frontiers in Genetics

ORIGINAL RESEARCH
published: 16 November 2020
doi: 10.3389/fgene.2020.598185

Prediction of Potential Associations Between miRNAs and Diseases Based on Matrix Decomposition

Pengcheng Sun, Shuyan Yang, Ye Cao, Rongjie Cheng and Shiyu Han**

Department of Obstetrics and Gynecology, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, China

It is known that miRNA plays an increasingly important role in many physiological processes. Disease-related miRNAs could be potential biomarkers for clinical diagnosis, prognosis, and treatment. Therefore, accurately inferring potential miRNAs related to diseases has become a hot topic in the bioinformatics community recently. In this study, we proposed a mathematical model based on matrix decomposition, named MFMDA, to identify potential miRNA–disease associations by integrating known miRNA and disease-related data, similarities between miRNAs and between diseases. We also compared MFMDA with some of the latest algorithms in several established miRNA disease databases. MFMDA reached an AUC of 0.9061 in the fivefold cross-validation. The experimental results show that MFMDA effectively infers novel miRNA–disease associations. In addition, we conducted case studies by applying MFMDA to three types of high-risk human cancers. While most predicted miRNAs are confirmed by external databases of experimental literature, we also identified a few novel disease-related miRNAs for further experimental validation.

Keywords: miRNA, matrix decomposition (MFMDA), endometrial cancer, miRNA–disease association, computational prediction model

OPEN ACCESS

Edited by:
Jialiang Yang,
Geneis (Beijing) Co. Ltd., China

Reviewed by:
Ali Salehzadeh-Yazdi,
University of Rostock, Germany
JunLin Xu,
Hunan University, China
Lan Yu,
Inner Mongolia People's Hospital, China

****Correspondence:***
Shiyu Han
shiyuhan62@163.com
Rongjie Cheng
rongjie_jie@126.com

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *24 August 2020*
Accepted: *22 October 2020*
Published: *16 November 2020*

Citation:
Sun P, Yang S, Cao Y, Cheng R and Han S (2020) Prediction of Potential Associations Between miRNAs and Diseases Based on Matrix Decomposition. Front. Genet. 11:598185. doi: 10.3389/fgene.2020.598185

INTRODUCTION

Non-coding RNA (ncRNA) is a type of RNA that cannot be translated into protein. Although ncRNA cannot be translated into protein, its target gene can be regulated at the post-transcriptional level, thereby affecting disease (Hammond, 2015). A large amount of research evidence indicates that mutations and disorders of ncRNA are important causes of disease. Therefore, the identification of disease-related ncRNA has become an important topic in the field of biological research in recent years. ncRNA is a huge family and can be divided into housekeeper ncRNA and regulatory ncRNA (Kapranov et al., 2007; Lindsay et al., 2017). Housekeeping ncRNA is closely related to cell function, mainly involved in gene translation, gene splicing, gene modification, etc. The main function of regulating ncRNA is to regulate the expression level of genes. As regulatory ncRNA, miRNA is a class of non-coding single-stranded RNA molecules with a length of 22 nucleotides encoded by endogenous genes. They participate in the regulation of post-transcriptional gene expression in animals and plants (Taft et al., 2007; Chen et al., 2015). So far, 28645 miRNA molecules have been found in animals, plants, and viruses. Most miRNA genes exist in the genome in the form of single copies, multiple copies, or gene clusters (Wang and Chang, 2011).

In recent years, more and more studies have shown that miRNA plays a huge role in the process of cell differentiation, biological development, and disease development, which has also attracted more researchers' attention (Xu et al., 2004; Jiang et al., 2012; Li et al., 2014; Kang et al., 2020). With further in-depth research on the mechanism of action of miRNA, and the use of the latest high-throughput technologies such as miRNA chips to study the relationship between miRNA and disease, people will make higher eukaryote gene expression regulation Network understanding has improved to a new level (Cui et al., 2006). This will also make miRNA a new biological marker for disease diagnosis; it may also make this molecule a drug target, or simulate this molecule for new drug development, which will likely provide a new treatment for human diseases (Goh et al., 2016).

However, using biological experiments to identify disease-associated miRNAs is expensive and time-consuming, and it is blind. Therefore, there is an urgent need for simple and effective computational prediction models for predicting disease-related miRNAs. With the rapid development of high-throughput sequencing technology, more and more omics data are published, which also provides data support for the study of computational prediction models (Yi et al., 2017).In recent years, many scholars have proposed some effective computational models for predicting miRNA related to complex diseases. According to their respective implementation strategies, we can roughly divide these methods into machine-based computational prediction methods and network-based computational prediction methods (Zou et al., 2016).

Machine learning-based computational prediction methods predict the association of potential miRNAs with the disease can be divided into supervised-based machine learning methods and semi-supervised-based machine learning methods. The method based on supervision is mainly based on labeling sample set and label-less sample set to construct a machine learning model. Jiang et al. extracted feature sets based on known and unknown associations for training support vector machine (SVM) classifiers to predict potential miRNAs and disease associations, and achieved comparative prediction performance through cross-validation (Maly et al., 2019). Qu et al. (Zou et al., 2015) developed a new calculation method based on the KATZ model to predict MiRNA disease association (KATZMDA) by integrating multiple data sources. Based on the known miRNA–disease association in the HMDD database, Li et al. (2017) developed a MiRNA–disease association prediction model (MCMDA) called the matrix completion algorithm. The MCMDA model uses a matrix completion algorithm to update the adjacency matrix of known miRNA–disease associations and further predict potential associations. Xu et al. (Chen et al., 2018) proposed a method based on low-rank matrix completion to predict miRNA–disease association (LRMCMDA). LRMCMDA first constructs negative samples based on known associations, and then uses a low-rank matrix to complete the model to infer all miRNA and disease associations. Cross-validation shows that the model has obtained reliable prediction performance. However, although this supervised machine learning method uses different ways to define negative sample data, it is difficult to deal with the actual situation in any way, which will affect the prediction performance. In order to overcome this limitation, Chen and Yan (2014) proposed a least-squares-based semi-supervised machine learning method for predicting the association of potential miRNAs with disease, referred to as RLSMDA for short. The RLSMDA method constructs a continuous classifier function, and the predicted value reflects the probability score between specific miRNAs and specific diseases. This method can obtain the predicted values of all miRNAs and diseases at the same time, and does not require negative sample data. In addition, the RLSMDA method can also predict miRNAs associated with isolated diseases. Xu et al. (2019) designed a set of probabilistic matrix decomposition algorithms by integrating the similarity of miRNAs with diseases, using known correlation matrices and integrated similarity matrices to identify miRNAs that are potentially related to diseases. Luo et al. (2017) proposed a semi-supervised method called KRLSM to reveal the association between miRNA and disease. Machine learning has been a hot topic in recent years, and some machine learning methods can be used to solve this problem. Despite the outstanding contributions made by existing methods, there is still room for improvement in prediction accuracy.

In addition to machine learning-based methods, network-based methods to predict disease-related miRNAs have also attracted the attention of many researchers. Such methods are mainly based on a common biological hypothesis, "miRNAs with similar functions are more likely to be associated with disease phenotypes with similar functions, and vice versa" (Jiang et al., 2010). Based on this basic assumption, Jiang et al. proposed a new method that uses Bayesian models to integrate genomic data to rank disease-related miRNAs. Chen et al. (2012) adopted the global network similarity measure and proposed an improved restart-based random walk model (RWRMDA) to predict the association between miRNAs and disease. Yet, this method is not suitable for predicting new disease-related miRNAs. Xuan et al. (2013) integrated the information entropy of disease entries and the similarity of disease phenotypes to measure the functional similarity of diseases and miRNAs, and gave greater weight to miRNAs belonging to the same family or the same cluster class, and proposed a k-nearest neighbor prediction model (HDMP) is used to predict disease-related miRNAs. This method has obtained reliable prediction performance, but also cannot predict miRNAs associated with isolated diseases. Later, Xuan et al. (Banys-Paluchowski et al., 2015) further proposed the MIDP method based on random walk. In this model, by assigning different weights to known and unknown nodes, the prior information of the topology is effectively integrated. In addition, the extended conversion on the double-layer network of miRNA diseases makes it possible to predict miRNAs associated with isolated diseases. You et al. (2017) proposed a path-based miRNA–disease association (PBMDA) prediction model by integrating known human miRNA–disease associations, miRNA functional similarities, disease semantic similarities, and Gaussian interaction profiles for miRNA and disease similarities. The model constructs a heterogeneous graph composed of three interrelated subgraphs, and further uses a depth-first search algorithm to infer potential miRNA–disease associations.

The results show that reliable performance is obtained. Gu et al. (2016) created a network consistency projection algorithm to identify potential associations (NCPMDA) by integrating similarity networks and association networks. The biggest advantage of these methods is that they can predict isolated miRNAs associated with disease, but the performance obtained is not very satisfactory.

Although research on miRNA disease association prediction models has made some progress, there is still room to further improve the prediction performance of the model. In this study, we propose a predictive model called matrix decomposition, which fully considers the similarity between miRNAs and the similarity between diseases. In order to evaluate the effectiveness of MFMDA, we tested it using a global fivefold and local LOOCV framework. MFMDA is superior to the benchmark algorithm used for comparison, and achieves reliable performance in the framework of fivefold CV and local LOOCV (AUC 0.9061 and 0.7933) in the HMDD (V2.0) data set. To further prove the superiority of MFMDA, we analyzed three common diseases. Based on the analysis of the test results, we can find that 18 of the top 30 potential miRNAs related to the three diseases predicted by MFMDA have been confirmed by other databases.

MATERIALS AND METHODS

Human Disease–miRNA Interactome Network

In the past few decades, as the technology has matured, a large number of omics data have been published, including a large number of pairs related to miRNA diseases. Here, we use the known miRNAs and disease-associated data set HMDD V2.0 as the benchmark dataset (Huang et al., 2019a). The data set contains 495 miRNAs and 383 diseases and 5430 experimentally verified human-disease-related pairs. We use the adjacency matrix A to represent this confirmed association. Specifically, if the disease $d\,(i)$ was previously associated with miRNA $m\,(j)$, the value of A_{ij} is 1; otherwise, the corresponding position is set to 0.

miRNA Functions Similarly

Based on previous research, it is not difficult to find that miRNAs with similar functions are more likely to be related to similar diseases (Wang et al., 2010). Under this assumption, the miRNA functional similarity score was calculated[1]. Therefore, we constructed a functional similarity matrix FS between miRNAs based on these data, where $FS(m\,(i), m\,(j))$ represents the similarity between miRNA $m\,(i)$ and another miRNA $m\,(j)$.

Disease Semantic Similarity

Semantic similarity is a common way to express the similarity of diseases in this field. MFMDA uses a layered directed acyclic graph (DAG) to calculate the similarity between two diseases (Wang et al., 2010). Specifically, for disease d, let $DAG_d = (d, T_d, E_d)$ be a DAG, where T_d represents the ancestor node set of d (including itself) and E_d represents the hierarchical connection between diseases defined by the MeSH disease tree structure of the National Library of Medicine. For any $t \in T_d$, MFMDA defines the semantic contribution of disease t to d as:

$$D_d\,(t) = \begin{cases} 1 & if\ t = d \\ max\left\{\Delta \times D_d(t^{'})|t^{'} \in children\ of\ t\right\} & if\ t \neq d \end{cases} \quad (1)$$

Where Δ is the semantic decay factor, which is set to 0.5 in the iterative equation according to previous researches (Dong et al., 2019; Marcuello et al., 2019). Therefore, the semantic similarity between the diseases d_1 and d_2 can be defined as:

$$D\left(d_i, d_j\right) = \frac{\sum_{t \in T_{d_i} \cap T_{d_j}} \left(D_{d_i}\,(t) + D_{d_j}\,(t)\right)}{\sum_{t \in T_{d_i}} D_{d_j}(t) + \sum_{t \in T_{d_j}} D_{d_j}(t)} \quad (2)$$

Gaussian Similarity of miRNA and Disease

Among various similarity measurement algorithms, Gaussian similarity is a very good measurement method, which has been widely used in various fields. Let $VP(m_i)$ be the vector related to miRNA $m_i$in Y, i.e., the i^{th} column of Y. Then, the Gaussian similarity between the diseases m_i and $m_j$is calculated as follows:

$$KM\left(r_i, r_j\right) = exp(-\gamma_m||VP\,(r_i) - VP(r_j)||^2) \quad (3)$$

Where γ_m is the adjustment parameter of the bandwidth (van Laarhoven et al., 2011). The update rule of parameter γ_m is as follows:

$$\gamma_m = \gamma_m^{'} / \left(\frac{1}{nm}\sum_{i=1}^{nm} ||VP\,(r_i)\,||^2\right) \quad (4)$$

Similarly, the Gaussian similarity between miRNAs can be defined as follows:

$$KD\left(d_i, d_j\right) = exp(-\gamma_d||VP\left(d_i\right) - VP(d_j)||^2) \quad (5)$$

$$\gamma_d = \gamma_d^{'} / \left(\frac{1}{nd}\sum_{i=1}^{nd} ||VP\left(d_i\right)\,||^2\right) \quad (6)$$

Integrated Similarity for Diseases and miRNAs

In order to obtain a more comprehensive disease similarity, the semantic similarity of the disease is combined with the Gaussian interactive contour kernel similarity through the following piecewise function to obtain the final similarity between the diseases:

$$S_d\left(d_i, d_j\right) = \begin{cases} D\left(d_i, d_j\right) & d_i\ and\ d_j\ has\ semantic\ similarity \\ KD\left(d_i, d_j\right) & otherwise \end{cases} \quad (7)$$

Similarly, the similarity between miRNAs can also be redefined as:

$$S_m\left(\mathrm{m}_i, m_j\right) = \begin{cases} FS\left(\mathrm{m}_i, m_j\right) & r_i\ and\ r_j\ has\ functional\ similarity \\ KM\left(\mathrm{m}_i, m_j\right) & otherwise \end{cases} \quad (8)$$

[1] http://www.cuilab.cn/files/images/cuilab/misim.zip

MFMDA

Matrix factorization (MF) is an effective technique that has been widely used in data representation (Huang and Zheng, 2006; Hosoda et al., 2009; Zheng et al., 2009; Xu et al., 2020). It aims to find two matrices whose product provides the best approximation to the original matrix. Given a miRNAs–diseases association matrix, MF can be decomposed into two matrices $Y = R^{n\times m}$, that is, $W \in R^{n\times k}$ and $H \in R^{m\times k}$,and $Y \approx UV^T$. Here, we use mathematical formulas to express the potential association prediction problem between diseases and miRNAs as the following objective function:

$$\min_{U,V} ||I\cdot(Y - WH^T)||_F^2 \tag{9}$$

where $||\cdots||_F^2$ represents the Frobenius norm and $\cdot$ denotes the Hadamard product of two matrices, that is, the multiplication of the corresponding elements of the matrix, and $I_{ij} = 0$ if the entry (i, j) in Y is missing, and 1 otherwise.

The standard MF in Eq. 2 is just to find two matrices, and their product tries to approximate the original matrix. However, the effects caused by the similarity between miRNAs and diseases are ignored. Suppose the functions of the two miRNAs are very similar, and at the same time, the diseases implicitly learned that they should have a similar distance in the vector space. The diseases dimension is the same. For the same reason, the miRNAs size can also use this idea to constrain the drug's implicit representation. That is, if the two diseases are similar, the distance of the miRNAs in the low-dimensional vector space should also be small.

$$\begin{aligned}\min_{U,V} ||I\cdot\left(Y - WH^T\right)||_F^2 + \lambda_l\left(||W||_F^2 + ||H||_F^2\right)\\ +\lambda_v\sum_{i,p=1}^{n}||w_i - w_p||^2 S_{i,p}^{m*}\\ +\lambda_d\sum_{j,k=1}^{m}||h_j - h_k||^2 S_{j,k}^{d*}\end{aligned} \tag{10}$$

where λ_l, λ_d, and λ_v are the regularization coefficients; w_i and h_j are the ith and jth rows of W and H, respectively. S^{v*} is the hidden social similarity between miRNAs and S^{d*} is the hidden social similarity between diseases.

Optimization

In order to solve the local optimal solution problem of Eq. 3, we use the gradient descent algorithm to solve. According to the nature of the Frobenius norm, the corresponding Lagrange function L_E of Eq. 2 can be redefined as:

$$\begin{aligned}L_E = Tr\left(I\cdot\left(YY^T - 2*YHW^T + WH^THW^T\right)\right) +\\ \lambda_l Tr\left(WW^T\right) + \lambda_l\, Tr\left(HH^T\right) + \lambda_m Tr\left(W^T L_m W\right) +\\ \lambda_d Tr\left(H^T L_d H\right) + Tr\left(\emptyset W^T\right) + Tr\left(\psi H^T\right)\end{aligned} \tag{11}$$

where $Tr(\cdots)$ represents the trace of a matrix; $L_m = D_m - S^{m*}$ and $L_d = D_d - S^{d*}$ are the graph Laplacian matrices for S^{m*} and S^{d*}, respectively; and D_m and D_m are the diagonal matrices whose entries are row (or column) sums of S^{m*} and S^{d*}, respectively.

The partial derivatives of the above functions with respect to W and H are:

$$\begin{aligned}\frac{\partial L_E}{\partial W} &= -2YH + 2WH^TH + 2\lambda_l W + 2\lambda_m L_m W + \emptyset\\ \frac{\partial L_E}{\partial H} &= -2Y^TW + 2HW^TW + 2\lambda_l H + 2\lambda_d L_d H + \psi\end{aligned} \tag{12}$$

According to the solution conditions of Karush–Kuhn–Tucker (KKT) (Facchinei et al., 2013), we can make $\emptyset_{ik}w_{ik} = 0$ and $\psi_{jk}h_{jk} = 0$, thus obtain the following equations for w and h:

$$\begin{aligned}-(YH)_{ik}\,w_{ik} + \left(WH^TH\right)_{ik}w_{ik} + (\lambda_l W)_{ik}\,w_{ik} +\\ \left(\lambda_m(D_m - S^{m*})W\right)_{ik}w_{ik} = 0\\ -\left(Y^TW\right)_{jk}h_{jk} + \left(HW^TW\right)_{jk}h_{jk} + (\lambda_l H)_{jk}\,h_{jk} +\\ \left(\lambda_d(D_d - S^{d*})H\right)_{jk}h_{jk} = 0.\end{aligned} \tag{13}$$

Therefore, we get the w_{ik} and h_{jk} update rules as follows:

$$\begin{aligned}w_{ik} &= w_{ik}\frac{(YH + \lambda_m S^{m*}W)_{ik}}{(WH^TH + \lambda_l W + \lambda_m D_m W)_{ik}}\\ h_{jk} &= h_{jk}\frac{(Y^TW + \lambda_d S^{d*}H)_{jk}}{(HW^TW + \lambda_l H + \lambda_d D_d H)_{jk}}\end{aligned} \tag{14}$$

The matrices W and H are updated based on Eq. 3 until convergence. Finally, we can obtain the predicted miRNAs–diseases association matrix as $Y^* = WH^T$, and determine the priority of potential miRNAs and disease according to the value in the matrix Y^*. In principle, the miRNAs with the highest grade in Y^* are more likely to be associated with the disease. The flow chart of MFMDA is shown in **Figure 1**.

RESULTS

Evaluation of Prediction Performance

There are many performance indicators for evaluating prediction models. In this field, ROC curve and AUC value, PR curve, and AUPR value are usually used to evaluate the performance of the algorithm (Chen and Huang, 2017; Chen et al., 2020).

The ROC curve, also called receiver operating characteristic curve or susceptibility curve, is a comprehensive indicator reflecting sensitivity and specificity. The ROC curve graphically reveals the correlation between sensitivity and specificity. By setting different thresholds, a series of corresponding sensitivities and specificities are calculated, and then plotted with the true positive rate on the ordinate and false positive rate on the abscissa curve. The simple assumption is that for binary classification problems (only two types, positive and negative samples), the

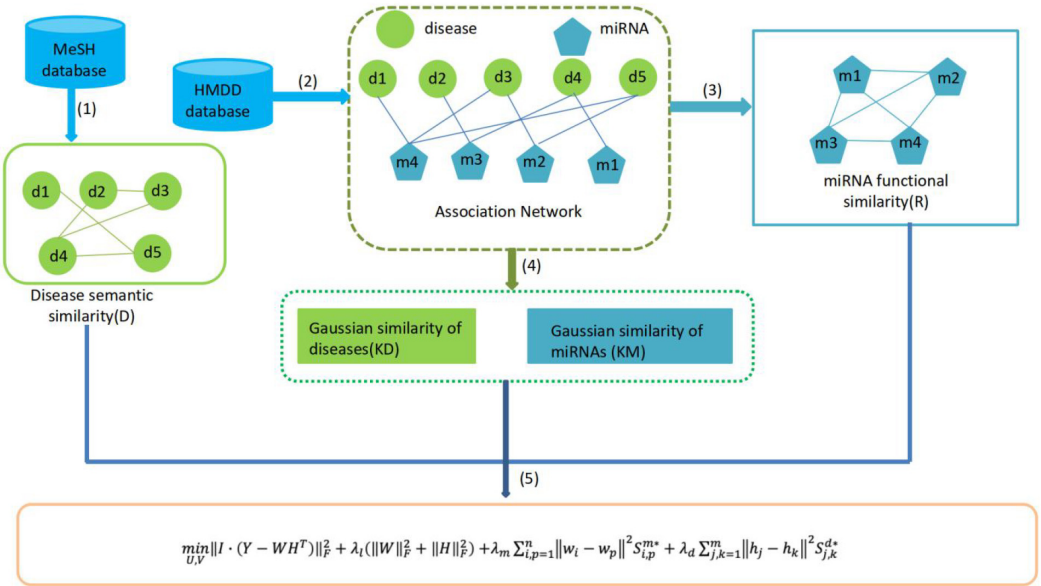

FIGURE 1 | Diagram of MFMDA for predicting potential miRNA–disease associations.

calculation methods of TPR and FPR are shown in Eq. 15.

$$\text{TPR} = \frac{TP}{TP + FN}, \ \text{FPR} = \frac{FP}{TN + FP} \tag{15}$$

TP refers to the number of positive samples that are correctly predicted, that is, the number of positive samples that are predicted as positive samples; FP refers to the number of positive samples that are incorrectly predicted, that is, the number of negative samples that are predicted to be positive samples; the number of negative samples correctly predicted, that is, the number of negative samples predicted as negative samples; FN refers to the number of negative samples that are incorrectly predicted, that is, the number of positive samples predicted as negative samples. The area under the line of the ROC curve is AUC. The more convex the ROC curve, the closer to the upper left corner. The larger the AUC value, the better the prediction performance. The AUC value is generally between 0.5 and 1. The AUC value of 0.5 is the effect of random prediction. The AUC value of 1 has the best performance and the perfect classifier, that is, it can correct all positive and negative classes.

The PR curve calculates a series of accuracy and recall by setting different thresholds, and then draws the curve as the precision ordinate and recall as the abscissa. The precision and recall are calculated into the formulas 16:

$$\text{precision} = \frac{TP}{TP + FP}, \ \text{FPR} = \frac{TP}{TP + FN}. \tag{16}$$

The PR curve reflects the correlation between accuracy and recall. The area under the PR curve is AUPR. The larger the AUPR value, the better the performance.

Comparison With Other Methods

We further compared the prediction performance of the MFMDA model with four benchmark prediction models (i.e., LRMCMDA, IMCMDA, NCPMDA, and RLSMDA). LRMCMDA and IMCMDA belong to the matrix completion algorithm, and have achieved good predictive performance in this field. NCPMDA is a network projection algorithm, which is one of the representatives of algorithms based on network prediction. RLSMDA is a semi-supervised learning method based on the Regularized Least Squares (RLS) framework, which represents a good opportunity to learn learning algorithms. Since the data used in this study are all from the public data set HMDD2.0, all the parameters of the comparison algorithm will also use the parameters given by the original author.

Performance on Predicting miRNA–Disease Association

We applied MFMDA, LRMCMDA, IMCMDA, NCPMDA, and RLSMDA to HMDD V2.0 miRNA–disease association data, which contains 5430 unique associations between 495 miRNAs and 383 diseases, and draws their ROC curves of the global fivefold CV in **Figure 2A**. As can be seen, the AUCs of MFMDA, LRMCMDA, IMCMDA, NCPMDA, and RLSMDA are 0.9061, 0.8883, 0.8364, 0.8637, and 0.8326, respectively, indicating that MFMDA performed best in predicting miRNA–disease associations.

However, considering the limited number of known and experimentally verified miRNA–disease associations, it is too arbitrary to use AUC to evaluate the performance of prediction methods. Therefore, we also include the exact recall (PR) curve and the AUPR in **Figure 2B** to supplement performance evaluation. As shown in **Figure 2B**, the AUPR of MMFDA,

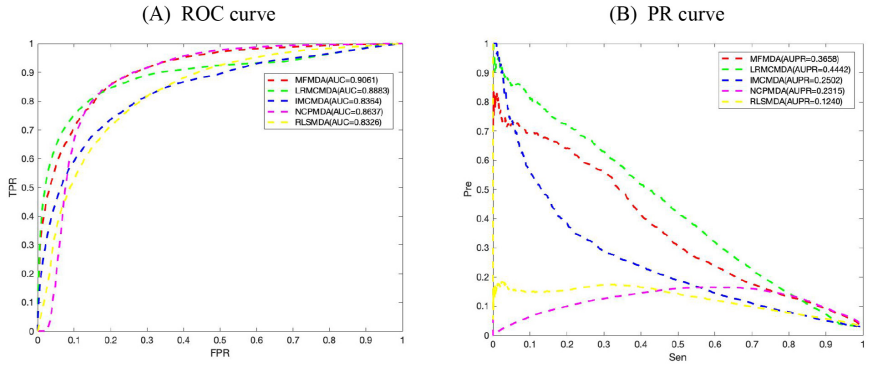

FIGURE 2 | Comparison of MFMDA with four best performers for miRNA–disease associations. **(A)** ROC curves for fivefold cross validation. **(B)** Precision–recall (PR) curve for fivefold cross validation.

LRMCMDA, IMCMDA, NCPMDA, and RLSMDA are 0.3658, 0.4442, 0.2502, 0.2315, and 0.1240, which again shows that MFMDA performs better than most algorithms in predicting miRNA–disease associations and can be a supplement to the existing computational prediction model.

Predicting Novel Disease-Related miRNAs

For a new disease, if it can find its related miRNAs, it will provide a great help for people to understand the pathogenesis of the disease. Therefore, we performed CV_d experiment to test the performance of MFMDA in predicting miRNAs associated to a novel disease d. In CV_d: CV on disease d_i, we remove all the known miRNA–disease association of the disease d_i (column vectors in matrix $Y \in R^{m \times n}$) and build prediction model (for inferring the deleted associations) using the remaining data. As shown in **Figure 3**, the AUC value obtained by MFMDA is second only to LRMCMDA, which also indicates that MFMDA is also relatively good at predicting miRNAs related to new diseases. Of course, although LRMCMDA is more effective at predicting new disease-related miRNAs, LRMCMDA uses network projection to construct negative samples. This method of constructing negative samples will be affected by the size of the data set, which will affect its prediction performance. Presumably, MFMDA is a semi-supervised algorithm, it does not need to construct negative samples and the prediction performance is relatively stable.

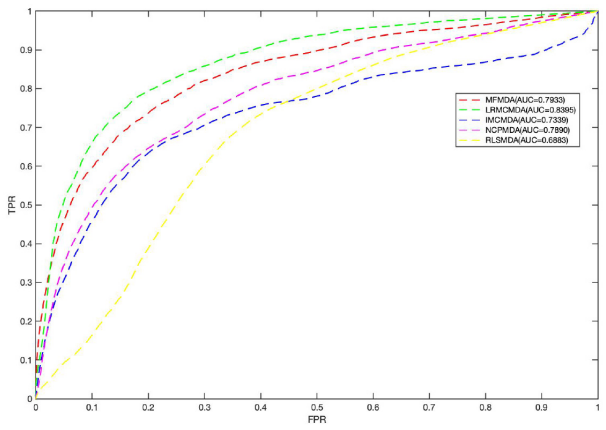

FIGURE 3 | Comparison between MFMDA and benchmark algorithms based on local LOOCV.

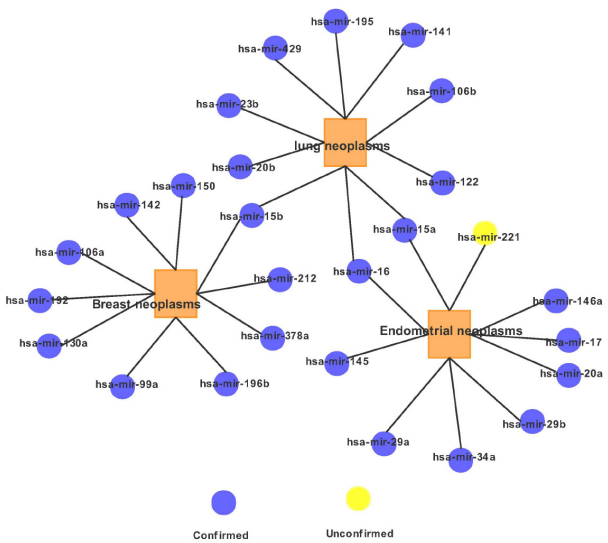

FIGURE 4 | Network of the top 10 predicted associations for the three diseases via MFMDA.

Finally, we explored the effect of the disease similarity and miRNA similarity on prediction performance. Specifically, we performed global fivefold CV with parameters $\lambda_m$or λ_d from 0.2 to 1 and a step size of 0.2 (**Table 1**). We can see that the two similarities really help predict performance. However, as the parameters continue to increase, the performance of the prediction is constantly decreasing.

TABLE 1 | Prediction AUCs of MFMDA at different choices of parameters.

MFMDA	$\lambda_m = \lambda_d = 0.2$	$\lambda_m = \lambda_d = 0.4$	$\lambda_m = \lambda_d = 0.6$	$\lambda_m = \lambda_d = 0.8$	$\lambda_m = \lambda_d = 1$
AUC	0.9061	0.9058	0.9013	0.8924	0.8912

TABLE 2 | The top 10 potential miRNA candidates detected by MFMDA for endometrial neoplasms.

Cancer	No. of confirmed miRNAs	Top 10 ranked predictions					
		Rank	miRNAs	Evidences	Rank	miRNAs	Evidences
Endometrial neoplasms	9	1	hsa-mir-146a	HMDD V3.0	6	hsa-mir-34a	HMDD V3.0
		2	hsa-mir-221	Unconfirmed	7	hsa-mir-29a	HMDD V3.0
		3	hsa-mir-20a	HMDD V3.0	8	hsa-mir-145	HMDD V3.0
		4	hsa-mir-17	HMDD V3.0	9	hsa-mir-15a	HMDD V3.0
		5	hsa-mir-16	HMDD V3.0	10	hsa-mir-29b	HMDD V3.0

Case Study

Next, three disease case studies were conducted to further validate the predictive power of the new miRNA disease pairs discovered by MFMDA. We first use the verified HMDD V2.0 pair as a training sample. For each predicted disease, the corresponding unverified miRNA is ranked according to the predicted score. Then, according to the other three well-known databases dbDEMC2.0 (Yang et al., 2017), miR2Disease (Jiang et al., 2009), and HMDD V3.0 (Huang et al., 2019b), the top 10 candidate miRNAs in the prediction list were examined.

Endometrial cancer is a group of epithelial malignant tumors that occur in the endometrium, and it occurs in perimenopausal and postmenopausal women. Endometrial cancer is one of the most common tumors of the female reproductive system. There are nearly 200,000 new cases each year, and it is the third most common gynecological malignant tumor that causes death. Earlier studies have shown that the differential expression of miRNA in endometrial adenocarcinoma can play a key auxiliary role in understanding the diagnosis and treatment of endometrial adenocarcinoma (Jurcevic et al., 2014). Therefore, in this study, we used MFMDA to identify potential miRNAs associated with endometrial adenocarcinoma. Nine of the top 10 miRNAs found were confirmed by at least one external database (see **Table 2**).

In the second case study, we still choose the tumor that belongs to women with high incidence, namely, breast tumor. Breast tumors are malignant tumors that occur in the epithelial tissue of the breast glands. Currently, the treatment is mainly based on clinical and pathological features. Targeted therapy and personalized therapy are the ultimate goals. Related studies have shown that the occurrence of breast tumors is also related to abnormalities of related miRNAs. For example, an abnormal increase in miR-22 may promote the occurrence and metastasis of breast cancer and lead to a higher degree of tumor malignancy.

TABLE 3 | The top 10 potential miRNA candidates detected by MFMDA for breast neoplasms.

Cancer	No. of confirmed miRNAs	Top 10 ranked predictions					
		Rank	miRNAs	Evidences	Rank	miRNAs	Evidences
Breast neoplasms	10	1	hsa-mir-150	dbDEMC 2.0	6	hsa-mir-130a	dbDEMC 2.0
		2	hsa-mir-142	dbDEMC 2.0	7	hsa-mir-99a	dbDEMC 2.0
		3	hsa-mir-15b	dbDEMC 2.0	8	hsa-mir-196b	dbDEMC 2.0
		4	hsa-mir-106a	dbDEMC 2.0	9	hsa-mir-378a	dbDEMC 2.0
		5	hsa-mir-192	dbDEMC 2.0	10	hsa-mir-212	dbDEMC 2.0

TABLE 4 | The top 10 potential miRNA candidates detected by MFMDA for lung neoplasms.

Cancer	No. of confirmed miRNAs	Top 10 ranked predictions					
		Rank	miRNAs	Evidences	Rank	miRNAs	Evidences
Lung neoplasms	9	1	hsa-mir-16	miR2Disease	6	hsa-mir-141	miR2Disease
		2	hsa-mir-122	dbDEMC 2.0	7	hsa-mir-195	miR2Disease
		3	hsa-mir-15a	dbDEMC 2.0	8	hsa-mir-429	miR2Disease
		4	hsa-mir-15b	Unconfirmed	9	hsa-mir-23b	dbDEMC 2.0
		5	hsa-mir-106b	dbDEMC 2.0	10	hsa-mir-20b	dbDEMC 2.0

Therefore, predicting miRNAs related to breast tumors through related algorithms will also provide corresponding help for human breast cancer treatment. As shown in **Table 3**, we found that the top 10 miRNAs predicted by MFMDA related to breast cancer have all been confirmed by relevant databases.

Finally, we conduct prediction studies on miRNAs associated with lung tumors. Lung cancer is one of the fastest growing morbidity and mortality rates, and the most threatening to the health and life of the population. In the past 50 years, many countries have reported that the incidence and mortality of lung cancer have increased significantly. The incidence and mortality of lung cancer in men accounted for the first place in all malignant tumors, the incidence in women accounted for the second place, and the mortality rate took the second place. Despite the important therapeutic value of chemotherapy, surgery is still the only way to treat lung cancer. There is an urgent need to find potential biomarkers that respond strongly to clinical observations. The researchers found that the expression level of miR-99a is related to the clinicopathological factors of lung cancer and lymph node metastasis. Identifying more miRNAs related to lung cancer helps to accurately assess clinical outcomes. Therefore, we conducted a lung cancer case study based on MFMDA. In the prediction list, nine of the top 10 predicted miRNAs confirmed their association with lung tumors (see **Table 4**).

For a clear view, we illustrate in **Figure 4** the association network of the top 10 predicted miRNA candidates for the three diseases. It is worth noting that some top candidates were found to be related to several diseases. For example: hsa-mir-15a has not only been shown to be related to the occurrence of endometrial neoplasms, but also has a certain relationship with lung neoplasms.

DISCUSSION

A large number of studies have shown that miRNA plays an increasingly important role in many physiological processes. Researchers are trying to identify disease-related miRNAs as valuable biomarkers that can be used for clinical measurement, diagnosis, prognosis, and treatment. Therefore, accurately inferring potential miRNAs related to diseases can help us study the pathogenesis of diseases and find more effective treatments. In this study, we proposed a mathematical model based on MF (MFMDA) to identify potential miRNA–disease associations. First, MFMDA not only uses known miRNA and disease-related data, but also integrates the similarities between miRNA and disease. Second, the model is a semi-supervised model, which does not rely on negative samples. Finally, in the process of solving the model, we use the alternating gradient descent algorithm to find the optimal solution to ensure a stable decomposition matrix. Experimental results show that, compared with other methods, MFDMA can effectively improve performance and is a powerful tool for discovering the association of potential diseases with miRNA. However, this method still has some limitations; we need to further optimize. For example, the similarity measure between diseases and miRNAs used by MFMDA is too single and may not be the best choice. How to integrate multiple omics information more effectively to improve prediction performance is also worthy of further research.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

SH and RC designed the study. PS collected and wrote the manuscript. SY and YC reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Heilongjiang Postdoctoral Fund (No. LBH-Z18190), Wutong Tree Foundation of The Fourth Affiliated Hospital of Harbin Medical University (No. HYDSYWTS201904), and Heilongjiang Youth Science Foundation (No. QC2018100).

REFERENCES

Banys-Paluchowski, M., Schneck, H., Blassl, C., Schultz, S., Meier-Stiegen, F., Niederacher, D., et al. (2015). Prognostic relevance of circulating tumor cells in molecular subtypes of breast cancer. *Geburtshilfe Frauenheilkd.* 75, 232–237. doi: 10.1055/s-0035-1545788

Chen, X., and Huang, L. (2017). LRSSLMDA: laplacian regularized sparse subspace learning for MiRNA-disease association prediction[J]. *PLoS Comput. Biol.* 13:e1005912. doi: 10.1371/journal.pcbi.1005912

Chen, X., Liu, M. X., and Yan, G. Y. (2012). RWRMDA: predicting novel human microRNA-disease associations. *Mol. Biosyst.* 8, 2792–2798. doi: 10.1039/c2mb25180a

Chen, X., Sun, L. G., and Zhao, Y. (2020). Ncmcmda: mirna–disease association prediction through neighborhood constraint matrix completion[J]. *Brief. Bioinform.* [Epub ahead of print].

Chen, X., Wang, L., Qu, J., Guan, N. N., and Li, J. Q. (2018). Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265.

Chen, X., Yan, C. C., Luo, C., Ji, W., Zhang, Y., Dai, Q., et al. (2015). Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* 5:11338.

Chen, X., and Yan, G. Y. (2014). Semi-supervised learning for potential human microRNA-disease associations inference. *Sci. Rep.* 4:5501.

Cui, Q., Yu, Z., Purisima, E. O., and Wang, E. (2006). Principles of microRNA regulation of a human cellular signaling network. *Mol. Syst. Biol.* 2:46. doi: 10.1038/msb4100089

Dong, J., Zhu, D., Tang, X., Qiu, X., Lu, D., Li, B., et al. (2019). Detection of circulating tumor cell molecular subtype in pulmonary vein predicting prognosis of stage i-iii non-small cell lung cancer patients. *Front. Oncol.* 9:1139. doi: 10.3389/fonc.2019.01139

Facchinei, F., Kanzow, C., and Sagratella, S. (2013). Solving quasi-variational inequalities via their KKT conditions. *Math. Programm.* 144, 369–412. doi: 10.1007/s10107-013-0637-0

Goh, J. N., Loo, S. Y., Datta, A., Siveen, K. S., Yap, W. N., Cai, W., et al. (2016). microRNAs in breast cancer: regulatory roles governing the hallmarks of cancer. *Biol. Rev. Camb. Philos. Soc.* 91, 409–428. doi: 10.1111/brv.12176

Gu, C., Liao, B., Li, X., and Li, K. (2016). Network consistency projection for human miRNA-disease associations inference. *Sci. Rep.* 6:36054.

Hammond, S. M. (2015). An overview of microRNAs. *Adv. Drug Deliv. Rev.* 87, 3–14. doi: 10.1007/978-3-319-03725-7_1

Hosoda, K., Watanabe, M., Wersing, H., Körner, E., Tsujino, H., Tamura, H., et al. (2009). A model for learning topographically organized parts-based representations of objects in visual cortex: topographic nonnegative matrix factorization. *Neural Comput.* 21, 2605–2633. doi: 10.1162/neco.2009.03-08-722

Huang, D., and Zheng, C. (2006). Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* 22, 1855–1862. doi: 10.1093/bioinformatics/btl190

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019a). HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* 47, D1013–D1017.

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019b). HMDD v3. 0: a database for experimentally supported human microRNA–disease associations[J]. *Nucleic Acids Res.* 47, D1013–D1017.

Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., et al. (2010). Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol.* 4:S2. doi: 10.1186/1752-0509-4-S1-S2

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* 37, D98–D104.

Jiang, W., Chen, X., Liao, M., Li, W., Lian, B., Wang, L., et al. (2012). Identification of links between small molecules and miRNAs in human cancers based on transcriptional responses. *Sci. Rep.* 2:282.

Jurcevic, S., Olsson, B., and Klinga-Levan, K. (2014). MicroRNA expression in human endometrial adenocarcinoma. *Cancer Cell Int.* 14:88.

Kang, B. J., Ra, S. W., Lee, K., Lim, S., Son, S. H., Ahn, J. J., et al. (2020). Circulating tumor cell number is associated with primary tumor volume in patients with lung adenocarcinoma. *Tuberc. Respir. Dis.* 83, 61–70. doi: 10.4046/trd.2019.0048

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488. doi: 10.1126/science.1138341

Li, J., Zhang, Y., Wang, Y., Zhang, C., Wang, Q., Shi, X., et al. (2014). Functional combination strategy for prioritization of human miRNA target. *Gene* 533, 132–141. doi: 10.1016/j.gene.2013.09.106

Li, J. Q., Rong, Z. H., Chen, X., Yan, G. Y., and You, Z. H. (2017). MCMDA Matrix completion for MiRNA-disease association. *Oncotarget* 8, 21187–21199. doi: 10.18632/oncotarget.15061

Lindsay, C. R., Faugeroux, V., Michiels, S., Pailler, E., Facchinetti, F., Ou, D., et al. (2017). A prospective examination of circulating tumor cell profiles in non-small-cell lung cancer molecular subgroups. *Ann. Oncol.* 28, 1523–1531. doi: 10.1093/annonc/mdx156

Luo, J., Xiao, Q., Liang, C., and Ding, P. (2017). Predicting MicroRNA-disease associations using kronecker regularized least squares based on heterogeneous omics data. *IEEE Access.* 5, 2503–2513. doi: 10.1109/access.2017.2672600

Maly, V., Maly, O., Kolostova, K., and Bobek, V. (2019). Circulating tumor cells in diagnosis and treatment of lung cancer. *In Vivo* 33, 1027–1037. doi: 10.21873/invivo.11571

Marcuello, M., Vymetalkova, V., Neves, R. P. L., Duran-Sanchon, S., Vedeld, H. M., Tham, E., et al. (2019). Circulating biomarkers for early detection and clinical management of colorectal cancer. *Mol. Aspects Med.* 69, 107–122.

Taft, R. J., Pheasant, M., and Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* 29, 288–299. doi: 10.1002/bies.20544

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241

Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell.* 43, 904–914. doi: 10.1016/j.molcel.2011.08.018

Xu, J., Cai, L., Liao, B., Zhu, W., Wang, P., Meng, Y., et al. (2019). Identifying potential miRNAs-disease associations with probability matrix factorization. *Front. Genet.* 10:1234. doi: 10.3389/fgene.2019.01234

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). CMF-Impute: an accurate imputation tool for single-cell RNA-seq data. *Bioinformatics* 36, 3139–3147. doi: 10.1093/bioinformatics/btaa109

Xu, P., Guo, M., and Hay, B. A. (2004). MicroRNAs and the regulation of cell death. *Trends Genet.* 20, 617–624. doi: 10.1016/j.tig.2004.09.010

Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., et al. (2013). Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS One* 8:e70204. doi: 10.1371/journal.pone.0070204

Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H., et al. (2017). dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. *Nucleic Acids Res.* 45, D812–D818.

Yi, Y., Zhao, Y., Li, C., Zhang, L., Huang, H., Li, Y., et al. (2017). RAID v2.0: an updated resource of RNA-associated interactions across organisms. *Nucleic Acids Res.* 45, D115–D118.

You, Z. H., Huang, Z. A., Zhu, Z., Yan, G. Y., Li, Z. W., Wen, Z., et al. (2017). PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005455. doi: 10.1371/journal.pcbi.1005455

Zheng, C., Huang, D. S., Zhang, L., and Kong, X. Z. (2009). Tumor clustering using nonnegative matrix factorization with gene selection. *IEEE Trans. Inform. Technol. Biomed.* 13, 599–607. doi: 10.1109/titb.2009.2018115

Zou, Q., Li, J., Hong, Q., Lin, Z., Wu, Y., Shi, H., et al. (2015). Prediction of MicroRNA-disease associations based on social network analysis methods. *Biomed. Res. Int.* 2015:810514.

Zou, Q., Li, J., Song, L., Zeng, X., and Wang, G. (2016). Similarity computation strategies in the microRNA-disease network: a survey. *Brief. Funct. Genomics* 15, 55–64.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.