DHNLDA: A novel deep hierarchical network based method for predicting lncRNA-disease associations

Fansen Xie, Ziqi Yang, Jinmiao Song*,Qiguo Dai, Xiaodong Duan

Abstract—Recent studies have found that lncRNA (long non-coding RNA) in ncRNA (non-coding RNA) is not only involved in many biological processes, but also abnormally expressed in many complex diseases. Identification of lncRNA-disease associations accurately is of great significance for understanding the function of lncRNA and disease mechanism. In this paper, a deep learning framework consisting of stacked autoencoder(SAE), multi-scale ResNet and stacked ensemble module, named DHNLDA, was constructed to predict lncRNA-disease associations, which integrates multiple biological data sources and constructing feature matrices. Among them, the biological data including the similarity and the interaction of lncRNAs, diseases and miRNAs are integrated. The feature matrices are obtained by node2vec embedding and feature extraction respectively. Then, the SAE and the multi-scale ResNet are used to learn the complementary information between nodes, and the high-level features of node attributes are obtained. Finally, the fusion of high-level feature is input into the stacked ensemble module to obtain the prediction results of lncRNA-disease associations. The experimental results of five-fold cross-validation show that the AUC of DHNLDA reaches 0.975 better than the existing methods. Case studies of stomach cancer, breast cancer and lung cancer have shown the great ability of DHNLDA to discover the potential lncRNA-disease associations.

Index Terms—deep learning, lncRNA-disease associations, multi-scale ResNet, autoencoder, node2vec.

1 Introduction

Recent transcriptomics and bioinformatics analyses have shown that only a small number of genes can encode protein sequences, and that more than 98% of the genes in the human genome cannot encode protein sequences [1]. Among them, non-coding RNA (ncRNA), especially long non-coding RNA (lncRNA) with a length greater than 200 nucleotides, plays an important biological role in chromatin remodeling, transcriptional regulation and post-transcriptional regulation [2]. A large number of studies have shown that lncRNA is related to many complex human diseases closely [3]. For example, the expression level of lncRNA PCA3 in prostate tumors cells is about 60 times higher than that in normal tissues cells [4], [5]. LncRNA BC088414 can promote apoptosis through some genes, thus aggravating hypoxic-ischemic injury of nerve cells [6], [7]. Therefore, predicting the potential lncRNA-disease associations can help us understand more about the pathogenesis of complex human diseases at the molecular level, as well as disease diagnosis, drug discovery and potential drug target research. However, the biological experiments, which a large number of literatures and databases are collected and processed usually costly and time-consuming. With the sequencing of various biological genomes has been completed, as well as the establishment and improvement of various relevant databases, it is become an important means to apply computational methods in the study of prediction of the potential lncRNA-disease associations [8].

Existing methods to study the potential lncRNA-disease associations can be divided into three categories. Job of the first category, the lncRNA-disease association prediction modal based on the biological theories that if two lncRNAs have similar functions with each other, they are often related to similar diseases. Ping et al. [9] proposed a bipartite network by calculating the lncRNA and disease similarity to predict the potential lncRNA-disease associations. Chen et al. [10] developed a calculation models to predict the potential lncRNA-disease associations based on the assumption that functionally similar lncRNAs are often associated with similar diseases. Sun et al. [11] proposed RWRlncD to predict the potential lncRNA-disease associations by implementing the random walk with restart method on a lncRNA functional similarity network. However, these methods rely too heavily on a set of seed genes that have been observed to be associated with disease and have little effect on new diseases for which there are unknown related genes.

The main job of the second category lies on integrating the multiple data sources to reveal the potential lncRNA-disease association. Yao et al. [12] proposed RFLDA, taking miRNA-disease association and lncRNA-disease association, disease semantic similarity, lncRNA functional similarity and lncRNA-miRNA interaction as input features. Then, a random forest regression model is trained to discover po-

- *Fansen Xie, Ziqi Yang, Qiguo Dai and Xiaodong Duan are with the Department of Computer Science and Engineering and the SEAC Key Laboratory of Big Data Applied Technology and also with the Dalian Key Lab of Digital Technology for National Culture, Dalian Minzu University, Dalian, 116600, China. E-mail:{7481342, 963346256}@qq.com, {daiqiguo, duanxd}@dlnu.edu.cn*
- *Jinmiao Song is with the SEAC Key Laboratory of Big Data Applied Technology and the Dalian Key Lab of Digital Technology for National Culture, Dalian Minzu University, Dalian, 116600, China. And also with the Department of Information Science and Engineering, XinJiang University, Urumqi, 830008, China. E-mail: sjm@dlnu.edu.cn*
- ** is the corresponding author.*

Manuscript received April 19, 2005; revised August 26, 2015.

tential lncRNA-disease associations. Gu et al. [13] developed GrwLDA, which integrates three associations include disease semantic similarities, lncRNA functional similarities, and known lncRNA-disease associations, and constructed the network to discover the potential associations. Wang et al. [14] proposed LncDisAP to predict potential lncRNA-disease associations based on multiple biological datasets. Marissa et al. [15] exploited the topology of multi-level networks consisted of the interaction of lncRNA, protein and disease, to propose the LION approach to identify lncRNA-disease associations. However, these methods are difficult to integrate heterogeneous data deeply from multiple sources.

Job of the third category lies on the deep learning, which better performs the learning ability on featured expressive information, so that to improve the prediction performance of the potential lncRNA-disease association. Xuan et al. [16] proposed a method based on the convolutional neural network with attention mechanism and convolutional autoencoder to predict the potential lncRNA-disease associations. Later, they proposed GCNLDA [17] baesd on the graph convolutional network and convolutional neural network (CNN) [18] to predict potential lncRNA-disease associations. Madhavan et al. [19] proposed DBNLDA to learn feature and to predict the potential lncRNA-disease association by using DBN and neural network. Sheng et al. [20] proposed VADLP, which used convolutional and variance autoencoders to discover potential lncRNA-disease associations. Zhang et al. [21] proposed GAAN for the potential lncRNA-disease associations prediction, which intergrate graph convolution networks and attention mechanism. However, the multiple data sources of some methods are still insufficient, and the complementary information between gene nodes is often ignored.

In this paper, we propose a deep learning prediction model DHNLDA by integrating multiple biological data sources, constructing feature matrices and learning node attribute feature representation to obtain node network structure information, complementary information and deep feature information. It is worth noting that the use of stacked ensemble module in the classification part makes DHNLDA achieve good classification effect. The contributions of our model included:

(1) The similarity and the interaction between lncRNA, miRNA and disease multiple biological data sources are integrated. The adjacency information matrix and the topological information matrix based on node attributes are obtained to supplement and enhance the attribute information.

(2) Deep feature representation of the nodes attributes complementary information can be learned adaptively by using SAE (stacked autoencoder) and multi-scale ResNet.

(3) The stacked ensemble module contains multiple shallow learning classifiers to further improve the performance of the model by combining the diversities of different classifiers.

2 MATERIALS AND METHODS

2.1 Dataset

The datasets used in this experiment are obtained from Fu et al. [22], which contained 240 lncRNAs, 495 miRNAs, and 412 diseases. lncRNA-disease associations (L-DAs) data are downloaded from the databases of LncRNADisease [23], Lnc2Cancer [24] and GeneRIF [25], 2697 known L-DAs are obtained. LncRNA-miRNA interactions (L-MIs) data and miRNA-disease associations (M-DAs) data are downloaded from starBase database [26] and HMDD (V2.0) [27] databases, respectively, where 1002 known L-MIs are obtained from Starbase and 13,562 known M-DAs are obtained from HMDD. Disease names come from the US National Library of Medicine (MeSH, http://www.ncbi.nlm.nihs.gov/mesh).

2.2 Construction of L-MS, M-DS and L-DA networks

In this section, Madhavan' s method [19] is used to construct lncRNA-miRNA similarity (L-MS) network, miRNA-disease similarity (M-DS) network and lncRNA-disease association (L-DA) network. In the construction of similarity matrix, the disease semantic similarity and the lncRNA function similarity are calculated by using the method of Fan et al. [28]. Among them, the disease semantic similarity beween two diseases is calculated by their DAGs. The disease semantic similarity matrix S^d can be defined as follows,

$$S^d = \begin{bmatrix} 1 & 0.1 & \cdots & 0.2 \\ 0 & 1 & \cdots & 0.35 \\ \vdots & \vdots & \vdots & \vdots \\ 0.45 & 0 & \cdots & 1 \end{bmatrix} \tag{1}$$

The lncRNA function similarity between two lncRNAs is calculated by the diseases which were related to them. The lncRNA function similarity matrix S^l can be defined as follows,

$$S^l = \begin{bmatrix} 1 & 0 & \cdots & 0.45 \\ 0.2 & 1 & \cdots & 0.15 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0.5 & \cdots & 1 \end{bmatrix} \tag{2}$$

In the construction of the association matrix, L-DAs matrix A^{l-d} is formed according to whether lncRNA is associated with disease. If a lncRNA is associated with a disease, the matrix element sets to 1, otherwise, 0. The association matrix A^{l-d} can be defined as follows,

$$A^{l-d} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 \end{bmatrix} \tag{3}$$

Similarly, L-MIs matrix A^{l-m} and M-DAs matrix A^{m-d} are constructed in the same way as A^{l-d}.

For the similarity network L-MS, the similarity network is constructed by using the lncRNA functional similarity and the lncRNA-miRNA interaction (L-MI). According to the type of connection edges, they are divided into intra-layer edges and inter-layer edges, and an edge is added by the inter-layer association matrix A^{l-m} and intra-layer similarity matrix S^l, respectively. Among them, if the matrix element > 0, an un-directed edge will be added. Similarly, M-DS is constructed by using the disease semantic similarity and the miRNA-disease association (M-DA). The L-DA network contains 240 lncRNAs and 412 diseases, in which an

undirected edge is used to represent the known association between lncRNAs and diseases.

2.3 Computing node embedding features from networks

In this experiment, node2vec [29], [30] is used to embed nodes in each network. Node2vec outputs node representation according to the neighborhood information of nodes to enhance the continuity between nodes, which can reflect the features of network and node neighbors. Different from Deepwalk [31], node2vec improves the generation mode of random walk, making the generated random walk reflect the two sampling characteristics of Depth-First Sampling (DFS) and Bread-First Sampling (BFS), thus improving the effect of network embedding, as shown in Fig 1.

The adjacency information matrix obtained in this paper is described as follows: first, the L-MS network is embedded through the node2vec, and the embedding matrix taking lncRNA as nodes is expressed as $L^s \in R^{nl\times e}$, where nl is the number of lncRNAs and e is the embedding dimension. D-MS network is embedded through the node2vec, and the adjacency information matrix taking disease as nodes is expressed as $D^s \in R^{nd\times e}$, where nd is the number of diseases. Then, for every lncRNA-disease pair in the datasets (including positive and negative samples), $L^s \in R^{nl\times e}$ and $D^s \in R^{nd\times e}$ are joined together to generate feature matrix $LD^s \in R^{2e\times n}$, which $LD^s \in R^{2e\times n}$ is expressed as $Y_1 \in R^{2e\times n}$, $Y_1 \in R^{2e\times n}$ can be expressed as follows,

$$Y_1 = [L^s; D^s] \in R^{2e\times n} \tag{4}$$

Where n is the total number of samples in the datasets. Finally, the adjacency information matrix of lncRNA and disease are obtained from L-DA network which are expressed as $L^a \in R^{nl\times e}$ and $D^a \in R^{nd\times e}$ respectively. In the same way as above, the vectors of $L^a \in R^{nl\times e}$ and $D^a \in R^{nd\times e}$ are joined together to generate the feature matrix $LD^a \in R^{2e\times n}$, which $LD^a \in R^{2e\times n}$ is expressed as $Y_2 \in R^{2e\times n}$, $Y_2 \in R^{2e\times n}$ can be expressed as follows,

$$Y_2 = [L^a; D^a] \in R^{2e\times n} \tag{5}$$

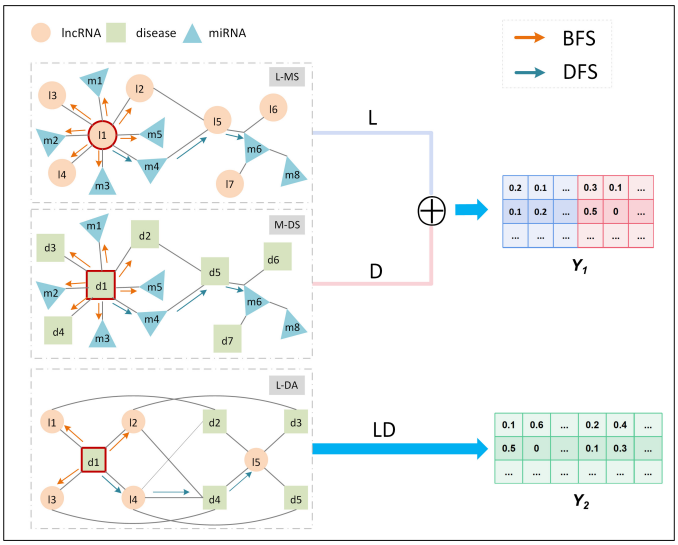

Fig. 1. Construction of the adjacency information matrix.

2.4 Heterogenous node attributes extraction by embedding strategy

The topology information obtained in this section is used to supplement the adjacency information, and enhance the node attribute information as a whole. Referring to the method of Xuan et al. [32], node attributes topology matrix can be divided into three part. Biological studies have shown that if a lncRNA and a disease have similar association with common lncRNAs (diseases or miRNAs), they are more possible to be associated with each other, and three submatrices are obtained respectively. The framework for constructing topology information matrix is shown in Fig 2. The n^{th} lncRNA ln and the m^{th} disease dm were taken as examples to illustrate the construction process of the topology information matrix. Firstly, the associations derived from common lncRNAs are embedded. Specifically, in the similarity matrix S^l, S_n^l denote the similarity between the lncRNA in row n^{th} and all lncRNAs. In the association matrix $A^{(l-d)^T}$, $A_m^{(l-d)^T}$ denote the association between the disease in column m and all lncRNAs. We stack S_n^l and $A_m^{(l-d)^T}$ together to obtain the submatrix X_1,

$$X_1 = \begin{bmatrix} S_n^l \\ A_m^{(l-d)^T} \end{bmatrix} \tag{6}$$

Secondly, the associations obtained from common diseases are embedded. Similar to the above, in the similarity matrix S^d, S_m^d denote the similarity between the disease in row m^{th} and all diseases. A_n^{l-d} denote the association between the lncRNA in row n^{th} and all diseases. A_n^{l-d} and S_m^d are stacked together to get the submatrix X_2,

$$X_2 = \begin{bmatrix} A_n^{l-d} \\ S_m^d \end{bmatrix} \tag{7}$$

Thirdly, the interactions derived from common miRNAs are embedded. In the association matrix A^{l-m},A_n^{l-m} denote the association between the lncRNA in row n^{th} and all miRNAs. In the association matrix $A^{(m-d)^T}$, $A_m^{(m-d)^T}$ denote the association between the disease in column m^{th} and all miRNAs. A_n^{l-m} and $A_m^{(m-d)^T}$ are stacked together to get the submatrix X_3,

$$X_3 = \begin{bmatrix} A_n^{l-m} \\ A_m^{(m-d)^T} \end{bmatrix} \tag{8}$$

Finally, lncRNA function similarity, disease semantic similarity, L-DA, L-MI and M-DA are integrated to construct the topology information matrix X between ln and dm, and the calculation formula is as follows,

$$X = \begin{bmatrix} X_1 & X_2 & X_3 \end{bmatrix} = \begin{bmatrix} S_n^l & A_n^{l-d} & A_n^{l-m} \\ A_m^{(l-d)^T} & S_m^d & A_m^{(m-d)^T} \end{bmatrix} \tag{9}$$

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/TCBB.2021.3113326, IEEE/ACM Transactions on Computational Biology and Bioinformatics

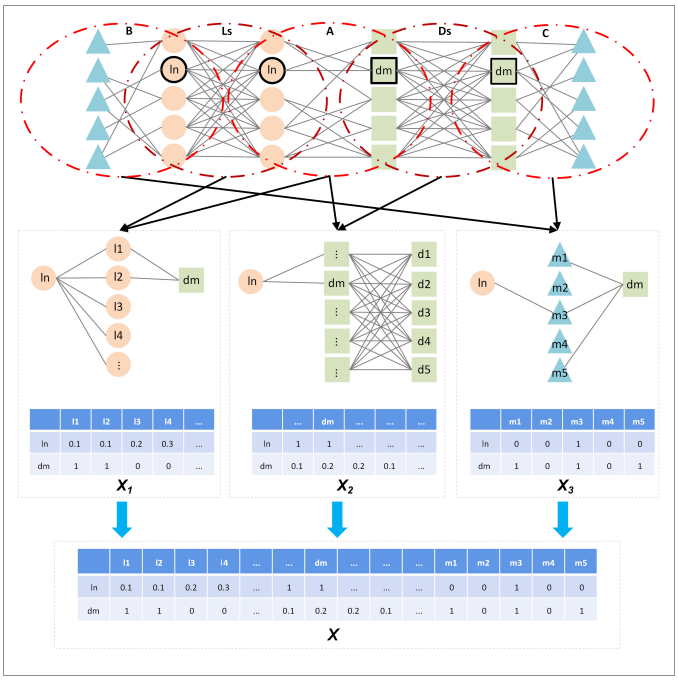

Fig. 2. Construction of the topology information matrix of ln and dm.

3 Deep Hierarchical Network

Deep neural networks can progressively learn deep feature informations along multiple layers. In this experiment, the whole framework of the model is composed of two modules. In the deep representation module, the adjacency information matrix and topology information matrix based on node attributes are constructed and input into the SAE and multi-scale ResNet to learn the dependencies of node attributes, respectively. And the outputs of networks are fused. In the stacked ensemble module, the high-level feature is input into the multiple shallow classifiers, and the probability score is determined by logistic regression. The overall model structure is shown in Fig 3.

3.1 Stacked autoencoder

In this study, stacked autoencoder [33] is introduced to capture the high-level features of adjacency information at multiple abstract levels. The features obtained by node2vec can be effectively embedded into the matrix as a linear projections through SAE to maximize the correlation between features. SAE training process includes two parts: encoder and decoder. Encoder is used to map input data to a potential representation, and decoder maps encoded features to reconstruct input data from the potential representation. A multi-layered SAE is built in this experiment to generate a deep learning architecture, both encoder and decoder consist of three fully connected layers. And the objective function of SAE parameters is optimized by $Adam$ algorithm using hierarchical learning method. The single-layer autoencoder is shown in the following formula. For the adjacency information matrix Y_1 (or Y_2), an encoder maps Y_1 to y with a nonlinear transform function $f(\cdot)$,

$$y = f(wY_1 + b) \tag{10}$$

where w and b are two parameters to be learned. A decoder is often used to reconstruct Y_1 from y from the following formula,

$$z = G\left(w^T y + b'\right) \tag{11}$$

where $G(\cdot)$ is a non-linear function. And the loss function is defined as follows,

$$l(Y_1, z) = ||Y_1 - z||^2 \tag{12}$$

Where, $l(\cdot)$is the function of minimizing the loss between Y_1 and z.

3.2 Deep multi-scale ResNet

Deep ResNet (Deep Residual Network, DRN) [34] is usually used in natural image classification. It can solve the problem of gradient disappearance in the training process of traditional neural network effectively. In this experiment, deep multi-scale ResNet is used to process the information of the topology information matrix, which is to capture the data features of the neighborhood by constructing the multi-scale CNN layer. Compared with the one-dimensional convolution layer, multi-scale CNN layers can obtain richer adjacent feature informations and local dependency informations. One-dimensional convolution calculated using the $Swish$ activation function is shown below,

$$Swish(x) = x \cdot (1 + \exp(-x))^{-1} \tag{13}$$

$$\tilde{o}_i = K * X_{i:i+k-1} = Swish\left(w \cdot X_{i:i+k-1} + b\right) \tag{14}$$

Where, $K\ (\in \mathbb{R}^k)$ denotes a convolution kernel, X denotes the topology information matrix, k is the range of the size of the convolution kernel determined along the matrix X, and b is a bias term. The residual block of DHNLDA performs the following computation:

$$O = Swish\left(f(X) + concatenate\left\{\tilde{o}_1, \tilde{o}_2, \tilde{o}_3\right\}\right) \tag{15}$$

Where, $f(\cdot)$ denotes one-dimensional convolutional layer and batch normalization (BN) layer, $concatenate\ \{\cdot\}$ denotes concatenate operation. Next, the average-pooling can be calculated as follows,

$$pooling(O)_j = average\left(O_j p, O_j p_{+1}, \ldots, O_j p{+}p - 1\right) \tag{16}$$

Where, j denotes the output position, and p is the pooling window size.

3.3 Stacked ensemble module

For the L-DA prediction, some classical shallow learning methods have also been used in research and practical application to solve classification problems [12]. Here, the high-level features obtained from the SAE and the multi-scale ResNet are fused, and RF [35], SVM [36], [37] and XGBoost [38] are used for training, which are respectively represented as RF^A, SVM^A and $XGBoost^A$. As expected, different classifiers have varying abilities to recognize categories. Considering the differences of classifiers, a stacked ensemble module is built to improve the prediction performance, which fuses multiple shallow classifiers to achieve higher performance. The stacked ensemble module is used to integrate the output of the three shallow learning models, then the logical regression (LR) is used as the fusion layer

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/TCBB.2021.3113326, IEEE/ACM Transactions on Computational Biology and Bioinformatics

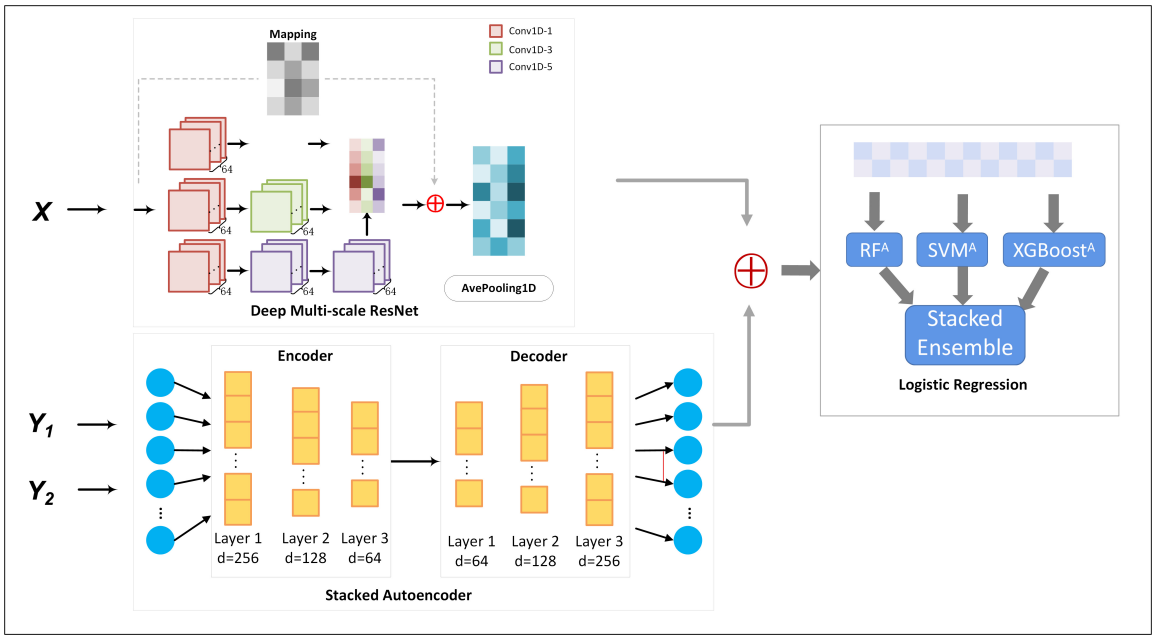

Fig. 3. The overall model structure.

to fuse the output of the three classifiers and obtained the fusion weight of each classifier. It is expressed as:

$$P_w(y = \pm 1 \mid x_{sum}) = \frac{1}{1 + e^{-yw^T x_{sum}}} \tag{17}$$

Where x_{sum} is the probability scores output from the three classifiers, y is the corresponding label of each association pair, and w is the weight vector of the three single classifiers. When the weight of each classifier of LR is judged to be the same, it will degenerate to an average strategy.

4 Experimental Results and Discussions

4.1 Experimental setup and evaluation metrics

The model was implemented on a GeForce RTX 2080 TI GPU 64G memory graphics card, based on TensorFlow 1.13 and Keras 2.0 architecture. Multi-scale ResNet layers of 64 channels with kernel sizes of 1, 3 and 5 were used as mapping connections for context feature vectors. After convolution, BN (Batch Normalization) was used for regularization. The hidden units of each layer of the stacked autoencoder layer were set as 128, 64, and 32. Meanwhile, dropout for each layer of the stacked autoencoder were added to avoid the risk of overfitting by randomly leaving out some neuron units. Next, the multi-scale ResNet layer and the stacked autoencoder layer were fused and input to the full connection layer of 32 hidden units, which was activated by $Swish$. All the layers in our network were trained simultaneously by $Adam$ with a batch size of 256, using the learning rate scheduler to control the learning rate. The details of parameters for our model as shown in Supplementary Table S4.

The experiment used the five-fold cross-validation to evaluate the performance of DHNLDA model and other L-DA prediction models. Here, TPR (True Positive Rate), FPR (False Positive Rate), AUPR (area under PR), PR (Precision-Recall) and AUC (area under ROC) were used to evaluate the performance of the model. In the process of five-fold cross-validation, all known L-DAs were taken as positive samples, and unknown associations with the same number as positive samples were selected as negative samples of the experiment according to random sampling. However, positive samples and negative samples are almost not balanced. Thus, we explore the performance of DHNLDA in the different sample ratios. Detailed negative sample analysis is shown in Supplementary Table S4. The average cross-validation AUC and AUPR were used to evaluate the ability of different L-DA prediction models. The formulas of false positive rate (FPR) and true positive rate (TPR) under different thresholds are as follows:

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN} \tag{18}$$

Where, TN, TP, FN, and FP denotes the number of true negative samples, true positive samples, false negative samples and false positive samples. The ROC curve was plotted by TPR and FPR, and the area under the curve (AUC) was calculated to evaluate the performance of different L-DA prediction models. PR can be obtained by calculating accuracy and recall rate, and its calculation formula is as follows:

$$precision = \frac{TP}{TP + FP}, recall = \frac{TP}{TP + FN} \tag{19}$$

Where, precision denotes the proportion of the correctly identified associated lncRNA-disease in all samples, while recall rate is the proportion of the truly associated lncRNA-disease in the total actually associated lncRNA-disease.

4.2 Performance evaluation using k-fold cross validation

To evaluate the performance of DHNLDA, k-fold cross-validations were employed. The AUC values were calculated to evaluate the performance of the model. In this experiment, two-fold, five-fold and ten-fold cross-validation were implemented to evaluate the performance. As shown

in Fig 4, five-fold cross-validation yielded good performance with average AUC of 0.975, which was higher than the average AUC of two-fold and ten-fold cross-validation. The result demonstrates that DHNLDA is effective to predict L-DAs on a large scale.

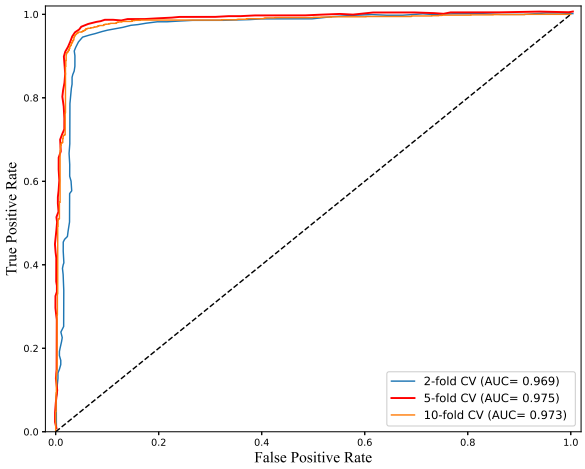

Fig. 4. AUC of DHNLDA in each fold cross-validation

4.3 Comparison between different learning strategies

In DHNLDA, a stacked ensemble module based on logistic regression was applied to learn the each weights of the three basic predictors (RF, SVM and XGBoost) for the final decision. To demonstrate the performance improvement of the stacked ensemble strategy, we compare it with single shallow classifier respectively, as shown in Table 1. On the L-DA benchmark datasets we constructed, the AUC of the LR-based stacked ensemble was 0.975, which was better than the other four classifiers. Experimental results show that the stacked ensemble strategy can integrate different predictors and improve the final performance by combining diversity, which is more powerful and flexible than single shallow classifier.

TABLE 1
THE AUCS AND AUPRS OF DIFFERENT LEARNING STRATEGIES.

Algorithm	AUC	AUPR
RF^A	0.972	0.965
SVM^A	0.971	0.967
$XGBoost^A$	0.969	0.963
DHNLDA	0.975	0.971

4.4 Comparison with other methods

In order to evaluate the performance of DHNLDA, we compared it with other existing methods such as SIMCLDA [39], Ping's Method [9], MFLDA [40], LDAP [41], CNNLDA [16], VADLP [20], DBNLDA [19] and RFLDA [12]. Table 2 shows the AUC and AUPR values of all prediction models. The ROC curves of different L-DA prediction models are shown in Fig 5.

As shown in Table 2 and Fig 5, the AUC of DHNLDA for all 412 diseases tested was 0.975, higher than that of all other methods except RFLDA. Its performance is better than SIMCLDA by 22.9%, Ping by 10.4%, MFLDA by 34.9%, LDAP by 11.2%, CNNLDA by 2.3%, VADLP by 1.9%,and DBNLDA by 1.5%, respectively. The AUC value of DHNLDA is closer to that of RFLDA (0.1% lower). However, DHNLDA is superior to all other methods in AUPR value, which the AUPR value of RFLDA is 0.779. These results indicate that DHNLDA can effectively predict the lncRNA-disease associations in unbalanced samples and could predict L-DAS better. The method in this paper combines the adjacency information between node attributes and the topology information between nodes, and integrates the stacked ensemble module for prediction, which makes the performance of DHNLDA better than that of other comparison methods.

TABLE 2
THE AUCS AND AUPRS OF DIFFERENT L-DA PREDICTION MODELS.

Algorithm	AUC	AUPR
SIMCLDA	0.746	0.095
Ping's Method	0.871	0.219
MFLDA	0.626	0.066
LDAP	0.863	0.166
CNNLDA	0.952	0.251
VADLP	0.956	0.449
DBNLDA	0.960	0.968
RFLDA	0.976	0.779
DHNLDA	0.975	0.971

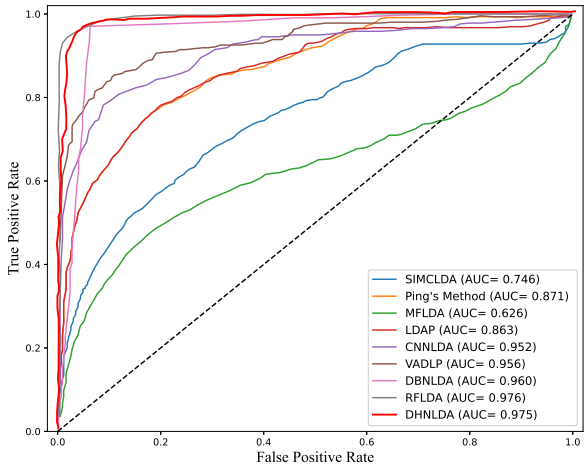

Fig. 5. The ROC curves of DHNLDA and other methods for all diseases

4.5 Case studies: stomach cancer, breast cancer and lung cancer

In order to further reflect DHNLDA ability to identify new disease-related genes, we constructed case studies of stomach cancer, breast cancer and lung cancer, sorted them according to association scores, and collected their top 15 candidate genes respectively.

TABLE 3
THE TOP 15 STOMACH CANCER, BREAST CANCER, LUNG CANCER RELATED CANDIDATE LNCRNAS.

Disease	Rank	LncRNA name	Source of verification	Rank	LncRNA name	Source of verification
stomach cancer	1	H19	C&D	9	BCYRN1	C&D
	2	HNF1A-AS1	C	10	HOTTIP	C&D
	3	BANCR	D	11	PANDAR	C&D
	4	MEG3	C&D	12	EWSAT1	D
	5	HULC	C&D	13	UCA1	C&D
	6	NEAT1	C&D	14	HCP5	unconfirmed
	7	XIST	C&D	15	MIR17HG	literature
	8	TUG1	C&D			
breast cancer	1	NEAT1	C&D	9	GAS5	C&D
	2	CCAT1	C&D	10	AFAP1-AS1	C
	3	DANCR	C&D	11	MIR17HG	D
	4	CDKN2B-AS1	C&D	12	LINC-ROR	C&D
	5	MALAT1	C&D	13	PCAT1	C&D
	6	MIR124-2HG	literature	14	TTTY15	D
	7	ZFAS1	C&D	15	NBAT1	C&D
	8	HOTTIP	C&D			
lung cancer	1	XIST	C&D	9	SPRY4-IT1	C&D
	2	H19	C&D	10	SOX2-OT	D
	3	ERICH1-AS1	C&D	11	ESRG	D
	4	HOTTIP	C&D	12	CDKN2B-AS1	literature
	5	LINC00675	unconfirmed	13	TINCR	C
	6	HULC	literature	14	ZFAS1	D
	7	PVT1	C&D	15	C5orf66-AS2	C&D
	8	GHET1	C			

'C' denotes the candidate lncRNA is included in the Lnc2Cancer database. 'D' denotes the candidate lncRNA is included in the LncRNADisease database. 'literature' denotess that the candidate lncRNA is included in the published literature.

In Table 3, 45 candidate genes associated with the three types of cancer are listed. Through LncRNADisease, Lnc2Cancer and records in the published literature to verify prediction of lncRNAs associated with stomach cancer, breast cancer and lung cancer. Lnc2Cancer is a lncRNA-cancer association database, which stores 4986 experimentally validated associations between 165 cancers and 1614 lncRNAs. The LncRNADisease database contains experimentally validated and latest method-predicted L-DAs.

First of all, as shown in Table 3, 14 of the top 15 genotypes predicted by DHNLDA related to stomach cancer have been verified by experimental data or published literature, and these databases have confirmed whether lncRNAs are associated with stomach cancer. In particular, it has been shown in the published literature that the expression of miR17HG in stomach cancer is negatively correlated with stomach cancer metastasis, thus promoting the dysregulation of cancer [42]. Secondly, among the top 15 candidate genes related to breast cancer given in Table 3, 12 candidate genes are included in the Lnc2Cancer database, and they are abnormally expressed in breast cancer. LncRNADisease included 13 candidates, which confirmed an association between these candidates and disease. Recent studies have showing that decreased expression of miR124-2HG enhances the proliferation of breast cancer cells targeting BECN1 [43], [44]. Finally, among all the top lncRNA candidate genes associated with lung cancer (Table 3), 10 of them have been confirmed by lncRNADisease. Among the top candidate genes, 9 cases are found in Lnc2Cancer, and their expression levels in lung cancer were significantly different from those in normal tissues. Two other candidate gene, supported by literature, demonstrated that HULC [45] and CDKN2B-AS1 [46] have dysregulations in non-small cell lung cancer and idiopathic pulmonary fibrosis, respectively. Fig 6 denotes the predicted association networks. According to the experiment results, 14, 15 and 14 of the top 15 candidate lncRNAs are verified to be related to stomach cancer, breast cancer and lung cancer. These results further illustrate that DHNLDA can effectively predict potential L-DAs.

In summary, among the top 45 lncRNAs predicted by DHNLDA to be associated with the three cancers, 43 were supported by experimental data from the Lnc2Cancer database, the LncRNADisease database or published literature. Therefore, these predictions indicate that the method proposed in this study can identify potential lncRNA-related diseases effectively.

4.6 Prediction of novel disease-related lncRNAs

After five-fold cross-validation and case studies to determine the performance of our model, we further applied the prediction model to the prediction of disease-related genes. The top 50 candidate lncRNAs predicted by our model are provided in Supplementary Table S5, which will provide valuable insights into the development of new L-DA predictions and further improve the performance of L-DA prediction models.

5 Conclusions

In this paper, a deep learning framework consisting of stacked autoencoder (SAE), multi-scale ResNet and stacked ensemble module, DHNLDA, is proposed to predict L-DAs. By integrating node attributes information and multiple biological data sources information, using SAE and multi-scale

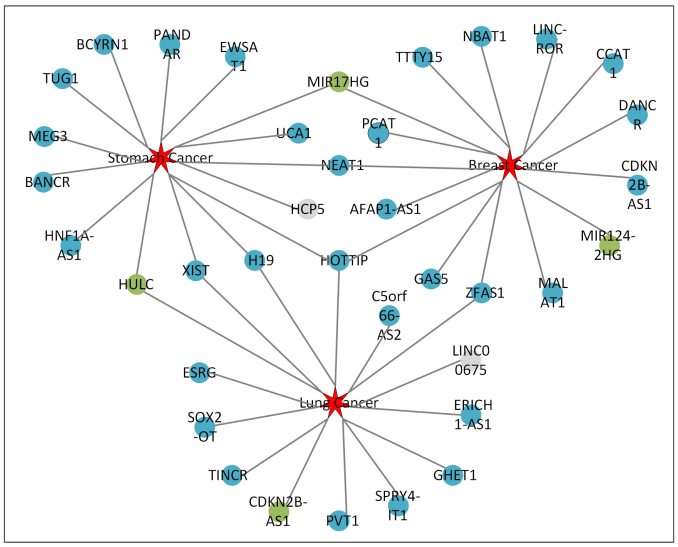

Fig. 6. Networks of the top 15 predicted associations for stomach cancer,breast cancer and lung cancer by DHNLDA. The red squares denote three cancers. The blue circles denote that the predicted associations and cancer-associated lncRNAs by experimental verification. The green circles denote that the predicted associations and cancer-associated lncRNAs by published literature, the gray circles denote that the predicted associations without experimental verification.

ResNet to learn deep representation, and combining the advantages of multiple shallow learning classifiers stacked ensemble, our method achieves good classification results. Case studies of three diseases confirm the ability of our model to predict potential disease-related NCDs. At present, we are studying the interaction of three gene molecules (lncRNA, miRNA and disease), while the influence of other molecules (such as proteins and drugs) on the association network is still incomplete. With the continuous improvement of a variety of biomolecule databases, studies on multi-molecular interactions also appeared. Next, the interaction between lncRNAs and drugs will become the main content of our study.

Acknowledgments

This research was supported by Dalian Young Science and Technology Star Project (No.2020RQ059), the National Natural Science Foundation of China (No.61701073) and Scientific Research Project of Education Department of Liaoning Province (No.LJKZ0028). We thank XuanPing and her colleagues for providing very valuable help and guidance for our research. The authors would like to thank all the editors and anonymous reviewers for their constructive advices.

Supporting Information

The following supporting information is available as part of the online article:

Table S1. The detailed values of two-fold are shown in Supplementary Table S1.

Table S2. The detailed values of five-fold are shown in Supplementary Table S2.

Table S3. The detailed values of ten-fold are shown in Supplementary Table S3.

Table S4. The experimental materials are shown in in Supplementary Table S4.

Table S5. The top 50 candidate lncRNAs predicted by our model are provided in Supplementary Table S5.

Abbreviations

L-MS: lncRNA-miRNA similarity; M-DS: miRNA-disease similarity; L-DA: lncRNA-disease association; L-DAs: lncRNA-disease associations; L-MI: lncRNA-miRNA interaction; L-MIs: lncRNA-miRNA interactions; M-DA: miRNA-disease association; M-DAs: miRNA-disease associations;

References

[1] Pertea M. The human transcriptome: an unfinished story[J]. Genes, 2012, 3(3): 344-360; https://doi.org/10.3390/genes3030344.

[2] Ahadi A. Functional roles of lncRNAs in the pathogenesis and progression of cancer[J]. Genes & Diseases, 2020; https://doi.org/10.1016/j.gendis.2020.04.009.

[3] Li Y, Shan G, Teng Z Q, et al. Non-coding RNAs and human diseases[J]. Frontiers in Genetics, 2020, 11: 523.

[4] Lemos A E G, da Rocha Matos A, Ferreira L B, et al. The long non-coding RNA PCA3: an update of its functions and clinical applications as a biomarker in prostate cancer[J]. Oncotarget, 2019, 10(61): 6589; 10.18632/oncotarget.27284.

[5] Liu Y, Zong Z H, Guan X, et al. The role of long non-coding RNA PCA3 in epithelial ovarian carcinoma tumorigenesis and progression[J]. Gene, 2017, 633: 42-47; https://doi.org/10.1016/j.gene.2017.08.027.

[6] Lim K H, Yang S, Kim S H, et al. Discoveries for Long Non-Coding RNA Dynamics in Traumatic Brain Injury[J]. Biology, 2020, 9(12): 458; https://doi.org/10.3390/biology9120458.

[7] Chen R, Xu X, Huang L, et al. The regulatory role of long noncoding RNAs in different brain cell types involved in ischemic stroke[J]. Frontiers in Molecular Neuroscience, 2019, 12: 61; https://doi.org/10.3389/fnmol.2019.00061.

[8] Chen X, Yan C C, Zhang X, et al. Long non-coding RNAs and complex diseases: from experimental results to computational models[J]. Briefings in bioinformatics, 2017, 18(4): 558-576; https://doi.org/10.1093/bib/bbw060.

[9] Ping P, Wang L, Kuang L, et al. A novel method for lncRNA-disease association prediction based on an lncRNA-disease association network[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2018, 16(2): 688-693; 10.1109/TCBB.2018.2827373.

[10] Chen X, Yan C C, Luo C, et al. Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity[J]. Scientific reports, 2015, 5: 11338; https://doi.org/10.1038/srep11338.

[11] Sun J, Shi H, Wang Z, et al. Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network[J]. Molecular BioSystems, 2014, 10(8): 2074-2081; 10.1039/c3mb70608g.

[12] Yao D, Zhan X, Zhan X, et al. A random forest based computational model for predicting novel lncRNA-disease associations[J]. BMC bioinformatics, 2020, 21: 1-18.

[13] Gu C, Liao B, Li X, et al. Global network random walk for predicting potential human lncRNA-disease associations[J]. Scientific reports, 2017, 7(1): 1-11; https://doi.org/10.1038/s41598-017-12763-z.

[14] Wang Y, Juan L, Peng J, et al. LncDisAP: a computation model for LncRNA-disease association prediction based on multiple biological datasets[J]. BMC bioinformatics, 2019, 20(16): 1-11; https://doi.org/10.1186/s12859-019-3081-1.

[15] Sumathipala M, Maiorino E, Weiss S T, et al. Network diffusion approach to predict lncRNA disease associations using multi-type biological networks: LION[J]. Frontiers in physiology, 2019, 10: 888; https://doi.org/10.3389/fphys.2019.00888.

[16] Xuan P, Cao Y, Zhang T, et al. Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncRNA genes. Front Genet 2019;10:416.https://doi.org/10.3389/fgene.2019.00416

[17] Xuan P, Pan S, Zhang T, et al. Graph convolutional network and convolutional neural network based method for predicting lncRNA-disease associations[J]. Cells, 2019, 8(9): 1012; https://doi.org/10.3390/cells8091012.

[18] Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks[J]. Pattern Recognition, 2018, 77: 354-377; https://doi.org/10.1016/j.patcog.2017.10.013.

[19] Madhavan M. Deep Belief Network based representation learning for lncRNA-disease association prediction[J]. arXiv preprint arXiv:2006.12534, 2020.

[20] Nan Sheng, Hui Cui, Tiangang Zhang, Ping Xuan, Attentional multi-level representation encoding based on convolutional and variance autoencoders for lncRNA–disease association prediction, Briefings in Bioinformatics, , bbaa067, https://doi.org/10.1093/bib/bbaa067.

[21] Zhang J, Jiang Z, Hu X, et al. A Novel Graph Attention Adversarial Network for Predicting Disease-related Associations[J]. Methods, 2020; https://doi.org/10.1016/j.ymeth.2020.05.010.

[22] Fu GY, Wang J, Domeniconi C, Yu GX. Matrix factorization-based data fusion for the prediction of lncRNA–disease associations. Bioinformatics. 2018;34(9):1529–37.https://doi.org/10.1093/bioinformatics/btx794

[23] Chen G, Wang ZY, Wang DQ, Qiu CX, Liu MX, Chen X, Zhang QP, Yan GY, Cui QH. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 2013;41(D1):D983–6.https://doi.org/10.1093/nar/gks1099

[24] Ning SW, Zhang JZ, Wang P, Zhi H, Wang JJ, Liu Y, Gao Y, Guo MN, Yue M, Wang LH, Li X. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. Nucleic Acids Res. 2016;44(D1):D980–5. https://doi.org/10.1093/nar/gkv1094

[25] Lu ZY, Coben KB, Hunter L. GeneRIF quality assurance as summary revision. Biocomputing. 2007;2007:269–80.https://doi.org/10.1142/9789812772435_0026

[26] Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. 2014;42(D1):D92–7. https://doi.org/10.1093/nar/gkt1248

[27] Li Y, Qiu CX, Tu J, Geng B, Yang JC, Jiang TZ, Cui QH. HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. Nucleic acids Res. 2014;42(D1):D1070–4. https://doi.org/10.1093/nar/gkt1023

[28] Fan, W Shang, J Li, F. et al. IDSSIM: an lncRNA functional similarity calculation model based on an improved disease semantic similarity method. BMC Bioinformatics 21, 339 (2020). https://doi.org/10.1186/s12859-020-03699-9

[29] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks,in: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 855–864.

[30] Peng J, Guan J, Shang X. Predicting Parkinson's disease genes based on node2vec and autoencoder[J]. Frontiers in genetics, 2019, 10: 226.

[31] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14). Association for Computing Machinery, New York, NY, USA, 701–710. DOI:https://doi.org/10.1145/2623330.2623732

[32] Xuan P, Sheng N, Zhang T, Liu Y, Guo Y. CNNDLP: A Method Based on Convolutional Autoencoder and Convolutional Neural Network with Adjacent Edge Attention for Predicting lncRNA-Disease Associations. Int J Mol Sci. 2019 Aug 30;20(17):4260. PMID: 31480319; PMCID: PMC6747450.doi: 10.3390/ijms20174260.

[33] Wang L, You Z H, Chen X, et al. A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network[J]. Journal of Computational Biology, 2018, 25(3): 361-373; https://doi.org/10.1089/cmb.2017.0135.

[34] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C].European conference on computer vision. Springer, Cham, 2016: 630-645; https://doi.org/10.1007/978-3-319-46493-0_38.

[35] Svetnik V . Random forest: a classification and regression tool for compound classification and QSAR modeling.[J]. Journal of Chemical Information & Computer Sciences, 2003, 43.https://doi.org/10.1021/ci034160g

[36] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.https://doi.org/10.1023/A:1022627411411

[37] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACMTrans Intell Syst Technol. 2011;2(3):1–27.https://doi.org/10.1145/1961189.1961199

[38] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. DOI:https://doi.org/10.1145/2939672.2939785

[39] Chengqian Lu, Mengyun Yang, Feng Luo, Fang-Xiang Wu, Min Li, Yi Pan, Yaohang Li, Jianxin Wang, Prediction of lncRNA–disease associations based on inductive matrix completion, Bioinformatics, Volume 34, Issue 19, 01 October 2018, Pages 3357–3364, https://doi.org/10.1093/bioinformatics/bty327

[40] Fu G, Wang J, Domeniconi C, et al. Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. Bioinformatics 2017; 34(9): 1529–37.https://doi.org/10.1093/bioinformatics/btx794

[41] Lan W, Li M, Zhao K, et al. LDAP: a web server for lncRNA-disease association prediction. Bioinformatics 2017; 33(3):458–60.https://doi.org/10.1093/bioinformatics/btw639

[42] Bahari, F., Emadi-Baygi, M., and Nikpour, P. (2015). miR-17-92 host gene,uderexpressed in gastric cancer and its expression was negatively correlated with the metastasis. Ind. J. Cancer 52, 22–25. doi: 10.4103/0019-509X.175605

[43] Huang R, Zhang Y, Han B, et al. Circular RNA HIPK2 regu-lates astrocyte activation via cooperation of autophagy and ER stress by targeting MIR124-2HG. Autophagy 2017; 13(10):1722–41.

[44] Lv X B, Jiao Y, Qing Y, et al. miR-124 suppresses multiple steps of breast cancer metastasis by targeting a cohort of pro-metastatic genes in vitro[J]. Chinese journal of cancer, 2011, 30(12): 821. doi: 10.5732/cjc.011.10289

[45] Zhang J, Lu S, Zhu J F, et al. Up-regulation of lncRNA HULC predicts a poor prognosis and promotes growth and metastasis in non-small cell lung cancer[J]. Int J Clin Exp Pathol, 2016, 9(12): 12415-12422.

[46] Du Y, Hao X, Liu X. Low expression of long noncoding RNA CDKN2B-AS1 in patients with idiopathic pulmonary fibrosis predicts lung cancer by regulating the p53-signaling pathway. Oncol Lett. 2018 Apr;15(4):4912-4918. doi: 10.3892/ol.2018.7910. Epub 2018 Jan 31. PMID: 29541247; PMCID: PMC5835920.

Fansen Xie received the BA degree in computer science and technology from Huanghuai University in 2015. He is currently pursuring the MA degree in computer science and engineering in the Dalian minzu University. His current research interest is bioinformatics.

Ziqi Yang received the BA degree in computer science and technology from Inner Mongolia Normal University in 2015. She is currently pursuring the MA degree in computer science and engineering in the Dalian minzu University. Her current research interest is bioinformatics.

Jinmiao Song received the M.S. degree in computer technology from North Minzu University in 2014. He is currently pursuing the Ph.D. degree in computer science and technology in the University of Xinjiang. His research interest includes bioinformatics and intelligence computing.

Qiguo Dai received the B.S., the M.S. and the Ph.D. degrees in computer science and technology from Hubei University of Automotive Technology in 2006, Beijing University of Technology in 2010 and Harbin Institute of Technology in 2015 respectively. Currently, he is an associate professor in the school of computer science and engineering, Dalian Minzu University. His research interests include bioinformatics and data mining.

Xiaodong Duan received the B.S. degrees in computer science and technology from Nankai University, Tianjin, P.R. China in 1985, and received the M.S. and Ph.D. degrees in applied mathematics and computer software and theory from Northeastern University, Shenyang, P.R. China in 1988 and 2001, respectively. Currently, he is a professor in the school of computer science and engineering, Dalian Minzu University. His research interests include pattern recognition and data mining.