Identifying Potential miRNAs–Disease Associations With Probability Matrix Factorization

Junlin Xu1, Lijun Cai1,* Bo Liao3*, Wen Zhu1, Peng Wang1, Yajie Meng1, Jidong Lang2, Geng Tian2 and Jialiang Yang3*

1 College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, 2 Department of Science, Genes Beijing Co., Ltd., Beijing, China, 3 School of Mathematics and Statistics, Hainan Normal University, Haikou, China

In recent years, miRNAs have been verified to play an irreplaceable role in biological processes associated with human disease. Discovering potential disease-related miRNAs helps explain the underlying pathogenesis of the disease at the molecular level. Given the high cost and labor intensity of biological experiments, computational predictions will be an indispensable alternative. Therefore, we design a new model called probability matrix factorization (PMFMDA). Specifically, we first integrate miRNA and disease similarity. Next, the known association matrix and integrated similarity matrix are utilized to construct a probability matrix factorization algorithm to identify potentially relevant miRNAs for disease. We find that PMFMDA achieves reliable performance in the frameworks of global leave-one-out cross validation (LOOCV) and 5-fold cross validation (AUCs are 0.9237 and 0.9187, respectively) in the HMDD (V2.0) dataset, significantly outperforming a few state-of-the-art methods including CMFMDA, IMCMDA, NCPMDA, RLSMDA, and RWMDA. In addition, case studies show that PMFMDA has good predictive performance for new associations, and the evidence can be identified by literature mining.

Keywords: diseases, miRNAs, probabilistic matrix factorization, association prediction, receiver operating characteristic curve (ROC)

INTRODUCTION

MicroRNAs are short non-coding RNAs. It plays a vital role in the regulation of many important biological processes (Bandyopadhyay et al., 2010; Hammond, 2015; Zhang et al., 2017). It has shown that human disease is associated with abnormal expression of miRNAs, whose analyses can guide the diagnosis, prognosis and treatment of certain diseases (Liang et al., 2019). However, identifying new miRNA–disease associations through bio-wet experiments not only has a high error rate, but also consumes huge financial resources (Feng et al., 2017). Therefore, in-silicon prediction of disease-associated miRNAs has become a critical step in prioritizing most confident targets for further experimental validation. Due to the growing power of sequencing technology, more and more omics data have been published (Yi et al., 2017), which provides a chance to reveal what role miRNAs play in physiology and pathology. Typical directions include miRNAs–disease interaction prediction, miRNA–miRNA regulatory module discovery, and so on (Chou et al., 2016). Undoubtedly, all these studies enrich our understanding of the functional regulation mechanisms of miRNA (Ha et al., 2019).

In recent years, in order to understand the pathogenesis of diseases, more and more computational models have been proposed by researchers to infer disease-related miRNAs, among which machine
learning-based and network-based methods are most popular (Luo et al., 2017a). Network-based methods are based on a common assumption that miRNAs associated with diseases using similar phenotypes are similar in function, and vice versa. For example, Jiang et al. (2010) proposed the priority of disease-associated miRNAs through human peptide–microRNAome networks to identify potential associations. However, this method relies too much on known associations to make its prediction performance less effective. Subsequently, Chen et al. (2012) implemented a random walk with restart (RWRMDA) on its network to identify potentially associated miRNAs by building a network of similarities between miRNAs. Similarly, Shi et al. (2013) conducted random walks through functional linkages between miRNA targets and disease genes to explore the relationship between human miRNA diseases. Peng et al. (2017) constructed a multiple biological network by integrating the two-way relationship among microRNA, disease and environmental factors, and realized the unbalanced random walk algorithm on this network to achieve the purpose of prediction. However, these methods cannot predict miRNAs associated with isolated diseases. Later, Chen and Zhang (2013) used a network of consistent reasoning methods to infer unknown miRNAs associated with disease. Gu et al. (2016) created a network consistent projection algorithm to identify latent associations by integrating similarity networks and associated networks. The biggest advantage of these methods is that they can predict isolated disease-associated miRNAs, but the performance achieved is not very satisfactory.

More recently, machine learning-based models have been implemented to improve classification accuracy and prediction performance (Gu et al., 2016). For example, Xu et al. (2011) designed a support vector machine (SVM) classifier that combines four topological features extracted from a miRNA target disease network to distinguish between prostate cancer-associated miRNAs and non-prostate cancer-associated miRNAs. To construct a negative sample, they randomly paired the miRNA with the disease and then removed the pair present in the positive sample set. It is clear that negative samples constructed in this way are prone to false positives. Chen and Yan (2014) introduced a normalized least square method to identify the association between potential miRNAs–diseases (RLSMDA), which does not require negative samples. In addition, Luo et al. (2017b) developed a Kronecker regularized least squares method to predict the potential association of miRNAs–disease by combining multiple omics data. Liu et al. (2019) converted the miRNAs–disease association prediction problem into a complete bipartite graph model, and proposed a prediction algorithm based on a restricted Boltzmann machine to improve prediction performance. Shen et al. (2017) introduced the cooperative matrix decomposition (CMFMDA) algorithm in the recommendation system to infer potential associations. Finally, Shen et al. (2018) introduced an induction matrix-completed algorithm to identify unknown associations. However, these methods do not perform well in predicting associations related to new diseases or miRNAs, and the prediction accuracy is not as satisfactory as associations with known diseases or miRNAs.

In order to achieve better predictive performance, we construct a new model called probability matrix factorization (PMFMDA) to predict unknown miRNAs–disease associations in this study. PMFMDA makes full use of miRNA disease association, miRNA similarity and disease similarity. To evaluate the effectiveness of PMFMDA, we test it using frameworks of global 5-fold CV and global LOOCV. In addition, a validation method called CVs is developed to estimate the performance in predicting novel diseases or miRNAs. Outperforming other state-of-the-arts methods, PMFMDA achieve reliable performance in the frameworks of global LOOCV and 5-fold CV (AUCs of 0.9237 and 0.9187, respectively) in the HMDD (V2.0) dataset (Li et al., 2014). To further demonstrate the superiority of PMFMDA, we conduct an analysis of three common diseases. According to the analysis of the test results, we can find that there are 20, 19 and 17 of 20 candidate miRNAs that are confirmed to be associated with esophageal neoplasms, breast neoplasms and lung neoplasms by dbDEMC and miRCancer, respectively.

### MATERIALS AND METHODS

The general workflow of PMFMDA is shown in Figure 1. We first use matrix Y to represent 5,430 experimentally validated associations after preprocessing the HMDD V2.0 database (Li et al., 2014). Specifically, Y is a 495 × 383 matrix with row denoting miRNAs and column denoting diseases; Yij = 1 if the ith miRNA is associated with the jth disease and 0 otherwise. We then calculate the disease similarity Sd and miRNA similarity Sm. Finally, a probability matrix factorization (PMF) model is proposed by integrating Y, Sd and Sm, the solution of which will recover unknown miRNAs–disease associations based on known ones.

#### Disease Semantic Similarity

The hierarchical directed acyclic graphs (DAGs), usually are obtained from the MeSH database, and are widely used to calculate the similarity between diseases (Gu et al., 2016). Specifically, for a disease d, let DAGd = (D, Td, Ed) represents its directed acyclic graph, where Td denotes the set of the ancestors of d, and Ed represents the set of links in the MeSH tree. So, the semantic contribution of disease t to disease d is defined as:

\[
D_d(t) = \begin{cases} 
1 & \text{if } t = d \\
\max \{\Delta \times D_d(t') | t' \in \text{children of } t\} & \text{if } t \neq d
\end{cases}
\]  

(1)

Where Δ is a predefined semantic contribution factor, the value of Δ in this study is set to 0.5. Therefore, we can calculate the semantic similarity of between diseases by formula (2).

\[
D(d, d') = \frac{\sum_{i} D_{d}(t) + D_{d'}(t)}{\sum_{i} D_{d}(t) + \sum_{i} D_{d'}(t)}
\]

(2)

#### miRNAs Functional Similarity

For the similarity between miRNAs, most studies use functional similarity measurements (Wang et al., 2010). Specifically, for any two miRNAs ri and rj, let DTi = {d1i, d2i, ..., di} and DTj = {d1j, d2j, ..., dj}
\( \{d_1, d_2, \ldots, d_l\} \) be their associated disease sets, respectively. Similar to Wang et al. we first use \( S(d, DT) = \max_{i \in DT} D(d, d_i) \) to represent the similarity between a disease \( d \) and \( DT \). Then the similarity between \( r_i \) and \( r_j \) is defined as

\[
R(r_i, r_j) = \frac{\sum_{m=1}^{k} S(d_m, DT_i) + \sum_{l=1}^{l} S(d_l, DT_j)}{k+l}
\]  

(3)

**The Gaussian Interaction Profile Kernel Similarity For Diseases and miRNAs**

In the similarity measurement algorithm, Gaussian interaction profile kernel similarity is also a good measurement algorithm, which is widely used in various fields (Lu et al., 2019). Let \( VP(d_i) \) be the vector associated with the disease \( d_i \) in \( Y \), i.e. the \( i \)th column of \( Y \). Then, the Gaussian interaction kernel similarity between disease \( d_i \) and \( d_j \) is calculated as:

\[
KD(d_i, d_j) = \exp\left(-\gamma_k \| VP(d_i) - VP(d_j) \|^2 \right)
\]  

(4)

where \( \gamma_k \) is the adjustment parameter of the kernel bandwidth. The parameter \( \gamma_k \) update rule is as follows:

\[
\gamma_k = \gamma_k / (\frac{1}{nk} \sum_{i=1}^{n} \| (d_i) \|^2)
\]  

(5)

where \( \gamma_k \) is usually set to 1.

Similarly, we can conclude that the Gaussian kernel similarity of miRNAs is as follows:

\[
KM(r_i, r_j) = \exp(-\gamma_m \| VP(r_i) - VP(r_j) \|^2)
\]  

(6)

\[
\gamma_m = \gamma_m / (\frac{1}{nm} \sum_{i=1}^{n} \| VP(r_i) \|^2)
\]  

(7)

Where \( \gamma_m \) is usually set to 1.

**Integrated Similarity For Diseases and miRNAs**

The similarity between disease \( d_i \) and disease \( d_j \) is constructed by combining the two similarities of the disease as follows:

\[
S_d(d_i, d_j) = \begin{cases} 
D(d_i, d_j) & d_i \text{ and } d_j \text{ has semantic similarity} \\
KD(d_i, d_j) & \text{otherwise}
\end{cases}
\]  

(8)

Similarly, the similarity between miRNAs \( r_i \) and \( r_j \) can be redefined as:

\[
S_r(r_i, r_j) = \begin{cases} 
R(r_i, r_j) & r_i \text{ and } r_j \text{ has functional similarity} \\
KM(r_i, r_j) & \text{otherwise}
\end{cases}
\]  

(9)

**FIGURE 1** The workflow of PMFMDA is used to infer disease-associated unknown miRNAs.
PMFMDA

Probability Matrix Decomposition (PMF) is a probabilistic linear model of Gaussian observation noise and has been widely used in data representation (Salakhutdinov and Mnih, 2008). Let $Y \in \mathbb{R}^{n \times m}$ be the known miRNAs–disease association matrix, $U_i$ and $V_j$ represent the D-dimensional miRNA-specific and disease-specific latent feature vectors, respectively. The conditional distribution of the observed associations $Y \in \mathbb{R}^{n \times m}$ (likelihood term) and the prior distribution of $U \in \mathbb{R}^{n \times D}$ and $V \in \mathbb{R}^{m \times D}$ are given by:

$$
P(Y | U, V, \alpha) = \prod_{i=1}^{N} \prod_{j=1}^{M} \mathcal{N}(Y_{ij} | U_i^T V_j, \alpha^{-1} I)$$  \tag{10}

$$
P(U(\alpha_U) = \prod_{i=1}^{N} \mathcal{N}(U_i | 0, \alpha_{U_i}^{-1} I)$$  \tag{11}

$$
P(V(\alpha_V) = \prod_{j=1}^{M} \mathcal{N}(V_j | 0, \alpha_{V_j}^{-1} I)$$  \tag{12}

Where $N(x | \mu, \sigma^{-1})$ denotes the Gaussian distribution, $I_i = 0$ if the entry $(i,j)$ in $Y$ is missing, and 1 otherwise.

The optimal model is obtained by maximizing the logarithmic posterior of miRNAs and disease characteristics using fixed hyperparameters:

$$
\ln P(U, V | Y, \alpha, \alpha_U, \alpha_V) = \ln P(Y | U, V, \alpha) + \ln P(U | \alpha_U) + \ln P(V | \alpha_V) + C
$$  \tag{13}

Where $C$ is a constant. So, using a quadratic regularization term to minimize the sum of squares of the error functions instead of maximizing the posterior distribution relative to $U$ and $V$:

$$
E_{\alpha} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{M} I_i (Y_{ij} - U_i^T V_j)^2 + \frac{\lambda_U}{2} \sum_{i=1}^{N} \|U_i\|_F^2 + \frac{\lambda_V}{2} \sum_{j=1}^{M} \|V_j\|_F^2$$  \tag{14}

Where $\lambda_U = \alpha_U / \alpha$ and $\lambda_V = \alpha_V / \alpha$ are regularization parameters, $\|\cdot\|_F$ denotes the Frobenius norm.

The standard PMF in Equation (10) does not consider the effect of similarity between miRNAs and the similarity between diseases. Since $U_i$ represents the D-dimensional miRNA-specific latent feature vectors, $U_i^T U$ denotes the weighted similarity matrix of the miRNAs. Similarly, $V_j^T V$ denotes the weighted similarity matrix of the disease. Thus, we propose a new objective function by integrating miRNAs similarity and diseases similarity named PMFMDA as follows:

$$
E_{\alpha} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{M} I_i (Y_{ij} - U_i^T V_j)^2 + \frac{\lambda_U}{2} \sum_{i=1}^{N} \|U_i\|_F^2 + \frac{\lambda_V}{2} \sum_{j=1}^{M} \|V_j\|_F^2 + \frac{\lambda_{U,U}}{2} \|U_i^T U - S_u\|_F^2 + \frac{\lambda_{V,V}}{2} \|V_j^T V - S_d\|_F^2$$  \tag{15}

where $S_u$ and $S_d$ have been calculated before.

Optimization

In order to obtain the local optimal solution of Equation (15), we use the gradient descent algorithm to solve (Xiao et al., 2018). According to the nature of the Frobenius norm, the corresponding Lagrange function $L_\alpha$ of Equation (15) is defined as:

$$
L_\alpha = \frac{1}{2} \text{Tr}(I - YY^T - VV^T) + \frac{\lambda_{U,U}}{2} \text{Tr}(UU^T) + \frac{\lambda_{V,V}}{2} \text{Tr}(VV^T) - 
\frac{\lambda_{U,U}}{2} \text{Tr}(SU U^T S_u) - 
\frac{\lambda_{V,V}}{2} \text{Tr}(SV V^T S_d) - 
\frac{\lambda_{U,U}}{2} \text{Tr}(SS U^T S_u) - 
\frac{\lambda_{V,V}}{2} \text{Tr}(SS V^T S_d) - 
\frac{\lambda_{U,U}}{2} \text{Tr}(UV U^T U) - 
\frac{\lambda_{V,V}}{2} \text{Tr}(VU V^T U) - 
\frac{\lambda_{U,U}}{2} \text{Tr}(UU U^T U) - 
\frac{\lambda_{V,V}}{2} \text{Tr}(VV V^T U) + \alpha \mathcal{L}(Y, \hat{Y})$$

where $\mathcal{L}(Y, \hat{Y})$ denotes the trace of a matrix, $\mathcal{O}=[\mathbf{q}_\alpha]$ and $\mathcal{P}=[\omega_{\beta}]$ are Lagrangian multipliers.

The partial derivatives of $U$ and $V$ are as follows:

$$
\frac{\partial L_\alpha}{\partial U} = I - YY^T + VV^T + \lambda_{U,U} U + 2\lambda_{U,U} (U^T S_u U) + \mathcal{O},
$$

$$
\frac{\partial L_\alpha}{\partial V} = I - YY^T + VV^T + \lambda_{V,V} V + 2\lambda_{V,V} (V^T S_d V) + \mathcal{P}
$$  \tag{17}

Finally, the Karush-Kuhn-Tucker (KKT) conditions $\mathbf{q}_{\alpha U_{i,j}=0}$ and $\omega_{\beta V_{i,j}=0}$ according to the gradient descent method. The following equations are obtained for $U_{i,k}$ and $V_{j,k}$:

$$
U_{i,k} = (I - YY^T + VV^T)_{i,k} + (\lambda_{U,U} I)_{i,k} U_{i,k} + (2\lambda_{U,U} (U^T S_u U))_{i,k} U_{i,k} = 0,
$$

$$
V_{j,k} = (I - YY^T + VV^T)_{j,k} + (\lambda_{V,V} I)_{j,k} V_{j,k} + (2\lambda_{V,V} (V^T S_d V))_{j,k} V_{j,k} = 0
$$  \tag{18}

Therefore, the updating rules for $U$ and $V$ as follows:

$$
U_{i,k}^{new} = U_{i,k} - \left[ (I - YY^T + VV^T)_{i,k} + (\lambda_{U,U} I)_{i,k} \right] U_{i,k} + (2\lambda_{U,U} (U^T S_u U))_{i,k} U_{i,k}
$$

$$
V_{j,k}^{new} = V_{j,k} - \left[ (I - YY^T + VV^T)_{j,k} + (\lambda_{V,V} I)_{j,k} \right] V_{j,k} + (2\lambda_{V,V} (V^T S_d V))_{j,k} V_{j,k}
$$  \tag{19}

Update $U$ and $V$ according to Equation (19) and Equation (20) until the local minimum of the objective function. Finally, the predicted miRNAs–disease association matrix is $\hat{Y} = U^T V$. The $i$th column of $\hat{Y}$ indicates the association score between
disease \( d_i \) and miRNAs, and the larger the score, the more relevant it is.

**Evaluation Methods**

In order to test the performance of PMFMDA, we utilize a 5-fold CV experiment and global LOOCV on the HMDD database and compare it with a few recent methods including CMFMDA, IMCMDA, NCPMDA, RLSMDA, and RWRMDA. In the 5-fold CV experiment of a single disease \( d_i \), known miRNAs associated with \( d_i \) (column vectors in matrix \( A \in \mathbb{R}^{m \times n} \)) are randomly divided into five subsets of equal size. Associations related to all other diseases together with 4 subsets (with respect to \( d_i \)) are taken as training samples and the remaining subset is considered as testing samples. The process is performed for 5 times until all the associations associated with \( d_i \) have been predicted once. Global LOOCV was used to evaluate the model’s global prediction ability for all miRNAs–disease association simultaneously. Specifically, we removed each known association in turn as a testing sample, with all remaining associations as training samples. We then predicted the removed entry and evaluated the performance. In addition, we perform \( CV_d \) experiment to test the performance of PMFMDA in predicting miRNAs associated to a novel disease \( d_i \).

**Parameter Tuning**

We cross-validate the training set to tune the parameters of PMFMDA. Specifically, the parameters \( \lambda_U, \lambda_V, \lambda_1, \) and \( \lambda_2 \) are increased from 0.001 to 1 with a step of 0.1 and the ones with the best AUC are selected. Since the other methods have also been tested on HMDD (V2.0) in published papers, we adopt the parameters provided by the authors. Specifically, \( W=0.9 \) for RLSMDA, \( \lambda_U = \lambda_V = 1, \lambda_1 = \lambda_2 = 0.005 \) for PMFMDA, \( \lambda_1 = \lambda_2 = 1 \) for IMCMDA, \( \lambda_1 = \lambda_2 = 1 \) for CMFMDA, \( r = 0.9 \) for RWRMDA and NCPMDA is parameter free.

**RESULTS**

**PMFMDA Outperforms Other Popular Methods In Predicting Potential Associations**

We apply PMFMDA, CMFMDA, IMCMDA, NCPMDA, RLSMDA, and RWRMDA into the HMDD database. Their receiver operating characteristic (ROC) curves and associated area under the curve (AUCs) of the global 5-fold CV and LOOCV are plotted in Figure 2. As can be seen, the AUCs of PMFMDA, CMFMDA, IMCMDA, NCPMDA, RLSMDA, and RWRMDA are 0.9187, 0.8928, 0.8372, 0.8792, 0.8333, and 0.8168, respectively. Furthermore, PMFMDA also achieve the best AUC (0.9237) on global LOOCV, indicating that PMFMDA perform best in predicting miRNAs–disease associations. However, considering the limited number of known miRNAs–disease associations, it might be insufficient to evaluate the performance of the methods by AUC alone. Thus, we also plotted the precise recall (PR) curve and calculated the area under the PR curve (AUPR) based on the global 5-fold CV experiment in Figure 3. In a PR-curve, the precision refers to the ratio of correctly predicted associations to all associations with scores higher than a given threshold; by contrast, the recall refers to the ratio of correctly predicted associations to all known miRNAs–disease associations. In general, the ROC curve and the PR curve show similar trend. As shown in Figure 3, the AUPRs of PMFMDA, CMFMDA, IMCMDA, NCPMDA, RLSMDA, and RWRMDA are 0.3535, 0.3428, 0.2509, 0.1176, 0.1234, and 0.1369 respectively, indicating that PMFMDA performed best in predicting miRNAs–disease associations. At the same time, in order to further prove the effectiveness of PMFMDA. We performed 10 times of global 5-fold CV and achieved an average AUC and AUPR of 0.9187
PMFMDA: Predict Disease Associated MiRNAs

Xu et al.

+/− 0.0013, 0.3535+/− 0.0015, respectively. This proves the reliability and stability of the PMFMDA algorithm.

PMFMDA Outperforms Other Popular Methods In Predicting miRNAs Associated With Novel Diseases

Besides global miRNAs–disease predictions, it is also critical to check the performance of the above methods on specific diseases. CVd is used to measure the ability of an algorithm to predict a new disease-associated miRNA. In order to compare the fairness of the test, we conduct CV tests on 8 common diseases (Xuan et al., 2015) and use the area under the accurate recall curve (AUPR) as an indicator of predictive performance. The reason is that AUPR severely penalizes highly ranked non-interactions, which is desirable here because in practice we do not want to recommend incorrect predictions (i.e., AUPR metrics severely penalize highly ranked false positives). The results for CVd are shown in Table 1. We can clearly see that the average AUPR of PMFMDA for the eight test diseases was 0.6687, which was significantly higher than IMCMDA (0.6377), CMFMDA (0.5091), NCPMDA (0.6121), and RLSMDA (0.5761). This also sufficient PMFMDA is also the best way to predict miRNAs associated with novel diseases.

Furthermore, in order to further evaluate our approach in predicting new diseases. We implement CVd experiments on the above 8 diseases. We show the calculation of the number of disease-associated miRNAs identified at different ranking thresholds in Table 2. For example: We delete all miRNAs associated with breast tumors, and then use PMFMDA to predict its related miRNAs. we can find that 91 of the top 100 predictions are accurately predicted through the test results. This is ample indication that our approach can yield high quality predictions for isolated disease-associated miRNAs. In order to better understand the predicted eight disease-related miRNAs, we listed the names and predicted scores of the top 100 candidates related to the eight diseases in the Supplementary Table S1.

Evaluate Performance on Different Data Sources

To further test the versatility of PMFMDA. We obtain 60,576 experimental validation correlation data by preprocessing the MNDR (V2.0) dataset (Cui et al., 2018). The data contains 887 diseases and 3,954 miRNAs. We apply PMFMDA, CMFMDA, IMCMDA, NCPMDA, RLSMDA, and RWRMDA on the MNDR (V2.0) database. As shown in Table 3, the AUC of PMFMDA was 0.9885, significantly higher than those of CMFMDA (0.9799), IMCMDA (0.9171), NCPMDA (0.9480), RLSMDA (0.9358), and RWRMDA (0.9055) with increases of about 0.86, 7.14, 4.05, 5.27, and 8.3% respectively. The AUPR of PMFMDA was 0.5174, significantly higher than those of CMFMDA (0.5047), IMCMDA (0.3865), NCPMDA (0.2045), RLSMDA (0.2818), and RWRMDA (0.1907). In conclusion, PMFDA has been proven to be effective in inferring related miRNAs with diseases in terms of AUC values and AUPR values.

Parameter Sensitivity Analysis

In machine learning, parameter tuning is critical for the performance of a model. Thus, we presented in Table 4 several sets of parameter settings based on the global 5-fold CV experiment on the HMDDV 2.0 dataset. We found that a better...
TABLE 2 | PMFMDA predicts the correct numbers of different ranking thresholds for 8 common diseases.

| Cancer                      | No. of known associated miRNAs | Ranking threshold |
|-----------------------------|-------------------------------|-------------------|
|                             |                               | 20    | 40    | 60    | 80    | 100   |
| Breast neoplasms            | 202                           | 20    | 38    | 54    | 74    | 91    |
| Colorectal neoplasms        | 147                           | 17    | 30    | 45    | 58    | 70    |
| Glioblastoma                | 96                            | 17    | 30    | 43    | 53    |       |
| Heart failure               | 120                           | 17    | 28    | 39    | 51    | 58    |
| Melanoma                    | 141                           | 19    | 35    | 51    | 63    | 77    |
| Prostatic neoplasms         | 118                           | 17    | 32    | 43    | 56    | 65    |
| Stomach neoplasms           | 173                           | 15    | 32    | 49    | 63    | 79    |
| Urinary bladder neoplasms   | 92                            | 18    | 31    | 42    | 51    | 55    |

TABLE 3 | The performance of PMFMDA and the baseline methods based on 5-fold CV on the MNDRV2.0 dataset.

| PMFMDA     | CMFMDA | IMCFMDA | NCPMDA | LSLMDA | RWRMDA |
|------------|--------|---------|--------|--------|--------|
| AUC        | 0.9885 | 0.9799  | 0.9171 | 0.9480 | 0.9358 |
| AUPR       | 0.5174 | 0.5047  | 0.3865 | 0.2045 | 0.2818 |

TABLE 4 | Parameter tuning for PMFMDA based on 5-fold CV.

| AUC        | $\lambda_u = \lambda_v = 1$ | $\lambda_u = \lambda_v = 0.1$ | $\lambda_u = \lambda_v = 0.01$ |
|------------|-----------------------------|-------------------------------|-------------------------------|
| $\lambda_1 = \lambda_2 = 1$ | 0.7905                       | 0.7728                        | 0.7588                        |
| $\lambda_1 = \lambda_2 = 0.1$ | 0.9040                       | 0.8507                        | 0.8381                        |
| $\lambda_1 = \lambda_2 = 0.01$ | 0.9185                       | 0.9032                        | 0.8692                        |

No. of miRNA similarity on prediction performance. Specifically, we confirms the effectiveness of seeking an optimal combination of parameters in improving performance.

Finally, we explore the effect of the disease similarity and miRNA similarity on prediction performance. Specifically, we perform global 5-fold CV with parameters $\lambda_1$ and $\lambda_2$ setting to zero (Figure 4) in the HMDD (V2.0) dataset. We can see that the two similarities do contribute to prediction performance. In addition, PMFMDA achieve good results even in the model without integrating disease and miRNA similarity. However, this model is not good in predicting the association of new diseases or new miRNAs.

Case Studies

Another aspect of PMFMDA’s strong predictive power is in case studies. Here, all the associations included in the HMDD (V2.0) database are used as training for the model, and the unincorporated associations are considered candidates for verification. In addition, miRCancer (Xie et al., 2013) and dbDEMC (Yang et al., 2010) were used to verify the correctness of the predictions. In this work, we mainly study three diseases including esophageal tumors, breast tumors, and lung tumors, and perform detailed analyses of the top 10 candidates predicted by PMFMDA in each disease (see Table 5).

Esophageal tumors are a disease with high morbidity and high mortality in the digestive system (Kano et al., 2010; He et al., 2012). Early diagnosis plays a crucial role in its treatment (Azmi, 2012). In this study, we use PMFMDA to identify potential miRNAs associated with esophageal tumors. The top 10 miRNAs to be all confirmed by the database were associated with esophageal tumors (see Table 5).

Breast neoplasm is the malignant tumor that is prone to occur in women, it is a systemic malignant disease, for which many related genes have been discovered (Venkatadri et al., 2016). MicroRNA (miRNA), as a kind of small RNA, can specifically bind to the 3’ untranslated region of its target mRNA, causing translational inhibition or degradation of target mRNA, and playing an oncogene in the process of cell growth and differentiation (Miller et al., 2008). Thus, MiRNAs present a new way for the study of pathogenic genes in breast neoplasms. As we can see from Table 5, 9 of the top 10 predictions have been confirmed by the relevant databases.

TABLE 5 | PMFMDA infers the top 10 miRNA candidates for the three selected diseases.

| Cancer                | Number of miRNAs identified by the literature | Top 10 |
|-----------------------|----------------------------------------------|--------|
|                       | Rank | miRNAs | Evidence | Rank | miRNAs | Evidence |
| Esophageal neoplasms  |      |        |          |      |        |          |
| 10                    | 1    | mir-17 | dbDEMC   | 6    | mir-1  | dbDEMC   |
|                       | 2    | mir-18a| dbDEMC   | 7    | mir-200b| dbDEMC   |
|                       | 3    | mir-221| dbDEMC   | 8    | mir-222| dbDEMC   |
|                       | 4    | mir-16 | dbDEMC   | 9    | mir-29a| dbDEMC   |
|                       | 5    | mir-19b| dbDEMC   | 10   | mir-133b| dbDEMC   |
| Breast neoplasms      | 9    |        |          |      |        |          |
| 10                    | 1    | mir-142| miRCancer| 6    | mir-138| dbDEMC   |
|                       | 2    | mir-150| dbDEMC, miRCancer | 7 | mir-15b| dbDEMC   |
|                       | 3    | mir-106a| dbDEMC   | 8    | mir-192| dbDEMC   |
|                       | 4    | mir-99a| dbDEMC, miRCancer | 9 | mir-378a| Unconfirmed|
|                       | 5    | mir-130a| dbDEMC   | 10   | mir-196b| dbDEMC   |
| Lung neoplasms        | 9    |        |          |      |        |          |
| 10                    | 1    | mir-16 | dbDEMC   | 6    | mir-99a| dbDEMC   |
|                       | 2    | hsa-mir-15a| dbDEMC | 7    | mir-429| dbDEMC, miRCancer |
|                       | 3    | hsa-mir-106b| dbDEMC | 8    | mir-302b| dbDEMC, miRCancer |
|                       | 4    | mir-195 | dbDEMC, miRCancer | 9 | mir-130a| dbDEMC   |
|                       | 5    | mir-141 | dbDEMC   | 10   | mir-296| Unconfirmed|
The death rate from lung neoplasms is extremely high. About 1.3 million people die of lung neoplasms every year, accounting for about one-third of all neoplasms deaths worldwide (Yu et al., 2015; Sun et al., 2016). miRNAs have been found as a tumor suppressor gene and lung neoplasms. For example, Gu et al. found that miR-99a was significantly expressed in lung cancer tissues and lung neoplasm cells. In addition, the expression level of miR-99a is correlated with clinicopathological factors, the clinical stage and lymph node metastasis of lung cancer patients. We use PMFMDA to predict potential related miRNAs in lung tumors. As shown in Table 5, we can find that only one of the top 10 related miRNAs predicted is unconfirmed.

For a clear view, we show the top 20 miRNAs associated networks predicting three tumors in Figure 5. It is worth noting that some miRNA candidates are usually associated with several diseases. For example, mir-15b and mir-130a are associated with both Prostatic lung and Breast Neoplasms. Has-mir-16 is associated with both Esophageal Neoplasms and lung Neoplasms.

**DISCUSSION**

It is known that miRNAs often play an irreplaceable role in biological processes related to human diseases (Shen et al., 2017).
Accurately inferring disease-related potential miRNAs is helpful for us to investigate the pathogenesis of the disease and find a more effective treatment. In this study, we construct a mathematical model based on probability matrix factorization (PMFMDA) to identifying potential miRNAs–disease associations. PMFMDA outperform a few state-of-the-art models in the HMDV2.0 database due to a few factors. First, PMFMDA not only uses known correlation data, but also integrates the similarities between miRNAs and between diseases. This has enabled PMFMDA to achieve good results in predicting isolated disease-associated miRNAs since theoretically similar miRNAs may associate with similar diseases. Second, the model is a semi-supervised model, which does not rely on negative samples. Thus, it is better than most machine learning algorithms with strong requirement for good negative samples. Finally, in the model solving process, we use the alternating gradient descent algorithm to find the optimal solution to ensure the reliability of disease feature vectors and miRNA feature vectors. In terms of experiment, PMFMDA achieves the highest AUC (0.9187, 0.9237, respectively) in 5-fold CV and global LOOCV, demonstrates its most reliable prediction performances. At the same time, we also perform CV_experiments to measure the ability of PMFMDA to predict miRNAs associated with novel diseases. We conduct CV testing on 8 common diseases, which have at least 80 associations are verified (Xuan et al., 2015). PMFMDA achieves the highest average AUPRs of 0.6687. Finally, to make the more comprehensive test of PMFMDA, we use the three most common diseases in humans for research. The number of other database validations in the top 20 predicted miRNAs for esophageal tumors, breast tumors, and lung tumors are found to be 20, 19, and 17, respectively. In conclusion, PMFMDA has achieved good results in predicting the potential association of miRNA disease and predicting new disease-associated miRNAs and can be used as a very useful supplement to existing prediction models.

Although quite satisfactory results have been achieved from PMFMDA, there are still some limitations to this approach. Firstly, we only use semantic similarity and the Gaussian kernel descent algorithm to find the optimal solution to ensure the reliability of disease feature vectors and miRNA feature vectors. In-depth research.

**DATA AVAILABILITY STATEMENT**

The program and data used in this study are publicly available at: https://github.com/xujunlin123/PMFMDA.git.

**AUTHOR CONTRIBUTIONS**

JY, JX, LC, GT and BL conceived the concept of the work. JX, PW, WZ, YM, and JL performed the experiments. JX and JY wrote the paper. GT helped in revising the manuscript.

**FUNDING**

This study is supported by National Nature Science Foundation of China (Grant Nos. 11171369, 61272395, 61307017, 61300128, 61472127, 61572178, 61672214, and 61772192), and the National Science Foundation of Hunan, China (Grant Nos. 2018JJ2461, 2018JJ3570).

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01234/full#supplementary-material

**REFERENCES**

Azmi, A. S. (2012). Systems biology in cancer research and drug discovery. Springer. doi: 10.1007/978-94-007-4819-4

Bandyopadhyay, S., Mitra, R., Maulik, U., and Zhang, M. Q. (2010). Development of the human cancer microRNA network. Silence 1, 6. doi: 10.1186/1755-8794-1-6

Chen, X., and Yan, G. Y. (2014). Semi-supervised learning for potential human microRNA-disease associations inference. Sci. Rep. 4, 5501. doi: 10.1038/srep05501

Chen, H., and Zhang, Z. (2013). Similarity-based methods for potential human microRNA-disease association prediction. BMC Med. Genomics 6. doi: 10.1186/1755-8794-6-12

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., 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. doi: 10.1093/bioinformatics/bty503

Chou, C. H., Chang, N. W., Shrestha, S., Hsu, S., Lin, Y. L., Lee, W. H., et al. (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res. 44, D239–D247. doi: 10.1093/nar/gkv1258

Cui, T., Zhang, L., Huang, Y., Yi, Y., Tan, P., Zhao, Y., et al. (2018). MNDR v2.0: an updated resource of ncRNA-disease associations in mammals. Nucleic Acids Res. 46, D371–D374. doi: 10.1093/nar/gkx1025

Feng, P., Zhang, J., Tang, H., Chen, W., and Lin, H. (2017). Predicting the organelle location of noncoding RNAs using pseudo nucleotide compositions. Interdiscip. Sci. Comput. Life Sci. 9, 540–544. doi: 10.1007/s12539-016-0193-4

Gu, C., Bo, L., Li, X., and Li, K. (2016). Network consistency projection for human miRNA-disease associations inference. Sci. Rep. 6, 36054. doi: 10.1038/srep36054

Ha, J., Park, C., and Park, S. (2019). PMAMCA: prediction of microRNA-disease association utilizing a matrix completion approach. BMC Syst. Biol. 13 (1), 1–13. doi: 10.1186/s12918-019-0700-4

Hammond, S. M. (2013). An overview of microRNAs. Adv. Drug Deliv. Rev. 87, 3–14. doi: 10.1016/j.addr.2015.05.001
He, B., Yin, B., Wang, B., Xia, Z., Chen, C., and Tang, J. (2012). MicroRNAs in esophageal cancer (review). Mol. Med. Rep. 6, 459–465. doi: 10.3892/mmr.2012.975
Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., et al. (2010). Prioritization of disease microRNAs through a human phenotype-microRNAome network. BMC Syst. Biol. 4, S2–S2. doi: 10.1186/1752-0509-4-S1-S2
Kano, M., Seki, N., Kikkawa, N., Fujimura, L., Hoshino, J., Akutsu, Y., et al. (2010). MiR-145, miR-133a and miR-133b: tumor-suppressive miRNAs target FSCN1 in esophageal squamous cell carcinoma. Int. J. Cancer 127, 2804–2814. doi: 10.1002/ijc.25284
Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2014). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res. 42, 1070–1074. doi: 10.1093/nar/gkt1023
Liang, C., Yu, S., and Luo, J. (2019). Adaptive multi-view multi-label learning for identifying disease-associated candidate miRNAs. PloS Comput. Biol. 15, e1006931. doi: 10.1371/journal.pcbi.1006931
Liu, Y., Luo, J., and Ding, P. (2019). Inferring MicroRNA targets based on restricted boltzmann machines. IEEE J. Biomed. Heal. In. 23, 427–436. doi: 10.1109/JBHI.2018.2814609
Lu, X., Qian, X., Li, X., Miao, Q., and Peng, S. (2019). DMCN: a data-adaptive mutation clustering method to identify cancer-related mutation clusters. Bioinformatics 35, 389–397. doi: 10.1093/bioinformatics/bty624
Luo, J., Ding, P., Liang, C., Cao, B., and Chen, X. (2017a). Collective prediction of disease-associated miRNAs based on transduction learning. IEEE/ACM Trans. Comput. Biol. Bioinforma. 14, 1468–1475. doi: 10.1109/TCBB.2016.2599866
Luo, J., Xiao, Q., Liang, C., and Ding, P. (2017b). 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
Miller, T. E., Ghoshal, K., Ramaswamy, B., Roy, S., Datta, J., Shapiro, C. L., et al. (2008). MicroRNA-221/222 confers tamoxifen resistance in breast cancer signaling pathway. J. Biol. Chem. 283, 29897–29903. doi: 10.1074/jbc.M804612200
Peng, W., Lan, W., Yu, Z., Wang, J., and Pan, Y. (2017). A framework for integrating multiple biological networks to predict MicroRNA-disease associations. IEEE Trans. Nanobiosci. 16, 100–107. doi: 10.1109/TNB.2016.2633276
Salakhutdinov, R., and Mnih, A. (2008). Bayesian probabilistic matrix factorization using markov chain monte carlo. in, 880–887. doi: 10.1145/1390156.1390267
Shen, Z., Zhang, Y.-H., Han, K., Nandi, A. K., Honig, B., and Huang, D.-S. (2017). miRNA-disease association prediction with collaborative matrix factorization. Complexity 2017, 1–9. doi: 10.1155/2017/2498957
Shi, H., Xu, J., Zhang, G., Xu, L., Li, C., Wang, L., et al. (2013). Walking the interactome to identify human miRNA-disease associations:through the functional link between miRNA targets and disease genes. BMC Syst. Biol. 7, 101. doi: 10.1186/1752-0509-7-101
Sun, M., Hong, S., Li, W., Wang, P., You, J., Zhang, X., et al. (2016). MiR-99a regulates RAS-mediated invasion and migration of lung adenocarcinoma cells by targeting NOX4. Oncol. Rep. 35, 2755–2766. doi: 10.3892/or.2016.4672
Venkatadri, R., Muni, T., Iyer, A. K. V., Yakissich, J. S., and Azad, N. (2016). Role of apoptosis-related miRNAs in resveratrol-induced breast cancer cell death. Cell Death Discov. 7, e2104. doi: 10.1038/cdd.2016.6
Wang, W., Wang, J., Lu, M., Song, F., and Cai, 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
Xiao, Q., Luo, J., Liang, C., Cai, J., and Ding, P. (2018). A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. Bioinformatics 34, 239–248. doi: 10.1093/bioinformatics/btx545
Xie, B., Ding, Q., Han, H., and Wu, D. (2013). MiRCancer: A microRNA-cancer association database constructed by text mining on literature. Bioinformatics 29, 638–644. doi: 10.1093/bioinformatics/btt014
Xu, J., Li, C.-X., Lv, J.-Y., Li, Y.-S., Xiao, Y., Shao, T.-T., et al. (2011). Prioritizing candidate disease miRNAs by topological features in the mirna target-dysregulated network: case study of prostate cancer. Mol. Cancer Ther. 10, 1857–1866. doi: 10.1158/1535-7163.MCT-11-0055
Xuan, P., Han, K., Guo, Y., Li, J., Li, X., Zhong, Y., et al. (2015). Prediction of potential disease-associated microRNAs based on random walk. Bioinformatics 31, 1805–1815. doi: 10.1093/bioinformatics/btv039
Yang, Z., Ren, F., Liu, C., He, S., Sun, G., Gao, Q., et al. (2010). dbDEMC: a database of differentially expressed miRNAs in human cancers. BMC Genomics 11, 1–8. doi: 10.1186/1471-2164-11-S3-11
Yi, Y., Zhao, Y., Li, G., 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. doi: 10.1093/nar/gkw1052
Yu, S. H., Zhang, C. L., Dong, F. S., and Zhang, Y. M. (2015). miR-99a suppresses the metastasis of human non-small cell lung cancer cells by targeting AKT1 signaling pathway. J. Cell. Biochem. 116, 268–276. doi: 10.1002/jcb.24965
Zhang, T., Tan, P., Wang, L., Jin, N., Li, Y., Zhang, L., et al. (2017). RNAlocate : a resource for RNA subcellular localizations. Nucleic Acids Res. 45, 135–138. doi: 10.1093/nar/gkw728

Conflict of Interest: The authors JL and GT were employed by Geneis Beijing Co., Ltd.

The remaining 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.

Copyright © 2019 Xu, Cai, Liao, Zhu, Wang, Meng, Lang, Tian and Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.