


METHODOLOGY ARTICLE

Open Access



# Bipartite graph-based collaborative matrix factorization method for predicting miRNA-disease associations

Feng Zhou, Meng-Meng Yin, Cui-Na Jiao, Zhen Cui, Jing-Xiu Zhao and Jin-Xing Liu\* 

\*Correspondence:  
sdcavell@126.com  
The School of Computer  
Science, Qufu Normal  
University, Rizhao 276826,  
China

## Abstract

**Background:** With the rapid development of various advanced biotechnologies, researchers in related fields have realized that microRNAs (miRNAs) play critical roles in many serious human diseases. However, experimental identification of new miRNA–disease associations (MDAs) is expensive and time-consuming. Practitioners have shown growing interest in methods for predicting potential MDAs. In recent years, an increasing number of computational methods for predicting novel MDAs have been developed, making a huge contribution to the research of human diseases and saving considerable time. In this paper, we proposed an efficient computational method, named bipartite graph-based collaborative matrix factorization (BGCMF), which is highly advantageous for predicting novel MDAs.

**Results:** By combining two improved recommendation methods, a new model for predicting MDAs is generated. Based on the idea that some new miRNAs and diseases do not have any associations, we adopt the bipartite graph based on the collaborative matrix factorization method to complete the prediction. The BGCMF achieves a desirable result, with AUC of up to  $0.9514 \pm (0.0007)$  in the five-fold cross-validation experiments.

**Conclusions:** Five-fold cross-validation is used to evaluate the capabilities of our method. Simulation experiments are implemented to predict new MDAs. More importantly, the AUC value of our method is higher than those of some state-of-the-art methods. Finally, many associations between new miRNAs and new diseases are successfully predicted by performing simulation experiments, indicating that BGCMF is a useful method to predict more potential miRNAs with roles in various diseases.

**Keywords:** MiRNA–disease associations association prediction, Matrix factorization, Bipartite graph, Gaussian interaction profile

## Background

MicroRNAs (miRNAs) are single-stranded small ncRNAs with a typical length of 19 ~ 25 nt [1]. Although they do not encode proteins, they play a significant role in regulating gene expression. They usually silence gene expression through translational repression or otherwise function as post-transcriptional gene regulators. In 1993, the first miRNA,



© The Author(s), 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

lin-4, was discovered by Victor Ambros et al. [2]. After seven years, biological researchers discovered the second miRNA, let-7 [3]. As miRNAs are increasingly identified as playing crucial roles, researchers have begun to focus more attention on identifying miRNAs.

Studies have found that miRNAs are crucial components in cells and can play roles in many important biological processes, including haematopoiesis, cell proliferation, development, differentiation, apoptosis, cell ageing, viral infection, embryonic development and organ formation [4–7]. Mutated and disordered miRNAs will lose the ability to control their target genes, leading to the development of various complex human diseases, such as cardiovascular diseases, nervous system diseases, tumours, metabolic diseases, and autoimmune diseases [8, 9]. As an example, a miR-133b defect is easily observed in the midbrain of patients with Parkinson's disease; miR-133b is thought to have a regulatory effect on the maturation and function of midbrain dopamine neurons [10]. In addition, Gao et al. found that the expression of miR-155 in the serum of lung cancer patients was much higher than that in normal samples by experimental PCR [11]. Furthermore, Takamizawa et al. have proved that the homology of let-7 is significantly reduced in the process of lung cancer [12]. However, discovering meaningful associations between miRNAs and diseases is a time-consuming process. Therefore, it is urgent to develop fast and efficient computational methods for predicting miRNA–disease associations.

In the last decade, a large number of methods and models have been proposed to identify potential relationships between miRNAs and diseases [13, 14]. These methods and models have mainly focused on solving the above problem by machine learning, network mining, combinatorial optimization, and related approaches [15–17]. For example, Jiang et al. used a support vector machine to extract data on positive samples from negative samples. This method extracted features from miRNA target data and phenotypic similarity data and achieved favourable results [18]. Chen et al. applied the random walk algorithm with a restart, which is also a classic network-based prediction model, to miRNA–disease association (MDA) prediction [19]. In 2013, Qabaja et al. proposed a protein–protein interaction network based on the lasso regression model. They first used the lasso regression model to identify miRNAs associated with markers of diseases and then integrated biological networks and multisource data to define the gene signatures of miRNAs and diseases. Finally, this method achieved good predictive performance [20]. Xuan et al. proposed a prediction method based on the K-nearest neighbour algorithm. This method constructs a similarity network by integrating the miRNA–disease phenotype similarity network, the family information of miRNAs, and the relationships between diseases and miRNAs identified by biological experiments. The disadvantage of this method is that it cannot be applied to the association prediction of diseases without any known related miRNAs [21].

In 2014, Chen et al. proposed a semi-supervised algorithm named regularized least squares (RLSMDA) to predict potential disease–miRNA associations. The advantage of this method is that it does not require negative MDAs information and can be applied to the prediction of isolated diseases. In 2017, Chen et al. proposed a predictive model for the associations between miRNAs and diseases based on Laplacian regularized sparse subspace learning (LRSSLMDA) [22]. They used Laplacian regularization to keep local information and then used the  $L_1$  norm to select important

miRNA/disease features, further improving the precision of the algorithm. Chen et al. proposed a computational method, ensemble learning and link prediction for miRNA-disease association (ELLPMDA), which combines both machine learning methods and similarity-based algorithms. This method is based on a globally similar measurement method for diseases without any associated miRNAs [23]. Algorithms such as neural networks are also used to predict miRNA-disease associations. In 2017, Fu et al. proposed a deep integration model (DeepMDA), which uses a stack-type autoencoder to extract high-level features from similar information and then predicts disease-miRNA associations through a three-layer neural network [24]. In addition, matrix factorization is also used to predict the association between miRNAs and diseases. In 2019, Gao et al. proposed a computational model, dual-network sparse graph regularized matrix factorization (DNSGRMF), for predicting miRNA-disease associations by integrating the miRNA functional similarity matrix, the disease semantic similarity matrix and Gaussian kernel similarities with the addition of the  $L_{2,1}$  norm. They used collaborative matrix factorization to predict miRNA-disease associations [25]. Later, a more efficient miRNA-disease associations prediction model, nearest profile-based collaborative matrix factorization (NPCMF), was proposed by Gao et al., which integrates Gaussian kernel similarity and the nearest profile, taking the nearest neighbour information into account. Finally, DNSGRMF and NPCMF achieved excellent predictive accuracy based on fivefold cross-validation [26].

Although there are many advanced methods to predict MDAs, they still have some shortcomings. For instance, several methods trigger bias to miRNAs (diseases). Moreover, the small number of known relationships cannot be utilized to predict new miRNAs and diseases. More importantly, in some methods, the nearest neighbour information of the miRNA and the disease is not considered. To address the limitations of previous methods, a computational method of bipartite graphs based on collaborative matrix factorization (BGC MF) is proposed. The specific contributions of our method include the following two aspects:

In our method, the miRNA similarity matrix and disease similarity matrix are constructed by combining Gaussian interaction kernel similarity, miRNA functional similarity, and disease semantic similarity. This could help to infer potential miRNA-disease associations.

It is worth noting that the bipartite graph algorithm (BG) is introduced to our method for maximum consideration of neighbouring information for miRNAs and diseases. Then, it calculates a weighted average of similarities between miRNAs and diseases to eliminate the bias on prediction.

In addition, there are quite a few missing associations in the original matrix  $\mathbf{Y}$ , and Weight K Nearest Known Neighbours (WKNKN) [27] is implemented as a pre-treatment step to minimize the error. Moreover, five-fold cross-validation is exploited in our method to evaluate our experimental results. We also introduce simulation experiments to further evaluate the performance of our method. Overall, the results demonstrate that our BGC MF method is superior to other existing advanced methods.

## Results

### Human miRNA–disease associations association dataset

In this study, the gold standard human miRNA–disease association dataset was downloaded from the Human microRNA Disease Database (HMDD) v2.0 [28]. HMDD v2.0 includes 495 miRNAs, 383 diseases and 5430 experimentally verified miRNA–disease associations. In this paper, the dataset includes three matrices: the adjacency matrix  $\mathbf{Y}$ , the miRNA functional similarity matrix  $\mathbf{S}_m$ , and the disease semantic similarity matrix  $\mathbf{S}_d$ . The information for the dataset is listed in Table 1.

We use an adjacency matrix  $\mathbf{Y} \in \mathbb{R}^{n \times m}$  to describe the associations between miRNAs and diseases that have been validated, where  $n$  represents the number of miRNAs and  $m$  represents the number of diseases. When  $M(i)$  and  $D(j)$  are associated,  $\mathbf{Y}(M(i), D(j))$  is set to 1; otherwise,  $\mathbf{Y}(M(i), D(j))$  is set to 0. The following is the expression of the matrix  $\mathbf{Y}$ :

$$\mathbf{Y}(M(i), D(j)) = \begin{cases} 1, & M(i) \text{ is associated with disease } D(j), \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

### Performance evaluation

In this study, we implement fivefold cross-validation to evaluate the prediction performance of each method. The principle of fivefold cross-validation is to randomly divide the known miRNA–disease associations into five subsets, one of which is used as a test set, and the rest are used as a training set. Then, five models are trained by cycling five times, and the average of the five evaluation results is calculated as the final score of the model. Finally, fivefold cross-validation was performed 100 times, and the final score was taken as the average. It is worth noting that in our BGCMF method, WKNKN is used as a preprocessing procedure to evaluate unknown MDAs. At the same time, the nearest neighbour information is applied to our method, and it has the advantage of taking into account the nearest neighbour information and improving the accuracy of the prediction.

To verify the effect of the prediction, the area under the curve (AUC) value was applied in this study, which is widely used in previous studies. Therefore, a receiver operating characteristic (ROC) curve was obtained. In this curve, the x-axis is the false-positive rate (FPR, specificity), and the y-axis is the true positive rate (TPR, sensitivity). The definitions for calculating specificity and sensitivity are as follows:

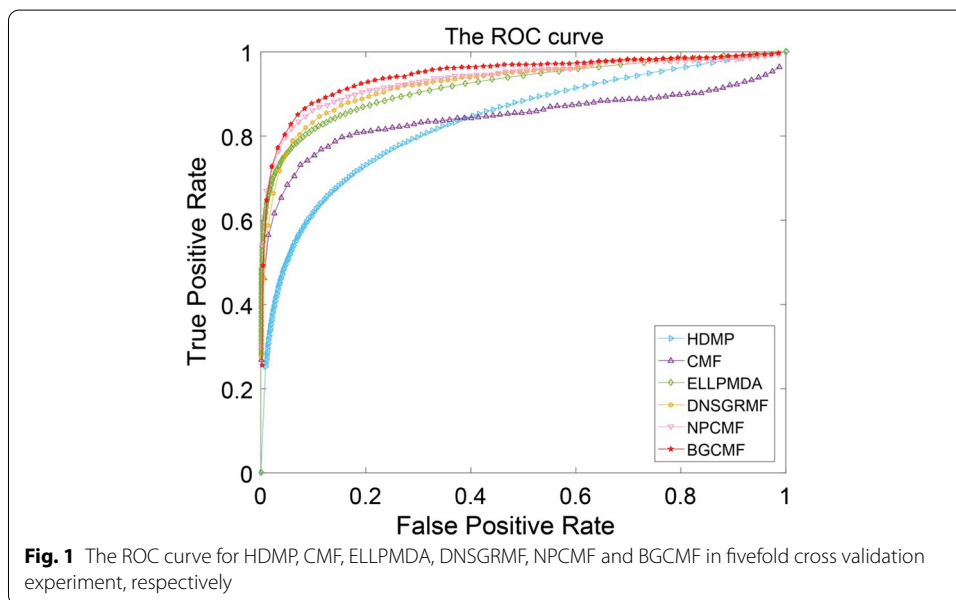
$$\text{Specificity} = \frac{TN}{TN + FP}, \quad (2)$$

**Table 1** The gold standard dataset of miRNAs, diseases and associations

Dataset	Gold standard dataset
MiRNAs	495
Diseases	383
Associations	5430

**Table 2** The AUC results of fivefold cross validation experiments

Methods	AUC
HDMP	0.8342(0.0010)
CMF	0.8697(0.0011)
ELLPMDA	0.9193(0.0002)
DNSGRMF	0.9304(0.0011)
NPCMF	0.9429(0.0011)
<b>BGCMF</b>	<b>0.9514(0.0007)</b>



$$Sensitivity = \frac{TP}{TP + FN}, \tag{3}$$

where *TP* represents the number of positive samples, *FP* represents the number of false-positive samples, *TN* represents the number of negative samples and *FN* represents the number of false-negative samples. An AUC value between 0.5 and 1 is considered feasible, with *AUC* = 1 representing the best predictive performance and *AUC* = 0.5 representing stochastic prediction.

**Comparison with other methods**

Based on experimentally confirmed associations between diseases and miRNAs, five-fold cross-validation is implemented in this paper to evaluate the predictive accuracy of BGCMF. We compared our method with other advanced methods, such as HDMP [21], CMF [29], ELLPMDA [23], DNSGRMF [25] and NPCMF [26]. The experimental results are listed in Table 2. More intuitively, the ROC curves are shown in Fig. 1. The AUCs of HDMP, CMF, ELLPMDA, DNSGRMF and NPCMF were 0.8342, 0.8697, 0.9193, 0.9304, and 0.9429, respectively, while the AUC of BGCMF was 0.9514. The best value is in bold.

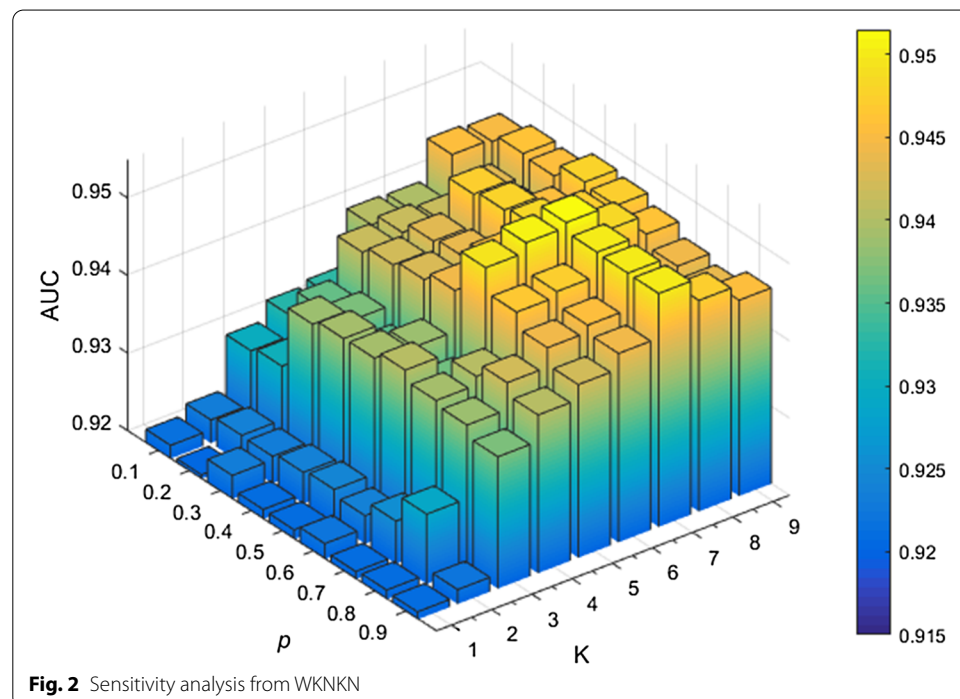
From the above statistical results, our method is 11.72% higher than the lowest value of HDMP. The BGCME is 2.1% and 0.85% higher than the values of DNSGRMF and NPCME, respectively. Therefore, we can conclude from the experimental results that the BGCME has excellent predictive performance.

### Sensitivity analysis from WKNKN

If a miRNA or a disease is known, it must have one or more associations. However, there are many missing unknown associations in the interaction matrix  $Y$ . WKNKN pre-processing is used to estimate the interaction possibilities to minimize the error. There are two parameters  $K$  and  $p$  in WKNKN, where  $K$  represents the number of known nearest neighbours and  $p$  represents the decay term for the neighbour. The value of  $p$  is between 0 and 1. As shown in Fig. 2, when  $K = 7$  and  $p = 0.6$ , the AUC value tends to be stable.

### Case study

The previous sections verify that our proposed method has outstanding predictive performance. Colon, prostate, and kidney are selected in the case study to further illustrate the superior performance of our BGCME. The known miRNA–disease associations in the standard dataset are used as a training set to predict potential disease-associated miRNAs. Our specific process first uses the BGCME method to predict these three diseases, and the choice of parameters is as described above. Then, the predicted score matrix is compared to the original miRNA–disease association matrix. The associations of predicted scores with changes are filtered and compared. Finally, we validated whether the predicted new miRNA–disease associations exist in the updated dbDEMOC [30], miR2Disease [31] and the HMDD v3.2 [32].



**Fig. 2** Sensitivity analysis from WKNKN

Colon neoplasms, also known as bowel cancer, are one of the three most common cancers, accounting for 10% of all cancer cases. Due to the low detection rate of colon tumours in the early stages, it creates a huge threat to people's lives. New biomarkers may help to improve the early detection of colon tumours. Recent studies have found that miRNA dysregulation can be used as a marker for colon tumour diagnosis in colon neoplasm cells. For example, miR-145 and miR-126 can inhibit the growth of colon neoplasm cells, and an increasing number of miRNAs associated with colon neoplasms have been found to be of great significance for improving the early detection of colon neoplasms. Here, the first type of case is colon neoplasms. In the dataset used in our experiments, there are 5 existing associations between miRNAs and colon neoplasms. After the simulation experiment is performed, the top 30 miRNAs of colon neoplasms are extracted, and all existing associations are successfully predicted. At the same time, 25 novel MDAs are predicted. Among these 25 new miRNAs, all of the miRNAs are validated by dbDEMOC, miR2Disease and HMDD v3.2. More importantly, fourteen of them are confirmed by the above three databases. For example, in 2007, Shi et al. found that the target gene of miR-145 is insulin receptor substrate-1 and can inhibit the growth of colon cancer cells [33]. In 2013, Wan et al. identified that patients with colon cancer with high expression of miR-199a-3p had a lower survival rate [34]. Table 3 lists the simulation results of colon tumours, and the known associations are shown in bold. *I, II, III* represent dbDEMOC, miR2Disease and the HMDD v3.2.

The next case is prostate neoplasms, which are the third most common cause of male cancer-related death. In our simulation experiments, we also select the top 30 miRNAs with the highest correlation scores, and seven known miRNAs associated with prostate neoplasms are successfully predicted. Among the 23 newly predicted miRNAs associated with prostate neoplasms, miR143, miR21, and miR126 are the highest ranked miRNAs, as confirmed by three databases at the same time. Only miR-200b is not confirmed in either dbDEMOCs or miR2Disease associated with prostate neoplasms, but it is confirmed by HMDD v3.2. Although a large number of miRNAs have been discovered,

**Table 3** Prediction MIRNAs for colon neoplasms

Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	<b>hsa-mir-145</b>	<b>Known</b>	16	hsa-mir-19b	<i>I,II</i>
2	<b>hsa-mir-1</b>	<b>Known</b>	17	hsa-let-7a	<i>I,II,III</i>
3	<b>hsa-mir-106a</b>	<b>Known</b>	18	hsa-mir-200b	<i>I,III</i>
4	<b>hsa-mir-126</b>	<b>Known</b>	19	hsa-mir-34a	<i>I,II,III</i>
5	<b>hsa-mir-17</b>	<b>Known</b>	20	hsa-mir-221	<i>I,II,III</i>
6	hsa-mir-143	<i>I,II,III</i>	21	hsa-mir-200c	<i>I,II,III</i>
7	hsa-mir-21	<i>I,II,III</i>	22	hsa-mir-146a	<i>I,III</i>
8	hsa-mir-155	<i>I,II,III</i>	23	hsa-mir-141	<i>I,II,III</i>
9	hsa-mir-20a	<i>I,II,III</i>	24	hsa-mir-19a	<i>I,II,III</i>
10	hsa-mir-125b	<i>I,III</i>	25	hsa-mir-29a	<i>I,II,III</i>
11	hsa-mir-9	<i>I,II</i>	26	hsa-mir-200a	<i>III</i>
12	hsa-mir-22	<i>I,III</i>	27	hsa-mir-142	<i>III</i>
13	hsa-mir-16	<i>I,III</i>	28	hsa-mir-7	<i>I</i>
14	hsa-mir-31	<i>I,II,III</i>	29	hsa-mir-92a	<i>III</i>
15	hsa-mir-18a	<i>I,II,III</i>	30	hsa-let-7b	<i>I,II,III</i>

knowledge regarding their function and physiological/pathological significance remains limited. Table 4 lists the details of the experiment and the existing associations.

The last case is kidney neoplasms. Kidney neoplasms, also known as kidney cancer, are cancers that originate in kidney cells and include several different types of tumours. Kidney neoplasms account for 3% of adult malignancies [35]. According to previous studies, a large number of miRNAs have been examined for kidney tumours. For example, circulating levels of miR-15b in patients with advanced kidney cancer are significantly reduced [36]. In addition, overexpression of miR-210 leads to the amplification of renal cancer cell centrosomes [37]. In this case, there are 9 miRNAs that have associations with kidney neoplasms. Nine known miRNAs are successfully predicted in our results. Simultaneously, 35 new miRNAs are predicted. Among the 35 new miRNAs, 29 miRNAs connected with kidney neoplasms have discovered experimental proof from three databases. For example, studies have found that miR-17 is carcinogenic and overexpressed in renal cell carcinoma. Although the predicted novel miRNAs, including miR205, miR125b, miR-7, miR-221, miR-31, and miR-92a, are unconfirmed by miR2Disease or dbDEMC, these miRNAs are closely associated with kidney neoplasms. Table 5 lists the simulation results of kidney neoplasms, and the known associations are shown in bold. To show the simulation experiment of BGCMF more intuitively, Cytoscape software was used to map the three predicted disease-miRNA association networks. As shown in Fig. 3, large ellipses indicate the three diseases, and small ellipses indicate the predicted miRNAs.

## Discussion

MiRNAs are involved in many physiological processes, such as organismal development, cell differentiation and proliferation, apoptosis, hormone secretion, and lipid metabolism. miRNAs are closely related to the occurrence and development of tumours, metabolic diseases, stress diseases, and cardiovascular diseases. With the development of miRNA bioinformatics, direction prediction and other advances in biological science

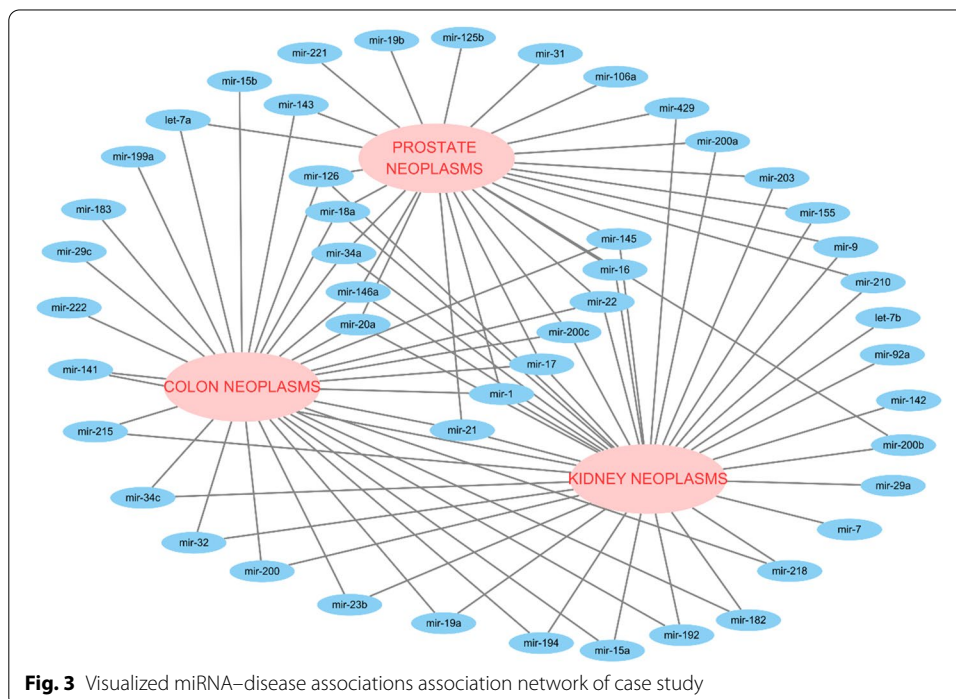
**Table 4** Prediction MIRNAs for prostate neoplasms

Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	<b>hsa-mir-125b</b>	<b>Known</b>	16	hsa-mir-100	I; III
2	<b>hsa-mir-1</b>	<b>Known</b>	17	hsa-mir-375	I; III
3	<b>hsa-mir-183</b>	<b>Known</b>	18	hsa-mir-20a	I; III
4	<b>hsa-mir-145</b>	<b>Known</b>	19	hsa-mir-31	I; III
5	<b>hsa-mir-99a</b>	<b>Known</b>	20	hsa-mir-7	I; III
6	<b>hsa-mir-9</b>	<b>Known</b>	21	hsa-mir-96	I; III
7	<b>hsa-mir-574</b>	<b>Known</b>	22	hsa-mir-200a	I; III
8	hsa-mir-143	I; II; III	23	hsa-mir-200b	III
9	hsa-mir-21	I; II; III	24	hsa-mir-34a	I; II; III
10	hsa-mir-126	I; II; III	25	hsa-mir-141	I; III
11	hsa-mir-182	I; II; III	26	hsa-mir-221	I; III
12	hsa-mir-133a	I; III	27	hsa-mir-155	I; III
13	hsa-mir-199a	I; II; III	28	hsa-mir-17	II; III
14	hsa-mir-223	I; II; III	29	hsa-mir-200c	I; III
15	hsa-mir-22	I; III	30	hsa-mir-146a	II; III



**Table 5** Prediction MIRNAs for kidney neoplasms

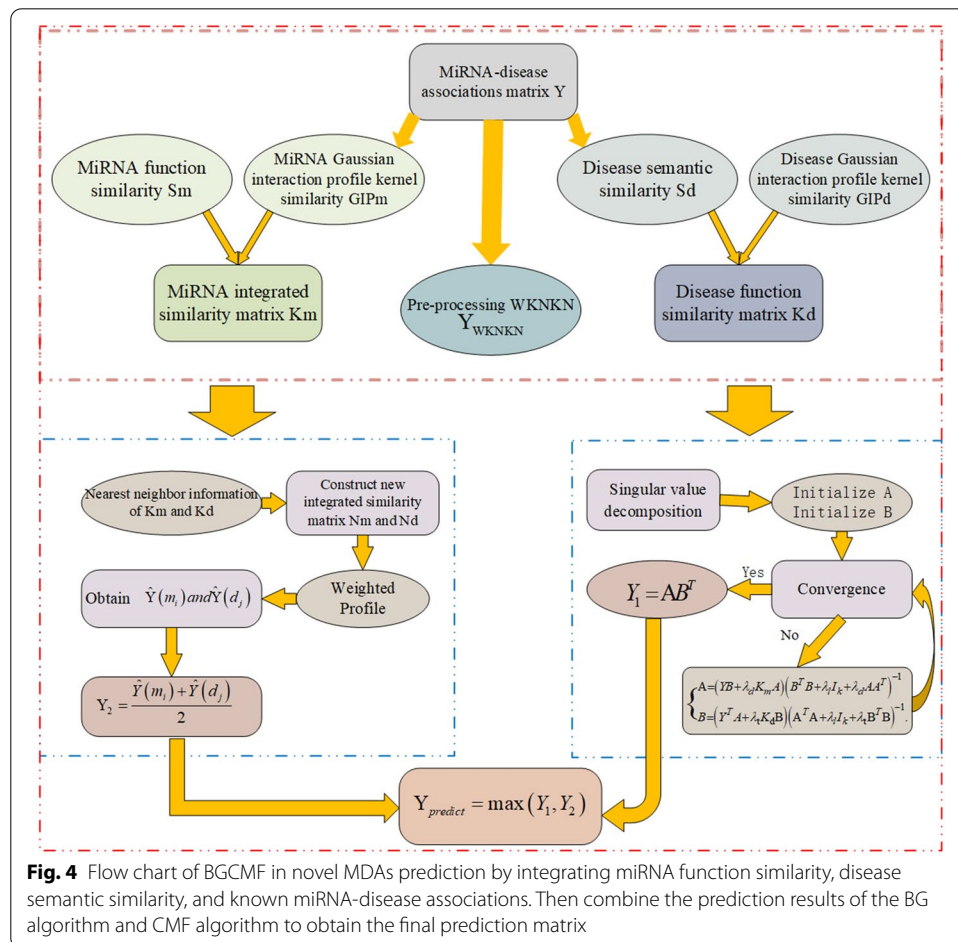
Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	hsa-mir-141	Known	23	hsa-mir-34a	/
2	hsa-mir-15a	Known	24	hsa-mir-126	I; II; III
3	hsa-mir-21	Known	25	hsa-mir-146a	/
4	hsa-mir-1	Known	26	hsa-mir-7	unconfirmed
5	hsa-mir-192	Known	27	hsa-mir-221	unconfirmed
6	hsa-mir-200c	Known	28	hsa-mir-17	II; III
7	hsa-mir-215	Known	29	hsa-mir-31	unconfirmed
8	hsa-mir-23b	Known	30	hsa-mir-92a	unconfirmed
9	hsa-mir-200	Known	31	hsa-mir-15b	/
10	hsa-mir-200b	I; II; III	32	hsa-mir-19a	/
11	hsa-mir-200a	I; III	33	hsa-mir-143	/
12	hsa-mir-9	/	34	hsa-mir-29c	I; III
13	hsa-mir-16	/	35	hsa-mir-183	I; III
14	hsa-mir-155	I; III	36	hsa-let-7a	/
15	hsa-mir-210	I; II	37	hsa-mir-222	/
16	hsa-mir-429	/	38	hsa-mir-199a	I; II; III
17	hsa-mir-203	/	39	hsa-mir-182	I; II
18	hsa-mir-205	Unconfirmed	40	hsa-mir-32	/
19	hsa-mir-125b	Unconfirmed	41	hsa-mir-18a	/
20	hsa-mir-20a	I; II	42	hsa-mir-194	/
21	hsa-mir-145	/	43	hsa-mir-34c	/
22	hsa-mir-22	/	44	hsa-mir-218	/



and technology, a large number of miRNAs have been discovered and verified. However, validating the associations between miRNAs and diseases through biological experiments is time-consuming and expensive. Therefore, it is absolutely necessary to develop new and effective computational models to predict potential associations between miRNAs and diseases. In this paper, an efficient and useful method to predict potential MDAs is developed. Our method can be divided into three parts. The entire calculation process is described in detail in Fig. 4. The first step in this method is to process the data for subsequent prediction. Then, we use the CMF algorithm and BG algorithm to make predictions based on the processed data separately. Finally, we combine the prediction results of the two algorithms to obtain the final prediction matrix. The BGCMP achieves an overall result that is better than the two results given by single models.

### Conclusions

The success of our method can be mainly attributed to several factors. First, we combined Gaussian interaction profile similarity with miRNA functional similarity and disease semantic similarity to obtain accurate information about miRNA pairs and disease pairs. In addition, WKNKN is an essential pretreatment process. It is worth noting that our largest contribution is combining the bipartite graph algorithm with the collaborative matrix



factorization model. This allows for maximum consideration of neighbouring information for miRNAs and diseases, preventing the network similarity of miRNAs and diseases from being affected. Finally, both fivefold cross-validation results and three kinds of case studies on colon neoplasms, prostate neoplasms, and kidney neoplasms demonstrated the reliable prediction performance of BGCMF.

In the future, an increasing number of useful methods will be applied to predict potential MDAs. We will continue to study this aspect of research. At the same time, more meaningful datasets are being published in online bio-databases. Therefore, our next work will focus on developing effective methods to predict novel miRNA–disease associations and to evaluate the effectiveness of the method on diverse datasets.

**Method**

This novel method is named bipartite graph-based collaborative matrix factorization (BGCMF). The method is divided into two major steps. First, the Gaussian interaction profile kernel (GIP) and nearest neighbour profile (NP) are introduced in our method to process the original miRNA matrix and the disease matrix to obtain their network information. At the same time, WKNKN is used to handle the original interaction matrix  $Y$  to minimize the error. Second, the BG algorithm is implemented to obtain prediction matrix  $Y_1$  and collaborative matrix factorization (CMF) to obtain the prediction matrix  $Y_2$ , respectively. Finally, the prediction matrix  $Y_{predict}$  is obtained by combining our two improved models. The flowchart of BGBMF is shown in Fig. 4.

**MiRNA functional similarity**

With the hypothesis that functionally similar miRNAs tend to be associated with phenotypically similar diseases, a computing method of miRNA functional similarity was presented by Wang et al. [10]. The functional similarity score matrix can be downloaded from <http://www.cuilab.cn/files/images/cuilab/misim.zip>. Here, the obtained functional similarity for miRNA is denoted by  $S_m \in \mathbb{R}^{n \times n}$ , and the value of entity  $S(M(i), M(j))$  measures the closeness between miRNA  $M(i)$  and  $M(j)$ .

**Disease semantic similarity**

A directed acyclic graph (DAG) is proposed to describe the relationships among various diseases. In addition, the disease  $D$  can be described by  $DAG(D) = (D, T(D), E(D))$ .  $T(D)$  is the node set and represents both its ancestor nodes and  $D$  itself.  $E(D)$  is used to represent all direct edges between child nodes and parent nodes. The semantic similarity value of disease  $D$  is as follows:

$$SV1(D) = \sum_{d \in T(D)} D1_D(d), \tag{4}$$

$$D1_D(d) = \begin{cases} 1 & \text{if } d = D, \\ \max \{ \Delta * D1_D(d') \mid d' \in \text{children of } d \} & \text{if } d \neq D, \end{cases} \tag{5}$$

where  $\Delta$  represents the semantic contribution factor and  $D1_D(d)$  is the contribution of disease  $d$ . For each disease  $d$ , its contribution to itself is 1, and the contribution of its

child node decreases with increasing distance. Obviously, when the two diseases have a larger shared part in their DAGs, they will obtain a greater similarity score.  $SV(d_i)$  and  $SV(d_j)$  represent the semantic similarity values of  $d_i$  and  $d_j$ , respectively. Thus, the semantic similarity score of the two diseases  $d_i$  and  $d_j$  can be calculated as follows:

$$S_d(d_i, d_j) = \frac{\sum_{t \in T(d_i) \cap T(d_j)} (D_{d_i}(t) + D_{d_j}(t))}{SV(d_i) + SV(d_j)}. \quad (6)$$

### Gaussian Interaction Profile Kernel for miRNAs and diseases

According to the previous work [38], the method is based on the idea that it relies on the topological structure of known miRNA–disease associations in a network to compute the similarity of diseases and miRNAs [26]. Here are two miRNAs  $m_i$  and  $m_j$  and two diseases  $d_i$  and  $d_j$ . The network similarity between them can be calculated with the following formulas:

$$GIP_{miRNA}(m_i, m_j) = \exp\left(-\gamma \|\mathbf{Y}(m_i) - \mathbf{Y}(m_j)\|^2\right), \quad (7)$$

$$GIP_{disease}(d_i, d_j) = \exp\left(-\gamma \|\mathbf{Y}(d_i) - \mathbf{Y}(d_j)\|^2\right), \quad (8)$$

where  $\gamma$  is an adjustable parameter that can control the bandwidth of the kernel. In addition,  $\mathbf{Y}(m_i)$  and  $\mathbf{Y}(m_j)$  are the miRNA interaction profiles of  $m_i$  and  $m_j$ , respectively. Similarly,  $\mathbf{Y}(d_i)$  and  $\mathbf{Y}(d_j)$  are the disease interaction profiles of  $d_i$  and  $d_j$ , respectively. Then, the network similarity matrix  $\mathbf{K}_m$  of miRNA and the  $\mathbf{K}_d$  of disease are obtained by combining the original matrix  $\mathbf{S}_m$  and  $\mathbf{S}_d$ . The detailed descriptions are as below:

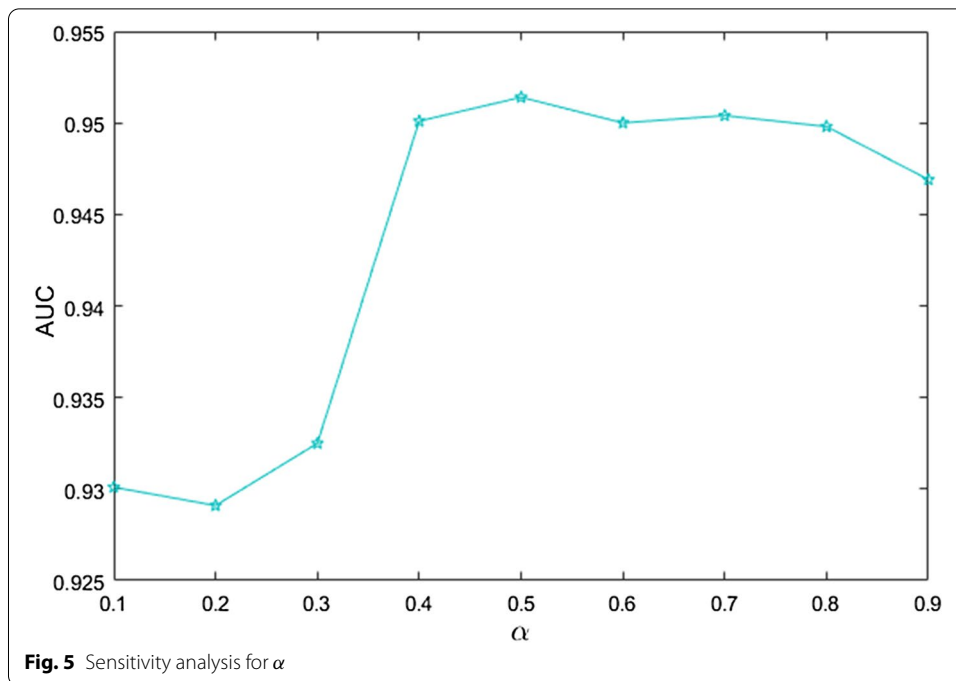
$$\mathbf{K}_m = \alpha \mathbf{S}_m + (1 - \alpha) \mathbf{GIP}_{miRNA}, \quad (9)$$

$$\mathbf{K}_d = \alpha \mathbf{S}_d + (1 - \alpha) \mathbf{GIP}_{disease}, \quad (10)$$

where  $\alpha$  is an adjustable parameter range in [0, 1], and  $\mathbf{K}_m$  represents the miRNA integrated similarity matrix, which is a linear combination of the Gaussian interaction profile kernel similarity for miRNA  $\mathbf{GIP}_{miRNA}$  and the miRNA functional matrix  $\mathbf{S}_m$ . Similar to  $\mathbf{K}_m$ ,  $\mathbf{K}_d$  represents the disease integrated similarity matrix, which is a linear combination of the Gaussian interaction profile kernel similarity for disease  $\mathbf{GIP}_{disease}$  and the disease semantic matrix  $\mathbf{S}_d$ . When  $\alpha$  is equal to 0.5, BGCMF achieves the highest AUC value. The sensitivity analysis of  $\alpha$  is shown in Fig. 5.

### Bipartite graph method

Based on the assumption that miRNAs that are similar will interact with similar diseases, the interaction profile for a new miRNA candidate could be inferred from the known interactions of their neighbours. MiRNAs with large similarities to new potential miRNAs are said to be their neighbours. Therefore, we introduce the nearest profile (NP) to our method [39]. Below are the formulas for calculating a new miRNA  $m_i$  and a new disease  $d_i$ .



$$\mathbf{N}_m(m_i) = \mathbf{K}_m(m_i, m_{nearest}) \times \mathbf{Y}(m_{nearest}), \tag{11}$$

$$\mathbf{N}_d(d_i) = \mathbf{K}_d(d_i, d_{nearest}) \times \mathbf{Y}(d_{nearest}), \tag{12}$$

where  $m_{nearest}$  and  $d_{nearest}$  are the miRNAs most similar to  $m_i$  and the diseases most similar to  $d_i$ , respectively.  $\mathbf{N}_m(m_i)$  and  $\mathbf{N}_d(d_i)$  are the association profiles of the miRNAs and diseases, respectively. The NP process in this method can be divided into four steps. First, remove the self-similarity of miRNA matrices  $\mathbf{K}_m$  and  $\mathbf{K}_d$ . Next, obtain the nearest neighbour for each miRNA and disease. Then, ignore all miRNA similarities and disease similarities. Finally, the miRNA nearest neighbour matrix  $\mathbf{N}_m$  and disease nearest neighbour matrix  $\mathbf{N}_d$  can be obtained.

### Weighted profile

The weighted profile (WP) is proposed as a simple predictive model in [39]. The idea of the weighted profile is to perform a similarity-weighted average of all other miRNAs or diseases to obtain the prediction matrix. For instance, the WP for a new miRNA  $m_i$  and a new disease are computed as:

$$\hat{\mathbf{Y}}(m_i) = \frac{\sum_{j=1}^{n_m} \mathbf{N}_m(m_i, m_j) \times \mathbf{Y}(m_j)}{\sum_{j=1}^{n_m} \mathbf{N}_m(m_i, m_j)}, \tag{13}$$

$$\hat{\mathbf{Y}}(d_i) = \frac{\sum_{j=1}^{n_d} \mathbf{N}_d(d_i, d_j) \times \mathbf{Y}(d_j)}{\sum_{j=1}^{n_d} \mathbf{N}_d(d_i, d_j)}, \tag{14}$$

where  $\mathbf{N}_m$  and  $\mathbf{N}_d$  are the nearest neighbour matrices we construct for miRNA and disease.  $\mathbf{Y}(m_i)$  and  $\mathbf{Y}(d_j)$  are association matrices of miRNA  $m_i$  and disease  $d_j$ , respectively. First, the BG algorithm is used to obtain the neighbour information about miRNAs and diseases, and then predictions from both miRNA and disease sides are averaged to obtain the final prediction matrix:

$$\mathbf{Y}_1 = \frac{\hat{\mathbf{Y}}(m_i) + \hat{\mathbf{Y}}(d_j)}{2}. \tag{15}$$

**BGCMF for MiRNA-disease associations association prediction**

The traditional collaborative matrix factorization (CMF) method is effective in predicting the underlying interactions between miRNAs and diseases [29]. The objective function of CMF method is defined as:

$$\min_{\mathbf{A}, \mathbf{B}} = \left\| \mathbf{Y} - \mathbf{A}\mathbf{B}^T \right\|_F^2 + \lambda_l \left( \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 \right) + \lambda_d \left\| \mathbf{S}_m - \mathbf{A}\mathbf{A}^T \right\|_F^2 + \lambda_t \left\| \mathbf{S}_d - \mathbf{B}\mathbf{B}^T \right\|_F^2, \tag{16}$$

where  $\lambda_l$ ,  $\lambda_d$ , and  $\lambda_t$  are non-parameters and  $\|\cdot\|_F^2$  represents the Frobenius norm. In this formula, the first item is used to find the low-rank matrices  $\mathbf{A}$  and  $\mathbf{B}$  of the reconstructed  $\mathbf{Y}$ . The second item is the Tikhonov regularization term. The last two items are regularization terms that demand potential feature vectors of similar miRNAs/diseases to be similar and potential feature vectors of dissimilar miRNAs/diseases to be dissimilar. However, traditional CMF does not take into account the network relationship between the miRNA and the disease, which will reduce the accuracy of predicting MDAs. Therefore, we introduce the Gaussian kernel similarity  $\mathbf{K}_m$  of miRNA and the  $\mathbf{K}_d$  of disease into CMF [40]. The objective function can be rewritten as:

$$\min_{\mathbf{A}, \mathbf{B}} = \left\| \mathbf{Y} - \mathbf{A}\mathbf{B}^T \right\|_F^2 + \lambda_l \left( \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 \right) + \lambda_d \left\| \mathbf{K}_m - \mathbf{A}\mathbf{A}^T \right\|_F^2 + \lambda_t \left\| \mathbf{K}_d - \mathbf{B}\mathbf{B}^T \right\|_F^2, \tag{17}$$

where  $\|\cdot\|_F^2$  is the Frobenius norm.  $\lambda_l$ ,  $\lambda_d$  and  $\lambda_t$  represent the positive parameters. In this study, the setting of the three parameters is done by cross-validation. The grid search is adopted to select the optimal parameters among these values:  $\lambda_l \in \{2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$ ,  $\lambda_d/\lambda_t \in \{2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$ . The association matrix  $\mathbf{Y}$  is decomposed into two low-rank matrices  $\mathbf{A}$  and  $\mathbf{B}$ , where  $\mathbf{Y} \approx \mathbf{A}\mathbf{B}^T$ . Tikhonov regularization is adopted to minimize the norms of both  $\mathbf{A}$  and  $\mathbf{B}$ . The roles of the third and fourth terms are to minimize the squared error  $\mathbf{S}_m \approx \mathbf{A}\mathbf{A}^T$  and  $\mathbf{S}_d \approx \mathbf{B}\mathbf{B}^T$ , respectively.

**Initialization of A and B**

In the CMF method, the first step is to initialize the adjacency matrix  $\mathbf{Y}$ . We use singular value decomposition (SVD) to decompose the input matrix  $\mathbf{Y} \in \mathbb{R}^{n \times m}$  into  $\mathbf{U}^{n \times k}$ ,  $\mathbf{S}^{k \times k}$  and  $\mathbf{V}^{k \times m}$ . Then, matrix  $\mathbf{A}$  and matrix  $\mathbf{B}$  are obtained by the following formula:

$$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = SVD(\mathbf{Y}, k), \quad \mathbf{A} = \mathbf{U}\mathbf{S}_k^{1/2}, \quad \mathbf{B} = \mathbf{V}\mathbf{S}_k^{1/2}, \tag{18}$$

where  $\mathbf{S}$  is a diagonal matrix and  $k$  represents the maximum number of singular values.

### Alternating least squares

In this study, alternating least squares is used to optimize  $\mathbf{A}$  and  $\mathbf{B}$  until convergence. Here,  $L$  is used to represent the objective function of BGCMF. Then,  $\mathbf{A}$  and  $\mathbf{B}$  are obtained by letting  $\partial L/\partial \mathbf{A} = 0$ , and  $\partial L/\partial \mathbf{B} = 0$ , respectively. Moreover, the optimal values of  $\lambda_l$ ,  $\lambda_d$  and  $\lambda_t$  are automatically obtained through a fivefold cross-validation experiment. The iterative formulas for  $\mathbf{A}$  and  $\mathbf{B}$  are represented by:

$$\mathbf{A} = (\mathbf{YB} + \lambda_d \mathbf{K}_m \mathbf{A}) (\mathbf{B}^T \mathbf{B} + \lambda_l \mathbf{I}_k + \lambda_d \mathbf{A} \mathbf{A}^T)^{-1}, \quad (19)$$

$$\mathbf{B} = (\mathbf{Y}^T \mathbf{A} + \lambda_t \mathbf{K}_d \mathbf{B}) (\mathbf{A}^T \mathbf{A} + \lambda_l \mathbf{I}_k + \lambda_d \mathbf{B}^T \mathbf{B})^{-1}. \quad (20)$$

Finally, the final prediction matrix  $\mathbf{Y}$  is obtained by combining both the BG algorithm and the optimized CMF model.

### Acknowledgements

Not applicable.

### Authors' contributions

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

### Funding

This work was supported in part by the grants of the National Natural Science Foundation of China, Nos. 62172254, and 61872220. The funder played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

### Availability of data and materials

The datasets that support the findings of this study are available in <https://github.com/zhoufeng-coder/>. The origins of the data used in the Case Studies in this paper are available on open-source data PMID: 24194601 (<http://www.cuilab.cn/hmdd>).

### Declarations

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

Received: 27 November 2020 Accepted: 17 November 2021

Published online: 27 November 2021

### References

1. Ambros V. The functions of animal microRNAs. *Nature*. 2004;431(7006):350–5.
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.
3. 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):901–6.
4. Chen CZ, Li L, Lodish HF, Bartel DP. MicroRNAs modulate hematopoietic lineage differentiation. *Science*. 2004;303(5654):83–6.
5. Ambros V. MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. *Cell*. 2003;113(6):673–6.
6. Miska EA. How microRNAs control cell division, differentiation and death. *Curr Opin Genet Dev*. 2005;15(5):563–8.
7. Alshalalfa A. Using context-specific effect of miRNAs to identify functional; associations between miRNAs and gene signatures. *BMC Bioinform*. 2013;14(S12):S1.
8. 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.

9. Latronico MV, Catalucci D, Condorelli G. Emerging role of microRNAs in cardiovascular biology. *Circ Res*. 2007;101(12):1225–36.
10. Bandyopadhyay S, Mitra R, Maulik U, Zhang MQ. Development of the human cancer microRNA network. *Silence*. 2010;1(1):6–6.
11. Feng G, Jingxia C, Huaqi W, Guojun Z. Potential diagnostic value of miR-155 in serum from lung adenocarcinoma patients. *Oncol Rep*. 2014;31(1):351–7.
12. Junichi T, Hiroyuki K, Kiyoshi Y, Shuta T, Hirotaka O, Hideki E, Tomoko H, Yasushi Y, Masato N, Yuji N. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Can Res*. 2004;64(11):3753–6.
13. Huang Z, Liu L, Gao Y, Shi J, Cui Q, Li J, Zhou Y. Benchmark of computational methods for predicting microRNA-disease associations. *Genome Biol*. 2019;20(1):1–13.
14. Gao Y, Jia K, Shi J, Zhou Y, Cui Q. A computational model to predict the causal miRNAs for diseases. *Front Genet*. 2019;10:935.
15. Xuan P, Han K, Guo M, Guo Y, Li J, Ding J, Liu Y, Dai Q, Li J, Teng Z. Correction: prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS ONE*. 2013;8(9):e70204.
16. Chen H, Zhang Z. Similarity-based methods for potential human microRNA-disease association prediction. *BMC Med Genom*. 2013;6(1):12.
17. Li J, Zhang S, Wan Y, Zhao Y, Shi J, Zhou Y, Cui Q. MISIM v2.0: a web server for inferring microRNA functional similarity based on microRNA-disease associations. *Nucl Acids Res*. 2019;47(W1):W536–41.
18. Jiang Q, Wang G, Jin S, Li Y, Wang Y. Predicting human microRNA-disease associations based on support vector machine. *Int J Data Min Bioinform*. 2013;8(3):282–93.
19. Chen H, Zhang Z. Prediction of associations between OMIM diseases and microRNAs by random walk on OMIM disease similarity network. *Sci World J*. 2013;2013(10):1–6.
20. Qabaja A, Alshalfalfa M, Bismar TA, Alhaji R. Protein network-based Lasso regression model for the construction of disease-miRNA functional interactions. *EURASIP J Bioinf Syst Biol*. 2013;2013(1):3–3.
21. 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.
22. Chen X, Huang L. LRSSLMDA: Laplacian regularized sparse subspace learning for miRNA-disease associations Association prediction. *PLoS Comput Biol*. 2017;13(12):e1005912.
23. Chen X, Zhou Z, Zh YA. ELLPMDA: ensemble learning and link prediction for miRNA-disease associations Association prediction. *RNA Biol*. 2018;15(6):807–18.
24. Fu L, Peng Q. A deep ensemble model to predict miRNA-disease associations association. *Sci Rep*. 2017;7(1):14482.
25. 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. *Mol Omics*. 2019;15(2):130–7.
26. 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 Bioinform*. 2019;20(1):353.
27. Ezzat A, Zhao P, Wu M, Li X-L, Kwok C-K. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans Comput Biol Bioinf*. 2016;14(3):646–56.
28. 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. *Nucl Acids Res*. 2014;42(D1):D1070–4.
29. Shen Z, Zhang Y-H, Han K, Nandi AK, Honig B, Huang D-S. miRNA-disease associations association prediction with collaborative matrix factorization. *Complexity*. 2017;2017:2498957.
30. Yang Z, Wu L, Wang A, Tang W, Zhao Y, Zhao H, Teschendorff AE. dbDEMOC 2.0: updated database of differentially expressed miRNAs in human cancers. *Nucl Acids Res*. 2017;45(D1):D812–8.
31. 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. *Nucl Acids Res*. 2009;37(suppl\_1):D98–104.
32. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q. HMDD v3. 0: a database for experimentally supported human microRNA-disease associations. *Nucl Acids Res*. 2019;47(D1):D1013–17.
33. Shi B, Seppolorenzino L, Prisco M, Linsley P, Deangelis T, Baserga R. Micro RNA 145 targets the insulin receptor substrate-1 and inhibits the growth of colon cancer cells. *J Biol Chem*. 2007;282(45):32582–90.
34. Wan D, He S, Gu W, Shen C, Hu Y. Aberrant expression of miR-199a-3p and its clinical significance in colorectal cancers. *Med Oncol*. 2013;30(1):378.
35. Ahmedin J, Rebecca S, Elizabeth W, Taylor M, Xu J, Carol S, Thun MJ. Cancer statistics, 2006. *CA Cancer J Clin*. 2006;56(2):106–30.
36. Wang H, Peng W, Ouyang X, Dai Y. Reduced circulating miR-15b is correlated with phosphate metabolism in patients with end-stage renal disease on maintenance hemodialysis. *Ren Fail*. 2012;34(6):685–90.
37. Nakada C, Tsukamoto Y, Matsuura K, Nguyen TL, Hijijya N, Uchida T, Sato F, Mimata H, Seto M, Moriyama M. Overexpression of miR-210, a downstream target of HIF1 $\alpha$ , causes centrosome amplification in renal carcinoma cells †. *J Pathol*. 2011;224(2):280–8.
38. Van LT, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics*. 2011;27(21):3036.
39. Yoshihiro Y, Michihiro A, Alex G, Wataru H, Minoru K. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*. 2008;24(13):i232–40.
40. Ding H, Takigawa I, Mamitsuka H, Zhu S. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Brief Bioinform*. 2013;15(5):734–47.

## Publisher's Note

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