

BMC Supplements

Prediction of LncRNA Functions using Deep Neural Networks based on multiple networks

--Manuscript Draft--

Manuscript Number:	SUPP-D-21-00223	
Full Title:	Prediction of LncRNA Functions using Deep Neural Networks based on multiple networks	
Article Type:	BMC Supplements Reviewed	
Funding Information:	National Natural Science Foundation of China (61972422)	Mr. Lei Deng
Abstract:	<p>More and more studies show that lncRNA is widely involved in various physiological processes of the organism. However, the functions of the vast majority of them continue to be unknown. In addition, data related to lncRNAs in biological databases are constantly increasing. Therefore, it is quite urgent to develop a computing method to make the utmost of these data. In this paper, we propose a new computational method based on global heterogeneous networks to predict the functions of lncRNAs, called DNGRGO. DNGRGO first calculates the similarities among proteins, miRNAs, and lncRNAs, and annotates the functions of lncRNAs according to its similar protein-coding genes, which have been labeled with gene ontology (GO). To evaluate the performance of DNGRGO, we manually annotated GO terms to lncRNAs and implemented our method on these data. Compared with the existing methods, the results of DNGRGO show superior predictive performance of maximum F-measure and coverage. DNGRGO is able to annotate lncRNAs through capturing the low-dimensional features of the heterogeneous network. Moreover, the experimental results show that integrating miRNA data can help to improve the predictive performance of DNGRGO.</p>	
Corresponding Author:	Jingpu Zhang Henan University of Urban Construction CHINA	
Corresponding Author E-Mail:	20181027@hncj.edu.cn	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	Henan University of Urban Construction	
Corresponding Author's Secondary Institution:		
First Author:	Lei Deng	
First Author Secondary Information:		
Order of Authors:	Lei Deng Shengli Ren Jingpu Zhang	
Order of Authors Secondary Information:		
Opposed Reviewers:		
Additional Information:		
Question	Response	
Please select the article type from the below list:	Meeting report	
Please select the name of the	ISBRA 2020	

Supplement you are submitting to from
the below list:

Click here to view linked References
Deng et al.

1
2
3
4
5
6
7
8
9
10
11
12

RESEARCH

Prediction of LncRNA Functions using Deep Neural Networks based on multiple networks

Lei Deng¹, Shengli Ren¹ and Jingpu Zhang^{2*}

Abstract

Background: More and more studies show that lncRNA is widely involved in various physiological processes of the organism. However, the functions of the vast majority of them continue to be unknown. In addition, data related to lncRNAs in biological databases are constantly increasing. Therefore, it is quite urgent to develop a computing method to make the utmost of these data.

Results: In this paper, we propose a new computational method based on global heterogeneous networks to predict the functions of lncRNAs, called DNGRGO. DNGRGO first calculates the similarities among proteins, miRNAs, and lncRNAs, and annotates the functions of lncRNAs according to its similar protein-coding genes, which have been labeled with gene ontology (GO). To evaluate the performance of DNGRGO, we manually annotated GO terms to lncRNAs and implemented our method on these data. Compared with the existing methods, the results of DNGRGO show superior predictive performance of maximum F-measure and coverage.

Conclusions: DNGRGO is able to annotate lncRNAs through capturing the low-dimensional features of the heterogeneous network. Moreover, the experimental results show that integrating miRNA data can help to improve the predictive performance of DNGRGO.

Keywords: Gene ontology; lncRNA functions; PPMI; SDAE; network representation

Background

LncRNA is an RNA molecule that is defined as endogenous molecules with a length of more than 200 nucleotides. More and more biologically-functioning lncRNAs are continually being found in various organisms. LncRNAs are widely involved in animal neurodevelopment, cell cycle regulation, cell regulation, tumorigenesis, and metastasis [1],[2]. Moreover, it is reported that human diseases and cancers are associated with mutations and dysregulations of lncRNAs [3]. Thus, identifying functions of lncRNAs has become increasingly important. During those years, a few functions of long non-coding RNAs (lncRNAs) have been annotated by the development of high-throughput next-generation sequencing techniques and lncRNA chip technology [4],[5],[6]. There are still a large number of lncRNAs need to be annotated.

Based on biological experiments, biologists can identify functions of lncRNAs through a variety of mechanisms, such as pIgR, CLIP, RAP, etc. However, the

experimental characterization of lncRNA functions often costs too much money while there also will be a slow process [7]. Besides some biological methods, recently, several approaches and tools have been designed to identify functions of lncRNAs. Genes with similar expression patterns across multiple conditions usually have close functional relationship or are associated with related biological pathways. Therefore, some researchers determined the lncRNA functions according to the co-expression patterns of genes. Guttman et al. used mouse microarray data and lncRNA-mRNA co-expression data to construct a network to predict the functions of lncRNAs [8]. Liao et al. also used those microarray expression profiles, as well as local information, to annotated functions of 340 lncRNAs, which were concluded by constructing the coexpression of encoding-non-coding[9]. In addition to these local methods, a bichromatic biological network was established to predict the functions of lncRNAs based on coexpression data and protein interaction data by Guo et al[10]. Recently, Jiang et al. have further proposed a method called LncRNA2Function, which was developed to identify the functions of 9625 lncRNAs by hypergeometric tests[11]. More recently, Zhang et

*Correspondence: zhanjp@csu.edu.cn

²School of Computer and Data Science, Henan University of Urban Construction, 467000, Pingdingshan, China

Full list of author information is available at the end of the article

al. have annotated lncRNAs with gene-ontology terminology based on KATZ measures [12].

Network learning is a set of techniques that aims to map data structures into latent spaces efficiently. Either for dimension reduction or for exploring semantic content, this type of feature embedding has proved to be robust for node classification. In this study, based on network representations, we developed a novel predictor named DNGRGO, in which we used GO terms as functional annotations for lncRNAs. In this method, we built a global heterogeneous network at first, which contained six networks, namely, lncRNA similarity network, lncRNA-protein association network, protein-protein interaction network, miRNA-lncRNA association network, miRNA-protein network, and miRNA-miRNA co-expression network. Then, we used random walk with restart(RWR) and stacked denoising autoencoder to calculate the low-dimensional features of each node in the network. Finally, we annotated lncRNAs by training an SVM classifier for each GO term based on these compact features and annotations of the protein. To evaluate DNGRGO, we run it on the manually organized independent test set, namely lncRNA2GO-68. Moreover, to illustrate the performance of our method, we compared our experimental results with the three latest methods, KATZGO [12], PLNRGO [13], and BIRWLGO [14]. The experimental results indicate that our method is better than others in terms of F-measure on the independent test set.

Results

Benchmark

We evaluated DNGRGO and compared it with other methods through independent validation. However, there was no functional annotation dataset for lncRNAs. Hence, we manually annotated each gene in lncRNA2GO-68 through the sequence, structural information, genomic background, expression, and other information about lncRNAs that had been experimentally verified in the literature.

Evaluation measures

We used the trained SVM model to make predictions for each lncRNA in the independent dataset. Each lncRNA is corresponding to several possible GO terms, and the score of each GO term is between 0 and 1. The higher the score, the more confident the prediction is. Therefore, we need to set a threshold of t to determine the final predicted term $p(t)$. We considered all GO items in each lncRNA which were greater than or equal to t as the prediction set $p(t)$, and each lncRNA manually annotated GO items as the experimental verification set T . To measure the performance of predictive

methods, we adopted three commonly used measurements, namely precision, recall, and F-measure. For a rank threshold, precision and recall are defined as followings:

$$Pr_i(t) = \frac{\sum_{f \in O} I(f \in P_i(t) \wedge f \in T_i)}{\sum_{f \in O} I(f \in P_i(t))} \quad (1)$$

and

$$Rc_i(t) = \frac{\sum_{f \in O} I(f \in P_i(t) \wedge f \in T_i)}{\sum_{f \in O} I(f \in T_i)} \quad (2)$$

Where, O denotes the data set of the entire gene ontology, and f represents a specific GO item in the entire ontology. $I(x)$ is the indicator function, which is described as:

$$I(x) = \begin{cases} 1 & x = \text{true} \\ 0 & x = \text{false} \end{cases} \quad (3)$$

After prediction, we expected to draw the PR curve by calculating the average precision and recall under different thresholds. A predicted lncRNA corresponds to several possible GO terms, and each GO term corresponds to a probability score. If at least one probability score is greater than or equal to the threshold, we put this lncRNA into the $m(t)(\leq N)$ dataset. Based on these $m(t)$ lncRNAs, we can calculate the average precision corresponding to each threshold t . Then we define the average precision as:

$$Pr(t) = \frac{1}{m(t)} * \sum_{i=1}^{m(t)} Pr_i(t) \quad (4)$$

Similarly, we can use the same way to calculate the average recall in the independent test dataset containing N lncRNAs. Then the average recall can be defined as:

$$Rc(t) = \frac{1}{N} * \sum_{i=1}^N Rc_i(t) \quad (5)$$

Different thresholds will lead to different precision and recall. In large-scale data sets, these two indicators are often mutually restrictive. When the threshold is larger, fewer GO terms are predicted of each lncRNA, which can get higher precision, but this will also lead to a lower recall. When the threshold is lower, more GO terms are predicted of each lncRNA, which can get a higher recall rate but also lead to lower precision. To solve this problem, we need to weigh these

two indicators (precision and recall) comprehensively, which is to calculate the maximum F-measure for all thresholds. It can be calculated as the following:

$$F_{max} = \max_t \left(\frac{2 * Pr(t) * Rc(t)}{Pr(t) + Rc(t)} \right) \quad (6)$$

Parameter tuning

In our method, we used RWR to extract the structural information of the global network. The RWR algorithm contains a parameter α , which denotes the restart probability. The setting α takes the value from 0 to 1. Assuming starting from a certain node, the larger the value of α , the greater the probability of returning to the starting node. To validate the influence of its different values, we increased α from 0.1 to 0.9 with step size 0.1. The demonstration shows that the performance is relatively stable when α is set to different values. In the experiment, we chose the restart probability α to be 0.5. After obtaining the topological features of the global network, SDAE was employed to reduce the dimension of features. In the SDAE network, there are many hyperparameters to be tuned. We set the same values as Cao et al.'s research [15]. The number of layers of the entire network was set to 5, and the number of nodes in each layer denoted as M equals [36863-10000-3000-1000-512].

We extracted features of different dimensions and calculated F_{max} on the lncRNA2GO-68 dataset for evaluation, because features of different dimensions may affect the prediction performance of DNGRGO. As shown in Figure 1, F_{max} increases first and then decreases gradually when the dimension increases. It comes to the max value when the dimension equals 512. Hence, we finally reduced the high-dimensional features to 512 dimensions and entered them as input to the classifier.

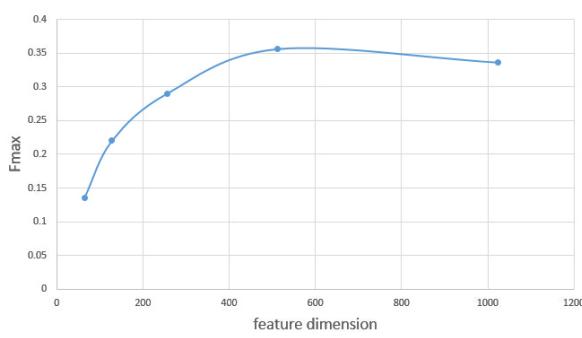


Figure 1: Influence of different feature dimension for function prediction.

Effect of Integrating miRNA Data

There have been several methods for investigating functions of lncRNAs through integrating multiple data sources, such as KATZLGO, BiRWLGO, PLNRGO. Compared with these methods, DNGRGO has newly added miRNA data. To validate the effectiveness of miRNA data, we evaluated DNGRGO on two different network configurations including: the network without miRNAs (miRNA-miRNA similarities, miRNA-protein interactions, and miRNA-lncRNA associations removed) and the entire network. These two configurations of DNGRGO were tested on the lncRNA2GO-68 dataset for precision, recall, and F_{max} . As shown in Table 1, the F_{max} score is 0.356 for the entire network, and 0.306 for the network without miRNAs. The results show that the entire network with integrated miRNA data can significantly better predict functions of lncRNAs than the network without integrated miRNAs .

Table 1: Performance comparision on two different network configurations: the network without miRNAs and the entire network.

Method	Recall	Precision	F_{max}
the entire network	0.395	0.324	0.356
the network without miRNAs	0.515	0.218	0.306

Performance Compared with Other Methods

At present, the most commonly used method for predicting functions of lncRNAs based on co-expression is "guilt-by-association". The conclusion is that if lncRNAs and the coding genes have similar expression patterns, they have similar functions [16]. The KATZ measure assigns different weights to neighboring nodes, giving larger weights to short paths and smaller weights to long paths. KATZLGO builds a global network of lncRNA and protein, then uses the KATZ measure to calculate the correlation scores between each pair of genes, and finally selects the GO term corresponding to the protein with the high correlation score as the functional annotation of lncRNAs [12]. In the BiRWLGO method, a global heterogeneous network of lncRNA and protein is constructed, and a double random walk is performed to calculate the probability scores between all lncRNA-protein pairs. A higher probability indicates a higher degree of association between the pair of genes. Then, the prediction of lncRNA functions can be achieved through the adjacent protein annotated with the GO terms. [14]. Same as the first two methods, PLNRGO first constructs a heterogeneous network of lncRNA-proteins, then uses

random walks to extract network features, and uses SVM to predict the functions of lncRNAs [13].

In this paper, our method DNGRGO is compared with the three methods in precision, recall, and F_{max} . The detailed comparison results are shown in Table 2. Moreover, the precision-recall curves of different methods are plotted in Figure 2. As shown, DNGRGO achieves the highest F_{max} score of 0.356, which performs better than the other three methods. Besides F_{max} , our method also gains the highest score of precision. Besides, the number of lncRNAs correctly annotated is shown in Figure 3. 66 lncRNAs of the manually organized 68 lncRNAs are correctly annotated by our method and PLNRGO, KATZLGO and BiRWGO follow with the numbers of 63 and 64.

Table 2: Performance comparison with other methods on the lncRNA2GO-68 dataset.

Method	Recall	Precision	F_{max}
DNGRGO	0.395	0.324	0.356
KATZLGO	0.382	0.241	0.297
BiRWGO	0.422	0.212	0.282
PLNRGO	0.535	0.220	0.312

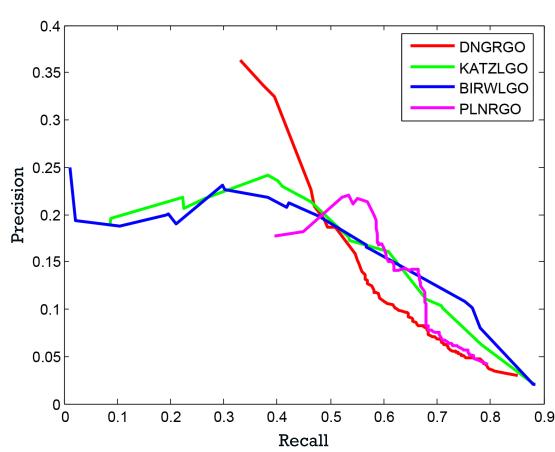


Figure 2: The precision-recall curve is used to estimate the overall performance.

Case study

To further illustrate the performance of our prediction method, we used the prediction results of NEAT1 as a case. NEAT1 is a long non-coding RNA that is critical to speckle integrity. Studies of gain-of-function

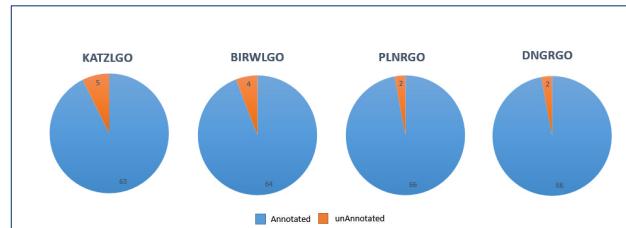


Figure 3: The numbers of lncRNAs that are annotated correctly by different methods, respectively.

or loss-of-function in C2C12 cells have shown that NEAT1 promotes myoblast proliferation but inhibits myoblast differentiation and fusion. [17]. NEAT1 is downregulated in acute promyelocytic leukemia, where it promotes leucocyte differentiation [18] [19]. The results show that the Wnt signaling pathway is activated by knockdown inactivation of NEAT1. And the Wnt signaling pathway is related to many important cell functions, such as cancer stem cells. [20]. NEAT1 knockdown cells produced smaller tumors, demonstrating that NEAT1 promotes tumor growth in vivo [21]. We used the DNGRGO method to predict 158 GO annotations for NEAT1, and then we ranked the GO terms in descending order of predicted scores, of which the first 30 GO terms are listed in Table 3. As predicted, many of them are related to metabolism, which are closely related to the development of cancer, such as GO: 0019222 (regulation of metabolic process), GO: 0044237 (cellular metabolic process), GO: 0032946 (positive regulation of Mononuclear cell proliferation), GO:0051493 (regulation of lipid metabolic process), GO: 0050794 (regulation of cellular process), GO:0046434 (organophosphate catabolic process). There are also a large number of GO terms related to signal channels, such as GO: 0007165 (signal transduction), GO: 0005102 (signaling receptor binding), GO: 0007167 (enzyme-linked receptor protein signaling pathway), GO: 0009755 (hormone-mediated signaling pathway), GO: 0016055 (Wnt signaling pathway).

Discussion and conclusion

Many studies have shown that lncRNA plays an important role in cell function. However, the functional annotation and prediction of lncRNAs have become a considerable challenge due to the non-conservative primary sequence and unstable secondary structure of lncRNAs. In our study, we proposed a global network-based approach, DNGRGO, for large-scale annotation of possible functions of lncRNAs. First, we built a global heterogeneous network containing lncRNAs,

Table 3: The top 30 predicted BP terms for lncRNA NEAT1 by DNNGO.

Rank	GO term	GO name
1	GO:0007166	cell surface receptor signaling pathway
2	GO:0017076	purine nucleotide binding
3	GO:0016192	vesicle-mediated transport
4	GO:0019222	regulation of metabolic process
5	GO:0044237	cellular metabolic process
6	GO:0008270	zinc ion binding
7	GO:0007165	signal transduction
8	GO:0003676	nucleic acid binding
9	GO:0046907	intracellular transport
10	GO:0016197	endosomal transport
11	GO:0009987	cellular process
12	GO:0000166	nucleotide binding
13	GO:0007159	leukocyte cell-cell adhesion
14	GO:0015711	organic anion transport
15	GO:0005102	signaling receptor binding
16	GO:0006936	muscle contraction
17	GO:0009991	response to extracellular stimulus
18	GO:0007167	enzyme linked receptor protein signaling pathway
19	GO:0046434	organophosphate catabolic process
20	GO:0009755	hormone-mediated signaling pathway
21	GO:0050794	regulation of cellular process
22	GO:0030522	intracellular receptor signaling pathway
23	GO:0030518	intracellular steroid hormone receptor signaling pathway
24	GO:0051493	regulation of cytoskeleton organization
25	GO:0051716	cellular response to stimulus
26	GO:0019216	regulation of lipid metabolic process
27	GO:0032946	positive regulation of mononuclear cell proliferation
28	GO:0016055	Wnt signaling pathway
29	GO:0007154	cell communication
30	GO:0046942	carboxylic acid transport

miRNAs, and proteins. Then, we extracted the high-dimensional topological features of each node by RWR, and captured vectors of low-density feature by stacked denoising autoencoder. Finally, we trained the SVM classifier by the previously calculated low-dimensional features and GO terms of the protein. Based on the manually annotated lncRNA2GO-68 dataset, we assessed the performance of DNNGO independently. Experimental results show that DNNGO scores 0.356 and 0.324 on F_{max} and precision, respectively, far higher than the other three methods. In addition, our experiments show that integrating miRNA data into the network can effectively improve the performance of lncRNA's functional prediction. In the end, we believe that DNNGO, as a supplement to biological protocols, will further enrich the study of lncRNA functions.

Methods

To predict the potential functions of lncRNAs, we proposed a new model named DNNGO, which consisted of four steps, as shown in Figure 4. First, we integrated the six networks into a large global network, which contained the protein-protein interaction network, protein-lncRNA association network, lncRNA similarity network, miRNA-miRNA co-expression network, miRNA-protein association network, and miRNA-lncRNA association network. Then we used the random walk with restart (RWR) to extract graph structure information and calculated the positive point of mutual information (PPMI) matrix. To extract low-dimensional features from the PPMI matrix, we used a stack denoising autoencoder to reduce the dimensions. Finally, we trained the SVM model based on topological features and annotation of protein-coding genes, and applied them to annotate the potential functions of lncRNAs.

Materials

lncRNA co-expression similarity

All human lncRNA co-expression data is obtained from NONCODE2016 database [22]. It contains the expression profiles of 90062 human lncRNAs. We calculated Pearson's correlation coefficient (PCC) between each pair of lncRNAs to represent the co-expression similarity of lncRNAs.

protein-protein interaction

We downloaded protein-protein interaction data from the STRING database V10.0[23]. The STRING database is a tool for searching for the relationship between genes and proteins. It contains 2031 species, 9,637,763 proteins, and 1,380,838,440 interactions. In the end, we obtained 17867232 PPI relationships from the database.

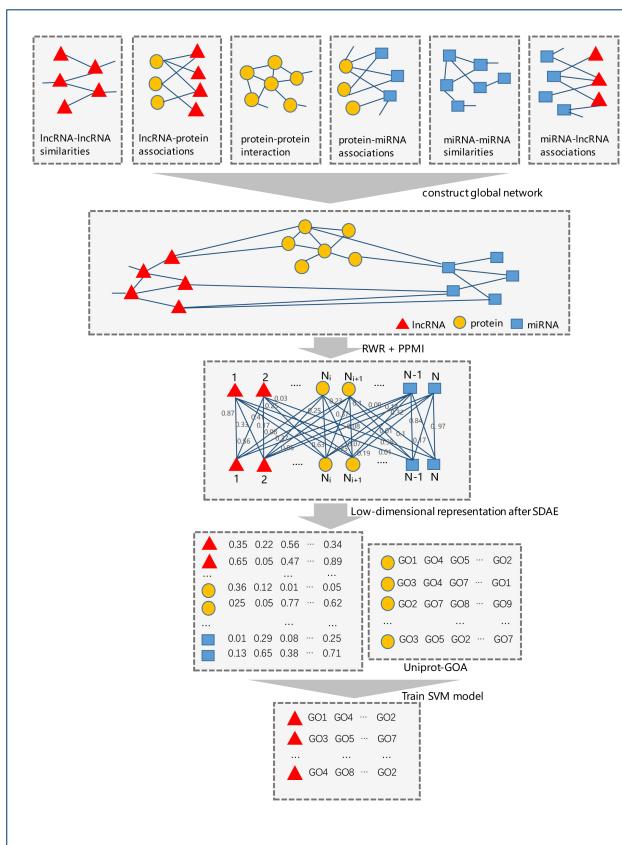


Figure 4: Flowchart of DNGRGO. It consists of four steps: (A) Build the global heterogeneous network composed of six component networks. (B) capture the topological feature of each node through running RWR algorithm on the global network, and calculate the PPMI according to these features. (C) Obtain the low-dimensional feature vectors through SDAE. (D) SVM models are built for different gene ontology terms.

lncRNA-protein associations

To obtain lncRNA protein-data for building a global network, we first downloaded all human lncRNA genes and protein-encoding genes from the GENCODE database of release 24[24]. After screening, a total of 15941 lncRNAs and 20284 proteins were extracted. To build the lncRNA-protein associations, we combined three data sources, which are as follows:

I. Co-expression data from COXPRESdb [25]. COXPRESdb is a database that provides co-expression information of 11 animal species. In COXPRESdb, we got a pre-processed lncRNA-protein co-expression dataset, which mainly refers to the Pearson correlation coefficient between human gene pairs. The specific cal-

culation is as follows:

$$S(l, p) = 1 - \prod_{n=1}^N (1 - S_n(l, p)) \text{ if } S_n(l, p) > 0 \quad (7)$$

where $S(l, p)$ represents the overall correlation between lncRNA l and the protein-coding gene p , $S_n(l, p)$ is the correlation score between l and p in the local data set n , and N is the number of $l-p$ gene pairs with positive correlation scores. We only considered positive correlation scores and removed negative correlation scores of gene pairs.

II. Co-expression data from ArrayExpress [26] and GEO [27]. Jiang et al.[11] processed the co-expression data in these two databases and built a web server for us to download. We used the Pearson correlation coefficient to indicate the degree of association between lncRNA and protein.

III. LncRNA-protein interactions from NPinter 3.0 [28]. The lncRNA-protein interactions of 'Homo sapiens' were downloaded from the NPinter database, which contains 491416 ncRNA interaction data with other biomolecules, and these data have been experimentally verified. If there are lncRNA-protein pairs in the interaction data set, we can set their interaction scores to 1, otherwise set to 0.

miRNA-miRNA co-expression similarities

The miRNA expressions involving 638 miRNAs are curated from the mimiRNA [29] database. We calculated the PCC score of each pair of miRNAs as the co-expression similarity of miRNAs.

miRNA-protein interactions

We downloaded the known miRNA-protein associations from RAID V2.0 [30], which covered more than 60 species and had more than 5.27 million RNA-related interactions, including more than 1.2 million RNA-protein interaction data. Then, we evaluate the reliability of each RNA interaction based on the comprehensive confidence score. After preprocessing, we finally obtained 2133 miRNA-protein associations.

miRNA-lncRNA associations

We downloaded miRNA-lncRNA associations from the starBase database [31], which provided the experimentally confirmed miRNA-lncRNA interactions. After removing the redundant items, we collected 4983 miRNA-lncRNA associations.

Construct the global network

Different types of biological data can be integrated to construct networks of biological interactions, thereby correlating potential function. Usually, combining

more interactions can be effective for the lncRNA annotations. The theoretical basis for this conclusion is that interacting protein, and lncRNAs tend to have the same or similar functions[32]. In addition, if genes have transcripts with similar expression patterns, they may share related biological pathways or have similar functions. Therefore, integrating multiple biological datasets can help annotate the functions of lncRNAs. In our work, we annotate the functions of lncRNAs by integrating six-component networks. Let L , P , M , LP , LM , PM represent the adjacency matrices for lncRNA similarity network, protein-protein interaction network, miRNA co-expression network, lncRNA-protein association network, lncRNA-miRNA association network, protein-miRNA association network, respectively. In addition, we represent the global network as the following:

$$G = \begin{bmatrix} L & LP & LM \\ LP^T & P & PM \\ LM^T & PM^T & M \end{bmatrix} \quad (8)$$

Where, T in LP^T , LM^T , PM^T denotes the transpose.

Obtain vector representations of nodes

To capture the topological information of the nodes in the global heterogeneous network, we adopted the DNGR model to obtain the vector representations of nodes [15]. In DNGR, the random walk with restart (RWR) algorithm was employed to extract the contextual information for the nodes. RWR considers not only the local but also the global structural information of the network. It measures the transition probability of the nodes on the graph, and the final distribution can be used to find out the correlations among the nodes. In the formula, G represents a weighted adjacency matrix, which is the global heterogeneous network we build. And A represents the transition matrix, and the sum of each column in the transition matrix is 1. Matrix A can be obtained by applying the column normalization of G . And, each entry $A_{i,j}$ in A represents the probability of walking from node i to node j , which is given by:

$$A_{i,j} = \frac{G_{i,j}}{\sum_k G_{k,i}}$$

RWR can be formulated as following recurrence relation:

$$P_k = \alpha * P_{k-1} * A + (1 - \alpha) * P_0 \quad (9)$$

Where, P_0 represents the identity matrix, each column in the matrix is a 1-hot code, that is, the j -th item

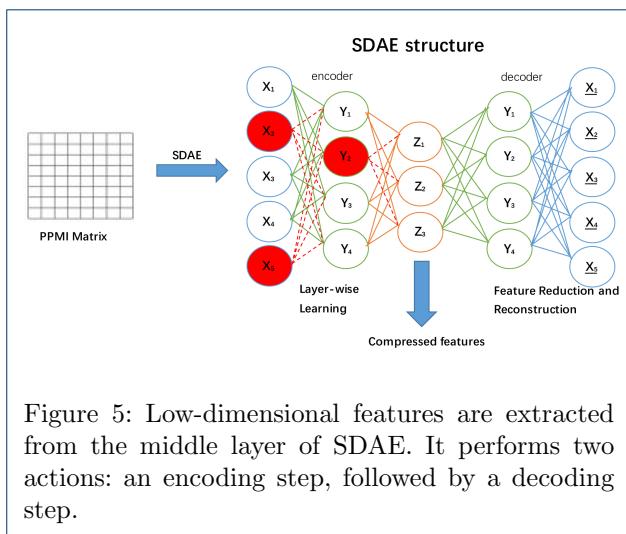
is 1, and the other items are 0. P_k represents the matrix obtained after k steps, and each row of the matrix represents the association between the current node and other nodes in the graph. Starting from a certain node in the graph, each step faces two choices, randomly selects neighboring nodes or returns to the starting node. And α is the probability of restart, means the probability of returning to the original node and restarting the random surfing procedure. $1-\alpha$ represents the probability of moving from the current node to a neighbor node. After multiple iterations, the probability distribution reaches a plateau, which is called the 'diffusion state'. Intuitively, the closer two nodes are, the more intimate the relationship they should have. This means they may have similar functions. Based on the matrix of the diffusion state, we refactor a vector representation of all the nodes in the global network by computing the PPMI matrix.

The PPMI matrix obtained from above approaches is highly dimensional when the network is large. As such, these features cannot be readily used for prediction. To extract the high-quality low-dimensional vector representation for nodes from the PPMI matrix, we employed stacked denosing autoencoder (SDAE) to generate compress low-dimensional vectors.

The stacked denosing autoencoder is based on the automatic encoder. We used the backpropagation algorithm to make the target value equal to the input value. An autoencoder can be divided into two parts: the encoder and the decoder. The autoencoder first receives the input vector, maps it to a low-dimensional latent representation space through a mapping function $f_{\theta_1}(\cdot)$, and then reconstructs the latent representation space into the original input vector by a reconstruction function $g_{\theta_2}(\cdot)$. It is assumed that $f_{\theta_1}(x) = \sigma(W_1x + b_1)$ and $g_{\theta_2}(x) = \sigma(W_2y + b_2)$, where $\sigma(\cdot)$ denotes the activation function, $\theta_1 = \{W_1, b_1\}$ and $\theta_2 = \{W_2, b_2\}$ are the weights in the encoder and the decoder, respectively. The aim is to find the optimal θ_1 and θ_2 by minimizing the loss function:

$$\min_{\theta_1, \theta_2} \sum_{i=1}^n L(x^{(i)}, g_{\theta_2}(f_{\theta_1}(\tilde{x}^{(i)}))) \quad (10)$$

Where, L is the standard squared loss. As shown in Figure 5, the PPMI matrix, which is denoted as x_i , is taken as the input into the SDAE model. y_i denotes the learned representations in the first layer, and z_i represents the learned representations in the second layer. We trained the model by minimizing the loss function, which can be optimized by the standard back-propagation algorithm. When the loss function comes to the minimum, we can extract the low-dimensional features from its bottleneck layer.



Train the SVM models

In this paper, we build a support vector machine (SVM) classifier for each GO term. And the compressed low-dimensional representations calculated in the previous step are taken as the input features. First, cluster proteins between 50 and 100 amino acids in length, and group proteins with a similarity greater than 90% into one class, then select a protein from each class as a representation. In these representations, we deleted GO annotated proteins that were not at least manually assigned. These GO-annotated proteins and the input features we calculated previously form a training set and are put into the SVM classifier. Finally, perform function prediction on lncRNAs according to the trained model.

Supplementary information

Abbreviations

GO: Gene Ontology
PPI: Protein-Protein interaction
SDAE: Stacked Denoise Autoencoder
PPMI: positive point of mutual information
PCC: Pearson's correlation coefficient
PR: Precision-Recall

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

SL designed and implemented the prediction model, performed the experiments, analyzed the obtained results, and wrote the paper. JPZ and LD reviewed and revised the paper. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Funding

This work was supported by National Natural Science Foundation of China [grants No. 61972422].

Availability of data and materials

The data sets of DNGRGO are freely available.

Author details

¹School of Computer Science and Engineering, Central South University, 410075, Changsha, China. ²School of Computer and Data Science, Henan University of Urban Construction, 467000, Pingdingshan, China.

References

- Diederichs, S., et al.: Long noncoding rna: "lncs" to cancer. *European Urology* **65**(6), 1152–1153 (2014)
- Shi, X., Sun, M., Liu, H., Yao, Y., Song, Y.: Long non-coding rnas: a new frontier in the study of human diseases. *Cancer letters* **339**(2), 159–166 (2013)
- Wapinski, O., Chang, H.Y.: Long noncoding rnas and human disease. *Trends in cell biology* **21**(6), 354–361 (2011)
- Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M., Gerstein, M.B.: Annotating non-coding regions of the genome. *Nature Reviews Genetics* **11**(8), 559 (2010)
- Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D., et al.: A comparative encyclopedia of dna elements in the mouse genome. *Nature* **515**(7527), 355 (2014)
- Nam, J.-W., Bartel, D.P.: Long noncoding rnas in c. elegans. *Genome research* **22**(12), 2529–2540 (2012)
- Zeng, X., Lin, W., Guo, M., Zou, Q.: A comprehensive overview and evaluation of circular rna detection tools. *PLoS computational biology* **13**(6), 1005420 (2017)
- Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al.: Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. *Nature* **458**(7235), 223 (2009)
- Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., Zhao, G., Luo, H., Bu, D., Zhao, H., et al.: Large-scale prediction of long non-coding rna functions in a coding–non-coding gene co-expression network. *Nucleic acids research* **39**(9), 3864–3878 (2011)
- Guo, X., Gao, L., Liao, Q., Xiao, H., Ma, X., Yang, X., Luo, H., Zhao, G., Bu, D., Jiao, F., et al.: Long non-coding rnas function annotation: a global prediction method based on bi-colored networks. *Nucleic acids research* **41**(2), 35–35 (2012)
- Jiang, Q., Ma, R., Wang, J., Wu, X., Jin, S., Peng, J., Tan, R., Zhang, T., Li, Y., Wang, Y.: Lncrna2function: a comprehensive resource for functional investigation of human lncrnas based on rna-seq data. In: *BMC Genomics*, vol. 16, p. 2 (2015). BioMed Central
- Zhang, Z., Zhang, J., Fan, C., Tang, Y., Deng, L.: Katzlgo: large-scale prediction of lncrna functions by using the katz measure based on multiple networks. *IEEE/ACM transactions on computational biology and bioinformatics* **16**(2), 407–416 (2017)
- Deng, L., Wu, H., Liu, C., Zhan, W., Zhang, J.: Probing the functions of long non-coding rnas by exploiting the topology of global association and interaction network. *Computational Biology & Chemistry* **74** (2018)
- Zhang, J., Zou, S., Deng, L.: Birwlgo: A global network-based strategy for lncrna function annotation using bi-random walk. In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 50–55 (2017). IEEE
- Cao, S., Lu, W., Xu, Q.: Deep neural networks for learning graph representations. In: *Thirtieth AAAI Conference on Artificial Intelligence* (2016)
- Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., Rinn, J.L.: Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. *Genes & development* **25**(18), 1915–1927 (2011)
- Wang, S., Zuo, H., Jin, J., Lv, W., Xu, Z., Fan, Y., Zhang, J., Zuo, B.: Long noncoding rna neat1 modulates myogenesis by recruiting ezh2. *Cell Death & Disease* **10**(7) (2019)
- Yu, X., Li, Z., Zheng, H., Chan, M.T., Wu, W.K.K.: Neat 1: A novel cancer-related long non-coding rna. *Cell proliferation* **50**(2), 12329 (2017)

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
19. Kong, X., Zhao, Y., Li, X., Tao, Z., Hou, M., Ma, H.: Overexpression of hif-2alpha-dependent neat1 promotes the progression of non-small cell lung cancer through mir-101-3p/sox9/wnt/beta-catenin signal pathway. *Cell Physiol Biochem* **52**, 368–381 (2019)
 20. Jiang, P., Xu, H., Xu, C., Chen, A., Chen, L., Zhou, M., ul Haq, I., Wu, X., Mariyam, Z., Feng, Q.: Neat1 contributes to the csc-like traits of a549/cddp cells via activating wnt signaling pathway. *Chemico-biological interactions* **296**, 154–161 (2018)
 21. Zhang, C., Li, J.-Y., Tian, F.-Z., Zhao, G., Hu, H., Ma, Y.-F., Yang, Y.-L.: Long noncoding rna neat1 promotes growth and metastasis of cholangiocarcinoma cells. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics* **26**(6), 879–888 (2018)
 22. Liu, C., Bai, B., Skogerbo, G., Cai, L., Deng, W., Zhang, Y., Bu, D., Zhao, Y., Chen, R.: Noncode: an integrated knowledge database of non-coding rnas. *Nucleic acids research* **33**(suppl_1), 112–115 (2005)
 23. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al.: String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research* **43**(D1), 447–452 (2014)
 24. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al.: The gencode v7 catalog of human long noncoding rnas: analysis of their gene structure, evolution, and expression. *Genome research* **22**(9), 1775–1789 (2012)
 25. Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T., Kinoshita, K.: Copresdb in 2015: coexpression database for animal species by dna-microarray and rnaseq-based expression data with multiple quality assessment systems. *Nucleic acids research* **43**(D1), 82–86 (2014)
 26. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., et al.: Arrayexpress—a public repository for microarray gene expression data at the ebi. *Nucleic acids research* **31**(1), 68–71 (2003)
 27. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Edgar, R.: Ncbi geo: mining tens of millions of expression profiles—database and tools update. *Nucleic acids research* **35**(suppl_1), 760–765 (2006)
 28. Hao, Y., Wu, W., Li, H., Yuan, J., Luo, J., Zhao, Y., Chen, R.: Npinter v3. 0: an upgraded database of noncoding rna-associated interactions. *Database* **2016** (2016)
 29. Ritchie, W., Flamant, S., Rasko, J.E.: mimirna: a microrna expression profiler and classification resource designed to identify functional correlations between micrornas and their targets. *Bioinformatics* **26**(2), 223–227 (2009)
 30. Yi, Y., Zhao, Y., Li, C., Zhang, L., Huang, H., Li, Y., Liu, L., Hou, P., Cui, T., Tan, P., et al.: Raid v2. 0: an updated resource of rna-associated interactions across organisms. *Nucleic acids research* **45**(D1), 115–118 (2016)
 31. Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., Yang, J.-H.: starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic acids research* **42**(D1), 92–97 (2013)
 32. Ferre, F., Colantoni, A., Helmer-Citterich, M.: Revealing protein–lncrna interaction. *Briefings in bioinformatics* **17**(1), 106–116 (2015)



Click here to access/download
Supplementary Material
68lncRNA_evaluate.goa.txt