

RESEARCH

Open Access



RCMF: a robust collaborative matrix factorization method to predict miRNA-disease associations

Zhen Cui¹, Jin-Xing Liu^{1,2*}, Ying-Lian Gao³, Chun-Hou Zheng² and Juan Wang^{1*}

From 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference

Wuhan and Shanghai, China. 15-18 August 2018, 3-4 November 2018

Abstract

Background: Predicting miRNA-disease associations (MDAs) is time-consuming and expensive. It is imminent to improve the accuracy of prediction results. So it is crucial to develop a novel computing technology to predict new MDAs. Although some existing methods can effectively predict novel MDAs, there are still some shortcomings. Especially when the disease matrix is processed, its sparsity is an important factor affecting the final results.

Results: A robust collaborative matrix factorization (RCMF) is proposed to predict novel MDAs. The $L_{2,1}$ -norm are introduced to our method to achieve the highest AUC value than other advanced methods.

Conclusions: 5-fold cross validation is used to evaluate our method, and simulation experiments are used to predict novel associations on Gold Standard Dataset. Finally, our prediction accuracy is better than other existing advanced methods. Therefore, our approach is effective and feasible in predicting novel MDAs.

Keywords: MiRNA-disease association prediction, $L_{2,1}$ -norm, Collaborative regularization, Matrix factorization

Background

A short class of non-coding RNAs called miRNAs, whose length is generally 19 to 25 nt. They usually regulate gene expression and protein production [1–7]. Since the first two miRNAs lin-4 and let-7 were discovered in 1993 and 2000, respectively [8, 9]. Thousands of miRNAs have been detected by biologists from nematodes to human eukaryotes [10, 11]. The latest miRNA database version miRBase contains 26,845 entries and more than 2000 human miRNAs are detected [12–14]. It is worth noting that with the development of bioinformatics, more researchers are starting to focus on the function of miRNAs. In addition, miRNAs begin to play an important role in biological processes such as proliferation, cell differentiation, viral infection, and signal transduction [15]. Moreover, some miRNAs are closely related to human diseases [16–18].

For example, mir-433 will upregulate the expression of GRB2 in gastric cancer, which is a known tumor-associated protein [19]. And in every pediatric brain tumor type, mir-25, mir-129, and mir-142 are differentially expressed [20]. Considering the strong association between miRNA and disease, all their potential associations should be explored [15, 21]. In medicine, the advantage is that it can promote the diagnosis and treatment of some complex diseases [22–25]. However, predicting MDAs is time-consuming and expensive. Only a few novel associations are discovered and used in clinical medicine each year, and most of the associations are not discovered by researchers. Therefore, it is imminent to improve the accuracy of prediction results.

In previous studies, functionally similar miRNAs always appear in similar diseases [26, 27]. Based on such theory, more and more computational methods and models are proposed for identifying novel miRNA-disease associations (MDAs) [13]. However, these methods have some shortcomings more or less. For example, Jiang et al.

* Correspondence: sdcavell@126.com; wangjuansdu@163.com

¹School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China

Full list of author information is available at the end of the article



proposed an improved disease-gene prediction model [28]. They introduced the principle and use of hypergeometric distribution. And then they analyzed the actual effect in the prediction model. Moreover, different datasets are used to predict novel MDAs, including the known human miRNA-disease data, miRNA functional similarity data and disease semantic similarity data. However, the shortcoming of this model is the excessive dependence on neighbor miRNA data. Chen et al. proposed a method HGIMDA (Heterogeneous Graph Inference miRNA-Disease Association) to predict novel MDAs [29]. It is worth noting the known miRNA-disease associations, miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity for diseases and miRNAs are integrated into this method. The benefit is that the accuracy of the algorithm is improved to some extent. The functional relationship between miRNA targets and disease genes in PPI (Protein-Protein Interaction) networks are considered by researchers. Shi et al. proposed a computational method to predict MDAs by performing random walk [30]. They used PPIs, the miRNA-target interactions and disease-gene associations to identify potential MDAs. However, the model strongly depended on the miRNA-target interactions with high rate of false-positive and high false-negative results [31]. Considering this disadvantage, Chen et al. proposed the RWRMDA (Random Walk with Restart for MiRNA-disease association) model [32]. Their approach was to map all miRNAs to a miRNA functional similarity network. Then, random walk with restart method was implemented until they got stable probability [33]. Finally, all candidate miRNAs will be sorted according to the probability of stability. Moreover, the method was the first global network-based method. Xuan et al. proposed a HDMP method [34]. The method was based on weighted k-nearest-neighbors. The phenotype similarity and semantic similarity between diseases were used to calculate the miRNAs functional similarity matrix. However, the simple ranking of k-nearest-neighbors was not always reliable for prediction. So Chen et al. proposed a new method of ranking-based KNN called RKNNMDA to identify potential MDAs [34]. These previously similarity-based sorted neighbors were re-ranked to get better prediction results. Recently, matrix factorization methods have been used to identify novel MDAs. The advantage is that these methods can better handle missing associations. Shen et al. proposed a matrix factorization model based collaborative matrix factorization to predict novel MDAs [10]. Matrix factorization method takes one input matrix and tries to obtain two other matrices, then the two matrices are multiplied to approximate the input matrix. Gao et al. proposed a dual-network sparse graph regularized matrix factorization method (DNSGRMF) to predict novel MDAs and obtained better experimental results [35]. However,

this method does not necessarily solve the overfitting problem very well. Chen et al. developed a computational model of ELLPMDA (Ensemble Learning and Link Prediction for miRNA-Disease Association) to predict novel MDAs [36]. The miRNA-disease association, miRNA functional similarity, disease semantic similarity and Gaussian profile kernel similarity for miRNAs and diseases were integrated, they used the integrated similarity network and utilized ensemble learning. Three classical algorithms based on similarity are combined to obtain better prediction results. However, even such an excellent method still has some shortcomings, such as excessive ensemble learning will bring more noise. Gao et al. proposed a Nearest Profile-based Collaborative Matrix Factorization (NPCMF) method to predict potential miRNA-disease associations [37]. More importantly, this method has achieved the highest prediction accuracy so far.

In this paper, a simple yet effective matrix factorization model is proposed. Its main function is to predict new MDAs based on existing MDAs. Considering that the missing associations will have a negative impact on the predictions, a pre-processing step is used to solve this problem. The main purpose of this pre-processing method is to try to weight K nearest known neighbors (WKNKN) [38, 39]. It is worth noting that the $L_{2,1}$ -norm is introduced in the collaborative matrix factorization (CMF) method. And the $L_{2,1}$ -norm can avoid over-fitting and eliminate some unattached disease pairs [40, 41]. We also use Gaussian interaction profile kernel similarity to get the network similarity of miRNAs and the network similarity of diseases. Therefore, the final prediction accuracy is greatly improved. Meanwhile, 5-fold cross validation is used to evaluate our experimental results. Our proposed method is superior to other methods. In addition, a simulation experiment is conducted to predict novel associations.

Materials

MDAs dataset

The information about associations between miRNA and disease is obtained from HMDD [42], including 383 diseases, 495 miRNAs and 5430 experimentally confirmed human miRNA-diseases associations. And it is a Gold Standard Dataset. The dataset contains three matrices: $\mathbf{Y} \in \mathbb{R}^{n \times m}$, $\mathbf{S}_m \in \mathbb{R}^{n \times n}$ and $\mathbf{S}_d \in \mathbb{R}^{m \times m}$. In addition, \mathbf{Y} is an adjacency matrix. In the adjacency matrix, there are n miRNAs as rows and m diseases as columns. If miRNA $D(i)$ is associated with disease $d(j)$, the entity $\mathbf{Y}(D(i), d(j))$ is 1, otherwise 0. The matrix \mathbf{Y} is used as the original input matrix. \mathbf{Y} is decomposed into two latent feature matrices, and the product of the two latent feature matrices is used to approximate \mathbf{Y} . Table 1 lists the specific information for the dataset.

MiRNA functional similarity

Considering the assumption that similarly functioning miRNAs have similar diseases, Wang *et.al.* proposed a method for calculating the similarity scores of RNA functions [26]. And the miRNA functional similarity scores are downloaded from [http:// www.cuilab.cn/files/ images/cuilab/misim.zip](http://www.cuilab.cn/files/images/cuilab/misim.zip). The matrix S_m is represented miRNA function similarity network. The functional similarity score between miRNA $m(i)$ and $m(j)$ can be represented $S_m(m(i), m(j))$. The S_m matrix is also used as an input matrix, which represents the functional similarity of miRNA pairs. Among them, each miRNA has a similarity score of 1 to itself.

Disease semantic similarity

In this work, Directed Acyclic Graph (DAG) is used to describe the diseases. $DAG(DD) = (d, T(DD), E(DD))$ is used to describe disease DD , where $T(DD)$ is the node set and $E(DD)$ is the corresponding links set [26]. S_d is represented disease semantic similarity network. The semantic value of disease DD in $DAG(DD)$ formula is defined as:

$$DV1(DD) = \sum_{d \in T(DD)} D1_{DD}(d), \tag{1}$$

$$D1_{DD}(d) = \begin{cases} 1 & \text{if } d = DD \\ \max\{\Delta * D1_{DD}(d') \mid d' \in \text{children of } d\} & \text{if } d \neq DD, \end{cases} \tag{2}$$

where Δ is represented the semantic contribution factor. Generally, the semantic contribution of disease DD to itself is 1. Based on previous research [43], we set Δ to 0.5. It is worth noting that the further the distance between DD and other disease, the smaller the semantic contribution score. Therefore, disease terms contribute the same score to the semantic value of the disease DD in the same layer. Finally, if the two diseases $d(i)$ and $d(j)$ have a larger common part of the DAGs, then the two diseases have a greater similarity score. The disease semantic similarity can be defined as follows:

$$S_d(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} (D1_{d(i)}(t) + D1_{d(j)}(t))}{DV1(d(i)) + DV1(d(j))}, \tag{3}$$

where S_d is the disease semantic similarity matrix. In addition, the S_d matrix is also used as an input matrix with Y and S_m . Similar to the S_m matrix, each disease

has its own semantic similarity score of 1. Therefore, the two feature matrices decomposed by Y are controlled by the S_m matrix and the S_d matrix.

Methodology

Problem formalization

Formally, the known associations $Y(m(i), d(j))$ of miRNA $m(i)$ associated with disease $d(j)$ are considered to be a matrix factorization model. First, the input associations matrix Y is decomposed into two low rank latent feature matrices A (for miRNAs) and B (for diseases). Then, some constraints are added to the two low rank matrices [44]. Specifically, the $L_{2,1}$ -norm is added to the latent feature matrix B (for diseases). Finally, the specific matrices of A and B are obtained by using some update rules. It is worth noting that we need a prediction matrix that is derived from the product of A and B . Considering the stronger association of miRNAs with diseases, the correlation score between them is higher. So, the miRNA-disease pairs $Y(m(i), d(j))$ are ranked from high to low.

Robust collaborative matrix factorization (RCMF)

The traditional CMF is an effective method for predicting novel MDAs [10]. Collaborative filtering is used by CMF. The objective function of CMF is given as follows:

$$\min_{A,B} = \|Y - AB^T\|_F^2 + \lambda_l (\|A\|_F^2 + \|B\|_F^2) + \lambda_d \|S_m - AA^T\|_F^2 + \lambda_t \|S_d - BB^T\|_F^2, \tag{4}$$

where $\|\cdot\|_F$ is Frobenius norm, λ_b , λ_d and λ_t are non-negative parameters.

However, although the B matrix is a low rank matrix, it is not sparse enough. In fact, B is indeed sparse. But we want to get the B matrix better, we use the $L_{2,1}$ -norm to constrain the latent feature matrix B of the disease. Because the $L_{2,1}$ -norm can achieve row sparse, the $L_{2,1}$ -norm can better remove the meaningless elements of the B matrix. For matrix B , overfitting problems may be generated to reduce the accuracy of the prediction in predicting novel MDAs.

Therefore, to overcome this problem, a robust collaborative matrix factorization method named RCMF is proposed to predict MDAs. The $L_{2,1}$ -norm is introduced to the RCMF method to solve over-fitting problems [45, 46]. In this paper, the dataset used in the experiment, the number of diseases is less than the number of miRNAs, we are more concerned about which miRNAs are likely to be associated with the diseases. Therefore, we apply the $L_{2,1}$ -norm on the potential feature matrix B of the disease to make the B matrix sparse. The advantage is that more miRNAs can be accurately matched to the disease to improve the accuracy of prediction. The interaction matrix Y is

Table 1 MiRNAs, Diseases, and Associations in Gold Standard Dataset

Datasets	MiRNAs	Diseases	Associations
Gold Standard Dataset	495	383	5430

decomposed into two matrices \mathbf{A} and \mathbf{B} , where $\mathbf{AB}^T \approx \mathbf{Y}$. RCMF uses two collaborative regularization terms to constrain \mathbf{A} and \mathbf{B} . Specifically, these two regularization terms require similar miRNAs or diseases potential feature vectors to be similar, and dissimilar miRNAs or diseases potential feature vectors are not similar, respectively [38]. Where $\mathbf{S}_m \approx \mathbf{AA}^T$ and $\mathbf{S}_d \approx \mathbf{BB}^T$. Therefore, the objective function of RCMF can be written as:

$$\min_{\mathbf{A}, \mathbf{B}} = \|\mathbf{Y} - \mathbf{AB}^T\|_F^2 + \lambda_l (\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2) + \lambda_l \|\mathbf{B}\|_{2,1} + \lambda_d \|\mathbf{S}_m - \mathbf{AA}^T\|_F^2 + \lambda_t \|\mathbf{S}_d - \mathbf{BB}^T\|_F^2, \tag{5}$$

where $\|\cdot\|_F$ is Frobenius norm, $\|\cdot\|_{2,1}$ is $L_{2,1}$ -norm, λ_l , λ_d and λ_t are non-negative parameters. Based on previous research [38], the grid search method is used to perform the selection of optimal parameters, where $\lambda_l \in \{2^{-2}, 2^{-1}, 2^0, 2^1\}$ and $\lambda_d/\lambda_t \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. In order to find the latent feature matrices \mathbf{A} and \mathbf{B} , an approximate model of the matrix \mathbf{Y} is constructed in the first term. In the second term, the Tikhonov regularization can minimize the norms of \mathbf{A} , \mathbf{B} . The $L_{2,1}$ -norm is applied on \mathbf{B} in the third term. And this advantage is able to increase the sparsity of the disease matrix and eliminate undesired disease pairs. The last two regularization terms represent the minimization of squared error between \mathbf{S}_m (\mathbf{S}_d) and \mathbf{AA}^T (\mathbf{BB}^T).

Initialization of A and B

\mathbf{A} and \mathbf{B} are initialized to use the SVD (Singular Value Decomposition) method for the input MDAs matrix \mathbf{Y} . The initialization formula can be written as:

$$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \text{SVD}(\mathbf{Y}, k), \mathbf{A} = \mathbf{US}_k^{1/2}, \mathbf{B} = \mathbf{VS}_k^{1/2}, \tag{6}$$

where \mathbf{S}_k is a diagonal matrix, which contains the k largest singular values.

Optimization algorithm

\mathbf{A} and \mathbf{B} are updated using least squares in this study. First, F is represented as the objection function of RCMF method. Then, $\partial F/\partial \mathbf{A}$ and $\partial F/\partial \mathbf{B}$ are set to be 0, respectively. \mathbf{A} and \mathbf{B} are continued to use the least squares until convergence. Figure 1 shows the convergence of the RCMF method. We perform the RCMF method on the dataset used in the experiment, where the x-axis represents the number of iterations and the y-axis represents the error. As can be seen from Fig. 1, after 50 iterations, the curve begins to converge on a straight line, which proves that our method begins to converge after 50 iterations. In addition, λ_l , λ_d and λ_t are automatically determined. The optimal parameter values are obtained when cross validating the training set. The update rules of \mathbf{A} and \mathbf{B} can be written as:

$$\mathbf{A} = (\mathbf{YB} + \lambda_d \mathbf{S}_m \mathbf{A})(\mathbf{B}^T \mathbf{B} + \lambda_l \mathbf{I}_k + \lambda_d \mathbf{AA}^T)^{-1}, \tag{7}$$

$$\mathbf{B} = (\mathbf{Y}^T \mathbf{A} + \lambda_t \mathbf{S}_d \mathbf{B})(\mathbf{A}^T \mathbf{A} + \lambda_l \mathbf{I}_k + \lambda_t \mathbf{B}^T \mathbf{B} + \lambda_t \mathbf{DI}_k)^{-1}, \tag{8}$$

where \mathbf{D} is a diagonal matrix with the i -th diagonal element as $d_{ii} = 1/2\|(\mathbf{B})^i\|_2$. Based on these update rules, we first calculate the maximum time complexity required to perform the iterative steps, and then we conclude that the final time complexity of RCMF method is $O(nmk)$, where n is the number of miRNAs, m is the number of diseases and k is the

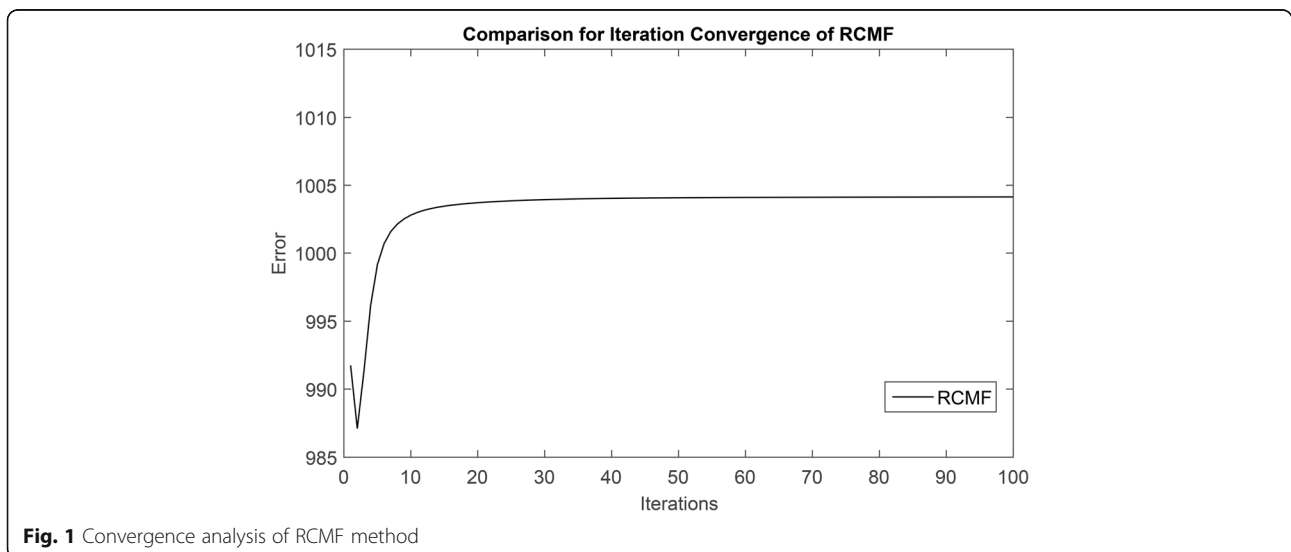


Fig. 1 Convergence analysis of RCMF method

number of singular values in the SVD. Therefore, the algorithm of RCMF is as follows:

Algorithm 1: RCMF	
Input:	MDA matrix $\mathbf{Y} \in R^{m \times n}$, miRNA similarity \mathbf{S}_m , and disease similarity \mathbf{S}_d
Parameters:	$K, p, k, \lambda_1, \lambda_2, \lambda_3$
Output:	prediction score matrix $\hat{\mathbf{Y}}$
Pre-processing:	$\mathbf{Y} = WKNKN(\mathbf{Y}, \mathbf{S}_m, \mathbf{S}_d, K, p)$
Initialization:	$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \text{SVD}(\mathbf{Y}, k), \mathbf{A} = \mathbf{U}\mathbf{S}_k^{1/2}, \mathbf{B} = \mathbf{V}\mathbf{S}_k^{1/2}$
Repeat	
Update \mathbf{A}	using Eq. (7)
Update \mathbf{B}	using Eq. (8)
Until convergence	
	$\hat{\mathbf{Y}} = \mathbf{A}\mathbf{B}$
Return	$\hat{\mathbf{Y}}$

Results

Cross validation experiments

In this study, our experiments are compared to the previous advanced methods CMF [10], HDMP [33], WBSMDA [47], MKRMDA [31], HAMDA [48] and ELLPMDA [36]. For each method, 5-fold cross validation is conducted 100 times. However, the WKNKN pre-processing steps is performed before running our method. This can solve the problem of missing unknown values. At the same time, it can also improve the accuracy of prediction to some extent.

In general, AUC (Area Under the Curve) is used as a reasonable indicator when evaluating the predictive performance of a method. The popular indicator of AUC is also used to evaluate our approach in this study. The area under the ROC (Receiver Operating Characteristic) curve is considered to be AUC. In other words, the value of this area will not be greater than 1. AUC values between 0.5 and 1 are normal and reasonable. Once below 0.5, the method will have no meaning at all. Before running cross validation, the miRNA-disease pairs are randomly removed in the input MDAs matrix \mathbf{Y} . Doing this is a comprehensive assessment of our approach by increasing the

Table 2 AUC Results of cross validation experiments

Methods	Gold Standard Dataset
WBSMDA	0.8185(0.0009)
HDMP	0.8342(0.0010)
CMF	0.8697(0.0011)
MKRMDA	0.8894(0.0015)
HAMDA	0.8965 (0.0012)
ELLPMDA	0.9193(0.0002)
RCMF	0.9345(0.0004)

Table 3 Predicted MiRNAs for Esophageal Neoplasms

Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	hsa-let-7a	known	16	hsa-mir-145	known
2	hsa-mir-100	known	17	hsa-mir-146a	known
3	hsa-mir-130a	known	18	hsa-mir-148a	known
4	hsa-let-7c	known	19	hsa-mir-617	known
5	hsa-mir-192	known	20	hsa-mir-758	known
6	hsa-mir-19a	known	21	hsa-mir-342	known
7	hsa-mir-21	known	22	hsa-mir-34a	known
8	hsa-mir-150	known	23	hsa-mir-34b	known
9	hsa-mir-205	known	24	hsa-mir-296	known
10	hsa-mir-22	known	25	hsa-mir-29c	known
11	hsa-mir-223	known	26	hsa-mir-215	dbDEMC
12	hsa-mir-25	known	27	hsa-mir-421	dbDEMC
13	hsa-mir-26a	known	28	hsa-mir-184	dbDEMC
14	hsa-mir-27a	known	29	hsa-mir-519a	Unconfirmed
15	hsa-mir-28	known	30	hsa-mir-610	Unconfirmed

difficulty of prediction [49]. This way is called CV-p (Cross Validation pairs).

Association prediction under CV-p

Table 2 lists the experimental results at CV-p. The AUC average of 100 times 5-fold cross validation is used as the final AUC score. It is worth noting that AUC is known to be insensitive to skewed class distributions [50]. The gold standard miRNA disease dataset is highly unbalanced in this study. One problem is that there are more negative factors than positive ones. Thus, AUC is a more suitable measure for other methods. As shown in

Table 4 Predicted MiRNAs for Liver Neoplasms

Rank	miRNA	Evidence
1	hsa-mir-372	known
2	hsa-mir-486	known
3	hsa-mir-10b	known
4	hsa-mir-122	known
5	hsa-mir-133b	known
6	hsa-mir-200a	known
7	hsa-mir-148b	known
8	hsa-mir-21	known
9	hsa-let-7b	known
10	hsa-mir-629	known
11	hsa-mir-24	known
12	hsa-mir-34c	known
13	hsa-mir-200b	dbDEMC
14	hsa-mir-15b	dbDEMC
15	hsa-mir-183	dbDEMC

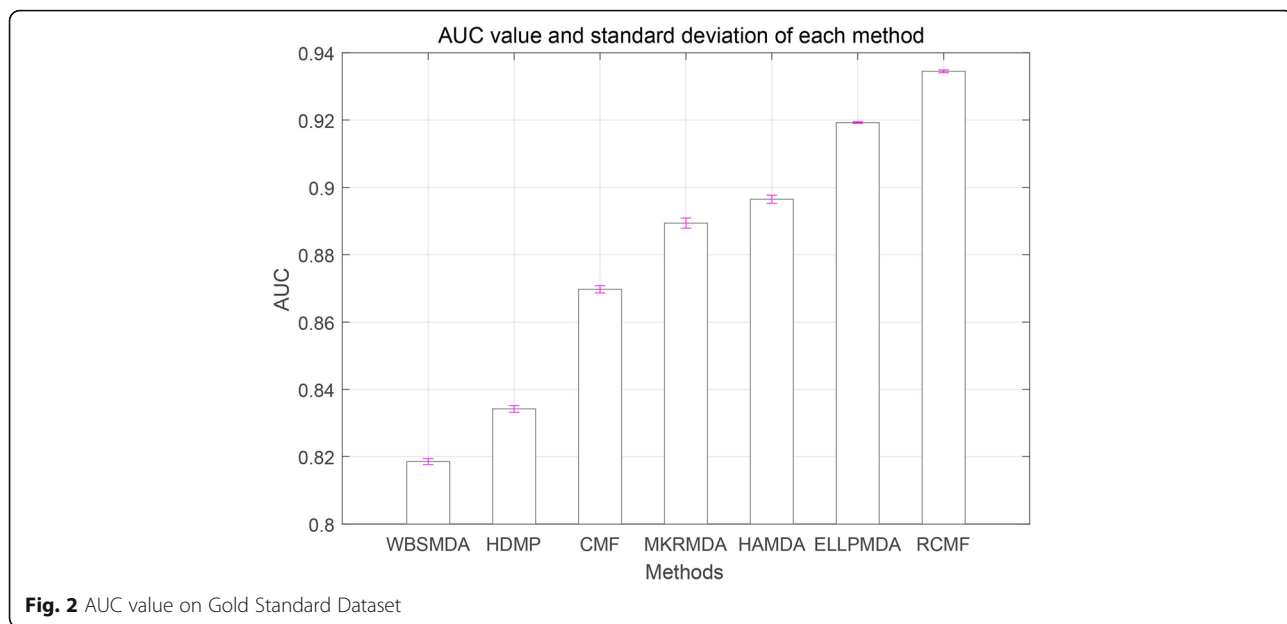
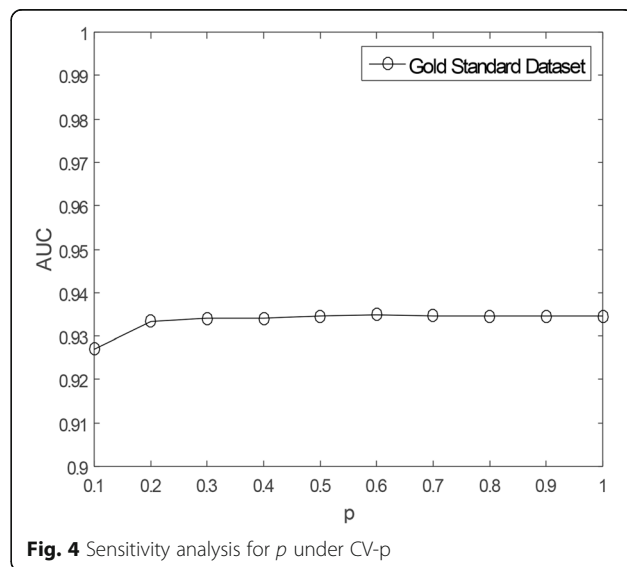
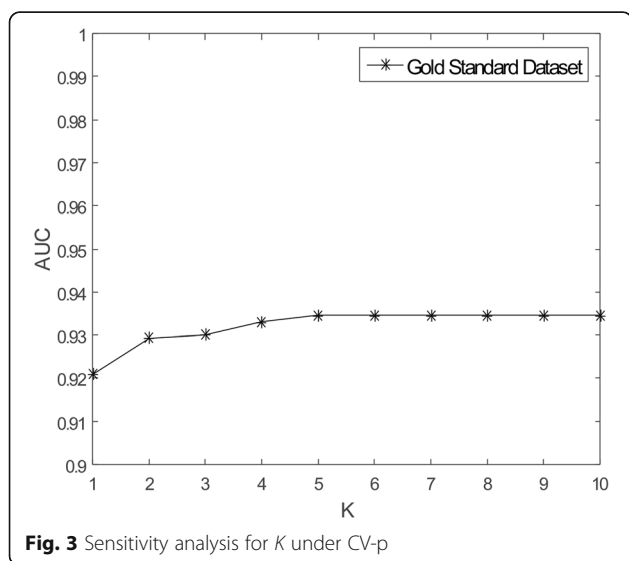


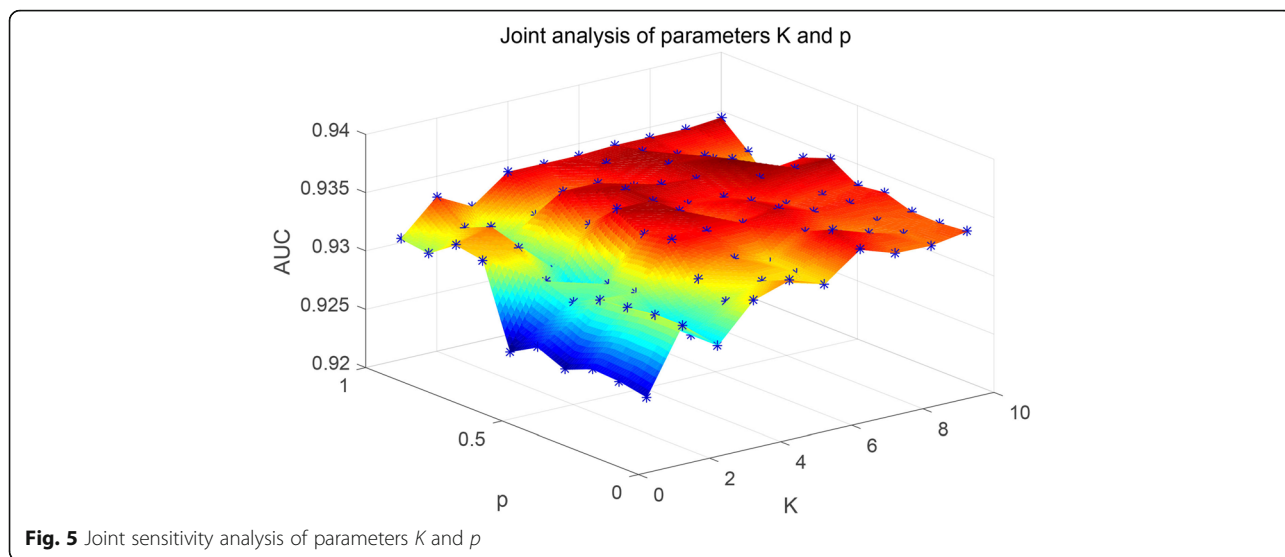
Table 2, the AUC values for each method are counted, including the highest AUC in bold, and standard deviations are given in (parentheses).

As listed in Table 2, our proposed method RCMF achieves an AUC of 0.9345 on Gold Standard Dataset, which is 1.52% higher than ELLPMDA with an AUC of 0.9193. The AUC value of the WBSMDA method is the lowest, and our method is 11.6% higher than it. Also, our method is 6.48% higher than the traditional CMF method. Therefore, our proposed is better than other existing methods. Figure 2 visually shows the AUC level of each method.

Comprehensive prediction for novel MDAs

A simulation experiment is conducted in this subsection. Two cases are tested by our method, one is Esophageal Neoplasms, the other is Liver Neoplasms. Esophageal Neoplasms is very common in many areas of China, especially in northern China [51]. More information about the disease are published in <http://www.omim.org/entry/133239>. For Esophageal Neoplasms, the 30 miRNAs associated with it are removed. Then, the simulation is conducted to get the final prediction score matrix. Based on the predicted scores for this disease, the miRNAs associated with this disease are ranked from





high to low. At the same time, whether the removed miRNA is successfully predicted and the novel associations are also needs to be counted. Twenty known associations are successfully predicted and five novel associations are predicted. Among the unknown associations, the three of five unknown associations are confirmed by the dbDEMC [52]. It is worth noting that hsa-mir-215 has the highest correlation with Esophageal Neoplasms. About hsa-mir-215, Fassan et al. have discovered this miRNA in 2011 related to Esophageal Neoplasms. They performed qRT-PCR and ISH analyses on two independent series of endoscopic biopsies (qRT-PCR) and esophagectomy specimens (ISH) [53]. In particular, hsa-mir-215 is significantly overexpressed during the pathogenesis of Esophageal Neoplasms. About hsa-mir-184, Kojima et al. discovered this miRNA in 2015 related to Esophageal Neoplasms. They conducted miRNA expression analysis by microarray [54].

By comparing Esophageal Neoplasms with normal samples, hsa-mir-184 is under-expressed in diseased samples. Table 3 lists the experimental results. And the known associations are in bold.

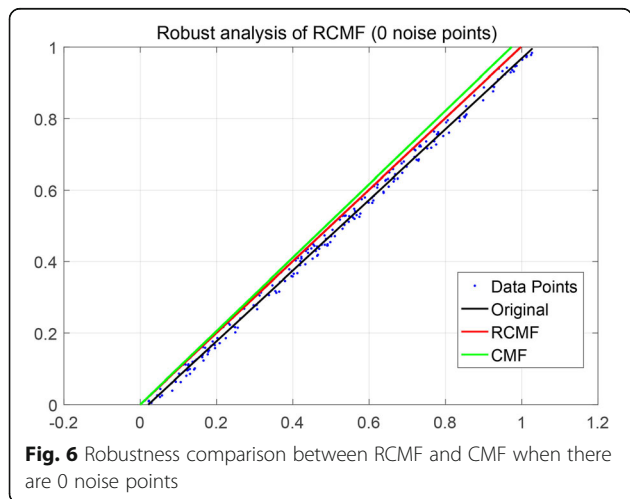
Another case is Liver Neoplasms. It is the fifth most common cancer and the third most common cause of death from cancer worldwide [55]. More information about the disease are published in <http://www.omim.org/entry/114550>. For Liver Neoplasms, fifteen miRNAs associated with it are removed from the dataset while running our method. Then based on the predicted scores for this disease, the miRNAs associated with this disease are ranked from high to low. Finally, twelve known associations are successfully predicted. At the same time, three novel associations are predicted. And, all three are confirmed by dbDEMC. About hsa-mir-200b, hsa-mir-15b and hsa-mir-183, Naoki et al. have discovered this miRNA in 2012 related to Liver Neoplasms. In particular, hsa-mir-200b, hsa-mir-15b, and hsa-mir-183 are significantly overexpressed during the pathogenesis of Liver Neoplasms [56]. Table 4 lists the experimental results.

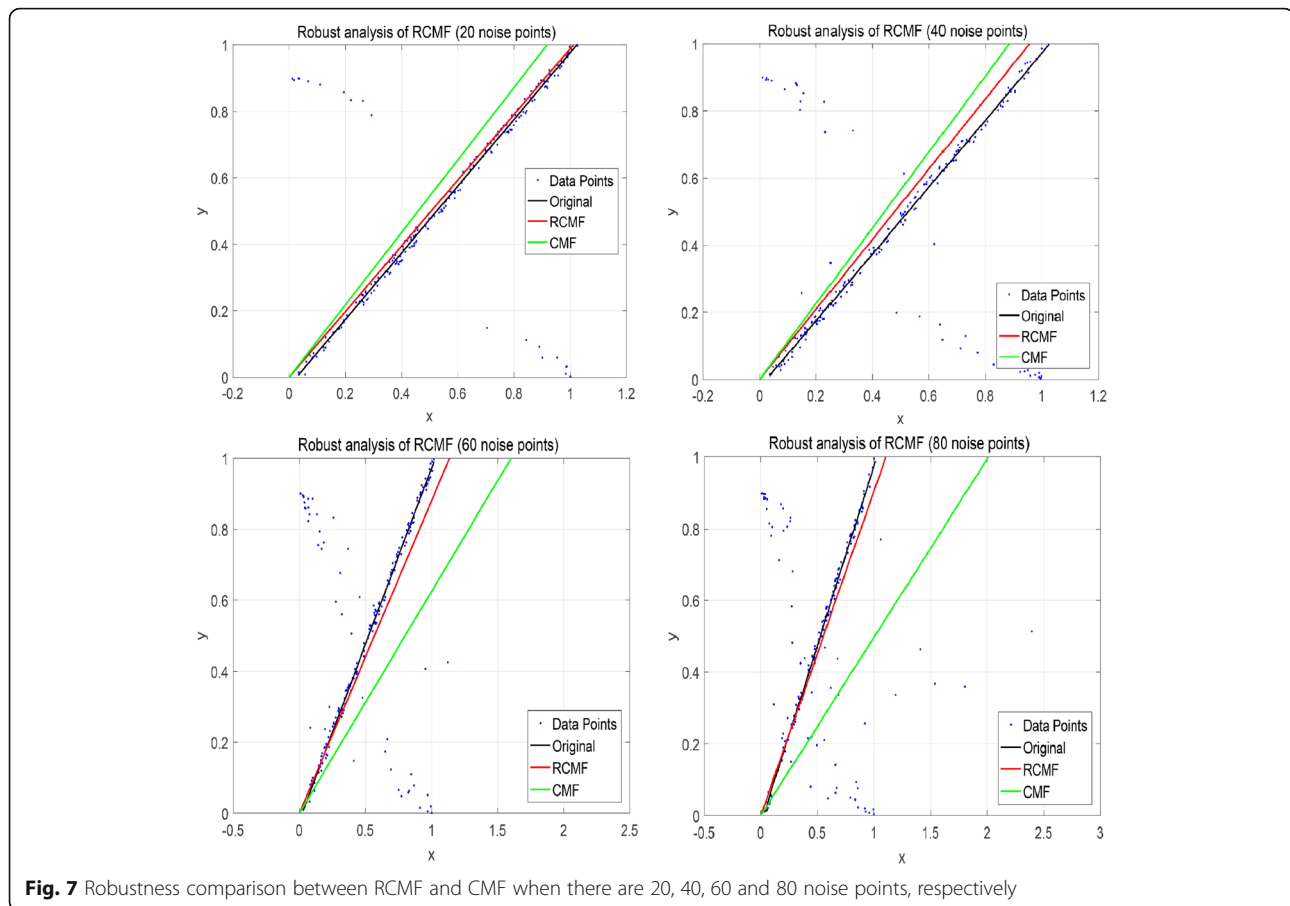
According to the above simulation results, most known miRNAs are predicted. At the same time, some unknown miRNAs are also confirmed by dbDEMC. Therefore, our method can be used to predict novel MDAs and achieve excellent predictions.

Discussion

Sensitivity analysis from WKNKN

As mentioned earlier in this study, there are some missing unknown associations in the matrix Y , so WKNKN method is used to minimize the error. K represents the number of nearest known neighbors and p represents a decay term where $p \leq 1$. Before running RCMF method,





the parameters K and p will be fixed. The sensitivity analysis of these two parameters is given by Figs. 3 and 4, respectively. It can be clearly seen from the figures that when $K=5$, $p=0.7$, the AUC tends to be stable. Furthermore, to more fully verify the sensitivity of these two parameters to AUC, their joint sensitivity analysis is shown in Fig. 5.

Robust analysis of our method

The $L_{2,1}$ -norm can increase the robustness of the algorithm. This is mainly reflected in the distinction between outliers in the dataset. In this section, we use a simulation dataset of 200 data points to verify the robustness of the algorithm. To illustrate RCMF's ability to learn a subspace, we apply RCMF on a synthetic dataset composed of 200 two-dimensional data points. It is worth noting that all data points are distributed in a one-dimensional subspace, i.e., a straight line ($y=x$). In addition, both RCMF and CMF are applied to the synthetic data set for comparison. Specifically, we add different numbers of noise points to the simulation dataset to compare RCMF and CMF. Figures 6 and 7 show the data distribution of 0 noise points, 20 noise points, 40 noise points, 60 noise points and 80 noise points, respectively. As can be seen from Fig. 6, both

RCMF and CMF remain stable when there are no noise points in the dataset. It can be seen from Fig. 7 that as the noise point increases, the CMF cannot continue to maintain stability but gradually shifts. It is worth noting that RCMF can still maintain the same state as the original data point due to the $L_{2,1}$ -norm. Even if the number of noise points is constantly increasing, RCMF is still unaffected by outliers. This proves that RCMF is robust.

Conclusions

Abnormal expression of miRNA has a crucial impact in the development of complex human diseases. More and more diseases are confirmed by biologists to have a close relationship with miRNAs. In this paper, a novel computational model is proposed to predict MDAs. The most valuable contribution is that the $L_{2,1}$ -norm is added to the CMF. AUC value is used as a reliable indicator to evaluate our approach. Meanwhile, the excellent results are generated by our method.

More importantly, WKNKN is used as a pre-processing method. This step plays a crucial role in predicting MDAs. The best predictions are achieved by dealing with missing unknown associations.

In the future, more and more novel MDAs will be predicted and more datasets will be available. At the same time, more valuable MDA information will be published in public databases. In fact, there are many other methods to predict MDAs. RCMF is hoped to be helpful for MDA prediction and relevant miRNA research from the computational biology. In future work, we will continue to study more effective methods to predict novel MDAs.

Abbreviations

AUC: Area under the curve; CMF: Collaborative matrix factorization method; CV: Cross-validation; CV-p: Cross validation pairs; MDAs: MIRNA-disease associations; RCMF: Robust collaborative matrix factorization method; SVD: Singular value decomposition

Acknowledgements

Not applicable.

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 20 Supplement 25, 2019: Proceedings of the 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference: bioinformatics*. The full contents of the supplement are available online at <https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-25>.

Authors' contributions

ZC and JXL jointly contributed to the design of the study. YLG designed and implemented the RCMF method, performed the experiments, and drafted the manuscript. CHZ participated in the design of the study and performed the statistical analysis. JW contributed to the data analysis. All authors read and approved the final manuscript.

Funding

Publication costs are funded by the NSFC under grant Nos. 61872220, and 61572284; by the Co-Innovation Center for Information Supply & Assurance Technology, Anhui University under grant No. ADXXBZ201704.

Availability of data and materials

The datasets that support the findings of this study are available in <https://github.com/cuizhensdws/L21-GRMF>.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China. ²Co-Innovation Center for Information Supply & Assurance Technology, Anhui University, Hefei 230601, China. ³Qufu Normal University Library, Qufu Normal University, Rizhao 276826, China.

Published: 24 December 2019

References

- Ambros V. microRNAs: tiny regulators with great potential. *Cell*. 2001;107(7):823–6.
- Ambros V. The functions of animal microRNAs. *Nature*. 2004;431(7006):350.
- Ibrahim R, Yousri NA, Ismail MA, El-Makky NM. miRNA and gene expression based cancer classification using self-learning and co-training approaches. *arXiv preprint arXiv:14014589*; 2014.
- Katayama Y, Maeda M, Miyaguchi K, Nemoto S, Yasen M, Tanaka S, Mizushima H, Fukuoka Y, Arai S, Tanaka H. Identification of pathogenesis-related microRNAs in hepatocellular carcinoma by expression profiling. *Oncol Lett*. 2012;4(4):817–23.
- Meister G, Tuschl T. Mechanisms of gene silencing by double-stranded RNA. *Nature*. 2004;431(7006):343.
- Deng S-P, Zhu L, Huang D-S. Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks. *BMC genomics*. 2015;16(Suppl 3):S4.
- Hu Y, Liu J-X, Gao Y-L, Li S-J, Wang J. Differentially expressed genes extracted by the tensor robust principal component analysis (TRPCA) method. *Complexity*. 2019;2019:1-13.
- Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*. 1993;75(5):843–54.
- Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G. The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*. 2000;403(6772):901.
- Shen Z, Zhang Y-H, Han K, Nandi AK, Honig B, Huang D-S. miRNA-disease association prediction with collaborative matrix factorization. *Complexity*. 2017;2017:1-9.
- Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, Cui Q. An analysis of human microRNA and disease associations. *PLoS One*. 2008;3(10):e3420.
- Chen X, Yan G-Y. Semi-supervised learning for potential human microRNA-disease associations inference. *Sci Rep*. 2014;4:5501.
- Bandyopadhyay S, Mitra R, Maulik U, Zhang MQ. Development of the human cancer microRNA network. *Silence*. 2010;1(1):6.
- Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*. 2013;42(D1):D68–73.
- Yi H-C, You Z-H, Huang D-S, Li X, Jiang T-H, Li L-P. A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information. *Mol Ther Nucleic Acids*. 2018;11:337–44.
- Peng C, Zou L, Huang D-S. Discovery of relationships between long non-coding RNAs and genes in human diseases based on tensor completion. *IEEE Access*. 2018;6:59152–62.
- Bao W, Jiang Z, Huang D-S. Novel human microbe-disease association prediction using network consistency projection. *BMC Bioinformatics*. 2017;18(16):543.
- Huang D-S, Zheng C-H. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics*. 2006;22(15):1855–62.
- Luo H, Zhang H, Zhang Z, Zhang X, Ning B, Guo J, Nie N, Liu B, Wu X. Down-regulated miR-9 and miR-433 in human gastric carcinoma. *J Exp Clin Cancer Res*. 2009;28(1):82.
- Patel V, Williams D, Hajarnis S, Hunter R, Pontoglio M, Somlo S, Igarashi P. miR-17~92 miRNA cluster promotes kidney cyst growth in polycystic kidney disease. *Proc Natl Acad Sci*. 2013;110(26):10765–70.
- Cui Z, Liu J-X, Gao Y-L, Zhu R, Yuan S-S. LncRNA-disease associations prediction using bipartite local model with nearest profile-based association inferring. *IEEE journal of biomedical and health informatics*; 2019.
- Zheng C-H, Huang D-S, Zhang L, Kong X-Z. Tumor clustering using nonnegative matrix factorization with gene selection. *IEEE Trans Inf Technol Biomed*. 2009;13(4):599–607.
- Sethupathy P, Collins FS. MicroRNA target site polymorphisms and human disease. *Trends Genet*. 2008;24(10):489–97.
- Li J, Liu Y, Xin X, Kim TS, Cabeza EA, Ren J, Nielsen R, Wrana JL, Zhang Z. Evidence for positive selection on a number of microRNA regulatory interactions during recent human evolution. *PLoS Genet*. 2012;8(3):e1002578.
- Yu N, Gao Y-L, Liu J-X, Shang J, Zhu R, Dai L-Y. Co-differential gene selection and clustering based on graph regularized multi-view NMF in cancer genomic data. *Genes*. 2018;9(12):586.
- Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microRNA functional network based on microRNA-associated diseases. *Bioinformatics*. 2010;26(13):1644–50.
- Gao M-M, Cui Z, Gao Y-L, Li F, Liu J-X. Dual Sparse Collaborative Matrix Factorization Method Based on Gaussian Kernel Function for Predicting LncRNA-Disease Associations. In: *International Conference on Intelligent Computing*. Cham, Springer; 2019. p. 318–26.
- Jiang Q, Hao Y, Wang G, Juan L, Zhang T, Teng M, Liu Y, Wang Y. Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst Biol*. 2010;4(1):S2.

29. Chen X, Yan CC, Zhang X, You Z-H, Huang Y-A, Yan G-Y. HGMDA: heterogeneous graph inference for miRNA-disease association prediction. *Oncotarget*. 2016;7(40):65257.
30. 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.
31. Chen X, Niu Y-W, Wang G-H, Yan G-Y. MKRMDA: multiple kernel learning-based Kronecker regularized least squares for MiRNA-disease association prediction. *J Transl Med*. 2017;15(1):251.
32. Chen X, Liu MX, Yan GY. RWRMDA: predicting novel human microRNA-disease associations. *Mol BioSyst*. 2012;8(10):2792–8.
33. Xuan P, Han K, Guo M, Guo Y, Li J, Ding J, Liu Y, Dai Q, Li J, Teng Z. Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS One*. 2013;8(8):e70204.
34. Chen X, Wu Q-F, Yan G-Y. RKNMMDA: ranking-based KNN for MiRNA-disease association prediction. *RNA Biol*. 2017;14(7):952–62.
35. Gao M-M, Cui Z, Gao Y-L, Liu J-X, Zheng C-H. Dual-network sparse graph regularized matrix factorization for predicting miRNA-disease associations. *Molecular omics*. 2019;15(2):130–7.
36. Chen X, Zhou Z, Zhao Y. ELLPMDA: ensemble learning and link prediction for miRNA-disease association prediction. *RNA Biol*. 2018;15(6):807–18.
37. Gao Y-L, Cui Z, Liu J-X, Wang J, Zheng C-H. NPCMF: nearest profile-based collaborative matrix factorization method for predicting miRNA-disease associations. *BMC Bioinformatics*. 2019;20(1):353.
38. Ezzat A, Zhao P, Wu M, Li X-L, Kwok C-K. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*. 2017;14(3):646–56.
39. Yin M-M, Cui Z, Gao M-M, Liu J-X, Gao Y-L. LWPCMF: logistic weighted profile-based collaborative matrix factorization for predicting MiRNA-disease associations. *IEEE/ACM transactions on computational biology and bioinformatics*; 2019.
40. Cui Z, Gao Y-L, Liu J-X, Wang J, Shang J, Dai L-Y. The computational prediction of drug-disease interactions using the dual-network L 2, 1-CMF method. *BMC Bioinformatics*. 2019;20(1):5.
41. Liu J-X, Wang D-Q, Zheng C-H, Gao Y-L, Wu S-S, Shang J-L. Identifying drug-pathway association pairs based on L 2, 1-integrative penalized matrix decomposition. *BMC Syst Biol*. 2017;11(6):119.
42. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q. HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res*. 2013;42(D1):D1070–4.
43. 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 Bioinform*. 2017;14(4):905–15.
44. Yuan L, Zhu L, Guo W-L, Zhou X, Zhang Y, Huang Z, Huang D-S. Nonconvex penalty based low-rank representation and sparse regression for eQTL mapping. *IEEE/ACM Trans Comput Biol Bioinform (TCBB)*. 2017;14(5):1154–64.
45. Cui Z, Gao Y-L, Liu J-X, Dai L-Y, Yuan S-S. L 2, 1-GRMF: an improved graph regularized matrix factorization method to predict drug-target interactions. *BMC bioinformatics*. 2019;20(8):287.
46. Yu N, Gao Y-L, Liu J-X, Wang J, Shang J. Hypergraph regularized NMF by L 2, 1-norm for Clustering and Com-abnormal Expression Genes Selection. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Madrid;IEEE. 2018;578–82.
47. Chen X, Yan CC, Zhang X, You Z-H, Deng L, Liu Y, Zhang Y, Dai Q. WBSMDA: within and between score for MiRNA-disease association prediction. *Sci Rep*. 2016;6:21106.
48. Chen X, Niu Y-W, Wang G-H, Yan G-Y. HAMDA: hybrid approach for MiRNA-disease association prediction. *J Biomed Inform*. 2017;76:50–8.
49. Ezzat A, Wu M, Li X-L, Kwok C-K. Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Brief Bioinform*. 2018;8:1337–57.
50. Ezzat A, Wu M, Li X-L, Kwok C-K. Drug-target interaction prediction via class imbalance-aware ensemble learning. *BMC Bioinformatics*. 2016;17(19):509.
51. Wang L-D, Zhou F-Y, Li X-M, Sun L-D, Song X, Jin Y, Li J-M, Kong G-Q, Qi H, Cui J. Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies a susceptibility locus at PLCE1. *Nat Genet*. 2010;42(9):759.
52. Yang Z, Ren F, Liu C, He S, Sun G, Gao Q, Yao L, Zhang Y, Miao R, Cao Y. dbDEMOC: a database of differentially expressed miRNAs in human cancers. *BMC Genomics*. 2010;11(Suppl 4):S5.
53. Fassan M, Volinia S, Palatini J, Pizzi M, Baffa R, De Bernard M, Battaglia G, Parente P, Croce CM, Zaninotto G. MicroRNA expression profiling in human Barrett's carcinogenesis. *Int J Cancer*. 2011;129(7):1661–70.
54. Kojima M, Sudo H, Kawachi J, Takizawa S, Kondou S, Nobumasa H, Ochiai A. MicroRNA markers for the diagnosis of pancreatic and biliary-tract cancers. *PLoS One*. 2015;10(2):e0118220.
55. Taniguchi K, Roberts LR, Aderca IN, Dong X, Qian C, Murphy LM, Nagorney DM, Burgart LJ, Roche PC, Smith DI. Mutational spectrum of β -catenin, AXIN1, and AXIN2 in hepatocellular carcinomas and hepatoblastomas. *Oncogene*. 2002;21(31):4863.
56. Oishi N, Kumar MR, Roessler S, Ji J, Forgues M, Budhu A, Zhao X, Andersen JB, Ye QH, Jia HL. Transcriptomic profiling reveals hepatic stem-like gene signatures and interplay of miR-200c and epithelial-mesenchymal transition in intrahepatic cholangiocarcinoma. *Hepatology*. 2012;56(5):1792–803.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

