LDICDL: LncRNA-disease association identification based on Collaborative Deep Learning

Article *in* IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM · October 2020

DOI: 10.1109/TCBB.2020.3034910

8 authors, including:

Wei Lan
Guangxi University
48 PUBLICATIONS **878** CITATIONS
SEE PROFILE

Qingfeng Chen
University of Technology Sydney
87 PUBLICATIONS **445** CITATIONS
SEE PROFILE

Baoshan Chen
Guangxi University
158 PUBLICATIONS **2,537** CITATIONS
SEE PROFILE

Jin Liu
Central South University
44 PUBLICATIONS **1,362** CITATIONS
SEE PROFILE

LDICDL: LncRNA-disease association identification based on Collaborative Deep Learning

Wei Lan, Dehuan Lai, Qingfeng Chen, Ximin Wu, Baoshan Chen, Jin Liu, Jianxin Wang, Yi-Ping Phoebe Chen

Abstract—It has been proved that long noncoding RNA (lncRNA) plays critical roles in many human diseases. Therefore, inferring associations between lncRNAs and diseases can contribute to disease diagnosis, prognosis and treatment. To overcome the limitation of traditional experimental methods such as expensive and time-consuming, several computational methods have been proposed to predict lncRNA-disease associations by fusing different biological data. However, the prediction performance of lncRNA-disease associations identification need to be improved. In this study, we propose a computational model (named LDICDL) to identify lncRNA-disease associations based on collaborative deep learning. It uses an automatic encoder to denoise multiple lncRNA feature information and multiple disease feature information, respectively. Then, the matrix decomposition algorithm is employed to predict the potential lncRNA-disease associations. In addition, to overcome the limitation of matrix decomposition, the hybrid model is developed to predict associations between new lncRNA (or disease) and diseases (or lncRNA). The ten-fold cross validation and de novo test are applied to evaluate the performance of method. The experimental results show LDICDL outperforms than other state-of-the-art methods in prediction performance.

Index Terms—lncRNA-disease associations, matrix factorization, stacked denoising autoencoder.

1 Introduction

It is well known that biological genetic information is primarily stored in protein-coding genes, and RNA is the intermediary between DNA sequences and proteins [1]. With the development of human genetic engineering, 2% of the genes have been confirmed to be protein-coding genes, and the remaining 98% of the genes have not or few protein encoding abilities [2]. These genes are usually transcribed into non-coding RNAs [3]. Non-coding RNAs have been regarded as the noise of genomic transcription for a long time [4], [5]. However, recent studies have shown that they play important regulatory roles in many biological processes of organism. In particular, long non-coding RNAs (lncRNAs) which are greater than 200 nucleotides in length have been unveiled to be related to a broad range of diseases [6]. For example, it has been found that HOTAIR is overexpressed in breast cancer, colon cancer, liver cancer and gastrointestinal stromal tumors [7]. Therefore, identifying lncRNA-disease association is helpful for biologist not only in understanding the underlying mechanisms of disease, but also disease prevention diagnosis and treatment [8], [9].

Many biological experimental studies have been developed to discover potential lncRNA-disease associations [10]. Although these methods can exactly discover lncRNA-disease association, they also have some limitations such as time-consuming and expensive. With the development of high-throughput sequencing technology, a large amount of lncRNA related data, such as the sequence, structure, function and expression, has been generated [11], [12]. Thus, many computation-based algorithms have been proposed to overcome these limitations for potential lncRNA-disease associations prediction [13]. These computational methods can be classified into two categories: (1) network-based methods that use similarity network to predict lncRNA-disease associations. For example, sun et al [14] proposed a computational method, RWRlncD, to identify lncRNA-disease associations based on lncRNA functional similarity network and the random walk with restart method. Chen et al [15] presented an algorithm, IRWRLDA, to predict lncRNA-disease associations in terms of lncRNA similarity network. They used various measures to calculate lncRNA similarity, and IRWRLDA could be used to diseases without any lncRNA-disease association. Zhou et al [16] developed a model, RWRHLD, for lncRNA-disease asso-

- *Wei Lan is School of Computer, Electronic and Information, Guangxi University, Nanning, Guangxi, 530004, China. E-mail: lanwei@gxu.edu.cn*
- *Dehuan Lai is School of Computer, Electronic and Information, Guangxi University, Nanning, Guangxi, 530004, China. E-mail: laidehuan@st.gxu.edu.cn*
- *Qingfeng Chen is School of Computer, Electronic and Information and State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, Guangxi University, Nanning, Guangxi, 530004, China. E-mail: qingfeng@gxu.edu.cn*
- *Ximin Wu is School of Computer, Electronic and Information, Guangxi University, Nanning, Guangxi, 530004, China. E-mail: wuximin@st.gxu.edu.cn*
- *Baoshan Chen is State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, Guangxi University, Nanning, Guangxi, 530004, China. E-mail:chenyaoj@gxu.edu.cn*
- *Jin Liu is Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, Hunan, 410083, China. E-mail:liujin06@mail.csu.edu.cn*
- *Jianxin Wang is Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, Hunan, 410083, China. E-mail: jxwang@mail.csu.edu.cn*
- *Yi-Ping Phoebe Chen is Department of Computer Science and Information Technology, La Trobe University, Melbourne Victoria 3086, Australia. E-mail:phoebe.chen@latrobe.edu.au*

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

ciation predictions by integrating three networks into one heterogeneous network. According to constructing a multi-level network of lncRNA-disease, Yao et al [17] proposed an algorithm, LncPriCNet, to prioritize candidate for lncRNA-disease associations. (2) machine learning-based methods that prioritize candidate lncRNAs by training disease related known lncRNAs and unknown lncRNAs. Lan et al [18] developed an online web server (LDAP) to identify new associations between lncRNAs and diseases based on positive-unlabeled (PU) learning. Chen et al [19] proposed a method (LRLSLDA) to infer lncRNA-disease associations based on the semi-Supervised learning. Wu et al [20] presented a computational method(GAMCLDA) to predict lncRNA-disease associations based on graph autoencoder matrix completion. Fu et al [21] developed a computational model (MFLDA) to predict the associations between lncRNA and disease based on multiple data fusion and matrix factorization (MF). Lu et al [22] presented a computational model (SIMCLDA) to prioritize candidate lncRNAs based on inductive matrix completion. Chen et al [23] proposed a computational framework, ILDMSF, for lncRNA-disease association identification based on multiple kernel fusion and Support Vector Machine (SVM).

These methods have achieved good performance in predicting the associations between lncRNAs and diseases. However, they do not make full use of the known lncRNA characteristic data and disease characteristic data, so there are limitations on the accuracy and prediction performance [24], [25]. This paper proposes a novel computational framework (LDICDL) to predict LncRNA-disease associations. It uses an automatic encoder to denoise multiple lncRNA feature information data and multiple disease feature information data. In addition, the matrix factorization algorithm is employed to predict the potential lncRNA-disease associations. Further, the hybrid model based on stacked denoising autoencoder and matrix factorization is developed to overcome the limitation of matrix factorization for de novo prediction. The experimental results demonstrate our method has better performance than other state-of-the-art methods.

2 METHODS

The task of identifying LncRNA-disease associations can be viewed as taking implicit feedback as the training and test data. The LncRNA-disease association matrix is represented by a matrix LD_{m*n} ,where m and n denotes the number of lncRNAs and diseases, respectively. The element of LD(i,j) is equal to 1 if lncRNA i is associated with disease j, otherwise 0. The lncRNA information is integrated into lncRNA feature matrix LF_{m*t}, where t denotes the number of features. The disease information is merged into disease feature matrix DF_{n*s}, where s denotes the number of disease features.

2.1 Stacked Denoising Autoencoder

The stacked denoising autoencoder (SDAE) is a kind of feedforward neural network which is widely used in recommend system [26]. In LDICDL, the SDAE is employed to select lncRNAs and diseases feature information, respectively. The original features of lncRNA and disease are t and s dimensions, respectively. In final, the lncRNAs and diseases feature information are reduced into k dimensions by using SDAE. The mini-batch gradient descent algorithm is used to train SDAE with the batch size=60.

2.2 Matrix Factorization for lncRNA-disease prediction

In the LncRNA-disease association matrix LD, the element $LD(i,j)$ is defined as follow:

$$LD(i,j) = \begin{cases} 1, & \text{if lncRNA i is related with disease j} \\ 0, & \text{if lncRNA i is not related with disease j} \end{cases} \quad (1)$$

Therefore, the loss function of matrix factorization for LncRNA-disease association prediction is defined as follow:

$$Loss = \Sigma_{i,j}\alpha_{i,j}(LD(i,j) - L(i,:)\cdot D(j,:)^T)^2 + \gamma(\Sigma_i\|L(i,:)\|^2 + \Sigma_j\|D(j,:)\|^2) \quad (2)$$

where γ denotes the regulation parameter. $L(i,:)$ and $D(j,:)$ denote lncRNA i subspace feature and disease j subspace feature, respectively. $\alpha_{i,j}$ is the parameter to show the confidence between lncRNA i and disease j where $\alpha_{i,j}$=1+$\theta(LD(i,j))$. $\|.\|^2$ denotes 2-norm.

2.3 Matrix Factorization with Implicit Feedback for LncRNA-disease association prediction

Considering that the lncRNA-disease associations prediction performance with matrix factorization is poor for the new lncRNA or disease, which called cold start problem [13], [27], the hybrid model is proposed to predict lncRNA and disease associations by combining matrix factorization with stocked denoising autoencoder. In our method, predicting association of new lncRNA or disease relies on the biological features of lncRNA and disease. The structure of hybrid model with three hidden layer of lncRNA is show in Figure 1. X_{input_l} is the input layer for lncRNA features (i.e. LF) and X_{encode_l} is the lncRNA features encoding. X_{out_l} is the output layer of lncRNA features.

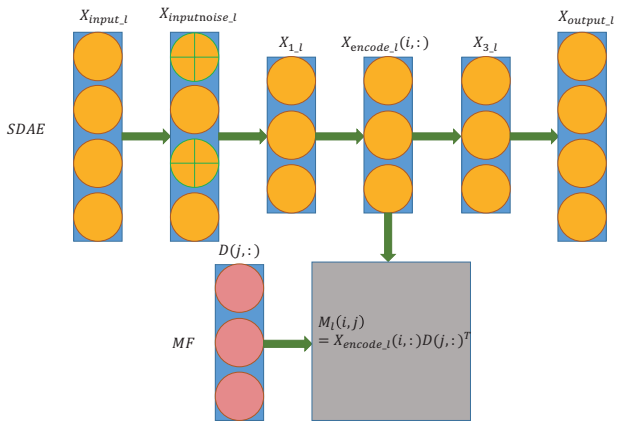

Fig. 1. The overview of hybrid model of lncRNA.

The loss function of hybrid model based on lncRNA features is defined as follow:

$$Loss = \Sigma_{i,j}\alpha_{i,j}(LD(i,j) - L(i,:)\cdot D(j,:)^T)^2 + \gamma(\Sigma_i\|L(i,:)\|^2 + \Sigma_j\|D(j,:)\|^2) + \gamma_l(\|L - X_{encode\_l}\|^2) + \gamma_n(\|X_{input\_l} - X_{out\_l}\|^2) + \Sigma_{layers}\gamma_w\|W\|^2 \quad (3)$$

where γ, γ_l, γ_n and γ_w denote regularization parameters. W denotes the weight matrix.

The loss function is minimized by block coordinate decent [28]. The $L(i,:)$ is updated in term of Eq.4 below in training step:

$$L(i,:) \leftarrow LD(i,:)C^{(i)}D(\gamma I + D^T C^{(i)} D)^{-1} \quad (4)$$

where $C^{(i)}$ is a diagonal matrix where $C^{(i)}$ (j,j)=$\alpha_{i,j}$.

The $D(:,j)$ is updated in term of Eq.5 below in training step:

$$D(:,j) \leftarrow LD(:,j)^T\tilde{C}^{(j)}L(\gamma I + L^T\tilde{C}^{(j)}L)^{-1} \quad (5)$$

where $\tilde{C}^{(j)}$ is a diagonal matrix where $\tilde{C}^{(j)}(i,i)$=$\alpha_{i,j}$.

For disease, the structure with three hidden layers is show in Figure 2. X_{input_d} is the input layer for disease features (i.e. DF) and X_{encode_d} represents the disease features encoding. X_{out_d} is the output layer of disease features.

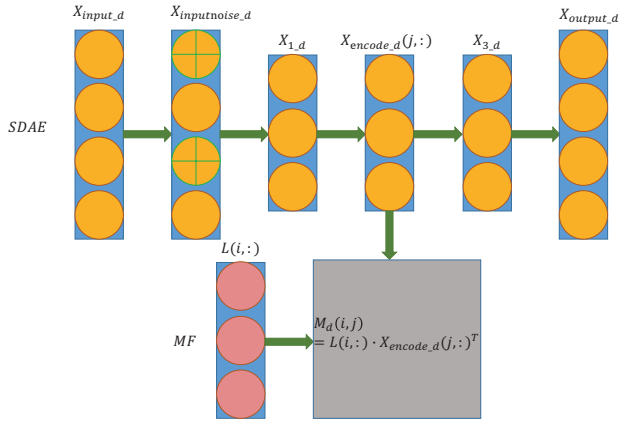

Fig. 2. The overview of hybrid model of disease.

The loss function of hybrid model based on disease feature information is defined as follow:

$$Loss = \Sigma_{i,j}\alpha_{i,j}(LD(i,j) - L(i,:)\cdot D(j,:)^T)^2 + \gamma(\Sigma_i\|L(i,:)\|^2 + \Sigma_j\|D(j,:)\|^2) + \gamma_d(\| D - X_{encode\_d}\|^2) + \gamma_n(\| X_{input\_d} - X_{out\_d}\|^2) + \Sigma_{layers}\gamma_w \| W \|^2 \quad (6)$$

where γ, γ_d, γ_n and γ_w denote regularization parameters. W denotes the weight matrix.

The final predicted score matrix S is calculated as follow:

$$S(i,j) = \frac{M_l(i,j) + M_d(i,j)}{2} \quad (7)$$

$$M_l(i,j) = X_{encode\_l}(i,:)\cdot D(j,:)^T \quad (8)$$

$$M_d(i,j) = L(i,:)\cdot X_{encode\_d}(j,:)^T \quad (9)$$

where $S(i,j)$ denotes the score between lncRNA i and disease j. X_{encode_l} and X_{encode_d} denote the sub-feature matrix of lncRNA and disease which are obtained by SDAE based on lncRNA and disease feature information, respectively. L and D denote the sub-feature matrix of lncRNA and disease obtained from matrix factorization.

The whole workflow of LDICDL is shown in Figure 3. In the first step, the lncRNA-disease association matrix is decomposed to lncRNA feature subspace and disease feature information is encoded by using SDAE. Meanwhile, the lncRNA-disease association matrix is decomposed to disease feature subspace and lncRNA feature information is encoded by SDAE. Then, the lncRNA-disease association score is predicted based on lncRNA feature matrix and disease encode matrix, and disease feature matrix and lncRNA encode matrix, respectively. Last, the final score of lncRNA-disease association is calculated by averaging the scores. The Block coordinate decent is used to minimize the loss function. Firstly, the L and D are updated by equation 4 and 5, respectively. Then, the parameters in the SDAE are updated using gradient decent with mini-batch. The mean square errors of output and encoding are used to adjust the gradient. It repeats the former steps for t times.

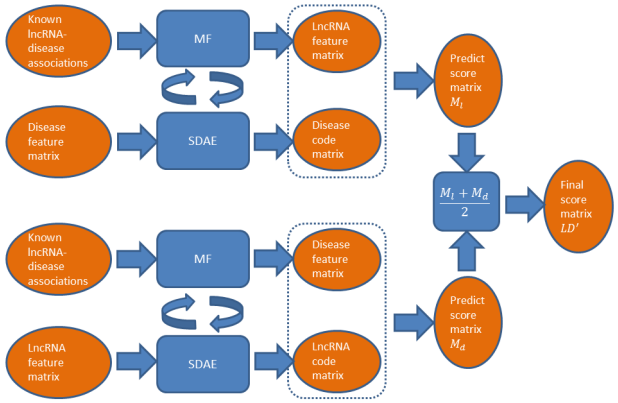

Fig. 3. The workflow of LDICDL.

3 Results

3.1 Datasets

The lncRNA-gene associations are downloaded from lncRNA2target [29] and lncRNA-gene function associations are collected from GeneRIF [30]. They are pre-processed using Open Biomedical Annotator [31]. The lncRNA-miRNA associations and disease-miRNA associations are downloaded

from starBase v2.0 [32] and HMDD [33], respectively. The disease-gene associations are downloaded from DisGeNET [34]. In final, 2697 associations between 240 lncRNAs and 412 diseases are obtained as gold-standard dataset. In addition, 6066-dimensions feature information of each lncRNA from lncRNA-related data and 10621-dimensions feature information of each disease from disease-related data are collected, respectively.

3.2 Performance evaluation

The ten-fold cross validation and *de novo* test are employed to evaluate the performance of different methods. In ten-fold cross validation, all known associations between lncRNAs and diseases are divided into ten folds randomly. In each test, one fold is selected as the test samples and other nine folds are treated as training samples. All known associations in test samples are removed by turns and all other known associations in training samples are used to train model. Then, the prediction algorithm is carried out to predict the scores of test samples and candidate samples. In the *de novo* test, for disease i, all known associations are removed as test samples, while all known associations between lncRNAs and other diseases are considered as training samples. Then, the scores of associations between lncRNAs and disease i are calculated by prediction method. After that, the scores of test and candidate samples are ranked with descending order and observe whether its ranking is greater than a specific threshold. If the rank of test sample is greater than the threshold, it is considered as true positive, otherwise false negative. If the rank of candidate sample is greater than the threshold, it is viewed as false positive, otherwise true negative. Further, the true positive rate (TPR) and false positive rate (FPR) are calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

$$FPR = \frac{FP}{FP + TN} \tag{11}$$

where TP denotes the number of true positive samples, TN denotes the number of true negative samples, FP denotes the number of false positive samples, and FN denotes the number of false negative samples. The receiver operating characteristic (ROC) curve is draw based on TPR and FPR at different thresholds and the Area under of ROC (AUC) is calculated to evaluate the performance of method. If the AUC equals to 1, it denotes that this method has perfect performance. If the AUC equals to 0.5, it denotes that the prediction of model is uncertain.

In addition, the precision and recall are also calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

where precision denotes the proportion of the true positive samples with rankings higher than the special threshold in the predicted positive samples, recall denotes the proportion of true positive samples whose ranking is lower than the special threshold in the whole positive samples. Then, Precision-Recall (PR) curve is plotted based on precision and recall. Finally, the area under of PR (AUPR) are computed to evaluate the performance of method.

3.3 Ten-fold cross validation

In order to evaluate the performance of LDICDL, the ten-fold cross validation is applied in the experiment. We compare LDICDL with two state-of-the-art methods based on matrix completion (SIMCLDA [22] and MFLDA [21]). The performance of different methods is evaluated in term of AUC. It can be observed from Figure 4 that LDICDL achieves the AUC of 0.8651, which is significantly higher than other methods (SIMCLDA 0.8259 and MFLDA 0.6430). It demonstrates that our method has higher performance than other methods. In addition, the AUPR is also utilized to compare the performance of different methods as shown in Figure 5. The AUPR of LDICDL is 0.0306 in contrast to 0.0227 and 0.0051 with SIMCLDA and MFLDA, respectively. It proves that our method is more effective than other two methods. Figure 6 shows the number of correctly retrieved known lncRNAdisease associations. It can be found that LDICDL outperforms other methods from top 10 to top 50 associations.

To prove our model can obtain deep latent representation of features, we conduct the experimental comparison between LDICDL and three classical feature extraction methods including Nonnegative Matrix Factorization(NMF), Principal Component Analysis(PCA) and Latent Dirichlet Allocation (LDA). The comparison result on different feature extraction methods is shown in Figure 7. It can be found from the result that LDICDL which is based on using the stacked denoising autoencoder outperforms other methods. Moreover, in order to show the effect of the combination of MF and SDAE, we compare it with MF and SDAE, respectively. The result is shown in Figure 8. The result demonstrates that the combination of MF and SDAE outperforms than single method (MF or SDAE). We also compared different regularization methods (L1, L21 and L2) on matrix factorization. The results are shown in Figure 9, and the L2 norm obtains the best performance.

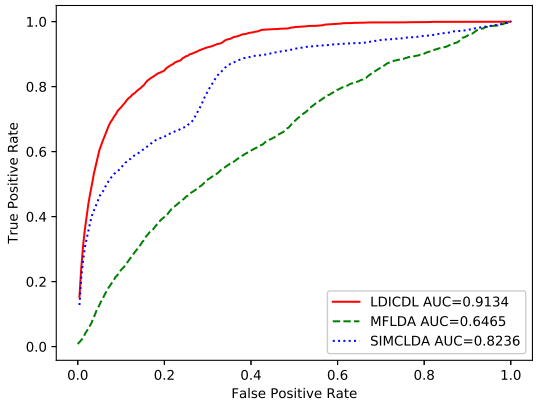

Fig. 4. The AUROC of LDICDL, SIMCLDA and MFLDA by using ten-fold cross validation.

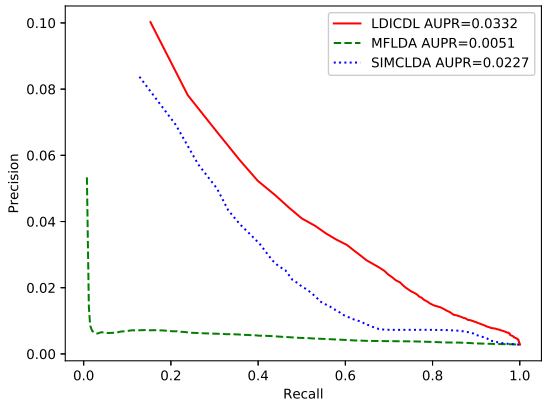

Fig. 5. The AUPR of LDICDL, SIMCLDA and MFLDA by using ten-fold cross validation.

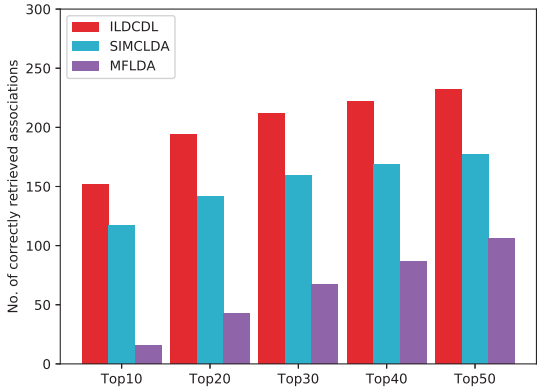

Fig. 6. Number of correctly retrieved known lncRNAdisease associations for specified rank thresholds based on ten-fold cross validation.

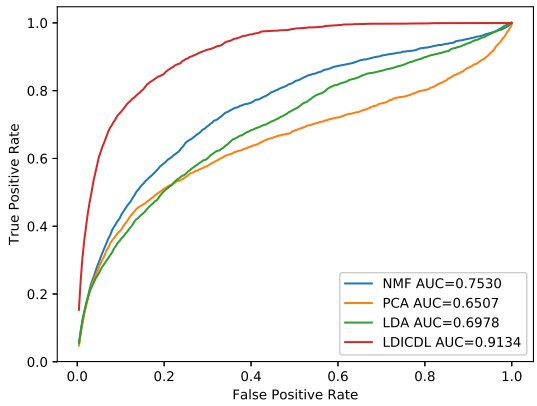

Fig. 7. The AUROC of LDICDL, PCA, LDA and NMF by using ten-fold cross validation.

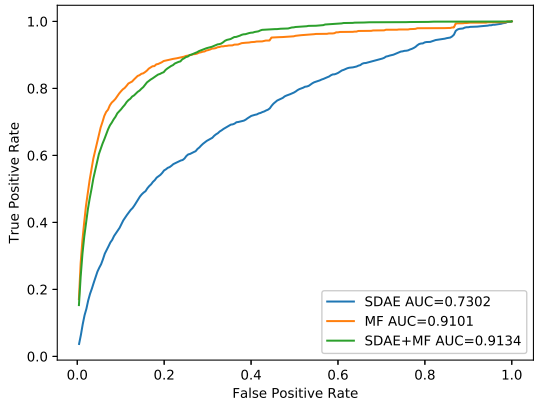

Fig. 8. The AUROC of MF, SDAE and SDAE+MF by using ten-fold cross validation.

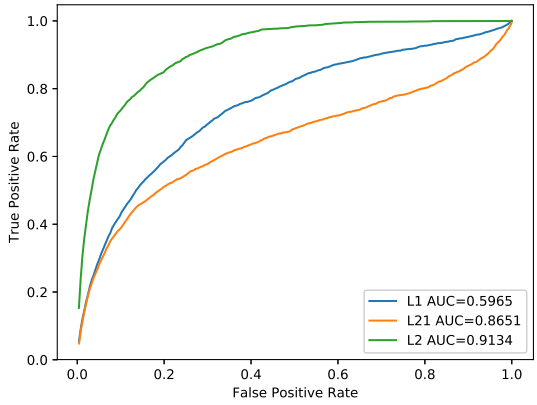

Fig. 9. The AUROC of L1, L21 and L2 in MF by using ten-fold cross validation.

3.4 De novo test

In order to validate the performance of LDICDL in identifying potential association for new diseases, the *de novo* test is conducted in the experiment. The *de novo* test removes all known associations with lncRNAs from each disease i as the test set each time. The potential associations between lncRNAs and disease i are predicted based on feature information. The results of AUROC and AUPR are shown in Figures 10 and 11, respectively. The LDICDL achieves the highest AUC and AUPR (0.8917 and 0.1666). Compared with other methods, our method is at least 0.09 higher than other methods in AUC (SIMCLDA 0.7923 and MFLDA 0.5952) and 0.04 higher than other methods in AUPR (SIMCLDA 0.1270 and MFLDA 0.0398). It demonstrates that our method is superior to other methods in prediction performance of *de novo* test. Figure 12 shows the number of correctly retrieved

known lncRNA-disease associations. It can be found that LDICDL outperforms than other methods for top 10 to top 50.

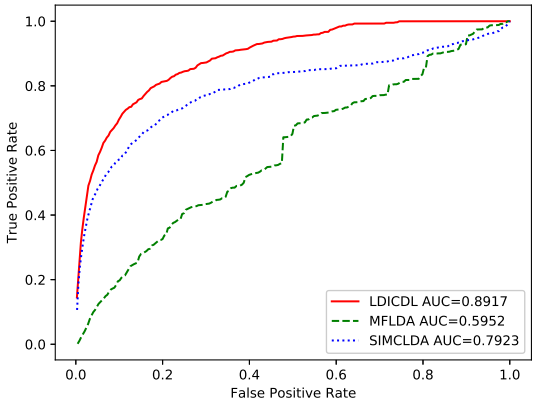

Fig. 10. The AUROC of LDICDL, SIMCLDA and MFLDA by using de novo cross validation.

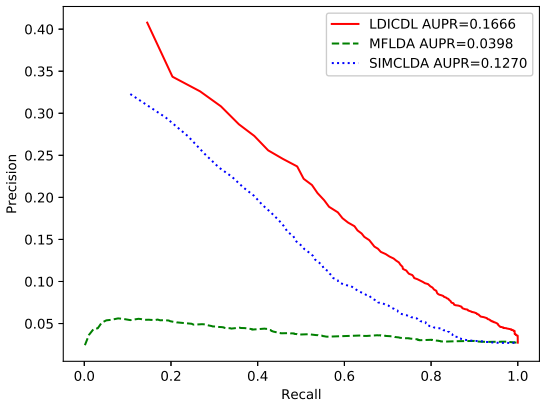

Fig. 11. The AUPR of LDICDL, SIMCLDA and MFLDA by using de novo cross validation.

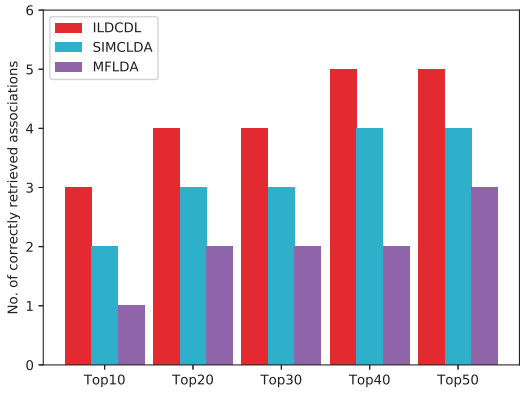

Fig. 12. The number of correctly retrieved known lncRNAdisease associations for specified rank thresholds based on de novo validation.

The hyperparameters are selected by random search proposed in [35]. γ and θ are chosen from $[0.1, 1, 10, 100, 200, 300, 500, 1000]$, $\gamma_l : \gamma_n$ and $\gamma_d : \gamma_n$ are both chosen from [1:1, 100:1, 200:1, 300:1, 400:1, 500:1, 600:1, 700:1, 800:1, 900:1, 1000:1] [36], γ_w is chosen from $[0.1, 0.3, 0.5, 0.7, 0.9]$. Then all hyperparameters are sampled from a uniform distribution over a set of possible values. In our experiment, we repeat the process 20 times to find the optimum parameters. The parameters are set as follows: $\theta = 100, \gamma = 300, \gamma_l : \gamma_n = \gamma_d : \gamma_n = 100 : 1, \gamma_w = 0.3$.

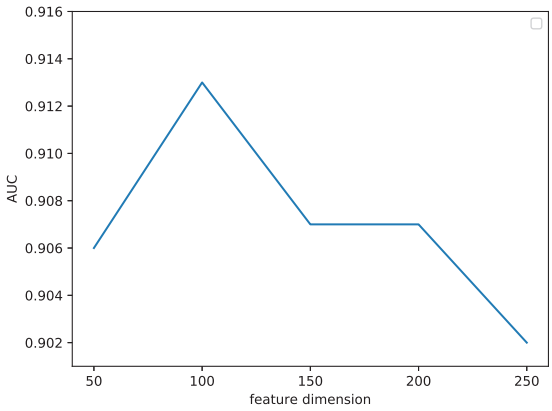

Fig. 13. The effect of feature dimension k.

3.5 The effects of parameters

In the SDAE, the feature information of lncRNA and disease are reduced into subspace. To test the effect of feature dimension k, we conduct the ten-fold cross validation by changing the feature dimension from 50 to 250 by increasing 50 each time. The result is shown in Figure 13. It is observed that the LDICDL achieves the best performance when the feature dimension is equal to 100. Therefore, 100 is applied for the feature dimension k in experiment. All the three hidden layers use non-linear activation functions tanh, and the output layer uses the sigmoid. The number of neurons of the auto-encoder are set to 130, 100 and 130, respectively.

3.6 Case study

To demonstrate the capability of LDICDL in identifying the potential lncRNA-disease associations, the osteosarcoma is selected as case study. In case study, all known associations between lncRNAs and osteosarcoma are treated as positive samples. Then, the potential associations are predicted by LDICDL. The predicted lncRNA of osteosarcoma is analyzed by consulting recent publication.

Osteosarcoma (osteogenic sarcoma) is a primary bone malignancy that often affects children and young adults

(approximately 3.4% of all childhood cancers) [37], [38]. This cancer is rare (less than 1% of all cancers diagnosed) and the pathogenesis is unknown. With the development of multi-agent chemotherapy regimens, the long-term survival rate is improved from 65% to 70% [39]. Unfortunately, the prognostic and treatment have no improved in several decades. Table 1 shows the top 10 lncRNA of osteosarcoma predicted by LDICDL. As shown in Table 1, 9 out of 10 lncRNAs are proved to relate with osteosarcoma by recent literatures. The H19 ranked in top 1 has been proved to be related with osteosarcoma [40]. The rs217727 in H19 can increase IGF2 cord blood level which has significantly associated with osteosarcoma. It has been proved that the long coding RNA PVT1 ranked at top 2 can promote cell apoptosis and inhibit cell proliferation, migration, and invasion in osteosarcoma cells by regulating the expression of miR-195 [41]. The GAS5 ranked at top 3 can promote the expression of aplasia Ras Homologue member I (ARHI) which suppresses Cell Growth and Epithelial-Mesenchymal Transition in Osteosarcoma by acting as molecular sponger to regulate the expression of miR-221 [42]. Recent research shows that the NEAT1 ranked at top 4 is significantly upregulated in osteosarcoma cell lines which has close association with higher clinical stage, distant metastasis and poorer prognosis. In addition, it can inhibit Ecadherin expression and promote the metastasis of osteosarcoma by relating with the G9a-DNMT1-Snail complex [43]. The long coding RNA KCNQ1OT1 ranked at top 5 has been discovered to associate with cell invasion, migration, growth, proliferation and apoptosis through enhancing WNT/beta-catenin signaling pathway activity in osteosarcoma tissue [44]. It has been discovered that AFAP1AS1 ranked at top 7 is significantly over-expressed and the knockdown of AFAP1-AS1 can strikingly inhibits the cell proliferation in osteosarcoma tissue. It demonstrates that AFAP1AS1 can promote cell proliferation in osteosarcoma via regulating miR-4695-5p/TCF4-β-catenin signaling [45]. The long Noncoding RNA XIST ranked at top 8 has been proved that it can bind to miR-320b and inhibit the expression of miR-320b in osteosarcoma cells. The miR-320b can target the Ras-Related Protein RAP2B and inhibit the expression of RAP2B which is involved in cell proliferation and invasion of osteosarcoma [46]. It has been revealed that the CCAT1 rank at top 9 is upregulated in osteosarcoma tissues and cells, and is related with the cell proliferation and migration of osteosarcoma by binding to miR-148a and regulating the signal pathway of phosphatidyl inositol 3-kinase interacting protein 1 (PIK3IP1) [47]. The recent evidences present that long coding RNA SPRY4-IT1 ranked at top 10 is over-expressed in osteosarcoma tissues and SPRY4-IT1 knockdown strikingly inhibits cells proliferation through inhibiting the expression of G1 [48]. In addition, some interesting lncRNAs such as MIR155HG are found by our method. The biological functions of these lncRNAs are still unknown. It deserves for biologist to validate by biological experiments.

TABLE 1
Top 10 lncRNA of osteosarcoma predicted by LDICDL

Rank	LncRNA	Evidence
1	H19	[40]
2	PVT1	[41]
3	GAS5	[42]
4	NEAT1	[43]
5	KCNQ1OT1	[44]
6	MIR155HG	Unknown
7	AFAP1-AS1	[45]
8	XIST	[46]
9	CCAT1	[47]
10	SPRY4-IT1	[48]

4 Discussion

It is well known than lncRNA is a kind of important non-coding RNA with the length more than 200 nucleotides [49]. Accumulating evidences show that lncRNA plays critical roles in various biological processes such as chromosome dosage compensation, genomic imprinting, epigenetic regulation, nuclear and cytoplasmic trafficking, cell proliferation, cell differentiation, cell growth, cell metabolism and cell apoptosis [50], [51]. In addition, increasing studies demonstrate that lncRNA has close relationship with various diseases including cancer [28]. Therefore, identifying LncRNA-disease associations benefits to understand the pathogenesis of disease, and further disease treatment and drug discovery.

In this study, we have proposed a computational method, called LDICDL, to predict LncRNA-disease associations based on collaborative deep learning. In this approach, the lncRNA-disease association matrix is decomposed to lncRNA feature subspace and disease feature information is encoded by using SDAE. Meanwhile, the lncRNA-disease association matrix is decomposed to disease feature subspace and lncRNA feature information is encoded by using SDAE. Then, the lncRNA-disease association score is predicted based on lncRNA feature matrix and disease encode matrix, and disease feature matrix and lncRNA encode matrix, respectively. The final score of lncRNA-disease association is calculated by averaging the scores. The results demonstrate LDICDL is competitive and often performs better than other state-of-the-art methods. In addition, our method may also be used to other biological entity prediction such as miRNA-disease association prediction [52], [53], [54], drug-target interaction prediction [55] and disease gene prediction [56].

Funding

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 Innovation Project of Guangxi Graduate Education (No. YCSW2020020), the Natural Science Foundation of Yunnan Province of China (No. 2019FA024), the Hunan Provincial Science and Technology Program (No. 2018WK4001), the scientific Research Foundation of Hunan Provincial Education Department (No.18B469).

REFERENCES

[1] G. L. Maor, A. Yearim, and G. Ast, "The alternative role of dna methylation in splicing regulation," *Trends in Genetics*, vol. 31, no. 5, pp. 274–280, 2015.

[2] W. Lan, J. Wang, M. Li, J. Liu, F.-X. Wu, and Y. Pan, "Predicting microrna-disease associations based on improved microrna and disease similarities," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 15, no. 6, pp. 1774–1782, 2018.

[3] E. Anastasiadou, L. S. Jacob, and F. J. Slack, "Non-coding rna networks in cancer," *Nature Reviews Cancer*, vol. 18, no. 1, p. 5, 2018.

[4] J. Ponjavic, C. P. Ponting, and G. Lunter, "Functionality or transcriptional noise? evidence for selection within long noncoding rnas," *Genome research*, vol. 17, no. 5, pp. 556–565, 2007.

[5] Q. Chen, W. Lan, and J. Wang, "Mining featured patterns of mirna interaction based on sequence and structure similarity," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 10, no. 2, pp. 415–422, 2013.

[6] K. C. Wang and H. Y. Chang, "Molecular mechanisms of long noncoding rnas," *Molecular cell*, vol. 43, no. 6, pp. 904–914, 2011.

[7] X. Xue, Y. A. Yang, A. Zhang, K. Fong, J. Kim, B. Song, S. Li, J. C. Zhao, and J. Yu, "Lncrna hotair enhances er signaling and confers tamoxifen resistance in breast cancer," *Oncogene*, vol. 35, no. 21, p. 2746, 2016.

[8] L. Yang, C. Lin, C. Jin, J. C. Yang, B. Tanasa, W. Li, D. Merkurjev, K. A. Ohgi, D. Meng, J. Zhang *et al.*, "lncrna-dependent mechanisms of androgen-receptor-regulated gene activation programs," *Nature*, vol. 500, no. 7464, p. 598, 2013.

[9] W. Lan, J. Wang, M. Li, W. Peng, and F. Wu, "Computational approaches for prioritizing candidate disease genes based on ppi networks," *Tsinghua Science and Technology*, vol. 20, no. 5, pp. 500–512, 2015.

[10] G. Yang, X. Lu, and L. Yuan, "Lncrna: a link between rna and cancer," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1839, no. 11, pp. 1097–1109, 2014.

[11] P.-J. Volders, K. Verheggen, G. Menschaert, K. Vandepoele, L. Martens, J. Vandesompele, and P. Mestdagh, "An update on lncipedia: a database for annotated human lncrna sequences," *Nucleic acids research*, vol. 43, no. D1, pp. D174–D180, 2014.

[12] Q. Jiang, J. Wang, X. Wu, R. Ma, T. Zhang, S. Jin, Z. Han, R. Tan, J. Peng, G. Liu *et al.*, "Lncrna2target: a database for differentially expressed genes after lncrna knockdown or overexpression," *Nucleic acids research*, vol. 43, no. D1, pp. D193–D196, 2014.

[13] W. Lan, L. Huang, D. Lai, and Q. Chen, "Identifying interactions between long noncoding rnas and diseases based on computational methods," in *Computational Systems Biology*. Springer, 2018, pp. 205–221.

[14] J. Sun, H. Shi, Z. Wang, C. Zhang, L. Liu, L. Wang, W. He, D. Hao, S. Liu, and M. Zhou, "Inferring novel lncrna–disease associations based on a random walk model of a lncrna functional similarity network," *Molecular BioSystems*, vol. 10, no. 8, pp. 2074–2081, 2014.

[15] X. Chen, Z.-H. You, G.-Y. Yan, and D.-W. Gong, "Irwrlda: improved random walk with restart for lncrna-disease association prediction," *Oncotarget*, vol. 7, no. 36, p. 57919, 2016.

[16] M. Zhou, X. Wang, J. Li, D. Hao, Z. Wang, H. Shi, L. Han, H. Zhou, and J. Sun, "Prioritizing candidate disease-related long non-coding rnas by walking on the heterogeneous lncrna and disease network," *Molecular BioSystems*, vol. 11, no. 3, pp. 760–769, 2015.

[17] Q. Yao, L. Wu, J. Li, L. guang Yang, Y. Sun, Z. Li, S. He, F. Feng, H. Li, and Y. Li, "Global prioritizing disease candidate lncrnas via a multi-level composite network," *Scientific reports*, vol. 7, p. 39516, 2017.

[18] W. Lan, M. Li, K. Zhao, J. Liu, F.-X. Wu, Y. Pan, and J. Wang, "Ldap: a web server for lncrna-disease association prediction," *Bioinformatics*, vol. 33, no. 3, pp. 458–460, 2016.

[19] X. Chen and G. Yan, "Novel human lncrna–disease association inference based on lncrna expression profiles," *Bioinformatics*, vol. 29, no. 20, pp. 2617–2624, 2013.

[20] X. Wu, W. Lan, Q. Chen, Y. Dong, J. Liu, and W. Peng, "Inferring lncrna-disease associations based on graph autoencoder matrix completion," *Computational Biology and Chemistry*, p. 107282, 2020.

[21] G. Fu, J. Wang, C. Domeniconi, and G. Yu, "Matrix factorization-based data fusion for the prediction of lncrna–disease associations," *Bioinformatics*, vol. 34, no. 9, pp. 1529–1537, 2017.

[22] C. Lu, M. Yang, F. Luo, F.-X. Wu, M. Li, Y. Pan, Y. Li, and J. Wang, "Prediction of lncrna–disease associations based on inductive matrix completion," *Bioinformatics*, vol. 34, no. 19, pp. 3357–3364, 2018.

[23] Q. Chen, D. Lai, W. Lan, X. Wu, B. Chen, Y.-P. P. Chen, and J. Wang, "Ildmsf: Inferring associations between long non-coding rna and disease based on multi-similarity fusion," *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[24] J. Han, L. Zheng, Y. Xu, B. Zhang, F. Zhuang, P. S. Yu, and W. Zuo, "Adaptive deep modeling of users and items using side information for recommendation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 3, pp. 737–748, 2020.

[25] H. Park, J. Jung, and U. Kang, "A comparative study of matrix factorization and random walk with restart in recommender systems," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 756–765.

[26] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning for recommender systems," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2015, pp. 1235–1244.

[27] A. Ramlatchan, M. Yang, Q. Liu, M. Li, J. Wang, and Y. Li, "A survey of matrix completion methods for recommendation systems," *Big Data Mining and Analytics*, vol. 1, no. 4, pp. 308–323, 2018.

[28] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *2008 Eighth IEEE International Conference on Data Mining*. Ieee, 2008, pp. 263–272.

[29] Q. Jiang, J. Wang, X. Wu, R. Ma, T. Zhang, S. Jin, Z. Han, R. Tan, J. Peng, G. Liu *et al.*, "Lncrna2target: a database for differentially expressed genes after lncrna knockdown or overexpression," *Nucleic acids research*, vol. 43, no. D1, pp. D193–D196, 2014.

[30] Z. Lu, K. BRETONNEL COHEN, and L. Hunter, "Generif quality assurance as summary revision," in *Biocomputing 2007*. World Scientific, 2007, pp. 269–280.

[31] C. Jonquet, N. H. Shah, and M. A. Musen, "The open biomedical annotator," *Summit on translational bioinformatics*, vol. 2009, p. 56, 2009.

[32] J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, and J.-H. Yang, "starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data," *Nucleic acids research*, vol. 42, no. D1, pp. D92–D97, 2013.

[33] Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, and Q. Cui, "Hmdd v2. 0: a database for experimentally supported human microrna and disease associations," *Nucleic acids research*, vol. 42, no. D1, pp. D1070–D1074, 2013.

[34] J. Pinero, N. Queralt-Rosinach, A. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong, "Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes," *Database*, vol. 2015, 2015.

[35] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281–305, 2012.

[36] H. Wang, N. Wang, and D. Y. Yeung, "Collaborative deep learning for recommender systems," 2014.

[37] P. A. Meyers and R. Gorlick, "Osteosarcoma," *Pediatric Clinics of North America*, vol. 44, no. 4, pp. 973–989, 1997.

[38] B. A. Lindsey, J. E. Markel, and E. S. Kleinerman, "Osteosarcoma overview," *Rheumatology and therapy*, vol. 4, no. 1, pp. 25–43, 2017.

[39] M. S. Isakoff, S. S. Bielack, P. Meltzer, and R. Gorlick, "Osteosarcoma: current treatment and a collaborative pathway to success," *Journal of clinical oncology*, vol. 33, no. 27, p. 3029, 2015.

[40] T. He, D. Xu, T. Sui, J. Zhu, Z. Wei, and Y. Wang, "Association between h19 polymorphisms and osteosarcoma risk," *Eur Rev Med Pharmacol Sci*, vol. 21, no. 17, pp. 3775–3780, 2017.

[41] Q. Zhou, F. Chen, J. Zhao, B. Li, Y. Liang, W. Pan, S. Zhang, X. Wang, and D. Zheng, "Long non-coding rna pvt1 promotes osteosarcoma development by acting as a molecular sponge to regulate mir-195," *Oncotarget*, vol. 7, no. 50, p. 82620, 2016.

[42] K. Ye, S. Wang, H. Zhang, H. Han, B. Ma, and W. Nan, "Long noncoding rna gas5 suppresses cell growth and epithelial–mesenchymal transition in osteosarcoma by regulating the mir-221/arhi pathway," *Journal of cellular biochemistry*, vol. 118, no. 12, pp. 4772–4781, 2017.

[43] Y. Li and C. Cheng, "Long noncoding rna neat1 promotes the metastasis of osteosarcoma via interaction with the g9a-dnmt1-snail complex," *American journal of cancer research*, vol. 8, no. 1, p. 81, 2018.

[44] C. Zhang, S. Du, and L. Cao, "Long non-coding rna kcnq1ot1 promotes osteosarcoma progression by increasing β-catenin activity," *RSC advances*, vol. 8, no. 66, pp. 37 581–37 589, 2018.
[45] R. Li, S. Liu, Y. Li, Q. Tang, Y. Xie, and R. Zhai, "Long noncoding rna afap1-as1 enhances cell proliferation and invasion in osteosarcoma through regulating mir-4695-5p/tcf4-β-catenin signaling," *Molecular medicine reports*, vol. 18, no. 2, pp. 1616–1622, 2018.
[46] G.-Y. Lv, J. Miao, and X.-L. Zhang, "Long noncoding rna xist promotes osteosarcoma progression by targeting ras-related protein rap2b via mir-320b," *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, vol. 26, no. 6, pp. 837–846, 2018.
[47] J. Zhao and L. Cheng, "Long non-coding rna ccat1/mir-148a axis promotes osteosarcoma proliferation and migration through regulating pik3ip1," *Acta biochimica et biophysica Sinica*, vol. 49, no. 6, pp. 503–512, 2017.
[48] J. Xu, R. Ding, and Y. Xu, "Effects of long non-coding rna spry4-it1 on osteosarcoma cell biological behavior," *American journal of translational research*, vol. 8, no. 12, p. 5330, 2016.
[49] C. P. Ponting, P. L. Oliver, and W. Reik, "Evolution and functions of long noncoding rnas," *Cell*, vol. 136, no. 4, pp. 629–641, 2009.
[50] R. Zheng, M. Li, X. Chen, S. Zhao, F. Wu, Y. Pan, and J. Wang, "An ensemble method to reconstruct gene regulatory networks based on multivariate adaptive regression splines," *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.
[51] A. Necsulea, M. Soumillon, M. Warnefors, A. Liechti, T. Daish, U. Zeller, J. C. Baker, F. Grützner, and H. Kaessmann, "The evolution of lncrna repertoires and expression patterns in tetrapods," *Nature*, vol. 505, no. 7485, p. 635, 2014.
[52] W. Peng, W. Lan, Z. Yu, J. Wang, and Y. Pan, "A framework for integrating multiple biological networks to predict microrna-disease associations," *IEEE transactions on nanobioscience*, vol. 16, no. 2, pp. 100–107, 2016.
[53] W. Lan, Q. Chen, T. Li, C. Yuan, S. Mann, and B. Chen, "Identification of important positions within mirnas by integrating sequential and structural features," *Current Protein and Peptide Science*, vol. 15, no. 6, pp. 591–597, 2014.
[54] W. Peng, W. Lan, J. Zhong, J. Wang, and Y. Pan, "A novel method of predicting microrna-disease associations based on microrna, disease, gene and environment factor networks," *Methods*, vol. 124, pp. 69–77, 2017.
[55] W. Lan, J. Wang, M. Li, J. Liu, Y. Li, F.-X. Wu, and Y. Pan, "Predicting drug–target interaction using positive-unlabeled learning," *Neurocomputing*, vol. 206, pp. 50–57, 2016.
[56] J. Liu, M. Li, W. Lan, F.-X. Wu, Y. Pan, and J. Wang, "Classification of alzheimer's disease using whole brain hierarchical network," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 2, pp. 624–632, 2016.