Computers in Biology and Medicine 110 (2019) 156–163

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

MAMDA: Inferring microRNA-Disease associations with manifold alignment

Fang Yan[a], Yuanjie Zheng[a,b,*], Weikuan Jia[a], Sujuan Hou[a], Rui Xiao[c]

[a] School of Information Science and Engineering at Shandong Normal University, Jinan, China
[b] Key Lab of Intelligent Computing & Information Security in Universities of Shandong, Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Institute of Biomedical Sciences, Shandong Normal University, Jinan, China
[c] Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

ARTICLE INFO

Keywords:
microRNA-disease association
Computational model
microRNA
Disease
Computational biology

ABSTRACT

Uncovering disease-related microRNAs (miRNAs) by inferring miRNA-disease associations is of critical importance for understanding the pathogenesis of disease and carrying out treatment and prevention. Recently developed computational models for inferring miRNA-disease associations assume that functionally related miRNAs are associated with phenotypically similar diseases and hence infer miRNA-disease associations by using miRNA-miRNA and disease-disease similarities, which are concretely determined by mining existing biological resources. From the perspective of manifold learning, miRNA-miRNA similarities and disease-disease similarities determine a low-dimensional manifold for miRNAs and diseases, respectively, and the basic assumption of current computational models is equivalent to consistency between the manifold structures of miRNA and disease. In this paper, we propose a novel microRNA-disease inference framework (MAMDA) that explicitly takes advantage of this consistency property and infers miRNA-disease associations by aligning the manifold structure of miRNA with that of disease together with supervision of experimentally verified miRNA-disease associations. Based on three aspects, experimental results show that the proposed framework outperforms several representative state-of-the-art techniques. First, AUC values using *k*-fold cross-validation indicate that our method acquires more reliable predictions than four classical techniques (HGIMDA, HDMP, RLSMDA, and NCPMDA). Second, 48/48 predicted associations between miRNAs and breast cancer are validated with the HMDD and dbDEMC to show the effectiveness of predicting isolated diseases with unknown miRNAs. Third, two case studies of colon neoplasms and lung neoplasms validate the superior accuracy of MAMDA, with 48/50 and 48/50 predicted associations in the HMDD and dbDEMC, respectively.

1. Introduction

As a class of noncoding RNAs, microRNAs (miRNAs) are single-stranded RNA molecules approximately 22 nucleotides in length that play pivotal roles in regulating gene expression [1]. They may fine-tune the protein-encoding expression of as many as 30% of all mammalian genes [2]. In addition, great discoveries in miRNAs have also shown that miRNAs acting as important new regulatory molecules in different human diseases have great potential in the diagnosis and treatment of many diseases. For instance, it is well documented that upregulation or downregulation of miRNAs occurs in various human cancers. Silencing, antisense blocking and modification of miRNAs are potential therapeutic treatments involving these miRNAs [3,4].

Uncovering miRNA-disease associations is of critical importance not only for investigating disease pathogenesis at the molecular level and facilitating diagnosis, treatment and prevention of disease [5–7] but also for formulating personalized treatment regimens [8]. Several experimental methods have been successfully exploited, such as microarray profiling and RT-PCR [9–11]. However, these experimental methods are expensive and time consuming [12]. To eliminate the drawbacks of these experimental techniques, computational methods have been developed to predict and rank disease-related miRNAs by inferring miRNA-disease associations [13].

The main purpose of these computational methods is to offer reliable miRNA candidates for biological experiments in combination with existing experimental data. This strategy can make up for the shortcomings of the experimental approaches and further improve the efficiency and success rate under the circumstances of many-to-many association maps between miRNAs and diseases. Computational strategies developed in recent years for predicting associations can be divided into the following two categories: relation-learning methods and similarity-computation approaches. They are both based on the assumption that

* Corresponding author. School of Information Science and Engineering at Shandong Normal University, Jinan, China.
E-mail address: yjzheng@sdnu.edu.cn (Y. Zheng).

https://doi.org/10.1016/j.compbiomed.2019.05.014
Received 31 December 2018; Received in revised form 17 May 2019; Accepted 17 May 2019

functionally related miRNAs tend to be associated with phenotypically similar diseases and vice versa [14]. Specifically, the relation-learning methods focus on learning an miRNA-disease relationship, while the similarity-computation approaches generate predictions/inferences by simultaneously considering the experimentally verified miRNA-disease associations, miRNA's functional similarities, and diseases' phenotypic similarities, among other relationships [14–21]. However, the existing studies on miRNA and disease association prediction still face the following problems. On the one hand, the relation-learning methods are subject to data samples that are distributed unevenly. Currently, only a relatively small number of miRNA-disease associations have been confirmed by experiments. The false positives in miRNA target gene prediction and the disease-related annotation information are not sufficient, and numerous undiscovered miRNA-disease associations still exist, leading to the problem of an uneven data distribution. On the other hand, the similarity-computation approaches suffer from inaccurate prediction caused by the sparsity of data and isolated diseases (diseases not related to other diseases and miRNAs). This is because the known diseases and miRNAs have some known associations, which can provide much prior knowledge for subsequently predicting potential relationships. Isolated diseases and unknown miRNAs, in contrast, can hardly provide prior knowledge and thus affect prediction accuracy.

The base assumption that miRNAs with similar functions are associated with diseases with similar phenotypes is equivalent to the consistency between the underlying low-dimensional manifold structure of miRNAs and that of diseases [22]. It is reasonable to assume that latent data for fully describing miRNAs or diseases are in a high-dimensional space, considering their high biological complexity. At the same time, the data should also form an underlying low-dimensional manifold, considering the established similarities between miRNA functions or disease phenotypes. This low-dimensional manifold can actually be reconstructed from the similarities between miRNAs or diseases using recent manifold learning techniques [23–25]. However, this invaluable information is not explicitly leveraged by most of the current miRNA-disease prediction techniques.

In this paper, we develop a novel framework for inferring miRNA-disease associations with manifold alignment (MAMDA). It establishes a regression between miRNAs and diseases by explicitly taking advantage of the consistency between their underlying low-dimensional manifold structures. Specifically, we first construct a graph Laplacian for both miRNAs and diseases, which is determined by the functional similarity of miRNAs or the phenotypic similarity of diseases. We then present a formulation that joins the graph representation of miRNAs and diseases by considering the experimentally verified miRNA-disease associations, miRNAs' functional similarities and diseases' phenotypic similarities together. This formulation results in common low-dimensional embedding over the joined graph and provides an alignment of the underlying low-dimensional manifold structures (benefitting from the low-dimensional manifold structure's consistency between miRNAs and diseases). The resulting manifold alignment offers an inference of the miRNA-disease associations. Experimental results show that the proposed manifold alignment-based technique for association prediction between miRNAs and diseases outperforms several representative state-of-the-art approaches.

2. Related works

Grounded in known miRNA-disease associations, miRNA functional similarities, disease phenotypic affinities, or protein-disease relations, recent progressions of computational methods put forward effective algorithms for systematically predicting miRNAs related to given diseases and screening candidates for molecular biological experiments, in turn reducing expenditures and curtailing experiment cycles [14,26,27]. Existing methods for predicting the type of miRNA-disease associations are classified mainly into relation-learning methods and similarity-computation methods.

In general, relation-learning methods focus on learning miRNA-disease relations. Ala Qabaja et al. [28] proposed the Lasso regression model-based protein network to infer associations between miRNAs and diseases. Xu et al. [29] and Jiang et al. [7] applied the support vector machine (SVM) to classify proven miRNA-disease associations and negative ones. Chen et al. [30] first introduced a decision tree learning-based model (EGBMMDA) to infer miRNA-disease associations. The authors used a gradient boosting model to train the regression tree based on the results of miRNA-disease relations calculated by statistical measures, graph theoretical measures and matrix factorization. As is well known, this type of method has difficulty in collecting negative training samples, which refer to unknown miRNA-disease relationships. To overcome these limitations, Zou et al. [21] employed CATAPULT, treating all negative associations as unlabeled data to solve the lack of negative samples. Therein, semisupervised techniques were introduced to solve the issue of imbalanced and unlabeled samples to uncover potential microRNA-disease association. Chen et al. [31] proposed an RLSMDA method combining the semisupervised strategy and regularized least squares framework to identify the miRNA-disease associations that have unknown related miRNAs. Chen et al. [32] developed the machine learning method-based restricted Boltzmann machine named RBMMMDA to infer multiple types of relationships between miRNA and disease on a large scale. In addition, Chen et al. [33] introduced a bipartite network projection-based method (BNPMDA) for predicting miRNA-disease relations. However, this method cannot infer isolated diseases without known related miRNAs. To solve this problem, three similar methods based on known associations and integrated miRNA similarity and disease similarity were introduced. Chen et al. [34] proposed another semisupervised model based on low-rank inductive matrix completion (IMCMDA) to identify the missing associations between miRNAs and diseases via known associations and integrated similarity for miRNA and disease in order to infer miRNA-disease associations. Chen et al. [35] developed a novel computational model (MDHGI) for the prediction of miRNA-disease associations through matrix decomposition-based sparse learning. Combining multiple feature spaces to achieve a single classifier, Chen et al. [36] projected the feature profile of miRNAs or diseases into a common subspace and predicted relationships between miRNAs and diseases via the Laplacian regularized sparse subspace learning method (LRSSLMDA) to acquire reliable predictions.

The similarity-computation approaches can be roughly classified into local and global network-based similarity-computation approaches. Moreover, the strategies used include various methods, such as hypergeometric distribution, random walk, graph theory, path-based approaches, and social network analysis. According to whether they use only miRNA neighbor information, some local network-based similarity-computation measures are summarized as follows. Jiang et al. [5] first constructed a Boolean miRNA network containing nodes and edges to infer miRNA-disease associations. Xuan et al. [37] proposed the HDMP technique based on the weighted k most-similar neighbors, which overcame the shortcoming that many false positives are produced in the Boolean network. Mørk et al. [27] presented protein-driven inference (miRPD) for prediction based on miRNA-disease and protein-disease associations. In contrast, Chen et al. [38] proposed a global network similarity strategy to predict miRNA-disease associations using three inference methods, namely, microRNA-, phenotype-, and network consistency-based similarity inference. Similarly, Chen et al. [39] also exploited a global network similarity strategy called the RWRMDA, inspired by a random walk. Gu et al. [16] introduced a network consistency projection approach called NCPMDA to predict the potential miRNA-disease associations in all diseases without known negative samples. The better performance of the global approach indicated that it may obtain higher accuracy than approaches using the local network similarity model, which focuses only on neighboring relationships; however, this method ignores the information in proteins. Then, Shi et al. [40] developed the RWR algorithm employing protein-

protein interaction (PPI) and disease/miRNA-gene association data and a random walk strategy to acquire the probability distribution for genes of disease and miRNA. Furthermore, Chen et al. [41] developed HGIMDA by adopting heterogeneous graph inference to uncover potential miRNA-disease associations and predict new diseases and miRNAs without any prior associations. With respect to graph theory, You et al. [42] adopted the path-based miRNA-disease association (PBMDA) method combining miRNA-disease relationships, miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity information. Based on the social network, Zou et al. [21] explored the computational module KATZ, inspired by social network analysis, to predict associations. Le [43] introduced a ranking method based on a network strategy to predict novel miRNA-disease associations.

3. Materials and methods

3.1. Datasets

The phenotypic similarities between diseases were acquired in a similar way as those in Ref. [31]. The computation was carried out based on a hierarchical directed acyclic graph (DAG) constructed from the MeSH database (http://www.ncbi.nlm.nih.gov) with 4663 diseases. The nodes of the DAG represent a disease, and a link from a parent node to its children denotes their relations. The similarity between two diseases is decided based on the important assumption that two more similar diseases share larger parts of the DAG.

The functional similarities between miRNAs were obtained from the webpage (http://www.lirmed.com/misim/Download). These similarities were computed based on the assumption that miRNAs with more similar functions tend to be associated with more similar diseases, as detailed in Ref. [44].

The experimentally verified miRNA-disease associations were collected from the latest version of the HMDD [45]. We chose 18732 high-quality miRNA-disease associations that include 1206 miRNAs and 894 diseases.

3.2. Manifold learning from the similarity network

We use the square matrices S_r and S_d to denote the miRNA-miRNA similarities and disease-disease similarities, respectively, in which the number of rows/columns equals the number of miRNAs and diseases that we collected. Note that S_r/S_d has nonzero values only for the detected miRNA/disease pair when determining the functional/phenotypic similarities. In addition, S_r and S_d are symmetric and have non-negative values. We built a graph Laplacian for miRNAs with the following equation:

$$L_r(i,j) = \begin{cases} d_i, & \text{if } i = j \\ -S_r(i,j), & \text{otherwise} \end{cases} \tag{1}$$

where d_i represents the summation of the i_{th} row/column of S_r. Similarly, we can build a graph Laplacian L_d for diseases.

Building a real-valued function $\bar{r}$ for miRNAs, an optimal low-dimensional embedding of miRNA can then be obtained by minimizing the following quadratic cost:

$$\bar{r}^{\intercal} L_r \bar{r} = \frac{1}{2} \sum_{i,j} [S_r(i,j)(\bar{r}_i - \bar{r}_j)^2] \tag{2}$$

where $\bar{r}_i$ indicates the real value for the i_{th} miRNA. Similarly, we can also build a real-valued function $\bar{d}$ for diseases and obtain an optimal low-dimensional embedding of a disease by minimizing the following quadratic cost:

$$\bar{d}^{\intercal} L_d \bar{d} = \frac{1}{2} \sum_{i,j} [S_d(i,j)(\bar{d}_i - \bar{d}_j)^2] \tag{3}$$

3.3. Manifold alignment for inferring miRNA-Disease associations (MAMDA)

Most of the existing computational models for miRNA-disease association inference are based on the assumption that functionally similar miRNAs are associated with phenotypically similar diseases. From the viewpoint of manifold learning, this assumption also means that the low-dimensional manifold $\bar{r}$ has a structure similar to that of $\bar{d}$. We therefore infer the miRNA-disease association by aligning these two low-dimensional manifolds [25], which can be carried out via optimization of the following objective function:

$$O(\bar{r}, \bar{d}) = \mu \sum_{(i,j)\in A} (\bar{r}_i - \bar{d}_j)^2 + \bar{r}^{\intercal} L_r \bar{r} + \bar{d}^{\intercal} L_d \bar{d} \tag{4}$$

In the above equation, A represents all experimentally verified miRNA-disease associations. An element of A is an index pair (i, j), which indicates that the i_{th} miRNA is associated with the j_{th} disease. The parameter μ adjusts the importance of the experimentally verified miRNA-disease associations based on the fidelity of the resulting embedding to the one specified by the graph Laplacians.

The optimization of the above objective function can be carried out by minimizing its Rayleigh quotient [25], which amounts to solving the below minimization:

$$\min_{\bar{x}} \frac{\bar{x}^{\intercal} L \bar{x}}{\bar{x}^{\intercal} \bar{x}} s.\ t.\ \bar{x}^{\intercal} \bar{e} = 0 \tag{5}$$

where $\bar{x} = [\bar{r}^{\intercal} \bar{d}^{\intercal}]^{\intercal}$, $\bar{e} = [1\ 1 \cdots 1]^{\intercal}$ and L is specified as $L = \begin{bmatrix} L_r + U_r & -U_{rd} \\ -U_{rd} & L_d + U_d \end{bmatrix}$. Here, U_r and U_d have nonzero elements (with a value μ) on the diagonal specified by i and j in $(i, j) \in A$, respectively. U_{rd} has nonzero elements (with a value μ) specified by $(i, j) \in A$ with the i and j index row and column, respectively.

The t-dimensional manifold embedding is obtained using the eigenvectors of L corresponding to the t largest eigenvalues, which result in a t-dimensional representation of the aligned manifold $R = [\bar{r}^1\ \bar{r}^2 \cdots \bar{r}^t]$ and $D = [\bar{d}^1\ \bar{d}^2 \cdots \bar{d}^t]$. Within the t-dimensional embedding, the association $\omega(i, j)$ between the i_{th} miRNA and the j_{th} disease is computed with the following equation:

$$\omega(i,j) = \frac{1 - e^{-\psi(i,j)}}{1 + e^{-\psi(i,j)}} \tag{6}$$

where $\psi(i,j) = \frac{1}{t} \sum_{k=1\cdots t} (\bar{r}_i^k - \bar{d}_i^k)^2$.

Algorithm 1 MAMDA algorithm.

Input: Matrices S_r, S_d, and A; parameters r and d;
Output: Predicted miRNA-disease associations matrix ω;

```
for each i, j ∈ [1, r] do
    Build a graph Laplacian L_r(i, j) for miRNAs by Equation (1);
    Build a quadratic cost for miRNA-manifold embedding by Equation (2);
end for
for each i, j ∈ [1, d] do
    Build a graph Laplacian L_d(i, j) for diseases as in Equation (1);
    Build a quadratic cost for disease-manifold embedding by Equation (3);
end for
Obtain manifold embedding of miRNAs and diseases by Equations (4) and (5);
for each i ∈ [1, r], j ∈ [1, d] do
    Obtain the predicted associations ω(i, j) by Equation (6);
end for
return ω;
```

4. Results

4.1. Evaluation and comparison

We validated the inference performance of the proposed framework by using k-fold cross-validation and comparing it with state-of-the-art techniques, namely, HGIMDA [41], HDMP [37], RLSMDA [31], and

NCPMDA [16]. In the k-fold cross-validation, the 18732 experimentally verified miRNA-disease associations were randomly partitioned into k subsets. In these subsets, there are $(k-1)$ subsets, each of which bears $\lceil 18732/k \rceil$ ($\lceil \cdot \rceil$ denotes ceil operation) associations, and 1 subset containing the remaining associations. The cross-validation process is repeated k times, with each of the k subsets used exactly once as the validation data and all the remaining $(k-1)$ subsets used as the training data. The k results from all the folds were then averaged to produce a single estimate for this random partition. This random partition was repeated 50 times (set empirically), and all the resulting estimates were averaged to produce the final result. Moreover, we used the receiver operating characteristic (ROC) curve (as estimated by the ROCKIT algorithm [46]) and the area under the ROC curve (AUC) values to quantitatively assess performance.

Regarding the parameters in the computational models, we empirically set $\mu = 1.8$ in our model and the one(s) in other methods as the number(s) used in the corresponding literature. In addition, the computation and usage of the miRNA-miRNA similarities, disease-disease similarities and miRNA-disease associations were carried out as in the corresponding literature.

As shown by the AUC values in Fig. 1, all approaches degrade when k decreases in the k-fold cross-validation. In addition, our approach outperforms all other methods disregarding the value of k. The best AUC value generated by our approach is 0.922, which occurs when $k = 1000$.

In Fig. 2, we adopted the precision-recall curve (PR curve) and ROC curve to compare the prediction accuracy of the experimental methods, with corresponding AUC values generated by HGIMDA [41], HDMP [37], RLSMDA [31], and NCPMDA [16]. Analysis of the PR curves reveals that our method is superior to all other models, considering that its curve is closer to the upper-right corner of the graph. According to the ROC curves, which characterize the complete performance when the thresholds (in ω of our approach or the score of classification for the other methods) vary from low values to high ones, our approach achieved AUC values of 0.925, 0.897, and 0.898 for global LOOCV, local LOOCV and 20-fold cross-validation, respectively. HGIMDA [41], HDMP [37], RLSMDA [31], and NCPMDA [16] acquired AUC values of 0.899, 0.889, 0.909 and 0.912 for global LOOCV, respectively. Moreover, for local LOOCV, these approaches achieved values of 0.860, 0.857, 0.870 and 0.890, respectively. Therefore, our algorithm has an excellent effect on the accuracy of miRNA-disease relationship prediction.

4.2. Prediction of isolated diseases with unknown associated miRNAs

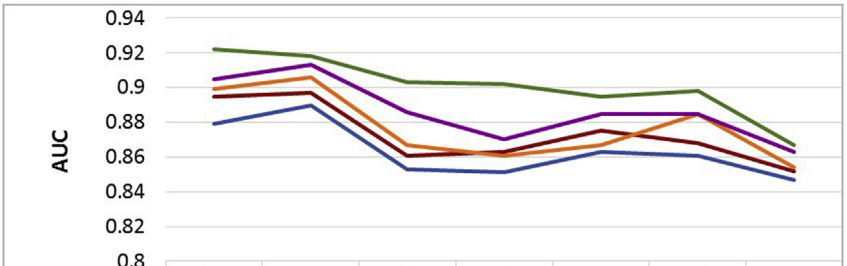
We also found that our approach can accurately predict diseases with no associated miRNAs. The only knowledge we acquired was disease-disease similarities and disease-miRNA similarities of other diseases. Specifically, we chose the disease “breast cancer” and excluded all related known miRNA-disease associations. This step guaranteed that only the experimentally verified associations that involved other diseases were included. We found that of the top 48 miRNAs inferred by our approach, all of them were consistent with the experimental verifications. The results shown in Table 1 indicate that MAMDA is valid for inferring associations with isolated diseases.

4.3. Case studies of colon neoplasms and lung neoplasms

Studies have shown that miRNA plays a dual role in oncogenes or tumor-suppressor genes, and their expression levels are closely related to tumor formation. In our work, we applied the MAMDA approach to predict the miRNA-disease associations for colon neoplasms and lung neoplasms. Herein, the predicted results were confirmed with the HMDD [45] and dbDEMC [47].

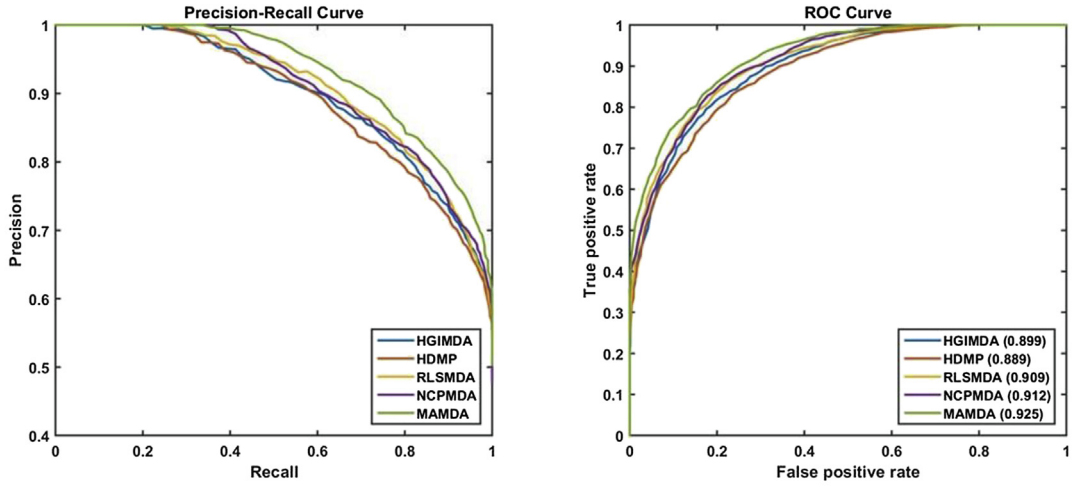
As the most common type of colorectal cancer, colon neoplasms are common malignant tumors in the gastrointestinal tract, and their incidence is second only to gastric and esophageal cancer. Recent studies have found that miRNAs are becoming increasingly significant in improving the detection of colon neoplasms; for instance, hsa-mir-126 is related to the growth of CN cells [48]. hsa-mir-186 may inhibit the proliferation, migration and invasion of colon cancer cells in vitro, so it plays a downregulatory role in colon cancer tissues [49], and [50] confirmed that the loss of mir-215 was associated with colon cancer. As another example, hsa-mir-140 inhibits the migration and invasion of colon cancer cells by downregulating Smad3 expression [51]. As shown in Table 2, we verified the accuracy of the MAMDA approach based on the disease. Specifically, 48 of the top 50 miRNA candidates were confirmed to be related to colon neoplasms. This performance shows that our method can predict colon neoplasms to some extent.

Lung neoplasms are one of the most common types of malignant tumors. In recent decades, the morbidity and mortality of lung tumors have increased significantly. Recent research has shown that miRNAs play a role in lung cancer development, epithelial-mesenchymal transition and therapeutic response [52]. For instance Ref. [53], demonstrated the inhibitory effect of hsa-let-7 on human lung cancer. hsa-mir-660 has been shown to be therapeutic in lung cancer cells [54]. In addition, hsa-mir-29 is closely related to gene expression in lung cancer [55]. The verification results can be viewed in Table 3, and 48 of the

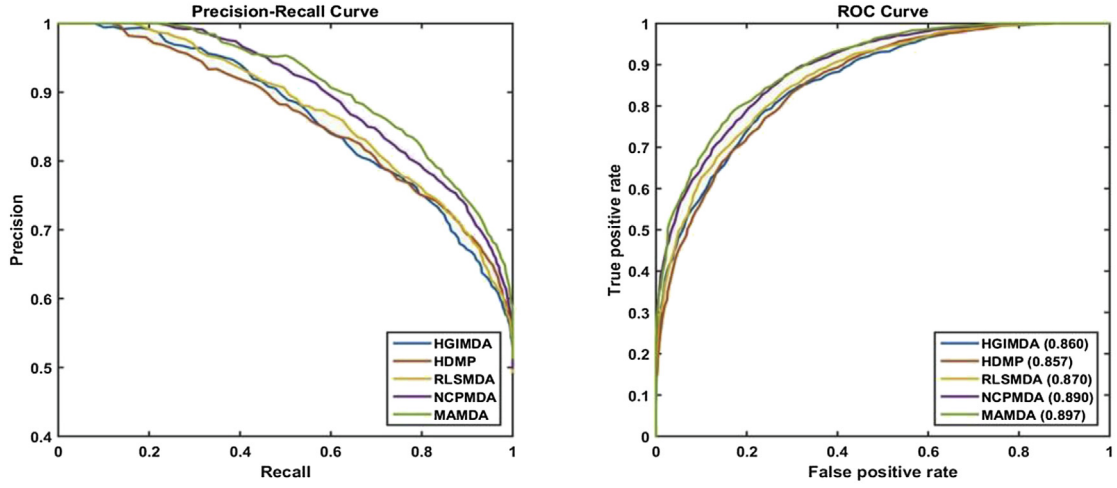


	1000	500	200	100	50	20	10
HGIMDA	0.879	0.89	0.853	0.851	0.863	0.861	0.847
HDMP	0.895	0.897	0.861	0.863	0.875	0.868	0.852
RLSMDA	0.899	0.906	0.867	0.861	0.867	0.885	0.854
NCPMDA	0.905	0.913	0.886	0.87	0.885	0.885	0.863
MAMDA	0.922	0.918	0.903	0.902	0.895	0.898	0.867

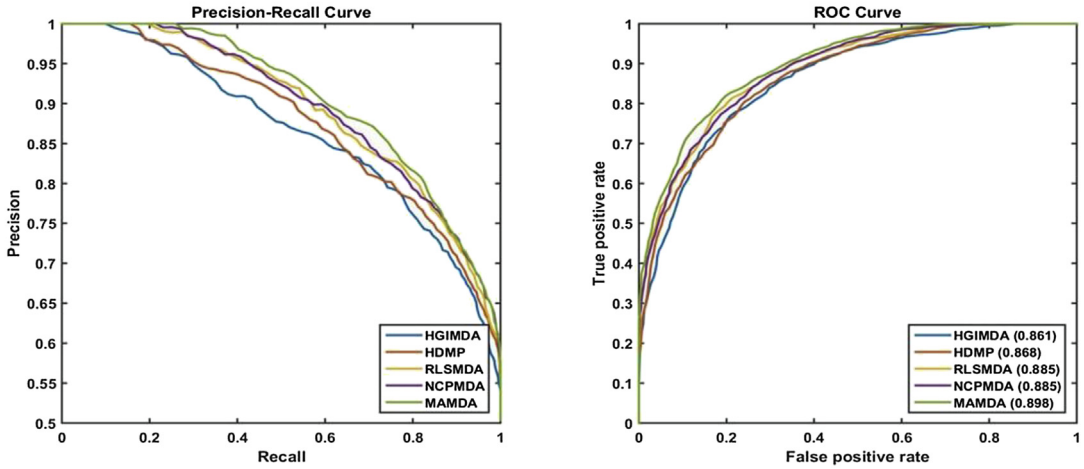
Fig. 1. Performance (measured with the AUC value) comparison among HGIMDA [41], HDMP [37], RLSMDA [31], NCPMDA [16] and MAMDA (our manifold alignment framework) when k (x-axis) takes different numbers in the k-fold cross-validation.



(a) Global LOOCV



(b) Local LOOCV



(b) 20-fold cross validation

Fig. 2. Precision-Recall (PR) curves and ROC curves with corresponding AUC values generated by HGIMDA [41], HDMP [37], RLSMDA [31], NCPMDA [16] and MAMDA (our manifold alignment framework).

Table 1
Prediction of isolated diseases with unknown miRNAs.

miRNA (1-24)	Evidence	miRNA (25-48)	Evidence
hsa-mir-15b	HMDD; dbDEMC	hsa-mir-92a	HMDD; dbDEMC
hsa-mir-19a	HMDD; dbDEMC	hsa-mir-145	HMDD; dbDEMC
hsa-mir-30e	HMDD; dbDEMC	hsa-mir-139	HMDD; dbDEMC
hsa-mir-103	HMDD; dbDEMC	hsa-mir-363	HMDD; dbDEMC
hsa-mir-200c	HMDD; dbDEMC	hsa-mir-153	HMDD; dbDEMC
hsa-mir-9	HMDD; dbDEMC	hsa-mir-621	HMDD; dbDEMC
hsa-mir-30c	HMDD; dbDEMC	hsa-let-7f-2	HMDD
hsa-let-7b	HMDD; dbDEMC	hsa-mir-25	HMDD; dbDEMC
hsa-mir-487b	dbDEMC	hsa-mir-93	HMDD; dbDEMC
hsa-mir-892a	HMDD; dbDEMC	hsa-mir-128-2	HMDD
hsa-mir-151	HMDD; dbDEMC	hsa-mir-190a	HMDD; dbDEMC
hsa-mir-124a-3	HMDD	hsa-mir-184	HMDD; dbDEMC
hsa-mir-513a-2	HMDD	hsa-mir-194-2	HMDD
hsa-mir-526b	HMDD; dbDEMC	hsa-mir-575	HMDD; dbDEMC
hsa-mir-573	HMDD; dbDEMC	hsa-mir-801	HMDD; dbDEMC
hsa-mir-203a	HMDD; dbDEMC	hsa-mir-19b	HMDD; dbDEMC
hsa-mir-191	HMDD; dbDEMC	hsa-mir-630	HMDD; dbDEMC
hsa-mir-132	HMDD; dbDEMC	hsa-mir-485	HMDD; dbDEMC
hsa-mir-411	HMDD; dbDEMC	hsa-let-7a-3	HMDD; dbDEMC
hsa-mir-328	HMDD; dbDEMC	hsa-mir-660	HMDD; dbDEMC
hsa-mir-20	HMDD; dbDEMC	hsa-mir-519c	HMDD; dbDEMC
hsa-mir-570	HMDD; dbDEMC	hsa-mir-563	dbDEMC
hsa-mir-506	HMDD; dbDEMC	hsa-mir-10	HMDD
hsa-mir-424	HMDD; dbDEMC	hsa-mir-520a	HMDD; dbDEMC

Table 2
The top 50 potential colon neoplasm-related miRNAs.

miRNA (1-25)	Evidence	miRNA (26-50)	Evidence
hsa-mir-328	HMDD; dbDEMC	hsa-mir-19b	HMDD; dbDEMC
hsa-mir-524	HMDD; dbDEMC	hsa-mir-208	unconfirmed
hsa-mir-1915	HMDD; dbDEMC	hsa-mir-143	HMDD; dbDEMC
hsa-mir-27a	HMDD; dbDEMC	hsa-mir-142	HMDD; dbDEMC
hsa-mir-135b	HMDD; dbDEMC	hsa-mir-34c	HMDD; dbDEMC
hsa-mir-3147	dbDEMC	hsa-mir-9	HMDD; dbDEMC
hsa-mir-30c-1	HMDD; dbDEMC	hsa-let-7d	HMDD; dbDEMC
hsa-mir-326	HMDD; dbDEMC	hsa-mir-18a	HMDD; dbDEMC
hsa-mir-106a	HMDD; dbDEMC	hsa-mir-222	HMDD; dbDEMC
hsa-mir-627	HMDD; dbDEMC	hsa-mir-126	HMDD; dbDEMC
hsa-mir-218-2	HMDD; dbDEMC	hsa-let-7a	HMDD; dbDEMC
hsa-mir-215	HMDD; dbDEMC	hsa-mir-34b	HMDD; dbDEMC
hsa-let-7f-1	HMDD; dbDEMC	hsa-mir-297	dbDEMC
hsa-mir-101-2	HMDD; dbDEMC	hsa-mir-191	HMDD; dbDEMC
hsa-mir-186	HMDD; dbDEMC	hsa-mir-223	HMDD; dbDEMC
hsa-mir-30c	HMDD; dbDEMC	hsa-mir-29b	HMDD; dbDEMC
hsa-mir-140	HMDD; dbDEMC	hsa-mir-205	HMDD; dbDEMC
hsa-mir-615	HMDD; dbDEMC	hsa-mir-16	HMDD
hsa-mir-370	dbDEMC	hsa-mir-200a	HMDD; dbDEMC
hsa-mir-92a	HMDD; dbDEMC	hsa-mir-29a	HMDD; dbDEMC
hsa-mir-886	HMDD; dbDEMC	hsa-mir-218	HMDD; dbDEMC
hsa-mir-363	HMDD; dbDEMC	hsa-mir-12	unconfirmed
hsa-mir-3613	HMDD; dbDEMC	hsa-let-7f	HMDD; dbDEMC
hsa-mir-551b	dbDEMC	hsa-mir-199a	HMDD
hsa-mir-101	HMDD; dbDEMC	hsa-mir-148a	HMDD; dbDEMC

Table 3
The top 50 potential lung neoplasm-related miRNAs.

miRNA (1-25)	Evidence	miRNA (26-50)	Evidence
hsa-let-7c	HMDD; dbDEMC	hsa-mir-183	HMDD; dbDEMC
hsa-let-132	HMDD; dbDEMC	hsa-mir-641	HMDD
hsa-mir-30c	HMDD; dbDEMC	hsa-mir-218	HMDD; dbDEMC
hsa-mir-185	HMDD; dbDEMC	hsa-mir-30c-2	HMDD; dbDEMC
hsa-mir-17	HMDD; dbDEMC	hsa-mir-320	HMDD; dbDEMC
hsa-mir-16	HMDD; dbDEMC	hsa-mir-7i	HMDD; dbDEMC
hsa-mir-18a	HMDD; dbDEMC	hsa-mir-328	HMDD
hsa-mir-146	HMDD; dbDEMC	hsa-mir-410	HMDD; dbDEMC
hsa-mir-660	HMDD; dbDEMC	hsa-mir-95	HMDD; dbDEMC
hsa-mir-25	HMDD; dbDEMC	hsa-mir-499b	HMDD
hsa-mir-1226	HMDD; dbDEMC	hsa-mir-124-2	HMDD
hsa-mir-144	HMDD; dbDEMC	hsa-mir-187	HMDD; dbDEMC
hsa-mir-216b	HMDD; dbDEMC	hsa-mir-107	HMDD; dbDEMC
hsa-mir-100	HMDD; dbDEMC	hsa-mir-206	HMDD; dbDEMC
hsa-mir-147b	dbDEMC	hsa-mir-200	HMDD
hsa-mir-1258	HMDD; dbDEMC	hsa-mir-138	HMDD; dbDEMC
hsa-mir-151a	unconfirmed	hsa-mir-126	HMDD; dbDEMC
hsa-mir-181c	HMDD; db DEMC	hsa-mir-638	HMDD; dbDEMC
hsa-mir-363	dbDEMC	hsa-mir-214	HMDD; dbDEMC
hsa-mir-1224	HMDD; dbDEMC	hsa-mir-743a	unconfirmed
hsa-mir-200b	HMDD; dbDEMC	hsa-mir-23a	HMDD; dbDEMC
hsa-mir-29	HMDD	hsa-mir-138-2	HMDD; dbDEMC
hsa-mir-608	HMDD	hsa-mir-335	HMDD; dbDEMC
hsa-mir-153	HMDD; dbDEMC	hsa-let-7	HMDD; dbDEMC
hsa-mir-630	HMDD; dbDEMC	hsa-mir-555	dbDEMC

top 50 miRNA candidates were confirmed to be associated with lung neoplasms. However, previous research found that hsa-mir-151a may be related to lung neoplasms [56]. Thus, our method can infer miRNA-disease relationships.

5. Discussion

Computational models for predicting miRNA-disease associations can address the drawbacks of related experimental verification techniques caused by their high cost and time consumption. These models benefit not only from the power of the inference techniques but also from various emerging resources, including but not limited to the experimentally verified miRNA-disease associations obtained via text mining, functional similarities between miRNAs and phenotypic similarities between diseases. Most of these computational models are built on the assumption that functionally similar miRNAs are associated with phenotypically similar diseases.

From the viewpoint of manifold learning, this basic assumption of current computational models for miRNA-disease inference is equivalent to the consistency between the underlying low-dimensional manifold structure of miRNAs and that of diseases. We hence proposed a novel computational framework for inferring miRNA-disease associations, which is distinguished from existing related models by its explicit leverage of the consistency between the low-dimensional manifold embedded in the space of miRNAs and that of diseases. In addition, this framework integrates the functional similarities of miRNA, phenotypic similarities of diseases and experimentally verified miRNA-disease associations.

We first constructed a graph Laplacian for not only miRNAs but also diseases. Then, we provided an objective function built for aligning the low-dimensional manifolds of miRNA and diseases. We finally offered a solution to minimize this objective function, which results in a practically invaluable measurement for assessing miRNA-disease associations.

Our experimental results were obtained by taking advantage of the existing resources of miRNA-miRNA similarities, disease-disease similarities and experimentally verified miRNA-disease associations. We compared the proposed model with a few representative state-of-the-art techniques using a series of *k*-fold cross-validations. We found that our approach outperforms other techniques not only for diseases when certain known miRNA-disease associations are provided but also when no associations are known. Success under the latter condition also indicates the capability of our approach to predict associated miRNAs for a biologically unknown disease.

The power of our approach can be attributed to several of its strengths. First, miRNA and disease manifolds are explicitly used. These manifolds are more reliable, especially when more miRNA-miRNA relations and disease-disease associations are verified. Second, structural consistency between the low-dimensional manifolds of miRNAs and diseases is explicitly leveraged, which is superior to the implicit use in some of the existing techniques. Third, the known miRNA-disease associations can be seamlessly embedded in the manifold alignment framework, which guides the matching result to comply with the known

associations. Finally, the solution to the proposed manifold alignment problem offers optimal optimization, resulting in better inference accuracies.

There are two main limitations of our work. First, this approach is sensitive to the number of data sources, which mainly affects the accuracy of manifold establishment. Future work will integrate more data into this method to improve the ability to predict miRNA-disease associations. Second, the approach in this paper relies strongly on miRNA-miRNA and disease-disease networks to construct manifolds. Therefore, future work should focus on the study of miRNA and disease networks.Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61572300; No. 81871508; No. 61773246; No. 61702313); Taishan Scholar Program of Shandong Province of China (No. TSHW201502038); Major Program of Shandong Province Natural Science Foundation (No. ZR2018ZB0419); and Natural Science Foundation of Shandong Province (No. ZR2016FQ20; No. ZR2017BC013).

Conflicts of interest

The authors declare that they do not have any potential conflict of interest.

References

[1] R.A. Espinoza-Lewis, D.Z. Wang, MicroRNAs in heart development, Curr. Top. Dev. Biol. 100 (2012) 279–317.
[2] D.P. Bartel, MicroRNAs: genomics, biogenesis, mechanism, and function, Cell 116 (2) (2004) 281–297.
[3] V.D.C. Martínez-Jiménez, A. Méndez-Mancilla, D.P. Portales-Pérez, miRNAs in nutrition, obesity, and cancer: the biology of miRNAs in metabolic disorders and its relationship with cancer development, Mol. Nutr. Food Res. 62 (1) (2018) 1600994.
[4] S.T. Aherne, S.F. Madden, D.J. Hughes, B. Pardini, A. Naccarati, M. Levy, et al., Circulating miRNAs miR-34a and miR-150 associated with colorectal cancer progression, BMC Canc. 15 (1) (2015) 329.
[5] Q. Jiang, Y. Hao, G. Wang, L. Juan, T. Zhang, M. Teng, et al., Prioritization of disease microRNAs through a human phenome-microRNAome network, BMC Syst. Biol. 4 (1) (2010) S2.
[6] G.A. Calin, C.M. Croce, MicroRNA signatures in human cancers, Nat. Rev. Canc. 6 (11) (2006) 857.
[7] Q. Jiang, G. Wang, S. Jin, Y. Li, Y. Wang, Predicting human microRNA-disease associations based on support vector machine, Int. J. Data Min. Bioinform. 8 (3) (2013) 282–293.
[8] J. Weidhaas, Using microRNAs to understand cancer biology, Lancet Oncol. 11 (2) (2010) 106–107.
[9] P. Jha, R. Agrawal, P. Pathak, A. Kumar, S. Purkait, S. Mallik, et al., Genome-wide small noncoding RNA profiling of pediatric high-grade gliomas reveals deregulation of several miRNAs, identifies downregulation of snoRNA cluster HBII-52 and delineates H3F3A and TP53 mutant-specific miRNAs and snoRNAs, Int. J. Cancer 137 (10) (2015) 2343–2353.
[10] X. Zhang, G. Lopez-Berestein, A.K. Sood, G.A. Calin, Profiling long noncoding RNA expression using custom-designed microarray, Methods Mol. Biol. 1402 (2016) 33.
[11] U. Jung, X. Jiang, S.H. Kaufmann, V. Patzel, A universal TaqMan-based RT-PCR protocol for cost-efficient detection of small noncoding RNA, RNA 19 (12) (2013) 1864–1873.
[12] O. Barad, E. Meiri, A. Avniel, R. Aharonov, A. Barzilai, I. Bentwich, et al., MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues, Genome Res. 14 (12) (2004) 2486–2494.
[13] Z. Yakhini, Cancer computational biology, BMC Bioinf. 12 (1) (2011) 120.
[14] Q. Zou, J. Li, L. Song, X. Zeng, G. Wang, Similarity computation strategies in the microRNA-disease network: a survey, Briefings in Functional Genomics 15 (1) (2015) 55–64.
[15] 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.
[16] C. Gu, B. Liao, X. Li, K. Li, Network consistency projection for human miRNA-disease associations inference, Sci. Rep. 6 (2016) 36054.
[17] W. Peng, W. Lan, J. Zhong, J. Wang, Y. Pan, A novel method of predicting microRNA-disease associations based on microRNA, disease, gene and environment factor networks, Methods 124 (2017) 69–77.
[18] Y. Liu, X. Zeng, Z. He, Q. Zou, Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources, IEEE ACM Trans. Comput. Biol. Bioinform 14 (4) (2017) 905–915.
[19] J. Luo, Q. Xiao, A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on heterogeneous network, J. Biomed. Inform. 66 (2017) 194–203.
[20] W. Peng, W. Lan, Z. Yu, J. Wang, Y. Pan, A Framework for integrating multiple biological networks to predict microRNA-disease associations, IEEE Trans. NanoBioscience 16 (2) (2017) 100–107.
[21] Q. Zou, J. Li, Q. Hong, Z. Lin, Y. Wu, H. Shi, et al., Prediction of microRNA-disease associations based on social network analysis methods, BioMed Res. Int. 2015 (10) (2015) 810514.
[22] 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.
[23] J.B. Tenenbaum, V.D. Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[24] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[25] J. Ham, D.D. Lee, L.K. Saul, Semisupervised alignment of manifolds, AISTATS, 2005, pp. 120–127.
[26] X. Chen, D. Xie, Q. Zhao, Z.H. You, MicroRNAs and complex diseases: from experimental results to computational models, Briefings Bioinf. 20 (2) (2019) 515–539.
[27] S. Mørk, S. Pletscher-Frankild, A. Palleja Caro, J. Gorodkin, L.J. Jensen, Protein-driven inference of miRNA–disease associations, Bioinformatics 30 (3) (2013) 392–397.
[28] A. Qabaja, M. Alshalalfa, T.A. Bismar, R. Alhajj, Protein network-based Lasso regression model for the construction of disease-miRNA functional interactions, EURASIP J. Bioinf. Syst. Biol. 2013 (1) (2013) 3–3.
[29] J. Xu, C.X. Li, J.Y. Lv, Y.S. Li, Y. Xiao, T.T. Shao, 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.
[30] X. Chen, L. Huang, D. Xie, Q. Zhao, EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction, Cell Death Dis. 9 (1) (2018) 3.
[31] X. Chen, G.Y. Yan, Semi-supervised learning for potential human microRNA-disease associations inference, Sci. Rep. 4 (2014) 5501.
[32] X. Chen, C.C. Yan, X. Zhang, Z. Li, L. Deng, Y. Zhang, et al., RBMMMDA: predicting multiple types of disease-microRNA associations, Sci. Rep. 5 (2015) 13877.
[33] X. Chen, D. Xie, L. Wang, Q. Zhao, Z.H. You, H. Liu, BNPMDA: bipartite network projection for MiRNA–disease association prediction, Bioinformatics 34 (18) (2018) 3178–3186.
[34] X. Chen, L. Wang, J. Qu, N.N. Guan, J. Li, Predicting miRNA-disease association based on inductive matrix completion, Bioinformatics 34 (24) (2018) 4256–4265.
[35] X. Chen, J. Yin, J. Qu, L. Huang, MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction, PLoS Comput. Biol. 14 (8) (2018) e1006418.
[36] X. Chen, L. Huang, LRSSLMDA: Laplacian regularized sparse subspace learning for MiRNA-disease association prediction, PLoS Comput. Biol. 13 (12) (2017) e1005912.
[37] P. Xuan, K. Han, M. Guo, Y. Guo, J. Li, J. Ding, et al., Prediction of microRNAs associated with human diseases based on weighted *k* most similar neighbors, PLoS One 8 (9) (2013) e70204.
[38] H. Chen, Z. Zhang, Similarity based methods for potential human microRNA-disease association prediction, BMC Med. Genomics 6 (1) (2013) 1–9.
[39] X. Chen, M.X. Liu, G.Y. Yan, RWRMDA: predicting novel human microRNA-disease associations, Mol. Biosyst. 8 (10) (2012) 2792–2798.
[40] H. Shi, J. Xu, G. Zhang, L. Xu, C. Li, L. Wang, et al., 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.
[41] X. Chen, Y.C. Clarence, X. Zhang, Z.H. You, Y.A. Huang, G.Y. Yan, HGIMDA: heterogeneous graph inference for miRNA-disease association prediction, Oncotarget 7 (40) (2016) 65257–65269.
[42] Z.H. You, Z.A. Huang, Z. Zhu, G.Y. Yan, Z.W. Li, Z. Wen, et al., PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction, PLoS Comput. Biol. 13 (3) (2017) e1005455.
[43] D.H. Le, Network-based ranking methods for prediction of novel disease associated microRNAs, Comput. Biol. Chem. 58 (2015) 139–148.
[44] J. Li, S. Zhang, Y. Wan, Y. Zhao, J. Shi, Y. Zhou, et al., MISIM v2. 0: a web server for inferring microRNA functional similarity based on microRNA-disease associations, Nucleic Acids Res. (2019), https://doi.org/10.1093/nar/gkz328 PMID: 31069374.
[45] Z. Huang, J. Shi, Y. Gao, C. Cui, S. Zhang, J. Li, et al., HMDD v3. 0: a database for experimentally supported human microRNA–disease associations, Nucleic Acids Res. 47 (1) (2018) 1013–1017.
[46] C.E. Metz, B.A. Herman, J.H. Shen, Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data, Stat. Med. 17 (9) (1998) 1033–1053.
[47] Z. Yang, F. Ren, C. Liu, S. He, G. Sun, Q. Gao, et al., dbDEMC: a database of differentially expressed miRNAs in human cancers, BMC Genomics 11 (Suppl 4) (2010) 1–8.
[48] J. Yin, Z. Bai, J. Song, Y. Yang, J. Wang, W. Han, et al., Differential expression of serum miR-126, miR-141 and miR-21 as novel biomarkers for early detection of liver metastasis in colorectal cancer, Chin. J. Canc. Res. 26 (1) (2014) 95.
[49] F. Chen, C. Zhou, Y. Lu, L. Yuan, F. Peng, L. Zheng, et al., Expression of hsa-miR-186 and its role in human colon carcinoma cells, J. South. Med. Univ. 33 (5) (2013) 654–660.
[50] Y. Sun, S. Koo, N. White, E. Peralta, C. Esau, N.M. Dean, et al., Development of a micro-array to detect human and mouse microRNAs and characterization of expression in human organs, Nucleic Acids Res. 32 (22) (2004) e188–e188.
[51] H. Pais, F.E. Nicolas, S.M. Soond, T.E. Swingler, I.M. Clark, A. Chantry, et al., Analyzing mRNA expression identifies Smad3 as a microRNA-140 target regulated only at protein level, RNA 16 (3) (2010) 489–494.
[52] P. Lin, S. Yu, P. Yang, MicroRNA in lung cancer, Br. J. Canc. 103 (8) (2010) 1144.

[53] J. Takamizawa, H. Konishi, K. Yanagisawa, S. Tomida, H. Osada, H. Endoh, et al., Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival, Cancer Res. 64 (11) (2004) 3753–3756.
[54] O. Fortunato, C. Verri, U. Pastorino, G. Sozzi, M. Boeri, MicroRNA profile of lung tumor tissues is associated with a high risk plasma miRNA signature, Microarrays 5 (3) (2016) 18.
[55] K. Piletič, T. Kunej, MicroRNA epigenetic signatures in human disease, Arch. Toxicol. 90 (10) (2016) 2405–2419.
[56] I. Daugaard, K.J. Sanders, A. Idica, K. Vittayarukskul, M. Hamdorf, J.D. Krog, et al., miR-151a induces partial EMT by regulating E-cadherin in NSCLC cells, Oncogenesis 6 (7) (2017) e366.

Fang Yan received a B.A. degree in Computer Science from Shandong Normal University, Jinan, Shandong, China, in 2016. She is currently a Ph.D. student at Shandong Normal University, Jinan, Shandong, China. Her research interests include biological information processing, image processing and machine vision.

Yuanjie Zheng received the Ph.D. degree in pattern recognition and intelligent systems from Shanghai Jiao Tong University, Shanghai, China in 2006. He is currently a professor at Shandong Normal University, Jinan, Shandong, China. He used to be a senior research investigator in the Perelman School of Medicine at the University of Pennsylvania and the primary contact of the Image Analysis Core at the Penn Vision Research Center. His research is in the fields of medical image analysis, translational medicine, computer vision, computational photography and information biology.

Weikuan Jia received the Ph.D. degree from Jiangsu University, Zhenjiang, Jiangsu, China, in 2016. He is currently a lecture at Shandong Normal University, Jinan, Shandong, China. His research interests include Artificial Intelligence, Intelligent Agriculture, Agricultural Information Technology and Equipment.

Sujuan Hou received her Ph.D. degree from Chongqing University, Chongqing, China, in 2015. Currently, she is an associate professor with the School of Information Science and Engineering, Shandong Normal University, Jinan, Shandong, China. Prior to that, she was a visitor with Faculty of Engineering and Information Technology (FEIT), University of Technology, Sydney (UTS) from 2013 to 2015. Her research interests include video representation, multimedia analysis and pattern recognition.

Rui Xiao received the PhD degree in biostatistics from the University of Michigan, Ann Arbor, in 2010. She is currently an assistant professor in the Department of Biostatistics and Epidemiology at the University of Pennsylvania School of Medicine. Her primary research interests are in the design, implementation, and analysis of neuroimaging related studies and high throughput genetic studies.