This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3168008, IEEE Journal of Biomedical and Health Informatics

Multi-relation graph embedding for predicting miRNA-target gene interactions by integrating gene sequence information

Jiawei Luo*, Wenjue Ouyang, Cong Shen, and Jie Cai

***Abstract*— Accumulated studies have found that miRNAs are in charge of many complex diseases such as cancers by modulating gene expression. Predicting miRNA-target interactions is beneficial for uncovering the crucial roles of miRNAs in regulating target genes and the progression of diseases. The emergence of large-scale genomic and biological data as well as the recent development in heterogeneous networks provides new opportunities for miRNA target identification. Compared with conventional methods, computational methods become a decent solution for high efficiency. Thus, designing a method that could excavate valid information from the heterogeneous network and gene sequences is in great demand for improving the prediction accuracy. In this study, we proposed a graph-based model named MRMTI for the prediction of miRNA-target interactions. MRMTI utilized the multi-relation graph convolution module and the Bi-LSTM module to incorporate both network topology and sequential information. The learned embeddings of miRNAs and genes were then used to calculate the prediction scores of miRNA-target pairs. Comparisons with other state-of-the-art graph embedding methods and existing bioinformatic tools illustrated the superiority of MRMTI under multiple criteria metrics. Three variants of MRMTI implied the positive effect of multi-relation. The experimental results of case studies further demonstrated the prominent ability of MRMTI in predicting novel associations.**

***Index Terms*— Heterogeneous information network, graph embedding, graph convolutional network, miRNA-target gene interactions**

I. Introduction

MICRORNAS (miRNAs) are small non-coding RNAs which play important roles in various biological processes, such as cell cycle control, cell growth, and cell differentiation[1]. They are first expressed as precursor RNAs, and then further processed into mature miRNAs. Mature miRNAs modulate gene expression post-transcriptionally by binding to 3' untranslated regions (3'UTRs) of target genes[2]. Abnormal miRNA expression can lead to dysfunctions of target genes, which in turn causes many complex diseases such as cancer. In the meantime, studies have shown that over one third of human genes appear to be conserved miRNA targets[3]. Therefore, identifying miRNA target genes is of great significance for revealing the regulatory mechanisms of miRNAs and their roles in the development of complex diseases.

Conventional biological methods for validating miRNA target interactions (MTIs) are based on molecular experiments or genome-wide screening, including western blot, quantitative real-time PCR (qPCR), and microarray experiments[4][5][6]. However, as the number of newly discovered miRNAs continues to increase, such experimental methods may be time-consuming, costly, and somewhat inefficient. In consequence, computational approaches are in great demand for facilitating miRNA target prediction.

Early computational researches are mainly based on biological features and principles. For instance, miRanda[7] filtered target genes on the basis of sequence complementarity, free energy calculation, and evolutionary conservation. TargetScan[8] combined cross-species conservation and thermodynamics-based modeling of RNA:RNA duplex interactions to predict targets of vertebrate miRNAs. As a refined version of TargetScan, TargetScanS[3] predicted targets with a conserved 6-nt seed match flanked by either an m8 match or a t1A anchor. Both miRanda and TargetScan were designed to predict targets containing multiple miRNA-recognition elements (MREs). In contrast, DIANA-microT[9] took into account the identifications of targets containing single MREs for human and mouse miRNAs. Although these sequence-based methods do provide a set of target candidates, they are likely to suffer from the high false positive rate.

With the accumulation of data, the construction of related databases including miRbase[10], HumanNet[11], and miRTarBase[12] provides relatively reliable data sources and makes it possible to develop machine learning methods for miRNA target prediction. Traditional machine learning methods including support vector machine[13][14], naïve bayes[15], and ensemble learning[16] have been frequently employed to make predictions at an early stage. MiTarget2[13] used features extracted from a public high-quality microarray dataset, and then leveraged the SVM classifier for prediction. TargetMiner[14] adopted the appropriate generation methods to obtain negative samples and identified miRNA targets based on SVM. Feature extraction was also required

This work was supported in part by the National Natural Science Foundation of China under Grant 61873089 and 62032007, and in part by the Hunan Provincial Innovation Foundation for Postgraduate under Grant CX20200436. *(Corresponding author: Jiawei Luo.)*

Jiawei Luo, Wenjue Ouyang, Cong Shen, and Jie Cai are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410000, China (e-mail: luojiawei@hnu.edu.cn; oywj@hnu.edu.cn; cshen@hnu.edu.cn; caijie@hnu.edu.cn).

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University Of South China. Downloaded on April 22,2022 at 15:55:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3168008, IEEE Journal of Biomedical and Health Informatics

for NBmiRTar[15], however, it utilized naïve bayes as the classification model rather than the SVM classifier. The aforementioned methods depend, to a great extent, on the artificially well-designed features, and usually show unsatisfactory results with respect to the false positive rate. Ensemble learning methods combine the outcomes of several prediction models to surpass the performance of each component model. Based on the idea of ensemble learning, SMILE[16] was designed by integrating six prediction tools and was demonstrated to achieve better generalization performance. Yet the major challenge of manual feature engineering still exists.

Due to the strength of representation learning, it has also been applied in a wide range of bioinformatic tasks. For example, NIMCGCN[17] first learned latent feature representations through graph convolutional networks[18] and then fed them into a matrix completion model to obtain association scores for miRNA-disease pairs. Another network-based model IDDkin[19] integrated graph convolution networks, graph attention networks, and adaptive weighting methods to effectively learn latent representations on the graph and subsequently enhanced the prediction of kinase inhibitors. Still, there are relatively few methods of miRNA-target identification that resort to representation learning. For instance, IMTRBM[20] built a weighted miRNA-target interaction network and then employed the restricted Boltzmann machine for extracting features and making predictions. Nevertheless, the sequence information was not involved in IMTRBM. SG-LSTM[21] generated both sequential and geometrical embeddings for miRNAs and genes, then an LSTM model was leveraged for predicting candidate targets. Apart from the weakness mentioned above, the existing graph embedding-based methods neglect the influence of relation types. However, modeling both structural and relational data in the heterogeneous network have been demonstrated to be beneficial[22][23]. Thus, preserving relational data along with structural and sequential features simultaneously is expected to help promote prediction performance.

To tackle the above challenges, we propose a novel graph-based model named MRMTI, which considers both network structure and sequential information for the task of predicting miRNA-target gene interactions. We construct a heterogeneous network that incorporates miRNA similarities, gene similarities, and miRNA-target interactions. Then the network embeddings of miRNAs and genes are trained through a multi-relation graph convolution module. Next, we leverage a Bi-LSTM model to extract deeper sequential features for genes. Afterward, the network structural embeddings and sequential embeddings are integrated to compute the association scores. The proposed MRMTI is compared with seven state-of-the-art methods under multiple metrics on miRNA-target prediction. We design the variants for validating the effectiveness of multi-relation. Furthermore, we conduct case studies to demonstrate the capability of MRMTI in predicting novel associations.

II. Methodology

A. Overview

The proposed model consists of three main parts. Firstly, a heterogeneous information network (HIN) is constructed by integrating miRNA similarity network, gene similarity network, and miRNA-gene bipartite network (Figure 1A). Then, on the basis of the HIN, the multi-relation graph convolution is adopted to embeds all the nodes in the network, incorporating both neighborhood information and multi-relational information (Figure 1B). In the meantime, the real-valued embeddings of genes generated by word2vec are fed into the Bi-LSTM module for excavating deeper sequential feature representations (Figure 1C). In the last period, we make predictions by inner product using the learned embeddings, and the model is trained in an end-to-end manner.

B. Construct heterogeneous information network

In this section, we introduce how to integrate information from different sources and construct a miRNA-gene heterogeneous network, which is the foundation of our method.

1) MiRNA-miRNA similarity network: The miRNA sequence data in miRbase[10] is mainly derived from author submission and wet experiments, hence it can be a reliable data source for directly mining functional relationships of miRNAs. Therefore, miRNA sequences are used to calculate the similarity scores based on the Needleman-Wunsch algorithm[24], which performs pairwise global alignment on two miRNA sequences. The miRNA-miRNA Network Net_m is denoted by the matrix $M = [M_{ij}] \in R^{N_m \times N_m}$, where M_{ij} represents the association between miRNA m_i and m_j. N_m denotes the number of miRNAs. To reduce the negative impact caused by redundant data, the top η_1 neighbors with the highest similarity scores are reserved for each miRNA. Let NS_{ij} represent the similarity score between the miRNA m_i and the miRNA m_j. Specifically, for the miRNA m_i, an edge connecting with m_j is added when NS_{ij} ranks top η_1 among all the neighbors of m_i. Hence, M_{ij} can be obtained as follows:

$$M_{ij} = \begin{cases} 1, & \text{if rank } (NS_{ij}) \leq \eta_1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

2) Gene-gene functional similarity network: Based on the assumption that genes with similar functions are more likely to be regulated by similar miRNAs, it is reasonable to integrate functional similarities of genes to construct the gene network. HumanNet v2[11] has been shown to be fairly useful for the task of disease gene prediction[25][26]. In this study, we use it to construct the gene-gene functional similarity network.

The associated log-likelihood-score of the interaction between gene g_i and gene g_j is denoted by LLS_{ij}, which measures the probability of the interaction representing a true functional connection. Similar to the construction of the miRNA-miRNA network, we preserve the top η_2 neighbors for each gene. The gene-gene network Net_g is represented by the matrix $G = [G_{ij}] \in R^{N_g \times N_g}$, where N_g denotes the number of genes and G_{ij} is defined as:

$$G_{ij} = \begin{cases} 1, & \text{if rank } (LLS_{ij}) \leq \eta_2 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University Of South China. Downloaded on April 22,2022 at 15:55:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3168008, IEEE Journal of Biomedical and Health Informatics

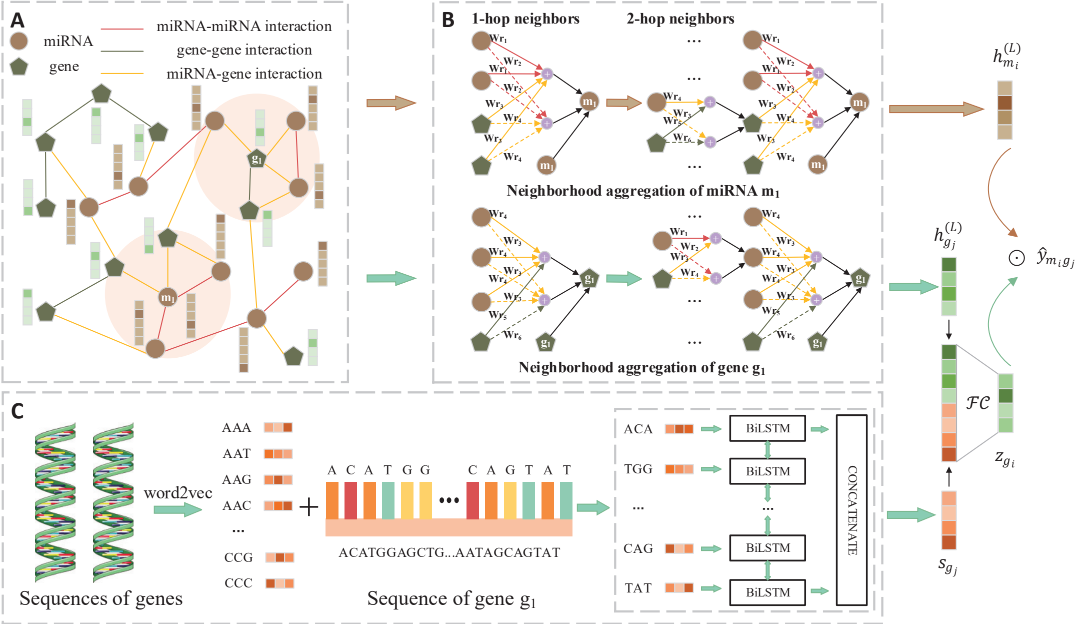

Fig. 1. The overall framework of the MRMTI model. (A) The construction of miRNA-gene heterogeneous information network. (B) The message passing procedure with multi-relation graph convolution for learning node embeddings. For each node, the embeddings are updated by incorporating neighborhood information under different relations. (C) The extraction of gene sequential feature utilizing word2vec and Bi-LSTM module. The final embeddings of miRNAs and genes are used for making predictions.

3) MiRNA-gene interaction network: We acquire the known human miRNA-gene associations from the experimentally validated miRNA-target association database miRTarBase[12]. The miRNA-gene interaction network Net_{mg} is represented by the matrix $A = [A_{ij}] \in R^{N_m \times N_g}$. If the miRNA m_i is associated with the gene g_j, the element $A_{ij} = 1$. Conversely, $A_{ij} = 0$ if the connection between the miRNA m_i and the gene g_j is unknown or unobserved.

After combining the miRNA-miRNA network Net_m, the gene-gene network Net_g and the miRNA-gene association network Net_{mg}, we finally construct a heterogeneous information network. In this study, a novel method named MRMTI is proposed based on the HIN for solving the problem of miRNA target identification.

C. Multi-relation graph convolution

A heterogeneous information network contains multiple types of relations between different nodes. In order to better integrate neighborhood information and capture the network structural feature, inspired by Decagon[23], we propose the multi-relation graph convolution network to obtain the representations of nodes. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$, where $\mathcal{V}$ denotes the set of vertices, $\mathcal{E}$ is the set of edges and $\mathcal{R}$ is the set of relations. As mentioned in the previous section, the heterogeneous information network includes miRNA-miRNA associations, gene-gene associations, and miRNA-gene interactions. Meanwhile, for a selected target node a, the process of passing information from node a to its neighboring nodes is considered as diffusion, while the process of passing information to node a from its neighbors is considered as fusion. These processes are different due to the different neighbor structure of node a and its neighboring nodes. Considering the process of fusion and diffusion, there are two types of relations between each pair of nodes. Thus, 6 kinds of edges exist in the graph, namely $\mathcal{R} = \{r_1, r_2, \ldots, r_6\}$. There is a d-dimensional initial embedding $e_i \in \mathbb{R}^d$ for each node $i \in \mathcal{V}$. We use one-hot vectors as the initial embeddings in this study. Noted that the dimension of the embeddings for miRNA nodes and gene nodes can be different.

The process of message passing is the core of graph convolution, including information fusion and diffusion. Since each node in the network has distinct local structure and neighborhood information, different message-passing schemas should be defined for different nodes and edge types. In other words, the updated node embeddings are supposed to be computed using relation-specific transformations. Specifically, by using learnable relational weights and the parameter sharing strategy, we are able to consider the local structure of nodes while taking into account the type and direction of edges, thereby obtaining more accurate representations.

Given node i, the local aggregation of neighborhood information under relation $r_1 \in \mathcal{R}$ in the l-th layer is described as follows:

$$U_N^{(l)} = \sum_{j \in N_i^{r_1}} \frac{1}{c_{ij}^{r_1}} W_{r_1}^{(l)} h_j^{(l)} \tag{3}$$

where $N_i^{r_1}$ represents the set of neighbors of node i under relation r_1 and node j is a neighboring node in $N_i^{r_1}$. $h_j^{(l)} \in \mathbb{R}^{d^{(l)}}$ represents the hidden state of node j in the l-th layer with $d^{(l)}$ denoting the dimension of embeddings, and $h_j^{(0)} = e_j$. $c_{ij}^{r_1}$ is a normalization constant which can either be learned or be set manually. Inspired by Decagon[23], the constant is set as $c_{ij}^{r_1} = \sqrt{|N_i^{r_1}||N_j^{r_1}|}$ in this study. It is particularly noteworthy

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University Of South China. Downloaded on April 22,2022 at 15:55:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3168008, IEEE Journal of Biomedical and Health Informatics

that $W_{r_1}^{(l)} \in \mathbb{R}^{d^{(l)} \times d^{(l+1)}}$ denotes a relation-specific weight matrix and is shared over the adjacent nodes connected by relation r_1 (i.e. nodes in $N_i^{r_1}$), thus the information of relation type is introduced into the computation. The self-loop of node i is defined as follows:

$$U_S^{(l)} = \frac{1}{c_i^{r_1}} h_i^{(l)} \tag{4}$$

where the normalization constant is defined as $c_i^{r_1} = |N_i^{r_1}|$. To sum up the input vectors of all neighboring nodes as well as the feature vector of node i itself, we compute $U_N^{(l)}$ and $U_S^{(l)}$ at the same time and integrate the results by adding them:

$$I_{r_1}^{(l)} = U_N^{(l)} + U_S^{(l)} \tag{5}$$

In consequence, we obtain the information computed utilizing the local graph connectivity structure under the given relation. The message passed from neighborhood under other relation types in the l-th layer such as $I_{r_2}^{(l)}$ are computed in the same way. Considering all edge types in the graph, the neural network propagation rule for updating the embeddings of node i can be written as follows:

$$h_i^{(l+1)} = \sigma\left(I_{r_1}^{(l)} + I_{r_2}^{(l)} + \cdots + I_{r_n}^{(l)}\right) \tag{6}$$

where $n = |\mathcal{R}|$ denotes the number of relation types and $\sigma(\cdot)$ is the ReLU activation function. The formulations directly demonstrate that for each node, different local neighborhood structures lead to different computational architectures. By stacking L graph convolution layers as defined, higher hop neighborhood information is incorporated into local neighbors. In this way, we can make better use of the network topology as well as the local structure of nodes.

D. Gene sequence feature extraction

Recently, the idea of Natural Language Processing (NLP) has been adapted and applied to many biological tasks, such as the identification of protein-protein interactions[27] and cancer prognostic genes[28]. Inspired by these ideas, in this study we first use the word2vec model to represent gene sequences. Specifically, we split the valid gene sequences retrieved from R package 'biomaRt' into k-mer segmentations which are regarded as "words", and then map them into real-valued embeddings through the pre-trained word2vec model. In our implementation, we set k to 3 and the size of embedding denoted as d_w is set to 64.

To fully exploit the latent feature of gene sequences, it is beneficial to use the bidirectional long short-term memory recurrent neural network (Bi-LSTM). Compared with the original LSTM, Bi-LSTM made an improvement by taking both previous and subsequent inputs into account. We utilize Bi-LSTM to capture "deep" sequential features, aiming for larger expressive capability. Formally, the sequential feature vector of gene g_i is updated as follows:

$$s_{g_i} = f(g_i) = \overrightarrow{\text{LSTM}}(w_j) \oplus \overleftarrow{\text{LSTM}}(w_j) \tag{7}$$

where $s_{g_i} \in \mathbb{R}^{d_f}$ denotes the output feature vector of gene g_i and d_f represents the dimensionality. We denote the j-th k-mer segment of a gene sequence as $w_j \in \mathbb{R}^{d_w}$, which is the output of the word2vec model, and j ranges from 1 to the number of k-mer segments existing in a gene sequence. $\overrightarrow{LSTM}(\cdot)$ and $\overleftarrow{LSTM}(\cdot)$ capture the underlying interactions in contexts from forward and backward directions. The symbol $\oplus$ represents the concatenation between the output of the forward and the backward LSTM cell. The obtained representations can serve as supplementary information for the task of miRNA target identification.

E. Information fusion and model prediction

To improve the accuracy of predicting miRNA-target interactions, we make an effort to extract and integrate effective information from different sources. Specifically, we use a learnable transformation matrix to project gene representations to the same embedding space as miRNA representations. As a result, the final embedding of gene g_i is formulated as:

$$z_{g_i} = W_p\left(\text{concat}\left(h_{g_i}^{(L)}, s_{g_i}\right)\right) \tag{8}$$

where W_p is the transformation matrix and $z_{g_i} \in \mathbb{R}^{d^{(L)}}$. $\text{concat}(\cdot)$ denotes the operation of concatenation.

MRMTI embeds both miRNAs and genes in a low dimensional latent space. The learned embedding matrices of miRNA and gene are represented as $H_m = \left[h_{m_i}^{(L)}\right] \in \mathbb{R}^{N_m \times d^{(L)}}$ and $H_g = [z_{g_i}] \in \mathbb{R}^{N_g \times d^{(L)}}$, respectively. Eventually, we adopt the inner product between the corresponding miRNA and gene embeddings to calculate the prediction score of the miRNA-target pair:

$$\hat{y}_{m_i g_j} = \sigma\left(h_{m_i}^{(L)} z_{g_j}^T\right) \tag{9}$$

where σ is the sigmoid function. $h_{m_i}^{(L)}$ is the embedding of miRNA m_i, which is the i-th row of the miRNA embedding matrix H_m. Likewise, z_{g_j} denotes the embedding of gene g_j that is the j-th row of the gene embedding matrix H_g.

We use the hinge loss for optimization, which is widely used in conventional binary classification tasks. Taking all relations into consideration, the loss function of MRMTI is designed as follows:

$$\mathcal{L} = \sum_{r \in \mathcal{R}} \sum_{i \in S^+ \cup j \in S^-} \text{ReLU}\left(\text{margin} - \hat{y}_i + \hat{y}_j\right) \tag{10}$$

where S^+ denotes the set of positive samples, and S^- denotes the set of negative samples that are randomly sampled with the same size as positive samples. The prediction scores of positive and negative sample are denoted as $\hat{y}_i$ and $\hat{y}_j$, respectively. The hyper-parameter margin is manually chosen in advance. The loss function encourages the model to score observed samples higher than the negative ones. The MRMTI model is then trained and optimized in an end-to-end manner. The data for MRMTI is available at https://github.com/ojinxj/MRMTI.

III. Results

A. Datasets

We downloaded the known human miRNA-target associations from miRTarBase[12], which contains 380,639 experimentally verified human miRNA-target association data,

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University Of South China. Downloaded on April 22,2022 at 15:55:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3168008, IEEE Journal of Biomedical and Health Informatics

TABLE I
STATISTICS OF THE DATASETS

Dataset	miRTarBase	HumanNet	miRBase	biomaRt
MicroRNA	2599	-	2656	-
Gene	15,064	17,929	-	13,416
Relation	380,639	525,537	-	-

involving 2599 mature miRNAs and 15,064 genes. The human mature miRNA sequences were acquired from miRBase database[10]. After removing the duplicate records, it contains 2656 mature miRNAs. The gene functional similarity data were derived from the HumanNet database[11], which contains 525,537 interactions and their log-likelihood score (LLS) among 17,929 genes. The 3'UTR gene sequences in FASTA format were retrieved from R package biomaRt[29].

A series of preprocessing operations were performed on the collected multi-source data. First, we calculated the average log-likelihood score (LLS) of all the records from the HumanNet and removed the entries with LLS below the average. As for gene sequences obtained from biomaRt, we removed the abnormal data and acquired 13,416 gene sequences. In view of the potential negative impact brought by network sparsity, we removed the nodes with degrees below 10 in the gene-gene network Net_g and consequently 7880 genes were retained, forming the list of genes for experimental use. Finally, after taking the intersection sets, our experimental dataset consists of 18,033 miRNA-miRNA associations, 127,772 gene-gene associations, and 211,111 miRNA-target interactions, involving 2546 miRNAs and 7880 genes. Statistical information about the dataset is listed in Table 1. The set of negative samples were randomly sampled with the same size as positive samples.

B. Baselines

We chose several state-of-the-art models as comparison methods for evaluating the performance on the dataset of MRMTI. These methods are roughly classified into two categories: novel computational bioinformatic methods and classic graph embedding methods.

In terms of existing computational methods in the field of bioinformatics, we compared MRMTI with SG-LSTM[21], IDDkin[19], and KATZ[30][31]. SG-LSTM is a deep learning framework for predicting miRNA-target gene interactions. It combines both geometric and sequential embeddings and utilizes LSTM as the classifier to get the prediction scores. We set the parameters as suggested in the original article. IDDkin is a network-based deep influence framework for predicting kinase inhibitors. The parameters L and p were chosen from {8,16,32,64,128} and {8,12,16,20,24}, respectively. We chose the parameter K from 6 to 14 with step 2. KATZ calculates the proximity of pair-wise nodes in the graph by integrating the information of different meta-paths. The parameter k is set to be 2, 3, and 4.

As for the classic graph embedding methods, we made comparisons with DeepWalk[32], LINE[33], GraRep[34], and SDNE[35]. DeepWalk is a graph embedding algorithm based on random-walk. The parameters, including walk length, walks per-vertex and the window size of skip-gram model were tuned meticulously for optimal performance. LINE makes an improvement in network embedding by taking into account both the first-order and second-order proximities. In accordance with the original paper, we concatenated the first-hop and second-hop representations as the final representation for better results. SDNE is a semi-supervised deep model for embedding graph vertices, which could capture the highly-nonlinear local-global network structure. The hyperparameters α and β were tuned in light of SDNE[35]. GraRep is a matrix factorization-based method for learning graph presentations with the advantage of incorporating global structural information. We chose maximum matrix k-step size K from 2 to 7.

C. Experimental settings

1) Parameter settings: MRMTI is an end-to-end model where all the trainable parameters in the model were trained jointly using Adam optimizer with a learning rate of 0.001. In practice, we randomly divided the data set into training, validation and test sets. Concretely, 80% of the known edges were used to train the model, and 10% were used to choose model parameters. The rest of edges were taken as the test set to evaluate the model performance. Note that in the experiment, we repeated for 10 times to take the average result for avoiding uncertainty. We implemented the MRMTI model in Tensorflow[36] and set the number of hidden layers L to 2. The output dimension of each layer was selected from {16,32,64,128,256} and {8,16,32,64,128}, respectively. The margin parameter of hinge loss was set to margin=0.3. The dropout ratio was set to 0.1. The parameters η_1 and η_2 were both selected from {5, 10, 20, 30, 40}. The output dimension of Bi-LSTM was selected from {8, 16, 32, 64}. The parameter k and d_w of word2vec model were chosen from {2, 3, 4} and {16, 32, 64}, respectively. We leveraged Xavier[37] as the initialization method for model parameters. Mini-batch was used during the training process and the batch size was fixed to 512.

2) Evaluation criteria: In this paper, AUC, AUPR, Precision, Recall, F1-score, and Balanced Accuracy were used for the evaluation of the performance of miRNA-target identification results. AUC is the area under the receiver operating characteristics (ROC) curve which is established by plotting the true positive rate (TPR) against the false positive rate (FPR) under changing threshold settings. TPR and FPR are calculated as follows: TPR=TP/(TP+FN), FPR=FP/(TN+FP), where TP and TN are used to represent the numbers of correctly identified positive and negative examples, FP and FN are the numbers of misidentified positive and negative samples. Similarly, AUPR is the area under the Precision-Recall (PR) curve, which is plotted using Precision as the vertical axis and Recall as the horizontal axis with various thresholds. The equations for computing Precision and Recall are as follows: Precision=TP/(TP+FP), Recall=TP/(TP+FN). F1-score is an evaluation metric that comprehensively considers the Precision and Recall: F1-score=2×Precision×Recall/(Precision+Recall). The evaluation metric Balanced Accuracy is especially useful

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University Of South China. Downloaded on April 22,2022 at 15:55:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3168008, IEEE Journal of Biomedical and Health Informatics

when dealing with imbalanced datasets, and it is calculated as: Balanced Accuracy=$\frac{1}{2}\times(TP/(TP+FN)+TN/(TN+FP))$.

D. Comparisons with baselines

Figure 2 shows the ROC curves and the PR curves of MRMTI and baseline models. It can be observed that MRMTI outperformed the state-of-the-art methods with AUC=0.9183 and AUPR=0.9204. Among the four popular graph embedding models, GraRep achieved the best performance with AUC=0.8468 and AUPR=0.8316, while the AUCs and AUPRs of DeepWalk, LINE, and SDNE were 0.8179, 0.8290, 0.8388 and 0.8106, 0.8058, 0.8221, respectively. In terms of computational bioinformatic methods, KATZ (AUC=0.8886, AUPR=0.8901) and IDDkin (AUC=0.8630, AUPR=0.8606) shown better results than SG-LSTM (AUC=0.8581, AUPR=0.8461). Since KATZ and IDDkin managed to capture the structural characteristics of the heterogeneous network, the results imply the importance of modeling network heterogeneity. Besides, SG-LSTM outperformed all the classic graph embedding methods, indicating that the combination of geometrical and sequential features had a positive effect on model performance. Our proposed MRMTI model achieved the best performance as it can not only take full advantage of the constructed heterogeneous network, but also considers both structural and sequential features. In order to further validate the performance of MRMTI, the AUCs and AUPRs of MRMTI and other methods with different runs were compared using paired t-test. As shown in the Table S1, the p-values were less than 0.05, suggesting that the differences between AUCs and AUPRCs were statistically significant.

Apart from the overall evaluation metrics (i.e., AUC and AUPR), we further analyzed the performance of MRMTI using Precision, Recall, F1-score and Balanced Accuracy with the top-ranked 10% and 20% predictions. As shown in Table 2, MRMTI outperformed most of the baseline models and the overall trend of the performances for different methods is similar as analyzed before. The reasons for the superiority of MRMTI are threefold. First of all, we integrated miRNA and gene similarities along with the miRNA-target interactions to build a heterogeneous network that was expected to be more informative. Secondly, MRMTI utilized the graph convolutional network to efficiently embed the nodes across the network, with both multi-relational information and network structure incorporated. In the meantime, MRMTI additionally excavated deep sequential features with Bi-LSTM under the inspiration of natural language processing.

E. Parameter analysis

In this section, we investigate the impacts of the embedding dimension $d^{(i)}$ of the i-th layer. AUC scores of the MRMTI model using only one graph convolution layer with varying output dimensions are shown in Figure 3(A). As shown, the experimental performance improves with the increase of dimension. However, as the dimension continues to increase from 128 to 256, the performance increases slightly and becomes relatively stable, implying that 128 is enough for obtaining information.

Accumulated researches indicated that multi-hop neighborhood information is beneficial for learning more accurate node representations, and also found similar trends that 2-hop higher-order graph structure may help achieve the best performance[18][38]. For such a reason, we added another convolution layer to acquire information from second-order neighbors. To further study the influence of the embedding size, we fixed the dimension of the first layer to 128 and then tuned the output dimension of the second layer. Results are presented in Figure 3(B). As the dimension ranging from 8 to 128, the performance first increases and then experiences a marginal decrease when the dimension reaches 128, which may due to the over-fitting or the introduction of noisy data. As a result, we set the dimension of the graph convolution layer to 128-64, aiming for the best performance. Other parameters including the output dimension of Bi-LSTM and the threshold of similarity network η were investigated, and the results are shown in Figure S1-S2 and Table S2 in the Supplementary Material.

F. The effect of multi-relation

In this section, to validate the effectiveness of multiple relations, we conducted an experiment to compare MRMTI with its three variants, each of which deliberately neglected one kind of inverse edge. The results of miRNA-target identification are reported in Table 3. MRMTI-MTI-inv means that the inverse edge between a miRNA and its target gene was ignored. Likewise, MRMTI-MM-inv and MRMTI-GG-inv respectively denote the omission of the inverse relation among miRNAs and genes.

As shown in Table 3, the original MRMTI model achieved the highest AUC and AUPR scores, which were 0.9183 and 0.9204, respectively. AUC scores of the three variants were 0.8979, 0.9131, and 0.9053, while the AUPR scores were 0.8983, 0.9086, and 0.8992. In contrast with the MRMTI model, there was a decline in both AUC score and AUPR score of the three variants, and the overall performance of MRMTI-MTI-inv dropped the most with its AUC score decreased by 2.2%.

A possible reason for the results may be that the local structure of two nodes connected by an edge is nonidentical, thus, considering the processes of data fusion and diffusion, the information conveyed through the original edge and inverse edge are dissimilar as well. Specifically, the local structure of miRNA nodes and gene nodes in the bipartite network is much more different than in homogenous networks, which may subsequently account for the bigger degradation in the performance of MRMTI-MTI-inv. Additionally, the performance of MRMTI-MM-inv was slightly better than MRMTI-GG-inv, which may be owing to the fact that the number of gene nodes (7880) in the heterogeneous network is much larger than that of miRNA nodes (2546) and the network structure of gene nodes has a greater effect accordingly. MRMTI comprehensively considered both data fusion and diffusion, so that it achieved the best performance.

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University Of South China. Downloaded on April 22,2022 at 15:55:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3168008, IEEE Journal of Biomedical and Health Informatics

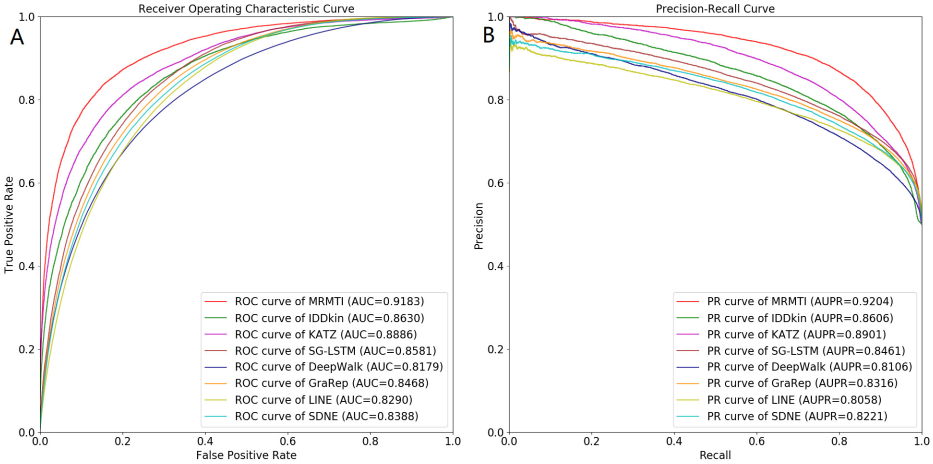

Fig. 2. Comparisons with baseline models by ROC curves (A) and PR curves (B).

TABLE II

COMPARISON WITH BASELINE METHODS ON OUR DATASET. PRE@0.1 DENOTES PRECISION AT 10%.

Methods	PRE@0.1	PRE@0.2	REC@0.1	REC@0.2	F1@0.1	F1@0.2	BA@0.1	BA@0.2
DeepWalk	0.9156	0.8751	0.1829	0.3500	0.3049	0.5000	0.5828	0.6494
GraRep	0.9195	0.8877	0.1839	0.3551	0.3065	0.5072	0.5831	0.6541
LINE	0.8908	0.8583	0.1782	0.3433	0.2969	0.4904	0.5770	0.6423
SDNE	0.9109	0.8807	0.1822	0.3523	0.3036	0.5033	0.5812	0.6497
SG-LSTM	0.9317	0.8969	0.1863	0.3587	0.3105	0.5125	0.5863	0.6587
KATZ	0.9869	0.9510	0.1973	0.3805	0.3288	0.5435	0.5973	0.6827
IDDkin	0.9619	0.9322	0.1924	0.3689	0.3206	0.5269	0.5924	0.6689
MRMTI	**0.9881**	**0.9722**	**0.1976**	**0.3878**	**0.3293**	**0.5544**	**0.5976**	**0.6889**

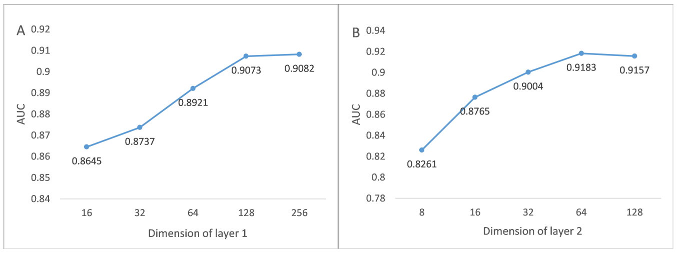

Fig. 3. Analysis of dimension. (A)The AUC scores of MRMTI with one convolution layer under different dimension. (B)The AUC scores under varying dimension of the second layer with the dimension of the first layer fixed.

TABLE III

THE RESULTS OF MRMTI AND ITS VARIANTS

Methods	AUC	Improvement	AUPR	Improvement
MRMTI	0.9183	—	0.9204	—
MRMTI-MTI-inv	0.8979	-2.2%	0.8983	-2.4%
MRMTI-MM-inv	0.9131	-0.6%	0.9086	-1.3%
MRMTI-GG-inv	0.9053	-1.4%	0.8992	-2.3%

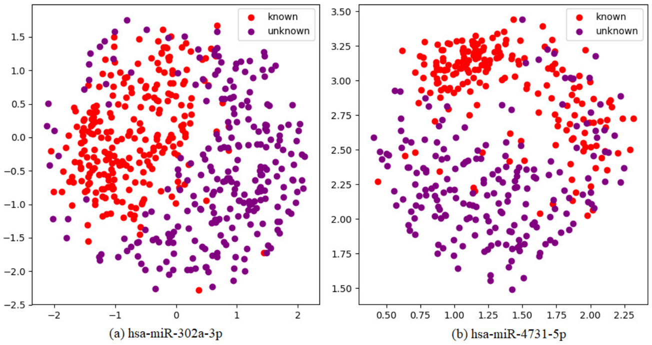

Fig. 4. 2D visualization of gene representations on miRTarBase.

G. Visualization

The functionality of the representations could be verified through visualization. Therefore, we first chose two miRNAs as cases, namely hsa-miR-302a-3p and hsa-miR-4731-5p. Then genes that are known to have associations with the miRNA are included in the experiment, while the same proportion of genes are selected from the unknown sets. Finally, we utilized the t-SNE algorithm (t-Distributed Stochastic Neighbor Embedding) to reduce the gene embeddings to 2 dimensions, and the visualization results obtained are shown in Figure 4. It can be observed that there is a distinct difference between the known and unknown associations. Thus, we consider the representations learned by MRMTI model to be effective.

H. Case studies

Case studies were carried out so as to further test the ability of our proposed MRMTI model in predicting unknown interactions. For predicting either potential targets related to a certain miRNA or miRNA candidates associated with a specific gene, we first trained the model with all known associations between miRNAs and genes to obtain the prediction results. Next, we sorted the results in descending order by the prediction scores

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University Of South China. Downloaded on April 22,2022 at 15:55:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3168008, IEEE Journal of Biomedical and Health Informatics

TABLE IV
THE PREDICTION RESULTS FOR HSA-MIR-155-5P AND HSA-MIR-335-5P

MiRNA	Top 10 ranked predictions		
	Rank	Target Gene	Evidence
hsa-miR-155-5p	1	IGF1R	PMID: 28613101
	2	DICER1	PMID: 31275058
	3	VPS33A	-
	4	LDLR	PMID: 26867493
	5	BTF3L4	-
	6	IGF1	PMID: 32256755
	7	JAG1	-
	8	XIAP	PMID: 26779627
	9	EIF5	-
	10	FKBP14	-
hsa-miR-335-5p	1	SMAD7	PMID: 31248450
	2	CHRDL1	-
	3	LMNB2	-
	4	HHIP	-
	5	H3F3B	PMID: 27347075
	6	DUSP4	-
	7	PTEN	PMID: 29307835
	8	BACH1	-
	9	STAT5A	PMID: 32791489
	10	IL6R	-

TABLE V
THE PREDICTION RESULTS FOR CDKN1A AND SMAD4

Gene	Top 10 ranked predictions		
	Rank	MiRNA	Evidence
CDKN1A	1	hsa-miR-608	-
	2	hsa-miR-363-5p	PMID: 30784290
	3	hsa-miR-29b-3p	-
	4	hsa-miR-423-5p	PMID: 32264887
	5	hsa-miR-6766-5p	-
	6	hsa-miR-193a-5p	PMID: 33352502
	7	hsa-miR-92a-3p	PMID: 26482648
	8	hsa-miR-103a-3p	-
	9	hsa-miR-299-3p	PMID: 28600498
	10	hsa-miR-125b-5p	-
SMAD4	1	hsa-miR-21-5p	PMID: 29943845
	2	hsa-miR-192-5p	PMID: 31293639
	3	hsa-miR-135a-5p	-
	4	hsa-miR-3662	-
	5	hsa-miR-135b-5p	PMID: 27422404
	6	hsa-miR-5692a	-
	7	hsa-miR-340-5p	PMID: 27229858
	8	hsa-miR-8063	-
	9	hsa-miR-9-5p	-
	10	hsa-miR-590-3p	PMID: 26498065

and removed all the existing entries in the original dataset. Finally, the top 10 candidates were extracted and validated manually in PubMed, which comprises copious biomedical literature.

Hsa-miR-155-5p is a well-known oncogenic miRNA that has been found to be linked to many complex diseases, such as breast cancer[39] and colorectal cancer[40]. Meanwhile, hsa-mir-335-5p is another important miRNA that is involved in numerous biological processes as a regulator[41][42]. Consequently, exploring potential targets of these two miRNAs could be meaningful. In like manner, CDKN1A and SMAD4 were chosen for their crucial functions in various cellular processes[43][44].

The prediction and validation results of the top 10 gene candidates and miRNA candidates are presented in Table 4 and Table 5, respectively. Noted that the known miRNA-target interactions we used for building the model were acquired from miRTarBase (version 8.0), which is a comparatively comprehensive dataset and was updated in 2020, thus the number of newly published experimental literature is limited. Under this condition, 5 out of the top 10 hsa-miR-155-5p-targeted genes and 4 out of the top 10 hsa-miR-335-5p-targeted genes were verified by literature. For instance, it has been proved by experimental means such as RT-qPCR that miR-335 activates TGFβ signaling pathway through targeting SMAD7[41]. Furthermore, as shown in Table 5, among the top 10 predicted miRNAs associated with CDKN1A or SMAD4, one-half of the associations were confirmed with supporting evidence provided by published papers.

In summary, the case studies indicate that the MRMTI model has the capability of predicting novel miRNA-target interactions.

IV. CONCLUSION

Identifying miRNA target genes is of great significance for improving our understanding in the regulatory roles of miRNAs. In this work, a unified graph-based model MRMTI was proposed for miRNA-target prediction. We constructed a heterogeneous information network (HIN) and then took advantage of the graph convolutional network along with relational data to obtain structural representations for miRNAs and genes. In the meantime, we made better use of gene sequences through word2vec and Bi-LSTM. Experimental results demonstrate that MRMTI could achieve superior performance in contrast with other state-of-the-art baseline models in most of the evaluation criteria. By comparing with three variants, we further validated the effect of multiple relations on the performance of MRMTI. Moreover, four important human miRNAs and genes were used as case studies to evaluate the ability of MRMTI in identifying novel miRNA-target interactions. After the integration of multi-source information, the graph convolutional network manages to efficiently extract latent features from the constructed HIN, which significantly contributes to the promotion of performance. The comprehensive consideration of relational data, network topology, and sequential information could be a reason for the excellent performance of MRMTI as well.

The superiority of MRMTI implies the potential of graph embedding methods and provides a new perspective for miRNA-target identification. Although MRMTI has shown outstanding prediction performance, it has certain limitations. As we used a concatenation operation for feature integration, other methods of feature integration such as the Hadamard product and attention mechanism are worth investigation. Meanwhile, other valuable biological information such as miRNA families and the Gene Ontology (GO) could be gathered for the construction of HIN. In future work, we would like to predict the interactions between miRNAs and genes for specific cancer lines, which could help understand the mechanism of miRNAs and target genes in the development of a certain disease. Moreover, more consideration will be given to the biological characteristics and comparisons will be made with traditional miRNA-target identification methods.

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University Of South China. Downloaded on April 22,2022 at 15:55:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3168008, IEEE Journal of Biomedical and Health Informatics

References

[1] R. A. Shivdasani, "Micrornas: regulators of gene expression and cell differentiation," *Blood*, vol. 108, no. 12, pp. 3646–3653, 2006.

[2] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, "Human microrna targets," *PLOS Biology*, vol. 2, no. 11, p. e363, 2004.

[3] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[4] X. Su, D. Chakravarti, M. S. Cho, L. Liu, Y. J. Gi, Y.-L. Lin, M. L. Leung, A. El-Naggar, C. J. Creighton, M. B. Suraokar *et al.*, "Tap63 suppresses metastasis through coordinate regulation of dicer and mirnas," *Nature*, vol. 467, no. 7318, pp. 986–990, 2010.

[5] F. M. Lang, A. Hossain, J. Gumin, E. N. Momin, Y. Shimizu, D. Ledbetter, T. Shahar, S. Yamashita, B. Parker Kerrigan, J. Fueyo *et al.*, "Mesenchymal stem cells as natural biofactories for exosomes carrying mir-124a in the treatment of gliomas," *Neuro-oncology*, vol. 20, no. 3, pp. 380–390, 2018.

[6] J. Yan, S.-B. Ng, J. L.-S. Tay, B. Lin, T. L. Koh, J. Tan, V. Selvarajan, S.-C. Liu, C. Bi, S. Wang *et al.*, "Ezh2 overexpression in natural killer/t-cell lymphoma confers growth advantage independently of histone methyltransferase activity," *Blood, The Journal of the American Society of Hematology*, vol. 121, no. 22, pp. 4512–4520, 2013.

[7] A. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. Marks, "Microrna targets in drosophila," *Genome biology*, vol. 4, no. 11, pp. 1–27, 2003.

[8] B. P. Lewis, I.-h. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge, "Prediction of mammalian microrna targets," *Cell*, vol. 115, no. 7, pp. 787–798, 2003.

[9] M. Kiriakidou, P. T. Nelson, A. Kouranov, P. Fitziev, C. Bouyioukos, Z. Mourelatos, and A. Hatzigeorgiou, "A combined computational-experimental approach predicts human microrna targets," *Genes & development*, vol. 18, no. 10, pp. 1165–1178, 2004.

[10] S. Griffiths-Jones, H. K. Saini, S. Van Dongen, and A. J. Enright, "mirbase: tools for microrna genomics," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D154–D158, 2007.

[11] S. Hwang, C. Y. Kim, S. Yang, E. Kim, T. Hart, E. M. Marcotte, and I. Lee, "Humannet v2: human gene networks for disease research," *Nucleic acids research*, vol. 47, no. D1, pp. D573–D580, 2019.

[12] H.-Y. Huang, Y.-C.-D. Lin, J. Li, K.-Y. Huang, S. Shrestha, H.-C. Hong, Y. Tang, Y.-G. Chen, C.-N. Jin, Y. Yu *et al.*, "mirtarbase 2020: updates to the experimentally validated microrna–target interaction database," *Nucleic acids research*, vol. 48, no. D1, pp. D148–D154, 2020.

[13] X. Wang and I. M. El Naqa, "Prediction of both conserved and nonconserved microrna targets in animals," *Bioinformatics*, vol. 24, no. 3, pp. 325–332, 2008.

[14] S. Bandyopadhyay and R. Mitra, "Targetminer: microrna target prediction with systematic identification of tissue-specific negative examples," *Bioinformatics*, vol. 25, no. 20, pp. 2625–2631, 2009.

[15] M. Yousef, S. Jung, A. V. Kossenkov, L. C. Showe, and M. K. Showe, "Naïve bayes for microrna target predictions—machine learning for microrna targets," *Bioinformatics*, vol. 23, no. 22, pp. 2987–2992, 2007.

[16] S. Yu, J. Kim, H. Min, and S. Yoon, "Ensemble learning can significantly improve human microrna target prediction," *Methods*, vol. 69, no. 3, pp. 220–229, 2014.

[17] J. Li, S. Zhang, T. Liu, C. Ning, Z. Zhang, and W. Zhou, "Neural inductive matrix completion with graph convolutional networks for mirna-disease association prediction," *Bioinformatics*, vol. 36, no. 8, pp. 2538–2546, 2020.

[18] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[19] C. Shen, J. Luo, W. Ouyang, P. Ding, and X. Chen, "Iddkin: network-based influence deep diffusion model for enhancing prediction of kinase inhibitors," *Bioinformatics*, vol. 36, no. 22-23, pp. 5481–5491, 2020.

[20] Y. Liu, J. Luo, and P. Ding, "Inferring microrna targets based on restricted boltzmann machines," *IEEE journal of biomedical and health informatics*, vol. 23, no. 1, pp. 427–436, 2018.

[21] W. Xie, J. Luo, C. Pan, and Y. Liu, "Sg-lstm-frame: a computational frame using sequence and geometrical information via lstm to predict mirna–gene associations," *Briefings in bioinformatics*, vol. 22, no. 2, pp. 2032–2042, 2021.

[22] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European semantic web conference*. Springer, 2018, pp. 593–607.

[23] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, 2018.

[24] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

[25] N. Natarajan and I. S. Dhillon, "Inductive matrix completion for predicting gene–disease associations," *Bioinformatics*, vol. 30, no. 12, pp. i60–i68, 2014.

[26] V. D. Tran, A. Sperduti, R. Backofen, and F. Costa, "Heterogeneous networks integration for disease–gene prioritization with node kernels," *Bioinformatics*, vol. 36, no. 9, pp. 2649–2656, 2020.

[27] Y. Wang, S. Zhang, Y. Zhang, J. Wang, and H. Lin, "Extracting protein-protein interactions affected by mutations via auxiliary task and domain pre-trained model," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 495–498.

[28] J. Choi, I. Oh, S. Seo, and J. Ahn, "G2vec: Distributed gene representations for identification of cancer prognostic genes," *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[29] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, "Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart," *Nature protocols*, vol. 4, no. 8, p. 1184, 2009.

[30] X. Chen, Y.-A. Huang, Z.-H. You, G.-Y. Yan, and X.-S. Wang, "A novel approach based on katz measure to predict associations of human microbiota with non-infectious diseases," *Bioinformatics*, vol. 33, no. 5, pp. 733–739, 2017.

[31] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[32] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 701–710.

[33] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 1067–1077.

[34] S. Cao, W. Lu, and Q. Xu, "Grarep: Learning graph representations with global structural information," in *Proceedings of the 24th ACM international on conference on information and knowledge management*, 2015, pp. 891–900.

[35] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 1225–1234.

[36] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[37] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[38] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 974–983.

[39] B. Pasculli, R. Barbano, A. Fontana, T. Biagini, M. P. Di Viesti, M. Rendina, V. M. Valori, M. Morritti, S. Bravaccini, S. Ravaioli *et al.*, "Hsa-mir-155-5p up-regulation in breast cancer and its relevance for treatment with poly [adp-ribose] polymerase 1 (parp-1) inhibitors," *Frontiers in Oncology*, vol. 10, p. 1415, 2020.

[40] X.-F. Zhang, R. Tu, K. Li, P. Ye, and X. Cui, "Tumor suppressor ptprj is a target of mir-155 in colorectal cancer," *Journal of cellular biochemistry*, vol. 118, no. 10, pp. 3391–3400, 2017.

[41] M. Kay, B. M. Soltani, F. H. Aghdaei, H. Ansari, and H. Baharvand, "Hsa-mir-335 regulates cardiac mesoderm and progenitor cell differentiation," *Stem cell research & therapy*, vol. 10, no. 1, pp. 1–13, 2019.

[42] L. Yao, M. Li, J. Hu, W. Wang, and M. Gao, "Mirna-335-5p negatively regulates granulosa cell proliferation via sgk3 in pcos," *Reproduction*, vol. 156, no. 5, pp. 439–449, 2018.

[43] N.-N. Kreis, F. Louwen, and J. Yuan, "The multifaceted p21 (cip1/waf1/cdkn1a) in cell differentiation, migration and cancer therapy," *Cancers*, vol. 11, no. 9, p. 1220, 2019.

[44] N. I. Fleming, R. N. Jorissen, D. Mouradov, M. Christie, A. Sakthianandeswaren, M. Palmieri, F. Day, S. Li, C. Tsui, L. Lipton *et al.*, "Smad2, smad3 and smad4 mutations in colorectal cancer," *Cancer research*, vol. 73, no. 2, pp. 725–735, 2013.

2168-2194 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University Of South China. Downloaded on April 22,2022 at 15:55:11 UTC from IEEE Xplore. Restrictions apply.