Molecular Therapy
Original Article

DF-MDA: An effective diffusion-based computational model for predicting miRNA-disease association

Hao-Yuan Li,[1,5] Zhu-Hong You,[2,5] Lei Wang,[2,3] Xin Yan,[1,4] 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; [4]School of Foreign Languages, Zaozhuang University, Zaozhuang, Shandong 277100, China

It is reported that microRNAs (miRNAs) play an important role in various human diseases. However, the mechanisms of miRNA in these diseases have not been fully understood. Therefore, detecting potential miRNA-disease associations has far-reaching significance for pathological development and the diagnosis and treatment of complex diseases. In this study, we propose a novel diffusion-based computational method, DF-MDA, for predicting miRNA-disease association based on the assumption that molecules are related to each other in human physiological processes. Specifically, we first construct a heterogeneous network by integrating various known associations among miRNAs, diseases, proteins, long non-coding RNAs (lncRNAs), and drugs. Then, more representative features are extracted through a diffusion-based machine-learning method. Finally, the Random Forest classifier is adopted to classify miRNA-disease associations. In the 5-fold cross-validation experiment, the proposed model obtained the average area under the curve (AUC) of 0.9321 on the HMDD v3.0 dataset. To further verify the prediction performance of the proposed model, DF-MDA was applied in three significant human diseases, including lymphoma, lung neoplasms, and colon neoplasms. As a result, 47, 46, and 47 out of top 50 predictions were validated by independent databases. These experimental results demonstrated that DF-MDA is a reliable and efficient method for predicting potential miRNA-disease associations.

INTRODUCTION

MicroRNAs (miRNAs) are a collection of small (about 23 nucleotides) non-coding RNAs.[1] They generally act as negative or positive regulators in biological processes by connecting with 3′ UTR sites of the mRNAs.[2] A great number of reports have demonstrated that miRNAs influence many critical biological processes, including cell diffusion,[3] growth,[4] divergence,[5] death,[6] and so on. Therefore, miRNAs have great effects on various biological progress.[7–9]

Recently, emerging evidence has shown that miRNAs are closely related to diseases and play an important role in complex human diseases.[10–12] It has become a research hotspot to predict miRNA-disease associations.[13–15] For example, Liu et al.[16] demonstrated that hsa-miR-124-3p could effectively regulate the SOCS3 (suppressor of cytokine signaling 3), a tumor suppressor in breast neoplasms cells. Kumarswamy et al.[17] detected that miR-21 is downregulated in almost all types of cancers, which led the miR-21 to become an attractive target for therapeutic strategies. Furthermore, miRNAs have been new biomarkers in human disease diagnosis, especially in the cancer field.[18] Xie et al.[19] discovered that miR-342-3p could inhibit lung cancer cell proliferation by targeting Ras-related protein Rap-2b, which may bring about a novel biomarker and treatment for lung cancer patients. Therefore, effectively identifying miRNA-disease associations could greatly promote the treatment of human complex diseases.[20,21]

With the development of biotechnology, a growing number of biological data were generated.[22] Multiple databases (e.g., the Human MicroRNA Disease Database [HMDD],[23] miR2Disease,[24] and Database of Differentially Expressed miRNAs in Human Cancers [dbDEMC][25]) are formed by collecting these biological data.[26] These databases supply a large amount of data verified by biological experiments, which makes predicting miRNA-disease associations by computational methods feasible.[27] An increasing number of researchers use these known data to predict the association between miRNAs and diseases by computational methods. The most likely relationship between miRNAs and diseases would need to be verified by biological experiment, which could eliminate a large number of wrong answers and save valuable experimental costs.[28–30] The best computational methods can even replace biological experiments and complete the prediction of the relationship between miRNAs and diseases with extremely high accuracy. For example, Jiang et al.[31] developed a novel computational model for the prediction of miRNAs and diseases. However, this method excessively

Received 7 September 2020; accepted 1 January 2021;
https://doi.org/10.1016/j.ymthe.2021.01.003.

[5]These authors contributed equally

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: Xin Yan, School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China.
E-mail: xinyanuzz@gmail.com

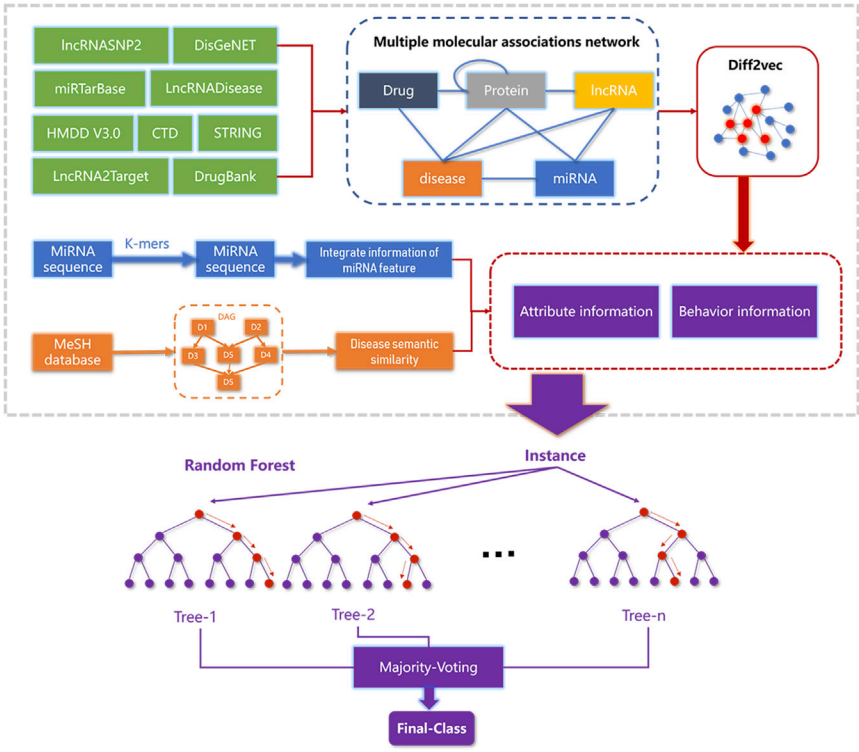

Figure 1. Flowchart of DF-MDA to predict potential miRNA-disease associations

relied on the relationship among miRNAs, which greatly impact the results. Xuan et al.[32] proposed a novel computational model of HDMP. Different from previous models, HDMP added weighted *k* most similarity neighbors of miRNAs, and the weight is determined by the similarity between miRNA and its neighbor, which could greatly improve the performance of model predictions. Nonetheless, HDMP becomes invalid to predict the diseases without any known related miRNAs. Chen et al.[33] presented WBSMDA for predicting miRNA-disease associations. This model connected the within score and between score of the relationship between miRNA and disease. WBSMDA greatly improved the scope and prediction of the model and suitable for predicting new disease-miRNA associations. In addition, You et al.[34] presented PBMDA to predict the potential relationship between miRNAs and diseases, which constructed a heterogeneous association network by integrating a large amount of biological data. PBMDA could well work for these new diseases with unknown related miRNAs and vice versa. What's more, this model takes advantage of the topology information of the heterogeneous network by depth-first calculating based on the path. A model of RKNNMDA proposed by Chen et al.[35] is a ranking-based K-nearest neighbor (KNN) method for predicting the association of miRNA and disease. These KNNs would be ranked by the support vector machine (SVM) ranking model to obtain the priority miRNA-disease relationships. In recent years, these proposed computational methods have made up for the time-consuming and costly traditional biological experiments to a certain degree.[36]

In this study, we developed a novel computational model, DF-MDA, for predicting miRNA-disease associations based on the assumption that molecules are related to each other in human physiological processes. The flow chart of DF-MDA is shown in Figure 1. More specifically, we first utilize a comprehensive molecular-associations network (MAN)[37] to integrate various biological data and learning the behavior feature information of miRNAs and diseases in the network by the diffusion model. Then, based on the known miRNA and disease association, we construct a comprehensive feature descriptor by integrating the above information with miRNA sequence information and disease semantic similarity information. Finally, these feature descriptors are trained by the Random Forest (RF) classifier to accurately classify and predict the association between miRNAs and disease. In the experiment, DF-MDA obtained outstanding performance in 5-fold cross-validation (the average area under the curve [AUC] of 0.9321) based on the HMDD v3.0 database. To further evaluate the performance of the model, we compared the proposed model with different classifiers and feature extraction models. In addition, we implemented the case studies of lung neoplasms, colon neoplasms, and lymphoma neoplasms. As a result, 47, 46, and 47 out of top 50 miRNA candidates were verified by independent databases, respectively. The above experiment results demonstrated that the DF-MDA is a reliable and effective model to predict the association of miRNA and disease.

RESULTS

Evaluation criteria

In order to more comprehensively evaluate the performance of the proposed model, we adopted a variety of evaluation criteria to assess DF-MDA, involving accuracy (Acc.), sensitivity (Sen.), specificity (Spec.), precision (Prec.), and Matthew's correlation coefficient (MCC). These formulae of criteria were calculated as follows:

$$Acc. = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sen. = \frac{TP}{TP + FN}$$

$$Prec. = \frac{TP}{TP + FP}$$

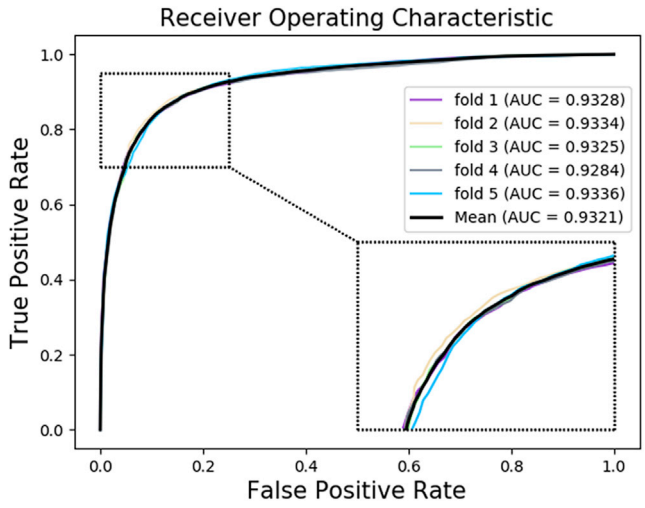

Figure 2. ROC curves performed by DF-MDA on HMDD dataset

$$Spec. = \frac{TN}{TN + FN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where true positive (TP), true negative (TN), false positive (FP) and false negative (FN) express the number of positive samples correctly predicted, negative samples correctly predicted, positive samples wrongly predicted, and negative samples wrongly predicted by the model, respectively. Furthermore, we also draw the receiver operating characteristic (ROC) curve and the AUC to describe the capability of DF-MDA.

Performance evaluation

In this study, we implemented the 5-fold cross-validation methods based on known database HMDD v3.0 to evaluate the DF-MDA model. In this paper, we choose these verified miRNA-disease associations to be the positive samples and randomly selected the same number of uncorrelated miRNA-disease associations to be the negative samples. These pairs of miRNA-disease would be split into five uncrossed subsets. For each validation, four-fifths of them were regarded as a train set and the other was test in the classifier. To avoid the leakage of test information, we used the train set to construct the whole network. As a result, the Acc., Sen., Spec., Prec., MCC, and AUC achieved 86.50%, 87.41%, 85.58%, 85.85%, 73.01%, and 0.9321, with standard deviations of 0.37%, 0.66%, 0.74%, 0.58%, 0.74%, and 0.0021, respectively. The detailed result of the model under 5-fold cross-validation on HMDD v3.0 is shown in Table 1. Furthermore, the ROC curves generated by DF-MDA are shown in Figure 2. The above experiment results indicate that DF-MDA is an efficacious model to predict the potential relationship of miRNA-disease.

Table 1. 5-fold cross-validation results performed by DF-MDA on HMDD dataset

Fold	Acc. (%)	Sen. (%)	Spec. (%)	Prec. (%)	MCC (%)	AUC
1	86.32	87.10	85.54	85.77	72.65	0.9328
2	87.13	88.01	86.24	86.48	74.27	0.9334
3	86.17	87.61	84.72	85.15	72.37	0.9325
4	86.40	86.40	86.40	86.40	72.79	0.9284
5	86.46	87.91	85.01	85.44	72.95	0.9336
Average	86.50 ± 0.37	87.41 ± 0.66	85.58 ± 0.74	85.85 ± 0.58	73.01 ± 0.74	0.9321 ± 0.0021

Comparison with DeepWalk model

In order to test the performance of the diffusion-based model, we compared the DF-MDA model with the DeepWalk model in the same dataset. DeepWalk[38] is a classic network embedding model based on a random walk. The algorithm has been widely used in bioinformatics and achieved excellent results.[39] In this experiment, we also used 5-fold cross-validation by the DeepWalk model based on HMDD v3.0. As a result, the DeepWalk model achieved an average AUC of 0.8929. As shown in Table 2, the performance of DF-MDA is better than that of DeepWalk for predicting miRNA-disease associations. The reason for this result is that the DeepWalk model is mainly concerned with the local characteristics of the network, while the DF-MDA model extracts the more comprehensive feature of nodes in the molecular association network. The accuracy of DF-MDA is 4.57% higher than the DeepWalk model, and the sensitivity of our method is 8.70% higher than the DeepWalk model. The performance comparisons in 5-fold cross-validation are shown in Figure 3.

Comparison with different feature descriptor models

In this study, every node is described by its inherent attribute information and the behavior information in the whole network. To test their influence on the performance of DF-MDA, we compared the different feature descriptors with only attribute information (DF-MDA_AI), only behavior information (DF-MDA_BI) and both of them (DF-MDA), respectively. We assume the attribute information of other nodes has almost no impact on the predictive performance of the proposed model. In this work, we only adopt the attribute feature information of miRNAs and diseases. The detailed result of different feature descriptor models based on HMDD v3.0 is shown in Table 3. As shown in the table, the AUC of DF-MDA is 0.0552 and 0.0138 higher than that of DF-MDA_AI and DF-MDA_BI, respectively, and the accuracy of DF-MDA is 5.23% and 0.74% higher than that of DF-MDA_AI and DF-MDA_BI, respectively. The reason for this

Table 2. The comparison results between DeepWalk and DF-MDA model by Random Forest classifier based on HMDD database

Model	Acc. (%)	Sen. (%)	Spec. (%)	Prec. (%)	MCC (%)	AUC
DeepWalk	81.93 ± 0.56	78.70 ± 0.71	85.15 ± 1.05	84.14 ± 0.92	63.99 ± 1.15	0.8929 ± 0.0047
DF-MDA	86.50 ± 0.37	87.41 ± 0.66	85.58 ± 0.74	85.85 ± 0.58	73.01 ± 0.74	0.9321 ± 0.0021

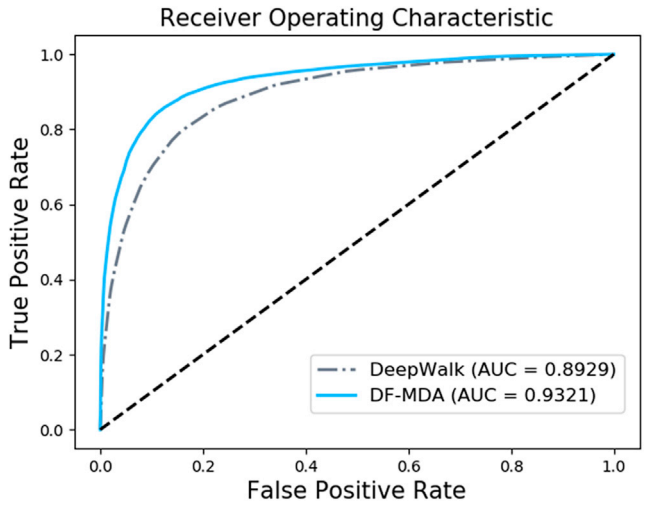

Figure 3. ROC curves performed by DeepWalk and DF-MDA by Random Forest classifier based on HMDD database

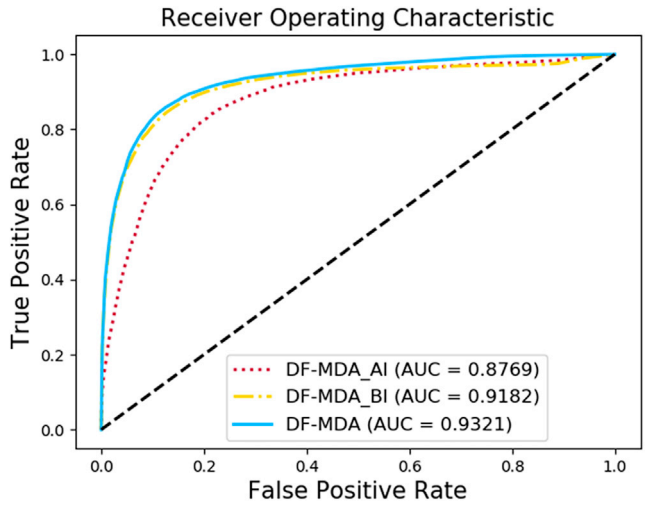

Figure 4. ROC curves performed by DF-MDA_AI model, DF-MDA_BI model, and DF-MDA model based on HMDD database

result is that the DF-MDA_AI predicts the relationships between miRNAs and diseases by the miRNA attribute information and disease semantic similarity, which could capture the properties of the node itself. However, the DF-MDA_AI ignores these known associations in the network, which is unable to provide comprehensive information for our prediction. The same situation exists in DF-MDA_BI, which lacks the attribute information of nodes. The ROC curves of the three experiments are shown in Figure 4. In conclusion, the performances in DF-MDA of AUCs are more outstanding than the model of feature descriptor with only one information.

Comparison with different classifier models

DF-MDA adopted the Random Forest classifier to train and classify the potential miRNA-disease associations. To evaluate the performance of the Random Forest model, we compared it with Bagging, LogisticRegression, Naive Bayes, and AdaBoost classifier models. We implemented 5-fold cross-validation by all above models on the same training set and test set. As a result, Random Forest yielded an average AUC of 5-fold cross-validation of 0.9321 and outperformed Bagging (0.9089), LogisticRegression (0.9124), Naive Bayes (0.8505), and AdaBoost (0.9153). The Random Forest is only worse than Bagging of Spec., and the accuracy is 2.54% higher than that of the second-highest method. The explanation for this phenomenon is that Random Forest can handle very high dimensional data. Furthermore, the model has strong generalization ability due to adopting the unbiased estimation for generalization error in creating a random forest. It is worth mentioning that the Naive Bayes model is lower than other algorithms. The reason is that the Naive Bayes assumes that the features are independent of each other, which is often not true in reality. The detailed results of different classifier models are shown in Table 4. Furthermore, we drew the ROC curve as shown in Figure 5. These results have demonstrated that DF-MDA by Random Forest classifier has achieved the best performance, particularly in Acc., MCC, and AUC. From the information above, the Random Forest classifier is more appropriate for DF-MDA.

Table 3. The comparison results between DF-MDA_AI model, DF-MDA_BI model and DF-MDA model based on HMDD database

Model	Acc. (%)	Sen. (%)	Spec. (%)	Prec. (%)	MCC (%)	AUC
DF-MDA_AI	81.27 ± 0.58	83.71 ± 0.75	78.84 ± 0.65	79.82 ± 0.57	62.63 ± 1.16	0.8769 ± 0.0052
DF-MDA_BI	85.74 ± 0.39	82.48 ± 0.67	88.99 ± 0.35	88.23 ± 0.35	71.63 ± 0.76	0.9183 ± 0.0030
DF-MDA	86.50 ± 0.37	87.41 ± 0.66	85.58 ± 0.74	85.85 ± 0.58	73.01 ± 0.74	0.9321 ± 0.0021

Comparison with related works

At present, an increasing number of computational methods have been proposed for predicting miRNA-disease associations. Therefore, in order to further evaluate the performance of our model, we compared the proposed method with seven previous works, including RWRMDA,[40] MTDN,[41] EGBMMDA,[42] LMTRDA,[43] DBMDA,[44] PBMDA,[45] and CGMDA.[46] Since these algorithms have not calculated multiple evaluation criteria, we only compared the AUC on the terms of the 5-fold cross-validation-based HMDD database. As shown in Table 5, the performance of DF-MDA is outstanding compared with other methods, and the proposed method is 0.0282 higher than the average AUC of all algorithms. This is because the proposed model combines the attribute information and behavior information of miRNAs and diseases and extracts the feature more comprehensively.

Case studies

In this work, we carry out three important human diseases (lung neoplasms, colon neoplasms, and lymphoma) by DF-MDA based on HMDD v3.0 to further evaluate its predictive power. These known

Table 4. The comparison results between Random Forest and other classifier models (Bagging, LogisticRegression, Naive Bayes, and AdaBoost) based on HMDD database

Model	Acc. (%)	Sen. (%)	Spec. (%)	Prec. (%)	MCC (%)	AUC
Random Forest	86.50 ± 0.37	87.41 ± 0.66	85.58 ± 0.74	85.85 ± 0.58	73.01 ± 0.74	0.9321 ± 0.0021
Bagging	83.90 ± 0.52	81.26 ± 1.06	86.53 ± 0.45	85.79 ± 0.40	67.89 ± 1.01	0.9089 ± 0.0012
LogisticRegression	83.96 ± 0.39	83.44 ± 0.80	84.48 ± 0.78	84.32 ± 0.61	67.93 ± 0.78	0.9124 ± 0.0014
Naive Bayes	78.67 ± 0.42	85.40 ± 0.75	71.94 ± 0.46	75.27 ± 0.33	57.86 ± 0.87	0.8515 ± 0.0079
AdaBoost	83.81 ± 0.17	83.32 ± 0.51	84.29 ± 0.62	84.14 ± 0.45	67.62 ± 0.34	0.9153 ± 0.0038

miRNA-disease associations are used as the training dataset, and the test dataset includes the association pairs of three diseases and all possible miRNAs. In this study, we verified the top 50 candidate miRNAs by two independent databases (dbDEMC[25] and miR2Disease[24]). In three case studies, most candidate-related miRNAs were confirmed, demonstrating that DF-MDA is a reliable model for predicting the association of miRNA and disease.

Lung neoplasm is the second most common cancer in humans (~13% of all) except for skin cancers, and the number of deaths caused by lung cancer is the highest (~24% of all).[47] In 2019, there are about 228,150 new lung cancer cases (116,440 of men and 111,710 of women) and 142,670 deaths for lung cancer (76,650 of men and 66,020 of women) in the United States. An increasing amount of research pays attention to the prediction of the potential relationship between miRNAs and lung neoplasms.[48] Therefore, we implemented a case study of lung neoplasms by DF-MDA for more miRNA based on HMDD v3.0, and the details of the result are shown in Table 6, in which 47 of top 50 candidates were verified based on the independent database.

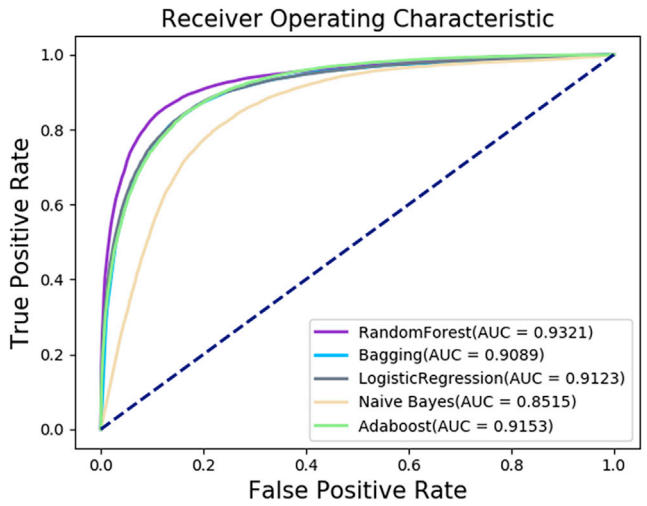

Figure 5. ROC curves performed by Random Forest and other classifiers (Bagging, LogisticRegression, Naive Bayes, and AdaBoost) based on HMDD database

Colon neoplasm is the third most common cancer in the United States (~8% of new cancer) except for skin cancer.[47] In 2019, it is expected that about 145,600 people will develop colon cancer (78,500 men and 67,100 women) and there will be about 51,020 deaths from colon cancer (27,640 men and 23,380 women). Recently, increasing researchers have indicated that miRNAs are related with colon neoplasms.[49] Thus, we used DF-MDA to predict more colon neoplasm-related miRNAs to verify its performance, and the details of the result is shown in Table 7, in which 46 of top 50 candidates were confirmed based on the independent database.

Lymphoma is one of the most common malignant cancers (~4% of all new cancers), especially in teenagers in the United States.[47] In 2019, it is estimated that there will be about 74,200 new cases of lymphoma (41,090 of men and 33,110 of women) and 19,970 deaths from lymphoma (11,510 men and 8,460 women). Lymphoma mainly contains two types of Hodgkin's lymphoma (HL) and non-HL (NHL).[50] Therefore, we selected lymphoma as a case study to verify the performance of DF-MDA. The details of the result are shown in Table 8, in which 47 of top 50 candidates were proved based on the independent database.

DISCUSSION

Recently, an accumulating amount of research demonstrated that miRNAs have a close link with diseases. In this work, we proposed the diffusion-based computational model DF-MDA for predicting miRNA-disease associations. This model can extract effective features of miRNAs and diseases from a complex heterogeneous network, including miRNA, disease, drug, protein, and long non-coding RNA (lncRNA), and the Random Forest classifier was adopted to classify the potential miRNA-disease associations. Compared with other classifiers and feature extraction models, DF-MDA shows excellent performance. In addition, in the case study of lung neoplasms, colon neoplasms, and lymphomas, 47, 46, and 47 of top 50 miRNA candidates predicted by DF-MDA were verified in the independent database, respectively. These results indicated that DF-MDA can be used as a valuable model for predicting miRNA-disease associations.

There are some reasons for the remarkable predictive power of DF-MDA. First, unlike previous studies, we combined multiple molecular-association datasets to construct a comprehensive network of more than just miRNAs and diseases. It is worth noting that DF-

Table 5. The comparison results between DF-MDA with other related works

Method	AUC
RWRMDA	0.8617
MTDN	0.8872
EGBMMDA	0.9048
LMTRDA	0.9054
DBMDA	0.9129
PBMDA	0.9172
CGMDA	0.9099
DF-MDA	0.9321

MDA not only uses the attribute information of miRNAs and diseases, but also adopts their behavior information for predicting the potential relationship between them. Second, the behavior information of biological molecular was extracted by the diffusion-based model, which could effectively detect the network structure by generating more informative traces. Additionally, DF-MDA is suitable for new diseases with unknown related miRNAs and new miRNAs with unknown related diseases. However, limitations also exist in the model. First, the relationship evidence of miRNAs and diseases are still insufficient for prediction. The prediction performance of DF-MDA would improve with the amount of biological data increasing in future work. Furthermore, the miRNA sequence information extraction method also influences the performance of our approach.

MATERIALS AND METHODS

Human miRNA-disease associations database

In this study, we implement the model on the HMDD v3.0[23] database. The HMDD database supplies plenty of experimentally verified miRNA-disease associations, which can be freely obtained from http://www.cuilab.cn/hmdd. Currently, HMDD has collected 32,281 verified miRNA-disease associations, including 1,102 miRNAs and 850 diseases from 17,412 papers. After removing redundancy and simplifying, we obtained 16,427 miRNA-disease associations, involving 1,023 miRNAs and 850 diseases. In the experiment, we use the adjacency matrix $AM(i, j)$ to represent the miRNA-disease association. When the miRNA $m(i)$ is confirmed to be related with disease $d(j)$, the $AM(i, j)$is equal to 1, otherwise 0.

MAN

In this experiment, we used the MAN to integrate multiple biological data. The MAN is a large heterogeneous network proposed by Guo et al.[37] This complex network consists of various nodes and edges based on the associations among them. It provides a novel frame to identify the potential association between any research object in the network. Through this molecular-association network, a comprehensive perspective is obtained to understand human biological progress and disease treatment. At present, MAN integrates five different kinds of molecules (miRNA, disease, lncRNA, protein, and drug) and associations between them. The details of different types of molecules are shown in Table 9, and associations between them are shown in Table 10.

Table 6. Prediction of the top 50 miRNAs related to lung neoplasms based on known miRNA-disease associations in HMDD database

Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	hsa-mir-106b-5p	dbDEMC	26	hsa-mir-302b-5p	dbDEMC
2	hsa-mir-204-5p	dbDEMC	27	hsa-mir-501-5p	dbDEMC
3	hsa-mir-181b-5p	dbDEMC	28	hsa-mir-302f	dbDEMC
4	hsa-mir-15b-5p	dbDEMC	29	hsa-mir-367-3p	dbDEMC
5	hsa-mir-16-1-3p	dbDEMC	30	hsa-mir-363-3p	dbDEMC
6	hsa-mir-193b-5p	dbDEMC	31	hsa-mir-449b-5p	dbDEMC
7	hsa-mir-23b-5p	dbDEMC	32	hsa-mir-429	dbDEMC
8	hsa-mir-424-5p	dbDEMC	33	hsa-mir-1271-5p	dbDEMC
9	hsa-mir-20b-5p	dbDEMC	34	hsa-mir-125b-2-3p	dbDEMC
10	hsa-mir-28-5p	dbDEMC	35	hsa-mir-484	dbDEMC
11	hsa-mir-296-5p	dbDEMC	36	hsa-mir-518b	dbDEMC
12	hsa-mir-452-5p	dbDEMC	37	hsa-mir-378a-5p	dbDEMC
13	hsa-mir-483-5p	dbDEMC	38	hsa-mir-376b-5p	dbDEMC
14	hsa-mir-329-3p	dbDEMC	39	hsa-mir-302a-5p	unconfirmed
15	hsa-mir-590-5p	dbDEMC	40	hsa-mir-450a-1-3p	unconfirmed
16	hsa-mir-383-5p	dbDEMC	41	hsa-mir-539-5p	dbDEMC
17	hsa-mir-211-5p	dbDEMC	42	hsa-mir-425-5p	dbDEMC
18	hsa-mir-491-5p	dbDEMC	43	hsa-mir-339-5p	dbDEMC
19	hsa-mir-373-3p	dbDEMC	44	hsa-mir-455-5p	dbDEMC
20	hsa-mir-302c-3p	dbDEMC	45	hsa-mir-128-1-5p	dbDEMC
21	hsa-mir-16-2-3p	dbDEMC	46	hsa-mir-500a-5p	dbDEMC
22	hsa-mir-19b-2-5p	unconfirmed	47	hsa-mir-370-5p	dbDEMC
23	hsa-mir-92a-2-5p	dbDEMC	48	hsa-mir-376a-5p	dbDEMC
24	hsa-mir-454-5p	dbDEMC	49	hsa-mir-345-5p	dbDEMC
25	hsa-mir-508-5p	dbDEMC	50	hsa-mir-584-5p	dbDEMC

Vector representation of miRNA sequences

To more comprehensively describe the features of miRNAs, we introduced the sequence information of the miRNA. In this study, we downloaded all miRNA sequences in MAN from the miRbase[60] and converted miRNA sequences to vectors by the k-mers method. The k-mers could divide the sequence into a train of subsequences with k bases.[61] Given a sequence of length m, the sequence could be divided into $m - k + 1$ k-mers. In this experiment, conjoint triads (3-mer) of miRNA were extracted from sequences. There are four bases of miRNA: A, C, G and U, therefore, 3-mers could split the sequence of miRNA into AAA, AAC, …, UUU. First, dividing the miRNA sequence into some conjoint triads was based on a slipping window. Then, we

Table 7. Prediction of the top 50 miRNAs related to colon neoplasms based on known miRNA-disease associations in HMDD database

Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	hsa-mir-182-5p	dbDEMC	26	hsa-mir-484	dbDEMC
2	hsa-mir-29c-5p	dbDEMC	27	hsa-mir-452-5p	dbDEMC
3	hsa-mir-193b-5p	dbDEMC	28	hsa-mir-27b-5p	dbDEMC
4	hsa-mir-206	dbDEMC	29	hsa-mir-30e-5p	dbDEMC
5	hsa-mir-122-5p	dbDEMC	30	hsa-mir-134-5p	dbDEMC
6	hsa-mir-214-5p	dbDEMC	31	hsa-mir-181c-5p	dbDEMC
7	hsa-mir-139-5p	dbDEMC	32	hsa-mir-99b-5p	dbDEMC
8	hsa-mir-497-5p	dbDEMC	33	hsa-mir-99a-5p	dbDEMC
9	hsa-mir-34c-5p	dbDEMC	34	hsa-mir-373-5p	dbDEMC
10	hsa-mir-183-5p	dbDEMC	35	hsa-mir-212-5p	dbDEMC
11	hsa-mir-423-5p	dbDEMC	36	hsa-mir-144-5p	dbDEMC
12	hsa-mir-100-5p	dbDEMC	37	hsa-mir-92a-2-5p	dbDEMC
13	hsa-mir-16-5p	dbDEMC	38	hsa-mir-92b-5p	dbDEMC
14	hsa-mir-9-5p	dbDEMC	39	hsa-mir-381-5p	unconfirmed
15	hsa-mir-149-5p	dbDEMC	40	hsa-mir-135a-5p	dbDEMC
16	hsa-mir-491-5p	dbDEMC	41	hsa-mir-10a-5p	dbDEMC
17	hsa-mir-124-5p	dbDEMC	42	hsa-mir-199b-5p	dbDEMC
18	hsa-mir-130b-5p	dbDEMC	43	hsa-mir-301a-5p	unconfirmed
19	hsa-mir-34b-5p	dbDEMC	44	hsa-mir-425-5p	dbDEMC
20	hsa-mir-146b-5p	dbDEMC	45	hsa-mir-542-5p	dbDEMC
21	hsa-mir-199a-5p	dbDEMC	46	hsa-mir-20b-5p	dbDEMC
22	hsa-mir-342-5p	dbDEMC	47	hsa-mir-340-5p	dbDEMC
23	hsa-mir-494-5p	dbDEMC	48	hsa-mir-181b-2-3p	unconfirmed
24	hsa-mir-26a-5p	dbDEMC	49	hsa-mir-338-5p	dbDEMC
25	hsa-mir-26b-5p	dbDEMC	50	hsa-mir-367-5p	unconfirmed

calculated the frequency of each sub-sequence and normalized these data. In this way, we converted the miRNA sequence information into a 64-dimensional numerical vector to represent miRNA attribute information.

Disease semantic similarity

To accurately describe the features of diseases, we obtained the disease semantic similarity information from the Medical Subject Headings (MeSH) database,[62] which provided an effective disease classification system. In this system, diseases could be represented by related directed acyclic graph (DAG).[63] The relationship between two diseases could be indicated by a directed edge pointing to child nodes by parent nodes. Suppose $DAG(D) = (D, N(D), E(D))$, where $N(D)$ indicates the node set containing all diseases of $DAG(D)$ and $E(D)$ indicates the edge set of all relationships of $DAG(D)$. The semantic value of disease D was contributed by disease T as the formula

$$\begin{cases} D_D(T) = 1 \quad if\ T = D \\ D_D(T) = max\{\theta * D_D(T') | T' \in childrenofT\}\ if\ T \neq D \end{cases} \quad (1)$$

Table 8. Prediction of the top 50 miRNAs related to lymphoma based on known miRNA-disease associations in HMDD database

Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	hsa-mir-34a-5p	dbDEMC	26	hsa-let-7b-5p	dbDEMC
2	hsa-mir-125b-5p	dbDEMC	27	hsa-mir-96-5p	dbDEMC
3	hsa-mir-107	dbDEMC	28	hsa-let-7g-5p	dbDEMC
4	hsa-mir-27a-5p	unconfirmed	29	hsa-mir-429	unconfirmed
5	hsa-mir-195-5p	dbDEMC	30	hsa-mir-192-5p	dbDEMC
6	hsa-mir-145-5p	dbDEMC	31	hsa-mir-125b-2-3p	dbDEMC
7	hsa-mir-106b-5p	dbDEMC	32	hsa-mir-30b-5p	dbDEMC
8	hsa-let-7a-5p	dbDEMC	33	hsa-mir-424-5p	dbDEMC
9	hsa-mir-29a-5p	dbDEMC	34	hsa-mir-146b-5p	dbDEMC
10	hsa-mir-182-5p	dbDEMC	35	hsa-mir-24-3p	dbDEMC
11	hsa-mir-34b-5p	dbDEMC	36	hsa-mir-339-5p	dbDEMC
12	hsa-mir-205-5p	dbDEMC	37	hsa-mir-148a-5p	dbDEMC
13	hsa-mir-9-5p	dbDEMC	38	hsa-mir-100-5p	dbDEMC
14	hsa-mir-183-5p	dbDEMC	39	hsa-mir-23a-5p	dbDEMC
15	hsa-mir-106a-5p	dbDEMC	40	hsa-mir-206	dbDEMC
16	hsa-let-7c-5p	dbDEMC	41	hsa-mir-199b-5p	dbDEMC
17	hsa-mir-218-5p	dbDEMC	42	hsa-mir-335-5p	dbDEMC
18	hsa-mir-141-5p	unconfirmed	43	hsa-mir-181b-5p	dbDEMC
19	hsa-mir-15b-5p	dbDEMC	44	hsa-mir-34c-5p	dbDEMC
20	hsa-mir-223-5p	dbDEMC	45	hsa-mir-214-5p	dbDEMC
21	hsa-mir-124-5p	dbDEMC	46	hsa-mir-30c-5p	dbDEMC
22	hsa-mir-30a-5p	dbDEMC, miR2Disease	47	hsa-mir-181d-5p	dbDEMC
23	hsa-mir-340-5p	dbDEMC	48	hsa-let-7e-5p	dbDEMC
24	hsa-mir-378a-5p	dbDEMC	49	hsa-mir-191-5p	dbDEMC
25	hsa-mir-196a-5p	dbDEMC	50	hsa-mir-125b-1-3p	dbDEMC

Table 9. The number of different types of nodes in MAN

Node	Number of nodes
miRNA	1,023
Disease	2,026
Drug	1,025
Protein	1,647
lncRNA	769
Total	6,528

Here, θ is the semantic contribution factor; the contribution value of D to itself is set as 1. Therefore, we can obtain the sum $DV(D)$ of D:

$$DV(D) = \sum_{T \in N_D} D_D(T) \quad (2)$$

According to the assumption that diseases with more same parts in their *DAG*s should hold higher similarity of them, we can obtain the semantic similarity among a and b by the following formula:

$$S(a,b) = \frac{\sum_{T \in N_a \cap N_b} (D_a(T) + D_b(T))}{DV(a) + DV(b)} \quad (3)$$

Then, we used the disease semantic similarity to express the attribute information of disease, and this process exists dimensional reduction by stacked autoencoder. The attribute information of diseases is also converted as a 64-dimensional vector.

Diffusion-based network embedding

In order to extract the comprehensive feature from the MAN, we adopted a diffusion-based network embedding. First of all, the complex heterogeneous network was constructed, including 6,528 nodes and 102,261 edges. Then, the 6,528-dimensional frequency vector before and after each node in the graph was obtained by the diffusion progress of the subgraph. To unify the dimensions of the feature vector, we used a neural network to process these frequency vectors. The input of the neural network is 6,528 one-hot vectors, and the output is the vector fusing the before and after frequency vector. Finally, we obtained a 64-dimensional vector to represent the behavior information of miRNAs and diseases.

Table 10. The number of different types of associations in MAN

Association	Database	Amounts of relationships
miRNA-disease	HMDD[51]	16,427
miRNA-protein	miRTarBase[52]	4,944
Drug-protein	DrugBank[53]	11,107
lncRNA-disease	lncRNADisease[54], lncRNASNP2[55]	1,264
Protein-protein	STRING[56]	19,237
miRNA-lncRNA	lncRNASNP2[55]	8,374
lncRNA-protein	lncRNA2Target[57]	690
Drug-disease	CTD[58]	18,416
Protein-disease	DisGeNET[59]	25,087
Total	–	105,546

The diffusion process for generating sequences

Previous studies have indicated that Random Walk is a depth-first algorithm that could repeatedly visit nodes. However, the original network structure is hardly reflected by the node similarity defined by a random walk. The diffusion could efficiently detect the network structure by generating more informative traces.[64]

Suppose a given graph is defined as $G(V, E)$, where V indicates the vertices set containing all nodes of G and E represents the edge set of G. The diffusion graph is defined as $\widehat{G}$, and the seed node is v_i. The maximal walking step is supposing as k. In every step, we chose a random node v_i from $\widehat{G}$ as diffusion source and v_j, a random neighbor of v_i, from G as diffusion object. Then, the diffusion object v_j and the edge (v_i, v_j) would be added to the diffusion graph $\widehat{G}$. We compared the generating sequence process of random walk and diffusion as shown in Figure 2. The walking step k is set as four in this example. In Figure 6 (1), the walker is starting from v_1, and the red node is the location of the walker. After a random walk process, it is obvious that the generated sequence is (v_1, v_2, v_4, v_2). In Figure 6 (2), we imitate the diffusion process by a graph with four nodes, and the initial diffusion source also is v_1. Unlike the single trace of random walk, all sampled nodes would be retained and may become a diffusion source in the next step in the diffusion process. The diffusion could generate a sequence of $((v_1), (v_1, v_2), (v_1, v_2, v_4), (v_1, v_2, v_4))$. As we can see in this example, if the step $k = 5$, the v_3 would not be visited in a random walk, which is a disadvantage of the single-trace algorithm. However, as the diffusion graph $\widehat{G}$ contained the neighbor of the node v_3, the v_3 is possible to be visited.

To obtain the sequence, we doubled each edge in the diffusion graph $\widehat{G}$. Then, the degree of each node is even, and there must be a Euler walk. This Euler walk would be the diffusion sequence that preserved the relationship of adjacent nodes. In this work, we set the walk length as 10, and the vertex-set-cardinality is equal to 80.

Feature extraction and network embedding

Given a set of node sequences, the feature is extracted by the sliding window. To more comprehensively detect the information of the network, we design the visit frequency vector M_1 by counting the frequency of other nodes before and after the node v_1. For example, there is a set of sequences with five nodes as follows:

$$v_1 - v_4 - v_5 - v_3 - v_4 - v_3 - v_2 - v_1$$

$$v_4 - v_5 - v_3 - v_1 - v_2 - v_3 - v_5 - v_4$$

$$v_2 - v_3 - v_2 - v_1 - v_4 - v_5 - v_4 - v_1$$

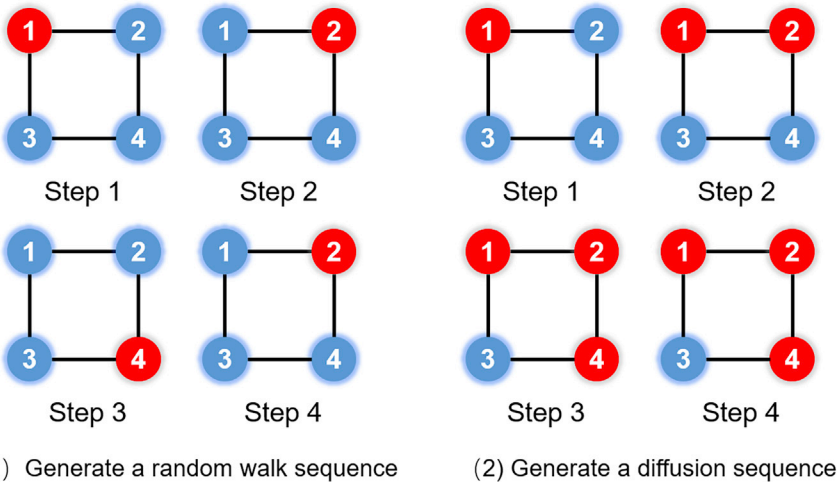

Figure 6. Example of generating sequence of random walk model and diffusion model

In this example, the size of the sliding window is set as 1, and we would demonstrate the visit frequency vector of v_3. As a result, the frequency vector of before and after the node v_3 is as follows:

$$M_3^{-1} = [0\ 2\ 0\ 1\ 2]$$

$$M_3^{+1} = [1\ 2\ 0\ 1\ 1]$$

where M_3^{-1} and M_3^{+1} represent the frequency of occurrence before and after v_3, respectively. Then, these two vectors would be concatenated as a visit frequency vector M_3.

In this study, we develop a neural network to learn an embedding from the feature. For each node v, we set as follows:

$$H_v = \alpha(\omega_{in} \times N_v + \beta_{in}) \quad (4)$$

Here, α and β_{in} are the regulated parameters, and ω_{in} is the incoming weight matrices of the hidden neurons. The output function is as shown:

$$\widehat{M}_v = \sigma(\omega_{out} \times H_v + \beta_{out}) \quad (5)$$

Then, we define the loss function as

$$L(M_v, \widehat{M}_v) = -M_v \log(\widehat{M}_v) \quad (6)$$

Finally, the minimization objective could be obtained as shown:

$$min \sum L(M_v, \sigma(\omega_{out} \times H_v + \beta_{out})) \quad (7)$$

Integration of feature information

To comprehensively describe the potential information of each node, we extracted feature descriptors from the two kinds of information of them. On the one hand is the attribute information, including the sequence information of miRNAs $KM(m(i))$ and the semantic similarity of diseases $SD(d(j))$. On the other hand, the behavior information $BM(m(i))$ and $BD(d(j))$ were extracted by the diffusion-based model. Finally, we integrated the above information into a comprehensive feature descriptor $F(m(i), d(j))$ based on known miRNA-disease associations from the HMDD v3.0 database. The feature descriptor can be defined by a 256-dimensional vector as follows:

$$F(m(i), d(j)) = [BM(m(i)), KM(m(i)), BD(d(j)), SD(d(j))] \quad (8)$$

Random Forest classifier

Random Forest is an important integrated machine learning algorithm proposed by Breiman et al.,[65] which can be used for classification and regression problems. The Random Forest has been widely used in bioinformatics with reliable performance.[66] The algorithm first randomly selects bootstrap samples from the original samples as the training set. Second, Random Forest randomly selects variables from each bootstrap sample and split nodes by the random subspace method. By this method, an unpruned classification tree is grown for each sample. Finally, Random Forest obtains prediction results by a majority vote according to these decision trees. Specifically, we adopted a 256-dimensional feature descriptor to represent each sample in the training set. In this study, we selected the optimal parameter n_{Tree} as 99 by the grid search method to implement the final classification prediction task.

ACKNOWLEDGMENTS

This work is supported is supported in part by the National Natural Science Foundation of China under grant 61702444; the West Light Foundation of the Chinese Academy of Sciences under grant 2018-XBQNXZ-B-008; the Chinese Postdoctoral Science Foundation under grant 2019M653804; the Tianshan Youth - Excellent Youth under grant 2019Q029; and by the Qingtan Scholar Talent Project of Zaozhuang University. The authors would like to thank all anonymous reviewers for their constructive advice.

AUTHOR CONTRIBUTIONS

H.-Y.L. wrote the paper; Z.-H.Y. and X.Y. designed the experiments; L.W. and Z.L. conducted the experiments.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Kloosterman, W.P., and Plasterk, R.H.A. (2006). The diverse functions of microRNAs in animal development and disease. Dev. Cell *11*, 441–450.
2. Ambros, V. (2001). microRNAs: tiny regulators with great potential. Cell *107*, 823–826.
3. Cheng, A.M., Byrom, M.W., Shelton, J., and Ford, L.P. (2005). Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. Nucleic Acids Res. *33*, 1290–1297.
4. Karp, X., and Ambros, V. (2005). Developmental biology. Encountering microRNAs in cell fate signaling. Science *310*, 1288–1289.
5. Miska, E.A. (2005). How microRNAs control cell division, differentiation and death. Curr. Opin. Genet. Dev. *15*, 563–568.
6. Xu, P., Guo, M., and Hay, B.A. (2004). MicroRNAs and the regulation of cell death. Trends Genet. *20*, 617–624.
7. Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.-H., and Liu, H. (2018). BNPMDA: bipartite network projection for MiRNA-disease association prediction. Bioinformatics *34*, 3178–3186.
8. Chen, X., Xie, D., Zhao, Q., and You, Z.-H. (2019). MicroRNAs and complex diseases: from experimental results to computational models. Brief Bioinform *20*, 515–539.
9. Chen, X., Wang, C.-.C., Yin, J., and You, Z.-H. (2018). Novel human miRNA-disease association inference based on random forest. Mol. Ther. Nucleic Acids *13*, 568–579.
10. Garzon, R., Marcucci, G., and Croce, C.M. (2010). Targeting microRNAs in cancer: rationale, strategies and challenges. Nat. Rev. Drug Discov. *9*, 775–789.
11. Gong, Y., Niu, Y., Zhang, W., and Li, X. (2019). A network embedding-based multiple information integration method for the MiRNA-disease association prediction. BMC Bioinformatics *20*, 468.
12. Huang, F., Yue, X., Xiong, Z., Yu, Z., Liu, S., and Zhang, W. (2020). Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations. Brief Bioinform *2020*, bbaa140.
13. Chen, X., Zhang, D.-H., and You, Z.-H. (2018). A heterogeneous label propagation approach to explore the potential associations between miRNA and disease. J. Transl. Med *16*, 348.
14. Zheng, K., You, Z.-H., Wang, L., and Guo, Z.-H. (2020). iMDA-BN: Identification of miRNA-Disease Associations based on the Biological Network and Graph Embedding Algorithm. Comp. Struct. Biotechnol. J *18*, 2391–2400.
15. Wong, L., You, Z.-H., Guo, Z.-H., Yi, H.-C., Chen, Z.-H., and Cao, M.-Y. (2020). MIPDH: A Novel Computational Model for Predicting microRNA–mRNA Interactions by DeepWalk on a Heterogeneous Network. ACS Omega *5*, 17022–17032.
16. Liu, Y.X., Wang, L., Liu, W.J., Zhang, H.T., Xue, J.H., Zhang, Z.W., and Gao, C.J. (2016). MiR-124-3p/B4GALT1 axis plays an important role in SOCS3-regulated growth and chemo-sensitivity of CML. J. Hematol. Oncol. *9*, 69.
17. Kumarswamy, R., Volkmann, I., and Thum, T. (2011). Regulation and function of miRNA-21 in health and disease. RNA Biol. *8*, 706–713.
18. Pereira, D.M., Rodrigues, P.M., Borralho, P.M., and Rodrigues, C.M. (2013). Delivering the promise of miRNA cancer therapeutics. Drug Discov. Today *18*, 282–289.
19. Xie, X., Liu, H., Wang, M., Ding, F., Xiao, H., Hu, F., Hu, R., and Mei, J. (2015). miR-342-3p targets RAP2B to suppress proliferation and invasion of non-small cell lung cancer cells. Tumour Biol. *36*, 5031–5038.
20. Chen, X., Gong, Y., Zhang, D.-H., You, Z.-H., and Li, Z.-W. (2018). DRMDA: deep representations-based miRNA–disease association prediction. J. Cell Mol. Med *22*, 472–485.
21. Zhang, W., Li, Z., Guo, W., Yang, W., and Huang, F. (2019). A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. IEEE/ACM Trans. Comput. Biol. Bioinform. Published online July 29, 2019. https://doi.org/10.1109/TCBB.2019/2931546.
22. Chen, X., Sun, L.-G., and Zhao, Y. (2020). NCMCMDA: miRNA–disease association prediction through neighborhood constraint matrix completion. Brief Bioinform *2020*, bbz159.
23. Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., Zhou, Y., and Cui, Q. (2019). HMDD v3.0: a database for experimentally supported human microRNA-disease associations. Nucleic Acids Res. *47* (D1), D1013–D1017.
24. 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.
25. Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H., and Teschendorff, A.E. (2017). dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. Nucleic Acids Res. *45* (D1), D812–D818.
26. Zheng, K., You, Z.H., Li, Y.R., Zhou, J.R., and Zeng, H.T. (2020). MISSIM: An incremental learning-based model with applications to the prediction of miRNA-disease association. IEEE/ACM Trans. Comput. Biol. Bioinform. Published online August 3, 2020. https://doi.org/10.1109/tcbb.2020.3013837.
27. Huang, Y.-A., You, Z.-H., Li, L.-P., Huang, Z.-A., Xiang, L.-X., Li, X.-F., and Lv, L.-T. (2017). EPMDA: an expression-profile based computational model for microRNA-disease association prediction. Oncotarget *8*, 87033–87043.
28. You, Z.-H., Wang, L.-P., Chen, X., Zhang, S., Li, X.-F., Yan, G.-Y., and Li, Z.-W. (2017). PRMDA: personalized recommendation-based MiRNA-disease association prediction. Oncotarget *8*, 85568–85583.
29. 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 *7*, 65257–65269.
30. Luo, J., and Xiao, Q. (2017). A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on. heterogeneous network *66*, 194–203.
31. Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., Liu, Y., and Wang, Y. (2010). Prioritization of disease microRNAs through a human phenome-microRNAome network. BMC Syst. Biol. *4* (*Suppl 1*), S2.
32. 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.
33. 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.
34. 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.
35. Chen, X., Wu, Q.-F., and Yan, G.-Y. (2017). RKNNMDA: ranking-based KNN for MiRNA-disease association prediction. RNA Biol. *14*, 952–962.
36. Ji, B.-Y., You, Z.-H., Cheng, L., Zhou, J.-R., Alghazzawi, D., and Li, L.-P. (2020). Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model. Sci. Rep. *10*, 6658.
37. Guo, Z.-H., Yi, H.-C., and You, Z.-H. (2019). Construction and Comprehensive Analysis of a Molecular Associations Network via lncRNA—miRNA—Disease—Drug—Protein Graph. Cells *8*, 866.
38. Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). DeepWalk: Online learning of social representations. KDD '14: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining *2014*, 701–710.
39. Chen, Z.-H., You, Z.-H., Guo, Z.-H., Yi, H.-C., Luo, G.-X., and Wang, Y.-B. (2020). Prediction of Drug–Target Interactions From Multi-Molecular Network Based on Deep Walk Embedding Model. Front. Bioeng. Biotechnol *8*, 338.
40. Chen, X., Liu, M.X., and Yan, G.Y. (2012). RWRMDA: predicting novel human microRNA-disease associations. Mol. Biosyst. *8*, 2792–2798.
41. 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 *10*, 1857–1866.

42. Chen, X., Huang, L., Xie, D., and Zhao, Q. (2018). EGBMMDA. extreme gradient boosting machine for miRNA-disease association prediction *9*, 3.
43. 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.
44. Zheng, K., You, Z.-H., Wang, L., Zhou, Y., Li, L.-P., and Li, Z.-W. (2020). DBMDA: A unified embedding for sequence-based miRNA similarity measure with applications to predict and validate miRNA-disease associations. Mol. Ther. Nucleic Acids *19*, 602–611.
45. 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.
46. 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.
47. Siegel, R.L., Miller, K.D., and Jemal, A. (2019). Cancer statistics, 2019. CA Cancer J. Clin. *69*, 7–34.
48. Yanaihara, N., Caplen, N., Bowman, E., Seike, M., Kumamoto, K., Yi, M., Stephens, R.M., Okamoto, A., Yokota, J., Tanaka, T., et al. (2006). Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell *9*, 189–198.
49. Ogata-Kawata, H., Izumiya, M., Kurioka, D., Honma, Y., Yamada, Y., Furuta, K., Gunji, T., Ohta, H., Okamoto, H., Sonoda, H., et al. (2014). Circulating exosomal microRNAs as biomarkers of colon cancer. PLoS ONE *9*, e92921.
50. Kwak, E.L., Bang, Y.-J., Camidge, D.R., Shaw, A.T., Solomon, B., Maki, R.G., Ou, S.H., Dezube, B.J., Jänne, P.A., Costa, D.B., et al. (2010). Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N. Engl. J. Med. *363*, 1693–1703.
51. Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., Zhou, Y., and Cui, Q. (2019). HMDD v3.0: a database for experimentally supported human microRNA-disease associations. Nucleic Acids Res. *47* (D1), D1013–D1017.
52. Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., Huang, W.-C., Sun, T.-H., Tu, S.-J., Lee, W.-H., et al. (2017). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res *46* (D1), D296–D302.
53. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res *46* (D1), D1074–D1082.
54. Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., and Cui, Q. (2013). LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res *41* (Database issue), D983–D986.
55. Miao, Y.R., Liu, W., Zhang, Q., and Guo, A.Y. (2018). lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. Nucleic Acids Res. *46* (D1), D276–D280.
56. Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res *45* (Database issue), D362–D368.
57. Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., Zhou, W., Liu, G., Jiang, H., and Jiang, Q. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Res. *47* (D1), D140–D144.
58. Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., McMorran, R., Wiegers, J., Wiegers, T.C., and Mattingly, C.J. (2019). The Comparative Toxicogenomics Database: update 2019. Nucleic Acids Res. *47* (D1), D948–D954.
59. Janet, P., Lex, B., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furling, L.I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res *45* (D1), D833–D839.
60. 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.
61. Pan, X., and Shen, H.-B. (2018). Learning distributed representations of RNA sequences and its application for predicting RNA-protein binding sites with a convolutional neural network. Neurocomputing *305*, 51–58.
62. Lipscomb, C.E. (2000). Medical subject headings (MeSH). Bull. Med. Libr. Assoc. *88*, 265–266.
63. Kalisch, M., and Bühlmann, P. (2012). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J. Mach. Learn. Res. *8*, 613–636.
64. Shi, Y., Lei, M., Zhang, P., and Niu, L. (2018). Diffusion Based Network Embedding. arXiv, arXiv.1805.03504v2.
65. Lawrence, R.L., Wood, S.D., and Sheley, R.L. (2006). Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest). Remote Sens. Environ *100*, 356–362.
66. Qi, Y. (2012). Random forest for bioinformatics. In Ensemble Machine Learning, C. Zhang and Y.Q. Ma, eds. (Springer), pp. 307–323.