Heterogeneous graph neural network for lncRNA-disease association prediction

Hong Shi
School of Information,Yunnan Normal University

Xiaomeng Zhang
School of Information,Yunnan Normal University

Lin Liu (liulinrache@163.com)
School of Information,Yunnan Normal University

Lin Tang
Key Laboratory of Educational Informatization for Nationalities Ministry of Education,Yunnan Normal University

Article

Keywords:

Posted Date: May 31st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1675450/v1

Heterogeneous graph neural network for lncRNA-disease association prediction

Hong Shi[1], **Xiaomeng Zhang**[1], **Lin Liu**[1,*], **and Lin Tang**[2]

[1]School of Information, Yunan Normal University, Kunming, 650092, China
[2]Key Laboratory of Educational Informatization for Nationalities Ministry of Education, Yunnan Normal University, Kunming, 650092, China
[*]liulinrache@163.com

ABSTRACT

The computational methods of lncRNA-disease association prediction are effective ways to solve the problem of expensive and time-consuming traditional biological experiments. However, existing methods still have challenges to make full use of network topology information to identify potential associations between lncRNA and disease in multi-source data. In this study, we propose a novel method called HGNNLDA for lncRNA-disease association prediction. First, HGNNLDA constructs a heterogeneous network composed of lncRNA similarity network, lncRNA-disease association network and lncRNA-miRNA association network; Then, on this heterogeneous network, various types of strong correlation neighbors with fixed size are sampled for each node by restart random walk; Next, the embedding information of lncRNA and disease in each lncRNA-disease association pair is obtained by the method of type-based neighbor aggregation and all types combination though heterogeneous graph neural network, in which attention mechanism is introduced considering that different types of neighbors will make different contributions to the prediction of lncRNA-disease association. As a result, the area under the receiver operating characteristic curve(AUC) and the area under the precision-recall curve(AUPR) under 5-fold cross-validation (5FCV) are 0.9786 and 0.8891, respectively. Compared with five state-of-art prediction models, HGNNLDA has better prediction performance. In addition, in two types of case studies, it is further verified that our method can effectively predict the potential lncRNA-disease associations, and have ability to predict new diseases without any known lncRNAs.

Introduction

Long non-coding RNAs (lncRNAs) are non-coding RNAs with more than 200nt (nucleotides) in length[1]. More and more studies have shown that lncRNAs participates in many important biological processes, including gene transcription, cell differentiation and genetic regulation[2]. Moreover, Complex diseases that seriously endanger human health are also inseparable from the abnormal expression of lncRNAs, including diabetes[3], cardiovascular diseases[4], HIV[5], mental disorders[6] and some cancers such as lung cancer[7], breast cancer[8] and prostate cancer[9]. Therefore, identifying the associations between lncRNA and disease contributes to understanding the pathogenesis and principles of the disease, and also provides help for the diagnosis, treatment and prevention of human disease. However, the traditional biological experiments take up a long time, cost much, and have some blindness, all of which will hinder the research process. In recent years, established lncRNA databases such as LncRNADisease2.0[10], Lnc2Cancer v2.0[11], NRED[12], MNDR[13], and GeneRIF[14] have made it possible to develop computational methods for predicting potential lncRNA-disease associations. According to the different ideas of algorithms, the existing methods for predicting lncRNA-disease association can be broadly classified into two categories. They are the method based on biological network and machine learning, respectively.

Computational methods based on biological networks often rely on the known associations information between lncRNA and disease to build heterogeneous networks. Then lncRNA-disease association prediction is carried out based on this heterogeneous network. For example, Sun et al.[15] proposed a network-based computational model RWRlncD that predicts potential associations between human lncRNA-disease by restart random walks in a lncRNA functional similarity network, but RWRlncD only randomly walks on a network and does not fuse rich biological data information. Gu et al.[16] proposed a model for a random walk on a global network (GrwLDA) that uses random walk in the lncRNA similarity network and the disease similarity network to predict potential lncRNA-disease associations. However, GrwLDA has difficulties in optimizing the model parameters. Wen at al.[17] proposed the Lap-BiRWRHLDA model, which Laplace normalized the similarity matrix before constructing the lncRNA and disease similarity networks, integrates the two similarity networks through known lncRNA-disease associations, and then implements the prediction of lncRNA-disease association using a double random walk on this heterogeneous network. Zhang et al.[18] propose a LncRDNetFlow based on a global network framework that integrates multisource networks, including lncRNA similarity networks, protein interaction networks, disease similarity networks, and

associations information among heterogeneous nodes. The model prioritizes disease-related lncRNA and is able to predict potential associations information for an isolated disease.

Computational methods based machine learning predict potential associations between lncRNA and disease by building lncRNA-disease association models, and train the model to improve accuracy using known lncRNA-disease association data. Chen et al.[19] assumed that similar diseases are often associated with functionally similar lncRNAs, and developed a model LRLSLDA based on a semi-supervised learning framework, where LRLSLDA effectively predicts potential lncRNA-disease associations by integrating known lncRNA-disease associations and lncRNA expression profiles. Nonetheless, LRLSLDA has the problem of optimize the model parameters. Subsequently, Chen et al.[20] proposed a new lncRNA-disease prediction model named LNCSIM. LNCSIM further improved LRLSLDA model by introducing lncRNA-disease prediction similarity score. However, this method still does not solve the problem of parameter selection of semantic contribution factors. Zhao et al.[21] developed a naive Bayesian-based computational approach that integrates various information of disease-related lncRNA, including genomic, regulome, transcriptome, which resulted in successfully predicting 707 potential cancer-associated potential lncRNAs. Lan et al.[22] proposed a novel computational method that uses Katcher means to fuse the lncRNA and disease similarity matrix of multiple data sources and predicts potential lncRNA-disease associations by the SVM classifier.

These two types of approaches still have methodological weaknesses. The methods based on biological network rely heavily on the constructed lncRNA-disease heterogeneous network. When network structure changes, this kind of method can't effectively deal with it. The problem of the method based on machine learning is how to select the optimal features. Most existing machine learning methods do not take full advantage of the rich topological information contained in heterogeneous networks. To make full use of the lncRNA and disease feature information and the local and global information on the lncRNA-disease association data, the graph neural network approach appears in some new studies recently. For example, Xuan et al.[23] used graph convolution network and convolutional neural network to learn the network structure information and the local network features of lncRNA-disease association pair. Wu et al.[24] used graph convolutional network(GCN) as encoder to obtain the features of lncRNA and disease on the heterogeneous network, and then calculated the interaction score between lncRNA and disease by using the inner product of two potential factor vector. Zhang et al.[25] utilized met apaths to represent complex semantic information between entities in the network and introduced attention mechanisms to learn the weights of each neighborhood under the metapath and finally aggregate the potential features they obtained from the GCN model. These graph neural network methods realize the capture and utilization of topological information in heterogeneous networks, but ignore the heterogeneity of nodes and edges in heterogeneous graphs.

Inspired by Zhang et al.[26], the heterogeneity of structure and content in the heterogeneous graph is considered. We propose a novel method for lncRNA-disease association prediction called HGNNLDA. First, a heterogeneous network is constructed, which is composed of the similar network of lncRNA, the known lncRNA-disease association network and the known lncRNA-miRNA association network. Then, a fixed-size sampling of strongly correlated neighbors is performed by restart random walk for each lncRNA and disease, and the sampled neighbors are grouped according to the types of nodes. Then, the feature vectors of sampled lncRNA, disease and miRNA are obtained by word2vec. The final embedding information of each lncRNA and disease is extracted by aggregating the sampling neighbors according to types and fusing different types, in which attention mechanism is introduced to indicate the importance of different types of neighbors. Finally, the embedding obtained from above steps of each lncRNA-disease association pair are used as the input of classifier, and the prediction score of association pair is calculated. The experimental results show that the AUC and AUPR values of HGNNLDA under 5-fold cross validation (5FCV) are 0.9786 and 0.8891, respectively, which is superior to other state-of-art methods. In addition, two case studies show that HGNNLDA has the ability to predict disease-related lncRNA without any known association.

Results

Performance evaluation

We considered 2697 known lncRNA-disease associations as positive samples, but the number of positive samples only accounts for 2.7% of the total number of samples, so we randomly selected negative samples with the same number of positive samples from all unknown association pairs. After constructing the training set of the model, 5-fold cross validation(5FCV) is used to evaluate the prediction performance of HGNNLDA. which is to divide the sample set into 5 disjoint subsets on average. In each cross-validation, each subset is regarded as a testing samples in turn and the rest as a training samples. Then, HGNNLDA model trained is used to obtain the score of each test sample. The higher the score, the more likely it is that this lncRNA sample is related to the disease. Next, all test samples are sorted in descending order according to their scores. On this basis, we calculate the true positive rate (TPR) and false positive rate (FPR), Precision and Recall under different thresholds. The specific

calculation is as follows:

$$TPR = \frac{TP}{TP+TN} \qquad FPR = \frac{FP}{TN+FP}$$
$$Precision = \frac{TP}{TP+FP} \qquad FPR = \frac{TP}{TN+FN} \tag{1}$$

Where TP (true positive) means that positive samples are correctly predicted as positive samples; FN (false negative) indicates that the positive sample is erroneously predicted as a negative sample; FP (false positive) means that the negative sample is erroneously predicted as a positive sample; TN (true negative) means that the negative sample is correctly predicted as a negative sample. Then, the ROC curve is drawn with TPR as the vertical axis and FPR as the horizontal axis, and the area under the ROC curve (AUC value) is used as the performance index to evaluate the prediction performance of the model. If the AUC value is larger, the prediction performance of this model is better. To improve the evaluation of the model performance when the positive and negative samples are seriously unbalanced, we also calculated AUPR value to evaluate the overall performance of the model.

Comparison with other models

In order to further evaluate the prediction performance of HGNNLDA method, we compared it with five state-of-art lncRNA-disease association prediction models, such as SIMCLDA[27], MFLDA[28], LDAP[22], CNNLDA[29] and GCNLDA[23]. Under the 5FCV, the average AUCs and AUPRs of all lncRNA-disease association prediction models as shown in Table 1. ROC curve of each cross-validation of HGNNLDA is shown in Fig. 1.

Method	AUC	AUPR
SIMCLDA	0.746	0.095
MFLDA	0.626	0.066
LDAP	0.863	0.166
CNNLDA	0.952	0.251
GCNLDA	0.959	0.223
HGNNLDA	0.9786	0.8891

Table 1. The mean AUCs and AUPRs of different methods.

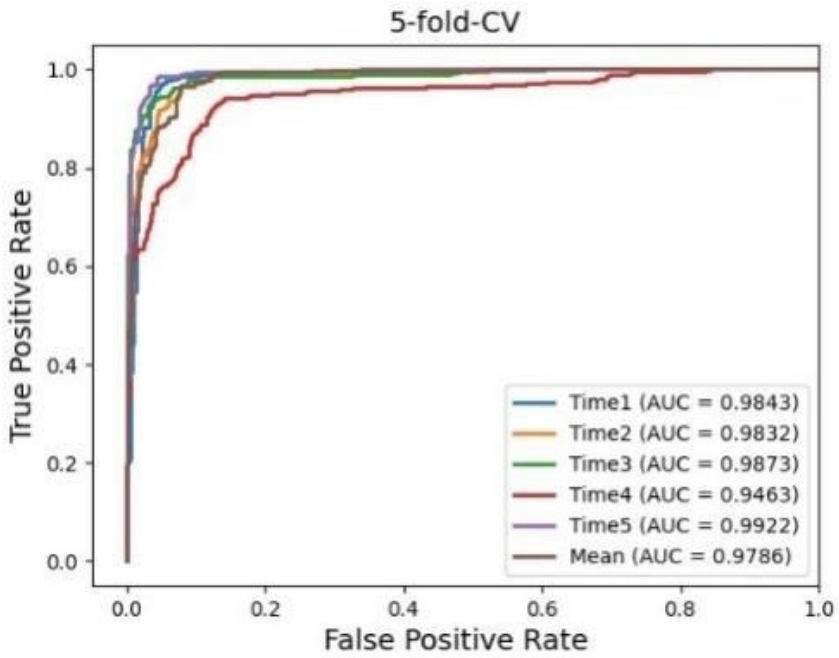

Figure 1. ROC curves of HGNNLDA based on 5-fold cross-validation.

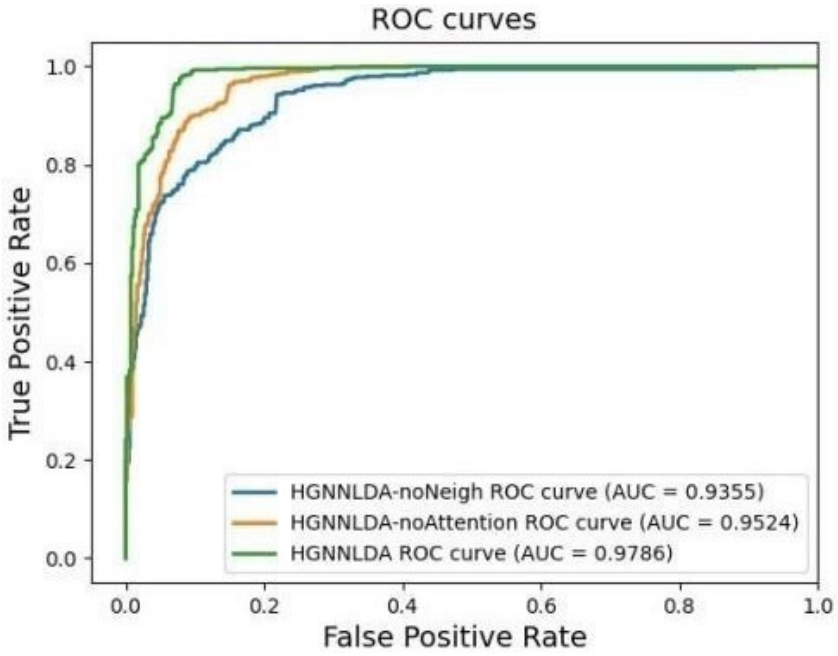

Figure 2. Performance of HGNNLDA and its variants.

Ablation study

To analyze the necessity of each component of our model, we adopted two variants of HGNNLDA (HGNNLDA-noNeigh and HGNNLDA-noAttention) as the comparison method. Specifically, HGNNLDA-noNeigh means that the embedded information of each node is only obtained by word2vec, and the information of any neighboring nodes is not aggregated. HGNNLDA-noAttention uses fully connected neural network instead of attention mechanism to aggregate the embedding of different types of neighbors, which means that different types of neighbor nodes are equally important for the final embedding of lncRNA and disease. Fig. 2 shows the average AUC obtained using HGNNLDA and two variant models. HGNNLDA has better performance than HGNNLDA-noNeigh, which indicates that aggregating the information of neighboring nodes can better generate the embedded information of nodes. HGNNLDA gets better results than HGNNLDA-noAttention, which shows that attention mechanism can capture the influence of different types of nodes.

The effects of embedding size

Embedding size plays an important role in HGNNLDA, which is able to directly affect the performance of the model. In the experiment, we set different embedding dimension d (i.e. 8, 16, 32, 64, 128, 256), and evaluate the prediction performance under different setting. As can be seen from Fig. 3, within a certain range, the larger the embedding dimension, the better the node representation can be learned, and the higher the AUC value. However, when the embedding dimension increase continuously, the AUC value will become stable or slightly worse, which may be caused by over-fitting. Accounting for this factor, the embedding size is set to 128 in this paper.

Case studies

To further verify the accuracy and effectiveness of HGNNLDA, we conducted two types of case studies.

For the first type of case study, we applied our proposed method to predict the potential lncRNA-disease associations of three common diseases (lung cancer, colon cancer and osteosarcoma). First, for a specific disease, we regarded all known associations lncRNA-disease as training samples and unknown associations with this disease as candidate samples. Then, we scored all unknown candidate samples of lncRNA-/lung cancer/colon cancer/osteosarcoma, then sorted the scores in descending order and select the top 10 candidate associations related to this disease. The prediction results were verified using two databases (LncRNADisease[10] database and LncRNA2Cancer[11] database) and published literatures. Table 2 showed the top 10 results of predicting the potential associations with lung cancer, colon cancer and osteosarcoma, the accuracy reached 100%, 80% and 80% respectively. The results showed that our method can effectively predict the potential lncRNA-disease associations.

For the second type of case study, We evaluated the ability of our proposed method to predict the new associations of diseases without any known related lncRNA. We took breast as an example in this case study. First, we set the known associations of breast cancer as unknown associations, and all lncRNAs were considered as candidate lncRNAs. The HGNNLDA was used

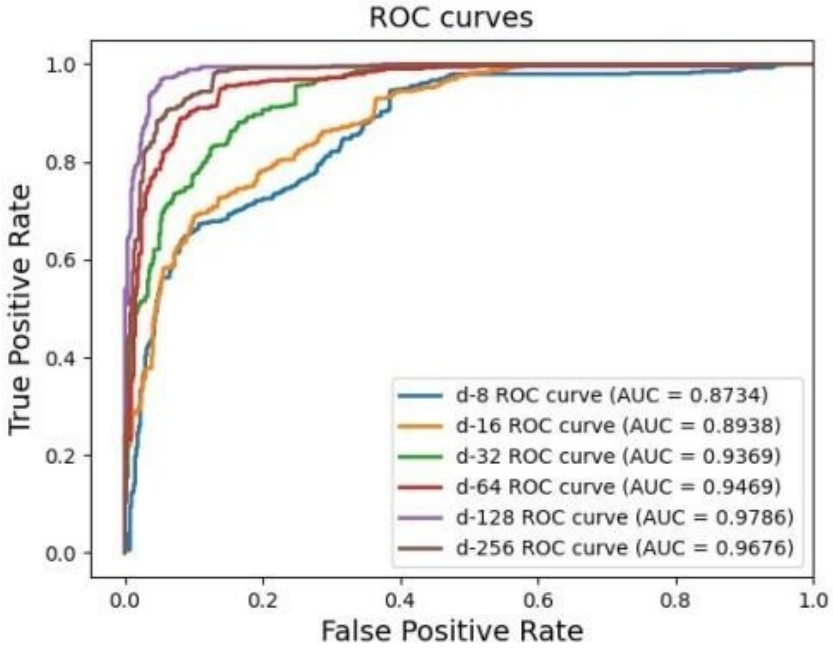

Figure 3. ROC curves of HGNNLDA based on 5-fold cross-validation.

to score these candidate lncRNAs associated with breast cancer. We found that 27 of the top 30 lncRNA were confirmed by LncRNAdisease database or LncRNA2Cancer database, as shown in Table 3. This result shows that HGNNLDA can effectively predict the potential associations of diseases without any known related lncRNAs.

Discussion

More and more studies show that lncRNA is an important regulatory factor of organisms and can be used as a diagnostic marker for many diseases. Exploring the associations between lncRNAs and disease are helpful for diagnosis, prognosis and treatment of these diseases. In this paper, we propose a novel method of HGNNLDA to predict the potential associations between lncRNAs and diseases. From the comparison of experimental results, it can be seen that HGNNLDA has superior performance for predicting lncRNA-disease associations. In addition, two types of cases also verify that HGNNLDA has the ability to identify potential lncRNA-disease associations, and can effectively predict new diseases without any known lncRNA.

The reliable performance of HGNNLDA is related to the following factors. First, the model integrates multiple sources of heterogeneous data useful for predicting lncRNA-disease associations to build a biological heterogeneous networks. Second, HGNNLDA gets all types of strong related neighbors of fixed size for each node by restarting random walk, which solves the defect that the direct related neighbors of some nodes are not representative enough. In addition, HGNNLDA is able to capture the strong correlation neighbor features of each node in this heterogeneous network, and fully exploiting the topology information of the heterogeneous network. Finally, HGNNLDA employs the attention mechanism to account for the differential impact of different types of nodes on lncRNA-disease association prediction. To sum up, HGNNLDA makes full use of the complex structural and semantic information of heterogeneous network, and so as to achieves good prediction of lncRNA-disease associations.

However, our method still has some limitations. First, the data we use to build heterogeneous networks may contain noise and some outliers. Second, we randomly select the unknown lncRNA-disease association pairs as negative samples for training, which can't guarantee that the lncRNA and disease in the unknown association pairs are completely unrelated, so it will have some influence on the prediction performance. Therefore, our future research will focus on how to overcome these problems.

Methods

The general overview of our proposed HGNNLDA framework is shown in Fig. 4, which consists of five key parts: (1) Construction of heterogeneous networks. First we downloaded lncRNA-disease associations, lncRNA-miRNA associations, and calculated the similarity between lncRNAs, and then constructed a heterogeneous network containing the three types of nodes of lncRNA, disease, and miRNA. (2) Sampling strong correlation neighbors and the feature representation of each

Rank	LncRNA	Disease	Evidence
1	GAS5	Lung cancer	LncRNAdisease
2	NEG8	Lung cancer	Lnc2Cancer
3	HOOTTIP	Lung cancer	Lnc2Cancer
4	LINC00472	Lung cancer	Lnc2Cancer
5	ZFAS1	Lung cancer	Lnc2Cancer
6	HULC	Lung cancer	Lnc2Cancer
7	BCAR4	Lung cancer	Lnc2Cancer
8	CASC15	Lung cancer	Lnc2Cancer
9	BCYRN1	Lung cancer	Lnc2Cancer
10	GHET1	Lung cancer	Lnc2Cancer
1	UCA1	Colon cancer	LncRNAdisease
2	OIP5-AS1	Colon cancer	Lnc2Cancer
3	HOTTIP	Colon cancer	LncRNAdisease
4	HOTAIR	Colon cancer	LncRNAdisease
5	LINC00319	Colon cancer	Unconfirmed
6	PVT1	Colon cancer	Lnc2Cancer
7	GAS5	Colon cancer	Lnc2Cancer
8	KCNQ10T1	Colon cancer	Lnc2Cancer
9	DANCR	Colon cancer	Lnc2Cancer
10	BANCR	Colon cancer	Unconfirmed
1	NEAT1	Osteosarcoma	Lnc2Cancer
2	XIST	Osteosarcoma	Lnc2Cancer
3	CCAT1	Osteosarcoma	LncRNAdisease
4	EWSAT1	Osteosarcoma	LncRNAdisease
5	AFAP1-AS1	Osteosarcoma	Unconfirmed
6	KCNQ10T1	Osteosarcoma	Lnc2Cancer
7	MIR155HG	Osteosarcoma	Unconfirmed
8	GAS5	Osteosarcoma	Lnc2Cancer
9	PVT1	Osteosarcoma	Lnc2Cancer
10	OIP5-AS1	Osteosarcoma	Lnc2Cancer

Table 2. The top 10 predicted lncRNAs associated with lung cancer, colon cancer, osteosarcoma.

neighbor. We sampled various types of fixed-sized neighbors for each lncRNA and disease by restart the random walk, and then extract the features of each neighbor node by word2vec. (3) Embedding learning. We used Bi-LSTM to obtain embedding for the three types of neighbors, lncRNA, disease, and miRNA. (4) Updating the node embedding. We introduced the attention mechanism, and aggregated the embedding of three types of neighbors and the embedding of nodes themselves based on the weights obtained. (5) LncRNA-disease association prediction. The embedding of lncRNA and disease were concatenated to get the embedding of lncRNA-disease association pair, then the prediction scores between lncRNA and disease were obtained by using fully connected and softmax layers, eventually optimizde the model by cross-entropy.

Datasets for lncRNA-disease associations prediction

Studies have shown that lncRNA can interact with the corresponding miRNA and perform biological functions together with miRNA[30]. Therefore, all useful biological information can be assembled to construct a heterogeneous network including the lncRNA-lncRNA similarity network, the experimentally validated lncRNA-disease association network, and the lncRNA-miRNA association network, so as to predict the potential lncRNA-disease associations. The data used in this paper were obtained from the previous study of lncRNA-disease association prediction by Fu et al.[28]. This dataset included 240 lncRNAs, 412 diseases, and 495 miRNAs. Among them, 2,697 verified lncRNA-disease associations are derived from LncRNADisease[10], Lnc2Cancer[11] and GeneRIF[14] databases. In addition, 1002 lncRNA-miRNA associations came from starBase database[31].

LncRNA functional similarity network

In this paper, the functional similarity of lncRNA is calculated by the method of Chen et al.[20]. LncRNA similarity is expressed by the similarity of lncRNA related diseases. Suppose that lncRNA $l(1)$ is associated with a group of diseases

LncRNA(1-15)	Evidence	LncRNA(16-30)	Evidence
H19	Lnc2Cancer	LSINCT5	Lnc2Cancer
HOTTIP	Lnc2Cancer	PVT1	Lnc2Cancer
CDKN2B-AS1	LncRNAdisease	ZFAS1	Lnc2Cancer
AFAP1-AS1	Lnc2Cancer	NCRUPAR	Unconfirmed
KCNQ1OT1	Lnc2Cancer	SOX2-OT	LncRNAdisease
LINC00472	Lnc2Cancer	TP53COR1	Unconfirmed
CASC16	LncRNAdisease	BCAR4	Lnc2Cancer
MALAT1	Lnc2Cancer	NPSR1-AS1	Unconfirmed
NEAT1	Lnc2Cancer	GHET1	Lnc2Cancer
LINC00583	LncRNAdisease	MIR17HG	LncRNAdisease
XIST	Lnc2Cancer	LINC-ROR	Lnc2Cancer
HOTAIR	Lnc2Cancer	NBAT1	Lnc2Cancer
CCAT2	Lnc2Cancer	BANCR	Lnc2Cancer
BCYRN1	LncRNAdisease	HOTAIRM1	Lnc2Cancer
SPRY4-IT1	Lnc2Cancer	DANCR	Lnc2Cancer

Table 3. The top 30 predicted lncRNAs associated with breast cancer.

$D(1)=\{d(11),d(12),\ldots,d(1m)\}$, lncRNA $l(2)$ is associated with a group of diseases $D(2)=\{d(21),d(22),\ldots,d(2n)\}$. Then the functional similarity between lncRNA $l(1)$ and $l(2)$ is represented by $S_{l(1),l(2)}$ as follows:

$$S_{l(1),l(2)}=\frac{\sum\limits_{1\le i\le m}\max\limits_{1\le j\le n}(DSS(d(1i),d(2j))+\sum\limits_{1\le j\le n}\max\limits_{1\le i\le m}(DSS(d(2j),d(1i))}{m+n} \tag{2}$$

$$LFS=\begin{pmatrix} S_{l(1),l(1)} & \cdots & S_{l(1),l(240)} \\ \vdots & \ddots & \vdots \\ S_{l(240),l(1)} & \cdots & S_{l(240),l(240)} \end{pmatrix} \tag{3}$$

where $DSS(d(1i),d(2j))$ represents the semantic similarity between disease $d(1i)$ and disease $d(2j)$, which adopts the method calculated by Wang et al.[32]; m and n represent the number of diseases in disease group $D(1)$ and $D(2)$, respectively; LFS is a functional similarity matrix of 240×240, and 240 represents the number of lncRNA.

LncRNA-disease associations and lncRNA-miRNA associations

The datasets includes 2697 experimentally verified lncRNA-disease associations and 1002 experimentally verified lncRNA-miRNA associations[28]. The associations between lncRNA and disease are expressed by a 240×412 adjacency matrix LD, $LD(l(i),l(j))=1$, if lncRNA $l(i)$ is related to disease $d(j)$, otherwise it is 0. Similarly, the associations between lncRNA and miRNA ar represented by an adjacency matrix LM of 240×495, $LM(l(i),m(j))=1$, if lncRNA $l(i)$ is related to miRNA $m(j)$, otherwise it is 0.

Heterogeneous network construction

As shown in Fig. 4(a), we construct a heterogeneous network based on lncRNA functional similarity LFS, lncRNA-disease association network LD and lncRNA-miRNA association network LM. Heterogeneous networks can be expressed as:

$$G=(N,E,NT,ET) \tag{4}$$

where N represents the node set, which contains three types of nodes, namely $NT=\{\ln cRNA, disease, miRNA\}$, E represents the edge set, which contains three types of edges, namely $ET=\{\ln cRNA-disease, \ln cRNA-\ln cRNA, \ln cRNA-miRNA\}$.

Sampling heterogeneous neighbors with Restart Random Walk

In heterogeneous networks, the neighbors of many nodes cannot include all types of nodes, and the number of neighbor nodes will vary[26]. For example, in Fig. 4(a), no disease node is directly connected to the miRNA node, and d_1 has two neighbor nodes, while l_2 has seven neighbor nodes. Therefore, to make full use of the information of heterogeneous networks, we introduce restart random walk (RRW) to sample three types of strongly correlated neighbors for each node. The sampling operation of RRW in lncRNA-disease heterogeneous network includes two steps:

- Selecting fixed size sampling length for *RRW*. Starting random walk from node $v \in N$, return to the starting node with probability p or iteratively move to the neighbor of the current node, where the probability q controls whether the walk is depth first select or breadth first select. When $q > 1$, random walk tends to give priority to breadth; when $q < 1$, random walk tends to give priority to depth. *RRW* runs until a fixed number of nodes are successfully collected, and the sampled nodes are denoted as $|RRW(v)|$. Moreover, the number of different types of nodes in $|RRW(v)|$ is constrained, which can ensure that all types of nodes are sampled.
- Grouping neighbor nodes of lncRNA, disease and miRNA-type. For each node type t, the top k_t nodes are selected based on the frequency of occurrence, and take them as the set of t-type correlated neighbors of node v.

In this way, three types of neighbors can be collected for each node, and classification by type is conducive to subsequently learn embedding of type.

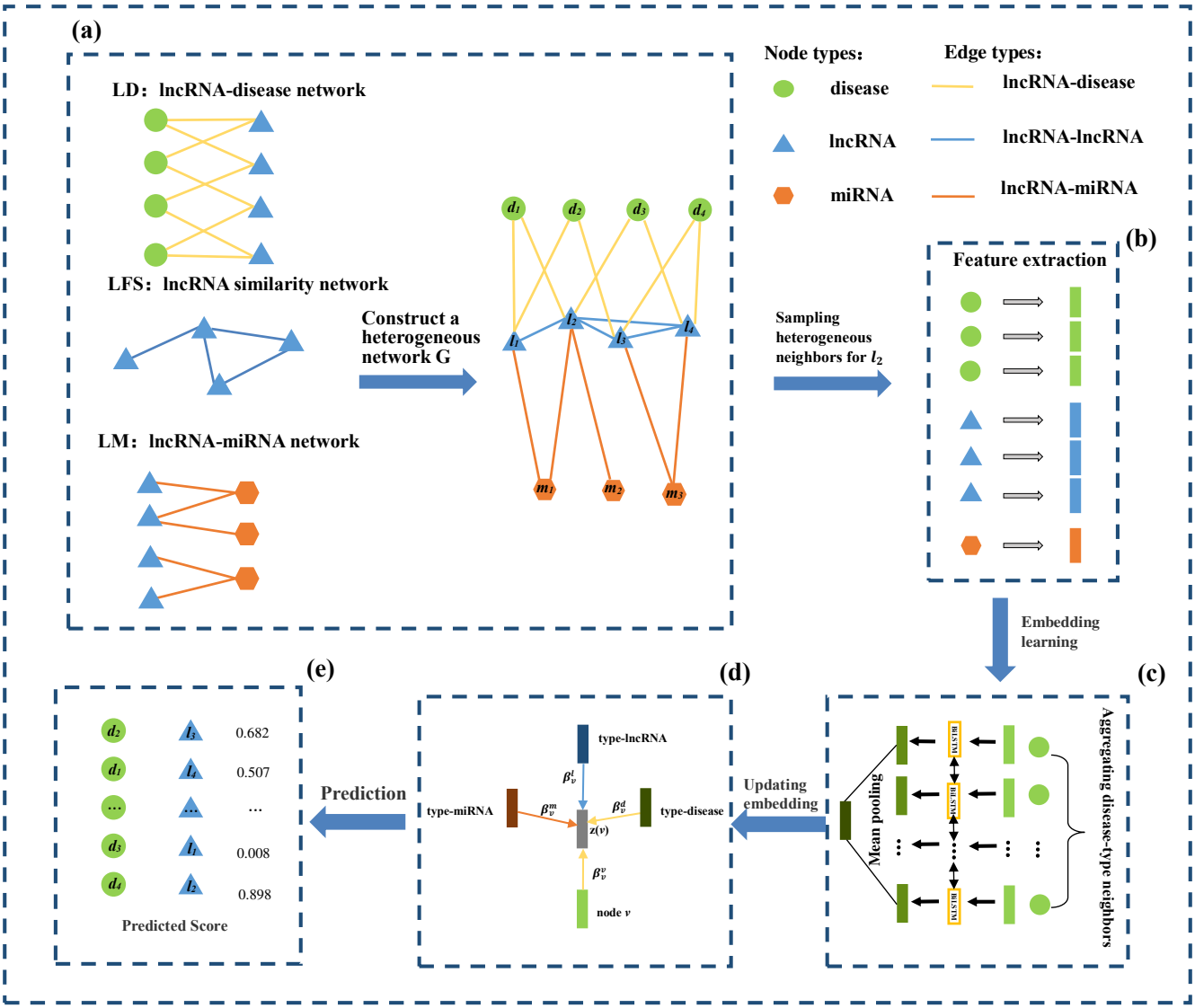

Figure 4. The framework of HGNNLDA.

Embedding learning

In the last step, three types of neighbors (lncRNA, disease, miRNA) of each node are sampled at a fixed size by using the strategy based on *RRW*. Therefore, to obtain the embedding of type, we can aggregate all the same type neighbors after sampling by using Bi-LSTM[33]. For example, we can express disease-type neighbors of node $v \in N$ in the heterogeneous network as $N_d(v)$. Next, the embedding of each disease-type neighbor node is obtained by word2vec[34], as shown in Fig. 4(b).

Then we utilize Bi-LSTM to aggregate the embeddings of all disease-type neighbors, as shown in Fig. 4(c). In the process of aggregating all lncRNA-type nodes, disease-type nodes and miRNA-type nodes, different Bi-LSTM are used to distinguish them. Bi-LSTM consists of a forward LSTM layer and a backward LSTM layer. The main structure of LSTM layer can be expressed as follows:

$$\begin{aligned} i_s &= \sigma\left(w_s \cdot f(s) + h_{s-1} \cdot w_{s'} + b_f\right) \\ f_s &= \sigma\left(w_f \cdot f(s) + h_{s-1} \cdot w_{f'} + b_f\right) \\ o_s &= \sigma\left(w_o \cdot f(s) + h_{s-1} \cdot w_{o'} + b_o\right) \\ \tilde{c}_s &= \sigma\left(w_c \cdot f(s) + h_{s-1} \cdot w_{c'} + b_c\right) \\ c_s &= i_s \otimes \tilde{c}_s + f_s \otimes c_{s-1} \\ h_s &= o_s \otimes \tanh(c_s) \end{aligned} \tag{5}$$

Where σ is sigmoid activation function; i, f, o and c represent input gate vector, forget gate, vector, output gate vector and memory unit respectively; h_s represents the output hidden vector by s-th node; w and b represent learnable parameters; $\otimes$ represents dot product operation. Two different middle layer representations can be obtained through calculation. Then, after splicing the two middle layers, the general embedding of all disease-type neighbor nodes of node v can be obtained through the average pool layer, as shown follow:

$$\begin{aligned} \overrightarrow{h_s} &= LSTM_d\left(\overrightarrow{h_{s-1}}, f(s)\right) \\ \overleftarrow{h_s} &= LSTM_d\left(\overleftarrow{h_s}, f(s)\right) \\ f^d(v) &= \frac{\sum_{s \in N_d(v)} \overrightarrow{h_s} \oplus \overleftarrow{h_s}}{|N_d(v)|} \end{aligned} \tag{6}$$

Where $f^d(v) \in \mathrm{R}^{d \times 1}$ is the general embedding of all disease-type neighbors of node v; $\overrightarrow{h_s}$ and $\overleftarrow{h_s}$ represent the forward and backward LSTM representations of s node respectively; the symbol $\oplus$ indicates the connection operation.

Updating the node embedding with attention mechanism

In the previous step, the general embedding of lncRNA-type, disease-type and miRNA-type will be generated. Different types of neighbors will have different influences on the final embedding of node v[26], for example, nodes of lncRNA, disease-type usually play a more important role in the prediction of lncRNA-disease association. So as to combine lncRNA-type, disease-type and miRNA-type general embeddings with node v embedding, we introduce the attention mechanism[35]. First, the importance of each type is learned, and then all heterogeneous types of nodes(including node v itself) are aggregated to form the final embedding of node v. For any $t \in N(v)$, $N(v) = \{v \cup NT\}$, the importance β_v^t of t-type relative to node v is expressed as:

$$\beta_v^t = \frac{\exp\left(\sigma\left(\mathrm{q}^{\mathrm{T}}\left[f(v) \parallel f^t(v)\right]\right)\right)}{\sum_{k \in N(v)}\left(\exp(\sigma\left(\mathrm{q}^{\mathrm{T}}\left[f(v) \parallel f^t(v)\right]\right))\right)} \tag{7}$$

Where σ is *ReLU* activation function; $\mathrm{q}^{\mathrm{T}} \in \mathrm{R}^{2d \times 1}$ represents the attention vector; $f(v)$ is that embedding of v obtained by word2vec; $f^t(v)$ is a general embedding based on t-type aggregating; $\parallel$ indicates the connection operation; $f^t(v) = f(v)$ when k equals v. Then, the final embedding of node v can be aggregated by various types of embedding based on the corresponding importance coefficient. The details are as follows:

$$z(v) = \sigma\left(\sum_{k \in N(v)} \beta_v^t f^k(v)\right) \tag{8}$$

Where $z(v) \in \mathrm{R}^{d \times 1}$ represents the final embedding. To better understand the aggregation process of various types of nodes, explanation is shown in Fig. 4(d).

LncRNA-disease association prediction

The final embedding of lncRNA l_i and the final embedding of disease d_j are spliced to constitute the vector representation $x_{i,j} = z(l_i) \otimes z(d_j)$ of the association pair $l_i - d_j$:

$$x_{i,j} = z(l_i) \otimes z(d_j) \tag{9}$$

Where $\otimes$ represents splicing operation. Then, each positive sample (there is an association between lncRNA and disease) is marked as 1, and each negative sample (there is no association between lncRNA and disease) is marked as 0. Then, we provide

the embedding of the association pair $l_i - d_j$ to the fully connected layer and the $softmax$ layer, and the score of association $s_{i,j} \in [0,1]$ between lncRNA l_i and disease d_j is obtained. The specific $s_{i,j}$ is expressed as follows:

$$s_{i,j} = softmax(Wx_{i,j} + b) \tag{10}$$

Where $W \in \mathrm{R}^{2\times 2d}$ is the parameter of the full connection layer and b is the bias; the larger the score of $s_{i,j}$, the greater the possibility of association between lncRNA l_i and disease d_j. In our model, the cross-entropy loss between prediction and real association is defined as follows:

$$Loss = -\sum_{i=1}^{T} y_i \log s_i \tag{11}$$

Where T is the number of training samples; s_i is the score of the association between lncRNA and disease of training sample; y_i is the label of real association between lncRNA and disease.

References

1. Kapranov, P. *et al.* Rna maps reveal new rna classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
2. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding rnas: insights into functions. *Nat. reviews genetics* **10**, 155–159 (2009).
3. Pasmant, E., Sabbagh, A., Vidaud, M. & Bièche, I. Anril, a long, noncoding rna, is an unexpected major hotspot in gwas. *The FASEB J.* **25**, 444–448 (2011).
4. Congrains, A. *et al.* Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of anril and cdkn2a/b. *Atherosclerosis* **220**, 449–455 (2012).
5. Zhang, Q., Chen, C.-Y., Yedavalli, V. S. & Jeang, K.-T. Neat1 long noncoding rna and paraspeckle bodies modulate hiv-1 posttranscriptional expression. *MBio* **4**, e00596–12 (2013).
6. Johnson, R. Long non-coding rnas in huntington's disease neurodegeneration. *Neurobiol. disease* **46**, 245–254 (2012).
7. Ji, P. *et al.* Malat-1, a novel noncoding rna, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**, 8031–8041 (2003).
8. Barsyte-Lovejoy, D. *et al.* The c-myc oncogene directly induces the h19 noncoding rna by allele-specific binding to potentiate tumorigenesis. *Cancer research* **66**, 5330–5337 (2006).
9. De Kok, J. B. *et al.* Dd3pca3, a very sensitive and specific marker to detect prostate tumors. *Cancer research* **62**, 2695–2698 (2002).
10. Bao, Z. *et al.* Lncrnadisease 2.0: an updated database of long non-coding rna-associated diseases. *Nucleic acids research* **47**, D1034–D1037 (2019).
11. Ning, S. *et al.* Lnc2cancer: a manually curated database of experimentally supported lncrnas associated with various human cancers. *Nucleic acids research* **44**, D980–D985 (2016).
12. Dinger, M. E. *et al.* Nred: a database of long noncoding rna expression. *Nucleic acids research* **37**, D122–D126 (2009).
13. Wang, Y. *et al.* Mammalian ncrna-disease repository: a global view of ncrna-mediated disease network. *Cell death & disease* **4**, e765–e765 (2013).
14. Lu, Z., BRETONNEL COHEN, K. & Hunter, L. Generif quality assurance as summary revision. In *Biocomputing 2007*, 269–280 (World Scientific, 2007).
15. Sun, J. *et al.* Inferring novel lncrna–disease associations based on a random walk model of a lncrna functional similarity network. *Mol. BioSystems* **10**, 2074–2081 (2014).
16. Gu, C. *et al.* Global network random walk for predicting potential human lncrna-disease associations. *Sci. reports* **7**, 1–11 (2017).
17. Wen, Y., Han, G. & Anh, V. V. Laplacian normalization and bi-random walks on heterogeneous networks for predicting lncrna-disease associations. *BMC systems biology* **12**, 11–19 (2018).
18. Zhang, J., Zhang, Z., Chen, Z. & Deng, L. Integrating multiple heterogeneous networks for novel lncrna-disease association inference. *IEEE/ACM transactions on computational biology bioinformatics* **16**, 396–406 (2017).

19. Chen, X. & Yan, G.-Y. Novel human lncrna–disease association inference based on lncrna expression profiles. *Bioinformatics* **29**, 2617–2624 (2013).

20. Chen, X. *et al.* Constructing lncrna functional similarity network based on lncrna-disease associations and disease semantic similarity. *Sci. reports* **5**, 1–12 (2015).

21. Zhao, T. *et al.* Identification of cancer-related lncrnas through integrating genome, regulome and transcriptome features. *Mol. BioSystems* **11**, 126–136 (2015).

22. Lan, W. *et al.* Ldap: a web server for lncrna-disease association prediction. *Bioinformatics* **33**, 458–460 (2017).

23. Xuan, P., Pan, S., Zhang, T., Liu, Y. & Sun, H. Graph convolutional network and convolutional neural network based method for predicting lncrna-disease associations. *Cells* **8**, 1012 (2019).

24. Wu, X. *et al.* Inferring lncrna-disease associations based on graph autoencoder matrix completion. *Comput. biology chemistry* **87**, 107282 (2020).

25. Zhang, J., Jiang, Z., Hu, X. & Song, B. A novel graph attention adversarial network for predicting disease-related associations. *Methods* **179**, 81–88 (2020).

26. Zhang, C., Song, D., Huang, C., Swami, A. & Chawla, N. V. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 793–803 (2019).

27. Lu, C. *et al.* Prediction of lncrna–disease associations based on inductive matrix completion. *Bioinformatics* **34**, 3357–3364 (2018).

28. Fu, G., Wang, J., Domeniconi, C. & Yu, G. Matrix factorization-based data fusion for the prediction of lncrna–disease associations. *Bioinformatics* **34**, 1529–1537 (2018).

29. Xuan, P., Cao, Y., Zhang, T., Kong, R. & Zhang, Z. Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncrna genes. *Front. genetics* **10**, 416 (2019).

30. Yang, G., Lu, X. & Yuan, L. Lncrna: a link between rna and cancer. *Biochimica et Biophys. Acta (BBA)-Gene Regul. Mech.* **1839**, 1097–1109 (2014).

31. Li, J.-H., Liu, S., Zhou, H., Qu, L.-H. & Yang, J.-H. starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic acids research* **42**, D92–D97 (2014).

32. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C.-F. A new method to measure the semantic similarity of go terms. *Bioinformatics* **23**, 1274–1281 (2007).

33. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780 (1997).

34. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. neural information processing systems* **26** (2013).

35. Veličković, P. *et al.* Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61862067), the Applied Basic Research Project in Yunnan Province (No.202101AT070132) and the NSFC-Yunnan Union Key Grant(No.U1902201).

Author contributions statement

L.T, L.L and H.S conceived the presented idea. H.S and X.M.Z carried out the experiment and wrote the draft. L.T and LL helped shape the research, analysis and manuscipt. All authors discussed the results and contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Data availability

The original datasets of our study was download from another lncRNA-disease association prediction study, the orginal datasets were available at https://github.com/ydkvictory/RFLDA. The processed data along with coded are available at https://github.com/hongshi940/HGNNLDA

Additional information

Correspondence and requests for materials should be addressed to L.L.