



WAFNRLTG: A Novel Model for Predicting LncRNA Target Genes Based on Weighted Average Fusion Network Representation Learning Method

Jianwei Li^{1,2*}, Zhenwu Yang¹, Duanyang Wang¹ and Zhiguang Li¹

¹School of Artificial Intelligence, Institute of Computational Medicine, Hebei University of Technology, Tianjin, China, ²Hebei Province Key Laboratory of Big Data Calculation, Hebei University of Technology, Tianjin, China

OPEN ACCESS

Edited by:

Liang Cheng,
Harbin Medical University, China

Reviewed by:

Li Xing,
University of Saskatchewan, Canada
Bingqiang Liu,
Shandong University, China

*Correspondence:

Jianwei Li
lijianwei@hebut.edu.cn

Specialty section:

This article was submitted to
Molecular and Cellular Pathology,
a section of the journal
Frontiers in Cell and Developmental
Biology

Received: 22 November 2021

Accepted: 14 December 2021

Published: 19 January 2022

Citation:

Li J, Yang Z, Wang D and Li Z (2022)
WAFNRLTG: A Novel Model for
Predicting LncRNA Target Genes
Based on Weighted Average Fusion
Network Representation
Learning Method.
Front. Cell Dev. Biol. 9:820342.
doi: 10.3389/fcell.2021.820342

Long non-coding RNAs (lncRNAs) do not encode proteins, yet they have been well established to be involved in complex regulatory functions, and lncRNA regulatory dysfunction can lead to a variety of human complex diseases. lncRNAs mostly exert their functions by regulating the expressions of target genes, and accurate prediction of potential lncRNA target genes would be helpful to further understanding the functional annotations of lncRNAs. Considering the limitations in traditional computational methods for predicting lncRNA target genes, a novel model which was named Weighted Average Fusion Network Representation learning for predicting lncRNA Target Genes (WAFNRLTG) was proposed. First, a novel heterogeneous network was constructed by integrating lncRNA sequence similarity network, mRNA sequence similarity network, lncRNA-mRNA interaction network, lncRNA-miRNA interaction network and mRNA-miRNA interaction network. Next, four popular network representation learning methods were utilized to gain the representation vectors of lncRNA and mRNA nodes. Then, the representations of lncRNAs and target genes in the heterogeneous network were obtained with the weighted average fusion network representation learning method. Finally, we merged the representations of lncRNAs and related target genes to form lncRNA-gene pairs, trained the XGBoost classifier and predicted potential lncRNA target genes. In five-cross validations on the training and independent datasets, the experimental results demonstrated that WAFNRLTG obtained better AUC scores (0.9410, 0.9350) and AUPR scores (0.9391, 0.9350). Moreover, case studies of three common lncRNAs were performed for predicting their potential lncRNA target genes and the results confirmed the effectiveness of WAFNRLTG. The source codes and all data of WAFNRLTG can be freely downloaded at <https://github.com/HGDYZW/WAFNRLTG>.

Keywords: lncRNA target genes prediction, weighted average fusion network representation learning, heterogeneous network, machine learning, XGBoost

INTRODUCTION

Long non-coding RNAs (lncRNAs) are important components of non-coding RNAs whose transcript lengths exceed 200 nucleotides (Ponting et al., 2009). lncRNAs generally exhibit low cross-species conservation, low expression levels and high tissue specificity (Mercer et al., 2008; Pauli et al., 2012), and do not have the functions of protein coding (Carninci and Hayashizaki, 2007). lncRNAs can interfere with the expressions of downstream genes through base complementary pairing, and participate in most biological processes, including cell proliferation, differentiation, chromatin remodeling, epigenetic regulation, genomic splicing, transcription, translation and other aspects (Lander et al., 2001; Guttman et al., 2009; Mercer et al., 2009; Wapinski and Chang, 2011). Due to the important role of lncRNAs in biological processes, their regulatory dysfunctions are commonly associated with a variety of human diseases, especially cancers (Gupta et al., 2010; Zhang et al., 2016). Recent studies have found that lncRNAs regulate many key biological processes by interacting with their target genes. For example, the binding of lncRNA BACE1-AS with its target gene BACE1 increases the stability of BACE1, which regulates BACE1 profile and subsequently affects BACE1 protein expression (Faghihi et al., 2008). In addition, lncRNAs can also be used as the competing endogenous RNA to indirectly regulate mRNA through the shared miRNAs. For example, lncRNA HULC can competitively regulate PRKACB by sharing the common binding site of miR-372, and induce the phosphorylation of CREB in liver cancer (Qi et al., 2015). Since lncRNAs have an important role in biological processes, it prompted researchers to develop computational methods to identify lncRNA regulatory functions. Currently, these computational methods are mainly classified into two categories based on their aims, in terms of lncRNA related diseases identification and lncRNA target genes identification.

In computational models aimed on diseases identification, they can be further divided into two groups: machine learning methods and biological network methods. In recent years, machine learning has been widely applied to predict lncRNA-disease associations. These methods extract the biological features of lncRNAs and diseases and then use machine learning classifiers to infer lncRNA-related diseases. Chen et al. (Chen and Yan, 2013) developed a novel model, LRLSLDA, which predicted potential disease-related lncRNAs in a semi-supervised learning framework. In addition, LRLSLDA requires only human lncRNA expression profiles and known lncRNA-disease associations without negative samples to produce reliable results. Lan et al. (2017) proposed LDAP model to predict lncRNA-disease associations by using a bagging SVM classifier based on lncRNA similarity and disease similarity. Yao et al. (2020) implemented a random forest and feature selection-based lncRNA-disease association prediction model, RFLDA. RFLDA integrated experimentally supported miRNA-disease associations, lncRNA-disease associations, disease semantic similarity, lncRNA functional similarity and lncRNA-miRNA interactions as input features. RFLDA selected the most useful features to train the prediction model by feature selection based

on the importance scores of random forest variables. Based on the hypothesis that similar diseases are more likely to be associated with similar lncRNAs, a number of biological network-based lncRNA-disease association prediction methods have been proposed. Sun et al. (2014) proposed a global network-based computational method named RWRlncD by integrating disease similarity network, lncRNAs functional similarity network and known lncRNA-disease associations. Zhou et al. (2015) proposed RWRHLD method to predict lncRNA-disease associations, which integrated three networks into a heterogeneous network and implemented a random walk on it. Deng et al. (2021) came up with a method, LDAH2V, for inferring lncRNA-disease associations by integrating lncRNA-disease associations, miRNA co-expression profiles, miRNA-disease associations, lncRNA-miRNA associations and lncRNA functional similarity. LDAH2V is a generic network-based link prediction model that can be applied to any number of entity networks.

The theoretical foundation for lncRNAs target genes prediction is the assumption that highly similar lncRNAs tend to have similar interaction. Many studies have shown that lncRNAs indirectly regulate gene expressions via adjusting expressions of miRNAs (Jones-Rhoades and Bartel, 2004). Therefore, exploring the interaction of lncRNA-miRNA would contribute to understand the complex functions of lncRNAs. Due to the rapid development of RNA sequencing technology, lncRNA-related and miRNA-related biological data are increasing rapidly. Predicting the interactions between lncRNAs and miRNAs through traditional experimental methods is very time-consuming and labor-intensive. Recently, many computational methods for predicting lncRNA-miRNA interactions have been proposed. For example, Wong et al. (2020) proposed the LNRLMI model, which constructed a bipartite network to predict potential lncRNA-miRNA interactions by combining the known interaction network and the similarity of the expression profile of lncRNA-miRNA. Zhou et al. (2019) proposed a GEEL model that constructed the lncRNA-miRNA interaction network based on the sequence features and known interactions of lncRNA and miRNA, and then used five different graph embedding methods to obtain the node representation of lncRNA/miRNA. Based on the embedding results, GEEL used individual graph embedding method-based model as basic predictors and build an ensemble model to predict the potential interactions between lncRNAs and miRNAs. Yang et al. (2020) proposed the lncMirNet model which predicted lncRNA-miRNA interactions based on hybrid sequence features. Based on these, Zhao et al. (2020) developed a method, named DeepLGP, for prioritizing lncRNA target genes via encoding gene and lncRNA features. These features were used by the convolutional neural network and were combined as the features of lncRNA-gene pairs. Finally, the CNN model was used to classify lncRNA-gene pairs into true or false pairs.

In this study, we developed a Weighted Average Fusion Network Representation Learning method-based model to improve the performance of lncRNA Target Genes prediction (WAFNRLTG). First, we constructed a heterogeneous network, which integrated two similar networks and three interaction networks. Next, the network representation learning method

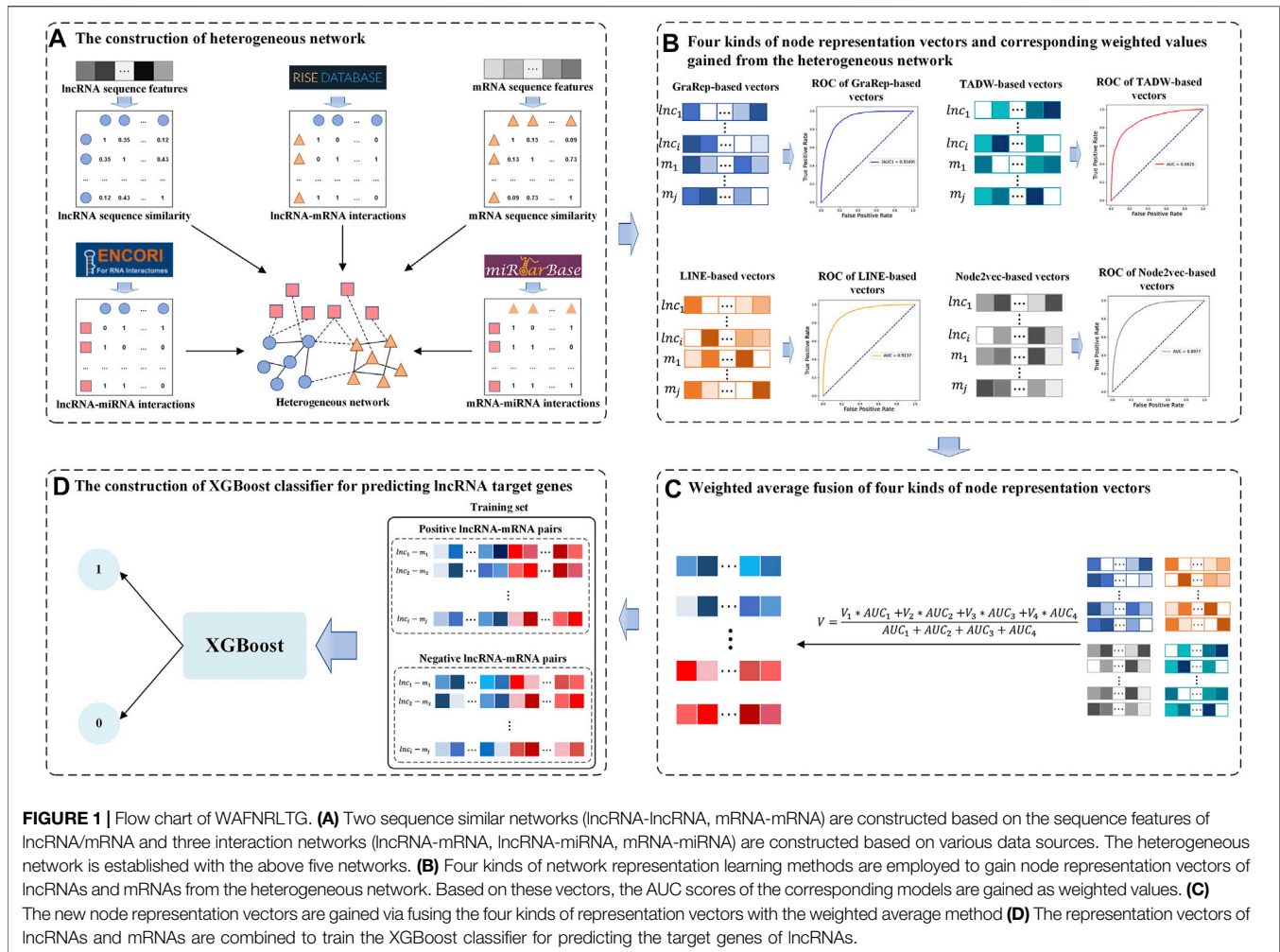


FIGURE 1 | Flow chart of WAFNRLTG. **(A)** Two sequence similar networks (IncRNA-IncRNA, mRNA-mRNA) are constructed based on the sequence features of lncRNA/mRNA and three interaction networks (IncRNA-mRNA, IncRNA-miRNA, mRNA-miRNA) are constructed based on various data sources. The heterogeneous network is established with the above five networks. **(B)** Four kinds of network representation learning methods are employed to gain node representation vectors of lncRNAs and mRNAs from the heterogeneous network. Based on these vectors, the AUC scores of the corresponding models are gained as weighted values. **(C)** The new node representation vectors are gained via fusing the four kinds of representation vectors with the weighted average method **(D)** The representation vectors of lncRNAs and mRNAs are combined to train the XGBoost classifier for predicting the target genes of lncRNAs.

was utilized to gain the representation vectors of lncRNA and mRNA nodes. Four popular network representation learning methods (GraRep, LINE, TADW, and Node2vec) were adopted in our model. Then, the weighted average method was further employed to fuse the corresponding representation vectors according to the AUC scores. The novel representation vectors were obtained which integrated different network structure information and improved the generalization ability of the model. Finally, the representation vectors of lncRNAs and the representation vectors of mRNAs were merged to form the lncRNA-gene pairs, and XGBoost classifier was built based on the merged representations of lncRNA-miRNA pairs. With five-fold cross-validations on training and independent dataset, the experimental results demonstrated that WAFNRLTG obtained AUC scores (0.9410, 0.9350) and AUPR scores (0.9391, 0.9350), and outperformed the individual network representation learning method-based models. Furthermore, three case studies were performed to evaluate the capability of WAFNRLTG. The results can be verified by the existing PubMed literatures. In conclusion, WAFNRLTG is an effective tool for predicting the potential lncRNA target genes. The flow chart of WAFNRLTG is shown in **Figure 1**.

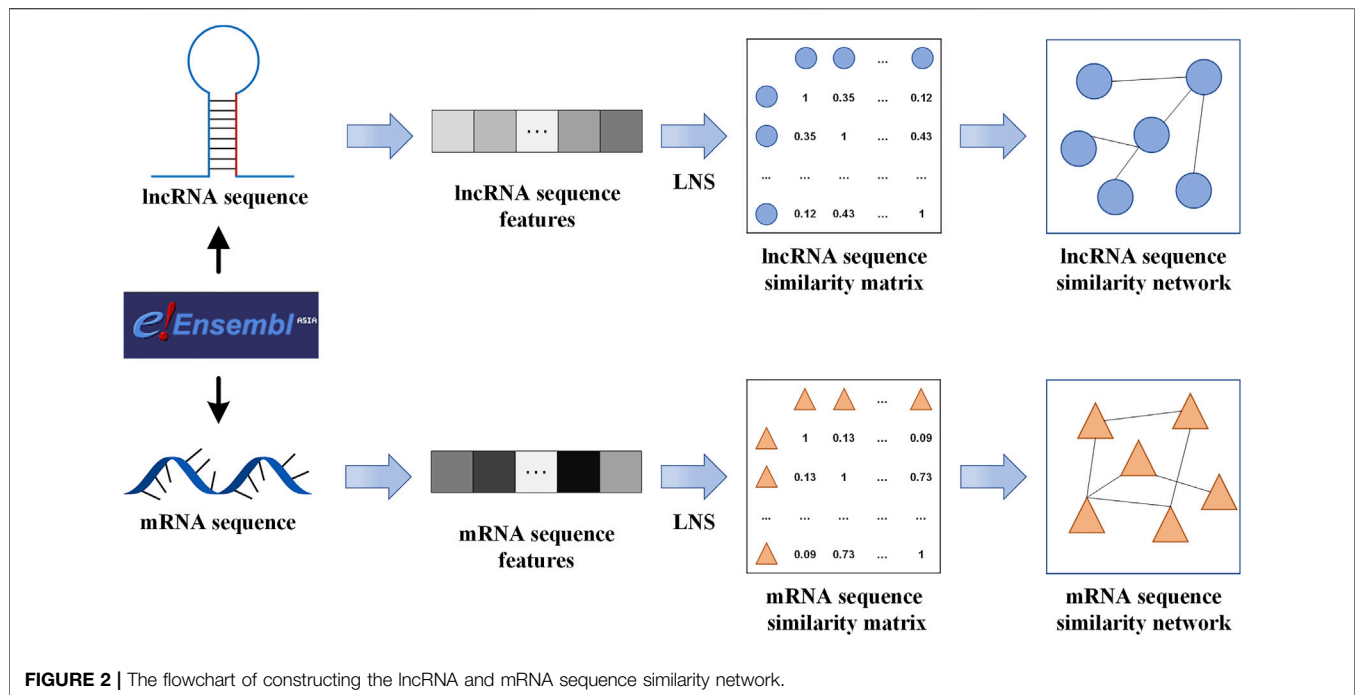
MATERIALS AND METHODS

Datasets

In this paper, there were 5,435 validated lncRNA-gene interactions used as positive samples which were obtained from the RISE database (Hubbard et al., 2002). Negative samples were randomly selected from all unknown lncRNA-gene interactions. Because the number of unknown lncRNA-gene interactions are far more than the number of positive samples, a total of 5,435 negative samples were generated with the same number of positive samples. Eventually, we obtained 5,435 positive samples and 5,435 negative samples as a new dataset. Subsequently, five-sixth of the new dataset were randomly selected as training samples to train the classifier, and the remaining samples were used as the independent samples to evaluate each classifier.

lncRNA and mRNA Sequence Similarity Networks

The corresponding sequences of lncRNAs and mRNAs were downloaded from the Ensembl database (Hubbard et al., 2002) according to their corresponding Ensembl IDs. Linear



neighborhood similarity measure (LNS) and two lncRNA (miRNA) sequence features [k-mer (Gupta et al., 2008) and CTD (Tong and Liu, 2019)] were employed to calculate the lncRNA similarities and miRNA similarities respectively. For each RNA sequence, its k-mer frequency distribution is usually defined as the occurrence frequency of corresponding k-length contiguous subsequences. LNS is a recently proposed similarity calculation method and has been widely used in the field of bioinformatics. The 30-dimensional CTD (composition, transition, and distribution) features are used to represent RNA structure information. In order to construct lncRNA and mRNA sequence similarity networks, the 3-mers features and CTD features of a lncRNA/mRNA sequence were merged into union vectors respectively. For the lncRNA sequence similarity network, the union vectors were used to construct the lncRNA similarity matrix by LNS (Zhang et al., 2017). For example, given a specified lncRNA, its top ten lncRNAs with similarity weights greater than 0 were considered to be linked with it. Based on this strategy, closely homologous lncRNAs were linked with this lncRNA for establishing the lncRNA sequence similarity network. For the mRNA sequence similarity network, its construction procedure was the same as those of the lncRNA sequence similarity network. After this, we gained a lncRNA sequence similarity network involving 2,249 lncRNAs and mRNA sequence similarity network involving 3,785 mRNAs. The flowchart of constructing the lncRNA and mRNA sequence similarity network is shown in **Figure 2**.

lncRNA-mRNA Interaction Network

In this study, experimentally validated lncRNA-mRNA interaction data was downloaded from the RISE database (Gong et al., 2018), and it included 10,941 lncRNA-mRNA

interactions. After removing the redundant data and nonhuman data, 5,435 associations involving 2,249 lncRNAs and 3,785 mRNAs were finally obtained. Therefore, the lncRNA-mRNA interaction network in our model was constructed based on these 5,435 interactions.

lncRNA-miRNA Interaction Network

First, the known lncRNA-miRNA interactions were downloaded from the ENCORI database (Li et al., 2014). Then, the duplicate interactions were removed and the only interactions between the lncRNAs which were from the lncRNA-mRNA interactions and miRNAs were preserved. In the end, the lncRNA-miRNA interaction network was constructed with 6,053 lncRNA-miRNA interactions between 2,249 lncRNAs and 636 miRNAs.

mRNA-miRNA Interaction Network

We downloaded the known mRNA-miRNA interactions from the miTarbase database (Huang et al., 2020). Then, the duplicate interactions were removed and the only interactions between the mRNAs which were from the lncRNA-mRNA interactions and the miRNAs which were from the lncRNA-miRNA interactions were retained. Ultimately, the constructed mRNA-miRNA network in our study contained 1983 mRNA-miRNA interactions between 305 mRNAs and 636 miRNAs.

Network Representation Learning of the Heterogeneous Network

Recently, many Network Representation Learning (NRL) methods have been proposed (Zhang et al., 2020), of which main purpose is to find a proper mapping function to map large-scale, high-dimensional, sparse vectors into a low-dimensional, dense

semantic space, while keep the proximity of these low-dimensional vector representations to the original network. NRL has attracted the attention of scholars in the fields of data mining of biological information data. The low-dimensional representation learned from the network representation is applied to downstream network analysis tasks, such as node classification (Tang et al., 2016a), link prediction, association mining (Zhao et al., 2019), information recommendation (Han et al., 2018) and network visualization (Tang et al., 2016b).

In our study, four state-of-the-art network representation learning methods [GraRep (Cao et al., 2015), LINE (Tang et al., 2015), TADW (Yang et al., 2015) and Node2vec (Grover and Leskovec, 2016)] were used to learn the representation vectors of lncRNA nodes and mRNA nodes for making full use of the various useful information in the heterogeneous network.

LINE maintains both first-order and second-order proximity during learning node representations. Given an undirected edge (i, j) , the joint probability of node v_i and node v_j is as follows:

$$p_1(v_i, v_j) = \frac{1}{1 + \exp(-\vec{u}_i^T \cdot \vec{u}_j)} \quad (1)$$

where $\vec{u}_j \in R^d$ is the low-dimensional vector representation of node v_j . The empirical probability of the distribution $p(\cdot, \cdot)$ on the space $V \times V$ is:

$$\hat{p}_1(i, j) = \frac{w_{ij}}{W} \quad (2)$$

where $W = \sum_{(i,j) \in E} w_{ij}$ and w_{ij} is the weight of edge (i, j) .

After optimizing the model by minimizing the KL scatter of the two distributions, the objective function is defined as follows:

$$O_1 = - \sum_{(i,j) \in E} w_{ij} \log p_1(v_i, v_j) \quad (3)$$

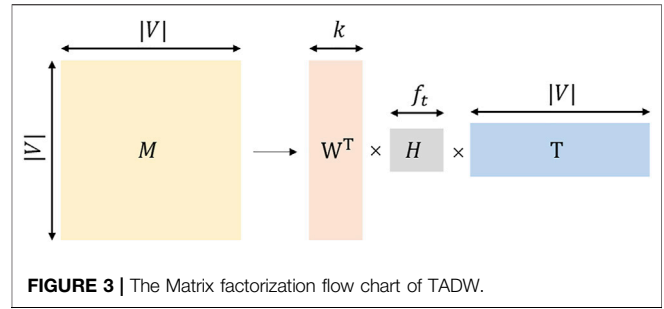
The second-order similarity scenario assumes that nodes sharing a large number of connections with other nodes are similar to each other, and each node is considered as a specific context, then nodes with similar distribution on the context are similar. Here, two vectors \vec{u}_i and \vec{u}_i' are introduced, where \vec{u}_i is the representation of v_i when it is treated as a vertex and \vec{u}_i' is the representation of v_i when it is treated as a specific "context". For an undirected edge (i, j) , the probability of generating the context v_j from v_i is:

$$p_2(v_j|v_i) = \frac{\exp(\vec{u}_j'^T \cdot \vec{u}_i)}{\sum_{k=1}^{|V|} \exp(\vec{u}_k'^T \cdot \vec{u}_i)} \quad (4)$$

where $|V|$ denotes the number of nodes or contexts. The empirical distribution $\hat{p}_2(\cdot|v_i)$ is defined as follows:

$$\hat{p}_2(v_j, v_i) = \frac{w_{ij}}{d_i} \quad (5)$$

where w_{ij} is the weight of edge (i, j) and d_i is the out-degree of node v_i . d_i is used in LINE as the importance of nodes λ_i . By using



the KL scatter while ignoring some constants, the objective function is obtained as follows:

$$O_2 = - \sum_{(i,j) \in E} w_{ij} \log p_2(v_j, v_i) \quad (6)$$

LINE also adopted negative sampling to optimize the model, while using the Alias method to accelerate the sampling process.

GraRep extends LINE by learning the k -order relational vector representations of the network nodes separately through matrix factorization and combines the k -order relational vector representations as the final representation. For a network G , the degree matrix D of the network is defined using the adjacency matrix S . The first-order transfer probability matrix is defined as follows:

$$A = D^{-1}S \quad (7)$$

where $A_{i,j}$ denotes the probability of transferring from v_i to v_j by one step.

Then using Skip-Gram and NCE (noise contrastive estimation) methods, for a transfer of order k , the model can be reduced to the decomposition problem with matrix $Y_{i,j}^k$.

For a transfer of order k , the model is then reduced to a decomposition problem of matrix $Y_{i,j}^k$ by using the Skip-Gram and NCE (noise contrastive estimation) methods.

$$Y_{i,j}^k = W_i^k \cdot C_j^k = \log \left(\frac{A_{i,j}^k}{\sum_t A_{t,j}^k} \right) - \log(\beta) \quad (8)$$

where $\beta = \lambda/N$, λ is the number of negative samples and N is the number of edges in network G .

Node2vec designs a biased random wander over a scalable node neighborhood and explores different node neighborhoods by using breadth first search (BFS) and depth first search (DFS), and then inputs the resulting node sequences as sentences into the skip-gram model to learn node representations. For a random wandering sequence $v_{i-w}, \dots, v_{i-1}, v_i, v_{i+1}, \dots, v_{i+m}$, centered on v_i with windows size w . Node2vec uses the Skip-Gram algorithm to optimize the model.

$$Pr(\{v_{i-w}, \dots, v_{i+w}\} | \Phi(v_i)) = \prod_{j=i-w, j \neq i}^{i+w} Pr(v_j | \Phi(v_i)) \quad (9)$$

TADW utilizes the inductive matrix decomposition method based on integrating the network topology information and

textual information. The Matrix factorization flow chart of TADW is shown in **Figure 3**, where V denotes the set of nodes, $W \in R^{k \times |V|}$, $M \in R^{|V| \times |V|}$, $H \in R^{k \times f_t}$, and $T \in R^{f_t \times |V|}$, and T is the text feature matrix. Therefore, the loss function of TADW as follows:

$$\min_{W,H} \|M - W^T H T\|_F^2 + \frac{\lambda}{2} (\|W\|_F^2 + \|H\|_F^2) \quad (10)$$

These four state-of-the-art network representation learning methods can capture valuable information of the structure and intrinsic properties of the heterogeneous network. The learned representations of lncRNAs and mRNAs were further utilized to construct WAFNRLTG.

Weighted Average Fusion Node Representation

The processes of fusing the node representations derived from the network representation learning methods (GraRep, LINE, TADW, and Node2vec) are described below. First, low-dimensional representations of lncRNAs and mRNAs were obtained from the heterogeneous network. Thus, the four representation vectors of nodes were obtained with different methods were V_1 , V_2 , V_3 and V_4 , and the corresponding prediction models were constructed based on these vectors. The AUC score of each model was calculated as the weighted value of the corresponding methods, they were labeled as AUC_1 , AUC_2 , AUC_3 and AUC_4 . To make full use of the information in node representation vectors and improve the generalization ability of the model, the new representation vector V was obtained by the weighted average fusion of the four representation vectors.

$$V = \frac{V_1 * AUC_1 + V_2 * AUC_2 + V_3 * AUC_3 + V_4 * AUC_4}{AUC_1 + AUC_2 + AUC_3 + AUC_4} \quad (11)$$

RESULTS

Parameter and Experimental Settings

In this study, four network representation learning methods (GraRep, LINE, TADW, and Node2vec) were employed to obtain node representation vectors from the heterogeneous network.

First, we took the dimensions of the node representation vectors as the common parameter of these four methods. Tuning dimensions d , and the d -dimensional features that produced the best AUC were selected. The experimental results under node representation vectors of different dimensions are shown in **Table 1**.

The other parameters of the network representation learning method are discussed in the following. GraRep has a parameter: k -step k , which indicates the k -step transfer matrix. Node2vec has four tunable parameters: number-walks n , indicates the number of random walks from each node; walk-length l , indicates the length of a random walk from each node; p and q control the

TABLE 1 | AUC scores of models with different dimension node representation vectors.

Method	$d = 16$	$d = 32$	$d = 64$	$d = 128$	$d = 256$
GraRep	0.8725	0.8991	0.9095	0.9147	0.8957
LINE	0.8375	0.8939	0.9177	0.9123	0.8828
TADW	0.7974	0.8466	0.8803	0.8397	0.9002
Node2vec	0.8709	0.8978	0.9131	0.9099	0.9115

TABLE 2 | Parameters settings for network representation learning methods.

Method	Optimal parameter value
GraRep	$d: 128, k: 4$
LINE	$d: 64, o: 3$
TADW	$d: 256, \lambda: 0.1$
Node2vec	$d: 64, n: 20, l: 80$

probability of the random walk to the next node. The combination of number-walks n and walk-length l were considered, and the rest of the parameter was set as defaults. TADW has a parameter: λ , which controls the weight of the regularization term. In LINE, a parameter: Order o , denotes the order of proximity was considered. Above all, we adjusted different parameter values and adopt the optimal values which produced the best AUC scores. The parameter settings for various of the network representation learning method are shown in **Table 2**. For more details, please see **Supplementary File S1**.

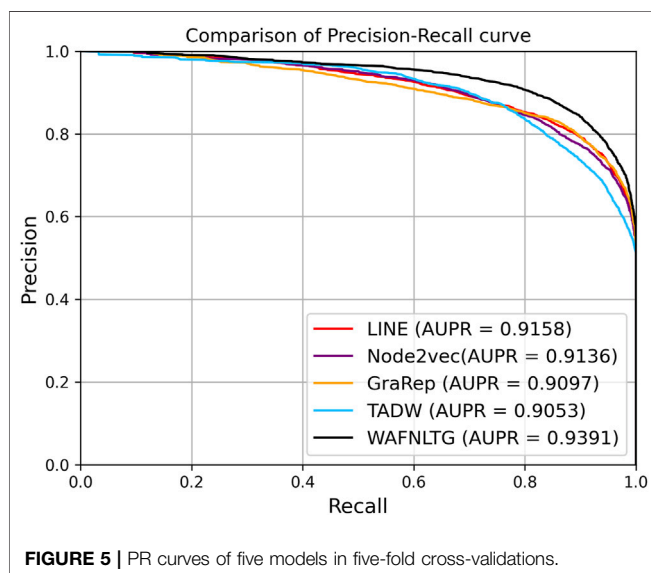
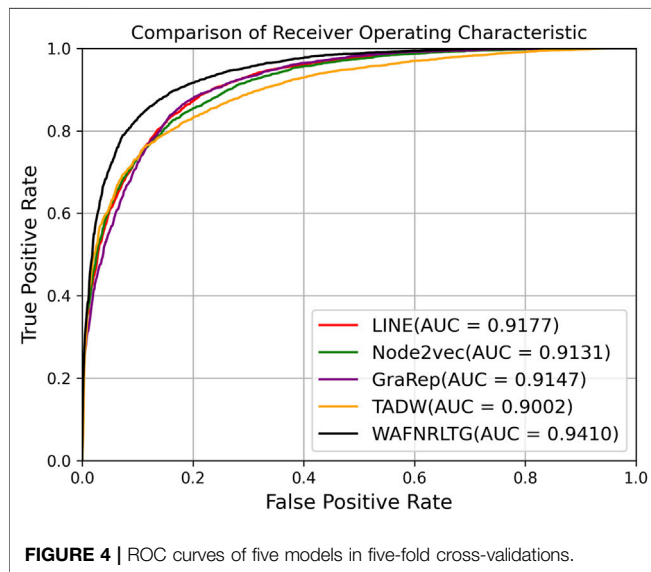
Comparison With Four Network Representation Learning Methods

In this section, we first constructed four models based on four representation network representation learning methods (GraRep, LINE, TADW, and Node2vec) and evaluated their effectiveness. In order to make full use of four network representation learning methods, the AUC scores of each method were adopted as weight values, and the four representation vectors were fused by the weighted average method. According to these constructed representation vectors, XGBoost classifier (Chen and Guestrin, 2016) was selected into WAFNRLTG.

Based on these four representations from the above network representation learning models, we subsequently fused them by the weighted average method to improve performance and generalization of WAFNRLTG. For assessing the prediction performance of WAFNRLTG, five-fold cross-validations experiments were used to evaluate the classification performances and the four network representation learning models. In our study, seven commonly metrics, namely Sensitivity (SN), Specificity (SP), Precision (PREC), Accuracy (ACC), Matthews correlation coefficient (MCC), AUC, and AUPR, were employed as evaluation metrics. ROC (receiver operating characteristic curve, ROC) and PR (Precision-Recall curve, PR) curves were plotted for showing the different performance of each model. The results of WAFNRLTG and

TABLE 3 | Performances of the five network representation learning models.

Method	Acc	Sen	Spec	Prec	MCC	AUC	AUPR
GraRep	0.8107	0.7920	0.8295	0.8229	0.6220	0.9147	0.9097
LINE	0.8358	0.8275	0.8441	0.8413	0.6718	0.9177	0.9158
TADW	0.8223	0.7798	0.8648	0.8523	0.6471	0.9002	0.9053
Node2vec	0.8305	0.8328	0.8282	0.8290	0.6611	0.9131	0.9136
WAFNRLTG	0.8683	0.8638	0.8728	0.8716	0.7362	0.9410	0.9391



the four network representation learning models are shown in **Table 3**. It can be seen from the table that WAFNRLTG achieves AUC score of 0.9410 and AUPR score of 0.9391, which outperforms GraRep (AUC score: 0.9147; AUPR score: 0.9097), LINE (AUC score: 0.9177; AUPR score: 0.9158), TADW (AUC score: 0.9002; AUPR score: 0.9053) and Node2vec (AUC

score: 0.9131; AUPR score: 0.9136). ROC and PR curves are plotted for the five models to further display their different performances, **Figure 4** shows five ROC curves and **Figure 5** shows five PR curves.

For comparative analysis, we also adopted the other two fusion methods (concatenate and average) to fuse the four node representation vectors similar to the weighted average method. The experimental results of the three fusion methods are recorded in **Table 4**. As shown in **Table 4**, the weighted average fusion method outperforms other methods.

Moreover, the information extracted from heterogeneous networks by WAFNRLTG are brought to the subsequent work for predicting of lncRNA target genes.

In order to evaluate the generalization ability of WAFNRLTG, we applied it on the training dataset and the independent dataset. Experimental results are shown in **Table 5**. As exhibited from **Table 5**, the results on the independent dataset are comparable to the results on the training dataset. The experimental results on the independent dataset demonstrate that WAFNRLTG is a robust and reliable model for predicting potential lncRNA target genes.

Comparison Among Different Classifiers

After the acquisition of weighted averages fused representation vectors, we compared five different machine learning methods and selected the most appropriate one. These five classification methods include K Nearest Neighbor (KNN), AdaBoost (Freund and Schapire, 1997), Support Vector Machine (SVM), Gradient Boosting Decision Tree (GBDT) and XGBoost, which are well known to perform well on a variety of tasks. They were evaluated by five-fold cross-validations. **Figure 6** shows their prediction performances. The AUC scores of KNN, AdaBoost, SVM, GBDT and XGBoost are 0.9082, 0.8717, 0.8959, 0.8462, and 0.9394, respectively. The detailed results of these five classifiers are shown in **Table 6**. From the analysis of above results, XGBoost model achieved the best performance.

Case Studies

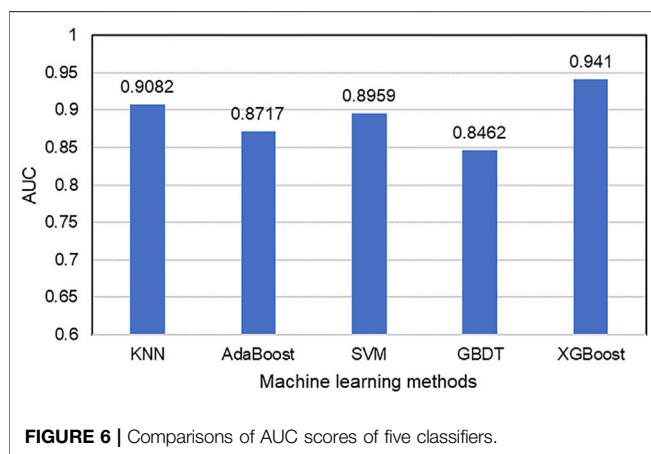
The main objective of this study is to screen potential target genes of lncRNAs and guide relevant researchers to explore novel target genes. For further evaluating the performance of WAFNRLTG model in practical applications, we selected three common lncRNAs (MALAT1, PVT1, and NEAT1) as case studies. The general processes of each of case studies were as following. First, all lncRNA-mRNA interactions from our dataset were utilized to construct the WAFNRLTG model. Then, the interactions between above three lncRNAs and the other mRNAs were adopted as the dataset in WAFNRLTG model. The model

TABLE 4 | The experimental results of the three fusion methods.

Method	ACC	SEN	SPEC	PREC	MCC	AUC	AUPR
Concatenate	0.86	0.8479	0.8722	0.8691	0.7203	0.9391	0.9307
Average	0.8515	0.8412	0.8618	0.8589	0.7033	0.9296	0.9277
Weighted average	0.8683	0.8638	0.8728	0.8716	0.7362	0.9410	0.9391

TABLE 5 | Performances of WAFNRLTG on the independent dataset and the training dataset.

Dataset	Acc	Sen	Spec	Prec	MCC	AUC	AUPR
independent dataset	0.8666	0.8567	0.8765	0.874	0.7333	0.9358	0.9350
training dataset	0.8683	0.8638	0.8728	0.8716	0.7362	0.9410	0.9391

**TABLE 6** | The detailed experimental results of the five classifiers.

Classifier	ACC	SEN	SPEC	PREC	MCC	AUC	AUPR
GBDT	0.8235	0.8037	0.8433	0.8369	0.6475	0.9082	0.9039
KNN	0.7197	0.7604	0.679	0.7037	0.4412	0.8012	0.8005
RF	0.785	0.7761	0.794	0.7907	0.5703	0.8718	0.8745
SVM	0.8063	0.784	0.8284	0.8204	0.6131	0.8959	0.8961
XGBoost	0.8683	0.8638	0.8728	0.8716	0.7362	0.9410	0.9391

outputted the predicting scores of the lncRNA-mRNA pairs. Finally, the top ten mRNAs were selected for literature mining on PubMed. We found that four of the top ten mRNAs corresponding to MALAT1 and NEAT1 were validated in the literature, and three of the top ten mRNAs corresponding to PVT1 were validated in the literatures.

MALAT1 is one of the first lncRNA discovered that are associated with human diseases. Many studies have demonstrated that the abnormal expression of MALAT1 is closely related to cancer pathophysiology, and has the potential to be translated clinically. MALAT1 regulates cancer processes by interacting with molecules, such as proteins, RNAs and DNAs, and further alters different signal pathways. To

TABLE 7 | The predicting target genes in the top ten for MALAT1.

Rank	Score	Target genes	PMID
1	0.9952035	MYC	33312756
2	0.953083	MT-ND4L	Not found
3	0.952141	LRRC2	Not found
4	0.947306	TET2	33165916
5	0.92575	ECT2	27313681
6	0.923501	OR1M1	Not found
7	0.917459	ATXN2L	Not found
8	0.917261	TP73AS1	32714991
9	0.915255	NFIC	Not found
10	0.908282	FASTK	Not found

demonstrate the ability of our model for predicting potential lncRNA target genes, we predicted the top ten mRNAs that interact with MALAT1, as shown in **Table 7**. With literature mining, we found that four mRNAs interact with MALAT1 and they can be used as the target genes of MALAT1. For example, abnormal expression of MALAT1 leads to reduced expression of TET2 thus causing neuronal damage, so there may be a targeting relationship between MALAT1 and TET2 (Li et al., 2021). The complete case study results of MALAT1 are available in **Supplementary File S2**.

NEAT1 expression is upregulated in many human malignancies, such as lung, esophageal and gastric cancers. In order to demonstrate that WAFNRLTG is effective in predicting potential target genes of lncRNAs, we predicted the top 10 mRNAs associated with NEAT1, and the results are shown in **Table 8**. After literature mining, four of them were shown to be target genes of NEAT1. For example, it was found that MYC-regulated NEAT1 promoted diffuse large B-cell lymphoma (DLBCL) proliferation via the miR-34b-5p-GLI1 pathway, which could provide a novel therapeutic target for DLBCL (Qian et al., 2020). The complete case study results of NEAT1 are available in **Supplementary File S3**.

Plasmacytoma variant translocation 1 (PVT1) is a newly discovered long non-coding RNA which preforms regulating functions as an oncogenic molecule in different cancers. In order to understand the functions of PVT1, WAFNRLTG

TABLE 8 | The predicting target genes in the top ten for NEAT1.

Rank	Score	Target genes	PMID
1	0.975445	SUSD6	Not found
2	0.961867	CPN2	Not found
3	0.952019	PEX26	Not found
4	0.951388	WNT9A	Not found
5	0.942409	SRP19	Not found
6	0.931132	MYC	32206038
7	0.926551	OAS3	33138195
8	0.925127	CPSF6	22960638
9	0.924981	QSOX2	Not found
10	0.910639	TET2	33987091

TABLE 9 | The predicting target genes in the top ten for PVT1.

Rank	Score	Target genes	PMID
1	0.971728	NANOG	34230224
2	0.9714675	GID4	Not found
3	0.970239	WNT9B	Not found
4	0.969168	TRPM3	Not found
5	0.966868	COL5A2	33750300
6	0.966181	RCC2	Not found
7	0.963472	RBM7	Not found
8	0.961825	WNT3A	32727463
9	0.957906	FAM101B	Not found
10	0.953643	CCDC115	Not found

predicted its target genes. We conducted a literature survey of the top ten mRNAs predicted by WAFNRLTG to interact with PVT1 and found that three mRNAs were verified to be its target genes, as shown in **Table 9**. For example, the abnormal expression of PVT1 affects the expression of NANOG and thus makes difference in the development of glioma (Gong et al., 2021). The complete case study results of PVT1 are available in **Supplementary File S4**.

DISCUSSION

lncRNAs and its target genes are involved in a variety of biological processes and are closely associated with serious human diseases. Predicting the potential lncRNA target genes can decipher complex biological mechanisms and reveal the functions of lncRNAs. In this paper, we firstly collected and processed multiple data from the multiple open databases, including lncRNA-mRNA interactions, lncRNA-miRNA interactions and mRNA-miRNA interactions. The lncRNA sequence similarity network and mRNA sequence similarity network were constructed based on sequence features. Then, we proposed a novel model, WAFNRLTG, to infer potential lncRNA target genes by integrating above data. In five-fold cross-validations on training and independent dataset, the experimental results demonstrated that WAFNRLTG obtained AUC scores (0.9410, 0.9350) and AUPR scores (0.9391, 0.9350). Three common lncRNAs (MALAT1, NEAT1, and PVT1) were introduced to WAFNRLTG model. Several target genes in the

predicted results were found according to experimental PubMed literatures.

Although WAFNRLTG has achieved satisfactory results in predicting lncRNA target genes, it is still necessary to point out there are still limitations in our model. For example, the negative samples randomly selected from unknown lncRNA-mRNA interactions may have the chance of becoming positive samples. This would have an impact on the accuracy of WAFNRLTG. The information extracted from the heterogeneous network is not comprehensive, some valuable information may be neglected. With the accumulation of biological data, more interaction records would be introduced to enrich the heterogeneous network and improve the prediction ability of WAFNRLTG.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JL conceived, designed the study, ZY and DW developed the algorithm and performed the statistical analysis, ZY and ZL wrote the codes. ZY drafted the original manuscript, JL revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work is supported by the National Natural Science Foundation of China under Grants No. 81672113, 62072154, 61806072.

ACKNOWLEDGMENTS

We thank members in our groups for their valuable discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcell.2021.820342/full#supplementary-material>

Supplementary File S1 | The experimental details of WAFNRLTG.

Supplementary File S2 | The prediction of MALAT1 target genes results of WAFNRLTG.

Supplementary File S3 | The prediction of NEAT1 target genes results of WAFNRLTG.

Supplementary File S4 | The prediction of PVT1 target genes results of WAFNRLTG.

REFERENCES

- Cao, S., Lu, W., and Xu, Q. (2015). "GraRep: Learning Graph Representations with Global Structural Information," in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (Melbourne, Australia: Association for Computing Machinery), 891–900. doi:10.1145/2806416.2806512
- Carninci, P., and Hayashizaki, Y. (2007). Noncoding RNA Transcription beyond Annotated Genes. *Curr. Opin. Genet. Development* 17 (2), 139–144. doi:10.1016/j.gde.2007.02.008
- Chen, T., and Guestrin, C. (2016). "XGBoost," in *Proceedings Of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining* (San Francisco: Association for Computing Machinery), 785–794. doi:10.1145/2939672.2939785
- Chen, X., and Yan, G.-Y. (2013). Novel Human lncRNA-Disease Association Inference Based on lncRNA Expression Profiles. *Bioinformatics* 29 (20), 2617–2624. doi:10.1093/bioinformatics/btt426
- Deng, L., Li, W., and Zhang, J. (2021). LDAH2V: Exploring Meta-Paths across Multiple Networks for lncRNA-Disease Association Prediction. *Ieee/acm Trans. Comput. Biol. Bioinf.* 18 (4), 1572–1581. doi:10.1109/TCBB.2019.2946257
- Faghghi, M. A., Modarresi, F., Khalil, A. M., Wood, D. E., Sahagan, B. G., Morgan, T. E., et al. (2008). Expression of a Noncoding RNA Is Elevated in Alzheimer's Disease and Drives Rapid Feed-Forward Regulation of β -secretase. *Nat. Med.* 14 (7), 723–730. doi:10.1038/nm1784
- Freund, Y., and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Computer Syst. Sci.* 55 (1), 119–139. doi:10.1006/jcss.1997.1504
- Gong, J., Shao, D., Xu, K., Lu, Z., Lu, Z. J., Yang, Y. T., et al. (2018). RISE: a Database of RNA Interactome from Sequencing Experiments. *Nucleic Acids Res.* 46 (D1), D194–D201. doi:10.1093/nar/gkx864
- Gong, R., Li, Z.-Q., Fu, K., Ma, C., Wang, W., and Chen, J.-C. (2021). Long Noncoding RNA PVT1 Promotes Stemness and Temozolomide Resistance through miR-365/ELF4/SOX2 Axis in Glioma. *Exp. Neurobiol.* 30 (3), 244–255. doi:10.5607/en20060
- Grover, A., and Leskovec, J. (2016). "node2vec," in *Proceedings Of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco: Association for Computing Machinery), 855–864. doi:10.1145/2939672.2939754
- Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., et al. (2010). Long Non-coding RNA HOTAIR Reprograms Chromatin State to Promote Cancer Metastasis. *Nature* 464 (7291), 1071–1076. doi:10.1038/nature08975
- Gupta, S., Dennis, J., Thurman, R. E., Kingston, R., Stamatoyannopoulos, J. A., and Noble, W. S. (2008). Predicting Human Nucleosome Occupancy from Primary Sequence. *Plos Comput. Biol.* 4 (8), e1000134. doi:10.1371/journal.pcbi.1000134
- Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., et al. (2009). Chromatin Signature Reveals over a Thousand Highly Conserved Large Non-coding RNAs in Mammals. *Nature* 458 (7235), 223–227. doi:10.1038/nature07672
- Han, X., Shi, C., Zheng, L., Yu, P. S., Li, J., and Lu, Y. (2018). "Representation Learning with Depth and Breadth for Recommendation Using Multi-View Data," in *Web and Big Data* (Springer International Publishing), 181–188. doi:10.1007/978-3-319-96890-2_15
- Huang, H.-Y., Lin, Y.-C.-D., Li, J., Huang, K.-Y., Shrestha, S., Hong, H.-C., et al. (2020). miRTarBase 2020: Updates to the Experimentally Validated microRNA-Target Interaction Database. *Nucleic Acids Res.* 48 (D1), D148–D154. doi:10.1093/nar/gkz896
- Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., et al. (2002). The Ensembl Genome Database Project. *Nucleic Acids Res.* 30 (1), 38–41. doi:10.1093/nar/30.1.38
- Jones-Rhoades, M. W., and Bartel, D. P. (2004). Computational Identification of Plant microRNAs and Their Targets, Including a Stress-Induced miRNA. *Mol. Cell* 14 (6), 787–799. doi:10.1016/j.molcel.2004.05.027
- Lan, W., Li, M., Zhao, K., Liu, J., Wu, F.-X., Pan, Y., et al. (2017). LDAP: a Web Server for lncRNA-Disease Association Prediction. *Bioinformatics* 33 (3), btw639–460. doi:10.1093/bioinformatics/btw639
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial Sequencing and Analysis of the Human Genome. *Nature* 409 (6822), 860–921. doi:10.1038/35057062
- Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2014). starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and Protein-RNA Interaction Networks from Large-Scale CLIP-Seq Data. *Nucl. Acids Res.* 42 (Database issue), D92–D97. doi:10.1093/nar/gkt1248
- Li, L., Miao, M., Chen, J., Liu, Z., Li, W., Qiu, Y., et al. (2021). Role of Ten Eleven Translocation-2 (Tet2) in Modulating Neuronal Morphology and Cognition in a Mouse Model of Alzheimer's Disease. *J. Neurochem.* 157 (4), 993–1012. doi:10.1111/jnc.15234
- Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long Non-coding RNAs: Insights into Functions. *Nat. Rev. Genet.* 10 (3), 155–159. doi:10.1038/nrg2521
- Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F., and Mattick, J. S. (2008). Specific Expression of Long Noncoding RNAs in the Mouse Brain. *Proc. Natl. Acad. Sci.* 105 (2), 716–721. doi:10.1073/pnas.0706729105
- Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., et al. (2012). Systematic Identification of Long Noncoding RNAs Expressed during Zebrafish Embryogenesis. *Genome Res.* 22 (3), 577–591. doi:10.1101/gr.133009.111
- Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and Functions of Long Noncoding RNAs. *Cell* 136 (4), 629–641. doi:10.1016/j.cell.2009.02.006
- Qi, X., Zhang, D.-H., Wu, N., Xiao, J.-H., Wang, X., and Ma, W. (2015). ceRNA in Cancer: Possible Functions and Clinical Implications. *J. Med. Genet.* 52 (10), 710–718. doi:10.1136/jmedgenet-2015-103334
- Qian, C.-S., Li, L.-J., Huang, H.-W., Yang, H.-F., and Wu, D.-P. (2020). MYC-regulated lncRNA NEAT1 Promotes B Cell Proliferation and Lymphomagenesis via the miR-34b-5p-GLI1 Pathway in Diffuse Large B-Cell Lymphoma. *Cancer Cell Int* 20, 87. doi:10.1186/s12935-020-1158-6
- Sun, J., Shi, H., Wang, Z., Zhang, C., Liu, L., Wang, L., et al. (2014). Inferring Novel lncRNA-Disease Associations Based on a Random Walk Model of a lncRNA Functional Similarity Network. *Mol. Biosyst.* 10 (8), 2074–2081. doi:10.1039/C3MB70608G
- Tang, J., Aggarwal, C., and Liu, H. (2016a). "Node Classification in Signed Social Networks," in *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM)* (Miami: Society for Industrial and Applied Mathematics), 54–62. doi:10.1137/1.9781611974348.7
- Tang, J., Liu, J., Zhang, M., and Mei, Q. (2016b). "Visualizing Large-Scale and High-Dimensional Data," in *Proceedings of the 25th International Conference on World Wide Web* (Montréal, Québec: International World Wide Web Conferences Steering Committee), 287–297. doi:10.1145/2872427.2883041
- Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). "LINE: Large-scale Information Network Embedding," in Proceedings of the 24th international conference on world wide web (Florence, Italy: International World Wide Web Conferences Steering Committee), 1067–1077. doi:10.1145/2736277.2741093
- Tong, X., and Liu, S. (2019). CPPred: Coding Potential Prediction Based on the Global Description of RNA Sequence. *Nucleic Acids Res.* 47 (8), e43. doi:10.1093/nar/gkz087
- Wapinski, O., and Chang, H. Y. (2011). Long Noncoding RNAs and Human Disease. *Trends Cell Biol.* 21 (6), 354–361. doi:10.1016/j.tcb.2011.04.001
- Wong, L., Huang, Y. A., You, Z. H., Chen, Z. H., and Cao, M. Y. (2020). LNRLMI: Linear Neighbour Representation for Predicting lncRNA-miRNA Interactions. *J. Cell Mol Med* 24 (1), 79–87. doi:10.1111/jcmm.14583
- Yang, C., Liu, Z., Zhao, D., Sun, M., and Chang, E. (2015). "Network Representation Learning with Rich Text Information," in *Proceedings of the 24th International Conference on Artificial Intelligence* (Buenos Aires, Argentina: AAAI Press), 2111–2117.
- Yang, S., Wang, Y., Lin, Y., Shao, D., He, K., and Huang, L. (2020). LncMirNet: Predicting lncRNA-miRNA Interaction Based on Deep Learning of Ribonucleic Acid Sequences. *Molecules* 25 (19), 4372. doi:10.3390/molecules25194372
- Yao, D., Zhan, X., Zhan, X., Kwoc, C. K., Li, P., and Wang, J. (2020). A Random forest Based Computational Model for Predicting Novel lncRNA-Disease Associations. *BMC Bioinformatics* 21 (1), 126. doi:10.1186/s12859-020-3458-1
- Zhang, D., Yin, J., Zhu, X., and Zhang, C. (2020). Network Representation Learning: A Survey. *IEEE Trans. Big Data* 6 (1), 3–28. doi:10.1109/TBDATA.2018.2850013

- Zhang, K., Luo, Z., Zhang, Y., Zhang, L., Wu, L., Liu, L., et al. (2016). Circulating lncRNA H19 in Plasma as a Novel Biomarker for Breast Cancer. *Cbm* 17 (2), 187–194. doi:10.1016/j.molcel.2004.05.02710.3233/cbm-160630
- Zhang, W., Yue, X., Liu, F., Chen, Y., Tu, S., and Zhang, X. (2017). A Unified Frame of Predicting Side Effects of Drugs by Using Linear Neighborhood Similarity. *BMC Syst. Biol.* 11 (6), 101–134. doi:10.1186/s12918-017-0477-2
- Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a Novel Deep Learning Method for Prioritizing lncRNA Target Genes. *Bioinformatics* 36 (16), 4466–4472. doi:10.1093/bioinformatics/btaa428
- Zhao, Z., Li, C., Zhang, X., Chiclana, F., and Viedma, E. H. (2019). An Incremental Method to Detect Communities in Dynamic Evolving Social Networks. *Knowledge-Based Syst.* 163, 404–415. doi:10.1016/j.knosys.2018.09.002
- Zhou, M., Wang, X., Li, J., Hao, D., Wang, Z., Shi, H., et al. (2015). Prioritizing Candidate Disease-Related Long Non-coding RNAs by Walking on the Heterogeneous lncRNA and Disease Network. *Mol. Biosyst.* 11 (3), 760–769. doi:10.1039/c4mb00511b
- Zhou, S., Yue, X., Xu, X., Liu, S., Zhang, W., and Niu, Y. (2019). “lncRNA-miRNA Interaction Prediction from the Heterogeneous Network through Graph Embedding Ensemble Learning,” in Proceeding of the 2019IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 8-21 Nov. 2019 (IEEE), 622–627. doi:10.1109/BIBM47256.2019.8983044
- 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.
- Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Li, Yang, Wang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.