OXFORD

Briefings in Bioinformatics, 2022, **23**, 1–11
https://doi.org/10.1093/bib/bbab464
Problem Solving Protocol

SGNNMD: signed graph neural network for predicting deregulation types of miRNA-disease associations

Guangzhan Zhang[†], Menglu Li[†], Huan Deng, Xinran Xu, Xuan Liu and Wen Zhang

Corresponding author: Wen Zhang, College of Informatics, Huazhong Agricultural University,Wuhan 430070, China. E-mail: zhangwen@mail.hzau.edu.cn
[†]These authors contributed equally to this work.

Abstract

MiRNAs are a class of small non-coding RNA molecules that play an important role in many biological processes, and determining miRNA-disease associations can benefit drug development and clinical diagnosis. Although great efforts have been made to develop miRNA-disease association prediction methods, few attention has been paid to in-depth classification of miRNA-disease associations, e.g. up/down-regulation of miRNAs in diseases. In this paper, we regard known miRNA-disease associations as a signed bipartite network, which has miRNA nodes, disease nodes and two types of edges representing up/down-regulation of miRNAs in diseases, and propose a **s**igned **g**raph **n**eural **n**etwork method (SGNNMD) for predicting deregulation types of **m**iRNA-**d**isease associations. SGNNMD extracts subgraphs around miRNA-disease pairs from the signed bipartite network and learns structural features of subgraphs via a labeling algorithm and a neural network, and then combines them with biological features (i.e. miRNA–miRNA functional similarity and disease–disease semantic similarity) to build the prediction model. In the computational experiments, SGNNMD achieves highly competitive performance when compared with several baselines, including the signed graph link prediction methods, multi-relation prediction methods and one existing deregulation type prediction method. Moreover, SGNNMD has good inductive capability and can generalize to miRNAs/diseases unseen during the training.

Keywords: miRNA-disease associations, subgraph, graph convolutional network, signed network

Introduction

The miRNAs are one class of small non-coding RNA molecules containing about 22 nucleotides [1, 2]. MiRNA plays important roles in various critical biological processes, such as cell proliferation, development and metabolism. More importantly, miRNAs are involved in a variety of human diseases, such as cancers, cardiovascular diseases, etc. The study on miRNAs becomes increasingly important not only for understanding the molecular mechanisms of physiology and pathology but also for discovering novel clinical biomarkers and therapeutic targets of complex diseases. Exploring miRNA-disease associations is the premise of studying how miRNAs are involved in diseases. However, wet methods that identify miRNA-disease associations are time-consuming and costly, and it is important to develop the high-efficient and high-accuracy computational methods for the miRNA-disease association prediction.

Recent years have witnessed the advances in biomedical link prediction [3, 4], and a growing number of computational methods have been proposed for the miRNA-disease association prediction by leveraging data from publicly available miRNA-disease association databases. For example, Shen et al. [5] developed the computational method of collaborative matrix factorization for miRNA-disease association prediction. Chen et al. [6] proposed a bipartite network projection recommendation algorithm to predict potential disease-related miRNAs. Zhao et al. [7] utilized a random sampling based on k-means clustering to balance the positive and negative samples, and then trained a classifier integrated with multiple decision trees to predict miRNA-disease associations. Zeng et al. [8] developed a structural perturbation method

Guangzhan Zhang is a student in the College of Informatics, Huazhong Agricultural University, People's Republic of China. His research interests include machine learning and data mining.
Menglu Li is a PhD student in the College of Informatics, Huazhong Agricultural University, People's Republic of China. Her research interests include machine learning and bioinformatics.
Huan Deng is a student in the College of Informatics, Huazhong Agricultural University, People's Republic of China. Her research interests include machine learning and data mining.
Xinran Xu is a student in the College of Informatics, Huazhong Agricultural University, People's Republic of China. Her research interests include machine learning and data mining.
Xuan Liu is a PhD student in the College of Informatics, Huazhong Agricultural University, People's Republic of China. His research interests include machine learning and bioinformatics.
Wen Zhang obtained the bachelor degree and the master degree in computational mathematics from Wuhan University in 2003 and 2006, and got the doctoral degree in computer science from Wuhan University in 2009. He is now a professor in the College of Informatics, Huazhong Agricultural University, People's Republic of China. His research interests include machine learning and bioinformatics.

Received: June 5, 2021. **Revised:** October 8, 2021. **Accepted:** October 11, 2021

based on the miRNA-disease bilayer network. Peng et al. [9] proposed a deep learning framework for miRNA-disease association prediction. Zhang et al. [10] built a similarity-based framework based on known miRNA-disease associations. Chen et al. [11] exploited the matrix completion algorithm to predict the potential associations between miRNAs and diseases. Xiao et al. [12] developed a semi-supervised multi-label graph convolutional network. Li et al. [13] presented a novel method of neural inductive matrix completion with graph convolutional network. All the aforementioned methods predict whether a miRNA-disease association exists or not. The recent study [14] classified the miRNA-disease associations into multiple types according to their sources: genetics, epigenetics, circulation and miRNA-target interactions, and researchers also proposed the methods for the multi-type miRNA-disease association prediction. For example, Chen et al. [15] used a restricted Boltzmann machine-based method; Huang et al. [16] represented miRNA-disease-type triples as a tensor and introduced tensor decomposition methods; Wang et al. [17] proposed a neural multi-category miRNA-disease association prediction method.

Although the roles of miRNAs in diseases are prominently diverged, the study [18] observed up-regulation or down-regulation of miRNAs in various types of human diseases, and exploring how genetic variants in miRNA genes affect the expression level of miRNAs and lead to diseases is one important issue. For example, down-regulation of miR-137 has an impact on Lung Neoplasms [19], while the up-regulation of let-7a contributes to the attenuation of insulin signaling [20]. Thus, it is difficult to fully understand the diseases implicated with miRNAs only through exploring the existence of miRNA-disease associations but without knowing the deregulation of miRNAs in diseases. Fortunately, HMDD v3.0 database [14] contains information about how deregulation of a miRNA (up-regulation or down-regulation) is involved in a disease, and annotates miRNA-disease associations as up-regulation or down-regulation. Based on the data from HMDD v3.0, Li et al [18] proposed a server MISIM v2.0, which made the first attempt to predict deregulation types of miRNA-disease associations. More specifically, for a target miRNA, the diseases associated with other miRNAs but not associated with the target miRNA are considered to be novel related diseases. MISIM v2.0 takes the value of the improved miRNA functional similarities as weighting factors and the reciprocals of the improved disease semantic values to score novel miRNA-disease associations, and positive/negative scores indicate the up-regulation/down-regulation associations respectively.

In the area of graph theory, a signed graph is a graph in which each edge has a positive or negative sign, and the positive or negative signs in signed graphs usually have opposed meanings, such as friend/enemy. As illustrated in Figure 1, miRNA-disease associations and their deregulation types can be naturally modeled as a signed graph, which has miRNA nodes, disease nodes and two types of edges representing up/down-regulation of miRNAs in diseases. Thus, predicting miRNA-disease associations and their deregulation types is formulated as the signed graph link prediction problem. In recent years, a number of methods have been proposed for the signed graph link prediction, such as signed network embedding [21], signed graph convolutional network [22], signed graph attention network [23] and self-attention graph convolutional network [24]. However, most of these methods are designed for the social networks and rely on the balance theory, which fails to be established in biomedical signed graphs.

In this paper, we propose a **s**igned **g**raph **n**eural **n**etwork method (SGNNMD) for predicting deregulation types of **m**iRNA-**d**isease associations. SGNNMD extracts subgraphs around miRNA-disease pairs from the signed graph and learns structural features of subgraphs via a labeling algorithm and a graph neural network, and use biological features (miRNA-miRNA functional similarity and disease–disease semantic similarity) as auxiliary information to build the prediction model. In the computational experiments, SGNNMD achieves highly competitive performance when compared with several baselines, including the signed graph link prediction methods, multi-relation prediction methods and one existing deregulation type prediction method. Our contributions are summarized as follows:

- We study how to predict the deregulation types of miRNA-disease associations, on which little attention has been paid previously. It benefits exploring how genetic variants in miRNA genes affect the expression level of miRNAs and cause diseases.
- We formulate the original problem as a signed graph link prediction task, and propose a graph neural network-based method SGNNMD to resolve it. In SGNNMD, a novel node labeling algorithm is designed for subgraphs from the signed graph, and it can better describe the structural information. SGNNMD can generalize to miRNAs/diseases unseen in the training set.
- SGNNMD leverages the structural information learned from subgraphs around miRNA-disease pairs as well as the biological information of miRNAs and diseases, and trains the prediction model in an end-to-end manner. The structural information leads to the high-accuracy prediction model, and the biological information further enhances the performances.

Materials

Dataset

In this paper, we compile our dataset from the HMDD v3.0 database [14], which has in-depth curation of miRNA-disease associations categorized as up-regulation or down-regulation. First, we download 5438 miRNA-disease associations with annotated deregulation types

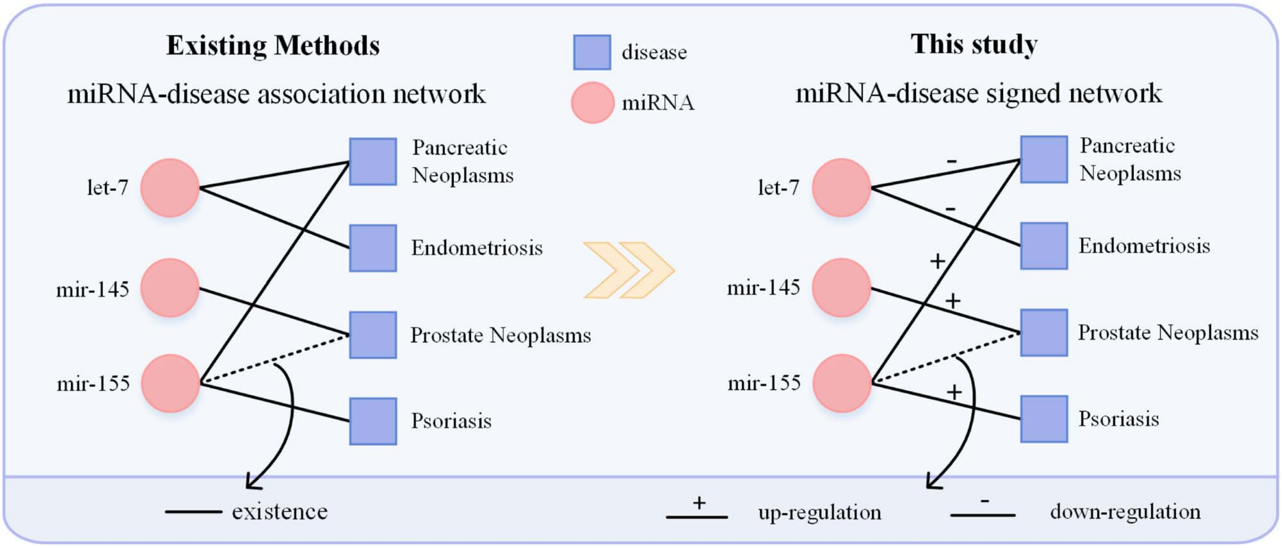

Figure 1. The link prediction in miRNA-disease binary network (left) and the link prediction in the miRNA-disease signed network (right). The solid lines represent known associations, the dashed lines represents unobserved associations that we want to predict.

from HMDD v3.0, and then filter them to ensure that each miRNA has at least one up-regulation and one down-regulation association with diseases. At last, we construct the benchmark dataset with 4264 miRNA-disease associations between 348 miRNAs and 210 diseases, including 2284 up-regulation and 1980 down-regulation associations.

Problem description

Given a set of miRNAs $U = \{u_1, u_2, \cdots, u_m\}$, a set of diseases $V = \{v_1, v_2, \cdots, v_n\}$ and a set of associations E belonging to two deregulation types, we can construct a signed bipartite network (signed graph) $G = (U, V, E)$, where $E = E^+ \cup E^-$, $E^+ = \{(v_i, v_j) \mid (v_i, v_j) = +1, v_i \in U, v_j \in V\}$ is the set of positive edges representing up-regulation associations, $E^- = \{(v_i, v_j) \mid (v_i, v_j) = -1, v_i \in U, v_j \in V\}$ is the set of negative edges representing down-regulation associations, and $E^+ \cap E^- = \emptyset$. In our study, an edge belong to one deregulation type, and it is different from the multi-relation miRNA-disease association prediction problem [16], in which one edge/association may have more than one relations/sources.

Although a number of miRNA-disease associations have been discovered, a large portion of associations still remains unobserved. The miRNA-disease associations provide partial information about how miRNAs causes diseases. For the comprehension of diseases implicated with miRNAs, we want to learn the patterns that determine the deregulation types of miRNA-disease associations from the signed graph G, and then predict deregulation types of novel miRNA-disease associations.

To enhance the performance of the prediction model, the biological features of miRNAs and diseases (i.e. disease semantic similarity and miRNA functional similarity) are used as auxiliary information, which are described as follows.

Disease semantic similarity

MeSH is a disease descriptor provided by the National Library of Medicine, and MeSH uses the directed a cyclic graphs (DAGs) to describe the relationship between diseases. As described in [25], the DAG structure of a disease d can be described as DAG(d) $= (N(d), E(d))$, where $N(d)$ is the set of node d and its ancestor nodes, and $E(d)$ is the set of links that connect parent nodes to their child nodes. Then, the contribution of a node n in DAG(d) to the semantic values of disease d is defined as:

$$C_d(n) = \begin{cases} 1 & if\ n = d \\ max\{C_d(n') * \Delta | n' \in children \quad of \quad n\} & if\ n \neq d \end{cases} \tag{1}$$

where Δ is the semantic contribution decay factor and is set to 0.5 in this study. It is assumed that diseases with more common ancestors in their DAGs have higher semantic similarities. Thus, the semantic similarity between diseases v_i and v_j is calculated by:

$$Sim^d(v_i, v_j) = \frac{\sum_{n \in N(v_i) \cap N(v_j)} (C_{v_i}(n) + C_{v_j}(n))}{\sum_{n \in N(v_i)} C_{v_i}(n) + \sum_{n \in N(v_j)} C_{v_j}(n)} \tag{2}$$

MiRNA functional similarity

We resort to the algorithm in [26] to calculate miRNA functional similarity. We first download the gene functional interaction network from HumanNet [27], which uses the associated log-likelihood scores (LLS) of each edge to measure the strength of interaction between any two genes. Then, the similarity between genes g_i and g_j is

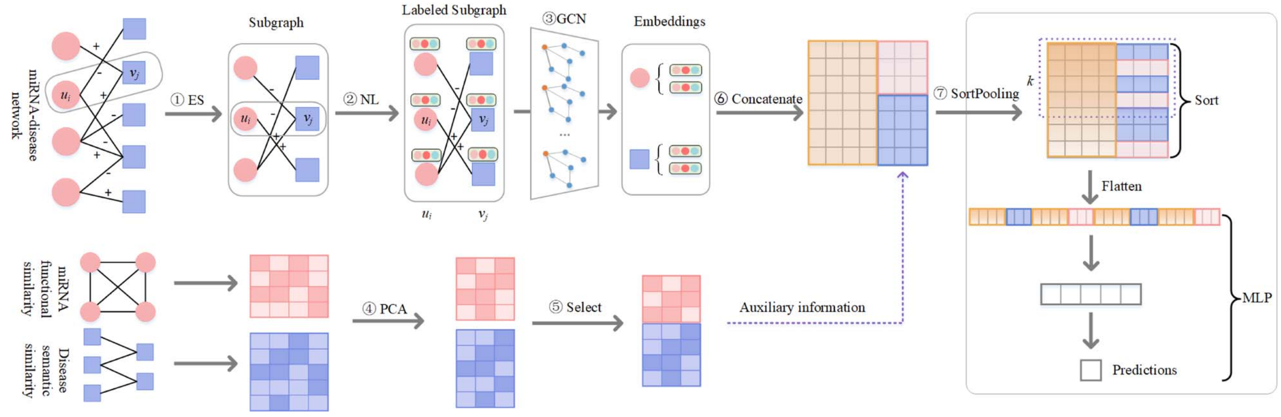

Figure 2. Overview of SGNNMD, ①**E**xtract **S**ubgraph for target nodes u_i, v_j, ② **N**ode **L**abeling, which assigns labels to nodes of the subgraph. ③ Learning topological features of the subgraph through **GCN**. ④ Perform **PCA** on miRNA and disease similarity to reduce dimensions. ⑤ **Select** miRNAs and diseases in the subgraph, and organize their biological features as a feature matrix. ⑥ The topological features and biological features are **concatenated**. ⑦ The concatenated feature matrix of subgraph nodes are passed into the **SortPooling** layer to meet the fixed size. Finally, embeddings of all nodes in the subgraph are flattened and fed into a MLP to make prediction.

defined as:

$$S(g_i, g_j) = \begin{cases} 0, & e(g_i, g_j) \notin \text{HumanNet} \\ \frac{LLS(g_i, g_j)}{LLSmax}, & e(g_i, g_j) \in \text{HumanNet} \\ 1, & g_i = g_j \end{cases} \tag{3}$$

where LLS_{max} represents the maximum LLS in HumanNet. Next, the similarity between gene g_i and gene set GS is defined as:

$$S(g_i, GS) = max_{g \in GS}(S(g_i, g)) \tag{4}$$

Finally, the functional similarity between miRNAs u_i and u_j is calculated by:

$$Sim^m(u_i, u_j) = \frac{\sum_{g \in GS_i} S(g, GS_j) + \sum_{g \in GS_j} S(g, GS_i)}{|GS_i| + |GS_j|} \tag{5}$$

where GS_i and GS_j are the gene sets related with the miRNAs u_i and u_j, and $|GS_i|$ and $|GS_j|$ are the total numbers of genes in GS_i and GS_j.

Methods

As shown in Figure 2, the signed graph neural network method for predicting deregulation types of miRNA-disease associations, SGNNMD, applies a labeling algorithm and a graph neural network to learn the topological features of subgraphs around miRNA-disease pairs, and uses miRNA functional similarity and disease semantic similarity as auxiliary information to build the prediction model.

Subgraph extraction

A miRNA-disease pair can be described by their local topological structure namely subgraph, which consists of the miRNA, the disease, and their neighbor nodes in the signed graph G. First, we define a subgraph around a miRNA-disease pair in an iterative manner.

Definition 1. (1-hop enclosing subgraph). In the miRNA-disease bipartite network $G = (U, V, E)$, given a miRNA node $u_i \in U$ and a disease node $v_j \in V$, the 1-hop subgraph of u_i and v_j is denoted as $G(1) = (U_1, V_1, E_1)$, where U_1 is the set including u_i and all miRNA nodes directly connected to v_j; V_1 is the set including v_j and all the miRNA nodes directly connected to u_i; $E_1 = E_1^+ \cup E_1^-$ is the set of edges formed by these nodes.

Definition 2. (h-hop enclosing subgraph, $h > 1$). $G(h) = (U_h, V_h, E_h)$ is the h-hop subgraph of the target nodes $u_i \in U$ and $v_j \in V$, where:

$$U_h = U_{h-1} \cup \{u \mid (u, v) \in E_{h-1}^+ \vee (u, v) \in E_{h-1}^-\} \tag{6}$$

$$V_h = V_{h-1} \cup \{v \mid (u, v) \in E_{h-1}^+ \vee (u, v) \in E_{h-1}^-\} \tag{7}$$

$E_h = E_h^+ \cup E_h^-$ are the edges formed by the related nodes.

With the above definition, we can obtain the h-hop subgraph $G(h)$ for the miRNA-disease pair (u_i, v_j). The subgraph implies beneficial information related to the target miRNA and disease. In the following content, we will describe how to learn topological features from subgraphs.

Node labeling for enclosing subgraph

Although the enclosing subgraph G_s can reflect the topological structure of the target miRNA-disease pair, we need the representation of the subgraph for the feature learning. Several issues should be taken into account for

Algorithm 1 Node Labeling for signed subgraphs

```
Input: Enclosing subgraph $G_s$ for miRNA $u_i$ and disease $v_j$, and hop number $H$, $l$=0;
Output: subgraph $G_s$ with labeled nodes.
assign integer 0, 1 to $u_i$ and $v_j$
for unlabeled miRNA nodes linked to $v_j$ do
    if edge is up-regulation then
        assign an integer 2 to it.
    else
        assign an integer 3 to it.
    end if
end for
for unlabeled disease nodes linked to $u_i$ do
    if edge is up-regulation then
        assign an integer 4 to it.
    else
        assign an integer 5 to it.
    end if
end for
all nodes in G(1) have been labeled.
for $h = 2, 3, \cdots, H$ do
    for miRNA nodes $\in G(h) - G(h-1)$ do
        let path denote the path linking it to target disease
        if all up-regulation edges in path then
            assign an integer $l + 0$ to it.
        else if all down-regulation edges in path then
            assign an integer $l + 1$ to it.
        else
            assign an integer $l + 2$ to it.
        end if
    end for
    for disease nodes $\in G(h) - G(h-1)$ do
        let path denote the path linking it to target miRNA
        if all up-regulation edges in path then
            assign a integer $l + 3$ to it.
        else if all down-regulation edges in path then
            assign a integer $l + 4$ to it.
        else
            assign a integer $l + 5$ to it.
        end if
    end for
    $l = l + 6$
end for
return subgraph $G_s$ with labeled nodes
```

the subgraph representation: (i) the relationship between neighbor nodes and target nodes (miRNA and disease) should be well reflected, (ii) the enclosing subgraph may contain noise that should be reduced and (iii) two types of edges should be differentiated. Inspired by previous studies [28, 29], we present a node labeling algorithm for the subgraph G_s, which assigns an integer to each subgraph node by considering the length of its path to the target nodes and edge types.

First, we respectively assign integers 0 and 1 to the target miRNA node u_i and the target disease node v_j in the 1-hop subgraph. We consider the miRNA nodes (except u_i) linked to v_j, and those nodes are classified into two classes according to the types of edges. Similarly, the disease nodes (except v_j) linking to the u_i are classified as two classes. Thus, we assign integers 2, 3, 4 and 5 to above nodes in two miRNA classes and two diseases classes. Then, given a (h −1)-hop subgraph whose nodes have been labeled from 0 to $l - 1$, we define the node labeling scheme for h-hop subgraph. Let U',V' denote the set of miRNA nodes and disease nodes that exist in the h-hop subgraph but do not exist in the (h-1)-hop subgraph. The nodes in U' have three types of paths linking to v_j: all positive edges, all negative edges, and hybrid edges. The nodes in V' also have three types of paths linking to u_i. Therefore, we need six integers for above six types of newly added nodes, and assign them the integers $l, l + 1, \ldots, l + 5$.

The proposed node labeling algorithm has several advantages: the nodes with different distances to target nodes are well discriminated; the nodes with the same distances are discriminated by the role of nodes (miRNA or disease) and types of paths to targets (all positive edges, all negative edges and hybrid edges). Therefore, the node labeling method can distinguish the target nodes from the neighbor nodes, miRNA nodes from disease nodes, and three path types, thus provides powerful information for the subgraph representation learning.

Graph neural network for link prediction

For the enclosing subgraph G_s around the miRNA-disease pair (u_i, v_j), let K denote the number of nodes and A denote the adjacency matrix of the subgraph, we use our proposed node labeling algorithm to label all nodes, and transform the labels of nodes into a label matrix X. Each row of the label matrix corresponds to a node, and a row vector is the one-hot vector for the label of the node. Then, we use X as the attributes of nodes, and employ the following two-layer graph convolutional network on G_s to learn its topological features:

$$Z_t = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}\sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}XW^{(0)})W^{(1)}\right) \qquad (8)$$

where $\tilde{A} = A + I$, I is identity matrix, and $\tilde{D} = \sum_j \tilde{A}_{ij}$ is the degree matrix of $\tilde{A}$. $W^{(0)}$ and $W^{(1)}$ are the weight matrices of two layers. $\sigma(\cdot)$ is the activation function, and is set as *tanh*. The GCN updates the embeddings of nodes from the initial node states X by aggregating information from neighbors. To extract multi-scale features, we stack node embeddings of two graph convolution layers to obtain the embedding matrix denoted as Z_t, which implicates the topological features of subgraph G_s, and we adopt Z_t for the downstream task.

Moreover, we pre-calculate miRNA functional similarity and disease semantic similarity for all miRNAs and diseases in our dataset. In the miRNA functional matrix, each row is the representative vector of a miRNA. Similarly, each row of the disease semantic similarity matrix corresponds to the representation of a disease. We perform principal component analysis (PCA) on these features, reduce the dimensions of functional similarity vectors and semantic similarity vectors to 128, and use these biological features as auxiliary information. For miRNA and disease nodes in the subgraph G_s, their biological features are organized as a feature matrix Z_b with K rows (number of nodes) and 128 columns.

For the subgraph G_s, we concatenate the topological feature matrix Z_t and biological feature matrix Z_b, $Z = concat(Z_t, Z_b)$. Since subgraphs usually have varied numbers of nodes, Z has different shapes, and thus we apply the SortPooling layer to Z, which sorts the final node states of subgraph according to their labels to achieve an isomorphism invariant node ordering [30]. In the SortPooling layer, a max-k pooling is used to unify the size of the sorted representations of different subgraphs. Finally, embeddings of all nodes in the subgraph are flattened into one-dimensional tensor of the fixed length to represent the miRNA-disease pair (u_i, v_j), and then are fed into a multi-layer perception (MLP) to classify the miRNA-disease pair as up-regulation or down-regulation.

Experiments

Experimental settings

We use the miRNA-disease association dataset described in Method section for computational experiments, and adopt 5-fold cross-validation (5-CV) to evaluate the performances of prediction models. To comprehensively investigate the performances of models, we consider two ways of implementing 5-CV, which have been adopted by related works [16, 17].

- CV_{type}: we randomly divide all miRNA-disease associations (up-regulation and down-regulation) into five equal-sized subsets. In each fold, one subset is used for testing, and the remaining four subsets are served as the training set. For each miRNA-disease association in the test set, we try to classify it as up-regulation or down-regulation. This setting tests how accurately the models can classify two deregulation types.
- $CV_{triplet}$: we take all known miRNA-disease-deregulation type triples as positives, and randomly select a fraction of unknown triples as negatives, the size of which is the same as the positive set. Then, we randomly divide all triples into five equal-sized subsets. In each fold, one subset is served as the test set and the rest triples are used for model training. This setting tests how accurately the models can predict novel associations.

In this study, we are more interested in the deregulation type prediction, and treat CV_{type} as the primary experimental setting. For CV_{type} and $CV_{triplet}$, we calculate several evaluation metrics: the area under the precision-recall curve (AUPR), the area under the ROC curve (AUC), F1, Precision (PR) and Recall (RE).

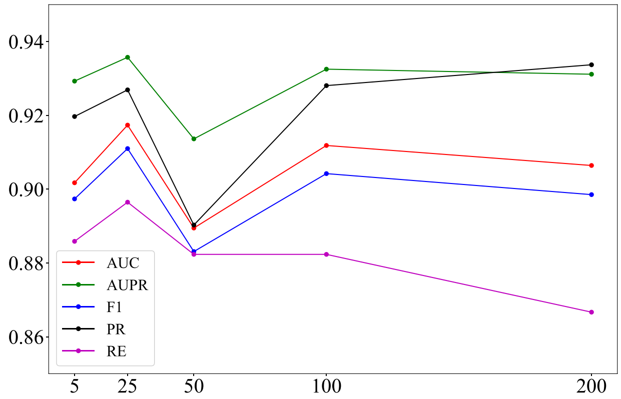

Figure 3. The performance of SGNNMD with different values for the truncation dimension k.

Parameter setting

In SGNNMD, we empirically set two embedding layers for the GCN, and use a MLP with two output nodes for deregulation type prediction. The cross-entropy loss function with L2 regulation is used to train SGNNMD in an end-to-end manner. For the Adam optimizer, the learning rate is set to 1×10^{-4}, and the weight decay is set to 5×10^{-4}. SGNNMD has two hyper-parameters: h the hop number of subgraphs around miRNA-disease pairs, k the truncation dimension for the shape of the output in the SortPooling layer. The hop number h determines the number of nodes that can be contained when extracting subgraphs. It should be noted that our miRNA-disease signed graph only has hundreds of nodes. As h increases, subgraphs will contain more nodes and lead to more similar topological structures, which makes it difficult to obtain the useful information for classifying two types of associations. To verify the viewpoint, we use a quarter of the experimental dataset and test the model performance under different hop numbers. The experimental results are shown in Supplementary Table S1. With the increase of the hop number, the performance of SGNNMD will decrease, which is consistent with the conclusion of Zhang's works [28–30]. Therefore, we set h to 1 as described in [29]. As shown in Figure 3, SGNNMD produces the robust performances to the parameter k, and we set k to 25 in the following computational experiments.

Comparison with other state-of-the-art methods

In this section, we compare SGNNMD with several baselines. Only one method named MISIM [18] has been proposed to predict deregulation types of miRNA-disease associations, and is thus adopted for comparison. Since we formulate the original problem as a signed graph link prediction task, we take two representative models SGCN [22] and SNEA [31] for comparison. Taking deregulation types (up-regulation and down-regulation) as two relations, we consider multi-relation prediction methods:

Table 1. The performances of different methods evaluated by CV_{type}

	AUPR	AUC	F1	PR	RE
SGCN	0.642	0.617	0.664	0.584	0.774
SNEA	0.675	0.731	0.694	0.683	0.706
GCMC	0.851	0.852	0.837	0.824	0.852
TDRC	0.889	0.920	0.807	0.954	0.875
NMCMDA	0.878	0.902	0.733	0.776	0.695
MISIM	0.875	0.850	0.791	0.774	0.809
SGNNMD	0.936	0.917	0.927	0.896	0.911

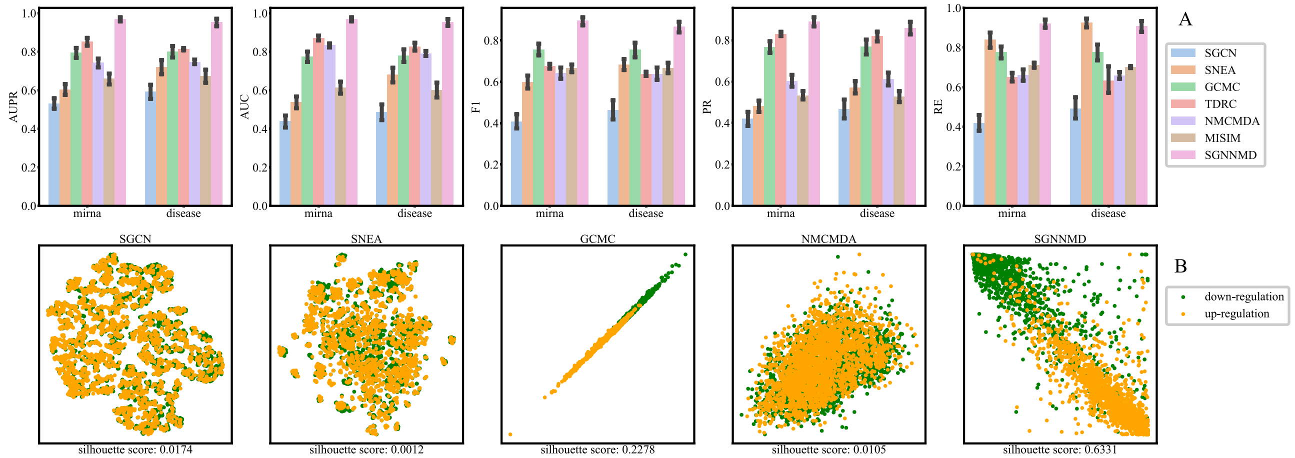

Figure 4. **A** The average performances of different models for all miRNAs/diseases in terms of AUPR, AUC, F1, PR and RE, where error bar represents the 0.95 confidence interval of different metrics. **B** The visualization of the representations of up-regulation and down-regulation associations learned by different GNN-based methods.

GCMC [32], TDRC [33] and NMCMDA [17]. It is worth mentioning that TDRC and NMCMDA have been developed for the similar task that predicts multiple sources of miRNA-disease associations. Therefore, six methods: MISIM, SGCN, SNEA, GCMC, TDRC and NMCMDA are used as baselines, and we replicate them according to their publications or use publicly available programs with the parameter settings recommended in the original papers.

First of all, we compare the performances of SGNNMD and six baselines in classifying up-regulation and down-regulation associations, which are evaluated by CV_{type}. The results in Table 1 demonstrate that SGNNMD outperforms baselines in terms of all metrics, achieving AUPR of 0.936 and AUC of 0.917, and two multi-relation prediction methods TDRC and NMCMDA are the second and third best, followed by GCMC. Moreover, we pay attention to the performances of prediction models on specific miRNAs and diseases. We calculate the evaluation metric scores based on the prediction for every miRNA/disease, and average scores of all miRNAs/diseases. As demonstrated in Figure 4A, SGNNMD produces better results than baselines for miRNAs/diseases, especially in terms of AUC and AUPR.

Further, we explore the effectiveness of the learned representation of miRNA-disease associations. We consider all up-regulation and down-regulation associations, and then extract the representation of these associations through trained GNN-based models: SGCN, SNEA, GCMC, NMCMDA and SGNNMD, and project them into 2D space using t-SNE. In addition, we calculate the silhouette score to further explore the inter-cluster and intra-cluster of the embedding results. As illustrated in Figure 4B, the proposed method can well distinguish the up-regulation associations (yellow) and down-regulation associations (green), and the silhouette scores also show that SGNNMD separates samples more clearly than other methods. In contrast, SGCN, SNEA and NMCMDA fail to clearly discriminate up-regulation associations from down-regulation associations, because the balance theory used by SGCN and SNEA are not established in biological networks and two types of opposed edges are not taken into account by NMCMDA.

Further, we compare the performances of SGNNMD and baselines in predicting miRNA-disease-deregulation type triples, evaluated by CV_{triple}. Here, we adopt the methods MISIM, TDRC and NMCMDA for the comparison, because they are suitable for the task and also produce superior performances in the CV_{type} evaluation. As shown in Table 2, SGNNMD produces better performances than TDRC and NMCMDA, because SGNNMD considers the difference between up-regulation and down-regulation and thus capture deeper information for prediction. Although MISIM produces better results than SGNNMD in terms of several evaluation metrics, we must point out that MISIM is different from TDRC, NMCMDA and SGNNMD, and its results could be exaggerated. Given a

Table 2. The performances of different methods evaluated by CV$_{triplet}$

	AUPR	AUC	F1	PR	RE
TDRC	0.732	0.673	0.669	0.673	0.665
NMCMDA	0.709	0.503	0.644	0.532	0.732
MISIM	0.781	0.705	0.718	0.688	0.750
SGNNMD	0.776	0.701	0.701	0.701	0.702

miRNA-disease pair, TDRC, NMCMDA and SGNNMD yield the scores about existence of up-regulation and down-regulation association simultaneously, while MISIM produces the score about existence of the association and then assigns a deregulation type to it. Therefore, we pay attention to those triples about miRNA-disease associations, and figure out how many deregulation types are correctly predicted. For fair comparison, TDRC, NMCMDA and SGNNMD assign the deregulation type with a greater score to each association. In general, MISIM correctly predicts deregulation types of 2547 associations out of 4264, and the numbers of deregulation types correctly predicted by TDRC, NMCMDA and SGNNMD are 4158, 3081 and 3918. The results demonstrate that SGNNMD is superior to MISIM and could better predict the deregulation types.

Discussion on SGNNMD

To investigate the importance of components in SGNNMD, we design the following variants of SGNNMD:

- **SGNNMD-Bio** only uses the biological features of miRNAs and diseases to build the prediction model.
- **SGNNMD-Topo** only uses the topological features of subgraphs learned by node labeling and GCN to build the prediction model.
- **SGNNMD-RGCN** uses the R-GCN encoder to learn the topological features from the signed network, and then builds prediction model.
- **SGNNMD-SL** applies a node labeling algorithm in IGMC [29] instead of the node labeling proposed in this paper, and then builds the prediction model.

We conduct the 5-CV experiments (CV$_{type}$) to evaluate the performances of four variants of SGNNMD, and results are shown in Figure 5A. The comparison of SGNNMD-Bio, SGNNMD-Topo and SGNNMD show that the topological features learned from the signed network can achieve high-accuracy performances. As auxiliary information, biological features are trivial, but it still improves the performances. R-GCN encoder can be applied to multi-relational graphs, but the comparison between SGNNMD-RGCN and SGNNMD reveals that R-GCN is incapable of dealing with the signed graph which has links with opposed meanings. Compared with SGNNMD, SGNNMD-SL also produces lower performances, indicating the proposed node labeling algorithm can perform better than the classic node labeling algorithm [29].

To further test the robustness of SGNNMD, we randomly remove/mask a percentage of links in the signed bipartite network, and then implement 5-CV to evaluate the performances of SGNNMD on the masked datasets. As shown in Figure 5B, SGNNMD will have decreased performance as masking more links, but can still produce AUPR/AUC scores greater than 0.8 when nearly half of the links are masked. As removing/masking a percentage of links, some subgraphs will only contain target nodes, and it becomes difficult to extract the topological information from these subgraphs. In the experiments, we have 3838, 3412, 2985 miRNA-disease associations and 0,1,0 subgraph only with target nodes when randomly removing 10, 20 and 30% links. More results are provided in Supplementary Table S2. After removing a large percentage of links, we still have plenty of associations for training, and those subgraphs only with target nodes take up a certain small proportion, and it ensures the robustness of SGNNMD.

The study demonstrates that our method SGNNMD can achieves good performances merely using structural information, which are learned from subgraphs around the miRNA-disease pairs. Moreover, SGNNMD relies on the subgraphs around the miRNA-disease pairs for model training and prediction, and thus miRNAs/diseases unseen during the training can be predicted.

Case studies

To further explore the ability of SGNNMD in practical applications, we conduct case studies on two diseases of the wide concerns: Schizophrenia and Lung Neoplasms. Schizophrenia is a psychiatric diagnosis characterized by continuous or relapsing episodes of psychosis. Lung cancer is a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung, which makes it the most common cause of cancer-related death in men and the second most common in women after breast cancer.

First, we select all miRNA-disease associations whose 'description' information is related to 'up-regulation' or 'down-regulation' from HMDD v3.0 database, then construct a signed bipartite network after filtering out miRNA and disease nodes without links, and then train the SGNNMD model. For a target disease, we take into account all miRNAs related to it, and predict associations and deregulation types between miRNAs and the target disease. For Schizophrenia, top 20 related miRNAs and deregulation types of their associations are listed in

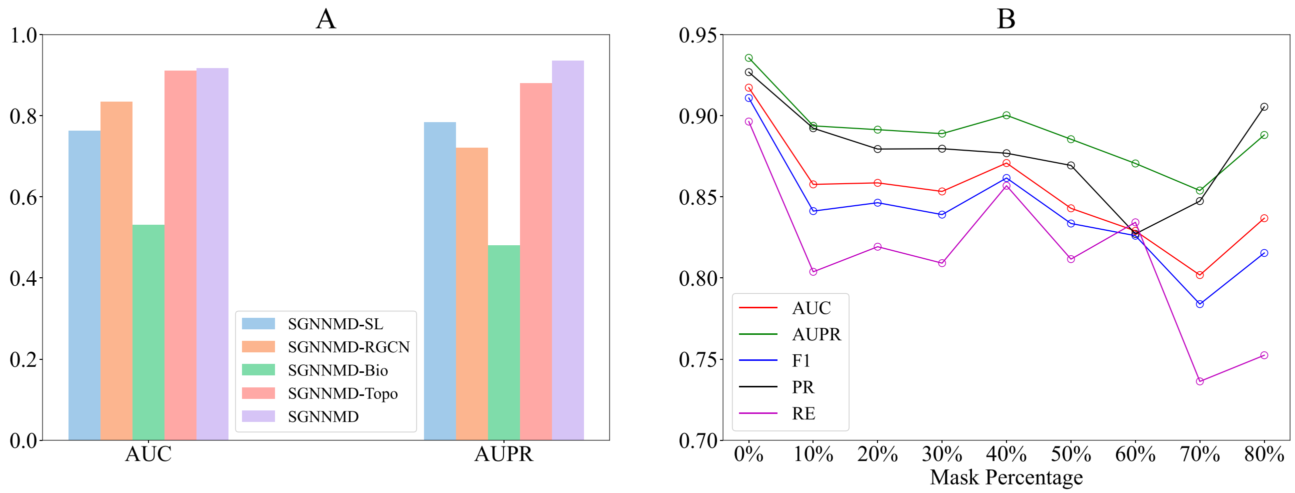

Figure 5. **A** The performances of SGNNMD variants in ablation analysis. **B** The performances of SGNNMD when masking links.

Table 3. Top 20 miRNAs associated with Schizophrenia and Lung Neoplasms and their types predicted by SGNNMD

Schizophrenia			**Lung Neoplasms**		
miRNA	**deregulation**	**Evidence (PMID)**	**miRNA**	**deregulation**	**Evidence (PMID)**
hsa-mir-29b-1	Down	17326821	hsa-mir-32	Down	22349819
hsa-mir-7-2	Down	17326821	hsa-mir-126	Down	22862169
hsa-mir-23b	Down		hsa-mir-29b-1	Down	17890317
hsa-let-7b	Down		hsa-mir-214	Down	16530703
hsa-mir-29b-2	Down		hsa-let-7a-1	Down	16712479
hsa-let-7g	Up	22094284	hsa-mir-30a	Down	26156018
hsa-mir-155	Down		hsa-mir-198	Down	23354517
hsa-mir-125a	Down		hsa-let-7c	Down	16712479
hsa-mir-195	Down	22094284	hsa-mir-17	Up	16266980
hsa-mir-92a-2	Down	17326821	hsa-let-7a-2	Down	16712479
hsa-mir-191	Down		hsa-let-7b	Down	16712479
hsa-mir-132	Down	25487174	hsa-mir-22	Down	22484852
hsa-mir-146b	Up	26173148	hsa-mir-33a	Down	25544258
hsa-mir-193a	Up	26183697	hsa-mir-130a	Up	20508945
hsa-mir-24-2	Down	17326821	hsa-mir-125b-1	Down	14973191
hsa-mir-212	Down	25487174	hsa-mir-222	Down	23974492
hsa-mir-30e	Down	25487174	hsa-mir-429	Up	
hsa-mir-26b	Down	26450699	hsa-mir-140	Down	26722475
hsa-mir-34	Up	26173148	hsa-mir-505	Down	
hsa-mir-92b	Down	17326821	hsa-mir-518e	Down	

Table 3. We find evidence from literature in PubMed to support our findings, and 14 out of 20 top predictions have been confirmed. For example, Perkins et al. [34] found that hsa-mir-29b-1 and hsa-mir-7-2 are expressed at lower levels in the individuals with Schizophrenia. For Lung Neoplasms, the top 20 predictions are shown in Table 3, among which hsa-mir-32 and hsa-let-7c have been confirmed by [35, 36]. Therefore, the case studies demonstrate the usefulness of SGNNMD in discovering novel miRNA-disease associations and the deregulation types of miRNAs in diseases.

Conclusion

In recent years, there have been numerous methods about the miRNA-disease association prediction, but few models have been developed to predict how the deregulation of miRNAs affects disease. In this paper, we propose a signed graph neural network-based method SGNNMD for predicting the deregulation types (up-regulation or down-regulation) of miRNA-disease associations. To classify miRNA-disease associations, SGNNMD presents a labeling algorithm for the signed subgraphs around miRNA-disease pairs to characterize them, and then learn the topological features of subgraphs via a graph neural network. Moreover, the biological features of miRNAs and diseases are incorporated into SGNNMD as auxiliary information to enhance performances. Owing to the subgraph learning, SGNNMD can generalize to miRNAs/diseases unseen in the training set. Extensive experiments show that SGNNMD has good performances under different experimental settings, and outperforms baselines. The case studies also demonstrate that SGNNMD can find the novel miRNA-disease associations and their deregulation types. The proposed method SGNNMD can be applied to various signed graph link prediction problems. However, SGNNMD extracts subgraphs around node pairs from the signed graph to train the prediction model, and it takes lots of training time if we have a number of subgraphs or a large signed graph. In this case, we have to restrict the number of subgraphs and sizes of subgraphs (number of nodes) to reduce the computational complexity.

Key Points

- We study how to predict the deregulation types of miRNA-disease associations, on which little attention has been paid previously. It benefits exploring how genetic variants in miRNA genes affect the expression level of miRNAs and lead to diseases.
- We formulate the original problem as a signed graph link prediction task, and propose a graph neural network-based method SGNNMD to resolve it. In SGNNMD, a novel node labeling algorithm is designed for subgraphs from the signed graph, and it can better describe the structural information. SGNNMD can generalize to miRNAs/diseases unseen in the training set.
- SGNNMD leverages the structural information learned from subgraphs around miRNA-disease pairs as well as the biological information of miRNAs and diseases, and trains the prediction model in an end-to-end manner. The structural information leads to the high-accuracy prediction model, and the biological information further enhances the performance.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Data availability

The datasets were derived from the following sources in the public domain, the miRNA-disease association from https://www.cuilab.cn/hmdd, the disease MeSH descriptors from https://meshb.nlm.nih.gov/ and the gene functional interaction network is downloaded from https://www.inetbio.org/humannet/download.php. The implementation of SGNNMD and the preprocessed data is available at https://github.com/bubblecode/SGNNMD and https://github.com/BioMedicalBigDataMiningLab/SGNNMD.

Author contributions statement

W.Z. conceived the project, G.Z. and H.D. conducted the experiment(s), W.Z., G.Z., X.X., M.L. and X.L. analyzed the results and wrote the manuscript.

Funding

The National Natural Science Foundation of China (62072206, 61772381); Huazhong Agricultural University Scientific & Technological Self-innovation Foundation; Fundamental Research Funds for the Central Universities (2662021JC008).

Acknowledgments

We thank anonymous reviewers for their valuable suggestions.

References

1. Llave C, Xie Z, Kasschau KD, et al. Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science* 2002; **297**(5589): 2053–6.
2. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004; **116**(2): 281–97.
3. Yue X, Wang Z, Huang J, et al. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics* 2020; **36**(4): 1241–51.
4. Zhang ZC, Zhang XF, Wu M, et al. A graph regularized generalized matrix factorization model for predicting links in biomedical bipartite networks. *Bioinformatics* 2020; **36**(11): 3474–81.
5. Shen Z, Zhang YH, Han K, et al. miRNA-Disease Association Prediction with Collaborative Matrix Factorization. *Complexity* 2017; **2017**:1–9.
6. Chen X, Xie D, Wang L, et al. BNPMDA: bipartite network projection for MiRNA–disease association prediction. *Bioinformatics* 2018; **34**(18): 3178–86.
7. Zhao Y, Chen X, Yin J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations. *Bioinformatics* 2019; **35**(22): 4730–8.
8. Zeng X, Liu L, Lü L, et al. Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* 2018; **34**(14): 2425–32.
9. Peng J, Hui W, Li Q, et al. A learning-based framework for miRNA-disease association identification using neural networks. *Bioinformatics* 2019; **35**(21): 4364–71.
10. Zhang W, Li Z, Guo W, et al. A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. *IEEE/ACM Trans Comput Biol Bioinform* 2021; **18**(2): 405–15.
11. Chen X, Wang L, Qu J, et al. Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* 2018; **34**(24): 4256–65.
12. Pan X, Shen HB. Inferring disease-associated microRNAs using semi-supervised multi-label graph convolutional networks. *Iscience* 2019; **20**:265–77.
13. Li J, Zhang S, Liu T, et al. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. *Bioinformatics* 2020; **36**(8): 2538–46.
14. Huang Z, Shi J, Gao Y, et al. HMDD v3. 0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res* 2019; **47**(D1): D1013–7.
15. Chen X, Clarence Yan C, Zhang X, et al. RBMMMDA: predicting multiple types of disease-microRNA associations. *Sci Rep* 2015; **5**(1): 13877.
16. Huang F, Yue X, Xiong Z, et al. Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations. *Brief Bioinform* 2020;Bbaa140.

17. Wang J, Li J, Yue K, et al. NMCMDA: neural multicategory MiRNA-disease association prediction. *Brief Bioinform* 2021; Bbab074.
18. Li J, Zhang S, Wan Y, et al. MISIM v2. 0: a web server for inferring microRNA functional similarity based on microRNA-disease associations. *Nucleic Acids Res* 2019; **47**(W1): W536–41.
19. Zhu X, Li Y, Shen H, et al. miR-137 inhibits the proliferation of lung cancer cells by targeting Cdc42 and Cdk6. *FEBS Lett* 2013; **587**(1): 73–81.
20. O'Toole TE, Abplanalp W, Li X, et al. Acrolein decreases endothelial cell migration and insulin sensitivity through induction of let-7a. *Toxicol Sci* 2014; **140**(2): 271–82.
21. Yuan S, Wu X, Xiang Y. SNE: signed network embedding. *Pacific-Asia conference on knowledge discovery and data mining Springer* 2017;183–95.
22. Derr T, Ma Y, Tang J. Signed graph convolutional networks. In: *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, 929–34.
23. Huang J, Shen H, Hou L, et al. Signed graph attention networks. In: *International Conference on Artificial Neural Networks*. Springer, 2019, 566–77.
24. Wang S, Aggarwal C, Tang J, et al. Attributed signed network embedding. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, 137–46.
25. Wang D, Wang J, Lu M, et al. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 2010; **26**(13): 1644–50.
26. Xiao Q, Luo J, Liang C, et al. A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* 2018; **34**(2): 239–48.
27. Lee I, Blom UM, Wang PI, et al. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res* 2011; **21**(7): 1109–21.
28. Zhang M, Chen Y. Link Prediction Based on Graph Neural Networks. In: Bengio S, Wallach H, Larochelle Het al. (eds). *Advances in Neural Information Processing Systems*, Vol. **31**. Curran Associates, Inc, 2018.
29. Zhang M, Chen Y. Inductive Matrix Completion Based on Graph Neural Networks. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. OpenReview.net, 2020.
30. Zhang M, Cui Z, Neumann M, et al. An End-to-End Deep Learning Architecture for Graph Classification. In: SA MI, Weinberger KQ (eds). *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018*. AAAI Press, 2018, 4438–45.
31. Heider F. Attitudes and cognitive organization. *J Psychol* 1946; **21**(1): 107–12.
32. van denBerg R, Kipf TN, Welling M. Graph Convolutional Matrix Completion. *CoRR* 2017; abs/1706.02263.
33. Huang F, Yue X, Xiong Z, et al. Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations. *Brief Bioinform* 2020;Bbaa140.
34. Perkins DO, Jeffries CD, Jarskog LF, et al. microRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder. *Genome Biol* 2007; **8**(2): 1–11.
35. Zhang S, Chen H, Zhao X, et al. REV3L 3'UTR 460 T>C polymorphism in microRNA target sites contributes to lung cancer susceptibility. *Oncogene* 2013; **32**(2): 242–50.
36. Tong AW. Small RNAs and non-small cell lung cancer. *Curr Mol Med* 2006; **6**(3): 339–49.