Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MCCMF: Collaborative matrix factorization based on matrix completion for predicting miRNA-disease associations

Tian-Ru Wu
Qufu Normal University - Rizhao Campus
Meng-Meng Yin
Qufu Normal University - Rizhao Campus
Cui-Na Jiao
Qufu Normal University - Rizhao Campus
Jin-Xing Liu (sdcavell@126.com)
Qufu Normal University - Rizhao Campus https://orcid.org/0000-0001-6104-2149
Ying-Lian Gao
Qufu Normal University - Rizhao Campus
Xiang-Zhen Kong
Qufu Normal University - Rizhao Campus

Research article

Keywords: MiRNA-disease association prediction, Matrix completion, Weight K nearest known neighbors, Matrix factorization

Posted Date: June 22nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-36602/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published on October 14th, 2020. See the published version at https://doi.org/10.1186/s12859-020-03799-6.

MCCMF: Collaborative matrix factorization based on matrix completion for predicting miRNA-disease associations

Tian-Ru Wu, Meng-Meng Yin, Cui-Na Jiao, Ying-Lian Gao, Xiang-Zhen Kong, Jin-Xing Liu*

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

* Corresponding author

Email addresses:

TRW: wutianru@126.com

MMY: yinmengmeng00@163.com

CNJ: jiaocuina123@163.com

YLG: yinliangao@126.com

XZK: kongxzhen@163.com

JXL: sdcavell@126.com

Abstract

Background: MicroRNAs (MiRNAs) are non-coding RNAs with regulatory functions. Many studies have shown that miRNAs are closely associated with human diseases. Among the methods to explore the relationship between the miRNA and the disease, traditional methods are time-consuming and the accuracy needs to be improved. In view of the shortcoming of previous models, a collaborative matrix factorization based on matrix completion (MCCMF) is proposed to predict the unknown miRNA-disease associations.

Results: The complete matrix of the miRNA and the disease is obtained by matrix completion. Moreover, Gaussian Interaction Profile (GIP) kernel is added to the miRNA functional similarity matrix and the disease semantic similarity matrix to form the GIP kernel similarity matrix. Then the Weight K Nearest Known Neighbors (WKNKN) method is used to pretreat the association matrix, so the model is close to the reality. Finally, collaborative matrix factorization (CMF) method is applied to obtain the prediction results. Therefore, the MCCMF obtains a satisfactory result in the five-fold cross-validation, with an AUC of 0.9569(0.0005).

Conclusions: The AUC value of MCCMF is higher than other advanced methods in the 5-fold cross validation experiment. In order to comprehensively evaluate the performance of MCCMF, f-measure and other evaluation indexes are also added. The final experimental results demonstrate that MCCMF outperforms other methods in prediction miRNA-disease associations. In the end, the effectiveness and practicability of MCCMF are further verified by researching three specific diseases.

Keywords: MiRNA-disease association prediction, Matrix completion, Weight K nearest known neighbors, Matrix factorization

Background

MicroRNAs (MiRNAs) are a class of non-coding single-stranded RNA molecules. Their lengths are usually 18 to 24 nucleotides. Instead of synthesizing proteins, miRNAs participate in post-transcriptional regulation of gene expression in eukaryotes and viruses [1]. In spite of the first miRNA Line-4 was discovered in 1993 [2], the diversity and prevalence of these genes were revealed in recent years. To date, 38,589 miRNA have been found in animals, plants and viruses [3]. At the same time, miRNAs were discovered to play an important role in cell proliferation [4], differentiation [5], senescence [6], apoptosis [7], and so on. A study indicated that more than one third of human genes are regulated by MiRNA [8]. Obviously, miRNA disorder could have severe impacts on humans.

Evidence shows that an increasing number of miRNAs are closely associated with diseases [9]. Since the first discovery of miR15 and miR16 deficiency in B cell chronic lymphocytic leukemia (B-CLL) [10], the research results of miRNA-disease associations are often reported. For example, the expression of miR-25 and miR-223 is significantly higher in patients with esophageal squamous cell carcinoma than the normal people, while the expression of miR-375 is significantly lower [11]. Studies show that miR-26a may be a regulatory factor that inhibits the progression and metastasis of c-Myc/EZH2 double height advanced HCC [12]. In addition, miR-340 has been suggested as a biomarker for cancer metastasis and prognosis [13]. At present,

the research on miRNAs and diseases is becoming more extensive. Researchers have also developed a number of data base to store miRNA and disease data, such as dbDEMC [14], HMDD v3.0 [15] and miR2Disease [16]. Unfortunately, the known correlation data is not enough. Moreover, traditional methods to identify new miRNA-disease associations are time-consuming and laborious.

With the improvement of information technology and the development of a large number of miRNA data sets, many effective methods for predicting miRNA-disease associations have been proposed. According to the hypothesis that functionally similar miRNAs may be associated with diseases with similar phenotypes [17], Jiang *et al*. first constructed a genetic data network, and then prioritized disease-related miRNAs to predict miRNA-disease associations [18]. However, due to the limited association information, this method is not quite effective. Based on heterogeneous omics data, the potential miRNA-disease associations were identified by using a Graph Regularized Non-negative Matrix Factorization (GRNMF) by Xiao *et al*. [19]. Similarly, Gao *et al.* put forward a method based on Double Network Sparse Graph Regularized Matrix Factorization (DNSGRMF), and added the $L_{2,1}-$norm and Gaussian interaction profile (GIP) kernel to improve the prediction ability [20]. In addition, considering the nearest neighbor information of the miRNA and the disease, Gao *et al.* introduced a method of Nearest Profile-based Collaborative Matrix Factorization (NPCMF) to predict miRNA-disease associations [21]. In order to protect the known correlation, Logistic Weighted Profile-based Collaborative Matrix Factorization (LWPCMF) method was proposed by Yin *et al.*, which effectively predicted miRNA-disease

associations [22].

In recent years, machine learning-based miRNA-disease association prediction methods are also popular. A support vector machine (SVM) classifier was developed by Xu *et al*. to extract features from the miRNA-disease network and miRNA expression levels [23]. Yet, the construction of miRNA target-dysregulated network (MTDN) is complex, so only direct miRNA target regulation can be predicted. Chen *et al*. used random walk to prioritize disease-related miRNAs to predict potential human miRNA-disease associations [24]. Like the problem of Jiang *et al*., their approach is also affected by limited disease-miRNA associations. A restricted Boltzmann machine was established by Chen *et al*. to predict miRNA-disease associations (RBMMMDA). Although RBMMMDA is effective, it may lead to bias in known miRNA-disease associations [25]. The above methods have made great contributions to predicting associations of miRNA-disease.

Since the shortcomings of the above methods, a novel method for predicting miRNA-disease associations with Collaborative Matrix Factorization based on Matrix Completion (MCCMF) is proposed in this paper. Firstly, human miRNA-disease association matrix, miRNA function similarity matrix and disease semantic similarity matrix are obtained from HMDD v2.0, but the obtained matrix is sparse. Therefore, the matrix completion method is used to complete the matrix. The matrix completion algorithm is mainly developed on the basis of Augmented Lagrange multiplier method (ALM) [26], Alternating Direction Method (ADM) [27] and Singular Value Threshold (SVT) operation [28]. Secondly, we integrate the completed matrix and the GIP kernel

similarly matrix of the disease and the miRNA. At the same time, the miRNA-disease association matrix is preprocessed by Weight K Nearest Known Neighbors (WKNKN) method to solve the problem of unknown missing values [29]. Finally, collaborative matrix factorization is used to predict associations between miRNAs and diseases. In the experiment, a five-fold cross validation on MCCMF is performed, and results show that our method is superior to the other four methods. In addition, we focus on the cases of Gastrointestinal Neoplasms, Retinoblastoma and Hepatoblastoma. Our method not only successfully verifies the known associations of miRNA-disease, but also finds many unknown associations. To sum up, MCCMF can avoid the inherent noise of the data set, with high-speed and high prediction accuracy.

Materials

Human miRNA-disease associations

The initial miRNA-disease association data is downloaded from HMDD v2.0 [30]. HMDD v2.0 is an experimental data set supporting human miRNA-disease associations, and storing 5430 experimentally verified miRNA-disease associations between 495 miRNAs and 383 diseases. In this paper, the adjacency matrix $\mathbf{MD}$ is used to represent the miRNA-disease association network. The adjacency matrix $\mathbf{MD}$ is a sparse matrix composed of 0 and 1. If $\mathbf{MD}\left(m_i, \mathrm{d}_j\right)$ is 1, disease d_j is correlated with miRNA m_i; otherwise irrelevant.

MiRNA function similarity

According to the hypothesis that functionally similar miRNAs are more likely to be associated with phenotypic diseases, a method for calculating the functional similarity of miRNAs is proposed by Wang *et al.* [31]. In this paper, we download the miRNA

function similarity from http://www.cuilab.cn/files/images/cuilab/misim.zip. And the matrix $\mathbf{MF}$ is used to represent the functional similarity network of the miRNA, in which the element $\mathbf{MF}(i, j)$ represents the similarity between miRNA m_i and miRNA m_j. The self-similarity of each miRNA is 1, so the diagonal elements of the matrix $\mathbf{MF}$ are 1.

Disease semantic similarity

The relationship between different diseases is obtained from the MeSH database (http://www.ncbi.nlm.nih.gov/). At the same time, we represent the disease D as a Directed Acyclic Graph, $DAG(D) = (D, T(D), E(D))$, where $T(D)$ is the set of both a node D and its ancestor nodes, and $E(D)$ is the set of edges that ancestor nodes pointing to node D. For ancestor node t in $DAG(A)$, its contribution to the semantic value of disease A is computed as follows:

$$D1_A(t) = \begin{cases} 1 & if\ \ t = A, \\ \max\left\{\Delta * D1_A\left(t'\right) \middle| t' \in children\ of\ \ t\right\} & if\ \ t \neq A. \end{cases} \quad (1)$$

In the above formula, Δ is a semantic contribution factor. Based on the method of Wang *et al*. [31], the value of Δ is set to 0.5. For the disease A, the contribution of itself to the disease A is 1, while the contribution of ancestor node t is decreasing with the increase of its layers.

Based on the contribution of ancestor diseases and disease A itself, the semantic value of disease A can be expressed as follows:

$$DV1(A) = \sum_{t \in T(A)} D1_A(t). \quad (2)$$

According to the hypothesis that the more shared part of the disease pairs in $DAGs$ is, the higher similarity is. The semantic similarity between disease A and disease B

is calculated as:

$$DS1(A,B)=\frac{\sum_{t\in T(A)\cap T(B)}\left(D1_A(t)+D1_B(t)\right)}{DV1(A)+DV1(B)}. \quad (3)$$

However, the above model is a little inadequacy, which is the setting of Δ that causes the same layer of diseases with the same semantic contribution. Obviously, the incidence of various diseases is different, and the contribution of diseases with high incidence should be less than those with low incidence. To improve the above model, we combine the method of Xuan *et al*. to define the semantic similarity calculation method [32]. In this method, the contribution of ancestor node t in $DAG(A)$ to the semantic value of disease A is as follows:

$$D2_A(t)=-\log\frac{the\ number\ of\ DAGs\ including\ t}{the\ number\ of\ diseases}. \quad (4)$$

The semantic value of disease A and the semantic similarity between the disease A and the disease B are calculated as:

$$DV2(A)=\sum_{t\in T(A)}D2_A(t), \quad (5)$$

$$DS2(A,B)=\frac{\sum_{t\in T(A)\cap T(B)}\left(D2_A(t)+D2_B(t)\right)}{DV2(A)+DV2(B)}. \quad (6)$$

Finally, in order to calculate the semantic similarity more comprehensive and rational, we combine the two models to get Equation (7).

$$\mathbf{DS}(A,B)=\frac{DS1(A,B)+DS2(A,B)}{2}. \quad (7)$$

Methods

We develop a novel method for predicting miRNA-disease associations with MCCMF. MCCMF is divided into four main steps: Firstly, we use the matrix completion

algorithm to complete the miRNA similarity matrix and the disease similarity matrix. Secondly, the new completion similarity matrix is integrated with existing miRNAs and disease similarity information. Thirdly, the WKNKN is used to convert the binary values of the miRNA-disease association matrix into the interaction likelihood values [29]. Finally, the Collaborative Matrix Factorization is used to predict the association of miRNA-disease. Fig. 1 shows the complete process for MCCMF.

Gaussian interaction profile kernel similarity for diseases and miRNAs

On the basis of the hypothesis that functionally similar miRNAs may be associated with similar diseases, and vice versa, the known miRNA-disease association network is used to construct the GIP kernel similarity for diseases and miRNAs [33]. GIP kernel similarity can increase the multiple and topological information of known correlations. The interaction profile of miRNA $m(i)$ is represented by the binary vector $M(i)$ of the $i-th$ column of the adjacency matrix **MD**. Similarly, the binary vector $D(\mathrm{i})$ of the $i-th$ row of the adjacency matrix **MD** denotes the interaction profile of disease $d(i)$. Hence, we can define the GIP kernel similarity for miRNAs and diseases as follows:

$$\mathbf{GM}(m(i), m(j)) = \exp(-\gamma_m \| M(i) - \mathrm{M(j)} \|^2), \tag{8}$$

$$\mathbf{GD}(\mathrm{d}(i), \mathrm{d}(j)) = \exp(-\gamma_d \| D(i) - D(\mathrm{j}) \|^2). \tag{9}$$

Here, γ_m and γ_d are parameters to control the kernel bandwidth and obtained by the following formulas:

$$\gamma_m = \frac{\delta_m}{\frac{1}{nm}\sum_{i=1}^{nm} \| M(i) \|^2}, \tag{10}$$

$$\gamma_d = \frac{\delta_d}{\frac{1}{nd}\sum_{i=1}^{nd} \| D(i) \|^2}, \tag{11}$$

where δ_m and δ_d are also bandwidth parameters and they are set to 1 according to the previous study [34]. The nm and nd mean the number of all the miRNAs and diseases.

Matrix completion

The miRNA functional similarity matrix and disease semantic similarity matrix calculated by the above operations are still sparse and incomplete, and there are some redundant associations (i.e. inherent noise). So we use the matrix completion method to solve the problem [35]. Suppose the incomplete matrix is $\mathbf{D}$, which can be represented as a linear combination of $\mathbf{D}$ and the noise matrix $\mathbf{N}$. The formula is as follows:

$$\mathbf{D} = \mathbf{DR} + \mathbf{N}, \tag{12}$$

where $\mathbf{DR}$ is a low-rank matrix, and specifically, it is a more refined or informative similarity matrix after removing noise from the existing similarity matrix.

In order to make $\mathbf{R}$ be low-rank, a nuclear norm on $\mathbf{D}$ is added. At the same time, the $L_{2,1}-$norm is used of the error term $\mathbf{N}$ to make noise matrix $\mathbf{N}$ more sparse. When the final low-rank matrix $\mathbf{DR}^*$ and sparse matrix $\mathbf{N}^*$ are calculated, $\mathbf{DR}^*$ or $\mathbf{D}$-$\mathbf{N}^*$ are used to describe a completed matrix. Therefore, a formula for solving convex optimization problem can be defined as follows:

$$\min_{\mathbf{R},\mathbf{N}} \| \mathbf{R} \|_* + \omega \| \mathbf{N} \|_{2,1} \; s.t. \mathbf{D} = \mathbf{DR} + \mathbf{N}. \tag{13}$$

Here, $\| \cdot \|_*$ represents the nuclear norm, $\omega \in (0,1)$ is the positive weighting parameter and $\| \cdot \|_{2,1}$ is the noise regularization term.

When solving optimization problems under equality constraints, the ALM method is

more effective [26]. Therefore, according to ALM, the Equation (13) can be rewritten as:

$$\min_{\mathbf{R},\mathbf{N},\mathbf{X}} \| \mathbf{X} \|_* + \omega \| \mathbf{N} \|_{2,1} \; s.t. \mathbf{D} = \mathbf{DR} + \mathbf{N}, \mathbf{R} = \mathbf{X}. \tag{14}$$

Then switch the Equation (14) to an unconstraint problem, which is the Lagrange function. The formula is as follows:

$$\begin{aligned} L_\beta(\mathbf{X},\mathbf{R},\mathbf{N}) = & \| \mathbf{X} \|_* + \omega \| \mathbf{N} \|_{2,1} \\ & + tr(Y_1^T(\mathbf{D} - \mathbf{DR} - \mathbf{N})) + tr(\mathrm{Y}_2^T(\mathbf{R} - \mathbf{X})) \\ & + \frac{\beta}{2}(\| \mathbf{D} - \mathbf{DR} - \mathbf{N} \|_F^2 + \| \mathbf{R} - \mathbf{X} \|_F^2), \end{aligned} \tag{15}$$

where $\beta > 0$ is the penalty parameter, and β is updated by $\beta = \min(\rho\beta, \max_\beta)$. Y_1 and Y_2 are the Lagrange multipliers.

The ADM method is used to solve the Equation (15) [27]. The ADM is a simple method to solve the decomposable convex optimization problem, especially in solving large-scale problems. The update iterations for ADM are as follows:

$$\begin{cases} \mathbf{X}^{k+1} = \arg\min_X L(\mathbf{X}, \mathbf{R}^k, \mathbf{N}^k, \beta), \\ \mathbf{R}^{k+1} = \arg\min_R L(\mathbf{X}^{k+1}, \mathbf{R}, \mathbf{N}^k, \beta), \\ \mathbf{N}^{k+1} = \arg\min_N L(\mathbf{X}^{k+1}, \mathbf{R}^{k+1}, \mathbf{N}, \beta). \end{cases} \tag{16}$$

Based on the singular value shrinkage operator [28], $\mathbf{X}^{k+1}$ and $\mathbf{N}^{k+1}$ are represented as follows:

$$\mathbf{X}^{k+1} = D_{\frac{1}{\beta}}(\mathbf{R} + \frac{Y_2}{\beta}) = \operatorname{argmin} \frac{1}{\beta} \| \mathbf{X} \|_* + \frac{1}{2} \| \mathbf{X} - (\mathbf{R} + \frac{Y_2}{\beta}) \|_F^2, \tag{17}$$

$$\mathbf{N}^{k+1} = D_{\frac{\omega}{\beta}}(\mathbf{D} - \mathbf{DR} + \frac{Y_1}{\beta}) == \operatorname{argmin} \frac{w}{\beta} \| \mathbf{N} \|_{2,1} + \frac{1}{2} \| \mathbf{N} - (\mathbf{D} - \mathbf{DR} + \frac{Y_1}{\beta}) \|_F^2, \tag{18}$$

while the minimization of $\mathbf{R}$ is a least squares problem, and its normal equation is as follows:

$$\mathbf{R} = (\mathbf{I} + \mathbf{D}^T\mathbf{D})^{-1}(\mathbf{D}^T\mathbf{D} - \mathbf{D}^T\mathbf{N} + \mathbf{X} + \frac{\mathbf{D}^T Y_1 - Y_2}{\beta}), \tag{19}$$

where $\mathbf{I}=\mathbf{D}\mathbf{D}^T$ is widely used in matrix completion.

Then $\mathbf{X}$, $\mathbf{R}$ and $\mathbf{N}$ are updated by changing the Lagrange multipliers Y_1 and Y_2. Moreover, Y_1 and Y_2 can be obtained by the following formulas:

$$Y_1 = Y_1 + \beta(\mathbf{D} - \mathbf{D}\mathbf{R} - \mathbf{E}), \tag{20}$$

$$Y_2 = Y_2 + \beta(\mathbf{R} - \mathbf{X}). \tag{21}$$

Finally, we can get the final low-rank matrix $\mathbf{R}^*$ and sparse matrix $\mathbf{N}^*$ until the convergence conditions $\|\mathbf{D} - \mathbf{D}\mathbf{R} - \mathbf{N}\|_\infty < \varepsilon$ and $\|\mathbf{R} - \mathbf{X}\|_\infty < \varepsilon$ are satisfied. Here, ε is an extremely low number (set as 1×10^{-8} in this paper). As mentioned above, the refined matrix $\mathbf{R}^*$ and noise matrix $\mathbf{N}^*$ can be used to describe a completed matrix in the form of $\mathbf{D}\times\mathbf{R}^*$ or $\mathbf{D}$-$\mathbf{N}^*$. The specific process of matrix completion is shown in the Fig. 2.

Based on the above matrix completion method, the disease semantic similarity matrix $\mathbf{DS}$ and miRNA functional similarity matrix $\mathbf{MF}$ are used as input matrices to replace matrix $\mathbf{D}$, so that we can obtain two refined similarity matrices $\mathbf{CD}$ and $\mathbf{CM}$, respectively.

Similarity Information Integrations

Subsequent work is to integrate the completed matrix with existing similarity matrices. Since similarity information integrations of diseases and miRNAs are similar, Fig. 3 only shows the process for integration of miRNA similarity.

The specific integration formulas are as follows:

$$\mathbf{IMS}(i,j)=\begin{cases}\dfrac{\mathbf{CM}(i,j)+\mathbf{GM}(i,j)}{2}, & \text{if } \mathbf{MF}(\mathrm{i,j})=0,\\ \dfrac{\mathbf{GM}(\mathrm{i,j})+\mathbf{CM}(i,j)+\mathbf{MF}(\mathrm{i,j})}{3}, & \text{otherwise,}\end{cases} \tag{22}$$

$$\mathbf{IDS}(i,j)=\begin{cases}\dfrac{\mathbf{CD}(i,j)+\mathbf{GD}(i,j)}{2}, & \text{if } \mathbf{DS}(\mathrm{i,j})=0,\\ \dfrac{\mathbf{GD}(\mathrm{i,j})+\mathbf{CD}(i,j)+\mathbf{DS}(\mathrm{i,j})}{3}, & \text{otherwise,}\end{cases} \tag{23}$$

WKNKN

In this paper, $\mathbf{MD}$ expresses the miRNA-disease association matrix, which only represents the association between the miRNA and the disease verified by human experiment at the current stage. And we simply stipulate that if the miRNA is associated with the disease, $\mathbf{MD}\left(m_i,\mathrm{d}_j\right)$ will be set to 1. However, there are still many unknown miRNAs and diseases in the world, and whether they can be used as a bridge between existing miRNAs and diseases or not are still unknown. Maybe existing miRNAs are correlated with existing diseases through these unknown miRNAs, so the $\mathbf{MD}$ regulation is obviously inappropriate. Therefore, by estimating these unknown conditions through the correlation of its known neighbors, the WKNKN method preprocesses the matrix $\mathbf{MD}$ to get the pre-processed matrix of $\mathbf{MD}$ ($\mathbf{PMD}$). WKNKN can be thought of as a voting or integration method: some potential classifiers (nearest neighbors) are aggregated by a (weight) majority vote, the results of which are used for prediction [29]. If $\mathbf{MD}\left(m_i,\mathrm{d}_j\right)=0$, WKNKN will give $\mathbf{MD}\left(m_i,\mathrm{d}_j\right)$ a value from 0 to 1 according to the corresponding similar information of miRNAs and diseases. The specific process of WKNKN is shown in Fig. 4.

MCCMF for MiRNA-Disease Association Prediction

The CMF method proposed by Shen *et al*. that can effectively predict the potential

interactions between miRNAs and diseases [36]. In this study, the idea of the CMF method is used to predict the miRNA-disease association. The specific steps of CMF are as follows: firstly, the input miRNA-disease association matrix **PMD** is decomposed into two low-rank matrices **A** and **B** by using singular value decomposition.

$$\begin{aligned} &[\mathbf{U},\mathbf{S},\mathbf{V}] = SVD(\mathbf{PMD},\mathrm{k}),\\ &\mathbf{A} = \mathbf{U}\mathbf{S}_k^{\frac{1}{2}},\\ &\mathbf{B} = \mathbf{V}\mathbf{S}_k^{\frac{1}{2}}, \end{aligned} \tag{24}$$

where **U** and **V** is the unitary matrix. **S** is a negative real diagonal matrix, and there are k singular values on the diagonal.

Secondly, we write the objection function of MCCMF according to the idea of CMF, as follows:

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B}} \| \mathbf{PMD} - \mathbf{A}\mathbf{B}^T \|_F^2 + \lambda_l (\| \mathbf{A} \|_F^2 + \| \mathbf{B} \|_F^2 \,) \\ + \lambda_m \| \mathbf{IMS} - \mathbf{A}\mathbf{A}^T \|_F^2 \\ + \lambda_d \| \mathbf{IDS} - \mathbf{B}\mathbf{B}^T \|_F^2 . \end{aligned} \tag{25}$$

Here, $\|\cdot\|_F$ is the Frobenius norm to ensure that the feature vectors of similar miRNAs and similar diseases are similar. λ_l, λ_m and λ_d are positive parameters, which are determined by the 5-fold cross validation, and $\lambda_l \in \{2^{-2}, 2^{-1}, 2^0, 2^1\}$, $\lambda_m / \lambda_d \in \{2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3, 2^4, 2^5\}$.

Thirdly, we use L to represent the equation (25), and derive two alternative update rules by setting $\frac{\partial L}{\partial \mathbf{A}} = 0$ and $\frac{\partial L}{\partial \mathbf{B}} = 0$.

$$\mathbf{A} = (\mathbf{PMD} * \mathbf{B} + \lambda_{\mathrm{m}} \mathbf{IMS} * \mathbf{A})(\mathbf{B}^T\mathbf{B} + \lambda_l \mathbf{I}_k + \lambda_m \mathbf{A}^T\mathbf{A})^{-1}, \tag{26}$$

$$\mathbf{B} = (\mathbf{PMD}^T * \mathbf{A} + \lambda_{\mathrm{d}} \mathbf{IDS} * \mathbf{B})(\mathbf{A}^T\mathbf{A} + \lambda_l \mathbf{I}_k + \lambda_d \mathbf{B}^T\mathbf{B})^{-1}, \tag{27}$$

where $\mathbf{I}_k$ is the $k \times k$ identity matrix.

Finally, we update $\mathbf{A}$ and $\mathbf{B}$ iteratively until they converge to get the final $\mathbf{A}$ and $\mathbf{B}$. By $\mathbf{A}*\mathbf{B}^T$, the prediction matrix for miRNA-disease associations is obtained. The detail process of MCCMF can be seen in Fig. 5.

The algorithm of MCCMF is summarized in Algorithm 1 and listed as follows:

Algorithm 1:MCCMF

Input: miRNA-disease matrix **MD**, miRNA function similarity **MF**, disease semantic similarity **DS**, the Gaussian interaction profile kernel similarity for miRNAs **GM**, the Gaussian interaction profile kernel similarity for diseases **GD**

Output: prediction score matrix $\mathbf{A}*\mathbf{B}^T$

Matrix completion:

Initialization: $D=0, N=0, Y_1=0, Y_2=0, \beta=10^{-4}, \max_\beta=10^{10}, \rho=1.1, \varepsilon=10^{-8}$

Repeat:

Update $\mathbf{X}$, $\mathbf{R}$ and $\mathbf{N}$ using Eq.(17), Eq.(18) and Eq.(19), respectively

Update the multiplier Y_1 and Y_2 using Eq.(20) and Eq.(21),respectively

Update parameter β by: $\beta=\min(\rho\beta, \max_\beta)$

Compare the convergence conditions: $\|\mathbf{D}-\mathbf{DR}-\mathbf{N}\|_\infty<\varepsilon$, $\|\mathbf{R}-\mathbf{X}\|_\infty<\varepsilon$

Until convergence

Return: $\mathbf{D}\times\mathbf{R}^*$ or $\mathbf{D}$-$\mathbf{N}^*$

Integrate:

Integrate miRNA and disease similarity information using Eq.(22) and Eq.(23),respectively

Return: IMS, **IDS**

Pre-processing: $\mathbf{PMD} = WKNKN(\mathbf{MD}, \mathbf{IMS}, \mathbf{IDS})$

CMF:

Initialization: $[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \mathrm{SVD}(\mathbf{Y}, k)$, $\mathbf{A} = \mathbf{U}\mathbf{S}_k^{1/2}, \mathbf{B} = \mathbf{V}\mathbf{S}_k^{1/2}$

Repeat

Update $\mathbf{A}$ and $\mathbf{B}$ using Eq.(26) and Eq.(27), respectively

Until convergence

Return: $\mathbf{A} * \mathbf{B}^T$

Results

Performance evaluation

In this section, AUC value and F-measure are used to evaluate the performance of MCCMF method. The AUC value is the area corresponding to the ROC curve. F-measure is a harmonic average combined by Precision and Recall.

Initially, we implement five-fold cross validation to objectively evaluate the predictive power of our method. The existing miRNA-disease associations are randomly divided into five groups, among which four groups are used as the training set and the remaining one as the test set. In addition, in order to demonstrate the high predictive capability of our method, the random deletion of the miRNA-disease association (i.e. Cross Validation pairs' mode) increases the difficulty of prediction before performing the cross validation [37]. Five-fold cross validation is repeated 10 times to prevent grouping from causing bias, and the average result of 10 times is used as the final evaluation result. The ROC curve is drawn to represent the predicted performance intuitively, and the AUC value is calculated to evaluate MCCMF

quantitatively. The ROC curve takes True Positive Rate (TPR) as the Y-axis and False Positive Rate (FPR) as the X-axis. TPR and FPR can be expressed as:

$$TPR = \frac{TP}{TP + FN}, \tag{28}$$

$$FPR = \frac{FP}{TN + FP}, \tag{29}$$

where TP is the number of samples that are actually positive samples and are also predicted to be positive. FN represents the number of samples that are actually negative and also predicted to be negative. However, TN and FP represent the number of samples for which the predicted results are inconsistent.

In order to make the performance evaluation more comprehensive, we also use other evaluation indicators, including the accuracy, precision, recall and f-measure. Their calculation formulas are defined as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{30}$$

$$precision = \frac{TP}{TP + FP}, \tag{31}$$

$$recall = \frac{TP}{TP + FN}, \tag{32}$$

$$f\text{-}measure = \frac{2 \times precision \times recall}{precision \times recall}, \tag{33}$$

where TP, FN, TN and FP have the same definition with Equation (28) and (29).

Comparison with Other Methods

The AUC value is generally between 0 and 1. The higher the AUC value is, the better the prediction result will be. MCCMF finally obtain an AUC value of 0.9563 in the five-fold cross validation. MCCMF is compared with four advanced methods such as WBNPMD [38], RLSMDA [39], GRNMF [19] and CMF [36], which proves the superior performance of our method. The ROC curves are drawn in Fig. 6, and the

comparison results are listed in Table 1.

In the Table 1, the highest value is highlighted in bold, with the standard deviation in parentheses. In the five-fold cross validation experiment, WBNPMD, RLSMDA, GRNMF, CMF and MCCMF obtain AUCs of 0.9173, 0.8389, 0.869, 0.8697 and 0.9569, respectively. Therefore, our method is superior to the other four methods.

WBNPMD with higher AUC value is selected for comparison with MCCMF, and accuracy, precision, recall and f-measure are presented as a bar graph in Fig. 7. Also, MCCMF is better than WBNPMD.

Case Studies

In the end, we carry out a simulation experiment to analyze the specific disease. First of all, the disease we want to explore are selected and the predicted score are ranked. Then, based on the predicted score after ranking, some miRNAs of high association degree with the disease are found. Moreover, by comparing with the original miRNA-disease association matrix, they are determined whether the associations of high prediction score are known. Finally, the unknown associations are verified by searching existing data sets. Here, we choose three diseases of Gastrointestinal Neoplasms, Retinoblastoma and Hepatoblastoma for analysis. In addition, three popular data sets, dbDEMC [14], HMDD v3.0 [15] and miRCancer [40] are used for validation. These data sets store miRNA-disease associations that have been experimentally confirmed by some researchers over the years.

Gastrointestinal Neoplasms is a very common gastrointestinal disease with a high incidence. However, there are no obvious symptoms in the early growth stage of the neoplasms, which is very dangerous to human beings. We successfully predict 31

known associations and 9 new associations, 7 of which are confirmed by HMDD v3.0 and miRCancer. For example, Tazawa *et al*. discovered the potential role of oncogenic miR-21 in Gastrointestinal Neoplasms [41]. Other confirmed miRNAs have been reported in relevant data sets, and they are not listed here. There are still two unconfirmed ones that need further research. Table 2 describes the simulation results, where known associations are shown in bold, confirmed new predictions are written to the corresponding database, and unconfirmed ones are shown in "unconfirmed".

Retinoblastoma is a malignant tumor that occurs in children under 3 years old, and has a familial predisposition. There are 38 known associations between the disease and miRNA in the known association data set, and 37 known associations are successfully predicted by us. At the same time, 23 new associations are predicted, of which seven are confirmed and the others are unconfirmed. Montoya *et al*. found that the expression of miR-31 in Retinoblastoma is significantly reduced, which promotes the development of targeted therapy for Retinoblastoma [42]. Table 3 shows the specific situation.

Hepatoblastoma is the most common intraabdominal malignant tumor after neuroblastoma and nephroblastoma in childhood. In the existing miRNA-disease association data set, there are 8 known miRNA-disease associations, and all of them have been predicted. Besides, we predicted 12 new associations, of which 7 are confirmed and 5 are not. We also find literatures confirming that miR-143 is a factor affecting Hepatoblastoma. The study of Zhang *et al*. showed that blocking miR-143 could significantly inhibit local liver metastasis [43]. Hepatoblastoma prediction results are shown in Table 4.

As can be seen from the simulation results above, most known miRNAs are successfully predicted, while a small number of unknown associations are in HMDDv3.0, miRCancer and dbDEMC data sets. Although a few have not been confirmed, they can be used as a reference for researchers. In addition, we used Cytoscape software to map the prediction network of these three diseases (Fig. 8). In the network, the ellipse represents miRNAs, and the remaining shapes represent diseases. The correlations are connected by line segments with arrows, and there are common miRNAs between diseases. According to the size of the predicted score, the color degree of the ellipse is set differently. The darker the color of the ellipse is set to, the stronger the correlation between miRNA and disease is.

Discussion and Conclusions

With the development of biotechnology, miRNAs have been found to be closely related to many human diseases. The prediction of disease-miRNA associations is conducive to the prevention or treatment of diseases. Therefore, in recent years, many methods have been proposed to predict the associations of miRNA-disease. However, traditional experiments are time-consuming and the accuracy of some prediction methods is not excellent. Thus, we propose a collaborative matrix factorization method based on matrix completion (MCCMF). Considering the sparse and incomplete similarity matrix of miRNA-disease, we use the matrix completion method to complete the matrix. Then the completed matrix is integrated with GIP kernel similarity to improve the data information and reduce the influence of noises. In addition, WKNKN is also introduced to pretreat the existing association matrix of miRNAs and diseases, so our method is

suitable to practical problems. Finally, the idea of CMF is adopted to construct the objective function and obtain the predicted results. The AUC value (0.9569) of MCCMF is higher than other advanced methods in the 5-fold cross validation experiment. In order to comprehensively evaluate the performance of MCCMF, accuracy, precision, recall and f-measure are applied, and results are 0.992, 0.779, 0.918 and 0.830, respectively. Compared with the other four methods, our method has the best performance. The analysis of Gastrointestinal Neoplasms, Retinoblastoma and Hepatoblastoma further verified the effectiveness of MCCMF. Since most of associations are unknown in reality, MCCMF can also be used to predict in this situation. However, the performance of our method needs to be further improved.

In the future, we will improve the technology to use the latest version of the data set, such as HMDD v3.0. There exists the better way to integrate data, rather than simply adding them together.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets that support the findings of this study are available in https://github.com/cuizhensdws.

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported in part by the grants of the National Science Foundation of China, Nos. 61872220, and 61702299.

Authors' contributions

TRW and MMY jointly contributed to the design of the study. TRW designed and implemented the BGCMF method, performed the experiments, and drafted the manuscript. MMY participated in the design of the study and performed the statistical analysis. YLG contributed to the data analysis. CNJ and JXL contributed to improving the writing of manuscripts. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

Reference

1. Alshalalfa M, Alhajj R: **Using context-specific effect of miRNAs to identify functional associations between miRNAs and gene signatures**. *BMC Bioinformatics* 2013, **14**(12):S1.
2. 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):843-854.
3. Kozomara A, Griffiths-Jones S: **miRBase: annotating high confidence microRNAs using deep sequencing data**. *Nucleic Acids Research* 2013, **42**(D1):D68-D73.
4. Cheng AM, Byrom MW, Shelton J, Ford LP: **Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis**. *Nucleic Acids Research* 2005, **33**(4):1290-1297.
5. Miska EA: **How microRNAs control cell division, differentiation and death**. *Current Opinion in Genetics & Development* 2005, **15**(5):563-568.
6. Bartel DP: **MicroRNAs: Target Recognition and Regulatory Functions**. *Cell* 2009, **136**(2):215-233.
7. Xu P, Guo M, Hay BA: **MicroRNAs and the regulation of cell death**. *Trends*

in Genetics 2004, **20**(12):617-624.

8. Lewis BP, Burge CB, Bartel DP: **Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets**. *Cell* 2005, **120**(1):15-20.
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.
10. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K *et al*: **Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia**. *Proc Natl Acad Sci U S A* 2002, **99**(24):15524-15529.
11. Wu C, Li M, Hu C, Duan H: **Clinical significance of serum miR-223, miR-25 and miR-375 in patients with esophageal squamous cell carcinoma**. *Molecular Biology Reports* 2014, **41**(3):1257-1266.
12. Zhang X, Zhang X, Wang T, Wang L, Zhijun T, Wei W, Yan B, Zhao J, Wu K, Yang A-G *et al*: **MicroRNA-26a is a key regulon that inhibits progression and metastasis of c-Myc/EZH2 double high advanced hepatocellular carcinoma**. *Cancer Letters* 2018, **426**.
13. Wu Z-s, Wu Q, Wang C-q, Wang X-n, Huang J, Zhao J-j, Mao S-s, Zhang G-h, Xu X-c, Zhang N: **miR-340 Inhibition of Breast Cancer Cell Migration and Invasion Through Targeting of Oncoprotein c-Met**. *Cancer* 2011, **117**:2842-2852.
14. Yang Z, Ren F, Liu C, he S, Sun G, Gao Q, Yao L, Zhang Y, Miao R, Cao Y *et al*: **DbDEMC: A database of differentially expressed miRNAs in human cancers**. *BMC genomics* 2010, **11 Suppl 4**:S5.
15. Huang Z, Jiangcheng S, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q: **HMDD v3.0: a database for experimentally supported human microRNA-disease associations**. *Nucleic acids research* 2018, **47**.
16. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y: **miR2Disease: a manually curated database for microRNA deregulation in human disease**. *Nucleic acids research* 2008, **37**:D98-104.
17. Zou Q, Li J, Song L, Zeng X, Wang G: **Similarity computation strategies in the microRNA-disease network: A survey**. *Briefings in functional genomics* 2015, **15**.
18. 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 Systems Biology* 2010, **4**(1):S2.
19. Xiao Q, Luo J, Liang C, Cai J, Ding P: **A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations**. *Bioinformatics* 2017, **34**(2):239-248.
20. 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):130-137.
21. 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.

22. 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, **PP**:1-1.
23. Xu J, Li C-X, Lv J-Y, Li Y-S, Xiao Y, Shao T-T, Huo X, Li X, Zou Y, Han Q-L *et al*: **Prioritizing Candidate Disease miRNAs by Topological Features in the miRNA Target–Dysregulated Network: Case Study of Prostate Cancer**. *Molecular Cancer Therapeutics* 2011, **10**(10):1857.
24. Chen H, Zhang Z: **Prediction of Associations between OMIM Diseases and MicroRNAs by Random Walk on OMIM Disease Similarity Network**. *TheScientificWorldJournal* 2013, **2013**:204658.
25. Chen X, Clarence Yan C, Zhang X, Li Z, Deng L, Zhang Y, Dai Q: **RBMMMDA: predicting multiple types of disease-microRNA associations**. *Scientific Reports* 2015, **5**(1):13877.
26. Lin Z, Chen M, Ma Y: **The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices**. *Mathematical Programming* 2010, **9**.
27. Yang J, Yuan X: **Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization**. *Mathematics of Computation* 2011, **82**.
28. Cai J-F, Candès EJ, Shen Z: **A Singular Value Thresholding Algorithm for Matrix Completion**. *SIAM Journal on Optimization* 2010, **20**:1956-1982.
29. Ezzat A, Zhao P, Wu M, Li X, Kwoh C: **Drug-Target Interaction Prediction with Graph Regularized Matrix Factorization**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2017, **14**(3):646-656.
30. 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 research* 2013, **42**.
31. 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 (Oxford, England)* 2010, **26**:1644-1650.
32. Xuan P, Han K, Guo M, Guo Y, Li J, Ding J, Liu Y, Dai Q, Li J, Teng Z *et al*: **Prediction of microRNAs Associated with Human Diseases Based on Weighted k Most Similar Neighbors**. *PloS one* 2013, **8**:e70204.
33. van Laarhoven T, Nabuurs SB, Marchiori E: **Gaussian interaction profile kernels for predicting drug–target interaction**. *Bioinformatics* 2011, **27**(21):3036-3043.
34. Chen X, Yan G-Y: **Novel human lncRNA–disease association inference based on lncRNA expression profiles**. *Bioinformatics* 2013, **29**(20):2617-2624.
35. Sheng-Peng Y, Liang C, Xiao Q, Li GH, Ding P, Luo JW: **MCLPMDA: A novel method for miRNA-disease association prediction based on matrix**

completion and label propagation. *Journal of Cellular and Molecular Medicine* 2018, **23**.

36. Shen Z, Zhang Y-H, Han K, Nandi A, Honig B, Huang D-S: **miRNA-Disease Association Prediction with Collaborative Matrix Factorization**. *Complexity* 2017, **2017**:1-9.
37. Ezzat A, Wu M, Li X-L, Kwoh C-K: **Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey**. *Briefings in Bioinformatics* 2018, **20**(4):1337-1357.
38. Xie G, Fan Z, Sun Y, Wu C, Ma L: **WBNPMD: weighted bipartite network projection for microRNA-disease association prediction**. *Journal of Translational Medicine* 2019, **17**:322.
39. Chen X, Yan G-Y: **Semi-supervised learning for potential human microRNA-disease associations inference**. *Scientific Reports* 2014, **4**(1):5501.
40. Xie B, Ding Q, Han H, Wu D: **MiRCancer: A microRNA-cancer association database constructed by text mining on literature**. *Bioinformatics (Oxford, England)* 2013, **29**.
41. Tazawa H, Kagawa S, Fujiwara T: **MicroRNAs as potential target gene in cancer gene therapy of gastrointestinal tumors**. *Expert opinion on biological therapy* 2011, **11**:145-155.
42. Montoya V, Fan H, Bryar P, Weinstein J, Mets M, Feng G, Martin J, Martin A, Jiang H, Laurie N: **Novel miRNA-31 and miRNA-200a-Mediated Regulation of Retinoblastoma Proliferation**. *PloS one* 2015, **10**:e0138366.
43. Zhang X, Liu S, Hu T, Liu S, He Y, Sun S: **Up-Regulated MicroRNA-143 Transcribed by Nuclear Factor kappa B Enhances Hepatocarcinoma Metastasis by Repressing Fibronectin Expression**. *Hepatology (Baltimore, Md)* 2009, **50**:490-499.

Tables

Table 1 AUC results of cross validation experiments

Methods	Gold Standard Dataset
RLSMDA	0.8389(0.0006)
GRNMF	0.869(0.00023)
CMF	0.8697(0.0011)
WBNPMD	0.9173(0.0005)
MCCMF	**0.9569(0.0005)**

Table 2 Predicted miRNAs for Gastrointestinal Neoplasms

Rank	MiRNA	Evidence	Rank	MiRNA	Evidence
1	has-mir-1	**known**	21	has-let-7a	**known**
2	has-mir-22	**known**	22	has-mir-152	**known**
3	has-mir-200	**known**	23	has-mir-497	**known**
4	has-mir-9	**known**	24	has-mir-21	HMDD v3.0
5	has-mir-221	**known**	25	has-mir-375	**known**
6	has-mir-146a	**known**	26	has-mir-107	**known**
7	has-mir-133b	**known**	27	has-mir-18b	**known**
8	has-mir-200c	**known**	28	has-mir-494	**known**
9	has-mir-200a	**known**	29	has-mir-150	miRCancer
10	has-mir-7	**known**	30	has-mir-208a	**known**
11	has-mir-200b	**known**	31	has-mir-98	**known**
12	has-mir-222	**known**	32	has-mir-141	miRCancer
13	has-mir-126	**known**	33	has-let-7g	**known**
14	has-mir-196a	**known**	34	has-mir-184	unconfirmed
15	has-mir-142	**known**	35	has-mir-210	miRCancer
16	has-mir-124	**known**	36	has-mir-486	HMDD v3.0
17	has-mir-148a	**known**	37	has-mir-338	**known**
18	has-mir-451a	**known**	38	has-mir-27a	miRCancer
19	has-mir-31	**known**	39	has-mir-146b	HMDD v3.0
20	has-mir-451	**known**	40	has-let-7c	unconfirmed

Table 3 Predicted microbes for Retinoblastoma

Rank	MiRNA	Evidence	Rank	MiRNA	Evidence
1	has-mir-1	**known**	31	has-mir-32	unconfirmed
2	has-mir-9	**known**	32	has-mir-200	unconfirmed
3	has-mir-17	**known**	33	has-mir-192	**known**
4	has-mir-20a	**known**	34	has-mir-513b	**known**
5	has-mir-18a	**known**	35	has-mir-135b	**known**
6	has-mir-29c	**known**	36	has-mir-513c	**known**

7	has-mir-92a	**known**	37	has-mir-22	miRCancer
8	has-let-7d	**known**	38	has-mir-31	HMDD v3.0
9	has-let-7f	**known**	39	has-mir-513a	**known**
10	has-let-7g	**known**	40	has-mir-30c	**known**
11	has-let-7a	**known**	41	has-mir-491	**known**
12	has-mir-19b	**known**	42	has-mir-135a	unconfirmed
13	has-let-7b	**known**	43	has-mir-125b	HMDD v3.0
14	has-mir-29b	**known**	44	has-mir-7	unconfirmed
15	has-let-7e	**known**	45	has-mir-181a	unconfirmed
16	has-let-7i	**known**	46	has-mir-223	unconfirmed
17	has-mir-124	**known**	47	has-mir-210	unconfirmed
18	has-let-7c	**known**	48	has-mir-376a	**known**
19	has-mir-19a	**known**	49	has-mir-30a	unconfirmed
20	has-mir-92	**known**	50	has-mir-145	miRCancer
21	has-mir-181b	**known**	51	has-mir-34b	HMDD v3.0
22	has-mir-34a	**known**	52	has-mir-155	unconfirmed
23	has-mir-29a	**known**	53	has-mir-133a	unconfirmed
24	has-mir-181	**known**	54	has-mir-137	unconfirmed
25	has-mir-24	**known**	55	has-mir-146b	unconfirmed
26	has-mir-142	**known**	56	has-mir-150	unconfirmed
27	has-mir-10b	**known**	57	has-mir-126	unconfirmed
28	has-mir-34c	**known**	58	has-mir-18b	unconfirmed
29	has-mir-125a	**known**	59	has-mir-221	miRCancer
30	has-mir-21	HMDDv3.0	60	has-mir-373	HMDD v3.0

Table 4 Predicted microbes for Hepatoblastoma

Rank	MiRNA	Evidence	Rank	MiRNA	Evidence
1	has-mir-1	**known**	11	has-mir-31	unconfirmed
2	has-mir-21	**known**	12	has-mir-126	dbDEMC
3	has-mir-150	**known**	13	has-mir-146a	HMDD v3.0
4	has-mir-199a	HMDD v3.0	14	has-mir-125b	unconfirmed
5	has-mir-143	HMDD v3.0	15	has-mir-148a	**known**
6	has-mir-145	**known**	16	has-mir-22	dbDEMC
7	has-mir-199b	**known**	17	has-mir-210	HMDD v3.0
8	has-mir-214	**known**	18	has-mir-138	unconfirmed
9	has-mir-125a	**known**	19	has-mir-133a	unconfirmed
10	has-mir-9	unconfirmed	20	has-mir-122	HMDD v3.0

Figures

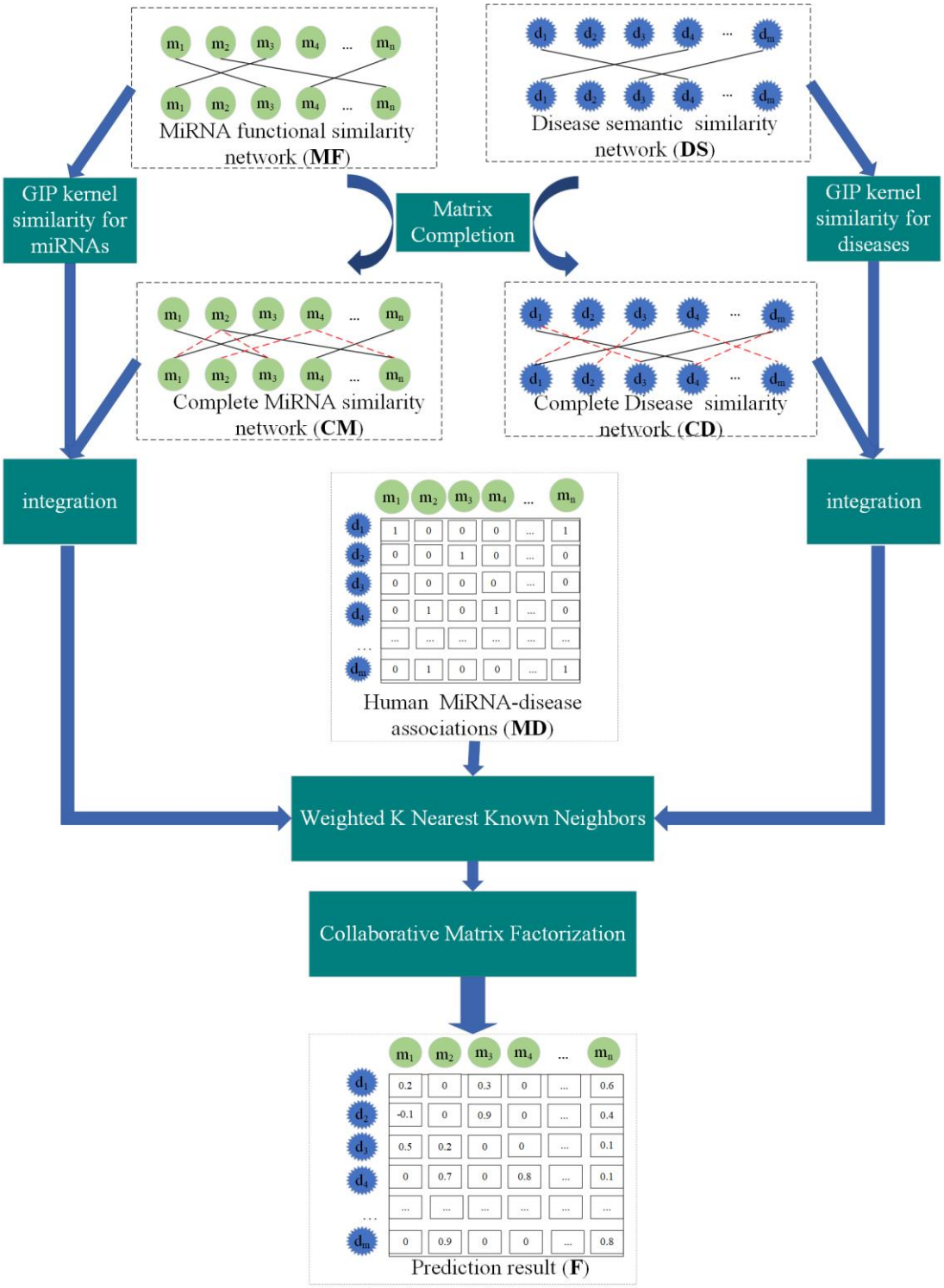

Fig. 1 The flowchart of MCCMF includes four steps: (1) completing matrix; (2) integrating similar information; (3) WKNKN; (4) collaborative matrix factorization.

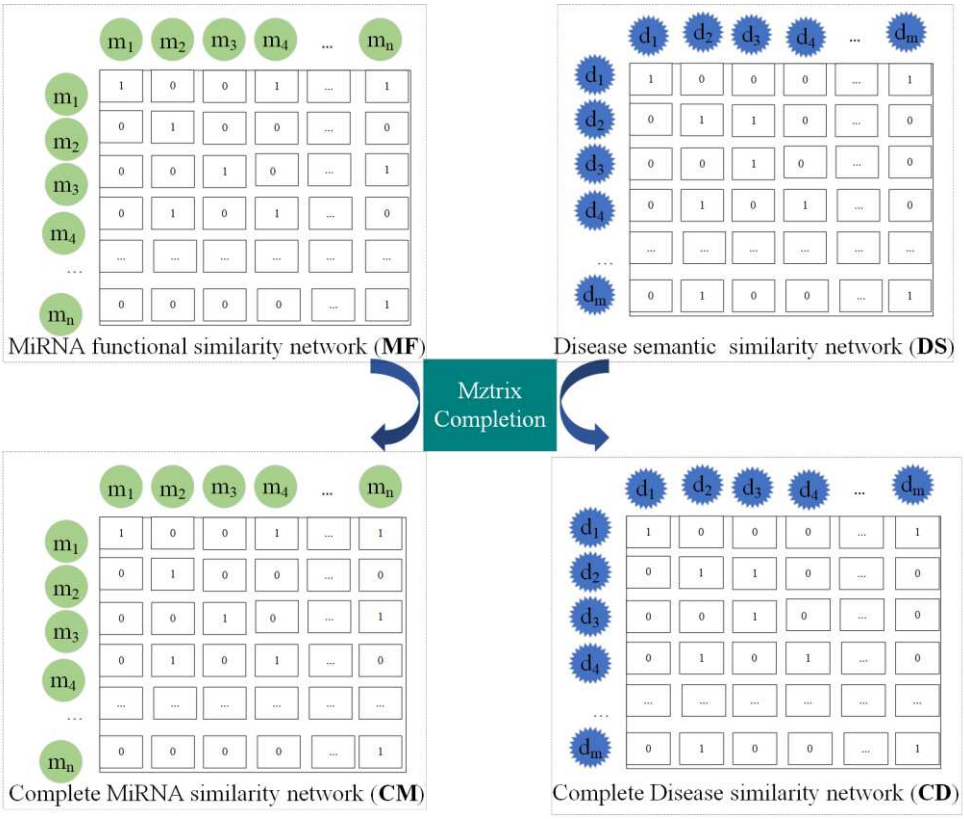

Fig. 2 The process of matrix completion

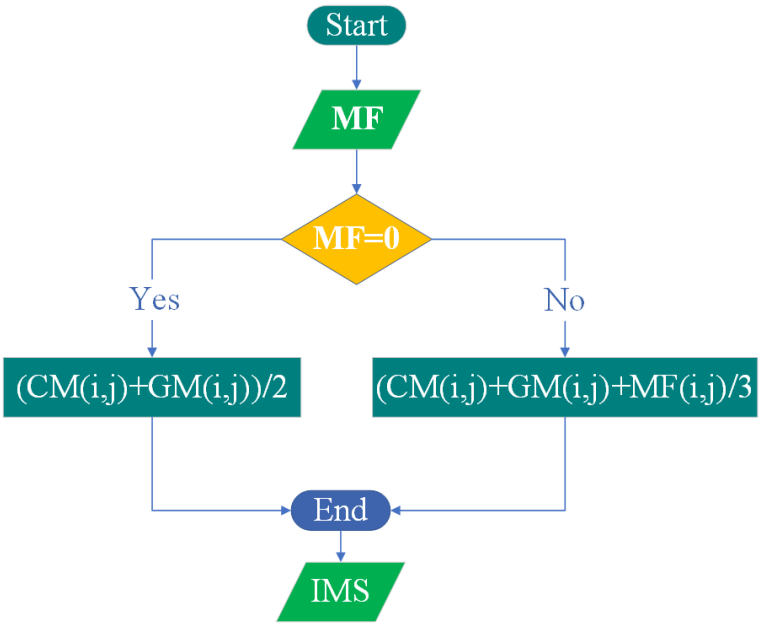

Fig. 3 The flowchart of an integration method. Matrix **MF** is the miRNA functional similarity matrix and **GM** is the GIP kernel similarity for miRNAs and **CM** is the Complete MiRNA similarity matrix and **IMS** is the integrated matrix.

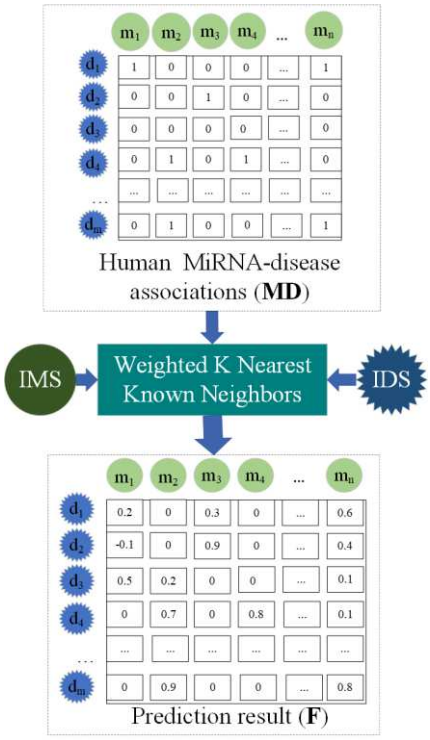

Fig. 4 The process of WKNKN. **IMS** and **IDS** are similar integration matrices of miRNA and disease, respectively.

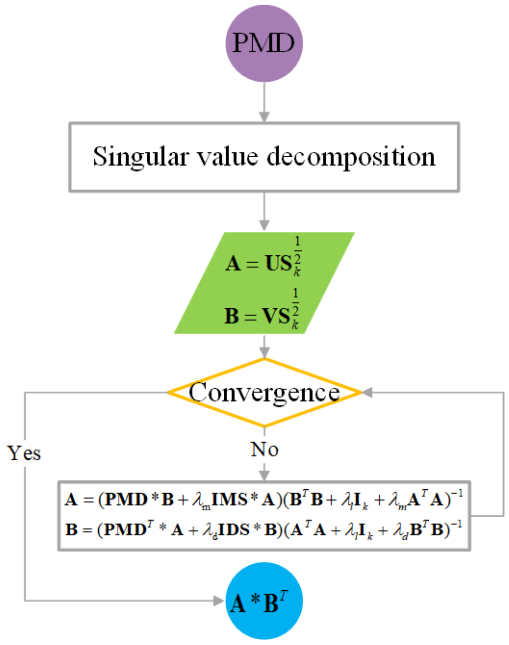

Fig. 5 The flowchart of CMF method. **PMD** is the pre-processed matrix of miRNA-disease association matrix. **A** and **B** are two low-rank matrices. $\mathbf{A} * \mathbf{B}^T$ is the predicted matrix for miRNA-disease associations.

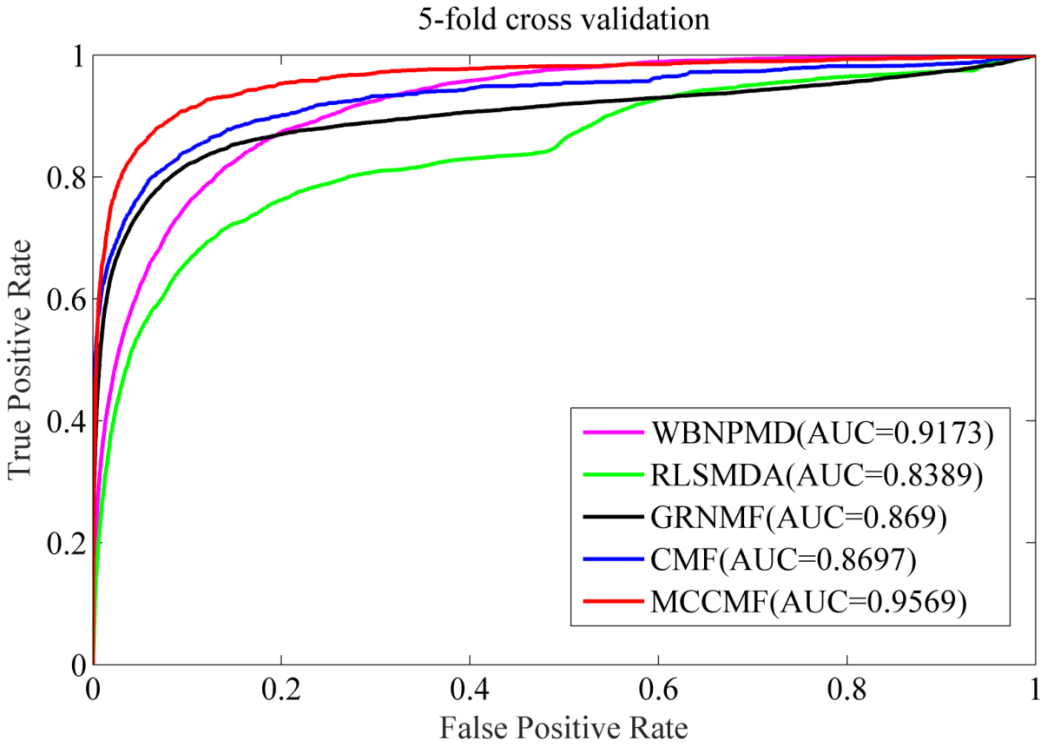

Fig. 6 The ROC curves for MCCMF and the other four methods in the 5-fold cross validation experiment.

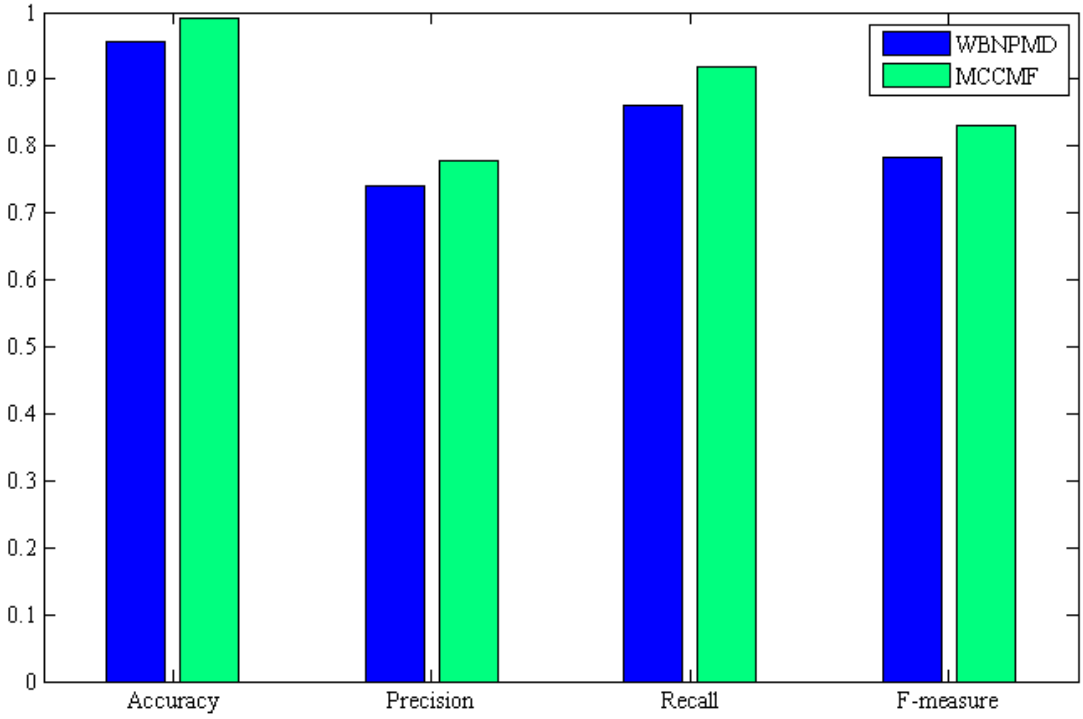

Fig. 7 Comparison of the accuracy, precision, recall and f-measure with WBNPMD.

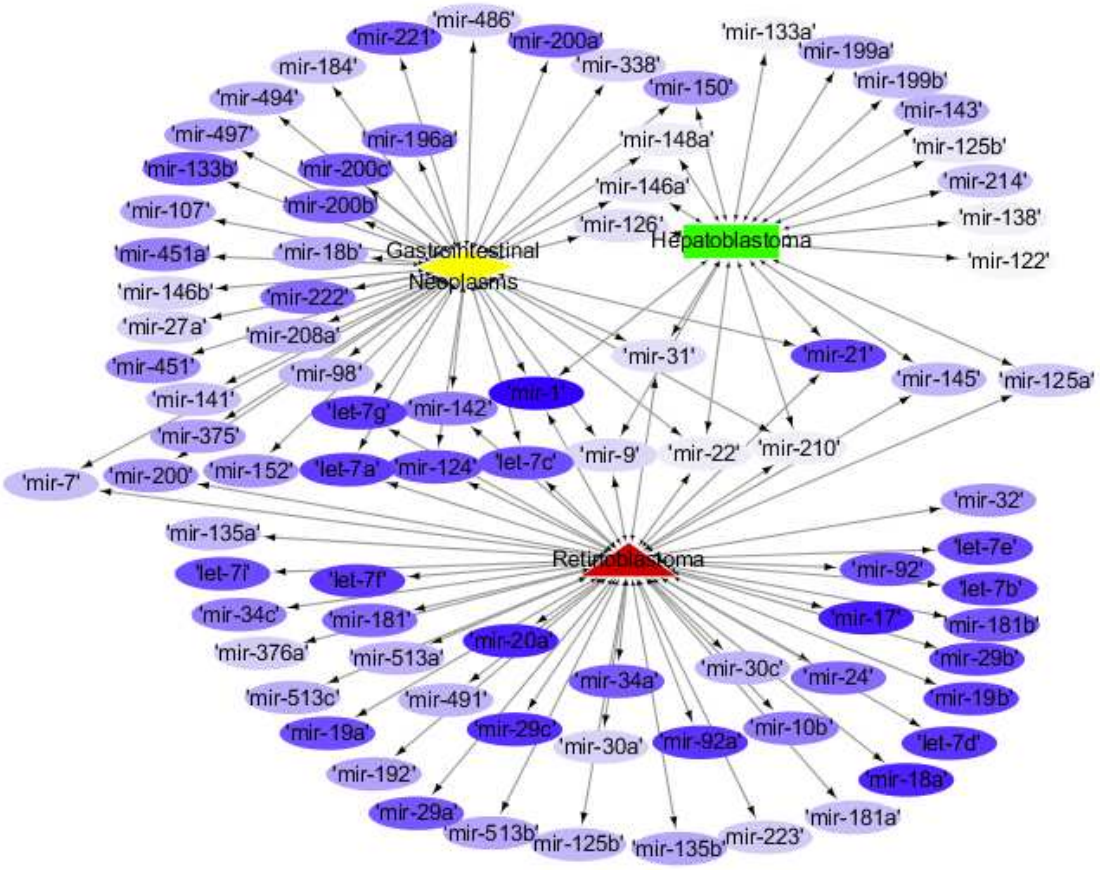

Fig. 8 The association network between disease and miRNA.

Figures

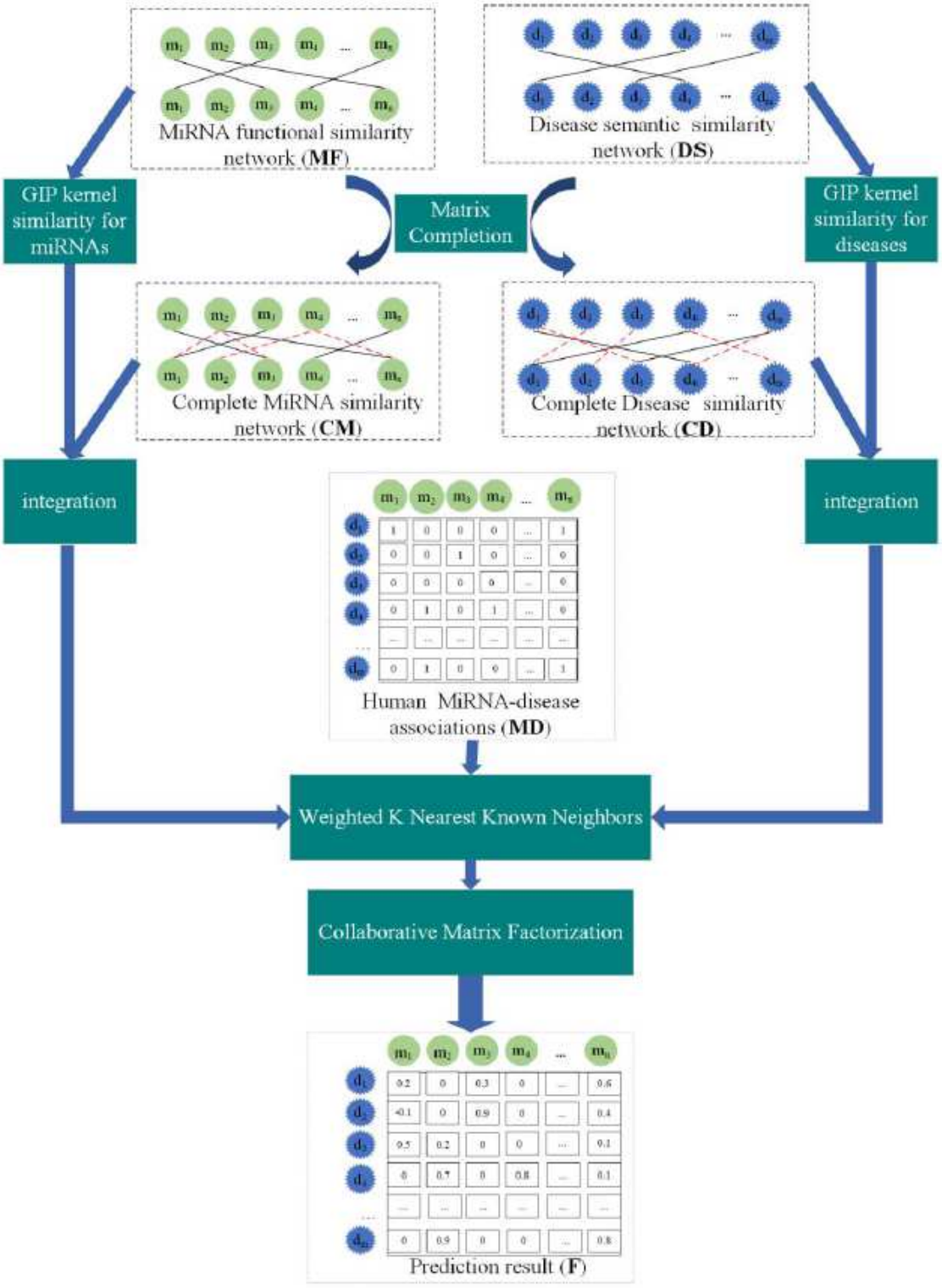

Figure 1

The flowchart of MCCMF includes four steps: (1) completing matrix; (2) integrating similar information; (3) WKNKN; (4) collaborative matrix factorizationThe process of matrix completion Fig.

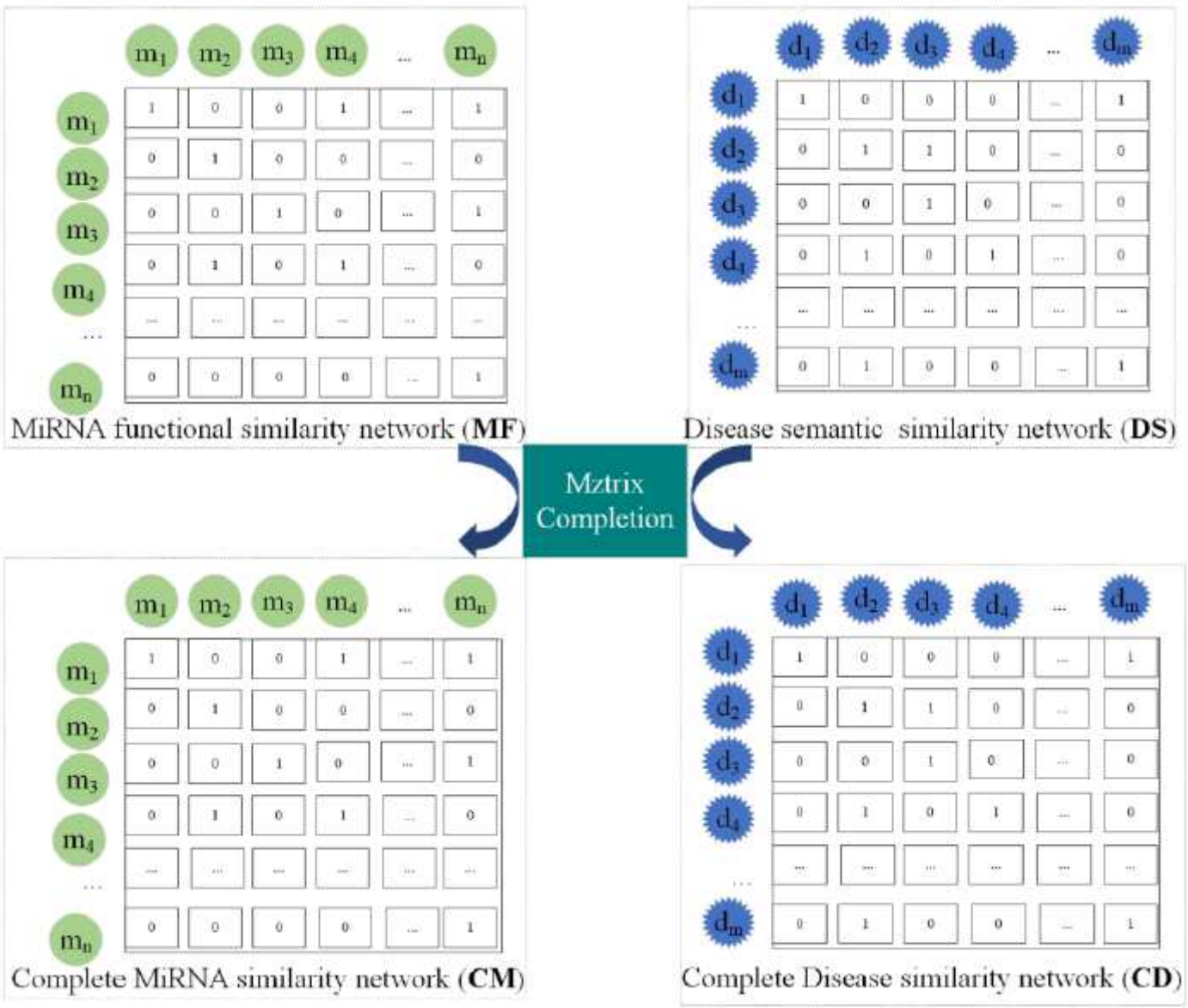

Figure 2

The process of matrix completion Fig.

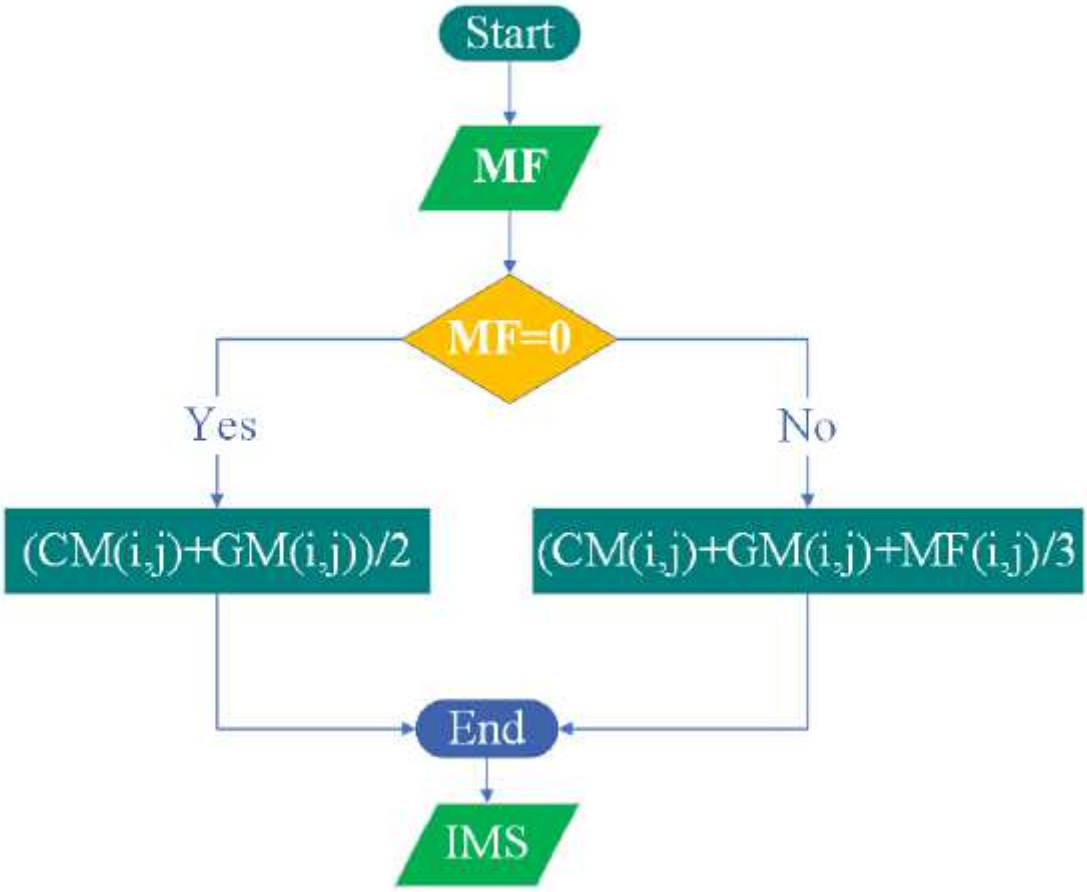

Figure 3

The flowchart of an integration method. Matrix MF is the miRNA functional similarity matrix and GM is the GIP kernel similarity for miRNAs and CM is the Complete MiRNA similarity matrix and IMS is the integrated matrix.

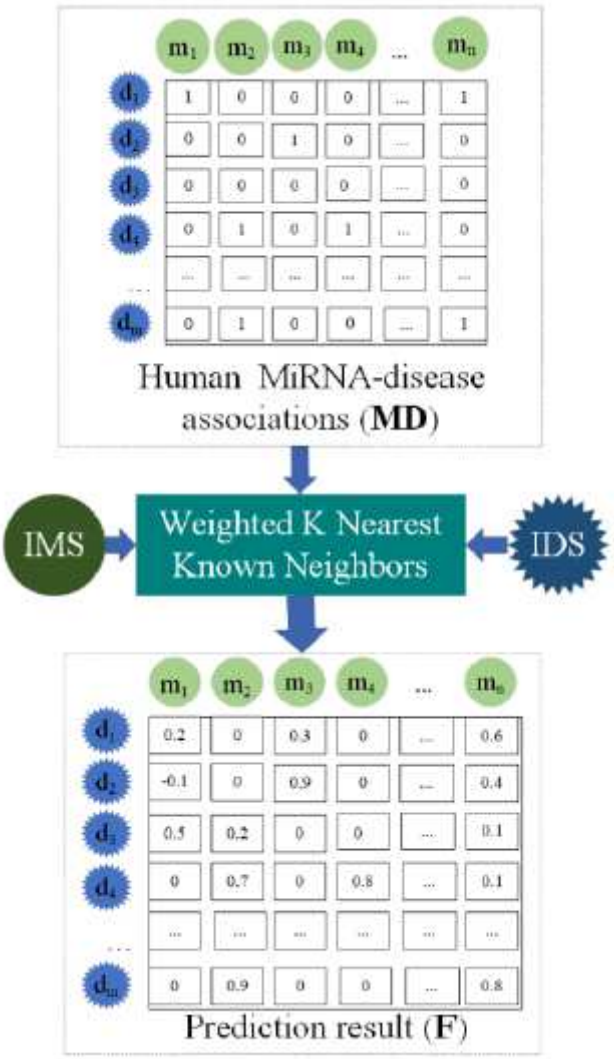

Figure 4

The process of WKNKN. IMS and IDS are similar integration matrices of miRNA and disease, respectively.

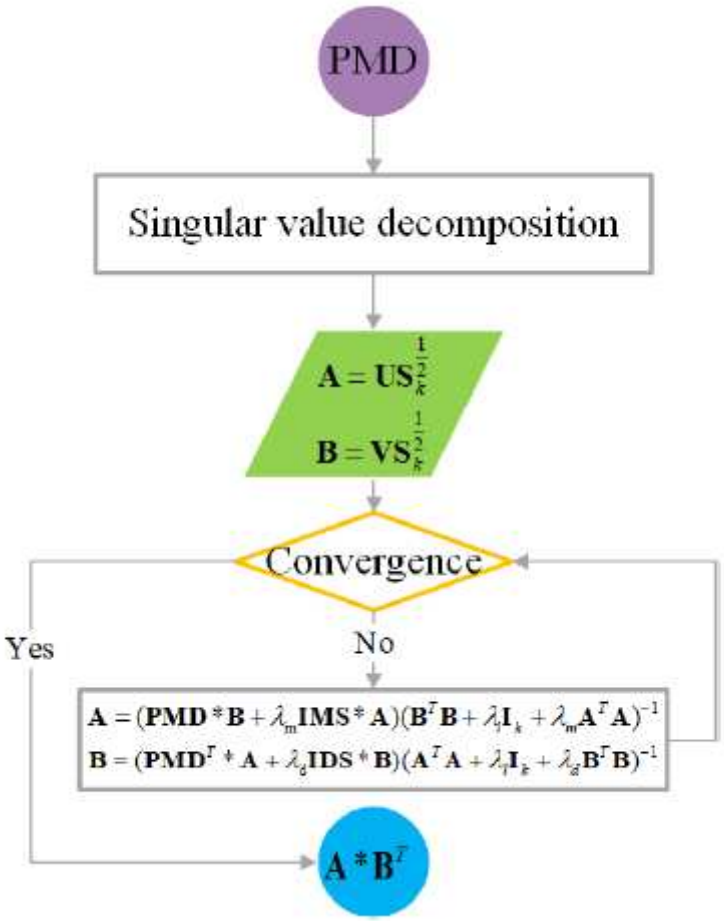

Figure 5

The flowchart of CMF method. PMD is the pre-processed matrix of miRNA-disease association matrix. A and B are two low-rank matrices. * T A B is the predicted matrix for miRNA-disease associations.

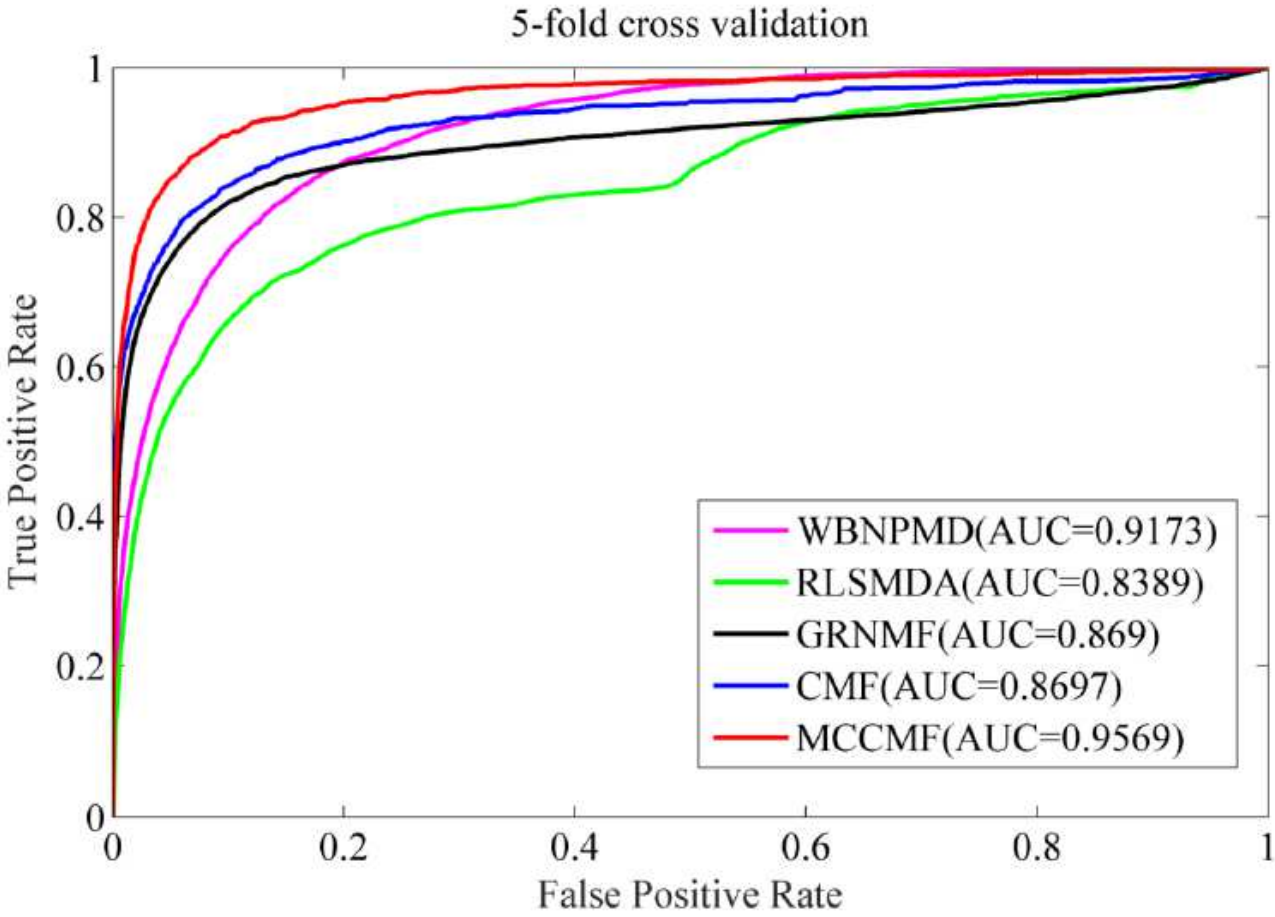

Figure 6

The ROC curve s for MCCMF and the other four methods in the 5 fold cross validation experiment Fig

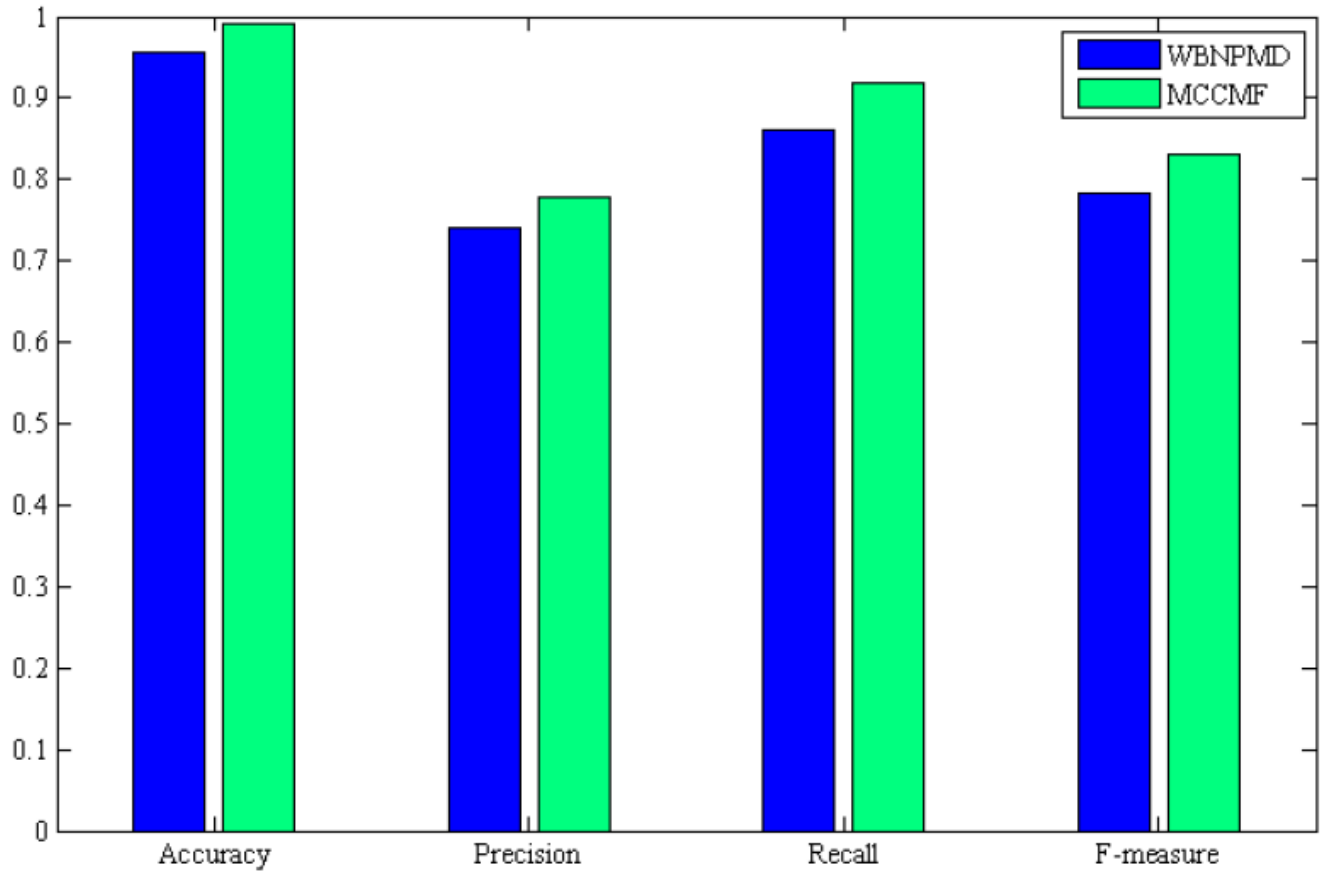

Figure 7

Comparison of the accuracy, precision, recall and f measure with WBNPMD.

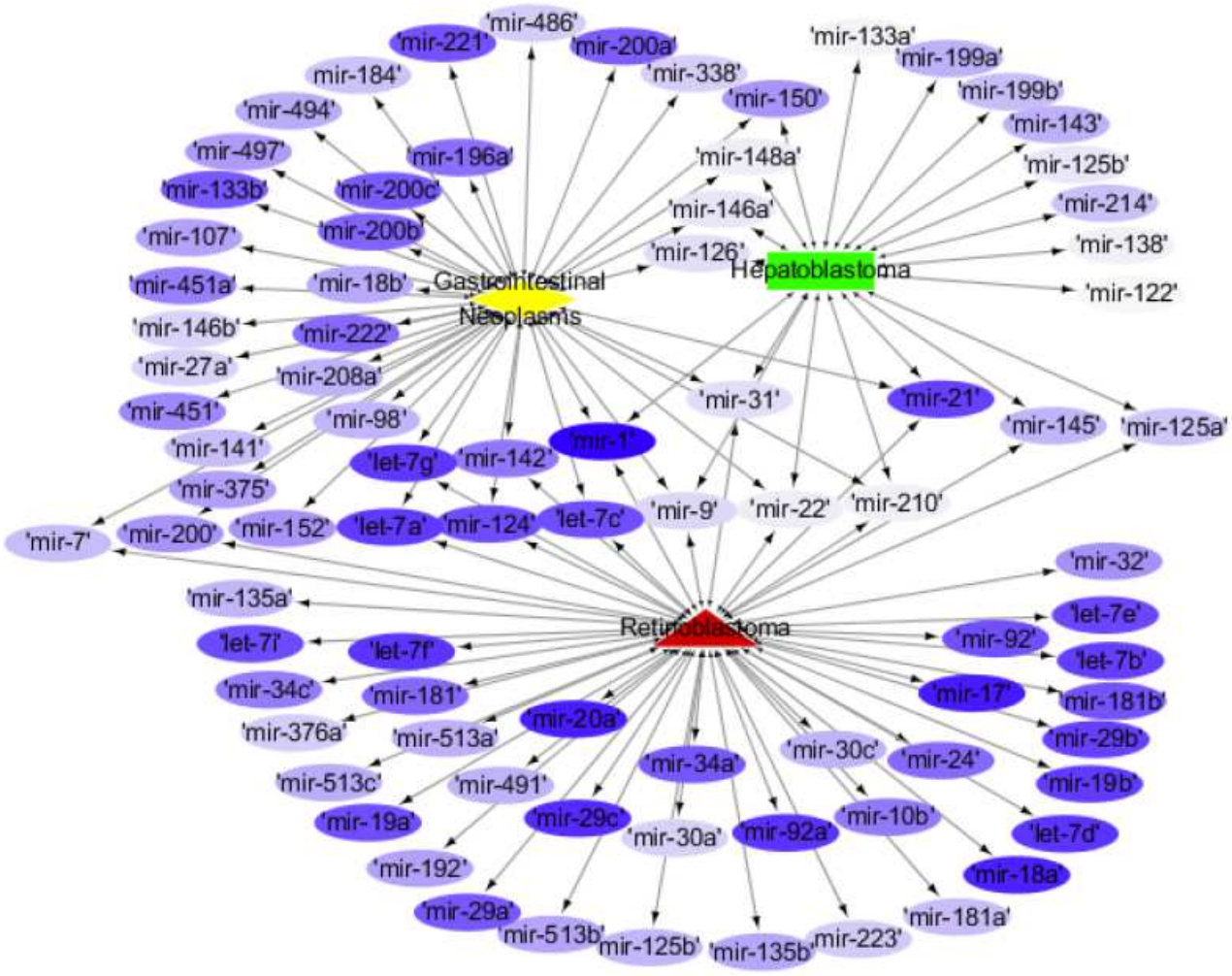

Figure 8

The association network between disease and miRNA.