LRMCMDA: Predicting miRNA-Disease Association by Integrating Low-Rank Matrix Completion With miRNA and Disease Similarity Information

JUNLIN XU, WEN ZHU, LIJUN CAI, BO LIAO, YAJIE MENG, JU XIANG, (Graduate Student Member, IEEE), DAWEI YUAN, GENG TIAN, AND JIALIANG YANG

1College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
2School of Mathematics and Statistics, Hainan Normal University, Haikou 570100, China
3School of Basic Medical Science, Changsha Medical University, Changsha 410219, China
4Genesis (Beijing) Company Ltd., Beijing 100102, China
5Academician Workstation, Changsha Medical University, Changsha 410219, China

Corresponding authors: Lijun Cai (ljcai@hnu.edu.cn), Bo Liao (dragonbw@163.com), and Jialiang Yang (yangjl@ genesis.cn)

This work was supported in part by the National Nature Science Foundation of China under Grant 11171369, Grant 61272395, Grant 61370171, Grant 61300128, Grant 61572178, Grant 61772192, and Grant 61702054, in part by the Natural Science Foundation of Hunan, China, under Grant 2018J12461, Grant 2018J3570, and Grant 2018J3568, and in part by the Training Program for Excellent Young Innovators of Changsha under Grant kq1905045.

ABSTRACT Identifying disease-related microRNAs (miRNAs) is crucial to understanding the etiology and pathogenesis of many diseases. However, existing computational methods are facing a few dilemmas such as lacking “negative samples” (i.e. confirmed unrelated miRNA-disease pairs). In this study, we proposed LRMCMDA, a low-rank matrix completion-based method to predict miRNA-disease associations. LRMCMDA firstly constructs a bipartite miRNA-disease graph from known associations and defines its R-projected miRNA graph, in which two miRNAs are connected if they are adjacent to the same disease in the bipartite graph. Similarly, we can define its D-projected disease graph. It then infers negative samples by assuming that connecting an unrelated miRNA-disease pair in the bipartite graph will change its R-projected miRNA graph and D-projected disease graph. Providing with both known miRNA-disease associations and negative samples, LRMCMDA infers associations between all miRNAs and diseases using a low-rank matrix completion model, in which miRNA similarity and disease similarity are incorporated into regularization terms. The assumption is that similar miRNAs will associate with similar diseases and vice versa. We compared LRMCMDA with a few state-of-the-art algorithms on several established miRNA-disease databases. LRMCMDA achieves an AUC of 0.8882 on the 5-fold cross-validation, significantly outperforming canonical methods when predicting miRNA-disease associations, and associating miRNAs with isolated diseases. The experimental results demonstrate that LRMCMDA effectively infers novel miRNA-disease associations. In addition, the case studies on cancers have further proven that LRMCMDA is useful in identifying potential cancer-associated miRNAs for experimental validation.

INDEX TERMS miRNA-disease association, low-rank matrix completion, receiver operating characteristic curve (ROC), area under curve (AUC), area under the precise recall curve (AUPR).

I. INTRODUCTION

As a class of RNAs critical in post-transcriptional gene regulations, microRNAs (miRNAs) are short non-coding RNAs of lengths approximately 22 nucleotides [1]–[3]. They can interact with many types of biomolecules including miRNAs, long non-coding RNAs (lncRNAs), and proteins to perform various biological functions. For example, the first discovered miRNA lin-4 plays a critical regulatory role in the nematode larvae development by regulating the expression of its target genes lin-14 and lin-28 [4].

One of the most fundamental goals in biomedical research is to understand the molecular, physiological and pathological mechanisms underlying complex human diseases [5], [6].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
With a deeper and deeper understanding of miRNA functions in recent years, studies on disease mechanisms have been extended from genes to miRNAs. Accumulating evidence has demonstrated that almost each miRNA interacts with hundreds of targets, and plays a role as ‘oncogenes’ or ‘tumor suppressor’ gene in tumorigenesis, metastasis, proliferation and differentiation of certain cancer cells. Therefore, miRNAs have become emerging cancer biomarkers, adding an auxiliary to predict and analyze several kinds of cancers [7]. For example, miR-137 is related to cell multiplication by targeting cell division cycle 42 (Cdc42), cyclin-dependent kinases 6 (Cdk6) in lung cancer cells [8]. In addition, miRNAs have been validated to be associated with many other human diseases such as Alzheimer’s disease [9]. Revealing the roles of miRNAs in human complex diseases would benefit not only the understanding of disease mechanisms, but also biomarker detections for disease diagnosis, development, prognosis and treatment.

In recent years, computational prediction of disease-related miRNAs has drawn more and more attention from the bioinformatics community due to the high cost and labor intensity for validating biological experiments [10]–[14]. With the accumulation of miRNA, disease, and miRNA-disease association data [15], it is now feasible to develop robust machine learning-based or network-based algorithms to infer novel miRNA-disease associations [16]. Machine learning-based methods generally assume that similar miRNAs will contribute to the same disease or close-related diseases. For example, Jiang et al. proposed a support vector machine (SVM)-based classification method to distinguish relevant miRNAs (positive samples) from irrelevant ones (negative samples) for specific diseases by integrating various miRNA genomic features [17]. Xu et al. proposed a prediction method by enhancing the function of the miRNA target disorder network [18]. Zeng et al. proposed two multipath methods for predicting disease-related genes based on gene-disease heterogeneous networks, and then applied them to predict miRNA-disease associations [19]. Chen et al. proposed to predict miRNA–disease association (IMCMDA) based on inductive matrix completion [20]. Xiao et al. proposed a graph regularized non-negative matrix factorization method to identify miRNA-disease associations and achieved good performances in cross-validation (CV) analyses [21]. Unfortunately, machine learning-based algorithms always face a dilemma in classifying negative samples, since usually, only positive miRNA-disease associations are reported. Alternatively, Chen et al. developed Regularized Least Squares for miRNA-Disease Association (RLSMDA) to prioritize potential miRNA-disease associations without using negative samples [22]. RLSMDA is a semi-supervised classification algorithm, however, the performance needs further improvement. Alaimo et al. developed a novel method called domain tuned-hybrid (DT-Hybrid) to identify drug-target [23], which was also used to identify miRNA-disease associations and achieved good performances. In addition, Peng et al. develop improved low-rank matrix recovery (ILRMR) for miRNA-disease association prediction [24]. Although, this method has achieved good results, it does not predict new disease-associated miRNAs.

To reduce the burden in explicitly constructing appropriate negative samples, network-based miRNA-disease association prediction methods were proposed recently. These methods usually calculate the likelihood of an association based on methods like the random walk on given networks involving miRNAs and diseases. The underlying assumption is “guilt by association”, that is, a miRNA is more likely to relate to a disease if its close neighbors are. Based on this assumption, Gu et al. proposed a network-consistent projection algorithm to predict miRNA-disease association (NCPMDA) by integrating known miRNA-disease associations and functional similarity between miRNAs [25]. Liu et al. calculated similarities between diseases and miRNAs by integrating multiple level disease and miRNA information, based on which to construct a heterogeneous miRNA-disease network and applied a random walk with restart on the network to predict novel associations [26]. Chen et al. used a generic network similarity measure to construct miRNA-miRNA functional networks and then proposed miRNA-disease association (RWRMDA) random walk with restart to predict potential miRNA-disease associations. Unfortunately, this method fails to associate any miRNAs to diseases without any known associated miRNAs, as the random walk with restart was unable to find a path to these “isolated” diseases [27]. In addition, Xuan et al. proposed an algorithm called HDMP to predict candidate miRNAs for a given disease by combining functional similarities and characteristics of miRNAs [28]. However, HDMP only considers k most similar neighbors in the candidate set and ignores the network topology of the neighbors. They also proposed a prediction method based on random walk using the characteristics of nodes and various topological ordering [29]. Similarly, Chen and Zhang proposed a Net-Consistency Based Reasoning (NetCBI) approach based on global network measurements [30]. NetCBI builds a global network of heterogeneous networks through integrating miRNA-similarity network, disease-similarity network, and known miRNA-disease association network. NetCBI is capable of predicting miRNAs associated with isolated diseases, however, the prediction performance is low. Alaimo et al. proposed a method called ncPred for the inference of novel ncRNA-disease association based tripartite network [31]. Finally, the readers are referred to Zou et al. for a summary of major prediction methods and future directions [32].

In summary, though tremendous progresses have been made in predicting miRNA-disease associations computationally, a few limitations still exist. First, we still do not have effective methods to identify negative samples, which has become a key bottleneck for machine learning-based methods. Second, most existing methods struggle when predicting miRNAs associated with isolated diseases. It is worth noting that after appropriate negative samples are inferred, the miRNA-disease association prediction problem
can be formulated as a canonical matrix completion problem [33]–[35].

In this study, we propose a low-rank matrix completion-based miRNA-disease association (LRMCMDA) prediction method. LRMCMDA first constructs a mapping network from the binary miRNA-disease relationships stored in known databases, and makes use of the invariant properties of the mapping network to construct negative samples. It then formulates the miRNA-disease association prediction as a low-rank matrix completion problem with miRNA similarity and disease similarity as regularization information, which was solved by an alternating gradient descent method. Finally, we compared LRMCMDA with a few best performing algorithms on several datasets collected from known miRNA-disease databases.

II. MATERIALS AND METHODS

We presented an overview of LRMCMDA in Fig. 1, which mainly consists of 4 steps: LRMCMDA first calculates (1) semantic similarity between diseases and (2) functional similarity between miRNAs; (3) It then constructs a miRNA-disease bipartite graph, through which to infer negative samples, i.e. unrelated miRNA-disease pairs; (4) LRMCMDA finally applies a low-rank matrix completion model integrating miRNA and disease similarities to estimate the associations between unrevealed miRNA-disease pairs and also adjust the known ones.

![Figure 1: The workflow of LRMCMDA.](image)

A. DISEASE SIMILARITY

LRMCMDA applies hierarchical directed acyclic graphs (DAGs) to calculate the similarity between two disease. Specifically, for a disease d, let \( \text{DAG}_d = (d, T_d, E_d) \) be its directed acyclic graph, where \( T_d \) represents the set of ancestral nodes of \( d \) including itself, and \( E_d \) denotes the hierarchical connections between the diseases as defined by MeSH disease tree structures from the National Library of Medicine (https://www.nlm.nih.gov/mesh/trees.html). For any \( t \in T_d \), LRMCMDA defines the semantic contribution of disease \( t \) to \( d \) as

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

where \( \Delta \in [0, 1] \) is a predefined semantic contribution factor to punish the distance from an ancestral node to \( d \) in \( \text{DAG}_d \). We set \( \Delta = 0.5 \) throughout this study as suggested by Wang et al. [36] The semantic similarity between two diseases \( d_1 \) and \( d_2 \) is then defined as

\[
D (d_1, d_2) = \frac{\sum_{t \in T_{d_1} \cap T_{d_2}} (D_{d_1} (t) + D_{d_2} (t))}{\sum_{t \in T_{d_1}} D_{d_1} (t) + \sum_{t \in T_{d_2}} D_{d_2} (t)}
\]

The assumption behind this definition is that two diseases are more similar if they share more ancestral diseases.

B. miRNA SIMILARITY

Previous studies have suggested that similar miRNAs are often associated with similar diseases [36], [37]. Thus, the functional similarity of two miRNAs could be roughly estimated by the similarity of their associated diseases. Specifically, for any two miRNAs \( r_a \) and \( r_b \), let \( \text{DT}_a = \{d_{a1}, d_{a2}, \ldots, d_{ak}\} \) and \( \text{DT}_b = \{d_{b1}, d_{b2}, \ldots, d_{bk}\} \) be their associated disease sets. Similar to Wang et al. [36], we first define the similarity between a disease \( d \) and a disease set \( \text{DT} \) as \( S (d, \text{DT}) = \max_{d \in \text{DT}} D (d, d) \). Then the similarity between \( r_a \) and \( r_b \) is defined as

\[
R (r_a, r_b) = \frac{\sum_{i=1}^k S (d_{ai}, \text{DT}_b) + \sum_{j=1}^l S (d_{bj}, \text{DT}_a)}{k + l}
\]

where \( k \) and \( l \) are the numbers of diseases in \( \text{DT}_a \) and \( \text{DT}_b \) respectively. By definition, the similarity of two miRNAs ranges from 0 to 1.

C. INFERRING NEGATIVE SAMPLES

We infer negative samples by exploiting the miRNA-disease bipartite graph. Specifically, the miRNA-disease associations were retrieved from the human miRNA disease database (HMDD V2.0), which curates 5430 unique associations between 495 miRNAs and 383 diseases [15]. Clearly, the interactions between miRNAs and diseases form a bipartite graph (see Fig. 2(A)).

![Figure 2: The miRNA-disease bipartite network and R/D- projected graphs.](image)

Let \( G = (R, D, E) \) be such graph, where \( R = \{r_1, r_2, \ldots, r_m\} \) and \( D = \{d_1, d_2, \ldots, d_n\} \) are the sets of miRNAs and diseases respectively, and \( E \) is the set of interactions between
R and D. Then, the adjacency matrix of G takes the following block off-diagonal form:

\[
A = \begin{bmatrix}
0_{m \times m} & A^T_{m \times n} \\
A_{m \times n} & 0_{n \times n}
\end{bmatrix},
\]

where \(A_{ij} = 1\) if \(r_i\) and \(d_j\) are associated and 0 otherwise, \(0_{m \times m}\) and \(0_{n \times n}\) are \(m \times m\) and \(n \times n\) all-zero matrices respectively, and \(A^T\) represents the transpose of A. Let \(N(r_i) = \{d \mid d \in D, (r_i, d) \in E\}\) be the set of neighboring diseases of \(r_i\). To construct negative samples, we first introduce two notations namely R-projected graph (see Fig. 2) and uncorrected node pair projected by R.

1) DEFINITION 1 (R-PROJECTED GRAPH).
Given a bipartite graph \(G=(R, D, E)\), its R-projected graph is defined as a unipartite graph \(G_R=(R, E_R)\), where \(E_R = \{ERR : nd : MbegChr = 0 \times 007 B, MendChr = 0 \times 007 C, nParams = 1\} \exists d_k \in D, s.t. d_j \in N (r_i) \cap N (r_j)\), that is, \(r_i\) and \(r_j\) are connected if they are associated with the same disease.

2) DEFINITION 2 (UNCORRECTED NODE PAIR PROJECTED BY R).
Let \(G = (R, D, E)\) be a bipartite graph and \(r_i \in R\) and \(d_j \in D\) be two nodes in G with \((r_i, d_j) \notin E\). By adding \((r_i, d_j)\) to E, we construct a new bipartite graph, say \(G'=(R, D, E')\) with \(E' = E \cup \{(r_i, d_j)\}\). Let \(G_R=(R, E_R)\) and \(G'_R=(R, E'_R)\) be the R-projected graphs of G and \(G'\) respectively. If \(G_R \neq G'_R\), then \((m_i, d_j)\) is an uncorrected node pair projected by R in graph \(G'\).

Similarly, we can define uncorrected node pairs projected by D. For a bipartite graph, structure invariance of projected graphs is a popular measurement to evaluate the contradiction of a new link to existing links. Thus, the intersection of uncorrected node pairs of D and R projections are usually taken as negative sample.

Compared to randomly selected negative samples or using all non-interacting miRNA-disease pairs as negative examples, negative sample inference by projected graph has a few advantages. First of all, random selection means the negative samples are undetermined, which will increase the randomness of the final prediction model. Secondly, random selection faces the choice of the proportion of negative samples. The optimal proportion of negative samples might be influenced by the dataset used. Third, using all non-interacting miRNA-disease pairs as negative examples will cause a huge imbalance between positive and negative samples, since only a small portion positive sample have been stored in popular databases. In contrast, network projection is also often used for link prediction of complex networks. Compared to single miRNA-disease associations, the information underlying the whole network structure is more robust and unprone to mistakes.

D. LOW RANK MATRIX COMPLETION ALGORITHM
Given negative samples, we revise the adjacency matrix \(A_{m \times n}\) with known interactions (positive samples) to be 1, the predicted negative samples (uncorrected node pair) to be \(10^{-30}\), and all other entries to be unknown (missing). We then perform a low rank matrix completion process to infer missing values from the specified ones by assuming A of low rank. Specifically, let \(X\) be the underlying true matrix of \(A\) and its rank with \(r \ll \min(m, n)\). Then the problem can be formulated as

\[
\min_{X, \lambda_1, \lambda_2, \lambda_3} \frac{1}{2} \| \text{W} \circ (X-A) \|_F^2 + \lambda_1 \| G(X) \|_F^2 + \lambda_2 \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij} \left\| X_{ij} - X_{ij}^T \right\|^2 + \lambda_3 \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij} \left\| X_{ij} - X_{ij}^T \right\|^2
\]

Such that, \(X = U_{m \times r} \Sigma_{r \times r} \Sigma_{n \times r} V_{n \times r}^T\). Where \(W_{ij} = 0\) if the entry \((i, j)\) in A is missing, and 1 otherwise, \(\circ\) denotes the Hadamard product of two matrices, that is, the multiplication of the corresponding elements of the matrix. \(\| \cdot \|_F\) denotes the Frobenius norm. The Frobenius norm of a matrix A is defined as the sum of the squares of the absolute values of the elements in matrix A.

\[
G(X) = \sum_{i=1}^{m} g \left( \frac{\| U_{i} \|^2}{3ar} \right) + \sum_{i=1}^{n} g \left( \frac{\| V_{i} \|^2}{3ar} \right)
\]

is a regularization term with \(g(z) = e^{(z-1)^2} - 1\) when \(z \geq 1\) and \(g(z) = 0\). otherwise, \(U_{i}^T \) and \(V_{i}^T\) denote the \(i^{th}\) row of \(U\) and \(V\) respectively and \(\alpha = \max(m, n)\). To explicitly formulate the hypothesis that similar miRNAs will have similar disease connection pattern, we add the restriction \(R_{ij} \left\| X_{ij} - X_{ij}^T \right\|^2\), where \(R\) is the semantic similarity matrix for miRNAs. As such, if miRNAs \(r_i\) and \(r_j\) are similar, i.e. \(R_{ij}\) is large, then \(X_{ij} - X_{ij}^T\) will be forced to be small, that is, \(r_i\) and \(r_j\) will have similar disease association pattern. Similarly, we add the last restriction to force similar diseases having similar miRNA association pattern. \(X_i\) represents the \(i^{th}\) row of \(X\) and \(X_i^T\) represents the \(i^{th}\) column of \(X\). \(\lambda_1, \lambda_2\) and \(\lambda_3\) are the regularization parameters balancing the contribution of regularization terms, which are trained by 5-fold cross-validation.

E. LOW RANK MATRIX COMPLETION ALGORITHM
To solve the optimization problem in Eq. (4), we propose an alternating gradient descent method (AGD) as follows:

**Initial value**: Let \(A^{(0)}\) be A by substituting all missing values with 0, and \(A^{(0)} = U \Sigma V \) be its singular value decomposition. Set \(U^{(0)} = U_0 \ast \sqrt{m}, V^{(0)} = V_0 \ast \sqrt{n}\), and \(\Sigma^{(0)} = \Sigma / \sqrt{mn}\) where \(U_0\) and \(V_0\) are consisted of the first \(r\) columns of \(U\) and \(V\) respectively.

**Update**: Repeat the following two steps until convergence or reaching a predetermined number of iterations.

- Fix \(U^{(k)}\) and \(V^{(k)}\) and calculate a matrix \(\Sigma^{(k)}\) to minimize Eq. (4).
The minimization problem could be solved by an iterating process,
\[ \Sigma_{ij}^{(k,a+1)} = \Sigma_{ij}^{(k,a)} \]
\[ \times \frac{W \circ \left((U^{(k)})^T A V^{(k)})_{ij}\right)}{W \circ \left((U^{(k)})^T U^{(k)} \Sigma_{ij}^{(k,a)} (V^{(k)})^T V^{(k)})_{ij}\right) + 2 \lambda_2 \phi_{ij} + 2 \lambda_3 \phi_{ij}}. \]

Here \( \phi = (U^{(k)})^T B (U^{(k)}) \Sigma_{ij}^{(k,a)} (V^{(k)})^T V^{k} \)
and, \( \varphi = (U^{(k)})^T T \Sigma_{ij}^{(k,a)} (V^{(k)})^T C (V^{(k)}) \),
where \( B = [b_{ij}]_{m \times m} \) and \( C = [c_{ij}]_{n \times n} \) are symmetric matrices with \( b_{ij} = \begin{cases} \sum_{k=1}^r R_{ik} & \text{if } i = j \\ -R_{ij} & \text{if } i \neq j \end{cases} \)
and \( c_{ij} = \begin{cases} \sum_{k \neq i} D_{ik} & \text{if } i = j \\ -D_{ij} & \text{if } i \neq j \end{cases} \).

The solution of \( \Sigma^{(k)} \) can be obtained iteratively by using the above updating rule. The iterative process terminates when the difference between the values of \( \Sigma^{(k)} \) at the \( a^{th} \) and at the \((a + 1)^{th}\) iterations is less than \( 10^{-7} \).

- Update \( U^{(k+1)} = U^{(k)} + \nabla U^{(k)} \) and \( V^{(k+1)} = V^{(k)} + \nabla V^{(k)} \).

The gradients of \( U \) and \( V \) are:
\[ \nabla U = W \circ \left( (U \Sigma V^T) - A \right) V \Sigma^T + U Q_U \]
\[ + \lambda_2 f \left( U, 2e(Q_{U1} - I_1)^2 (Q_{U1} - I_1) \right) \]
\[ + 2 B U \left( V^T V \right) \Sigma^T + 2 \lambda_3 (U \Sigma V^T) CV \Sigma^T \]
and
\[ \nabla V = W \circ \left( (U \Sigma V^T) - A \right)^T U \Sigma + V Q_V \]
\[ + \lambda_2 f \left( V, 2e(Q_{V1} - I_2)^2 (Q_{V1} - I_2) \right) \]
\[ + 2 \lambda_3 V^T U^T B U \Sigma + 2 \lambda_3 CV \Sigma^T (U^T U) \Sigma, \]

where
\[ Q_U = \frac{1}{m} U^T \mathbb{W} \circ \left( A - (U \Sigma V^T) \right) V \Sigma^T, \]
\[ Q_V = \frac{1}{n} V^T \mathbb{W} \circ \left( A - (U \Sigma V^T) \right)^T U \Sigma, \]
\[ I_1 = (1, 1, \ldots)^{T}_{m \times 1}, I_2 = (1, 1, \ldots)^{T}_{n \times 1}, \]
\[ Q_{U1} = \frac{\sum_{j=1}^r U_{1j}^2}{3 \alpha r} \]
\[ + \frac{\sum_{j=1}^r U_{2j}^2}{3 \alpha r} \]
\[ + \ldots \]
\[ + \frac{\sum_{j=1}^r U_{mj}^2}{3 \alpha r} \]
and
\[ Q_{V1} = \frac{\sum_{j=1}^r V_{1j}^2}{3 \alpha r} \]
\[ + \frac{\sum_{j=1}^r V_{2j}^2}{3 \alpha r} \]
\[ + \ldots \]
\[ + \frac{\sum_{j=1}^r V_{nj}^2}{3 \alpha r} \]
\[ \alpha = \max(m,n) \frac{f(A_{mxr}, B_{mx1})}{Z_{mxr}} \]
\[ Z_{ij} = \begin{cases} \frac{1}{\alpha r} A_{ij} * B_i & \text{if } B_i > 0 \\ 0 & \text{otherwise} \end{cases}. \]

The calculation of gradients are similar to Keshavan et al. [38], Cai et al. [39] and our previous works [33]–[35]. The detailed information of the gradient calculation are provided in the supplementary information (S1 Text Additional file 1). It is of note that both classical gradient-decent (CGD) and AGD work for solving the optimization problem, and will achieve similar optimization results given appropriate initial values. The main reason we use AGD is that it is faster. Jointly minimizing the objective function with respect to the three matrices is slow since we need to calculate gradients and perform line search for each gradient, which makes CGD not scalable as the dimension of the matrix increases. In contrast by fixing one or two matrices, one converts the non-convex optimization problem into a relatively easy quadratic problem, which could be solved more efficiently. In addition, real data might not be very sparse, which increases the computational burden. Actually, AGD is widely used in solving low-rank matrix completion problems.

### III. EVALUATION METHODS

The area under receiver operating characteristic (ROC) curve (ROC or AUC) is a widely used statistical measure to evaluate a prediction method [40]. We compare different methods by their AUCs on 3 different evaluation strategies including (1) a global 5-fold cross-validation (CV) experiment, (2) a 5-fold CV experiment on individual diseases and (3) prediction on isolated diseases.

Finally, the top predictions are usually selected by the biologists to further validate by wet laboratory experiments. Thus, it is also helpful to check the precision, recall and 1-specificity of the top k selected candidates. Specifically, precision calculates the proportion of true miRNA-disease association in the top k candidates. Recall calculates the proportion of true associations (within the top k candidates) among all true associations. Specificity measures the proportion of negative samples (not selected associations) correctly identified. Therefore, 1-Specificity reflects the proportion of negative samples is misidentified.

#### A. GLOBAL 5-FOLD CV

In a global 5-fold CV experiment, all known miRNA-disease associations are randomly divided into five disjoint parts with equal sizes. We combined data from four of these parts to train a prediction model, and evaluated the performance of said model with the remaining set. The process is repeated 5 times until all samples are predicted once.

#### B. 5-FOLD CV OF A SINGLE DISEASE

In the 5-fold CV experiment of a single disease d, known miRNAs associated with d (column vectors in matrix \( A \in \mathbb{R}^{m \times r} \)) are randomly divided into five subsets of the equal size. Associations related to all other diseases together with 4 subsections are taken as training samples and the left subset is consider.
TABLE 1. Comparing prediction AUCs for eight diseases among LRMCMDA, NCPMDA, IMCMDA, DT-hybrid and RLSMDA.

| Disease name       | LRMCMDA | NCPMDA | RLSMDA | IMCMDA | DT-Hybrid |
|--------------------|---------|--------|--------|--------|-----------|
| Breast Neoplasms   | 0.8262  | 0.8230 | 0.7667 | 0.7971 | 0.8155    |
| Colorectal         | 0.8304  | 0.7971 | 0.7584 | 0.7771 | 0.8201    |
| Neoplasms          | 0.8149  | 0.8140 | 0.7544 | 0.8114 | 0.9315    |
| Glioblastoma       | 0.9132  | 0.9066 | 0.8561 | 0.8492 | 0.8198    |
| Lung Neoplasms     | 0.8309  | 0.8275 | 0.7868 | 0.8251 | 0.8231    |
| Melanoma           | 0.9010  | 0.9003 | 0.8784 | 0.8602 | 0.8387    |
| Ovarian            | 0.9036  | 0.8967 | 0.8341 | 0.8573 | 0.9028    |
| Prostatic Neoplasms| 0.8436  | 0.8637 | 0.7781 | 0.8381 | 0.8524    |

C. PREDICTION ON ISOLATED DISEASES

In addition, we performed CV\textsubscript{d} experiment to test the performance of LRMCMDA in predicting miRNAs associated to a novel disease \textit{d}. In CV\textsubscript{d}; CV on disease \textit{d}, we remove all the known associations of the disease \textit{d} (column vectors in matrix $Y \in \mathbb{R}^{m \times n}$) and build prediction model (for inferring the deleted associations) using the remaining data.

D. MODEL PARAMETERS

The parameters $\lambda_1$ controlling the contribution of regularization term, $\lambda_2$ controlling the contribution of miRNA similarity, and $\lambda_3$ controlling the contribution of disease similarity of LRMCMDA were trained using the cross-validation process. Specifically, the three parameters are increased from 0.1 to 1 with a step of 0.1. Parameters that lead to the highest AUC are selected. To ensure a fair comparison, the parameters in the comparison method, i.e. RLSMDA, are also set such that the AUCs are the highest. Specifically, $\lambda_M = \lambda_D = 1$ and $W = 0.9$ for RLSMDA, $\lambda_1 = \lambda_2 = 1$ for IMCMDA, $\lambda = 0.8$, $\alpha = 0.7$ for DT-Hybrid, $\lambda_1 = 0.1, \lambda_2 = \lambda_3 = 1$ for LRMCMDA, and NCPMDA is parameter free. We also tuned the low rank $r$ from 1 to $\min(m, n)$ using global 5-fold cross-validation. Initial tests revealed $r=3$ was optimal, so it was subsequently used.

IV. RESULTS AND DISCUSSIONS

A. PERFORMANCE ON PREDICTING miRNA-disease ASSOCIATION

We applied LRMCMDA, NCPMDA, RLSMDA, IMCMDA and DT-Hybrid into the HMDD V2.0 miRNA-disease association data, which curates 5430 unique associations between 495 miRNAs and 383 diseases [15] and plotted their ROC curves of the global 5-fold CV in Fig. 3(A). As can be seen, the AUCs of LRMCMDA, NCPMDA, RLSMDA, IMCMDA and DT-Hybrid are 0.8822, 0.8637, 0.8415, 0.8364 and 0.8863 respectively, indicating that LRMCMDA performed best in predicting miRNA-disease associations. However, considering the limited number of known and experimentally verified miRNA-disease associations, using only AUC to evaluate the performance of the predictive method was too arbitrary. Therefore, we also included the precision-recall (PR) curve and the area under PR curve (AUPR) in Fig. 3(B) to complement the performance evaluation. Generally, if the ROC curve and the PR curve show similar variation at different thresholds, and if the AUPR is closer to 1, the prediction performance is better.

Besides global miRNA-disease predictions, it is also very important to check the performance of prediction methods on specific diseases. Similar to [29], we selected 8 common diseases associated with at least 80 verified associations and plotted their ROC curves of the global 5-fold CV in Fig. 3(A). As can be seen, the AUCs of LRMCMDA, NCPMDA, RLSMDA, IMCMDA and DT-Hybrid are 0.8822, 0.8637, 0.8415, 0.8364 and 0.8863 respectively, indicating that LRMCMDA performed best in predicting miRNA-disease associations.

### Table 2. The performance of LRMCMDA and other four methods based on 5-fold CV on the HMDD3.0 dataset.

| Methods         | AUC   | AUPR |
|-----------------|-------|------|
| LRMCMDA         | 0.8353 | 0.3419 |
| NCPMDA          | 0.7844 | 0.0403 |
| RLSMDA          | 0.6130 | 0.0340 |
| IMCMDA          | 0.7432 | 0.2180 |
| DT-Hybrid       | 0.8196 | 0.3178 |
TABLE 3. Prediction results of LRMCMDA and the other methods for the novel diseases.

| Disease name        | AUC   | NCPMDA | RLSMDA | IMCMDA |
|---------------------|-------|--------|--------|--------|
| Breast Neoplasms    | 0.8144| 0.7707 | 0.7182 | 0.7980 |
| Colorectal          | 0.8254| 0.7762 | 0.7264 | 0.7838 |
| Neoplasms           | 0.8211| 0.7877 | 0.7406 | 0.7482 |
| Glioblastoma        | 0.9048| 0.8243 | 0.8188 | 0.8474 |
| Lung Neoplasms      | 0.8189| 0.7770 | 0.7801 | 0.8166 |
| Melanoma            | 0.8908| 0.8615 | 0.8493 | 0.7739 |
| Ovarian             | 0.8900| 0.8276 | 0.8007 | 0.8554 |
| Prostatic Neoplasms | 0.8313| 0.8101 | 0.7866 | 0.7985 |

FIGURE 3. Comparison of LRMCMDA with four best performers for miRNA-disease associations. (A) ROC curves for 5-fold cross validation (B) precision-recall (PR) curve for 5-fold cross validation.

(Additional file 1) and S3 Fig (Additional file 1), respectively. In all the comparisons, LRMCMDA is the best among the 5 methods.

TABLE 4. AUC and AUPR predicted by LRMCMDA with different negative sample selections.

| Negative sample inference method | AUC   | AUPR   |
|----------------------------------|-------|--------|
| Network projection               | 0.8882| 0.4598 |
| all of the non-interaction pairs | 0.8789| 0.4301 |

TABLE 5. Prediction AUCs of LRMCMDA at different choices of parameters.

| Parameters | AUC   |
|------------|-------|
| $\lambda_1 = 0$ | 0.8867 |
| $\lambda_2 = 0$ | 0.8827 |
| $\lambda_3 = 0$ | 0.8843 |
| $\lambda_4 = 0$ | 0.8210 |
| $\lambda_1 = 0.1, \lambda_2 = \lambda_3 = 1$ | 0.8882 |

To further test the versatility of LRMCMDA, we downloaded 35548 experimentally verified data by downloading the HMDD3.0 dataset [41]. As the database is still being improved, we removed repeated associated data, and finally obtained 18732 valid associations, which contains 894 diseases and 1206 miRNAs. We applied LRMCMDA, NCPMDA, RLSMDA, IMCMDA and DT-Hybrid into the database. As shown in Table 2. The AUCs of LRMCMDA, NCPMDA, RLSMDA, IMCMDA and DT-Hybrid are 0.8353, 0.7844, 0.6130, 0.7432 and 0.8196, respectively, indicating that LRMCMDA performed best in predicting miRNA-disease associations. In addition, we also calculated the AUPR value of the 5 methods. The AUPR values obtained by NCPMDA and RLSMDA are less than 0.1, which indicates that these two algorithms are not suitable for relatively large datasets. The DT-Hybrid algorithm achieves an AUPR of 0.3178, slightly worse than that of LRMCMDA.
However, unlike other methods, DT-Hybrid cannot predict miRNA associated with new diseases. In addition, we performed a paired t-test between LRMCMDA and each of the previous models (NCPMDA $p = 1.6020e-65$, RLSMDA $p = 1.5418e-59$, IMCMDA $p = 1.6020e-65$, DT-Hybrid $p = 6.5315e-108$). Therefore, LRMCMDA is significantly better than previous methods on the eight diseases.

**B. PREDICTING NOVEL DISEASE-RELATED miRNAs**

To assess the performance of LRMCMDA in new diseases without any known miRNAs, we removed all known associations related to a disease and evaluated the resulting predictions. Since DT-Hybrid could not predict miRNA candidates for those isolated diseases, we only compared LRMCMDA with NCPMDA, RLSMDA and IMCMDA. We listed the prediction AUCs on the previously mentioned 8 diseases (Table 3). LRMCMDA consistently out-performs the other three algorithms in all eight diseases. The average AUCs for all evaluated algorithms are: 0.8496 (LRMCMDA), 0.7776 (NCPMDA), 0.8044 (RLSMDA) and 0.8027 (IMCMDA). In addition, LRMCMDA has the highest average precision (S4 Fig (Additional file 1)), highest average recall (S5 Fig (Additional file 1)), and the lowest average 1-specificity value (S6 Fig (Additional file 1)) at the top 20, 40, 60, 80 and 100 predictions.

**C. EFFECT OF NEGATIVE SAMPLE SELECTION ON LRMCMDA PERFORMANCE**

To test the effectiveness of constructing negative samples through network projection, we compared it with using all non-interacting miRNA disease pairs as negative samples. As shown in Table 4, we can find that the AUC and AURP of the 5-CV experiment by constructing negative samples through network projection have increased by 0.0093 and 0.0297, respectively. The implementation proves that it is necessary to fully consider the characteristics of the network structure to select negative samples.
Finally, we explored the effect of the disease similarity and miRNA similarity on prediction performance. Specifically, we performed global 5-fold CV with parameters $\lambda_1$, $\lambda_2$ or $\lambda_3$ set to zero, respectively (Table 5). We can see that the two similarities do contribute to prediction performance. The biological contribution of disease similarity and miRNA similarity seems to be similar. Integrating both only marginally improves the model’s performance. This observation is reasonable since the miRNA similarity is calculated through disease similarity. We will explore other measures to calculate miRNA similarity in the future, which may allow for us to account for increasingly complex biology.

D. CASE STUDIES

To further illustrate the method, we used HMDD V2.0 as the training data to construct the bipartite graph and prediction model and three public datasets PhenomiR [36], miRCancer [37] and dbDEMC2.0 [38] for testing the model. Specifically, we inferred the best prediction model using all known miRNA-disease associations in HMDD V2.0, and tuned model parameters through the global 5-fold cross-validation. The association values among all miRNA-disease pairs were predicted through this model and we further conducted case studies on breast, colorectal and lung tumors. Finally, we used three orthogonal public databases including dbDEMC2.0, PhenomiR and miRCancer to confirm our prediction of potential disease-related miRNAs. Table 6 lists the top 10 predicted miRNA candidates for the three selected diseases. Interestingly, 29 out of the 30 associations were confirmed to be true by the 3 databases. For a clear view, we illustrated in Fig. 4 the association networks of the top 15 predicted miRNA candidates for the three diseases.

![Network of the top 15 predicted associations for the three diseases via LRMCMDA.](image)

Figure 4, Network of the top 15 predicted associations for the three diseases via LRMCMDA.

It is worth noting that some top candidates are observed to be associated with several diseases. For example, has-mir-142 is associated with both Prostatic Neoplasms and Breast Neoplasms. Has-mir-142 or MIR142 is an RNA gene associated with many diseases and is known to have effect on the M2 macrophage and therapeutic efficacy against cancers like murine glioblastoma [42].

V. CONCLUSION

There is accumulating evidence suggesting that miRNAs play critical roles in the development of diseases, especially cancers. The identification of disease-associated miRNAs helps understand the mechanisms as well as the treatment of diseases. Machine learning and network-based models are widely used to predict miRNA-disease associations. However, there are limitations. For example, network-based models often cannot predict miRNAs related to new diseases. Based on machine learning models, the negative training samples are difficult to obtain, and the prediction accuracy remains to be improved. To solve these problems, we proposed a low-rank matrix completion-based miRNA-disease association (LRMCMDA) prediction method. Compared to other algorithms, LRMCMDA not only uses known correlation data, but also integrates the similarities between miRNAs and between diseases. This has enabled LRMCMDA to achieve good results in predicting isolated disease-associated miRNAs since theoretically similar miRNAs may associate with similar diseases, and vice versa. Secondly, LRMCMDA constructs a mapping network based on the binary miRNA-disease relationship stored in the known database, and then uses the invariant properties of the mapping network to construct a negative sample, thereby solving the problem of 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.

The performance of our method is validated by cross-validation and case studies of the collected data sets. Compared with other methods, LRMCMDA has higher prediction accuracies on global 5-fold cross-validation, 5-fold cross-validation on specific diseases and de-novo miRNA-disease association prediction. In addition, we identified a few novel disease-associated miRNAs for further experimental validation. However, there are some limitations for LRMCMDA. First, the similarity measure of LRMCMDA may not be optimal. We used disease semantic similarity and defined miRNA similarity through diseases. However, there are other similarity measures. For example, one can define similarity of diseases by their shared causal genes. In the future, the method will be further improved by incorporating other sources of disease and miRNA similarities. Second, it is worth noticing that LRMCMDA only makes use of disease and miRNA information. Since both diseases and miRNAs are related to other molecules like mRNA, long non-coding RNA and proteins. It will be interesting to incorporating this additional information to further improve the performance of the prediction.

**ABBREVIATIONS**

CV: Cross validation; HMDD: Human microRNA disease database; DAGs: directed acyclic graphs
**ADDITIONAL FILES**

**Additional file 1:** S1 Text. Derivation of the gradients for terms $\Sigma$.

**Additional file 2:** S1 Fig. Prediction precisions across the eight diseases at different top k values.

**Additional file 3:** S2 Fig. Prediction recalls across the eight diseases at different top k values.

**Additional file 4:** S3 Fig. Prediction 1-specificities across the eight diseases at different top k values.

**Additional file 5:** S4 Fig. Prediction precisions across the eight novel diseases at different top k values.

**Additional file 6:** S5 Fig. Prediction recalls across the eight novel diseases at different top k values.

**Additional file 7:** S6 Fig. Prediction 1-specificities across the eight novel diseases at different top k values.

**AVAILABILITY OF DATA AND MATERIALS**

The code and all supporting data files are available from https://github.com/JunlinXu/LRMCMDA.

**COMPETING INTERESTS**

GT and JY are employed by Geneis (Beijing) Co. Ltd.. The other authors declare that they have no competing interests.

**ACKNOWLEDGMENT**

The authors thank the members for their support and guidance.

**REFERENCES**

[1] S. M. Hammond, “An overview of microRNAs,” Adv. Drug Del. Rev., vol. 87, pp. 3–14, Jun. 2015.

[2] G. Meister and T. Tuschl, “Mechanisms of gene silencing by double-stranded RNA,” Nature, vol. 431, no. 7006, pp. 343–349, Sep. 2004.

[3] Z. Yang, L. Wu, A. Wang, W. Tang, Y. Zhao, H. Zhao, and A. E. Teschendorff, “DbDEMC 2.0: Updated database of differentially expressed miRNAs in human cancers,” Nucleic Acids Res., vol. 45, no. D1, pp. D812–D818, Jan. 2017.

[4] R. C. Lee, R. L. Feinbaum, and V. Ambros, “The C. Elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14,” Cell, vol. 75, no. 5, pp. 843–854, Dec. 1993.

[5] X. Chen, L.-G. Sun, and Y. Zhao, “NCMCMDA: MiRNA–disease association prediction through neighborhood constraint matrix completion,” Briefings Bioinf., Jan. 2020, doi: 10.1093/bib/bbz159.

[6] X. Chen, C.-C. Zhu, and J. Yin, “Ensemble of decision tree reveals potential miRNA-disease associations,” PLOS Comput. Biol., vol. 15, no. 7, 2019, Art. no. e1007209.

[7] X. Chen, D. Xie, Q. Zhao, and Z.-H. You, “MicroRNAs and complex diseases: From experimental results to computational models,” Briefings Bioinf., vol. 20, no. 2, pp. 515–539, Mar. 2019.

[8] X. Zhu, Y. Li, H. Shen, H. Li, L. Long, L. Hui, and W. Xu, “miR-137 inhibits the proliferation of lung cancer cells by targeting Cdc42 and Cdk6,” FEBS Lett., vol. 587, no. 1, pp. 73–81, Jan. 2013.

[9] A. Esquela-Kerscher and F. J. Slack, “Oncomirs—microRNAs with a role in cancer,” Nature Rev. Cancer, vol. 6, no. 4, pp. 259–269, Apr. 2006.

[10] J. Luo, C. Pan, G. Xiang, and Y. Yin, “A novel cluster-based computational method to identify miRNA regulatory modules,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 16, no. 2, pp. 681–687, Mar. 2019.

[11] Q. Xiao, J. Luo, C. Liang, G. Li, J. Cai, P. Ding, and Y. Liu, “Identifying lncRNA and miRNA co-expression modules from matched expression data in ovarian cancer,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 17, no. 2, pp. 623–634, Mar./Apr. 2020.

[12] X. Chen and H. Huang, “LRSSLMDA: Laplacian regularized sparse subspace learning for MiRNA-disease association prediction,” PLOS Comput. Biol., vol. 13, no. 12, 2017, Art. no. e1005912.
J. Xu, L. Cai, B. Liao, W. Zhu, P. Wang, Y. Meng, J. Lang, G. Tian, and J. Yang, “Identifying potential miRNAs–disease associations with probability matrix factorization,” Frontiers Genet., vol. 10, p. 1234, Dec. 2019.

D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, “Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases,” Bioinformatics, vol. 26, no. 13, pp. 1644–1650, Jul. 2010.

Y. Li and K. V. Kowdley, “MicroRNAs in common human diseases,” Genomics, Proteomics Bioinf., vol. 10, no. 5, pp. 246–253, Oct. 2012.

R. H. Keshavan, A. Montanari, and S. Oh, “Matrix completion from a few entries,” IEEE Trans. Inf. Theory, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

Z. Cai, T. Zhang, and X.-F. Wan, “A computational framework for influenza antigenic cartography,” PLoS Comput. Biol., vol. 6, no. 10, 2010, Art. no. e1000949.

P. Ding, J. Luo, C. Liang, Q. Xiao, and B. Cao, “Human disease MiRNA inference by combining target information based on heterogeneous manifolds,” J. Biomed. Informat., vol. 80, pp. 26–36, Apr. 2018.

Z. Huang, J. Shi, Y. Gao, C. Cui, S. Zhang, J. Li, Y. Zhou, and Q. Cui, “HMDD v3. 0: A database for experimentally supported human microRNA-disease associations,” Nucleic acids Res., vol. 47, no. D1, pp. D1013–D1017, 2019.

S. Xu, J. Wei, F. Wang, L.-Y. Kong, X.-Y. Ling, E. Nduom, and K. Gabrusiewicz, “Effect of miR-142-3p on the M2 macrophage and therapeutic efficacy against murine glioblastoma,” J. Nat. Cancer Inst., vol. 106, no. 8, Aug. 2014.