNA-lncRNA Jaccard similarity and taking the mean value, respectively. The protein-protein similarity network was constructed after summing protein-protein Pearson similarity, protein-protein Jaccard similarity and taking the mean value.

Next, the number of Graphlets is traversed on the lncRNA-lncRNA similarity network and the protein-protein similarity network to train the model. This process yields the weight coefficient V_L for the lncRNA similarity network and the weight coefficient V_P for the protein similarity network. Subsequently, the scores for the test set and the candidate set are calculated to determine the relationships between lncRNAs and proteins.

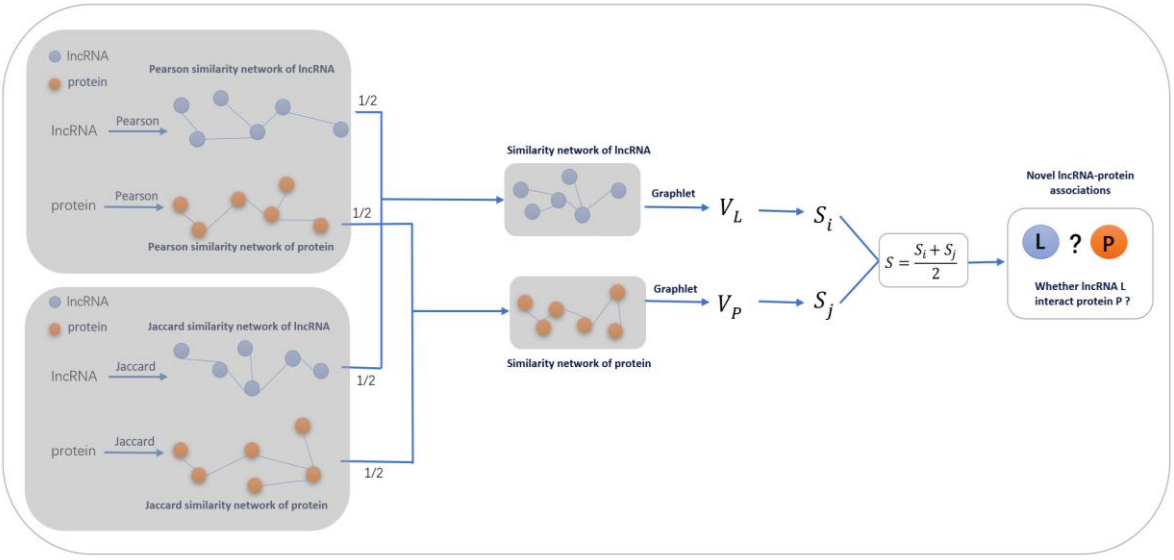

FIGURE 1: Framework of GILPI

2.3 Network Construction

2.3.1 Construction of lncRNA-lncRNA Pearson Similarity Network

We used the 14,892-dimensional vectors extracted with the Pyfeat Python tool to calculate the Pearson similarity between lncRNAs using the Pearson correlation coefficient. This resulted in an lncRNA-lncRNA Pearson similarity network. The formula is as follows:

$$\rho_{x,x_1} = \frac{cov(x,x_1)}{\sigma_x \sigma x_1} \tag{2}$$

Where x and $x_1$represent different lncRNAs, $cov(x, x_1)$ is the covariance between two lncRNAs, and $\sigma_x \sigma x_1$ are the standard deviations of the two lncRNAs. The value of ρ_{x,x_1} ranges from -1 to 1, with values less than 0 indicating negative correlation and values greater than 0 indicating positive correlation. The Pearson similarity network for lncRNAs in the five datasets was calculated using this formula.

2.3.2 Protein-protein Pearson similarity network construction:

In order to construct the Pearson similarity network between proteins, we use 10029-dimensional vectors extracted by BioTriangle software and also go through the Pearson correlation coefficient to calculate the Pearson similarity between two proteins two by two to get the protein-protein Pearson similarity network. The formula is the same as shown in (3) above.Just where x and $x_1$denote different proteins respectively, $cov(x, x_1)$ is the covariance between two proteins, andρ_{x,x_1} is the standard deviation between two proteins.

2.3.3 Calculation of lncRNAs and protein Jaccard similarity:

In order to fully explore the biological properties of lncRNAs and proteins, this paper not only used the Pearson correlation coefficient to calculate the similarity between lncRNAs and lncRNAs and between proteins and proteins, but also introduced the Jaccard similarity to measure the relationship between lncRNAs and lncRNAs and between proteins and proteins.

Jaccard similarity is a popular approximation metric for calculating the similarity between two objects. It can be used to find the similarity between two asymmetric binomial vectors or to find the similarity between two sets. Jaccard coefficient is usually used between texts that are sequence order insensitive. The higher the value of Jaccard coefficient, the more similar the samples are. In LPI network, based on known lncRNAs and proteins, we calculated lncRNA-lncRNA similarity and protein-protein similarity by using Jaccard similarity principle. jaccard coefficient is defined as the size of intersection of the sample sets divided by the size of the merged set. For example, L_i and $L_j$are two lncRNA datasets. The Jaccard similarity between any two sets of lncRNAs is calculated as follows:

$$J(L_i, \ L_j) = \frac{|L_i \cap L_j|}{|L_i U L_j|} \tag{3}$$

The Jaccard similarity of proteins was calculated identically to lncRNA.

2.3.4 lncRNAs and protein similarity network construction:

The lncRNA-lncRNA Pearson similarity and protein-protein Pearson similarity were obtained by Pearson correlation calculation, and the lncRNA-lncRNA Jaccard similarity and protein-protein Jaccard similarity were obtained by Jaccard calculation, and then the lncRNA-lncRNA Pearson similarity was added, lncRNA-lncRNA Jaccard similarity and protein-protein Pearson similarity, protein-protein Jaccard similarity were averaged after addition to get the final lncRNA-lncRNA similarity network, protein-protein similarity network.

2.4 Introduction to Graphlet and Graphlet Interaction:

Graphlets are small non-isomorphic connected subgraphs, and a complete large network is composed of Graphlets. In this paper, we only consider Graphlets with no more than 4 nodes, as shown in Fig2 below. In the figure, from G1 to G9 are the 9 types of the corresponding Graphlet, the nodes in the Graphlet occupy different positions called self-isomorphic orbits, and the nodes on the same self-isomorphic orbit have the same local topological features in the Graphlet, and there are 15 self-isomorphic orbits for these 9 types of Graphlets.

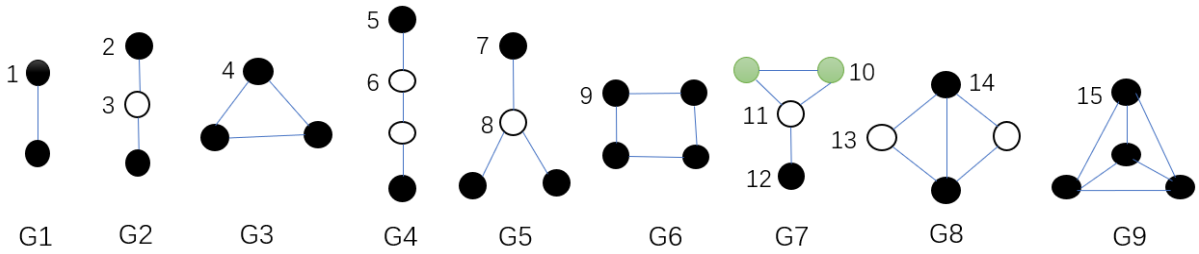

FIGURE 2: Graphlet diagram

Graphlet interaction describes the relationship between 2 nodes. There is a Graphlet interaction between two nodes in the same Graphlet, when there is a Graphlet interaction between node i and node j of graph H, the following equation is satisfied:

$$\exists G \subseteq H, \quad and \ \mathrm{i} \in G, \quad j \in G \tag{4}$$

Where G is a Graphlet in the graph H and V (G) is the set of nodes of G.

In Fig 3 below, the black and light green nodes represent nodes i and j with Graphlet interactions. therefore, different types of relationships exist between two nodes based on their different self-isomorphic orbits. Different types of relationships between two nodes are called Graphlet interaction isomers. For example, Graphlet interaction isomers I_2、$I_3$and I_4. nodes i and j are in different self-isomorphic orbits and are viewed as different Graphlet interaction isomers. graphlet interactions are a vector, where each element denotes the number of corresponding Graphlet interaction isomers. graphlet interaction vector has 28 elements corresponding to 28 Graphlet interaction isomers. In this paper, only Graphlet interactions with no more than 4 nodes are considered, and there are a total of 28 Graphlet interaction isomers, labeled $I_1$to I_{28}.

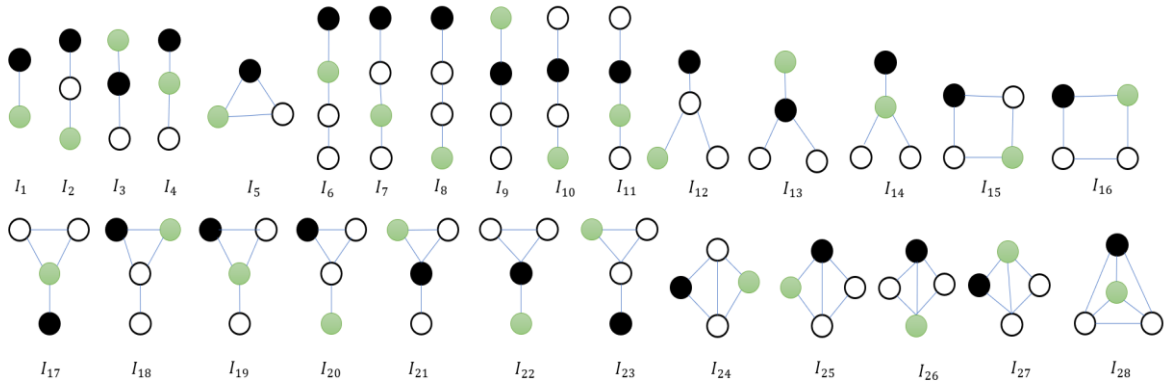

FIGURE 3: Graphlet interactions

2.5 Graphlet interaction computation:

Denote the graph H by the adjacency matrix A= (a_{ij}) . in the graph, a_{ij}=1 if there is an edge between node i and node j, and a_{ij}=0 if there is no edge connecting node i and node j. In the calculation of Graphlet interactions between nodes i and j, the number of isomers I_k is calculated as follows in Eq:

$$N_{ij}(I_k) = \sum_{l \in V(G)} \sum_{m \in V(G)} b_{ij} b_{il} b_{im} b_{jl} b_{jm} b_{lm} \tag{5}$$

To make the above equation clearer, b is a variable and is calculated as follows:

$$b(i,j) = \begin{cases} a_{st}, & \text{s and t are linked in } I_k \\ 1 - a_{st}, & \text{s and t are not linked in } I_k \end{cases} \tag{6}$$

In the above equation, $N_{ij}(I_k)$denotes the number of isomers $I_k$between nodes i and j. l and m denote the other 2 nodes other than nodes i and j. i, j, l, and m are unequal. The above equation calculates the total number of isomers from node i to j. The higher the number of isomers I_k, the closer the relationship between node i and node j is indicated.

Since formula (5) is time consuming to compute, plus in practice isomers are computed from vectors of adjacency matrices such as a_i and a_j, so formula (5) can be rewritten as:

$$N_{ij}(I_k) = a_i * a_j \tag{7}$$

where $*$ denotes the inner product of the two vectors a_i and a_j.

Graphlet has directionality. When calculating the Graphlet interactions of two nodes i and j, the Graphlet interactions from node i to node j are not equal to the Graphlet interactions from node j to node i. The Graphlets have symmetry. However, Graphlets have symmetry, such asI_3 and I_4, I_{17} and I_{22} in Fig3. where $N_{ij}(I_3)$=$N_{ij}(I_4)$ means that the 3rd element of the Graphlet interaction vector from node i to node j is equal to the 4th element of the Graphlet interaction vector from node j to node i.

2.6 Sorting LPIs for unknown associations based on Graphlet interaction scores

Graphlet interactions were used to categorize LPIs, and PLIs with unknown associations were sorted based on Graphlet interaction scores. The higher the score, the more closely the LPI in which the lncRNA is likely to be related to the protein. Below is the formula for calculating the Graphlet interaction score in the protein-protein similarity network:

$$S_j = \sum_k v_k \sum_{i \in P} norm\left(N_{ij}(I_k)\right) \tag{8}$$

In the above equation, S_j denotes the protein-to-protein fraction in the network, P denotes the set of points with known associations for a particular class of proteins, v_k denotes the corresponding weight coefficients, and $norm\left(N_{ij}(I_k)\right)$ denotes the Graphlet interactions normalized from node i to node j. The normalization formula is as follows:

$$norm\left(N_{ij}(I_k)\right) = \frac{N_{ij}(I_k)}{N_i(I_k)} \tag{9}$$

Where $N_{ij}(I_k)$ denotes the number of Graphlet interaction isomersI_k between node i to node j, and $N_i(I_k)$ is the total number of Graphlet interaction isomers I_k between node i to all other nodes. $N_i(I_k)$ is calculated as follows:

$$N_i(I_k) = \sum_{j \in C} N_{ij}(I_k) \tag{10}$$

Where C denotes the set of unknown associations of a certain protein.

The weight v_k in Eq. (8), we use linear regression to calculate. In order to validate the performance of the proposed algorithmic model in this paper, we divide the data into training set and test set. The training set is put through regression to get the weights, and then the test set is used to validate the algorithmic model.

Rewrite equation (8) as:

$$S_j = \sum_k v_k x_{jk} \tag{11}$$

where x_{jk} is calculated by the following equation:

$$x_{jk} = \sum_{i \in P} norm\left(N_{ij}(I_k)\right) \tag{12}$$

At the time of training data, S_j and x_{jk} in Eq. (11) are known, so it is possible to calculate v_k, which is given below:

$$V = (XX^T)^{-1}XS \tag{13}$$

Similarly, in the lncRNA network, the lncRNA-to-lncRNA Graphlet interaction score is given by:

$$S_i = \sum_k v_k \sum_{j \in R} norm\left(N_{ij}(I_k)\right) \tag{14}$$

In Eq. S_i denotes the lncRNAs to lncRNAs score, R denotes the set of points with known associations for a particular type of lncRNA, v_k denotes the corresponding weight coefficients, and $norm\left(N_{ij}(I_k)\right)$ denotes the node i to node j normalized Graphlet interaction. Similarly, Eq. (14) can be rewritten as:

$$S_i = \sum_k v_k x_{ik} \tag{15}$$

The following equation calculates x_{ik}:

$$x_{ik} = \sum_{j \in R} norm\left(N_{ij}(I_k)\right) \tag{16}$$

Finally, the protein-to-protein fraction S_j and the lncRNA-to-lncRNA fraction S_i were used to take the mean value as the calculated protein-to-lncRNA fraction S, calculated as follows:

$$S = \frac{S_i + S_j}{2} \tag{17}$$

III. RESULTS

3.1 Performance Evaluation

To evaluate the performance of the GILPI model, we use five-fold cross-validation to rank the test samples and candidate samples on each of the five datasets, calculate the AUC and AUPR, and repeat the experiment 10 times. First, inside the LPI matrix, which contains known association part 1 and unknown association part 0, the known association part is randomly disrupted and then divided into 5 parts, where the number of data in the last part is slightly less than that in the remaining 4 parts, and the data between every two parts are not repeated. Then 1 part is selected as the test set, and the remaining 4 parts are used as the training set, and so on, until each part of the data is used as the test set and the training set. Similarly, we take the unknown lncRNA-protein as a candidate sample, and then calculate the scores of the test sample and the candidate sample. We compare the score of each test sample with the score of the candidate sample in turn. The prediction is considered successful only when the rank of the test sample exceeds a given threshold.

The AUC values were then calculated by calculating the true positive rate TPR (sensitivity) and false positive rate FPR (specificity) for different thresholds, where sensitivity refers to the percentage of test samples above a given threshold that are positive cases and specificity refers to the percentage of pseudo-cases of lncRNA-protein associations that are below a given threshold. AUC=1 indicates that the model correctly predicted all test samples. AUC=0.5 indicates that the model is randomly predicted. AUPR refers to the area enclosed by the precision and recall versus PR curves. In these two metrics, the higher the value, the better the performance of the GILPI model, and the average value is taken as the final evaluation criterion after repeating the experiments for 10 times. The values of AUC and AUPR calculated by repeating the experiments 10 times for the GILPI model are shown in Table 2:

TABLE 2
AUC AND AUPR VALUES CORRESPONDING TO THE 5 DATA SETS

Dataset	AUC	AUPR
Data1	0.9477	0.9349
Data2	0.9496	0.9305
Data3	0.8986	0.8867
Data4	0.9706	0.8205
Data5	0.9757	0.9715
Ave.	0.9484	0.9088

As can be seen from Table 2, except for dataset 3, the AUC values of the rest of the data are all above 0.9, and dataset 5 is as high as 0.9757. Not only that, the AUPR values are also very good, except for datasets 3 and 4, which are all above 0.9, and dataset 5 is as high as 0.9715. On the whole, dataset 5 has the highest AUC and AUPR values of all datasets, and dataset 3 has the lowest AUC and AUPR values among all datasets. of all the datasets, and dataset 3 has the lowest AUC and AUPR values among all the datasets. Overall, the AUC and AUPR values of the five datasets perform very well, indicating that the GILPI model has good prediction performance.

Further, we compare the model proposed in this paper with five advanced LPI prediction models. These five models are LPI-deepGBDT [22], LPI-DLDN [23], LPI-EnANNDeep [24], LPI-EnEDT [25], and LPI-HyADBS [26], in order to measure the classification ability of the GILPI model. Their comparison results are shown in Table 3 below:

TABLE 3
COMPARISON OF AUC VALUES FOR THE 6 MODELS

Dataset	GILPI	LPI-deepGBDT	LPI-DLDN	LPI-EnANNDeep	LPI-EnEDT	LPI-HyADBS
data1	0.9477	0.9354	0.9404	0.9473	0.9297	**0.9488**
data2	0.9496	0.9423	0.9447	0.9556	0.9474	**0.9583**
data3	**0.8986**	0.8526	0.8301	0.8597	0.8235	0.8593
data4	**0.9706**	0.8542	0.9099	0.8648	0.8866	0.9162
data5	**0.9757**	0.9523	0.9302	0.9557	0.9458	0.9672
Ave.	**0.9484**	0.9074	0.9111	0.9166	0.9066	0.93

In Table 3, the GILPI model is the model proposed in this paper, and the black bolded ones are the highest values in each dataset. From the table, it can be seen that the GILPI model has the highest average AUC value of 0.9484, which was higher than LPI-deepGBDT, LPI-DLDN, LPI-EnANNDeep, LPI-EnEDT, and LPI-HyADBS by 4.51%, 4.09%, 3.46%, 4.61%, and 1.98%, respectively. The AUCs of the GILPI model are the highest in Data 3 through Data 5, with an AUC value of 0.9757 in Data 5. In Data 1 through Data 2, the AUCs of the GILPI model are slightly lower compared to the other models at 0.9477 and 0.9496, which are only 0.11% and 0.92% lower than the highest at 0.9488 and 0.9583, but most of them are higher than the other models.

TABLE 4
COMPARISON OF AUPR VALUES FOR THE 6 MODELS

Dataset	GILPI	LPI-deepGBDT	LPI-DLDN	LPI-EnANNDeep	LPI-EnEDT	LPI-HyADBS
data1	**0.9349**	0.9043	0.9282	0.9283	0.9001	0.93
data2	0.9305	0.9242	0.9292	0.9408	0.9262	**0.9423**
data3	**0.8867**	0.8016	0.8099	0.8356	0.8005	0.8354
data4	0.8205	0.8488	0.9001	0.8683	0.8767	**0.9098**
data5	**0.9715**	0.9457	0.9246	0.954	0.9374	0.9653
Ave.	0.9088	0.8849	0.8984	0.9054	0.8882	**0.9166**

As can be seen from Table 4, the AUPR values of the GILPI model are the highest in datasets 1, 3, and 5, which are 0.9349, 0.8867, and 0.9715, respectively. with a high of 0.9715 for dataset 5, which is higher than the AUPR values of LPI-deepGBDT, LPI-DLDN, LPI-EnANNDeep, LPI-EnEDT, and LPI- HyADBS by 2.72%, 5.07%, 1.83%, 3.64%, and 0.64%. In Data 2 and Data 4, the AUPR of the GILPI model is slightly lower 0.9305 and 0.8205. In Data 2, it is only 1.27% lower than the highest 0.9423, but all of them are higher than the LPI-deepGBDT , LPI-DLDN, and LPI-EnEDT models. In Data 5, it is only 2% lower than the highest 0.9653, but all higher than the LPI-deepGBDT, LPI-DLDN, and LPI-EnEDT models.

The five LPI prediction methods, LPI-deepGBDT, LPI-DLDN, LPI-EnANNDeep, LPI-EnEDT, and LPI-HyADBS, are the most advanced and classical prediction models, however, the GILPI model proposed in this paper is far better than these five models. The comparative results show that the GILPI model has a powerful classification performance and is capable of mining the potential interactions between lncRNAs and proteins.

3.2 Case Studies

3.2.1 Discovery of proteins that interact with novel lncRNAs:

lncRNAs are a class of long chain RNA molecules that do not code for proteins, and he plays an important role in a variety of biological processes such as gene expression regulation and cell differentiation. In this paper, three lncRNAs, NONHSAT021830, n385685, and NONHSAT098243, which interact with 15, 16, and 19 proteins, respectively, were selected from the human dataset. In order to find out the proteins interacting with these three lncRNAs, the interaction information between the proteins associated with these three lncRNAs is masked out and these three lncRNAs are taken as the new lncRNAs in the neighboring matrix Y. Then the potential proteins are found with the GILPI modeling algorithm proposed in this paper, and the top 5 proteins predicted are shown in Table 5 below. It can be found that a total of 6 proteins were confirmed in the three datasets. Among them, in Data 1, Data 2 and Data 3, NONHSAT021830 with Q9H9S0, n385685 with Q07955, and NONHSAT098243 with P25490 were not confirmed, but they were ranked first, indicating that these three pairs of lncRNAs and proteins are likely to be associated with one another, but this is only a speculation, which needs to be further biological experiments to prove it. In summary, these results reconfirm the classification performance of GILPI. Therefore, GILPI is suitable for predicting proteins interacting with novel lncRNAs.

TABLE 5
PROTEINS INTERACTING WITH NEW LNCRNAS

Dataset	lncRNA	Protein	Confirmed	GILPI
Data1	NONHSAT021830	Q9H9S0	No	1
		P48431	No	2
		Q12968	No	3
		Q5S007	No	4
		Q8NDV7	Yes	5
Data2	n385685	Q07955	No	1
		Q9UKV8	Yes	2
		Q9UPQ9	Yes	3
		Q9HCJ0	Yes	4
		Q8NDV7	Yes	5
Data3	NONHSAT098243	P25490	No	1
		Q13285	No	2
		P60484	No	3
		Q96PU8	Yes	4
		O43251	No	5

3.2.2 Discovery of lncRNAs that interact with novel proteins:

Proteins are extremely important macromolecules in living organisms, which play key roles in physiological activities such as signaling, immune defense, cell growth and differentiation. In this paper, three proteins, O00425, Q9Y6M1 and O00425, were selected from three human datasets, which interacted with 443, 342, and 463 lncRNAs in dataset 1, dataset 2, and dataset 3, respectively. In order to find out the lncRNAs interacting with these 3 proteins, all the interaction information between the lncRNAs associated with these 3 proteins are masked out in the neighbor-joining matrix Y, and these 3 proteins are treated as new proteins to discover the potential lncRNAs with the GILPI modeling algorithm proposed in this paper. the top 5 predicted lncRNAs are shown in Table 6 below. It can be found that most of the lncRNAs were confirmed in the 3 datasets.

In dataset 1, O00425 was not confirmed with NONHSAT112460, but its ranking was 1st, indicating that O00425 and NONHSAT112460 are greatly likely to interact, but further biological proof is needed. Overall, GILPI can be used for LPI prediction of new proteins.

TABLE 6
lncRNAs THAT INTERACT WITH NOVEL PROTEINS

Dataset	Protein	lncRNA	Confirmed	GILPI
Data1	O00425	NONHSAT112460	No	1
		NONHSAT008249	No	2
		NONHSAT052575	No	3
		NONHSAT112472	No	4
		NONHSAT066972	Yes	5
Data2	Q9Y6M1	n338605	Yes	1
		n377669	Yes	2
		n345648	Yes	3
		n381041	Yes	4
		n342241	Yes	5
Data3	O00425	NONHSAT016408	Yes	1
		NONHSAT093392	No	2
		NONHSAT124481	Yes	3
		NONHSAT041141	Yes	4
		NONHSAT025390	Yes	5

3.2.3 Finding new LPIs based on known LPIs:

Immediately after that, based on the known LPIs, we use the model GILPI proposed in this paper to discover new LPIs. the top 50 lncRNA-protein pairs with the highest scores on the five datasets are filtered as shown below, where the circle represents the lncRNA, the hexagon represents the protein, the ones with known associations are connected by a solid line, the ones with unknown associations are connected by a dashed line, and the ones connected by a light blue color with a light green color are with known associations, and yellow and light green are connected with unknown associations, and these top 50 contain lncRNA-protein pairs with known associations and unknown associations.

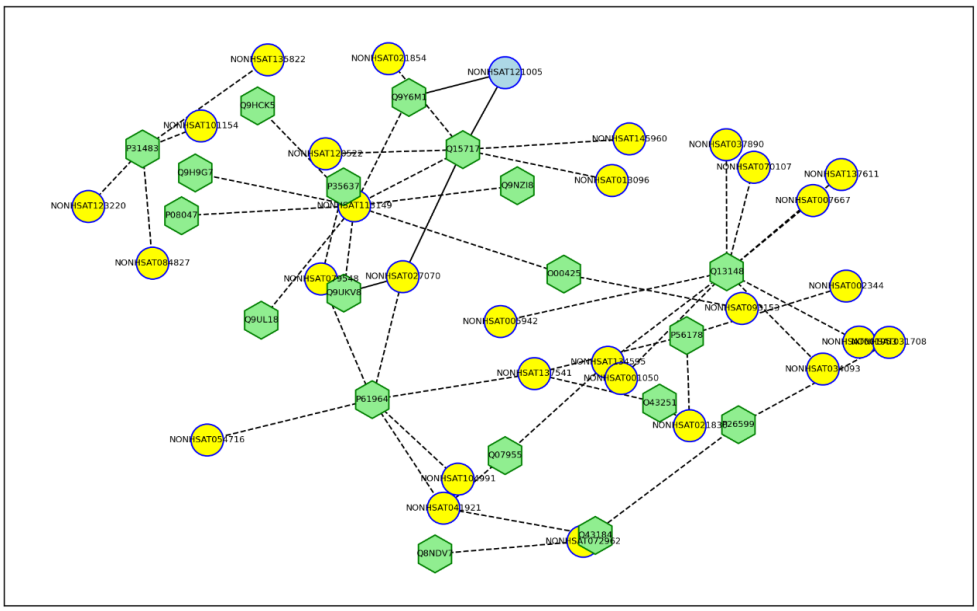

FIGURE 4: Top 50 lncRNA-protein pairs with the highest scores in Data 1

In Data 1, there are a total of 55,165 lncRNA-protein pairs. In the calculated top 50, there are a total of 4 lncRNA-protein pairs with known associations and 46 pairs with unknown associations, e.g., NONHSAT113149 is associated with Q15717 and NONHSAT137541 is associated with P61964 with unknown associations, but these two lncRNA-protein pairs are ranked as

the 2nd and 3rd out of 55165 pairs, so there is a high probability that they are interacting, but this needs to be proved by further biological experiments.

In Data 2, there are a total of 74,340 lncRNA-protein pairs, and there are 44 pairs of unknown associations in the top 50 calculated as n385685 with O14746, n385685 with O95793, etc., but they are all ranked very high, and even the 2 pairs of unknown associations, n385685 with O14746 and n385685 with O95793, ranked 1st and 3rd, respectively, inside the 74340 pairs ranked 1st and 3rd respectively, so there is a great possibility that these 2 pairs of lncRNAs and proteins are interacting. In the 6 known association pairs, their rankings are 2, 14, 18, 29, 45, 46 respectively.

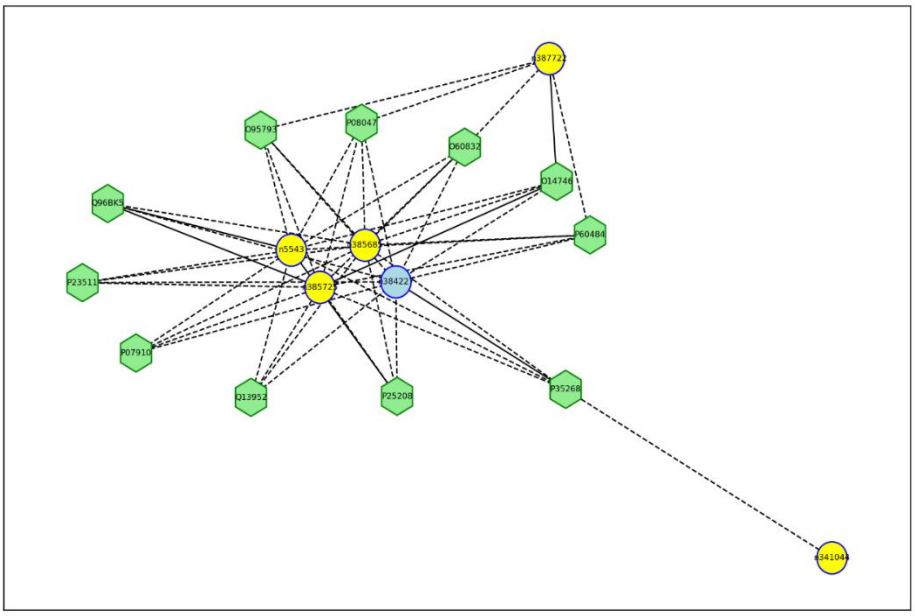

FIGURE 5: Top 50 lncRNA-protein pairs with the highest scores in Data 2

In Data 3, there are a total of 26,730 pairs of lncRNAs and proteins, with 4 pairs of known associations and 46 pairs of unknown associations in the top 50 calculated. Although the number of unknown associations is high, they are all ranked highly. For example, NONHSAT121765 with P60484 and NONHSAT121765 with O43251 are ranked 1st and 4th, respectively, for these unknown associations, and there may be interactions between them, but this is only a guess, and further experiments are needed to prove it.

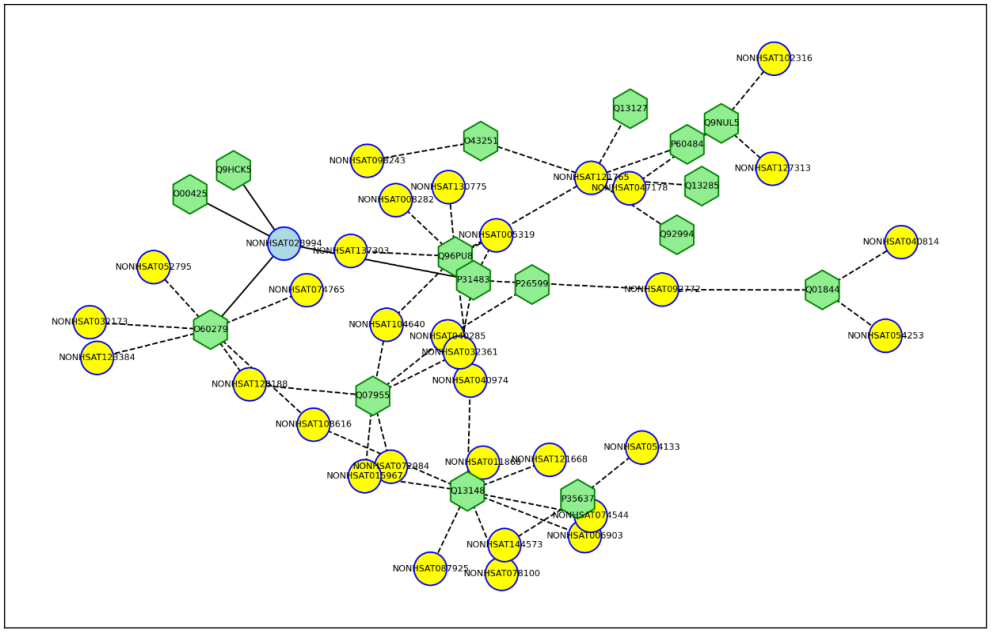

FIGURE 6: Top 50 lncRNA-protein pairs with the highest scores in Data 3

In Data 4, there are a total of 3,815 lncRNA-protein pairs, and among the top 50 calculated pairs, there are 30 pairs with known associations and 20 pairs with unknown associations, for example, the 3 unknown associations of AthlncRNA229 with 15229884, AthlncRNA227 with 79326195, and AthlncRNA302 with 18423684 , ranked 8th, 11th and 15th inside all 3815 lncRNA-protein pairs, suggesting that there may be interactions between these lncRNA-protein pairs.

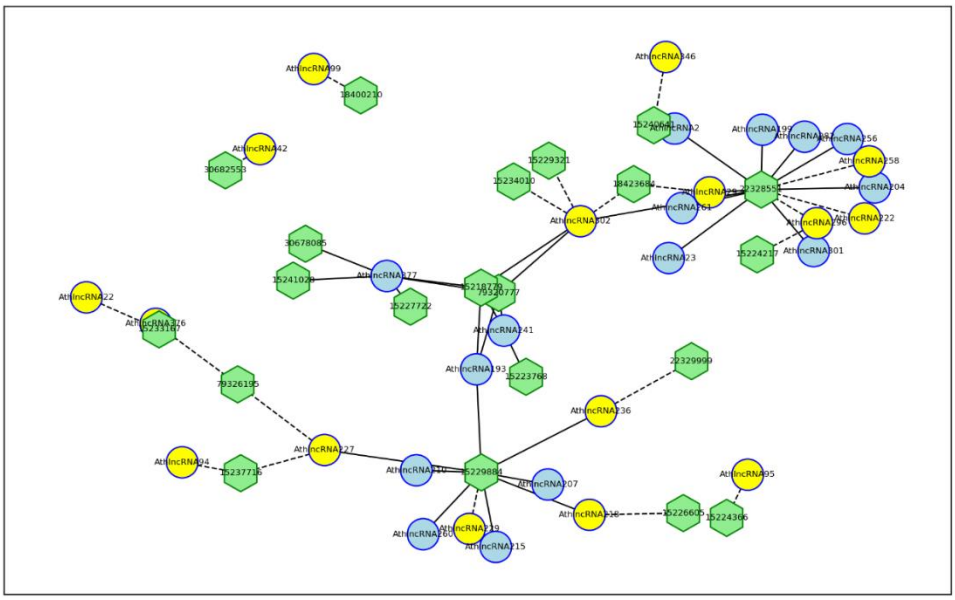

FIGURE 7: Top 50 lncRNA-protein pairs with the highest scores in Data 4

In dataset 5, there are a total of 22,133 pairs of lncRNAs with proteins, which is the most in the five datasets. In the top 50 calculated pairs, there are 24 pairs with known associations and 26 pairs with unknown associations. The known associations are basically ranked at the top of the 50 pairs, and the unknown associations are ranked a little bit later. However, the ranking of unknown associations is also very good in all the 22,133 pairs. For example, ZmalncRNA1062 with B4FLX0 and ZmalncRNA1263 with C0PF88 ranked 9th and 15th respectively, so these two pairs and other unknown associations of lncRNAs and proteins in the 50 pairs, they may interact with each other.

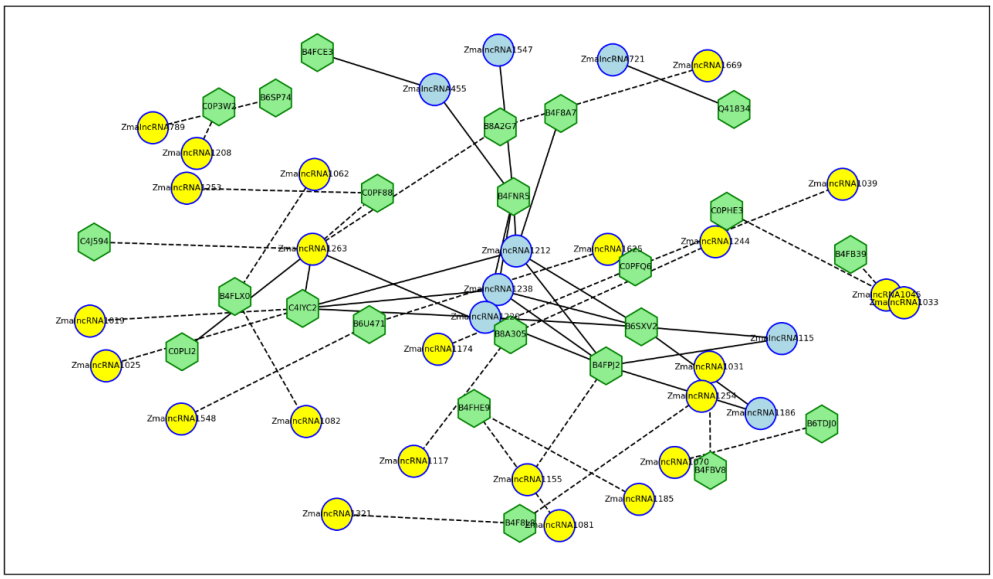

FIGURE 8: Top 50 lncRNA-protein pairs with the highest scores in Data 5

IV. DISCUSSION

Characterization of lncRNA-protein interaction relationships helps to discover the function and mechanism of action of lncRNAs. In this paper, we developed a prediction model GILPI incorporating Pearson similarity, Jaccard similarity to classify lncRNA-protein interaction relationships. The experiment was repeated 10 times to train the model using five-fold cross-

validation and compared with other state-of-the-art LPI prediction models. The experimental results show that the GILPI prediction model proposed in this paper is able to classify lncRNA-protein interaction relationships more accurately and can be used to discover new LPIs.

Under the five-fold cross-validation, most of the performances of the five prediction models, LPI-deepGBDT, LPI-DLDN, LPI-EnANNDeep, LPI-EnEDT, and LPI-HyADBS, are much lower than that of the GILPI model proposed in this paper. For training, among the five data, after randomly disrupting the known associations, 80% is selected as the training set, 20% as the test set, and the remaining unknown associations as the candidate samples, and then the test set and the candidate samples are scored and ranked. In addition, it is further shown in the case study that the GILPI prediction model proposed in this paper can mine useful information for new lncRNAs or new proteins.

The GILPI prediction model proposed in this paper demonstrates a powerful LPI classification capability. It incorporates Pearson similarity and Jaccard similarity to fully mine the complex biological information between lncRNA-protein, and then utilizes the characteristics of Graphlet interaction direct connection and indirect connection on lncRNA-protein network to deeply mine the hidden features between lncRNA and protein. It greatly enriches the features when the model is trained, and makes the prediction performance of the model more accurate and powerful. Although the GILPI model can accurately identify new LPIs, it also has the following problems: one is that the network-based method has a defect that it cannot predict separate lncRNAs and proteins, so the GILPI model proposed in this paper can not predict single lncRNAs and proteins. Second, the Graphlet interactions used in this paper have the number of nodes within 4 nodes, so the information beyond 4 nodes is ignored, resulting in insufficiently rich training features obtained. Third, the time complexity of this model is high. It takes a long time for the model to run once, and repeating the experiment 10 times in this paper takes a lot of time.

V. CONCLUSIONS

lncRNAs play a crucial role in many biological activities, such as gene transcription, translation and other processes. Not only that, lncRNAs also affect numerous diseases, so recognizing the lncRNA and protein interaction relationship can be a good grasp of the biological function of lncRNAs, which is important for the treatment of disease therapy, diagnosis and so on.

First, five datasets were collected; second, features of lncRNAs and proteins were extracted from the sequence data using pyfeat and BioTriangle, respectively. Third, these features were analyzed by Pearson's correlation coefficient to calculate the similarity between lncRNAs and the similarity between proteins. Fourth, the Jaccard similarity between lncRNAs and proteins was calculated based on the LPI network, and then the corresponding Pearson similarity and Jaccard similarity were averaged to construct the lncRNA-lncRNA similarity network and protein-protein similarity network. The experiment was repeated 10 times, and GILPI was compared with five state-of-the-art LPI prediction methods, namely, LPI-deepGBDT, LPI-DLDN, LPI-EnANNDeep, LPI-EnEDT, and LPI-HyADBS, and the results showed that the GILPI prediction model had a strong LPI classification performance.The GILPI prediction model in the case study also achieved good results.

In future studies, we will first integrate various lncRNA and protein related datasets from different data sources. Secondly, mining the secondary and tertiary structures of proteins fused into lncRNA-protein pairs makes it possible to predict the relationship between a single lncRNA-protein pair. Then secondly, other nodes than the four nodes are considered in Graphlet interactions to make the acquired features more complete and rich. Finally, the computational efficiency is optimized by utilizing high-performance computing resources such as GPU acceleration and distributed computing to reduce the time of a single run, developing more efficient algorithms to handle large-scale datasets with less computational redundancy, and optimizing and automating the tuning of the model parameters to reduce the time needed to manually adjust the parameters.

ACKNOWLEDGEMENTS

This work was supported by Science and Technology Planning Project of Guizhou Province of China (No. Qian Ke He Ji Chu -ZK[2021] Yi Ban 315), Science Foundation of Guizhou University of Finance and Economics (No. 2021KYYB21).

REFERENCES

[1] KHALIL A M, RINN J L. RNA–protein interactions in human health and disease [J]. Seminars in Cell & Developmental Biology, 2011, 22(4): 359-65.

[2] Tiwari A, Srivastava R. A survey of computational intelligence techniques in protein function prediction. Int J Proteomics. 2014;2014: 845479.

[3] Tamang S, Acharya V, Roy D, Sharma R, Aryaa A, Sharma U, Khandelwal A, Prakash H, Vasquez KM, Jain A (2019) Snhg12: an lncRNA as a potential therapeutic target and biomarker for human cancer. Front Oncol 9:901.

https://doi.org/10.3389/fonc.2019.00901

[4] Mao Z, Li H, Du B, Cui K, Xing Y, Zhao X, Zai S (2017) LncRNA dancr promotes migration and invasion through suppression of lncRNA-let in gastric cancer cells. Biosci Rep. https://doi.org/10.1042/BSR20171070

[5] Batista PJ, Chang HY. Long noncoding RNAs: cellular address codes in development and disease. Cell. 2013;152(6):1298–307.

[6] Selth LA, Gilbert C, Svejstrup JQ. RNA immunoprecipitation to determine RNA–protein associations in vivo. Cold Spring Harbor Potoc. 2009;2009(6):pdb–prot5234.

[7] Liu H, Ren G, Hu H, Zhang L, Ai H, Zhang W, Zhao Q (2017) Lpi-nrlmf: lncrna-protein interaction prediction by neighborhood regularized logistic matrix factorization. Oncotarget. https://doi.org/10.18632/oncotarget.21934

[8] Zhang T, Wang M, Xi J, 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(1):189–197.

https://doi.org/10.1109/TCBB.2018.2861009

[9] Ma Y, He T, Jiang X (2019) Projection-based neighborhood nonnegative matrix factorization for lncRNA-protein interaction prediction. Front Genet 10:1148. https://doi.org/10.3389/fgene.2019.01148

[10] Zhao Q, Yu H, Ming Z, Hu H, Ren G, Liu H (2018) The bipartite network projection-recommended algorithm for predicting long non-coding RNA-protein interactions. Mol Therapy Nucleic Acids 13:464–471. https://doi.org/10.1016/j.omtn.2018.09.020

[11] Ge M, Li A, Wang M (2016) A bipartite network-based method for prediction of long non-coding RNA-protein interactions. Genom Proteom Bioinform 14(1):62–71. https://doi.org/10.1016/j.gpb.2016.01.004

[12] Jia L, Luan Y. Multi-feature Fusion Method Based on Linear Neighborhood Propagation Predict Plant LncRNA-Protein Interactions. Interdiscip Sci. 2022 Jun;14(2):545-554. doi: 10.1007/s12539-022-00501-7. Epub 2022 Jan 17. PMID: 35040094.

[13] Li A, Ge M, Zhang Y, Peng C, Wang M (2015) Predicting long noncoding RNA and protein interactions using heterogeneous network model. BioMed Res Int. https://doi.org/10.1155/2015/ 671950

[14] Yuan J, Wu W, Xie C, Zhao G, Zhao Y, Chen R (2014) Npinter v2. 0: an updated database of ncrna interactions. Nucleic Acids Res 42(D1):D104–D108. https://doi.org/10.1093/nar/gkt1057

[15] Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, Zhu W, Wu W, Chen R, Zhao Y (2014) Noncodev4: exploring the world of long noncoding RNA genes. Nucleic Acids Res 42(D1):D98–D103. https:// doi.org/10.1093/nar/gkt1222

[16] Consortium U (2019) Uniprot: a worldwide hub of protein knowledge. Nucleic Acids Res 47(D1):D506–D515. https://doi.org/10. 1093/nar/gky1049

[17] Zheng X, Wang Y, Tian K, Zhou J, Guan J, Luo L, Zhou S (2017) Fusing multiple protein-protein similarity networks to efectively predict lncRNA-protein interactions. BMC Bioinform 18(12):11– 18. https://doi.org/10.1186/s12859-017-1819-1

[18] Zhang W, Qu Q, Zhang Y, Wang W (2018) The linear neighborhood propagation method for predicting long non-coding RNAprotein interactions. Neurocomputing 273:526–534. https://doi. org/10.1016/j.neucom.2017.07.065

[19] Bai Y, Dai X, Ye T, Zhang P, Yan X, Gong X, Liang S, Chen M (2019) PLNCRNADB: a repository of plant LNCRNAS and LNCRNA-RBP protein interactions.Curr Bioinform 14(7):621–627. https://doi.org/10.2174/1574893614666190131161002

[20] R. Muhammod, S. Ahmed, D. Md Farid, S. Shatabda, A. Sharma, and A.Dehzangi, "PyFeat: A python-based effective feature generation tool for DNA, RNA and protein sequences," Bioinformatics, vol. 35, no. 19, pp. 3831–3833, 2019.

[21] J. Dong et al., "Biotriangle: A web-accessible platform for generating various molecular representations for chemicals, proteins, DNAs/RNAs and their interactions," J. Cheminformat., vol. 8,no. 1, pp. 1–13, 2016.

[22] Zhou L, Wang Z, Tian X, Peng L. LPI-deepGBDT: a multiple-layer deep framework based on gradient boosting decision trees for lncRNA-protein interaction identification. BMC Bioinformatics. 2021 Oct 4;22(1):479. doi: 10.1186/s12859-021-04399-8. PMID: 34607567; PMCID: PMC8489074.

[23] Peng L, Wang C, Tian X, Zhou L, Li K. Finding lncRNA-Protein Interactions Based on Deep Learning With Dual-Net Neural Architecture. IEEE/ACM Trans Comput Biol Bioinform. 2022 Nov-Dec;19(6):3456-3468. doi: 10.1109/TCBB.2021.3116232. Epub 2022 Dec 8. PMID: 34587091.

[24] Peng L, Tan J, Tian X, Zhou L. EnANNDeep: An Ensemble-based lncRNA-protein Interaction Prediction Framework with Adaptive k-Nearest Neighbor Classifier and Deep Models. Interdiscip Sci. 2022 Mar;14(1):209-232. doi: 10.1007/s12539-021-00483-y. Epub 2022 Jan 10. PMID: 35006529.

[25] Peng L, Yuan R, Shen L, Gao P, Zhou L. LPI-EnEDT: an ensemble framework with extra tree and decision tree classifiers for imbalanced lncRNA-protein interaction data classification. BioData Min. 2021 Dec 3;14(1):50. doi: 10.1186/s13040-021-00277-4. PMID: 34861891; PMCID: PMC8642957.

[26] Zhou L, Duan Q, Tian X, Xu H, Tang J, Peng L. LPI-HyADBS: a hybrid framework for lncRNA-protein interaction prediction integrating feature selection and classification. BMC Bioinformatics. 2021 Nov 26;22(1):568. doi: 10.1186/s12859-021-04485-x. PMID: 34836494; PMCID: PMC8620196.