# LOMDA: Linear optimization for miRNA-disease association prediction

Ratha Pech[a,b], Yan-Li Lee[a,b], Dong Hao[a,b], Maryna Po[c], Tao Zhou[a,b,*]

[a] *CompleX Lab, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China.*
[b] *Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China.*
[c] *Department of Chemistry and Biochemistry, George Mason University, Virginia 22030, USA*

## Abstract

MicroRNAs (miRNAs) have been playing a crucial role in many important biological processes. Currently, the validated associations between miRNAs and diseases are insufficient comparing to all underlying associations. To identify these hidden associations based on biological experiment is expensive, laborious and time consuming. Therefore, computationally inferring the potential associations from biological data for further biological experiment has attracted increasing interests from different communities ranging from biological to computational science. In this work, we propose an effective and flexible method to predict the associations between miRNAs and diseases, namely linear optimization (LOMDA). The proposed method is capable of predicting the associations in three manners e.g., extra information such as miRNA functional similarity, gene functional similarity and known miRNA-disease associations are available; only some associations are known; and new miRNAs or diseases that do not have any known associations at all. The average AUC obtained from LOMDA over 15 diseases in a 5-fold-cross validation is 0.997, while the AUC of 5-fold cross validation on all diseases is 0.957. Moreover, the average AUC on leave-one-out cross validation is 0.866. We compare LOMDA with the state-of-the-art meth-

[*] Corresponding author
*Email address:* zhutou@ustc.edu (Tao Zhou)

ods and the results show that LOMDA outperforms the others in both cases, e.g., extra information is combined and only known associations are used.

*Keywords:* MicroRNA-disease association prediction, linear optimization, link prediction

## 1. Introduction

MicroRNAs (miRNAs) are the short non-coding RNAs about 22 nucleotides that regulate the gene expression of the target post-transcriptional level [1, 2, 3, 4]. In the last few decades, accumulative evidences show that miRNA has strong relationships with many critical life processes including early cell growth, proliferation, apoptosis, differentiation and metabolism [5, 6, 7, 8, 9]. Moreover, miRNA dysregulation has also been shown to have close relations with many human complex diseases [10, 11, 12, 13, 14, 15] including lung cancer [16], breast cancer [17, 18] and cardiovascular diseases [19] and so on. Therefore, studying the associations of miRNA and disease from biological data has become a significant problem in biomedical research which not only helps in the investigation of the pathogenesis [20], but also assists the diagnosis, treatment and preventions. Obviously, a lot of cancers can be much easier treated at the initial stage, therefore, identifying all the novel associations of miRNA and cancer can play an important role for early investigation and treatment [21, 22, 23, 24]. Biological experiments to verify the new associations one by one would require huge amount of time and labor, hence, an effective and flexible tool for selecting a small portion which are the most likely associations among a large pool of associations is needed for scientists to further experiment.

Because of the importance of the miRNA-disease associations and to help the researchers study the associations between miRNAs and diseases, different comprehensive databases of miRNA-disease associations have been constructed, e.g., human miRNA-association disease database (HMDD) [25] collecting human miRNA and disease associations which are experiment-supported, dbDEMC [26] containing different expressions of miRNAs in human cancers detected by highthroughput methods and miR2Disease [27] containing comprehensive resource of miRNA deregulations in various human diseases. These databases have facilitated the researchers and scientists in understanding the disease pathogenesis, furthermore, they are the main resources for the association identification research. Although there are rich collections of miRNA-disease association databases, these known associations are still limited comparing to all potential miRNA and disease associations. Moreover, it is believed that one miRNA can be associated with multiple diseases and vice versus. Laboratory experiments on searching for such underlying associations are very costly and time-consuming. Motivated by and based on the collected biological data, the machine learning and network-based methods are the appropriate and effective tools to predict the most likely potential associations for further laboratory experiments.

In the last few years, there is an emerge of research [28] in predicting the association between miRNA and disease by utilizing machine learning methods [29, 30, 31, 32, 33] and network-based methods [34, 35, 36, 37, 38, 39, 40]. In particular, Jiang *et al.* [41] proposed a model by integrating miRNA functional similarity network, disease phenotype similarity, known disease-miRNA association network and a discrete probability distribution named hypergeometric to predict the potential associations. Xuan *et al.* [42] combined the miRNA-disease associations with miRNA functional similarity, disease semantic similarity and disease phenotype similarity in a model called HDMP to obtain the association scores between miRNA and disease by summing up the sub-scores for the miRNA's $k$ neighbors. Then sub-score of a neighbor was computed by multiplying the weight of the neighbor with the functional similarity between the neighbor and the miRNA. Chen *et al.* [35] developed a method named RWRMDA to predict the novel human miNRA-disease associations by using random walk with restart based on network global similarity. In their proposed method, two data are utilized including miRNA-disease network and miRNA functional network similarity matrix. RWRMDA cannot predict the associations between miRNAs and diseases that do not have any known associations. Chen *et al.* [36] proposed three inference methods namely miRNA-based similarity inference, phenotype-based similarity inference and network-consistency-based inference. The three methods also utilize the global network similarity to predict new miRNA-disease associations. Chen *et al.* [43] developed a method by using semi-supervised learning namely the regularized least squares for miRNA-disease association (RLSMDA) by integrating miRNA-disease associations, disease-disease similarity and miRNA-miRNA associations. Chen *et al.* [33] proposed a method based on restricted Boltzmann Machine named RBMMMDA where a two-layered undirected miRNA-disease graph was built. RBMMMDA is unable to handle new diseases or miRNAs that do not have known associations. Pasquier *et al.* [29] proposed a vector space model to predict the miRNA-disease associations by utilizing miRNA-target associations, miRNA-word associations and miRNA-family associations to form a large matrix. The large matrix is later on decomposed by using singular value decomposition (SVD) as the dimensionality reduction. Finally, the decomposed matrices are utilized to reconstruct a new matrix which is embedded in a much lower dimensional space. Gu *et al.* [44] developed a method named network consistency projection for miRNA-disease association by integrating the miRNA family similarity network, the disease semantic similarity network, the known miRNA-disease associations and the miRNA family information to predict the potential association. [45] proposed a model namely extreme gradient boosting machine (EGBMMDA) by taking the feature vector obtaining from feature extraction on the miRNA functional similarity, disease semantic similarity and known miRNA-disease associations as input to the model. Xuan *et al.* [46] proposed a method namely MIDP based on random walk on the miRNA and disease networks. Chen *et al.* [47] developed a method based on a recommendation system on the network integrating of known miRNA-disease associations, disease semantic similarity, miRNA functional similarity and Gaussian interaction profile kernel similarity. Ding *et al.* [48] proposed a method called DMHM by using graph-based regularization on the manifold heterogenous networks integrating target information. You *et al.* presented a method called path-based miRNA-disease association (PBMDA) prediction by integrating known human miRNA-disease associations, miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity for miRNAs and diseases. Chen *et al.* [49] proposed a method called MDHGI by integrating the predicted association probability obtained from matrix decomposition through sparse learning, miRNA functional similarity, disease similarity and Gaussian interaction profile kernel similarity. Zeng [50] *et al.* propose a method using structural perturbation method (SPM) on the integration of miRNA-disease association network, miRNA similarity network and disease similarity network. Zeng [51] *et al.* proposed a method namely neural network model for miRNA-disease association prediction (NNMDA) on also the heterogenous network by integrating neighborhood information in the neural network which also consider the imbalance of datasets. This model predicts miRNA-disease associations by integrating multiple biological data resources.

Some algorithms use only the known miRNA-disease associations. For instance, Sun *et al.* [37] proposed a method, namely network topological similarity (NTSMDA), which used only the known miRNA-disease associations as bipartite network. Firstly, they constructed the adjacency matrix representing the association between miRNA and disease. Two matrices, e.g., miRNA network topological similarity and disease network topological similarity matrix, were constructed by using Gaussian interaction profile kernel on the constructed adjacency matrix. The two topological similarity matrices then were integrated with the original adjacency matrix as linear combinations. Finally, the resource allocation was utilized on the two matrices to improve the network-inference by incorporating them together. Li *et al.* [30] developed a matrix completion technique, namely MCMDA, to predict the associations between miRNA and disease on the miRNA-disease adjacency matrix. MCMDA cannot predict new disease or miRNA that do not have any known association at all. Methods that exploit different sources of information normally perform better than the methods that use only known associations. However, it is noteworthy that the extra information about the characteristics of the disease and miRNA is not always available. Therefore, an effective method utilizing only the known associations which are already experimentally validated is still a very necessary bioinformatics tool. Intuitively, the effective computational methods should not be only capable of producing accurate predicting associations, but also be able to predict potential associations of diseases or miRNAs that do not have any known association at all, as well as predicting the associations of miRNAs and diseases that have only known associations. We illustrate some of miRNA-disease association prediction methods and their characteristics in table 1.

Although many methods have been proposed to predict the associations between miRNAs and diseases, their performances are still not satisfying. Hence there is still a room to grow. In this study, we proposed an effective and flexible method namely linear optimization for miRNA-disease association (LOMDA). On one hand, by using merely known associations LOMDA performs much better than the other methods. On the other hand, extra information of miRNA functional similarity and gene functional similarity boost up the performance of LOMDA. The proposed model can also predict the miRNAs or diseases that do not have any known associations, but have some characteristics information. Moreover, the study cases demonstrate the effectiveness of LOMDA on predicting the novel associations between miRNAs and diseases. In a word, LOMDA would be a promising bioinformatics tool for biomedical researches. The Matlab source code of the proposed method and data implemented in this work can be obtained at `https://github.com/rathapech/LOMDA`.

## 2. Method

It is believed that similar diseases are associated with similar miRNAs and vice versus [46, 50]. Embedding this information, e.g., disease similarity, gene functional similarity, target information and so on, into the model will boost the prediction performance. In this work, we propose a method based on linear optimization namely LOMDA to solve miRNA-disease association prediction problem. Whenever extra information about the characteristics of disease and gene functions are available, we can embed them into the model. In particular,

Table 1: Some of miRNA-disease association prediction methods and their characteristics.

| Method | Extra information | No extra information | New miRNA/disease |
|---|---|---|---|
| NTSMDA | | ✓ | |
| MCMDA | | ✓ | |
| RWRMDA | ✓ | | |
| HDMP | | ✓ | |
| RLSMDA | ✓ | | ✓ |
| MIDP | ✓ | | ✓ |
| SPM | ✓ | | ✓ |
| NNMDA | ✓ | | ✓ |
| LOMDA | ✓ | ✓ | ✓ |

we integrate miRNA functional similarity and gene functional similarity and miRNA-disease association to obtain a heterogenous matrix as

$$\mathbf{A} = \begin{bmatrix} MS & MD \\ MD' & DS \end{bmatrix}, \tag{1}$$

where $MS \in \mathbb{R}^{m\times m}$ is the miRNA functional similarity matrix in which $m$ is the number of miRNA, $DS \in \mathbb{R}^{d\times d}$ is the disease semantic similarity matrix in which $d$ is the number of disease and $MD \in \mathbb{R}^{m\times d}$ is the known miRNA-disease association matrix. After integrating the three types of information into a heterogenous matrix, we obtain a square matrix $\mathbf{A} \in \mathbb{R}^{(m+d)\times(m+d)}$. On the other hand, when only the associations between miRNAs and diseases are available, we can set $\mathbf{A} = \mathbf{MD}$.

Disease similarity is obtained from HumanNet database [52] which contains the log likelihood score ($LLS$) of each interaction between genes. Zeng *et al.* computed the gene functional similarity [50] as

$$DS = \begin{cases} \frac{\sum_{x\in S(d_i)} LLS(x,S(d_j)) + \sum_{y\in S(d_j)} LLS(y,S(d_i))}{|S(d_i)|+|S(d_j)|}, & |S(d_i)| + |S(d_j)| \neq 0 \\ 0, & otherwise \end{cases} \tag{2}$$

where $S(d_i)$ and $S(d_j)$ are the the gene sets that related to disease $d_i$ and $d_j$, respectively. $|S|$ is the cardinality of set $S$. $LLS(x, S(d_i))$ is the $LLS$ between gene $x$ and gene set $S(d_i)$, where $x \in d_j$. The miRNA functional similarity [50] is obtained from four source of information including verified miRNA-target associations ($RST$), miRNA family information ($RSF$), cluster information ($RSC$) and verified miRNA-disease associations ($RSD$).

$$MS(r_i, r_j) = \eta.RST(r_i, r_j) + \beta.RSF(r_i, r_j) + \gamma.RSC(r_i, r_j) + \theta.RSD(r_i, r_j) \quad (3)$$

where $\eta, \beta, \gamma$ and $\theta$ are the parameter to adjust the four weights and were set as $\alpha = 0.2$, $\beta = 0.1$, $\gamma = 0.2$ and $\theta = 0.5$.

### *2.1. LOMDA*

Denoting the integration matrix by $\mathbf{A}$ as shown in Eq. (1), we assume that the likelihood of the associations between miRNA and disease can be written as a linear combination of $\mathbf{A}$ and weighting matrix $\mathbf{Z}$ as

$$\mathbf{S} = \mathbf{AZ}. \quad (4)$$

Since $\mathbf{S}$ and $\mathbf{Z}$ are unknown, the problem of Eq. (4) has infinite solutions. However, in order to obtain the likelihood $\mathbf{S}$ containing the existing and predicted associations, $\mathbf{S}$ should be intuitively and reasonably close to $\mathbf{A}$. Then we can write

$$||\mathbf{A} - \mathbf{AZ}|| < \epsilon, \quad (5)$$

where $\epsilon$ is the threshold parameter. Moreover, to avoid the model to be overfitted and simultaneously to constrain the magnitudes of $\mathbf{Z}$, we can relax the Eq. (5) as

$$\mathbf{E} = \alpha||\mathbf{A} - \mathbf{AZ}|| + ||\mathbf{Z}||, \quad (6)$$

where $\alpha$ is the positive free parameter greater than 0 and $||.||$ is the matrix norm. Without losing the generality [53], we use Frobenius norm and raise the two terms with power 2. We can have

$$\mathbf{E} = \alpha||\mathbf{A} - \mathbf{AZ}||_F^2 + ||\mathbf{Z}||_F^2, \quad (7)$$

where Frobenius norm is denoted as $||\mathbf{Z}||_F = \sqrt{\text{trace}(\mathbf{Z}^T\mathbf{Z})} = \sqrt{\sum_{i=1}^{min\{p,q\}} \sigma_i^2}$, e.g., $\sigma_i$ is the singular value, $p$ is the number of row, and $q$ is the number of column of $\mathbf{Z}$. The expansion of Eq. (7) reads

$$\begin{aligned}\mathbf{E} &= \alpha\mathbf{Tr}[(\mathbf{A}-\mathbf{A}\mathbf{Z})^T(\mathbf{A}-\mathbf{A}\mathbf{Z})]+\mathbf{Tr}(\mathbf{Z}^T\mathbf{Z}) \\ &= \alpha\mathbf{Tr}(\mathbf{A}^T\mathbf{A}-\mathbf{A}^T\mathbf{A}\mathbf{Z}-\mathbf{Z}^T\mathbf{A}^T\mathbf{A}+\mathbf{Z}^T\mathbf{A}^T\mathbf{A}\mathbf{Z})+\mathbf{Tr}(\mathbf{Z}^T\mathbf{Z}),\end{aligned} \tag{8}$$

with its partial derivative being

$$\frac{\partial\mathbf{E}}{\partial\mathbf{Z}} = \alpha(-2\mathbf{A}^T\mathbf{A}+2\mathbf{A}^T\mathbf{A}\mathbf{Z})+2\mathbf{Z}. \tag{9}$$

Setting $\partial\mathbf{E}/\partial\mathbf{Z}=0$, we can obtain the optimal solution of $\mathbf{Z}$ as

$$\mathbf{Z}^* = \alpha(\alpha\mathbf{A}^T\mathbf{A}+\mathbf{I})^{-1}\mathbf{A}^T\mathbf{A}, \tag{10}$$

where $\mathbf{I}$ is the identity matrix. The likelihood matrix $\mathbf{S}$ can be obtained as

$$\mathbf{S} = \mathbf{A}\mathbf{Z}^*. \tag{11}$$

Finally, $\mathbf{S}$ is utilized as scores of each pair of the miRNA and disease association after disregarding the scores of the known associations.

## 3. Results

### *3.1. Performance evaluation*

To evaluate our proposed method against others, we adopt cross validation techniques including leave-one-out cross validation and 5-fold cross validation which are the widely used evaluation methods. In the LOOCV, we randomly remove one association of each disease as testing samples and use the remaining associations as the training samples. For the 5-fold cross validation, we randomly divide all the known associations of diseases and miRNAs into 5 subsets. We utilize the four subsets of the five subsets as training samples and leave the remaining subset as testing samples. We repeatedly and independently do this for five times until all the five subsets are utilized as testing samples exactly once. The database contains 336 diseases, 577 miRNAs and 6441 associations.

For the LOOCV and 5-fold cross validation, we use the area under the receiver operating characteristic (ROC) curve (AUC) to evaluate the predicting

Table 2: The AUC values from different methods by using 5-fold cross validation on 15 diseases.

| Disease | RWRMDA | HDMP | RLSMDA | MIDP | SPM | NNMDA | LOMDA-**MD** | LOMDA-**A** |
|---|---|---|---|---|---|---|---|---|
| Breast neoplasms | 0.785 | 0.801 | 0.832 | 0.838 | 0.932 | 0.968 | 0.963 | **0.998** |
| Hepatocellular carcinoma | 0.749 | 0.759 | 0.794 | 0.807 | 0.918 | 0.966 | 0.936 | **0.999** |
| Renal cell carcinoma | 0.815 | 0.833 | 0.839 | 0.862 | 0.901 | 0.912 | 0.978 | **0.995** |
| Squamous cell carcinoma | 0.819 | 0.820 | 0.849 | 0.870 | 0.899 | 0.924 | 0.950 | **0.997** |
| Colorectal neoplasms | 0.793 | 0.802 | 0.831 | 0.845 | 0.885 | 0.927 | 0.988 | **0.998** |
| Glioblastoma | 0.680 | 0.700 | 0.714 | 0.786 | 0.840 | 0.911 | 0.978 | **0.996** |
| Heart disease | 0.722 | 0.770 | 0.738 | 0.821 | 0.950 | 0.945 | 0.979 | **0.997** |
| Acute myeloid leukemia | 0.839 | 0.858 | 0.853 | 0.915 | 0.957 | 0.916 | 0.987 | **0.997** |
| Lung neoplasms | 0.827 | 0.835 | 0.855 | 0.876 | 0.892 | 0.943 | 0.984 | **0.998** |
| Melanoma | 0.784 | 0.790 | 0.807 | 0.837 | 0.951 | 0.949 | 0.985 | **0.999** |
| Ovarian neoplasms | 0.882 | 0.884 | 0.909 | 0.923 | 0.949 | 0.928 | 0.991 | **0.998** |
| Pancreatic neoplasms | 0.871 | 0.895 | 0.887 | 0.945 | 0.954 | 0.954 | 0.986 | **0.997** |
| Prostatic neoplasms | 0.823 | 0.854 | 0.841 | 0.882 | 0.928 | 0.936 | 0.975 | **0.996** |
| Stomach neoplasms | 0.779 | 0.787 | 0.797 | 0.821 | 0.859 | 0.955 | 0.980 | **0.999** |
| Urinary bladder neoplasm | 0.821 | 0.850 | 0.845 | 0.897 | 0.898 | 0.920 | 0.984 | **0.997** |
| Average $AUC$ | 0.799 | 0.816 | 0.826 | 0.862 | 0.914 | 0.937 | 0.986 | **0.997** |

Table 3: The prediction of the top 30 predicted miRNAs associated with breast neoplasms based on known associations in HMDD database. The first column records top 1-15 related miRNA, while the third column records top 16-30.

| miRNA | Evidence | miRNA | Evidence |
|---|---|---|---|
| hsa-mir-106a | dbDEMC | hsa-mir-138 | dbDEMC |
| hsa-mir-142 | *unconfirmed* | hsa-mir-574 | *unconfirmed* |
| hsa-mir-130a | dbDEMC | hsa-mir-15b | dbDEMC |
| hsa-mir-99a | dbDEMC | hsa-mir-19b | dbDEMC |
| hsa-mir-150 | dbDEMC | hsa-mir-30e | *unconfirmed* |
| hsa-mir-378a | *unconfirmed* | hsa-mir-542 | *unconfirmed* |
| hsa-mir-186 | dbDEMC | hsa-mir-650 | dbDEMC |
| hsa-mir-92b | dbDEMC | hsa-mir-98 | dbDEMC;miR2Disease |
| hsa-mir-185 | dbDEMC | hsa-mir-370 | dbDEMC |
| hsa-mir-130b | dbDEMC | hsa-mir-99b | dbDEMC |
| hsa-mir-192 | dbDEMC | hsa-mir-95 | dbDEMC |
| hsa-mir-212 | dbDEMC | hsa-mir-32 | dbDEMC |
| hsa-mir-449b | dbDEMC | hsa-mir-196b | dbDEMC |
| hsa-mir-372 | dbDEMC | hsa-mir-449a | dbDEMC |
| hsa-mir-330 | dbDEMC | hsa-mir-181c | dbDEMC |

Table 4: The prediction of the top 30 predicted miRNAs associated with colon neoplasms based on known associations in HMDD database. The first column records top 1-15 related miRNA, while the third column records top 16-30.

| miRNA | Evidence | miRNA | Evidence |
|---|---|---|---|
| hsa-mir-200a | dbDEMC | hsa-mir-224 | dbDEMC |
| hsa-mir-92a | dbDEMC | hsa-mir-429 | dbDEMC |
| hsa-mir-29b | dbDEMC | hsa-mir-34b | *unconfirmed* |
| hsa-mir-222 | dbDEMC | hsa-mir-199a | *unconfirmed* |
| hsa-mir-34c | *unconfirmed* | hsa-mir-148a | dbDEMC |
| hsa-mir-100 | dbDEMC | hsa-mir-27a | dbDEMC;miR2Disease |
| hsa-mir-210 | dbDEMC | hsa-mir-181b | dbDEMC;miR2Disease |
| hsa-mir-182 | dbDEMC | hsa-mir-195 | dbDEMC |
| hsa-mir-375 | dbDEMC | hsa-mir-20b | dbDEMC |
| hsa-mir-1-2 | dbDEMC | hsa-mir-150 | dbDEMC |
| hsa-mir-203 | dbDEMC;miR2Disease | hsa-mir-103a | *unconfirmed* |
| hsa-mir-181a | dbDEMC;miR2Disease | hsa-mir-181a-2 | *unconfirmed* |
| hsa-mir-30d | dbDEMC | hsa-mir-146b | dbDEMC |
| hsa-mir-99a | dbDEMC | hsa-mir-25 | dbDEMC;miR2Disease |
| hsa-mir-29c | dbDEMC | hsa-mir-151a | *unconfirmed* |

accuracy. The curve displays true positive rate (sensitivity) versus false positive rate (1-specificity) at different values of thresholds. On one hand, sensitivity is the percentage of the test samples in which rank higher than a given threshold. On the other hand, specificity is the percentage the test samples that fall below the threshold. AUC = 1 indicates the perfect prediction, while AUC = 0.5 indicates random performance.

### *3.2. Comparison with other methods*

A great number of the methods that are developed to predict the potential associations between miRNAs and diseases utilized different sources of information including gene expressions, miRNA functional similarity, disease similarity and miRNA family information. Moreover, it is believed that the methods that combine different sources of information perform better. In this study, we test the proposed method on two manners e.g., the information of miRNA and disease are embedded in the model (called LOMDA-**A**) and only known associations are utilized (called LOMDA-**MD**). We compare LOMDA with many methods including RWRMDA, HDMP, RLSMDA, MIDP, SPM and NNMDA [51].

Table 2 illustrates the predicting performances of the proposed method and others on different diseases. The highest values generating from any methods are shown in boldface. The average AUC values among the 15 disease, as shown in bottom row of the table 2, of RWRMDA, HDMP, RLSMDA, MIDP, SPM, NNMDA, LOMDA-**MD** and LOMDA-**A** are 0.799, 0.816, 0.826, 0.862, 0.914, 0.937, 0.986 and 0.997, respectively. In other words, LOMDA–**A** and LOMDA–**MD** perform higher than the others. LOMDA–**A** specifically outperforms RWRMDA, HDMP, RLSMDA, MIDP, SPM and NNMDA in all the 15 diseases by 24.7%, 22.1%, 20.7%, 15.6%, 9.1% and 6.4%, respectively. The highest AUC values generating from LOMDA-**A** on three diseases e.g., hepatocellular carcinoma, melanoma and stomach neoplasms are up to 0.999. Moreover, LOOCV obtained from LOMDA-**A**, NNMDA, SPM, HDMP and RLSMDA are 0.866, 0.843, 0.811, 0.770 and 0.695, respectively.

### *3.3. Case studies*

In order to verify the effectiveness of LOMDA, case studies of three critical diseases including breast neoplasms, colon neoplasms and kidney neoplasms have been investigated to predict the potential associations. Breast neoplasms is the most common malignant tumor in women accounting for 25% [54] followed by prostate and colon cancer [55, 56]. Moreover, colon neoplasm is one of the common cancers which has high death rate [57].

In these case studies, all the known associations are utilized as training samples and the unknown are the testing samples. First of all, we compute the scores of all the unknown associations then sorted in descending order corresponding to each disease. After that, we select the top 30 association scores of the interested disease and manually verify the existences of the associations in other two miRNA-disease databases, e.g., dbDEMC [26] and miR2Disease

Table 5: The prediction of the top 30 predicted miRNAs associated with kidney neoplasms based on known associations in HMDD database. The first column records top 1-15 related miRNA, while the third column records top 16-30.

| miRNA | Evidence | miRNA | Evidence |
|---|---|---|---|
| hsa-mir-21 | dbDEMC;miR2Disease | hsa-mir-34c | dbDEMC |
| hsa-mir-146a | dbDEMC | hsa-mir-30a | dbDEMC |
| hsa-mir-17 | dbDEMC;miR2Disease | hsa-mir-638 | dbDEMC |
| hsa-mir-200a | dbDEMC;miR2Disease | hsa-mir-320a | dbDEMC |
| hsa-mir-451a | dbDEMC | hsa-mir-499a | *unconfirmed* |
| hsa-mir-200b | dbDEMC;miR2Disease | hsa-mir-1207 | *unconfirmed* |
| hsa-mir-433 | *unconfirmed* | hsa-let-7b | dbDEMC |
| hsa-mir-192 | dbDEMC | hsa-mir-362 | dbDEMC |
| hsa-mir-198 | dbDEMC | hsa-mir-148b | *unconfirmed* |
| hsa-mir-205 | dbDEMC;miR2Disease | hsa-mir-199a | dbDEMC;miR2Disease |
| hsa-mir-423 | dbDEMC | hsa-mir-15a | dbDEMC;miR2Disease |
| hsa-mir-200c | dbDEMC;miR2Disease | hsa-mir-181a | dbDEMC |
| hsa-mir-30b | dbDEMC | hsa-mir-206 | dbDEMC |
| hsa-mir-10a | dbDEMC | hsa-mir-182 | dbDEMC;miR2Disease |
| hsa-mir-377 | dbDEMC | hsa-mir-196b | dbDEMC |

[27]. The predicting results of the three diseases are shown in table 3, 4 and 5, respectively.

In addition, to verify the prediction of LOMDA on the disease without any known associations, we remove all the associations of hepatocellular carcinoma and lung neoplasms with all the miRNAs. Then we compute the likelihood scores of these diseases with all the miRNAs by using LOMDA-**A**. Finally, we select the top 30 candidates and manually check these candidates in the three databases including HMDD, dbDEMC and miR2Disease. All these predicted candidates belonging to kidney neoplasms can be confirmed in at least one of the three databases, while 29 among 30 predicted associations of lung neoplasms are also confirmed. These predicted results are shown in table 6 and 7, respectively.

Table 6: The top 30 miRNAs associated with hepatocellular carcinoma were predicted by LOMDA with hiding all known related miRNAs.

| miRNA | Evidence | miRNA | Evidence |
|---|---|---|---|
| hsa-mir-21 | HMDD;dbDEMC;miR2Disease | hsa-mir-18a | HMDD;dbDEMC;miR2Disease |
| hsa-mir-155 | HMDD;dbDEMC | hsa-mir-19b-1 | HMDD;dbDEMC |
| hsa-mir-146a | HMDD;dbDEMC;miR2Disease | hsa-mir-221 | HMDD;dbDEMC;miR2Disease |
| hsa-mir-17 | HMDD;dbDEMC | hsa-mir-29b-2 | HMDD |
| hsa-mir-20a | HMDD;dbDEMC;miR2Disease | hsa-mir-16-2 | HMDD;dbDEMC |
| hsa-mir-125b-1 | HMDD;miR2Disease | hsa-mir-223 | HMDD;dbDEMC;miR2Disease |
| hsa-mir-34a | HMDD;miR2Disease | hsa-mir-19a | HMDD;dbDEMC;miR2Disease |
| hsa-mir-29a | HMDD;dbDEMC | hsa-mir-29c | HMDD;dbDEMC |
| hsa-mir-125b-2 | HMDD;dbDEMC;miR2Disease | hsa-mir-181a-1 | HMDD;miR2Disease |
| hsa-mir-15a | HMDD;dbDEMC;miR2Disease | hsa-mir-150 | HMDD;dbDEMC |
| hsa-mir-29b-1 | HMDD;dbDEMC | hsa-mir-1-1 | HMDD |
| hsa-mir-92a-1 | HMDD | hsa-mir-181a-2 | HMDD;dbDEMC;miR2Disease |
| hsa-mir-126 | HMDD;dbDEMC;miR2Disease | hsa-mir-92a-2 | HMDD |
| hsa-mir-145 | HMDD;dbDEMC;miR2Disease | hsa-mir-122 | HMDD;dbDEMC;miR2Disease |
| hsa-mir-16-1 | HMDD | hsa-mir-1-2 | HMDD;dbDEMC;miR2Disease |

Table 7: The top 30 miRNAs associated with lung neoplasms were predicted by LOMDA with hiding all known related miRNAs.

| miRNA | Evidence | miRNA | Evidence |
|---|---|---|---|
| hsa-mir-21 | HMDD;dbDEMC;miR2Disease | hsa-mir-146a | HMDD;dbDEMC;miR2Disease |
| hsa-mir-34c | HMDD;dbDEMC | hsa-mir-30e | HMDD;dbDEMC |
| hsa-mir-155 | HMDD;dbDEMC | hsa-mir-152 | HMDD;dbDEMC |
| hsa-mir-486 | HMDD;dbDEMC | hsa-mir-219-2 | *unconfirmed* |
| hsa-mir-15a | HMDD;dbDEMC | hsa-mir-223 | HMDD |
| hsa-mir-34a | HMDD;dbDEMC | hsa-mir-520c | HMDD |
| hsa-mir-148b | HMDD;dbDEMC | hsa-mir-137 | HMDD;dbDEMC |
| hsa-mir-126 | HMDD;dbDEMC | hsa-mir-134 | dbDEMC |
| hsa-mir-29a | HMDD;dbDEMC;miR2Disease | hsa-mir-125b | HMDD;dbDEMC;miR2Disease |
| hsa-mir-200b | HMDD;dbDEMC;miR2Disease | hsa-mir-204 | HMDD;dbDEMC |
| hsa-mir-326 | HMDD;dbDEMC | hsa-mir-374a | HMDD |
| hsa-mir-34b | HMDD;dbDEMC | hsa-mir-7b | HMDD;miR2Disease |
| hsa-mir-337 | HMDD;dbDEMC | hsa-mir-375 | HMDD;dbDEMC |
| hsa-mir-25 | HMDD;dbDEMC | hsa-mir-92b | HMDD;dbDEMC |
| hsa-mir-122 | HMDD;dbDEMC | hsa-mir-28 | HMDD;dbDEMC |

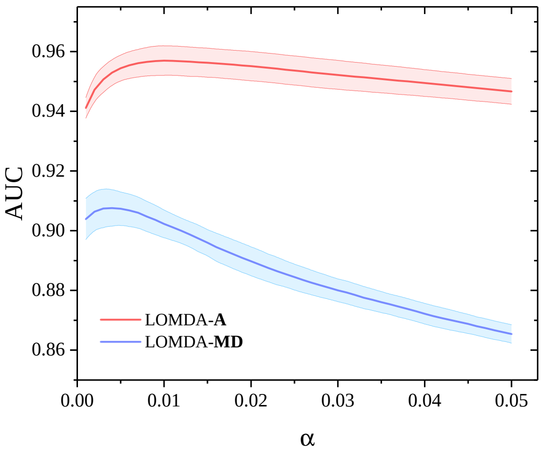

Figure 1: The five-fold cross validation AUCs of LOMDA corresponding to different $\alpha$ for the integration matrix (**A**) and the only known miRNA-disease association matrix (**MD**).

### *3.4. Parameter $\alpha$*

In the proposed method, there contains a parameter $\alpha$ playing a role to control the residual and to avoid overfitting of the model. The optimal value of $\alpha$ for the integration matrix $\mathbf{A}$ is 0.01, while the optimal $\alpha$ for the only known interaction matrix, e.g., $\mathbf{A} = \mathbf{MD}$, is 0.001. As shown in Figure 1, $\alpha$ is not sensitive to the performance, especially when extra information are embedded into the model. That means more information provided, the more stable the model. We train the model to obtain the optimal value of this parameter by using cross-validation technique and tune $\alpha$ from 0.001 to 0.05 with 0.001 step. The optimal $\alpha$ is the one that produces the highest accuracy.

## 4. Conclusion

Predicting the novel associations between miRNA and disease helps scientists firstly focus on the most likely associations rather than blindly check on all the possible associations which is extremely costly and laborious. Moreover, it can help researchers enhance their understanding toward the molecular mechanisms of disease at the miRNA level. This prediction also plays an important role in understanding the pathogenesis of human disease at the early stage, therefore, it can help in diagnose, treatment and prevention. Motivated by the necessity of the identifying the novel associations between miRNAs and diseases, in this work we proposed a method, namely linear optimization for miRNA-disease association (LOMDA), to predict the potential associations between miRNAs and diseases. The proposed method utilizes the heterogenous matrix by integrating the miRNA functional similarity, disease gene similarity and known miRNA-disease associations. In case only known associations are available, the method can also be applied. Moreover, the method can also predict the associations of new miRNAs (or diseases) by using miRNA functional similarity (or disease semantic similarity). According to the cross validation evaluated by AUC, the proposed method has been shown to perform very satisfied. Thus, LOMDA is an effective and flexible tool for predicting miRNA-disease associations. Firstly, the researchers can apply this method to predict the potential associations by computing the association scores and finally choose the most promising associations for further biological experiment.

LOMDA contains a parameter $\alpha$. In order to find the optimal value of this $\alpha$, one might need to apply cross validation by dividing the known association into training samples and testing samples, then the AUC is computed based on different values of parameters. The optimal $\alpha$ is the one that produces highest AUC. When the data is large enough, the optimal parameter of this division is approximately the same as that of the whole dataset.

## References

[1] D. P. Bartel, MicroRNAs: genomics, biogenesis, mechanism, and function, Cell 116 (2) (2004) 281–297.

[2] S. Chatterjee, H. Großhans, Active turnover modulates mature microRNA activity in caenorhabditis elegans, Nature 461 (7263) (2009) 546.

[3] C. Llave, Z. Xie, K. D. Kasschau, J. C. Carrington, Cleavage of scarecrow-

like mRNA targets directed by a class of arabidopsis miRNA, Science 297 (5589) (2002) 2053–2056.

[4] A. Eulalio, E. Huntzinger, E. Izaurralde, Getting to the root of miRNA-mediated gene silencing, Cell 132 (1) (2008) 9–14.

[5] L. Zhu, J. Zhao, J. Wang, C. Hu, J. Peng, R. Luo, C. Zhou, J. Liu, J. Lin, Y. Jin, et al., MicroRNAs are involved in the regulation of ovary development in the pathogenic blood fluke schistosoma japonicum, PLoS Pathogens. 12 (2) (2016) e1005423.

[6] T. R. Fernando, N. I. Rodriguez-Malave, D. S. Rao, MicroRNAs in B cell development and malignancy, J Hematol. Oncol. 5 (1) (2012) 7.

[7] M. Lize, S. Pilarski, M. Dobbelstein, E2F1-inducible microRNA 449a/b suppresses cell proliferation and promotes apoptosis, Cell Death and Differ. 17 (3) (2010) 452.

[8] A. Esquela-Kerscher, F. J. Slack, Oncomirs—microRNAs with a role in cancer, Nat. Rev. Cancer 6 (4) (2006) 259.

[9] R. F. Duisters, A. J. Tijsen, B. Schroen, J. J. Leenders, V. Lentink, I. van der Made, V. Herias, R. E. van Leeuwen, M. W. Schellings, P. Barenbrug, et al., mir-133 and mir-30 regulate connective tissue growth factor: implications for a role of microRNAs in myocardial matrix remodeling, Circ. Res. 104 (2) (2009) 170–178.

[10] N. Lynam-Lennon, S. G. Maher, J. V. Reynolds, The roles of microRNA in cancer and apoptosis, Biol. Rev. 84 (1) (2009) 55–71.

[11] A. Etheridge, I. Lee, L. Hood, D. Galas, K. Wang, Extracellular microRNA: a new source of biomarkers, Mutat. Res. - Fund. Mol. M. 717 (1-2) (2011) 85–90.

[12] C. E. S. Espinosa, F. J. Slack, Cancer issue: the role of microRNAs in cancer, Yale J. Biol. Med. 79 (3-4) (2006) 131.

[13] G. A. Calin, C. M. Croce, MicroRNA signatures in human cancers, Nat. Rev. Cancer 6 (11) (2006) 857.

[14] J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, et al., MicroRNA expression profiles classify human cancers, Nature 435 (7043) (2005) 834.

[15] M. S. Weinberg, M. J. Wood, Short non-coding RNA biology and neurodegenerative disorders: novel disease targets and therapeutics, Hum. Mol. Genet. 18 (R1).

[16] A. Markou, E. G. Tsaroucha, L. Kaklamanis, M. Fotinou, V. Georgoulias, E. S. Lianidou, Prognostic value of mature microRNA-21 and microRNA-205 overexpression in non–small cell lung cancer by quantitative real-time RT-PCR, Clin. Chem. 54 (10) (2008) 1696–1704.

[17] M. V. Iorio, M. Ferracin, C.-G. Liu, A. Veronese, R. Spizzo, S. Sabbioni, E. Magri, M. Pedriali, M. Fabbri, M. Campiglio, et al., MicroRNA gene expression deregulation in human breast cancer, Cancer Res. 65 (16) (2005) 7065–7070.

[18] Q. Huang, K. Gumireddy, M. Schrier, C. Le Sage, R. Nagel, S. Nair, D. A. Egan, A. Li, G. Huang, A. J. Klein-Szanto, et al., The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis, Nat. Cell Biol. 10 (2) (2008) 202.

[19] M. V. Latronico, D. Catalucci, G. Condorelli, Emerging role of microRNAs in cardiovascular biology, Circ. Res. 101 (12) (2007) 1225–1236.

[20] R. W. Chen, L. T. Bemis, C. M. Amato, H. Myint, H. Tran, D. K. Birks, S. G. Eckhardt, W. A. Robinson, Truncation in CCND1 mRNA alters miR-16-1 regulation in mantle cell lymphoma, Blood 112 (3) (2008) 822–829.

[21] F. Xin, M. Li, C. Balch, M. Thomson, M. Fan, Y. Liu, S. M. Hammond, S. Kim, K. P. Nephew, Computational analysis of microRNA profiles and

their target genes suggests significant involvement in breast cancer antiestrogen resistance, Bioinformatics 25 (4) (2008) 430–434.

[22] J. Xu, C.-X. Li, J.-Y. Lv, Y.-S. Li, Y. Xiao, T.-T. Shao, X. Huo, X. Li, Y. Zou, Q.-L. Han, et al., Prioritizing candidate disease miRNAs by topological features in the miRNA target–dysregulated network: Case study of prostate cancer, Mol. Cancer Ther. 10 (10) (2011) 1857–1866.

[23] Z. Yu, Z. Li, N. Jolicoeur, L. Zhang, Y. Fortin, E. Wang, M. Wu, S.-H. Shen, Aberrant allele frequencies of the SNPs located in microRNA target sites are potentially associated with human cancers, Nucleic Acids. Res. 35 (13) (2007) 4535–4541.

[24] Y. Xiao, J. Guan, Y. Ping, C. Xu, T. Huang, H. Zhao, H. Fan, Y. Li, Y. Lv, T. Zhao, et al., Prioritizing cancer-related key miRNA–target interactions by integrative genomics, Nucleic Acids. Res. 40 (16) (2012) 7653–7665.

[25] Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, Q. Cui, HMDD v2. 0: a database for experimentally supported human microRNA and disease associations, Nucleic Acids. Res. 42 (D1) (2013) D1070–D1074.

[26] Z. Yang, L. Wu, A. Wang, W. Tang, Y. Zhao, H. Zhao, A. E. Teschendorff, dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers, Nucleic Acids. Res. 45 (D1) (2016) D812–D818.

[27] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, Y. Liu, miR2Disease: a manually curated database for microRNA deregulation in human disease, Nucleic Acids. Res. 37 (suppl_1) (2008) D98–D104.

[28] X. Zeng, X. Zhang, Q. Zou, Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks, Brief. Bioinform. 17 (2) (2015) 193–203.

[29] C. Pasquier, J. Gardès, Prediction of miRNA-disease associations with a vector space model, Sci. Rep. 6 (2016) 27036.

[30] J.-Q. Li, Z.-H. Rong, X. Chen, G.-Y. Yan, Z.-H. You, MCMDA: Matrix completion for MiRNA-disease association prediction, Oncotarget 8 (13) (2017) 21187.

[31] N. Omranian, J. M. Eloundou-Mbebi, B. Mueller-Roeber, Z. Nikoloski, Gene regulatory network inference using fused LASSO on multiple data sets, Sci. Rep. 6 (2016) 20533.

[32] Q. Jiang, G. Wang, S. Jin, Y. Li, Y. Wang, Predicting human microRNA-disease associations based on support vector machine, International Journal of Data Mining and Bioinformatics 8 (3) (2013) 282–293.

[33] X. Chen, C. C. Yan, X. Zhang, Z. Li, L. Deng, Y. Zhang, Q. Dai, RBM-MMDA: predicting multiple types of disease-microRNA associations, Sci. Rep. 5 (2015) 13877.

[34] H. Shi, J. Xu, G. Zhang, L. Xu, C. Li, L. Wang, Z. Zhao, W. Jiang, Z. Guo, X. Li, Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes, BMC Syst. Biol. 7 (1) (2013) 101.

[35] X. Chen, M.-X. Liu, G.-Y. Yan, RWRMDA: predicting novel human microRNA–disease associations, Mol. Biosyst. 8 (10) (2012) 2792–2798.

[36] H. Chen, Z. Zhang, Similarity-based methods for potential human microRNA-disease association prediction, BMC Med. Genet. 6 (1) (2013) 12.

[37] D. Sun, A. Li, H. Feng, M. Wang, NTSMDA: prediction of miRNA–disease associations by integrating network topological similarity, Mol. Biosyst. 12 (7) (2016) 2224–2232.

[38] D. Wang, J. Wang, M. Lu, F. Song, Q. Cui, Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases, Bioinformatics 26 (13) (2010) 1644–1650.

[39] Q. Zou, J. Li, Q. Hong, Z. Lin, Y. Wu, H. Shi, Y. Ju, Prediction of microRNA-disease associations based on social network analysis methods, BioMed. Res. Int. 2015.

[40] M. Chen, X. Lu, B. Liao, Z. Li, L. Cai, C. Gu, Uncover miRNA-disease association by exploiting global network similarity, PloS ONE 11 (12) (2016) e0166509.

[41] Q. Jiang, Y. Hao, G. Wang, L. Juan, T. Zhang, M. Teng, Y. Liu, Y. Wang, Prioritization of disease microRNAs through a human phenome-microRNAome network, BMC Syst. Biol. 4 (1) (2010) S2.

[42] P. Xuan, K. Han, M. Guo, Y. Guo, J. Li, J. Ding, Y. Liu, Q. Dai, J. Li, Z. Teng, et al., Prediction of microRNAs associated with human diseases based on weighted $k$ most similar neighbors, PloS ONE 8 (8) (2013) e70204.

[43] X. Chen, G.-Y. Yan, Semi-supervised learning for potential human microrna-disease associations inference, Sci. Rep. 4 (2014) 5501.

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

[45] X. Chen, L. Huang, D. Xie, Q. Zhao, EGBMMDA: extreme gradient boosting machine for miRNA-disease association prediction, Cell Death & Disease 9 (1) (2018) 3.

[46] P. Xuan, K. Han, Y. Guo, J. Li, X. Li, Y. Zhong, Z. Zhang, J. Ding, Prediction of potential disease-associated microRNAs based on random walk, Bioinformatics 31 (11) (2015) 1805–1815.

[47] X. Chen, Y.-W. Niu, G.-H. Wang, G.-Y. Yan, HAMDA: hybrid approach for mirna-disease association prediction, J. Biomed. Inform. 76 (2017) 50–58.

[48] P. Ding, J. Luo, C. Liang, Q. Xiao, B. Cao, Human disease MiRNA inference by combining target information based on heterogeneous manifolds, J. Biomed. Inform. 80 (2018) 26–36.

[49] X. Chen, J. Yin, J. Qu, L. Huang, MDHGI: Matrix decomposition and heterogeneous graph inference for mirna-disease association prediction, PLoS Comp. Biol. 14 (8) (2018) e1006418.

[50] X. Zeng, L. Liu, L. Lü, Q. Zou, Prediction of potential disease-associated microRNAs using structural perturbation method, Bioinformatics 34 (14) (2018) 2425–2432.

[51] X. Zeng, W. Wang, G. Deng, Bing, Z. Quan, Prediction of potential disease-associated MicroRNAs by using neural networks, Mol. Ther. Nucleic Acids 16 (2019) 566–575.

[52] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, E. M. Marcotte, Prioritizing candidate disease genes by network-based boosting of genome-wide association data, Genome Res. 21 (7) (2011) 1109–1121.

[53] R. Pech, D. Hao, Y.-L. Lee, Y. Yuan, T. Zhou, Link prediction via linear optimization, Physica A (2019) 121319.

[54] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, D. Forman, Global cancer statistics, CA Cancer J. Clin. 61 (2) (2011) 69–90.

[55] M. Al-Hajj, M. S. Wicha, A. Benito-Hernandez, S. J. Morrison, M. F. Clarke, Prospective identification of tumorigenic breast cancer cells, Proc. Natl. Acad. Sci. U.S.A. 100 (7) (2003) 3983–3988.

[56] J. D. Potter, M. L. Slattery, R. M. Bostick, S. M. Gapstur, Colon cancer: a review of the epidemiology, Epidemiol. Rev. 15 (2) (1993) 499–545.

[57] R. L. Siegel, K. D. Miller, A. Jemal, Cancer statistics, 2016, CA Cancer J. Clin. 66 (1) (2016) 7–30.