Interdisciplinary Sciences: Computational Life Sciences (2021) 13:547–558
https://doi.org/10.1007/s12539-021-00454-3

ORIGINAL RESEARCH ARTICLE

Metapath-Based Deep Convolutional Neural Network for Predicting miRNA-Target Association on Heterogeneous Network

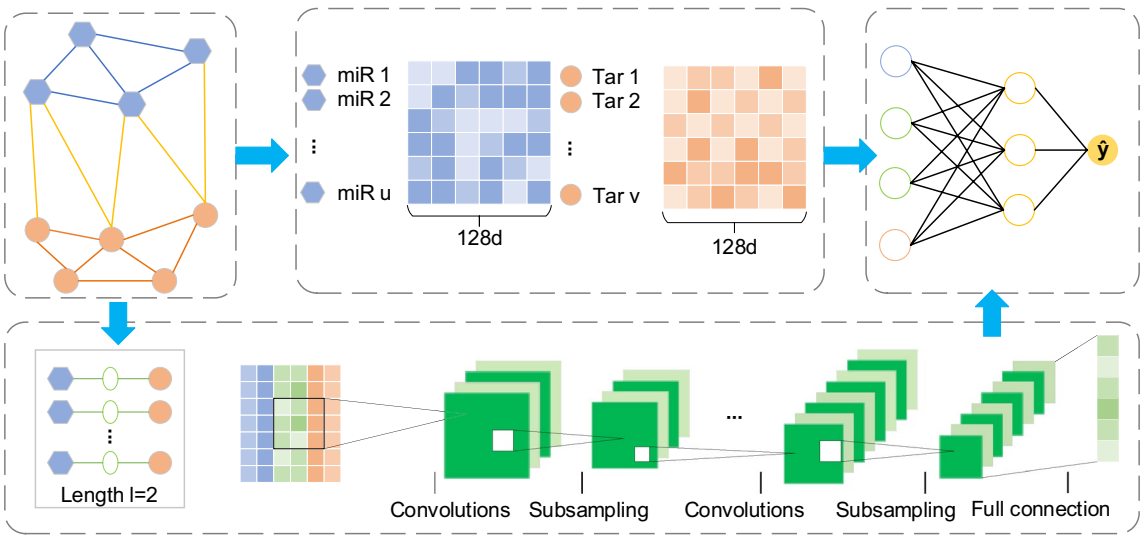
Jiawei Luo[1] · **Yaoting Bao**[1] · **Xiangtao Chen**[1] · **Cong Shen**[1]

Received: 3 February 2021 / Revised: 17 June 2021 / Accepted: 17 June 2021 / Published online: 25 June 2021

Abstract

Predicting the interactions between microRNAs (miRNAs) and target genes is of great significance for understanding the regulatory mechanism of miRNA and treating complex diseases. The emergence of large-scale, heterogeneous biological networks has offered unprecedented opportunities for revealing miRNA-associated target genes. However, there are still some limitations about automatically learn the feature information of the network in the existing methods. Since network representation learning can self-adaptively capture structure information of the network, we propose a framework based on heterogeneous network representation, MDCNN (Metapath-Based Deep Convolutional Neural Network), to predict the associations between miRNAs and target genes. MDCNN samples the paths between the node pairs in the form of meta-path based on the heterogeneous information network (HIN) about miRNAs and target genes. Then the node feature and the path feature which is learned by the Deep Convolutional Neural Network (DCNN) are spliced together as the representation of the miRNA-target gene, to predict the miRNA-target gene interactions. The experiment results indicate that the performance of MDCNN outperforms other methods in multiple validation metrics by fivefold cross validation. We set an ablation study to identify the necessity of miRNA similarity and target gene similarity for improving the prediction ability of MDCNN. The case studies on hsa-miR-26b-5p and CDKN1A further demonstrates that MDCNN can successfully predict potential miRNA-target gene interactions.

Graphic abstract



Keywords miRNA-target gene associations · Network representation · Deep learning · Meta-path

Extended author information available on the last page of the article

1 Introduction

MicroRNAs (miRNAs) are a kind of endogenous small RNAs with about 20 nucleotides [1]. As one of the most important components in cells, miRNA can cause gene degradation or inhibit gene translation by complementary pairing with 3'UTRs of mRNA [2]. Biological experiments have confirmed that miRNAs are widely involved in a large number of cell processes, and are closely related to the occurrence and development of diseases [3, 4]. Up to now, there are 2,656 mature human miRNAs in miRbase [5]. Relevant studies have shown that such a small amount of miRNA regulates nearly one-third of human genes. Therefore, the prediction of miRNA-target interaction (MTI) is of great significance for understanding miRNA function and regulatory mechanism, preventing and treating human diseases.

In general, a miRNA can regulate multiple genes, and a gene can also be regulated by multiple miRNAs. Biological experimental methods can directly identify the existence of genes, but they are costly, time-consuming and unable to achieve a large scale of MTIs. Fortunately, with the rapid development of computing technology, researchers have proposed computational prediction methods for the MTI, many of which have been proved to be effective in promoting the design of biological experiments.

Early researchers proposed algorithms based on seed matching, thermal stability, species conservatism, and other principles [6], such as TargetScan [7], MiRanda [8], and PITA [9]. Although these methods have a low data dependency, due to the complex regulatory relationship between miRNA and its target genes, this kind of algorithm has a high false positive. To improve the accuracy of MTI, many methods based on machine learning have been proposed, such as ensemble learning [10], support vector machine (SVM) [11] and so on. MiTarget [12] designed a location-based feature, combining features based on structural and thermodynamics, and SVM was proposed to build miRNA-target classifiers. MiREE [13] believed the joint optimization of the Ab-Initio and machine learning parts can lead to better results. It generated a set of candidate sites upon a genetic algorithmic approach and the SVM learning module evaluated the influence of microRNA on target genes. Due to the development of crosslinking ligation and sequencing of hybrids (CLASH) experiments, new features of miRNA target sites may be inferred. TarPmiR [14] selected 13 important features from 18 potential features and predicted the target sites of miRNA based on the random forest method. In general, methods based on traditional machine learning can solve the problem of false positives to a certain extent, but these methods inevitably require manual extraction of feature data. Manual feature extraction is very time-consuming, and the constructed features may not be suitable for machine learning.

In recent years, with the increase of data and computing power, amounts of methods based on network representation learning have been proposed which could solve the problem of manual feature extraction. This type of method aims to learn low-dimensional representations of nodes in the network which retains the structure and inherent property information of the graph, and has been successfully applied in many domains, such as social network, recommendation system, natural language processing, and so on [15]. In the beginning, some approaches learn the representation of network nodes by randomly walking over homogeneous networks, such as DeepWalk [16], LINE [17], node2vec [18], GraRep [19]. Later, considering that most networks are heterogeneous, some methods for heterogeneous networks were proposed, such as metapath2vec [20], HIN2vec [21], and so on. At present, the association prediction methods based on network embedding have been successfully applied in the field of bioinformatics [22, 23] and computational pharmacology [24–26]. For example, IMTRBM [27] constructed a weighted miRNA-target bipartite network based on the prediction results of multiple single classification algorithms and applied Restricted Boltzmann Machine (RBM) to predict MTIs. SG-LSTM-FRAME [28] used Doc2vec to construct the network, then role2vec was used to learn the characteristics of nodes in the network. Finally, a Long Short-Term Memory (LSTM) module was trained to predict MTIs. BIRWMDA [29] first integrated a variety of similarity networks to obtain disease network and miRNA network and then carried out a bi-random walk on the network to predict miRNA-disease association. IDDkin [30] used a graph convolution network (GCN) to fuse neighbor information into heterogeneous networks. Then, combined with graph attention network (GAT) and adaptive weighting to predict kinase inhibitor association. The methods based on network representation learning can adaptively learn the information of network nodes, which not only solve the shortcomings of manual data extraction but also improve the prediction performance. However, most of these methods learn the representation of nodes separately and ignore the path information between node pairs.

This paper, we proposed an end-to-end deep learning framework, called MDCNN, to predict the MTIs. Considering that the meta-path can effectively represent the relationship between different types of nodes in the network and different meta-paths contain different information, we combined the representation of the node pair and path as the embedding of the miRNA-target gene by learning the representation of the meta-path between pairs of nodes. The MLP was used to predict the miRNA-target gene interactions. The evaluation results show that MDCNN is superior

to some MTI algorithms and other algorithms based on network embedding in fivefold cross-validation. Besides, the case studies further show the powerful ability of MDCNN in predicting MTIs.

2 Preliminary

Before introducing our method, we give the notations we used in our paper in Table 1, and the definitions related to our method as follows:

A heterogeneous information network (HIN) is an information network which contains many kinds of objects or links.

Definition 2.1 Heterogeneous Information Network [31]. A HIN is a graph $G = (V, E)$ with an entity type mapping function $\phi : V \rightarrow A$ and a link type mapping function $\varphi : E \rightarrow R$. A and R, respectively denotes the predefined entity set and link type set, where $|A| + |R| > 2$.

Example In this paper, we construct a HIN, including two types of objects (miRNA and target) and three types of relationships (miRNA-miRNA, target-target, and miRNA-target).

Due to the complexity of the HIN, meta-paths are used to describe the semantic relations between two nodes.

Definition 2.2 Meta-path [31]. A meta-path p is defined as a path in the form of $A_1 \overset{R_1}{\rightarrow} A_2 \overset{R_2}{\rightarrow} \cdots \overset{R_l}{\rightarrow} A_{l+1}$ (abbreviated as $A_1 A_2 \ldots A_{l+1}$), which describes a composite relation $R = R_1 \circ R_2 \circ \cdots \circ R_l$ between object A_1 and A_{l+1}, where $\circ$ denotes the composition operator on relations.

Table 1 Notations and explanations

Notation	Definition
$G = (V, E)$	A heterogeneous information network
$V = \{M \bigcup T\}$	The set of two types of nodes
$E = \{\text{EM} \cup \text{ET} \cup \text{EA}\}$	The set of three types of links
P	Meta-path
h	Initial node feature
e	Projected node feature
L	Path length set
r_p	Feature embedding of path p
RM	Path embedding matrix
f	The output of the convolutional layer
g	The output of the pooling layer
s	The output of the fully connected layer
$\chi_{\mathbb{P}}$	Path feature
Z	The final embedding representation
$\hat{y}$	The predicted label

Example In miRNA-target HIN, two nodes can be connected by multiple meta-paths, e.g., miRNA-target-miRNA (MTM) and target-miRNA-target (TMT). Each meta-path has its semantics. For example, the MTM means two miRNAs co-regulate the same target while TMT means two targets are regulated by the same miRNA.

3 Materials and Method

3.1 Datasets

The miRNA sequence information was download from miRbase [5] and we extracted 2,656 mature miRNAs as experimental data. We downloaded 509,664 association data between 17,929 genes from HumanNet v2 [32]. And the MTIs are downloaded from the known experimental database, mirTarBase [33]. After unioning and removing duplicates, we got 237,574 associations including 2547 miRNAs and 9096 target genes.

3.2 Method Overview

In our method, we first generated the miRNA similarity network by calculating the sequence similarity. Meanwhile, we generated the target gene similarity network from the database. The miRNA-target gene HIN is composed of the miRNA-miRNA network, target-target network, and known miRNA-target gene association network. To solve the problem that the existing methods ignore the path information between node pairs, we introduced meta-paths to capture node information in HIN. We gathered paths with different lengths between miRNAs and target genes, extracted the important information of the paths by a DCNN, and learned the path representation. Finally, the combination of miRNA representation, target gene representation, and path representation was used as the input of MLP to predict the MTIs. The overall workflow of MDCNN is depicted in Fig. 1.

3.3 Construction of the miRNA-Target Heterogeneous Network

The existing network-based methods tend to focus on the intrinsic characteristics of miRNAs and target genes while ignoring the heterogeneous information of biological networks. The structure and semantic information of HIN can help us obtain richer node information and improve the accuracy of the model. Hence, we integrate miRNA sequence data, target similarity data, and miRNA-target interaction data to construct the miRNA-target gene HIN.

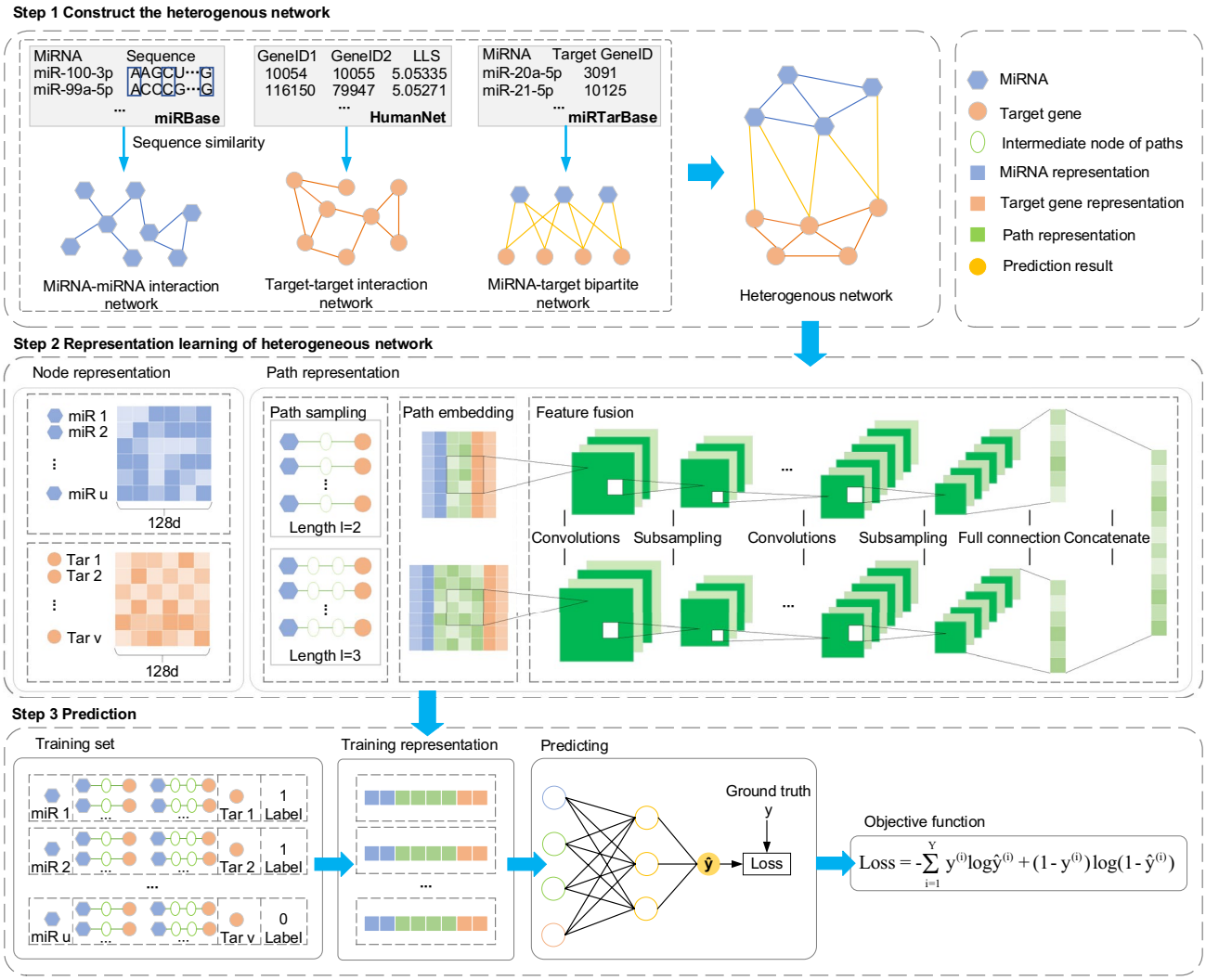


Fig. 1 The overview of MDCNN. Step 1: construct the heterogeneous network by combining the miRNA network, the target network, and the miRNA-target pair network. Step 2: learn the representation of nodes and paths based on different path lengths. Step 3: predict the interaction of miRNAs and target genes

3.3.1 MiRNA Similarity Network

We obtain the sequence information of miRNAs from miRbase [5] and calculate the similarity of miRNAs by the Needleman Wunsch algorithm [34]. The Needleman–Wunsch algorithm uses the principle of dynamic programming to match the sequences globally and optimizes the measurement to determine the similarity between the two miRNA sequences. However, there may be some unknown noise data in biological data, which will affect the experimental results. Therefore, for each miRNA, the top δ miRNAs with similarity scores are selected as an association to improve the reliability of the miRNA similarity network. Hence, let M be the set of miRNAs, and the edges EM in the miRNA network could be defined as follow:

$$\mathrm{EM} = \left\{ \left(m_i, m_j \right) | \mathrm{rank}_{m_i} \left(m_j \right) \leq \delta \right\}, \tag{1}$$

where $\mathrm{rank}_{m_i}\left(m_j\right) \leq \delta$ represents miRNA m_j is the top δ similarity of miRNA m_i, δ is a hyper-parameter.

3.3.2 Target Similarity Network

We download the human gene functional data from HumanNet v2 [32] to construct the target gene similarity network. Each interaction in HumanNet v2 represents the probability of interaction between two genes. Let T be the set of genes, the edges ET in the target gene network could be defined as follow:

$$\mathrm{ET} = \left\{ \left(t_i, t_j\right) | \mathrm{rank}_{t_i}\left(t_j\right) \le \delta \text{ and } \mathrm{LLS}\left(t_i, t_j\right) > \mathrm{avg(LLS)} \right\}, \tag{2}$$

where $\mathrm{rank}_{t_i}\left(t_j\right) \le \delta$ represents gene t_j is the top δ similarity of gene t_i, and avg(LLS) is the average score of the gene functional similarity.

3.3.3 MiRNA-Target Interaction Network

The edges EA in the miRNA-target network are obtained from the known experimental database, mirTarBase [33], which was defined as follow:

$$\mathrm{EA} = \left\{ \left(m_i, t_j\right) | m_i \in M \text{ and} t_j \in T \right\}, \tag{3}$$

where $\left(m_i, t_j\right)$ are the association of experimental verification in the database.

Finally, we integrate the miRNA similarity network, target similarity network, and miRNA-target interaction network to construct a HIN of miRNAs and target genes. The network could be represented as $G = (V, E)$, where $V = \{M \cup T\}$ and $E = \{\mathrm{EM} \cup \mathrm{ET} \cup \mathrm{EA}\}$. All the edges in the HIN represent the close relationship between nodes, and the weights are 1.

3.4 Representation Learning of Heterogeneous Network

In this part, we use the HIN constructed in the previous chapter as input to establish a deep learning model, which does not need to manually extract the representation of network nodes. Considering that most methods only utilize the structural information of the network and ignore the semantic information between nodes, we connect network nodes in the form of meta-paths and then use deep convolutional neural networks to learn the effective information in the paths.

3.4.1 Node Embedding

A large amount of network-based methods use one-hot encoding for the embedding of nodes. However, one-hot coding requires each category to be independent of each other, and its dimension depends on the size of the dataset. Thus, one-hot coding is not a wise choice in our research, in which miRNA-target is a complex network. Consequently, we use a transformation matrix to convert this one-hot embedding into a dense feature. The specific operation is as follows:

$$e_i = X \cdot h_i, \tag{4}$$

where $h_i \in \mathbb{R}^{(|\boldsymbol{M}|+|\boldsymbol{T}|)\times 1}$ is the original features of node i and $e_i \in \mathbb{R}^{d\times 1}$ (d: embedding size of nodes) is the projected features of node i. $\mathrm{X} \in \mathbb{R}^{\boldsymbol{d}\times(|\boldsymbol{M}|+|\boldsymbol{T}|)}$ is the transformation matrix to project nodes into a low dimensional continuous vector space.

3.4.2 Paths Sampling

We regard meta-paths between miRNAs and target genes as contexts for MTIs and assume they contain useful semantic information for the prediction of MTI. Therefore, we collect meta-paths between all miRNA-target pairs with different lengths. In this paper, we only concern with the regulation of miRNAs on target genes. So only meta-paths from miRNAs to target genes are selected in this paper. Table 2 shows meta-paths under different lengths.

Due to the sparsity and complexity of miRNA-target HIN, not all node pairs have all types of meta-paths. Therefore, we consider merging paths with the same length into one set. Let $L = \left\{l_1, l_2, \ldots, l_q\right\}$ denotes the set of different lengths. For each length $l \in L$, we collect a set of paths P^l and $\mathbb{P} = \left\{P^{l_1}, P^{l_2}, \ldots, P^{l_q}\right\}$ is the set of all paths.

3.4.3 Path Embedding

The information in the path is mainly reflected in the arrangement of the nodes. Therefore, we splice the node embedding vectors together as the representation of the path. Given a path $p = A_1 \to A_2 \to \cdots \to A_{l+1}$, the embedding could be shown as follow:

$$r_p = e_1 \parallel e_2 \parallel \cdots \parallel e_{l+1} \tag{5}$$

where $r_p \in \mathbb{R}^{((l+1)\cdot d)\times 1}$ is the feature embedding of path p, and $\parallel$ denotes the concatenate operation.

Then the paths of the same length are arranged in turn to form the path embedding matrix. Given a path set P^l, the matrix is formulated as follow:

$$\mathrm{RM}_l = \left[r_1, r_2, \ldots, r_K\right]^{\mathrm{T}}, \tag{6}$$

where $\mathrm{RM}_l \in \mathbb{R}^{K\times((l+1)\cdot d)}$, K is a her-parameter which denotes the size of the path set P^l.

3.4.4 Path Feature Fusion

Convolutional Neural Networks (CNN) can automatically extract features from basic data, which is a very useful feature fusion method. Study [35] shows that deep architectures is usually better than shallow architecture in dealing with complex learning problems. Hence, we apply Deep

Table 2 The description of the meta-path

Index	Meta-path	Length	Index	Meta-path	Length
P_1	$M \to T$	1	P_5	$M \to M \to T \to T$	3
P_2	$M \to T \to T$	2	P_6	$M \to M \to M \to T$	3
P_3	$M \to M \to T$	2	P_7	$M \to T \to T \to T$	3
P_4	$M \to T \to M \to T$	3	…	…	…

Convolutional Neural Networks (DCNN) to fuse features in path embedding matrix. The DCNN contains multiple convolutional layers and pooling layers. The output of the convolutional layer at location (x, y) in the jth feature tensor of ith layer is noted as $f_{i,j}^{x,y}$ and formally expressed according to:

$$f_{i,j}^{x,y} = \mathrm{ReLU}\left(\beta_{i,j} + \sum_{c} \sum_{\mathrm{ch}_1=0}^{\mathrm{CH}_i-1} \sum_{\mathrm{cw}_1=0}^{\mathrm{CW}_i-1} \alpha_{i,j,c}^{\mathrm{ch,cw}} \cdot f_{(i-1)c}^{x+\mathrm{ch},y+\mathrm{cw}}\right), \quad (7)$$

where $\alpha_{i,j,k}^{kh,kw}$ denotes the weight of the jth feature map of the ith convolutional layer, $\beta_{i,j}$ is the bias parameter, KH_i and KW_i are the size of the filter matrix, c is the index of the feature map. ReLU is the activation function, $\mathrm{ReLU}(x) = \max(0, x)$.

To avoid over-fitting, the average-pooling operator is used for each feature map. The output of the average-pooling layer noted as $g_{i,j}^{x,y}$, is formulated as follow:

$$g_{i,j}^{x,y} = \sum_{c} \left[\left(\sum_{\mathrm{ph}_1=0}^{\mathrm{PH}_i-1} \sum_{\mathrm{pw}_1=0}^{\mathrm{PW}_i-1} f_{(i-1)c}^{x+\mathrm{ph},y+\mathrm{pw}}\right) / \left(\mathrm{PH_i} \cdot \mathrm{PW}_i\right)\right], \quad (8)$$

where PH_i and PW_i are the height and width of the pooling matrix, respectively.

After alternatively stacking multiple convolutional layers and pooling layers, a fully connected layer is applied subsequently to transform the extracted feature map into a one-dimensional array. Finally, the fused features from different path embedding matrices are concatenated as the features of meta-paths set $\mathbb{P} = \left\{P^{l_1}, P^{l_2}, \ldots, P^{l_q}\right\}$. The feature of paths, $\chi_{\mathbb{P}}$, is formulated as follow:

$$\chi_{\mathbb{P}} = s_1 \parallel s_2 \parallel \cdots \parallel s_q, \quad (9)$$

where s_l is the output of lth path embedding matrix after the fully connected layer.

3.5 Prediction

We combine the features of miRNAi, target j, and paths between i and j, and then use the MLP to make the final relationship prediction. Let Z denotes the embedding after combination,

$$Z = e_i \parallel \chi_{\mathbb{P}} \parallel e_j. \quad (10)$$

The details of the MLP are as follows:

$$\mu_1 = \mathrm{ReLU}\left(W_1 \cdot Z + b_1\right), \quad (11)$$

$$\mu_i = \mathrm{ReLU}\left(W_i \cdot \mu_{i-1} + b_i\right) \quad (i = 2, \ldots, N) \quad (12)$$

$$\hat{y} = \mathrm{sigmoid}\left(W_N \cdot \mu_{N-1} + b_N\right), \quad (13)$$

where W and b are, respectively, represented the weight matrix and bias of the network layer, σ denotes the activation function used by the network layers, N is the number of layers, and $\hat{y}$ represents the label predicted by our model.

3.6 Objective Function

Cross-entropy is a common loss function of the classification problem. We use the cross-entropy loss as the objective function to train our model:

$$\mathrm{Loss} = -\sum_{i=1}^{Y} y^{(i)} \log \hat{y}^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right), \quad (14)$$

where Y refers to the training set, and y denotes the truth label. In this paper, we appeal to Adaptive Moment Estimation (Adam) [36] to minimize the total loss.

4 Results and Discussion

4.1 Experiment Settings

We applied fivefold cross-validation to evaluate the performance of the miRNA target gene relationship prediction model. Specifically, the data is randomly divided into five parts, one of which is selected as the test samples each time, and the remaining parts are used as the training samples to train the model. Calculate the average result of 5 folds and use it as the performance index of the model under fivefold cross-validation. In the experiment, known miRNA target gene association data are considered as positive samples. Since the unknown interactions are far more than the known interactions, we randomly select unassociated samples as negative samples with an equal number of positive samples in both the training and testing phase. We set the dimension d for path from {4,8, 16,32,64,128,256}, the length set L from {{2}, {3}, {2,3}, {2,3, 4}}, the threshold of similarity network δ from {5,10,20,30,40}, the number of layers from {1,2, 3,4}, the dimension of each layer in MLP is {128,64,1}, the path number K is 10, and the learning rate of Adam is 0.001. True positive (TP) and true negative (TN) represent the number of positive samples and negative samples correctly identified, respectively. False positive (FP) and false negative (FN) denote the number of positive samples and negative samples misidentified, respectively. The ROC curves are drawn by plotting the true positive rate (TPR) and the false positive rate (FPR), which are calculated as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}, \quad (15)$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN+FP}}, \tag{16}$$

The area under the receiver operating characteristics (ROC) curve (AUC) is used to evaluate the overall performance of the methods. Further, the area under the precision-recall (PR) curve (AUPR), the precision, the recall, and the F1 score are also shown in the form of tables calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{(TP+FP)}}, \tag{17}$$

$$\text{Recall} = \frac{\text{TP}}{\text{(TP+FN)}}, \tag{18}$$

$$\text{F1 score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{(precision+recall)}}. \tag{19}$$

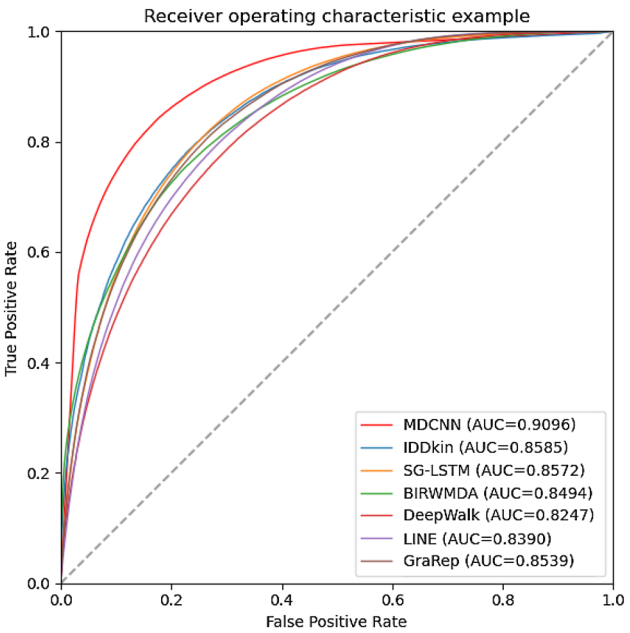


Fig. 2 The ROC curves of comparison methods

4.2 Comparison with Other Methods

For comparison, we compare our approach with those in the field of bioinformatics, computational pharmacology, and traditional representation learning methods. In the experiment, all the comparison methods are tested on our data and adjusted to their best parameters.

Figure 2 shows the ROC curves of MDCNN, SG-LSTM [28], BIRWMDA [29], IDDkin [30], DeepWalk [16], LINE [17], GraRep [19] obtained with fivefold cross-validation, respectively. As shown in Fig. 2, the AUC of MDCNN is 0.9096, and the AUC of other comparison methods are 0.8572, 0.8494, 0.8585, 0.8247, 0.8390, and 0.8539, respectively. The AUPR further evaluates the overall performance of the model which is shown in Table 3. From Table 3, the AUPR of MDCNN is 0.9143 which is higher than other methods [SG-LSTM (AUPR = 0.8385), BIRWMDA (AUPR = 0.8537), IDDkin (AUPR = 0.8566), LINE (AUPR = 0.8154), DeepWalk (AUPR = 0.8079), and GraRep (AUPR = 0.8375)]. More specifically, compared with the second-ranked method IDDkin, the performance of MDCNN is 5.77% higher than it. Subsequently, we compare the precision, recall, and F1 scores of the top 10%, top 20% and top 50% prediction results. Since recall in the case of the top 50% are the same as those in the case of F1 score and precision, only the recall value in the case of top 50% is shown here. As shown in Table 3, the recall at top 50% of MDCNN is 0.8326 and the precision, recall and F1 scores at top 10% and top 20% of MDCNN are 0.9730, 0.1946, 0.3244, 0.9725, 0.2890 and 0.5557, respectively. Compared with other methods, MDCNN is the best in all evaluation metrics, indicating that MDCNN achieves the best performance among these competing algorithms.

Three reasons may explain the superiority of MDCNN. First, MDCNN takes into account the attribute information of miRNA and gene, such as sequence information and functional similarity information, and constructs a more reliable heterogeneous information network of the miRNA target gene. Second, using meta-paths, the intermediate nodes are used as bridges to connect miRNAs and target genes which fully capture the structure and semantic

Table 3 The performance of MDCNN and other models using multiple evaluation metrics

Method	SG-LSTM	BIRWMDA	IDDkin	DeepWalk	LINE	GraRep	MDCNN
AUPR	0.8385	0.8520	0.8566	0.8079	0.8154	0.8375	0.9143
Recall	0.7730	0.7648	0.7742	0.7414	0.7543	0.7692	0.8326
Precision (top 10%)	0.9229	0.9723	0.9565	0.9044	0.8989	0.9251	0.9730
Recall (top 10%)	0.1846	0.1945	0.1913	0.1809	0.1798	0.1850	0.1946
F1 score (top 10%)	0.3076	0.3241	0.3189	0.3015	0.2996	0.3084	0.3244
Precision (top 20%)	0.8919	0.9219	0.9161	0.8651	0.8718	0.8937	0.9725
Recall (top 20%)	0.3568	0.3688	0.3665	0.3460	0.3487	0.3575	0.3890
F1 score (top 20%)	0.5097	0.5268	0.5235	0.4943	0.4982	0.5107	0.5557

information of the network. Thirdly, MDCNN is an end-to-end framework, which optimizes the parameters of the model while training the model, so that the overall performance of the model can be improved.

4.3 Ablation Study

In MDCNN, we used three types of networks. To prove the effectiveness of these networks in improving MDCNN's performance, we conducted the following ablation experiments with three designed variant models. Here MDCNN-1 indicates that experiments are only performed on the miRNA-target gene binary network, that is, only miRNA-target gene-related data is considered. MDCNN-2 and MDCNN-3 represent the addition of the miRNA network and the gene network to the miRNA-target gene association network, respectively.

Figure 3 shows the impact of different networks on the performance of the model. The performance of variants MDCNN-2 and MDCNN-3 are better than that of MDCNN-1, which proves that adding similarity data could improve the performance. Considering both the miRNA data and the gene data, the model MDCNN in this article has better performance than the models with only add a single network, i.e., MDCNN-2, and MDCNN-3. The reason may be that the constructed heterogeneous information network contains more structural information and semantic information than the binary network, which has a positive impact on the improvement of model performance.

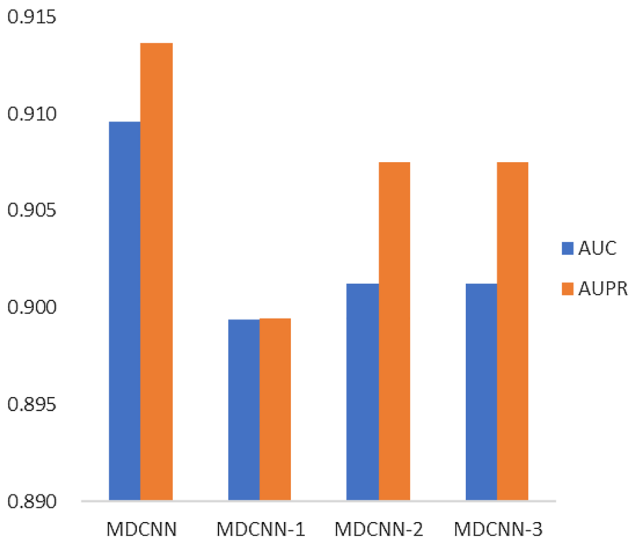


Fig. 3 The impact of different networks on the performance of the MDCNN

4.4 Parameter Sensitivity

In this section, we investigate the influence of different parameters on the performance of the model. We report AUC to analyze the predicted results, as shown in Fig. 4.

Figure 4a shows how the embedding dimension affects the performance of MDCNN. We set the embedding dimension from {4,8, 16,32,64,128,256}. With the increase of dimension, the performance of MDCNN first increases and then remains stable. This means that too large embedded dimensions may introduce noise data, which makes MDCNN unable to capture more useful information.

The path length set is another important parameter of MDCNN, so we experiment on the variable of the path set L. We consider the L from $\{\{2\},\{3\},\{2,3\},\{2,3,4\}\}$. It can be seen from Fig. 4b that it is better to consider path lengths of 2 and 3 at the same time than to consider only one of them. At the same time, although the performance of path combination $\{2,3,4\}$ is better than $\{2,3\}$, it is not obvious. Therefore, considering the efficiency of the model, we choose $\{2,3\}$ as the final path combination.

Different thresholds δ lead to different heterogeneous networks, and will affect the performance. We set the similarity network threshold δ from {5,10,20,30,40} to analyze their impact on the performance of the model. As can be seen from Fig. 4c, the model achieves the best performance when $\delta = 10$. With the increase of δ, the model performance may decline due to the introduction of more noise. Therefore, $\delta = 10$ is finally determined as the final threshold of the model. At the same time, we can see that when $\delta > 10$, the AUC of the model does not change much, indicating that the model has strong robustness. Therefore, choosing $\delta = 10$ can not only make the model get the best performance, but also save the resources needed for the model running.

Finally, we consider the influence of the number of layers of the convolutional neural networks on the performance of the model. We set layers from 1 to 4 to analyze the model. It can be seen from Fig. 4d that the performance of the model tends to improve with the superposition of layers, but the improvement effect is not obvious. We consider that with the increase of the number of layers, the time complexity of the model also increases exponentially, we choose 3 as the final number of layers.

4.5 Analysis of Negative Sampling

In this experiment, we train the model by sampling the same number of unknown correlation samples as positive samples. The positive and negative proportion of real training samples is unbalanced, so we study the influence of the different positive and negative proportion of training sets on the performance of the model in this section.

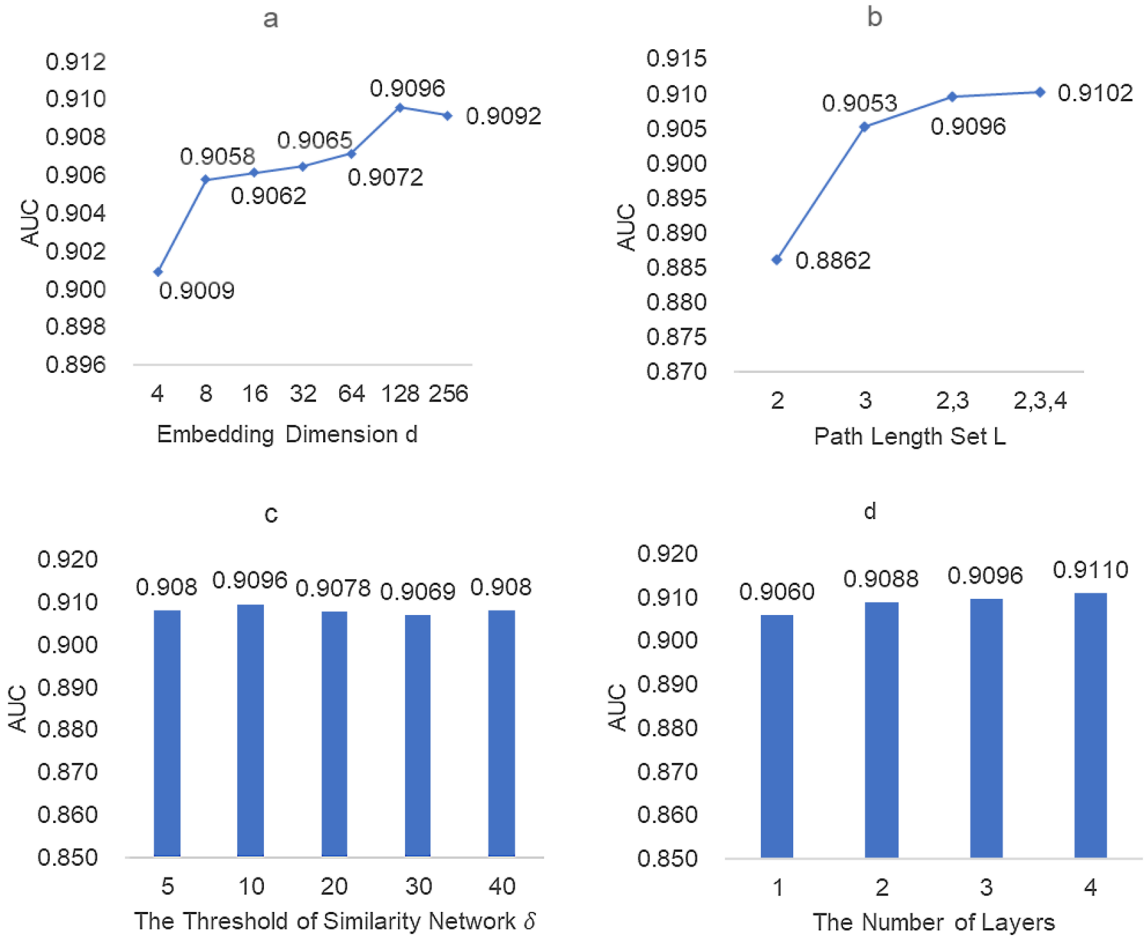


Fig. 4 The effect of parameters change on the MDCNN. **a** The effects of embedding size. **b** The effects of path length set. **c** The effects of thresholds in similarity network. **d** The effects of the number of layers

Table 4 The influence of a different number of negative samples on the MDCNN

Ratios	Positive samples size	Negative samples size	AUC
1:1	237,574	237,574	0.9096
1:2	237,574	475,148	0.9234
1:3	237,574	712,722	0.9327
1:4	237,574	950,296	0.9379
1:5	237,574	1,187,870	0.9429

Specifically, we studied the positive and negative ratios of {1 : 1,1 : 2,1 : 3,1 : 4,1 : 5}, respectively. Table 4 shows the experimental results. The results indicate that MDCNN can still produce good results even if the dataset is unbalanced. If we use more negative samples, our model can still get good performance.

4.6 Case Studies

Case studies are conducted to further verify the capability of MDCNN to detect novel miRNA-target gene associations. Breast cancer is one of the most common cancers in women worldwide. Experiment [37] has shown that hsa-miR-26b-5p, as one of the miRNAs with the largest number of gene associations, has a close relationship with breast cancer. CDKN1A is one of the genes with the largest increase in the number of associations experimentally verified in the past two years, and it is closely related to hepatocellular carcinoma [38]. Therefore, we did case studies on a miRNA (hsa-miR-26b-5p) and a target gene (CDKN1A), respectively. For hsa-mir-26b-5p, we used all of the known positive samples in the dataset and equal-size negative samples which include all unknown entries of hsa-mir-26b-5p to train MDCNN. Table 5 shows the top 10 miRNA-target gene relationships predicted by MDCNN and the predicted results were verified by searching literature in PubMed. As shown in Tables 4 and 5 candidate target genes were supported by

Table 5 The top 10 potential candidates of hsa-mir-26b-5p detected by MDCNN

Rank	Target gene	Target gene (Entrez gene ID)	Score	Evidence
1	KCTD5	54442	0.9882	–
2	ABRAXAS1	84142	0.9748	–
3	KIAA1468	57614	0.9229	PMID: 26123714
4	IKBKG	8517	0.8808	–
5	CHRDL1	91851	0.8624	PMID: 32832554
6	CYP3A4	1576	0.7893	PMID: 27756246
7	SLC5A6	8884	0.7872	–
8	GAS1	2619	0.7792	PMID: 30647817
9	CHEK1	1111	0.7249	–
10	PRPF4B	8899	0.6326	–

Table 6 The top 10 potential candidates of CDKN1A detected by MDCNN

Rank	MiRNA	Score	Evidence
1	hsa-miR-1182	0.9045	–
2	hsa-miR-92a-3p	0.7774	PMID: 26482648
3	hsa-miR-8066	0.7316	–
4	hsa-miR-1304-3p	0.7070	–
5	hsa-miR-432-5p	0.6926	PMID: 25762502
6	hsa-miR-30b-5p	0.6815	PMID: 31463131
7	hsa-miR-7153-5p	0.6572	–
8	hsa-miR-4716-5p	0.6479	–
9	hsa-miR-150-5p	0.6029	PMID: 26644403
10	hsa-miR-4790-3p	0.6006	–

the literature. For example, Grilli et al. [39] found that hsa-mir-26b-5p and KIAA1468 is one of the most interesting couples after using the intersection of prediction and correlation approaches. Furthermore, Chen et al. [40] pointed out that CYP3A4 may play critical roles in the development through the regulation of hsa-miR-26b-5p. It is worth noting that although the difference of predicted scores is relatively large, the difference of scores is reasonable because of the different degrees of each node in HIN and the different effective information captured by each node.

For CDKN1A, we used all of the known positive samples in the dataset and equal-size negative samples which includes all unknown entries of CDKN1A to train MDCNN. Then the pairs of CDKN1A are used for prediction. Table 6 shows the top 10 potential miRNAs of CDKN1A predicted by MDCNN. In Tables 4 and 6 of 10 miRNAs were identified by PubMed. Zhao et al. [41] suggested that the low expressions of miR-92 families, which results in high expressions of CDKN1A. And in the study of the interaction between mir-30b-5p and esophageal squamous cell carcinoma, Xu et al. [42] found that mir-30b-5p could down-regulate the expression of the CDKN1A gene.

In summary, the prediction results further indicated that the effectiveness of MDCNN covering potential miRNA-target gene associations.

5 Conclusions

In conclusion, the novel computational framework MDCNN we proposed is superior to other state-of-art methods. In the experiment, we confirmed that the information from multiple sources of data is helpful to the improvement of model performance. Therefore, we will consider more miRNA and target gene information in the next research, such as miRNA family information, gene sequence information, to construct a more biologically meaningful miRNA target gene heterogeneous information network. In addition, we randomly select data with the same size as the positive sample from the unknown correlation sample as the negative sample in the experiment, which may cause the performance of the model to have certain volatility. Thus, we will consider proposing a new and properer negative sampling method to improve the stability of the model in the future.

Acknowledgements This work has been supported by the National Natural Science Foundation of China (Grant nos. 62032007, 61873089), Hunan Provincial Innovation Foundation for Postgraduate (Grant no. CX20200436)

References

1. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116(2):281–297. https://doi.org/10.1016/S0092-8674(04)00045-5
2. Ambros V (2004) The functions of animal microRNAs. Nature 431(7006):350–355. https://doi.org/10.1038/nature02871
3. Xia W, Cao G, Shao N (2009) Progress in miRNA target prediction and identification. Sci China Ser C Life Sci 52(12):1123–1130. https://doi.org/10.1007/s11427-009-0159-4
4. Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, Cui Q (2008) An analysis of human microRNA and disease associations. PLoS ONE 3(10):e3420. https://doi.org/10.1371/journal.pone.0003420
5. Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) miRBase: from microRNA sequences to function. Nucleic Acids Res 47(D1):D155–D162. https://doi.org/10.1093/nar/gky1141
6. Wei L, Huang Y, Qu Y, Jiang Y, Zou Q (2012) Computational analysis of miRNA target identification. Curr Bioinform 7(4):512–525. https://doi.org/10.2174/157489312803900974
7. Lewis BP, Shih I-H, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of mammalian microRNA targets. Cell 115(7):787–798. https://doi.org/10.1016/S0092-8674(03)01018-3
8. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004) Human microRNA targets. PLoS Biol 2(11):e363. https://doi.org/10.1371/journal.pbio.0020363
9. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. Nat Genet 39(10):1278–1284. https://doi.org/10.1038/ng2135

10. Opitz D, Maclin R (1999) Popular ensemble methods: an empirical study. J Artif Intell Res 11:169–198. https://doi.org/10.1613/jair.614
11. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297. https://doi.org/10.1007/BF00994018
12. Kim S-K, Nam J-W, Rhee J-K, Lee W-J, Zhang B-T (2006) miTarget: microRNA target gene prediction using a support vector machine. BMC Bioinform 7(1):1–12. https://doi.org/10.1186/1471-2105-7-411
13. Reyes-Herrera PH, Ficarra E, Acquaviva A, Macii E (2011) miREE: miRNA recognition elements ensemble. BMC Bioinform 12(1):454. https://doi.org/10.1186/1471-2105-12-454
14. Ding J, Li X, Hu H (2016) TarPmiR: a new approach for microRNA target site prediction. Bioinformatics 32(18):2768–2775. https://doi.org/10.1093/bioinformatics/btw318
15. Chen H, Perozzi B, Al-Rfou R, Skiena S (2018) A tutorial on network embeddings. arXiv preprint arXiv:1808.02590 [v1]
16. Perozzi B, Al-Rfou R, Skiena S (2014) DeepWalk: online learning of social representations. ACM. https://doi.org/10.1145/2623330.2623732
17. Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q (2015) LINE: large-scale information network embedding. WWW. https://doi.org/10.1145/2736277.2741093
18. Grover A, Leskovec J (2016) node2vec: scalable feature learning for networks. ACM. https://doi.org/10.1145/2939672.2939754
19. Cao S, Wei L, Xu Q (2015) GraRep: learning graph representations with global structural information. ACM. https://doi.org/10.1145/2806416
20. Dong Y, Chawla NV, Swami A (2017) metapath2vec: scalable representation learning for heterogeneous networks. ACM. DOI 10(1145/3097983):3098036
21. Fu TY, Lee WC, Zhen L (2017) HIN2Vec: explore meta-paths in heterogeneous information networks for representation learning. ACM. DOI 10(1145/3132847):3132953
22. do Valle ÍF, Menichetti G, Simonetti G, Bruno S, Zironi I, Durso DF, Mombach JC, Martinelli G, Castellani G, Remondini D (2018) Network integration of multi-tumour omics data suggests novel targeting strategies. Nat Commun 9(1):1–10. https://doi.org/10.1038/s41467-018-06992-7
23. Wang L, Nie R, Yu Z, Xin R, Zheng C, Zhang Z, Zhang J, Cai J (2020) An interpretable deep-learning architecture of capsule networks for identifying cell-type gene expression programs from single-cell RNA-sequencing data. Nat Mach Intell 2(11):693–703. https://doi.org/10.1038/s42256-020-00244-4
24. Shen C, Luo J, Ouyang W, Ding P, Wu H (2020) Identification of small molecule–miRNA associations with graph regularization techniques in heterogeneous networks. J Chem Inf Model 60(12):6709–6721. https://doi.org/10.1021/acs.jcim.0c00975
25. Shen C, Luo J, Lai Z, Ding P (2020) Multiview joint learning-based method for identifying small-molecule-associated MiRNAs by integrating pharmacological, genomics, and network knowledge. J Chem Inf Model 60(8):4085–4097. https://doi.org/10.1021/acs.jcim.0c00244
26. Luo J, Shen C, Lai Z, Cai J, Ding P (2020) Incorporating clinical, chemical and biological information for predicting small molecule-microRNA associations based on non-negative matrix factorization. IEEE/ACM Trans Comput Biol Bioinform PP(99):1–1. https://doi.org/10.1109/TCBB.2020.2975780
27. Liu Y, Luo J, Ding P (2018) Inferring MicroRNA targets based on restricted Boltzmann machines. IEEE J Biomed Health Inform 23(1):427–436. https://doi.org/10.1109/JBHI.2018.2814609
28. Xie W, Luo J, Pan C, Liu Y (2020) SG-LSTM-FRAME: a computational frame using sequence and geometrical information via LSTM to predict miRNA–gene associations. Brief Bioinform 22(2):2032–2042. https://doi.org/10.1093/bib/bbaa022
29. Zhu Q, Fan Y, Pan X (2020) Fusing multiple biological networks to effectively predict miRNA-disease associations. Curr Bioinform 16(3):371–384. https://doi.org/10.2174/1574893615999200715165335
30. Shen C, Luo J, Ouyang W, Ding P, Chen X (2020) IDDkin: Network-based influence deep diffusion model for enhancing prediction of kinase inhibitors. Bioinformatics 36(22–23):5481–5491. https://doi.org/10.1093/bioinformatics/btaa1058
31. Sun Y, Han J, Yan X, Yu PS, Wu T (2011) Pathsim: meta path-based top-k similarity search in heterogeneous information networks. Proc VLDB Endow 4(11):992–1003. https://doi.org/10.14778/3402707.3402736
32. Sohyun H, Yeong KC, Yang S, Eiru K, Traver H, Marcotte EM, Insuk L (2018) HumanNet v2: human gene networks for disease research. Nucleic Acids Res D1:D573–D580. https://doi.org/10.1093/nar/gky1126
33. Chou C-H, Shrestha S, Yang C-D, Chang N-W, Lin Y-L, Liao K-W, Huang W-C, Sun T-H, Tu S-J, Lee W-H (2018) miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res 46(D1):D296–D302. https://doi.org/10.1093/nar/gkx1067
34. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453. https://doi.org/10.1016/0022-2836(70)90057-4
35. Khan A, Sohail A, Zahoora U, Qureshi AS (2020) A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 53(8):5455–5516. https://doi.org/10.1007/s10462-020-09825-6
36. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 [v4]
37. Liu X-X, Li X-J, Zhang B, Liang Y-J, Zhou C-X, Cao D-X, He M, Chen G-Q, He J-R, Zhao Q (2011) MicroRNA-26b is underexpressed in human breast cancer and induces cell apoptosis by targeting SLC7A11. FEBS Lett 585(9):1363–1367. https://doi.org/10.1016/j.febslet.2011.04.018
38. Fornari F, Milazzo M, Chieco P, Negrini M, Marasco E, Capranico G, Mantovani V, Marinello J, Sabbioni S, Callegari E (2012) In hepatocellular carcinoma miR-519d is up-regulated by p53 and DNA hypomethylation and targets CDKN1A/p21, PTEN, AKT3 and TIMP2. J Pathol 227(3):275–285. https://doi.org/10.1002/path.3995
39. Grilli A, Sciandra M, Terracciano M, Picci P, Scotlandi K (2015) Integrated approaches to miRNAs target definition: time-series analysis in an osteosarcoma differentiative model. BMC Med Genom 8(1):34. https://doi.org/10.1186/s12920-015-0106-0
40. Chen Z, Wu H, Wang G, Feng Y (2016) Identification of potential candidate genes for hypertensive nephropathy based on gene expression profile. BMC Nephrol 17(1):149. https://doi.org/10.1186/s12882-016-0366-8
41. Zhao J, Fu W, Liao H, Dai L, Jiang Z, Pan Y, Huang H, Mo Y, Li S, Yang G (2015) The regulatory and predictive functions of miR-17 and miR-92 families on cisplatin resistance of non-small cell lung cancer. BMC Cancer 15(1):1–14. https://doi.org/10.1186/s12885-015-1713-z
42. Xu J, Lv H, Zhang B, Xu F, Zhu H, Chen B, Zhu C, Shen J (2019) miR-30b-5p acts as a tumor suppressor microRNA in esophageal squamous cell carcinoma. J Thorac Dis 11(7):3015. https://doi.org/10.21037/jtd.2019.07.50

Authors and Affiliations

Jiawei Luo[1] · Yaoting Bao[1] · Xiangtao Chen[1] · Cong Shen[1]

✉ Xiangtao Chen
xtchen2009@sina.cn

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha 410083, China