Molecular Therapy
Nucleic Acids
Original Article

DBMDA: A Unified Embedding for Sequence-Based miRNA Similarity Measure with Applications to Predict and Validate miRNA-Disease Associations

Kai Zheng,[1,4] Zhu-Hong You,[2,4] Lei Wang,[2,3] Yong Zhou,[1] Li-Ping Li,[2] and Zheng-Wei Li[1]

[1]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China; [2]Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China; [3]College of Information Science and Engineering, Zaozhuang University, Zaozhuang 277100, China

MicroRNAs (miRNAs) play a critical role in human diseases. Determining the association between miRNAs and disease contributes to elucidating the pathogenesis of liver diseases and seeking the effective treatment method. Despite great recent advances in the field of the associations between miRNAs and diseases, implementing association verification and recognition efficiently at scale presents serious challenges to biological experimental approaches. Thus, computational methods for predicting miRNA-disease association have become a research hotspot. In this paper, we present a new computational method, named distance-based sequence similarity for miRNA-disease association prediction (DBMDA), that directly learns a mapping from miRNA sequence to a Euclidean space. The notable feature of our approach consists of inferring global similarity from region distances that can be figured by chaos game representation algorithm based on the miRNA sequences. In the 5-fold cross-validation experiment, the area under the curve (AUC) obtained by DBMDA in predicting potential miRNA-disease associations reached 0.9129. To assess the effectiveness of DBMDA more effectively, we compared it with different classifiers and former prediction models. Besides, we constructed two case studies for prostate neoplasms and colon neoplasms. Results show that 39 and 39 out of the top 40 predicted miRNAs were confirmed by other databases, respectively. BDMDA has made new attempts in sequence similarity and achieved excellent results, while at the same time providing a new perspective for predicting the relationship between diseases and miRNAs. The source code and datasets explored in this work are available online from the University of Chinese Academy of Sciences (http://220.171.34.3:81/).

INTRODUCTION

MicroRNA (miRNA) is a short group of noncoding RNA (ncRNA) constructed from about 22 nt that can combine designated messenger RNA by base pairing and control the translation and stability.[1] Since the first miRNA was discovered by Victor Ambros in 1993, a large number of found miRNAs accumulated at a high level during the past 20 years from a far-ranging variety of species.[2,3] The study found that miRNA plays an important influence on biological processes, such as cell development, proliferation, and apoptosis,[4] and the regulation functions of miRNA are related to some particular gene expressions in the post-transcriptional stage.[5] Based on the above findings, more and more miRNAs have been validated in connection with the development of complex diseases in humans.[6] For instance, miR-137 controlled the mitotic progression of lung cancer cells by targeting Cdc42 and Cdk6.[7] In von Brandenstein et al.'s[8] study, miR-15a is a potential biological marker for differentiating benign and malignant renal tumors in biopsy and urine samples. The progression of head and neck carcinomas could also be boosted by miR-211 through combining transforming growth factor-β receptor 2 (TGF-βR2).[9] However, the biological experimental conditions for verifying the association between miRNA and disease are harsh and have time-consuming and laborious disadvantages. Therefore, the computational algorithms for forecasting the potential miRNA-disease associations have become a hot topic, and more studies attach importance to it. Correspondingly, computational methods can more effectively assist biological experiments to validate disease-associated miRNAs by predicting results.[10]

Over the years, an increasing number of studies constructed computational models for predicting miRNA-disease association.[11–17] There are two main types of computational models based on similarity and based on machine learning. To be specific, methods based on similarity figure the correlation intension through the miRNA and disease network. For example, Chen et al.[18] proposed Random Walk with Restart for MiRNA-Disease Association prediction (RWRMDA) is a method for calculating global network similarity by combining matrices of miRNA functional similarity. Li et al.[19] presented a computational method to predict potential associations by calculating

Received 8 July 2019; accepted 10 December 2019;
https://doi.org/10.1016/j.omtn.2019.12.010.

[4]These authors contributed equally to this work.

Correspondence: Zhu-Hong You, Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China.
E-mail: zhuhongyou@ms.xjb.ac.cn
Correspondence: Lei Wang, Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China.
E-mail: leiwang@ms.xjb.ac.cn
Correspondence: Kai Zheng, School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China.
E-mail: zhengkai951211@gmail.com

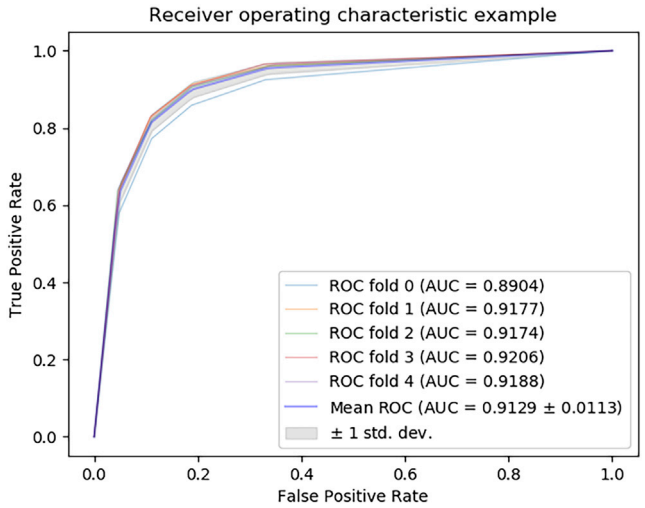

Figure 1. The ROCs of DBMDA and AUCs Based on 5-Fold Cross-Validation

functional consistency score (FCS) of target genes and disease-related genes. The main progress of heterogeneous graph inference for miRNA-disease association prediction (HGIMDA) was to calculate the optimal solution set through an iterative process, given by Chen et al.[20] On the other hand, the methods based on machine learning predict the potential miRNA-disease association by using the known miRNA-disease association training model.[21] For example, Xu et al.[22] used a support vector machine (SVM) classifier to identify positive and negative associations in a miRNA-target-dysregulated network. Chen and Yan[23] proposed a method for predicting new disease-related RNA without negative correlation named Regularized Least-squares for MiRNA-Disease Association according to semi-supervised learning. Restricted Boltzmann machine for multiple types of miRNA-disease association prediction (RBMMMDA) is a method developed by Chen et al.,[24] whose main improvement is the acquisition of several types of new associations.

In this study, we build a distance-based sequence similarity for miRNA-disease association prediction (DBMDA) based on chaos game representation (CGR). DBMDA combines the information of miRNA sequence, miRNA function, confirmed association, and disease semantic. The motivation for this approach is to map miRNA sequences to Euclidean space, where the regional distance directly corresponds to a measure of miRNA sequence similarity. In detail, we first obtained miRNA and disease similarity matrices based on miRNA sequence information and disease semantic information. Second, the similarity matrices obtained in the previous step are combined with the Gaussian profile kernel similarity matrices of miRNA and disease to get the integrated similarity matrices. Third, each nucleotide directly learns the mapping from miRNA sequences to Euclidean space through CGR techniques. To be specific, the CGR plane is divided into 8×8 grids, and the average coordinates of each grid are calculated. Also, the regional distance between miRNAs is used to quantify the similarity of miRNA functions to construct a miRNA sequence similarity matrix and integrate the similar information obtained in the second step into a comprehensive feature. Finally, the integrated feature vector is placed in the rotation forest (RoF) classifier to predict the potential association. The following experiments have been designed to evaluate the reliability of the method. We use the 5-fold cross-validation to assess the performance of DBMDA in the Human microRNA Disease Database (HMDD) v.3.0 dataset. The AUC of 5-fold cross-validation was 0.9129 ± 0.0113 in result. Moreover, two case studies on prostate neoplasms and colon neoplasms have been applied. As a result, 39 (prostate tumors) and 39 (colon tumors) of the top 40 predicted miRNAs, respectively, were verified by other datasets. It shows that DBMDA is an efficient predicting potential miRNA-disease associations method.

RESULTS

Performance Evaluation

Evaluation Criteria

We follow the widely used evaluation measure by means of classification accuracy (Accu.), sensitivity (Sen.), precision (Prec.), and F1 score to assess the performance of DBMDA as defined, respectively, by:

$$Accu. = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Sen. = \frac{TP}{TP + FN} \quad (2)$$

$$Prec. = \frac{TP}{TP + FP} \quad (3)$$

$$F_1 = \frac{Prec. \times Sen.}{Prec. + Sen.}, \quad (4)$$

where TP, FP, TN, and FN represent the true positive, false positive, true negative, and false negative, respectively. In addition, the receiver operating characteristic (ROC) curve and the area under the curve (AUC) can be used to show the performance of the model generally.

Prediction of miRNA-Disease Association

We have used the 5-fold cross-validation to assess the performance of DBMDA based on confirmed associations in HMDD v.3.0.[25] Li et al.[25] selected 17,412 papers and extracted 32,281 known miRNA-disease associations constructed by 1,102 miRNAs and 850 diseases. Because some information of miRNA cannot be judged by the public database miRBase, we have removed it. After screening, the associations confirmed by miRbase have been chosen as positive samples.[26] Meanwhile, negative samples are constructed by possible miRNA-disease association pairs from all possible miRNA-disease pairs.

Figure 1 lists the performance of the 5-fold cross-validation obtained by DBMDA. We can see from the table that DBMDA has gained an average prediction AUC of 0.9129 ± 0.0113. The AUC of the five

Table 1. The Comparison Results of DBMDA Based on 5-Fold Cross-Validation

Testing Set	Accuracy	Sensitivity	Precision	F1-Score
1	83.14%	81.55%	84.23%	82.87%
2	86.21%	86.83%	85.77%	86.30%
3	85.57%	86.42%	84.99%	85.70%
4	86.22%	87.07%	85.63%	86.34%
5	85.66%	86.83%	84.85%	85.83%
Average	85.36% ± 1.27%	85.74% ± 2.35%	85.09% ± 0.62%	85.40% ± 1.44%

experiments is 0.8904 (fold 1), 0.9177 (fold 2), 0.9174 (fold 3), 0.9206 (fold 4), and 0.9188 (fold 5), respectively. The yielded averages of accuracy, sensitivity, precision, and F1-score come to be 85.36%, 85.74%, 85.09%, and 85.40% as in Table 1.

Comparison with Different Classifier Models

In the 5-fold cross-validation, our proposed method achieved good results in the HMDD v.3.0 dataset using the RoF classifier. The RoF as part of the proposed method was compared with SVM, random forest (RF), and decision tree (DT) in this experiment to illustrate why it was chosen. The accuracies of the four experiments are 85.00% (RoF), 83.73% (SVM), 82.06% (RF), and 80.33% (Decision Tree). Their AUCs are 91.15% (RoF), 89.01% (SVM), 90.77% (RF), and 80.29% (Decision Tree), which are shown in Figure 2. The accuracy, sensitivity, precision, and F1-score have been shown in Table 2. From the experimental results, the performance of the rotating forest classifier in terms of sensitivity is not the highest among the four classifiers However, the best results were obtained in other evaluation criteria, especially the AUC that represents the overall performance of the model. In general, rotating forests is the best classifier for the features we build.

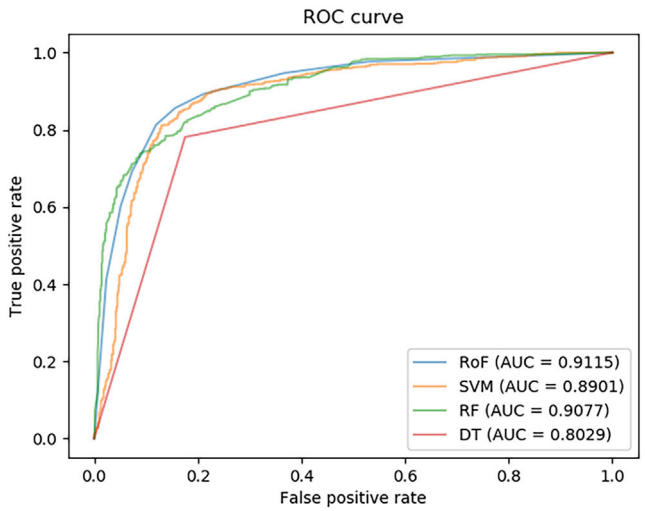

Figure 2. The ROCs of Four Different Classifiers, which Are RoF, SVM, Random Forest, and Decision Tree

Comparison with Related Methods

Many studies in the past have explored the field of the associations between miRNAs and diseases. To evaluate performance, we compared it with eight state-of-the-art methods. Because the database versions used are not the same, we compare only the AUC values reported in the article. Compared with the AUC of RLSMDA, PBSI, MBSI, NetCBI, MaxFlow, miRGOFS, HGIMDA, MDHGI, and LMTRDA, DBMDA performs better, as shown in Table 3.[20,23,27–31] There are manifold reasons why DBMDA is more outstanding than traditional miRNA similarity. First, the sequence information of miRNAs contains attribute features and is an excellent source of knowledge reflecting essential information. Second, the miRNA similarity obtained based on limited knowledge resources may have errors caused by information loss. Third, our approach inferring global similarity from regional distances also helps improve performance.

Case Studies

Here DBMDA will be applied to two kinds of human diseases, including prostate neoplasms and colon neoplasms. It further evaluates the effectiveness of DBMDA based on the associations identified in the HMDD v.3.0 database. The test samples are miRNA-disease associations consisting of two diseases and all possible miRNAs. We confirmed prediction results with top 40 ranks in dbDEMC v.2.0[32] and dbDEMC v.2.0.[33]

In the United States, prostate cancer has caused more than 20,000 deaths and has become one of the hidden dangers of men's health today. Age is a major cause of prostate cancer, and older people may have a higher rate. However, an increasing number of younger men were diagnosed with prostate neoplasms. Prostate neoplasms may pass to other areas of the human body, such as surrounding tissue like regional lymph nodes. Therefore, we took prostate neoplasms as an example to evaluate the performance of DBMDA. The results are shown in Table 4. Thirty-nine of the top 40 predicted miRNAs were identified by the two datasets mentioned above.

In the United States, colon neoplasms have the third highest morbidity and third highest fatality rate, which is defined as a type of common malignant cancer. A study showed that more than 135,000 individuals would be diagnosed with colon neoplasms and rectum neoplasms. Therefore, we chose colon neoplasms as a case study to evaluate the performance of DBMDA. As a result, 39 of the top 40 potential miRNAs that associate with colon neoplasms were confirmed by experimental findings recorded in dbDEMC v.2.0 and miR2Disease as shown in Table 5.

DISCUSSION

Sequence-based miRNA similarity can aid in predicting miRNA-disease associations, extract biological property information, and enhance the analytical quality of high-throughput sequencing data. However, most existing methods do not involve sequence information, and according to current information sources (miRNA-disease association), the relationship between miRNAs is not directly reflected. Therefore, this paper proposed a predictive model

Table 2. Performance Comparison among Four Different Classifiers, which are Rotation Forest, SVM, Random Forest, and Decision Tree

Method	Accuracy	Sensitivity	Precision	F1-Score
SVM	83.73%	83.56%	83.33%	83.45%
RF	82.06%	76.49%	85.43%	80.72%
DT	80.33%	78.12%	81.10%	79.58%
RoF	85.00%	85.60%	84.11%	84.85%

Table 3. The Comparison with Related Models

Methods	AUC Scores
RLSMDA[a]	86.17%
PBSI[b]	54.02%
MBSI[b]	74.83%
NetCBI[b]	80.66%
MaxFlow[c]	86.93%
miRGOFS[d]	87.70%
HGIMDA[e]	87.81%
MDHGI[f]	87.94%
LMTRDA[g]	90.54%
DBMDA	91.29%

[a]The results of the method are reported in Chen and Yan.[23]
[b]The results of the method are reported in Chen and Zhang.[27]
[c]The results of the method are reported in Yu et al.[28]
[d]The results of the method are reported in Yang et al.[30]
[e]The results of the method are reported in Chen et al.[20]
[f]The results of the method are reported in Chen et al.[31]
[g]The results of the method are reported in Wang et al.[44]

for inferring miRNA similarity based on sequence information, called DBMDA. The improvement of the method was to directly learn the mapping from miRNA sequence to Euclidean space. In Euclidean space, the regional distance directly corresponds to the measure of miRNA sequence similarity. Excellent experimental results indicate that DBMDA had performed well in predicting disease-associated miRNAs with the support of new algorithms and sequence information. In addition, sequence information has sufficient coverage for human miRNAs, and DBMDA is universal in functional analysis.

Conclusions

Sequence-based miRNA similarity can aid in predicting miRNA-disease associations, extract biological property information, and enhance the analytical quality of high-throughput sequencing data. However, most existing methods do not involve sequence information, and according to current information sources (miRNA-disease association), the relationship between miRNAs is not directly reflected. Therefore, this paper proposed a predictive model for inferring miRNA similarity based on sequence information, called DBMDA. The improvement of the method was to directly learn the mapping from miRNA sequence to Euclidean space. In Euclidean space, the regional distance directly corresponds to the measure of miRNA sequence similarity. Excellent experimental results indicate that DBMDA had performed well in predicting disease-associated miRNAs with the support of new algorithms and sequence information. In addition, sequence information has sufficient coverage for human miRNAs, and DBMDA is universal in functional analysis.

MATERIALS AND METHODS

Human miRNA-Disease Associations

We downloaded the confirmed associations data from the HMDD dataset in this experiment.[25] The last update of HMDD v.3.0 was October 9, 2018, which includes 32,281 experimentally known associations about 850 diseases and 1,102 miRNAs from 17,412 papers. Based on it, an adjacency matrix $X \in R^{nM \times nD}$ is built to reshape the associations, where nD and nM are the number of the diseases and miRNAs in HMDD v.3.0. X_{ij} is equal to 1 if miRNA m_i had been confirmed to associate with a disease d_j, otherwise equal to 0.[34]

miRNA Functional Similarity

Wang et al.[35] proposed a method for quantifying miRNA functional similarity between miRNAs based on the hypothesis that functionally similar miRNAs are more likely to affect the same disease and pathologically similar diseases are more likely to be affected by the same miRNA. The miRNA function information is uploaded to http://www.cuilab.cn/files/images/cuilab/misim.zip. A 495 rows × 495 columns matrix, $MF(m_a, m_b)$, can be defined to represent the miRNA functional similarity, and the element is the similarity score between the miRNA m_a and the miRNA m_b.

Disease Semantic Similarity Model

We built a directed acyclic graph (DAG) to define the relationship among diseases based on the method proposed by Wang et al.,[35] which is according to the Medical Subject Headings (MeSH) descriptors.[36] The MeSH descriptors can be downloaded from the U.S. National Library of Medicine database (https://www.nlm.nih.gov/). The disease d_i can be defined as $DAG_{d_i} = D, N_{d_i}, E_{d_i}$, where N_{d_i} is a node set including the information of disease d_i and its ancestor diseases, and E_{d_i} is an edge set including the information of the corresponding edges. Based on the DAG, the contribution values of disease o in DAG_{d_i} to the semantic value of disease d_i was calculated as:

$$\begin{cases} D_{d_i}(o) = 1 & if\ o = d_i \\ D_{d_i}(o) = max\left\{\partial * D_{d_i}(o') \middle| o' \in children\ of\ o\right\} & if\ o \neq d_i \end{cases}, \quad (5)$$

where the semantic contribution decay factor is ∂, which is set to 0.5 according to previous studies.[29] Furthermore, if disease o is not disease d_i, it will decrease the contribution of disease o. If disease o is disease d_i, the contribution of disease d_i is defined as 1. Besides, we described the semantic value $DV(d)$ as follows:

Table 4. Prediction of the Top 40 Predicted miRNAs Associated with Prostate Neoplasms Based on Known Associations in dbDEMC v.2.0 and miR2Database

miRNA	dbDEMC	miR2D	miRNA	dbDEMC	miR2D
hsa-mir-192	confirmed	unconfirmed	hsa-mir-181a-2	confirmed	unconfirmed
hsa-let-7i	confirmed	unconfirmed	hsa-mir-196a	confirmed	unconfirmed
hsa-mir-140	confirmed	unconfirmed	hsa-mir-208a	confirmed	unconfirmed
hsa-mir-199b	confirmed	confirmed	hsa-mir-337	confirmed	unconfirmed
hsa-mir-144	confirmed	unconfirmed	hsa-mir-1246	confirmed	unconfirmed
hsa-mir-372	confirmed	unconfirmed	hsa-mir-30	confirmed	unconfirmed
hsa-let-7e	confirmed	confirmed	hsa-mir-184	confirmed	confirmed
hsa-let-7f	confirmed	confirmed	hsa-mir-509	unconfirmed	unconfirmed
hsa-mir-10b	confirmed	confirmed	hsa-mir-9-3	confirmed	unconfirmed
hsa-mir-129	confirmed	unconfirmed	hsa-let-7f-2	confirmed	unconfirmed
hsa-mir-9-1	confirmed	unconfirmed	hsa-mir-202	confirmed	confirmed
hsa-mir-206	confirmed	unconfirmed	hsa-mir-33a	confirmed	unconfirmed
hsa-mir-125a	confirmed	confirmed	hsa-mir-451a	confirmed	unconfirmed
hsa-mir-30b	confirmed	confirmed	hsa-let-7f-1	confirmed	unconfirmed
hsa-mir-362	confirmed	unconfirmed	hsa-mir-186	confirmed	unconfirmed
hsa-mir-133	confirmed	unconfirmed	hsa-mir-302b	confirmed	unconfirmed
hsa-mir-139	confirmed	unconfirmed	hsa-mir-328	confirmed	unconfirmed
hsa-mir-137	confirmed	unconfirmed	hsa-mir-383	confirmed	unconfirmed
hsa-mir-181b-2	confirmed	unconfirmed	hsa-mir-431	confirmed	unconfirmed
hsa-mir-338	confirmed	unconfirmed	hsa-mir-103a-2	confirmed	unconfirmed

$$DV(d_i) = \sum_{t \in N_{d_i}} D_{d_i}(o). \quad (6)$$

If disease d_i and d_j have more shared segments of their DAGs, they will have a larger similarity score. The semantic similarity score could be defined as follows:

$$Sim(d_i, d_j) = \frac{\sum_{t \in N_{d_i} \cap N_{d_j}} \left(D_{d_i}(o) + D_{d_j}(o)\right)}{DV(d_i) + DV(d_j)}. \quad (7)$$

The *Sim* is defined as the 850 rows and 850 columns semantic similarity matrix, and the element $Sim(d_i, d_j)$ is the semantic similarity of d_i and d_j based on disease semantic similarity model 1.

According to the above formula, diseases in the same layer in *DAGs* will have the same contribution value. However, a higher value should be contributed by a definite disease that appears in fewer *DAGs*. Hence the contribution of disease o in $DAG(d)$ to the semantic value of disease d is described based on the method built by Xuan et al.[29] as follows:

$$D'_{d_i}(o) = -log\left(\frac{number\ of\ DAGs\ including\ t}{number\ of\ disease}\right), \quad (8)$$

where o is a disease of all the diseases in our method. Also, the semantic similarity between disease d_i and d_j is described as sim', which is based on the shared ancestor nodes and all the ancestor nodes. To be specific, the disease semantic similarity can be computed as follows:

$$Sim'(d_i, d_j) = \frac{\sum_{o \in E_{d_i} \cap E_{d_j}} \left(D'_{d_i}(o) + D'_{d_j}(o)\right)}{DV(d_i) + DV(d_j)}, \quad (9)$$

where $DV(d_i)$ and $DV(d_j)$ are the semantic score of d_i and d_j, and can be computed the same as for Equation 2.

GIP Similarity for Diseases and miRNA

The HMDD v.3.0 dataset provides plenty of correlation information.[37] Based on the hypothesis that the pathologically similar disease may be affected by the same miRNA and vice versa, we calculate the disease and miRNA similarity by Gaussian interaction profile kernel (GIP) similarity. The binary vector $V(d_i)$ is the i-th row vector of adjacency matrix X. The disease GIP similarity $GD(d_i, d_j)$ between d_i and d_j was computed by:

$$GD(d_i, d_j) = exp\left(-\gamma_d * \|V(d_i) - V(d_j)\|^2\right), \quad (10)$$

Table 5. Prediction of the Top 40 Predicted miRNAs Associated with Colon Neoplasms Based on Known Associations in dbDEMC v.2.0 and miR2Database

miRNA	dbDEMC	miR2D	miRNA	dbDEMC	miR2D
hsa-mir-26a	confirmed	confirmed	hsa-mir-497	confirmed	confirmed
hsa-mir-182	confirmed	confirmed	hsa-mir-92a-2	confirmed	unconfirmed
hsa-mir-342	confirmed	confirmed	hsa-mir-124	confirmed	confirmed
hsa-mir-483	confirmed	unconfirmed	hsa-mir-129	confirmed	confirmed
hsa-mir-139	confirmed	unconfirmed	hsa-mir-133a-1	confirmed	confirmed
hsa-mir-372	confirmed	unconfirmed	hsa-mir-181b-1	confirmed	confirmed
hsa-mir-181b-2	confirmed	confirmed	hsa-mir-26a-1	confirmed	confirmed
hsa-mir-181a-2	confirmed	confirmed	hsa-mir-373	confirmed	unconfirmed
hsa-mir-124-1	confirmed	confirmed	hsa-mir-423	confirmed	unconfirmed
hsa-mir-193a	confirmed	unconfirmed	hsa-mir-499	unconfirmed	unconfirmed
hsa-mir-193b	confirmed	unconfirmed	hsa-mir-128	confirmed	confirmed
hsa-mir-26b	confirmed	unconfirmed	hsa-mir-16	confirmed	unconfirmed
hsa-mir-34b	confirmed	unconfirmed	hsa-mir-212	confirmed	unconfirmed
hsa-mir-1	confirmed	confirmed	hsa-mir-340	confirmed	unconfirmed
hsa-mir-133a-2	confirmed	confirmed	hsa-mir-98	confirmed	unconfirmed
hsa-mir-199b	confirmed	unconfirmed	hsa-mir-100	confirmed	unconfirmed
hsa-mir-27b	confirmed	confirmed	hsa-mir-124-3	confirmed	confirmed
hsa-mir-29c	confirmed	unconfirmed	hsa-mir-133	confirmed	confirmed
hsa-mir-451a	confirmed	unconfirmed	hsa-mir-183	confirmed	confirmed
hsa-mir-144	confirmed	unconfirmed	hsa-mir-370	confirmed	unconfirmed

where adjustment coefficient γ_d was used to adjust the kernel bandwidth, which was computed via normalizing original parameter γ_d' as follows:

$$\gamma_d = \frac{1}{\gamma_d'\left(\frac{1}{nd}\sum_{i=1}^{nd}\|IP(d_i)\|^2\right)}. \quad (11)$$

Similarly, GIP similarity for miRNA $GM(m_i, m_j)$ between miRNA m_i and miRNA m_j can be calculated as follows:

$$GM(m_i, m_j) = exp(-\gamma_m * \| V(m_i) - V(m_j) \|^2) \quad (12)$$

$$\gamma_m = \frac{1}{\gamma_m'\left(\frac{1}{nm}\sum_{i=1}^{nm}\|IP(m_i)\|^2\right)}, \quad (13)$$

where binary vector $V(m_i)$ [or $V(m_j)$)] is the interaction profile of miRNA m_i (or m_j) by observing whether m_i (or m_j) has association with each of the 850 diseases and is equivalent to the i-th (or j-th) column vector of adjacency matrix X.

Multi-source Feature Fusion

By combining the semantic similarity of the disease with the GIP similarity constructed above, a comprehensive similarity matrix incorporating heterogeneous information is computed.[38] The element $DS(d_i, d_j)$ represented combined similarity between disease d_i and d_j, and was described as follows:

$$DS(d_i, d_j) = \begin{cases} \frac{Sim(d_i, d_j) + Sim'(d_i, d_j)}{2} & if\ d_i, d_j\ in\ Sim1 and Sim2 \\ GD(d_i, d_j) & others \end{cases}. \quad (14)$$

The miRNA similarity matrix MS is constructed from miRNA functional similarity MF and miRNA GIP similarity GM. The miRNA similarity matrix [r(i), r(j)] formula for miRNA r(i) and miRNA r(j) is as follows:

$$MS(m_i, m_j) = \begin{cases} MF(m_i, m_j) & if\ m_i, m_j\ in\ FS \\ GM(m_i, m_j) & others \end{cases}. \quad (15)$$

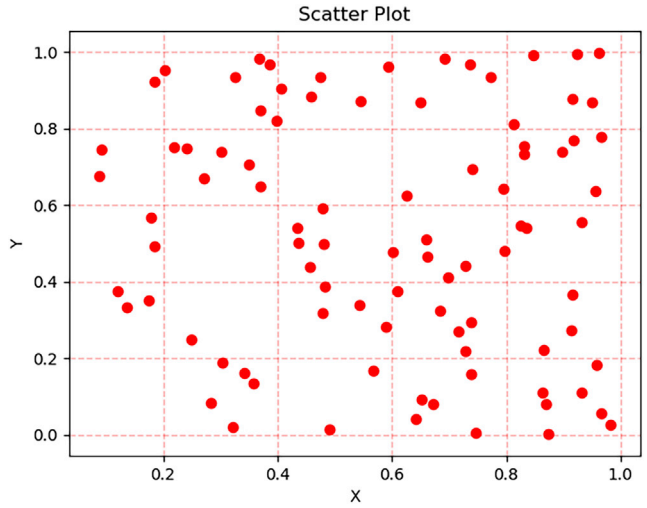

Figure 3. CGR of the miRNA Named hsa-mir-135

CGR

In this study, based on the research of Jessime et al.[39] that homologs can be effectively detected even if all positions of ncRNA are treated equally, we introduced CGR to map RNA sequences. In 1990, Jeffrey[40] built a scale-independent representation for RNA sequences named CGR. CGR is an iterative mapping that can be traced back to chaos theory and is the basis of statistical mechanics. However, studies never fully explore identifying the resulting sequence scheme as representing the nucleotide sequence by the CGR format. RNA sequences can be mapped into the CGR space, which is planar. The four possible nucleotides confine the CGR space as vertices of a binary square (Figure 3).

$$nt_i = \quad nt_{i-1} + \theta * (\, nt_{i-1} - \lambda_i) \tag{16}$$

$$\lambda_i = \begin{cases} (0,0) & if \;\; nucleotide = A \\ (0,1) & if \;\; nucleotide = C \\ (1,1) & if \;\; nucleotide = G \\ (1,0) & if \;\; nucleotide = U \end{cases}, \tag{17}$$

where nt_i is the CGR positions, N_{seq} is the length of the sequence, λ_i is the nucleotide coefficient, parameter θ is the decay factor, and we define $i = 1 \ldots N_{seq}$ and $nt_0 = (0.5, 0.5)$.

Sequence Similarity for miRNAs

Information on miRNA sequences is mapped to Euclidean space, and its region distance is utilized to quantify the similarity of miRNA sequence. It will be easy to implement assignments such as miRNA sequence recognition, verification, and clustering using standard methods with DBMDA embeddings as feature vectors, if this space has been built. First, we downloaded 1,057 miRNA precursor sequences from the miRBase. Second, each nucleotide is mapped to a Euclidean space, and the CGR space is separated from the appropriately sized grid. After

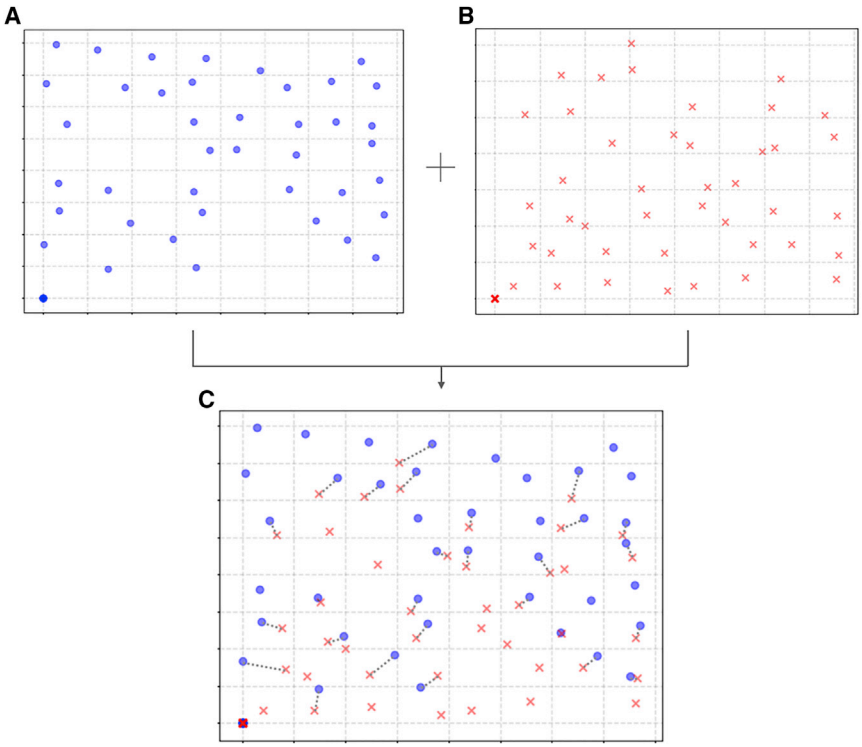

Figure 4. The Flowchart of Quantify the Sequence Similarity Utilizing its Regional Distance

(A) The CGRs of hsa-mir-27a are plotted with the average coordinates for each 8 × 8 quadrant represented. (B) The CGRs of hsa-mir-651 are plotted with the average coordinates for each 8 × 8 quadrant represented. (C) Figuring the region distances of hsa-mir-27a and hsa-mir-651.

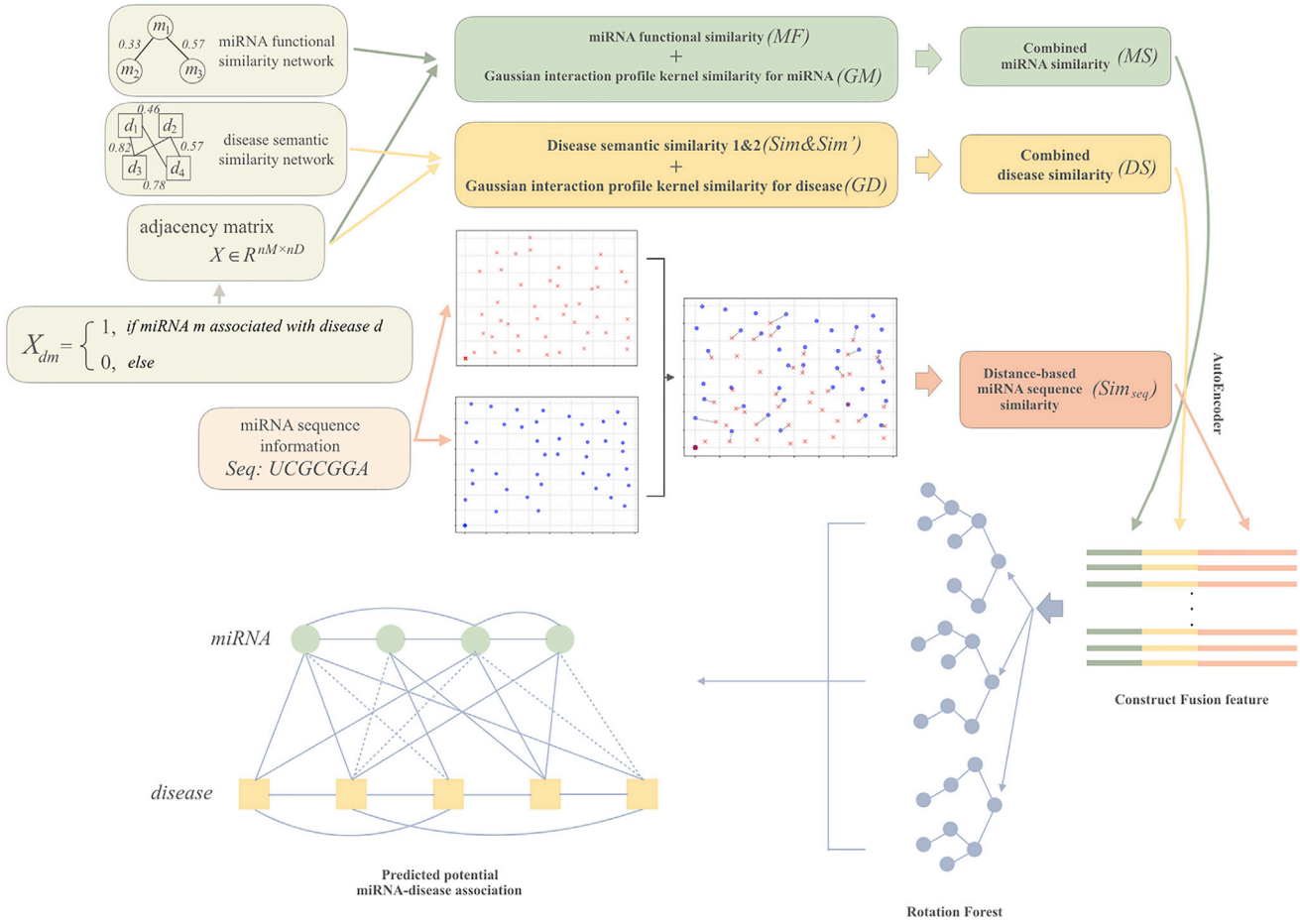

Figure 5. The Workflow of DBMDA Model to Predict Potential miRNA-Disease Associations

that, average coordinate in each quadrant is figured (Figure 4). Third, the regional distance between each miRNA and other miRNAs is calculated. The region distance as $DR_{m_i m_j}(g(i))$ was figured by:

$$DR_{m_i m_j}(g(i)) = \begin{cases} \sqrt{\left(x_{m_i} - x_{m_j}\right)^2 + \left(y_{m_i} - y_{m_j}\right)^2} & \text{if } m_i \text{ and } m_i \text{ both have average coordinate in } g(i) \\ 0 & \text{if } m_i \text{ and } x_{m_j} \text{ both don't have average coordinate in } g(i) \\ \alpha & \text{else} \end{cases}, \quad (18)$$

where $g(i)$ indicates the i-th grid, α represents the penalty parameter, and x_{m_i}, y_{m_i} is the average coordinate of m_i in $g(i)$. Fourth, the calculation of the similarity between sequences at any scale was based on the region distance $DR_{m_i m_j}(g(i))$, defined as follows:

$$Sim_{seq}(m_i, m_j) = \sum_{t=1}^{n_c^2} DR_{m_i m_j}(g(t)). \quad (19)$$

Finally, we used a $2^{n_c} \times 2^{n_c}$ grid to get the distance-based similarity matrix of nucleotide length n_c. (1057×1057). Therefore, each miRNA sequence could be described by a 1,057-dimensional vector:

$$F_{seq} = \left(f_1, f_2, f_3, \ldots, f_{1056}, f_{1057}\right). \quad (20)$$

Rotation Forest

Rotation Forest() independently trains decision trees using different extraction feature sets.[41,42] Rodríguez et al.[42] defined $F = [f_1, \ldots, f_n]^T$ as n features (attributes), which is an $N \times n$ matrix that represents the training and $T = [T_1, \ldots, T_r]$ as the

ensemble of r classifiers. Each bootstrap sample is trained separately for the independent classifiers. The improvement of RoF is extracting a feature and rebuilding a complete training set for each decision tree in T. Specifically, the RoF randomly divides the training set into e subsets and runs principal-component analysis (PCA) separately. The data are mapped into the new feature space and use it to train classifier T_i. Different subsets will extract different features that improve the diversity through the bootstrap sampling.

Method Overview

A DBMDA was built. It assumes that functionally similar diseases have relation to similar miRNAs, which is also used to compute the association between target proteins and drug. DBMDA has four main processes: first, choosing positive examples and negative examples; second, gathering complex feature vectors by miRNA and disease similarity matrixes; and third, building an effective prediction model to figure potential miRNA-disease pairs. Specifically speaking, we will introduce each process in more detail.

First, we constructed the training examples. Specifically, we analyzed HMDD v.3.0 and selected the known miRNA-disease associations as positive samples. Then, we clustered all of the positive samples with negative samples to build a training set. There are three steps of selecting negative samples: (1) we selected a disease from all known diseases (850) randomly, (2) chose a miRNA in the same way, and (3) combined the miRNA and disease if miRNA and disease pair is not in positive samples as a negative sample.

Second, we built the feature set. In particular, we gathered three disease matrixes, which are a Gaussian profile kernel similarity matrix and two semantic similarity matrixes, into feature vectors as disease features. Feature vector of disease is described as follows:

$$DS(d_i) = (\eta_1, \eta_2, \eta_3, \ldots \eta_{849}, \eta_{850}), \tag{21}$$

where the i-th row vector of matrix DS is described as $DS(d_i)$, and the combined similarity value between disease d_i and d_j is defined as η_j. In the same way, we calculated each of 1,057 similarity values to construct a 1,057-dimensional feature vector by Gaussian interaction kernel profile similarity matrix as follows:

$$MS(m_a) = (\varphi_1, \varphi_2, \varphi_3, \ldots \varphi_{1056}, \varphi_{1057}), \tag{22}$$

where the a-th column vector of matrix MS is described as $MS(m_a)$, and the gathered similarity value of miRNA m_a and m_b is described as φ_b. Each miRNA-disease sample can be described as 1,907-dimensional vector as follows:

$$F_{sim} = (DS(d_i), MS(m_a)). \tag{23}$$

$F_{sim} = (\eta_1, \eta_2, \eta_3, \ldots \varphi_{1906}, \varphi_{1907})$, where $(\eta_1, \eta_2, \eta_3, \ldots \eta_{850})$ are the 850 gathered similarity values of the disease and $(\varphi_{851}, \varphi_{852}, \varphi_{853}, \ldots \varphi_{1907})$ stands for the 1,057 combined similarity values of the miRNAs. After that, we resized F_{sim} by autoencoder (AE) from 1,097 to 32, and the sequence feature matrixes F_{seq} is resized in same way from 1,057 to 32.[43] We defined each miRNA-disease sample as a 64-dimensional vector as follows:

$$F = \left(F_{sim}', F_{seq}'\right). \tag{24}$$

Finally, we used RoF to build the prediction model by training set. In particular, we got 64-dimensional vectors in steps 2 and 3 and used them as training set. Then, training samples were put into RoF, and a predicting potential miRNA-disease associations model was built. The workflow of the DBMDA model is shown in Figure 5.

Availability of Data and Materials

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

AUTHOR CONTRIBUTIONS

K.Z. conceived the algorithm, analyzed it, conducted the experiment, and wrote the manuscript. K.Z. and L.W. prepared the dataset. L.-P.L., Z.-W.L., and Y.Z. analyzed the experiment. The final draft was read and approved by all authors.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

Z.-H.Y. was supported by National Natural Science Foundation of China Grant 61572506, the Pioneer Hundred Talents Program of Chinese Academy of Sciences, and the CCF-Tencent Open Fund. L.W. was supported by National Natural Science Foundation of China Grant 61702444, Chinese Postdoctoral Science Foundation Grant 2019M653804, and West Light Foundation of the Chinese Academy of Sciences Grant 2018-XBQNXZ-B-008. Z.-W.L. was supported by National Natural Science Foundation of China Grant 61873270. The authors would like to thank all anonymous reviewers for their constructive advice.

REFERENCES

1. Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350–355.
2. Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281–297.
3. Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843–854.
4. Ambros, V. (2001). microRNAs: tiny regulators with great potential. Cell *107*, 823–826.
5. Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell *75*, 855–862.
6. Meola, N., Gennarino, V.A., and Banfi, S. (2009). microRNAs and genetic diseases. PathoGenetics *2*, 7.
7. Zhu, X., Li, Y., Shen, H., Li, H., Long, L., Hui, L., and Xu, W. (2013). miR-137 inhibits the proliferation of lung cancer cells by targeting Cdc42 and Cdk6. FEBS Lett. *587*, 73–81.
8. von Brandenstein, M., Pandarakalam, J.J., Kroon, L., Loeser, H., Herden, J., Braun, G., Wendland, K., Dienes, H.P., Engelmann, U., and Fries, J.W. (2012). MicroRNA 15a,

inversely correlated to PKCα, is a potential marker to differentiate between benign and malignant renal tumors in biopsy and urine samples. Am. J. Pathol. *180*, 1787–1797.

9. Chu, T.-H., Yang, C.C., Liu, C.J., Lui, M.T., Lin, S.C., and Chang, K.W. (2013). miR-211 promotes the progression of head and neck carcinomas by targeting TGFβRII. Cancer Lett. *337*, 115–124.
10. Bandyopadhyay, S., Mitra, R., Maulik, U., and Zhang, M.Q. (2010). Development of the human cancer microRNA network. Silence *1*, 6.
11. Chen, X., Yan, C.C., Zhang, X., and You, Z.-H. (2016). Long non-coding RNAs and complex diseases: from experimental results to computational models. Brief Bioinform *18*, 558–576.
12. Chen, X., You, Z.H., Yan, G.Y., and Gong, D.W. (2016). IRWRLDA: improved random walk with restart for lncRNA-disease association prediction. Oncotarget *7*, 57919–57931.
13. Chen, X., Huang, Y.A., Wang, X.S., You, Z.H., and Chan, K.C. (2016). FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model. Oncotarget *7*, 45948–45958.
14. You, Z.H., Huang, Z.A., Zhu, Z., Yan, G.Y., Li, Z.W., Wen, Z., and Chen, X. (2017). PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. PLoS Comput. Biol *13*, e1005455.
15. Zheng, K., You, Z.H., Wang, L., Li, Y.R., Wang, Y.B., and Jiang, H.J. (2019). MISSIM: improved miRNA-disease association prediction model based on chaos game representation and broad learning system. In Intelligent Computing Methodologies: 15th International Conference, ICIC 2019, D.S. Huang, Z.K. Huang, and A. Hussain, eds. (Springer), pp. 392–398.
16. Zheng, K., You, Z.H., Wang, L., Zhou, Y., Li, L.P., and Li, Z.W. (2019). MLMDA: a machine learning approach to predict and validate MicroRNA-disease associations by integrating of heterogenous information sources. J. Transl. Med. *17*, 260.
17. Zheng, K., Wang, L., and You, Z.-H. (2019). CGMDA: An Approach to Predict and Validate MicroRNA-Disease Associations by Utilizing Chaos Game Representation and LightGBM. IEEE Access *7*, 133314–133323.
18. Chen, X., Liu, M.-X., and Yan, G.-Y. (2012). RWRMDA: predicting novel human microRNA-disease associations. Mol. Biosyst. *8*, 2792–2798.
19. Li, X., Wang, Q., Zheng, Y., Lv, S., Ning, S., Sun, J., Huang, T., Zheng, Q., Ren, H., Xu, J., et al. (2011). Prioritizing human cancer microRNAs based on genes' functional consistency between microRNA and cancer. Nucleic Acids Res. *39*, e153.
20. Chen, X., Yan, C.C., Zhang, X., You, Z.H., Huang, Y.A., and Yan, G.Y. (2016). HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction. Oncotarget *7*, 65257–65269.
21. Xuan, P., Han, K., Guo, Y., Li, J., Li, X., Zhong, Y., Zhang, Z., and Ding, J. (2015). Prediction of potential disease-associated microRNAs based on random walk. Bioinformatics *31*, 1805–1815.
22. Xu, J., Li, C.X., Lv, J.Y., Li, Y.S., Xiao, Y., Shao, T.T., Huo, X., Li, X., Zou, Y., Han, Q.L., et al. (2011). Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. Mol. Cancer Ther. *10*, 1857–1866.
23. Chen, X., and Yan, G.-Y. (2014). Semi-supervised learning for potential human microRNA-disease associations inference. Sci. Rep. *4*, 5501.
24. Chen, X., Yan, C.C., Zhang, X., Li, Z., Deng, L., Zhang, Y., and Dai, Q. (2015). RBMMMDA: predicting multiple types of disease-microRNA associations. Sci. Rep *5*, 13877.
25. Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., and Cui, Q. (2014). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res. *42*, D1070–D1074.
26. Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J. (2008). miRBase: tools for microRNA genomics. Nucleic Acids Res. *36*, D154–D158.
27. Chen, H., and Zhang, Z. (2013). Similarity-based methods for potential human microRNA-disease association prediction. BMC Med. Genomics *6*, 12.
28. Yu, H., Chen, X., and Lu, L. (2017). Large-scale prediction of microRNA-disease associations by combinatorial prioritization algorithm. Sci. Rep. *7*, 43792.
29. Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., Liu, Y., Dai, Q., Li, J., Teng, Z., and Huang, Y. (2013). Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. PLoS ONE *8*, e70204.
30. Yang, Y., Fu, X., Qu, W., Xiao, Y., and Shen, H.-B. (2018). MiRGOFS: a GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA-disease association. Bioinformatics *34*, 3547–3556.
31. Chen, X., Yin, J., Qu, J., and Huang, L. (2018). MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. PLoS Comput. Biol. *14*, e1006418.
32. Yang Z, Ren F, Liu C, He S, Sun G, Gao Q, Yao L, Zhang Y, Miao R, Cao Y, et al., dbDEMC: a database of differentially expressed miRNAs in human cancers, BMC Genomics, 11, https://doi.org/10.1186/1471-2164-11-S4-S5.
33. Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. *37*, D98–D104.
34. Chen, L., Liu, B., and Yan, C. (2018). DPFMDA: Distributed and privatized framework for miRNA-Disease association prediction. Pattern Recognit. Lett. *109*, 4–11.
35. Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics *26*, 1644–1650.
36. Lord, P.W., Stevens, R.D., Brass, A., and Goble, C.A. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics *19*, 1275–1283.
37. van Laarhoven, T., Nabuurs, S.B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics *27*, 3036–3043.
38. Chen, X., Yan, C.C., Zhang, X., You, Z.H., Deng, L., Liu, Y., Zhang, Y., and Dai, Q. (2016). WBSMDA: within and between score for MiRNA-disease association prediction. Sci. Rep. *6*, 21106.
39. Kirk, J.M., Kim, S.O., Inoue, K., Smola, M.J., Lee, D.M., Schertzer, M.D., Wooten, J.S., Baker, A.R., Sprague, D., Collins, D.W., et al. (2018). Functional classification of long non-coding RNAs by k-mer content. Nat. Genet. *50*, 1474–1482.
40. Jeffrey, H.J. (1990). Chaos game representation of gene structure. Nucleic Acids Res. *18*, 2163–2170.
41. Kuncheva, L.I., and Rodríguez, J.J. (2007). An experimental study on Rotation Forest ensembles. In Multiple Classifier Systems: MCS 2007. Lecture Notes in Computer Science, *Volume 4472*, M. Haindl, J. Kittler, and F. Roli, eds. Multiple Classifier Systems: MCS 2007. Lecture Notes in Computer Science (Springer), pp. 459–468.
42. Rodríguez, J.J., Kuncheva, L.I., and Alonso, C.J. (2006). Rotation forest: A new classifier ensemble method. IEEE Trans. Pattern Anal. Mach. Intell. *28*, 1619–1630.
43. Deng, L., et al. Eleventh Annual Conference of the International Speech Communication Association.
44. Wang, L., You, Z.H., Chen, X., Li, Y.M., Dong, Y.N., Li, L.P., and Zheng, K. (2019). LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities. PLoS Comput. Biol. *15*, e1006865.