Bioinformatics, 34(2), 2018, 239–248
doi: 10.1093/bioinformatics/btx545
Advance Access Publication Date: 1 September 2017
Original Paper

Systems biology

A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations

Qiu Xiao[1], Jiawei Luo[1,*], Cheng Liang[2], Jie Cai[1] and Pingjian Ding[1]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410083, China and [2]College of Information Science and Engineering, Shandong Normal University, Jinan 250000, China

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on January 9, 2017; revised on July 21, 2017; editorial decision on August 30, 2017; accepted on August 31, 2017

Abstract

Motivation: MicroRNAs (miRNAs) play crucial roles in post-transcriptional regulations and various cellular processes. The identification of disease-related miRNAs provides great insights into the underlying pathogenesis of diseases at a system level. However, most existing computational approaches are biased towards known miRNA-disease associations, which is inappropriate for those new diseases or miRNAs without any known association information.

Results: In this study, we propose a new method with graph regularized non-negative matrix factorization in heterogeneous omics data, called GRNMF, to discover potential associations between miRNAs and diseases, especially for new diseases and miRNAs or those diseases and miRNAs with sparse known associations. First, we integrate the disease semantic information and miRNA functional information to estimate disease similarity and miRNA similarity, respectively. Considering that there is no available interaction observed for new diseases or miRNAs, a preprocessing step is developed to construct the interaction score profiles that will assist in prediction. Next, a graph regularized non-negative matrix factorization framework is utilized to simultaneously identify potential associations for all diseases. The results indicated that our proposed method can effectively prioritize disease-associated miRNAs with higher accuracy compared with other recent approaches. Moreover, case studies also demonstrated the effectiveness of GRNMF to infer unknown miRNA-disease associations for those novel diseases and miRNAs.

Availability and implementation: The code of GRNMF is freely available at https://github.com/XIAO-HN/GRNMF/.

Contact: luojiawei@hnu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

MicroRNAs (miRNAs) are a kind of important regulators that play critical roles in post-transcriptional regulations and many important biological processes (Jopling *et al.*, 2005; Xu *et al.*, 2011a,b). Previous studies have shown that the aberrant expression of miRNAs was related to human diseases (Chou *et al.*, 2016; Li *et al.*, 2014a). Experimental determination of new miRNA-disease associations is tremendously expensive and laborious and has a high-failure rate. Therefore, identifying disease-related miRNAs through computational approaches will contribute to the exploration of molecular mechanisms and greatly facilitate disease diagnosis and treatment (Zeng *et al.*, 2016; Luo and Xiao, 2017).

With the development of high-throughput techniques, a vast amount of omics data are publicly available, which create opportunities to decipher the underlying roles of miRNA-associated activities in physiologic and pathologic conditions, such as miRNA-target interaction prediction (Liu *et al.*, 2014b; Chen *et al.*, 2013), transcription factor (TF) and miRNA co-regulatory motif identification (Liang *et al.*, 2015) and miRNA-mRNA regulatory module discovery (Li *et al.*, 2014b; Liang *et al.*, 2016). Undoubtedly, all these studies have greatly expanded our understanding of miRNA functions and their coordinated regulatory mechanisms.

In recent years, to elucidate the initiation and progression of tumorigenesis, considerable efforts have also been made to prioritize disease-associated miRNAs using in silico prediction models. Li *et al.* (2011) developed an approach to identify the potential disease miRNAs through calculation of the relevance between the known disease genes and the target genes. Xu *et al.* (2011a) first constructed a miRNA-target gene dysregulated network and then applied support vector machine classifier to distinguish positive disease miRNAs from negative ones based on the topological properties. Zhao *et al.* (2015) recently utilized gene expression data and miRNA-gene regulations to discover disease miRNA candidates. However, these models based on the miRNA targets encountered difficulties in the achievement of a significant performance because these target prediction databases have relatively high false-negative and false-positive rates (Ritchie *et al.*, 2009; Zhu *et al.*, 2015). Meanwhile, some other approaches have also been developed to discover miRNA-disease associations based on the hypothesis that miRNAs with similarity functions are often associated with similar diseases and vice versa (Ding *et al.*, 2016; Zeng *et al.*, 2016). Chen *et al.* (2012) used the miRNA similarity network to present a new method for identifying disease miRNAs by using random walk with restart. Mørk *et al.* (2014) predicted the potential associations between diseases and miRNAs by combining the linkages among miRNAs, proteins and diseases. Luo *et al.* (2016a) developed a transduction learning-based algorithm to prioritize disease-related miRNAs, especially for those diseases that are associated with sparse known miRNAs. By fully exploiting the characteristic of miRNAs in the constructed miRNA network, Xuan *et al.* (2015) presented a framework called MIDP. Their framework assigned different weights for the different categories of nodes and adopted random walk to predict disease-related miRNAs. Later on, they extended their method to identify candidate miRNAs for those novel diseases (Xuan *et al.*, 2015). In addition, Chen *et al.* (2016) proposed another method based on the between-scores and within-scores of each miRNA-disease pair to prioritize disease miRNAs. More recently, Luo and Qiu (2017) developed a Kronecker regularized least squares-based method by integrating heterogeneous omics data for identifying disease miRNAs. However, most of these approaches strongly rely on the known association information, and only a handful of them could be applied to uncover the potential associations involving novel diseases or miRNAs. Moreover, due to the lack of sufficient experimentally validated interactions, it still remains a challenge to achieve significant performance for prioritization of disease miRNAs.

Here, we develop a novel framework called GRNMF to infer the unknown miRNA-disease associations in heterogeneous omics data, which could work for both new diseases and miRNAs. GRNMF fully exploits the semantic associations between diseases, the weighted gene network, and the experimentally validated miRNA-target gene interactions to quantify the similarities for diseases and miRNAs. Distinct from previous approaches, to extend our method to new diseases and new miRNAs, the weighted nearest neighbor interaction profiles are constructed based on prior information to assist both novel diseases or miRNAs and those with sparse known associations for prediction of potential miRNA-disease associations. Accumulated studies (Luo *et al.*, 2016b; Hernando *et al.*, 2016) have demonstrated that matrix factorization technique is an effective tool that has been successfully applied to recommender systems. Motivated by these, we formulate approved miRNAs, diseases, and miRNA-disease associations as a recommender system, and transform the problem of disease-related miRNA identification into a recommender task. Thus, we adopt a graph regularized non-negative matrix factorization (GRNMF) framework for potential miRNA-disease association inference. The experimental results indicate that GRNMF achieves superior performance compared with other methods and can be effectively applied in the discovery of missing or potential associations for novel diseases and miRNAs.

2 Materials and methods

2.1 Methods overview

To detect miRNA-disease associations that remain undiscovered, we propose a novel method called GRNMF, which consists of three steps (Fig. 1). First, the similarities for miRNAs and diseases are calculated based on the collected data sources. Second, to extend GRNMF to new miRNAs and diseases, a preprocessing step is performed to reduce reliance on validated miRNA-disease associations through the addition of edges with intermediate interaction probability values based on the weighted *K* nearest neighbor profiles (WKNNP). Finally, the framework of graph regularized non-negative matrix factorization is used to infer the potential associations.

2.2 Similarity measures

2.2.1 Disease similarity measure

In this work, we use the hierarchical directed acyclic graphs (DAGs) to calculate the similarities between disease pairs as the same way in Wang *et al.* (2010). $DAG_d = (d, T_d, E_d)$ is a hierarchical DAG graph

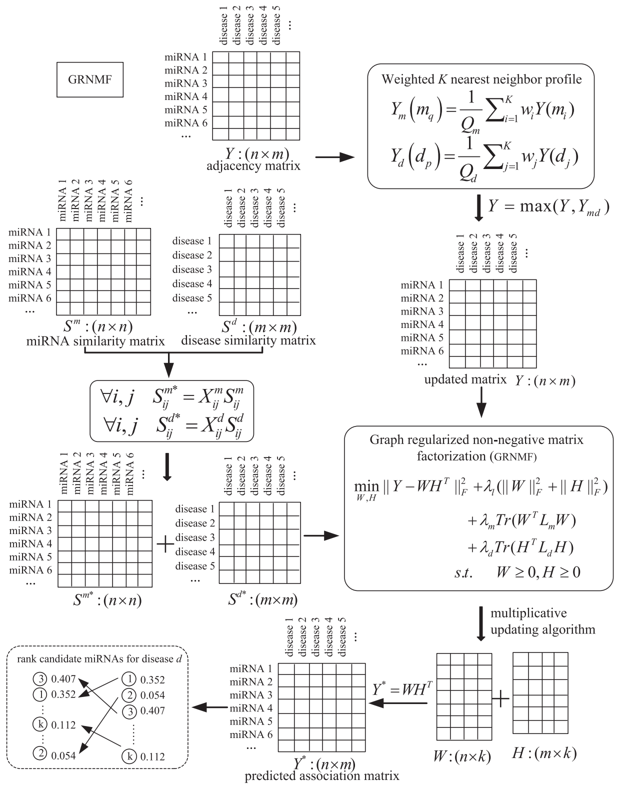
GRNMF
disease 1
disease 2
disease 3
disease 4
disease 5
...
miRNA 1
miRNA 2
miRNA 3
miRNA 4
miRNA 5
miRNA 6
...
$Y:(n\times m)$
adjacency matrix
Weighted K nearest neighbor profile
$Y_m(m_q)=\frac{1}{Q_m}\sum_{i=1}^{K} w_i Y(m_i)$
$Y_d(d_p)=\frac{1}{Q_d}\sum_{j=1}^{K} w_j Y(d_j)$
$Y=\max(Y,Y_{md})$
$S^m:(n\times n)$
miRNA similarity matrix
$S^d:(m\times m)$
disease similarity matrix
updated matrix $Y:(n\times m)$
$\forall i,j \quad S_{ij}^{m*}=X_{ij}^m S_{ij}^m$
$\forall i,j \quad S_{ij}^{d*}=X_{ij}^d S_{ij}^d$
Graph regularized non-negative matrix factorization (GRNMF)
$\min_{W,H} \| Y-WH^T \|_F^2 + \lambda_l(\| W \|_F^2 + \| H \|_F^2)$
$+\lambda_m Tr(W^T L_m W)$
$+\lambda_d Tr(H^T L_d H)$
$s.t. \quad W \geq 0, H \geq 0$
$S^{m*}:(n\times n)$
$S^{d*}:(m\times m)$
multiplicative updating algorithm
rank candidate miRNAs for disease d
3 0.407
1 0.352
k 0.112
2 0.054
1 0.352
2 0.054
3 0.407
k 0.112
$Y^*=WH^T$
$Y^*:(n\times m)$
predicted association matrix
$W:(n\times k)$
$H:(m\times k)$

Fig. 1. Overall workflow of GRNMF for discovering potential miRNA-disease associations

of disease d, where T_d denotes the set of diseases, and E_d denotes the set of links in the graph. The disease DAGs are obtained from MeSH database. Subsequently, we can calculate the semantic contribution of disease t to disease d as follows:

$$D_d(t) = \max\{\Delta * D_d(t')|t' \in childrenof(t)\}, \tag{1}$$

where Δ denotes the semantic contribution factor ($\Delta = 0.5$) (Wang *et al.*, 2010). If a disease pair share a large part of DAGs, they would likely obtain a higher similarity between them. The following equation is used to evaluate the semantic similarity between disease d_i and disease d_j:

$$S^d(d_i, d_j) = \frac{\sum_{t \in T_{d_i} \cap T_{d_j}} \left(D_{d_i}(t) + D_{d_j}(t) \right)}{\sum_{t \in T_{d_i}} D_{d_i}(t) + \sum_{t \in T_{d_j}} D_{d_j}(t)}, \tag{2}$$

where $D_{di}(t)$ and $D_{dj}(t)$ are the semantic values of disease t related to diseases d_i and d_j, respectively.

2.2.2 MiRNA similarity measure

As a matter of fact, most of the existing miRNAs similarity measurements are based on the overlap of miRNA-related diseases. Besides, they also strongly rely on the available association information between miRNAs and diseases (Zeng et al., 2016), which is usually not applicable to novel miRNAs. To solve these limitations, we develop a new measurement to quantify the miRNA similarity by effectively integrating the experimentally verified miRNA-gene interactions as well as the weighted gene functional interaction network. The gene functional interaction network is downloaded from HumanNet (Lee et al., 2011), which uses the associated log-likelihood scores (LLS) of each edge to measure the strength of interaction between any two genes. First, we normalize LLS(g_i, g_j) based on min-max normalization and obtain the normalized similarity LLSN(g_i, g_j) between genes g_i and g_j as follows:

$$LLS_N(g_i, g_j) = \frac{LLS(g_i, g_j) - LLS_{\min}}{LLS_{\max} - LLS_{\min}}, \tag{3}$$

where LLS_{min} and LLS_{max} represent the minimum and maximum LLS in HumanNet, respectively. Consequently, the similarity between genes g_i and g_j is given as follows:

$$S(g_i, g_j) = \begin{cases} 1, & g_i = g_j \\ 0, & e(g_i, g_j) \notin \text{HumanNet} \\ LLS_N(g_i, g_j), & e(g_i, g_j) \in \text{HumanNet} \end{cases}, \tag{4}$$

where $e(g_i, g_j)$ denotes the linkage between genes g_i and g_j. Subsequently, we obtain the similarity between gene g_t and gene set $G = \{g_{t1}, g_{t2}, \ldots, g_{tk}\}$ as follows:

$$S(g_t, G) = \max_{1 \le i \le k} (S(g_t, g_{ti})). \tag{5}$$

After that, the functional similarity between miRNAs m_i and m_j is defined in accordance with the BMA method (Wang *et al.*, 2007), which is calculated as follows:

$$S^m(m_i, m_j) = \frac{\sum_{g \in G_i} S(g, G_j) + \sum_{g \in G_j} S(g, G_i)}{|G_i| + |G_j|}, \tag{6}$$

where G_i and G_j represent the gene sets associated with m_i and m_j, respectively; and $|G|$ represents the number of genes in G.

2.3 Weighted *K* nearest neighbor profiles for miRNAs and diseases

Let $M = \{m_1, m_2, \ldots, m_n\}$ and $D = \{d_1, d_2, \ldots, d_m\}$ denote the set of n miRNAs and m diseases, respectively. $Y \in R^{n \times m}$ represent the adjacency matrix of the original association network, where Y_{ij}=1 if miRNA m_i has a known association with disease d_j; otherwise Y_{ij}=0. The ith row vector of matrix Y, $Y(m_i) = (Y_{i1}, Y_{i2}, \ldots, Y_{im})$, denotes the interaction profile for miRNA m_i. The jth column vector of matrix Y, $Y(d_j) = (Y_{1j}, Y_{2j}, \ldots, Y_{nj})$, indicates the interaction profile for disease d_j. It is obvious that the values in these interaction profiles of the novel miRNAs or diseases are all zeros, which may lead to unsatisfactory performance in the prediction of the potential associations between miRNAs and diseases.

Here, we perform a procedure for the construction of new interaction profiles to address the above-mentioned problem. For each miRNA m_q, its similarity with other K nearest known miRNAs (with at least one experimentally verified association) and their corresponding K interaction profiles are utilized to obtain the following interaction profile:

$$Y_m(m_q) = \frac{1}{Q_m} \sum_{i=1}^{K} w_i Y(m_i), \tag{7}$$

where m_1 to m_K are the miRNAs sorted in descending order based on their similarity to m_q; w_i is the weight coefficient, and w_i=α^{i-1}*$S^m(m_i, m_q)$, which means that a higher weight is assigned if m_i is more similar to m_q. $\alpha \in [0, 1]$ is a decay term, and Q_m=$\sum_{1 \le i \le K} S^m(m_i, m_q)$ is the normalization term. In the same manner, the new interaction profile for each disease d_p can be determined as follows:

$$Y_d(d_p) = \frac{1}{Q_d} \sum_{j=1}^{K} w_j Y(d_j), \tag{8}$$

where d_1 to d_K are the diseases sorted in descending order based on the their similarity to d_p; w_j is the weight coefficient, and w_j=α^{j-1} *$S^d(d_j, d_p)$. Q_d is a normalization term, and Q_d=$\sum_{1 \le j \le K} S^d(d_j, d_p)$.

Thereafter, we combine the above two matrices, Y_m and Y_d, obtained from different data spaces, replace Y_{ij}=0 with an associated likelihood score, and then update the original adjacency matrix Y as follows:

$$Y = \max(Y, Y_{md}), \tag{9}$$

where

$$Y_{md} = (a_1 Y_m + a_2 Y_d) / \sum a_i (i = 1, 2),$$

and a_i is the weight coefficient. For simplicity, we assign the same weight to the two parts, namely $a_1 = a_2 = 1$.

2.4 Graph regularized non-negative matrix factorization for prediction of disease-associated miRNAs

2.4.1 Standard NMF

Non-negative matrix factorization (NMF) is an effective technique and has been widely used for data representation (Hosoda *et al.*, 2009; Zheng *et al.*, 2009; Huang and Zheng, 2006). It aims to find two non-negative matrices whose product provides an optimal approximation to the original matrix. Given the miRNA-disease matrix $Y \in R^{n \times m}$, NMF can be decomposed into two matrices, that is, $W \in R^{n \times k}$ and $H \in R^{m \times k}$ ($k \ll \min(n, m)$), and $Y \approx WH^T$. Here, we mathematically formulate the problem of disease-related miRNA prediction as the following objective function:

$$\min_{W,H} ||Y - WH^T||_F^2 \, s.t. \quad W \ge 0, H \ge 0, \tag{10}$$

where $||.||_F$ represents the Frobenius norm. The above objective function can be minimized using the iterative update algorithm proposed by Lee *et al.* (1999).

2.4.2 GRNMF

The standard NMF in Eq. (10) performs the learning in the Euclidean space, which fails to discover the intrinsic geometrical and discriminating structure of the data space (Li *et al.*, 2016; Wang *et a*l., 2012; Yuan *et al.*, 2016). To prevent overfitting and significantly enhance the learning performance, we present a new objective function through incorporation of the Tikhonov (L_2) and graph Laplacian regularization terms into the standard NMF framework for miRNA-disease association prediction. The Tikhonov regularization is used to ensure the W and H smoothness (Guan *et al.*, 2011), and the graph regularization mainly aims to guarantee a part-based representation by fully exploiting the data geometric structure (Cai *et al.*, 2011). The optimization problem of GRNMF can be formularized as follows:

$$\begin{aligned} \min_{W,H} ||Y - WH^T||_F^2 &+ \lambda_l(||W||_F^2 + ||H||_F^2) \\ &+ \lambda_m \sum_{i,p=1}^{n} ||w_i - w_p||^2 S_{ip}^{m*}, \\ &+ \lambda_d \sum_{j,q=1}^{m} ||h_j - h_q||^2 S_{jq}^{d*} \\ s.t. \quad & W \geq 0, H \geq 0 \end{aligned} \tag{11}$$

where λ_l, λ_m and λ_d are the regularization coefficients; w_i and h_j are the ith and jth rows of W and H, respectively. S^{d*} and S^{m*} are the sparse weight matrices, which are established using the geometrical information of disease and miRNA data spaces (S^d and S^m), and could effectively avoid noisy information to achieve more accurate results. Then, Eq. (11) can be transformed into:

$$\begin{aligned} \min_{W,H} ||Y - WH^T||_F^2 &+ \lambda_l\left(||W||_F^2 + ||H||_F^2\right) \\ &+ \lambda_m Tr\left(W^T L_m W\right) \\ &+ \lambda_d Tr\left(H^T L_d H\right), \\ s.t. \quad & W \geq 0, H \geq 0 \end{aligned} \tag{12}$$

where Tr (.) represents the trace of a matrix; L_m=D_m-S^{m*} and L_d=D_d-S^{d*} are the graph Laplacian matrices for S^{m*} and S^{d*} (Liu *et al.*, 2014a), respectively; D_m and D_d are the diagonal matrices whose entries are row (or column) sums of S^{m*} and S^{d*}, respectively.

Recent studies on spectral graph and manifold learning theories have demonstrated that local geometric structure can be effectively modeled through the nearest neighbor graph on a scatter of data points (Cai *et al.*, 2011; Li *et al.*, 2016; You *et al.*, 2010). Meanwhile, the miRNAs or diseases located in the same cluster tend to behave more similarly. Therefore, we construct the graphs (S^{m*} and S^{d*}) for miRNA and disease spaces based on the p nearest neighbors and clustering information. As a graph clustering method, ClusterONE (Nepusz *et al.*, 2012) is utilized to detect clusters. The graph for miRNA space is constructed, and the weight matrix X^m is generated based on the miRNA similarity matrix S^m as follows:

$$X_{ij}^m = \begin{cases} 1 & i \in N(m_j) \& j \in N(m_i) \ \& \ m_i, m_j \in C \\ 0 & i \notin N(m_j) \& j \notin N(m_i) \ \& \ m_i, m_j \notin C, \\ 0.5 & \qquad\qquad otherwise \end{cases} \tag{13}$$

where $N(m_i)$ and $N(m_j)$ are the sets of p nearest neighbors of m_i and m_j, respectively; C represents any one of the clusters obtained through ClusterONE. Subsequently, we determined the matrix $S^{m*}$for miRNAs as follows:

$$\forall i, j S_{ij}^{m*} = X_{ij}^m S_{ij}^m. \tag{14}$$

By applying the same procedure for diseases, the matrix S^{d*} can be obtained based on the disease similarity matrix S^d.

2.4.3 Optimization

To solve the optimization problem in Eq. (12), let $\Phi = [\varphi_{ik}]$ and $\Psi = [\psi_{jk}]$ be the Lagrange multipliers for the constrains w_{ik}≥0 and h_{jk} 0, respectively. The corresponding Lagrange function L_f of Eq. (12) is defined as:

$$\begin{aligned} L_f = Tr\left(YY^T\right) - 2Tr\left(YHW^T\right) &+ Tr\left(WH^T HW^T\right) \\ + \lambda_l Tr\left(WW^T\right) &+ \lambda_l Tr\left(HH^T\right) \\ + \lambda_m Tr\left(W^T L_m W\right) &+ \lambda_d Tr\left(H^T L_d H\right). \\ + Tr\left(\Phi W^T\right) &+ Tr\left(\Psi H^T\right) \end{aligned} \tag{15}$$

The partial derivatives of the above function with respect to W and H are:

$$\begin{aligned} \frac{\partial L_f}{\partial W} &= -2YH + 2WH^T H + 2\lambda_l W + 2\lambda_m L_m W + \Phi \\ \frac{\partial L_f}{\partial H} &= -2Y^T W + 2HW^T W + 2\lambda_l H + 2\lambda_d L_d H + \Psi. \end{aligned} \tag{16}$$

Using the Karush–Kuhn–Tucker (KKT) conditions (Facchinei *et al.*, 2014) $\varphi_{ik}w_{ik}$=0 and $\psi_{jk}h_{jk}$=0, the following equations are obtained for w_{ik} and h_{jk}:

$$\begin{aligned} -(YH)_{ik}w_{ik} + \left(WH^T H\right)_{ik}w_{ik} &+ (\lambda_l W)_{ik}w_{ik} \\ + [\lambda_m(D_m - S^{m*})W]_{ik}w_{ik} &= 0 \\ -\left(Y^T W\right)_{jk}h_{jk} + \left(HW^T W\right)_{jk}h_{jk} &+ (\lambda_l H)_{jk}h_{jk} \\ + \left[\lambda_d\left(D_d - S^{d*}\right)H\right]_{jk}h_{jk} &= 0. \end{aligned} \tag{17}$$

Therefore, we determine the updating rules as follows:

$$\begin{aligned} w_{ik} &\leftarrow w_{ik}\frac{(YH + \lambda_m S^{m*} W)_{ik}}{(WH^T H + \lambda_l W + \lambda_m D_m W)_{ik}} \\ h_{jk} &\leftarrow h_{jk}\frac{\left(Y^T W + \lambda_d S^{d*} H\right)_{jk}}{(HW^T W + \lambda_l H + \lambda_d D_d H)_{jk}}. \end{aligned} \tag{18}$$

The nonnegative matrices W and H are updated based on Eq. (18) until convergence. Finally, we obtain the predicted miRNA-disease association matrix as Y*=WH^T, and prioritize the disease-related miRNAs based on the entities in matrix Y*. In principle, the top-ranked miRNAs in each column of Y* are more likely to be related to the corresponding disease.

Algorithm 1 summarizes the procedure of GRNMF for miRNA-disease association prediction.

3 Results and discussion

3.1 Data collection and preprocessing

We downloaded the relationships among diseases from MeSH (https://www.nlm.nih.gov/mesh/), which includes 4663 diseases and are adopted to estimate the disease semantic similarity based on their hierarchical structures. Due to the relatively high false negative and false positive rates of some miRNA target prediction programs, we acquired the miRNA-gene interactions from experimentally verified databases, including miRTarBase (version 4.5) (Chou *et al.*, 2016), TarBase (version 6.0) (Vergoulis *et al.*, 2012) and miRecords (version 4.0) (Xiao *et al.*, 2009). After the removal of duplicate interactions, a total of 38 089 interactions between 12 422 genes and 477 miRNAs are retained. We obtained the weighted gene network from HumanNet (Lee *et al.*, 2011), including 476 399 interactions between 16 243 genes. The experimentally verified associations between miRNAs and diseases are derived from the latest version of HMDD v2.0 (Li *et al.*, 2014a), where 5424 associations involving 378 diseases and 495 miRNAs are obtained after combing multiple miRNA transcripts with the same mature miRNA as done in Xuan *et al.* (2015). With the elimination of several irregularly named diseases according to MeSH database and removal of those miRNAs absent from the aforementioned three miRNA target databases, 327 diseases and 351 miRNAs are retained for prediction. Finally, the disease similarity matrix $S^d \in R^{327\times 327}$, the miRNA similarity matrix $S^m \in R^{351\times 351}$, and the adjacency matrix $Y \in R^{351\times 327}$ (including 4887 known miRNA-disease associations) are obtained for GRNMF-based disease miRNA prediction.

3.2 Comparison with other methods

3.2.1 Experimental settings

To systematically evaluate GRNMF performance on the collected datasets, we perform 5-fold cross validation (CV) experiments and compare it with the following methods: RLSMDA (Chen and Yan, 2014), MIDP (Xuan *et al.*, 2015), MIDPE (Xuan *et al.*, 2015) and RWRMDA (Chen *et al.*, 2012). In each 5-fold CV repetition, for a given disease *d*, the known *d*-related miRNAs are randomly divided into five subsets of equal size; then one subset is used as the test set, and the remaining four subsets are utilized as the training set.

On the other hand, we implement other two types of CV experiments under the following scenarios as in Pahikkala *et al.* (2015) to investigate the prediction capability of our method in inferring potential associations for those novel miRNAs and diseases with no known association information: (1) CV_d: CV on diseases, where all disease interaction profiles (column vectors in matrix $Y \in R^{n\times m}$) are divided into 5-folds, 20% of columns in Y are used as the test data, while the remaining columns served as the training set in each round; (2) CV_m: CV on miRNAs, where all miRNA interaction profiles (row vectors in matrix $Y \in R^{n\times m}$) are divided into 5-folds, 20% of rows in Y are used as the test data, while the remaining rows are utilized as the training data. Note that CV_d and CV_m are mainly focused on the predictions for novel diseases and miRNAs, respectively.

In this paper, we perform CV experiments on the training dataset to estimate the parameters. All parameter combinations are considered based on grid search. The optimal combination is determined from the following values: {50, 100} for k and {2^{-2}, 2^{-1}, 2°, 2^{1}} for λ_l. Subsequently, we set $\lambda_m = \lambda_d$ and choose the two parameters from {0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}}. For WKNNP, the decay value α is chosen from {0.1, …, 0.9, 1}, and the neighborhood size K is selected from {1, 2, …, 5}. Meanwhile, we set $p = 5$ when the graphs for miRNA and disease spaces are constructed based on Cai *et al.* (2011) and Li *et al.* (2016). The Supplementary Material illustrates more detailed information about the parameters. To ensure a fair comparison, the parameters in the compared methods are set to their default values in accordance with the authors' recommendations ($\eta_M = \eta_D = 1$ and $w = 0.9$ for RLSMDA, $r_Q = 0.4$ and $r_U = 0.1$ for MIDP, $\alpha = 0.9$ and $\gamma = 0.8$ for MIDPE and $r = 0.9$ for RWRMDA).

Algorithm 1 GRNMF Algorithm

```
Input: Matrices $Y \in R^{n\times m}$, $S^m \in R^{n\times n}$ and $S^d \in R^{m\times m}$, decay term $\alpha$, sub-space dimensionality $k$, neighborhood sizes $K$ and $p$, regularization coefficients $\lambda_l$, $\lambda_m$ and $\lambda_d$.
Output: Predicted association matrix $Y^*$.
randomly initialize two non-negative matrices $W \in R^{n\times k}$ and $H \in R^{m\times k}$.
$M = \{m_1, \ldots, m_n\}, D = \{d_1, d_2, \ldots, d_m\}$;
for each miRNA $m_q \in M$ do
    $V = KNN(m_q, K, S^m)$; //$KNN(m_q, K, S^m)$ is the function to obtain the $K$ known nearest neighbors of $m_q$ in matrix $S^m$ in descending order
    for $i \leftarrow 1$ to $K$ do
        $w_i = \alpha^{i-1} S^m(m_q, m_i)$; //$m_i \in V$
    end for
    $Q_m = \sum_{i=1}^{K} S^m(m_q, m_i)$;
    $Y_m(m_q) = \sum_{i=1}^{K} w_i Y(m_i)/Q_m$;
end for
for each disease $d_p \in D$ do
    $U = KNN(d_p, K, S^d)$;
    for $j \leftarrow 1$ to $K$ do
        $w_j = \alpha^{j-1} S^d(d_p, d_j)$; // $d_j \in U$
    end for
    $Q_d = \sum_{j=1}^{K} S^d(d_p, d_j)$;
    $Y_d(d_p) = \sum_{j=1}^{K} w_j Y(d_j)/Q_d$;
end for
$Y_{md} = (a_1 Y_m + a_2 Y_d)/\sum a_i (i = 1, 2)$;
$Y = \max(Y, Y_{md})$;
construct matrices $S^m*$, $S^d*$ from $S^d$, $S^m$;
repeat
    update $W$ and $H$ by the following rules:
    $w_{ik} \leftarrow w_{ik} \frac{(YH + \lambda_m S^{m*} W)_{ik}}{(WH^T H + \lambda_l W + \lambda_m D_m W)_{ik}}$
    $h_{jk} \leftarrow h_{jk} \frac{(Y^T W + \lambda_d S^{d*} H)_{jk}}{(HW^T W + \lambda_l H + \lambda_d D_d H)_{jk}}$
until convergence
$Y^* = WH^T$;
return $Y^*$.
```

We utilize the latest version databases and recalculate the similarity of any disease pairs or miRNA pairs considering that the compared methods adopted different database versions.

3.2.2 Cross validation

To obtain a fair and convincing comparison, we test 13 common diseases associated with at least 80 verified associations under CV setting as done in Xuan *et al.* (2015). Here, the construction of new interaction profiles as introduced in Section 2.3 is related to the known miRNA-disease associations. Therefore, it is necessary to recalculate them and obtain different weighted matrix Y_{md} to update original adjacency matrix Y in each repetition of CV experiment. As shown in Figure 2A, the AUC values of GRNMF, RLSMDA, MIDP, MIDPE and RWRMDA are 0.869, 0.762, 0.825, 0.813 and 0.802, respectively. GRNMF achieves the best performance, and its average AUC values are 10.7, 4.4, 5.6 and 6.7% higher compared with the other four computational methods. Table 1 lists the AUC values of the 13 diseases, and GRNMF outperforms other methods for all the 13 diseases. Meanwhile, Supplementary Figure S1 displays the results measured through AUC within the top k candidates, and the performance of GRNMF is superior to other models. Moreover, we estimate the performance using recall, which calculates the percentage of correctly discovered true associations at different top-ranked thresholds under CV. Figure 2B and Supplementary Table S1 exhibit the comparison results, which also obtained the same result. In addition, we first check the AUC normality with regard to the selected diseases using Shapiro-Wilk test, and the QQ (quantile-quantile) plots were given in Supplementary Figure S2. We then measure the statistical significance based on the paired t-tests. Table 2 displays the P-values. The result demonstrates that GRNMF is significantly better than other methods (P-value < 0.05).

In principle, the top-ranked prediction results are more important than those obtained from the other portions. We use all the known associations as the training data, and count the number of correctly retrieved known associations under various top-ranked thresholds. Usually, the prediction model is more effective if most true associations are obtained from the top portions. As shown in Figure 3A, as expected, our proposed method is superior compared with RLSMDA, MIDP, MIDPE and RWRMDA under different thresholds. For example, among all of the 4887 known associations, GRNMF correctly predicted 75.2% (or 3673) and 92.6% (or 4524) of them at the top 50 and 100, whereas the values of the second best method (MIDP) are 73.6 and 89.8%, respectively, suggesting that GRNMF is more efficient in recovering experimentally validated associations with a lower false positive rate.

In summary, these results demonstrated the powerful ability of GRNMF in prioritizing disease-related miRNAs. This finding is reasonable because our method adopts the novel similarity measurements, which are totally independent from these known miRNA-disease associations and different from the measurements utilized by the compared approaches. In addition, our method performs a preprocessing step (WKNNP) to reconstruct the matrix Y before non-negative matrix factorization, which could supply additional association information and help to substantially improve the prediction results.

3.2.3 Performance on predicting miRNA-disease associations for novel diseases and miRNAs

The cross validation experiments under CV_d and CV_m are performed to further verify the prediction ability of GRNMF for novel diseases and miRNAs. Given that RWRMDA and MIDP could not predict miRNA candidates for those novel diseases, we only compare GRNMF with RLSMDA and MIDPE under the CV_d setting. On the other hand, the miRNA similarity measurements of all the compared methods invariably make use of the known miRNA-disease associations, which result in those novel miRNAs becoming isolated, and thus, they could not be used for prediction by the compared methods. Meanwhile, our proposed method could work for both new diseases and miRNAs. Therefore, the experiment under the CV_m setting for GRNMF is conducted to further investigate its ability to discover potential associations for novel miRNAs.

Table 2. P-values obtained through paired t-test of the AUCs of GRNMF and other compared methods for the 13 diseases

	RLSMDA	MIDP	MIDPE	RWRMDA
P-values	5.49e-07	9.41e-07	9.92e-06	6.24e-06

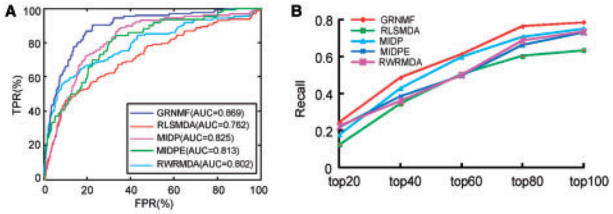

Fig. 2. (**A**) ROC curves for GRNMF and other approaches in miRNA-disease association prediction for 5-fold cross validation. (**B**) The average recalls of all the selected diseases at different top k ranking lists

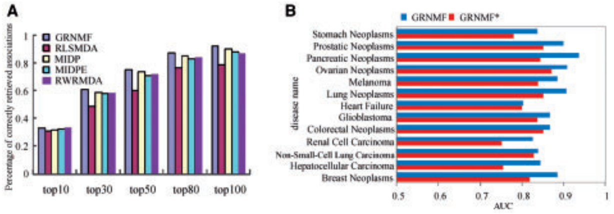

Fig. 3. (**A**) Percentage of correctly retrieved known associations between miRNAs and diseases for various ranking thresholds. (**B**) Performance comparison of selected diseases in terms of AUC values

Table 1. AUC values of GRNMF and other four compared methods for the 13 diseases

Disease name	AUC				
	GRNMF	RLSMDA	MIDP	MIDPE	RWRMDA
Breast Neoplasms	0.885	0.791	0.831	0.815	0.816
Hepatocellular Carcinoma	0.844	0.722	0.769	0.751	0.753
Non-Small-Cell Lung Carcinoma	0.839	0.806	0.798	0.837	0.832
Renal Cell Carcinoma	0.826	0.716	0.809	0.780	0.787
Colorectal Neoplasms	0.866	0.765	0.824	0.824	0.813
Glioblastoma	0.868	0.775	0.814	0.773	0.772
Heart Failure	0.803	0.671	0.780	0.762	0.765
Lung Neoplasms	0.905	0.843	0.879	0.879	0.800
Melanoma	0.885	0.799	0.845	0.809	0.814
Ovarian Neoplasms	0.909	0.742	0.870	0.879	0.876
Pancreatic Neoplasms	0.936	0.786	0.877	0.876	0.832
Prostatic Neoplasms	0.898	0.730	0.827	0.814	0.810
Stomach Neoplasms	0.837	0.762	0.804	0.767	0.757

Table 3. The average recalls for various methods under different top *k* thresholds

Method		Ranking threshold				
		Top20	Top40	Top60	Top80	Top100
CV_d	GRNMF	**22.96%**	**45.03%**	**55.83%**	**68.03%**	71.05%
	RLSMDA	18.60%	26.14%	34.28%	40.35%	47.12%
	MIDPE	21.19%	43.60%	51.86%	60.30%	**72.56%**
CV_m	GRNMF	50.06%	65.47%	75.51%	80.25%	83.75%

Supplementary Figure S3 displays the results obtained under CV_d for novel diseases and under CV_m for novel miRNAs. As shown in Supplementary Figure S3A, the average AUC values of GRNMF and MIDPE are 0.802 and 0.809, respectively. MIDPE achieves slightly better performance than GRNMF, and RLSMDA obtains the worst AUC value of 0.674. In this case, all the three methods are not as good as their own prediction performances under the CV setting. In addition, GRNMF achieves higher recall values than RLSMDA and MIDPE from the top 20 to the top 80 (Table 3), which imply that our method can identify more known associations in the top-ranked prediction portions. On the other hand, our method still achieves good performance with an average AUC value of 0.863 for those novel miRNAs under CV_m, as displayed in Supplementary Figure S3B. This could be attributed to the fact that our proposed method fully exploits the interaction profile information of other diseases with known related miRNAs to discover candidate miRNAs for a new disease, and the interaction profile information of other miRNAs with known associated diseases is integrated to predict potential associations for a new miRNA as well. The above results imply that GRNMF has a powerful ability in uncovering miRNA-disease associations for both new diseases and miRNAs.

3.3 The effects of WKNNP on performance

We also investigate the effectiveness of the preprocessing step (WKNNP) of GRNMF. The performances of the two methods (GRNMF and GRNMF*) are evaluated for the 13 diseases (as mentioned in Section 3.2) under CV. For GRNMF, we use WKNNP to update the original association matrix *Y* before implementing the graph regularized non-negative matrix factorization, which aims to add more interaction information to assist in the prediction of novel miRNAs or diseases and those miRNAs or diseases with sparse known associations. For GRNMF*, we directly perform the matrix factorization for prioritizing disease miRNAs and ignore the preprocessing step of WKNNP. The average AUC obtained by GRNMF and GRNMF* are 0.869 and 0.821 (Supplementary Fig. S4), respectively. Figure 3B illustrates the comparison of the 13 diseases, and we can find that the performance of GRNMF is superior compared with GRNMF*. For example, the AUC values achieved by GRNMF for stomach neoplasms and prostatic neoplasms are 0.837 and 0.898, respectively, whereas the AUC values obtained by GRNMF* for those diseases are 0.780 and 0.852, respectively. The comparison results indicate that the WKNNP based on the nearest neighbor information exhibits high influence on the prediction performance.

3.4 Parameter sensitivity analysis

In terms of the machine learning algorithm, the optimal parameters combination may differ from one experiment scenario to another, which makes the sensitivity analysis for parameters more complicated. In this section, we mainly focus on the subspace dimensionality *k* and the neighborhood size *K* for GRNMF and perform the experiment under the CV setting. Supplementary Table S2A and Supplementary Figure S5A show the impact of the subspace dimensionality *k* on the performance. We find that a better prediction result will be achieved when the value of *k* is larger. Furthermore, the average AUC value rapidly improves until $k = 60$ and then becomes almost stable with the increase of the dimensionality *k*. As for *K*, we vary its value from 1 to 10, and the results are shown in Supplementary Table S2B and Supplementary Figure S5B. The average AUC of GRNMF is 0.841 when *K* is set as 1, and the best performance (AUC = 0.869) is obtained when $K = 5$. This result further confirms the effectiveness of WKNNP in improving the performance.

In addition, we explore the effect of the maximum number of iterations on performance. As shown in Supplementary Table S3, the results indicate that GRNMF tends to converge within a few rounds of iterations, and very limited improvement will be achieved if the iterations are further increased. Similarly, the percentage of correctly retrieved known associations with different number of iterations for various rank thresholds also demonstrated a similar outcome (Supplementary Fig. S6).

3.5 Case studies

Case studies are conducted to further verify the capability of GRNMF to detect novel miRNA–disease associations. Here, all the known association information is used to make predictions, and the unknown associations are treated as candidate set for validation. Subsequently, the optimal parameters under the CV setting are adopted to perform the experiment for GRNMF. For each disease, the candidate miRNAs are ranked based on the prediction scores. The potential miRNAs of all diseases predicted by our method are provided in Supplementary Table S4. We use two public databases, namely, dbDEMC (Yang *et al.*, 2017) and miRCancer (Xie *et al.*, 2013), to confirm the predicted potential miRNAs for the selected disease. Supplementary Table S5 lists the top 10 predicted miRNA candidates for the three selected diseases. There are 8, 9 and 7 of 10 candidate miRNAs are confirmed to be associated with breast neoplasms, lung neoplasms and prostatic neoplasms by dbDEMC and miRCancer, respectively. In addition, some candidates also had high rankings in other methods. For instance, 7, 4 and 1 miRNAs have also been identified through MIDP within the top 10 for the three diseases, respectively (Supplementary Table S6). This finding suggests that these miRNAs are expected to be associated with the diseases. Meanwhile, some potential miRNAs are validated by the literature. For example, hsa-mir-138 overexpression affected cell proliferation in breast cancer (Denis *et al.*, 2016), and the hsa-mir-122 expression was shown to be dysregulated in lung neoplasms (Keller *et al.*, 2011). The association network of the top 20 predicted miRNA candidates for the three diseases is shown in Figure 4, in which some top-ranked candidates are observed to be related to one or more diseases.

Moreover, we divided the candidate miRNAs of each disease into two groups ('top-ranked group' vs. 'bottom-ranked group') according to their rankings, and we then use the Fisher's exact test to evaluate the statistical significance of the differences between the two groups. As shown in Figure 5, we found that 77.6% and 47.1% of miRNAs in the top-ranked group and bottom-ranked group were reported to be involved in breast neoplasms by the two aforementioned public databases, respectively. Additionally, the number of confirmed miRNAs in the top-ranked predictions was significant than those from the bottom-ranked predictions (P-value = 6.46e-05). Last but not least, the results shown that the number of confirmed predictions in the top-ranked groups of lung neoplasms (P-value = 2.88e-08) and prostatic neoplasms (P-value = 1.45e-12) were also significantly higher than the number of confirmed predictions in the bottom-ranked groups. In summary, the prediction instances further indicated that the effectiveness of GRNMF in discovering potential miRNA-disease associations.

3.6 Predicting novel miRNA-disease associations

To further demonstrate the actual potential for miRNA-disease discovery of GRNMF, we performed an additional experiment based on the older version databases, and then adopted the latest version of HMDD v2.0 as mentioned in Section 3.1 to validate those predicted potential miRNA-disease associations. We downloaded the older version of HMDD (September-2009 Version) from the supple mentary material of Wang et al. (2010), and obtained MeSH (version 2009) from its online website. After preprocessing, 1326 known associations between 228 miRNAs and 137 diseases are retained for prediction. All the predicted candidate miRNAs for 137 diseases are provided in Supplementary Table S7. Intriguingly, as shown in Supplementary Figure S7, we found that most of the top-ranked disease miRNA candidates could be directly confirmed by the latest associations in HMDD. For example, 7 out of the top 10 predicted miRNAs of breast neoplasms have been validated by HMDD (Table 4). The above observations imply that the proposed method could effectively discover those experimentally validated miRNA-disease associations in the latest version database.

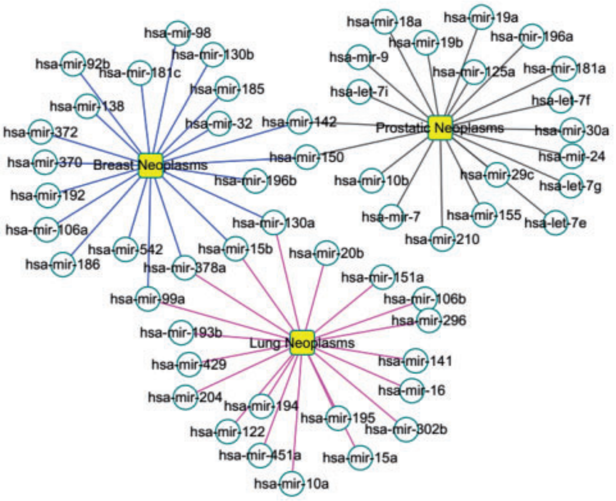

Fig. 4. Network of the top 20 predicted associations for breast neoplasms, lung neoplasms and prostatic neoplasms via GRNMF

4 Conclusions

Identifying disease-associated miRNAs contributes to decipher the underlying pathogenesis of human diseases. In this study, we have proposed a computational method, called GRNMF, for miRNA-disease association prediction. Unlike other conventional computational approaches, GRNMF could effectively discover potential associations for new diseases (or miRNAs) without any known related miRNAs (or diseases). The main contribution of our work is the development of novel similarity metrics through effective incorporation of multiple heterogeneous information and the implementation of a preprocessing step, WKNNP, to replace the elements (Y_{ij}=0) in the original miRNA–disease matrix with likelihood scores. Meanwhile, a novel integrative framework based on the graph regularized matrix factorization has been proposed to predict

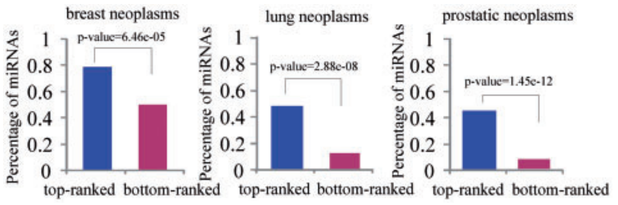

Fig. 5. Percentage of candidate miRNAs in the top-ranked groups and bottom-ranked groups that have been experimentally confirmed to be involved in the selected diseases

Table 4. The top 10 potential miRNA candidates detected by GRNMF based on the older version (2009) databases for the three selected diseases

Cancer	No. of miRNAs confirmed by the latest HMDD	Top 10 ranked predictions					
		Rank	miRNAs	Evidences	Rank	miRNAs	Evidences
Breast Neoplasms	7	1	hsa-mir-126	HMDD*	6	hsa-mir-7b	HMDD*
		2	hsa-mir-223	HMDD*	7	hsa-mir-150	miRCancer, dbDEMC
		3	hsa-mir-130a	miRCancer, dbDEMC	8	hsa-mir-181a	HMDD*
		4	hsa-mir-16	HMDD*	9	hsa-mir-106a	miRCancer, dbDEMC
		5	hsa-mir-7e	HMDD*	10	hsa-mir-101	HMDD*
Lung Neoplasms	5	1	hsa-mir-195	miRCancer	6	hsa-mir-200b	HMDD*
		2	hsa-mir-106a	dbDEMC	7	hsa-mir-107	HMDD*
		3	hsa-mir-221	HMDD*	8	hsa-mir-16	miRCancer, dbDEMC
		4	hsa-mir-92b	dbDEMC	9	hsa-mir-15b	miRCancer, dbDEMC
		5	hsa-mir-127	HMDD*	10	hsa-mir-222	HMDD*
Prostatic Neoplasms	5	1	hsa-mir-155	miRCancer, dbDEMC	6	hsa-mir-24	miRCancer, dbDEMC
		2	hsa-mir-34a	HMDD*	7	hsa-mir-29a	HMDD*
		3	hsa-mir-372	miRCancer, dbDEMC	8	hsa-mir-18a	miRCancer
		4	hsa-mir-143	HMDD*	9	hsa-mir-150	miRCancer, dbDEMC
		5	hsa-mir-15b	HMDD*	10	hsa-mir-200b	HMDD*

Note: HMDD* represent the newest version of HMDD (http://www.cuilab.cn/hmdd).

disease-associated miRNAs, which also can be easily reused and adapted in other relevant prediction problems (e.g. miRNA-gene and disease–gene relationships).

The performance of our method is validated through cross validations and case studies on the collected datasets. The experiment results indicate that GRNMF can effectively improve performance compared with other methods. Moreover, the findings of the experiments under $\mathrm{CV_d}$ and $\mathrm{CV_m}$ also demonstrate that our method is a powerful tool in uncovering potential associations for these novel diseases and miRNAs. However, there are still some limitations that require further research. First, it is a non-trivial work to determine the optimal parameter combination for different biological datasets. Second, our similarity measurement for GRNMF might not be optimal in certain circumstances. Finally, the process on how to more reasonably integrate different biological information to improve prediction performance deserves further research.

Funding

This work has been supported by the National Natural Science Foundation of China (Grant no.61572180, 61602283) and Shandong Provincial Natural Science Foundation, China (Grant no. ZR2016FB10).

Conflict of Interest: none declared.

References

Cai,D. *et al.* (2011) Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **33**, 1548–1560.

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

Chen,X. *et al.* (2016) WBSMDA: within and between score for MiRNA-disease association prediction. *Sci. Rep. UK*, **6**, 21106.

Chen,X.W. *et al.* (2013) Joint analysis of expression profiles from multiple cancers improves the identification of microRNA-gene interactions. *Bioinformatics*, **29**, 2137–2145.

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

Chou,C.H. *et al.* (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*, **44**, D239–D247.

Denis,H. *et al.* (2016) MicroRNAs regulate KDM5 histone demethylases in breast cancer cells. *Mol. Biosyst.*, **12**, 404–413.

Ding,P.J. *et al.* (2016) A path-based measurement for human miRNA functional similarities using miRNA-disease associations. *Sci. Rep. UK*, **6**, 32533.

Facchinei,F. *et al.* (2014) Solving quasi-variational inequalities via their KKT conditions. *Math. Program*, **144**, 369–412.

Guan,N.Y. *et al.* (2011) Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. *IEEE Trans. Image Process.*, **20**, 2030–2048.

Hernando,A. *et al.* (2016) A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model. *Knowl. Based Syst.*, **97**, 188–202.

Hosoda,K. *et al.* (2009) A model for learning topographically organized parts-based representations of objects in visual cortex: topographic nonnegative matrix factorization. *Neural Comput.*, **21**, 2605–2633.

Huang,D.S. and Zheng,C.H. (2006) Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics*, **22**, 1855–1862.

Jopling,C.L. *et al.* (2005) Modulation of hepatitis C virus RNA abundance by a liver-specific microRNA. *Science*, **309**, 1577–1581.

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

Lee,D. *et al.* (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

Lee,I. *et al.* (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.*, **21**, 1109–1121.

Li,X. *et al.* (2011) Prioritizing human cancer microRNAs based on genes' functional consistency between microRNA and cancer. *Nucleic Acids Res.*, **39**, e153.

Li,X. *et al.* (2016) Graph regularized non-negative low-rank matrix factorization for image clustering. *IEEE Trans. Cybern*, doi: 10.1109/TCYB.2016.2585355.

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

Li,Y. *et al.* (2014b) Mirsynergy: detecting synergistic miRNA regulatory modules by overlapping neighbourhood expansion. *Bioinformatics*, **30**, 2627–2635.

Liu,X.M. *et al.* (2014a) Progressive image denoising through hybrid graph Laplacian regularization: a unified framework. *IEEE Trans. Image Process.*, **23**, 1491–1503.

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

Liang,C. *et al.* (2015) A novel motif-discovery algorithm to identify co-regulatory motifs in large transcription factor and microRNA co-regulatory networks in human. *Bioinformatics*, **31**, 2348–2355.

Liang,C. *et al.* (2016) A novel method to detect functional microRNA regulatory modules by bicliques merging. *IEEE ACM Trans. Comput. Biol.*, **13**, 549–556.

Luo,J. *et al.* (2016a) Collective prediction of disease-associated miRNAs based on transduction learning. *IEEE/ACM Trans. Comput. Biol. Bioinform*, doi: 10.1109/TCBB.2016.2599866.

Luo,X. *et al.* (2016b) A nonnegative latent factor model for large-scale sparse matrices in recommender systems via alternating direction method. *IEEE Trans. Neural Net. Learn.*, **27**, 579–592.

Luo,J.W. *et al.* (2017) Predicting microRNA-disease associations using kronecker regularized least squares based on heterogeneous omics data. *IEEE Access*, **5**, 2503–2513.

Luo,J.W. and Xiao,Q. (2017) A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on heterogeneous network. *J. Biomed. Inf.*, **66**, 194–203.

Mørk,S. *et al.* (2014) Protein-driven inference of miRNA-disease associations. *Bioinformatics*, **30**, 392–397.

Nepusz,T. *et al.* (2012) Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods*, **9**, 471–U481.

Pahikkala,T. *et al.* (2015) Toward more realistic drug-target interaction predictions. *Brief. Bioinf.*, **16**, 325–337.

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

Vergoulis,T. *et al.* (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res.*, **40**, D222–D229.

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

Wang,J.Y. *et al.* (2012) Adaptive graph regularized nonnegative matrix factorization via feature selection. *Int. C Patt. Recog.*, 963–966.

Wang,J.Z. *et al.* (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.

Xiao,F.F. *et al.* (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.

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

Xu,J. *et al.* (2011a) Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. *Mol. Cancer Ther.*, **10**, 1857–1866.

Xu,J.A. *et al.* (2011b) MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. *Nucleic Acids Res.*, **39**, 825–836.

Xuan,P. *et al*. (2015) Prediction of potential disease-associated microRNAs based on random walk. *Bioinformatics*, **31**, 1805–1815.

Yang,Z. *et al*. (2017) dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. *Nucleic Acids Res*., **45**, D812–D818.

You,Z.H. *et al*. (2010) Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics*, **26**, 2744–2751.

Yuan,L. *et al*. (2016) Nonconvex penalty based low-rank representation and sparse regression for eQTL mapping. *IEEE/ACM Trans. Comput. Biol. Bioinf*, doi: 10.1109/TCBB.2016.2609420.

Zhao,X.M. *et al*. (2015) Identifying cancer-related microRNAs based on gene expression data. *Bioinformatics*, **31**, 1226–1234.

Zeng,X.X. *et al*. (2016) Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief Bioinf*., **17**, 193–203.

Zheng,C.H. *et al*. (2009) Tumor clustering using nonnegative matrix factorization with gene selection. *IEEE Trans. Inf. Technol. Biomed*., **13**, 599–607.

Zhu,L. *et al*. (2015) A two-stage geometric method for pruning unreliable links in protein-protein networks. *IEEE Trans. Nanobiosci*., **14**, 528–534.