MISSIM: An Incremental Learning-Based Model With Applications to the Prediction of miRNA-Disease Association

Kai Zheng, Zhu-Hong You, Lei Wang, Yi-Ran Li, Ji-Ren Zhou, and Hai-Tao Zeng

Abstract—In the past few years, the prediction models have shown remarkable performance in most biological correlation prediction tasks. These tasks traditionally use a fixed dataset, and the model, once trained, is deployed as is. These models often encounter training issues such as sensitivity to hyperparameter tuning and "catastrophic forgetting" when adding new data. However, with the development of biomedicine and the accumulation of biological data, new predictive models are required to face the challenge of adapting to change. To this end, we propose a computational approach based on Broad learning system (BLS) to predict potential disease-associated miRNAs that retain the ability to distinguish prior training associations when new data need to be adapted. In particular, we are introducing incremental learning to the field of biological association prediction for the first time and proposed a new method for quantifying sequence similarity. In the performance evaluation, the AUC in the 5-fold cross-validation was 0.9400 +/- 0.0041. To better assess the effectiveness of MISSIM, we compared it with various classifiers and former prediction models. Its performance is superior to the previous method. Besides, the case study on identifying miRNAs associated with breast neoplasms, lung neoplasms and esophageal neoplasms show that 34, 36 and 35 out of the top 40 associations predicted by MISSIM are confirmed by recent biomedical resources. These results provide ample convincing evidence of this approach have potential value and prospect in promoting biomedical research productivity.

Index Terms—miRNA-disease association, heterogenous information sources, broad learning system, sequence information, incremental learning

1 Introduction

MicroRNAs (miRNA) regulate gene expression in some physiological processes, such as apoptosis and differentiation of cells, through complementary base pairing with messenger RNA (mRNA) [1], [2], [3]. Line-4 and let-7 are miRNAs which are known as characterizations of genes in the past 20 years [4], [5]. Since then, the number of discovered miRNAs accumulated quickly by various biological experimental methods [6]. Furthermore, abundant experimental studies have shown that miRNA is closely related to human diseases. Exploring the influence mechanisms of miRNA in diseases will boost the transformation diagnosis and treatment model. For instance, the combination of miR-211 and TGFbeta R2 accelerated the cancerization of head and neck [7]. By targeting c-Met, migration and invasion of breast cancer cell were inhibited by mir-340 [8]. Gao *et al*. have found miRNAs are dysfunctional at an early stage by researching the expression changes of miRNAs which is related with disease in prime of HBV-associated hepatocarcinogenesis [9]. The miR-145 was a tumor suppressor candidate miRNA and could give a major push to the development of HCC indicated by their results in the meantime [10]. However, the evolution may be blocked by the high-cost, long cycle experiment and sensitivity of noise. Finding a more credible miRNA–disease association prediction method becomes an important research hotspot.

In the past five years, traditional prediction models have been proposed to solve biological problems [11], [12], [13], [14], [15], [16], [17], [18], [19]. They are based primarily on similarity or on machine learning [20]. A miRNA prioritization approach was built by Xu *et al*. [21]. The potential associations were distinguished by the target-miRNA interactions and genes of known disease. Liu *et al*. predicted miRNA-disease associations by a heterogeneous network [22]. Later, a method was proposed by Zeng *et al*. which gathers social network analysis to forecast the relationship between miRNAs and diseases [23]. Zou *et al*. predicted disease-specific miRNAs by a supervised machine learning [24]. They used bootstrap aggregating algorithm to train the biased SVM classifier.

In the traditional prediction model, all training data are presented to the classifier [25], [26], [27], [28], [29], [30]. However, under the condition that the miRNA regulation mechanism has not been thoroughly explored, all biological information can hardly be acquired at the same time, but gradually collected by the database. Therefore, incremental learning is of great value. In this work, we propose a prediction model called

- *Zheng Kai is the with the School of Computer Science and Engineering, Central South University, Changsha 410083, China. E-mail: zhengkai951211@gmail.com.*
- *Zhu-Hong You, Lei Wang, and Ji-Ren Zhou are with the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China. E-mail: {zhuhongyou, leiwang, zhoujr}@ms.xjb.ac.cn.*
- *Yi-Ran Li is with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221000, China. E-mail: lyrram@163.com.*
- *Hai-Tao Zeng is with the School of Mechanical Electronic & Information Engineering, China University of Mining and Technology-Beijing, Beijing 100083, China. E-mail: haitao.zeng@vipl.ict.ac.cn.*

Manuscript received 8 Apr. 2020; revised 21 June 2020; accepted 30 July 2020. Date of publication 4 Aug. 2020; date of current version 7 Oct. 2021. (Corresponding authors: Zhu-Hong You and Lei Wang) Digital Object Identifier no. 10.1109/TCBB.2020.3013837

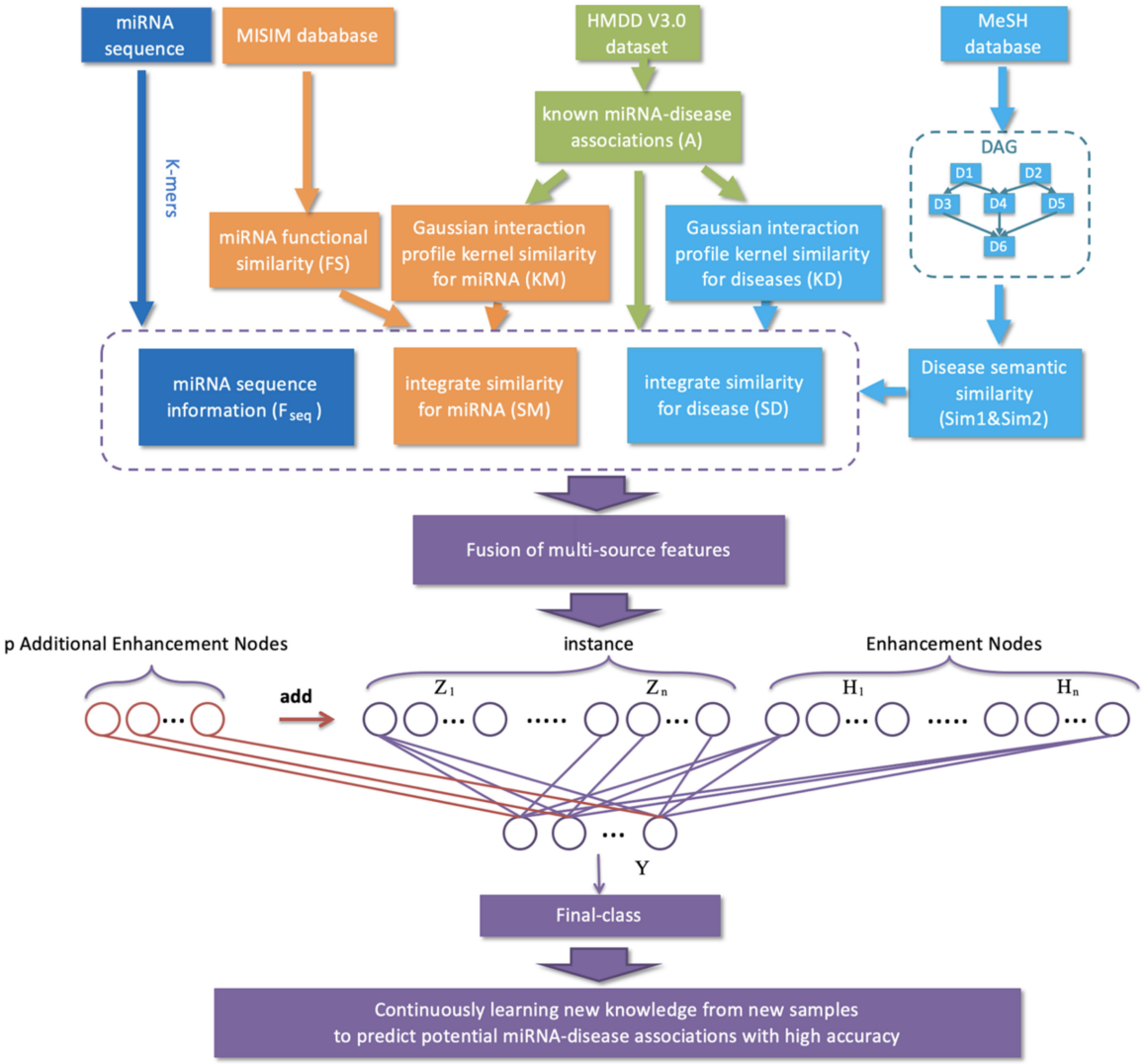


Fig. 1. The workflow of MISSIM model to predict potential miRNA-disease associations.

MISSIM to solve the problem of learning such incremental available data in biological association prediction. In addition, another innovation of the proposed method is to propose an algorithm for quantifying sequence similarity. Specifically, according to the miRNA functional data and disease semantic data, we first obtain the similarity between miRNAs and diseases. Second, the feature information of the miRNA sequence can be abstracted by the Chaos Game Representation (CGR) technology [31]. We compute the relative similarity between any pair of miRNAs by Pearson's correlation to build the miRNA sequence similarity matrix. Third, we construct a feature descriptor which gathered the similarity matrixes of sequence and association. Finally, the processed feature vectors are placed in the broad learning system classifier and potential miRNA-disease associations are obtained. To assess the performance of MISSIM in the HMDD V3.0 data set [32], we computed the AUC of 5-fold cross-validation (0.9400+/-0.0041). Moreover, we verified MISSIM by three disease including Breast Neoplasms, Lung Neoplasms and Esophageal Neoplasms. As a result, 34, 36 and 35 out of the top 40 predicted miRNAs were respectively verified by other association database. These results provide ample convincing evidence to demonstrate the effectiveness of the method. Fig. 1 shows the workflow of the proposed method.

2 Results

2.1 Evaluation Criteria

Accuracy ($Acc.$), sensitivity ($Sen.$), precision ($Pre.$) and F_1 score are used to assess the performance of MISSIM, which are defined by:

$$Acc. = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Sen. = \frac{TP}{TP + FN} \tag{2}$$

$$Pre. = \frac{TP}{TP + FP} \tag{3}$$

$$F_1 = \frac{Prec. \times Sen.}{Prec. + Sen.}, \tag{4}$$

Where TP is the true positive. FP is the false positive. TN is the true negative and FN is false negative.

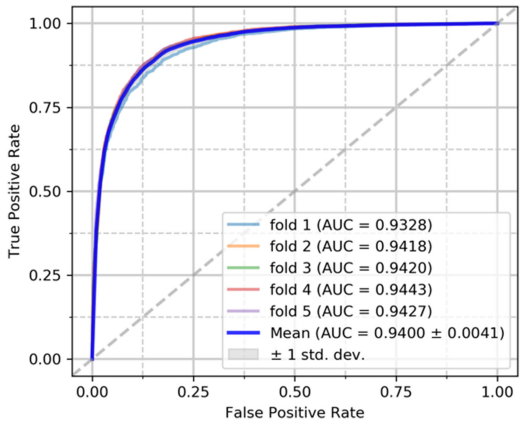


Fig. 2. ROC curves performed by MISSIM on HMDD v3.0 dataset.

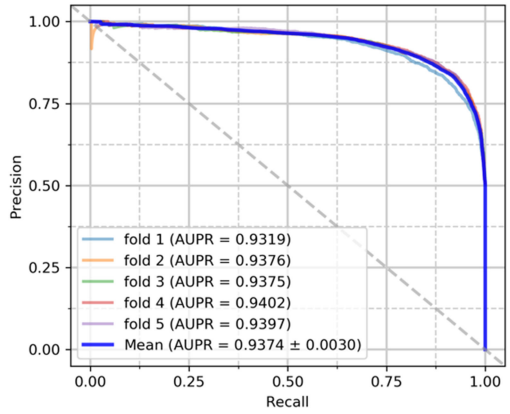


Fig. 3. PR curves performed by MISSIM on HMDD v3.0 dataset.

2.2 Performance Evaluation

In HMDD v3.0 dataset, 1102 miRNAs and 850 diseases build the dataset with 32281 known miRNA-disease associations from 17412 papers. Some of the associations whose information is unreliable that judged by the public database miRBase and we have removed it [33]. After screening, positive samples were constructed from 32226 miRNA-disease associations, and we randomly selected the same number pairs from unproven miRNA-disease pairs as negative samples.

TABLE 1
5-Fold Cross-Validation Results Performed by Proposed Model on HMDD v3.0

Testing set	Acc.(%)	Sen.(%)	Pre.(%)	F1(%)	AUC(%)
1	85.65	86.83	84.83	85.82	93.28
2	86.94	86.54	87.23	86.88	94.18
3	86.99	91.44	83.98	87.55	94.20
4	87.23	90.11	85.20	87.59	94.43
5	87.42	88.63	86.54	87.57	94.27
Average	86.85±0.62	88.71±1.88	85.56±1.18	87.08±0.69	94.00±0.41

Prediction of miRNA-Disease Association. Fig. 2 lists the performance of MISSIM and it has gained an average AUC of 0.9400+/-0.0041. The AUC of the five experiments is 0.9328, 0.9418, 0.9420, 0.9443 and 0.9427 respectively. And, the AUPR of the five experiments is 0.9319, 0.9376, 0.9375, 0.9402 and 0.9397 respectively (Fig. 3). Table 1 shows the average accuracy, sensitivity, accuracy, and f1 scores of 0.8685, 0.8871, 0.8556, and 0.8708, respectively. In accordance with the results of experiment, our approach is feasible, reliable and comes to the result of the expectation. It is a powerful tool for predicting potential miRNA-disease association.

Comparison With Different Classifier Models. The MISSIM model has excellent performance on the HMDD 3.0 database using the BLS classifier. Here, Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) are selected to compare with it [34], [35], [36]. The accuracy of the four experiments are 0.8685 (Broad Learning System [37]), 0.8200 (SVM), 0.8233 (Random forest) and 0.8080 (Decision Tree). Their AUCs are 0.9400 (Broad Learning System), 0.8839 (SVM), 0.9150 (Random forest) and 0.8078 (Decision Tree) shown as Fig. 4. The accuracy, sensitivity, precision and f1-score have been shown in Table 2. It can be directly observed that the MISSIM model based on the BLS classifier achieves the highest results in all four evaluation criteria, which indicates that the performance of MISSIM is better than the other three, especially in the AUC which represents the overall performance of the model. The results show that the "mapped feature" adopted by the BLS can effectively extract the deep features of the data and help to improve the performance of the model.

Comparison With Related Method. The performance of MISSIM is compared with five most advanced prediction factors

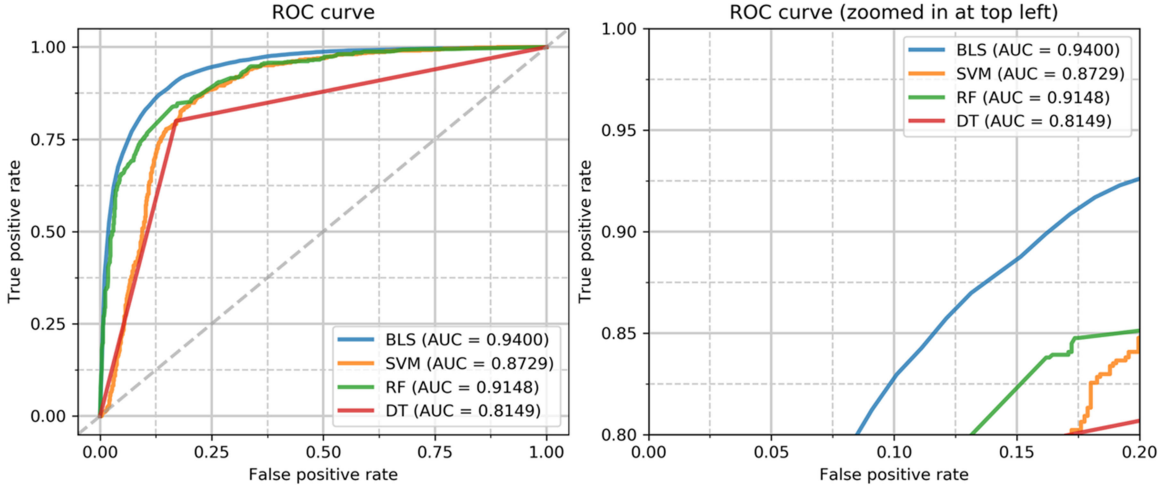


Fig. 4. The ROCs of four different classifiers which are BLS, SVM, Random Forest and Decision Tree.

TABLE 2
Performance Comparison of Different Classifiers on HMDD v3.0 Dataset

Method	Accuracy(%)	Sensitivity(%)	Precision(%)	F1-score(%)
SVM	82.00%	81.62%	81.84%	81.73%
RF	82.33%	76.22%	86.37%	80.98%
DT	80.80%	79.19%	81.39%	80.27%
BLS	86.85%	88.71%	85.56%	87.08%

to further verify the effectiveness of the model. The performance of MISSIM is further evaluated by comparison with eight state-of-the-art predictors including PBMDA [12], VAEMDA [38], LMTRDA [39], MLMDA [40], MDA-CNN [41], EPMDA [42], DBMDA [43], CGMDA [44]. Table 3 lists the performance of various predictors. In detail the AUCs of PBMDA, VAEMDA, LMTRDA, MLMDA, MDA-CNN, EPMDA, DBMDA, CGMDA are 0.9172, 0.9091, 0.9054, 0.9172, 0.8897, 0.9371, 0.9129, and 0.9099, respectively. Obviously, MISSIM is superior to other methods, indicating that the similarity of sequence information based on chaos game and efficient incremental learning by lateral expansion can improve the prediction performance of miRNA-disease association.

2.3 Case Studies

To further evaluate the effectiveness of MISSIM, we applied MISSIM to three human diseases, including breast, lung, and esophageal Neoplasms. Among them, the test sample was established by the miRNA-disease associations about these three diseases and all possible miRNAs. We confirmed the top 40 predictions in dbDEMC v2.0 and miR2Disease [45], [46].

Breast neoplasms which occur in breast tissue takes up about 66 percent of breast disease. Breast cancer is a malignant breast tumor that develops from the uncontrolled growth of freak breast cells. Malignant neoplasms can invade and destroy surrounding tissue and spread to other parts of the body. The reason of most malignant breast tumors is unknown, however, a small of them tend to group in families. So, in the first case study, we took it to assess the performance of MISSIM. As shown in Table 4, 34 associations were confirmed.

The main culprit behind lung cancer is the uncontrolled growth of cells in lung tissue. Here, lung tumors were selected as the second case study. After the candidate miRNAs were sorted according to the predicted score, the first 40 were validated. Of these, 36 associations were confirmed to be associated with lung tumors. (See Table 5). As shown in Table 6, 35 of the top 40 Esophageal Neoplasms-associated miRNAs predicted by the proposed model were validated.

TABLE 3
The Comparison With Related Models

Method	AUC
PBMDA	0.9172
VAEMDA	0.9091
LMTRDA	0.9054
MLMDA	0.9172
MDA-CNN	0.8897
EPMDA	0.9371
CGMDA	0.9129
DBMDA	0.9099
MISSIM	0.9400

TABLE 4
Prediction of the Top 40 Predicted miRNAs Associated With Breast Neoplasms Based on Known Associations in dbDEMC v2.0 and miR2Database

miRNA	dbDEMC	miR2D	miRNA	dbDEMC	miR2D
hsa-mir-921	**confirmed**	N/A	hsa-mir-604	**confirmed**	N/A
hsa-mir-600	N/A	N/A	hsa-mir-220a	**confirmed**	N/A
hsa-mir-662	**confirmed**	N/A	hsa-mir-518d	**confirmed**	N/A
hsa-mir-596	**confirmed**	N/A	hsa-mir-3926	N/A	N/A
hsa-mir-548i	**confirmed**	N/A	hsa-mir-544	**confirmed**	N/A
hsa-mir-602	**confirmed**	N/A	hsa-mir-1268a	**confirmed**	N/A
hsa-mir-769	**confirmed**	N/A	hsa-mir-4772	N/A	N/A
hsa-mir-1468	**confirmed**	N/A	hsa-mir-1282	**confirmed**	N/A
hsa-mir-1237	**confirmed**	N/A	hsa-mir-548l	**confirmed**	N/A
hsa-mir-615	**confirmed**	N/A	hsa-mir-1284	N/A	N/A
hsa-mir-521-1	**confirmed**	N/A	hsa-mir-768	**confirmed**	N/A
hsa-mir-145a	**confirmed**	N/A	hsa-mir-1285-2	**confirmed**	N/A
hsa-mir-623	**confirmed**	N/A	hsa-mir-9a	**confirmed**	**confirmed**
hsa-mir-612	N/A	N/A	hsa-mir-3928	**confirmed**	N/A
hsa-mir-4301	**confirmed**	N/A	hsa-mir-7152	N/A	N/A
hsa-mir-4753	**confirmed**	N/A	hsa-mir-644b	**confirmed**	N/A
hsa-mir-654	**confirmed**	N/A	hsa-mir-1286	**confirmed**	N/A
hsa-mir-1293	**confirmed**	N/A	hsa-mir-1914	**confirmed**	N/A
hsa-mir-518a-1	**confirmed**	N/A	hsa-mir-5010	**confirmed**	N/A
hsa-mir-518a-2	**confirmed**	N/A	hsa-mir-583	**confirmed**	N/A

3 Materials and Methods

3.1 Data Set

HMDD [47]. In the proposed method, HMDD v3.0 provides the known experimentally verified human miRNA-disease association. The experimental data can be downloaded from the homepage of the dataset, http://www.cuilab.cn/hmdd. After pretreatment, 32226 miRNA-disease associations were obtained, including 1057 miRNA and 850 diseases.

miRBase [48]. The database provides all-round data on miRNA, including miRNA sequence annotation, prediction of gene targets and other information. In this work, the

TABLE 5
Prediction of the Top 40 Predicted miRNAs Associated With Lung Neoplasms Based on Known Associations in dbDEMC v2.0 and miR2Database

miRNA	dbDEMC	miR2D	miRNA	dbDEMC	miR2D
hsa-mir-515	**confirmed**	N/A	hsa-mir-3170	**confirmed**	N/A
hsa-mir-513a	**confirmed**	N/A	hsa-mir-617	**confirmed**	N/A
hsa-mir-658	**confirmed**	N/A	hsa-mir-633	**confirmed**	N/A
hsa-mir-3200	**confirmed**	N/A	hsa-mir-3201	N/A	N/A
hsa-mir-642	**confirmed**	N/A	hsa-mir-562	**confirmed**	N/A
hsa-mir-507	**confirmed**	N/A	hsa-mir-517b	**confirmed**	N/A
hsa-mir-526a	**confirmed**	N/A	hsa-mir-122a	**confirmed**	N/A
hsa-mir-550b-1	**confirmed**	N/A	hsa-mir-1301	**confirmed**	N/A
hsa-mir-1269	**confirmed**	N/A	hsa-mir-4534	N/A	N/A
hsa-mir-4449	**confirmed**	N/A	hsa-mir-514	**confirmed**	N/A
hsa-mir-147a	**confirmed**	N/A	hsa-mir-654	**confirmed**	N/A
hsa-mir-3117	**confirmed**	N/A	hsa-mir-1292	**confirmed**	N/A
hsa-mir-1274b	**confirmed**	N/A	hsa-mir-649	**confirmed**	N/A
hsa-mir-587	N/A	N/A	hsa-mir-1277	**confirmed**	N/A
hsa-mir-626	**confirmed**	N/A	hsa-mir-889	**confirmed**	N/A
hsa-mir-1293	**confirmed**	N/A	hsa-mir-941-1	**confirmed**	N/A
hsa-mir-548f	**confirmed**	N/A	hsa-mir-450a-1	**confirmed**	N/A
hsa-mir-1273c	N/A	N/A	hsa-mir-1469	**confirmed**	N/A
hsa-mir-365-1	**confirmed**	N/A	hsa-mir-591	**confirmed**	N/A
hsa-mir-1260	**confirmed**	N/A	hsa-mir-933	**confirmed**	N/A

TABLE 6
Prediction of the Top 40 Predicted miRNAs Associated With Esophageal Neoplasms Based on Known Associations in dbDEMC v2.0 and miR2Database

miRNA	dbDEMC	miR2D	miRNA	dbDEMC	miR2D
hsa-mir-6886	**confirmed**	N/A	hsa-mir-614	**confirmed**	N/A
hsa-mir-3131	**confirmed**	N/A	hsa-mir-3610	**confirmed**	N/A
hsa-mir-644b	N/A	N/A	hsa-mir-4290	**confirmed**	N/A
hsa-mir-550b-1	**confirmed**	N/A	hsa-mir-4257	**confirmed**	N/A
hsa-mir-4298	**confirmed**	N/A	hsa-mir-1269a	**confirmed**	N/A
hsa-mir-4489	**confirmed**	N/A	hsa-mir-4496	**confirmed**	N/A
hsa-mir-1468	**confirmed**	N/A	hsa-mir-3656	**confirmed**	N/A
hsa-mir-9a	**confirmed**	N/A	hsa-mir-668	**confirmed**	N/A
hsa-mir-5193	**confirmed**	N/A	hsa-mir-4478	**confirmed**	N/A
hsa-mir-2355	**confirmed**	N/A	hsa-mir-4706	N/A	N/A
hsa-mir-453	**confirmed**	N/A	hsa-mir-4279	**confirmed**	N/A
hsa-mir-1972	**confirmed**	N/A	hsa-mir-3185	**confirmed**	N/A
hsa-mir-4419b	**confirmed**	N/A	hsa-mir-4649	**confirmed**	N/A
hsa-mir-1254	**confirmed**	N/A	hsa-mir-514a-3	**confirmed**	N/A
hsa-mir-1206	N/A	N/A	hsa-mir-374	**confirmed**	N/A
hsa-mir-1302-5	**confirmed**	N/A	hsa-mir-514	**confirmed**	N/A
hsa-mir-3130-2	**confirmed**	N/A	hsa-mir-3158	**confirmed**	N/A
hsa-mir-450a-2	**confirmed**	N/A	hsa-mir-1283	N/A	N/A
hsa-mir-3622a	**confirmed**	N/A	hsa-mir-365-1	**confirmed**	N/A
hsa-mir-6880	**confirmed**	N/A	hsa-mir-933	N/A	N/A

miRNA sequence information is downloaded from the homepage of miRBase (http://www.mirbase.org).

3.2 miRNA Functional Similarity

Wang *et al*. built a method for computing miRNA functional similarity scores between different miRNAs in the scenario that phenotypically similar diseases tend to relate with functional similarity miRNAs, and uploaded the information at www.cuilab.cn/files/images/cuilab/misim.zip [49], [50], [51], [52], [53]. In this method, we downloaded it and constructed a 495 rows $\times$ 495 columns matrix *FS* where an entity *FS(m(a), m(b))* is degree of comparability between miRNA *m(a)* and *m(b)*. This data is only used in case studies.

3.3 Disease Semantic Similarity

Disease Semantic Similarity Model 1. We downloaded the disease semantic information from MeSH database (https://www.nlm.nih.gov/). In the system, we used the Directed Acyclic Graph (DAG) to describe the association between diseases. Each the direct edge connects to two nodes which represent disease from parent to child nodes. We defined disease D as $\mathrm{DAG}_d = D,\ T_d,\ E_d$ where T_d is a nodal set consisting of disease D and E_d is a set consisting of the corresponding edges [49]. Here, Xuan *et al*. offered a method to figure disease semantic similarity by MeSH diseases descriptors [54]. Particularly, the degree of semantic contribution is described as follows:

$$\begin{cases} D_d\ (t) = 1 & if\ t = D \\ D_d\ (t) = max\{\Delta * D_d(t') | t' \in children\ of\ t\} & if\ t \neq D \end{cases}, \tag{5}$$

Δ is the semantic contribution coefficient. According to the semantic contribution, the semantic value $DV(D)$ of disease D can be described as follows:

$$DV\ (D) = \sum_{t \in T_d} D_d(t)\ . \tag{6}$$

If the diseases $d(i)$ and $d(j)$ share more DAG, then the two diseases are more semantically similar. According to this assumption, semantic comparability is defined as follows:

$$Sim1(d(i), d(j)) = \frac{\sum_{t \in T_{d(i)} \cap T_{d(j)}} \left(D_{d(i)}(t) + D_{d(j)}(t)\right)}{DV(d(i)) + DV(d(j))}. \tag{7}$$

$Sim1$ is a semantic comparability matrix of disease which has 850 rows and 850 columns. The element $Sim1(d(i), d(j))$ is regarded as the semantic similarity of $d(i)$ and $d(j)$.

Disease Semantic Similarity Model 2. Hence, the effectiveness of prediction model can be improved by retaining the specificity of disease terms. Because the information content can measure the particularity of disease term effectively, we used it in common ancestor nodes and the closest leaf nodes. First, the information content of all diseases can be figured by the negative log possibility of each term. And we can define disease term t's information content as follow [54]:

$$D2_d(t) = -log\left(\frac{number\ of\ DAGs\ including\ t}{number\ of\ disease}\right). \tag{8}$$

Next step, the degree of semantic comparability between diseases $d(i)$ and $d(j)$ can be figured as below:

$$Sim2(d(i), d(j)) = \frac{\sum_{t \in T_{d(i)} \cap T_{d(j)}} \left(D2_{d(i)}(t) + D2_{d(j)}(t)\right)}{DV(d(i)) + DV(d(j))}, \tag{9}$$

Where $DV(d(i))$ and $DV(d(j))$ are the semantic score of $d(i)$ and (j), and can be figured in same way as formula (6).

3.4 Gaussian Interaction Profile Kernel Similarity

Gaussian Interaction Profile Kernel Similarity for Diseases. According to previous studies [55], we marked miRNAs which can associate with $d(a)$ to describe binary vector $IP(d(a))$ that represents the interaction profiles of disease $d(a)$. We described $KD(d(a), d(b))$ between $d(a)$ and $d(b)$ as follow:

$$KD(d(a), d(b)) = \exp\left(-\gamma_d * \|IP(d(a)) - IP(d(b))\|^2\right), \tag{10}$$

Where parameter γ_d is a coefficient of the kernel bandwidth and nd is the number of matrix A's row. γ_d is designed as follows:

$$\gamma_d = \frac{1}{nd} \sum_{i=1}^{nd} \|IP(d(i))\|^2. \tag{11}$$

Gaussian Interaction Profile Kernel Similarity for miRNAs. The column vector of the adjacency matrix A is defined as $IP(m(a))$ or $IP(m(b))$ and nm is the number of matrix A's column.

$$KM(m(a), m(b)) = \exp\left(-\gamma_m * \|IP(m(a)) - IP(m(b))\|^2\right) \tag{12}$$

$$\gamma_m = \frac{1}{nm} \sum_{i=1}^{nm} \|IP(m(i))\|^2. \tag{13}$$

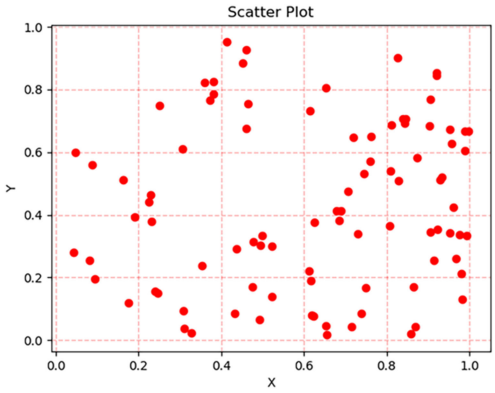


Fig. 5. CGR of the miRNA named hsa-mir-449.

3.5 Integrated Similarity

Integrated Similarity for Diseases. For getting the utmost out of $Sim1(d(i), d(j))$, $Sim2(d(i), d(j))$ and $KD(d(a), d(b))$, we built a gathered disease similarity matrix SD combined above similarities [56]. The element $SD(d(a), d(b))$ is integrated similarity between disease $d(a)$ and $d(b)$. It can be described as follows:

$$SD(d(a), d(b)) = \begin{cases} \frac{Sim1(d(a),d(b))+Sim2(d(a),d(b))}{2} & if\ d(a), d(b)\ in\ Sim1 and Sim2 \\ KD(d(a), d(b)) & others \end{cases}. \tag{14}$$

Integrated Similarity for miRNAs. $FS(m(a), m(b))$ and $KM(m(a), m(b))$ were used to build miRNA similarity:

$$SM(m(a), m(b)) = \begin{cases} FS(m(a), m(b)) & if\ m(a), m(b)\ in\ FS \\ KM(m(a), m(b)) & others \end{cases}. \tag{15}$$

3.6 Sequence Similarity for miRNAs

In 1990, Jeffrey built a mapping method for genomic sequences named Chaos Game Representation [57]. CGR is an iterative mapping derived from statistical mechanics, especially chaos theory. And, this method maps gene sequences to two-dimensional space uniquely. However, previous studies did not adequately explore the possibility of extracting potential features of a sequence through CGR. We set the four possible nucleotides in the miRNA sequence to the four vertices of a binary square (Fig. 5).

$$CGR_i = CGR_{i-1} + \theta*(CGR_{i-1} - g_i), \tag{16}$$

Where g_i is the nucleotide coefficient, and when the nucleotides are A, C, G and U, the corresponding nucleotide coefficients are (0, 0), (0, 1), (0, 1) and (1, 0), respectively. According to previous research, parameter θ is set to 0.5. In addition, we define $i = 1 \ldots n_G$ and $CGR_0 = (0.5, 0.5)$. n_G is the length of a miRNA sequence.

The positional representation CGR_i of each nucleotide can be described as follows:

Recently, a number of tools have been proposed to analyze DNA, RNA and protein sequences at the sequence level [58], [59], [60], which has inspired us. However, we found few ways to uniquely map sequence information to the euclidean space. In this work, we were inspired by previous research and quantified the nonlinear sequence information [28], [30], [61], [62], [63], [64], [65], [66], [67]. The miRNA sequence containing a large amount of information is converted into a numerical vector to more fully represent the characteristics of the miRNA. First, we downloaded the precursor sequences of the desired miRNAs from miRBase owing to they contain richer epigenetic information. Second, the sequence of miRNA can be mapped into the CGR space with equally divided areas and the number of occurrences of each area is calculated. We used $2^{n_c} \times 2^{n_c}$ grid to get the frequency matrix of nucleotide length n_c. The nucleotide frequency matrix in Fig. 6 defined as chaos game contents is transformed from CGR drew in Fig. 5. Third, using miRNA chaos game contents shown in Fig. 6 as feature vectors to describe miRNA. Finally, according to the miRNA feature vector, the Pearson correlation coefficient was used to calculate the sequence similarity between miRNAs. We used similarity to build sequence similarity matrix (1057×1057). Therefore, each miRNA sequence could be described by a 1057-dimensional vector:

$$F_{seq} = (f_1, f_2, f_3, \ldots, f_{1056}, f_{1057}). \tag{17}$$

3.7 Broad Learning System

Broad Learning System based on Random Vector Functional Link Neural Network (RVFLNN) effectively eliminates the shortcoming of too long training process, and also ensures excellent generalization ability [37]. The core of BLS is incremental learning algorithm, which will not affect the global model by modifying a part of the parameter space and can avoid the problem of "catastrophic forgetting" [68]. Broad Learning system is a flat network, where the original inputs A are placed as 'mapped feature' and the network is expanded in the 'enhancement nodes'. The ith mapped feature F_i can be project as $\emptyset_i(AW_{ei} + \beta_{ei})$. And the connection of all the first i group of mapping feature can be donated as $F^i \equiv [F_1, \ldots, F_i]$. By fine-tuning the initial W_{ei}, the model can get better feature. Meanwhile, the jth group of enhancement nodes, $\gamma_j(F^i W_{hj} + \beta_{hj})$ can be present as E_j, and $E^i \equiv [E_1, \ldots, E_j]$ can be donated as the first j set of enhancement nodes. The weight of the feature maps W_{ei} and the weight of the enhancement nodes W_{hj} are random weight with the proper dimension. The bias β_{ei} and β_{hj} are randomly generated.

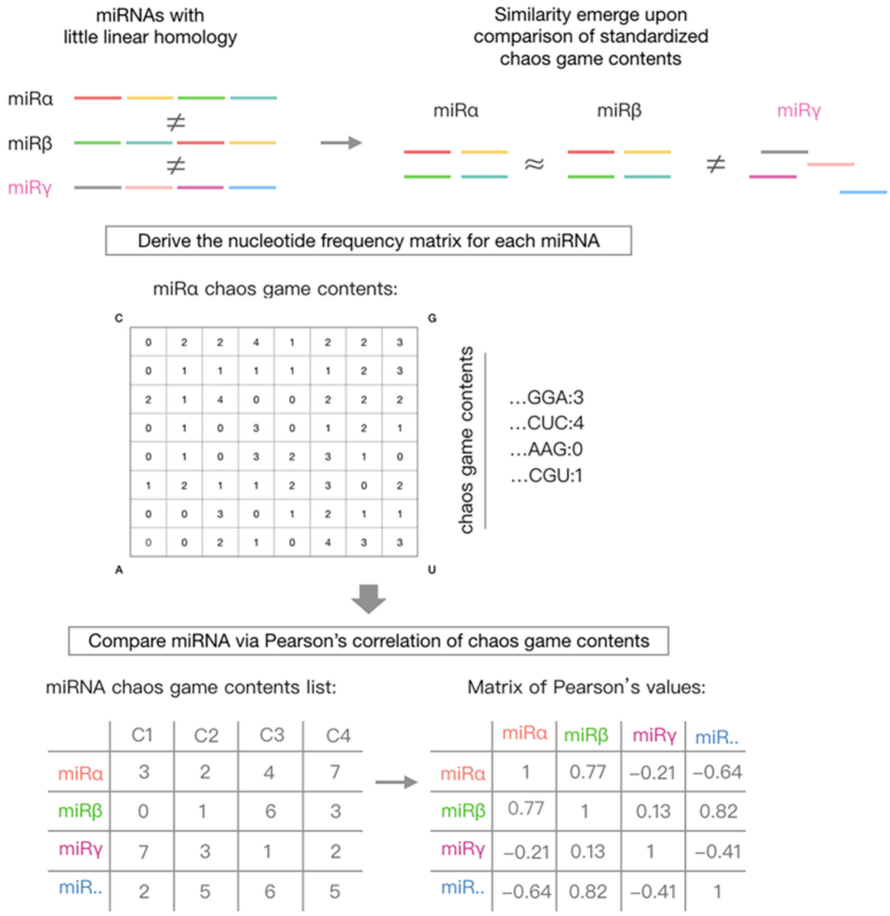
Assuming that B is the output matrix, and the input data A has N samples with M dimension, a broad learning system with n feature mappings and m groups of enhancement nodes can be present as below.

$$F_i = \emptyset_i \left(AW_{e_i} + \beta_{e_i}\right), i = 1, \ldots, n. \tag{18}$$

Donate all n groups of feature nodes as $F^i \equiv [F_1, \ldots, F_i]$, then the j set of enhancement nodes $(j = 1, \ldots, m)$ can be presented as:

$$E_j \equiv \gamma_j \left(F^i W_{h_j} + \beta_{h_j}\right). \tag{19}$$

In this way, the broad learning system can be presented as the equation:



	C1	C2	C3	C4
miRα	3	2	4	7
miRβ	0	1	6	3
miRγ	7	3	1	2
miR..	2	5	6	5

	miRα	miRβ	miRγ	miR..
miRα	1	0.77	−0.21	−0.64
miRβ	0.77	1	0.13	0.82
miRγ	−0.21	0.13	1	−0.41
miR..	−0.64	0.82	−0.41	1

Fig. 6. The workflow of calculating the miRNA sequence similarity.

$$B = \left[F_1, \ldots, F_n | \gamma\left(F^n W_{h_1} + \beta_{h_1}\right), \ldots, \gamma\left(F^n W_{h_m} + \beta_{h_m}\right) W^m\right] = [F_1, \ldots, F_n | E_1, \ldots, E_m] \quad W^m = [F^n | E^m]\ W^m. \tag{20}$$

where the $W^m = [F^n|E^m]^+ B$ and $a^+ = \lim_{\lambda \to 0} (\lambda I + aa^T)^{-1} a^T$ according to ridge regression learning algorithms. Sometimes, the result of learning cannot live up to our expectation. One solution is to insert additional enhancement node in order to get better accuracy. In this way, the algorithm only needs to compute the additional enhancement nodes. To generate new additional enhancement nodes, we donated $\mathrm{X}^m \equiv [F^n|E^m]$ and X^{m+1} can be presented as:

$$\mathrm{X}^{m+1} \equiv \left[\mathrm{X}^m | \gamma\left(F^n W_{h_{m+1}} + \beta_{h_{m+1}}\right)\right]. \tag{21}$$

And the pseudoinverse of the new matrix can be deduced as:

$$\left(\mathrm{X}^{m+1}\right)^+ = \begin{pmatrix} (\mathrm{X}^m)^+ - DY^T \\ Y^T \end{pmatrix}. \tag{22}$$

3.8 Overview

The method according to the hypothesis that functionally similar miRNAs have relation to similar diseases is also used in calculating the association between drugs and target proteins. MISSIM is mainly composed of four parts: 1. selecting positive set and negative set; 2. combining feature vectors of miRNA and disease; 3. lessening the size of combined features; 4. building the better forecast model to calculate potential associations. Here in below, we will go into detail of every process.

First, we built the training set. To be specific, we extracted the 32226 corroborative miRNA-disease pairs from HMDD v3.0 as positive samples. Then, we combined them with and negative samples to construct training set. Random selection of negative samples is composed of three steps. To be specific, choosing a disease from the 850 diseases discretionarily; selecting one of the 1057 miRNAs in same way; building a negative sample by combining the miRNA and disease which are not in positive samples.

Second, we described the associations as feature vectors. In detail, SD is integrated as a feature vector to represent each disease as a feature. Disease's feature vector SD is defined as follow:

$$SD(d(a)) = (v_1, v_2, v_3, \ldots v_{849}, v_{850}). \tag{23}$$

By the same method, the feature vector of the miRNA SM can be defined as follows:

$$SM(m(a)) = (w_1, w_2, w_3, \ldots w_{1056}, w_{1057}). \tag{24}$$

Based on the above described feature vector of disease and miRNA, the similarity feature vector F_{sim} of each miRNA-disease pair can be defined by the 1907-dimensional

vector as follows:

$$F_{sim} = (SD(d(a)), SM(m(a))). \tag{25}$$

After that, we adjust F_{sim} from 1097 to 32 through the automatic encoder. Similarly, the feature matrix F_{seq} is adjusted from 64 to 32 in the same way. We defined the final descriptor for each miRNA-disease pair as a 64-dimensional vector as follow:

$$F = \left(F_{sim}', F_{seq}'\right). \tag{26}$$

Finally, the extensive learning system is trained by the final descriptor to obtain a predictive model. If the sample is the positive sample, we define the label as 1. And if it is negative samples set, the label is defined as 0. Then, we put the data of training set into broad learning system and gained a predicting potential miRNA-disease association's model. In our prediction model, if a miRNA and a disease get the higher score, they tend to have a relationship.

4 Conclusion

In this study, we propose a model based on incremental learning to predict miRNA-disease associations, called MISSIM. This method integrated miRNA sequence information, disease semantic information, and similarity information calculated from miRNA and disease associations. In particular, we introduced incremental learning into the field of bio-association prediction for the first time to learn the biological incrementally available data, thus overcoming the problem of "catastrophic forgetting" and modifying the parameter space to affect the global model. In addition, a new method of quantifying sequences was proposed, which provided a new perspective for the characterization of sequence information. In the performance evaluation, the AUC was 0.9400 +/- 0.0041. To better evaluate the effectiveness of MISSIM, it is compared with various classifiers and previous prediction models. Its performance is superior to the previous method. Besides, the case study on identifying miRNAs associated with breast neoplasms, lung neoplasms and esophageal neoplasms show that 34, 36 and 35 out of the top 40 associations predicted by MISSIM are confirmed by recent biomedical resources. These results provide sufficient convincing evidence that MISSIM can provide researchers with powerful and useful computational support that providing large-scale disease-related miRNA candidates to promote biomedical research productivity and the development of complex disease treatment. The next task is to explore how to better characterize the biological sequence data in order to obtain better predictive model performance.

Acknowledgments

This work was supported in part by the Awardee of the NSFC Excellent Young Scholars Program, under Grant 61722212, in part by the National Natural Science Foundation of China, under Grants 61702444, 61572506, in part by the Pioneer Hundred Talents Program of Chinese Academy of Sciences, in part by the Chinese Postdoctoral Science Foundation, under Grant 2019M653804, in part by the West Light Foundation of the Chinese Academy of Sciences, under Grant 2018-XBQNXZ-B-008.

References

[1] V. Ambros, "The functions of animal microRNAs," *Nature*, vol. 431, no. 7006, pp. 350–355, 2004.
[2] D. P. Bartel, "MicroRNAs: Genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.
[3] V. Ambros, "microRNAs: Tiny regulators with great potential," *Cell*, vol. 107, no. 7, pp. 823–826, 2001.
[4] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14," *Cell*, vol. 75, no. 5, pp. 843–854, 1993.
[5] B. J. Reinhart *et al.*, "The 21-nucleotide let-7 RNA regulates developmental timing in caenorhabditis elegans," *Nature*, vol. 403, no. 6772, pp. 901–906, 2000.
[6] M. Lu *et al.*, "An analysis of human microRNA and disease associations," *PLoS One*, vol. 3, no. 10, 2008, Art. no. e3420.
[7] T.-H. Chu, C.-C. Yang, C.-J. Liu, M.-T. Lui, S.-C. Lin, and K.-W. Chang, "miR-211 promotes the progression of head and neck carcinomas by targeting TGFβRII," *Cancer Lett.*, vol. 337, no. 1, pp. 115–124, 2013.
[8] Z. S. Wu *et al.*, "miR-340 inhibition of breast cancer cell migration and invasion through targeting of oncoprotein c-Met," *Cancer*, vol. 117, no. 13, pp. 2842–2852, 2011.
[9] P. Gao, C. C.-L. Wong, E. K.-K. Tung, J. M.-F. Lee, C.-M. Wong, and I. O.-L. Ng, "Deregulation of microRNA expression occurs early and accumulates in early stages of HBV-associated multistep hepatocarcinogenesis," *J. Hepatol.*, vol. 54, no. 6, pp. 1177–1184, 2011.
[10] K. R. Cordes *et al.*, "miR-145 and miR-143 regulate smooth muscle cell fate and plasticity," *Nature*, vol. 460, no. 7256, pp. 705–710, 2009.
[11] X. Chen, C. C. Yan, X. Zhang, and Z.-H. You, "Long non-coding RNAs and complex diseases: From experimental results to computational models," *Briefings Bioinf.*, vol. 18, no. 4, pp. 558–576, 2016.
[12] Z.-H. You *et al.*, "PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction," *PLoS Comput. Biol.*, vol. 13, no. 3, 2017, Art. no. e1005455.
[13] L. Wang, Z.-H. You, D.-S. Huang, and F. Zhou, "Combining high speed ELM learning with a deep convolutional neural network feature encoding for predicting protein-RNA interactions," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2018.
[14] Y.-B. Wang, Z.-H. You, L.-P. Li, D.-S. Huang, F.-F. Zhou, and S. Yang, "Improving prediction of self-interacting proteins using stacked sparse auto-encoder with PSSM profiles," *Int. J. Biol. Sci.*, vol. 14, no. 8, pp. 983–991, 2018.
[15] W. Bao, Z.-H. You, and D.-S. Huang, "CIPPN: Computational identification of protein pupylation sites by using neural network," *Oncotarget*, vol. 8, no. 65, pp. 108867–108879, 2017.
[16] X. Chen *et al.*, "A novel computational model based on super-disease and miRNA for potential miRNA–disease association prediction," *Mol. BioSyst.*, vol. 13, no. 6, pp. 1202–1212, 2017.
[17] Y.-A. Huang, X. Chen, Z.-H. You, D.-S. Huang, and K. C. Chan, "ILNCSIM: Improved lncRNA functional similarity calculation model," *Oncotarget*, vol. 7, no. 18, pp. 25902–25914, 2016.
[18] Z. Shen, Y.-H. Zhang, K. Han, A. K. Nandi, B. Honig, and D.-S. Huang, "miRNA-disease association prediction with collaborative matrix factorization," *Complexity*, vol. 2017, pp. 1–9, 2017.
[19] Z. Shen, S. P. Deng, and D. Huang, "Capsule network for predicting RNA-protein binding preferences using hybrid feature [J]," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2019.
[20] P. Xuan *et al.*, "Prediction of potential disease-associated microRNAs based on random walk," *Bioinformatics*, vol. 31, no. 11, pp. 1805–1815, 2015.
[21] C. Xu *et al.*, "Prioritizing candidate disease miRNAs by integrating phenotype associations of multiple diseases with matched miRNA and mRNA expression profiles," *Mol. Biosyst.*, vol. 10, no. 11, pp. 2800–2809, 2014.
[22] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 4, pp. 905–915, Jul./Aug. 2017.
[23] X. Zeng, X. Zhang, and Q. Zou, "Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks," *Briefings Bioinf.*, vol. 17, no. 2, pp. 193–203, 2015.

[24] Q. Zou, J. Li, L. Song, X. Zeng, and G. Wang, "Similarity computation strategies in the microRNA-disease network: A survey," *Briefings Functional Genomics*, vol. 15, no. 1, pp. 55–64, 2015.

[25] L. Wang, X. Yan, M.-L. Liu, K.-J. Song, X.-F. Sun, and W.-W. Pan, "Prediction of RNA-protein interactions by combining deep convolutional neural network with feature selection ensemble method," *J. Theor. Biol.*, vol. 461, pp. 230–238, 2019.

[26] L. Wang, H.-F. Wang, S.-R. Liu, X. Yan, and K.-J. Song, "Predicting protein-protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation forest," *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 9848.

[27] Q. Zhang, Z. Shen, and D.-S. Huang, "Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network," *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, 2019.

[28] Z. Shen, S. P. Deng, and D. S. Huang, "RNA-Protein binding sites prediction via multi scale convolutional gated recurrent unit networks [J]," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2019.

[29] Q. Zhang, L. Zhu, and D.-S. Huang, "High-order convolutional neural network architecture for predicting DNA-protein binding sites," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 4, pp. 1184–1192, Jul./Aug. 2018.

[30] Z.-W. Li *et al.*, "Accurate prediction of protein-protein interactions by integrating potential evolutionary information embedded in PSSM profile and discriminative vector machine classifier," *Oncotarget*, vol. 8, no. 14, pp. 23638–23649, 2017.

[31] H. J. Jeffrey, "Chaos game representation of gene structure," *Nucleic Acids Res.*, vol. 18, no. 8, pp. 2163–2170, 1990.

[32] Z. Huang *et al.*, "HMDD v3. 0: A database for experimentally supported human microRNA–disease associations," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1013–D1017, 2018.

[33] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright, "miRBase: Tools for microRNA genomics," vol. 36, no. suppl_1, pp. D154–D158, 2007.

[34] V. J. I. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[35] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[36] T. Menzies, and Y. J. C. Hu, "Data mining for very busy people," *Computer*, vol. 36, no. 11, pp. 22–29, 2003.

[37] C. P. Chen and Z. Liu, "Broad learning system: An effective and efficient incremental learning system without the need for deep architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 10–24, Jan. 2018.

[38] C. Liang, S. Yu, and J. Luo, "Adaptive multi-view multi-label learning for identifying disease-associated candidate miRNAs," *PLoS Comput. Biol.*, vol. 15, no. 4, 2019, Art. no. e1006931.

[39] L. Wang *et al.*, "LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities," *PLoS Comput. Biol.*, vol. 15, no. 3, 2019, Art. no. e1006865.

[40] K. Zheng, Z.-H. You, L. Wang, Y. Zhou, L.-P. Li, and Z.-W. Li, "MLMDA: A machine learning approach to predict and validate MicroRNA–disease associations by integrating of heterogenous information sources," *J. Translational Med.*, vol. 17, no. 1, pp. 1–14, 2019.

[41] X. Chen, L. Huang, D. Xie, and Q. Zhao, "EGBMMDA: Extreme gradient boosting machine for MiRNA-disease association prediction," *Cell Death Dis.*, vol. 9, no. 1, 2018, Art. no. 3.

[42] X. Zeng, L. Liu, L. Lü, and Q. Zou, "Prediction of potential disease-associated microRNAs using structural perturbation method," *Bioinformatics*, vol. 34, no. 14, pp. 2425–2432, 2018.

[43] K. Zheng, Z.-H. You, L. Wang, Y. Zhou, L.-P. Li, and Z.-W. Li, "DBMDA: A unified embedding for sequence-based miRNA similarity measure with applications to predict and validate miRNA-disease associations," *Mol. Therapy-Nucleic Acids*, vol. 19, pp. 602–611, 2020.

[44] K. Zheng, L. Wang, and Z.-H. You, "CGMDA: An approach to predict and validate MicroRNA-disease associations by utilizing chaos game representation and LightGBM," *IEEE Access*, vol. 7, pp. 133314–133323, 2019.

[45] Z. Yang *et al.*, "dbDEMC: A database of differentially expressed miRNAs in human cancers," *BMC Genomics*, vol. 11, 2010, Art. no. S5.

[46] Q. Jiang *et al.*, "miR2Disease: A manually curated database for microRNA deregulation in human disease," *Nucleic Acids Res.*, vol. 37, no. suppl_1, pp. D98–D104, 2008.

[47] Y. Li *et al.*, "HMDD v2. 0: A database for experimentally supported human microRNA and disease associations," *Nucl. Acids Res.*, vol. 42, no. D1, pp. D1070–D1074, 2013.

[48] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "miRBase: From microRNA sequences to function," *Nucl. Acids Res.*, vol. 47, no. D1, pp. D155–D162, 2018.

[49] D. Wang, J. Wang, M. Lu, F. Song, and Q. J. B. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010.

[50] P. W. Lord, R. D. Stevens, A. Brass, and C. A. J. B. Goble, "Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation," *Bioinformatics*, vol. 19, no. 10, pp. 1275–1283, 2003.

[51] M. Lu *et al.*, "An analysis of human microRNA and disease associations," *Plos One*, vol. 3, no. 10, 2008, Art. no. e3420.

[52] G. L. Papadopoulos, M. Reczko, V. A. Simossis, P. Sethupathy, and A. G. Hatzigeorgiou, "The database of experimentally supported targets: A functional update of TarBase," *Nucl. Acids Res.*, vol. 37, no. suppl_1, pp. D155–D158, 2008.

[53] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies," *PLoS Comput. Biol.*, vol. 5, no. 7, 2009, Art. no. e1000443.

[54] P. Xuan *et al.*, "Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors," *Plos One*, vol. 8, no. 8, 2013, Art. no. e70204.

[55] T. van Laarhoven, S. B. Nabuurs, and E. J. B. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[56] X. Chen *et al.*, "WBSMDA: Within and between score for MiRNA-disease association prediction," *Sci. Rep.*, vol. 6, 2016, Art. no. 21106.

[57] H. J. Jeffrey, "Chaos game representation of gene structure," *Nucl. Acids Res.*, vol. 18, no. 8, pp. 2163–2170, 1990.

[58] B. Liu, "BioSeq-Analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches," *Briefings Bioinf.*, vol. 20, no. 4, pp. 1280–1294, 2019.

[59] Z. Chen *et al.*, "iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data," *Briefings Bioinf.*, vol. 21, pp. 1047–1057, 2019.

[60] B. Liu, X. Gao, and H. Zhang, "BioSeq-Analysis2.0: An updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches," *Nucleic Acids Res.*, vol. 47, no. 20, 2019, Art. no. e127.

[61] J.-Y. An *et al.*, "Identification of self-interacting proteins by exploring evolutionary information embedded in PSI-BLAST-constructed position specific scoring matrix," *Oncotarget*, vol. 7, no. 50, pp. 82440–82449, 2016.

[62] J.-Y. An, Z.-H. You, X. Chen, D.-S. Huang, G. Yan, and D.-F. Wang, "Robust and accurate prediction of protein self-interactions from amino acids sequence using evolutionary information," *Mol. Biosyst.*, vol. 12, no. 12, pp. 3702–3710, 2016.

[63] L. Zhu, S.-P. Deng, Z.-H. You, and D.-S. Huang, "Identifying spurious interactions in the protein-protein interaction networks using local similarity preserving embedding," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 2, pp. 345–352, Mar./Apr. 2017.

[64] L. Zhu, Z.-H. You, and D.-S. Huang, "Increasing the reliability of protein–protein interaction networks via non-convex semantic embedding," *Neurocomputing*, vol. 121, pp. 99–107, 2013.

[65] L. Zhu, Z.-H. You, D.-S. Huang, and B. Wang, "t-LSE: A novel robust geometric approach for modeling protein-protein interaction networks," *PLoS One*, vol. 8, no. 4, 2013, Art. no. e58368.

[66] Y.-K. Lei, Z.-H. You, Z. Ji, L. Zhu, and D.-S. Huang, "Assessing and predicting protein interactions by combining manifold embedding with multiple information integration," *BMC Bioinf.*, vol. 13, 2012, Art. no. S3.

[67] Z.-H. You, Y.-K. Lei, J. Gui, D.-S. Huang, and X. Zhou, "Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data," *Bioinformatics*, vol. 26, no. 21, pp. 2744–2751, 2010.

[68] I. J. Goodfellow *et al.*, "An empirical investigation of catastrophic forgetting in gradient-based neural networks [J]," 2013, *arXiv:1312.6211*.

Kai Zheng received the BE degree in computer science and technology from Central South University, Changsha, China, in 2017. He is currently working toward the PhD degree in Central South University. His current research interests include data mining, pattern recognition, recommender systems, machine learning, deep learning, intelligent information processing and its applications in bioinformatics.

Zhu-Hong You (Member, IEEE) received the BE degree in electronic information science and engineering from Hunan Normal University, Changsha, China, in 2005, and the PhD degree in control science and engineering from the University of Science & Technology of China (USTC), Hefei, China, in 2010. From June 2008 to November 2009, he was a visiting research fellow at the Center of Biotechnology and Information, Cornell University. He is currently a professor with the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Ürümqi, China. His current research interests include neural networks, intelligent information processing, sparse representation, and its applications in bioinformatics.

Lei Wang received the PhD degree from the School of Computer Science Technology at China University of Mining and Technology, Jiangsu, China, in 2018. He is currently a postdoctoral with the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Ürümqi, China. His research interests include data mining, pattern recognition, machine learning, deep learning, computational biology, and bioinformatics. He acted as reviewers for many international journals, such as *Scientific Reports*, *Current Protein & Peptide Science*, *Computational Biology and Chemistry*, *Soft Computing*, and *Journal of Computational Biology*.

Yi-Ran Li received the bachelor's degree in electrical engineering and automation from the China University of Mining and Technology, Xuzhou, China, in 2017. She is currently working toward the master's degree in the school of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China. Her current research interests include hyperspectral detection.

Ji-Ren Zhou received the bachelor's degree in civil engineering from the China University of Mining and Technology, Xuzhou, China, in 2019. Currently, he is working toward the master's degree in the Hong Kong University of Science and Technology. His research interests have turned to data mining, machine learning, deep learning, and bioinformatics.

Hai-Tao Zeng received the BS degree from the school of geomatics, Shandong University of Science and Technology, Qingdao, China, in 2017. He is currently working toward the graduate degree in computer science in the School of Mechanical Electronic and Information Engineering, China University of Mining and Tech- nology, Beijing, China. He is also an intern with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, and image processing.