Bioinformatics, 34(24), 2018, 4256–4265
doi: 10.1093/bioinformatics/bty503
Advance Access Publication Date: 22 June 2018
Original Paper

OXFORD

Systems biology

Predicting miRNA–disease association based on inductive matrix completion

Xing Chen[1,*], Lei Wang[1], Jia Qu[1], Na-Na Guan[2] and Jian-Qiang Li[2]

[1]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China and [2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on February 4, 2018; revised on June 17, 2018; editorial decision on June 18, 2018; accepted on June 20, 2018

Abstract

Motivation: It has been shown that microRNAs (miRNAs) play key roles in variety of biological processes associated with human diseases. In Consideration of the cost and complexity of biological experiments, computational methods for predicting potential associations between miRNAs and diseases would be an effective complement.

Results: This paper presents a novel model of Inductive Matrix Completion for MiRNA–Disease Association prediction (IMCMDA). The integrated miRNA similarity and disease similarity are calculated based on miRNA functional similarity, disease semantic similarity and Gaussian interaction profile kernel similarity. The main idea is to complete the missing miRNA–disease association based on the known associations and the integrated miRNA similarity and disease similarity. IMCMDA achieves AUC of 0.8034 based on leave-one-out-cross-validation and improved previous models. In addition, IMCMDA was applied to five common human diseases in three types of case studies. In the first type, respectively, 42, 44, 45 out of top 50 predicted miRNAs of Colon Neoplasms, Kidney Neoplasms, Lymphoma were confirmed by experimental reports. In the second type of case study for new diseases without any known miRNAs, we chose Breast Neoplasms as the test example by hiding the association information between the miRNAs and Breast Neoplasms. As a result, 50 out of top 50 predicted Breast Neoplasms-related miRNAs are verified. In the third type of case study, IMCMDA was tested on HMDD V1.0 to assess the robustness of IMCMDA, 49 out of top 50 predicted Esophageal Neoplasms-related miRNAs are verified.

Availability and implementation: The code and dataset of IMCMDA are freely available at https://github.com/IMCMDAsourcecode/IMCMDA.

Contact: xingchen@amss.ac.cn

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

1 Introduction

MicroRNAs (miRNAs) are a category of small non-coding RNAs (~22 nt) which could function in the post-transcriptional regulation of gene expression through binding to the 3′-UTRs of the target mRNAs (Bartel, 2004; Meister and Tuschl, 2004; Victor, 2001, 2004). The first miRNA lin-4 was found in the early 1990s in *Caenorhabditis elegans* by Lee *et al.* (1993), since then, thousands of currently annotated miRNAs have been discovered in variety of species from plants, animals to viruses (Jopling *et al.*, 2005; Kozomara and Griffiths-Jones, 2011). The latest version of miRBase collects 28 645 entries among 223 species (1881 human miRNAs) (Kozomara and Griffiths-Jones, 2014). Furthermore, more and more studies have pointed out that miRNAs could influence multiple stages of the biological processes (Lee *et al.*, 1993), including early cell growth, proliferation (Cheng, 2005), differentiation (Miska, 2005), development (Karp and Ambros, 2005), aging (Bartel, 2009), apoptosis, viral infection (Miska, 2005) and so on. In addition, miRNAs are suitable to be drug targets as they have

several attractive features, such as specific secondary structures and conserved sequences (Chen *et al.*, 2017a,d). Therefore, it is obvious that miRNAs have critical impact on the human diseases. For example, Mayr *et al.* (2007) demonstrated that a chromosomal translocation at 12q5 influences the expression of let-7 and consequently curtails repress of the oncogene High Mobility Group A2 (Hmga2). Another example is that the expression of let-7b was proved not only to be related to luminal tumors but also to show an independent crucial positive prognostic value according to experiments (Quesne *et al.*, 2012). Zhang *et al.* (2011) pointed that hsa-miR-31 could be chosen as a potential biomarker of Esophageal Squamous Cell Carcinoma (ESCC) diagnosis and prevention. Hsa-miR-31 was indicated to be up-regulated in most of the ESCC tissues and the vitro experiments showed that miR-31 stimulated the ESCC colony formation, migration and invasion. Besides, the overexpression of miR-124 can effectively inhibit the invasion ability of glioma cell in vitro (Xia *et al.*, 2012). Identification of disease-related miRNAs would contribute to the research of disease pathological mechanisms and identification of disease biomarkers (Chen *et al.*, 2017e; Goh *et al.*, 2016). Computational models are developed to search the most potential disease-related miRNAs for further biological experiments, decrease the time and money for miRNA–disease association identification, and therefore reduce the difficult of disease biomarkers detection (Calin and Croce, 2006).

Over the past few year, a plenty of models have been developed for miRNA–disease associations prediction (Xuan *et al.*, 2015; You *et al.*, 2017; Zeng *et al.*, 2016; Zou *et al.*, 2016). For example, Jiang *et al.* (2010) proposed a novel computational method to predict potential miRNA–disease associations by applying a scoring system to the miRNA functional similarity network and human phenome microRNAome network to assess the probability that a miRNA may be involved in a specific disease. However, this model did not achieve a satisfactory prediction accuracy, because only the information of miRNA neighbors has been adopted. Shi *et al.* (2013) proposed a computational model by considering the functional associations between miRNAs targets and diseases genes in protein–protein interaction network. The miRNA targets and disease genes were used as seeds for implementing random walk on the protein–protein interaction network to compute the *P*-value and assess the potential association between the miRNA and disease. If the *P*-value exceeded the threshold, the corresponding miRNA and disease were considered to have a link. Pasquier and Gardès (2016) presented a model of MiRAI to identify potential miRNA–disease associations. For each miRNA, MiRAI exploited five critical information: its known related diseases, its target mRNAs, its family members, the distance to its neighbors, and the abstracts of related studies in text format to construct a high-dimensional vector space. Furthermore, diseases and miRNA were represented by vector in the vector space. After dimensionality reduction, MiRAI can obtain a ranked list of the miRNAs that are related to disease *d* by computing their distance to the vector of disease *d*. The limitation of MiRAI is that MiRAI cannot be applied for a sparse database. Mork *et al.* (2014) devised a novel model of miRNA–Protein–Disease (miRPD) to uncover potential miRNA–disease associations in consideration of not only the experimental and computationally predicted miRNA–protein associations but also the text mined protein–disease associations. However, these methods could not provide satisfactory prediction results because of the high rate of false positive and false negative samples in the miRNA–target interactions. The first global network similarity-based model of Random Walk with Restart for MiRNA–Disease Association (RWRMDA) was proposed by Chen *et al.* (2012). First, RWRMDA assigned an initial probability for each miRNA in the miRNA functional similarity network (MFSN). Then, a random walk algorithm was introduced to MFSN until the probability of each miRNA get stable. The stable probability of miRNA is used to assess the potential association between the miRNA and given disease. RWRMDA has shown a superior performance to previous local network-based methods. However, RWRMDA is not suitable for those new diseases without any known related miRNAs. Yu *et al.* (2017) has proposed a method by modifying the existing maximizing information flow (Maxflow) to infer novel miRNA–disease associations by exploiting multiple sources of information including the miRNA functional similarity network, the disease semantic and phenotypic similarity network and the miRNA–disease association network and combining them to form a directed miRNAome-phenome network graph.

Xuan *et al.* (2013) developed a computational model of Human Disease-related MiRNA Prediction (HDMP) by considering the *k* most similar neighbors of each miRNA. The *k* nearest neighbors of each miRNA and miRNA functional similarity were combined to estimate more reliable relevance scores of the unlabeled miRNAs. Besides, HDMP assigned higher weight to the miRNAs in the same miRNA family or cluster. However, HDMP cannot be implemented to infer the potential related miRNAs for those diseases with few known related miRNAs or without any known related miRNAs. Chen *et al.* (2017b) proposed a model of Ranking-based K-Nearest Neighbors for miRNA–Disease Association prediction (RKNNMDA) by searching the *k*-nearest-neighbors both for miRNAs and diseases. After using the SVM ranking model to re-rank these k-nearest-neighbors, they can obtain the final ranking of all miRNA–disease pairs through weighted voting. Liu *et al.* (2016) developed a model to predict miRNA–disease association by introducing random walk to the heterogeneous network which was constructed by multiple data sources (disease semantic similarity, Disease functional similarity, Gene similarity, miRNA–target gene associations, miRNA–lncRNA associations, lncRNA similarity). Zeng *et al.* (2018) applied structural perturbation method (SPM) on the miRNA–disease bilayer network to predict potential miRNA–disease associations. Zou *et al.* (2015) introduced two computational methods of KATZ and CATAPULT to make prediction for miRNA–disease pairs based on social network analysis methods. Chen *et al.* (2016a) released a new reliable model named Within and Between Score for MiRNA–Disease Association prediction (WBSMDA). The Within-Score was defined to capture the similarities between disease-related miRNAs and the similarities between miRNA-related diseases, and Between-Score was defined to capture the similarities between disease-uncorrelated miRNAs and the similarities between miRNA-uncorrelated diseases. Then WBSMDA integrated the Within-Score and Between-Score to calculate the miRNA–disease association prediction score. Recently, another model Heterogeneous Graph Inference for MiRNA–Disease Association prediction (HGIMDA) was proposed by Chen *et al.* (2016b). In HGIMDA, the integrated similarity networks and the known miRNA–disease association network were combined to generate a heterogeneous graph, in which an iterative equation was used for the prediction for potential miRNA–disease associations. HGIMDA has a superior performance to previous methods, but the selection of the parameter is still not well solved.

Machine learning has been applied in extensive scientific fields, and it is highly effective to solve most of the research problems (Chen *et al.*, 2016c, 2017c). For example, Xu *et al.* (2011) proposed a computational method by integrating miRNA–target interactions and expression levels of miRNAs and mRNAs. In addition, they constructed a support vector machine (SVM) classifier based on

feature vectors including four features (the number of dysregulated target genes, the number of its coregulators, the proportion of prostate cancer miRNAs in its coregulator set, the fraction of targets which are coregulated by itself and other prostate cancer miRNAs). However, it's almost impossible to get the negative samples in this filed. So negative sample sets were usually artificially constructed by randomly pairing miRNAs and diseases and then removing the pairs existed in the positive sample sets. Such negative samples are not real negative samples, some of which may be positive samples which are not experimentally validated, and therefore decreases the prediction accuracy of this SVM-based model. Chen and Yan (2015) presented a semi-supervised learning-based model of Regularized Least Squares for MiRNA–Disease Association (RLSMDA) based on semi-supervised learning framework so that negative samples were not required in this model. The drawback of RLSMDA is how to choose the appropriate parameters and way to combine the classifiers trained from the miRNA space and disease space. Chen *et al.* (2015) also released a model of Restricted Boltzmann Machine for Multiple types of MiRNA–Disease Association prediction (RBMMMDA) which was based on Restricted Boltzmann Machine (RBM). The RBMMMDA was not only a model that could uncover novel miRNA–disease associations but also the first model which could estimate the corresponding types of miRNA–disease associations. Li *et al.* (2017) released an effective computational model of Matrix Completion for MiRNA–Disease Association prediction (MCMDA). A matrix completion algorithm of high efficiency is adopted in MCMDA, which updated the low-rank miRNA–disease association matrix. But it cannot be applied to predicting the potential miRNAs associated with the new diseases without any known related miRNAs and potential diseases associated with new miRNAs without any known related diseases. The optimal parameters of MCMDA is still hard to choose.

In this study, we proposed a novel computational model of Inductive Matrix Completion for MiRNA–Disease Association prediction (IMCMDA). This model exploited not only the known miRNA–disease associations but also the integrated similarity for miRNA and disease. To evaluate the effectiveness of IMCMDA, Leave-one-out cross validation (LOOCV) was carried on the known miRNA–disease association data downloaded from HMDD V2.0 (Li *et al.*, 2014). Furthermore, three types of case studies were carried on five high-risk human diseases. In the first type of case study, three diseases (Colon Neoplasms, Kidney Neoplasms and Lymphoma) were used to evaluate the prediction ability of implementing IMCMDA on the data collected from HMDD V2.0. All the candidate miRNAs of these three diseases were ranked according to their prediction score, respectively. Then the top 50 predicted miRNAs of these three diseases were examined in dbDEMC (Yang *et al.*, 2010) and miR2Disease (Jiang *et al.*, 2009). As a result, 42, 44 and 45 out of the top 50 potential related miRNAs of Colon Neoplasms, Kidney Neoplasms, Lymphoma were respectively confirmed by recent experimental discoveries. In the second case study, Breast Neoplasms was selected to evaluate the performance of IMCMDA for the new disease without any known related miRNAs. Here, we artificially removed all the experimentally validated Breast Neoplasms-related miRNAs so that Breast Neoplasms could be considered as a new disease. As a result, 50 out of top 50 predicted Breast Neoplasms-related miRNAs were included in one of the three databases HMDD V2.0, dbDEMC and miR2Disease. In the third Case, Esophageal Neoplasms was chosen as the test example to demonstrate the robustness of IMCMDA by testing the model on the database HMDD V1.0. Finally, 49 out of the top 50 predicted potential Esophageal Neoplasms-related miRNAs were included in one of the three databases: dbDEMC, miR2Disease and HMDD V2.0. In conclusion, the model of IMCMDA with a reliable performance could be help for miRNA–disease association prediction.

2 Materials and methods

2.1 Human miRNA–disease associations

The data of known human miRNA–disease associations, which we used in this article were retrieved from the HMDD V2.0 database (June, 2014). After sorting and standardizing the downloaded data, we obtained 5430 experimentally verified human miRNA–disease associations between 383 diseases and 495 miRNAs. An $nd \times nm$ adjacency matrix A was defined as:

$$\begin{cases} A(d(i),\mathrm{m}(j))=1 \ \textit{diased}(i)\textit{has association with miRNA}m(j) \\ A(d(i),m(j))=0 \ \textit{diased}(i)\textit{has no association with miRNA}m(j) \end{cases} \tag{1}$$

2.2. miRNA functional similarity

The miRNA functional similarity was calculated based on a basic assumption that functionally similar miRNAs tend to connect with similar diseases, and vice versa (Goh *et al.*, 2007; Lu *et al.*, 2008). Thanks to the excellent work of Wang *et al.* (2010), we can download the miRNA functional similarity data from http://www.cuilab.cn/files/images/cuilab/misim.zip. With these data, we constructed a$nm \times nm$matrix FS to represent the miRNA functional similarity. The element $FS(\mathrm{m}(i), \mathrm{m}(j))$ denotes the functional similarity between miRNA $m(i)$ and $m(j)$.

3 Methods

3.1 Disease semantic similarity model 1

A Directed Acyclic Graph (DAG) was constructed to describe a disease based on the MeSH descriptors downloaded from the National Library of Medicine (Lipscomb, 2000). The DAG of disease D included not only the ancestor nodes of D and D itself but also the direct edges from parent nodes to child nodes.

The semantic score of disease D could be defined by the following equation:

$$DV1(D) = \sum_{d \in T(D)} D_D1(d) \tag{2}$$

where we defined the contribution score of disease d in DAG(D) to the disease D by:

$$\begin{cases} D_D1(d) = 1 & \textit{if } d = D \\ D_D1(d) = \max\{\Delta * D_D1(d') | d' \in \textit{children of } d\} & \textit{if } d \neq D \end{cases} \tag{3}$$

Here, Δ is the semantic contribution factor. The contribution score of disease is decreased as the distance between D and other diseases increases.

Based on the assumption that two diseases with larger shared area of their DAGs may have greater similarity score, the semantic similarity score between disease $d(i)$ and disease $d(j)$ could be defined by the following equation:

$$\mathrm{SS1}(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} (D_{d(i)}1(t) + D_{d(j)}1(t))}{DV1(d(i)) + DV1(d(j))} \tag{4}$$

3.2 Disease semantic similarity model 2

Supposing for two diseases in the same layer, if one disease appears in less disease DAGs than the other disease, obviously we can conclude that the first disease would have a greater contribution to the semantic value of disease D than the second disease. In conclusion, different disease terms in the same layer of DAG(D) may have the different contribution to the semantic value of disease D. Considering about the above factor, we use a new model to describe the contribution of a disease d in DAG(D) to the semantic value of disease D:

$$D_D2(d) = -\log\left[\frac{\textit{the number of DAGs including } d}{\textit{the number of disease}}\right] \tag{5}$$

The semantic value of disease D is defined as follows:

$$DV2(D) = \sum_{d \in T(D)} D_D2(d) \tag{6}$$

Based on the assumption that two diseases with larger shared area of their DAGs may have higher similarity score, we defined the semantic similarity score between disease $d(i)$ and disease $d(j)$ as following:

$$\mathrm{SS2}(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} \left(D_{d(i)}2(t) + D_{d(j)}2(t)\right)}{DV2(d(i)) + DV2(d(j))} \tag{7}$$

3.3 Gaussian interaction profile kernel similarity for diseases and miRNAs

Based on the basic assumption that similar diseases tend to be associated with functionally similar miRNAs and vice versa (Bandyopadhyay, 2010; Goh *et al.*, 2007; Lu *et al.*, 2008; Wang *et al.*, 2010), we calculated Gaussian interaction profile kernel similarity to represent the miRNA similarity and disease similarity. Firstly, we used vector $IP(d(i))$ to represent the interaction profile of disease $d(i)$ by observing whether there is known association between disease $d(i)$ and each miRNA or not. Then, Gaussian interaction kernel similarity between disease $d(i)$ and $d(j)$ was calculated as follows.

$$\mathrm{KD}(d(i), d(j)) = \exp\left(-\gamma_d \|IP(d(i)) - IP(d(j))\|^2\right) \tag{8}$$

where, γ_d is used to control kernel bandwidth which is obtained by normalizing a new bandwidth parameter γ'_d by the average number of associations with miRNAs for all the diseases. γ_d is defined as follows:

$$\gamma_d = \gamma'_d \Big/ \left(\frac{1}{nd}\sum_{i=1}^{nd} \|IP(d(i))\|^2\right) \tag{9}$$

Similarly, Gaussian interaction profile kernel similarity between miRNA $m(i)$ and $m(j)$ is defined in a similar way:

$$\mathrm{KM}(m(i), m(j)) = \exp\left(-\gamma_m \|IP(m(i)) - IP(m(j))\|^2\right) \tag{10}$$

$$\gamma_{\mathrm{m}} = \gamma'_m \Big/ \left(\frac{1}{nm}\sum_{i=1}^{nm} \|IP(m(i))\|^2\right) \tag{11}$$

3.4 Integrated similarity for diseases and miRNAs

In fact, we could not get DAGs for all diseases. That is to say, for the specific disease without DAG, we could not calculate the semantic similarity score between the disease and other diseases. Therefore, for those disease pairs with semantic similarity score, we used the semantic similarity score to denote the disease similarity, for the others, the Gaussian interaction profile kernel similarity score was used to denote the disease similarity. The disease similarity matrix between disease $d(i)$ and disease $d(j)$ is constructed as follows:

$$S_d(d(i), d(j)) = \begin{cases} \dfrac{SS1(d(i), d(j)) + SS2(d(i), d(j))}{2} & d(i) \textit{ and } d(j) \textit{ has semantic similarity} \\ KD(d(i), d(j)) & \textit{otherwise} \end{cases} \tag{12}$$

Similarly, the new similarity matrix between miRNA $m(i)$ and $m(j)$ is defined as follows:

$$S_m(m(i), m(j)) = \begin{cases} FS(m(i), m(j)) & m(i) \textit{ and } m(j) \textit{ has functional similarity} \\ KM(m(i), m(j)) & \textit{otherwise} \end{cases} \tag{13}$$

3.5 IMCMDA

In this paper, we presented a novel matrix completion-based model named IMCMDA for miRNA–disease associations prediction. This model of IMCMDA was implemented based on the known miRNA–disease associations, disease semantic similarity, miRNA functional similarity, Gaussian interaction profile kernel similarity for miRNAs and diseases. The specific implementation process of IMCMDA is shown in Figure 1.

After data collection and similarity calculation, we can obtain the human miRNA–disease association matrix $A \in R^{nd \times nm}$, disease similarity matrix $S_d \in R^{nd \times nd}$, miRNA similarity matrix $S_m \in R^{nm \times nm}$. Obviously, adjacency matrix A is a very sparse matrix (Matrix density is 0.028) because only 5430 experimentally verified human miRNA–disease associations were collected between 383

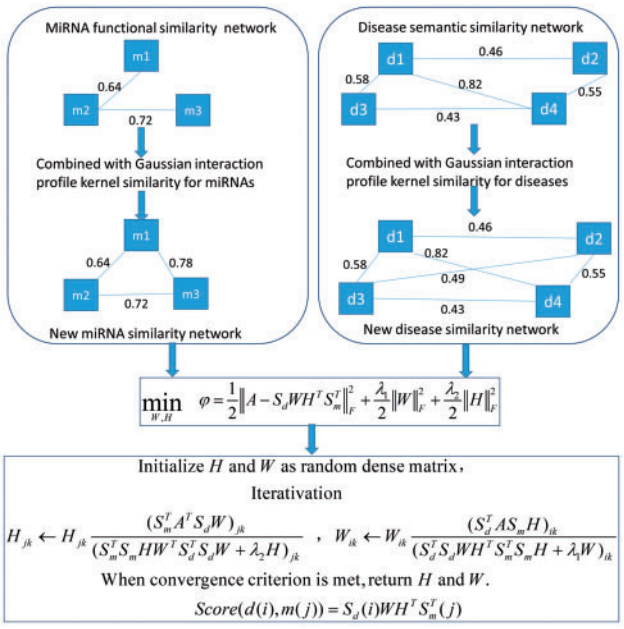

Fig. 1. Flowchart of IMCMDA model to infer the potential miRNA–disease associations

diseases and 495 miRNAs. Our goal is to complete the missing entries of A. Here, we chose disease similarity matrix $S_d \in R^{nd\times nd}$ and miRNA similarity matrix $S_m \in R^{nm\times nm}$ as the feature matrix of nd diseases and nm miRNAs, respectively, $S_d(i)$ denote the feature vector of disease $d(i)$, and $S_m(j)$ denote the feature vector of miRNA $m(j)$. The main idea of IMC is to recover a matrix $Z \in R^{nd\times nm}$ using the known entries from the miRNA–disease associations matrix A, the form of Z is $Z = WH^T$, where $W \in R^{nd\times r}$ and $H \in R^{nm\times r}$, r is the desired rank which is equal to min(rank(W), rank(H)). The parameter r will affect the convergence speed of the inductive matrix completion algorithm, but the impact on the result is very small. The element $Score(d(i), m(j))$ is calculated to denote the predicted association possibility between disease $d(i)$ and miRNA $m(j)$. The matrices W and H can be obtained as the solutions to the following optimization problem:

$$\min_{W,H} \quad \varphi = \frac{1}{2}\left\|A - S_d W H^T S_m^T\right\|_F^2 + \frac{\lambda_1}{2}\left\|W\right\|_F^2 + \frac{\lambda_2}{2}\left\|H\right\|_F^2 \qquad (14)$$
$$such\ that, \qquad W \geq 0, H \geq 0.$$

where λ_1, λ_2 is the regularization parameters and usually we set $\lambda_1 = \lambda_2 = 1$ $\|\bullet\|_F$ is the Frobenius norm of matrix. $\frac{1}{2}\|A - S_d W H^T S_m^T\|_F^2$ is the least square cost function, $\frac{\lambda_1}{2}\|W\|_F^2$ and $\frac{\lambda_2}{2}\|H\|_F^2$ are set to overcome over-fitting problem, we can solved the minimum problem with a method proposed by Jain and Dhillon (2013). Firstly, we set W and H as random dense matrix, then we updated W and H using an iterative equation, the iterative process should stop when the convergence criterion is met, usually we set the convergence criterion as 10^{-6}. The detail algorithm steps to solve the minimum problem is given in Figure 1 with an iterative equation. We can use W and H to calculate the prediction score between disease $d(i)$ and miRNA $m(j)$ by the following equation

$$Score(d(i), m(j)) = S_d(i) W H^T S_m^T(j) \qquad (15)$$

If we have a new disease $newd(i)$ without any known related miRNAs, the entry $Score(newd(i), j)$ still can be computed for all miRNAs as long as we have the feature vector of disease $newd(i)$.

4 Results

4.1 Performance evaluation

To evaluate the prediction accuracy of IMCMDA, we implemented LOOCV frameworks on the known miRNA–disease associations in the following way: For the disease $d(i)$, each known miRNA–disease pair (take miRNA–disease pair ($m(j)$-$d(i)$) as an example) was selected in turn as test sample, while all the other known miRNA–disease pairs were considered as training samples. Firstly, we artificially changed the known miRNA–disease pair ($m(j)$-$d(i)$) into an unverified miRNA–disease pair. The unverified miRNAs-disease pairs of disease $d(i)$ were considered as candidate samples, and then we ranked the predicted score of the test miRNA–disease pair ($m(j)$-$d(i)$) with the candidate samples. If the rank of the test miRNA–disease pair ($m(j)$-$d(i)$) exceeded the given threshold, the model could be considered to be successful in predicting the miRNA–disease pair ($m(j)$-$d(i)$). We have compared our method with RLSMDA, HDMP, RWRMDA, MCMDA, MiRAI and Maxflow based on the framework of LOOCV. The known miRNA–disease association dataset used for this comparison was the same, i.e. the 5430 known associations between 495 miRNAs and 383 diseases in the HMDD v2.0 database. As for other input datasets required by these six methods, we either downloaded the corresponding data from the supplementary files in the methods' literatures or collected the data from the sources specified in the literatures.

To validate the performance of our method, we compared it with a number of baseline methods. The details of baselines were provided as follows: **Maxflow** (Yu *et al.*, 2017): exploited multiple sources of information including the miRNA functional similarity network, the disease semantic and phenotypic similarity network and the miRNA–disease association network. Subsequently, these three networks were further combined to form a directed miRNAome-phenome network graph (the parameters we used for comparison are $\alpha = 0.1$, $\beta = 0.6$, $\gamma = 100$, $\eta = 6$, $\sigma = 10$). **HDMP** (Xuan *et al.*, 2013): The k nearest neighbors of each miRNA and miRNA functional similarity were combined to estimate more reliable relevance scores of the unlabeled miRNAs. Besides, HDMP assigned different weight to the miRNAs based on miRNA family or cluster (the parameters we used for comparison are $\alpha = 4$, $\beta = 4$, $k = 20$). **MCMDA** (Li *et al.*, 2017): MCMDA introduced the matrix completion algorithm on the known miRNA–disease association matrix A to predict potential miRNA–disease associations (the parameters we used for comparison are $\varepsilon = 10^{-4}$, max_iter = 500). **RLSMDA** (Chen and Yan, 2015): the method has combined two classifiers trained from the miRNA space and the disease space respectively based on the framework of regularized least squares algorithm (the parameters we used for comparison are $\eta_M = 1$, $\eta_D = 1$, $\omega = 0.9$). **RWRMDA** (Chen *et al.*, 2012): Chen has introduced random walk on the miRNA functional similarity network (MFSN) to predict potential miRNAs for disease (the parameters we used for comparison are $r = 0.2$, threshold $= 10^{-6}$). **MiRAI** (Pasquier and Gardès, 2016): This model exploited multiple sources of information including miRNA-neighbor associations MN, miRNA–target associations MT, miRNA–word associations MW, miRNA–family associations matrix MF, the miRNA–disease association MD (the parameter we used for comparison is $r = 400$).

A Receiver Operating Characteristics (ROC) curve is plotted by using the result of LOOCV. The X-axis of the ROC graph is the true positive rate (TPR) while the Y-axis is the false positive rate (FPR). The ROC curve based on LOOCV have been plotted in Figure 2. From the ROC curve, Area under curve (AUC) could be calculated as an evaluation metric for the model. As a result, IMCMDA, MCMDA, RWRMDA, Maxflow, HDMP, RLSMDA and MiRAI had obtained AUCs of 0.8034, 0.7718, 0.7891, 0.7774, 0.7702, 0.6953 and 0.6229 in the LOOCV, respectively. The core of MiRAI is collaborative filtering so that this method cannot be well applied for a sparse database, our training database is very sparse, that's why MiRAI did not have satisfying AUCs as shown in their studies (Pasquier and Gardès, 2016). Therefore, in comparison with the previous models, we can intuitively observe the improvement in predicting the miRNA–disease associations with IMCMDA.

Furthermore, Precision-Recall (PR) curve is plotted by using the result of LOOCV in Figure 3. As showed in the PR curve, obviously, our method outperformed MCMDA, Maxflow, RLSMDA, RWRMDA and MiRAI, but underperformed HDMP. However, a great drawback of model HDMP is that HDMP cannot be used to predict miRNAs for new diseases, conversely, IMCMDA is performed well in this context.

Especially, the differences of inference capability of these algorithms were further analyzed by paired t-test. The paired t-test was performed on the result of LOOCV. We can observe the significance difference between IMCMDA and previous models (MCMDA, HDMP, MiRAI, RLSMDA, RWRMDA, Maxflow), with the P-value of 2.14E-15, 5.67E-27, 2.20E-233, 2.07E-195, 9.01E-10, 9.85E-19, respectively.

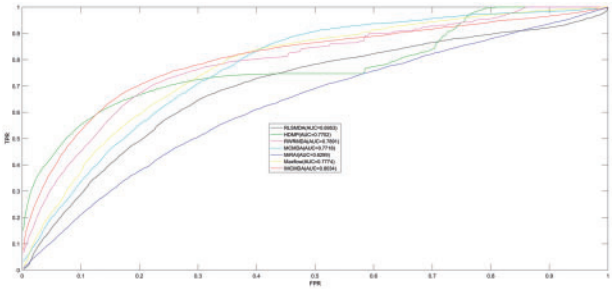

Fig. 2. Performance comparisons between IMCMDA and baseline methods (RLSMDA, HDMP, RWRMDA, MCMDA, MiRAI, Maxflow) in terms of AUC based on LOOCV. As a result, IMCMDA achieved an AUC of 0.8034, outperforming the previous models

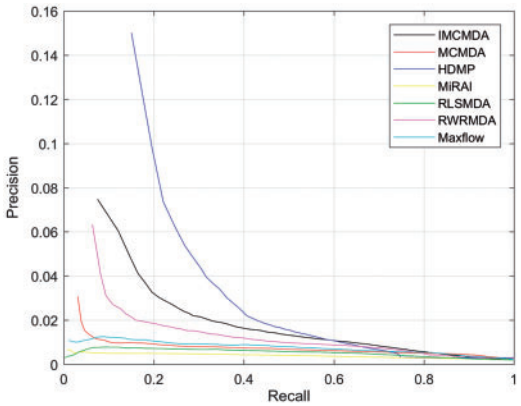

Fig. 3. Performance comparisons between IMCMDA and baseline models (RLSMDA, HDMP, RWRMDA, MCMDA, MiRAI, Maxflow) in terms of PR curve.

4.2 Performance of models on new disease

An important criterion for evaluating the usefulness of the model is whether the model can be used to predict potential related miRNA for new disease. We adopted the cumulative distribution of the ranks as a measure for comparing the performances of different models for predicting potential related miRNA for new disease. Several articles (Fantine and Jean-Philippe, 2011; Natarajan and Dhillon, 2014; Singh-Blom *et al.*, 2013) have adopted this performance measure for evaluation. The motivation for using this performance measure is to distinguish methods based on the probability of recovering a true association in the top-*r* predictions for a new disease. Step 1, we chose a test disease like Breast Neoplasms in the second type of case study by setting all the known associations between miRNAs and test disease as unknown ones. Step 2, we implemented different models on this test set to obtain the ranks of miRNAs which were actually related to the test disease. Step 3, we chose all 383 diseases as test diseases to repeat steps 1 and 2 in turn. After that, we can plot the cumulative distribution function (CDF) in Figure 4, *x*-axis denotes the top-*r* predicted miRNAs, y-axis denotes the probability of recovering a true association in the top-*r* predictions. Since the models of MCMDA, RWRMDA, HDMP cannot be applied for new diseases, we compared the rest three models of IMCMDA, MiRAI and RLSMDA. We can see a significant increase in the performance of IMCMDA around top 50–100 predictions in Figure 4 with almost 60% chance to recovering a true association in the top-100 predictions. We can conclude that IMCMDA is more effective for predicting miRNAs for new disease.

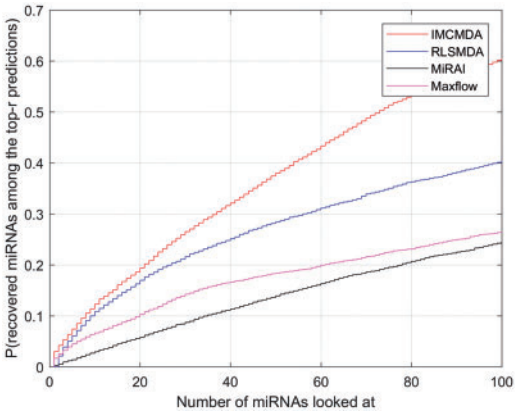

Fig. 4. Performance comparisons between IMCMDA and baseline models (RLSMDA, MiRAI, Maxflow) in predicting potential miRNAs for new diseases

4.3 Permutation test to assess contributions of different data sources

We have performed permutation test to assess the contributions of miRNA and disease similarity matrix and miRNA–disease association matrix for the improvement of performance, respectively. The main idea of permutation test is to randomize one of the three matrix and keep the other two matrices unchanged based on the LOOCV framework. For each type of data matrix, we will randomize it 50 times and calculate the median of the resulting AUCs for permutation tests. If a particular data type contributes more for the performance of model, then the result of permutation test based on this data will be more similar to a random prediction. As shown in Table 1, the average AUC of our model based on randomized miRNA–disease association matrix is much lower than another two types permutation, which indicates that the contribution of miRNA–disease association is the most important. Furthermore, we can conclude that the miRNA similarity matrix is more important than disease similarity matrix for the improvement of performance.

4.4 Case studies

Three different types of case studies were implemented to demonstrate the accuracy of IMCMDA for novel miRNA–disease association prediction. All of them had shown excellent results. The first case study included three common human diseases (Colon Neoplasms, Kidney Neoplasms and Lymphoma). The predicted miRNAs of those diseases were examined according to two databases: dbDEMC and miR2Disease. Through the case study, we can further validate the effectiveness of the IMCMDA. And then we observed the number of the verified miRNAs in the top 10 and top 50 ones which are related to the three diseases respectively according to the two databases.

Colon Neoplasms is a most common malignancy in the gastrointestinal tract (Jemal *et al.*, 2011; Ogata-Kawata *et al.*, 2014). In 2018, there are 97 220 estimated new cases and 50 630 estimated deaths from Colon Neoplasms in U.S. (Siegel *et al.*, 2018). Several Colon Neoplasms-related miRNAs have been confirmed by recent biological experiments. For example, the expression level of miR-106a in Colon Neoplasms line is lower than in normal human colon epithelia (Díaz *et al.*, 2008). It also has been demonstrated that mir-145 could down-regulates the IRS-1 protein in Colon Neoplasms cells and thereby inhibit the growth of Colon Neoplasms cells through targeting the IRS-1 3′-untranslated region (UTR) (Shi *et al.*, 2007). In this case study, IMCMDA was implemented to predict

potential Colon Neoplasms-related miRNAs. As a result, 10 out of the top 10 and 42 out of the top 50 predicted Colon Neoplasms-related miRNAs were included by either dbDEMC or miR2Disease (see Table 2).

Kidney Neoplasms is one of the common human genitourinary malignancies which accounts for 3% of adult malignancies, with more than 250 000 new cases of Kidney Neoplasms diagnosed every year (Jemal *et al.*, 2006; Seigel *et al.*, 2012). Some miRNAs could be help for the treatment of Kidney Neoplasms. For example, the expression of miR-141 is significantly lower in Kidney Neoplasms cells when compared to normal kidney cells (Senanayake *et al.*, 2012). Furthermore, Proline oxidase (POX) is known as a tumor suppressor that can effectively inhibit cell proliferation and induce cell apoptosis and recent study shows that the expression of miR-23b and POX protein is negatively correlated. So down-regulation of miR-23b could be a considerable way to inhibit Kidney Neoplasms cell growth (Liu *et al.*, 2010). We introduced IMCMDA to uncover the potential Kidney Neoplasm-related miRNAs. As a result, 9 out of the top 10 candidates and 44 out of the top 50 candidates of Kidney Neoplasms-related miRNAs were confirmed by either dbDEMC or miR2Disease (see Supplementary Table S1).

Table 1. Permutation test was implemented 50 times to assess the contributions of miRNA and disease similarity matrix and miRNA–disease association matrix, respectively

Randomized Data sources	Average value of AUCs
miRNA–disease association matrix	0.4988 ± 0.005639
miRNA similarity matrix	0.6497 ± 0.009217
Disease similarity matrix	0.7985 ± 0.000291

Note: The corresponding average values are shown in the second columns.

Lymphoma is a malignant tumor originating in the lymphoid hematopoietic system (Wan and Tian, 2014). Lymphoma is divided into non-Hodgkin's Lymphoma (NHL) and Hodgkin's lymphoma (HL) according to tumor cells (Harrison, 2013). About 90 percent of people with Lymphoma would be non-Hodgkin's Lymphoma (McDuffie *et al.*, 2009). Researchers have pointed that the deletion or downregulation of mir-15a results in increased expression of the antiapoptotic B cell Lymphoma 2 (BCL2), and overexpression of BCL2 protein has been reported in many cases of Lymphoma (Cimmino *et al.*, 2005). Lymphoma was chosen as the third case studies. As a result, 9 out of the top 10 candidates and 45 out of the top 50 predicted Lymphoma-related miRNAs were included by either dbDEMC or miR2Disease (see Supplementary Table S2).

In the second type of case study, we want to evaluate the performance when IMCMDA was implemented to the new disease without any known related miRNAs. Breast Neoplasms was used as an example in our experiment. Therefore, we hid the association information between the miRNAs and the Breast Neoplasms by setting all the known associations between them as unknown ones. Then, we implemented IMCMDA to obtain the ranking list of the miRNA-Breast Neoplasms association prediction scores. We show the result of Breast Neoplasms in Table 3. We can conclude that 10 out of the top 10 and 50 out of the top 50 predicted miRNAs were confirmed by at least one of the three databases HMDD V2.0, dbDEMC and miR2Disease. For example, hsa-mir-21 was ranked first and recent research has confirmed that hsa-mir-21 is strongly expressed in numerous cancers like Breast Neoplasms, Glioblastoma and Pancreatic Neoplasms (Wiemer, 2007).

In the third type of case study, we chose Esophageal Neoplasms as the test disease to validate the robustness of IMCMDA based on the database HMDD V1.0. There are respectively 10 and 49 out of the top 10 and 50 predicted Esophageal Neoplasms-related miRNAs included by dbDEMC, miR2Disease and HMDD V2.0

Table 2. Prediction results of the top 50 predicted Conlon Neoplasms-related miRNAs based on known associations in HMDD V2.0

miRNA	Evidence	miRNA	Evidence
hsa-mir-21	dbdeMC; miR2Disease	hsa-mir-16	dbDEMC
hsa-mir-155	dbdeMC; miR2Disease	hsa-mir-127	dbdeMC; miR2Disease
hsa-mir-19b	dbdeMC; miR2Disease	hsa-mir-29b	dbdeMC; miR2Disease
hsa-mir-18a	dbdeMC; miR2Disease	hsa-mir-146b	unconfirmed
hsa-mir-20a	dbdeMC; miR2Disease	hsa-mir-101	unconfirmed
hsa-let-7a	dbdeMC; miR2Disease	hsa-mir-92b	unconfirmed
hsa-mir-19a	dbdeMC; miR2Disease	hsa-mir-9	dbdeMC; miR2Disease
hsa-mir-143	dbdeMC; miR2Disease	hsa-mir-214	dbDEMC
hsa-mir-125b	dbDEMC	hsa-mir-1	dbdeMC; miR2Disease
hsa-mir-34a	dbdeMC; miR2Disease	hsa-mir-30c	dbdeMC; miR2Disease
hsa-let-7e	dbDEMC	hsa-mir-181b	dbdeMC; miR2Disease
hsa-let-7d	dbDEMC	hsa-mir-191	dbdeMC; miR2Disease
hsa-mir-223	dbdeMC; miR2Disease	hsa-mir-222	dbDEMC
hsa-let-7c	dbDEMC	hsa-let-7g	dbdeMC; miR2Disease
hsa-let-7b	dbdeMC; miR2Disease	hsa-mir-106b	dbdeMC; miR2Disease
hsa-mir-132	miR2Disease	hsa-mir-210	dbDEMC
hsa-let-7f	dbdeMC; miR2Disease	hsa-mir-205	dbDEMC
hsa-mir-199a	unconfirmed	hsa-mir-203	dbdeMC; miR2Disease
hsa-mir-92a	unconfirmed	hsa-mir-24	miR2Disease
hsa-let-7i	dbDEMC	hsa-mir-20b	unconfirmed
hsa-mir-125a	dbdeMC; miR2Disease	hsa-mir-29a	dbdeMC; miR2Disease
hsa-mir-200b	dbDEMC	hsa-mir-34c	miR2Disease
hsa-mir-146a	dbDEMC	hsa-mir-150	unconfirmed
hsa-mir-141	dbdeMC; miR2Disease	hsa-mir-200a	unconfirmed
hsa-mir-221	dbdeMC; miR2Disease	hsa-mir-34b	dbdeMC; miR2Disease

Note: The first column records top 1–25 related miRNAs. The third column records the top 26–50 related miRNAs.

Table 3. Prediction results of the top 50 predicted Breast Neoplasms-related miRNAs when we set the known associations of Breast Neoplasms as unknown ones

miRNA	Evidence	miRNA	Evidence
hsa-mir-21	dbdemc; miR2Diseaes; hmdd	hsa-mir-200b	dbdemc; miR2Diseaes; hmdd
hsa-mir-17	miR2Disease; hmdd	hsa-mir-221	dbdemc; miR2Diseaes; hmdd
hsa-mir-155	dbdemc; miR2Diseaes; hmdd	hsa-mir-141	dbdemc; miR2Diseaes; hmdd
hsa-mir-19b	dbdemc; hmdd	hsa-mir-29b	dbdemc; miR2Diseaes; hmdd
hsa-mir-145	dbdemc; miR2Diseaes; hmdd	hsa-mir-16	dbdemc; hmdd
hsa-mir-18a	dbdemc; miR2Diseaes; hmdd	hsa-mir-146b	dbdemc; miR2Diseaes; hmdd
hsa-mir-20a	miR2Disease; hmdd	hsa-mir-127	dbdemc; miR2Diseaes; hmdd
hsa-let-7a	dbdemc; miR2Diseaes; hmdd	hsa-mir-92b	dbDEMC
hsa-mir-19a	dbdemc; hmdd	hsa-mir-9	dbdemc; miR2Diseaes; hmdd
hsa-let-7e	dbdemc; hmdd	hsa-mir-101	dbdemc; miR2Diseaes; hmdd
hsa-mir-34a	dbdemc; hmdd	hsa-mir-106a	dbDEMC
hsa-mir-125b	miR2Disease; hmdd	hsa-let-7g	dbdemc; hmdd
hsa-mir-223	dbdemc; hmdd	hsa-mir-106b	dbdemc; hmdd
hsa-mir-126	dbdemc; miR2Diseaes; hmdd	hsa-mir-210	dbdemc; miR2Diseaes; hmdd
hsa-mir-92a	hmdd	hsa-mir-191	dbdemc; miR2Diseaes; hmdd
hsa-let-7d	dbdemc; miR2Diseaes; hmdd	hsa-mir-200c	dbdemc; miR2Diseaes; hmdd
hsa-mir-143	dbdemc; miR2Diseaes; hmdd	hsa-mir-29a	dbdemc; hmdd
hsa-let-7c	dbdemc; hmdd	hsa-mir-222	dbdemc; miR2Diseaes; hmdd
hsa-let-7b	dbdemc; hmdd	hsa-mir-181b	dbdemc; miR2Diseaes; hmdd
hsa-let-7f	dbdemc; miR2Diseaes; hmdd	hsa-mir-20b	hmdd
hsa-mir-146a	dbdemc; miR2Diseaes; hmdd	hsa-mir-150	dbDEMC
hsa-mir-199a	dbdemc; hmdd	hsa-mir-30c	dbdemc; hmdd
hsa-mir-132	dbdemc; hmdd	hsa-mir-214	dbdemc; hmdd
hsa-let-7i	dbdemc; miR2Diseaes; hmdd	hsa-mir-24	dbdemc; hmdd
hsa-mir-125a	dbdemc; miR2Diseaes; hmdd	hsa-mir-203	dbdemc; miR2Diseaes; hmdd

Note: The first column records top 1–25 related miRNAs. The third column records the top 26–50 related miRNAs.

(see Table 4). It is worth noting that only the candidate miRNAs for Esophageal Neoplasms were ranked and confirmed by experimental evidences. As has been defined, the candidate miRNAs are the miRNAs which was unassociated with the Esophageal Neoplasms according to HMDD v1.0. Therefore, none of the top 50 predictions existed in HMDD v1.0 so that validation of the predictions using HMDD v2.0 was completely independent of this training database HMDD v1.0.

The results in independent case studies on five major human diseases have indicated excellent prediction performance of IMCMDA. We have provided the prediction scores of all the unknown human miRNA–disease pairs (see Supplementary Table S3). We hope the prediction results could be helpful in searching potential disease-related miRNAs in the future.

5 Discussion

The researches for potential miRNA–disease associations prediction would help us understand the pathogenesis of disease and promote the treatment of diseases. In this paper, we developed a model of Inductive Matrix Completion for MiRNA–Disease Association prediction (IMCMDA). In model of IMCMDA, the known miRNA–disease associations and the integrated miRNA similarity and disease similarity were combined to calculate the prediction score of each miRNA–disease pair. The AUC of IMCMDA is 0.8034 based on LOOCV, which showed a better performance than previous methods. Furthermore, the predicted disease-related miRNAs of five major human diseases: Colon Neoplasms, Kidney Neoplasms, Lymphoma, Breast Neoplasms and Esophageal Neoplasms were respectively confirmed by the experimental literatures.

The reasons of the reliable performance of IMCMDA are as follows: IMCMDA predicts the miRNA–disease associations by using

Table 4. Prediction results of the top 50 predicted Esophageal Neoplasms -related miRNAs based on known associations in HMDD V1.0

miRNA	Evidence	miRNA	Evidence
hsa-mir-20a	dbdemc; hmdd	hsa-mir-127	dbDEMC
hsa-mir-17	dbDEMC	hsa-mir-222	dbDEMC
hsa-mir-18a	dbDEMC	hsa-mir-106b	dbDEMC
hsa-mir-155	dbdemc; hmdd	hsa-mir-9	dbDEMC
hsa-mir-19a	dbdemc; hmdd	hsa-mir-25	dbdemc; hmdd
hsa-mir-19b	dbDEMC	hsa-mir-125a	dbDEMC
hsa-mir-92a	hmdd	hsa-mir-29a	dbDEMC
hsa-mir-221	dbDEMC	hsa-mir-146b	dbDEMC
hsa-let-7a	dbdemc; hmdd	hsa-mir-141	dbdemc; hmdd
hsa-mir-146a	dbdemc; hmdd	hsa-mir-132	dbDEMC
hsa-mir-34a	dbdemc; hmdd	hsa-mir-191	dbDEMC
hsa-let-7e	dbDEMC	hsa-let-7g	dbDEMC
hsa-mir-145	dbdemc; hmdd	hsa-mir-92b	dbDEMC
hsa-mir-223	dbdemc; miR2Diseaes; hmdd	hsa-mir-214	dbdemc; hmdd
hsa-let-7b	dbdemc; hmdd	hsa-mir-93	dbDEMC
hsa-mir-29b	dbDEMC	hsa-mir-34c	dbdemc; hmdd
hsa-let-7d	dbDEMC	hsa-mir-181b	dbDEMC
hsa-let-7f	unconfirmed	hsa-mir-15a	dbdemc; hmdd
hsa-let-7i	dbDEMC	hsa-mir-20b	dbDEMC
hsa-mir-199a	dbdemc; hmdd	hsa-mir-200a	dbdemc; hmdd
hsa-mir-125b	dbDEMC	hsa-mir-101	dbdemc; hmdd
hsa-mir-126	dbdemc; hmdd	hsa-mir-24	dbDEMC
hsa-let-7c	dbdemc; hmdd	hsa-mir-30c	dbDEMC
hsa-mir-200b	dbDEMC	hsa-mir-34b	dbdemc; hmdd
hsa-mir-16	dbDEMC	hsa-mir-18b	dbDEMC

Note: The first column records top 1–25 related miRNAs. The third column records the top 26–50 related miRNAs.

the low-rank inductive matrix completion (IMC) algorithm. A crucial advantage of IMC is that it utilizes disease similarity and miRNA similarity as the feature of disease and miRNA to complete the missing miRNA–disease association. It means that we can use the feature vector of a new disease without any known related miRNAs to predict the relevance-scores between this new disease and all miRNAs. That's why IMCMDA can be applied to new disease without any known related miRNAs (Natarajan and Dhillon, 2014). In addition, searching the optimal solution with an alternating gradient descent algorithm made sure the reliability of the disease eigenvectors and the miRNA eigenvectors. Finally, the model is a semi- supervised model. The advantage of semi-supervised model is that it doesn't rely on negative samples. It only needs positive samples and unlabeled samples, which greatly reduces the difficulty of building models. Therefore, this model fits in with our current research topic (almost no negative samples).

Yet, there are some limitations that influence the performance of IMCMDA. Firstly, the materials we used including human miRNA–disease associations, disease semantic similarity and miRNA functional similarity possibly contains noise and outliers. Secondly, IMCMDA uses the least square error function which is well known to be unstable with noises and outliers. Besides, due to the limitations of laboratory conditions, we are not able to do wet experiments to verify the predictions. As the laboratory conditions allow, we will certainly supplement the relevant experiments in the future. We have provided the prediction results in Supplementary Table S3, we expect to receive validation from other teams on biological experiments. Finally, to our knowledge, the hallmarks of cancer are one of the most widely acknowledged organizing principles for research on cancer. Some literatures pointed out that there are some associations between cancer hallmarks and genes (Gao *et al.*, 2016). For example, hsa-mir-21 obtained the highest score in the case studies of Colon Neoplasms and Breast Neoplasms, while according to data from NanoString's Hallmarks of Cancer Panel collection (https://www.nanostring.com/), the target of hsa-mir-21, APP, has been identified to be associated with Hallmark inflammation. In the future, new biological information, such as the types of disease–miRNA associations (Chen *et al.*, 2015), cancer hallmark-gene associations and gene sequence information (Wang *et al.*, 2015), could be also incorporated into our future research. We should exploit the information of disease-microRNA associations, cancer hallmark-gene associations and gene sequence to establish miRNA–disease similarity networks thus improve the accuracy of our model. With the huge amount of biological data, the prediction of models could be more reliable and useful.

Acknowledgements

We thank anonymous reviewers for very valuable suggestions.

Funding

X.C. was supported by National Natural Science Foundation of China under Grant No. 61772531.

Conflict of Interest: none declared.

References

Bandyopadhyay,S. (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.

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

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

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

Chen,X. *et al.* (2015) RBMMMDA: predicting multiple types of disease-microRNA associations. *Sci. Rep.*, **5**, 13877.

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

Chen,X. *et al.* (2016b) HGIMDA: heterogeneous graph inference for miRNA–disease association prediction. *Oncotarget*, **7**, 65257–65269.

Chen,X. *et al.* (2016c) Drug-target interaction prediction: databases, web servers and computational models. *Brief. Bioinf.*, **17**, 696–712.

Chen,H. *et al.* (2017a) miRDDCR: a miRNA-based method to comprehensively infer drug–disease causal relationships. *Sci. Rep.*, **7**, 15921.

Chen,X. *et al.* (2017b) RKNNMDA: ranking-based KNN for MiRNA–disease association prediction. *RNA Biol.*, **14**, 952–962.

Chen,X. *et al.* (2017c) Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinf.*, **18**, 558–576.

Chen,X. *et al.* (2017d) NRDTD: a database for clinically or experimentally supported non-coding RNAs and drug targets associations. *Database*, **2017**, bax057.

Chen,X. *et al.* (2017e) MicroRNAs and complex diseases: from experimental results to computational models. *Brief. Bioinf.*

Chen,X. and Yan,G.-Y. (2015) Semi-supervised learning for potential human microRNA-disease associations inference. *Sci. Rep.*, **4**, 5501.

Cheng,A.M. (2005) Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic Acids Res.*, **33**, 1290–1297.

Cimmino,A. *et al.* (2005) miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc. Natl. Acad. Sci. USA*, **102**, 13944–13949.

Díaz,R. *et al.* (2008) Deregulated expression of miR-106a predicts survival in human colon cancer patients. *Genes Chromosomes Cancer*, **47**, 794–802.

Fantine,M. and Jean-Philippe,V. (2011) ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinformatics*, **12**, 389.

Gao,S. *et al.* (2016) Identification and construction of combinatory cancer hallmark-based gene signature sets to predict recurrence and chemotherapy benefit in stage II colorectal cancer. *JAMA Oncol.*, **2**, 37–39.

Goh,J.N. *et al.* (2016) microRNAs in breast cancer: regulatory roles governing the hallmarks of cancer. *Biol. Rev. Camb. Philos. Soc.*, **91**, 409–428.

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

Harrison,J.S. (2013) Leukemia and lymphoma society. *Soc. Sci. Electronic Publish.*, **21**, 3699–3707.

Jain,P. and Dhillon,I.S. (2013) Provable inductive matrix completion.

Jemal,A. *et al.* (2011) Global cancer statistics. *CA Cancer J. Clin.*, **61**, 69–90.

Jemal,A. *et al.* (2006) Cancer statistics, 2006. *CA Cancer J. Clin.*, **56**, 106–130.

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

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

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

Karp,X. and Ambros,V. (2005) Encountering microRNAs in cell fate signaling. *Science*, **310**, 1288–1289.

Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.

Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.

Lee,R.C. *et al.* (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.

Li,J.-Q. *et al*. (2017) MCMDA: matrix completion for MiRNA–disease association prediction. *Oncotarget*, **8**, 21187–21199.

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

Lipscomb,C.E. (2000) Medical subject headings (MeSH). *Bull. Med. Libr. Assoc*., **88**, 265–266.

Liu,W. *et al*. (2010) MicroRNA-23b* targets proline oxidase, a mitochondrial tumor suppressor protein in renal cancer. *Oncogene*, **29**, 4914–4924.

Liu,Y. *et al*. (2016) Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans. Comput. Biol. Bioinf*., **99**, 1–1.

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

Mayr,C. *et al*. (2007) Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation. *Science*, **315**, 1576–1579.

McDuffie,H.H. *et al*. (2009) Clustering of cancer among families of cases with Hodgkin Lymphoma (HL), Multiple Myeloma (MM), Non-Hodgkin's Lymphoma (NHL), Soft Tissue Sarcoma (STS) and control subjects. *BMC Cancer*, **9**, 70.

Meister,G. and Tuschl,T. (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature*, **431**, 343–349.

Miska,E.A. (2005) How microRNAs control cell division, differentiation and death. *Curr. Opin. Genet. Dev*., **15**, 563–568.

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

Natarajan,N. and Dhillon,I.S. (2014) Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, **30**, i60–i68.

Ogata-Kawata,H. *et al*. (2014) Circulating exosomal microRNAs as biomarkers of colon cancer. *PLoS One*, **9**, e92921.

Pasquier,C. and Gardès,J. (2016) Prediction of miRNA–disease associations with a vector space model. *Sci. Rep*., **6**, 27036.

Quesne,J.L. *et al*. (2012) Biological and prognostic associations of miR-205 and let-7b in breast cancer revealed by in situ hybridization analysis of micro-RNA expression in arrays of archival tumour tissue. *J. Pathol*., **227**, 306–314.

Senanayake,U. *et al*. (2012) miR-192, miR-194, miR-215, miR-200c and miR-141 are downregulated and their common target ACVR2B is strongly expressed in renal childhood neoplasms. *Carcinogenesis*, **33**, 1014–1021.

Shi,B. *et al*. (2007) Micro RNA 145 targets the insulin receptor substrate-1 and inhibits the growth of colon cancer cells. *J. Biol. Chem*., **282**, 32582–32590.

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.

Seigel,R. *et al*. (2012) Cancer statistics 2012. *CA Cancer J. Clin*., **62**, 10–29.

Siegel,R.L. *et al*. (2018) Cancer statistics, 2018, 68, 7–30.

Singh-Blom,U.M. *et al*. (2013) Correction: prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLoS One*, **8**, e58977.

Victor,A. (2001) microRNAs: tiny regulators with great potential. *Cell*, **107**, 823–826.

Victor,A. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.

Wan,W. and Tian,L. (2014) Research progress of Hodgkin lymphoma. *J. Leuk. Lymp*., 308–311.

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

Wang,E. *et al*. (2015) Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Semin. Cancer Biol*., **30**, 4.

Wiemer,E.A. (2007) The role of microRNAs in cancer: no small matter. *Eur. J. Cancer*, **43**, 1529–1544.

Xia,H. *et al*. (2012) Loss of brain-enriched miR-124 microRNA enhances stem-like traits and invasiveness of glioma cells. *J. Biol. Chem*., **287**, 9962–9971.

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

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

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

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

You,Z.H. *et al*. (2017) PBMDA: a novel and effective path-based computational model for miRNA–disease association prediction. *PLoS Comput. Biol*., **13**, e1005455.

Yu,H. *et al*. (2017) Large-scale prediction of microRNA-disease associations by combinatorial prioritization algorithm. *Sci. Rep*., **7**, 43792.

Zeng,X. *et al*. (2018) Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics (Oxford, England)*.

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

Zhang,T. *et al*. (2011) The oncogenetic role of microRNA-31 as a potential biomarker in oesophageal squamous cell carcinoma. *Clin. Sci*., **121**, 437.

Zou,Q. *et al*. (2015) Prediction of MicroRNA-disease associations based on social network analysis methods. *Biomed. Res. Int*., **2015**, 1.

Zou,Q. *et al*. (2016) Similarity computation strategies in the microRNA-disease network: a survey. *Brief. Funct. Genomics*, **15**, 55.