Computational Method Using Heterogeneous Graph Convolutional Network Model Combined With Reinforcement Layer For MiRNA-Disease Association Prediction

Dang Huang
China University of Mining and Technology
JiYong An (ajy@cumt.edu.cn)
China University of Mining and Technology
Lei Zhang
China University of Mining and Technology
BaiLong Liu
China University of Mining and Technology

Research Article

Keywords: miRNA and disease interactions, Graph Convolutional Network

Posted Date: September 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-737865/v1

Computational method using Heterogeneous Graph Convolutional Network Model Combined with Reinforcement Layer for MiRNA-disease association prediction

Dang Huang[1], JiYong An[1,*], Lei Zhang[1,*], BaiLong Liu[1]
[1]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou Jiangsu 21116, China
[*]Corresponding author: ajy@cumt.edu.cn, zhanglei@cumt.edu.cn

Abstract

Background: A large number of evidences from biological experiments have confirmed that miRNAs play an important role in the progression and development of various human complex diseases. However, the traditional experiment methods are expensive and time-consuming. Therefore, it is a challenging task that how to develop more accurate and efficient methods for predicting potential associations between miRNA and disease.

Results: In the study, we developed a computational model that combined Heterogeneous Graph Convolutional Network with Enhanced Layer for miRNA-Disease Association prediction (HGCNELMDA). The major improvement of our method lies in through restarting the random walk optimized the original features of nodes and adding a Reinforcement layer to the hidden layer of graph convolutional network retained similar information between nodes in the feature space. In addition, the proposed approach recalculated the influence of neighborhood nodes on target nodes by introducing the attention mechanism. The reliable performance of the HGCNELMDA was certified by the AUC of 93.47% in global leave-one-out cross-validation (LOOCV), and the average AUCs of 93.01% in fivefold cross-validation. Meanwhile, we compared the HGCNELMDA with the state-of-the-art methods. Comparative results indicated that o the HGCNELMDA is very promising and may provide a cost-effective alternative for miRNA-Disease Association prediction. Moreover, we applied HGCNELMDA to 3 different case studies to predict potential miRNAs related to lung cancer, prostate cancer, and pancreatic cancer. Results showed that 48, 50, and 50 of the top 50 predicted miRNAs were supported by experimental association evidence. Therefore, the HGCNELMDA is a reliable method for predicting disease-related miRNAs.

Conclusions: The results of the HGCNELMDA method in the LOOCV (leave-one-out cross validation, LOOCV) and 5-cross validations were 93.47% and 93.01%, respectively. Compared with other typical methods, the performance of HGGCNMA is higher. Three cases of lung cancer, prostate cancer, and pancreatic cancer were studied. Among the predicted top 50 candidate miRNAs, 48, 50, and 50 were verified in the biological database HDMMV2.0.Therefore; this further confirms the feasibility and effectiveness of our method. To facilitate extensive studies for future disease-related miRNAs research, we developed a freely available web server called HGCNELMDA is available at http://132.232.17.50:8080/HGCNELMDA.jsp.

KEYWORD: miRNA and disease interactions, Graph Convolutional Network

1、Background

As a kind of non-coding RNA with regulatory properties and highly conserved in the evolutionary process, miRNA is approximately 20-24 nucleotides in length. Researchers that have been studying miRNA [1] have found that it plays a vital role in biological processes such as cell growth, proliferation, metabolism, differentiation and apoptosis. Moreover, the abnormal expression of miRNA has also been proved to be closely related to some diseases, such as chronic lymphocytic leukemia, tumor, gastric cancer, cardiomyopathy, etc. Therefore, identifying the correlation between miRNA and diseases has become a critical step in biological research recently [2]. However, the traditional biological experiments take up a long time, cost much, and have some blindness, all of which would stall the research process. Therefore, many researchers are devoted to designing computational methods to discover the interaction between unidentified miRNAs and diseases to make up for the shortcomings

of traditional experimental approaches [3].

Currently, researchers have established a series of effective calculation models for miRNA-disease association prediction, which can be roughly divided into two categories according to the methods used: similarity measurement-based and machine learning-based. For similarity measurement [4], the miRNA-disease association is predicted by measuring the degree of similarity between nodes using different statistical methods. The machine-learning approach trains other models by learning features and then predicting miRNA-disease associations based on the trained models. The above two methods have different theoretical bases and innovations, and thus making outstanding contributions to future research. For example, Jiang et al. [5] determined the functional correlation of two miRNAs by calculating the number of familiar neighbors and the shortest path length of two miRNAs and constructing two miRNAs' functional correlation information. For the first time, Jiang et al. combined disease phenotype information with miRNA function information to predict miRNA-disease association [6], contributing significantly to the future research. Subsequently, for each predicted disease, they designed a hypergeometric distribution-based scoring system [7] to score the diseases and all of the miRNAs associated with them. However, this method comes with some limitations too. Because only the direct neighbors of the miRNA were considered as the criterion for miRNA functional similarity score, the prediction effect was limited. To increase the accuracy of miRNA-disease association prediction, Xuan et al. [8] proposed the weighted k-nearest neighbor method (HDMP). They suggested that members of the same miRNA family may be involved in diseases with related phenotypes. According to the association state of the nearest neighbor [9], members of the miRNA family and miRNA cluster can obtain more weight, which improves the prediction performance of the model to some extent. However, it is difficult to manually select the optimal parameter K that classifies the number of members in each miRNA family and miRNA cluster [10], and this method cannot predict new diseases that do not have known miRNA associations. Pasquier et al. [11] formed a matrix with higher dimensions based on miRNA-disease association, miRNA target association, miRNA word association, miRNA family association and miRNA neighbor association state data. Using the singular value matrix decomposition method to decompose the matrix, Pasquier et al. successfully obtained miRNA vectors and disease vectors [18]. They took the cosine distance between the miRNA node vector and the disease node vector as the degree of association between the nodes. However, due to the false-positive rate and false-negative rate between miRNA and target, the model's prediction performance is affected to a certain extent.

In addition to similarity-based approaches, machine learning algorithms aiming at exploring potential miRNA disease interactions are also an essential academic approach in this field. Unlike the method of directly calculating the similarity between nodes in the network based on similarity itself, the machine learning approach [19] is devoted to extracting inherent features and designing practical classification algorithms to find miRNA and disease associations. As an early method based on machine learning, Jiang et al. [20] first extracted feature vectors from disease similarity and miRNA function similarity. Then, they randomly selected 270 samples from unknown miRNA disease pairs as negative data, as missing negative instances in the actual data set [21]. Finally, they chose the SVM (Support Vector Machine) as the classifier [22]. However, this artificial method randomly selected negative samples, impacting on the model's accuracy. A different approach conducted by Chen et al. [23] constructed a semi-supervised classifier with regularized least squares. Although the model does not require negative samples, and the possibility of unknown associations is confirmed, this method also has some limitations: the predicted results of fusion miRNA and disease are strongly dependent on parameters [24], and thus it is difficult to choose the optimal parameters. Chen et al. [25] proposed the DRMDA method to use stacked autoencoders for feature extraction to obtain low-dimensional and high-resolution feature vectors and then used SVM to score candidate miRNAs. This method eliminated a lot of noise in similar unprocessed data and achieved good performance results. Graph neural network has attracted extensive attention from researchers due to its high precision. Also, biological information networks such as disease and miRNA have complex topological structures, so it is suitable for graphical modelling [26]. For graph data, graph convolutional networks (GCN) have better performance than inhomogeneous networks (such as classification). Therefore, researchers have been trying to apply GCN in heterogeneous networks to predict the association between miRNA and disease [27]. For example, Li et al. [28] extracted node features from the protein-protein interaction network and put them into the graph convolutional network following the

Node2VEC algorithm. Finally, each node was embedded in the graph convolutional layer, and the miRNA-disease association was obtained by multiplying the miRNA-gene adjacency matrix by the disease-gene adjacency matrix [29]. This method provides a new perspective for the field of miRNA-disease association prediction. Then, Li et al. [30] proposed the FCGCNMDA method based on a fully connected graph. They extracted the aggregation of node features by using a two-layer graph convolution layer in miRNA functional similarity network and disease semantic similarity network to make end-to-end prediction [31]. However, the GCN model considers all neighbors equally, and the similarity information of nodes cannot be retained when learning node embedding.

Although the existing methods have good performances in predicting miRNA-disease associations, we can still improve some aspects of them. On the one hand, some methods produce inevitable data noise during feature extraction, affecting the prediction effect. On the other hand, some graph convolution methods fail to retain the similarity information of nodes so that similar nodes have similar feature representations in the feature space to enhance the spatial node features of the topology graph [36]. This paper is based on strengthening layer figure convolution heterogeneous network model HGCNELMDA (Heterogeneous Graph Convolutional Network model with Enhanced Layer to predict miRNA – Disease Associations) to extract node features from the level of the graph. To reduce the data noise of the similarity matrix calculation, the random reboot walk is used to get the original features of nodes from the similarity matrix. Graph convolution aggregates node information according to edge information and represents new node features. Before the figure of convolution model, GCN (Graph Convolutional Network) will consider all equal neighbors, and thus being unable to retain when learning node embedded nodes similarity information. The enhancement layer added in the GCN hidden layer is used to strengthen the similar representation of similar nodes (miRNAs or diseases) in the feature space and enhance the eigenvector aggregation of similar nodes to retain similar information between nodes. First, we constructed an miRNA-disease heterogeneous network based on the proven miRNA-disease association, disease semantic similarity and miRNA functional similarity. Second, to reduce the data noise of extracting the original feature vectors of miRNA and disease nodes and better capture the structural relationship between different types of nodes in heterogeneous graphs, the method based on restart random walk is used for extracting node features from similarities. Third, the miRNA-disease heterogeneous graph and the miRNA-disease feature matrix are gathered through graph convolution to gather the information of neighbor nodes on the layer, and an attention-based reinforcement layer is added to the hidden layer. In the miRNA-disease heterogeneous graph, to strengthen similar nodes (miRNA or disease) for similar representations in the feature space, a reinforcement layer is added to the GCN hidden layer, enhancing the feature vectors of similar aggregate retain similar information between nodes. The attention mechanism is introduced in the reinforcement layer, and more critical topological neighborhood nodes are merged, and miRNA and disease node features are extracted from the spatial topological structure of heterogeneous graphs to predict associations. The results of the HGCNELMDA method in LOOCV (Leave-One-Out cross-validation) and 5-fold cross-validations were 93.47% and 93.01%, respectively. Compared with other typical methods, the HGGCNMA has a better performance. Four cases of lung cancer, prostate cancer and pancreatic cancer were used for research. Among the predicted top 50 candidate miRNAs, 48, 50, and 50 were verified in the biological database HDMI V2.0. Therefore, the result further confirms the feasibility and effectiveness of our method.

2、Results

First, we present the experimental methods and evaluation indexes. The performance of the HGCNELMDA approach is then compared with the following four existing approaches. Finally, we used the HGCNELMDA method to determine the accuracy of the predictive association based on three cases of prostate tumor, lung tumor and pancreatic tumor.

2.1 Experimental Approaches and Evaluation Criteria

We collected 5430 known miRNA-disease associations from HMDD V2.0 as the experimental data set. Based on experimentally verified associations between miRNAs and diseases, we implemented global LOOCV and 5-fold CV to evaluate the predictive accuracy of HGCNELMDA. In LOOCV evaluation, every confirmed association was regarded as a test sample in turn, while the rest associations were treated as training samples. Candidate samples included all of the miRNA-disease pairs that

experimental studies had not verified. After executing HGCNELMDA, every miRNA-disease pair will obtain an association score. A higher score means a higher likelihood for a link to exist between a pair. In global LOOCV, we compared the score of the test sample with the scores of all the candidate samples. Furthermore, we drew Receiver Operating Characteristics (ROC) curve by plotting the actual positive rate (TPR, sensitivity) against the false positive rate (FPR, 1-specificity) at different thresholds. Sensitivity denotes the percentage of miRNA-disease test samples with ranks exceeded the given point, while specificity represents the percentage of negative miRNA-disease associations with ranks lower than the threshold. AUC was further calculated to demonstrate the prediction ability of HGCNELMDA. The model has perfect prediction performance when AUC reaches exactly 1. If AUC equals 0.5, it suggests that the model only has random prediction performance.

Moreover, we exploited 5-fold CV to examine the predictive accuracy further. 5-fold cross-validation was also implemented to further estimate the prediction accuracy of the HGCNELMDA model by randomly dividing the known associations equally into five groups and treating each one of them as test samples in turn by removing the associations of the current test samples simultaneously. Afterwards, every test sample would be scored and compared with the candidate miRNA–disease pairs to obtaining the rankings. We repeated this procedure 50 times to get a more accurate average AUC value.

2.2 Compare with other methods

In order to verify the accuracy of our method, the HGCNELMDA method was compared with the following four existing methods, namely FCGCNMDA[32], CNMDA[33], EDTMDA[34] and RKNNMDA[35], for five-fold cross-validation.As shown in Table 1, the AUC of FCGCNMDA, CNMDA, EDTMDA and RKNNMDA were 92.85%, 85.33%, 91.92% and 82.21%, respectively. Among them, the AUC of HGCNELMDA was the highest under five-fold cross-validation, with a value of 93.01%.Therefore, HGCNELMDA was proved to be reliable in miRNA-disease association.

Table 1 Comparison of HGCNELMDA and other models for five-fold cross-validation

Control group	AUC（%）
HGCNELMDA	**93.01**
FCGCNMDA	92.85
CNMDA	85.33
EDTMDA	91.92
RKNNMDA	82.21

2.3 Comparison of results with or without reinforcement layer

Fig. 1 and Fig. 2 respectively show the influence of HGCNELMDA on the model performance with or without reinforcement layer under one-fold cross-validation and five-fold cross-validation. In the experiment, the reinforcing layer is removed and replaced by the common hidden layer of GCN.The results showed that the AUC value with the reinforcement layer was higher than that without the hidden layer, because the similar miRNA(or disease) nodes in the reinforcement layer were similar in the feature space, and the attention mechanism was used to focus on the aggregation of similar important neighbor nodes in the reinforcement layer, and the similar information of nodes was retained.

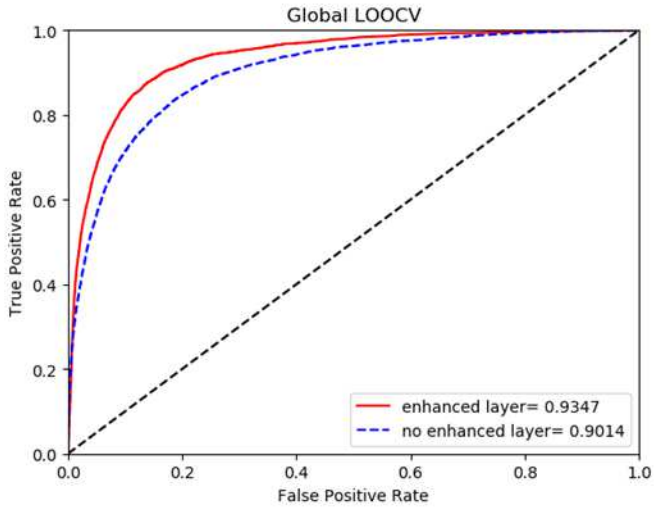

Fig. 1 Comparison of left cross-validation with or without reinforcement layer

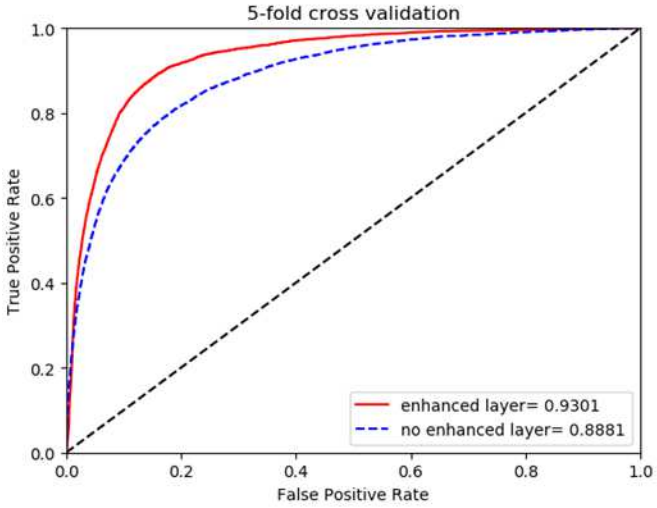

Fig. 2 Comparison of 50 fold cross validation with or without reinforcement layer

2.4 Comparison of results with or without Random Walk With Restart

Fig. 3 and Fig. 4 respectively show the influence of HGCNELMDA on the results by using RWR to extract node features under one-fold and five-fold cross validation. No experiments using RWR were used directly SM and SD a row or a column of is used as the eigenmatrix of nodes. As shown in the figure, it is better to use RWR as the initial feature of the node, because RWR can select adjacent nodes to travel or return to the initial node, thus reducing the influence of data noise in node feature extraction.

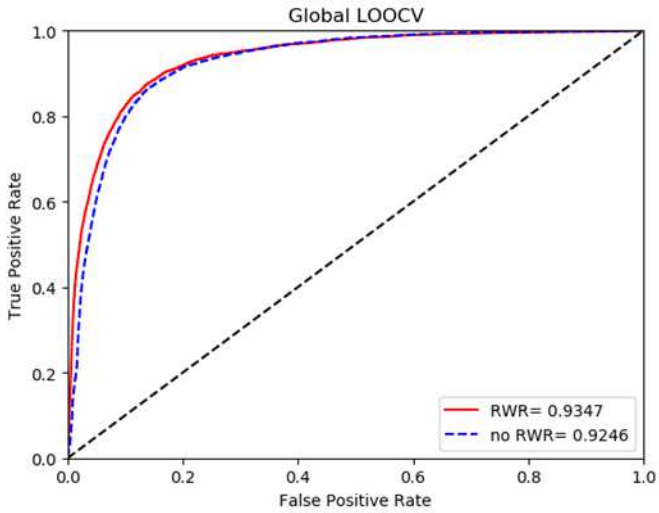

Fig. 3 Comparison of residual cross-validation with or without RWR

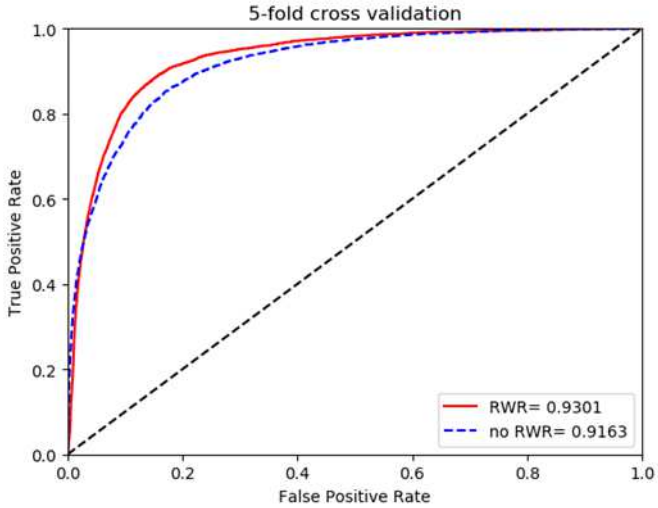

Fig. 4 Comparison of 50% fold cross validation with or without RWR

2.5 Comparison of parameter sensitivities

Layer node embedding dimension is the node embedding parameter in GCN hidden layer h , Different parameter values will affect the experimental results. As shown in Figure 5, define h as [32, 64, 128, 256, 512], Compared with the AUC results, The validation methods of one-left cross-validation and five-fold cross-validation show that the AUC value presents an upward trend with the increase of node embedding dimension h .The performance of the HGCNELMDA approach is highest when the embedding dimension h is defined as 256.

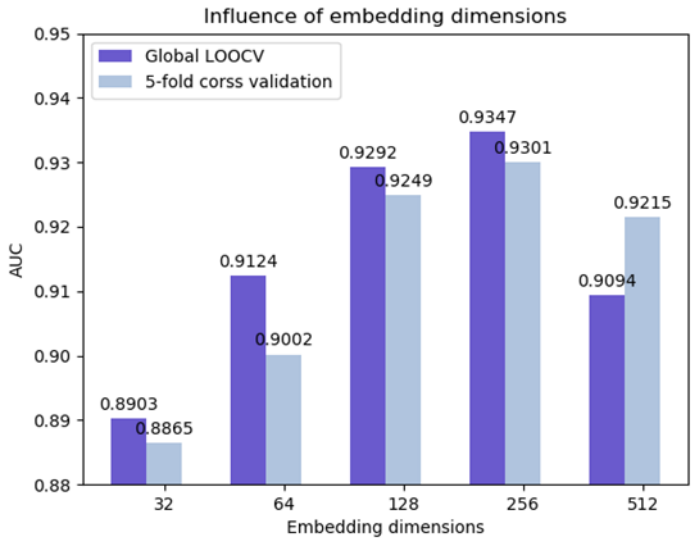

Fig. 5 Comparison of different embedded dimensions

2.6 Cases Studies

The HGCNELMDA method was used to determine the accuracy of the predictive association based on three cases of prostate cancer, lung cancer and pancreatic cancer. We compared the predicted candidate miRNAs with DBDEMC and Phenomir, two public databases, to verify their accuracy.

In the first case study, the selected prostate tumors are used to test whether our approach is suitable for novel diseases with unsupported miRNAs or not. This case selected prostate tumors because this is the most common cancer happening on males worldwide. In 2018, more than 100,000 males died of prostate cancer in Europe alone [37]. This case study first set all miRNA-disease associations related to prostate neoplasms from HMDD 2.0 to zero. Then, M2GMDA was performed to identify the associated miRNAs for prostate neoplasms. Table 2 lists the top 50 candidate miRNAs for HGCNELMDA prediction associated with prostate tumors. The first 50 miRNAs were verified by DBDEMC and Phenomir databases. The results show that the above two databases could verify the first 50 miRNAs.

Researchers found that the second-ranked HAS-miR-96b was found to regulate apoptosis of prostate cancer cells by inhibiting the FoxO1 transcription factor, indicating that the HGCNELM subsequently validates the predictive ability of HGCNELMDA in new diseases without any known linked miRNAs. To further investigate, we set up a special case study. In this case, we examined HGCNELMDA on Lung Neoplasms, a common human cancer with many experimentally verified related miRNAs. We utilized the experimentally verified miRNA-disease associations from the HMDD v2.0 database as the initial training set. However, we removed all the associations, including lung neoplasms, from the training set this time. Hence, lung neoplasms could be regarded as a disease without any known related miRNAs. Lung tumors are devastating and fatal, causing many deaths in both males and females worldwide [38]. The survival rate of lung tumors is as low as five years, so early diagnosis is critical to save patients' lives [39]. Therefore, lung tumors, in which miRNAs have become a promising tool in diagnosing and treating process, were selected in this case. HGCNELMDA is used to predict candidate miRNAs associated with lung tumors. The validations of the first 50 related miRNAs are listed in Table 3. Two databases confirmed 49 miRNAs, and only one miRNA was not verified. In addition, the ectopic expression of miR-494-3p in A549 lung cancer cells promoted the tumor-initiating population and enhanced the motor ability of cancer cells and the expression of stem cell-related genes, suggesting that HGCNELMDA can help the diagnosis and treatment of lung tumors. HGCNELMDA method has good accuracy in predicting prostate tumor-associated miRNA.

Table 2 Top 50miRNAs associated with prostate tumors

miRNA	dbDEMC	PhenomiR	miRNA	dbDEMC	PhenomiR
hsa-mir-10a	confirmed	confirmed	hsa-mir-297	confirmed	confirmed
hsa-mir-96b	confirmed	confirmed	hsa-mir-23a	confirmed	confirmed
hsa-mir-186	confirmed	confirmed	hsa-mir-27a	confirmed	confirmed
hsa-mir-194	confirmed	confirmed	hsa-mir-33b	confirmed	confirmed
hsa-mir-15a	confirmed	confirmed	hsa-mir-19a	confirmed	confirmed
hsa-mir-26b	confirmed	confirmed	hsa-mir-1	confirmed	confirmed
hsa-let-7d	confirmed	confirmed	hsa-mir-27b	confirmed	confirmed
hsa-mir-20a	confirmed	confirmed	hsa-mir-218	confirmed	confirmed
hsa-mir-301a	confirmed	confirmed	hsa-let-7e	confirmed	confirmed
hsa-mir-363	confirmed	Not confirmed	hsa-mir-373	confirmed	confirmed
hsa-mir-23b	confirmed	confirmed	hsa-mir-16	confirmed	confirmed
hsa-mir-92	confirmed	confirmed	hsa-mir-197	confirmed	confirmed
hsa-mir-302d	confirmed	confirmed	hsa-mir-181b	confirmed	confirmed
hsa-mir-195	confirmed	confirmed	hsa-mir-23b	confirmed	confirmed
hsa-mir-130b	confirmed	confirmed	hsa-mir-101	confirmed	confirmed
hsa-let-7i	confirmed	confirmed	hsa-mir-26a	confirmed	confirmed
hsa-let-7c	confirmed	confirmed	hsa-mir-17	confirmed	confirmed
hsa-mir-92a	confirmed	confirmed	hsa-mir-146a	confirmed	confirmed
hsa-mir-184	confirmed	confirmed	hsa-mir-182	confirmed	confirmed
hsa-mir-130a	confirmed	confirmed	hsa-mir-122	confirmed	confirmed
hsa-mir-155	confirmed	confirmed	hsa-mir-93	confirmed	confirmed
hsa-mir-20b	confirmed	confirmed	hsa-mir-10b	confirmed	confirmed
hsa-mir-29a	confirmed	confirmed	hsa-mir-31	confirmed	confirmed
hsa-mir-191	confirmed	confirmed	hsa-let-7g	confirmed	confirmed
hsa-mir-137	confirmed	confirmed	hsa-mir-181d	confirmed	confirmed

Table 3 Top 50miRNAs associated with lung tumors

miRNA	dbDEMC	PhenomiR	miRNA	dbDEMC	PhenomiR
hsa-mir-320a	confirmed	confirmed	hsa-mir-28	confirmed	confirmed
hsa-mir-494	confirmed	confirmed	hsa-mir-141	confirmed	confirmed
hsa-mir-23b	confirmed	confirmed	hsa-mir-329	confirmed	Not confirmd
hsa-mir-15a	confirmed	confirmed	hsa-mir-320e	confirmed	Not confirmd
hsa-mir-107	confirmed	confirmed	hsa-mir-378	confirmed	confirmed
hsa-mir-122	confirmed	confirmed	hsa-mir-15b	confirmed	confirmed
hsa-mir-422a	confirmed	confirmed	hsa-mir-371	confirmed	confirmed
hsa-mir-377	confirmed	confirmed	hsa-mir-153	confirmed	confirmed
hsa-mir-383	confirmed	confirmed	hsa-mir-663	Not confirmed	confirmed
hsa-mir-141	confirmed	confirmed	hsa-mir-374b	confirmed	confirmed
hsa-mir-342	confirmed	confirmed	hsa-mir-584	confirmed	confirmed
hsa-mir-425	confirmed	confirmed	hsa-mir-202	confirmed	confirmed
hsa-mir-377	confirmed	confirmed	hsa-mir-10a	confirmed	confirmed

miRNA	dbDEMC	PhenomiR	miRNA	dbDEMC	PhenomiR
hsa-mir-423	confirmed	confirmed	hsa-mir-16	confirmed	confirmed
hsa-mir-130b	confirmed	confirmed	hsa-mir-181d	confirmed	confirmed
hsa-mir-328	confirmed	confirmed	hsa-mir-129	confirmed	confirmed
hsa-mir-515	Not confirmd	Not confirmd	hsa-mir-147b	confirmed	confirmed
hsa-mir-320d	confirmed	confirmed	hsa-mir-410	Not confirmed	confirmed
hsa-mir-323b	confirmed	Not confirmd	hsa-mir-421	confirmed	confirmed
hsa-mir-92	confirmed	confirmed	hsa-mir-189	confirmed	Not confirmed
hsa-mir-105	confirmed	confirmed	hsa-mir-17	confirmed	confirmed
hsa-mir-34c	confirmed	confirmed	hsa-mir-99a	confirmed	confirmed
hsa-mir-187	confirmed	confirmed	hsa-mir-20b	confirmed	confirmed
hsa-mir-149	confirmed	confirmed	hsa-mir-92	confirmed	confirmed
hsa-mir-124a	confirmed	confirmed	hsa-mir-302d	confirmed	confirmed

For the third disease case we chose pancreatic tumor as the new disease case. When the known miRNA and disease association matrix is set to zero, the column of pancreatic tumor indicates that no related miRNA is associated with it, as a new disease [40].HGCNELMDA is used to predict candidate miRNAs associated with pancreatic tumors, and the top 50 related miRNAs are listed in Table 4.The DBDEMC and Phenomir databases validated the first 50miRNAs.Studies have shown that increased serum miR-193b is a potential new biomarker for pancreatic neuroendocrine tumors (PNEN).The results indicate that HGCNELMDA plays an important role in predicting new diseases

Table 4 Top 50miRNAs associated with pancreatic tumors

miRNA	dbDEMC	PhenomiR	miRNA	dbDEMC	PhenomiR
hsa-mir-18a	confirmed	confirmed	hsa-mir-199a	confirmed	confirmed
hsa-let-7a	confirmed	confirmed	hsa-mir-210	confirmed	confirmed
hsa-mir-193b	confirmed	confirmed	hsa-mir-34c	Not confirmed	confirmed
hsa-mir-155	confirmed	confirmed	hsa-mir-15a	confirmed	confirmed
hsa-mir-143	confirmed	confirmed	hsa-let-7c	confirmed	Not confirmed
hsa-mir-19a	confirmed	confirmed	hsa-mir-29c	confirmed	confirmed
hsa-mir-29a	confirmed	confirmed	hsa-mir-9	confirmed	confirmed
hsa-mir-200c	confirmed	confirmed	hsa-mir-200a	confirmed	confirmed
hsa-mir-200b	confirmed	confirmed	hsa-mir-146b	confirmed	confirmed
hsa-mir-31	confirmed	confirmed	hsa-mir-182	confirmed	confirmed
hsa-mir-21	confirmed	confirmed	hsa-mir-181b	confirmed	confirmed
hsa-mir-155	confirmed	Not confirmed	hsa-let-7d	confirmed	confirmed
hsa-mir-146a	confirmed	confirmed	hsa-mir-30a	confirmed	confirmed
hsa-mir-17	confirmed	confirmed	hsa-mir-142	confirmed	confirmed
hsa-mir-145	confirmed	confirmed	hsa-mir-106b	confirmed	confirmed
hsa-mir-20a	confirmed	confirmed	hsa-mir-218	Not confirmed	confirmed
hsa-mir-34a	confirmed	confirmed	hsa-mir-223	confirmed	confirmed
hsa-mir-125b	confirmed	confirmed	hsa-let-7b	confirmed	confirmed
hsa-mir-126	confirmed	confirmed	hsa-let-7e	confirmed	confirmed
hsa-mir-221	Not confirmed	confirmed	hsa-mir-34b	confirmed	confirmed
hsa-mir-92a	confirmed	confirmed	hsa-mir-205	confirmed	confirmed
hsa-mir-16	confirmed	confirmed	hsa-mir-7	confirmed	confirmed

miRNA	dbDEMC	PhenomiR	miRNA	dbDEMC	PhenomiR
hsa-mir-222	confirmed	confirmed	hsa-mir-148a	confirmed	confirmed
hsa-mir-181a	confirmed	confirmed	hsa-mir-195	Not confirmed	confirmed
hsa-mir-29b	confirmed	confirmed	hsa-mir-133b	confirmed	confirmed

For the results of the four case studies, our method was effective when predicting unvalidatedmiRNA and disease interactions

3、Discussion

Compared with five classic methods based on Global LOOCV and 5-fold cross-validation, the experimental results show that HGCNELMDA has better predictive performance. In addition, three case studies also support the results of our method. First, we constructed a heterogeneous network of miRNA-disease based on the proven miRNA-disease association, disease semantic similarity and miRNA functional similarity. Second, we used the restart random walk method to extract node features from similarity, aiming at reducing the data noise of extracting the original feature vectors of miRNA and disease nodes and better capturing the structural relationship between different types of nodes in the heterogeneous graph. In the miRNA-disease heterogeneous graph, to reinforce that similar nodes (miRNAs or diseases) have identical representations in the feature space, a reinforcement layer was added to the GCN hidden layer, enhancing the eigenvector aggregation of similar nodes, to preserve similar information between nodes. The attention mechanism was introduced in the reinforcement layer, more important topological neighborhood nodes were integrated, and miRNA and disease node features were extracted from the spatial topology of heterogeneous graphs to predict associations. In summary, the HGCNELMDA method makes full use of the complex structure and semantic information of the miRNA-disease heterogeneous network to achieve good predictions.

4、Conclusion

This paper mainly describes the enhancement layer based heterogeneous graph convolutional network model (HGCNELMDA) to predict miRNA-disease association method. First, by restarting the random walk between the miRNA and the disease phase
The eigenvectors of miRNA and disease nodes were obtained from the similarity network. Secondly, the heterogeneous graph of miRNA-disease was input into GCN, and a reinforcement layer was added into the hidden layer of GCN to make similar nodes have similar feature representations in the feature space. The attention mechanism was used to update the influence of important adjacent nodes in the reinforcement layer on the target node. Thirdly, the association matrix between miRNA and disease was reconstructed by bilinear encoder, and the cross-entropy loss function was used to train the model. Finally, HGCNELMDA performance was evaluated by four sets of experiments, which were left one-fold cross-validation and five-fold cross-validation, compared with other methods, ablation test, parameter sensitivity test and three disease case studies. The results indicated that HGCNELMDA method had a good predictive effect in the prediction of miRNA-disease association.

5、Methods

In order to reduce the data noise of extracting original features, make similar nodes have similar feature representation in feature space, and enhance the spatial node feature aggregation of topology map, this paper constructs a heterogeneous graph convolutional network model based on reinforcement layer to predict miRNA-disease association. The model framework is shown in Figure 6.

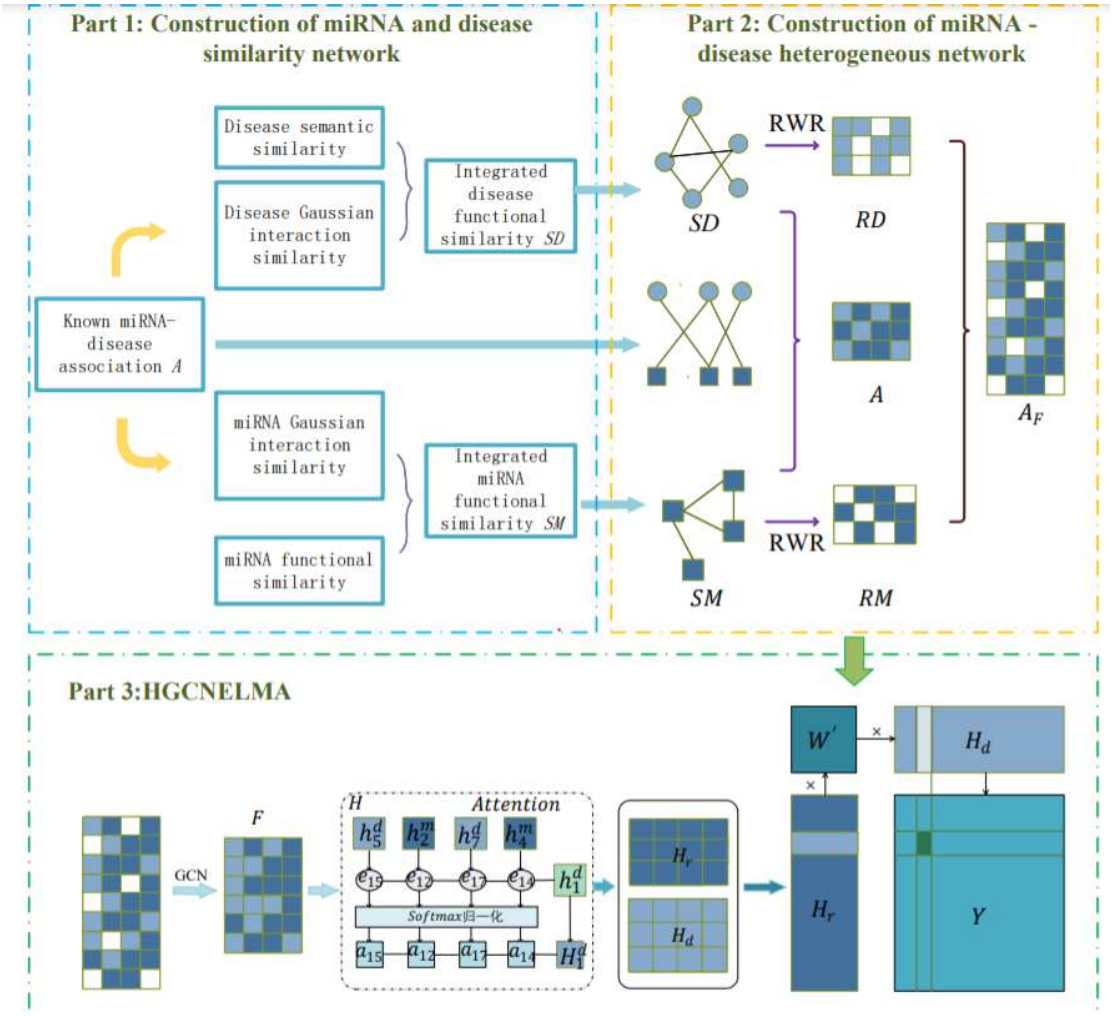

Fig. 6 HGCNELMDA model

(1) Step 1: Build miRNA-disease isomerization map according to literature [41].Through integrated disease semantic similarity network SD, The known miRNA-disease association matrix is the same A and an integratedmiRNA functional similarity network SM constructed into a miRNA-disease heterogeneous map A_H, as shown in Formula (1) :

$$A_H = \begin{bmatrix} SM & A \\ A^T & SD \end{bmatrix} \quad (1)$$

Among them, $A_H \in \mathbb{R}^{(m+n)\times(m+n)}$, m and n are respectively the number of miRNA and disease.

(2) Step 2: node feature extraction based on restart random walk. In order to reduce the influence of data noise on the original features of nodes, restart the random walk is used to represent the original features of nodes.

(3) Step 3: node embedding based on GCN.The information of neighbor nodes of each layer is aggregated through GCN to form an embedding of miRNA or disease node features.

(4) Step 4: reinforcement layer based on attention mechanism. Since the previous GCN considered neighbor nodes equally, the text adds an attention-based reinforcement layer to the GCN hidden layer H.

(5) Step 5: Use the decoder to reconstruct the association matrix between miRNA and disease. The node feature embedding matrix is obtained by the reinforcement layer$H$, The Eigen matrix of miRNA is $H_R \in \mathbb{R}^{m\times h}$, The characteristic matrix of disease is $H_d \in \mathbb{R}^{n\times h}$, h is layer embedding dimension, Since $sigmod$ is often used as the activation function of dichotomy, It can be used to reconstruct miRNA-disease association matrix Y, As shown in Equation (2) :

$$Y = sigmod(H_r W' H_d^T) \quad (2)$$

Where, The element in the row of the matrix Y represents the correlation prediction score $y_{ij}{}'$ between miRNA r_i and diseased_j, $W' \in \mathbb{R}^{X \times X}$ is a trainable matrix.

(6) Step 6: In order to make the predicted results as close as possible to the actual results, cross entropy is used as the loss function to carry out end-to-end back propagation, as shown in Formula (3) :

$$\mathcal{L}_{cross} = -\sum_{i,j \in y \cup y^-} (y_{ij} log\, y_{ij}{}' + (1 - y_{ij}) \log(1 - y_{ij}{}')) \quad (3)$$

Among them, y represents an associated miRNA-disease positive sample,y^- represents a negative sample with an unknown relationship.

Feature extraction based on Random Walk with Restart

The M2GMDA and CEMDA methods assign each row or column in the SM (or SD) similarity matrix to represent an eigenvector of amiRNA(or disease). Literature [39] believes that the limitation of similarity calculation method may lead to some data noise in the direct extraction of original node features. Therefore, in order to optimize the original feature vectors of miRNA and disease nodes and better capture the structural relationship between different types of nodes in heterogeneous graphs, the text reference [40] uses a method based on R(Random Walk With Restart ,RWR) to extract node features from similarity. Restarting the random walk starts from a node, and each step can randomly select adjacent nodes or return to the starting node. Assume that there are n nodes, Right at the start node e, Then the probability of appearing at any node i in the next move is r, as shown in Formula (4) :

$$r^0[i] = W[i] \cdot e \quad (4)$$

Here, $W[i]$ represents the i row of the transition probability matrix W, that is, the transition probability from all nodes to node i, In the next move, the probability of the node i is shown in formula (5) :

$$r^1[i] = W[i] \cdot r^0 \quad (5)$$

After considering restarting, after k iterations, it still returns to node i with probability$c$. After the $k+1$ iteration is stable, r_i is the probability score of reaching node i, which is the similarity feature vector of node i, as shown in formula (6):

$$r_i^{k+1} = cWr_i^k + (1 - c)e_i \quad (6)$$

Here,$c \in (0, 1)$ represents the restart probability,$W[i,j] \in \mathbb{R}^{n \times n}$ represents the probability from i to j ,and $e_i \in \mathbb{R}^{n \times 1}$ is the i-th node Initial probability vector. If i is equal to j, then e_{ij} is 1, otherwise it is 0. This paper replaces W with SM or SD respectively, and obtains the probability distribution matrix of the node (miRNA or disease) based on the restart random walk, and normalizes the feature matrix as the miRNA feature matrix $RM \in \mathbb{R}^{m \times m}$ And the characteristic matrix of the disease$RD \in \mathbb{R}^{n \times n}$. By restarting the random walk, the similarity between two points can be obtained, and the global structure of the graph can be better captured. According to $RM$and RD, the characteristic matrix of miRNA-disease $A_F \in \mathbb{R}^{(m+n) \times (m+n)}$ is obtained, as shown in formula (7):

$$A_F = \begin{bmatrix} 0 & RM \\ RD & 0 \end{bmatrix} \quad (7)$$

GCN-based node cutting

Graph convolution aggregates node information according to edge information and represents new node features. The two feature

extraction methods of graphs are spatial domain and Spectral domain. According to the explanation in Literature [41], the spatial method means that the neighbor nodes connected with the vertex are directly used to extract features. But the spectral method hopes to realize the convolution operation on the graph with the help of the graph theory, and studies the properties of the graph with the eigenvalues and eigenvectors of the Laplace matrix of the graph. Laplacian matrices are symmetric matrices, and GCN can perform feature decomposition. Common Laplacian matrix is symmetric normalized Laplacian, each node is the purpose of the normalized Laplacian matrix by foreign transfer the same amount of information, the more edge nodes exist, the less the amount of information transmitted each edge. The definition of the symmetric normalization Laplace matrix is shown in Eq. (8) :

$$\hat{L} = D^{-\frac{1}{2}} \cdot L \cdot D^{-\frac{1}{2}} \tag{8}$$

Here, D represents the degree matrix of the vertex, also called the diagonal matrix, and the definition of the elements of the L matrix is shown in formula (9):

$$L_{ij} = \begin{cases} 1 & i = j \ \ and \ \ diag(v_i) \neq 0 \\ -\frac{1}{\sqrt{diag(v_i)diag(v_j)}} & i \neq j \ \ and \ \ v_i \ \ is \ \ adjacent \ \ to \ \ v_j \\ 0 & otherwise \end{cases} \tag{9}$$

According to the heterogeneous map A_H of miRNA-disease, the normalized Laplacian matrix is constructed as shown in formula (10):

$$\hat{A}_H = D^{-\frac{1}{2}} \ A_H \ D^{-\frac{1}{2}} \tag{10}$$

Literature [42] indicates that Laplace matrix and Fourier transform are the two theoretical foundations of GCN. The Fourier transform of the graph expresses the arbitrary vector f defined on the graph as a linear combination of the eigenvectors of the Laplacian matrix, as shown in formula (11):

$$f = \hat{f}(1)u_1 + \ \hat{f}(2)u_2 + \cdots \hat{f}(n)u_n \tag{11}$$

$(u_1, u_2, \cdots u_n)$ is a set of orthogonal bases formed by n linearly independent vectors. The relationship between Fourier transform and Laplace matrix: The eigenvector of Laplace matrix is the base of Fourier transform, Get the graph convolution network, as shown in formula (12):

$$f(X, A) = ReLU(\hat{A}XW) \tag{12}$$

Here, X represents the feature matrix of the node, $\hat{A}$ represents the normalized adjacency matrix, and W is the weight matrix from the input layer to the hidden layer, which is equivalent to using a fully connected network to combine the feature connections.

According to the miRNA-disease heterogeneous map A_H and the miRNA-disease feature matrix A_F, the initial embedding of miRNA and disease nodes is formed through GCN. Make GCN directly connect and gather the information of neighbor nodes on each layer through the graph, as the input of the next layer, as shown in formula (13):

$$F = f(A_F, A_H) = ReLU(\ \hat{A}_H A_F W^{(0)}) \tag{13}$$

Here,$W^{(0)} \in \mathbb{R}^{(m+n)\times h}$, h embeds dimensions for layers.

Reinforcement layer based on attention mechanism

In order to make similarmiRNA (or disease) nodes similar in the feature space, this paper added an attentional strengthening layer H into the GCN hiding layer. The initial reinforcement layer H was defined as F, and an attention mechanism was introduced to consider all neighbor nodes. The attention mechanism is used to measure the influence of the feature vector H of adjacent nodes in the reinforcement layer on the feature vector H of nodes.a_{ij} represents the attention coefficient between

nodes, as shown in Formula (14), (15) and (16):

$$e_{ij} = ReLU\left(Wh_i, Wh_j\right) \quad (14)$$

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{j\in\mathcal{N}_i}\exp(e_{ix})} \quad (15)$$

$$H_i = \sum_{j\in\mathcal{N}_i} a_{ij}h_i \quad (16)$$

Where, $\mathcal{N}_i$ is the set of neighborhood nodes of node i. $ReLU$ is the activation function and $W \in \mathbb{R}^{(m+n)\times X}$ is a trainable matrix.

Next, define the $Loss$ function $\mathcal{L}_H$ of the reinforcement layer. In order to make the feature vector of node H_i on the reinforcement layer H focus on the feature vector H_j of important similar neighbor nodes, so that the feature vector of node i can be better iterated and updated, $Loss$ function is defined as follows, as shown in Eq. (17):

$$Loss(H_i) = \sum_{i=1}^{m+n}\sum_{j\in N_i} a_{ij}\,|H_i - H_j|^2 \quad (17)$$

Among them, m and n represent the number of miRNAs and diseases.

Abbreviations

HGCNELMDA: Heterogeneous Graph Convolutional Network model with Enhanced Layer to predict miRNA-Disease Associations;
GCN: Graph Convolutional Network;
RWR: Random Walk with Restart;
LOOCV: Global Leave-one-out Cross Validation;
miRNAs: Micro ribonucleic acids;
AUC: Area under the Curve

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The datasets that support the findings of this study are available in https://github.com/liubailong/HGCNELMDA.

Competing interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and publication of this article.

Funding

This work was supported by 'the Fundamental Research Funds for the Central Universities, No. 2019XKQYMS88." The funder had no role in study design, data collection and preparation of the manuscript.

Author Contributions

HD and AJY conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript; ZL and LBL designed, performed and analyzed experiments and wrote the manuscript; all authors read and approved the final manuscript.

Acknowledgements

We thank the editor and the anonymous reviewers for their comments and suggestions.

References

[1] Huang HY, Lin YCD, Li J, Huang KY, Shrestha S, Hong HC, et al. miRTarBase 2020: updates to the experimentally valid-ted microRNA-target interaction database. Nucleic Acids Research. 2020:145-148.

[2] Leland H. Genetics: From Gene to Genomes[M]. US: McGraw-Hill Higher Education, 2021.

[3] Cantile M, Di B M, Tracey D B M, et al. Functional Interaction among lncRNA HOTAIR and MicroRNAs in Cancer and Other Human Diseases[J]. Cancers, 2021, 13(3): 570.

[4] Shefa U, Jung JY. Comparative study of microarray and experimental data on Schwann cells in peripheral nerve degeneration and regeneration: big data analysis. Neural Regeneration Research. 2019;14(6): 1099.

[5] Chen X, Xie D, Zhao Q, You ZH. MicroRNAs and complex diseases: from experimental results to computational models. Briefings in Bioinformatics. 2019;20(2): 515-539.

[6] Zhang H, Liang Y, Han SY, Peng C, Li Y. Long Noncoding RNA and Protein Interactions: From Experimental Results to Computational Models Based on Network Methods. International Journal of Molecular Sciences. 2019; 20(6):1284.

[7] Blanca O Q. Extracellular MicroRNAs as Intercellular Mediators and Noninvasive Biomarkers of Cancer[J]. Cancers, 2020, 12(11): 3455.

[8] Wang X, Chai Y B, Li H, et al. Link prediction in heterogeneous information networks: An improved deep graph convolution approach[J]. Decision Support Systems, 2021, 141: 113448.

[9] Chen M, Liao B, Li ZJ. Global similarity method based on a two-tier random walk for the prediction of microRNA–disease association. Scientific Reports. 2018; 8(1): 1-16.
[10] Zhang W, Li Z S, Guo W Z, et al. A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations[C]. IEEE/ACM transactions on computational biology and bioinformatics, 2021: 1-1.
[11] Zhao HC, Kuang LN, WangL, et al. Prediction of MicroRNA-Disease Associations Based on Distance Correlation Set. BMC Bioinformatics. 19, 141, 2018. DOI：10.1186/s12859-018-2146-x
[12] Yue X, Wang Z, Huang J, et al. Graph embedding on biomedical networks: methods, applications and evaluations[J]. Bioinformatics, 2020, 36(4): 1241-1251.
[13] Chen X, Sun L G, Zhao Y. NCMCMDA:miRNA-disease association prediction through neighborhood constraint matrix completion[J]. Briefings in Bioinformatics, 2020, 22(1):485-496.
[14] Liang C, Yu SP, Luo JW. Adaptive multi-view multi-label learning for identifying disease-associated candidatemiRNAs. PLoS Computational Biology. 2019; 15(4): e1006931.
[15]Chen X, Sun LG, Zhao Y. NCMCMDA:miRNA–disease association prediction through neighborhood constraint matrix completion. Briefings in Bioinformatics. 2020.
[16]Zhao Y, Chen X, Yin J. Adaptive boosting-based computational model for predicting potentialmiRNA-disease associations. Bioinformatics. 2019; 35(22): 4730-4738.
[17]Chen X, Wang CC, Yin J, You ZH. Novel humanmiRNA-disease association inference based on random forest. Molecular Therapy-Nucleic Acids. 2018;13:568–79.
[18]Chen X, Wang L, Qu J, Guan NN, Li JQ. PredictingmiRNA–disease association based on inductive matrix completion. Bioinformatics. 2018; 34(24): 4256-4265.
[19]Jiang YT, Liu BT, Yu LH, Yan CG, Bian HJ. Predict miRNA-disease association with collaborative filtering. Neuroinformatics. 2018; 16(3-4): 363-372.
[20]Mao G, Wang S L, Zhang W. Prediction of Potential Associations Between MicroRNA and Disease Based on Bayesian Probabilistic Matrix Factorization Model. Journal of Computational Biology, 2019, 26(9): 1030-1039.
[21]Chen ZH, Wang XK, Gao P, Liu HJ, Song BS. Predicting Disease Related microRNA Based on Similarity and Topology. Cells. 2019; 8(11): 1405.
[22]Zeng XX, Wang W, Deng GS, Bing JX, Zou Q. Prediction of potential disease-associated MicroRNAs by using neural networks. Molecular Therapy-Nucleic Acids. 2019; 16: 566-575.
[23] Gong YC, Niu YQ, Zhang W, Li XH. A network embedding-based multiple information integration method for the miRNA-disease association prediction. BMC Bioinformatics. 2019; 20(1): 468.
[24] Zhang C, Chao Huang, Lu Yu, et al. Camel: Content-Aware and Meta-path Augmented Metric Learning for Author Identification. WWW. 2018
[25] Wang Y, Zheng FS, Wang ZB, Lu JB, Zhang HY. Circular RNA circ-SLC7A6 acts as a tumor suppressor in non-small cell lung cancer through abundantly sponging miR-21. Cell cycle. 2020. 19(17):2235-2246.

[26] Zhang XJ, Li YL, Qi PF, Ma ZL. Biology of MiR-17-92 Cluster and Its Progress in Lung Cancer. International journal of medical Sciences. 2018: 15(13):1443-1448.
[27] Fu X, Zhang J, Meng Z, et al. MAGNN: meta path aggregated graph neural network for heterogeneous graph embedding[C]. The Web Conference, 2020: 2331-2341.
[28] Song X Y, Liu T, Qiu Z Y, et al. Prediction of lncRNA-disease associations from heterogeneous information network based on deepwalk embedding model[C]. Intelligent Computing Methodologies, 2020: 291-300.
[29] Minh Nguyen Thi, Yi-Hung Wu. Integrating Meta-Path Similarity with User Preference for Top-N Recommendation. International Conference on Technologies and Applications of Artificial Intelligence (TAAI). 2019:1-6.
[30] Chen Z H, You Z H, Guo Z H, et al. Predicting drug-target interactions by node2vec node embedding in molecular associations network[C]. Intelligent Computing Theories and Application, 2020: 348-358.
[31] Li SR, Xie MZ, Liu XQ. A novel approach based on bipartite network recommendation and KATZ model to predict potential micro-disease associations. Frontiers in Genetics. 2019; 10:1147.
[32] Li Z W, Li J S, Nie R, et al. A graph auto-encoder model for miRNA-disease associations prediction[J]. Briefings in Bioinformatics, 2020, 1: 1-13.
[33] [81] Ding Y, Tian L P, Lei X, et al. Variational graph auto-encoders for miRNA-disease association prediction[J/OL]. Methods, 2020. https://doi.org/10.1016/j.ymeth.2020.08.004.
[34] [94] Li J S, Li Z W, Nie R, et al. FCGCNMDA: predictingmiRNA-disease associations by applying fully connected graph convolutional networks[J]. Molecular Genetics and Genomics, 2020, 295(5):1197-1209.
[35] Xiao WD, Zhong YC, Wu LL, Yang DX, Ye SQ, Zhang M. Prognostic value of microRNAs in lung cancer: A systematic review and meta-analysis. Molecular and clinical oncology. 2019; 10(1): 67-77.
[36] Li YX, Cui XM, Li YD, Zhang TT, Li SY. Upregulated expression of miR-421 is associated with poor prognosis in non-small-cell lung cancer. 2018: 10:2627-2633.
[37] Mansoori B, Mohammadi A, Ghasabi M, Shirjang S, Dehghan R, Montazeri V, et al. MiR-142-3p as tumor suppressormiRNA in the regulation of tumorigenicity, invasion and migration of human breast cancer by targeting Bach-1 expression. Journal of Cellular Physiology. 2019; 234(6): 9816-9825.
[38] He YJ, Deng F, Zhao SJ, Zhong SL, Zhao JH, Wang DD, et al. Analysis of miRNA–mRNA network reveals miR-140-5p as a suppressor of breast cancer glycolysis via targeting GLUT1. Epigenomics. 2019.11(9):1021-1036.
[39]Voss G, Haflidadóttir B S, Järemo H, Persson M, Ivkovic CT, Wikström P, Ceder Y. Regulation of cell–cell adhesion in prostate cancer cells by microRNA-96 through upregulation of E-Cadherin and EpCAM. Carcinogenesis. 2019. 41(7):865-874.
[40] Huang Z, Shi JC, Gao YX, Cui CM, Zhang S, Li JW, et al. HMDD v3.0: a database for

experimentally supported human microRNA-disease associations. Nucleic Acids Research.2018. 47(D1):D1013-D1017:47(D1).
[41] Zhang L, Liu B L, Li Z W, Zhu X Y, Liang Z Z, An J Y. PredictingmiRNA-disease associations by multiple meta-paths fusion graph embedding Model[J]. BMC Bioinformatics, 2020, 21: 470.