IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 8, 2017, accepted October 17, 2017, date of publication October 26, 2017,
date of current version November 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2766758

Predicting MicroRNA-Disease Associations Using Network Topological Similarity Based on DeepWalk

GUANGHUI LI[1], JIAWEI LUO[2], QIU XIAO[2], CHENG LIANG[3], PINGJIAN DING[2], (Student Member, IEEE), AND BUWEN CAO[4]

[1]School of Information Engineering, East China Jiaotong University, Nanchang 330013, China
[2]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
[3]College of Information Science and Engineering, Shandong Normal University, Jinan 250000, China
[4]College of Information and Electronic Engineering, Hunan City University, Yiyang 413000, China

Corresponding author: Jiawei Luo (luojiawei@hnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61572180 and Grant 61602283, in part by the Key Project of the Education Department of Hunan Province under Grant 17A037, and in part by the Hunan Provincial Innovation Foundation for Postgraduate under Grant CX2017B102.

ABSTRACT Recently, increasing experimental studies have shown that microRNAs (miRNAs) involved in multiple physiological processes are connected with several complex human diseases. Identifying human disease-related miRNAs will be useful in uncovering novel prognostic markers for cancer. Currently, several computational approaches have been developed for miRNA-disease association prediction based on the integration of additional biological information of diseases and miRNAs, such as disease semantic similarity and miRNA functional similarity. However, these methods do not work well when this information is unavailable. In this paper, we present a similarity-based miRNA-disease prediction method that enhances the existing association discovery methods through a topology-based similarity measure. DeepWalk, a deep learning method, is utilized in this paper to calculate similarities within a miRNA-disease association network. It shows superior predictive performance for 22 complex diseases, with area under the ROC curve scores ranging from 0.805 to 0.937 by using five-fold cross-validation. In addition, case studies on breast cancer, lung cancer, and prostatic cancer further justify the use of our method to discover latent miRNA-disease pairs.

INDEX TERMS Deep learning, disease-related microRNAs, microRNA-disease association, similarity measure.

I. INTRODUCTION

MicroRNAs (miRNAs), as a class of short non-coding RNA molecules (19∼24 nt), act as negative regulators of gene expression by binding to the 3'-UTRs of target mRNAs [1], [2]. Recently, increasing evidence has indicated that mutation and functional disorders of miRNAs are connected with the development and progression of various complex human diseases [3]–[5]. Consequently, identifying disease-related miRNAs will be beneficial for investigating mechanisms of pathogenicity and promoting the diagnosis and treatment of human disease.

It has proved effective using biomedical technologies such as microarrays and PCR to identify the miRNAs associated with individual diseases. However, these biological experimental methods can be costly and time-consuming. Encouragingly, more and more experimentally verified disease-miRNA association databases have become available, for example, the Human miRNA Disease Database (HMDD) [6] and miR2Disease [7]. The establishment of these miRNA-related biological datasets has created a foundation for predictive research. Therefore, there exists a strong need to develop efficient computational models to predict new types of disease-related miRNAs on a large scale.

Many computational models have been put forward to excavate latent miRNA-disease associations, under the assumption that functionally similar miRNAs are likely to be linked with similar diseases, and vice versa [8]–[12]. For instance, Jiang *et al*. [13] introduced a computational

model that integrates miRNA functional similarity data, phenotype similarity data, and experimentally verified disease-miRNA association data to predict latent interactions between miRNAs and diseases by using a hypergeometric distribution. Unfortunately, the efficacy of this method was seriously restrained by predicted miRNA-target interactions with high false-positive and false-negative rates [14], [15]. Xuan *et al*. [16] proposed a highly performing prediction algorithm called HDMP based on weighted k most similar neighbors, which computed miRNA functional similarity by utilizing semantic similarity and phenotype similarity data of their associated diseases. These aforementioned methods only considered local miRNA and disease information in their computational models but did not utilize global network association information, which can significantly enhance prediction performance. Thus, researchers have presented many global network similarity-based computational modes based on random walks. Chen *et al*. [17] developed the first global model called RWRMDA by performing random walks restarting on a constructed miRNA-miRNA functional similarity network to infer new miRNA-disease pairs. In addition, Shi *et al*. [18] used random walk analysis to rank miRNA-disease pairs by searching for functional associations between disease-related genes and miRNA targeted genes in a protein-protein interaction network. Similarly, Xuan *et al*. [19] presented another new model called MIDP based on random walks, which assigned different transition matrices to labeled and unlabeled miRNAs of a specific disease during the iterative process. Luo and Xiao [20] also introduced a novel approach that performed unbalanced bi-random walks on bipartite subgraphs to identify disease-miRNA interactions. Liu *et al*. [21] extended random walk with restart on a constructed heterogeneous network to infer the relationship between disease and miRNA. To further enhance the prediction accuracy, Chen *et al*. [22] developed a computational model named WBSMDA, which took advantage of within and between scores of each candidate disease-miRNA pair to discover disease-miRNA associations. WBSMDA improved disease semantic similarity and miRNA functional similarity by integrating Gaussian interaction profile kernel similarity. Recently, Chen *et al*. [23] has further proposed a heterogeneous graph-based model called HGIMDA, which infers the potential association likelihood of each candidate disease-miRNA pair by counting all routes of length three. In addition, some machine learning-based models have been proposed to identify latent relationships between diseases and miRNAs [24], [25]. However, none of the above methods have satisfactory performance, and most of them rely on heterogeneous omics data. On the other hand, in a miRNA-disease bipartite network, each miRNA-disease pair is validated by biological experiments, which provides important prior information and produces direct benefits to the prediction of novel disease-miRNA pairs.

In this study, we propose a similarity-based miRNA-disease prediction method that adopts a deep learning algorithm, DeepWalk [26], to extract features of vertices in the miRNA-disease bipartite network, which can be adapted to compute the topological similarities of two vertices [27]. The resulting similarity measure is used to infer disease-related miRNAs based on a rule-based inference method [28] that uses disease-disease similarities as the input for miRNA-disease prediction. The experimental results of five-fold cross-validation and case studies support the ability of our method to infer novel miRNA-disease pairs, which may be of great use in further biological experiments.

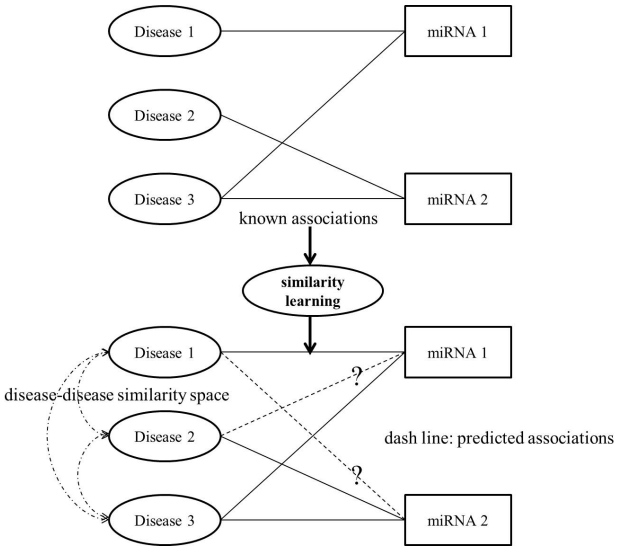

FIGURE 1. Overall workflow of our method for identifying latent miRNA-disease pairs.

II. MATERIALS AND METHODS

A. METHOD OVERVIEW

The method proposed in this study is based on the topological structure of a miRNA-disease bipartite network. The association discovery method can be separated into three steps: (i) data collection, (ii) similarity learning, and (iii) association discovery. First, a bipartite network containing the topological interactions of existing miRNAs and diseases is constructed. Second, similarity scores of disease-disease pairs are learned based on the topology of this network. Finally, predictions and evaluations of new disease-miRNA pairs are made based on these similarities. Fig. 1 illustrates the overall workflow of our method.

B. HUMAN miRNA-DISEASE ASSOCIATION DATASET

The disease-miRNA association dataset was downloaded from the HMDD v2.0 database. There were 5424 distinct experimentally confirmed associations between 378 diseases and 495 miRNAs after filtering out duplicate records. Briefly, the number of diseases and miRNAs are represented by variables nd and nm, respectively. In addition, two other public databases (i.e., dbDEMC [29] and PhenomiR2.0 [30]) were adopted to assess the candidate miRNA predictions with case studies.

C. SIMILARITY LEARNING

DeepWalk [26], a deep learning method, vectorizes the vertices (e.g., diseases and miRNAs) of the network for similarity computation. This method utilizes local information from truncated random walks to learn vertex representation by maximizing the probability of observing vertex v_i in view of all vertices previously visited up to the current point in the random walk. DeepWalk has two main components. First, for each vertex v_i, γ random walks with length t are conducted, with v_i as the starting vertex. Second, for each walk, the vertex representation is updated with the SkipGram algorithm [31]. SkipGram maximizes the co-occurrence likelihood of the vertices that come into view within a window w using an independent assumption as follows:

$$\Pr(\{v_{i-w}, \cdots, v_{i+w}\} \backslash v_i | \Phi(v_i)) = \prod_{j=i-w, j\neq i}^{i+w} \Pr(v_j | \Phi(v_i)) \quad (1)$$

where Φ denotes the latent topological representation associated with each vertex v_i. Φ is represented by a $|V| \times d$ matrix, where $|V|$ is the cardinality of vertex set V, and d is the dimension of the vertex vector. To speed up the training time, $\Pr(v_j|\Phi(v_i))$ is factorized with Hierarchical Softmax [32] by allocating the vertices to the leaves of a binary tree, and $\Pr(v_j|\Phi(v_i))$ is then computed as follows:

$$\Pr(v_j | \Phi(v_i)) = \prod_{l=1}^{\lceil \log |V| \rceil} 1/(1 + e^{-\Phi(v_i)\cdot\psi(b_l)}) \quad (2)$$

where $\psi\ (b_l)$ represents the parent of tree node b_l. $(b_0, b_1, \ldots, b_{\log|v|})$ is the sequence of tree nodes to identify v_j, where $b_0 =$ root and $b +_{\log|v|} + = v_j$.

After completing the training, the output of DeepWalk is a latent topological representation (i.e., d-dimensional vector) of vertices in the network. Therefore, the similarity of two vertices u and v can be computed by using cosine similarity as follows:

$$sim(u, v) = \frac{\sum_{k=1}^{d} u_k v_k}{\sqrt{\sum_{k=1}^{d} u_k^2} \sqrt{\sum_{k=1}^{d} v_k^2}} \quad (3)$$

where d is the dimension, and u_i and v_i are the components of vector u and v, respectively.

D. DISEASE-BASED SIMILARITY INFERENCE

We adapted a rule-based inference method, drug-based similarity inference (DBSI) [28], which was derived from complex network theory [33] to predict disease-related miRNA candidates with disease-disease similarities. The main idea of DBSI is as follows: if a disease is associated with a miRNA, then other diseases similar to that disease will also possibly be associated with the miRNA. In terms of the pair (d_i, m_j), a linkage between disease d_i and miRNA m_j is decided by the following prediction score:

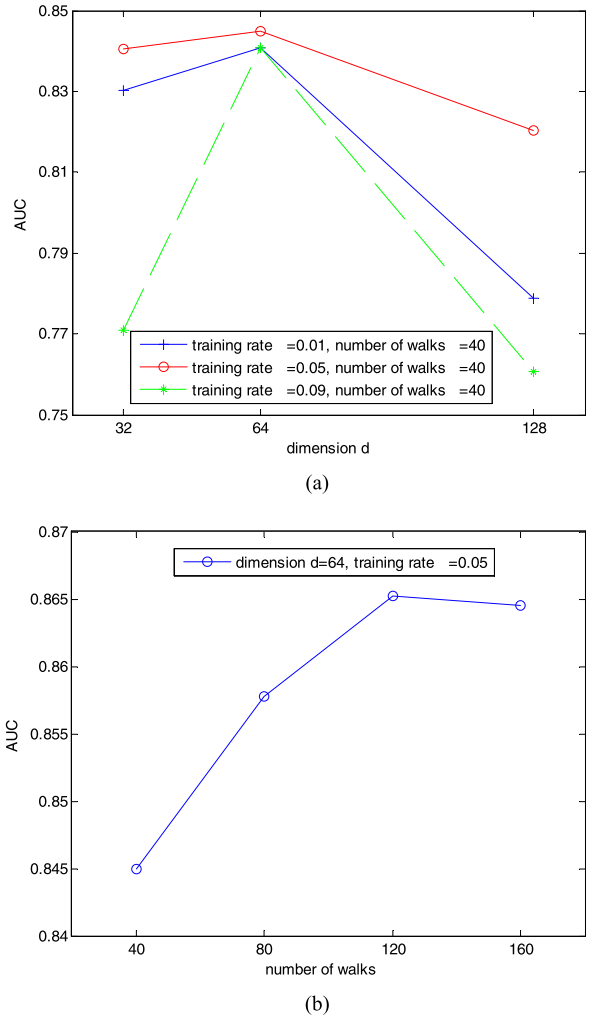

FIGURE 2. The average AUCs of varying the parameters. (a) The effect of different dimensionality and training ratio. (b) The effect of different number of walks.

$$Score_{DBSI}(d_i, m_j) = \frac{\sum_{l=1, l\neq i}^{nd} sim(d_i, d_l) a_{lj}}{\sum_{l=1, l\neq i}^{nd} sim(d_i, d_l)} \quad (4)$$

where $sim(d_i, d_l)$ is the similarity between disease d_i and disease d_l obtained from (3), and $a_{lj} = 1$ if an association between disease d_l and miRNA m_j is known; otherwise, $a_{lj} = 0$.

Operationally, for a disease d_i as the input query, each associated score is normalized as follows:

$$Score^*_{DBSI}(d_i, m_j) = \frac{Score_{DBSI}(d_i, m_j) - Min(d_i, \cdot)}{Max(d_i, \cdot) - Min(d_i, \cdot)} \quad (5)$$

where $Max(d_i, \cdot)$ and $Min(d_i, \cdot)$ represent the maximum and minimum associated score, respectively, of disease d_i with miRNAs that have no known association with d_i.

TABLE 1. Prediction results of our method, RWRMDA, MIDP, and WBSMDA over five-fold cross-validation.

Disease name	No. of associated miRNAs	AUC			
		Our Method	RWRMDA	MIDP	WBSMDA
Breast Neoplasms	202	0.861	0.801	0.808	0.827
Hepatocellular Carcinoma	214	0.825	0.753	0.762	0.792
Non-Small-Cell Lung Carcinoma	95	0.890	0.817	0.846	0.841
Renal Cell Carcinoma	107	0.835	0.782	0.809	0.826
Squamous Cell Carcinoma	80	0.877	0.839	0.870	0.842
Colonic Neoplasms	78	0.884	0.799	0.844	0.791
Colorectal Neoplasms	147	0.854	0.793	0.810	0.764
Endometriosis	62	0.840	0.777	0.792	0.795
Esophageal Neoplasms	74	0.842	0.742	0.865	0.828
Glioblastoma	96	0.838	0.771	0.809	0.818
Glioma	71	0.887	0.860	0.887	0.844
Head and Neck Neoplasms	64	0.886	0.831	0.867	0.852
Heart Failure	120	0.805	0.762	0.782	0.795
Leukemia, Myeloid, Acute	64	0.856	0.778	0.846	0.841
Lung Neoplasms	132	0.937	0.863	0.898	0.864
Medulloblastoma	62	0.842	0.770	0.795	0.816
Melanoma	141	0.860	0.770	0.816	0.822
Ovarian Neoplasms	114	0.900	0.877	0.892	0.866
Pancreatic Neoplasms	99	0.911	0.861	0.888	0.864
Prostatic Neoplasms	118	0.888	0.804	0.829	0.883
Stomach Neoplasms	174	0.857	0.773	0.781	0.790
Urinary Bladder Neoplasms	92	0.860	0.787	0.836	0.866
Average AUC		0.865	0.801	0.833	0.829

III. EXPERIMENTS AND RESULTS

A. EVALUATION METRICS

To systematically evaluate the prediction accuracy of our method, five-fold cross-validation was implemented on the basis of disease-miRNA pairs obtained from the HMDD database. In the 5-fold cross validation framework, for a given disease d, the labeled d-associated miRNAs are partitioned into five disjoint subsections at random: one subsection is used for testing and the other four subsections for training through multiple iterations. The similarity computation for diseases is connected with known disease-miRNA pairs; thus, disease-disease similarities are recalculated in each repetition of the cross-validation experiments. The area under the ROC curve (AUC) was used to assess the quality of the predicted associations.

B. EFFECT OF PARAMETERS ON THE PERFORMANCE OF OUR METHOD

There are five parameters in DeepWalk. In this article, we fixed the window size $w = 10$ and the walk length $t = 40$ to highlight the local structure according to a previous study [26]. The three other parameters were determined by a grid search over the parameter ranges specified in previous work (i.e., dimension $d = \{32, 64, 128\}$, training rate $\alpha = \{0.01, 0.05, 0.09\}$, and number of walks $\gamma = \{40, 80, 120, 160\}$) [26]. Here, to study the effect of these three parameters on the prediction accuracy, we varied the values of d, α, and γ in 5-fold cross-validation experiments. Fig. 2 presents the average AUC values obtained from our method for different values of d, α, and γ. As is shown in the figure, the best prediction performance is achieved at $d = 64$, $\alpha = 0.05$, and $\gamma = 120$. Therefore, we set $d = 64$, $\alpha = 0.05$, and $\gamma = 120$ as default values in our experiment.

C. PREDICTION PERFORMANCE EVALUATION

We compared our method with RWRMDA [17], MIDP [19], and WBSMDA [22], which serve as advanced computational prediction models to discover potential candidate miRNAs. Since RWRMDA and MIDP were developed based on the association data from the previous version of HMDD, the similarity of diseases or miRNAs pairs was recalculated with the latest version of HMDD. Many diseases have connections with only a few miRNAs; hence, the performance of five-fold cross-validation may not be sufficient for them. Consequently, we only considered 22 diseases associated with at least 60 miRNAs, as confirmed by our experiments.

As is shown in Table 1, our method achieves the best performance for all the 22 diseases except esophageal neoplasms and urinary bladder neoplasms, which performed better with MIDP and WBSMDA, respectively. The average AUC value achieved by our tool was 0.865, with a minimum of 0.805 for heart failure and a maximum of 0.937 for lung neoplasms, whereas the respective AUCs of RWRMDA, MIDP, and WBSMDA were 0.801, 0.833, and 0.829. The average AUCs obtained by our method were 6.4%, 3.2%, and 3.6% higher than those of the other three methods. The ROC curves of each method using five-fold cross-validation are shown in Fig. 3.

Moreover, for purpose of comparison we select the top 10, top 30, top 50, top 80 and top 100 predicted associations for each disease as potential candidates. For these selections,

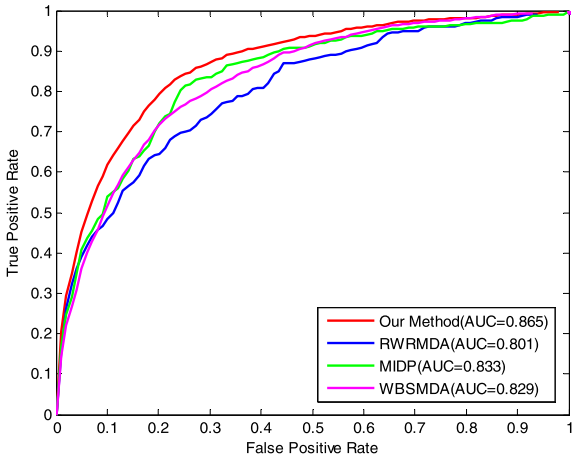

FIGURE 3. The ROC curves and average AUCs of each method for 22 diseases.

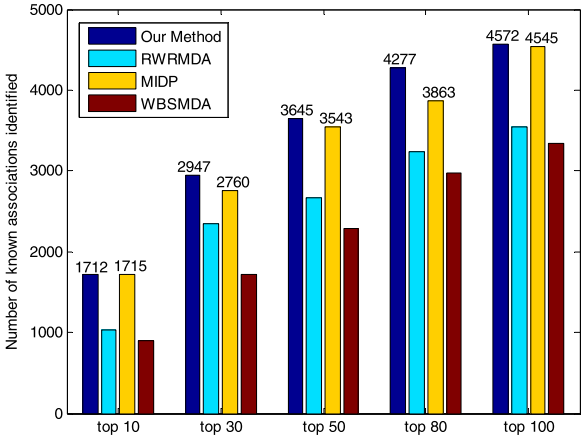

FIGURE 4. Comparison of the number of known associations identified by different methods.

the number of known disease-miRNA associations detected by our method and three other methods are shown in Fig. 4. From Fig. 4, we can see that the result of our method performs better than the other three algorithms in all selections, except in top 10 which is only slightly lower than that of MIDP. These prediction results illustrate that our method achieves reliable prediction performance, especially because our method only depends on the topological structure of the miRNA-disease bipartite network.

D. DEEPWALK-BASED VS. GAUSSIAN INTERACTION PROFILE KERNEL-BASED SIMILARITY MEASURES

We compared DeepWalk-based to Gaussian interaction profile kernel-based topology similarity measures [34], predicting miRNA-disease association for the 22 aforementioned diseases. As for the latter, we first adopted the Gaussian kernel to compute the similarity of each disease-disease pair based on the topological structure of the miRNA-disease bipartite network. We then assembled similarity measures with the rule-based inference method, DBSI, to obtain the

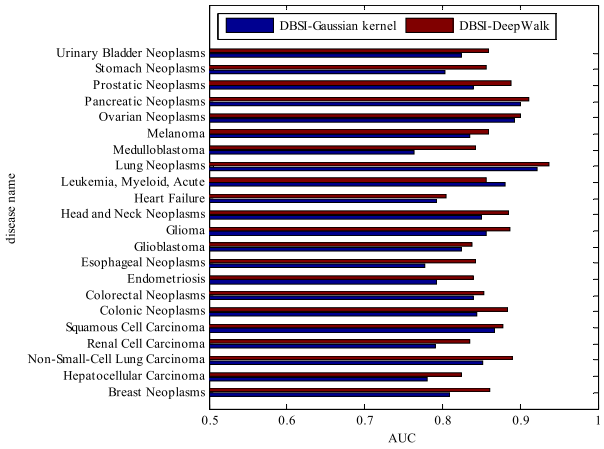

FIGURE 5. Comparison of average AUCs (5-fold validation) using different similarity measures.

association probability of each candidate disease-miRNA pair. Fig. 5 illustrates that DeepWalk is superior to the method based on Gaussian kernel in terms of AUC for these diseases, except for acute myeloid leukemia. For example, the AUC scores achieved by DeepWalk for breast cancer, lung cancer, and prostatic cancer are 0.861, 0.937, and 0.888, respectively, whereas the respective AUCs obtained by Gaussian kernel are 0.810, 0.922, and 0.839. The comparison results also demonstrate that applying DeepWalk for similarity computation can improve prediction accuracy.

E. CASE STUDIES

Additionally, in an attempt to assess the ability of our method to uncover potential disease-associated miRNAs, case studies of three important complex human diseases were investigated by considering all known associations included in the HMDD database as a training set. The prediction-associated miRNAs for each selected disease were validated based on two independent databases, dbDEMC [29] and PhenomiR2.0 [30], and experimental literature.

Particularly in developed countries, women's cancer deaths are primarily caused by breast cancer. Recently, accumulating evidence has shown that many miRNAs are related to the formation of diverse cancers comprised of breast neoplasms. For instance, hsa-mir-205 regulates ErbB3 by binding to its 3'-UTR, which is significantly under-expressed in breast tumors [35]. Discovering more miRNAs associated with breast cancer will aid in accurately assessing clinical results. The case study of breast neoplasms was implemented with our method. As a result, 14 and 26 out of the top 15 and top 30 potentially related miRNAs have been directly shown to be linked with breast neoplasms through dbDEMC and PhenomiR2.0 databases (see Table 2). Furthermore, some predicted miRNAs were verified by previously published literature. Specifically, hsa-mir-378a (6th in the prediction list) represses the expression of two genes in breast neoplasms, ERR γ and GABPA [36]. Hsa-mir-574 (18th in the prediction list) has been identified as a potentially novel prognostic

TABLE 2. The top 30 predicted breast neoplasm-associated miRNAs.

Rank	miRNAs	Evidences	Rank	miRNAs	Evidences
1	hsa-mir-106a	dbDEMC, PhenomiR2.0	16	hsa-mir-449b	dbDEMC
2	hsa-mir-142	PhenomiR2.0	17	hsa-mir-130b	dbDEMC, PhenomiR2.0
3	hsa-mir-138	dbDEMC	18	hsa-mir-574	literature
4	hsa-mir-99a	dbDEMC, PhenomiR2.0	19	hsa-mir-362	literature
5	hsa-mir-150	dbDEMC, PhenomiR2.0	20	hsa-mir-30e	PhenomiR2.0
6	hsa-mir-378a	literature	21	hsa-mir-144	dbDEMC
7	hsa-mir-130a	dbDEMC, PhenomiR2.0	22	hsa-mir-181d	dbDEMC, PhenomiR2.0
8	hsa-mir-15b	dbDEMC, PhenomiR2.0	23	hsa-mir-196b	dbDEMC, PhenomiR2.0
9	hsa-mir-192	dbDEMC, PhenomiR2.0	24	hsa-mir-372	dbDEMC, PhenomiR2.0
10	hsa-mir-92b	dbDEMC	25	hsa-mir-330	dbDEMC, PhenomiR2.0
11	hsa-mir-99b	dbDEMC, PhenomiR2.0	26	hsa-mir-449a	dbDEMC, PhenomiR2.0
12	hsa-mir-212	dbDEMC, PhenomiR2.0	27	hsa-mir-542	literature
13	hsa-mir-98	dbDEMC, PhenomiR2.0	28	hsa-mir-181c	dbDEMC, PhenomiR2.0
14	hsa-mir-185	dbDEMC, PhenomiR2.0	29	hsa-mir-211	dbDEMC, PhenomiR2.0
15	hsa-mir-186	dbDEMC, PhenomiR2.0	30	hsa-mir-95	dbDEMC, PhenomiR2.0

TABLE 3. The top 30 predicted lung neoplasm-associated miRNAs.

Rank	miRNAs	Evidences	Rank	miRNAs	Evidences
1	hsa-mir-16	dbDEMC	16	hsa-mir-151a	literature
2	hsa-mir-15a	dbDEMC, PhenomiR2.0	17	hsa-mir-10a	dbDEMC, PhenomiR2.0
3	hsa-mir-195	dbDEMC, PhenomiR2.0	18	hsa-mir-92b	dbDEMC, PhenomiR2.0
4	hsa-mir-106b	dbDEMC, PhenomiR2.0	19	hsa-mir-378a	literature
5	hsa-mir-15b	dbDEMC, PhenomiR2.0	20	hsa-mir-204	dbDEMC, PhenomiR2.0
6	hsa-mir-99a	dbDEMC, PhenomiR2.0	21	hsa-mir-194	dbDEMC
7	hsa-mir-429	dbDEMC	22	hsa-mir-208a	PhenomiR2.0
8	hsa-mir-130a	dbDEMC, PhenomiR2.0	23	hsa-mir-149	dbDEMC, PhenomiR2.0
9	hsa-mir-141	dbDEMC, PhenomiR2.0	24	hsa-mir-708	dbDEMC
10	hsa-mir-122	dbDEMC, PhenomiR2.0	25	hsa-mir-196b	dbDEMC, PhenomiR2.0
11	hsa-mir-23b	dbDEMC, PhenomiR2.0	26	hsa-mir-129	dbDEMC
12	hsa-mir-193b	dbDEMC, PhenomiR2.0	27	hsa-mir-302b	dbDEMC, PhenomiR2.0
13	hsa-mir-20b	dbDEMC, PhenomiR2.0	28	hsa-mir-625	dbDEMC
14	hsa-mir-296	PhenomiR2.0	29	hsa-mir-152	dbDEMC, PhenomiR2.0
15	hsa-mir-451a	dbDEMC	30	hsa-mir-342	dbDEMC, PhenomiR2.0

TABLE 4. The top 30 predicted prostatic neoplasm-associated miRNAs.

Rank	miRNAs	Evidences	Rank	miRNAs	Evidences
1	hsa-mir-18a	dbDEMC, PhenomiR2.0	16	hsa-mir-142	PhenomiR2.0
2	hsa-mir-155	dbDEMC, PhenomiR2.0	17	hsa-mir-429	unconfirmed
3	hsa-mir-19b	dbDEMC	18	hsa-mir-24	dbDEMC
4	hsa-mir-210	dbDEMC, PhenomiR2.0	19	hsa-mir-103a	unconfirmed
5	hsa-mir-29c	dbDEMC, PhenomiR2.0	20	hsa-let-7i	dbDEMC, PhenomiR2.0
6	hsa-mir-19a	dbDEMC, PhenomiR2.0	21	hsa-mir-30a	PhenomiR2.0
7	hsa-mir-10b	dbDEMC, PhenomiR2.0	22	hsa-mir-181a	dbDEMC
8	hsa-mir-9	literature	23	hsa-mir-30b	dbDEMC, PhenomiR2.0
9	hsa-let-7f	dbDEMC	24	hsa-mir-196a	dbDEMC
10	hsa-let-7e	dbDEMC, PhenomiR2.0	25	hsa-mir-125a	dbDEMC, PhenomiR2.0
11	hsa-mir-199b	dbDEMC, PhenomiR2.0	26	hsa-mir-192	dbDEMC
12	hsa-mir-138	literature	27	hsa-mir-18b	dbDEMC
13	hsa-let-7g	dbDEMC, PhenomiR2.0	28	hsa-mir-135a	dbDEMC
14	hsa-mir-7	dbDEMC	29	hsa-mir-625	dbDEMC
15	hsa-mir-150	dbDEMC, PhenomiR2.0	30	hsa-mir-451a	dbDEMC

indicator of breast cancer, which is significantly down-regulated in tumor samples [37]. Hsa-mir-362 (19th in the prediction list) has been shown to be differentially expressed in MCF-7 human breast cancer cells [38]. It has also been shown that hsa-mir-542 (27th in the prediction list) is significantly down-regulated in breast cancer cells [39].

Lung cancer is one of the pervasive malignant tumors with the highest mortality. The top 30 predicted lung neoplasm-associated miRNAs are listed in Table 3. From the table, 15 out of the top 15 and 28 out of the top 30 potentially associated miRNAs were validated by the two aforementioned databases. In addition, the other two candidates are supported by published literature. Specifically, hsa-mir-151a (16th in the prediction list) is markedly up-regulated in non-small cell lung carcinoma compared with non-tumorous tissue [40]. Hsa-mir-378a (19th in the prediction list) is

significantly overexpressed in squamous cell carcinoma when compared with lung adenocarcinoma [41].

Prostatic cancer is the second major cause of male cancer-related deaths in developed countries. We implemented our method to prioritize candidate prostatic neoplasm-associated miRNAs, and results show that 13 and 26 out of the top 15 and top 30 predicted miRNAs were contained in dbDEMC and PhenomiR2.0 (see Table 4). Two candidates, hsa-mir-9 [42] and hsa-mir-138 [43], were verified to be correlated with prostatic neoplasms by experimental literature. In addition, hsa-mir-429 (17th in the prediction list) is the second ranked miRNA by RWRMDA and Jiang's method. Hsa-mir-103a (19th in the prediction list) is ranked No. 3 by KRLSM [12], which indirectly confirms that it is probably associated with prostatic cancer.

In summary, the results of cross-validation and case studies of several common diseases fully illustrate that our method achieves excellent prediction performance. Therefore, we have further used our method to rank potential miRNAs for each human disease contained in HMDD (shown in Supplementary Table S1), in the hope that these prediction results can be verified in future scientific research.

IV. CONCLUSION

Identifying novel miRNA-disease associations is important for exploring disease pathogenesis and to further improve human medicine. In this paper, a similarity-based method was designed to identify latent miRNA-disease pairs. First, we adopted a deep learning algorithm, DeepWalk, to determine the similarity of each disease-disease pair based on a known disease-miRNA bipartite network. Then, with a rule-based inference method, DBSI, similarity measures were assembled to compute the association likelihood of each candidate disease-miRNA pair. To validate the prediction accuracy of our approach, five-fold cross-validation was implemented with a miRNA-disease association dataset. In addition, case studies on breast cancer, lung cancer, and prostatic cancer were done, and 30, 30, and 28 of the top 30 predicted miRNAs for each of these three principal human diseases have been verified by the latest experimental literature and two independent databases.

Despite this successful exploitation of bipartite network topology through application of DeepWalk for similarity computation in miRNA-disease interaction prediction, there are also some inevitable limitations expected to be improved in future research. To begin with, the proposed method fails to predict associations for new diseases or miRNAs that do not exist within the network because our method is only informed by known miRNA-disease associations. To solve this problem, a hybrid similarity measure that includes both topological and non-topological features, like disease semantic similarity data and miRNA functional similarity data, may facilitate application of this methodology to predict new diseases or miRNAs. Second, the currently known miRNA-disease associations are insufficient. Therefore, a heterogeneous network that integrates additional disease-gene and miRNA-gene associations can be used for similarity learning, which may potentially improve prediction results. Finally, there are five parameters in our method, and the selection of appropriate parameters for different diseases needs to be properly addressed.

REFERENCES

[1] D. P. Bartel, "MicroRNAs: Genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, Jan. 2004.
[2] S. Chatterjee and H. Großhans, "Active turnover modulates mature microRNA activity in *Caenorhabditis elegans*," *Nature*, vol. 461, no. 7263, pp. 546–549, Sep. 2009.
[3] I. Alvarez-Garcia and E. A. Miska, "MicroRNA functions in animal development and human disease," *Development*, vol. 132, no. 21, pp. 4653–4662, 2005.
[4] N. Lynam-Lennon, S. G. Maher, and J. V. Reynolds, "The roles of microRNA in cancer and apoptosis," *Biol. Rev.*, vol. 84, no. 1, pp. 55–71, Feb. 2009.
[5] N. Meola, V. A. Gennarino, and S. Banfi, "MicroRNAs and genetic diseases," *Pathogenetics*, vol. 2, no. 1, p. 7, 2009.
[6] Y. Li *et al.*, "HMDD v2.0: A database for experimentally supported human microRNA and disease associations," *Nucl. Acids Res.*, vol. 42, pp. D1070–D1074, Jan. 2014.
[7] Q. Jiang *et al.*, "miR2Disease: A manually curated database for microRNA deregulation in human disease," *Nucleic Acids Res.*, vol. 37, pp. D98–D104, Jan. 2009.
[8] X. Zeng, X. Zhang, and Q. Zou, "Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks," *Briefings Bioinform.*, vol. 17, no. 2, pp. 193–203, Jun. 2015.
[9] W. Lan, J. Wang, M. Li, J. Liu, F.-X. Wu, and Y. Pan, "Predicting microRNA-disease associations based on improved microRNA and disease similarities," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2016.2586190.
[10] Q. Zou, J. Li, L. Song, X. Zeng, and G. Wang, "Similarity computation strategies in the microRNA-disease network: A survey," *Briefings Funct. Genomics*, vol. 15, no. 1, pp. 55–64, Jan. 2016.
[11] J. Luo, P. Ding, C. Liang, B. Cao, and X. Chen, "Collective prediction of disease-associated miRNAs based on transduction learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2016.2599866.
[12] J. Luo, Q. Xiao, C. Liang, and P. Ding, "Predicting microRNA-disease associations using Kronecker regularized least squares based on heterogeneous omics data," *IEEE Access*, vol. 5, pp. 2503–2513, Feb. 2017.
[13] Q. Jiang *et al.*, "Prioritization of disease microRNAs through a human phenome-microRNAome network," *BMC Syst. Biol.*, vol. 4, p. S2, Jun. 2010.
[14] W. Ritchie, S. Flamant, and J. E. J. Rasko, "Predicting microRNA targets and functions: Traps for the unwary," *Nature Methods*, vol. 6, no. 6, pp. 397–398, Jun. 2009.
[15] B. Liu, J. Li, and M. J. Cairns, "Identifying miRNAs, targets and functions," *Briefings Bioinform.*, vol. 15, no. 1, pp. 1–19, Jan. 2014.
[16] P. Xuan *et al.*, "Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors," *PLoS ONE*, vol. 8, no. 8, p. e70204, Aug. 2013.
[17] X. Chen, M.-X. Liu, and G.-Y. Yan, "RWRMDA: Predicting novel human microRNA–disease associations," *Mol. Biosyst.*, vol. 8, no. 10, pp. 2792–2798, 2012.
[18] H. Shi *et al.*, "Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes," *BMC Syst. Biol.*, vol. 7, p. 101, Dec. 2013.
[19] P. Xuan *et al.*, "Prediction of potential disease-associated microRNAs based on random walk," *Bioinformatics*, vol. 31, no. 11, pp. 1805–1815, Jun. 2015.
[20] J. Luo and Q. Xiao, "A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on heterogeneous network," *J. Biomed. Inform.*, vol. 66, pp. 194–203, Feb. 2017.
[21] 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.

[22] X. Chen *et al.*, "WBSMDA: Within and between score for MiRNA-disease association prediction," *Sci. Rep.*, vol. 6, Apr. 2016, Art. no. 21106.
[23] X. Chen, C. C. Yan, X. Zhang, Z.-H. You, Y.-A. Huang, and G.-Y. Yan, "HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction," *Oncotarget*, vol. 7, no. 40, pp. 65257–65269, 2016.
[24] X. Chen and G.-Y. Yan, "Semi-supervised learning for potential human microRNA-disease associations inference," *Sci. Rep.*, vol. 4, no. 1, 2014, Art. no. 5501.
[25] X. Chen *et al.*, "RBMMMDA: Predicting multiple types of disease-microRNA associations," *Sci. Rep.*, vol. 8, no. 5, 2015, Art. no. 13877.
[26] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2014, pp. 701–710.
[27] N. Zong, H. Kim, V. Ngo, and O. Harismendy, "Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations," *Bioinformatics*, vol. 33, no. 15, pp. 2337–2344, 2017.
[28] F. Cheng *et al.*, "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Comput. Biol.*, vol. 8, no. 5, p. e1002503, May 2012.
[29] Z. Yang *et al.*, "dbDEMC: A database of differentially expressed miRNAs in human cancers," *BMC Genomics*, vol. 11, p. S5, Sep. 2010.
[30] A. Ruepp *et al.*, "PhenomiR: A knowledgebase for microRNA expression in diseases and biological processes," *Genome Biol.*, vol. 11, no. 1, p. R6, 2010.
[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean. (2013). "Efficient estimation of word representations in vector space." [Online]. Available: https://arxiv.org/abs/1301.3781
[32] A. Mnih and G. E. Hinton, "A scalable hierarchical distributed language model," in *Advances in Neural Information Processing Systems*, vol. 21. Cambridge, MA, USA: MIT Press, 2009.
[33] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang, "Bipartite network projection and personal recommendation," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, no. 4, p. 046115, 2007.
[34] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *Bioinformatics*, vol. 27, pp. 3036–3043, Sep. 2011.
[35] H. Wu, S. Zhu, and Y.-Y. Mo, "Suppression of cell growth and invasion by *miR*-205 in breast cancer," *Cell Res.*, vol. 19, no. 4, pp. 439–448, 2009.
[36] L. J. Eichner *et al.*, "miR-378* mediates metabolic shift in breast cancer cells via the PGC-1β/ERRγ transcriptional pathway," *Cell Metabolism*, vol. 12, no. 4, pp. 352–361, 2010.
[37] P. Krishnan *et al.*, "Next generation sequencing profiling identifies miR-574-3p and miR-660-5p as potential novel prognostic markers for breast cancer," *BMC Genomics*, vol. 16, no. 1, p. 735, 2015.
[38] S. E. Lee *et al.*, "MicroRNA and gene expression analysis of melatonin-exposed human breast cancer cell lines indicating involvement of the anticancer effect," *J. Pineal Res.*, vol. 51, no. 3, pp. 345–352, 2011.
[39] Y. Yamamoto *et al.*, "An integrative genomic analysis revealed the relevance of microRNA and gene expression for drug-resistance in human breast cancer cells," *Mol. Cancer*, vol. 10, p. 135, Nov. 2011.
[40] P. Leidinger, A. Keller, and E. Meese, "MicroRNAs—Important molecules in lung cancer research," *Frontiers Genetics*, vol. 2, p. 104, Jan. 2011.
[41] Y. Lu *et al.*, "MicroRNA profiling and prediction of recurrence/relapse-free survival in stage I lung cancer," *Carcinogenesis*, vol. 33, no. 5, pp. 1046–1054, May 2012.
[42] L. Wang *et al.*, "Gene networks and microRNAs implicated in aggressive prostate cancer," *Cancer Res.*, vol. 69, no. 24, pp. 9490–9497, 2009.
[43] K. Erdmann *et al.*, "Elevated expression of prostate cancer-associated genes is linked to down-regulation of microRNAs," *BMC Cancer*, vol. 14, no. 1, p. 82, 2014.

GUANGHUI LI received the Ph.D. degree in computer science from Hunan University in 2015. He is currently an Assistant Professor with the School of Information Engineering, East China Jiaotong University. His research interests include data mining and computational biology.

JIAWEI LUO received the Ph.D. degree in computer science from Hunan University in 2008. She is currently a Professor with the College of Computer Science and Electronic Engineering, Hunan University. She has authored about 50 research papers in various international journals and proceedings of conferences. Her research interests include graph theory, data mining, computational biology, and bioinformatics.

QIU XIAO received the master's degree in computer science from Hunan University in 2013, where he is currently pursuing the Ph.D. degree with the College of Computer Science and Electronic Engineering, Hunan University. His research interests include data mining and computational biology.

CHENG LIANG received the Ph.D. degree in computer science from Hunan University in 2015. She was with the Donnelly Centre, University of Toronto, from 2012 to 2014, as a joint Ph.D. Student. She is currently an Assistant Professor with the College of Information Science and Electronic Engineering, Shandong Normal University. Her research interests include data mining and computational biology.

PINGJIAN DING (S'16) received the master's degree in computer science from Hunan University in 2015, where he is currently pursuing the Ph.D. degree with the College of Computer Science and Electronic Engineering, Hunan University. His research interests include data mining and computational biology.

BUWEN CAO received the master's degree in computer science from Hunan Normal University in 2007, and the Ph.D. degree from the College of Computer Science and Electronic Engineering, Hunan University, in 2016. He is currently with Hunan City University. His research interests include data mining and computational biology.