PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2022, 49(7): 1369~1380
www.pibb.ac.cn

NELDA: Prediction of LncRNA−disease Associations With Network Embedding*

LI Wei-Na, FAN Xiao-Nan, ZHANG Shao-Wu**
(*School of Automation, Key Laboratory of Information Fusion Technology of Ministry of Education, Northwestern Polytechnical University, Xi'an* 710072, China)

Abstract **Objective** Long non-coding RNAs (lncRNAs) participate in a variety of vital biological processes and closely relate with various human diseases. The prediction of lncRNA-disease associations can help to understand the mechanisms of human disease at the molecular level, and also contribute to diagnosis and treatment of diseases. Most existing methods of predicting the lncRNA-disease associations ignore the deep embedding features hiding in lncRNA/disease network topological structures. Moreover, randomly selecting the negative samples will affect the robustness of predictors. **Methods** Here we first set up a high quality dataset by using an effective strategy to select the negative samples (*i.e.*, pairs of non lncRNA-disease association) with relatively higher quality instead of randomly selecting the negative samples, then proposed a novel method (called NELDA) to predict the potential lncRNA-disease associations by building 4 deep auto-encoder models to learn the low dimensional network embedding features from the lncRNA/disease similarity networks, and lncRNA-disease association network, respectively. NELDA takes the lncRNA/disease similarity network embedding features as the input of one support vector machine (SVM) classifier, and the lncRNA/disease association network embedding features as the input of another SVM classifier. The prediction results of these two SVM classifiers are fused by the weighted average strategy to obtain the final prediction results. **Results** In 10-fold cross-validation (10 CV) test, the AUC of NELDA achieves 0.982 7 on high quality dataset, which is 0.062 7 and 0.020 7 higher than that of other two state-of-the-art methods of LDASR and LDNFSGB, respectively. In the case studies of stomach cancer and breast cancer, 29/40 (72.5%) novel predicted lncRNAs associated with stomach and breast cancers are supported by recent literatures and public datasets. **Conclusion** These experimental results demonstrate that NELDA is a superior method for predicting the potential lncRNA-disease associations. It has the ability to discover the new lncRNA-disease associations.

Key words lncRNA-disease association, network embedding, deep auto-encoder, high quality negative samples
DOI: 10.16476/j.pibb.2021.0132

Accumulated evidences reveal that more than 70% of the human genome can be transcribed, but only less than 2% of the genome is able to be translated into proteins[1-2] , and the RNAs which do not encode proteins exist in the form of non-coding RNAs[3]. Long non-coding RNAs (lncRNAs) with the length more than 200 nucleotides account for a large proportion of non-coding RNAs[4], participating in a variety of vital biological processes[5-6]. Multiple lines of evidence have linked dysregulations and mutations of lncRNAs to diverse human diseases[7], such as bladder cancer[8], lung cancer[9], gastric cancer[10], and breast cancer[11]. For example, upregulated MALAT-1 contributes to bladder cancer cell migration by inducing epithelial-to-mesenchymal transition[8]. LncRNA H19 expression was elevated in the lung cancer cell lines and tissues. H19 promotes lung cancer metastasis and proliferation by inhibiting the function of miR-200a[9]. Therefore, identifying the potential human disease-related lncRNAs will help to understand the mechanisms of human disease at the molecular level, and also provide the potential biomarkers for human disease diagnosis and treatment[12].

* This work was supported by grants from The National Natural Science Foundation of China (61873202, 62173271).
** Corresponding author.
Tel: 86-29-88431308, E-mail: zhangsw@nwpu.edu.cn
Received: May 10, 2021 Accepted: September 2, 2021

In recent years, computational methods have been developed to predict the lncRNA-disease associations by utilizing diversity of biological data, which can provide the candidate disease-associated lncRNAs for biological experiment verification, reducing the time-consuming and costs in biological experiments[13-15]. Existing computational methods for predicting lncRNA-disease associations can mainly be divided into network-based methods[16-25] and machine learning-based methods[12, 26-31]. The network-based methods, such as RWRlncD[18] and IDHI-MIRW[16], usually constructed the lncRNA similarity network, or lncRNA-disease heterogeneous network by integrating one or more of lncRNA similarity, disease similarity, lncRNA-protein interactions, lncRNA-miRNA interactions, disease-protein associations, disease-miRNA associations, the known lncRNA-disease associations and so on, then adopted the random walk, flow propagation and other algorithms to predict the lncRNA-disease associations. For example, RWRlncD[18] implemented the random walk with restart (RWR) algorithm to predict the potential lncRNA-disease associations by constructing the lncRNA functional similarity network based on the known lncRNA-disease associations. IDHI-MIRW[16] constructed a large-scale lncRNA-disease heterogeneous network by integrating lncRNA expression profiles, lncRNA-miRNA interactions, lncRNA-protein interactions, disease ontology, disease-miRNA associations, disease-protein associations and known lncRNA-disease associations, then used RWR to predict the potential lncRNA-disease associations. However, because most of the functions and mechanisms of lncRNAs are still unclear, the lncRNA similarity or lncRNA-disease heterogeneous network built in existing network-based methods would be noisy and information missing.

The machine learning-based methods, such as LRLSLDA[12], LDAP[26], LDNFSGB[27] and LDASR[28], usually used lncRNA similarities, disease similarities, known lncRNA-disease associations and other information to represent lncRNA-disease pairs, then used the machine learning algorithms, such as Laplacian Regularized Least Squares, Bagging SVM, gradient boosting and rotation forest, to predict the lncRNA-disease associations. For example, LRLSLDA[12] adopted the Laplacian Regularized Least Squares to predict the lncRNA-disease associations in the semi-supervised learning framework by using the information of lncRNA expression profiles and known lncRNA-disease associations. LDAP[26] employed 2 lncRNA similarity methods to calculate the similarities between lncRNAs, 5 disease similarity methods to calculate the similarities between diseases, and then utilized the Karcher mean of matrices to fuse similarity matrices of lncRNA and disease, respectively, and finally used the bagging SVM classifier to predict the potential lncRNA-disease associations. However, these two methods ignore the deep embedding features. LDNFSGB[27] and LDASR[28] built an auto-encoder model to learn the hidden abstract representation for lncRNA-disease pairs based on lncRNA and disease similarities, then adopted the rotation forest and gradient boosting algorithm to predict the potential lncRNA-disease associations, respectively. Although these two methods learn the hidden abstract representation for lncRNA-disease pairs, they also ignore the deep embedding features which preserve the network structures.

In this work, we proposed a novel machine learning-based method (called NELDA) to predict the potential lncRNA-disease associations. Based on the known lncRNA-disease associations, lncRNA expression profiles and disease ontology, NELDA first constructs 3 networks of the lncRNA-disease association network, the lncRNA similarity network and the disease similarity network. Then, 4 deep auto-encoder models are built to extract the lncRNA similarity network embedding, disease similarity network embedding, lncRNA association network embedding, and disease association network embedding from lncRNA similarity network, disease similarity network, and lncRNA-disease association network, respectively. Based on the lncRNA-disease similarity network embedding (*i. e.*, concatenating lncRNA similarity network embedding and disease similarity network embedding) and lncRNA-disease association network embedding (*i. e.*, concatenating lncRNA association network embedding and disease association network embedding), NELDA designs 2 support vector machine (SVM) classifiers to separately predict the lncRNA-disease associations. The final prediction result of NELDA is obtained by fusing the results of 2 SVM classifiers with the weighted average strategy. In order to generate the robust prediction results, we also set up a higher

quality dataset by choosing the higher quality non lncRNA-disease association samples instead of randomly selecting the non lncRNA-disease association samples. The performance of NELDA in 10-fold cross validation (10 CV) test shows that NELDA is superior to other 2 methods of LDASR and LDNFSGB for predicting the lncRNA-disease associations. The case studies of stomach and breast cancers indicate that NELDA has the power to predict the novel lncRNA-disease associations, and it can provide the candidates for further biological experimental validations.

1 Methods

1.1 Datasets

To effectively validate the performance of NELDA, we first set up a higher quality lncRNA-disease association dataset $D_{rel} = D^{+} \cup D_{rel}^{-}$. For constructing the lncRNA-disease association sample set (*i. e.*, D^{+}), we downloaded the known lncRNA-disease associations from the previous work[16], which collected the known lncRNA-disease associations from LncRNAdisease[32], Lnc2Cancer[33] and GeneRIF[34], then deleted the lncRNAs/diseases with less than two associations. We finally obtained 1 824 known lncRNA-disease associations to form the positive sample set D^{+}, which contains 151 lncRNAs and 233 diseases. For constructing the high-quality non lncRNA-disease association sample set (*i. e.*,D_{rel}^{-}), instead of randomly pairing the lncRNAs and diseases, we took the idea from literature[35] to select the high-quality non lncRNA-disease association samples (*i. e.*, negative samples). The procedures of constructing D_{rel}^{-} are as follows:

(1) Randomly pairing the lncRNAs and diseases in the 1 824 known lncRNA-disease associations. Then, based on the known lncRNA-disease associations and the disease semantic similarity, calculating the association score S for each randomly paired lncRNA-disease pair without any association evidence.

$$S\left(l_i,d_j\right) = \frac{S'\left(l_i,d_j\right) - S'_{\min}}{S'_{\max} - S'_{\min}} \tag{1}$$

$$S'\left(l_i,d_j\right) = \sum_{k=1}^{N_d} A_{LD}\left(i,k\right) S_{D1}\left(j,k\right) \tag{2}$$

where, $A_{LD} \in R^{N_l \times N_d}$ is the adjacency matrix of the known lncRNA-disease associations; N_l and N_d are the number of lncRNAs and diseases in D^{+}, respectively; if the pair of the lncRNA l_i and the disease d_k belongs to D^{+}, $A_{LD}(i,k) = 1$; otherwise, $A_{LD}(i,k) = 0$. $S_{D1}(j,k)$ is the disease semantic similarity between j-th disease d_j and k-th disease d_k, which can be calculated by "doSim" function from R package "DOSE" according to the structure of the directed acyclic graph in Disease Ontology[36-37]. $S'_{\max}$ and $S'_{\min}$ are the maximum and minimum S' of all randomly paired lncRNA-disease pairs without association evidences, respectively.

(2) According to the association scores S, ranking all randomly paired lncRNA-disease pairs without association evidences (*i. e.*, unconfirmed lncRNA-disease pairs) in ascending order.

(3) Randomly selecting a certain number of unconfirmed lncRNA-disease pairs with association score S less than 0.02 to form the non lncRNA-disease association set D_{rel}^{-}, in which the number of non lncRNA-disease association pairs is same as the number of lncRNA-disease association pairs in D^{+}. So far, we build a higher quality lncRNA-disease association dataset $D_{rel} = D^{+} \cup D_{rel}^{-}$ with 1 824 known lncRNA-disease associations and 1 824 high-quality non lncRNA-disease association pairs.

In addition, to verify the strategy effectiveness of constructing the high-quality non lncRNA-disease association pairs, we also constructed another dataset $D_{ran} = D^{+} \cup D_{ran}^{-}$ by randomly pairing the lncRNAs and diseases in D^{+}, removing the known lncRNA-disease association pairs and selecting the same number of the non lncRNA-disease association pairs as D^{+} to form the non lncRNA-disease association set D_{ran}^{-}. The distributions of the association score of positive samples in D^{+}, the high-quality non lncRNA-disease association samples in D_{rel}^{-} and the non lncRNA-disease association samples in D_{ran}^{-} are shown in Figure S1 in **Supplementary**.

1.2 Overview of NELDA algorithm

NELDA algorithm mainly consists of the following 3 phases: (1) Constructing 3 networks of the lncRNA similarity network, the disease similarity network and the lncRNA-disease association network based on the lncRNA expression similarity, lncRNA Gaussian interaction profile kernel similarity, disease semantic similarity, disease Gaussian interaction profile kernel similarity, and known the lncRNA-disease associations. (2) Building the deep auto-encoder models to extract lncRNA similarity network embedding, disease similarity network embedding,

lncRNA association network embedding, disease association network embedding, respectively. (3) Generating two representation vectors for each lncRNA-disease pair by concatenating the lncRNA similarity network embedding features and the disease similarity network embedding features, concatenating the lncRNA association network embedding features and the disease association network embedding features, respectively. Then 2 representation vectors are inputted into 2 SVM classifiers for predicting the lncRNA-disease associations, respectively. In the end, the outputs of 2 SVM classifiers are fused by the weighted average strategy to get the final prediction results. Figure 1 is the flowchart of our NELDA.

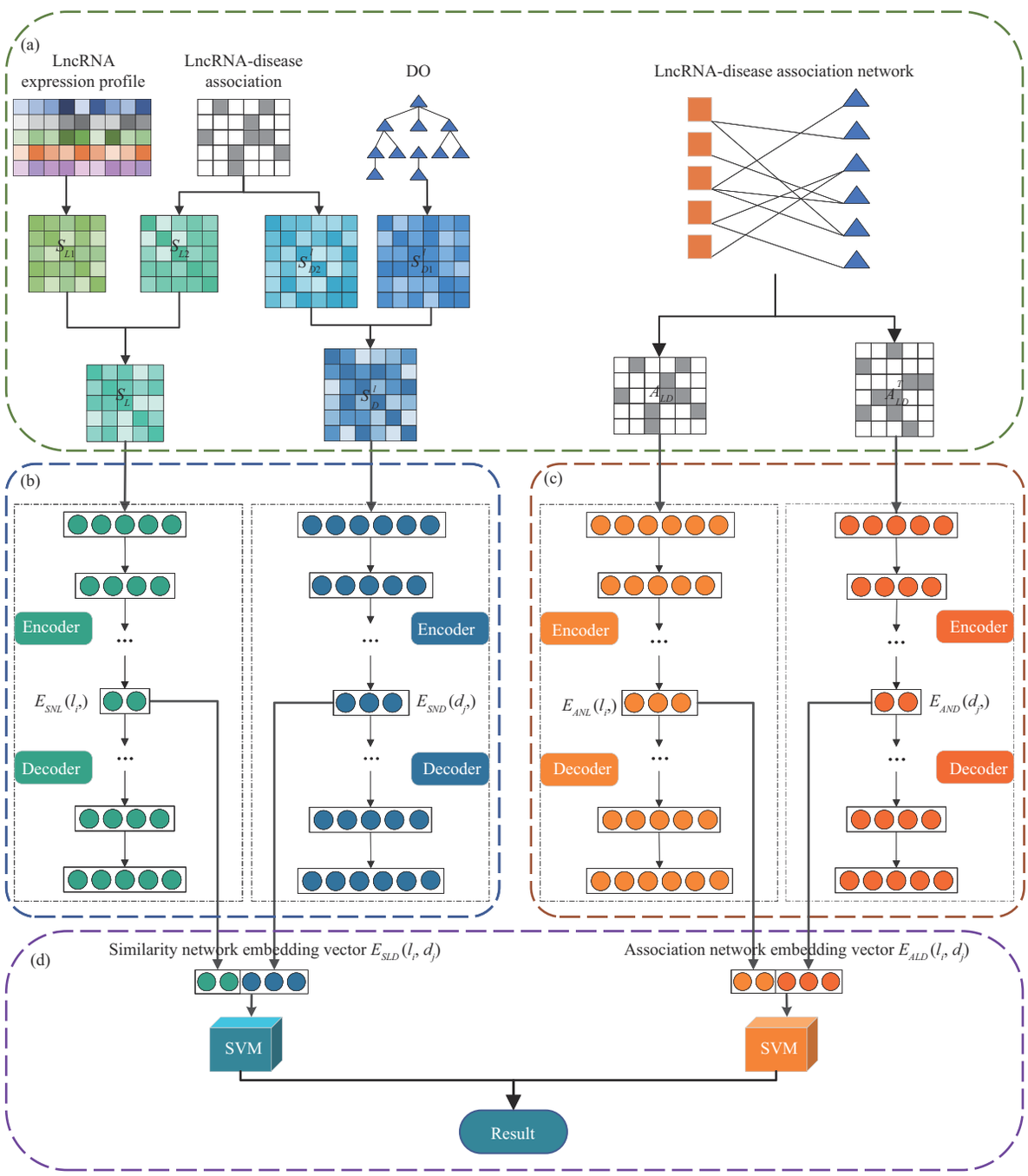

Fig. 1 Flowchart of NELDA

(a) Constructing the lncRNA similarity network, the disease similarity network and the lncRNA-disease association network. (b) Extracting the lncRNA similarity network embedding and disease similarity network embedding based on lncRNA expression profile, known lncRNA-disease associations and Disease Ontology by building 2 deep auto-encoder models. (c) Extracting lncRNA association network embedding and disease association network embedding from lncRNA-disease association network by building 2 deep auto-encoder models. (d) Constructing 2 support vector machine classifiers, and using the weighted average strategy to get the final prediction results.

1.3 LncRNA/disease similarity networks

We generated 2 lncRNA similarity matrices S_{L1} and S_{L2}. $S_{L1} \in R^{N_l \times N_l}$ is a lncRNA expression similarity matrix generated by calculating the absolute value of Spearman correlation coefficient of any lncRNA pair from their expression profiles, which were downloaded from the previous work[16]. $S_{L2} \in R^{N_l \times N_l}$ is a lncRNA Gaussian interaction profile kernel similarity matrix [12] (calculated with Equation (1) in **Supplementary**). Two lncRNA similarity matrices of S_{L1} and S_{L2} are combined to generate the lncRNA integrated similarity matrix$S_L \in R^{N_l \times N_l}$ for constructing the lncRNA similarity network. The integrated similarity between lncRNA l_i and lncRNA l_j is defined as:

$$S_L(i,j) = \begin{cases} \dfrac{S_{L1}(i,j) + S_{L2}(i,j)}{2} & if\ l_i \text{ and } l_j \text{ have the expression profiles and } S_{L2}(i,j) \neq 0 \\ S_{L1}(i,j) & if\ l_i \text{ and } l_j \text{ have the expression profiles and } S_{L2}(i,j) = 0 \\ S_{L2}(i,j) & \text{otherwise} \end{cases} \quad (3)$$

where, $S_{L1}(i,j)$ and $S_{L2}(i,j)$ are the expression similarity and the Gaussian interaction profile kernel similarity between lncRNA l_i and lncRNA l_j, respectively.

In addition, we also generated the disease semantic similarity matrix S_{D1} and the disease Gaussian interaction profile kernel similarity matrix $S_{D2} \in R^{N_d \times N_d}$[12] (Equation (3) in **Supplementary**).

Two disease similarity matrixes S_{D1} and S_{D2} are combined to generate the disease integrated similarity matrix$S_D \in R^{N_d \times N_d}$ for constructing the disease similarity network. Considering that there is semantic similarity between any two diseases in the known lncRNA-disease associations, the integrated similarity between disease d_i and disease d_j is defined as:

$$S_D(i,j) = \begin{cases} \dfrac{S_{D1}(i,j) + S_{D2}(i,j)}{2} & if\ S_{D2}(i,j) \neq 0 \\ S_{D1}(i,j) & if\ S_{D2}(i,j) = 0 \end{cases} \quad (4)$$

1.4 LncRNA/disease embedding features

1.4.1 Extracting the lncRNA/disease similarity network embedding

Network embedding can learn the low-dimensional representations of vertexes in networks, aiming to capture and preserve the network structure. For example, the structural deep network embedding (SDNE) is able to capture the highly non-linear network structure[38]. Inspired by SDNE in which the deep auto-encoder model is used to preserve the global network structure, we first built a deep auto-encoder model to extract the lncRNA similarity network embedding matrix E_{SNL} from the lncRNA similarity network. In deep auto-encoder model, the encoder consists of multiple non-linear functions that map the input data to the representation space, and the decoder consists of multiple non-linear functions that map the representations in representation space to reconstruction space[38].

We inputted the lncRNA integrated similarity matrix S_L to the deep auto-encoder model, that is to say, for every lncRNA l_i, the i-th row $S_L(i,:)$ in lncRNA integrated similarity matrix S_L is used as the input vector. The output of encoder, h_i^K, is the final low-dimensional representation of lncRNA l_i, that is, the lncRNA similarity network embedding $E_{SNL}(l_i,:)$ of lncRNA l_i. The decoder is built to reconstruct the input. The detail of the deep auto-encoder model[38] is shown in the supplementary file and the loss function is shown as follows:

$$L = L_{con} + \alpha L_{reg} \quad (5)$$

$$L_{con} = \sum_{i=1}^{N_l} \left\| S_L(i,:) - \hat{S}_L(i,:) \right\|_2^2 = \left\| S_L - \hat{S}_L \right\|_F^2 \quad (6)$$

where L_{reg} is a L_2-norm regularization term, which is used to prevent overfitting[38].

Similarly, we also built another deep auto-encoder model to extract the disease similarity network embedding matrix E_{SND} from the disease integrated similarity network.

1.4.2 Extracting the lncRNA/disease association network embedding

According to the process of extracting the lncRNA/disease similarity network embedding, we built other 2 deep auto-encoder models to extract the lncRNA association network embedding matrix E_{ANL} and the disease association network embedding matrix E_{AND} from the lncRNA-disease association network A_{LD}, respectively. Considering that the lncRNA-disease association network is sparse, we redefined the reconstruction loss function to alleviate the sparse issue by imposing more penalty to the reconstruction error of the non-zero elements[38]. For examples, the

reconstruction loss function for extracting lncRNA association network embedding is defined as follow[38]:

$$L_{con} = \sum_{i=1}^{N_l} \left\| \left(A_{LD}(i,:) - \hat{A}_{LD}(i,:) \right) \odot b_i \right\|_2^2 = \left\| \left(A_{LD} - \hat{A}_{LD} \right) \odot B \right\|_F^2 \quad (7)$$

where $A_{LD}(i,:)$ is the *i-th* row of lncRNA-disease association matrix, $\hat{A}_{LD}(i,:)$ is the *i-th* row of reconstruction matrix of lncRNA-disease association by decoder. If $A_{LD}(i,j) = 0$, then $b(i,j) = 1$, else $b(i,j) = \beta > 1$, $B = \{ b(i,j) \} \in R^{N_l \times N_d}$. $\odot$represents Hadamard product.

Referring to SDNE method, we used the Deep Belief Network model to pretrain the parameters of deep auto-encoder models for extracting lncRNA association network embedding and disease association network embedding.

1.5 Network embedding feature combination and decision−level weighted fusion

For each lncRNA-disease pair l_i - d_j, we generated 2 embedding vectors to represent it. One is the similarity network embedding vector $E_{SLD}(l_i,d_j)$ formed by concatenating the lncRNA similarity network embedding vector $E_{SNL}(l_i,:)$ and disease similarity network embedding vector $E_{SND}(d_j,:)$. Another is the association network embedding vector $E_{ALD}(l_i,d_j)$ formed by concatenating the lncRNA association network embedding vector $E_{ANL}(l_i,:)$ and disease association network embedding vector $E_{AND}(d_j,:)$.

$$E_{SLD}(l_i,d_j) = [E_{SNL}(l_i,:), E_{SND}(d_j,:)] \quad (8)$$

$$E_{ALD}(l_i,d_j) = [E_{ANL}(l_i,:), E_{AND}(d_j,:)] \quad (9)$$

$E_{SLD}(l_i,d_j)$ and $E_{ALD}(l_i,d_j)$ are inputted into 2 SVM classifiers to output the prediction results of p_{sim} and p_{ass}, respectively. Thus, the results of p_{sim} and p_{ass} are fused to get the final prediction results by using the following weighted average strategy.

$$p(l_i,d_j) = w \cdot p_{sim}(l_i,d_j) + (1 - w) \cdot p_{ass}(l_i,d_j) \quad (10)$$

where w is the weight.

1.6 Assessment of the prediction system

The 10-fold cross-validation (10-CV) test is used to evaluate the performance of NELDA. In 10-CV test, the positive sample set and negative sample set are randomly divided into 10 subsets with the almost equal size, respectively. For each fold in 10-CV test, 9 subsets are used as the training samples, and the remaining 1 subset are used as the testing samples. For all the following 10-CV test experiments, all the known lncRNA-disease associations to be used as the testing samples in each fold testing subset were removed, and then we recalculated the lncRNA similarities, disease similarities and lncRNA-disease association network by using the remaining known lncRNA-disease associations.

Accuracy (ACC), F1-score and Matthew's correlation coefficient (MCC) are used as the evaluation metrics to assess the prediction system. We also use AUC and AUPR to evaluate the prediction system. AUC is the area under the receiver operating characteristic (ROC) curve, and AUPR is the area under the precision-recall curve.

2 Results and discussion

2.1 Comparison with other methods

We compared our NELDA method with the state-of-the-art methods of LDASR[28] and LDNFSGB[27] on D_{rel} dataset in 10-CV test (Table S1 lists the main parameters of NELDA in **Supplementary**), and all experiments are implemented on Ubuntu system with a NVIDIA TITAN V GV100. The prediction results of NELDA, LDASR and LDNFSGB are shown in Table 1, from which we can see that the AUC of NELDA is 0.982 7, which is 0.062 7 and 0.020 7 higher than that of LDASR and LDNFSGB, respectively. The AUPR of NELDA is 0.987 4, which is 0.044 9 and 0.014 6 higher than that of LDASR and LDNFSGB, respectively. The ACC, F1 and MCC of NELDA are 0.950 6, 0.948 5 and 0.904 0, which are 0.055 2, 0.055 9 and 0.111 9 higher than that of LDASR, and 0.029 7, 0.031 1 and 0.059 1 higher than that of LDNFSGB, respectively. These results show that NELDA can effectively predict the lncRNA-disease

Table 1 Results of NELDA，LDASR and LDNFSGB on D_{rel} dataset in 10−CV test

Method	ACC	F1	MCC	AUPR	AUC
LDASR	0.895 4	0.892 6	0.792 1	0.942 5	0.920 0
LDNFSGB	0.920 9	0.917 4	0.844 9	0.972 8	0.962 0
NELDA	**0.950 6**	**0.948 5**	**0.904 0**	**0.987 4**	**0.982 7**

Above results are the average results of running three 10-CV tests.

associations.

To evaluate the effect of the strategy of selecting high-quality non lncRNA-disease association pairs, we compared the performance of NELDA, LDASR and LDNFSGB on D_{rel} dataset and D_{ran} dataset in 10-CV test. The results of NELDA, LDASR and LDNFSGB are shown in Table 2, from which we can see that ACC, F1, MCC, AUPR and AUC of NELDA on D_{rel} dataset are 0.950 6, 0.948 5, 0.904 0, 0.987 4 and 0.982 7, respectively, which are higher than that of NELDA on D_{ran} dataset. In addition, the measure metrics values of LDASR and LDNFSGB on the D_{rel} dataset are also higher than those on D_{ran} dataset. These results indicate that the strategy of selecting the high-quality non lncRNA-disease association samples to construct the training dataset indeed help to improve the performance of predictors.

Table 2 Results of NELDA，LDASR and LDNFSGB on D_{rel} and D_{ran} datasets in 10−CV test

Method	Dataset	ACC	F1	MCC	AUPR	AUC
LDASR	D_{ran}	0.755 4	0.737 8	0.515 6	0.829 9	0.794 1
	D_{rel}	**0.895 4**	**0.892 6**	**0.792 1**	**0.942 5**	**0.920 0**
LDNFSGB	D_{ran}	0.809 4	0.796 9	0.623 7	0.900 5	0.894 7
	D_{rel}	**0.920 9**	**0.917 4**	**0.844 9**	**0.972 8**	**0.962 0**
NELDA	D_{ran}	0.842 2	0.833 7	0.688 2	0.924 6	0.923 4
	D_{rel}	**0.950 6**	**0.948 5**	**0.904 0**	**0.987 4**	**0.982 7**

Above results are the average results of running three 10-CV tests.

2.2 Influence of the decision−level weighted fusion strategy

To evaluate the influence of decision-level weighted fusion strategy, we also designed another two predictors of NELDA-SIM and NELDA-ASS. For one lncRNA-disease pair, NELDA-SIM inputs the concatenating feature of its lncRNA similarity network embedding feature and its disease similarity network embedding feature to a SVM classifier, and NELDA-ASS inputs the concatenating feature of its lncRNA/disease embedding feature derived from the known lncRNA-disease association network. The results of NELDA, NELDA-SIM and NELDA-ASS on D_{ran} and D_{rel} datasets in 10-CV test are shown in Table 3, from which we can see that the AUC and AUPR of NELDA on D_{rel} dataset are 0.982 7 and 0.987 4, which are 0.004 7 and 0.003 9 higher than that of NELDA-SIM, respectively, and 0.008 5 and 0.005 7 higher than that of NELDA-ASS, respectively; the AUC and AUPR of NELDA on D_{ran} dataset are 0.923 4 and 0.924 6, which are 0.018 9 and 0.031 8 higher than that of NELDA-SIM, and 0.019 6 and 0.013 9 higher than that of NELDA-ASS, respectively. These results show that the strategy of decision-level weighted fusion can effectively improve the performance of NELDA.

Table 3 Results of NELDA，NELDA−SIM and NELDA−ASS on D_{ran} and D_{rel} datasets in 10−CV test

Dataset	Predictor	ACC	F1	MCC	AUPR	AUC
D_{ran}	NELDA-SIM	0.838 4	**0.833 7**	0.678 2	0.892 8	0.904 5
	NELDA-ASS	0.786 2	0.742 6	0.607 9	0.910 7	0.903 8
	NELDA	**0.842 2**	**0.833 7**	**0.688 2**	**0.924 6**	**0.923 4**
D_{rel}	NELDA-SIM	0.944 4	0.942 9	0.890 2	0.983 5	0.978 0
	NELDA-ASS	0.942 6	0.940 0	0.888 7	0.981 7	0.974 2
	NELDA	**0.950 6**	**0.948 5**	**0.904 0**	**0.987 4**	**0.982 7**

Above results are the average results of running three 10-CV tests.

To analyze the effect of different fusion weight used in the decision-level fusion strategy, we implemented NELDA with different fusion weights on D_{rel} dataset in 10-CV test. As shown in Table 4, all the measurement metrics first increase and then decrease with the increase of w. When w = 0.5, the performance of NELDA is optimal. Therefore, we set w = 0.5 for NELDA on D_{rel} dataset.

Table 4 Results of NELDA using different fusion weights on D_{rel} dataset in 10−CV test

w	ACC	F1	MCC	AUPR	AUC
0.9	0.946 9	0.945 4	0.895 4	0.985 5	0.980 6
0.8	0.947 5	0.945 9	0.896 7	0.986 4	0.981 6
0.7	0.949 0	0.947 3	0.900 0	0.986 9	0.982 3
0.6	0.950 3	**0.948 5**	0.902 9	0.987 2	0.982 6
0.5	**0.950 6**	**0.948 5**	**0.904 0**	**0.987 4**	**0.982 7**
0.4	0.949 2	0.946 9	0.901 5	0.987 3	0.982 6
0.3	0.946 6	0.944 2	0.896 6	0.987 1	0.982 4
0.2	0.944 6	0.942 1	0.892 7	0.986 6	0.981 9
0.1	0.943 7	0.941 1	0.890 9	0.985 7	0.980 8

Above results are the average results of running three 10-CV.

In addition, we also compared the performance of using the similarity network raw features and its embedding features, association network raw features and its embedding features. By separately concatenating the lncRNA similarity network raw features and the disease similarity network raw features, the lncRNA similarity network embedding features and the disease similarity network embedding

features, the lncRNA association network raw features and the disease association network raw features, the lncRNA association network embedding features and the disease association network embedding features to generate 4 vectors for representing each lncRNA-disease pair, then we input them into 4 SVM models to predict the lncRNA-disease associations. The comparison results of raw and embedding features are shown in Table S2 in **Supplementary**, from which we can see that association network embedding features achieve a better performance than their corresponding raw features, but the prediction results of similarity network embedding features are slightly lower than that of their corresponding raw features. The reason may be that lncRNA/disease similarity networks do not contain the label information of lncRNA-disease associations, applying the unsupervised deep auto-encoder model to dimension reduction will cause information loss; while lncRNA/disease association networks contain the label information of lncRNA-disease associations, thus its lower dimension embedding features preserve the class separability information, which can improve the prediction performance. Furthermore, we also built another model of ANELDA (Figure S2 in **Supplementary**) by fusing the similarity network raw features and association network embedding features to predict the lncRNA-disease associations. The results of ANELDA and NELDA on D_{rel} dataset in 10-CV test are shown in Table S3 in **Supplementary**, from which we can see that the performance of ANELDA is better than that of NELDA.

2.3 Case studies

To evaluate the power of NELDA for predicting the novel lncRNA-disease associations, we adopted the stomach and breast cancers as the cases to implement our NELDA to predict their potential associated lncRNAs. Stomach cancer is the fifth leading cancer and third most common cause of cancer-related deaths worldwide[39]. For stomach cancer, among all the 20 top lncRNAs (Table 5) predicted by NELDA, 15 of them have the corresponding evidences to verify the associations with stomach cancer. For example, DANCR promotes the progression of stomach cancer, and it has the potential to act as a novel diagnostic biomarker[40]. ZEB1-AS1 acts as the oncogenic roles in the regulation of stomach cancer cells migration, invasion and EMT process *via* modulating ZEB1[10]. EGOT serves as an oncogene in stomach cancer, and it could be useful as a conceivable diagnostic and prognostic

Table 5 Top 20 lncRNAs predicted with NELDA for stomach cancer

LncRNA	Evidences	Rank
HOTAIRM1	MNDR v3.1, Lnc2Cancer 3.0	*1*
DANCR	MNDR v3.1, LncRNADisease 2.0, Lnc2Cancer 3.0	*2*
PCAT1	MNDR v3.1, LncRNADisease 2.0, Lnc2Cancer 3.0	*3*
KCNQ1OT1	MNDR v3.1, LncRNADisease 2.0, Lnc2Cancer 3.0	*4*
CRNDE	MNDR v3.1, LncRNADisease 2.0, Lnc2Cancer 3.0	*5*
ZEB1-AS1	MNDR v3.1, LncRNADisease 2.0, Lnc2Cancer 3.0	*6*
LINC00687	Unconfirmed	*7*
LINC00602	Unconfirmed	*8*
HCCAT5	Unconfirmed	*9*
SNHG1	MNDR v3.1, LncRNADisease 2.0, Lnc2Cancer 3.0	*10*
SNHG12	MNDR v3.1, LncRNADisease 2.0, Lnc2Cancer 3.0	*11*
EGOT	MNDR v3.1, LncRNADisease 2.0, Lnc2Cancer 3.0	*12*
C5orf66-AS1	MNDR v3.1, Lnc2Cancer 3.0	*13*
RMST	Unconfirmed	*14*
LINC00473	MNDR v3.1, Lnc2Cancer 3.0	*15*
SOX2-OT	MNDR v3.1, LncRNADisease 2.0	*16*
LUCAT1	MNDR v3.1, Lnc2Cancer 3.0	*17*
HCG27	Unconfirmed	*18*
HCP5	MNDR v3.1, Lnc2Cancer 3.0	*19*
CBR3-AS1	MNDR v3.1, LncRNADisease 2.0, Lnc2Cancer 3.0	*20*

biomarker for stomach cancer tumorigenesis[41].

Breast cancer is the most frequently diagnosed cancer and leading cause of cancer death in women[42]. For breast cancer, among all the 20 top lncRNAs (Table 6) predicted by NELDA, 14 of them have the corresponding evidences to verify the associations with breast cancer. For example, HULC is overexpressed in breast cancer cell lines and tissues compared with normal breast cell line and normal healthy breast tissues[43]. In addition, HULC promotes the development of breast cancer *via* regulating LYPD1 expression through sponging miR-6754-5p. Deletion of HNF1A-AS1 suppresses the malignant phenotypes of breast cancer cells *in vitro* and *in vivo* by targeting miRNA-20a-5p/TRIM32 axis[44]. HNF1A-AS1 could be a promising treatment target for breast cancer[44]. SNHG1 functions as a novel oncogene in breast cancer through the SNHG/miR573/LMO4 axis [11].

Table 6 Top 20 lncRNAs predicted with NELDA for breast cancer

LncRNA	Evidences	Rank
HULC	MNDR v3.1, Lnc2Cancer 3.0	*1*
NPTN-IT1	MNDR v3.1	*2*
WT1-AS	MNDR v3.1, LncRNADisease 2.0	*3*
PCAT1	MNDR v3.1, Lnc2Cancer 3.0	*4*
HNF1A-AS1	Reference[44]	*5*
SNHG1	MNDR v3.1, Lnc2Cancer 3.0	*6*
ZEB1-AS1	MNDR v3.1, Lnc2Cancer 3.0	*7*
LINC00687	Unconfirmed	*8*
LINC00602	Unconfirmed	*9*
HCCAT5	Unconfirmed	*10*
TUSC7	MNDR v3.1	*11*
CBR3-AS1	MNDR v3.1	*12*
PCGEM1	Unconfirmed	*13*
CASC2	MNDR v3.1, LncRNADisease 2.0, Lnc2Cancer 3.0	*14*
GHET1	MNDR v3.1, Lnc2Cancer 3.0	*15*
DRAIC	MNDR v3.1, LncRNADisease 2.0	*16*
MIR17HG	Unconfirmed	*17*
HCP5	Unconfirmed	*18*
HOTAIRM1	MNDR v3.1, Lnc2Cancer 3.0	*19*
BANCR	MNDR v3.1, Lnc2Cancer 3.0	*20*

In summary, 29 (14 for breast cancer, 15 for stomach cancer) out of 40 predicted lncRNAs associated with breast and stomach cancers have been supported by recent literatures and public database. Results of these 2 case studies show that our NELDA can effectively predict the potential association lncRNAs for a disease.

3 Conclusion

LncRNAs participate in a variety of vital biological processes and closely relate with various human diseases. The prediction of lncRNA-disease association can help to understand the mechanisms of human disease at the molecular level, and contribute to diagnosis and treatment of diseases. Most existing lncRNA-disease association prediction methods ignored the deep embedding features hided in the network topological structures. In this work, we presented a novel method of NELDA to predict the potential lncRNA-disease associations by extracting the lncRNA/disease deep embedding features with 4 deep auto-encoder models. NELDA first constructs 3 networks of a lncRNA similarity network, a disease similarity network and a lncRNA-disease association network based on the lncRNA expression profiles, disease ontology and the known lncRNA-disease associations, then builds 4 deep auto-encoder models to extract the lncRNA/disease similarity network embedding features and the lncRNA/disease

association network embedding features, respectively. In the end, NELDA adopts the weighted fusion strategy to fuse the outputs of 2 SVM classifiers for identifying whether a lncRNA is associated with a disease. The experimental results on two datasets in 10-CV test show that the performance of our NELDA is superior to other state-of-the-art methods of LDASR and LDNFSGB. The strategies of the weighted fusion in decision level and selecting the higher quality non lncRNA-disease association pairs for building the high-quality training set can effectively improve the performance of predictors. In addition, results of two cases studies on stomach and breast cancers indicate that NELDA is powerful to find the novel association lncRNAs for one disease, which provides the candidates for further biological experimental validation.

Although NELDA achieves good performance for predicting the lncRNA-disease associations, there are still some issues needing to be improved and further studied in the future. On one hand, there are many biological resources about lncRNAs and diseases, but how to effectively integrate these biological resources is a direction worthy to discuss and further research in the future. On the other hand, we expect to explore more effective strategies to select higher quality non lncRNA-disease association pairs for constructing high quality training dataset to further enhance the accuracy of predicting the lncRNA-disease associations.

Supplementary PIBB_20210132_S1. pdf is available online (http://www.pibb.ac.cn or http://www.cnki.net).

References

[1] Djebali S, Davis C A, Merkel A, *et al*. Landscape of transcription in human cells. Nature, 2012, **489**(7414): 101-108

[2] Fan X N, Zhang S W, Zhang S Y, *et al*. LncRNA_Mdeep: an alignment-free predictor for distinguishing long non-coding rnas from protein-coding transcripts by multimodal deep learning. Int J Mol Sci, 2020, **21**(15): 5222

[3] Mattick J S, Makunin I V. Non-coding RNA. Hum Mol Genet, 2006, **15**(1): R17-R29

[4] Shi Y G, Wang K M. Long noncoding RNA in digestive system neoplasms. Chinese Journal of Clinical Oncology, 2013, **40**(15): 938-940
石永国, 王科明 . 中国肿瘤临床, 2013, **40**(15): 938-940

[5] Mercer T R, Dinger M E, Mattick J S. Long non-coding RNAs: insights into functions. Nat Rev Genet, 2009, **10**(3): 155-159

[6] Li J, Xuan Z, Liu C. Long non-coding RNAs and complex human diseases. Int J Mol Sci, 2013, **14**(9): 18790-18808

[7] Wapinski O, Chang H Y. Long noncoding RNAs and human disease. Trends Cell Biol, 2011, **21**(6): 354-361

[8] Ying L, Chen Q, Wang Y W, *et al*. Upregulated MALAT-1 contributes to bladder cancer cell migration by inducing epithelial-to-mesenchymal transition. Mol Biosyst, 2012, **8**(9): 2289-2294

[9] Zhao Y, Feng C J, Li Y J, *et al*. LncRNA H19 promotes lung cancer proliferation and metastasis by inhibiting miR-200a function. Mol Cell Biochem, 2019, **460**(1-2):1-8

[10] Li Y L, Wen X W, Wang L G, *et al*. LncRNA ZEB1-AS1 predicts unfavorable prognosis in gastric cancer. Surg Oncol, 2017, **26**(4): 527-534

[11] Xiong X, Feng Y, Li L, *et al*. Long noncoding RNA SNHG1 promotes breast cancer progression by regulation of LMO4. Oncol Rep, 2020, **43**(5): 1503-1515

[12] Chen X, Yan G Y. Novel human lncRNA-disease association inference based on lncRNA expression profiles. Bioinformatics, 2013, **29**(20): 2617-2624

[13] Ding L, Wang M, Sun D, *et al*. TPGLDA: novel prediction of associations between lncRNAs and diseases *via* lncRNA-disease-gene tripartite graph. Sci Rep, 2018, **8**:1065

[14] Chen X, Yan C C, Zhang X, *et al*. Long non-coding RNAs and complex diseases: from experimental results to computational models. Brief Bioinform, 2017, **18**(4): 558-576

[15] Chen X, Sun Y Z, Guan N N, *et al*. Computational models for lncRNA function prediction and functional similarity calculation. Brief Funct Genomics, 2019, **18**(1): 58-82

[16] Fan X N, Zhang S W, Zhang S Y, *et al*. Prediction of lncRNA-disease associations by integrating diverse heterogeneous information sources with RWR algorithm and positive pointwise mutual information. BMC Bioinformatics, 2019, **20**:87

[17] Zhang J, Zhang Z, Chen Z, *et al*. Integrating multiple heterogeneous networks for novel lncRNA-disease association inference. IEEE/ACM Trans Comput Biol Bioinform, 2019, **16**(2): 396-406

[18] 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. Mol Biosyst, 2014, **10**(8): 2074-2081

[19] Zhou M, Wang X, Li J, *et al*. Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. Mol Biosyst, 2015, **11**(3): 760-769

[20] Chen X. KATZLDA: KATZ measure for the lncRNA-disease association prediction. Sci Rep, 2015, **5**: 16840

[21] Wang Y, Juan L, Peng J, *et al*. LncDisAP: a computation model for lncRNA-disease association prediction based on multiple biological datasets. BMC Bioinformatics, 2019, **20**(Suppl 16): 582

[22] Li J, Li X, Feng X, *et al*. A novel target convergence set based random walk with restart for prediction of potential lncRNA-disease associations. BMC Bioinformatics, 2019, **20**(1): 626

[23] Chen X, You Z H, Yan G Y, *et al*. IRWRLDA: improved random walk with restart for lncRNA-disease association prediction. Oncotarget, 2016, **7**(36): 57919-57931

[24] Zhao X, Yang Y, Yin M. MHRWR: Prediction of lncRNA-disease associations based on multiple heterogeneous networks. IEEE/ACM Trans Comput Biol Bioinform, 2020, **18**(6):2577-2585

[25] Liu Y, Feng X, Zhao H C, *et al*. A novel network-based computational model for prediction of potential lncRNA-disease association. Int J Mol Sci, 2019, **20**:1549

[26] Lan W, Li M, Zhao K J, *et al*. LDAP: a web server for lncRNA-disease association prediction. Bioinformatics, 2017, **33**(3): 458-460

[27] Zhang Y, Ye F, Xiong D P, *et al*. LDNFSGB: prediction of long non-coding RNA and disease association using network feature similarity and gradient boosting. BMC Bioinformatics, 2020, **21**:377

[28] Guo Z H, You Z H, Wang Y B, *et al*. A learning-based method for lncRNA-disease association identification combing similarity information and rotation forest. iScience, 2019, **19**: 786-795

[29] Yao D, Zhan X, Zhan X, *et al*. A random forest based computational model for predicting novel lncRNA-disease associations. BMC Bioinformatics, 2020, **21**(1): 126

[30] Guo Z H, You Z H, Li L P, *et al*. Combining High Speed ELM With a CNN Feature Encoding to Predict LncRNA-disease Associations. Heidelberg: Springer, Cham, 2019

[31] Fu G, Wang J, Domeniconi C, *et al*. Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. Bioinformatics, 2018, **34**(9): 1529-1537

[32] Chen G, Wang Z Y, Wang D Q, *et al*. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res, 2013, **41**(D1): D983-D986

[33] Ning S W, Zhang J Z, Wang P, *et al*. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. Nucleic Acids Res, 2016, **44**(D1): D980-D985

[34] Lu Z Y, Cohen K B, Hunter L. GeneRIF quality assurance as summary revision. Pac Symp Biocomput, 2007: 269-280

[35] Cheng Z, Huang K, Wang Y, *et al*. Selecting high-quality negative samples for effectively predicting protein-RNA interactions. BMC Syst Biol, 2017, **11**(Suppl 2): 9

[36] Yu G, Wang L G, Yan G R, *et al*. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics, 2015, **31**(4): 608-609

[37] Wang J Z, Du Z, Payattakool R, *et al*. A new method to measure the semantic similarity of GO terms. Bioinformatics, 2007, **23**(10): 1274-1281

[38] Wang D X, Cui P, Zhu W W. Structural Deep Network Embedding// Kdd'16. Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. Heidelberg: Springer Cham, 2016: 1225-1234

[39] Balakrishnan M, George R, Sharma A, *et al*. Changing trends in stomach cancer throughout the world. Currt Gastroenterol Rep, 2017, **19**(8): 36

[40] Pan L, Liang W, Gu J M, *et al*. Long noncoding RNA DANCR is activated by SALL4 and promotes the proliferation and invasion of gastric cancer cells. Oncotarget, 2018, **9**(2): 1915-1930

[41] Peng W, Wu J Z, Fan H, *et al*. LncRNA EGOT promotes tumorigenesis *via* Hedgehog pathway in gastric cancer. Pathol Oncol Res, 2019, **25**(3): 883-887

[42] Donahue H J, Genetos D C. Genomic approaches in breast cancer research. Brief Funct Genomics, 2013, **12**(5): 391-396

[43] Wang N, Zhong C, Fu M, *et al*. Long non-coding RNA HULC promotes the development of breast cancer through regulating LYPD1 expression by sponging miR-6754-5p. Onco Targets Ther, 2019, **12**:10671-10679

[44] Meng Q, Wang L, Lv Y, *et al*. Deletion of HNF1A-AS1 suppresses the malignant phenotypes of breast cancer cells *in vitro* and *in vivo* through targeting miRNA-20a-5p/TRIM32 axis. Cancer Biother Radiopharm, 2020, **36**(1):23-35

NELDA：基于网络嵌入的lncRNA–疾病关联关系预测*

李维娜 樊校楠 张绍武**

（西北工业大学自动化学院，信息融合技术教育部重点实验室，西安 710072）

摘要 目的 长非编码RNA（lncRNAs）参与多种重要的生物学过程并与各种人类疾病密切相关，因此，lncRNA-疾病关联预测研究有助于疾病的诊断、治疗和在分子水平理解人类疾病的发生发展机制。目前，大多数lncRNA-疾病关联预测方法倾向于浅层整合lncRNA和疾病的相关信息，忽略网络拓扑结构中的深层嵌入特征；另外通过随机选取lncRNA-疾病非关联对构建负样本训练集合，影响预测方法的鲁棒性。**方法** 本文提出一种基于网络嵌入的NELDA方法，预测潜在的lncRNA-疾病关联关系。NELDA首先利用lncRNA表达谱、疾病本体论和已知的lncRNA-疾病关联关系，构建lncRNA相似性网络、疾病相似性网络和lncRNA-疾病关联网络。然后，通过设计4个深度自编码器分别从lncRNA/疾病的相似性网络、lncRNA-疾病关联网络学习lncRNA和疾病的低维网络嵌入特征。串联lncRNA和疾病的相似性网络嵌入特征及lncRNA和疾病的关联网络嵌入特征，分别输入两个支持向量机分类器预测lncRNA-疾病关联。最后，采用加权融合策略融合两个支持向量机分类器的预测结果，给出lncRNA-疾病关联关系的最终预测结果。另外，根据已知的lncRNA-疾病关联对和疾病语义相似性，设计一种负样本选取策略构建可信度相对较高的lncRNA-疾病非关联对样本集，用以改善分类器的鲁棒性，该策略通过设计一种打分函数为每对lncRNA-疾病进行打分，选取得分较低的lncRNA-疾病对作为lncRNA-疾病非关联对样本（即负样本）。**结果** 十折交叉验证实验结果表明：NELDA能够有效预测lncRNA-疾病关联关系，其AUC达到0.982 7，比现有LDASR和LDNFSGB方法分别提高了0.062 7和0.020 7。另外，负样本选取策略与决策级加权融合策略能够有效改善NELDA预测性能。胃癌和乳腺癌案例研究中，29/40（72.5%）预测的与胃癌和乳腺癌关联lncRNAs，在近期文献和公共数据库中能够发现相关的支撑证据。**结论** 这些实验结果表明，NELDA是一种有效的lncRNA-疾病关联关系预测方法，具有挖掘潜在lncRNA-疾病关联关系的能力。

关键词 lncRNA-疾病关联，网络嵌入，深度自编码器，高质量负样本选取

中图分类号 TP391 **DOI**：10.16476/j.pibb.2021.0132

* 国家自然科学基金（61873202，62173271）资助项目。

** 通讯联系人。

Tel：029-88431308，E-mail：zhangsw@nwpu.edu.cn

收稿日期：2021-05-10，接受日期：2021-09-02