Computational Biology and Chemistry 87 (2020) 107282

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac

Research Article

Inferring LncRNA-disease associations based on graph autoencoder matrix completion

Ximin Wu[a], Wei Lan[a,b,*], Qingfeng Chen[a,*], Yi Dong[a], Jin Liu[b], Wei Peng[c]

[a] School of Computer, Electronic and Information, Guangxi University, Nanning, China
[b] Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, China
[c] The Network Center, Kunming University of Science and Technology, Kunming, China

ARTICLE INFO

Keywords:
LncRNA-disease association
Matrix completion
Graph convolutional network
Inner product

ABSTRACT

Accumulating studies have indicated that long non-coding RNAs (lncRNAs) play crucial roles in large amount of biological processes. Predicting lncRNA-disease associations can help biologist to understand the molecular mechanism of human disease and benefit for disease diagnosis, treatment and prevention. In this paper, we introduce a computational framework based on graph autoencoder matrix completion (GAMCLDA) to identify lncRNA-disease associations. In our method, the graph convolutional network is utilized to encode local graph structure and features of nodes for learning latent factor vectors of lncRNA and disease. Further, the inner product of lncRNA factor vector and disease factor vector is used as decoder to reconstruct the lncRNA-disease association matrix. In addition, the cost-sensitive neural network is utilized to deal with the imbalance between positive and negative samples. The experimental results show GAMLDA outperforms other state-of-the-art methods in prediction performance which is evaluated by AUC value, AUPR value, PPV and F1-score. Moreover, the case study shows our method is the effectively tool for potential lncRNA-disease prediction.

1. Introduction

Recent studies have found that the protein-coding genes only account for approximately 1.5% of the entire human genome (Consortium et al., 2001; Lan et al., 2014). It means that more than 98% of the human genome do not participate in encoding protein sequences. These non-coding RNAs, especially long non-coding RNAs (lncRNAs) with lengths exceed 200 nucleotides (nts), have been shown to play important roles in a variety of biological processes such as gene transcription, cell differentiation, immune responses, epigenetic regulation and so on (Wang and Chang, 2011; Wapinski and Chang, 2011; Derrien et al., 2012; Guttman and Rinn, 2012). In addition, the dysregulation and mutation of lncRNAs are implicated in numerous complex human diseases such as bladder cancer (Zhang et al., 2012), diabetes (Pasmant et al., 2011), cardiovascular diseases (Congrains et al., 2012), breast cancer (Godinho et al., 2012) and so on. Therefore, predicting potential associations between lncRNAs and diseases can help us understand human disease and contribute to disease diagnosis and therapy (Chen et al., 2017, 2019a).

Due to identify lncRNA-disease associations based on biological experiment method is still time-consuming and expensive, computational methods provide a significant way to prioritize disease-related lncRNAs. In the past few years, plenty of computational methods have been proposed to infer the relationships between lncRNAs and diseases based on assumption that phenotypically similar diseases tend to be associated with functionally similar lncRNAs. In general, these methods can be divided into three categories (Lan et al., 2018a). The first class of methods are information propagation-based methods. Sun et al. (Sun et al., 2014) presented a network method (named RWRlncD) to predict lncRNA-disease interactions based on the random walk with restart method. This method integrates the lncRNA functional similarity network, disease similarity network and experimentally validated lncRNA-disease associations. Similar work has been done by Zhou et al. (Zhou et al., 2015), they constructed an heterogeneous network to infer potential human lncRNA-disease associations. Chen et al. (Chen et al., 2016c) introduced a model, IRWRLDA, to predict the associations between lncRNAs and diseases. The second kind of methods are based on machine learning. Chen et al. (Chen and Yan, 2013) proposed a semi-supervised learning framework, LRLSLDA, to identify associations between lncRNAs and diseases by utilizing Laplacian Regularized Least Squares method. This method does not need negative samples to train model, but how to more reasonably select

* Corresponding author.
E-mail addresses: ximinwu@qq.com (X. Wu), lanwei@gxu.edu.cn (W. Lan), qingfeng@gxu.edu.cn (Q. Chen), dongyi@st.gxu.edu.cn (Y. Dong), liujin06@csu.edu.cn (J. Liu), weipeng1980@gmail.com (W. Peng).

https://doi.org/10.1016/j.compbiolchem.2020.107282
Received 11 December 2019; Received in revised form 1 April 2020; Accepted 9 May 2020
Available online 20 May 2020

parameters and combine different classifiers are not solved. Chen et al. (Chen et al., 2019) presented a novel framework (ILDMSF) to identify the potential lncRNA-disease associations by using similar network fusion and support vector machine. Fu et al. (Fu et al., 2017) proposed a computational model called MFLDA to infer potential lncRNA-disease associations by weighting different data sources and using the optimized low-rank matrices to reconstruct the lncRNA-disease association matrix. The problem of this method is that the noises of original features are not processed which may affect the prediction performance. Lan et al. (Lan et al., 2016a) developed a web server tool (LDAP) based on positive-unlabeled (PU) learning for lncRNA-disease associations inference by fusing multiple data resources. Lu et al. (Lu et al., 2018) presented a new computational method, SIMCLDA, to identify lncRNA-disease associations by using inductive matrix completion. This method integrates known lncRNA-disease interactions, disease-gene interactions and gene-gene interactions. The third type of methods are statistics methods. Chen et al. (Chen, 2015) designed a novel inference model, HGLDA, to predict lncRNA-disease interactions by using HyperGeometric distribution. Li et al. (Li et al., 2019a) developed a computational method based on hypergeometric distribution to predict the related lncRNAs of gastric cancer.

In this paper, we introduce a computational framework (GAMCLDA) to infer the association of lncRNA-disease based on graph autoencoder matrix completion. In GAMCLDA, the graph convolutional networks is used as the encoder to obtain latent factor vectors of lncRNAs and diseases. Further, the scores of lncRNA-disease interactions are calculated by using the inner product of two latent factor vectors. The experimental results demonstrate that GAMCLDA has a better performance than other four state-of-the-art approaches. Furthermore, case study shows that GAMCLDA can be an effective method for potential lncRNA-disease interaction prediction.

2. Materials and methods

2.1. Problem formulation

Given n lncRNAs and m diseases, the matrix A is constructed to represent associations between lncRNAs and diseases. $A_{ij} = 1$ denotes the lncRNA i is associate with disease j, and $A_{ij} = 0$ denotes their relationship is unobserved. The identification task of potential lncRNA-disease associations can be viewed as completing the matrix A. The example of lncRNA-disease graph structure is shown in Fig. 1.

2.2. Graph convolution networks encoder

The graph convolutional networks (GCNs) is proposed by (Kipf and Welling, 2016), whose purpose is to learn the representations of node attributes and graph structure information.

The vectors $l = [l_{f1}, l_{f2}, \ldots, l_{fu}]$ and $d = [d_{f1}, d_{f2}, \ldots, d_{fv}]$ are constructed to represent original features of lncRNAs and diseases, respectively. The features of lncRNA and disease are reduced to the same dimension by using Multilayer Perceptron (MLP). Then, $l_h = [l_{f1}, l_{f2}, \ldots, l_{fc}]$ and $d_h = [d_{f1}, d_{f2}, \ldots, d_{fc}]$ are obtained as the new lncRNA feature vector and disease feature vector, respectively.

The graph convolutional network encoder is employed to combine the lncRNA-disease association matrix A and the feature vectors of lncRNAs and diseases. The $X = \begin{bmatrix} l_h \\ d_h \end{bmatrix}$ is constructed as the input of encoder. The graph adjacency matrix M is constructed as:

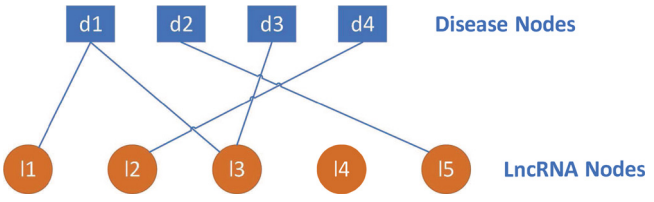

Fig. 1. The example of lncRNA-disease bipartite graph.

$$M = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \tag{1}$$

We use graph convolutional network to compute the latent representation of lncRNAs and diseases, respectively. The latent factor vector of lncRNA L and disease D is calculated as follows:

$$\begin{bmatrix} L \\ D \end{bmatrix} = \delta(\widehat{M} \cdot \delta(\widehat{M} \cdot X \cdot W) \cdot W^{(2)}) \tag{2}$$

$$\widehat{M} = \widetilde{D}^{-\frac{1}{2}} \widetilde{M} \widetilde{D}^{-\frac{1}{2}} \tag{3}$$

$$\widetilde{M} = M + I \tag{4}$$

where W and $W^{(2)}$ denote two weight matrices, I denotes identity matrix. $\widetilde{D}$ denotes a diagonal matrix with $\widetilde{D}_{ii} = \sum_j \widetilde{M}_{ij}$. $\delta(\cdot)$ denotes the activation function.

The propagation of different layers is utilized Kipf method (Kipf and Welling, 2016) which is defined as follows:

$$H^{(l)} = \delta(\widehat{M} H^{(l-1)} W^{(l)}) \tag{5}$$

where $H^{(l)}$ is the matrix of activation in layer l. $W^{(l)}$ denotes the weight matrix for the lth neural network layer. In the input layer ($l = 0$), the matrix of activation $H^{(0)}$ is the feature matrix X, in which row i contains the features of node i.

2.3. Inner product decoder and loss function

As the latent embedding vectors of lncRNAs and diseases cover both content and structure information, the inner product decoder is employed to identify potential lncRNA-disease associations. Let L_i and D_j denote the latent vector of lncRNA i and disease j. Based on Matrix Factorization (MF) (Koren et al., 2009; Ahmed et al., 2018; Ramlatchan et al., 2018), the score of lncRNA-disease interaction is calculated as follows:

$$\widehat{A}_{ij} = \text{sigmoid}(L_i \cdot D_j^T) \tag{6}$$

During model training, we minimize the following logistic loss of matrix $\widehat{A}$ and $L2$ regularization:

$$\ell = \ell_0 + \frac{\lambda}{2} \cdot \sum_{\theta \in \{W, W^{(2)}\}} \|\theta\|^2 \tag{7}$$

$$\text{with} \quad \ell_0 = -\frac{1}{N} \sum_{i,j} (A_{ij} \log \widehat{A}_{ij} + (1 - A_{ij}) \log(1 - \widehat{A}_{ij})) \tag{8}$$

where N denotes the size of adjacency matrix.

2.4. Imbalance problem

Due to the imbalance between positive and negative samples of lncRNA-disease associations, the cost-sensitive neural networks (Kukar et al., 1998) is utilized to tackle this problem. The cost-sensitive neural networks have been widely used in the imbalance learning problem. Here, we modify the loss function as follows:

$$\ell_0 = -\frac{1}{N} \sum_{i,j} W_{ij} \cdot (A_{ij} \log \widehat{A}_{ij} + (1 - A_{ij}) \log(1 - \widehat{A}_{ij})) \tag{9}$$

where W_{ij} denotes the learnable label weight matrix where the dimension is the same as the association matrix. The weight matrix is updated by the feedforward network.

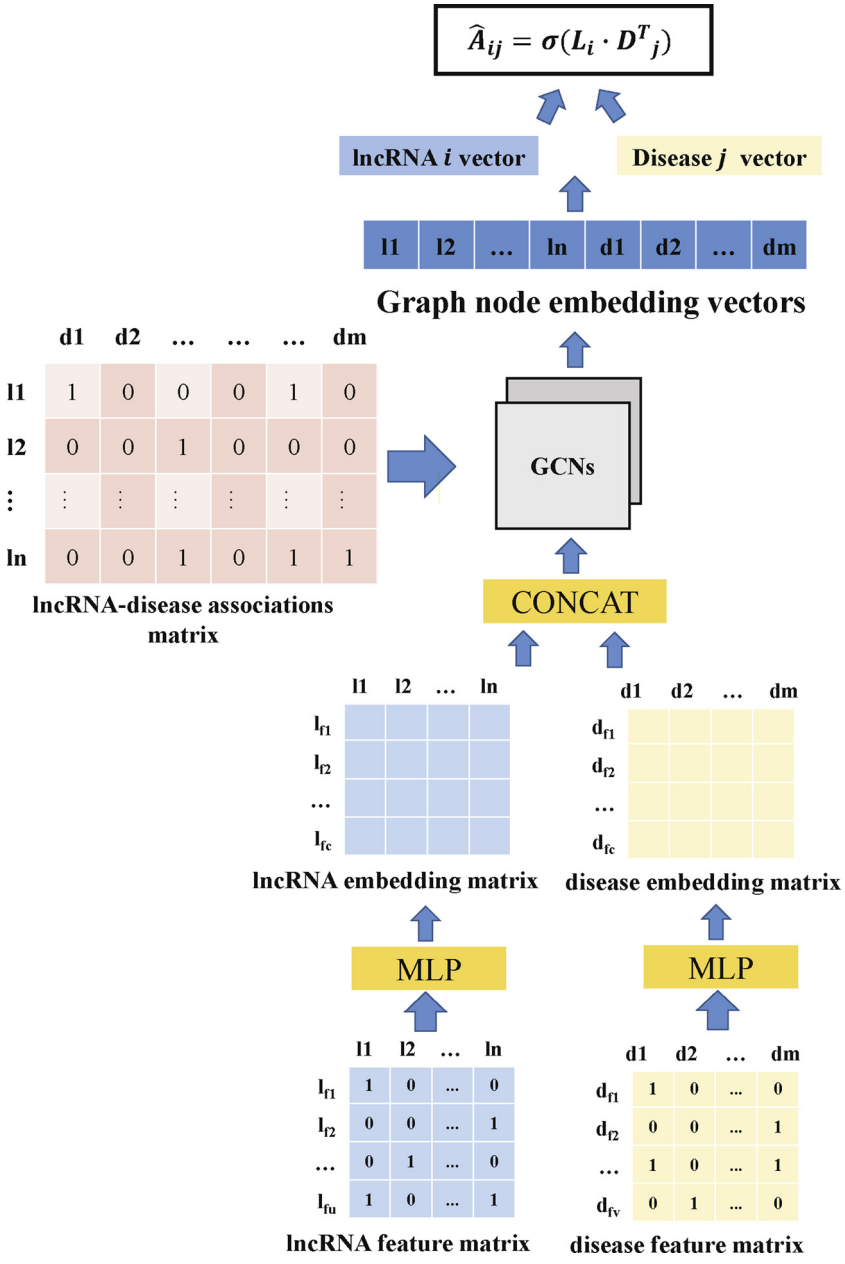

Fig. 2. The flowchart of GAMCLDA.

2.5. GIP Kernel Similarity

Based on assumption that similar diseases exhibit a similar interaction or non-interaction with lncRNAs, the GIP kernel (van Laarhoven et al., 2011) is used to calculate the similarity between the new disease d with other diseases as follows:

$$\mathrm{Sim}(d(i), d(j)) = \exp(-\gamma \|d(i) - d(j)\|^2) \tag{10}$$

$$\text{with} \quad \gamma = \left(\frac{1}{N}\sum_{i=1}^{N} \|d(i)\|^2\right) \tag{11}$$

where $d(i)$ and $d(j)$ denotes the ith and jth original feature of diseases, respectively. The N denotes the numbers of diseases.

The flowchart of GAMCLDA for predicting lncRNA-disease associations is shown in Fig. 2.

3. Experimental results

3.1. Data collection

In order to obtain comprehensive features of lncRNAs and diseases, we combine six different related data sources from public databases, including: (1) LncRNA – gene function associations are collected from GeneRIF (Lu et al., 2007); (2) LncRNA – disease associations are collected from LncRNADisease (Chen et al., 2012), lnc2Cancer (Ning et al., 2015) and GeneRIF (Lu et al., 2007); (3) LncRNA – miRNA associations are downloaded from starBase v2.0 (Li et al., 2014a); (4) MiRNA – disease associations are downloaded from HMDD (Li et al., 2014b); (5) LncRNA – gene associations are downloaded from lncRNA2target (Jiang et al., 2015); (6) Gene – disease associations are downloaded from DisGeNet (Pinero et al., 2015; Lan et al., 2015). In addition to lncRNA-disease interactions, these datasets are fused into lncRNA feature matrix or disease feature matrix. If the lncRNA i is related with entity j, the element (i, j) of lncRNA feature matrix is set to 1, otherwise

0. The disease feature matrix is obtained in term of similar way. In final, 2697 experimentally validated lncRNA-disease associations among 240 lncRNAs and 412 diseases are obtained as gold standard dataset. In addition, 6066 lncRNA features and 10621 disease features are collected from those databases.

3.2. Evaluation metrics

In experiment, the 10-fold cross validation is conducted to evaluate the prediction performance of model in inferring potential associations between lncRNAs and diseases. In the 10-fold cross validation, the known associations are divided into ten sets and each set is removed in turn as test samples, whereas the rest known associations are regarded as training samples for model learning. The scores of unknown associations between lncRNA and disease and test samples are calculated by prediction algorithm. Then, unknown associations and test samples are sorted in descending order based on predicted score. The higher the score is, the closer association is. After that, the true positive rate (TPR) and false positive rate (FPR) are calculated as follows:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{12}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

where TP indicates true positives. FP indicates false positives. TN indicates true negatives. FN indicates false negatives. TPR represents the proportion of positive samples that are correctly identified. FPR indicates that the proportion of real negative samples misclassified as positive samples in the total number of negative samples. Then, the receive operating characteristic (ROC) curve is plotted by varying rank thresholds. The value of area under the ROC Curve (AUC) is computed to evaluate the performance. The higher AUC value is, the better performance is.

In addition, the area under Precision-Recall curve (AUPR) is used to evaluate the performance of prediction model. The precision and recall are defined as follows:

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{14}$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

where precision represents the fraction of true positive samples predicted in predicted positive samples. Recall represents the proportion of true positive samples that are correctly identified. The higher value of AUPR is, the better performance is.

The positive predictive value (PPV) and F1 score are also used to evaluate the performance of method, which are defined as follows:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{16}$$

$$F1\ \mathrm{score} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{17}$$

3.3. Ten-fold cross validation

In order to evaluate the predictive ability of GAMCLDA in lncRNA-disease associations identification, we compare our method with four state-of-the-art methods: MFLDA (Fu et al., 2017), SIMCLDA (Lu et al., 2018), RWRlncD (Sun et al., 2014) and HGLDA (Chen, 2015). The result shows that our approach is significantly superior to those state-of-the-art approaches in terms of AUC and AUPR, respectively. Fig. 3 shows the AUC of different methods. It can be observed that GAMCLDA obtained an averaged AUC of 0.9071, which is significantly higher than other methods (MFLDA 0.6465, SIMCLDA 0.8236, RWRlncD 0.8666 and HGLDA 0.6952). Fig. 4 shows the AUPR of different methods. It can

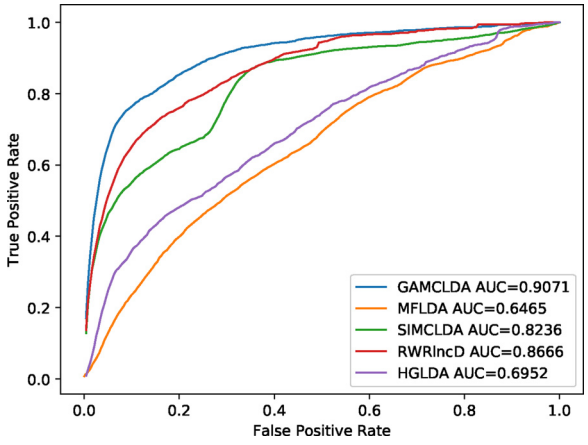

Fig. 3. The performance comparison of GAMCLDA, MFLDA, SIMCLDA, RWRlncD and HGLDA in term of AUC value based on 10-fold cross validation.

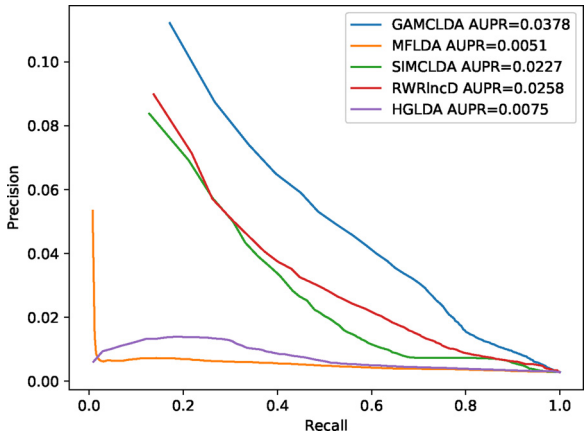

Fig. 4. The performance comparison of GAMCLDA, MFLDA, SIMCLDA, RWRlncD and HGLDA in term of AUPR value based on 10-fold cross validation.

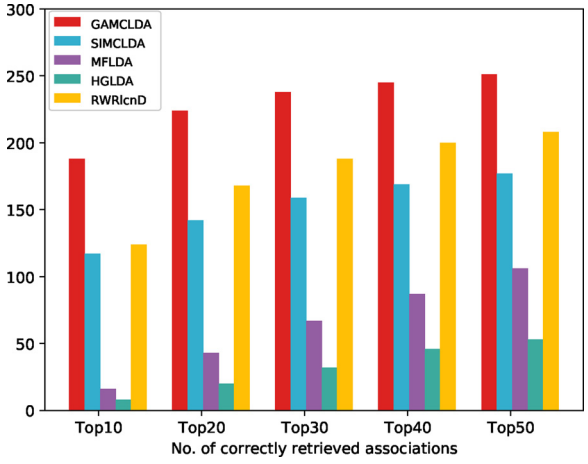

Fig. 5. Number of correctly retrieved known lncRNA-disease associations of top-k based on 10-fold cross validation.

be found that GAMCLDA obtains the averaged AUPR of 0.0378 which is superior to MFLDA (0.0051), SIMCLDA (0.0227), RWRlncD (0.0258) and HGLDA (0.0075). Furthermore, Fig. 5 shows the numbers of correctly retrieved lncRNA−disease associations. In practice, the top−k predicted lncRNAs of each disease are selected as predicted positive sets in ten-fold cross validation. Then, the correctly retrieved number is counted based on the number of true positives. It can be discovered that

Table 1
The performance comparison of GAMCLDA, MFLDA, SIMCLDA, RWRlncD and HGLDA in term of average PPV value and F1 score based on 10-fold cross validation.

Metric	GAMCLDA	MFLDA	SIMCLDA	RWRlncD	HGLDA
PPV	0.009887	0.004578	0.008027	0.008664	0.005138
F1 score	0.019561	0.009091	0.015900	0.017157	0.010202

GAMCLDA outperforms the other methods among top 10 to top 50. Table 1 shows the performance comparison of GAMCLDA, MFLDA, SIMCLDA, RWRlncD and HGLDA in term of average PPV value and F1 score based on 10-fold cross validation. It can be observed that GAMCLDA has a better performance than other methods.

3.4. De novo test of lncRNA − disease prediction

In order to evaluate the prediction performance of GAMCLDA for new disease, the *de novo* test is implemented in our experiment. In the *de novo* test, all known associations of disease d are removed in each time and the rest of lncRNA-disease associations are treated as training set. Then, the top-k similar diseases are selected based on disease similarity. The associations between these diseases and all lncRNAs are treated as known associations between lncRNAs and disease i. Further, the GAMCLDA is utilized to predict lncRNA-disease interactions.

We compare GAMCLDA with other four methods (MFLDA, SIMCLDA, RWRlncD and HGLDA) in term of AUC and AUPR. It can be found from Fig S1 (see Supplementary Materials), GAMCLDA achieves the averaged AUC value of 0.8710, which is much higher than MFLDA (AUC = 0.5952), SIMCLDA (AUC = 0.7923), RWRlncD (AUC = 0.4699) and HGLDA (AUC = 0.6086). Fig S2 (see Supplementary Materials) shows the AUPR of different prediction methods. It can be found that GAMCLDA achieves the highest averaged AUPR value of 0.1640, which is higher than MFLDA (AUPR = 0.0398), SIMCLDA (AUPR = 0.1270), RWRlncD (AUPR = 0.0234) and HGLDA (AUPR = 0.0508). Furthermore, Fig S3 (see Supplementary Materials) shows the numbers of correctly retrieved lncRNA-disease associations. In practice, the top − k predicted lncRNAs of each disease are selected as predicted positive sets in *de novo* test. Then, the correctly retrieved number is counted based on the average number of true positive of each disease. It can be observed that GAMCLDA outperforms other methods among top 10 to top 50. The Table S1 (see Supplementary Materials) shows the performance comparison of GAMCLDA, MFLDA, SIMCLDA, RWRlncD and HGLDA in term of average PPV value and F1 score based on de novo test. It can be found that GAMCLDA outperforms than other state-of-the-art methods.

3.5. Parameters setting

In the experiment, the weights of graph convolution networks are initialized as (He et al., 2015). Moreover, the ReLU activation functions is used in hidden layers. The parameters are set as follows: we train our model for a maximum of 1000 epochs (training iterations) using Adam (Kingma and Ba, 2014). In addition, the learning rate and weight decay parameter (λ) are set as 0.001 and 0.005, respectively. In order to analyze the affect of dimension of latent feature vector, we tested the dimension from 8 to 64. The result is shown in Fig. 6. It can be found that the AUC is the best when the value of dimension is equal to 32. Fig. 7 shows the performance of the top-k similar diseases are selected in *de novo* test where k ranges from 1 to 10. It can be observed that the AUC is the best when k is set to 8.

3.6. Case study

In order to further validate the predictive ability of GAMCLDA, the case study is implemented on breast cancer. Breast cancer (BC) is the major cause of cancer death among women in worldwide (Torre et al., 2017; Lan et al., 2018b). Identifying lncRNAs related with BC may contribute to the prevention and treatment of BC. In the case study, all known lncRNA-disease interactions are treated as training samples. The candidate interactions are predicted by using GAMCLDA. The predicted lncRNAs of breast cancer are ranked according to predicted scores. The top-20 predicted lncRNA are selected, which is shown in Table 2. Then, we verify those associations by manually retrieving databases and literatures.

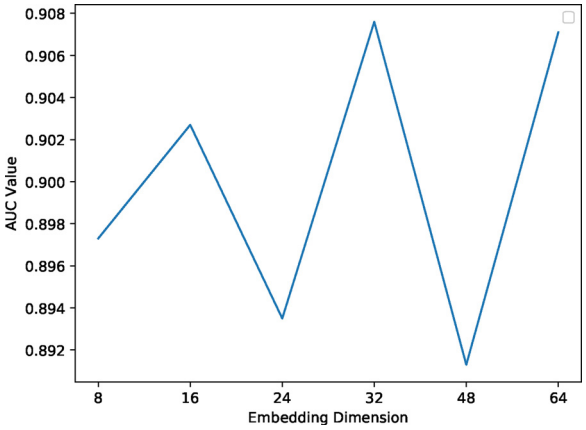

Fig. 6. Average AUC value on different dimensions of the embedding.

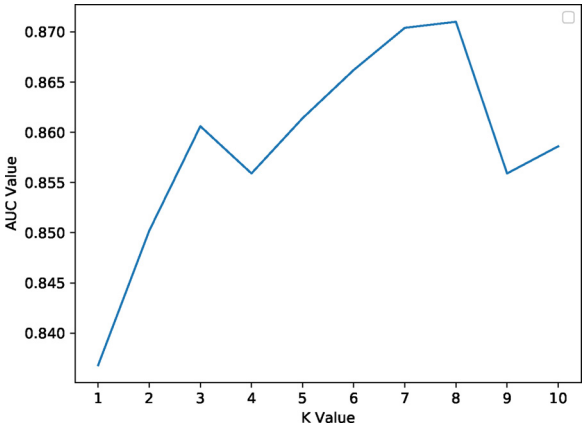

Fig. 7. Average AUC value on different k value.

It is found that 16 out of 20 lncRNAs including CCAT1, CCAT2, HOTTIP, SPRY4-IT1, GAS5, PANDAR, AFAP1-AS1, CASC16, MALAT1, EGOT, LINC-ROR, NBAT1, BCAR4, KCNQ1OT1, LSINCT5 and XIST have been verified by recent publications. The expression of CCAT1 ranked at top 1 in breast cancer tissues is significantly higher than normal tissues (Zhang et al., 2015). In addition, the expression of CCAT2 ranked at top 3 is up-regulated in breast cancer tissues (Deng et al., 2017). It has been discovered that comparing to adjacent non-tumor tissues, the expression of HOTTIP ranked at top 4 is higher in breast cancer tissues (Han et al., 2019). It has been proved that the lncRNA SPRY4-IT1 ranked at top 5 is highly expressed in breast cancer cells (Wu et al., 2018). It has been discovered that GAS5 ranked at top 6 is down-regulated in tamoxifen resistant breast cancer cells (Gu et al., 2018). It has been found that the expression of PANDAR ranked at top 7 is higher in breast cancer tissues and cells than normal tissues and the normal mammary epithelial cell line (Li et al., 2019b). The expression of AFAP1-AS1 ranked at top 8 is up-regulated in human breast cancer tissue and malignancy status. In addition, the high expression of AFAP1-AS1 has the poor prognosis in breast cancer patients (Ma et al., 2019). Recent literatures show CASC16 ranked at top 10 increased the

Table 2
The top 20 predicted results of breast cancer-related lncRNAs based on GAMCLDA.

Rank	LncRNA	Evidence
1	CCAT1	PMID: 26464701
2	SOX2-OT	unknown
3	CCAT2	PMID: 28531944
4	HOTTIP	PMID: 30676763
5	SPRY4-IT1	PMID: 30104400
6	GAS5	PMID: 29969658
7	PANDAR	PMID: 31104011
8	AFAP1-AS1	PMID: 29974352
9	MIR124-2HG	unknown
10	CASC16	PMID: 31655495
11	CDKN2B-AS1	unknown
12	MALAT1	PMID: 28915533
13	EGOT	PMID: 26159853
14	LINC-ROR	PMID: 29041978
15	NBAT1	PMID: 26378045
16	BCAR4	PMID: 26016563
17	KCNQ1OT1	PMID: 30157476
18	LSINCT5	PMID: 21532345
19	XIST	PMID: 29550489
20	BCYRN1	unknown

risk of lymph node metastasis in breast cancer patients (Sun et al., 2019). It has been discovered that the expression of MALAT1 ranked at top 12 increases in triple-negative breast cancer tissues and cell lines (Zuo et al., 2017). It has been proved that the expression of EGOT ranked at top 13 in breast cancer is lower than the adjacent non-cancerous tissues. In addition, the low expression of EGOT expression is significantly correlated with tumor size, lymph node metastasis and Ki-67 expression (Xu et al., 2015). It has been proved that Linc-RoR ranked at top 14 can act as an onco-lncRNA to promote estrogen-independent growth of ER+ breast cancer (Peng et al., 2017b). It has been discovered that NBAT1 ranked at top 15 is downregulated in invasive breast cancer (Hu et al., 2015). It has been found that the expression of BCAR4 ranked at top 16 is dramatic upregulated in breast cancer tissues (Xing et al., 2015). It has been demonstrated that KCNQ1OT1 ranked at top 17 is highly expressed in BRCA tissues and cells (Feng et al., 2018). It has been proved that the expression of LSINCT5 ranked at top 18 is higher in the breast cancer cell lines than the normal breast epithelial cell line (Silva et al., 2011). The lncRNA XIST ranked at top 19 is significantly down-regulated in breast cancer tissues and cell lines (Zheng et al., 2018). In addition, some interesting lncRNAs are also found such as SOX2-OT, MIR124-2HG, CDKN2B-AS1 and BCYRN1. Although the function of these lncRNAs is still unknown. It deserves biologists to validate it by using experimental methods.

4. Conclusion

Accumulating evidences have indicated that identifying lncRNA-disease associations can not only help biologists to understand the complex diseases at the molecular level, but also contribute to disease diagnosis, treatment, prognosis and prevention (Peng et al., 2017a; Lan et al., 2018b). In this article, we propose a computational method, GAMCLDA, to infer associations between lncRNAs and diseases based on graph auto-encoder matrix completion. In our method, it employs GCNs to encode the feature vectors of lncRNAs and diseases. Further, inner product is utilized to predict potential lncRNA-disease associations. In addition, the cost-sensitive learning is employed to solve the problem of data imbalance based on multilayer feed forward neural networks. Compared with other state-of-the-art methods (MFLDA, SIMCLDA, RWRlncD and HGLDA), our method achieves a better performance in AUC value, which reaches to 0.9071. In addition, GAMCLDA outperforms other methods in the *de novo* test, which obtained an AUC of 0.8710. Furthermore, the case study shows that GAMCLDA can effectively identify potential lncRNA-disease interactions.

Recently, predicting disease related biological entities interactions have attracted more and more attention such as miRNA-disease association (Chen et al., 2018, 2019b), drug-target interaction (Chen et al., 2016b; Lan et al., 2016b) and synergistic drug combination (Chen et al., 2016a).The GAMCLDA also can be used to predict these interactions. In addition, motived by (Liu et al., 2016, 2018; Li et al., 2017; Zheng et al., 2019), the feature engineering and ensemble learning can be integrated to improve the performance of GAMCLDA.

Author statement

Ximin Wu: Conceptualization, Methodology, Writing - Original Draft.:Wei Lan: Conceptualization, Methodology, Writing- Reviewing and Editing. Qingfeng Chen: Conceptualization, Methodology. Yi Dong: Visualization, Validation. Jin Liu: Validation. Wei Peng: Writing-Reviewing and Editing.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Nos. 61702122, 61963004 and 61972185), the Natural Science Foundation of Guangxi (Nos. 2017GXNSFDA198033 and 2018GXNSFBA281193), the Key Research and Development Plan of Guangxi (No. AB17195055), the foundation of Guangxi University (Nos. 20190240 and XBZ180479), the Natural Science Foundation of Yunnan Province of China (2019FA024), the scientific Research Foundation of Hunan Provincial Education Department (No.18B469), the Hunan Provincial Science and Technology Program (No. 2018WK4001).

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.compbiolchem.2020.107282.

References

Ahmed, N.M., Chen, L., Wang, Y., Li, B., Li, Y., Liu, W., 2018. Deepeye: link prediction in dynamic networks based on non-negative matrix factorization. Big Data Mining and Analytics 1, 19–33.
Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., Cui, Q., 2012. Lncrnadisease: a database for long-non-coding rna-associated diseases. Nucleic acids research 41, D983–D986.
Chen, Q., Lai, D., Lan, W., Wu, X., Chen, B., Chen, Y.P.P., Wang, J., 2019. Ildmsf: Inferring associations between long non-coding rna and disease based on multi-similarity fusion. IEEE/ACM Transactions on Computational Biology and Bioinformatics 1-1.
Chen, X., 2015. Predicting lncrna-disease associations and constructing lncrna functional similarity network based on the information of mirna. Scientific reports 5, 1–11.
Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., Yan, G., 2016a. Nllss: predicting synergistic drug combinations based on semi-supervised learning. PLoS computational biology 12.
Chen, X., Sun, Y.Z., Guan, N.N., Qu, J., Huang, Z.A., Zhu, Z.X., Li, J.Q., 2019a. Computational models for lncrna function prediction and functional similarity calculation. Briefings in functional genomics 18, 58–82.
Chen, X., Wang, L., Qu, J., Guan, N.N., Li, J.Q., 2018. Predicting mirna-disease association based on inductive matrix completion. Bioinformatics 34, 4256–4265.
Chen, X., Xie, D., Zhao, Q., You, Z.H., 2019b. Micrornas and complex diseases: from experimental results to computational models. Briefings in bioinformatics 20, 515–539.
Chen, X., Yan, C.C., Zhang, X., You, Z.H., 2017. Long non-coding rnas and complex diseases: from experimental results to computational models. Briefings in bioinformatics 18, 558–576.
Chen, X., Yan, C.C., Zhang, X., Zhang, X., Dai, F., Yin, J., Zhang, Y., 2016b. Drug-target interaction prediction: databases, web servers and computational models. Briefings in bioinformatics 17, 696–712.
Chen, X., Yan, G.Y., 2013. Novel human lncrna-disease association inference based on lncrna expression profiles. Bioinformatics 29, 2617–2624.
Chen, X., You, Z.H., Yan, G.Y., Gong, D.W., 2016c. Irwrlda: improved random walk with restart for lncrna-disease association prediction. Oncotarget 7, 57919–57931.
Congrains, A., Kamide, K., Oguro, R., Yasuda, O., Miyata, K., Yamamoto, E., Kawai, T.,

Kusunoki, H., Yamamoto, H., Takeya, Y., et al., 2012. Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of anril and cdkn2a/b. Atherosclerosis 220, 449–455.
Consortium, I.H.G.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860.
Deng, X., Zhao, Y., Wu, X., Song, G., 2017. Upregulation of ccat2 promotes cell proliferation by repressing the p15 in breast cancer. Biomedicine & Pharmacotherapy 91, 1160–1166.
Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al., 2012. The gencode v7 catalog of human long noncoding rnas: analysis of their gene structure, evolution, and expression. Genome research 22, 1775–1789.
Feng, W., Wang, C., Liang, C., Yang, H., Chen, D., Yu, X., Zhao, W., Geng, D., Li, S., Chen, Z., Sun, M., 2018. The dysregulated expression of kcnq1ot1 and its interaction with downstream factors mir-145/ccne2 in breast cancer cells. Cellular Physiology and Biochemistry 49, 432–446.
Fu, G., Wang, J., Domeniconi, C., Yu, G., 2017. Matrix factorization-based data fusion for the prediction of lncrna-disease associations. Bioinformatics 34, 1529–1537.
Godinho, M.F., Wulfkuhle, J.D., Look, M.P., Sieuwerts, A.M., Sleijfer, S., Foekens, J.A., Petricoin III, E.F., Dorssers, L.C., van Agthoven, T., 2012. Bcar4 induces antioestrogen resistance but sensitises breast cancer to lapatinib. British journal of cancer 107, 947.
Gu, J., Wang, Y., Wang, X., Zhou, D., Shao, C., Zhou, M., He, Z., 2018. Downregulation of lncrna gas5 confers tamoxifen resistance by activating mir-222 in breast cancer. Cancer Letters 434, 1–10.
Guttman, M., Rinn, J.L., 2012. Modular regulatory principles of large non-coding rnas. Nature 482, 339.
Han, S., Jin, X., Liu, Z., Xing, F., Han, Y., Yu, X., He, G., Qiu, F., 2019. The long noncoding rna hottip promotes breast cancer cell migration, invasiveness, and epithelial-mesenchymal transition via the wnt-ß-catenin signaling pathway. Biochemistry and Cell Biology 97, 655–664.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision 1026–1034.
Hu, P., Chu, J., Wu, Y., Sun, L., Lv, X., Zhu, Y., Li, J., Guo, Q., Gong, C., Liu, B., 2015. Nbat1 suppresses breast cancer metastasis by regulating dkk1 via prc2. Oncotarget 6, 32410–32425.
Jiang, Q., Wang, J., Wu, X., Ma, R., Zhang, T., Jin, S., Han, Z., Tan, R., Peng, J., Liu, G., et al., 2015. Lncrna2target: a database for differentially expressed genes after lncrna knockdown or overexpression. Nucleic acids research 43, D193–D196.
Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
Koren, Y., Bell, R., Volinsky, C., 2009. Matrix factorization techniques for recommender systems. IEEE Computer 42, 30–37.
Kukar, M., Kononenko, I., et al., 1998. Cost-sensitive learning with neural networks. in: ECAI, pp. 445–449.
van Laarhoven, T., Nabuurs, S.B., Marchiori, E., 2011. Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics 27, 3036–3043.
Lan, W., Chen, Q., Li, T., Yuan, C., Mann, S., Chen, B., 2014. Identification of important positions within mirnas by integrating sequential and structural features. Current Protein and Peptide Science 15, 591–597.
Lan, W., Huang, L., Lai, D., Chen, Q., 2018a. Identifying interactions between long noncoding rnas and diseases based on computational methods. Computational Systems Biology 205–221.
Lan, W., Li, M., Zhao, K., Liu, J., Wu, F.X., Pan, Y., Wang, J., 2016a. Ldap: a web server for lncrna-disease association prediction. Bioinformatics 33, 458–460.
Lan, W., Wang, J., Li, M., Liu, J., Li, Y., Wu, F.X., Pan, Y., 2016b. Predicting drug-target interaction using positive-unlabeled learning. Neurocomputing 206, 50–57.
Lan, W., Wang, J., Li, M., Liu, J., Wu, F.X., Pan, Y., 2018b. Predicting microrna-disease associations based on improved microrna and disease similarities. IEEE/ACM Transactions on Computational Biology and Bioinformatics 15, 1774–1782.
Lan, W., Wang, J., Li, M., Peng, W., Wu, F., 2015. Computational approaches for prioritizing candidate disease genes based on ppi networks. Tsinghua Science & Technology 20, 500–512.
Li, J.H., Liu, S., Zhou, H., Qu, L.H., Yang, J.H., 2014a. starbase v2.0: decoding mirna-cerna, mirna-ncrna and protein-rna interaction networks from large-scale clip-seq data. Nucleic acids research 42, D92–D97.
Li, M., Zheng, R., Li, Y., Wu, F.X., Wang, J., 2017. Mgt-sm: a method for constructing cellular signal transduction networks. IEEE/ACM transactions on computational biology and bioinformatics 16, 417–424.
Li, Y., He, Y., Han, S., Liang, Y., 2019a. Identification and functional inference for tumor-associated long non-coding rna. IEEE/ACM transactions on computational biology and bioinformatics 16, 1288–1301.
Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., Cui, Q., 2014b. Hmdd v2.0: a database for experimentally supported human microrna and disease associations. Nucleic acids research 42, D1070–D1074.
Li, Y., Su, X., Pan, H., 2019b. Inhibition of lncrna pandar reduces cell proliferation, cell invasion and suppresses emt pathway in breast cancer. Cancer Biomarkers 1–8.
Liu, J., Li, M., Lan, W., Wu, F.X., Pan, Y., Wang, J., 2016. Classification of alzheimer's disease using whole brain hierarchical network. IEEE/ACM transactions on computational biology and bioinformatics 15, 624–632.
Liu, J., Pan, Y., Li, M., Chen, Z., Tang, L., Lu, C., Wang, J., 2018. Applications of deep learning to mri images: A survey. Big Data Mining and Analytics 1, 1–18.
Lu, C., Yang, M., Luo, F., Wu, F.X., Li, M., Pan, Y., Li, Y., Wang, J., 2018. Prediction of lncrna-disease associations based on inductive matrix completion. Bioinformatics 34, 3357–3364.
Lu, Z., BRETONNEL COHEN, K., Hunter, L., 2007. Generif quality assurance as summary revision. Biocomputing 2007 269–280.
Ma, D., Cheng, C., Jun, W., Honglei, W., Duoming, W., 2019. Up-regulated lncrna afap1-as1 indicates a poor prognosis and promotes carcinogenesis of breast cancer. Breast Cancer 26, 74.
Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., Gao, Y., Guo, M., Yue, M., Wang, L., et al., 2015. Lnc2cancer: a manually curated database of experimentally supported lncrnas associated with various human cancers. Nucleic acids research 44, D980–D985.
Pasmant, E., Sabbagh, A., Vidaud, M., Bièche, I., 2011. Anril, a long, noncoding rna, is an unexpected major hotspot in gwas. The FASEB Journal 25, 444–448.
Peng, W., Lan, W., Zhong, J., Wang, J., Pan, Y., 2017a. A novel method of predicting microrna-disease associations based on microrna, disease, gene and environment factor networks. Methods 124, 69–77.
Peng, W.X., Huang, J.G., Yang, L., Gong, A.H., Mo, Y.Y., 2017b. Linc-ror promotes mapk/erk signaling and confers estrogen-independent growth of breast cancer. Molecular Cancer 16, 161.
Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., Furlong, L.I., 2015. Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. Database 2015, bav028.
Ramlatchan, A., Yang, M., Liu, Q., Li, M., Wang, J., Li, Y., 2018. A survey of matrix completion methods for recommendation systems. Big Data Mining and Analytics 1, 308–323.
Silva, J.M., Boczek, N.J., Berres, M.W., Ma, X., Smith, D.I., 2011. Lsinct5 is over expressed in breast and ovarian cancer and affects cellular proliferation. Rna Biology 8, 496–505.
Sun, J., Shi, H., Wang, Z., Zhang, C., Liu, L., Wang, L., He, W., Hao, D., Liu, S., Zhou, M., 2014. Inferring novel lncrna-disease associations based on a random walk model of a lncrna functional similarity network. Molecular BioSystems 10, 2074–2081.
Sun, Y., Chen, P., Wu, J., Xiong, Z., Liu, Y., Liu, J., Li, H., Li, B., Jin, T., 2019. Association of polymorphisms in loc105377871 and casc16 with breast cancer in the northwest chinese han population. Journal of Gene Medicine 1–7.
Torre, L.A., Islami, F., Siegel, R.L., Ward, E.M., Jemal, A., 2017. Global cancer in women: Burden and trends. Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 26, 444–457.
Wang, K.C., Chang, H.Y., 2011. Molecular mechanisms of long noncoding rnas. Molecular cell 43, 904–914.
Wapinski, O., Chang, H.Y., 2011. Long noncoding rnas and human disease. Trends in cell biology 21, 354–361.
Wu, H., Wang, Y., Chen, T., Li, Y., Wang, H., Zhang, L., Chen, S., Wang, W., Yang, Q., Chen, C., 2018. The n-terminal polypeptide derived from vmip-ii exerts its anti-tumor activity in human breast cancer by regulating lncrna spry4-it1. Bioscience reports 38 BSR 20180411.
Xing, Z., Park, P.K., Lin, C., Yang, L., 2015. Lncrna bcar4 wires up signaling transduction in breast cancer. Rna Biology 12, 681–689.
Xu, S.P., Zhang, J.F., Sui, S.Y., Bai, N.X., Gao, S., Zhang, G.W., Shi, Q.Y., You, Z.L., Zhan, C., Pang, D., 2015. Downregulation of the long noncoding rna egot correlates with malignant status and poor prognosis in breast cancer. Tumor Biology 36, 1–6.
Zhang, X.F., Liu, T., Li, Y., Li, S., 2015. Overexpression of long non-coding rna ccat1 is a novel biomarker of poor prognosis in patients with breast cancer. International Journal of Clinical & Experimental Pathology 8, 9440–9445.
Zhang, Z., Hao, H., Zhang, C., Yang, X., He, Q., Lin, J., 2012. Evaluation of novel gene uca1 as a tumor biomarker for the detection of bladder cancer. Zhonghua yi xue za zhi 92, 384–387.
Zheng, R., Li, M., Chen, X., Zhao, S., Wu, F., Pan, Y., Wang, J., 2019. An ensemble method to reconstruct gene regulatory networks based on multivariate adaptive regression splines. IEEE/ACM transactions on computational biology and bioinformatics 99 1–1.
Zheng, R., Lin, S., Guan, L., Yuan, H., Zhang, R., 2018. Long non-coding rna xist inhibited breast cancer cell growth, migration, and invasion via mir-155/cdx1 axis. Biochemical & Biophysical Research Communications 498, 1002.
Zhou, M., Wang, X., Li, J., Hao, D., Wang, Z., Shi, H., Han, L., Zhou, H., Sun, J., 2015. Prioritizing candidate disease-related long non-coding rnas by walking on the heterogeneous lncrna and disease network. Molecular BioSystems 11, 760–769.
Zuo, Y., Li, Y., Zhou, Z., Ma, M., Fu, K., 2017. Long non-coding rna malat1 promotes proliferation and invasion via targeting mir-129-5p in triple-negative breast cancer. Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie 95, 922.