BMC Genomics

RESEARCH Open Access

MicroRNA-disease association prediction by matrix tri-factorization

Huiran Li, Yin Guo, Menglan Cai and Limin Li*

From The 18th Asia Pacific Bioinformatics Conference
Seoul, Korea. 18-20 August 2020

Abstract

Background: Biological evidence has shown that microRNAs(miRNAs) are greatly implicated in various biological progresses involved in human diseases. The identification of miRNA-disease associations(MDAs) is beneficial to disease diagnosis as well as treatment. Due to the high costs of biological experiments, it attracts more and more attention to predict MDAs by computational approaches.

Results: In this work, we propose a novel model MTFMDA for miRNA-disease association prediction by matrix tri-factorization, based on the known miRNA-disease associations, two types of miRNA similarities, and two types of disease similarities. The main idea of MTFMDA is to factorize the miRNA-disease association matrix to three matrices, a feature matrix for miRNAs, a feature matrix for diseases, and a low-rank relationship matrix. Our model incorporates the Laplacian regularizers which force the feature matrices to preserve the similarities of miRNAs or diseases. A novel algorithm is proposed to solve the optimization problem.

Conclusions: We evaluate our model by 5-fold cross validation by using known MDAs from HMDD V2.0 and show that our model could obtain the significantly highest AUCs among all the state-of-art methods. We further validate our method by applying it on colon and breast neoplasms in two different types of experiment settings. The new identified associated miRNAs for the two diseases could be verified by two other databases including dbDEMC and HMDD V3.0, which further shows the power of our proposed method.

Keywords: micoRNA-disease association prediction, Matrix tri-factorization

Background

MicroRNAs(miRNAs), a class of small, endogenous, non-coding RNAs including approximately 22 nucleotides, could regulate post-transcription of gene expression and RNA silencing by binding specific target messenger RNAs through base-pairing interactions [1, 2]. Since the first miRNA named lin-4 was found twenty years ago by Victor Ambros [3], with the development of technology, an increasing number of studies found that miRNAs play important roles in various stages of biological processes [3], such as cell development [4], proliferation [5] and viral infection [6]. Meanwhile, biological experiments indicate that miRNAs are involved in close relationships with the emergence and development processes of various human diseases [7]. For example, the study in [8] showed that a chromosomal translocation at 12q5 could influence the expression of let-7 and finally could cause the repress of the oncogene High Mobility Group A2(Hmga2). Another example is that mir-7 could influence epidermal growth factor receptor (EGFR) expression and protein kinase B activity in head and neck cancer(HNC) [9]. Furthermore, the work in [10] showed that mir-15a is a potential marker to differentiate between benign and malignant

*Correspondence: liminli@mail.xjtu.edu.cn
School of Mathematics and Statistics, Xi'an Jiaotong University, Xianning West 28, Xi'an, China

renal tumors in biopsy and urine samples. It is very important to identify miRNA-disease associations (MDAs) for the research on disease mechanism and discovering disease biomarkers. Due to the high costs of the current biological technologies, computational methods are useful by prioritizing candidate miRNAs for specific diseases.

The main information used for predicting MDAs mainly includes miRNA similarities, disease similarities and the known MDAs. Generally, miRNA similarities could be computed by using functional or sequence information of miRNAs, disease similarities could be obtained by using the phenotype terms, and the known MDAs could be obtained from databases such as HMDD [11]. The main challenge in MDA prediction is how to optimally utilize these information and predict MDAs with a high accuracy. Based on these information, many computational models are proposed to predict new MDAs.

The existing methods follow two lines. The first line is to determine the link probabilistically by using random walk. For example, RWRMDA [12] adopts random walk on the miRNA functional similarity network. It first gives each miRNA an initial probability in the miRNA functional similarity network(MFSN), and then use a random walk algorithm until the probability get stable. However, this method cannot predict new disease without any known related miRNAs. Thus Shi et al. [13] use the random walk algorithm on miRNA target and disease genes at the same time to map the protein-protein interaction (PPI) network, and then they construct a bipartite miRNA-disease network by using *p*-values in the PPI network and identify co-regulated modules by hierarchical clustering analysis. Later Xuan et al. [14] develop the MIDP method by using the prior information of nodes. They first divide the diseases related to the miRNAs into labeled nodes and unlabeled nodes, and establish the transition matrices for the two categories of nodes. Then by using the random walk algorithm on the two weighted transition matrices, the final miRNA ranking could be obtained. Liu et al. [15] proposed a random walk method to predict the associations by combining the multiple data sources.

The second line is to formulate the problem as machine learning problems such as classification, matrix completion. For classification formulation, examples include the RLSMDA [16] and the MTDN [17] methods. RLSMDA [16] develops Regularized Least Squares algorithm by training two classifiers from the miRNA space and the disease space. However, how to choose the parameter of RLSMDA and how to combine the classifiers need to be studied furthermore. Xu et al. [17] introduce the MTDN approach based on miRNA target-dysregulated network to prioritize novel disease miRNAs. The method first constructs the network by combining computational target prediction with miRNA and mRNA expression profiles in tumor and non-tumor tissues, and then applies a support vector machine classifier to distinguish positive miRNA-disease associations from negative ones by extracting the feature of network topologic information. However, it is hard to obtain the negative miRNA-disease associations. Another option using machine learning is matrix completion such as MCLPMDA [18] , IMCMDA [19], CMFMDA [20] and PMAMCA [21]. MCLPMDA [18] constructs new miRNA and disease similarity matrices by matrix completion algorithm firstly, and then uses label propagation algorithm to predict miRNAs. Chen et al. [19] propose a method named IMCMDA based on nonnegative matrix factorization, whose main idea is to complete the missing miRNA-disease association based on the known associations and miRNA and disease similarity. CMFMDA [20] and PMAMCA [21] both factorize the association matrix into two matrices which representing the features for miRNAs and diseases, respectively.

In this study, we propose a novel computational method MTFMDA to predict new MDAs by matrix tri-factorization, to follow the idea of matrix completion. The main idea of MTFMDA is that we factorize the complete MDA matrix to three matrices, a feature matrix P for miRNAs, a feature matrix Q for diseases, and a low-rank matrix D representing relationships between miRNA features and disease features. Laplacian regularizers are used for the feature matrices P and Q by using two types of miRNA similarities, and two types of disease similarities, respectively. Optimal matrices P, D and Q are learnt by using the known MDAs and the Laplacian regularizers, and then the MDA matrix is completed by PDQ^T and thus new MDAs can be identified.

The contributions in this work are listed as follows:

1. We propose a new MDA prediction model by matrix tri-factorization model, which combines the two types of miRNA similarities, two types of disease similarities, and the known miRNA-disease associations, and predict new MDAs by completing the MDA matrix. We develop an algorithm for solving the optimization problem.
2. We evaluate our MTFMDA model by 5-fold cross-validation and obtain higher accuracies than other state-of-art methods.
3. We apply our method on two diseases to identify related miRNAs, and our prediction results could be supported by other databases. This further validates the effectiveness of our model MTFMDA.

Materials and methods

Datasets

Human miRNA-disease associations

We collect the known human miRNA-disease associations from HMDD V2.0 database (June, 2014) [11], and obtain 3693 associations among 368 miRNAs and 383 diseases.

MiRNA functional similarity and sequence similarity

The functional similarities among miRNAs can be calculated by the method proposed in [22], and we download the similarity data from http://www.cuilab.cn. Since miRNA's function is closely relevant to the miRNA sequence [23], we also obtain the miRNA sequence similarity from http://www.mirbase.org/ftp.shtml. The integrated similarities among miRNAs are defined as the average of the functional similarity and the sequence similarity, and the integrated similarity matrix for miRNAs is denoted as S_m.

Two disease semantic similarities

To calculate disease semantic similarities, Wang [22] and Xuan [24] propose two methods based on the Medical Subject Headings (MeSH) descriptors which could be downloaded from the National Library of Medicine (http://www.nlm.nih.gov/).

Wang's method [22] first calculates the semantic value and contribution value of a disease, and then uses these two values to compute the semantic similarity between two diseases. Unlike Wang's method, Xuan et al. [25] improves the calculation method of semantic value. It also uses semantic value and contribution value to calculate the semantic similarity. We use the integrated similarity in our work by averaging the two types of semantic similarities, and denote the integrated similarity matrix for diseases as S_d.

Our proposed method via matrix tri-factorization

Problem statement and notations

We are now given the integrated similarity matrix $S_m \in R^{n_m \times n_m}$ among n_m miRNAs $\left\{m_1, \cdots, m_{n_m}\right\}$, and the integrated similarities $S_d \in R^{n_d \times n_d}$ among n_d diseases $\left\{d_1, \cdots, d_{n_d}\right\}$. We are also given the miRNA-disease association (MDA) indicator matrix $A \in R^{n_m \times n_d}$ defined as follows

$$A(i,j) = \begin{cases} 1, & i\text{-th miRNA } m_i \text{ is associated with } j\text{-th disease } d_j, \\ 0, & \text{association between } i\text{-th miRNA } m_i \text{ and } j\text{-th disease } d_j \text{ is unknown.} \end{cases}$$

We denote $\Omega = \left\{(i,j)\,|A_{ij} = 1\right\}$ to be the indices for the miRNA-disease pairs which are known to be associated, and $\Omega^c = \left\{(i,j)\,|A_{ij} = 0\right\}$ to be all the pairs whose associations are unknown. For any matrix M, we denote $\mathcal{R}_\Omega(M)$ by only keeping its Ω part and forcing its Ω^c part to be zeros, that is,

$$\mathcal{R}_\Omega(M)_{ij} = \begin{cases} M_{ij}, & if\,(i,j) \in \Omega \\ 0, & if\,(i,j) \in \Omega^c. \end{cases}$$

Our aim in this work is to complete the Ω^c part in matrix A, and recover the complete matrix $\tilde{A}$.

MTFMDA model

We propose our MTFMDA method by considering the following three aspects. First, the unknown complete miRNA-disease association (MDA) matrix $\tilde{A}$ can be factorized into three matrices, a feature matrix for miRNAs $P \in R^{n_m \times r_m}$, a feature matrix for diseases $Q \in R^{n_d \times r_d}$, and the feature relationship matrix $D \in R^{r_m \times r_d}$. The factorization $\tilde{A} = PDQ^T$ implies that the column vectors in $\tilde{A}$ lie in the subspace spanned by the column vectors in P, and the row vectors in $\tilde{A}$ lie in the subspace spanned by the column vectors in Q. D is generally required to be low rank, and P and Q are orthonormal matrices satisfying $P^TP = I$ and $Q^TQ = I$. Second, the complete $\tilde{A}$ should recover the known associations between miRNAs and diseases, i.e, the Ω part of the difference matrix $\left(A - \tilde{A}\right)$ should be zero or as small as possible. Third, the feature vectors in P and Q should preserve the similarity information hidden in the S_m and S_d, respectively, and thus two Laplacian regularizers should be used for preserving the geometric structure. By considering the above three aspects, we propose the following MTFMDA model

$$\begin{aligned} &\min_{P,D,Q} \left\|\mathcal{R}_\Omega\left(A - PDQ^T\right)\right\|_F^2 + \lambda_1 tr\left(P^TL_mP\right) + \lambda_2 tr\left(Q^TL_dQ\right) + \lambda_3\|D\|_* \\ &s.t. \quad P^TP = I\,, Q^TQ = I, \end{aligned} \tag{1}$$

where λ_1, λ_2 and λ_3 are the regularization parameters to control the trade-offs. The first term is to recover the known MDAs in A. In the second term, $L_m = D_m - S_m$ is the Laplacian matrix for the miRNAs, where D_m is a diagonal matrix with the i-th diagonal element being the sum of i-th row in S_m. In the third term, L_d is the Laplacian matrix for diseases, defined in the same way as L_m. Once the optimal P, D and Q are solved in the optimization problem, the completed MDA matrix $\tilde{A}$ can be obtained by $\tilde{A} = PDQ^T$. The flowchart of our method is shown in Fig.1.

Optimization algorithm

In order to solve optimization problem above, we develop an alternate iteration algorithm to update P, D and Q alternately.

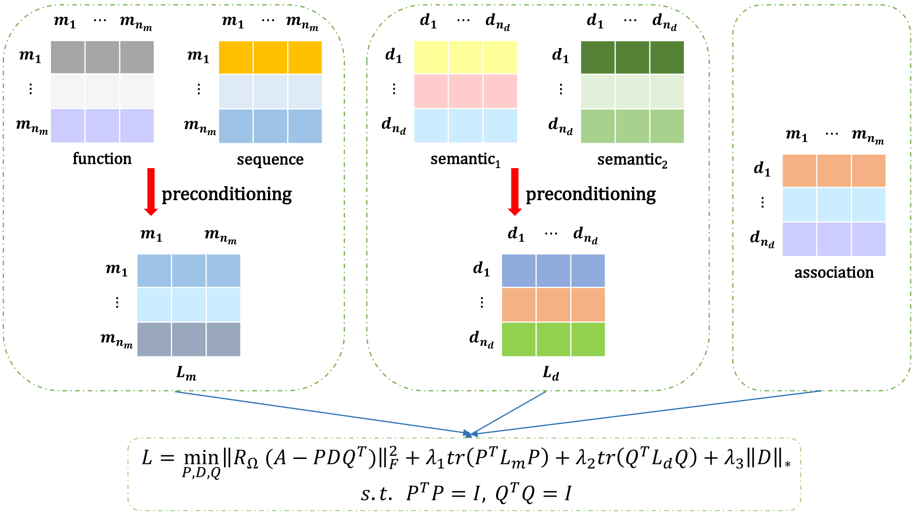
Step 1: Fix P and Q, solve D

By fixing P and Q in the optimization problem (1), the subproblem to solve D can be obtained as follows:

$$\min_D \left\|\mathcal{R}_\Omega\left(A - PDQ^T\right)\right\|_F^2 + \lambda_3\|D\|_*. \tag{2}$$

The sub-problem can be solved by an accelerated gradient descent algorithm [26] with the following iterations,

$$\begin{aligned} &Y_k = D_k + \gamma_k\left(\gamma_{k-1}^{-1} - 1\right)\left(D_k - D_{k-1}\right), \\ &D_{k+1} = \arg\min\nolimits_D \lambda_3 \left|\left|D\right|\right|_* + \frac{s}{2}\left|\left|D - \left(Y_k - \frac{1}{s}\,\triangledown f\left(Y_k\right)\right)\right|\right|_F^2, \\ &\gamma_{k+1} = \left(\sqrt{\gamma_k^4 + 4\gamma_k^2} - \gamma_k^2\right)/2, \end{aligned} \tag{3}$$



$$L = \min_{P,D,Q} \|R_\Omega\left(A - PDQ^T\right)\|_F^2 + \lambda_1 tr(P^T L_m P) + \lambda_2 tr(Q^T L_d Q) + \lambda_3 \|D\|_*$$
$$s.t.\ \ P^T P = I,\ Q^T Q = I$$

Fig. 1 Flowchart of MTFMDA model to infer the potential miRNA-disease associations

where s is a proximal parameter for estimating the second-order gradient of $f(Y)$. The second equation in (3) can be solved by using the linearized Bregman iteration as a special form of Uzawa's algorithm proposed in Cai et al. [27].

Step 2: Fix D and Q, solve P.

By fixing D and Q in optimization problem (1), we obtain the sub-problem of P as follows:

$$\min_P \left\|\mathcal{R}_\Omega\left(A - PDQ^T\right)\right\|_F^2 + \lambda_1 tr\left(P^T L_m P\right). \tag{4}$$

Similarly to solving D in step 1, we could also use the accelerated gradient descent (APG) model to update P as follows:

$$\begin{aligned} &\hat{Y}_k = P_k + \gamma_k\left(\gamma_{k-1}^{-1} - 1\right)\left(P_k - P_{k-1}\right), \\ &P_{k+1} = \arg\min_P \lambda_1 tr\left(P^T L_m P\right) + \frac{s}{2}\left\|P - \left(\hat{Y}_k - \frac{1}{s}\nabla f\left(\hat{Y}_k\right)\right)\right\|_F^2, \\ &\gamma_{k+1} = \left(\sqrt{\gamma_k^4 + 4\gamma_k^2} - \gamma_k^2\right)/2. \end{aligned} \tag{5}$$

Wen's algorithm proposed in [28] is used to solve the second Eq. in (5).

Step 3: Fix D and P, solve Q.

By fixing D and P in optimization problem (1), we obtain the sub-problem of Q as follows:

$$\min_Q \left\|\mathcal{R}_\Omega\left(A - PDQ^T\right)\right\|_F^2 + \lambda_2 tr\left(Q^T L_d Q\right). \tag{6}$$

It can be seen that the sub-problem (6) to solve for Q is the same with the sub-problem (4) to solve for P. Thus we skip the details.

Overall, the framework of our algorithm is shown as follows:

Results

In this section, we will first evaluate our method on the known associations collected from HMDD V2.0 database by 5-fold cross validation. Then we further evaluate our method by using the probability of recovering a true association in the top-t predictions for new diseases. We also estimate the contribution of the performance for the known association matrix, integrated miRNA and disease similarity.

Comparing methods

We compare our MTFMDA method with the following seven methods.

Algorithm 1 Algorithm MTFMDA

```
Inputs:
        MiRNA-disease association matrix $A \in R^{n_m \times n_d}$;
        MiRNA similarity matrix $S_m \in R^{n_m \times n_m}$;
        Disease similarity matrix $S_d \in R^{n_d \times n_d}$;
        Parameters $\lambda_1 = 10, \lambda_2 = 10, \lambda_3 = 1$.
Outputs: $\tilde{A} = PDQ^T$.
1. Construct Laplacian matrices $L_m = D_m - S_m$,
   $L_d = D_d - S_d$.
2. Initialize $P, Q, D$.
3. while $\tilde{A}, P, Q$ and $D$ not converged
4.    Fixed $P$ and $Q$, update $D$ by using Eq. (3);
5.    Fixed $D$ and $Q$, update $P$ by using Eq. (5);
6.    Fixed $P$ and $D$, update $Q$ by using the same way
      with updating $P$;
7.    $\tilde{A} = PDQ^T$;
8. end
```

- **RLSMDA** [16].The model is a semi-supervised learning method, and it develops a Regularized Least Squares algorithm by training two classifiers from the miRNA space and the disease space (we use parameter $\omega = 0.9$).
- **RWRMDA** [12]. The model is a random walk method which infers potential miRNA-disease interactions by implementing random walk on the miRNA-miRNA functional similarity network (we use parameters $r = 0.2$, threshold = 10^{-6}).
- **IMCMDA** [19]. The method is a matrix completion model by nonnegative matrix factorization using the same datasets with our method (we use parameter r=100).
- **NCPMDA** [29]. The method is a non-parametric universal network-based method that combines miRNA space and disease space.
- **KBMFMDA** [30]. The model combines kernel-based nonlinear dimensionality reduction, matrix factorization and binary classification. The main idea of the method is to project miRNAs and diseases into a subspace and estimate the association network in the subspace.
- **CMFMDA** [20]. The model factorizes the association matrix into two parts which represent miRNA and disease information, respectively. SVD factorization is used to initialize the two parts.
- **PMAMCA** [21]. The method divides the association matrix into two latent matrices, and solve the matrix factorization by using the recommend system. Note that this method doesn't use miRNA and disease similarity matrices.

Evaluating our method by cross-validation

We first evaluate the performance of our MTFMDA method by the 5-fold cross validation framework. We use the data with 3693 associations between 368 miRNAs and 383 diseases collected from the HMDD V2.0 database. For the 5-fold cross validation, we divide 368 miRNAs into five folds. We take one fold as the test set, and take the rest as the training set. Each fold is taken as the test set once in turn. After obtaining the complete MDA matrix $\tilde{A}$, we rank the scores for all the test pairs of miRNA-diseases. If the rank of an miRNA-disease pair exceeds a given threshold, then the pair is considered to have an association. In our method, we set the parameters as $\lambda_1 = \lambda_2 = 10$ and $\lambda_3 = 1$. The dimension parameters r_m and r_d are set as the one sixth of n_m and n_d, respectively.

We first plot Receiver Operation Characteristics (ROC) curves for all the methods to check the true positive rates and false positive rates. In the ROC curve, the x-axis is the true positive rate (TPR) and the y-axis is the false positive rate (FPR). The ROC curves for the all the methods are plotted in Fig. 2. We can see that our MTFMDA could obtain the best ROC curve. We then perform 50 runs of 5-fold cross validation, and calculate the AUC (area under curve) values. The average AUC values of MTFMDA, RLSMDA, RWRMDA, IMCMDA, NCPMDA, KBMFMDA, CMFMDA and PMAMCA are reported in Table 1. The results show that our method

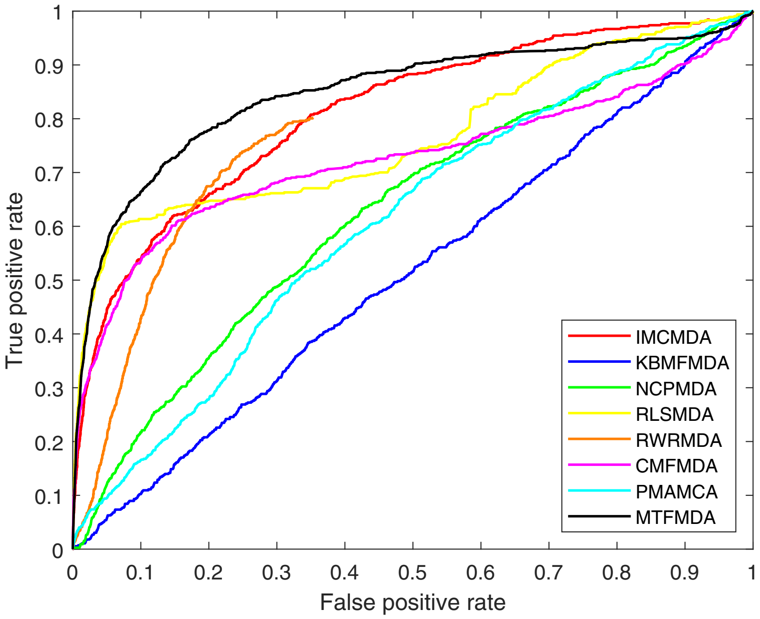


Fig. 2 Performance comparisons between our method MTFMDA and baseline methods(IMCMDA, KBMFMDA, NCPMDA, RLSMDA, RWRMDA, CMFMDA, PMAMCA) in terms of AUC based on 5-fold cross validation

Table 1 Performance comparisons between our method MTFMDA and baseline methods in terms of AUCs based on 5-fold cross validation and paired t-test *p*-values which compare our method with other methods

	IMCMDA	KBMFMDA	NCPMDA	RLSMDA	RWRMDA	CMFMDA	PMAMCA	MTFMDA
AUC	0.8114	0.5110	0.6263	0.7657	0.7778	0.7502	0.6061	0.8484
p-value	8.43e-94	7.87e-196	1.36e-167	7.76e-125	1.88e-119	8.11e-046	2.99e-57	

achieves the highest AUC value and performs better than other methods. We further analyze the differences of inference capability between our method and others. Note that for each method we obtain 50 AUC values for the 50 runs of 5-fold cross validation. Thus, the paired t-test can be used to check whether our method is significantly better than other methods. The *p*-values between our method and other five methods are reported in Table 1. The results show that our method MTFMDA performs significantly better than other methods. We further plot the Precision-Recall curves in Fig. 3 for all the methods, and we can also see that our method performs better than all the other comparing methods.

Probability to recover true associated miRNAs for new diseases

We further evaluate our MTFMDA method by the probability of recovering a true association in the top-*t* predictions for a new disease. The probability can measure whether the method can predict potential related miRNA for a new disease. The measurement has been used in many other publications such as [19, 31, 32]. In detail, for each test disease, we first mask its known associated miRNAs as zero in matrix A, and then apply our model to obtain the ranks of the masked true associated miRNAs. Thus for all the 383 diseases, we could obtain the ranks of the true associated miRNAs among all the miRNAs. We could then plot the cumulative distribution function (CDF), where x-axis represents the top-*t* predicted miRNAs, and y-axis represents the probability of recovering a true association in the top t predictions. Other methods could also plot the curves, except the RWRMDA, which cannot predict the new diseases. The CDFs for the five methods are shown in Fig. 4. From the figure, we can see that though NCPMDA performs better than ours for the top 13 miRNAs, the probability of recovering a true association in the top t predictions does not change much when t from 13 to 40. When t is from 13 to 40, our method could recover true associated miRNAs with highest probabilities.

Contributions of different data sources

We use three data sets in our model, miRNA similarity S_m, disease similarity S_d and known miRNA-disease association matrix A. To examine their contributions to the performance of our model, we first change each of the

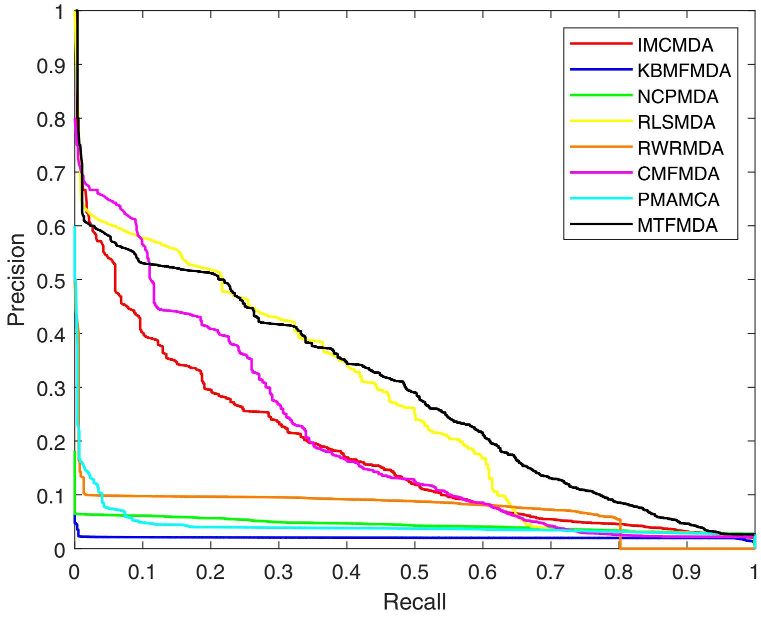


Fig. 3 Performance comparisons between MTFMDA and baseline methods(IMCMDA, KBMFMDA, NCPMDA, RLSMDA,RWRMDA, CMFMDA, PMAMCA) in terms of PR curve

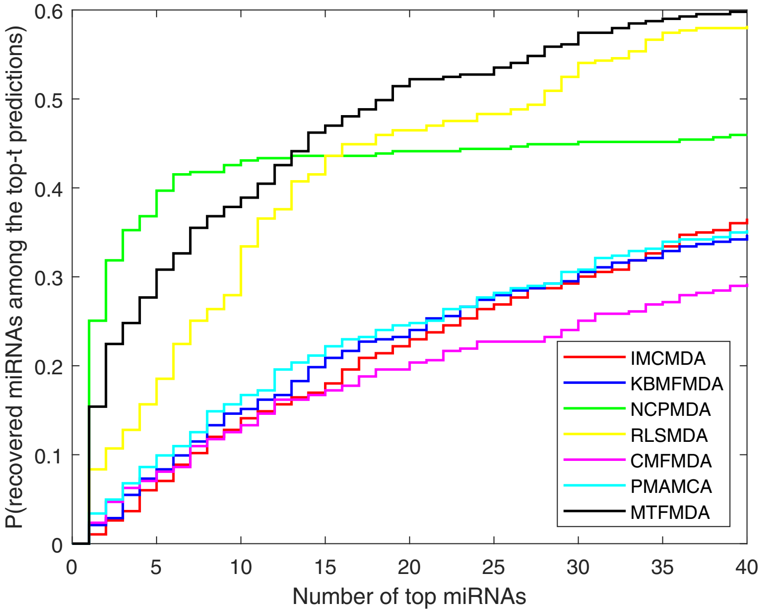


Fig. 4 Performance comparisons between MTFMDA and baseline methods(IMCMDA, KBMFMDA, NCPMDA, RLSMDA, CMFMDA, PMAMCA) in predicting potential miRNAs for new diseases

three matrices to a random matrix, and then apply our MTFMDA model to check the AUC values based on 5-fold cross validation. If one data type contributes the most, then the corresponding AUC should decrease a lot when changing to the data to be random. We change S_m, S_d and A to be random in turn, and the resulting AUCs are reported in Table 2. As shown in Table 2, the average AUC value based on the random miRNA similarity matrix is much lower than the other two types in our model, and thus the miRNA similarity contributes the most to the performance of our model. We also see that the disease similarity contributes the least in our MTFMDA model.

Discussion

We apply two diseases including colon and breast neoplasms to verify the effectiveness of the MTFMDA for miRNA-disease association prediction. For each disease, we apply our MTFMDA model to predict the top-t associated miRNAs, and then examine these predictions by two datasets: dbDEMC [33] and HMDD V3.0 [34].

Table 2 AUCs of our model by changing one of the three data sources to be random in turn

Randomized Data sources	Average value of AUCs
Known miRNA-disease associations	0.5025
MiRNA similarities	0.1951
Disease similarities	0.8360

Table 3 shows the predicted top 50 miRNAs for the colon neoplasms using our MTFMDA model. Colon neoplasms is an out-of-control cell growth, and it is the second leading cause of death in cancer and the most common tumor in the gastrointestinal tract [35, 36]. Many factors will cause the neoplasms, such as old age, unhealthy lifestyle and heredity [37]. Recently, more and more evidence proved that some miRNAs are related to the colon neoplasms. For example, Zhang et al. [38] showed that mir-21, mir-17 and mir-19a promote the metastasis and spread of colon neoplasms. Coincidentally, the expression levels of miR-106a of normal human are higher than colon cancer patients [39]. Shi et al. [40] found that mir-145 could down-regulate the IRS-1 protein in the colon neoplasms cells and inhibit cells growth through targeting the IRS-1 3'-untranslated region. From our prediction results shown in Table 3, we can see that all the predicted top 10 miRNAs by our method can be confirmed by both the dbDEMC and HMDD V3.0 databases, and 46 among the top 50 miRNAs can be confirmed by dbDEMC and HMDD V3.0 databases. This validates the effectiveness of our MTFMDA method.

For the breast neoplasms, we evaluate our method in another way. We mask all the known associated miRNAs with breast neoplasms and apply our MTFMDA method to obtain the predicted top 50 associated miRNAs for the breast neoplasms, shown in Table 4. From this table, we can see that, all the top 10 miRNAs are confirmed by the two databases, and 49 of the top 50 miRNAs can be confirmed by the two databases. Through the Table 4,

Table 3 The predicted top 50 miRNAs associated with colon neoplasms and the evidence from databases HMDD V3.0 and dbDEMC

miRNA	Evidence	miRNA	Evidence
hsa-mir-155	dbDEMC;HMDD	hsa-let-7b	dbDEMC;HMDD
hsa-mir-146a	dbDEMC;HMDD	hsa-mir-143	dbDEMC;HMDD
hsa-mir-328	dbDEMC;HMDD	hsa-mir-200a	HMDD
hsa-mir-29a	dbDEMC;HMDD	hsa-mir-195	dbDEMC;HMDD
hsa-mir-17	dbDEMC;HMDD	hsa-let-7c	dbDEMC;HMDD
hsa-mir-34a	dbDEMC;HMDD	hsa-mir-326	dbDEMC;HMDD
hsa-mir-20a	dbDEMC;HMDD	hsa-mir-23a	dbDEMC;HMDD
hsa-mir-15a	dbDEMC;HMDD	hsa-mir-210	dbDEMC;HMDD
hsa-mir-221	dbDEMC;HMDD	hsa-mir-141	dbDEMC;HMDD
hsa-mir-145	dbDEMC;HMDD	hsa-mir-320a	dbDEMC
hsa-mir-19a	dbDEMC;HMDD	hsa-let-7i	dbDEMC;HMDD
hsa-mir-206	dbDEMC	hsa-mir-214	dbDEMC
hsa-mir-29c	dbDEMC	hsa-let-7d	dbDEMC;HMDD
hsa-mir-593	unconfirmed	hsa-mir-34b	unconfirmed
hsa-mir-150	dbDEMC;HMDD	hsa-mir-133b	dbDEMC;HMDD
hsa-mir-18a	dbDEMC;HMDD	hsa-mir-146b	dbDEMC
hsa-mir-222	dbDEMC;HMDD	hsa-let-7e	dbDEMC;HMDD
hsa-mir-15b	dbDEMC;HMDD	hsa-mir-663a	dbDEMC
hsa-mir-142	HMDD	hsa-mir-200c	HMDD
hsa-mir-223	dbDEMC;HMDD	hsa-mir-148a	dbDEMC;HMDD
hsa-mir-200b	dbDEMC;HMDD	hsa-mir-193a	unconfirmed
hsa-mir-483	HMDD	hsa-mir-574	dbDEMC
hsa-mir-30b	dbDEMC;HMDD	hsa-mir-106a	dbDEMC;HMDD
hsa-mir-34c	unconfirmed	hsa-let-7g	dbDEMC;HMDD
hsa-mir-106b	dbDEMC;HMDD	hsa-mir-335	dbDEMC;HMDD

Table 4 The predicted top 50 miRNAs associated with breast neoplasms and the evidence from databases HMDD V3.0 and dbDEMC

miRNA	Evidence	miRNA	Evidence
hsa-mir-155	dbDEMC;HMDD	hsa-mir-106b	dbDEMC;HMDD
hsa-mir-146a	dbDEMC;HMDD	hsa-mir-143	dbDEMC;HMDD
hsa-mir-328	dbDEMC;HMDD	hsa-let-7c	dbDEMC;HMDD
hsa-mir-29a	dbDEMC;HMDD	hsa-mir-195	dbDEMC;HMDD
hsa-mir-15a	dbDEMC;HMDD	hsa-mir-326	dbDEMC;HMDD
hsa-mir-17	dbDEMC;HMDD	hsa-mir-141	dbDEMC;HMDD
hsa-mir-20a	dbDEMC;HMDD	hsa-let-7i	dbDEMC;HMDD
hsa-mir-145	dbDEMC;HMDD	hsa-let-7d	dbDEMC;HMDD
hsa-mir-34a	dbDEMC;HMDD	hsa-mir-23a	dbDEMC;HMDD
hsa-mir-221	dbDEMC;HMDD	hsa-mir-210	dbDEMC;HMDD
hsa-mir-19a	dbDEMC;HMDD	hsa-mir-34b	dbDEMC;HMDD
hsa-mir-206	dbDEMC;HMDD	hsa-let-7e	dbDEMC;HMDD
hsa-mir-29c	dbDEMC;HMDD	hsa-mir-335	dbDEMC;HMDD
hsa-mir-150	dbDEMC;HMDD	hsa-mir-146b	HMDD
hsa-mir-18a	dbDEMC;HMDD	hsa-mir-214	dbDEMC;HMDD
hsa-mir-142	HMDD	hsa-mir-200c	dbDEMC;HMDD
hsa-mir-15b	dbDEMC;HMDD	hsa-mir-320a	dbDEMC;HMDD
hsa-mir-200b	dbDEMC;HMDD	hsa-let-7g	dbDEMC;HMDD
hsa-mir-222	dbDEMC;HMDD	hsa-mir-133b	dbDEMC;HMDD
hsa-mir-223	dbDEMC;HMDD	hsa-mir-106a	dbDEMC;HMDD
hsa-let-7b	dbDEMC;HMDD	hsa-mir-451a	dbDEMC;HMDD
hsa-mir-34c	dbDEMC;HMDD	hsa-mir-193a	dbDEMC;HMDD
hsa-mir-30b	dbDEMC;HMDD	hsa-mir-663a	HMDD
hsa-mir-200a	dbDEMC;HMDD	hsa-mir-152	dbDEMC;HMDD
hsa-mir-483	unconfirmed	hsa-mir-92b	dbDEMC;HMDD

we found that hsa-mir-155 ranks the first, and it has been found that this miRNA could affect many cancers in recent studies, such as breast neoplasms, colon neoplasms and esophageal neoplasms [41–43].

Overall, the case studies on colon and breast neoplasms further validate the effectiveness of our MTFMDA method for predicting miRNA-disease associations.

Conclusion

Identifying potential miRNA-disease associations could help understand the pathogenesis of the disease from a genetic perspective. In this work, we propose a computational method MTFMDA to predict new MDAs by using an idea of matrix tri-factorization. Different from other matrix completion methods, we factorize the complete MDA matrix to three matrices including a feature matrix for miRNAs, a feature matrix for diseases and a low-rank matrix representing the relationships between miRNA features and disease features. Experiments show that our method performs better for predicting miRNAs associated with new diseases. As we have shown, based on the 5-fold cross validation, the comparisons on the ROC curves, AUCs and Precision-Recall curves show that our MTFMDA performs better than the other methods. Furthermore, the experiments to predict associated miRNAs for colon and breast neoplasms also demonstrate the effectiveness of our method. However, this research only takes the average of two types of similarities of miRNAs and diseases, but not consider how to combine the two similarities optimally. This could be our future topic to work on.

Abbreviations

miRNA: MicroRNA; MDA: MicroRNA disease association; MTFMDA: MicroRNA disease association prediction by matrix tri-factorization; HNC: Head and neck cancer; PPI: Protein-protein interaction network; HMDD V2.0: Human microRNA disease database version 2.0; MeSH: Medical subject headings; CDF: Cumulative distribution function; IRS: Insulin receptor substrate; dbDEMC: Database of differentially expressed miRNAs in human cancers

About this supplement
This article has been published as part of *BMC Genomics Volume 21 Supplement 10, 2020: Selected articles from the 18th Asia Pacific Bioinformatics Conference (APBC 2020): genomics*. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-21-supplement-10.

Authors' contributions
HL designed the optimization algorithms, conducted the experiments, wrote the manuscript. YG and MC wrote the manuscript. LL designed the model and the experiments, and wrote the manuscript. All author(s) revised and approved the manuscript.

Funding
The publication charges for this article were funded partly by the Fundamental Research Funds for the Central Universities in China, and NSFC projects 11631012, Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01). The funding body did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
Human miRNA-disease associations were downloaded from HMDD V2.0 database [11]. MiRNA functional similarity was downloaded from http://www.cuilab.cn. MiRNA sequence data was downloaded from http://www.mirbase.org/ftp.shtml. Disease semantic similarity were downloaded from National Library of Medicine http://www.nlm.nih.gov.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Published: 18 November 2020

References

1. Ines AG, Miska EA. Microrna functions in animal development and human disease. Development. 2005;132(21):4653–62.
2. Ines AG, Miska EA. Micrornas: genomics, biogenesis, mechanism, and function. Cell. 2004;116(21):281–97.
3. Lee RC, Feinbaum RL, Ambros, Victor. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. Cell. 1993;75(5):843–54.
4. Karp X. Encountering micrornas in cell fate signaling. Science. 2005;310(5752):1288–9.
5. Cheng AM. Antisense inhibition of human mirnas and indications for an involvement of mirna in cell growth and apoptosis. Nucleic Acids Res. 2005;33(5752):1290–7.
6. Miska EA. How micrornas control cell division, differentiation and death. Curr Opin Genet Dev. 2005;15(5):563–8.
7. Jopling CL, Minkyung Y, Lancaster AM, Lemon SM, Peter S. Modulation of hepatitis c virus rna abundance by a liver-specific microrna. Science. 2005;309(5740):1577–81.
8. Christine M, Hemann MT, Bartel DP. Disrupting the pairing between let-7 and hmga2 enhances oncogenic transformation. Science. 2007;315(5818):1576–9.
9. Felicity C Kalinowski PACAACGMRERJWPJL Keith M Giles. Regulation of epidermal growth factor receptor signaling and erlotinib sensitivity in head and neck cancer cells by mir-7. Plos ONE. 2012;7(10):e47067–576.
10. Melanie VB, Pandarakalam JJ, Lukas K, Heike L, Jan H, Gabriele B, Katherina W, Dienes HP, Udo E, Ullrich E. Microrna 15a, inversely correlated to pkcalpha, is a potential marker to differentiate between benign and malignant renal tumors in biopsy and urine samples. Am J Pathol. 2012;180(5):1787–97.
11. Yang L, Chengxiang Q, Jian T, Bin G, Jichun Y, Tianzi J, Qinghua C. Hmdd v2.0: a database for experimentally supported human microrna and disease associations. Nucleic Acids Res. 2014;42:D1070–4.
12. Chen X, Liu MX, Yan GY. Rwrmda: predicting novel human microrna-disease associations. Mol BioSyst. 2012;8(10):2792–8.
13. Shi H, Xu J, Zhang G, Xu L, Li C, Wang L, Zhao Z, Jiang W, Guo Z, Li X. Walking the interactome to identify human mirna-disease associations through the functional link between mirna targets and disease genes. BMC Syst Biol. 2013;7(1):101.
14. Ping X, Ke H, Yahong G, Jin L, Xia L, Yingli Z, Zhaogong Z, Jian D. Prediction of potential disease-associated micrornas based on random walk. Bioinformatics. 2015;31(11):1805.
15. Liu Y, Zeng X, He Z, Zou Q. Inferring microrna-disease associations by random walk on a heterogeneous network with multiple data sources. IEEE/ACM Trans Comput Biol Bioinforma. 2017;14(4):1–1.
16. Chen X, Yan GY. Semi-supervised learning for potential human microrna-disease associations inference. Sci Rep. 2014;4:5501.
17. Xia L, Xu J, Li Y. Prioritizing candidate disease miRNAs by topological features in the miRNA-Target dysregulated network. Syst Biol Cancer Res Drug Disc. 2011;10:1857–66.
18. Yu S, Liang C, Xiao Q, Li G, Ding P, Luo J. Mclpmda: A novel method for miRNA-disease association prediction based on matrix completion and label propagation. J Cell Mol Med. 2019;23:1–12.
19. Chen X, Wang L, Qu J, Guan N, Li JQ. Predicting mirna-disease association based on inductive matrix completion. Bioinformatics. 2018;34(24):4256–65.
20. Zhen S, You-Hua Z, Kyungsook H, K. NA, Barry H, De-Shuang H. mirna-disease association prediction with collaborative matrix factorization. Complexity. 2017;2017:1–9.
21. Ha J, Park C, Park S. Pmamca: prediction of microrna-disease association utilizing a matrix completion approach. BMC Syst Biol. 2019;13(1):1–13.
22. Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. Bioinformatics. 2010;26(13):1644–50.
23. Mitra CK, Korla K. Functional, structural, and sequence studies of microrna. Methods Mol Biol. 2014;1107(1107):189.
24. Lipscomb CE. Medical subject headings (mesh). Bull Med Libr Assoc. 2000;88(3):265–6.
25. Xuan P, Han K, Guo M, Guo Y, Li J, Ding J, Liu Y, Dai Q, Li J, Teng Z. Correction: Prediction of micrornas associated with human diseases based on weighted k most similar neighbors. Plos ONE. 2013;8(9):e70204.
26. Toh, Kim-Chuan, Yun, Sangwoon. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. Pac J Optim. 2010;6(3):615–40.
27. Cai JF, Candès EJ, Shen Z. A singular value thresholding algorithm for matrix completion. Siam Jo Optim. 2010;20(4):1956–82.
28. Wen, Zaiwen. A feasible method for optimization with orthogonality constraints. Math Program. 2013;142(1-2):397–434.
29. Gu Cea. Network consistency projection for human mirna-disease associations inference. Sci Rep. 2016;6(24):36054.
30. Chen X YJWC, Li SX. Potential mirna-disease association prediction based on kernelized bayesian matrix factorization. Genomics. 2019;112(1): 809–19.
31. Mordelet F. Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. BMC Bioinformatics. 2011;12(1):389.
32. Nagarajan N, Dhillon IS. Inductive matrix completion for predicting gene-disease associations. Bioinformatics. 2014;30(12):i60–68.
33. Yang Z, Wu L, Wang A, Tang W, Zhao Y, Zhao H, Teschendorff AE. dbdemc 2.0: updated database of differentially expressed mirnas in human cancers. Nucleic Acids Res. 2017;45:D812–8.
34. Zhou Huang YGCCSZJLYZQC JiangchengShi. Hmdd v3.0: a database for experimentally supported human microrna-disease associations. Nucleic Acids Res. 2019;47:D1013–7.
35. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. Ca Cancer J Clin. 2011;61(2):69–90.
36. Hiroko OK, Masashi I, Daisuke K, Yoshitaka H, Yasuhide Y, Koh F, Toshiaki G, Hideki O, Hiroyuki O, Hikaru S. Circulating exosomal micrornas as biomarkers of colon cancer. Plos ONE. 2014;9(4):e92921.
37. Schwartz MK. Enzymes in colon cancer. general information. Cancer. 2015;36(S6):2334–6.
38. Zhang J, Xiao Z, Lai D, Sun J, He C, Chu Z, Ye H, Chen S, Wang J. mir-21, mir-17 and mir-19a induced by phosphatase of regenerating

liver-3 promote the proliferation and metastasis of colon cancer. Br J Cancer. 2012;107(2):352–9.

39. Raquel D, Javier S, García JM, Yolanda L, Vanesa G, Cristina P, et al. Deregulated expression of mir-106a predicts survival in human colon cancer patients. Genes Chromosome Cancer. 2010;47(9):794–802.
40. Bin S, Laura SL, Marco P, Peter L, Tiziana D, Renato B. Micro rna 145 targets the insulin receptor substrate-1 and inhibits the growth of colon cancer cells. J Biol Chem. 2007;282(45):32582–90.
41. Davoren PA, Mcneill RE, Lowery AJ, Kerin MJ, Miller N. Identification of suitable endogenous control genes for microrna gene expression analysis in human breast cancer. BMC Mol Biol. 2008;9(1):76.
42. Sarver AL, French AJ, Borralho PM, Thayanithy V, Oberg AL, Silverstein KA, Morlan BW, Riska SM, Boardman LA, Cunningham JM. Human colon cancer profiles show differential microrna expression depending on mismatch repair status and are characteristic of undifferentiated proliferative states. BMC Cancer. 2009;9(1):401.
43. Yong G, Zhaoli C, Liang Z, Fang Z, Susheng S, Xiaoli F, Baozhong L, Xin M, Xi M, Mingyong L. Distinctive microrna profiles relating to patient survival in esophageal squamous cell carcinoma. Cancer Res. 2008;68(1): 26.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.