Molecular Therapy
Nucleic Acids
Original Article

Novel Human miRNA-Disease Association Inference Based on Random Forest

Xing Chen,[1] Chun-Chun Wang,[1] Jun Yin,[1] and Zhu-Hong You[2]

[1]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China; [2]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Ürümqi 830011, China

Since the first microRNA (miRNA) was discovered, a lot of studies have confirmed the associations between miRNAs and human complex diseases. Besides, obtaining and taking advantage of association information between miRNAs and diseases play an increasingly important role in improving the treatment level for complex diseases. However, due to the high cost of traditional experimental methods, many researchers have proposed different computational methods to predict potential associations between miRNAs and diseases. In this work, we developed a computational model of Random Forest for miRNA-disease association (RFMDA) prediction based on machine learning. The training sample set for RFMDA was constructed according to the human microRNA disease database (HMDD) version (v.)2.0, and the feature vectors to represent miRNA-disease samples were defined by integrating miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity. The Random Forest algorithm was first employed to infer miRNA-disease associations. In addition, a filter-based method was implemented to select robust features from the miRNA-disease feature set, which could efficiently distinguish related miRNA-disease pairs from unrelated miRNA-disease pairs. RFMDA achieved areas under the curve (AUCs) of 0.8891, 0.8323, and 0.8818 ± 0.0014 under global leave-one-out cross-validation, local leave-one-out cross-validation, and 5-fold cross-validation, respectively, which were higher than many previous computational models. To further evaluate the accuracy of RFMDA, we carried out three types of case studies for four human complex diseases. As a result, 43 (esophageal neoplasms), 46 (lymphoma), 47 (lung neoplasms), and 48 (breast neoplasms) of the top 50 predicted disease-related miRNAs were verified by experiments in different kinds of case studies. The results of cross-validation and case studies indicated that RFMDA is a reliable model for predicting miRNA-disease associations.

INTRODUCTION

MicroRNAs (miRNAs) are a series of endogenous small non-coding RNAs (about 22 nt), which can suppress the expression of target genes by inducing the cutting degradation of mRNA, translation inhibition, or other modality regulation mechanism.[1–4] Line-4 and let-7, the first two miRNAs that were discovered more than 20 years ago,[5–7] act as positive regulators coincidently. Since then, as many studies on miRNAs have been carried out, a mass of miRNAs was found in viruses, green algae plants, and animals.[8] Furthermore, several studies have demonstrated that miRNAs are in connection with many important biological processes, such as cell growth,[9] cell death,[10] cell proliferation,[11] immune reaction,[12] signal transduction,[13] tumor invasion,[14] and viral infection.[15] Hence, it is no surprise that miRNAs could be associated with different kinds of diseases.[16]

With the development of biotechnology and accumulation of theories, more and more associations between miRNAs and diseases have been discovered. Yao et al.[17] found that miRNA-103 and miRNA-107 inhibit the translation of cofilin. Moreover, the decrease of miRNA-103 or miRNA-107 levels and the increase of cofilin protein levels happened at the same time in a transgenic mouse model of Alzheimer's disease. Gao et al.[18] discovered the phenomenon that deregulation of miRNA-145 and miRNA-199 expression happened in previous stages of hepatitis B virus (HBV)-associated multi-step hepatocarcinogenesis. In addition, miRNA-155 plays an important role in the induction of chronic gastritis and colitis and the T cell-mediated control of *Helicobacter pylori* infection.[19] A recent study also showed that miRNA-23, miRNA-24, and miRNA-27 contained underlying therapeutic factors in ischemic heart and vascular disorders disease.[20] Hence, there is no doubt that obtaining and taking advantage of association information between miRNAs and diseases could improve the treatment level for complex diseases. However, since experimental methods may consume plenty of time, numerous materials, and a lot of labor to find associations, proposing efficient computational methods based on existing databases is expected to significantly reduce the workload. Indeed, a growing number of computational models have been developed to predict potential associations between miRNAs and diseases in recent years.[21–26]

On the basis of a reasonable conjecture that functionally similar miRNAs tend to link with diseases that have similar phenotypes,

Received 28 May 2018; accepted 5 October 2018;
https://doi.org/10.1016/j.omtn.2018.10.005.

Correspondence: Xing Chen, School of Information and Control Engineering, China University of Mining and Technology, 1 Daxue Road, Xuzhou 221116, China.
E-mail: xingchen@amss.ac.cn
Correspondence: Zhu-Hong You, Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Ürümqi 830011, China.
E-mail: zhuhongyou@ms.xjb.ac.cn

lots of computational methods are proposed for predicting associations between miRNAs and diseases.[27–29] Jiang et al.[30] created a hypergeometric distribution-based computational method by integrating miRNA functional similarity network, disease phenotype similarity network, and known human miRNA-disease association network. However, only using the information of miRNA neighbors resulted in an unsatisfactory prediction performance of this model. Later, Shi et al.[31] built a computational model to identify unknown miRNA-disease associations by using the algorithm of random walk on the protein-protein interaction (PPI) network. Under the conjecture that a miRNA may be associated with a certain disease when target genes of the miRNA have connection with this disease, they predicted novel miRNA-disease associations by integrating PPI network, miRNA-target interaction network, and gene-disease interaction network. Mørk et al.[32] proposed a miRNA-protein-disease (miRPD) association prediction model to judge whether miRNAs link with diseases via considering the underlying proteins. After combining the known miRNA-disease associations, text-mined disease-protein associations, and predicted miRNA-protein associations into a scoring framework, they finally ranked the miRNA-disease pairs to infer miRNA-disease associations. Xu et al.[33] developed a comprehensive prioritization method for prioritizing miRNAs associated with disease, without using any known miRNA-disease associations, by integrating a few diseases' phenotypes with suited mRNAs and miRNA expression profiles. However, all of the above computational models have a common shortcoming because they relied much on miRNA-target interactions with high false-positive and false-negative ratios.

There were also some classical calculation models without depending on miRNA-target interactions. A prediction method named "human disease-related miRNA prediction" (HDMP) was presented.[34] To get a good performance, they computed the functional similarity of two miRNAs by integrating information of phenotype similarity between diseases and description of disease terms and distributing higher weights to miRNAs that were members of a miRNA cluster and/or family. Nevertheless, HDMP could not work for novel diseases without known associated miRNAs because the prediction process was mainly based on miRNAs' neighbors. Chen et al.[35] proposed a method named "random walk with restart for miRNA-disease association" (RWRMDA) to predict miRNA-disease associations by applying random walk with restart to find candidate miRNAs for the concerned disease. RWRMDA achieved a satisfactory performance, but it still failed to seek potential associated miRNAs for new diseases without any known related miRNAs. Later, Xuan et al.[36] presented a prediction method named "miRNAs associated with disease prediction" (MIDP), still based on random walk, which utilized different kinds of topologies and features of nodes. They extended the work on a miRNA-disease bilayer network so that their model could be used to predict candidate diseases without any known associated miRNAs. Chen et al.[37] further proposed a model named "within and between score for miRNA-disease association" (WBSMDA) prediction, which could avoid the above limitation as well. They calculated within scores, based on the information of known associated miRNA-disease pairs, and between scores, according to the information of unlabeled miRNA-disease pairs. WBSMDA could infer not only potential miRNAs for novel disease but also potential diseases for novel miRNA. Another computational method named "heterogeneous graph inference for miRNA-disease association" (HGIMDA) prediction was built by Chen et al.[38] to predict miRNA-disease associations through an iterative process. They obtained a convergent association probability matrix after some steps, since the two similarity matrices for miRNAs and diseases were respectively normalized properly. Similarly, HGIMDA could also work for new diseases as well as new miRNAs.

Li et al.[39] proposed a matrix completion algorithm named "matrix completion for miRNA-disease association" (MCMDA) prediction by updating the adjacency matrix more efficiently based on the known miRNA-disease association information only. Yu et al.[40] applied the maximizing information flow approach (MaxFlow) for the first time to predict miRNA-disease associations through integrating disease phenotypic and semantic similarity network, miRNA functional similarity network, and known miRNA-disease association network into a phenome-microRNAome network and maximizing network information flow. Chen et al.[41] presented a computational model of ranking-based K-nearest neighbor (KNN) for miRNA-disease association prediction (RKNNMDA). In the model, the K-nearest neighbor algorithm was used to search for k-nearest neighbors by combining known similarity networks. This model could also work for diseases (or miRNAs) without any known associated miRNAs (or diseases).

In addition, prediction models based on machine learning have also been frequently utilized to search for potential miRNA-disease associations. Xu et al.[42] proposed a model named "miRNA target-dysregulated network" (MTDN) based on support vector machine (SVM) to prioritize candidate disease-related miRNAs for prostate cancer. The algorithm MTDN was utilized to define four features of a miRNA, and then the SVM classifier divided miRNAs into two categories, positive and negative, as the result of prediction. Later, Chen and Yan[43] developed a semi-supervised learning method named "regularized least-squares for miRNA-disease association" (RLSMDA) prediction to infer miRNA-disease associations. Negative samples are not necessary in the model, but it's still hard to select optimal parameter values. Next Chen et al.[44] proposed a computational model called "restricted Boltzmann machine for multiple types of miRNA-disease association" (RBMMMDA) to predict different types of associations between miRNAs and diseases. Compared with former methods, which could only predict binary miRNA-disease associations, RBMMMDA could obtain both new miRNA-disease associations and corresponding association types.

As we know, Random Forest has been successfully applied to a wide range of bioinformatics problems, including protein or peptide identification,[45] *in vivo* transcription factor-binding prediction,[46] enhancer identification,[47] and functional annotation of non-coding SNPs.[48,49] In this study, we developed a novel efficient computational model of Random Forest for miRNA-disease association (RFMDA) prediction (motivated by the study of Cheng et al.[50]). First, we

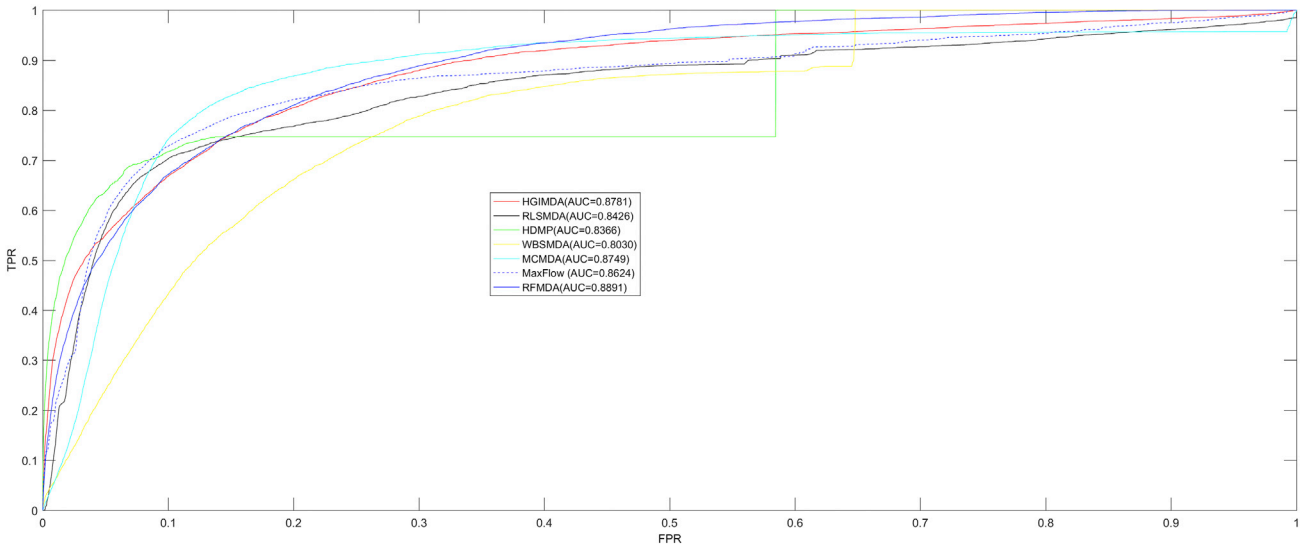

Figure 1. AUCs of RFMDA and HGIMDA, RLSMDA, HDMP, WBSMDA, MaxFlow, and MCMDA under Global LOOCV
As one can see, RFMDA achieved AUCs of 0.8891 under global LOOCV, which were higher than those of previous models.

constructed training samples for data preparation. Second, RFMDA made full use of biological information of miRNAs and diseases by integrating the miRNA functional similarity network or disease semantic similarity network with the Gaussian interaction profile kernel similarity network for miRNAs or diseases. Each miRNA-disease pair (m(i), d(j)) was represented by a feature vector based on the above similarity networks. Then, a feature selection method was used to cut down the dimensionality of feature vectors for efficiently distinguishing associated miRNA-disease pairs from unassociated miRNA-disease pairs and decreasing computational cost. Finally, a Random Forest prediction model could be obtained by training the samples mentioned in the first step.

To evaluate the performance of RFMDA, local and global leave-one-out cross-validations (LOOCVs) as well as 5-fold cross-validation were implemented. As a result, the areas under the curve (AUCs) of global and local LOOCVs were 0.8891 and 0.8323, respectively; and, the AUC obtained from 5-fold cross-validation was 0.8818 ± 0.0014. Besides, we implemented three types of case studies on esophageal neoplasms, lymphoma, lung neoplasms, and breast neoplasms. The top 10 and top 50 candidate miRNAs associated with these four diseases obtained from RFMDA were verified by experimental reports in some representative databases. As a result, 86% (esophageal neoplasms), 92% (lymphoma), 94% (lung neoplasms), and 96% (breast neoplasms) of the top 50 predicted miRNAs were respectively verified by recent experimental results. The data demonstrated that RFMDA is an excellent method to predict potential miRNA-disease associations.

RESULTS

Performance Evaluation

To evaluate the performance of RFMDA, LOOCVs and 5-fold cross-validation were utilized based on the known miRNA-disease associations in the human microRNA disease database (HMDD) version (v.) 2.0. The dataset contains 5,430 known miRNA-disease associations between 495 miRNAs and 383 diseases. In our model, all known miRNA-disease associations were treated as positive samples, while the unknown miRNA-disease pairs were treated as unlabeled samples. Global LOOCV and local LOOCV are two categories of LOOCV. For global LOOCV, each positive sample would be left out in turn as a test sample, and other positive samples were used to train the model. RFMDA would give a predicted score to the test sample and each unlabeled sample. After sorting all scores in decreasing order, we could obtain the ranking of the test sample. Finally, 5,430 rankings could be obtained by this way.

Then we drew the Receiver Operating Characteristic (ROC) curves by plotting the true positive rate (TPR, sensitivity) against the false positive rate (FPR, 1-specificity) with different thresholds. Sensitivity shows the percentage of test samples that was ranked in front of the given threshold, while specificity demonstrates the percentage of negative miRNA-disease associations whose ranks were lower than the given threshold. The ROC AUC was regarded as a standard for performance evaluation. A higher AUC indicates more excellent prediction performance of a prediction model. What makes a difference in local LOOCV is that the test sample was ranked with miRNAs that have no known association with the investigated disease according to the prediction scores.

As shown in Figures 1 and 2, RFMDA achieved an AUC of 0.8891, which is higher than AUCs of 0.8781 (HGIMDA), 0.8749 (MCMDA), 0.8624 (MaxFlow), 0.8426 (RLSMDA), 0.8366 (HDMP), and 0.8030 (WBSMDA) in global LOOCV. Besides, in local LOOCV, HGIMDA, MCMDA, MaxFlow, RLSMDA, HDMP, WBSMDA, MIDP, MiRAI, and RWRMDA obtained AUCs of 0.8077, 0.7718, 0.7774, 0.6953, 0.7702, 0.8031, 0.8196, 0.6299, and 0.7891, respectively. The AUCs

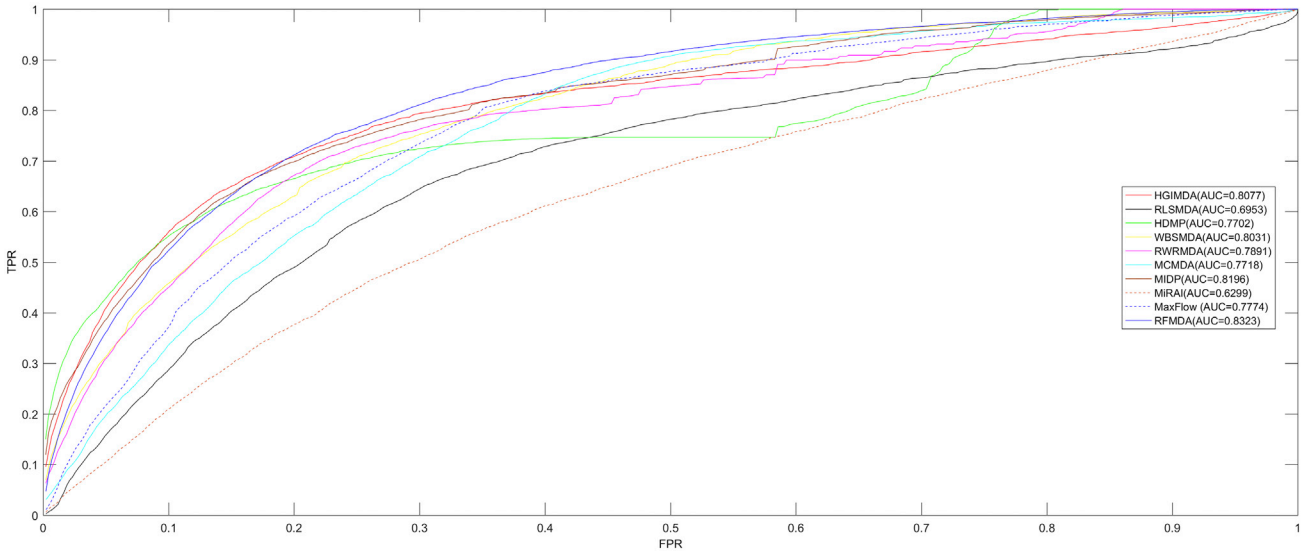

Figure 2. AUCs of RFMDA and HGIMDA, RLSMDA, HDMP, WBSMDA, RWRMDA, MaxFlow, MCMDA, MIDP, and MiRAI under Local LOOCV
As one can see, RFMDA achieved AUCs of 0.8323 under local LOOCV, which were higher than those of previous models.

of all the nine models were lower than RFMDA's AUC of 0.8323. As we can see, MIDP, RWRMDA, and MiRAI only appeared in local LOOCV comparison. On the one hand, MIDP and RWRMDA were based on random walk, which is a local method so that these two methods could not be used to predict for all diseases simultaneously. On the other hand, the association score between a disease (miRNA) and its candidate miRNAs (diseases) computed by MiRAI was extremely correlated with how many known miRNAs (diseases) associated with the disease (miRNAs). For a disease (miRNA) with more known associated miRNAs (diseases), the prediction scores between the disease (miRNA) and its candidate miRNAs (diseases) would be higher. Therefore, it is unfair to compare the prediction scores obtained from different diseases. MiRAI obtained a lower AUC after being implemented on our training dataset because the model suffered from the data sparsity problem. In our training dataset, the majority of 383 diseases (495 miRNAs) were associated with only a few miRNAs (diseases). However, MiRAI was implemented on the dataset that included 83 diseases with at least 20 known associated miRNAs for each in the original literature.[51] It is obvious that AUCs of RFMDA were higher than all of the previous methods mentioned above both in local and global LOOCVs. Both the global and local LOOCVs showed the excellent prediction performance of our model.

As for 5-fold cross-validation, we evenly divided positive samples into 5 parts, and each part would be treated as test samples in turn; and, each miRNA-disease pair in the test sample would be ranked with all unlabeled samples based on their prediction scores. The whole process was repeated 100 times to avoid evaluation bias. RFMDA achieved an AUC of 0.8818 ± 0.0014 in 5-fold cross-validation. The average AUC of 0.8818 under 100 cross-validation is still higher than the average AUCs of 0.8767 (MCMDA), 0.8579 (MaxFlow), 0.8569 (RLSMDA), 0.8342 (HDMP), and 0.8185 (WBSMDA). The average AUC revealed the superiority of our model, and the SD of 0.0014 demonstrated the stability of RFMDA.

Case Studies

To further evaluate the prediction performance of our model, we carried out three types of case studies on four diseases. The first type was implemented on esophageal neoplasms and lymphoma. Here, all known miRNA-disease associations in the HMDD v.2.0 were put into the training set of RFMDA. We selected the top 50 predicted miRNAs associated with the investigated disease based on their prediction scores, and then we validated them in another two databases, namely, dbDEMC[52] and miR2Disease.[53]

Esophageal neoplasms ranked eighth in the most common cancers and sixth in cancer mortality all over the world, according to the literature.[54] In the United States, about 10 in 100,000 people die of esophageal neoplasms every year, and the number of male patients was about four times as many as female patients.[55] Diagnosing the disease in the early stages would enhance the survival rate of patients.[56] Many experiments have confirmed that there are many miRNAs related to esophageal neoplasms. For example, the methylation ratios of miRNA-34a, miRNA-34b/c, and miRNA-129-2 are 66.67%, 40.74%, and 96.30%, respectively, in esophageal squamous cell carcinoma, which are obviously higher than those in non-tumor tissues.[57] We selected esophageal neoplasm as an example in the first type of case study, then RFMDA was implemented to predict miRNAs potentially associated with the disease. As a result, 10 of the top 10 and 43 of the top 50 predicted miRNAs were verified by experimental data in dbDEMC or miR2Disease (see Table 1).

Lymphoma is a cancer that starts in the lymphocytes or white blood cells.[58] These cells play a critical role in the immune system, which can help us fight against various diseases in the human body.[59]

Table 1. Top 50 miRNAs Associated with Esophageal Neoplasms Were Predicted by RFMDA Based on Known Associations in the HMDD v.2.0

miRNA	Evidence	miRNA	Evidence
hsa-mir-127	dbDEMC	hsa-mir-30a	dbDEMC
hsa-let-7g	dbDEMC	hsa-mir-125b	dbDEMC
hsa-mir-222	dbDEMC	hsa-mir-7	dbDEMC
hsa-mir-221	dbDEMC	hsa-mir-18a	dbDEMC
hsa-mir-30c	dbDEMC	hsa-mir-95	dbDEMC
hsa-mir-146b	dbDEMC	hsa-let-7i	dbDEMC
hsa-mir-372	dbDEMC	hsa-mir-181a	dbDEMC
hsa-mir-181b	dbDEMC	hsa-mir-204	dbDEMC
hsa-mir-10b	dbDEMC	hsa-mir-107	unconfirmed
hsa-mir-93	dbDEMC	hsa-mir-451	dbDEMC and miR2Disease
hsa-mir-16	dbDEMC	hsa-mir-122	dbDEMC
hsa-mir-200b	dbDEMC	hsa-mir-335	unconfirmed
hsa-mir-142	dbDEMC	hsa-let-7f	dbDEMC
hsa-mir-191	dbDEMC	hsa-mir-29a	unconfirmed
hsa-mir-9	dbDEMC	hsa-mir-139	dbDEMC
hsa-mir-199b	dbDEMC	hsa-mir-140	dbDEMC
hsa-mir-137	dbDEMC	hsa-mir-218	dbDEMC
hsa-mir-20b	dbDEMC	hsa-mir-135a	unconfirmed
hsa-mir-132	dbDEMC	hsa-mir-125a	dbDEMC
hsa-mir-18b	dbDEMC	hsa-mir-194	dbDEMC
hsa-mir-449a	unconfirmed	hsa-mir-29b	dbDEMC and miR2Disease
hsa-mir-449b	unconfirmed	hsa-mir-30e	dbDEMC
hsa-mir-106a	dbDEMC	hsa-mir-27b	unconfirmed
hsa-mir-373	dbDEMC and miR2Disease	hsa-mir-193b	dbDEMC
hsa-mir-224	dbDEMC	hsa-mir-195	dbDEMC

The top 1–25 related miRNAs are recorded in the first column, and the top 26–50 related miRNAs are recorded in the third column. As we can see 10, 19, and 43 of the top 10, top 20, and top 50 were verified by databases.

Table 2. Top 50 miRNAs Associated with Lymphoma Were Predicted by RFMDA Based on Known Associations in the HMDD v.2.0

miRNA	Evidence	miRNA	Evidence
hsa-let-7b	dbDEMC	hsa-mir-199a	dbDEMC
hsa-mir-125b	unconfirmed	hsa-mir-106a	dbDEMC and miR2Disease
hsa-let-7a	dbDEMC	hsa-mir-29a	dbDEMC
hsa-let-7c	dbDEMC	hsa-mir-182	dbDEMC
hsa-mir-34a	dbDEMC	hsa-mir-23b	dbDEMC
hsa-let-7e	dbDEMC and miR2Disease	hsa-mir-15b	dbDEMC
hsa-mir-106b	dbDEMC	hsa-mir-29b	dbDEMC
hsa-let-7d	dbDEMC	hsa-mir-27a	dbDEMC
hsa-mir-145	dbDEMC and miR2Disease	hsa-mir-141	dbDEMC
hsa-let-7i	dbDEMC	hsa-mir-22	dbDEMC
hsa-mir-143	dbDEMC and miR2Disease	hsa-mir-195	dbDEMC
hsa-mir-222	dbDEMC	hsa-mir-30a	dbDEMC
hsa-mir-221	dbDEMC and miR2Disease	hsa-mir-196a	dbDEMC
hsa-mir-223	dbDEMC	hsa-mir-1	dbDEMC
hsa-mir-127	dbDEMC and miR2Disease	hsa-mir-32	dbDEMC
hsa-mir-25	dbDEMC	hsa-mir-95	dbDEMC and miR2Disease
hsa-mir-30c	dbDEMC	hsa-mir-34b	dbDEMC
hsa-mir-146b	unconfirmed	hsa-mir-148a	dbDEMC
hsa-let-7f	dbDEMC	hsa-mir-183	dbDEMC
hsa-mir-181b	dbDEMC	hsa-mir-10b	dbDEMC
hsa-mir-214	dbDEMC	hsa-mir-132	dbDEMC
hsa-mir-191	dbDEMC	hsa-mir-133a	dbDEMC
hsa-let-7g	dbDEMC	hsa-mir-199b	dbDEMC
hsa-mir-34c	unconfirmed	hsa-mir-335	dbDEMC
hsa-mir-100	dbDEMC	hsa-mir-372	unconfirmed

The top 1–25 related miRNAs are recorded in the first column, and the top 26–50 related miRNAs are recorded in the third column. As we can see 9, 18, and 46 of the top 10, top 20, and top 50 were verified by databases.

Hodgkin lymphoma and non-Hodgkin lymphoma are two main kinds of lymphoma that can happen in both children and adults.[60] Recently, many miRNAs have been verified to be associated with lymphoma in different mechanisms. For example, plasma miRNA-92a values in non-Hodgkin lymphoma were about 5%, which were far less than those in heathy subjects.[61] Besides, several miRNAs were discovered significantly overexpressed in splenic marginal zone lymphoma, including miRNA-21, miRNA-155, and miRNA-146a.[62] We took lymphoma as another example in the first type of case study, and we utilized RFMDA to predict lymphoma-associated miRNAs. Finally, 9 of the top 10 and 46 of the top 50 predicted miRNAs were confirmed by experimental data in dbDEMC or miR2Disease (see Table 2).

RFMDA was also implemented to predict potential miRNAs for 381 other diseases in the HMDD v.2.0 apart from esophageal neoplasms and lymphoma. The whole prediction list is shown in Table S1. The table includes three kinds of information: the disease, the miRNA, and the predicted association score.

To prove the ability of our model in predicting new diseases without known associated miRNAs, we selected lung neoplasm as an example in the second type of case study. Here, before training the model, we removed all known associations of lung neoplasms. Then, we ranked all the 495 miRNAs based on their predicted association scores, and we validated the top 50 miRNAs in the HMDD v.2.0, dbDEMC, and miR2Disease. As a result, 10 of the top 10 and 47 of the top 50 miRNAs were confirmed by these databases (see Table 3).

Table 3. Top 50 miRNAs Associated with Lung Neoplasms Were Predicted by RFMDA after Hiding All Known Associations about Lung Neoplasms Based in the HMDD v.2.0

miRNA	Evidence	miRNA	Evidence
hsa-mir-133a	dbDEMC and HMDD	hsa-mir-192	dbDEMC, miR2Disease, and HMDD
hsa-mir-150	dbDEMC, miR2Disease, and HMDD	hsa-mir-130a	dbDEMC and miR2Disease
hsa-mir-196a	dbDEMC and HMDD	hsa-mir-10a	dbDEMC
hsa-mir-210	dbDEMC, miR2Disease, and HMDD	hsa-mir-200c	dbDEMC, miR2Disease, and HMDD
hsa-mir-182	dbDEMC, miR2Disease, and HMDD	hsa-mir-148a	dbDEMC and HMDD
hsa-mir-204	miR2Disease	hsa-mir-17	miR2Disease and HMDD
hsa-mir-100	dbDEMC and HMDD	hsa-mir-146a	dbDEMC, miR2Disease, and HMDD
hsa-mir-199b	dbDEMC, miR2Disease, and HMDD	hsa-mir-206	HMDD
hsa-mir-196b	dbDEMC	hsa-mir-203	dbDEMC, miR2Disease, and HMDD
hsa-mir-31	dbDEMC, miR2Disease, and HMDD	hsa-mir-20a	dbDEMC, miR2Disease, and HMDD
hsa-mir-335	miR2Disease and HMDD	hsa-mir-26a	dbDEMC, miR2Disease, and HMDD
hsa-mir-30a	miR2Disease and HMDD	hsa-mir-302b	dbDEMC
hsa-mir-1	dbDEMC, miR2Disease, and HMDD	hsa-mir-224	dbDEMC, miR2Disease, and HMDD
hsa-mir-296	dbDEMC	hsa-mir-302c	dbDEMC
hsa-mir-205	dbDEMC, miR2Disease, and HMDD	hsa-mir-181a	dbDEMC and HMDD
hsa-mir-27a	dbDEMC and HMDD	hsa-mir-221	dbDEMC and HMDD
hsa-mir-21	dbDEMC, miR2Disease, and HMDD	hsa-mir-95	miR2Disease and HMDD
hsa-mir-183	dbDEMC, miR2Disease, and HMDD	hsa-mir-143	dbDEMC, miR2Disease, and HMDD
hsa-mir-22	miR2Disease and HMDD	hsa-mir-32	miR2Disease and HMDD
hsa-mir-200a	dbDEMC, miR2Disease, and HMDD	hsa-mir-135b	dbDEMC and HMDD
hsa-mir-181b	dbDEMC and HMDD	hsa-mir-135a	dbDEMC and HMDD
hsa-mir-146b	miR2Disease and HMDD	hsa-mir-302a	unconfirmed
hsa-mir-107	dbDEMC and HMDD	hsa-mir-7	miR2Disease and HMDD
hsa-mir-34c	dbDEMC and HMDD	hsa-mir-218	dbDEMC, miR2Disease, and HMDD
hsa-mir-372	unconfirmed	hsa-mir-491	unconfirmed

The top 1–25 related miRNAs are recorded in the first column, and the top 26–50 related miRNAs are recorded in the third column. As we can see 10, 20, and 47 of the top 10, top 20, and top 50 were verified by databases.

We took breast neoplasms (BNs) as an example in the third type of case study to evaluate the performance of RFMDA using another miRNA-disease association database. Here, we used the associations in the HMDD v.1.0 as our training set. The HMDD v.1.0 contains 1,395 known associations between 271 miRNAs and 137 diseases. The whole prediction process was similar to the first type of case study. Finally, the data showed that 10 of the top 10 and 48 of the top 50 predicted miRNAs were confirmed by experimental data recorded in the HMDD v.2.0, dbDEMC, and miR2Disease (see Table 4).

DISCUSSION

With the development of experimental technologies and computational tools, more and more miRNAs have been discovered in recent years. Various associations between miRNAs and diseases have attracted the attention of researchers. These associations play an important role in the prevention, diagnosis, and treatment of complex human diseases. However, using traditional experimental methods to find miRNA-disease associations may be expensive and inefficient. Thus, we proposed an efficient computational model of RFMDA based on machine learning to predict potential miRNA-disease associations.

Positive samples and negative samples were selected from known miRNA-disease associations and unlabeled miRNA-disease pairs, respectively, according to the HMDD v.2.0. We represented each miRNA-disease pair as a feature vector by integrating information of miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity of miRNA and disease. RFMDA predicted unknown miRNA-disease associations by labeling them with scores after implementing the Random Forest algorithm. Based on LOOCV and 5-fold cross-validation, RFMDA obtained more excellent performance than lots of previous reliable computational models, such as MiRAI, MIDP, RWRMDA, MCMDA, WBSMDA, HDMP, RLSMDA, MaxFlow, and HGIMDA. In addition, the results of three types of case studies for four complex human

Table 4. Top 50 miRNAs Associated with Breast Neoplasms Were Predicted by RFMDA Based on Known Associations in the HMDD v.1.0

miRNA	Evidence	miRNA	Evidence
hsa-mir-223	dbDEMC and HMDD	hsa-mir-520b	dbDEMC and HMDD
hsa-mir-24	dbDEMC and HMDD	hsa-mir-23b	dbDEMC and HMDD
hsa-let-7b	dbDEMC and HMDD	hsa-mir-148a	dbDEMC, miR2Disease, and HMDD
hsa-mir-126	dbDEMC, miR2Disease, and HMDD	hsa-mir-135a	dbDEMC and HMDD
hsa-mir-373	dbDEMC, miR2Disease, and HMDD	hsa-mir-182	dbDEMC, miR2Disease, and HMDD
hsa-mir-32	dbDEMC	hsa-mir-142	unconfirmed
hsa-mir-16	dbDEMC and HMDD	hsa-let-7i	dbDEMC, miR2Disease, and HMDD
hsa-let-7c	dbDEMC and HMDD	hsa-mir-128b	miR2Disease
hsa-mir-150	dbDEMC	hsa-mir-335	dbDEMC, miR2Disease, and HMDD
hsa-mir-29c	dbDEMC, miR2Disease, and HMDD	hsa-mir-15b	dbDEMC
hsa-mir-372	dbDEMC	hsa-mir-98	dbDEMC and miR2Disease
hsa-mir-101	dbDEMC, miR2Disease, and HMDD	hsa-mir-181a	dbDEMC, miR2Disease, and HMDD
hsa-let-7e	dbDEMC and HMDD	hsa-mir-183	dbDEMC and HMDD
hsa-mir-106a	dbDEMC	hsa-mir-26a	dbDEMC, miR2Disease, and HMDD
hsa-let-7g	dbDEMC and HMDD	hsa-mir-100	dbDEMC and HMDD
hsa-mir-99b	dbDEMC	hsa-mir-107	dbDEMC and HMDD
hsa-mir-192	dbDEMC	hsa-mir-224	dbDEMC and HMDD
hsa-mir-30e	unconfirmed	hsa-mir-92b	dbDEMC
hsa-mir-199b	dbDEMC and HMDD	hsa-mir-95	dbDEMC
hsa-mir-27a	dbDEMC, miR2Disease, and HMDD	hsa-mir-22	dbDEMC, miR2Disease, and HMDD
hsa-mir-130a	dbDEMC	hsa-mir-196b	dbDEMC
hsa-mir-195	dbDEMC, miR2Disease, and HMDD	hsa-mir-191	dbDEMC, miR2Disease, and HMDD
hsa-mir-30a	miR2Disease and HMDD	hsa-mir-18b	dbDEMC and HMDD
hsa-mir-203	dbDEMC, miR2Disease, and HMDD	hsa-mir-186	dbDEMC
hsa-mir-92a	HMDD	hsa-mir-424	dbDEMC

The top 1–25 related miRNAs are recorded in the first column, and the top 26–50 related miRNAs are recorded in the third column. As we can see 10, 19, and 48 of the top 10, top 20, and top 50 were verified by databases.

diseases (esophageal neoplasms, lymphoma, lung neoplasms, and breast neoplasms) further demonstrated that RFMDA was a reliable prediction model.

There were several important factors that contributed to the satisfying performance of RFMDA. First, RFMDA made full use of biological information, including known miRNA-disease associations, miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity of miRNA and disease when we constructed feature vectors for miRNA-disease pairs. Second, each miRNA-disease pair was represented by a feature vector, and then a feature selection method was used to cut down the dimension of feature vector. The method could select robust features for each miRNA-disease pair. In this way, RFMDA could efficiently distinguish related miRNA-disease pairs from unrelated miRNA-disease pairs. Finally, the model of RFMDA had a good generalization ability, which benefitted from utilizing an unbiased estimator for generalization error in the Random Forest algorithm and that the parameters of Random Forest were easy to select.

However, some limitations still exist in the model of RFMDA. RFMDA requires training samples, including both positive samples and negative samples. As we know, it is difficult or even impossible to obtain reliable negative samples. We utilized a random selection method to select negative samples based on unknown miRNA-disease associations. That may influence the final result of prediction. Besides, in the HMDD v.2.0, there are only 5,430 known miRNA-disease associations, which are far less than unknown associations between 383 diseases and 495 miRNAs. Finally, not just similarity information can be used to the constructed feature vector, with a deeper understanding of mechanisms both for miRNAs and diseases. Thus, we believe that the performance of RFMDA will be much better in the future.

MATERIALS AND METHODS

Human miRNA-Disease Associations

The information of 5,430 known human miRNA-disease associations between 383 diseases and 495 miRNAs was obtained from the HMDD v.2.0.[63] We constructed an adjacency matrix A with 383 (nd) rows and 495 (nm) columns to briefly store the information of

known and unknown miRNA-disease associations between 383 diseases and 495 miRNAs. The element $A(d(i), m(j))$ is equal to 1 when miRNA $m(j)$ had been verified to be associated with disease $d(i)$, otherwise 0.

miRNA Functional Similarity

Under the assumption that functionally similar miRNAs tend to link with phenotypically similar diseases, Wang et al.[28] developed a method to compute the miRNA functional similarity score of two miRNAs. We downloaded the miRNA functional similarity scores from http://www.cuilab.cn/. Then, we constructed the miRNA functional similarity matrix FS with 495 rows and 495 columns, where the element $FS(m(i), m(j))$ denotes the functional similarity score between miRNA $m(i)$ and miRNA $m(j)$.

Disease Semantic Similarity Model 1

Medical subject headings (MeSH) disease descriptors were downloaded from the National Library of Medicine (https://www.nlm.nih.gov/), which furnished a rigorous system for disease classification. In the system, each disease could be described by a directed acyclic graph (DAG), in which the nodes represent diseases and each of the direct edges connects two nodes from parent node to child node. A disease D can be described as $DAG_D = (D, T_D, E_D)$, where T_D is a node set containing disease D and its ancestor diseases and E_D is an edge set containing the corresponding edges.[28] Actually, Xiang et al.[64] utilized MeSH gene descriptors to calculate dissimilarity between genes by the GenoMeSH algorithm. Here, we computed disease semantic similarity based on MeSH disease descriptors in another method according to previous study.[34] Specifically, we defined the contribution of disease t to the semantic value of disease D as follows.

$$\begin{cases} D1_D(t) = 1 & \text{if } t = D \\ D1_D(t) = \max\{\Delta * D1_D(t') \mid t' \in \text{children of } t\} & \text{if } t \neq D \end{cases} \quad \text{(Equation 1)}$$

where Δ is the semantic contribution decay factor. It will reduce the contribution of disease t if t is different from D. Besides, the contribution of disease D to its own semantic value is equal to 1.

Moreover, the semantic value $DV1(D)$ of disease D was defined as follows.

$$DV1(D) = \sum_{t \in T_D} D_D(t) \quad \text{(Equation 2)}$$

The semantic similarity value between disease $d(i)$ and $d(j)$ could be computed based on a conjecture that two diseases will be more similar if they share a larger part of their DAGs,

$$SS1(d(i), d(j)) = \frac{\sum_{t \in T_{d(i)} \cap T_{d(j)}} \left(D1_{d(i)}(t) + D1_{d(j)}(t)\right)}{DV1(d(i)) + DV1(d(j))}, \quad \text{(Equation 3)}$$

where $SS1$ is a disease semantic similarity matrix with 383 rows and 383 columns and the element $SS1(d(i),d(j))$ represents the semantic similarity of $d(i)$ and $d(j)$ based on disease semantic similarity model 1.

Disease Semantic Similarity Model 2

Each disease can be described as a hierarchical DAG in which the parent node represents a more general disease and the child node represents a more specific disease. According to disease semantic similarity model 1, the contributions of different diseases in the same layer of DAG_D to the semantic value of D are at a same level. However, these diseases may appear in other DAGs, and the number of DAGs in which they appear may be different. Thus, we believe that the contributions of these diseases should be distinguished. The contributions of diseases appearing in other DAGs more frequently should be less than specific diseases that appear in fewer DAGs. According to previous study,[34] the contribution of disease t to the semantic value of disease D can be calculated as follows.

$$D2_D(t) = -\log\left(\frac{\text{the number of DAGs including } t}{\text{the number of diseases}}\right) \quad \text{(Equation 4)}$$

The semantic similarity value between disease $d(i)$ and $d(j)$ was calculated similarly to the disease semantic similarity model 1 as follows,

$$SS2(d(i), d(j)) = \frac{\sum_{t \in T_{d(i)} \cap T_{d(j)}} \left(D2_{d(i)}(t) + D2_{d(j)}(t)\right)}{DV2(d(i)) + DV2(d(j))}, \quad \text{(Equation 5)}$$

where $DV2(d(i))$ and $DV2(d(j))$ are semantic values of $d(i)$ and $d(j)$, respectively, which can be calculated similarly to Equation 2. $SS2$ is another disease semantic similarity matrix with 383 rows and 383 columns, and the element $SS2(d(i), d(j))$ represents the semantic similarity of $d(i)$ and $d(j)$ based on disease semantic similarity model 2.

Gaussian Interaction Profile Kernel Similarity for Diseases

Under the assumption that similar diseases are more likely to be related with functionally similar miRNAs and vice versa, the Gaussian interaction profile kernel similarity for diseases can be computed.[65] We defined binary vector $IP(d(u))$ to represent the interaction profiles of disease $d(u)$ by observing whether $d(u)$ is associated with each of the 495 miRNAs. The binary vector $IP(d(u))$ is equivalent to the u-th row vector of adjacency matrix A. Then the Gaussian interaction profile kernel similarity between $d(u)$ and $d(v)$ was defined as follows,

$$KD(d(u), d(v)) = \exp\left(-\alpha_d \| IP(d(u)) - IP(d(v)) \|^2\right), \quad \text{(Equation 6)}$$

where parameter α_d was implemented to tune the kernel bandwidth, which was calculated via normalizing the original parameter α'_d as follows.

$$\alpha_d = \alpha'_d \Big/ \left(\frac{1}{nd} \sum_{i=1}^{nd} \| IP(d(i)) \|^2\right) \quad \text{(Equation 7)}$$

Gaussian Interaction Profile Kernel Similarity for miRNAs

The Gaussian profile kernel similarity between miRNAs was calculated similarly to the method of disease Gaussian interaction profile kernel similarity computation:

$$KM(m(u), m(v)) = \exp\left(- \alpha_m \| IP(m(u)) - IP(m(v)) \|^2 \right) \quad \text{(Equation 8)}$$

$$\alpha_m = \alpha'_m \Big/ \left(\frac{1}{nm} \sum_{i=1}^{nm} \| IP(m(i)) \|^2 \right), \quad \text{(Equation 9)}$$

where binary vector $IP(m(u))$ (or $IP(m(v))$) represents the interaction profiles of miRNA $m(u)$ (or $m(v)$) by observing whether $m(u)$ (or $m(v)$) is associated with each of the 383 diseases and is equivalent to the u-th (or v-th) column vector of adjacency matrix A.

Integrated Similarity for Diseases

To make full use of disease semantic similarity 1, disease semantic similarity 2, and disease Gaussian interaction profile kernel similarity, an integrated disease similarity matrix SD was constructed by integrating the above similarities. According to previous study,[37] the element $SD(d(u), d(v))$ represented integrated similarity between disease $d(u)$ and $d(v)$ and was defined as follows,

$$SD(d(u), d(v)) = \begin{cases} \dfrac{SS1(d(u), d(v)) + SS2(d(u), d(v))}{2} & \text{if } d(u) \text{ and } d(v) \text{ have semantic similarity} \\ KD(d(u), d(v)) & \text{otherwise} \end{cases}, \quad \text{(Equation 10)}$$

where $d(u)$ and $d(v)$ have semantic similarity if both $d(u)$ and $d(v)$ have their own DAGs.

Integrated Similarity for miRNAs

We integrated miRNA functional similarity and miRNA Gaussian interaction profile kernel similarity in a similar way into the integrated miRNA similarity. Thus, the integrated similarity between miRNA $m(i)$ and $m(j)$ was calculated as follows.

$$SM(m(i), m(j)) = \begin{cases} FS(m(i), m(j)) & \text{if } m(i) \text{ and } m(j) \text{ have functional similarity} \\ KM(m(i), m(j)) & \text{otherwise} \end{cases} \quad \text{(Equation 11)}$$

RFMDA

The RFMDA model was constructed based on Random Forest for predicting miRNA-disease associations, which can be divided into four steps (see Figure 3; motivated by the study of Cheng et al.[50]): (1) selecting positive samples and negative samples; (2) constructing feature vectors to represent samples; (3) reducing the dimension of feature; (4) constructing the final prediction model and predicting miRNA-disease associations. The details of each step are described below.

First, we constructed a training sample set by selecting both positive samples and negative samples according to the ratio of 1:1. The 5,430 known associated miRNA-disease pairs were extracted from the HMDD v.2.0 to compose the positive sample set. Based on the assumption that if there is no confirmed association between a miRNA and a disease, then the miRNA and the disease constitute a negative sample, we randomly selected 5,430 negative samples to compose a negative sample set. The process of randomly selecting negative samples can be roughly divided into three steps. Above all, we randomly selected one of the 383 diseases; then we randomly selected a miRNA from the 495 miRNAs; finally, the disease and the miRNA constitute a negative sample if the miRNA-disease pair isn't a member of the 5,430 known miRNA-disease associations. We repeated this process until 5,430 negative samples were obtained. Our training sample set consisted of positive and negative sample sets.

Second, we represented each miRNA-disease pair by a feature vector. The semantic similarity 1, semantic similarity 2, and Gaussian interaction profile kernel similarity between each disease can be calculated.

For each disease, there were 383 integrated similarity values. We used integrated semantic similarity values as features to represent each disease by a 383-dimensional feature vector. For example, we represented disease $d(u)$ by a feature vector,

$$SD(d(u)) = (a_1, a_2, a_3, \ldots, a_{383}), \quad \text{(Equation 12)}$$

where $SD(d(u))$ is the u-th row vector of matrix SD, and a_v is the integrated similarity value between disease $d(u)$ and $d(v)$.

For each miRNA, we could obtain 495 integrated similarity values through integrating miRNA functional similarity and Gaussian interaction kernel profile similarity between the miRNA and all 495 miRNAs, including itself. A miRNA $m(i)$ could be represented by a 495-dimensional feature vector in a similar way to disease,

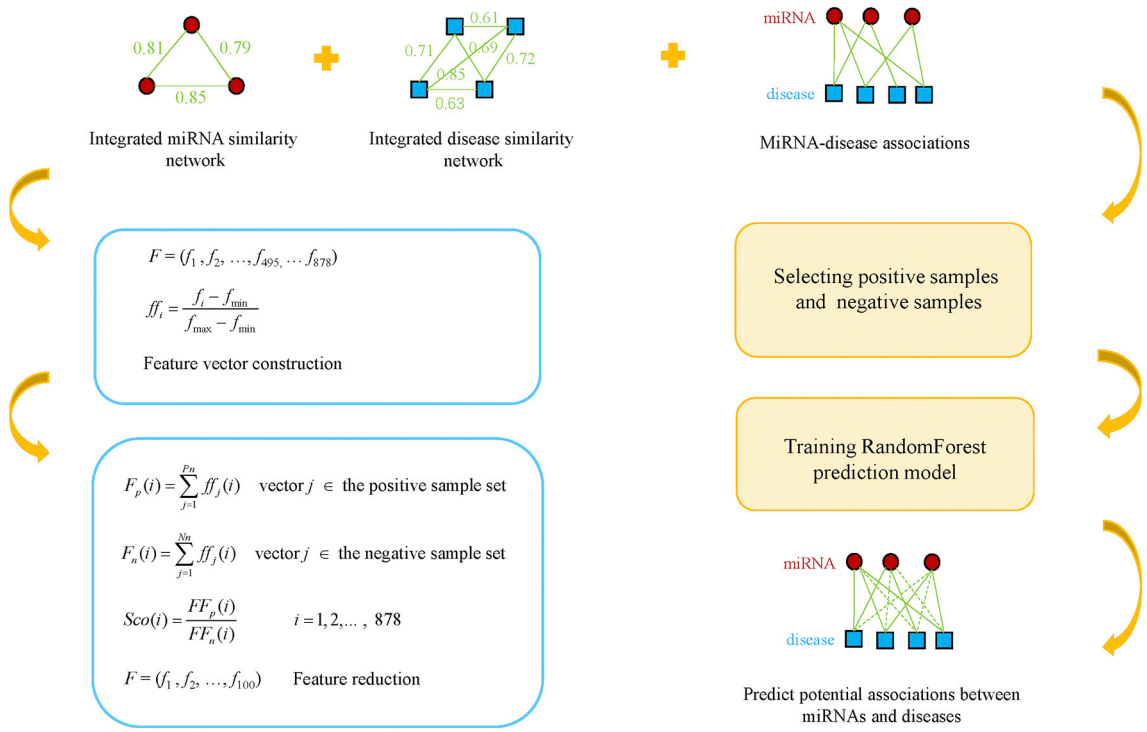

Figure 3. Flowchart of RFMDA Model to Predict Potential Associations between miRNAs and Diseases

$$SM(m(i)) = (b_1, b_2, b_3, \ldots, b_{495}), \quad \text{(Equation 13)}$$

where $SM(m(i))$ is the i-th column vector of matrix SM, and b_j is the integrated similarity value between miRNA $m(i)$ and $m(j)$.

Therefore, each miRNA-disease sample could be described by an 878-dimensional vector based on integrated similarity for the miRNA and integrated similarity for the disease,

$$F = (SM(m(i)), SD(d(u))) \quad \text{(Equation 14)}$$

$F = (f_1, f_2, \ldots, f_{495}, \ldots, f_{878})$, where $(f_1, f_2, \ldots, f_{495})$ represents the 495 integrated similarity values of the miRNA, and $(f_{496}, f_{497}, \ldots, f_{878})$ represents the 383 integrated similarity values of the disease. Then we normalized f_i to ff_i as follows,

$$ff_i = \frac{f_i - f_{\min}}{f_{\max} - f_{\min}}, \quad \text{(Equation 15)}$$

where $f_{\max}$ and $f_{\min}$ are the maximum and the minimum of all f_i ($i = 1, 2, \ldots, 878$).

Third, we reduced the dimension of feature vectors to reduce the computational cost and obtain more effective features. Our purpose was to select those discriminative features that either frequently appear in the positive sample set but seldom appear in the negative sample set or frequently appear in the negative sample set but seldom appear in the positive sample set. We used $ff_j(i)$ to denote the i-th feature of j-th vector and let $F_p(i)$ and $F_n(i)$ respectively represent the feature occurrence frequency in the positive sample set and negative sample set. $F_p(i)$ and $F_n(i)$ can be computed as follows,

$$F_p(i) = \sum_{j=1}^{Pn} ff_j(i) \text{ vector } j \in \text{the positive sample set} \quad \text{(Equation 16)}$$

$$F_n(i) = \sum_{j=1}^{Nn} ff_j(i) \text{ vector } j \in \text{the negative sample set}, \quad \text{(Equation 17)}$$

where P_n and N_n are the number of positive samples and negative samples, respectively.

$F_p(i)$ and $F_n(i)$ are futher normalized to $FF_p(i)$ and $FF_n(i)$ in a similar way to Equation 15. Then the final score of every feature can be calculated by

$$Sco(i) = \frac{FF_p(i)}{FF_n(i)} \quad i = 1, 2, \ldots, 878. \quad \text{(Equation 18)}$$

To achieve our purpose, $Sco(i)$ can be utilized to judge whether the i-th feature is effective, as $Sco(i)$ measures the relative enrichment of the i-th feature in the positive samples over the negative samples. For a feature i, when $FFp(i)$ in P is large but $FFn(i)$ in N is small, $Sco(i)$ will be large. On the contrary, $Sco(i)$ will be small when

FFp(*i*) in *P* is small but *FFn*(*i*) in *N* is large. The most effective feature has the largest or smallest score. In this study, we cut down the dimension of feature vector from 878 to 100. Thus, 50 features with the largest scores and 50 features with the smallest scores were selected to represent each sample by a 100-dimensional vector, which could improve the ability of our model to distinguish positive miRNA-disease associations from negative associations.

Finally, RandomForestRegressor, an algorithm package of Random Forest, was implemented to train the prediction model by training sample set. More specifically, each of the samples in the training set was represented by a 100-dimensional vector according to step 2 and step 3. Each sample in the positive sample set was given a label of 1, and each sample in the negative sample set was given a label of 0. Then we put these training samples' data into the package of Random Forest. After training, we obtained a prediction model that could infer potential miRNA-disease associations by scoring miRNA-disease samples. The higher score of a miRNA-disease sample indicates that the miRNA is more likely to be associated with the disease. It's worth noting that the max_features, n_estimators, and min_samples_leaf, main parameters of RandomForestRegressor, were set to 0.2, 100, and 10, respectively, according to empirical data.

SUPPLEMENTAL INFORMATION

Supplemental Information includes one table and can be found with this article online at https://doi.org/10.1016/j.omtn.2018.10.005.

AUTHOR CONTRIBUTIONS

All authors contributed important elements to the work presented herein. X.C. conceived the project, developed the prediction method, designed the experiments, analyzed the results, and wrote the paper. C.-C.W. implemented the experiments, analyzed the results, and wrote the paper. J.Y. and Z.-H.Y. analyzed the results and revised the paper. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

ACKNOWLEDGMENTS

X.C. was supported by the National Natural Science Foundation of China under grant 61772531.

REFERENCES

1. Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281–297.
2. Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350–355.
3. Meister, G., and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. Nature *431*, 343–349.
4. Ambros, V. (2001). microRNAs: tiny regulators with great potential. Cell *107*, 823–826.
5. 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.
6. Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature *403*, 901–906.
7. 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.
8. 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.
9. Ambros, V. (2003). MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. Cell *113*, 673–676.
10. Xu, P., Guo, M., and Hay, B.A. (2004). MicroRNAs and the regulation of cell death. Trends Genet. *20*, 617–624.
11. Zhang, K., and Guo, L. (2018). MiR-767 promoted cell proliferation in human melanoma by suppressing CYLD expression. Gene *641*, 272–278.
12. Taganov, K.D., Boldin, M.P., Chang, K.J., and Baltimore, D. (2006). NF-kappaB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. Proc. Natl. Acad. Sci. USA *103*, 12481–12486.
13. Cui, Q., Yu, Z., Purisima, E.O., and Wang, E. (2006). Principles of microRNA regulation of a human cellular signaling network. Mol. Syst. Biol. *2*, 46.
14. Ma, L., Teruya-Feldstein, J., and Weinberg, R.A. (2007). Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. Nature *449*, 682–688.
15. Miska, E.A. (2005). How microRNAs control cell division, differentiation and death. Curr. Opin. Genet. Dev. *15*, 563–568.
16. Calin, G.A., Dumitru, C.D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., et al. (2002). Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc. Natl. Acad. Sci. USA *99*, 15524–15529.
17. Yao, J., Hennessey, T., Flynt, A., Lai, E., Beal, M.F., and Lin, M.T. (2010). MicroRNA-related cofilin abnormality in Alzheimer's disease. PLoS ONE *5*, e15546.
18. Gao, P., Wong, C.C., Tung, E.K., Lee, J.M., Wong, C.M., and Ng, I.O. (2011). Deregulation of microRNA expression occurs early and accumulates in early stages of HBV-associated multistep hepatocarcinogenesis. J. Hepatol. *54*, 1177–1184.
19. Oertli, M., Engler, D.B., Kohler, E., Koch, M., Meyer, T.F., and Müller, A. (2011). MicroRNA-155 is essential for the T cell-mediated control of Helicobacter pylori infection and for the induction of chronic Gastritis and Colitis. J. Immunol. *187*, 3578–3586.
20. Bang, C., Fiedler, J., and Thum, T. (2012). Cardiovascular importance of the microRNA-23/27/24 family. Microcirculation *19*, 208–214.
21. Chen, X., Wang, L., Qu, J., Guan, N.N., and Li, J.Q. (2018). Predicting miRNA-disease association based on inductive matrix completion. Bioinformatics. Published online June 22, 2018. https://doi.org/10.1093/bioinformatics/bty503.
22. 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.
23. Chen, X., and Huang, L. (2017). LRSSLMDA: Laplacian Regularized Sparse Subspace Learning for MiRNA-Disease Association prediction. PLoS Comput. Biol. *13*, e1005912.
24. 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 *13*, e1005455.
25. Chen, X., Huang, L., Xie, D., and Zhao, Q. (2018). EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction. Cell Death Dis. *9*, 3.
26. Chen, X., Zhou, Z., and Zhao, Y. (2018). ELLPMDA: Ensemble learning and link prediction for miRNA-disease association prediction *15*, 807–818.
27. Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabási, A.L. (2007). The human disease network. Proc. Natl. Acad. Sci. USA *104*, 8685–8690.
28. 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.

29. Chen, X., Xie, D., Zhao, Q., and You, Z.H. (2017). MicroRNAs and complex diseases: from experimental results to computational models. Brief. Bioinform. Published online October 17, 2017. https://doi.org/10.1093/bib/bbx130.

30. 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.

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

32. Mørk, S., Pletscher-Frankild, S., Palleja Caro, A., Gorodkin, J., and Jensen, L.J. (2014). Protein-driven inference of miRNA-disease associations. Bioinformatics *30*, 392–397.

33. Xu, C., Ping, Y., Li, X., Zhao, H., Wang, L., Fan, H., Xiao, Y., and Li, X. (2014). Prioritizing candidate disease miRNAs by integrating phenotype associations of multiple diseases with matched miRNA and mRNA expression profiles. Mol. Biosyst. *10*, 2800–2809.

34. 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.

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

36. 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.

37. 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.

38. 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.

39. Li, J.Q., Rong, Z.H., Chen, X., Yan, G.Y., and You, Z.H. (2017). MCMDA: Matrix completion for MiRNA-disease association prediction. Oncotarget *8*, 21187–21199.

40. Yu, H., Chen, X., and Lu, L. (2017). Large-scale prediction of microRNA-disease associations by combinatorial prioritization algorithm. Sci. Rep. *7*, 43792.

41. Chen, X., Wu, Q.F., and Yan, G.Y. (2017). RKNNMDA: Ranking-based KNN for MiRNA-Disease Association prediction. RNA Biol. *14*, 952–962.

42. 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.

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

44. 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.

45. Ulintz, P.J., Zhu, J., Qin, Z.S., and Andrews, P.C. (2006). Improved classification of mass spectrometry database search results using newer machine learning approaches. Mol. Cell. Proteomics *5*, 497–509.

46. Xu, T., Li, B., Zhao, M., Szulwach, K.E., Street, R.C., Lin, L., Yao, B., Zhang, F., Jin, P., Wu, H., and Qin, Z.S. (2015). Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. Nucleic Acids Res. *43*, 2757–2766.

47. Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M., and Ren, B. (2013). RFECS: a random-forest based algorithm for enhancer identification from chromatin state. PLoS Comput. Biol. *9*, e1002968.

48. Ritchie, G.R., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. Nat. Methods *11*, 294–296.

49. Chen, L., Jin, P., and Qin, Z.S. (2016). DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. Genome Biol. *17*, 252.

50. Cheng, Z., Huang, K., Wang, Y., Liu, H., Guan, J., and Zhou, S. (2017). Selecting high-quality negative samples for effectively predicting protein-RNA interactions. BMC Syst. Biol. *11* (*Suppl 2*), 9.

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

52. Yang, Z., Ren, F., Liu, C., He, S., Sun, G., Gao, Q., Yao, L., Zhang, Y., Miao, R., Cao, Y., et al. (2010). dbDEMC: a database of differentially expressed miRNAs in human cancers. BMC Genomics *11* (*Suppl 4*), S5.

53. 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.

54. Parkin, D.M., Bray, F., Ferlay, J., and Pisani, P. (2005). Global cancer statistics, 2002. CA Cancer J. Clin. *55*, 74–108.

55. Bosetti, C., Levi, F., Ferlay, J., Garavello, W., Lucchini, F., Bertuccio, P., Negri, E., and La Vecchia, C. (2008). Trends in oesophageal cancer incidence and mortality in Europe. Int. J. Cancer *122*, 1118–1129.

56. Daly, J.M., Fry, W.A., Little, A.G., Winchester, D.P., McKee, R.F., Stewart, A.K., and Fremgen, A.M. (2000). Esophageal cancer: results of an American College of Surgeons Patient Care Evaluation Study. J. Am. Coll. Surg. *190*, 562–572.

57. Chen, X., Hu, H., Guan, X., Xiong, G., Wang, Y., Wang, K., Li, J., Xu, X., Yang, K., and Bai, Y. (2012). CpG island methylation status of miRNAs in esophageal squamous cell carcinoma. Int. J. Cancer *130*, 1607–1613.

58. Nayak, L.M., and Deschler, D.G. (2003). Lymphomas. Otolaryngol. Clin. North Am. *36*, 625–646.

59. Barajas Torres, R.L., Domínguez Cruz, M.D., Borjas Gutiérrez, C., Ramírez Dueñas, Mde.L., Magaña Torres, M.T., and González García, J.R. (2016). 1,2:3,4-Diepoxybutane Induces Multipolar Mitosis in Cultured Human Lymphocytes. Cytogenet. Genome Res. *148*, 179–184.

60. Intlekofer, A.M., and Younes, A. (2014). Precision therapy for lymphoma–current state and future directions. Nat. Rev. Clin. Oncol. *11*, 585–596.

61. Ohyashiki, K., Umezu, T., Yoshizawa, S., Ito, Y., Ohyashiki, M., Kawashima, H., Tanaka, M., Kuroda, M., and Ohyashiki, J.H. (2011). Clinical impact of down-regulated plasma miR-92a levels in non-Hodgkin's lymphoma. PLoS ONE *6*, e16408.

62. Peveling-Oberhag, J., Crisman, G., Schmidt, A., Döring, C., Lucioni, M., Arcaini, L., Rattotti, S., Hartmann, S., Piiper, A., Hofmann, W.P., et al. (2012). Dysregulation of global microRNA expression in splenic marginal zone lymphoma and influence of chronic hepatitis C virus infection. Leukemia *26*, 1654–1662.

63. 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.

64. Xiang, Z., Qin, T., Qin, Z.S., and He, Y. (2013). A genome-wide MeSH-based literature mining system predicts implicit gene-to-gene relationships and networks. BMC Syst. Biol. *7* (*Suppl 3*), S9.

65. van Laarhoven, T., Nabuurs, S.B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics *27*, 3036–3043.