Research Paper

Computational prediction of human disease-related microRNAs by path-based random walk

Israel Mugunga[1], **Ying Ju**[1], **Xiangrong Liu**[1] **and Xiaoyang Huang**[1]

[1]Department of Computer Science, Xiamen University, Xiamen, 361005, China

Correspondence to: *Xiaoyang Huang,* ***email:*** *xyhuang@xmu.edu.cn*

Keywords: *systems biology, microRNA, disease-related microRNA prediction, path-based random walk*

Received: *February 08, 2017* ***Accepted:*** *March 22, 2017* ***Published:*** *April 19, 2017*

ABSTRACT

MicroRNAs (miRNAs) are a class of small, endogenous RNAs that are 21–25 nucleotides in length. In animals and plants, miRNAs target specific genes for degradation or translation repression. Discovering disease-related miRNA is fundamental for understanding the pathogenesis of diseases. The association between miRNA and a disease is mainly determined via biological investigation, which is complicated by increased biological information due to big data from different databases. Researchers have utilized different computational methods to harmonize experimental approaches to discover miRNA that articulates restrictively in specific environmental situations. In this work, we present a prediction model that is based on the theory of path features and random walk to obtain a relevancy score of miRNA-related disease. In this model, highly ranked scores are potential miRNA-disease associations. Features were extracted from positive and negative samples of miRNA-disease association. Then, we compared our method with other presented models using the five-fold cross-validation method, which obtained an area under the receiver operating characteristic curve of 88.6%. This indicated that our method has a better performance compared to previous methods and will help future biological investigations.

INTRODUCTION

MicroRNAs (miRNAs) are a class of small, endogenous RNAs that are 21–25 nucleotides long. MiRNAs are found in plants, animals, and some viruses, and function in RNA silencing and post-transcriptional regulation of gene expression [1, 2]. MiRNAs are involved in many diverse biological processes, such as development, differentiation, apoptosis, and viral infection [3]. Increasing evidence implicates miRNAs in human disease development, progression, prognosis, diagnosis, and evaluation of treatment response [4–6]. Since the first miRNA (lin-4) was discovered from C. *elegans* 20 years ago, many miRNAs have been annotated in various species with experimental and computational methods [7]. However, the discovery of disease-related miRNA via existing biological experimental methods is expensive and time-consuming [8]. Thus, computational prediction methods are significant techniques for identifying the most promising miRNA-disease associations prior to additional experimental examinations. Many databases have been developed to accumulate miRNA data. The Human MiRNA Disease Database (HMDD) [9] is an online database that provides complete information on miRNA deregulation in various human diseases. MiRbase [10] is a comprehensive miRNA database that contains the sequences of precursor miRNAs, mature miRNAs, miRNA hairpin structure, and miRNA targets. OncomiRDB is a manually curated miRNA-cancer association database that contains more than 300 miRNAs and 2259 miRNA-cancer associations [11]. dbDEMC [12] is an integrated database that is designed to store and display differentially expressed miRNAs in human cancers. MiR2disease [13] is a manually curated database that provides a comprehensive resource from miRNA deregulation in various human diseases. The identification of miRNAs that underlie human diseases is an important goal of biomedical researchers. However, one major issue in miRNA studies is the lack of enough bioinformatics methods that predict potential miRNA-disease associations.

Computational methods for predicting miRNA and related diseases have been proposed to overcome this major issue. Computational predictions indicate that miRNAs, which account for at least 1% of human protein-coding genes, regulate protein production for thousands of human genes. Most computational predictive methods for miRNA-disease associations are based on the hypothesis that miRNAs with similar functions tend to associate with a common disease and that diseases with shared similar phenotypes likely share common miRNA [14–16]. Therefore, different methods have been proposed to predict miRNA and disease association such as random walk-based methods. In 2008, Xu et al. [17] proposed miRank, a ranking algorithm based on random walk. They tested their method on *Homo sapiens* genomes and achieved a good accuracy. In 2012, Chen et al. [18] adopted a global network similarity measure and developed Random Walk with Restart for MiRNA-disease Association (RWRMDA) to infer potential miRNA-disease interactions by implementing random walk on a miRNA–miRNA functional similarity network. Prediction of disease-related miRNA has proposed in various ways. A robust regularization path for *v*-support vector classification based on lower upper decomposition with pivoting is a method present in [19]. This method uses a regularization parameter *v* as the assistance of adjusting the number of support vectors and margin errors.

In 2014, Liu et al. [20] developed a random walk with restart method. Using this model, they first revealed the limitations of previous computational methods, and then implemented random walk with restart on a heterogeneous network to infer potential miRNA-disease associations. To declare these limitations, the first limitation has stated as the use of single dataset, second limitation as an inadequacy of disease semantic similarity and the third limitation as an overestimation of the predictive accuracy. They applied their method to diseases with no known related miRNAs. The majority of the top 30 candidates were confirmed by various databases.

In addition, many researchers have proposed different methods of predicting miRNA-related diseases. In 2009, Jiang et al. [21] proposed the miRNA-disease association prediction technique, wherein miRNA–miRNA and human phenome–miRNA functional similarity networks were constructed. The score for all miRNAs was computed using cumulative hypergeometric distribution [21]. Chen et al. [22] anticipated HGIMDA method as a graph inference by integrating miRNAs similarity, diseases similarity and Gaussian interaction profile kernel similarity into a heterogeneous graph. To analyze, predict, and providing possible lncRNA-disease associations is another problem stated by Chen et al. [23]. Improved random walk with restart to predict lncRNA with related disease method has been proposed to improve the traditional limitations of random walk with restart method. It also based on an integration of data similarities such as diseases and lncRNA, and applying a random walk to predict novel lncRNA-disease.

In 2015, Xuan et al. [24] proposed MIDP, a predictive model for disease-related miRNA. In this method, the prediction process is modeled as random walk on a miRNA network that is derived from miRNA-associated diseases. For a specific disease with some related miRNAs, a random walker starts at a known related miRNA node with equal probability. An extension method, MIDPE, was specifically proposed for diseases without any known related miRNAs. This model obtained a higher predicted accuracy [24]. In addition to random walk-based methods, many other methods have been developed to support biological examinations.

Xuan et al. [25] presented HDMP, a method that is based on weighed *k* and most similar neighbors, for miRNA-disease association prediction. The functional similarity of their method combined the similarity of information content of disease terms and phenotype similarity between diseases. Based on the hypothesis that miRNAs with similar functions tend to associate with a common disease and that diseases with shared similar phenotypes likely share common miRNA [14–16], Chen et al. [26] developed WBSMDA to predict relationship among miRNAs and diseases. By combining data from disease-related miRNAs, similarity miRNAs, disease similarity and similarity of Gaussian interaction profile kernel, a score has been calculated to predict new miRNA-disease associations. The model evaluation has been conducted based on Leave-one-out cross validation method to verify its performance. However, this method attained a better performance, it has a limitation; such that it was developed based on the hypothesis in [14–16], which may cause unfairness to miRNAs which have a lot number of related diseases.

Chen et al. [27] used three network-based similarities (miRNA-based similarity inference, phenotype-based similarity inference and network-consistency-based inference) to infer potential miRNA-disease associations. Another model proposed based on matrix completion algorithm has been proposed to predict miRNA-disease associations based on the known diseases and miRNAs. Li et al. [28], proposed MCMDA model which based on updating the adjacency matrix feature for known related miRNAs and disease.

Zou et al. [29] presented Pretata, a predictive method to identify TATA-binding proteins. This model is an applied advanced machine learning methods which facilitate to identify TBP (TATA-binding proteins) and protein detection from primary sequences.

Le [30] constructed a miRNA function network based on common targets and applied four network-based methods to infer novel disease-related miRNAs. Mørk et al. [31] presented miRPD to infer miRNA-disease associations by coupling known and predicted miRNA–protein associations with protein–disease associations. Zeng et al. [32, 33] suggested two predictive methods on multipath to predict miRNA-disease associations based on heterogeneous network of diseases and miRNAs. To calculate the similarity between objects, HeteSim_Multipath (HSMP) method was proposed to combine scores form heterogeneous similarity paths with constant. The second method, HeteSim_SVM (HSSVM) uses machine learning method to combine heterogeneous similarity to measure scores without a constant.

Zou et al. [34] proposed KATZ and CATAPULT, which are based on social network analysis, to infer potential miRNA-disease associations. Gu et al. [35] developed a method based on the *v*-support vector regression was proposed to overcome the challenge that presented by the method of Schölkopf et al. [36]. The initial adjustments proposed in this work as an additional term to overcome this challenge where the parameter *v* has the role of monitoring the support vectors based on Karush–Kuhn–Tucker (KKT) conditions to make an initial solution for the incremental learning. Chen et al. [37] introduced a research review about lncRNA and disease associations. In this review, Chen et al., first introduced functions of lncRNA, five important of lncRNA-disease associations, precarious disease-related lncRNA and finally available important lncRNA-database related to sequence, expression, and function. Therefore, a state-of-the-art computational method has been proposed to identify disease and lncRNA.

Chen et al. [38] developed a machine-learning method, Regularized Least Square for MiRNA-disease Association (RLSMDA), to discover the potential relationships between diseases and miRNAs. Based on the hypothesis that lncRNAs with similar functions tend to be related with diseases that have similar functions, Chen et al. [39] developed FMLNCSIM, a model that computing a functionally similarity between lncRNA on a large scale by combining information content and fuzzy measure' concept to the directed acyclic graphs disease. Zeng et al. [40] first presented three general problems in the prediction of miRNA-disease associations. The problems were styled as follows; the lack of similarity among miRNAs stated to be the first challenge in their research, and the second problem defined as inadequate miRNA-disease associations and lastly, the third problem distinct as not enough availability number of negative samples for investigating miRNA-disease relationship. Therefore, they proposed a matrix completion method which is based on integrating multiple feature sets to conquer the challenges [40].

Besides, another computational model to infer miRNA-disease association type pairs has been developed by Chen et al. [41]. RBMMMDA developed based on fact that all the previous computational techniques may only predict binary associations among miRNAs and diseases. Therefore, this method has been develop to conquer this challenge and can predict four different multiple types of miRNA-disease relationship.

Based on the hypothesis that functionally related microbes share the similar interaction and non-interaction pattern diseases, Chen et al [42] developed a model of KATZ measure for Human Microbe–Disease Association prediction (KATZHMDA). Therefore, this model was constructed based on a Gaussian interaction profile kernel similarity from disease and related microbe. To state this method thus, the similarity network derived from microbe-disease for Gaussian interaction profile kernel was contained by two main steps. Step one is that, every interaction profile of microbe indicated as binary vector which converts the existence or lack of the relationship among disease and microbe. In step two, based on interaction profiles, the Gaussian interaction profile kernel among a pair of microbe was computed. MiRNAs involve in various biological processes. Fungal infection is one of the infections that developed through miRNAs and predicting fungus-causing diseases became also attractive. Therefore, Chen et al. [43] proposed a computational model NLLSS, a semi-supervised learning technique used to combine different kind of information to predict potential synergistic drug combinations.

One the hand, we cannot ignore classification in this research; it helps in designing classifiers in real-world problems. By using two finite mixture models to capture the structural information from different classes, Gu et al. [44] presented a structural minimax probability machine (SMPM) to solve a sequence of SOCP (second order cone programming) problems through a binary search procedure. The method has further proposed a nonlinear SMPM model based on linear SMPM by using kernelization techniques. Predicting disease-related miRNA has been reported in different ways.

Although the proposed methods have successfully predicted miRNA-disease associations, these methods have some drawbacks. The most commonly reported drawback is the poor dataset quality of some methods, which causes poor performance. Some methods evaluate disease–miRNA similarity, miRNA similarity, and disease similarity by using disease data only, which possibly results cause bias that disregards miRNA features.

Therefore, we propose a computational prediction method for potential miRNA-disease associations. Our proposed model is based on path-based features and random walk. MiRNA functional similarity network [45], disease similarity network [46], and known miRNA-disease associations [9] were integrated in our work which ranked every miRNA of a given disease. High prediction ranks were predicted to have high probabilities as potential candidates of a given disease. Experimental results demonstrated that our method effectively predicted potential miRNA-related disease candidates. Known miRNA-disease association databases were applied to evaluate the performance of our proposed method.

RESULTS

Evaluation of prediction performance

We present a computational prediction method that is based on random walk and graph theory. We introduced the five-fold cross validation method to evaluate the strength of our method based on random walk and the construction of RDnet (miRNA and disease network). The datasets were randomly divided into *n* subsets, where *n–1* subsets were used for model construction to predict potential candidates. The remaining subset was used to test the performance of our model. A receiver operating characteristic (ROC) curve was applied to evaluate the strength of our proposed method. A ROC is designed by varying the achieved threshold. The numeric representation of ROC is the area under the curve (AUC). The ROC is shown in Figure 1. The horizontal axis in the plot represents false positive rate (1–specificity) and the vertical axis represents true positive rate (sensitivity). The true positive rate (TPR) refers to the percentage of true associations with scores that are higher than the given threshold. On the other hand, the false positive rate (FPR) is the ratio of the successfully predicted miRNA-disease associations to all known miRNA-disease associations. FPR refers to the percentage of associations with scores that are lower than the given threshold. Mathematical formulas of TPR and FPR are given below:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{1}$$

$$\mathrm{FPR} = \frac{FP}{\mathrm{TN}+\mathrm{FP}} \tag{2}$$

where true positive (TP) denotes the number of known associations with scores are higher than the given threshold. By contrast, false negative (FN) denotes the number of known associations with scores that are lower than the given threshold. True negative (TN) denotes the number of unknown associations with scores that are lower than the given threshold. False positive (FP) denotes the number of unknown associations with scores that are higher than the given threshold.

Comparison with other prediction methods

Models by Xuan et al. [24] (MIDP and MIDPE, 2015), Chen and Yan [38] (RLSMDA, 2014), Chen and Zhang [27] (Chen's method, 2013), Shi et al [47] (Shi's method, 2013), Xuan et al. [25] (HDMP, 2013) and Chen

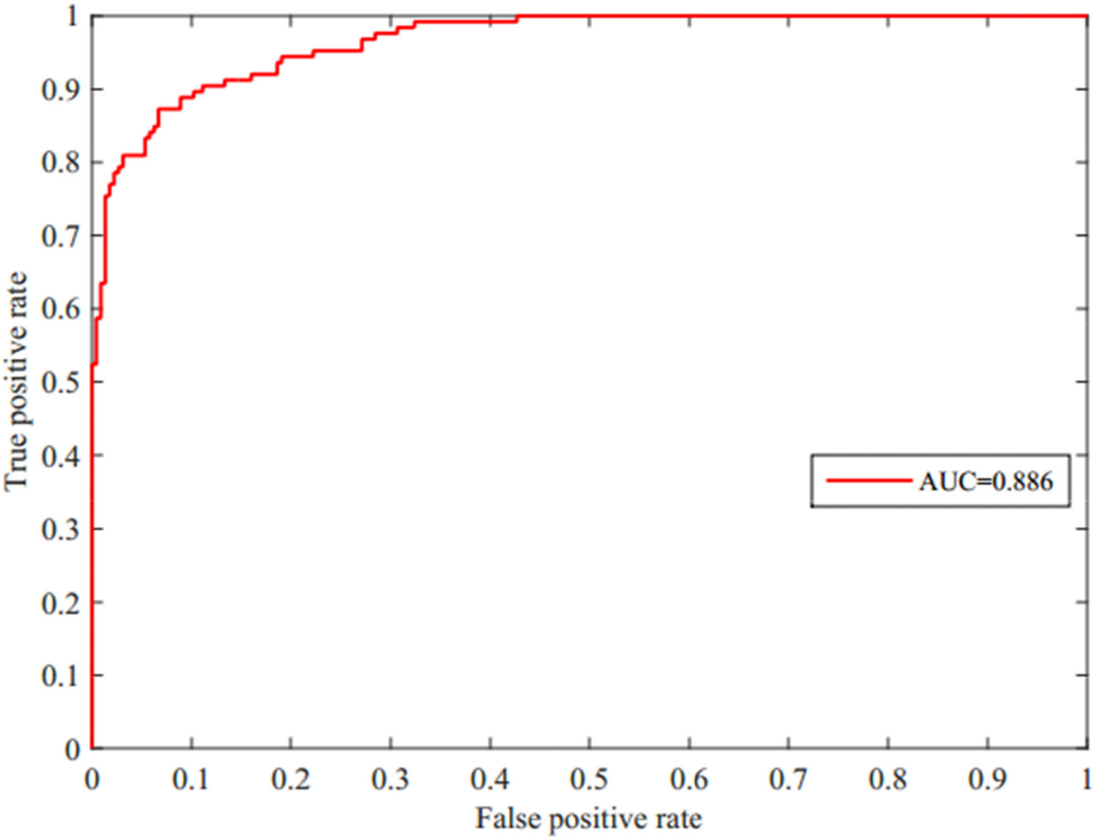

Figure 1: ROC curve and AUC=0.886 value of our predictive model for miRNA-disease associations by five-fold cross validation.

Table 1: Prediction results of our method and other methods for 11 diseases with more than 100 related miRNAs in terms of accuracy (%) using five-fold cross validation

Diseases name	Our method	MIDP	RWRMDA	HDMP	Chen's method
Breast Neoplasms	**0.865**	0.854	0.785	0.801	0.653
Colorectal Neoplasms	**0.879**	0.845	0.793	0.802	0.662
Glioblastoma	**0.821**	0.786	0.68	0.70	0.607
Heart failure	**0.843**	0.821	0.722	0.77	0.761
Hepatocellular Carcinoma	**0.832**	0.807	0.749	0.759	0.613
Lung Neoplasms	**0.904**	0.876	0.876	0.835	0.606
Melanoma	**0.869**	0.837	0.784	0.79	0.642
Ovarian Neoplasms	**0.947**	0.923	0.882	0.884	0.644
Pancreatic Neoplasms	**0.968**	0.945	0.871	0.895	0.684
Prostatic Neoplasms	**0.898**	0.882	0.823	0.854	0.629
Stomach Neoplasms	**0.854**	0.821	0.779	0.787	0.628

From the above, we compared the accuracies of our method with MIDP [24], RWRMDA [18], HDMP [25], and Chen's method [27] for 11 diseases with more than 100 related miRNAs. The comparative analytical results of our method are presented in bold numbers.

Table 2: Different parameters used in the prediction of miRNA-related disease

Methods	Our method	MIDP	RWRMDA	HDMP	Chen's method
Range Parameter	0.1-0.9 (**M**)	0.1-0.9($\mathbf{r}_Q$,$\mathbf{r}_U$)	0.1-0.9 (**r**)	1-50 (**k**)	0.1-0.9 (**r**)

et al. [18] (RWRMDA, 2012) comprise the majority of predictive computational methods that have been presented in previous years. MIDP, RWRMDA, and HDMP have exhibited considerably superior performances, but were developed based on the association data from earlier versions of HMDD [9]. Although we cannot ignore the performances of these methods, our proposed method is compared with MIDP as the only method that was implemented based on association data from the latest database [9] in 2013. MIDP [24] had a good predictable performance, extending to miRNA with no known related disease. However, the performances of these methods have not ignored. The comparison of our method with these methods has shown in Table 1 which confirms a higher predictive accuracy of our method compared with MIDP [24], RWRMDA [18], HDMP [25], and Chen's method [27].

In the analysis of all methods our method, MIDP [24], RWRMDA [18], HDMP [25], and Chen's method [27] held different parameters for tuning as follow: for our computational prediction method, the parameter M was set in the range of 0.1 to 0.9. Our method achieved a better performance when M=0.7. The highest scores from

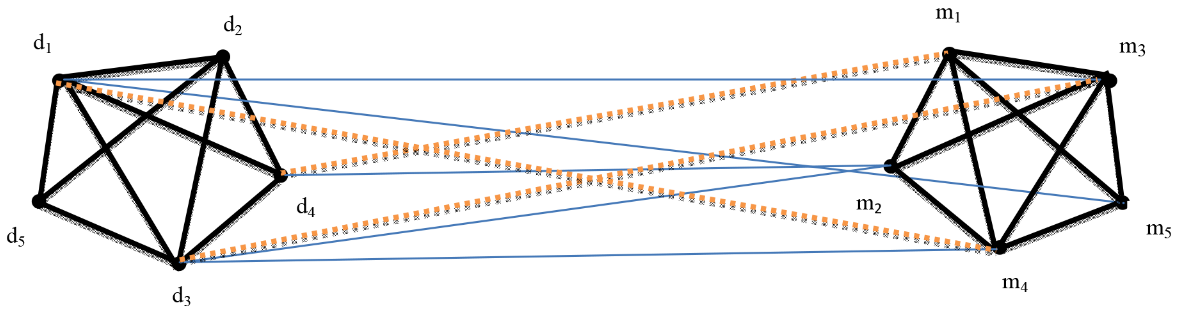

Figure 2: Illustration of the proposed method based on random walk and graph theory derived from RDnet.

Table 3: The top 10 highest scores miRNAs potential candidates related to Hepatocellular carcinoma as confirmed by public databases

MiRNAs	Confirmation	Ranks
hsa-mir-507	dbDEMC2.0	1
hsa-mir-30e	dbDEMC2.0	2
hsa-mir-9-2	dbDEMC2.0;Mir2desease	3
hsa-mir-520f	dbDEMC2.0	4
hsa-mir-132	dbDEMC2.0	5
hsa-mir-424	dbDEMC2.0	6
hsa-mir-431	dbDEMC2.0	7
hsa-mir-34b	dbDEMC2.0	8
hsa-mir-149	dbDEMC2.0	9
hsa-mir-185	dbDEMC2.0	10

a given disease with miRNAs were confirmed as potential candidate miRNA-related diseases. The parameters r_Q and r_U of MIDP [24] were selected from (0.1,0.2,…,0.9) such that r_Q was greater than r_U. The parameter *r* of RWRMDA [18] ranged from 0.1 to 0.9. The parameter *k* of HDMP [25] diverged from 1 to 50. The parameter *r* for Chen's method [27] varied from 0.1 to 0.9. The summary of these parameters is shown in Table 2 below.

DISCUSSION

We present a computational prediction method for ranking all unknown miRNAs that are related to a given disease. The high accuracy of our method has shown that supporting biological tests can reliably predict miRNA and related diseases. Comparing our method with other prediction models proved that its predictive capabilities. We first constructed a miRNA-disease network(RDnet) derived from [9]. Then, we built our model based on random walk and graph theory where the walker starts walking form a known miRNA-related disease with equal probability. From known associations, the walker's goal was to reach the neighbor of a targeted disease, which is the unknown association to calculate the similarity score as the relationship. The weights were set as similarity scores between unknown miRNA and a given disease. Thus, the highest scores were set as potential candidates of miRNA-disease associations, as proven by public databases. The most challenging problem is that the majority of miRNAs have no known related diseases. Moreover, many diseases also have a small amount of related miRNAs. This means that a high number of false negatives exists massively challenging this research field. The construction of consistent prediction models with better accuracies will help to overcome these challenges. Furthermore, using different features to discover the similarities between miRNA-disease associations will be useful in future studies.

MATERIALS AND METHODS

We combined different data set to overcome the results of poor performances. The datasets that were utilized in this study were downloaded from the following databases as follows.

Disease phenotype similarity data

We downloaded the disease phenotype similarity scores from the MimMiner which was developed by Van Driel et al. [46]. They computed a phenotype similarity score per phenotype pair via text-mining analysis of their phenotype descriptions in the Online Mendelian Inheritance in Man database [48]. The phenotypic similarity scores were successfully used to predict and prioritize disease-related protein-coding genes [49, 50].

The human miRNA-disease association data

We downloaded experimentally verified miRNA-disease association in [9]. HMDDv2 is a database that contains experimentally supported data for miRNA-disease associations. These data are manually curated from publications. In our experiment, we used 578 miRNAs and 382 diseases from the 2013 released of HMDD, which had 10,381 entries [9].

MiRNA–miRNA functional similarity data

The miRNA–miRNA functional similarity scores were downloaded from http://www.cuilab.cn/files/images/cuilab/misim.zip [45]. In this dataset, a functional similarity score per miRNA pair is calculated based on the observation that genes with similar function are often associated with similar diseases [14–16].

Construction of RDnet (miRNA and disease network)

The key to network-based predictions of potentially disease-related miRNA is the calculation of similarity network among disease and miRNA over networks. The construction of miRNA and disease similarity network significantly affects the prediction of miRNA-related disease. The network was derived from [9]. The graph theory was used as a method for connecting different nodes and the walker to measure similarity nodes. Therefore, for all known disease-related miRNAs, if the miRNA m_i (i=1, 2, 3… n) is related to any disease $d_{i,}$ this relationship between miRNA and disease was set to be 1, otherwise 0. This helps us to obtain the number of miRNAs that is connected to disease (d). Then, we extract all unknown miRNAs for the prediction. We ranked all unknown miRNAs. The higher-ranked miRNAs were confirmed as potential candidates of a given disease (d).

$$\begin{cases} 1 \text{ if } m_i = d_{i,} \; i = 1, 2, 3 ... n \\ 0 \qquad \text{otherwise} \end{cases} \tag{3}$$

From the dataset, the large numbers were unknown miRNA that caused greater challenges for researchers in the prediction and discovery of new miRNA-disease associations. Moreover, several diseases have no related miRNA. Therefore, random walk and path theory were combined to overcome these challenges.

Random walk and graph theory

A random walk consists of a sequence of vertices, which are generated from a start vertex by first selecting an edge, then traversing the edge to a new vertex, and finally repeating the process. We constructed a strongly connected graph from RDnet network. Then the fraction of time the walker spends at various vertices of the graph converges to a stationary probability distribution. After constructing RDnet, we solved our problem using random walk to calculate the miRNA score for a given disease. Then, high-ranked miRNAs were stated as potential candidates for miRNA-related disease. Initial states of the walker are highly dependent on the start state of the walk. Therefore, the walker starts walking from a known disease node until it reaches to related miRNA nodes with equal probability for each miRNA. Nevertheless, the probability that the walker arrives at each miRNA node is equal, the walker prefers the neighboring miRNA nodes in the network. Then, from the disease-related miRNA node, the walker will continue walking to unknown miRNAs of a given disease. The probability that unknown miRNA relates with a given disease is considered by the similarity score between unknown miRNA with a given disease.

Two nodes are related if the weight of an edge between two vertices is higher than a given threshold. However, in random walk, the next step does not depend on the previous history of steps, only the current position/node of the moving walker. This is the case for random walk on a directed graph (randomly selecting an outgoing edge out of d… disease to leave from) and walks on a weighted graph (select an edge with probability proportional to its weight). In our method, we modeled the score as a weight of our graph after calculating the similarity score. Here, we have our network as the undirected graph.

Path-based random walk

We design our experiment based on random walk and graph theory. For a given disease d, a known miRNA, miRNA-related disease candidates, and their relationship were modeled as a weighted graph G. Given a weighted graph G = (V, E), where $v \in V$ in G represents the vertices of a known miRNA-related disease, miRNA–miRNA similarity, disease–disease similarity, and miRNA-disease potential candidates, $e \in E \subseteq V \times V$ captures the relationship between two vertices that are linked by the edge. Moreover, the weight w of an edge e measures the relationship between a given disease and unknown miRNAs after calculating the similarity score. The higher the w, value is, the better probability that two vertices are associated with a group of similar diseases. We aim to calculate the ranking order by measuring the similarity score of all unknown nodes to obtain the potential miRNA candidates of a given disease. From a given disease-related miRNA nodes, the walker starts walking to their neighboring nodes with equal probability. The proposed method is illustrated below.

MiRNAs that possess similar functions are implicated in similar diseases and vice versa [14–16]. The graph was labeled as undirected; from the Figure 2, disease network and miRNAs network were combined to form a miRNA-related disease network that is derived from RDnet (miRNA and disease network). Blue straight lines from diseases to miRNAs signify that disease nodes and miRNA nodes are related. The walker starts walking from known related miRNA-disease with equal probability. The probability that the walker starts from diseases (d_1, d_2, d_3… d_n) to miRNAs (m_1, m_2, m_3…m_k) is equal to every known disease-miRNA node. Therefore, the walker proceeds to the closest neighbor of a related miRNA; the closer it is to the node, the higher it is likely to be connected. From Figure 2, m_5 to m_1, and m_5 to m_3 are clearly shown as neighbors in the miRNA network. In addition, disease's network and miRNA's network have been constructed based on their functional similarity. Then, from RDnet, d_1 is strongly related to m_5, d_3 is associated with m_2 and d_1 is also associated with m_3. From the assumption that miRNAs share the same functional similarity and are related to similar diseases and vice versa [14–16]. Thus, the walker moves from m_5 and walks to its neighbor $m_{4.}$ Path-based random walk states that the walker follows the walking path from miRNA-related disease until it reaches to its neighbor. Therefore, the similarity score between a given diseases with any unknown miRNAs was calculated through known disease-associated miRNAs and disease network. The similarity score confirmed that d_1 and m_4 are likely to be associated as shown by the brown dotted lines in Figure 2. Graph theory and random walk were used interchangeably as the walker has to reach each node of the graph until converging; d_4 is associated with m_2 and m_2 has m_1 as its directly connected vertex neighbor. This means that when the walker reaches to m_2, the probability of moving to its neighbors is equal to each unknown miRNA node; then the walker will stay in the unknown node to measure the similarity score. As the walker starts to walk from known nodes d_3–m_2, the walker will continue to the neighbor vertex m_3 to predict an association between vertices d_3 and m_3 as shown by the brown dotted line from nodes d_3 to m_3 in the Figure 2. Here, *m* stands for miRNA and *d* stands for disease. We thus calculated the similarity score from unknown miRNAs with a given query disease. The higher the similarity scores, the better the prediction that miRNA is a potential candidate of a given disease. A comparison of the top 10 miRNAs potential candidates are described in Table 3 as discussed above in section two (Results).

Predictive algorithm for miRNA-related disease similarity score of potential candidates

```
Inputs: RDnet, denoted as $G (V, E, W)$; specific disease $d$.
Outputs: Top ranked $d$-related miRNA candidates.
1. Obtain diseases and miRNAs associations to form RDnet.
2. For $i^{th}$ element of the initial probability $d(0)$, $d_i(0)$ ($1 \leq i \leq N$, $N$ is the number of vertices in the graph G);
3. If the $i^{th}$ vertex, $d_i - m_j$ is a known disease-related miRNA node (where $d_i$ is any disease node and $m_j$ is any miRNA node);
4. $d_i - m_j$ =1
5. else
6. $d_i - m_j$ =0
7. End If
8. End For
9. Initialize the walker vector $X=d(0)$
10. For each vertex $d_i$ ($1 \leq i \leq N$, $d_i$ is the number of disease vertices in the graph G).
11. For each vertex $m_j$ ($1 \leq j \leq N$, $m_j$ is the number of miRNA vertices in the graph G).
12. The walker is walking from $d_i - m_j$ nodes until reaches to unknown neighboring miRNA nodes of disease-related miRNA node.
13. Calculate the sum of the probability that the walker arrives at each unknown miRNA node.
14. The steady-state that the walker will stay at the $m_j$ unknown miRNA node is used as the relevancy score between unknown miRNAs and a disease node.
15. End For
16. End For
17. All the unknown miRNAs nodes are ranked by their scores.
18. The unknown miRNAs with higher ranks are confirmed as potential candidates of $d$-related miRNA.
```

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (61202011 and 61272152), and President's fund of Xiamen University (20720170054).

CONFLICTS OF INTEREST

The authors declare no conflict(s) of interest.

REFERENCES

1. Ambros V. The functions of animal microRNAs. Nature. 2004; 431:350–55.
2. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004; 116:281–97.
3. Miska EA. How microRNAs control cell division, differentiation and death. Curr Opin Genet Dev. 2005; 15:563–68.
4. Calin GA, Croce CM. MicroRNA signatures in human cancers. Nat Rev Cancer. 2006; 6:857–66.
5. Markou A, Tsaroucha EG, Kaklamanis L, Fotinou M, Georgoulias V, Lianidou ES. Prognostic value of mature microRNA-21 and microRNA-205 overexpression in non-small cell lung cancer by quantitative real-time RT-PCR. Clin Chem. 2008; 54:1696–704.
6. Slack FJ, Weidhaas JB. MicroRNA in cancer prognosis. N Engl J Med. 2008; 359:2720–22.
7. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993; 75:843–54.
8. Zeng X, Zhang X, Zou Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. Brief Bioinform. 2016; 17:193–203.
9. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res. 2014; 42:D1070–74.
10. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014; 42:D68–73.

11. Wang D, Gu J, Wang T, Ding Z. OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs. Bioinformatics. 2014; 30:2237–38.

12. Yang Z, Wu L, Wang A, Tang W, Zhao Y, Zhao H, Teschendorff AE. dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. Nucleic Acids Res. 2017; 45:D812–18.

13. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. 2009; 37:D98–104.

14. Bandyopadhyay S, Mitra R, Maulik U, Zhang MQ. Development of the human cancer microRNA network. Silence. 2010; 1:6.

15. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci USA. 2007; 104:8685–90.

16. Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, Cui Q. An analysis of human microRNA and disease associations. PLoS One. 2008; 3:e3420.

17. Xu Y, Zhou X, Zhang W. MicroRNA prediction with a novel ranking algorithm based on random walks. Bioinformatics. 2008; 24:i50–58.

18. Chen X, Liu MX, Yan GY. RWRMDA: predicting novel human microRNA-disease associations. Mol Biosyst. 2012; 8:2792–98.

19. Gu B, Sheng VS. A robust regularization path algorithm for ν-support vector classification. IEEE Trans Neural Netw Learn Syst. 2016. https://doi.org/10.1109/TNNLS.2016.2527796.

20. Liu Y, Zeng X, He Z, Zou Q. Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. IEEE/ACM Trans Comput Biol Bioinformatics. 2016. https://doi.org/10.1109/TCBB.2016.2550432.

21. Jiang Q, Hao Y, Wang G, Juan L, Zhang T, Teng M, Liu Y, Wang Y. Prioritization of disease microRNAs through a human phenome-microRNAome network. BMC Syst Biol. 2010; 4:S2.

22. Chen X, Yan CC, Zhang X, You ZH, Huang YA, Yan GY. HGIMDA: heterogeneous graph inference for miRNA-disease association prediction. Oncotarget. 2016; 7:65257–69. https://doi.org/10.18632/oncotarget.11251.

23. Chen X, You ZH, Yan GY, Gong DW. IRWRLDA: improved random walk with restart for lncRNA-disease association prediction. Oncotarget. 2016; 7:57919–31. https://doi.org/10.18632/oncotarget.11141.

24. Xuan P, Han K, Guo Y, Li J, Li X, Zhong Y, Zhang Z, Ding J. Prediction of potential disease-associated microRNAs based on random walk. Bioinformatics. 2015; 31:1805–15.

25. Xuan P, Han K, Guo M, Guo Y, Li J, Ding J, Liu Y, Dai Q, Li J, Teng Z, Huang Y. Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. PLoS One. 2013; 8:e70204.

26. Chen X, Yan CC, Zhang X, You ZH, Deng L, Liu Y, Zhang Y, Dai Q. WBSMDA: within and between score for MiRNA-disease association prediction. Sci Rep. 2016; 6:21106.

27. Chen H, Zhang Z. Similarity-based methods for potential human microRNA-disease association prediction. BMC Med Genomics. 2013; 6:12.

28. Li JQ, Rong ZH, Chen X, Yan GY, You ZH. MCMDA: matrix completion for MiRNA-disease association prediction. Oncotarget. 2017; 8:21187–99. https://doi.org/10.18632/oncotarget.15061.

29. Zou Q, Wan S, Ju Y, Tang J, Zeng X. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. BMC Syst Biol. 2016; 10:114.

30. Le DH. Network-based ranking methods for prediction of novel disease associated microRNAs. Comput Biol Chem. 2015; 58:139–48.

31. Mørk S, Pletscher-Frankild S, Palleja Caro A, Gorodkin J, Jensen LJ. Protein-driven inference of miRNA-disease associations. Bioinformatics. 2014; 30:392–97.

32. Zeng X, Liao Y, Zou Q. Prediction and validation of disease genes using HeteSim Scores. IEEE/ACM Trans Comput Biol Bioinformatics. 2016. https://doi.org/10.1109/TCBB.2016.2520947.

33. Zeng X, Zhang X, Liao Y, Pan L. Prediction and validation of association between microRNAs and diseases by multipath methods. Biochimica et Biophysica Acta (BBA)-General Subjects. 2016; 1860: 2735-2739. https://doi.org/10.1016/j.bbagen.2016.03.016.

34. Zou Q, Li J, Hong Q, Lin Z, Wu Y, Shi H, Ju Y. Prediction of microRNA-disease associations based on social network analysis methods. BioMed Res Int. 2015; 2015: 810514. https://doi.org/10.1155/2015/810514.

35. Gu B, Sheng VS, Wang Z, Ho D, Osman S, Li S. Incremental learning for ν-Support Vector Regression. Neural Netw. 2015; 67:140–50.

36. Schölkopf B, Smola AJ, Williamson RC, Bartlett PL. New support vector algorithms. Neural Comput. 2000; 12:1207–45.

37. Chen X, Yan CC, Zhang X, You ZH. Long non-coding RNAs and complex diseases: from experimental results to computational models. Brief Bioinform. 2016. https://doi.org/10.1093/bib/bbw060.

38. Chen X, Yan GY. Semi-supervised learning for potential human microRNA-disease associations inference. Sci Rep. 2014; 4:5501.

39. Chen X, Huang YA, Wang XS, You ZH, Chan KC. FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model. Oncotarget. 2016; 7:45948–58. https://doi.org/10.18632/oncotarget.10008.

40. Zeng X, Ding N, Rodríguez-Patón A, Lin Z, Ju Y. Prediction of MicroRNA–disease Associations by Matrix Completion. Curr Proteomics. 2016; 13:151–57.

41. Chen X, Yan CC, Zhang X, Li Z, Deng L, Zhang Y, Dai Q. RBMMMDA: predicting multiple types of disease-microRNA associations. Sci Rep. 2015; 5:13877.

42. Chen X, Huang YA, You ZH, Yan GY, Wang XS. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics. 2017; 33:733–39.

43. Chen X, Ren B, Chen M, Wang Q, Zhang L, Yan G. NLLSS: predicting synergistic drug combinations based on semi-supervised learning. PLOS Comput Biol. 2016; 12:e1004975.

44. Gu B, Sun X, Sheng VS. Structural minimax probability machine. IEEE Trans Neural Netw Learn Syst. 2016. https://doi.org/10.1109/TNNLS.2016.2544779.

45. Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics. 2010; 26:1644–50.

46. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. Eur J Hum Genet. 2006; 14:535–42.

47. Shi H, Xu J, Zhang G, Xu L, Li C, Wang L, Zhao Z, Jiang W, Guo Z, Li X. Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. BMC Syst Biol. 2013; 7:101.

48. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005; 33:D514–17.

49. Ala U, Piro RM, Grassi E, Damasco C, Silengo L, Oti M, Provero P, Di Cunto F. Prediction of human disease genes by human-mouse conserved coexpression analysis. PLOS Comput Biol. 2008; 4:e1000043.

50. Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. Mol Syst Biol. 2008; 4:189.