Bioinformatics, 31(11), 2015, 1805–1815
doi: 10.1093/bioinformatics/btv039
Advance Access Publication Date: 23 January 2015
Original Paper

OXFORD

Data and text mining

Prediction of potential disease-associated microRNAs based on random walk

Ping Xuan[1], Ke Han[2], Yahong Guo[3,*], Jin Li[4], Xia Li[4,*], Yingli Zhong[1], Zhaogong Zhang[1] and Jian Ding[1]

[1]School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China, [2]School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150028, China, [3]School of Information Science and Technology, Heilongjiang University, Harbin 150080, China and [4]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

*To whom correspondence should be addressed.

Associate Editor: Igor Jurisica

Received on August 27, 2014; revised on December 29, 2014; accepted on January 18, 2015

Abstract

Motivation: Identifying microRNAs associated with diseases (disease miRNAs) is helpful for exploring the pathogenesis of diseases. Because miRNAs fulfill function via the regulation of their target genes and because the current number of experimentally validated targets is insufficient, some existing methods have inferred potential disease miRNAs based on the predicted targets. It is difficult for these methods to achieve excellent performance due to the high false-positive and false-negative rates for the target prediction results. Alternatively, several methods have constructed a network composed of miRNAs based on their associated diseases and have exploited the information within the network to predict the disease miRNAs. However, these methods have failed to take into account the prior information regarding the network nodes and the respective local topological structures of the different categories of nodes. Therefore, it is essential to develop a method that exploits the more useful information to predict reliable disease miRNA candidates.

Results: miRNAs with similar functions are normally associated with similar diseases and vice versa. Therefore, the functional similarity between a pair of miRNAs is calculated based on their associated diseases to construct a miRNA network. We present a new prediction method based on random walk on the network. For the diseases with some known related miRNAs, the network nodes are divided into labeled nodes and unlabeled nodes, and the transition matrices are established for the two categories of nodes. Furthermore, different categories of nodes have different transition weights. In this way, the prior information of nodes can be completely exploited. Simultaneously, the various ranges of topologies around the different categories of nodes are integrated. In addition, how far the walker can go away from the labeled nodes is controlled by restarting the walking. This is helpful for relieving the negative effect of noisy data. For the diseases without any known related miRNAs, we extend the walking on a miRNA-disease bilayer network. During the prediction process, the similarity between diseases, the similarity between miRNAs, the known miRNA-disease associations and the topology information of the bilayer network are exploited. Moreover, the importance of information from different layers of network is considered. Our method achieves superior performance for 18 human diseases with AUC values ranging from 0.786 to 0.945. Moreover, case studies on breast neoplasms, lung neoplasms, prostatic neoplasms and 32 diseases further confirm the ability of our method to discover potential disease miRNAs.

© The Author 2015. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/31/11/1805/2364959
by Xiamen University user
on 15 July 2018

Availability and implementation: A web service for the prediction and analysis of disease miRNAs is available at http://bioinfolab.stx.hk/midp/.
Contact: guoyahong_hlju@163.com or lixia@hrbmu.edu.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

MicroRNAs (miRNAs) are small non-coding RNAs that play important roles in gene regulation by targeting mRNAs for cleavage or translational repression (Bartel, 2004; Chatterjee and Grobhans, 2009). Recently, accumulating evidence has indicated that miRNA dysregulation is closely related to the development, progression and prognosis of various human diseases (Alvarez-Garcia *et al.*, 2005; Lynam-Lennon *et al.*, 2009; Meola *et al.*, 2009). Therefore, identifying the miRNAs associated with diseases (disease miRNAs) contributes to the exploration of the pathogenesis of diseases.

Recently, computational prediction methods have been used to obtain reliable disease miRNA candidates for further experimental studies. Since miRNAs fulfill their functions via the regulation of target mRNAs (target genes) expression, several methods based on the targets have been presented. Jiang *et al.* (2010a) estimated the similarity between miRNAs by measuring the similarity of their associated target genes. The miRNA network based on targets was combined with a disease phenotype network to infer the correlation scores between miRNAs and diseases. In addition, they improved the score calculation by further integrating the similarities of miRNAs with the phenotype similarities of diseases (Jiang *et al.*, 2010b). Li *et al.* (2011) also collected miRNA targets, and measured the function consistence score (FCS) between the target genes and the disease-related genes. However, when calculating the FCS, this method ignored the topological structure that is composed of the targets and disease genes. Moreover, as the current number of targets that have been verified by biological experiments is insufficient, both Jiang and Li obtained the target genes by using target prediction programs, such as TargetScan (Lewis *et al.*, 2003) and PITA (Kertesz *et al.*, 2007). Because these programs have high false-positive and false-negative rates (Bartel, 2009; Liu *et al.*, 2014; Ritchie *et al.*, 2009), it is difficult for the methods based on targets to achieve high prediction performance. Shi *et al.* (2013) exploited the functional link between miRNA targets and disease genes in protein–protein interaction network to identify miRNA-disease associations. This method was also affected by the low accuracy of target prediction. In addition, it ignored the functional similarity between genes.

It is well known that miRNAs with similar functions are normally implicated in similar diseases and vice versa (Bandyopadhyay *et al.*, 2010; Goh *et al.*, 2007; Lu *et al.*, 2008). Chen's method (Chen and Zhang, 2013) focused on the phenotype similarity between diseases and the associations between miRNAs and diseases. For a specific miRNA m_i, the novel diseases that are similar to the known m_i-related diseases were obtained as m_i-related candidates. Due to not considering the similarity between miRNAs, this method could not achieve excellent performance. As the functionally related miRNAs tend to be associated with similar diseases, the functional similarity of two miRNAs has been successfully estimated based on the semantic similarities of their associated diseases (Wang *et al.*, 2010). HDMP integrated the functional similarities with the characteristics of miRNAs to predict candidates associated with a given disease (Xuan *et al.*, 2013). It only considered the k most similar neighbors of a candidate and ignored the topology formed by the neighbors. To construct a miRNA network derived from miRNA-associated diseases, the functional similarity between any two miRNAs was used as the weight of edge connecting them. By integrating known miRNA-disease associations, the similarity between diseases and the miRNA network, RLSMDA (Chen and Yan, 2014) developed the prediction method based on regularized least squares to uncover potential miRNA candidates for a specific disease. RLSMDA achieved excellent prediction performance not only for the diseases with some related miRNAs but also for the diseases without any known related miRNAs. Unfortunately, it ignored the topology information of the miRNA network. RWRMDA (Chen *et al.*, 2012) obtained the putative disease miRNAs that have similar functions to known disease miRNAs via random walk through the miRNA network. However, the network is composed of two categories of nodes. For a specific disease, some nodes are validated by biological experiments to be implicated in the disease, but the other nodes have no evidence to verify their association with the disease. Unfortunately, the prior information regarding the two categories of nodes and their respective local topological structures are not considered. Therefore, we propose a novel prediction method based on random walk, which exploits the characteristics of the nodes and the various ranges of topologies. In addition, we extend the walk on a miRNA-disease bilayer network to predict candidates specially for the diseases without any known related miRNAs.

2 Methods

2.1 Process of predicting disease miRNAs

We model the prediction process as random walk on a miRNA network derived from miRNA-associated diseases. Our method for **mi**RNAs associated with **d**iseases **p**rediction is referred to as **MIDP**. For a specific disease d with some related miRNAs, a random walker starts at one of known d-related miRNA nodes with equal probability. If a neighbor of the current node is more likely to be associated with d, the walker transmits to it with a greater proportion. After the iterative walking process is converged, the steady-state probability with which the walker stays at a candidate node (possible d-related miRNA) is defined as its relevance score. In this way, the candidates with higher scores are more likely to be associated with d.

For the disease d, the known d-related miRNAs, the d-related miRNA candidates and their relationship form a network, which is denoted as a weighted graph $G(V, E, W)$. Each vertex ($v \in V$) represents a d-related miRNA or a candidate. Each edge ($e \in E$) captures the relationship between the two vertices that are linked by edge e. The weight w of e is set as the functional similarity which quantifies the relationship. A greater w means that the two vertices are more likely to be associated with a group of similar diseases.

In the network, the known d-related miRNAs are referred to as the *labeled nodes*. The remaining miRNAs, which have no evidence to verify that they are associated with d so far, are the *unlabeled nodes*. Because the unlabeled nodes are probably associated with d, the prediction goal is to rank all the unlabeled nodes and obtain the potential disease miRNAs. In our method, we correlate an unlabeled node u_i with a relevance score $RScore(u_i)$. A higher $RScore(u_i)$ represents a more probable association between u_i and d.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/31/11/1805/2364959
by Xiamen University user
on 15 July 2018

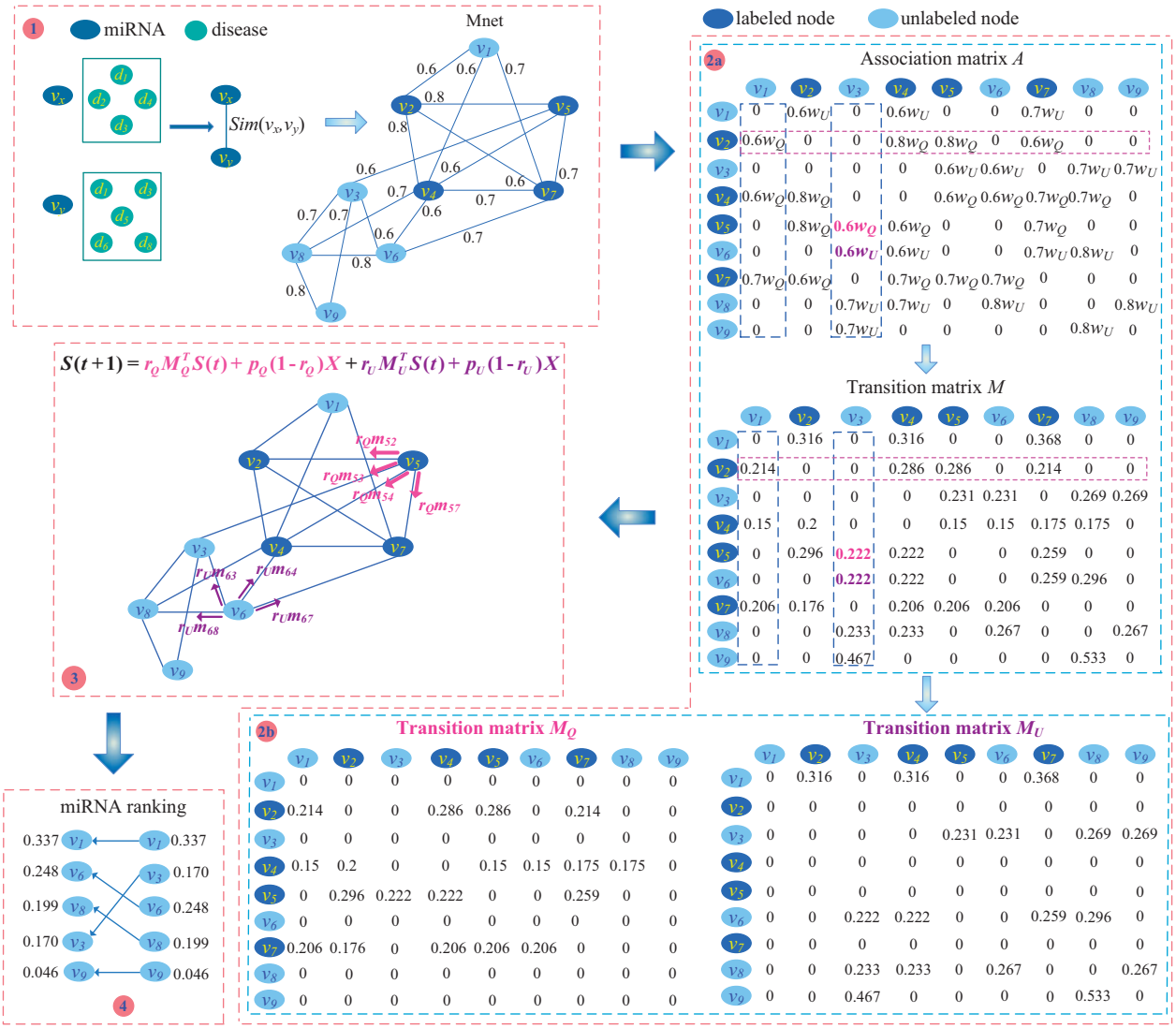


Fig. 1. Illustration of the process of predicting disease *d*-related miRNA candidates. (**1**) calculate the functional similarity between miRNAs based on their associated diseases and construct the miRNA network derived from miRNA-associated diseases (Mnet). (**2a**) construct the association matrix (*A*) to obtain the transition matrix (*M*). (**2b**) separate the matrix *M* into two matrices, M_Q (transition matrix of the labeled nodes) and M_U (transition matrix of the unlabeled nodes). (**3**) a novel prediction model is established to estimate the relevance scores of candidates. (**4**) rank all the unlabeled nodes and select the potential candidates

The process of predicting *d*-related miRNAs by MIDP is illustrated in Figure 1. First, the functional similarity of any two miRNAs is calculated by measuring the semantic similarities of their associated diseases. The miRNA network derived from miRNA-associated diseases (**Mnet**) is constructed by defining the functional similarity of two miRNAs as their edge weight. Second, we divide the nodes within Mnet into labeled nodes and unlabeled nodes. Based on the similarity between the nodes and the prior information regarding the nodes, the transition matrix of the labeled nodes (M_Q) and that of the unlabeled nodes (M_U) are created. Third, we construct a novel prediction model based on random walk to estimate the relevance score between each unlabeled node and *d*. Finally, all the unlabeled nodes are ranked by their scores, and the top ranked nodes are selected as potential candidates.

2.2 Construction of Mnet

Accurate calculation of the functional similarity between miRNAs is the basis for constructing Mnet. It has been observed that miRNAs associated with similar diseases normally have similar functions and vice versa (Bandyopadhyay *et al.*, 2010; Goh *et al.*, 2007; Lu *et al.*, 2008). Therefore, the functional similarity of two miRNAs is successfully estimated by measuring the semantic similarities of their associated diseases (Wang *et al.*, 2010). For instance, as shown in Figure 1, miRNA v_x is associated with disease d_1, d_2, d_3 and d_4, and miRNA v_y is associated with d_1, d_3, d_5, d_6 and d_8. Wang *et al.* calculated the semantic similarity between $\{d_1, d_2, d_3, d_4\}$ and $\{d_1, d_3, d_5, d_6, d_8\}$ as the similarity of v_x and v_y, denoted as $Sim(v_x,v_y)$.

However, when they calculated the similarity, Wang *et al.* extracted the disease information associated with the miRNAs from an earlier version of the human miRNA-disease database (HMDD released in September 2009, Lu *et al.*, 2008). The latest version of HMDD was released in June 2014 (Li *et al.*, 2014). We recalculated the similarity with the latest data. If the similarity of two miRNAs is greater than 0, an edge is added to link them. Furthermore, the weight of their edge is set as the similarity. Thus, Mnet (denoted as a weighted graph *G*) is constructed.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/31/11/1805/2364959
by Xiamen University user
on 15 July 2018

2.3 Two one-step transition matrices

For a given disease d, the prediction of the d-related miRNAs is modeled as random walk on the weighted graph G. A random walker starts at the labeled nodes and walks around them. Assume the walker stays at vertex v_x now. As far as v_x is concerned, if a neighbor of v_x is more likely to be associated with d, we make the walker transmit to the neighbor with a higher probability. Furthermore, the walker iteratively walks to sufficiently exploit the topological structural information of graph G. After the iterative walking process is converged, the steady-rate probability for a vertex indicates the possibility that the walker will finally stay at the vertex. A greater probability reveals a closer association relationship between the vertex and d.

The key to our prediction model is to determine the one-step transition probability of the walker. As far as a vertex (such as v_i) within graph G is concerned, if one of its neighbors (such as v_j) is more likely to be associated with d, we assign v_j with higher transition probability (m_{ij}). Therefore, the transition probability assignment mainly includes two steps. First, we estimate the degree of association (a_{ij}) between the neighbor v_j and d with respect to v_i. Second, we make the transition probability m_{ij} proportional to a_{ij}. The detailed probability assignment strategy is illustrated as follows.

The association matrix $A(a_{ij})_{N\times N}$ is constructed first. a_{ij} represents the degree of association between vertex v_j and d, with respect to v_i. A greater a_{ij} means a more probable association between v_j and d, and N is the number of vertices within graph G. The elements of matrix A are then classified into the following four cases:

1. Assume Q is a set composed of all the labeled nodes and $v_i \in Q$. As the neighbors of v_i have similar function with v_i, they are likely to be associated with a group of similar diseases. Therefore, if v_i is associated with disease d, its neighbors are also likely to be associated with d. For instance, as shown in Figure 1, v_7 is a labeled node (d-related miRNA) and v_1 is one of its neighbors. The higher the functional similarity between v_7 and v_1 [denoted as $Sim(v_7,v_1)$], the greater the association possibility between v_1 and d. In addition, the association possibility when v_7 is a labeled node, is greater than that when v_7 is an unlabeled node. Therefore, the former and the latter are multiplied by w_Q and w_U, respectively. $w_Q \in (0,1)$ represents the weight of the association information from the labeled nodes, and $w_U \in (0,1)$ represents that from the unlabeled nodes. Obviously, w_Q is greater than w_U. Thus, with respect to v_7, the degree of association between v_1 and d (a_{71}) is set as $Sim(v_7,v_1)\cdot w_Q = 0.7w_Q$. In this way, a_{72}, a_{74}, a_{75} and a_{76} are $0.6w_Q$, $0.7w_Q$, $0.7w_Q$ and $0.7w_Q$, respectively.
2. Assume U is a set composed of all the unlabeled nodes (possible d-related miRNAs) and $v_i \in U$. As v_i has similar function with its neighbors, v_i and its neighbors tend to be associated with similar diseases. In other words, if v_i is likely to be associated with d, its neighbors are also likely to be associated with d. For instance, v_8 is an unlabeled node and it is possibly associated with d. For one thing, the higher the functional similarity between v_8 and its neighbor v_3 [denoted as $Sim(v_8,v_3)$], the greater the association possibility between v_3 and d. For another, the weight of the association information from the unlabeled nodes is w_U. So as far as v_8 is concerned, the association degree between v_3 and d (a_{83}) is set as $Sim(v_8,v_3)\cdot w_U = 0.7w_U$. In this way, a_{84}, a_{86} and a_{89} are $0.7w_U$, $0.8w_U$ and $0.8w_U$, respectively.
3. If there is no edge between two miRNAs, such as v_i and v_j, a_{ij} is set as 0. For instance, as v_1 is not connected with v_5, a_{15} is 0.
4. The a value from a miRNA to itself, such as a_{ii}, is set as 0. For instance, a_{11} is 0.

We can construct the association matrix $A(a_{ij})_{N\times N}$ according to the above rules. The formal definition of a_{ij} is as follows,

$$a_{ij} = \begin{cases} Sim(v_i,v_j)\cdot w_Q, & v_i \in Q,\ (v_i,v_j) \in E \\ Sim(v_i,v_j)\cdot w_U, & v_i \in U,\ (v_i,v_j) \in E \\ 0, & (v_i,v_j) \notin E \\ 0, & v_i = v_j \end{cases} \quad (1)$$

where v_i is a vertex and v_j is one of its neighbors.

Subsequently, assume that the walker stays at vertex v_i now. Here, v_i might be a labeled node or an unlabeled node. To make the probability of the walker reaching each its neighbors proportional to the degree of association between the neighbor and d, we must construct the transition matrix $M(m_{ij})_{N\times N}$. Matrix A is row-normalized by the following equation to obtain the one-step transition probability matrix $M(m_{ij})_{N\times N}$.

$$m_{ij} = a_{ij} / \sum_{j=1}^{N} a_{ij} \quad (2)$$

m_{ij} represents the transition probability from v_i to v_j. Therefore, the higher the degree of association between a neighbor of v_i and d, the greater the transition probability from v_i to the neighbor. For example, v_3, v_4, v_7 and v_8 are the neighbors of v_6. As v_6 is an unlabeled node, and when only considering v_6, the association degree between its neighbors and d are $0.6w_U$, $0.6w_U$, $0.7w_U$ and $0.8w_U$, respectively. The transition probabilities are 0.222, 0.222, 0.259 and 0.296 after row-normalization.

However, after the rows of A are normalized, the weights (w_Q and w_U) used to differentiate the association information from the labeled nodes and the unlabeled nodes are lost. For example, when normalizing the 6th row, the process of transforming a_{63} into m_{63} is as follows.

$$\begin{aligned} m_{63} &= \frac{0.6w_U}{0.6w_U + 0.6w_U + 0.7w_U + 0.8w_U} \\ &= \frac{0.6}{0.6 + 0.6 + 0.7 + 0.8} = 0.222 \end{aligned} \quad (3)$$

No matter what the value of w_U is, its values in both the numerator and the denominator can be removed. Therefore, the effect of w_U is lost. In addition, when normalizing the 5th row, the process of transforming a_{53} into m_{53} is as follows.

$$\begin{aligned} m_{53} &= \frac{0.6w_Q}{0.8w_Q + 0.6w_Q + 0.6w_Q + 0.7w_Q} \\ &= \frac{0.6}{0.8 + 0.6 + 0.6 + 0.7} = 0.222 \end{aligned} \quad (4)$$

In this way, the effect of w_Q is also lost. As a result, the influence of the prior information as to whether a vertex is associated with d, is ignored.

To solve this problem, we separate matrix M into two matrices, M_Q and M_U. M_Q represents the transition matrix of the labeled nodes and M_U represents that of the unlabeled nodes. All the rows of the labeled nodes in M_Q are consistent with the corresponding rows in M. The remaining rows of M_Q are set as 0. In M_U, all the rows of unlabeled nodes are consistent with the corresponding rows in M. The remaining rows are set as 0.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/31/11/1805/2364959
by Xiamen University user
on 15 July 2018

2.4 Establishing MIDP prediction model

Based on the transition matrices M_Q and M_U, we can further establish the prediction model of MIDP as follows.

$$S(t+1) = r_Q M_Q^T S(t) + p_Q(1-r_Q)X + r_U M_U^T S(t) + p_U(1-r_U)X \quad (5)$$

First, $S(t+1)$ is a probability vector that indicates the walker arrives at the ith vertex with the probability $S_i(t+1)$ at time $t+1$. Similarly, $S(t)$ demonstrates the probability of arriving at each vertex at time t. For a given disease d, since the vertices near the labeled nodes are more likely to be associated with d than those near the unlabeled nodes, a walker starts from a labeled node. $S(0)$ is the initial probability vector that indicates the walker starts walking from one of the labeled nodes with equal probability at time 0. The ith element of $S(0)$, $S_i(0)$, is set as follows.

$$S_i(0) = \begin{cases} 1/|Q| & if\ v_i \in Q \\ 0 & otherwise \end{cases} \quad (6)$$

Second, to exploit the prior information, the labeled nodes and the unlabeled nodes are assigned different weights, r_Q $(0 < r_Q < 1)$ and r_U $(0 < r_U < 1)$. Furthermore, r_Q is greater than r_U. Actually, r_Q and r_U have same functions as w_Q and w_U. They are used to tune the weights of the association information from the labeled nodes and the unlabeled nodes. The ultimate goal is to make the transition probability proportional to the degree of association between each neighboring vertex and d. For instance, the actual transition probability from labeled node v_5 to v_3 is $r_Q \cdot m_{53}$, and that from unlabeled node v_6 to v_3 is $r_U \cdot m_{63}$. The former is higher than the latter when both m_{53} in M_Q and m_{63} in M_U are 0.222.

In addition, r_Q and r_U also affect how far the walker can go away from the labeled nodes and the unlabeled nodes. This means that the greater the r_Q (r_U), the wider the range of topology around the labeled nodes (unlabeled nodes) involved in the walking process. For instance, when considering ($r_Q = 0.4$, $r_U = 0.1$) and ($r_Q = 0.8$, $r_U = 0.2$), the proportion of r_Q:r_U for both is 4:1. However, for the latter, the walker can go farther from the labeled nodes (and unlabeled nodes) than the former.

Third, if the walker locates a labeled node, it will return to the starting vertices (labeled nodes) with the probability $p_Q(1-r_Q)$ at time $t+1$ and restart walking. p_Q is the sum of the probability that the walker arrives at each labeled node at time t. This can be defined as,

$$p_Q = \sum_{v_i \in Q} S_i(t) \quad (7)$$

In the same way, if the walker locates an unlabeled node, it will go back to the starting vertices with the probability $p_U(1-r_U)$ in the next time. p_U represents the sum of the probability that the walker arrives each unlabeled node at time t, which equals $1-p_Q$. X is a vector defining the nodes that are chose by the walker to go back and restart walking. Since the walker restarts from one of the labeled nodes, X equals $S(0)$.

The advantage of restarting walking is to control the overall degree to which the walker can go far from the labeled nodes. According to the above formula, the walker restarts walking with the probability $p_Q(1-r_Q)+p_U(1-r_U)$. The closer to 0 the probability is, the more nodes around the labeled nodes the walker can reach. If the probability is too small, the walker goes too far away from the labeled nodes, which results in the inclusion of noisy data. It is not helpful for improving the prediction performance. If the probability is too great, there is not enough data to accurately estimate the relevance score. Therefore, it is essential to select suitable values for r_Q and r_U. In our experiments, r_Q ranges from 0.9 to 0.1, and for a specific r_Q, r_U ($r_U < r_Q$) ranges from ($r_Q - 0.1$) to 0.1. For instance, if r_Q is 0.5, r_U is 0.4, 0.3, 0.2 or 0.1. Our prediction model achieves the best performance when $r_Q = 0.4$ and $r_U = 0.1$.

The walker starts from the labeled nodes and begins to walk iteratively. The iterative process stops when the convergence condition is satisfied. The convergence condition means that the L_1-norm between the two consecutive vectors, $S(t)$ and $S(t+1)$, is less than 10^{-10}. The steady-state probability with which the walker stays at an unlabeled node, is defined as its relevance score. All the unlabeled nodes are ranked by their scores. A higher score indicates a more probable association between an unlabeled node and the given disease d. The algorithm for predicting d-related miRNAs by MIDP is demonstrated in Figure 2.

2.5 Prediction based on the bilayer network

For the diseases with some known related miRNAs, cross-validation (Section 3.3) demonstrates the superior performance of MIDP. However, for the diseases without any known related miRNAs, almost all the previous methods and MIDP could not be applied to them. To provide the potential miRNA candidates for them, we further propose an extension method based on MIDP, referred to as MIDPE.

MIDPE is also motivated by the observation that miRNAs with similar functions tend to be associated with similar diseases and vice versa. Besides Mnet, a network composed of diseases is exploited to construct a miRNA-disease bilayer network. Given a specific disease d without known related miRNAs, the walker starts from the node d and walks on the bilayer network. If a disease node (d_i) and the node d are more possibly related with similar miRNAs, or a neighbor (m_j) of a miRNA node is more likely to associate with d, the walker transmits to d_i or m_j with greater possibility. After the walking process is converged, the greater probability the walker arriving at a miRNA node indicates the closer association between the miRNA and d.

2.5.1 Construction of bilayer network and its association matrix

The miRNA-disease bilayer network is composed of the miRNA network derived from the miRNA-associated diseases (Mnet), the disease network (Dnet), and the edges between two networks. If a miRNA has been verified to associate with a disease, an edge is added to connect them. In Dnet, each node d_i represents a disease. If the similarity between two nodes d_i and d_j is more than 0, an edge is added to connect the two nodes. The weight of the edge between d_i and d_j is set as the similarity between them. The disease similarity was calculated in the same way as the literature (Wang *et al.*, 2010). Wang's method represented a disease with a directed acyclic graph (DAG), and the DAG contained all the annotation terms relative to the disease. Furthermore, these annotation terms were obtained from the U.S. National Library of Medicine (MeSH, http://www.nlm.nih.gov/mesh). The similarity between diseases was measured based on their DAGs. When constructing the association matrix for the bilayer network, the following three aspects should be considered.

1. The similar diseases are more likely to associate with functionally related miRNAs. For instance, the disease node d_1 in Figure 3 has no any known associated miRNAs. Since d_1 is similar to d_2 and d_2 has been associated with the miRNA m_6, d_1 is possibly associated with m_6. Furthermore, the greater similarity between d_1 and d_2 means the more possible association between

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/31/11/1805/2364959
by Xiamen University user
on 15 July 2018

ALGORITHM: Algorithm of predicting disease *d*-related miRNAs
Input: Mnet, denoted as $G=(V, E, W)$; a specific disease d
Output: ranked *d*-related miRNA candidates and their relevance scores
1 obtain the known *d*-related miRNAs to form the labeled node set Q
2 For the ith element of the initial probability vector $S(0)$, $S_i(0)$ $(1\le i\le N$, N is the number of vertices in the graph $G)$
3 If the ith vertex, v_i, is a labeled node
4 $S_i(0)=1/|Q|$
5 else
6 $S_i(0)=0$
7 End If
8 End For
9 Initialize the restart vector $X = S(0)$
10 For each vertex v_i $(1\le i\le N)$
11 For each vertex v_j $(1\le j\le N)$
12 $a_{ij}=\begin{cases} Sim(v_i,v_j) & if\ (v_i,v_j)\in E \\ 0 & otherwise \end{cases}$
13 End For
14 End For
15 The matrix $A(a_{ij})_{N\times N}$ is row-normalized as the transition matrix $M(m_{ij})_{N\times N}$
16 For each vertex v_i $(1\le i\le N)$
17 If v_i is a labeled node
18 all the elements of the ith row in the matrix M_Q are equal to the corresponding elements of the ith row in M
19 else
20 all the elements of ith row in M_Q are set as 0
21 End If
22 End For
23 For each vertex v_j $(1\le j\le N)$
24 If v_j is an unlabeled node
25 all the elements of the jth row in the matrix M_U are equal to the corresponding elements of the jth row in M
26 else
27 all the elements of jth row in M_U are set as 0
28 End If
29 End For
30 While (S has not converged)
31 calculate the sum of the probability that the walker arrives at each labeled node (p_Q) and the sum for the walker arriving at unlabeled nodes (p_U) at time t
32 $S(t+1)=r_Q M_Q^T S(t)+p_Q(1-r_Q)X+r_U M_U^T S(t)+p_U(1-r_U)X$
33 End While
34 The steady-state probability that the walker stays at the kth unlabeled node u_k, is used as the relevance score between u_k and d
35 All the unlabeled nodes are ranked by their scores
36 The unlabeled nodes with higher ranks are the potential *d*-related candidates

Fig. 2. Algorithm for predicting the miRNA candidates associated with disease *d*

d_1 and m_6. Therefore, assume A_{DD} is the association matrix of Dnet. $(a_{ij})_{DD}$ is set as the similarity between diseases d_i and d_j.

2. The miRNAs with similar functions are more possibly related to a group of similar diseases. Therefore, if there is an edge connecting a disease node and a miRNA node, the walker can jump from Dnet to Mnet along the edge. Through walking on the Mnet, the miRNAs associated with d_1 can be further searched. For example, the walker transmits from d_1 to d_2 and then jumps to m_6. Here, m_6 has been inferred to probably associate with d_1 in the last paragraph. Moreover, m_6 and its neighbors (m_2 and m_4) have similar functions. So m_2 and m_4 are also possibly associated with d_1. Therefore, the walker is allowed to jump from Dnet to Mnet. A_{DM} is a matrix indicating the jumping case. If the disease node d_i is connected with the miRNA node m_j, $(a_{ij})_{DM}$ is set as 1. Otherwise, it is set as 0.
3. During the process of walking on Mnet, if a neighbor (m_2 or m_4) has greater functional similarity with the current node (m_6), it has higher possibility to associate with d_1. Therefore, A_{MM} describes the association matrix of Mnet and $(a_{ij})_{MM}$ is set as the similarity between m_i and m_j.

At last, assume Mnet consists of N miRNAs and Dnet consists of R diseases, and the association matrix of the bilayer network is A_B.

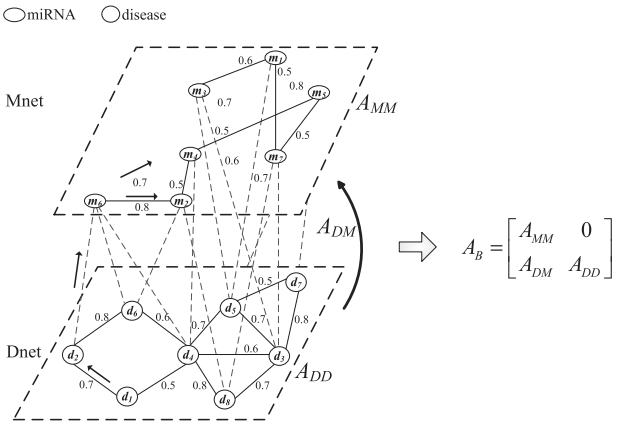


Fig. 3. The miRNA-disease bilayer network and its association matrix A_B

As shown in Figure 3, A_B is composed of A_{MM}, 0 ($0 \in \Re^{N\times R}$), A_{DM}, and A_{DD}.

2.5.2 Establishing MIDPE prediction model

For a random walk on the bilayer network, the transition matrix M_B is defined as follows,

$$M_B=\begin{bmatrix} A^*_{MM} & 0 \\ \alpha A^*_{DM} & (1-\alpha)A^*_{DD} \end{bmatrix} \tag{8}$$

where A^*_{MM}, A^*_{DM}, and A^*_{DD} are the matrices obtained by row-normalizing A_{MM}, A_{DM} and A_{DD} respectively. α controls the relative importance of the information from Mnet or Dnet.

Based on the transition matrix M_B, the prediction model of MIDPE is defined as follows.

$$P(t+1)=(1-\gamma)M_B^T P(t)+\gamma Y \tag{9}$$

Let $P(t)$ be the probability distribution that the walker arrives at the $N+R$ nodes of the bilayer network at time t. It is defined as follows.

$$P(t)=\begin{bmatrix} PM(t) \\ PD(t) \end{bmatrix} \tag{10}$$

The walker arrives at the ith miRNA node with the probability $PM_i(t)$ and arrives at the jth diseases node with $PD_j(t)$ at time t. As the walker starts from a specific disease node, such as d_i, only the ith element of $PD(0)$ is set as 1 and other elements are 0. Moreover, all the elements of $PM(0)$ are set as 0. Since the walker restarts from the node d_i, Y equals $P(0)$. The restart probability $\gamma \in (0,1)$ controls how far the walker can go away from the starting node. After the walking process is converged, the ith probability element of PM with which the walker stays at m_i, is defined as its relevance score.

Both α for adjusting the importance of network information and the restart probability γ range from 0.1 to 0.9. MIDPE achieved the best prediction performance when $\alpha=0.9$ and $\gamma=0.8$. $\alpha=0.9$ indicates that the information of Mnet accounts for the majority in prediction process. It is consistent with that Mnet plays more important role in the miRNA-disease association prediction (Chen and Yan, 2014).

3 Results

3.1 Data preparation

The human miRNA-disease database (HMDD) collected the experimentally validated associations between miRNAs and diseases by

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/31/11/1805/2364959
by Xiamen University user
on 15 July 2018

text mining (Li *et al.*, 2014). For one thing, as done in previous studies (Wang *et al.*, 2010; Chen and Yan, 2014), we merged different miRNA-disease copies that produce the same mature miRNA. For another, there are a few associations with irregular disease names which are not included by the U.S. National Library of Medicine (MeSH, http://www.nlm.nih.gov/mesh). After merging the redundant data and eliminating the irregular data, the latest version of HMDD (updated in June 2014) contains 5100 associations between 490 miRNAs and 326 diseases.

3.2 Evaluation of prediction performance

A specific disease d was used as a query to rank the d-related candidates by their relevance scores. To evaluate the performance of MIDP and other methods, fivefold cross-validation was performed. In the fivefold cross-validation, the labeled nodes were randomly divided into five subsets, four of which were used as known information to predict potential candidates, while the omitted subset was added to the d-related testing dataset as *positive samples*. The dataset also consisted of all the unlabeled nodes (*negative samples*). The relevance score of each node in the testing dataset was calculated, and these nodes were ranked by their scores. The higher the positive samples were ranked, the better the performance.

If a labeled node has higher score than a given threshold θ, it is regarded as a successfully identified positive sample. If an unlabeled node has a score lower than θ, it is regarded as a correctly identified negative sample. By varying θ, the true positive rate (TPR) and the false positive rate (FPR) were calculated to obtain the receiver operating characteristic (ROC) curves.

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{TN + FP} \tag{11}$$

TP and TN are the number of positive samples and the number of negative samples that were correctly identified, respectively. FP and FN are the number of positive samples and the number of negative samples that were identified incorrectly. The area under the ROC curve (AUC) was used as the global prediction performance.

As the top portion of the prediction results is more important, we measured the AUC within the top 30, 60, … , 210 and 240 candidates of the ranking list. In addition, we also reported the recall rate, which measures how many positive samples are successfully identified within the top k.

In the latest association data of HMDD, as the majority of the diseases were associated with only a few miRNAs, they were not sufficient to evaluate the performance of the prediction methods. Therefore, we tested 15 diseases associated with at least 80 miRNAs.

3.3 Comparison with other methods

RWRMDA (Chen *et al.*, 2012), HDMP (Xuan *et al.*, 2013), RLSMDA (Chen and Yan, 2014), Shi's method (Shi *et al.*, 2013), and Chen's method (Chen and Zhang, 2013) are the-state-of-art methods. According to the literatures, RWRMDA and HDMP have achieved significantly better performance than Li's method (Li *et al.*, 2011) and Jiang's method (Jiang *et al.*, 2010a). In addition, Shi's method concentrated on the functional links between miRNA targets and disease-related genes in PPI network. It exploited the interactions between disease-related genes, the associations between miRNAs and their targets and the protein interactions, which were completely different from the datasets used in our method (MIDP). Therefore, Shi's method could not be compared with MIDP in a reasonable and fair way. On the basis of this consideration, we compared MIDP with RWRMDA, HDMP, RLSMDA and Chen's method.

As RWRMDA, HDMP and RLSMDA were developed based on the association data in the earlier versions of HMDD, we implemented these methods with the latest data. Specially, Chen's method focused on the earlier miRNA-disease associations and the phenotype similarity information between diseases. The disease names in Chen's method came from the Online Mendelian Inheritance in Man (OMIM) database (Hamosh *et al.*, 2005), while those in the associations of HMDD came from U.S. National Library of Medicine (MeSH, http://www.nlm.nih.gov/mesh). Therefore, we firstly mapped the disease names contained by HMDD into OMIM disease names, and then updated Chen's method with the latest association data.

When analyzing all the methods, MIDP, RWRMDA, HDMP and Chen's method contained parameters that need to be fine-tuned. The parameters r_Q and r_U of MIDP were chosen from {0.1,0.2, … ,0.9}. Furthermore, r_Q should be greater than r_U. The parameter r of RWRMDA ranged from 0.1 to 0.9. The parameter k of HDMP varied from 1 to 50. The parameter r of Chen's method changed from 0.1 to 0.9. The results of each method that were produced by using the optimum parameters are illustrated in Table 1 and Figure 4 ($r_Q = 0.4$ and $r_U = 0.1$ for MIDP, $r = 0.9$ for RWRMDA, $k = 20$ for HDMP, $r = 0.8$ for Chen's method). The detailed parameter tuning for our method (MIDP) is demonstrated in supplementary Figure S1.

As shown in Table 1, the average AUC values of MIDP, RWRMDA, HDMP, RLSMDA and Chen's method for all 15 diseases, are 86.2, 79.9, 81.6, 82.6 and 65.2%, respectively. MIDP achieved the best prediction performance, and its average AUC is 6.3, 4.6, 3.6 and 21% higher than other four methods, respectively. Although RWRMDA, HDMP and RLSMDA also obtained high performance, MIDP consistently outperformed them in all measures for the 15 diseases. Chen's method produced inferior result. The possible reason is that Chen's method ignored the similarity information between miRNAs. Furthermore, the method strongly depended on how much of the known disease information related to a given miRNA. Currently, more than 30% of miRNAs are only associated with 1–2 diseases. It is difficult to infer accurate disease candidates for these miRNAs.

The results measured by AUC and recall within the top k candidates are shown in Figures 5 and 6, respectively. MIDP also performed the best for the top k ranking list, and ranked approximately 44% of positive testing samples in the top 30 and 86% in the top 120. HDMP ranked approximately 34% in the top 30 and 82% in the top 120, which is not as high as MIDP but better than RWRMDA. However, in the top 180, top 210 and top 240, HDMP had lower recall values than RWRMDA. Similarly, RLSMDA ranked 32.8% in the top 30, which is lower than HDMP. Nevertheless, for the result from the top 60 to the top 240, RLSMDA achieved higher recall values than HDMP. The possible reason is that HDMP only considered the local information regarding the k most similar neighbors. Therefore, the positive testing samples located in the neighborhood of the labeled nodes could be easily identified, while it was difficult to discover those located far from the labeled nodes. MIDP, RWRMDA, HDMP and RLSMDA worked much better than Chen's method because they exploited multiple kinds of information of Mnet while Chen *et al.* ignored the information.

In addition, a paired t-test was used to measure the statistical significance that MIDP's AUCs about multiple diseases were higher than another method. The P-values are listed in Table 2. The results

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/31/11/1805/2364959
by Xiamen University user
on 15 July 2018

Table 1. Prediction results for MIDP and the other methods using 5-fold cross-validation

Disease name	AUC				
	MIDP	RWR-MDA	HDMP	RLS-MDA	Chen's method
Acute myeloid leukemia	0.913	0.839	0.858	0.853	0.716
Breast neoplasms	0.838	0.785	0.801	0.832	0.653
Colorectal neoplasms	0.845	0.793	0.802	0.831	0.662
Glioblastoma	0.786	0.680	0.700	0.714	0.607
Heart failure	0.821	0.722	0.770	0.738	0.761
Hepatocellular carcinoma	0.807	0.749	0.759	0.794	0.613
Lung neoplasms	0.876	0.827	0.835	0.855	0.606
Melanoma	0.837	0.784	0.790	0.807	0.642
Ovarian neoplasms	0.923	0.882	0.884	0.909	0.644
Pancreatic neoplasms	0.945	0.871	0.895	0.887	0.684
Prostatic neoplasms	0.882	0.823	0.854	0.841	0.629
Renal cell carcinoma	0.862	0.815	0.833	0.839	0.627
Squamous cell carcinoma	0.870	0.819	0.820	0.849	0.676
Stomach neoplasms	0.821	0.779	0.787	0.797	0.628
Urinary bladder neoplasms	0.897	0.821	0.850	0.845	0.632

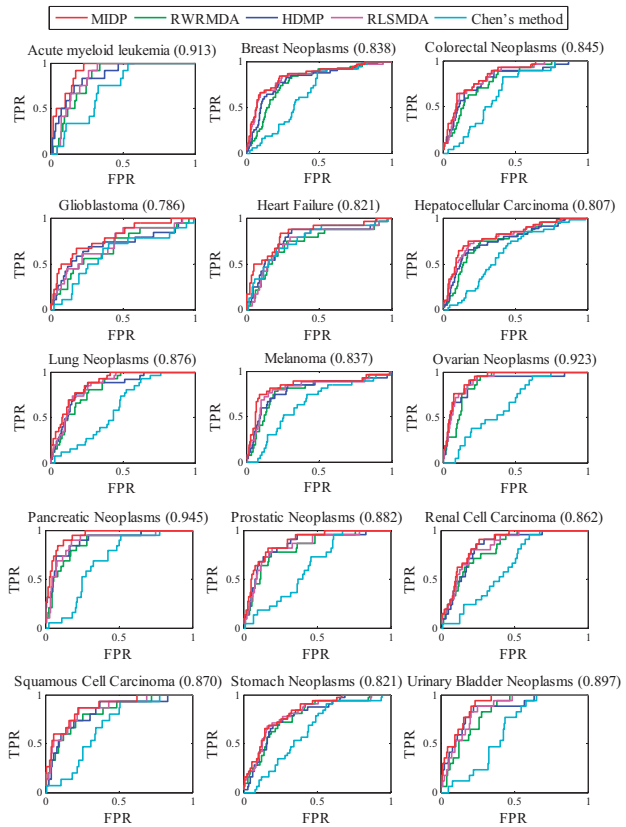


Fig. 4. The ROC curves for MIDP and other methods for 15 diseases. The value in the bracket is the area under MIDP's ROC curve for the specific disease

confirmed that MIDP performed significantly better than the other methods at the significance level 0.05.

To estimate the performance of MIDPE for the diseases without any known related miRNAs, we implemented case studies for three diseases including *breast neoplasms* (*BN*), *colorectual neoplasms* (*CN*) and *hepatocellular carcinoma* (*HC*). For a given disease *d*, we removed all the known associations related to *d*. This operation ensured that predicting *d*-related miRNA candidates only exploited the information of other diseases with known related miRNAs and the information of Mnet. All the removed *d*-rated associations were used as positive samples of the testing dataset.

In the previous methods, only RLSMDA can be applied to the diseases without known related miRNAs. Therefore, MIDPE was compared with RLSMDA. For *BN*, *CN* and *HC*, we removed 202 known *BN*-related miRNA-disease associations, 147 *CN*-related associations and 214 *HC*-related associations, respectively.

For MIDPE, AUC values corresponding to the 3 diseases are 0.821, 0.829 and 0.804. For RLSMDA, AUC values are 0.803, 0.812 and 0.789 (Fig. 7). MIDPE achieved slightly better performance than RLSMDA. The reason is that MIDPE additionally considered the topology between miRNA nodes in Mnet and that between disease nodes in Dnet, relative to RLSMDA.

3.4 Case studies: BN, lung neoplasms, prostatic neoplasms and 32 diseases

To further demonstrate MIDP's ability for discovering potential disease miRNA candidates, case studies of BN, lung neoplasms and prostatic neoplasms were analyzed. We took the lung neoplasms-related candidates as examples and analyzed the top 50 candidates in detail.

First, we used miR2Disease database, which contains manually curated miRNAs that have abnormal regulation in various human diseases (Jiang *et al.*, 2009). This database contains 3273 associations between 349 miRNAs and 163 diseases. As shown in Table 3, 9 of 50 candidates are included in miR2Disease, which indicates their dysregulation in lung neoplasms. This confirms that they are indeed associated with the disease.

Next, the database of differentially expressed miRNAs in human cancers, dbDEMC (Yang *et al.*, 2008), was constructed to provide potential cancer-related miRNAs. With analysis of the microarray datasets, dbDEMC identified 607 miRNAs which were more likely to have different expression levels in 14 types of cancer when compared with normal tissues. 31 of 50 candidates were contained by dbDEMC, indicating that they are potentially upregulated or downregulated in lung cancer (malignant lung neoplasms).

In Table 3, there are 13 candidates labeled with 'literature' (for details see Supplementary Table S1). Several studies confirmed that 8 of 13 miRNAs are significantly upregulated or downregulated in human

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/31/11/1805/2364959
by Xiamen University user
on 15 July 2018

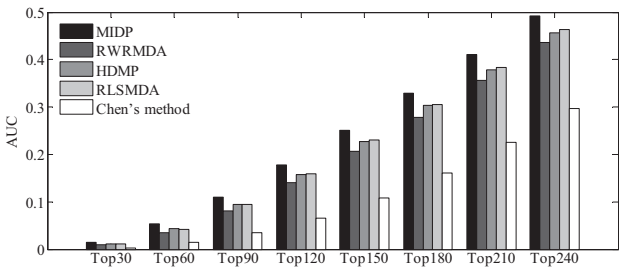


Fig. 5. The average AUCs across all the tested diseases at different top *k* cutoffs

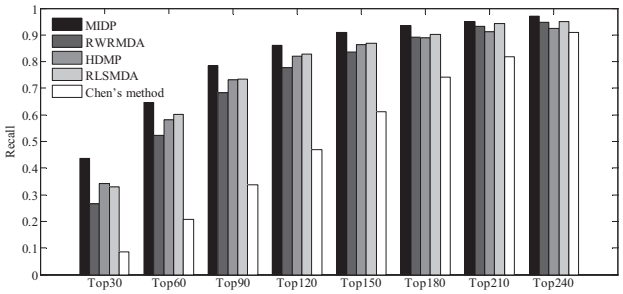


Fig. 6. The average recalls across all the tested diseases at different top *k* cutoffs

Table 2. A pairwise comparison using a paired *t*-test of the ranking results based on the AUCs

	RWRMDA	HDMP	RLSMDA	Chen's method
P-value between MIDP and another method	3.24e–09	1.817e–09	2.589e–05	3.222e–10

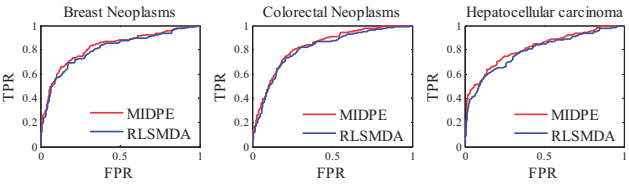


Fig. 7. The ROC curves for MIDPE and RLSMDA for 3 diseases which known related miRNAs were removed

lung neoplasms versus normal lung tissue. In addition, hsa-mir-708 was found to be associated with poor survival in lung adenocarcinomas of non-smokers (Jang *et al.*, 2012). Hsa-mir-302a was identified to regulate epidermal growth factor receptor which is frequently found to be activated by mutation or amplification in lung cancer (Chan *et al.*, 2012). Furthermore, the down-regulation of hsa-mir-320a contributes to the up-regulation of PFKm in lung cancer (Tang *et al.*, 2012). Hsa-mir-452 was significantly upregulated in current smokers compared with former smokers (Mascaux *et al.*, 2013). Hsa-mir-122 was upregulated in five of eight lung cancer patients after diagnosis (Keller et al., 2011). The above analysis indicates that these five miRNAs are likely to be associated with lung cancer as well.

Finally, we used PhenomiR, which is a systematic and manually curated resource that demonstrates miRNA dysregulation in diseases and biological processes (Ruepp *et al.*, 2010). It recorded that hsa-mir-92b and hsa-mir-449b had abnormal expression in lung cancer. Aside from PhenomiR, GeneCards provides compendium for non-coding RNA genes, including the miRNAs (Belinky *et al.*, 2013). The association between hsa-mir-211 and lung cancer has been contained by GeneCards. In addition, it has been reported that miRNAs are often found in genomic clusters (Baskerville *et al.*, 2005). As the clustered miRNAs are usually transcribed together and coordinately regulated, they are more likely to associate with similar diseases (Wang *et al.*, 2010; Xuan *et al.*, 2013). Hsa-mir-373 and hsa-mir-372 are clustered within a 10-kb region on chromosome 19, and the latter has been associated with lung cancer (Yu *et al.*, 2008). This shows that the former is likely to participate in the lung cancer-related biological process.

In terms of BN, the top 50 candidates are listed in Supplementary Table S2. Two candidates were included by miR2disease. Seven candidates were supported by the literature to have dysregulation in BN, and four additional candidates were also supported by the literature to be related to breast cancer-related proteins or transcription factors. dbDEMC identified 37 candidates as potential upregulated or downregulated miRNAs in breast cancer (malignant BN). PhenomiR reported that 1 candidate had abnormal expression in breast cancer. The genes-to-systems breast cancer database (G2SBC) is usually used for studying breast cancer (Mosca *et al.*, 2010). Furthermore, miRNAs execute their functions by regulating target genes. For 4 candidates, G2SBC showed that at least 11 of their top 100 predicted target genes were breast cancer-related genes.

In terms of prostatic neoplasms, the top 50 candidates are demonstrated in Supplementary Table S3. Twelve candidates were included by miR2disease. Eight candidates were supported by the literature to be upregulated or downregulated in prostate neoplasms, and five candidates were reported by the literature to have abnormal expression in metastatic prostate cancer xenografts, relative to their non-metastatic counterparts. dbDEMC identified 30 candidates as potential miRNAs which have abnormal expression levels in prostate cancer. PhenomiR reported that one candidate was overexpressed in prostate cancer. Three candidates were also ranked higher by Li's method, RWRMDA, and Jiang's method, which indirectly confirms that they are more likely to be associated with the disease.

In addition, as RLSMDA was applied to the 32 diseases without any known related miRNAs, MIDPE also investigated these diseases. In the top 3 potential miRNA candidates predicted by MIDPE, 44 miRNA-disease associations were confirmed by the recent literatures and 3 associations have been included by miR2Disease (See supplementary Table S4). In summary, these case studies demonstrate that both MIDP and MIDPE are powerful for discovering potential disease miRNAs.

3.5 Predicting novel miRNA-disease associations

After the superior performance of MIDP and MIDPE was confirmed by cross-validation and case studies, MIDP and MIDPE were further applied to the diseases with known related miRNAs and those without known related miRNAs, respectively. During the prediction process, all known miRNA-disease associations were utilized to construct the prediction model. All the potential disease miRNA candidates were listed in supplementary table S5. MIDP and MIDPE will be useful in generating reliable candidates for subsequent experimental research.

4 Conclusion

A new method based on random walk (MIDP) was developed to predict potential miRNA candidates for the diseases with known

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/31/11/1805/2364959
by Xiamen University user
on 15 July 2018

Table 3. The top 50 lung neoplasms-related candidates

Rank	miRNA name	Description	Rank	miRNA name	Description
1	hsa-mir-130a	dbDEMC, miR2disease	26	hsa-mir-129	literature[1]
2	hsa-mir-151a	literature[1]	27	hsa-mir-20b	dbDEMC
3	hsa-mir-193b	dbDEMC	28	hsa-mir-23b	dbDEMC
4	hsa-mir-302b	dbDEMC	29	hsa-mir-367	literature[1]
5	hsa-mir-16	dbDEMC, miR2disease	30	hsa-mir-92b	PhenomiR
6	hsa-mir-451a	dbDEMC, miR2disease	31	hsa-mir-302a	literature[2]
7	hsa-mir-195	dbDEMC, miR2disease	32	hsa-mir-215	dbDEMC
8	hsa-mir-106b	dbDEMC	33	hsa-mir-320a	literature[2]
9	hsa-mir-139	dbDEMC	34	hsa-mir-302d	dbDEMC
10	hsa-mir-708	literature[2]	35	hsa-mir-328	dbDEMC
11	hsa-mir-99a	dbDEMC, miR2disease	36	hsa-mir-452	literature[2]
12	hsa-mir-296	dbDEMC	37	hsa-mir-345	dbDEMC
13	hsa-mir-149	dbDEMC	38	hsa-mir-339	dbDEMC
14	hsa-mir-429	dbDEMC, miR2disease	39	hsa-mir-449a	literature[1]
15	hsa-mir-625	literature[1]	40	hsa-mir-153	dbDEMC
16	hsa-mir-302c	dbDEMC	41	hsa-mir-342	dbDEMC
17	hsa-mir-15a	dbDEMC	42	hsa-mir-130b	dbDEMC
18	hsa-mir-141	dbDEMC, miR2disease	43	hsa-mir-148b	dbDEMC
19	hsa-mir-378a	literature[1]	44	hsa-mir-196b	dbDEMC
20	hsa-mir-152	dbDEMC	45	hsa-mir-449b	PhenomiR
21	hsa-mir-10a	dbDEMC	46	hsa-mir-122	literature[2]
22	hsa-mir-15b	dbDEMC	47	hsa-mir-211	GeneCards
23	hsa-mir-373	cluster	48	hsa-mir-425	dbDEMC
24	hsa-mir-204	miR2disease	49	hsa-mir-99b	literature[1]
25	hsa-mir-194	literature[1]	50	hsa-mir-372	miR2disease

(1) 'literature[1]' means that there is a literature to support that a miRNA is upregulated or downregulated in human lung neoplasm, as compared with normal lung tissue. (2) 'literature[2]' represents that a miRNA is related to some important factors affecting the development of lung neoplasms. (3) With analysis of the microarray data sets, a miRNA is considered to potentially have different express levels in lung cancer when compared with normal tissues. This kind of miRNAs is labeled by 'dbDEMC'. (4) 'miR2Disease' means that a miRNA is included in the manually curated miRNA-disease association database, miR2Disease. (5) 'PhenomiR' means a miRNA has dysregulation expression in lung cancer. (6) GeneCards provides comprehensive information on all known human genes, including miRNAs. 'GeneCards' means the database recorded that a miRNA was associated with lung cancer. (7) 'cluster' represents a miRNA and another miRNA are clustered, and the latter was associated with lung neoplasms.

related miRNAs. We constructed the miRNA network derived from miRNA-associated diseases to integrate the similarities between nodes, the prior information of nodes, and the local topological structure. Based on the characteristics of the labeled and unlabeled nodes, their transition matrices were established. The transition probability between the nodes was proportional to the similarity between them. The labeled nodes were assigned higher transition weight than the unlabeled nodes, which efficiently exploited the prior information of nodes and the various ranges of topologies. The degree to which the walker could deviate from the labeled nodes was controlled by restarting the walking. This effectively relieved the negative effect of noisy data.

In addition, an extension method (MIDPE) was proposed specially for the diseases without any known related miRNAs. The miRNA-disease bilayer network was constructed to integrate the information in miRNA network derived from miRNA-associated diseases (Mnet) and the information in disease network (Dnet). The transition matrix for the bilayer network was constructed according to the association degree between the miRNA node (or the disease node) and a given disease. At the same time, the information in Mnet and that in Dnet were assigned different weights to balance their relative importance. Similarly, the restart probability in the bilayer network controlled how far the walker can go away from the starting node.

MIDP was compared with RWRMDA, HDMP, RLSMDA and Chen's method. MIDPE was compared with RLSMDA which was the single method applied to the diseases without known related miRNAs before. The results demonstrated that MIDP and MIDPE have superior prediction performance. The cross-validation and the case studies indicated that MIDP and MIDPE are powerful not only for capturing known disease miRNAs but also for discovering potential candidates. It will be useful for providing reliable candidates for future studies of miRNA involvement in the pathogenesis of diseases.

Acknowledgements

The work was supported by the Natural Science Foundation of China (61302139, 61402138), China Postdoctoral Science Foundation (2014M550200, 2014M561350), the Science and Technology Innovation Team Construction Project of Heilongjiang Province College (2013TD012), the Natural Science Foundation of Heilongjiang Province (F201324, E201452), the Postdoctoral Foundation of Heilongjiang Province (LBH-Z14152), the Young Innovative Talent Research Foundation of Harbin Science and Technology Bureau (2012RFQXS094), the Support Program for Young Academic Key Teacher of Higher Education of Heilongjiang Province (1254G030), and the Distinguished Youth Foundation of Heilongjiang University (JCL201405).

Conflict of Interest: none declared.

References

Alvarez-Garcia,I. and Miska,E.A. (2005) MicroRNA functions in animal development and human disease. *Development*, **132**, 4653–4662.

Bandyopadhyay,S. *et al*. (2010) Development of the human cancer microRNA network. *Silence*, **1**, 6.

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

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/31/11/1805/2364959
by Xiamen University user
on 15 July 2018

Bartel,D.P. (2009) MicroRNAs: Target recognition and regulatory functions. *Cell*, **136**, 215–233.

Baskerville,S. and Bartel,D.P. (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, **11**, 241–247.

Belinky,F. *et al*. (2013) Non-redundant compendium of human ncRNA genes in GeneCards. *Bioinformatics*, **29**, 255–261.

Chan,L.W. *et al*. (2012) Genomic sequence analysis of EGFR regulation by microRNAs in lung cancer. *Curr. Top. Med. Chem.*, **12**, 920–926.

Chatterjee,S. and Grobhans,H. (2009) Active turnover modulates mature microRNA activity in caenorhabditis elegans. *Nature*, **461**, 546–549.

Chen,H. and Zhang Z. (2013) Prediction of associations between OMIM Diseases and microRNAs by random walk on OMIM disease similarity network. *Sci. World J.*, **2013**, 204658.

Chen,X. *et al*. (2012) RWRMDA: predicting novel human microRNA-disease associations. *Mol. BioSyst.*, **8**, 2792–2798.

Chen,X. and Yan G.Y. (2014) Semi-supervised learning for potential human microRNA-disease association inference. *Sci. Rep.*, **4**, 5501.

Goh,K.I. *et al*. (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Hamosh,A. *et al*. (2005) Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, 514–517.

Jang,J.S. *et al*. (2012) Increased miR-708 expression in NSCLC and its association with poor survival in lung adenocarcinoma from never smokers. *Clin. Cancer Res.*, **18**, 3658–3667.

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

Jiang,Q. *et al*. (2010a) Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol.*, **4**, S2.

Jiang,Q. *et al*. (2010b) Weighted network-based inference of human microRNA-disease associations. In: *Fifth International Conference on Frontier of Computer Science and Technology, Changchun, Jilin Province, 18-22 Aug. 2010*. IEEE, pp. 431–435.

Keller,A. *et al*. (2011) Stable serum miRNA profiles as potential tool for non-invasive lung cancer diagnosis. *RNA Biol.*, **8**, 506–516.

Kertesz,M. *et al*. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.

Lewis,B.P. *et al*. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.

Li,X. *et al*. (2011) Prioritizing human cancer microRNAs based on genes' functional consistency between microRNA and cancer. *Nucleic Acids Res.*, **39**, 1–10.

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

Liu,B. *et al*. (2014) Identifying miRNAs, targets and functions. *Brief. Bioinf.*, **15**, 1–19.

Lu,M. *et al*. (2008) An analysis of human microRNA and disease associations. *PLoS One*, **3**, e3420.

Lynam-Lennon,N. *et al*. (2009) The roles of microRNA in cancer and apoptosis. *Biol. Rev. Camb. Philos. Soc.*, **84**, 55–71.

Mascaux,C. *et al*. (2013) Endobronchial miRNAs as biomarkers in lung cancer chemoprevention. *Cancer Prev. Res.*, **6**, 100.

Meola,N. *et al*. (2009) MicroRNAs and genetic diseases. *PathoGenetics*, **2**, 7.

Mosca,E. *et al*. (2010) A multilevel data integration resource for breast cancer study. *BMC Syst. Biol.*, **4**, 76.

Ritchie,W. *et al*. (2009) Predicting microRNA targets and functions: traps for the unwary. *Nat. Methods*, **6**, 397–398.

Ruepp,A. *et al*. (2010) PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. *Genome Biol.*, **11**, R6.

Shi,H. *et al*. (2013) Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. *BMC Syst. Biol.*, **7**, 101.

Tang,H. *et al*. (2012) Oxidative stress-responsive microRNA-320 regulates glycolysis in diverse biological systems. *FASEB J.*, **11**, 197467.

Wang,D. *et al*. (2010) Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics*, **26**, 1644–1650.

Xuan,P. *et al*. (2013) Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS ONE*, **8**, e70204.

Yang,Z. *et al*. (2008) dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC Genomics*, **11**, S5.

Yu,S.L. *et al*. (2008) MicroRNA signature predicts survival and relapse in lung cancer. *Cancer Cell*, **13**, 48–57.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/31/11/1805/2364959
by Xiamen University user
on 15 July 2018