The Prediction of MiRNA-disease Associations Using Random Walk with Restart *

Jheng-Yan Lyu[a], Kuo-Tsung Tseng[b] and Chang-Biau Yang[a†]
[a]Department of Computer Science and Engineering
National Sun Yat-sen University, Kaohsiung, Taiwan
[†]cbyang@cse.nsysu.edu.tw
[b]Department of Shipping and Transportation Management
National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan

Abstract

In the miRNA-disease association prediction problem*, we are given disease semantic similarity, miRNA similarity and some known miRNA-disease associations. Then, its goal is to predict unknown potential miRNA-disease associations. Liu et al. [18] used the random walk with restart* (*RWR*) *to determine the association ranks of all possible miRNAs with a query disease, so that the unknown possible miRNAs associated with the query disease can be found out. The miRNA similarity used by them is the miRNA functional similarity. In this paper, we follow the method of Liu et al. To enhance the prediction performance, we add the target gene similarities of miRNAs, and apply the threshold concept to the transition matrix of RWR to reduce noise interference. The experimental dataset comes from HMDD. The average area under the curve* (*AUC*) *of our method is* 0.8836*, which is better than* 0.8266 *of the method of Liu et al. For a subset of 15 diseases proposed by Liu et al., we find out 76% associated miRNAs in average under the threshold 30%, which is higher than 72% of the method of Liu et al.*

Keywords: miRNA, miRNA-disease association, functional similarity, target gene, random walk with restart, leave-one-out cross-validation

1 Introduction

A microRNA (abbreviated as miRNA) is a non-coding RNA molecule with about 21 to 23 nucleotides, which may be the regulation of a certain gene expression. miRNAs can be found in plants, animals, and some viruses. In early 2000, several researchers [21, 22] revealed the regulatory functions of miRNAs. They showed that different groups of miRNAs may have different expression in different cell types and species. Besides, the miRNAs may have multiple functions in diseases and other biological processes [3, 6, 12, 20, 25], such as cancer [4], nervous system diseases, cell proliferation, apoptosis, growth, and differentiation [10, 19]. Therefore, identifying associations between miRNAs and diseases is an important issue on biomedical researches.

Some researchers have developed various methods to calculate the similarities among miRNAs or diseases, such as MISIM [17], MiRGOFS [28] and so on. MISIM calculates the disease similarities with directed acyclic graphs (DAGs) based on MeSH disease terms. And the miRNA functional similarities are calculated by measuring the similarities between miRNAs by associated disease expressed in the DAG structure. In MiRGOFS, the miRNA functional similarities are calculated by measuring GO [2] semantic similarity metric, where each GO term is weighted with its statistical significance.

The cost of experiments for confirming the association relationships between miRNAs and diseases is very expensive. Several studies tried to develop prediction methods to find potential associations among miRNAs and diseases.

We follow the method of Liu *et al.* [18] to solve the miRNA-disease association prediction problem. In addition to disease semantic similar-

*This research work was partially supported by the Ministry of Science and Technology of Taiwan under contract MOST 108-2221-E-110-031.
†Corresponding author.

ity, miRNA functional similarity and some known miRNA-disease associations, we add the target gene similarities between miRNAs to enhance the prediction results. The target genes of miRNA mean that the genes are regulated by miRNA in organism. Since miRNA can regulate gene expression, it is thought that the similar genes shared by two miRNAs could have the similar functions. And we also apply thresholds θ_d and θ_m to the transition matrix in random walk with restart to reduce noise interference.

The dataset used for our experiments comes from HMDD [24], in which the numbers of miRNAs, diseases and miRNA-disease associations are 806, 629 and 13534, respectively. The area under the curve (AUC) value with leave-one-out cross-validation (LOOCV) is 0.8836. The AUC values of other methods are obtained from their published articles, including 0.7580 of Jiang *et al.* [15], 0.8049 of Liu *et al.* [18], 0.8770 of Yang *et al.* [28], and 0.8460 of Chen *et al.* [7]. Note that the datasets used in different methods are different. The AUC of the method of Liu *et al.* [18] with our dataset is 0.8266.

The organization of this paper is structured as follows. In Section 2, we introduce the miRNA-disease association prediction problem, disease semantic similarity, miRNA functional similarity, and random walk with restart as background knowledge. Section 3 presents our method, including creating the transition matrix of random walk with restart, and combining the results of two methods measuring miRNA similarities (miRNA functional similarity and target gene similarities of miRNAs). Our experimental results are shown in Section 4. Finally, we give our conclusions in Section 5.

2 Preliminaries

2.1 The Definition of the MiRNA-disease Association Prediction Problem

Figure 1 illustrates an example of miRNA-disease association. Each circle/square is a disease/miRNA. The name of each disease is denoted as d_i, and the name of each miRNA is denoted as m_i. Solid lines represent the known associations between miRNAs and diseases. The numbers beside lines represent similarity scores. Dashed lines are some unknown possible miRNA-disease associations.

For example, the disease d_1 is associated with miRNAs m_1 and m_2. The similarity score between m_1 and m_3 is 6, which is very high. It implies that m_1 may "recommend" m_3 as a possible potential association with disease d_1. In addition, the disease d_2 is close to the disease d_1 with high score 3, so that miRNAs associated with d_2 may also associate with d_1. m_1, m_3 and m_4 are associated with d_2. It implies that d_2 may "recommend" m_3 and m_4 as possible potential associations with disease d_1. m_3 is recommended by both m_1 and d_2, so m_3 may be a strong potential association with d_1.

Formally, the problem is defined as follows.

Definition 1. [7–9, 18, 26, 29] *Given a disease similarity matrix, a miRNA similarity matrix and a matrix of miRNA-disease associations, the* miRNA-disease association prediction *problem aims to predict possible potential* (*unknow*) *miRNA-disease associations with these three matrices.*

2.2 Disease Semantic Similarity

The disease semantic similarity calculation problem was proposed by Wang *et al.* [24], and their method for solving the problem is based on the MeSH [1] descriptor. This descriptor describes the structure of a hierarchical directed acyclic graph (DAG) with disease similarity. For example, the DAG of liver neoplasms is shown in Figure 2. Each node on layer i is a child of another node on layer $i+1$ if there is a link between them.

First, the semantic value of a disease is calculated with Equations 1 and 2. For a disease d and its ancestor t, if $t = d$, which means that disease term is the same, the semantic value is 1. Otherwise, we calculate the maximal value of children of t. Then we accumulate the values of $S_d(t)$ as the semantic value of d. T_d in Equation 2 denotes the set of all ancestors of d.

$$S_d(t) = \begin{cases} 1 & \text{if } t = d, \\ max\{0.5 \times S_d(t') | t' \text{ is a child of } t\} & \text{if } t \neq d. \end{cases} \quad (1)$$

$$DV(d) = \sum_{t \in T_d} S_d(t). \quad (2)$$

For example, in Figure 2, suppose that we would like to calculate the semantic value of the disease term *liver neoplasms*. The calculation is given as follows. $DV(liver\ neoplasms) = 1.0 \times S_d(liver\ neoplasms) + 0.5 \times S_d(digestive\ system\ neoplasms) + 0.5 \times S_d(liver\ diseases) +$

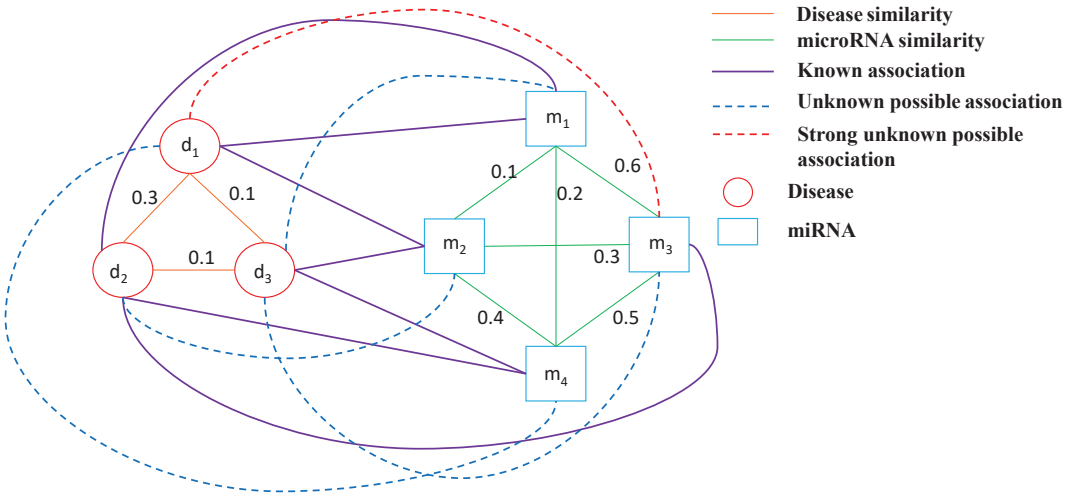

Figure 1: An example of miRNA-disease associations [29].

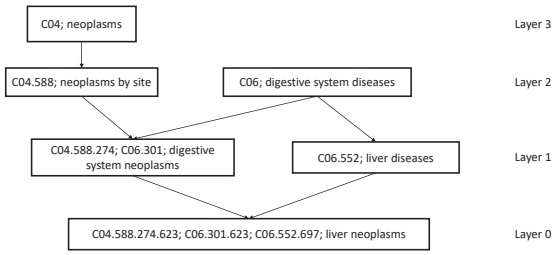

Figure 2: The DAG of liver neoplasms [18].

$0.25 \times S_d(digestive\ system\ diseases) + 0.25 \times S_d(neoplasms\ by\ site) + 0.125 \times S_d(neoplasms) = 2.625$.

Given two diseases d_i and d_j, the disease semantic similarity of d_i and d_j is calculated as follows.

$$DS(d_i, d_j) = \frac{\sum_{t \in T_{d_i} \cap T_{d_j}} (S_{d_i}(t) + S_{d_j}(t))}{DV(d_i) + DV(d_j)}. \quad (3)$$

2.3 The miRNA Functional Similarity

Suppose that n_i and n_j diseases are associated with two given miRNAs m_i and m_j, respectively. The method of Wang *et al.* [24] for measuring the functional similarity of m_i and m_j is given as follows.

Step 1: Use Equations 1 and 2 to calculate the semantic value of each disease which is associated with m_i or m_j.

Step 2: Use Equation 3 to calculate the semantic similarity of each pair of diseases.

Step 3: Calculate the functional similarity of m_i and m_j based on the disease semantic similarity. The details are given in the following.

Let D_i and D_j denote the set of diseases associated with miRNAs m_i and m_j, respectively, where $|D_i| = n_i$ and $|D_j| = n_j$. The miRNA functional similarity (MFS) of m_i and m_j is calculated as follows.

$$S(d, D_j) = \max_{d' \in D_j} \{DS(d, d')\}, \text{ for } d \in D_i. \quad (4)$$

$$MFS(m_i, m_j) = \frac{\sum_{d \in D_i} S(d, D_j) + \sum_{d' \in D_j} S(d', D_i)}{n_i + n_j}. \quad (5)$$

2.4 Liu *et al.*'s Transition Matrix of Random Walk with Restart

Random walk with restart (RWR) [5] was originally proposed for the task of image segmentation. In random walk, it is assumed that starting from a certain node, each step is random. If each node is assigned a probability, it can be viewed as the importance weight of the node in the network. RWR is different from random walk since RWR allows the walker to restart from the source node every time at any node with a probability γ. RWR has been used in numerous researches, such as predicting disease-associated miRNAs [9, 18, 27], prioritizing candidate disease genes [16], predicting drug-target interactions [8, 23]. The random walk with restart is defined in Equation 6.

$$P_{t+1} = (1-\gamma)W^T P_t + \gamma P_0. \quad (6)$$

In Equation 6, γ is the restart probability; W is the transition matrix; P_0 is the initial vector; P_t is the vector at time t.

The transition matrix of Liu *et al.* [18] in random walk with restart is described as follows.

$$\boldsymbol{W}_{DD}(i,j) = \begin{cases} D(i,j)/\sum_{z=1}^{n} D(i,z) & \text{if } \sum_{z=1}^{m} B(i,z) = 0, \\ (1-\lambda)D(i,j)/\sum_{z=1}^{n} D(i,z) & \text{otherwise,} \end{cases} \quad (7)$$

$$\boldsymbol{W}_{MM}(i,j) = \begin{cases} M(i,j)/\sum_{z=1}^{m} M(i,z) & \text{if } \sum_{z=1}^{n} B(z,i) = 0, \\ (1-\delta)M(i,j)/\sum_{z=1}^{m} M(i,z) & \text{otherwise,} \end{cases} \quad (8)$$

$$\boldsymbol{W}_{DM}(i,j) = \begin{cases} 0 & \text{if } \sum_{z=1}^{m} B(i,z) = 0, \\ \lambda B(i,j)/\sum_{z=1}^{m} B(i,z) & \text{otherwise,} \end{cases} \quad (9)$$

$$\boldsymbol{W}_{MD}(i,j) = \begin{cases} 0 & \text{if } \sum_{z=1}^{n} B(z,i) = 0, \\ \delta B(j,i)/\sum_{z=1}^{n} B(z,i) & \text{otherwise,} \end{cases} \quad (10)$$

In the above, $\boldsymbol{W}_{DD}$, $\boldsymbol{W}_{MM}$, $\boldsymbol{W}_{DM}$ and $\boldsymbol{W}_{MD}$ are transition matrices, with sizes $n \times n$, $m \times m$, $n \times m$ and $m \times n$, respectively. In $\boldsymbol{W}_{DD}$, each entry represents the probability that a disease is associated with another disease. In $\boldsymbol{W}_{MM}$, each entry represents the probability that a miRNA is associated with another miRNA. In $\boldsymbol{W}_{DM}$, each entry represents the probability that a disease is associated with a miRNA. In $\boldsymbol{W}_{MD}$, each entry represents the probability that a miRNA is associated with a disease. $\boldsymbol{B}$ is an $n \times m$ miRNA-disease association matrix, which contains the known associations.

3 Our Method

For solving the miRNA-disease association prediction problem, our method first calculates the disease semantic similarities and miRNA functional similarities, as described in Sections 2.2 and 2.3. Then, we calculate the target gene similarity between two miRNAs. In addition to the miRNA functional similarity, the target gene similarity is another view of miRNA similarity. The target gene similarity may enhance the performance of our prediction. These three similarity matrices are used as the input to establish the transition matrix of RWR for predicting the potential miRNA-disease associations. The flowchart of our method is shown in Figure 3.

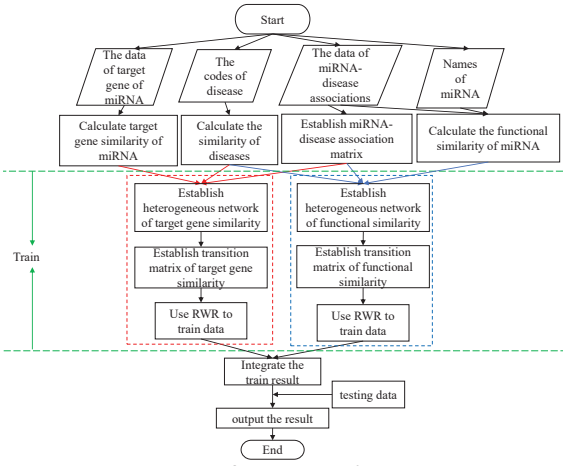

Figure 3: The flowchart of our method.

To calculate the target gene similarity between two miRNAs, we download miRNA-target interaction information from MiRTarBase [11], which contains 806 miRNAs, 15022 genes, 206712 miRNA-target interactions. Equation 11 calculates the target gene similarity of two miRNAs m_i and m_j, where G_i and G_j represent the gene set of m_i and m_j, respectively.

$$TG(m_i, m_j) = \frac{|G_i \cap G_j|}{|G_i \cup G_j|}. \quad (11)$$

3.1 The Transition Matrix of Random Walk with Restart

To use random walk with restart for predicting potential miRNA-disease associations, we have to prepare the transition matrix and P_0. Let m and n denote the numbers of miRNAs and diseases, respectively. The transition matrix is represented as Equation 12, provided by Liu *et al.* [18].

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_{DD} & \boldsymbol{W}_{DM} \\ \boldsymbol{W}_{MD} & \boldsymbol{W}_{MM} \end{bmatrix}. \quad (12)$$

In this paper, we modify the calculation of $\boldsymbol{W}_{DD}$, $\boldsymbol{W}_{MM}$, $\boldsymbol{W}_{DM}$ and $\boldsymbol{W}_{MD}$, as described in Equations 13, 14, 15, and 16.

$$\boldsymbol{W}_{DD}(i,j) =$$

$$\begin{cases} D(i,j)/\sum_{z=1}^{n} D(i,z) & \text{if } D(i,j) \geq \theta_d, \\ (1-\lambda)D(i,j)/\sum_{z=1}^{n} D(i,z) & \text{otherwise,} \end{cases} \tag{13}$$

$$\boldsymbol{W}_{MM}(i,j) =$$

$$\begin{cases} M(i,j)/\sum_{z=1}^{m} M(i,z) & \text{if } M(i,j) \geq \theta_m, \\ (1-\delta)M(i,j)/\sum_{z=1}^{m} M(i,z) & \text{otherwise,} \end{cases} \tag{14}$$

$$\boldsymbol{W}_{DM}(i,j) =$$

$$\begin{cases} 0 & \text{if } \sum_{z=1}^{m} A(i,z) = 0, \\ \lambda A(i,j)/\sum_{z=1}^{m} A(i,z) & \text{otherwise,} \end{cases} \tag{15}$$

$$\boldsymbol{W}_{MD}(i,j) =$$

$$\begin{cases} 0 & \text{if } \sum_{z=1}^{n} A(z,i) = 0, \\ \delta A(j,i)/\sum_{z=1}^{n} A(z,i) & \text{otherwise,} \end{cases} \tag{16}$$

In the above, $\boldsymbol{A}$ is an $n \times m$ miRNA-disease association matrix, which contains the known associations. λ is the probability that the disease similarity network moves to the miRNA similarity network; δ is the probability that the miRNA similarity network move to the disease similarity network; θ_d and θ_m are the threshold values of disease similarity and miRNA similarity, respectively.

To calculate $\boldsymbol{W}_{DD}$ in Equation 13 and $\boldsymbol{W}_{MM}$ in Equation 14, we made modification to the equations provided by Liu *et al.* [18]. We add thresholds θ_d and θ_m into the two equations, respectively. We think if disease semantic similarity or miRNA similarity is higher than the threshold, the transition probability should be higher than the low disease semantic similarity or miRNA similarity. Thus, if the similarity is below the threshold, we decrease the transition probability by λ or δ, respectively. P_0 is the initial probability vector in the RWR, represented as Equation 17 [18].

$$P_0 = \begin{bmatrix} (1-\eta)U_0 \\ \eta V_0 \end{bmatrix}. \tag{17}$$

The two vectors U_0 of $n \times 1$ and V_0 of $m \times 1$ are the initial probabilities of the disease similarity network and the miRNA similarity network, respectively. The parameter η balances the importance of the disease similarity network and the miRNA similarity network.

When a query disease d_i is given to test whether it has associations with a potential miRNA m_j, we first remove the association of d_i and m_j. In other words, we simulate that the association of d_i and m_j is the unknown potential association to be found.

The initial values of U_0 and V_0 are set as follows. $U_0(i) = 1$ if d_i is a query disease; $U_0(i) = 0$ if otherwise. If the known miRNAs associated with d_i are m_a, m_b, m_c and m_d, then the value of each $V_0(a)$, $V_0(b)$, $V_0(c)$ and $V_0(d)$ is set to $1/4 = 0.25$. The sum of these associated miRNAs is 1. When the L_1 norm between P_t and P_{t+1} is less than 10^{-6}, the iteration stops.

3.2 Integration of Two Methods

Our method for the prediction of miRNA-disease association is described as follows.

Step 1: Calculate the disease semantic similarities with Equation 3, the functional similarities of miRNAs with Equation 5, the target gene similarities of miRNAs with Equation 11, and establish miRNA-disease associations. In the establishment of miRNA-disease association, if a miRNA is associated with a disease, we set the corresponding value to 1; 0 if otherwise.

Step 2: Establish the transition matrix of functional similarities of miRNAs with Equation 12, and the transition matrix of target gene similarities of miRNAs with Equation 12, respectively. In our experiments, the threshold values in the transition matrix of functional similarities of miRNAs are set as $\theta_d = \theta_m = \theta_f \in \{0.3, 0.5, 0.6\}$, and the transition matrix of target gene similarities of miRNAs are set as $\theta_d = \theta_m = \theta_g \in \{0.3, 0.4, 0.5\}$. Therefore, we get six transition matrices with different values of θ_f and θ_g.

Step 3: These six transition matrices are used in RWR independently, and we get six prediction results from the six matrices.

Step 4: Set the weight of results obtained from miRNA functional similarities as w_f, and target gene similarities of miRNAs as $1 - w_f$.

Step 5: Merge these six results with their weights and get the association ranks between the

query disease d_i and all miRNAs. If the rank of d_i and the target miRNA m_j falls in the top θ_{md} (such as 20%), and d_i and m_j are verified as a real association, then it is regarded as a true positive case.

4 Experimental Results

4.1 Datasets

Table 1 illustrates the datasets for our experiments. The main experimental dataset was extracted from HMDD [24]. With the gene interaction information in MiRTarBase [11], we can calculate the target gene similarities between miRNAs. With the directed acyclic graphs (DAGs) in MeSH [1], we can calculate the disease semantic similarities between diseases. The counts of times cited for the datasets come from google scholar.

To ensure that each miRNA has both information of target genes and diseases, we choose the miRNAs which simultaneously appear in two datasets HMDD and MiRTarBase, and remove the others. In our experimental dataset, after extraction and removal from HMDD, the number of miRNAs, diseases and miRNA-disease associations are 806, 629 and 13534, respectively. Note that originally, there are 1206 miRNAs, 893 diseases and 35547 miRNA-disease associations in HMDD.

4.2 Leave-One-Out Cross Validation

There are seven parameters in our method, where $0 < \lambda, \delta, \gamma, \eta, \theta_d, \theta_m, w_f < 1$. Here, we always set $\theta_d = \theta_m$. For simplifying the parameter tuning step, the values of these parameters are limited in $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. Then, we use the leave-one-out cross validation (LOOCV) to test most ranges of parameters to find better performance in our experiments. Experimental results show that the better values of these parameters are $\lambda = 0.9$, $\delta = 0.9$, $\gamma = 0.5$, $\eta = 0.5$ and $w_f = 0.5$, $\theta_d = \theta_m = \theta_f \in \{0.3, 0.5, 0.6\}$ for functional similarities and $\theta_d = \theta_m = \theta_g \in \{0.3, 0.4, 0.5\}$ for target gene similarities.

In LOOCV, one of the original samples (known miRNA-disease associations) is picked up as the only testing sample, and the others are used as training samples. The step is repeated until each sample in the original samples is regarded as a testing sample exactly once. In our experiments, for a query disease d_i, we choose a miRNA m_j, which is known to be associated with d_i, as our target miRNA, and the other miRNAs associated with disease d_i are as known information. All miRNAs known to be associated with d_i is chosen as the target miRNA exactly once. If the association rank of d_i and m_j in the prediction is in top θ_{md} (set as 20% here) among all miRNAs, then it is regarded as a true positive result; otherwise, the prediction of d_i and m_j is a false negative result.

For another miRNA $m_{j'}$, which has no association with d_i, if the association rank of d_i and $m_{j'}$ in the prediction is in top θ_{md} among all miRNAs, then it is regarded as a false positive result; otherwise, the prediction of d_i and $m_{j'}$ is a true negative result.

4.3 The Results of General Experiments

We use the *receiver operating characteristic* (ROC) curve [30] and *area under the curve* (AUC) [13] to evaluate the performance of various methods. To calculate the ROC curve, we need two evaluating formulas, true positive rate (TPR) and false positive rate (FPR), given as follows.

$$TPR = \frac{TP}{TP + FN}, \tag{18}$$

$$FPR = \frac{FP}{FP + TN}. \tag{19}$$

In the above, TP (true positive) represents the number of the target miRNAs are correctly predicted to be positive samples; FP (false positive) represents the number of the target miRNAs are incorrectly predicted to be positive samples; TN (true negative) represents the number of the target miRNAs are correctly predicted to be negative samples; FN (false negative) represents the number of the target miRNAs are incorrectly predicted to be negative samples.

Table 2 shows the AUC values of our method and other methods. The result shows that the AUC value of our method is higher than other methods. However, the comparison is not so fair, since the sizes of datasets used in different methods are different. The numbers of miRNAs and diseases in our experiment are greater than other methods, which may be more statistically meaningful.

In order to compare the prediction performances of our method and Liu *et al.*'s method [18] with the same dataset (our dataset), we try to implement their method by ourselves, since we do

Table 1: The datasets used for experiments.

Year	Dataset(s)	Description	Times Cited
2018	MiRTarBase [11]	A dataset that experimentally verified miRNA-target interaction, accumulating more than 360,000 miRNA-target interactions.	729
2019	HMDD v3.2 [14]	A web server that contains 35547 miRNA-disease associations, including 1206 miRNAs, and 893 diseases from 19280 papers.	618
2019	MeSH [1]	A web page that used for indexing, cataloging, and searching of biomedical and health-related information. among them biomedical tree structure, including 59746 indexes.	578

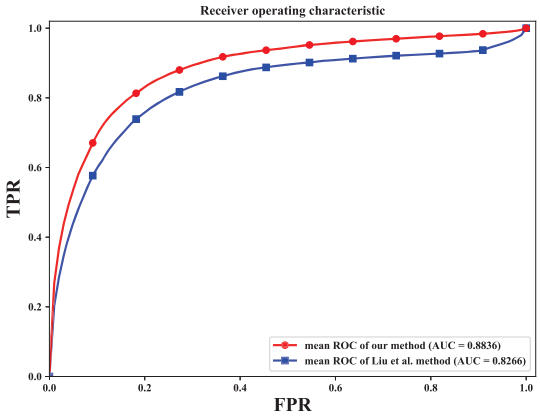

Figure 4: The ROC curves and AUC values of our method and the method of Liu *et al.* with our dataset.

not have the source code of Liu *et al.*'s method [18]. As shown in the last two rows of Table 2, by using the same dataset, the AUC of our method is 0.8836, which is better that 0.8266 of Liu *et al.*'s method.

Figure 4 shows the ROC curves and AUC values of our method and Liu *et al.*'s method, with the same dataset (our dataset).

4.4 The Results of 15 Diseases

In the study of Liu *et al.* [18], they chose 15 diseases for further observing the prediction performance. To fairly compare the prediction performances of our method and Liu *et al.*'s method, we also conduct the same experiment with the same 15 diseases and the same dataset. Note that Liu *et al.*'s method is implemented by ourselves.

At each time, we try to find the miRNAs associated with one of the 15 diseases. The associations of miRNAs with the one queried disease are all removed. That is, it is to simulate that these associations are unknown. The prediction results of our method with various values of the threshold θ_{md} are shown in Table 3. In these 15 diseases, most of the miRNAs can be found within the top 30% of the association ranks.

The prediction results of Liu *et al.*'s method with these 15 diseases are shown in Table 4. By comparing these two tables, our method finds more numbers of associated miRNAs in most of diseases, such as 'breast neoplasms', 'colorectal neoplasms', 'carcinoma, hepatocellular', 'lung neoplasms', 'melanoma', 'ovarian neoplasms', 'pancreatic neoplasms', and 'stomach neoplasms'. It is clear that our method can predict the potential associations of miRNAs and diseases with higher accuracies.

Figure 5 shows the visual comparison of Tables 3 and 4. We can see that when $\theta_{md} \leq 30\%$, our method finds more numbers of miRNAs than Liu *et al.*'s method. That one method can find more numbers of miRNAs under same θ_{md} indicates its superior performance.

5 Conclusion

To improve the performance of the miRNA-disease association prediction, we add the target gene similarity of miRNA, and modify the transition matrix of RWR. We use LOOCV to evaluate the performance of our method. In our experiments, the AUC value of our method is 0.8836, which is better than the previously pub-

Table 2: The AUC values of our method and other methods, with various datasets of various sizes.

Year	Method or authors	# of miRNAs	# of diseases	AUC
2010	Jiang *et al.* [15]	120	53	0.7580
2017	Liu *et al.* [18]	476	341	0.8049
2018	MiRGOFS [28]	267	137	0.8770
2019	Chen *et al.* [7]	404	362	0.8460
2017	Liu *et al.* [18] with our dataset	**806**	**629**	0.8266
2021	Our method	**806**	**629**	**0.8836**

Table 3: The numbers of associated miRNAs found by our method for the 15 diseases with various threshold θ_{md}. In each entry, the percentage represents the ratio of found miRNAs, and the number within parentheses denotes the number of found miRNAs.

Diseases	# of miRNAs	θ_{md}				
		10%	20%	30%	40%	50%
leukemia, myeloid, acute	130	42%(55)	70%(91)	88%(115)	93%(121)	96%(125)
breast neoplasms	382	21%(81)	42%(159)	59%(226)	71%(273)	82%(313)
colorectal neoplasms	295	27%(80)	49%(145)	68%(202)	79%(234)	86%(255)
glioblastoma	196	31%(60)	59%(116)	77%(151)	83%(162)	88%(173)
heart failure	28	68%(19)	86%(24)	93%(26)	96%(27)	96%(27)
carcinoma, hepatocellular	392	20%(80)	41%(159)	59%(231)	73%(285)	83%(325)
lung neoplasms	264	29%(76)	54%(142)	72%(189)	84%(221)	90%(237)
melanoma	213	34%(73)	60%(128)	76%(161)	83%(177)	86%(184)
ovarian neoplasms	216	34%(74)	63%(135)	78%(169)	89%(192)	94%(204)
pancreatic neoplasms	181	41%(75)	67%(121)	82%(148)	89%(161)	93%(169)
prostatic neoplasms	274	28%(76)	49%(135)	65%(177)	76%(209)	82%(224)
carcinoma, renal cell	163	36%(59)	64%(104)	80%(130)	91%(149)	93%(152)
carcinoma, squamous cell	26	46%(12)	81%(21)	96%(25)	96%(25)	96%(25)
stomach neoplasms	314	25%(79)	47%(147)	67%(209)	80%(251)	88%(277)
urinary bladder neoplasms	203	35%(71)	64%(130)	75%(153)	84%(171)	89%(180)
Average		35%	60%	76%	85%	90%

Table 4: The numbers of associated miRNAs found by Liu *et al.* for the 15 diseases with various threshold θ_{md}. In each entry, the percentage represents the ratio of found miRNAs, and the number within parentheses denotes the number of found miRNAs.

Diseases	# of miRNAs	θ_{md}				
		10%	20%	30%	40%	50%
leukemia, myeloid, acute	130	38%(49)	68%(89)	80%(104)	88%(115)	95%(124)
breast neoplasms	382	20%(78)	39%(150)	55%(211)	69%(262)	78%(298)
colorectal neoplasms	295	26%(76)	48%(141)	64%(190)	77%(226)	85%(250)
glioblastoma	196	30%(59)	53%(104)	72%(142)	85%(166)	90%(177)
heart failure	28	68%(19)	82%(23)	93%(26)	93%(26)	96%(27)
carcinoma, hepatocellular	392	20%(78)	39%(154)	57%(224)	69%(269)	78%(306)
lung neoplasms	264	27%(72)	52%(136)	67%(176)	78%(205)	86%(226)
melanoma	213	30%(64)	54%(115)	68%(144)	79%(169)	85%(180)
ovarian neoplasms	216	33%(72)	61%(131)	74%(160)	86%(185)	91%(197)
pancreatic neoplasms	181	40%(72)	64%(116)	77%(139)	88%(160)	91%(165)
prostatic neoplasms	274	27%(75)	47%(129)	62%(169)	73%(199)	82%(224)
carcinoma, renal cell	163	36%(58)	62%(101)	77%(125)	88%(143)	93%(151)
carcinoma, squamous cell	26	46%(12)	73%(19)	92%(24)	96%(25)	96%(25)
stomach neoplasms	314	25%(78)	46%(144)	61%(192)	75%(235)	84%(264)
urinary bladder neoplasms	203	35%(71)	59%(119)	73%(148)	82%(166)	89%(180)
Average		33%	57%	72%	82%	88%

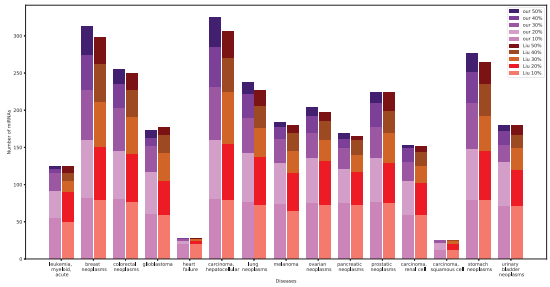

Figure 5: The numbers of associated miRNAs for the 15 diseases predicted by our method and the method of Liu *et al.* Here, each percentage represents the value of θ_{md}.

lished methods. We also find that our method has acceptable performance even if all associations of specific diseases are removed.

It may be interesting to investigate other similarity measuring methods for raising the prediction performance. So far, we use statistical methods to solve the problem. We may use biological features to enhance the measuring methods, such as miRNA families and clusters, sequence similarity, deep sequencing, disease gene, and disease phenotype, etc. In the future, the miRNA-disease association prediction methods may involve machine learning, convolutional neural network, etc.

References

[1] "Medical subject headings." `https://meshb.nlm.nih.gov/`, 2019.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, Vol. 25, No. 1, pp. 25–29, May 2000.

[3] J. Brennecke, D. R. Hipfner, A. Stark, R. B. Russell, and S. M. Cohen, "bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in drosophila," *Cell*, Vol. 113, No. 1, pp. 25–36, Apr. 2003.

[4] G. A. Calin and C. M. Croce, "MicroRNA signatures in human cancers," *Nature Reviews Cancer*, Vol. 6, No. 11, pp. 857–866, Nov. 2006.

[5] T. Can, O. Çamoğlu, and A. K. Singh, "Analysis of protein-protein interaction networks using random walks," *Proceedings of the 5th international workshop on Bioinformatics*, pp. 61–68, Aug. 2005.

[6] C. Z. Chen, L. Li, H. F. Lodish, and D. P. Bartel, "MicroRNAs modulate hematopoietic lineage differentiation," *Science*, Vol. 303, No. 5654, pp. 83–86, Jan. 2004.

[7] H. Chen, Z. Zhang, and D. Feng, "Prediction and interpretation of miRNA-disease associations based on miRNA target genes using canonical correlation analysis," *BMC Bioinformatics*, Vol. 20, No. 404, Dec. 2019.

[8] X. Chen, M. X. Liu, and G. Y. Yan, "Drug–target interaction prediction by random walk on the heterogeneous network," *Molecular BioSystems*, Vol. 8, pp. 1970–1978, Mar. 2012.

[9] X. Chen, M. X. Liu, and G. Y. Yan, "RWRMDA: predicting novel human microRNA-disease associations," *Molecular BioSystems*, Vol. 8, pp. 2792–2798, July 2012.

[10] A. M. Cheng, M. W. Byrom, J. Shelton, and L. P. Ford, "Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis," *Nucleic Acids Research*, Vol. 33, No. 4, pp. 1290–1297, Jan. 2005.

[11] C. H. Chou, S. Shrestha, C. D. Yang, N. W. Chang, Y. L. Lin, K. W. Liao, W. C. Huang, T. H. Sun, S. J. Tu, W. H. Lee, M. Y. Chiew, C. S. Tai, T. Y. Wei, T. R. Tsai, H. T. Huang, C. Y. Wang, H. Y. Wu, S. Y. Ho, P. R. Chen, C. H. Chuang, P. J. Hsieh, Y. S. Wu, W. L. Chen, M. J. Li, Y. C. Wu, X. Y. Huang, F. L. Ng, W. Buddhakosai, P. C. Huang, K. C. Lan, C. Y. Huang, S. L. Weng, Y. N. Cheng, C. Liang, W. L. Hsu, and H. D. Huang, "miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions," *Nucleic Acids Research*, Vol. 46, No. D1, pp. D296–D302, Jan. 2018.

[12] T. L. Cuellar and M. T. McManus, "MicroRNAs and endocrine biology," *Journal of Endocrinology*, Vol. 187, No. 3, pp. 327–332, Dec. 2005.

[13] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, Vol. 7, pp. 1–30, Dec. 2006.

[14] Z. Huang, J. Shi, Y. Gao, C. Cui, S. Zhang, J. Li, Y. Zhou, and Q. Cui, "HMDD v3.0: a database for experimentally supported hu-

man microRNA-disease associations," *Nucleic Acids Research*, Vol. 47, No. D1, pp. D1013–D1017, Jan. 2019.

[15] Q. Jiang, Y. Hao, G. Wang, L. Juan, T. Zhang, M. Teng, Y. Liu, and Y. Wang, "Prioritization of disease microRNAs through a human phenome-microRNAome network," *BMC Systems Biology volume*, Vol. 4, May 2010.

[16] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *The American Journal of Human Genetics*, Vol. 82, No. 4, pp. 949–958, Apr. 2008.

[17] J. Li, S. Zhang, Y. Wan, Y. Zhao, J. Shi, Y. Zhou, and Q. Cui, "MISIM v2.0: a web server for inferring microRNA functional similarity based on microRNA-disease associations," *Nucleic Acids Research*, Vol. 47, No. W1, pp. W536–W541, July 2019.

[18] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 14, No. 4, pp. 905–915, July 2017.

[19] E. A. Miska, "How microRNAs control cell division, differentiation and death," *Current Opinion in Genetics and Development*, Vol. 15, No. 5, pp. 563–568, Oct. 2005.

[20] M. N. Poy, L. Eliasson, J. Krutzfeldt, S. Kuwajima, X. Ma, P. E. MacDonald, S. Pfeffer, T. Tuschl, N. Rajewsky, P. Rorsman, and M. Stoffel, "A pancreatic islet-specific microRNA regulates insulin secretion," *Nature*, Vol. 432, No. 7014, pp. 226–230, Nov. 2004.

[21] M. L. Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl, "Identification of novel genes coding for small expressed RNAs," *Science*, Vol. 294, No. 5543, pp. 853–858, Oct. 2001.

[22] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun, "The 21-nucleotide let-7 RNA regulates developmental timing in caenorhabditis elegans," *Nature*, Vol. 403, No. 6772, pp. 901–906, Feb. 2000.

[23] A. Seal, Y. Y. Ahn, and D. J. Wild, "Optimizing drug–target interaction prediction based on random walk on heterogeneous networks," *Journal of Cheminformatics*, Vol. 7, No. 40, Aug. 2015.

[24] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, Vol. 26, No. 13, pp. 1644–1650, May 2010.

[25] B. R. Wilfred, W. X. Wang, and P. T. Nelson, "Energizing miRNA research: A review of the role of miRNAs in lipid metabolism, with a prediction that miR-103/107 regulates human metabolic pathways," *Molecular Genetics and Metabolism*, Vol. 91, No. 3, pp. 209–217, July 2007.

[26] P. Xuan, K. Han, M. Guo, Y. Guo, J. Li, J. Ding, Y. Liu, Q. Dai, J. Li, Z. Teng, and Y. Huang, "Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors," *PLOS ONE*, Vol. 8, No. 8, Aug. 2013.

[27] P. Xuan, K. Han, Y. Guo, J. Li, X. Li, Y. Zhong, Z. Zhang, and J. Ding, "Prediction of potential disease-associated microRNAs based on random walk," *Bioinformatics*, Vol. 31, No. 11, pp. 1805–1815, June 2015.

[28] Y. Yang, X. Fu, W. Qu, Y. Xiao, and H. B. Shen, "MiRGOFS: a GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA-disease association," *Bioinformatics*, Vol. 34, No. 20, pp. 3547–3556, Oct. 2018.

[29] Y. Zhong, P. Xuan, X. Wang, T. Zhang, J. Li, Y. Liu, and W. Zhang, "A non-negative matrix factorization based method for predicting disease-associated miRNAs in miRNA-disease bilayer network," *Bioinformatics*, Vol. 34, No. 2, pp. 267–277, Jan. 2018.

[30] K. H. Zou, A. J. O'Malley, and L. Mauri, "Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models," *Circulation*, Vol. 115, No. 5, pp. 654–657, Feb. 2007.