Bioinformatics, 38(8), 2022, 2226–2234
https://doi.org/10.1093/bioinformatics/btac077
Advance Access Publication Date: 12 February 2022
Original Paper

Systems biology

PDMDA: predicting deep-level miRNA–disease associations with graph neural networks and sequence features

Cheng Yan[1,2,†], Guihua Duan[2,†], Na Li[2], Lishen Zhang[2], Fang-Xiang Wu [3] and Jianxin Wang [2,*]

[1]School of Information Science and Engineering, Hunan University of Chinese Medicine, Changsha 410208, China, [2]School of Computer Science and Engineering, Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha 410083, China and [3]Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, Saskatoon SK S7N5A9, Canada

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Teresa Przytycka

Received on July 24, 2021; revised on January 18, 2022; editorial decision on January 30, 2022; accepted on February 5, 2022

Abstract

Motivation: Many studies have shown that microRNAs (miRNAs) play a key role in human diseases. Meanwhile, traditional experimental methods for miRNA–disease association identification are extremely costly, time-consuming and challenging. Therefore, many computational methods have been developed to predict potential associations between miRNAs and diseases. However, those methods mainly predict the existence of miRNA–disease associations, and they cannot predict the deep-level miRNA–disease association types.

Results: In this study, we propose a new end-to-end deep learning method (called PDMDA) to predict deep-level miRNA–disease associations with graph neural networks (GNNs) and miRNA sequence features. Based on the sequence and structural features of miRNAs, PDMDA extracts the miRNA feature representations by a fully connected network (FCN). The disease feature representations are extracted from the disease–gene network and gene–gene interaction network by GNN model. Finally, a multilayer with three fully connected layers and a softmax layer is designed to predict the final miRNA–disease association scores based on the concatenated feature representations of miRNAs and diseases. Note that PDMDA does not take the miRNA–disease association matrix as input to compute the Gaussian interaction profile similarity. We conduct three experiments based on six association type samples (including circulations, epigenetics, target, genetics, known association of which their types are unknown and unknown association samples). We conduct fivefold cross-validation validation to assess the prediction performance of PDMDA. The area under the receiver operating characteristic curve scores is used as metric. The experiment results show that PDMDA can accurately predict the deep-level miRNA–disease associations.

Availability and implementation: Data and source codes are available at https://github.com/27167199/PDMDA.

Contact: jxwang@mail.csu.edu.cn

1 Introduction

MicroRNAs (miRNAs) are single-stranded small non-coding RNAs that are typically 22 nucleotides long, which can regulate genes at the post-transcription level to affect the translation of mRNAs to proteins (Kim, 2005). Therefore, many studies have revealed that miRNAs play important roles in a wide range of biological processes, such as cell proliferation, cell death, metabolism, apoptosis, developmental timing, neuronal gene expression and so on (Calin and Croce, 2006). miRNA let-7 is the first known miRNA that regulates the heterochronic genes in the developmental timing of the nematode *Caenorhabditis elegans* (Friedman *et al.*, 2009). Period protein homolog lin-42 has a regulating microRNA (miRNA) biogenesis function at the transcriptional level, which is a complex gene with four isoforms and multiple functions including the regulation of molting, developmental timing and entry into dauer (Van Wynsberghe and Pasquinelli, 2014).

Meanwhile, many associations between miRNAs and human diseases have also been revealed. Normal human articular cartilage expressed miR-140 is significantly reduced in osteoarthritic cartilage (Miyaki *et al.*, 2009). The expression levels of miRNAs (miRs)-143 and -145 are reduced in colon cancers and various kinds of

established cancer cell lines (Akao *et al.*, 2007). MiR-346 and GRID1 are associated with schizophrenia, the expression is lower in schizophrenia patients than that in controls (Zhu *et al.*, 2009). Therefore, based on the validated experiments from existing literature, miRNA–disease association databases that have been constructed (Li *et al.*, 2014; Wang *et al.*, 2014; Xie *et al.*, 2013; Yang *et al.*, 2010). The miRCancer curates 196 human cancer diseases and 9080 associations between miRNAs and diseases from more than 7288 existing literature (Xie *et al.*, 2013). The dbDEMC is a miRNA–disease association database for human cancer diseases, and the newest version (dbDEMC 2.0) contains 49 202 miRNA-cancer associations across 36 cancer types and 73 subtypes (Yang *et al.*, 2010). The miR2disease is also a manually curated database containing relationships between human diseases and miRNAs, which also provides detailed information on miRNA–disease relationships, such as microRNA ID, disease name and a brief description of the microRNA–disease relationship (Jiang *et al.*, 2009). The current version of HMDD (HMDD 3.2) includes 1206 miRNA genes, 893 diseases and 35 547 miRNA–disease associations (Huang *et al.*, 2019; Li *et al.*, 2014), which also provides the types of miRNA–disease associations, such as genetics, epigenetics, circulating miRNAs and miRNA–target interactions.

With the development of miRNA–disease association benchmark datasets, many computational methods have been developed to predict potential miRNA–disease associations. Yan *et al.* (2019) proposed a method (called DNRLMF-MDA) to predict miRNA–disease associations based on dynamic neighborhood regularized logistic matrix factorization. The Neighborhood Constraint Matrix Completion for miRNA–disease association prediction (NCMCMDA) was also applied to predict potential miRNA–disease associations (Chen *et al.*, 2021). EDTMDA was an integrating ensemble learning and dimensionality reduction computational framework to predict potential miRNA–disease associations based on an ensemble of Decision Tree (Chen *et al.*, 2019). Based on the Logistic Model Tree, LMTRDA (Wang *et al.*, 2019) was provided to predict miRNA–disease association, which also fuses multi-source information including miRNA sequences, miRNA functional similarity, disease semantic similarity and known miRNA–disease associations. By integrating the miRNA functional similarity, the disease semantic similarity and known miRNA–disease associations, an Extreme Gradient Boosting Machine model (called EGBMMDA) was also proposed to predict potential miRNA–disease associations (Chen *et al.*, 2018). By performing random sampling based on k-means clustering on negative samples, an adaptive boosting method (ABMDA) was also developed to predict potential miRNA–disease associations (Zhao *et al.*, 2019). MDAPCOM was a miRNA–disease association prediction method which extracted hybrid feature representation in the heterogeneous network that includes the known miRNA–disease association network (Liu *et al.*, 2020). DeepMDA was a deep ensemble model that extracts high-level features from similarity information using stacked auto encoders and then predicts miRNA–disease associations by adopting a 3-layer neural network (Fu and Peng, 2017). A network embedding-based heterogeneous information integration method has also been provided to predict miRNA–disease associations (Ji *et al.*, 2020). The heterogeneous network feature of these methods were based on miRNA–disease association network, which results that they cannot predict the types of associations and their candidate miRNA–disease pairs are restricted in the network. In addition, the Graph Convolutional Network model has also been used to predict miRNA–disease associations (Chu *et al.*, 2021; Pan *et al.*, 2019) and lncRNA-disease associations (Xuan *et al.*, 2019). These methods extract the feature from the known miRNA–disease association network or the heterogeneous network based on known miRNA–disease association network. They are also successful applications of graph neural networks (GNNs) in predicting miRNA–disease association.

Above computational methods have obtained good prediction results in miRNA–disease associations prediction. However, these methods can only handle miRNA–disease association matrix containing either 1 or 0 (1 represents known association, 0 represents unknown association), and some limitations that should be studied further: (i) those methods mainly predict the existence of associations between miRNAs and diseases, and they cannot predict the types of associations; (ii) they take the miRNA–disease association matrix as input which limits the association prediction to only candidate miRNA–disease pairs in the association matrix. The miRNA–disease association contains many types, such as genetics, epigenetics, circulation and target. Each type represents the different relation mechanism between miRNA and disease. For type genetics (SNP or deletion), the microRNA genes functioning as tumor suppressors can be down-regulated because of deletions, epigenetic silencing or loss of the expression of one or more transcription factors (Croce, 2008). MiR15 and miR16 are located at chromosome 13q14, a region deleted in more than half of B cell chronic lymphocytic leukemias (B-CLL) (Calin *et al.*, 2002). In addition, for type epigenetics (the methylation of CpG islands in their promoters), the epigenetic alterations affected the expression of tumor suppressor genes and lead to diseases (Calin and Baylin, 2006). The therapies directed toward the reversal of the epigenetic changes have developed in various malignancies, azacitidine (Vidaza; Celgene) has been used to reactivate tumor suppressor genes that have been silenced by the methylation of promoter CpG islands. For type circulation (associations identified from blood samples), these miRNAs are optimal biomarkers owing to high stability under storage and handling conditions and their presence in blood, urine and other body fluids (Schwarzenbach *et al.*, 2014). Additionally, circulating miRNAs are correlated with the degree of tumor progression and present differently at different stages of cancer (Cui *et al.*, 2019). A tremendous number of publications proposed that the circulating miRNAs, especially in serum and plasma, could potentially be used as diagnostic, prognostic and predictive biomarkers for different types of tumors (Armand-Labit and Pradines, 2017). For type target, miRNAs can bind to the UTR region of the mRNA and induce its degradation or repress its translation that leads to disease (Esteller, 2011). The miR-21 can directly target the MAP2K3 gene which is a tumor repressor gene, and inhibit its expression during the carcinogenesis of hepatocellular carcinoma, at both transcriptional and post-translational levels (Xu *et al.*, 2013). In addition, the studies also showed that targeting miR-21 in human PDA (pancreatic ductal adenocarcinoma (PDA) cell lines using lentiviral vectors (LVs) may impede tumor growth. Targeting these miRNAs for the disease therapy is a new opportunity (Sicard *et al.*, 2013).

Furthermore, the HMDD database has provided the association types between miRNAs and diseases, such as circulation, epigenetics, genetics and target. The identification of miRNA–disease association types is important to systematically understand the mechanism of miRNAs and diseases, which then improves the disease diagnosis and treatment efficiency. Specifically, for the association type circulation, the plasma concentrations of miRNAs miR-17-5p are found to be significantly higher in gastric cancer patients, so miRNA miR-17-5p can be considered as a biomarker of gastric cancer diagnosis (Tsujiura *et al.*, 2010). Moreover, the association type between miR-127 and colon cancer is epigenetics, miR-127 is embedded in a CpG island and silenced in colon cancer, which makes it to be considered as a possible tumor suppressor gene (Cahill *et al.*, 2007). The association type of genetics also can be a possible biomarker of human diseases, such as the association between let-7 and lung cancer (Williams, 2008).

In this study, by considering the importance of miRNA–disease association type and the limit of current methods, we propose an end-to-end deep learning method, namely PDMDA, to predict deep-level miRNA–disease associations based on the GNNs and sequence features. By integrating the disease–gene network and gene–gene interaction network, PDMDA extracts the disease feature representation based on GNN. In addition, based on the sequence and structural features of miRNAs, we extract miRNA feature representation by a fully connected network (FCN). After obtaining the feature representations of diseases and miRNAs, PDMDA concatenates them to predict the final association's label by a multilayer with three fully connected (FC) layers and a softmax layer. The fivefold cross-validation (5CV) is used to evaluate the prediction performances of

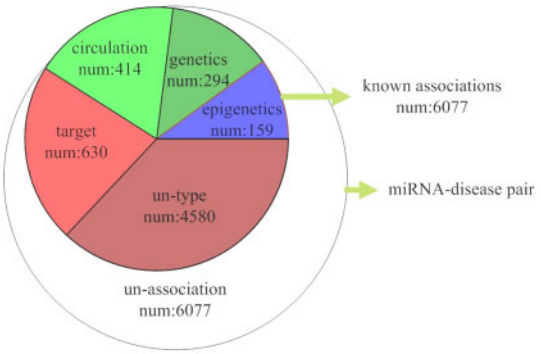

Fig. 1. The composition of the miRNA–disease associations. The un-type represents known associations but do not know types; the un-association represents randomly selected unknown associations; the num is number

our method. The area under the receiver operating characteristic curve (AUC) scores is used as the evaluation metrics. We conduct 5CV on three tasks for predicting miRNA–disease association type. The experiment results show that PDMDA can effectively predict deep-level miRNA–disease associations.

2 Materials

In this study, all known miRNA–disease associations are downloaded from the HMDD 2.0 database. The disease–gene associations are retrieved from the DisGeNET database which is one of the largest available collections of associations between human diseases and genes. The gene–gene interactions are downloaded from the HumanNet database which contains 16 243 genes and 476 399 interactions. We also obtain the pre-miRNA sequences and mature sequences from the miRbase database which is a primary repository and database resource for miRNA data (Griffiths-Jones *et al.*, 2008).

After sorting and projecting the downloaded data, we obtain the miRNA–disease association dataset, including 546 miRNAs, 333 diseases and 6077 miRNA–disease associations. In addition, the miRNA–disease association type dataset includes 414 circulations, 159 epigenetics, 630 target and 294 genetics association samples. We obtain 4580 un-type samples by removing these 4 types of association samples. In addition, by considering the balance of known association samples and unknown association samples, we randomly choose 6077 un-association samples from all unknown association samples. Figure 1 summarizes the miRNA–disease association composition and the number of different type samples. So, we use six types (circulations, epigenetics, target, genetics, un-type and un-association) of samples in this study.

3 Methods

The miRNA–disease association type prediction is the multi-class classification problem. Specifically, for the association type prediction, the association type scores of a miRNA–disease pair are in ranged [0, 1]. As shown in Figure 2, PDMDA includes three parts. The initialization layer takes the disease–gene interaction and gene–gene interaction networks to initialize the origin features of diseases and miRNAs. In GNN and FC layer, the feature representations of diseases and miRNAs are learned by GNN and FC layer models, respectively. Finally, after concatenating the feature vectors of miRNAs and diseases of GNN and FC layer, a multilayer with three FC layers and a softmax layer is designed to predict association types. In addition, the subfigure of GNN of diseases is shown in Figure 3, which includes four main processes: (i) the embedding process based on disease–gene interaction and gene–gene interaction network; (ii) the extraction process of subgraphs with radius r; (iii) the fusion of subgraphs; (iv) obtaining the disease feature representation based on the node vector.

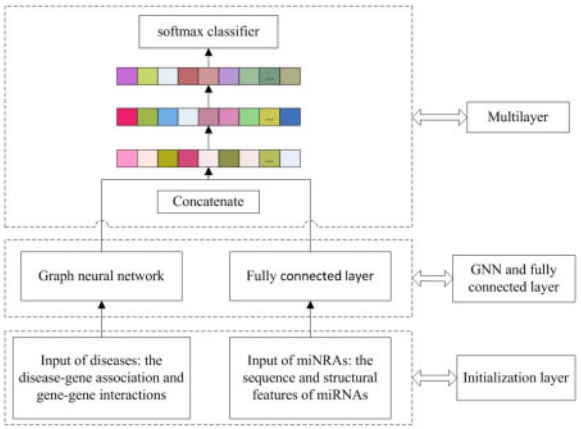

Fig. 2. The overview of the PDMDA approach

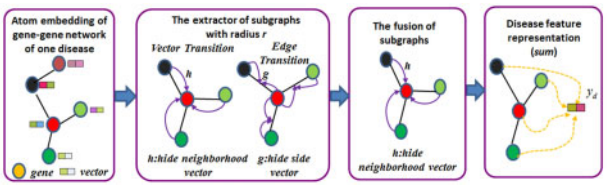

Fig. 3. The data flow subfigure of GNN of diseases

3.1 FCN for miRNA sequence

In this section, we describe the process of the miRNA feature representation. We all know that the production process of miRNAs includes two steps: (i) nuclear RNase III Drosha produces precursors (pre-miRNAs) from long primary miRNAs (Bartel, 2004), (ii) a pre-miRNA is cleaved to produce mature miRNAs. Therefore, by comprehensively considering the production process of mature miRNAs, we use the features of both mature-miRNAs and pre-miRNAs. Therefore, we extract 53 original features of miRNA as the input of a FCN to obtain miRNA feature representations. Table 1 summarizes the detail of all original miRNA features. Those features are divided into three categories including base sequence, ratio of base sequence and structure feature. The number of single base of pre-miRNAs, the Dinucleotide pairs in pre-miRNAs, the length of pre-miRNAs and the aggregate Dinucleotide pairs are the same scale and represented by integers. The MFE and nMFE are the same scale and represented by negative numbers. In addition, the ratio of base content of pre-miRNAs, the structure feature of pre-miRNAs, the ratio of Dinucleotide pairs in pre-miRNAs and the ratio of aggregate Dinucleotide pairs are the same scale and represented by floats. After obtaining the original miRNA features, we concatenate them to one vector and further extract the final miRNA features by an FC layer.

After obtaining the original miRNA features, we further extract the final miRNA features by an FC layer. This layer has 10 neurons, which contains a linear transformation structure applied to each position in the original feature vector.

$$FC(x) = W_m x + b_m, \tag{1}$$

where $W_m \in R^{d*d_{in}}$ and $b_m \in R^d$ are the transformation parameters and biases, respectively. x is the original miRNA feature. Therefore, we obtain the miRNA feature vector $\mathbf{y}_{mi}$. In this study, d and d_{in} are set to be 10 and 53, respectively.

3.2 GNN for disease graph

According to previous studies (Tsubaki *et al.*, 2019), a graph G can be mapped to a vector $\mathbf{y} \in R^d$ with two functions (transition and output), so we use a GNN to obtain the disease vector representations by the disease-related gene–gene interaction graph. In this

Table 1. The origin miRNA feature description

Category	Description	Number of features
The number of single base of pre-miRNAs	The number of single base X in pre-miRNAs, $X \in \{A, U, C, G\}$	4
The ratio of single base of pre-miRNAs	The $\%X$ ratio in pre-miRNAs, $X \in \{A, U, C, G\}$	4
The structure feature of pre-miRNAs	Normalized base-pairing propensity ($P(s)$), Normalized base-pairing propensity divided by its length ($nP(s)$), Normalized Shannon entropy ($Q(s)$), Normalized Shannon entropy divided by its length ($nQ(s)$), Normalized base-pair distance ($D(s)$), Normalized base-pair distance divided by its length ($nD(s)$)	6
Dinucleotide pairs in pre-miRNAs	The Dinucleotide pairs XY in pre-miRNAs, $X, Y \in \{A, U, C, G\}$	16
The ratio of Dinucleotide pairs in pre-miRNAs	The $\%XY$ ratio in pre-miRNAs, $X, Y \in \{A, U, C, G\}$	16
MFE and nMFE	The minimum free energy of pre-miRNA secondary structures and it is divided by its length	2
The length of pre-miRNAs	The sequence length of pre-miRNAs	1
Aggregate Dinucleotide pairs	The aggregate Dinucleotide $X + Y$ in pre-miRNAs, $X + Y \in \{A + U, C + G\}$	2
The ratio of Aggregate Dinucleotide pairs	The ratio of aggregate Dinucleotide $\%X + Y$ in pre-miRNAs, $X + Y \in \{A + U, C + G\}$	2

study, genes and their interactions are represented by vertices and edges in a graph G.

The iteration process of a GNN consists of two parts. The first one updates each vertex's information in consideration of its neighboring vertices and edges in graph. The other maps the set of vertices to vector. In this study, all parameters of these two processes are learned by backpropagation. It is an end-to-end learning method, the iteration times are set to be 100 based on the previous studies and experiment results. According to the GNN model, the transition function updates each vertex's information with its neighboring vertices and edges, and the output function maps to the set of vertices to vector $\mathbf{y}$. Both functions are achieved via neural networks.

Let V be the set of vertices, E be the set of edges in a graph $G = (V, E)$. In a particular disease, v_i is the ith related gene and e_{ij} is the interaction of the ith and jth genes. For a graph G, we first embed all genes and edges in a d-dimensional real-valued vector space in consideration of these types (Tsubaki *et al.*, 2019). By considering the fact that there is only one type in edges to affect the effect of representation learning, we divide the edges into 10 types according to their values range from 0 to 1 with 0.1 increments, and then take the r-radius subgraph to address this problem (Costa and De Grave, 2010). In this model, the feature of a vertex is induced by the neighboring vertices and edges within r radius. Specifically, $N(i, r)$ is a set of all neighboring vertex indices of the ith vertex within r radius. For vertex v_i, the r-radius subgraph is defined as follows:

$$v_i^{(r)} = (V_i^{(r)}, E_i^{(r)}), \tag{2}$$

where

$$\begin{aligned} V_i^{(r)} &= \{v_j | j \in N(i, r)\}, \\ E_i^{(r)} &= \{e_{mn} \in E | (m, n) \in N(i, r) * N(i, r-1)\}. \end{aligned} \tag{3}$$

Then, we also define the r-radius subgraph of e_{ij} as follows:

$$e_{ij}^{(r)} = (V_i^{(r)} \cup V_j^{(r)}, \quad E_i^{(r)} \cup E_j^{(r)}). \tag{4}$$

We assign an embedding (vector) to each r-radius vertex and r-radius edge, such as $V_i^{(r)} \in R^d$ and $\varepsilon_i^{(r)} \in R^d$. They are randomly initialized and subsequently learned during the training process.

After assigning embedding vectors to the r-radius vertex and r-radius edge, we use two transition functions to obtain their final vectors simultaneously. All vertex embeddings can gradually gather more global information on a graph. Specifically, for vertex v_i, its embedding at time step t is defined as $\mathbf{v}_i^{(t)} \in R^d$. Then the transition function of vertex is defined as follows:

$$\mathbf{v}_i^{(t+1)} = \sigma(\mathbf{v}_i^{(t)} + \textstyle\sum_{j \in N(i)} \mathbf{h}_{ij}^{(t)}), \tag{5}$$

where σ is the element-wise sigmoid function, and $\mathbf{h}_{ij}^{(t)} \in R^d$ is the hidden neighborhood vector. The computed progress of $\mathbf{h}_{ij}^{(t)}$ is based on the neighboring vertex v_i and edge e_{ij}, which is defined as follows:

$$\mathbf{h}_{ij}^{(t)} = f(\mathbf{W}_{ne}[\begin{matrix}\mathbf{v}_i^{(t)} \\ \mathbf{e}_{ij}^{(t)}\end{matrix}] + \mathbf{b}_{ne}). \tag{6}$$

where f is the activation function. In this study, we use $ReLU(f(x) = max(0, x))$ as the activation function. In addition, $\mathbf{W}_{ne} \in R^{d*2d}$ and $\mathbf{b}_{ne} \in R^d$ are the weight matrix and the bias vector, respectively. The $\mathbf{e}_i^{(t)}$ is the edge embedding of edge e_{ij} at time step t.

The iterative process of edge embedding is also defined in a similar manner. Specifically, the transition function of the edge is defined as follows:

$$\mathbf{e}_{ij}^{(t+1)} = \sigma(\mathbf{e}_{ij}^{(t)} + \mathbf{g}_{ij}^{(t)}), \tag{7}$$

where $\mathbf{g}_{ij}^{(t)} \in R^d$ is the hidden side vector. The update process of $\mathbf{g}_{ij}^{(t)}$ is based on the side vertex embeddings $\mathbf{v}_i^{(t)}$ and $\mathbf{v}_j^{(t)}$, and is defined as follows:

$$\mathbf{g}_{ij}^{(t)} = f(\mathbf{W}_{si}(\mathbf{v}_i^{(t)} + \mathbf{v}_j^{(t)}) + \mathbf{b}_{si}), \tag{8}$$

where $\mathbf{W}_{si} \in R^{d*d}$ and $\mathbf{b}_{si} \in R^d$ are the weight matrix and bias vector, respectively.

After obtaining the vertex vectors by the transition function, we can compute the final disease vector representation by the average of the vertex vectors, and the specific process is defined as follows:

$$\mathbf{y}_{di} = \frac{1}{|V|} \sum_{i=1}^{|V|} \mathbf{v}_i^{(t)}, \tag{9}$$

where $|V|$ is the number of vertices in the disease-related gene–gene interaction graph.

3.3 Deep-level miRNA–disease association prediction

After obtaining the final miRNA feature vector $\mathbf{y}_{mi}$ and disease feature vector $\mathbf{y}_{di}$, we concatenate them as the input of FC model. The final feature vector $\mathbf{y}_c = [\mathbf{y}_{mi}; \mathbf{y}_{di}] \in R^{2d}$ of the miRNA–disease pair is computed by FC model with three layers, which is also used as input to the miRNA–disease association classifier:

Table 2. The used data in predicting deep-level types of miRNA–disease association

The used different type samples on three tasks						
Task	Circulations	Epigenetics	Target	Genetics	Un-type	Un-association
Task1	414	159	630	294	0	0
Task2	414	159	630	294	4580	0
Task3	414	159	630	294	4580	6077

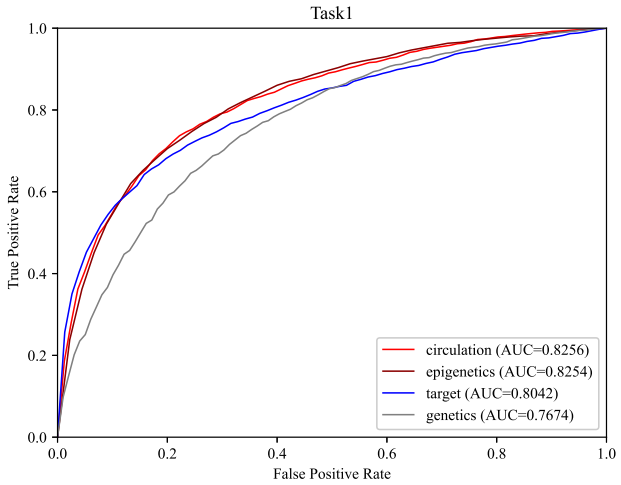

Fig. 4. The ROC curves for association type prediction of task1

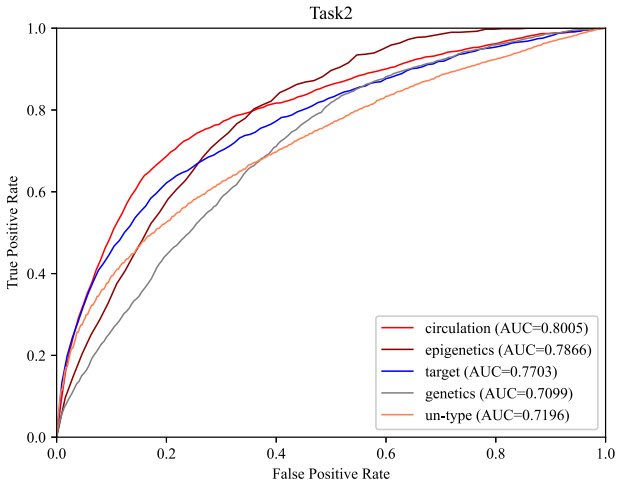

Fig. 5. The ROC curves for association type prediction of task2

$$z = \mathbf{W}_o \mathbf{y}_c + \mathbf{b}_o, \tag{10}$$

where $\mathbf{W}_o \in R^{k*2d}$ and $\mathbf{b}_o \in R^k$ are the weight matrix and the bias vector, respectively.

In this study, k is the number of miRNA–disease association types. We set k to 4, 5 and 6 when conducting task1, task2 and task3, respectively. Finally, based on the output vector $z = [o_0, o_1, \ldots, o_{k-1}]$, the miRNA–disease association types probability can be computed by a softmax layer, which is defined as follows:

$$p_l = \frac{\exp(o_l)}{\sum_i o_i}, \tag{11}$$

where $l \in \{0, 1, \ldots, k-1\}$ is the label and p_l is the probability of label l. Therefore, we can use the softmax function to predict miRNA–disease association types.

In addition, we use the cross-entropy loss as the loss function which is defined as follows:

$$loss_{cross} = -\frac{1}{N}\sum_{i}^{N}\sum_{l}^{k-1} y_{i,l} \log(p_{i,l}), \tag{12}$$

where $y_{i,l}$ and $p_{i,l}$ are the real and predicted one-hot representation on label l of ith sample, respectively. If the ith sample belongs to lable l, then $y_{i,l} = 1$, otherwise $y_{i,l} = 0$. N is the number of miRNA–disease pairs in the training dataset. Therefore, the training objective is to minimize the loss function *loss* and is defined as follows:

$$loss(\theta) = -\frac{1}{N}\sum_{i}^{N}\sum_{l}^{k-1} y_{i,l} \log(p_{i,l}) + \frac{\lambda}{2}||\theta||_2^2, \tag{13}$$

where θ is the set of all parameters in the model, including weight matrices, bias vectors, embeddings of miRNA and embeddings of disease. The parameter λ is the regularization hyper-parameter. Similarly, the backpropagation algorithm is used to learn θ.

4 Results

4.1 Experiments

In this study, we conduct deep-level miRNA–disease association type prediction to evaluate the performance of PDMDA. We divide the predicting deep-level types of miRNA–disease association into three tasks to evaluate the prediction ability among four known association type samples, un-type samples and un-associationocation samples. Task1 is predicting association type among four known association type samples, task2 is predicting type among four known association type samples and un-type samples, task3 is predicting the type among four known association type samples, un-type samples and un-association samples. Table 2 shows that task1, task2 and task3 contain four, five and six different types of samples, respectively. The miRNA–disease association type prediction experiment is conducted by the fivefold cross-validation (5CV). The AUC is used as metrics to evaluate the prediction performance. ADAM is chosen as the optimizer of the neural networks for our proposed method, which is one of the SGD-based algorithms. The radius r is chosen from set {1, 2, 3}. The vector dimensionality d of vertices and edges is chosen from set {5, 10, 20}. The regularization parameter λ is chosen from set $\{10^{-5}, 10^{-6}, 10^{-7}\}$. The default values of these parameters are set by conducting 5CV.

4.2 The prediction performances of experiments

Predicting deep-level miRNA–disease association is important to systemically understand the association mechanism between miRNAs and diseases.

Figure 4 shows the ROC curves of PDMDA based on miRNA–disease association type samples of task1 in 5CV. AUC values of PDMDA are more than 0.8 on three types: circulation, epigenetics and target. The average AUC of all miRNA–disease association type also reaches 0.8056 based on miRNA–disease association type samples of task1 in 5CV. Besides, the AUC value reaches 0.7674 on genetics which indicates that our method is effective in predicting miRNA–disease association type based on the association type samples of task1.

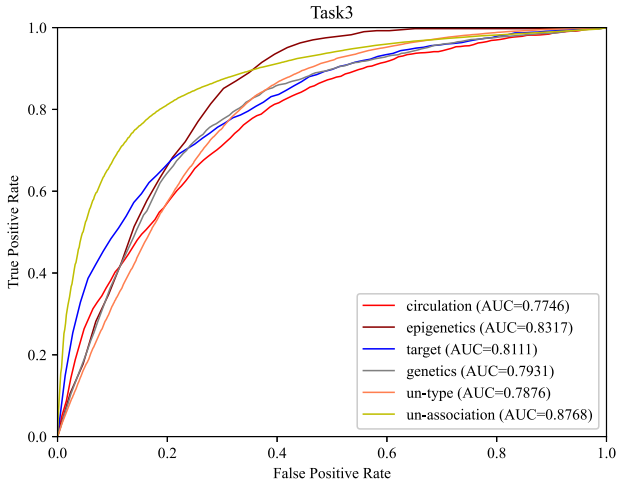

Fig. 6. The ROC curves for association type prediction of task3

Table 3. The performance of three methods on miRNA–disease association prediction

Method	AUC	Precision	Recall	F1
PDMDA	0.8863	0.8057	0.8223	0.8140
EGBMMDA	0.8823	0.8118	0.8139	0.8126
ABMDA	0.8720	0.9488	0.2314	0.3708

Table 4. The performance of *de novo* validation for PDMDA and other comparative methods

Method	AUC	Precision	Recall	F1
PDMDA	0.8995	0.1441	0.7828	0.2027
EGBMMDA	0.7217	0.0487	0.5810	0.0898
ABMDA	0.7960	0.2423	0.1585	0.1385

Figure 5 also shows the prediction performances of PDMDA based on miRNA–disease association type of task2 in 5CV. We can see from Figure 5 that PDMDA obtains AUC values of 0.8005, 0.7866, 0.7703 and 0.7099, on circulation, epigenetics, target and genetics, respectively. The average AUC of all miRNA–disease association type reaches 0.7573 based on miRNA–disease association type samples of task2 in 5CV. In addition, the AUC value of un-type reaches 0.7196.

Figure 6 shows the prediction performances of PDMDA based on association samples of task3 in 5CV. On circulation, epigenetics, target and genetics, the AUC values are 0.7746, 0.8317, 0.8111 and 0.7931, respectively. In addition, on un-type and un-association, AUC values also reach 0.7876 and 0.8768, respectively. The average AUC of all miRNA–disease association type reaches 0.8124 based on miRNA–disease association type samples of task3 in 5CV.

In summary, PDMDA obtains effective prediction performance on three tasks, the average AUC values reach 0.8056, 0.7573 and 0.8124, respectively. However, the imbalance of different association type samples has maybe influence on their predictive performance. Especially on task2 and task3, the number of un-type samples and un-association samples are 4580 and 6077 which are larger than numbers of circulation, epigenetics, target and genetics type samples, respectively.

4.3 miRNA–disease association prediction

PDMDA is also able to predict miRNA–disease associations. In addition, we conduct miRNA–disease association prediction to evaluate the performance of PDMDA and other compared methods by the 5CV and *de novo* validation.

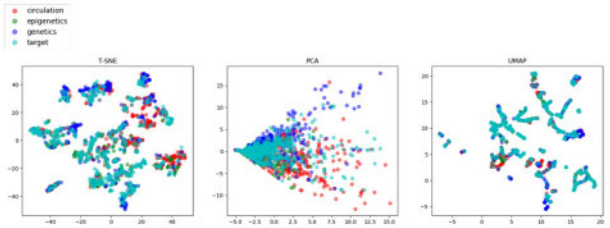

Fig. 7. The feature vectors of the test sets in the final multilayer are visualized after dimensionality reduction by t-SNE, PCA and UMAP on task1. The red circle, green circle, blue circle and cyan circle represent the circulation, epigenetics, genetics and target, respectively

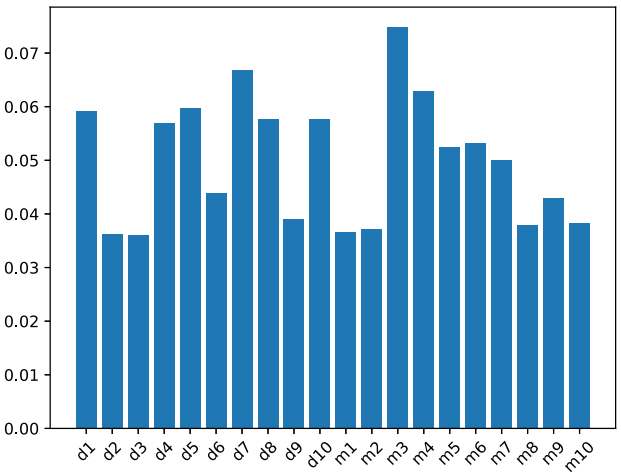

Fig. 8. The relative importance of the feature of final multilayer on the four known association type samples of task1

We conduct miRNA–disease association prediction by the 5CV based on the 6077 known association samples and 6077 unknown association samples which are randomly selected. Table 3 shows the performances of three methods on miRNA–disease association prediction with 5CV. We can see from Table 3 that PDMDA is comparable to the comparative methods in the miRNA–disease association prediction in terms of AUC. Besides, PDMDA can also obtain the highest F1-score, and its precision and recall scores are more than 0.8. It also shows that although PDMDA is dedicated to predicting deep-level miRNA–disease associations, it is comparable in miRNA–disease association prediction with 5CV.

In addition, *de novo* miRNA validation is also an important part to evaluate the performance of computational methods. We further conduct *de novo* miRNA validation with miRNA–disease associations. The used dataset for *de novo* validation includes 546 miRNAs, 333 diseases and 6077 miRNA–disease associations. We randomly choose 50 miRNAs to conduct *de novo* validation in miRNA–disease association prediction to avoid the computation time too long. We conduct *de novo* validation on each miRNA in turn for these 50 miRNAs. In *de novo* validation of an miRNA, the known disease associations of this miRNA are removed, and the miRNA will have no association information during the process of prediction. Then the existing associations of other miRNAs are used as training samples, the removed associations of this miRNA are used for evaluation. Table 4 illustrates the performance of *de novo* miRNA validation for PDMDA and the other two comparative methods. We can see that PMMDA obtains better prediction performance than ABMDA and EGBMMDA in terms of AUC and F1-score values.

5 Model and parameter analysis

In this study, we also analyze the feature learning ability, the relative importance of the feature and the prediction performance influence of parameters. The feature learning ability is analyzed in 5CV of the final multilayer under three association type prediction experiment condition. The relative importance of the feature is based on the features of final multilayer on the miRNA–disease association type samples of task1. The prediction performance influence of parameters is analyzed in 5CV with miRNA–disease association type prediction under the association type samples of task1. We project the feature vectors derived from each layer to two-dimensional feature space and visualize the result of miRNA–disease pair association classification to illustrate the feature learning ability of PDMDA. Figure 7 shows the visualized results of the feature vectors of the test sets in the final multilayer after dimensionality reduction by t-SNE (Maaten and Hinton, 2008), PCA (Abdi and Williams, 2010) and UMAP (McInnes *et al.*, 2018) on task1. We can see from Figure 7 that although the four known association type samples are not completely balanced, our method also can distinguish them. In addition, among these three dimensionality reduction methods, the t-SNE method is relatively more obvious in distinguishing various types of associations.

Furthermore, to demonstrate the extracted features in the prediction method, we further analyze the relative importance of final multilayer on the four known association type samples of task1. Figure 8 plots the relative importance of the features, which is computed by the XGBoost package. We can see from Figure 8 that eight features are relatively obvious, and all features worked. It also demonstrates that the extracted features can reflect the intrinsic characteristics of miRNA–disease pair.

To evaluate the influence of parameters in PDMDA, we analyze parameters radius r and dimensionality d which are used in GNN, and regularization parameter λ by conducting 5CV in miRNA–disease association type prediction based on association type samples of task1. We assign the default value (10^{-6}) to λ when analyzing parameters r and d. Similarly, we also assign the default values (2 and 10) to r and d when analyzing parameter λ.

Table 5 shows the average AUC scores of parameters r and d via fivefold cross-validation with miRNA–disease association type prediction based on association type samples of task1. We conduct a grid searching method to analyze them. We can see from Table 5 that the average AUC has the slightly increases when r from 1 to 2, and decrease from 2 to 3. It also illustrates that the radius r of embedding process has effect to improve the prediction performance of PDMDA. Furthermore, the results also show that r from 1 to 2 and d from 5 to 10 have little effect on prediction performance of PDMDA. Our method obtains best prediction performance when r and d are set to 2 and 10, respectively. In addition, when λ ranges from 10^{-5} to 10^{-7}, the AUC values of PDMDA are 0.8001, 0.8056 and 0.8023, respectively. Therefore, in this study, we set the default value of r and d to 2 and 10, respectively.

Table 5. The performance of PDMDA in different values of radius r and dimensionality d in miRNA–disease association type prediction based on association type samples of task1

	r		
	***r* = 1**	***r* = 2**	***r* = 3**
D			
$d = 5$	0.8017	0.8011	0.7881
$d = 10$	0.8023	**0.8056**	0.8033
$d = 20$	0.7885	0.7902	0.7893

The bold in the table means the best results.

6 Case study

To further evaluate the performance of our method in practical application, we validate the predicted new miRNAs associated with Colorectal Neoplasm (Colorectal Cancer) based on the other three independent databases (dbDEMC, miRCancer, mir2disease) and previous studies. Case studies are conducted to the predicted association type validation of diseases Colorectal Neoplasm.

Table 6 shows the validation result of the top five predicted miRNAs of each association type for Colorectal Neoplasm by PDMDA. We can see from Table 6 that all associations of top five related miRNAs of epigenetics, targets and genetics are validated in databases or previous studies. For example, up-regulation of hsa-mir-592 correlates with tumor progression and poor prognosis in patients with colorectal cancer. In addition, hsa-mir-519d is also up-regulated in Colorectal Neoplasm. However, hsa-mir-126, hsa-mir-15b, hsa-mir-212, hsa-mir-375 and hsa-mir-1247 are down-regulated in Colorectal Neoplasm.

Furthermore, there are only nine association types of top five related miRNAs are validated in miRCancer, dbDEMC 2.0 and

Table 6. The result of the top five predicted miRNAs of each association type for Colorectal Neoplasm by PDMDA

Association types	Rank	Association miRNA	Evidence (association)	Evidence (association types)
Circulation	1	hsa-mir-592	miRCancer, dbDEMC 2.0	dbDEMC 2.0
	2	hsa-mir-375	miRCancer, dbDEMC 2.0	Unknown
	3	hsa-mir-1247	dbDEMC 2.0	dbDEMC 2.0
	4	hsa-mir-498	miRCancer, dbDEMC 2.0	dbDEMC 2.0
	5	hsa-mir-767	dbDEMC 2.0	dbDEMC 2.0
Epigenetics	1	hsa-mir-202	miRCancer, dbDEMC 2.0	Unknown
	2	hsa-mir-126	miRCancer, dbDEMC 2.0	Unknown
	3	hsa-mir-34c	miRCancer, dbDEMC 2.0	dbDEMC 2.0
	4	hsa-mir-15b	miRCancer, dbDEMC 2.0	Unknown
	5	hsa-mir-212	miRCancer, dbDEMC 2.0	dbDEMC 2.0
Target	1	hsa-mir-186	dbDEMC 2.0	Literature (Islam *et al.*, 2017)
	2	hsa-mir-527	dbDEMC 2.0	Unknown
	3	hsa-mir-548d-2	dbDEMC 2.0	Unknown
	4	hsa-mir-320d-1	dbDEMC 2.0	Unknown
	5	hsa-mir-519d	miRCancer, dbDEMC 2.0	dbDEMC 2.0
Genetics	1	hsa-mir-1302-1	dbDEMC 2.0	Unknown
	2	hsa-mir-144	dbDEMC 2.0	Unknown
	3	hsa-mir-200b	miRCancer	miRCancer
	4	hsa-mir-1302-8	dbDEMC 2.0	Unknown
	5	hsa-mir-1-2	miRCancer	Unknown

previous studies. Based on the analysis of blood from colorectal cancer, circulating miRNAs hsa-mir-592, hsa-mir-1247, hsa-mir-498 and hsa-mir-767 could serve as biomarkers for the accurate detection of colorectal cancer. In addition, epigenetics type of hsa-miR-34c and hsa-mir-34c are also validated by dbDEMC 2.0. For hsa-mir-34c, the interrupted E2F1-miR-34c-SCF negative feedback loop by hyper-methylation promotes colorectal cancer cell proliferation. In addition, genetic and epigenetic down-regulation of miRNA-212 also promotes Colorectal Tumor Metastasis via Dysregulation of MnSOD. However, hsa-miR-186 serves as a promoter for the migration of colon cancer cells by targeting RETREG1 (Islam *et al.*, 2017). Has-mir-519d inhibits cell proliferation and migration by targeting TROAP in colorectal cancer. Therefore, the association types of Hsa-miR-186 and has-miRNA-519d are target.

The results of the case study show that the predicted miRNA–disease associations are more easily verified than predicted miRNA–disease association types based on constructed databases and previous studies. This is because that previous studies only describe the up-regulated or down-regulated associations but do not describe association types.

7 Conclusions

In this study, we have proposed a new GNN-based framework, named PDMDA, for predicting deep-level miRNA–disease associations. Firstly, the miRNA feature representation of miRNAs is extracted by a FCN based on the sequence and structural features of miRNAs. Then the disease feature representation is extracted based on the GNN by integrating the disease–gene network and protein-protein interaction network. Finally, the association label of miRNA–disease pairs is predicted by a multiplayer network. PDMDA is the first time to use GNN to extract disease feature representation from disease–gene association and PPI network. It is noteworthy that PDMDA neither takes the miRNA–disease association matrix as input nor calculates miRNA and disease GIP similarities.

Although we provide an effective method to predict deep-level miRNA–disease associations, there is still room for improvement. Firstly, more biological information should be considered and analyzed during the prediction process, such as miRNA-target associations and disease ontology. Besides, other new deep learning technology should be reviewed and implemented, such as the attention mechanism. In conclusion, we would like to develop a more effective method for predicting deep-level miRNA–disease associations by using biological information and new deep learning models.

Funding

This work was supported by NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization [U1909208]; National Natural Science Foundation of China [61962050 and 62072473]; 111 Project [B18059]; the Science and Technology Foundation of Guizhou Province of China ([2020]1Y264).

Conflict of Interest: none declared.

References

Abdi,H. and Williams,L.J. (2010) Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.*, **2**, 433–459.

Akao,Y. *et al.* (2007) Downregulation of microRNAs-143 and -145 in B-cell malignancies. *Cancer*, **98**, 1914–1920.

Armand-Labit,V. and Pradines,A. (2017) Circulating cell-free microRNAs as clinical cancer biomarkers. *Biomol. Concepts*, **8**, 61–81.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Cahill,S. *et al.* (2007) Effect of BRAF V600E mutation on transcription and post-transcriptional regulation in a papillary thyroid carcinoma model. *Mol. Cancer*, **6**, 21.

Calin,G.A. and Croce,C.M. (2006) MicroRNA signatures in human cancers. *Nat. Rev. Cancer*, **6**, 857–866.

Calin,G.A. *et al.* (2002) Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. USA*, **99**, 15524–15529.

Chen,X. *et al.* (2018) EGBMMDA: extreme gradient boosting machine for MiRNA–disease association prediction. *Cell Death Dis.*, **9**, 3–16.

Chen,X. *et al.* (2019) Ensemble of decision tree reveals potential miRNA–disease associations. *PLoS Comput. Biol.*, **15**, e1007209.

Chen,X. *et al.* (2021) Ncmcmda: mirna–disease association prediction through neighborhood constraint matrix completion. *Brief. Bioinf.*, **22**, 485–496. doi:10.1093/bib/bbz159.

Chu,Y. *et al.* (2021) MDA-GCNFTG: identifying miRNA–disease associations based on graph convolutional networks via graph sampling through the feature and topology graph. *Brief. Bioinf.*, **22**, bbab165.

Costa,F. and De Grave,K. (2010) Fast neighborhood subgraph pairwise distance kernel. In: *Proceedings of the 26th International Conference on Machine Learning*. Omnipress, Madison, WI, USA, pp. 255–262.

Croce,C.M. (2008) Oncogenes and cancer. *N. Engl. J. Med.*, **358**, 502–511.

Cui,M. *et al.* (2019) Circulating microRNAs in cancer: potential and challenge. *Front. Genet.*, **10**, 626.

Esteller,M. (2011) Non-coding RNAs in human disease. *Nat. Rev. Genet.*, **12**, 861–874.

Friedman,R.,C. *et al.* (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.

Fu,L. and Peng,Q. (2017) A deep ensemble model to predict miRNA–disease association. *Sci. Rep.*, **7**, 1–13.

Griffiths-Jones,S. *et al.* (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.

Huang,Z. *et al.* (2019) HMDD v3. 0: a database for experimentally supported human microRNA disease associations. *Nucleic Acids Res.*, **47**, D1013–D1017.

Islam,F. *et al.* (2017) MicroRNA-186-5p overexpression modulates colon cancer growth by repressing the expression of the FAM134B tumour inhibitor. *Exp. Cell Res.*, **357**, 260–270.

Jiang,Q. *et al.* (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, **37**, D98–D104.

Ji,B.Y. *et al.* (2020) Predicting miRNA–disease association from heterogeneous information network with GraRep embedding model. *Sci. Rep.*, **10**, 1–12.

Kim,V.N. (2005) MicroRNA biogenesis: coordinated cropping and dicing. *Nat. Rev. Mol. Cell Biol.*, **6**, 376–385.

Liu,M. *et al.* (2020) Predicting miRNA–disease associations using a hybrid feature representation in the heterogeneous network. *BMC Med. Genomics*, **13**, 1–11.

Li,Y. *et al.* (2014) HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.*, **42**, D1070–D1074.

Maaten,L. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

McInnes,L. *et al.* (2018) Umap: uniform manifold approximation and projection for dimension reduction.*arXiv e-Print arXiv*:1802.03426.

Miyaki,S. *et al.* (2009) MicroRNA-140 is expressed in differentiated human articular chondrocytes and modulates interleukin-1 responses. *Arthritis Rheum.*, **60**, 2723–2730.

Pan,X. *et al.* (2019) Inferring disease-associated microRNAs using semi-supervised multi-label graph convolutional networks. *Iscience*, **20**, 265–277.

Schwarzenbach,H. *et al.* (2014) Clinical relevance of circulating cell-free microRNAs in cancer. *Nat. Rev. Clin. Oncol.*, **11**, 145–156.

Sicard,F. *et al.* (2013) Targeting miR-21 for the therapy of pancreatic cancer. *Mol. Ther.*, **21**, 986–994.

Tsubaki,M. *et al.* (2019) Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, **35**, 309–318.

Tsujiura,M. *et al.* (2010) Circulating microRNAs in plasma of patients with gastric cancers. *Br. J. Cancer*, **102**, 1174–1179.

Van Wynsberghe,P.M. and Pasquinelli,A.E. (2014) Period homolog LIN-42 regulates miRNA transcription to impact developmental timing. *Worm*, **3**, e974453.

Wang,D. *et al.* (2014) OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs. *Bioinformatics*, **30**, 2237–2238.

Wang,L. *et al.* (2019) LMTRDA: using logistic model tree to predict MiRNA–disease associations by fusing multi-source information of sequences and similarities. *PLoS Comput. Biol.*, **15**, e1006865.

Williams,A.E. (2008) Functional aspects of animal microRNAs. *Cell. Mol. Life Sci.*, **65**, 545–562.

Xie,B. *et al.* (2013) miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinformatics*, **29**, 638–644.

Xuan,P. *et al.* (2019) Graph convolutional network and convolutional neural network based method for predicting lncRNA-disease associations. *Cells*, **8**, 1012.

Xu,G. *et al.* (2013) MicroRNA-21 promotes hepatocellular carcinoma HepG2 cell proliferation through repression of mitogen-activated protein kinase-kinase 3. *BMC Cancer*, **13**, 469.

Yang,Z. *et al.* (2010) dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC Genomics*, **11**, S5.

Yan,C. *et al.* (2019) DNRLMF-MDA: predicting microRNA-disease associations based on similarities of microRNAs and diseases. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **16**, 233–243.

Zhao,Y. *et al.* (2019) Adaptive boosting-based computational model for predicting potential miRNA–disease associations. *Bioinformatics*, **35**, 4730–4738.

Zhu,Y. *et al.* (2009) A microRNA gene is hosted in an intron of a schizophrenia-susceptibility gene. *Schizophrenia Res.*, **109**, 86–89.