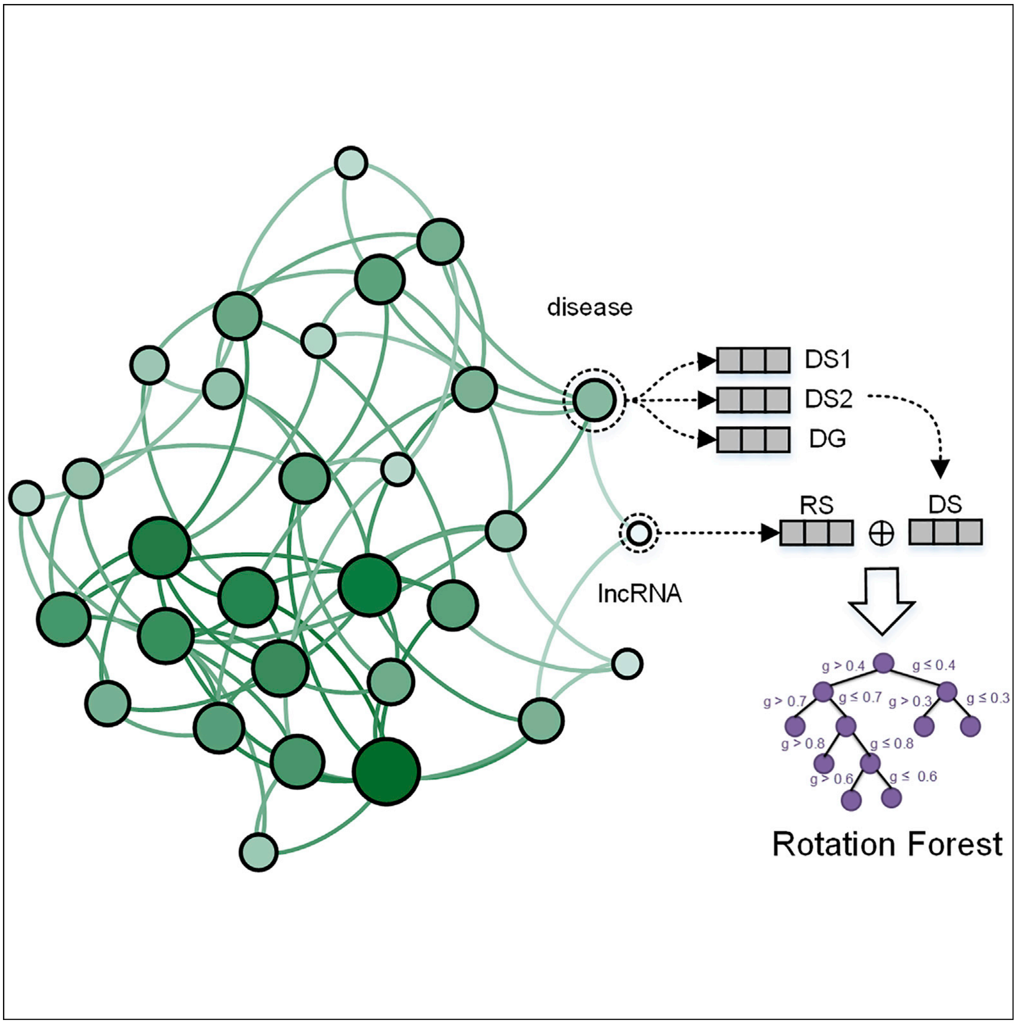
iScience

CellPress

Article

A Learning-Based Method for LncRNA-Disease Association Identification Combing Similarity Information and Rotation Forest



Zhen-Hao Guo,
Zhu-Hong You,
Yan-Bin Wang,
Hai-Cheng Yi,
Zhan-Heng Chen

zhuhongyou@ms.xjb.ac.cn
(Z.-H.Y.)
wangyanbin15@mails.ucas.ac.cn (Y.-B.W.)

HIGHLIGHTS

We propose a similarity-based characterization method for RNA-disease associations

The model automatically captures important association features

This method determines the prospects of machine learning techniques on such problems

Guo et al., iScience 19, 786–795
September 27, 2019 © 2019
The Authors.
https://doi.org/10.1016/j.isci.2019.08.030

iScience

CellPress

Article

A Learning-Based Method for LncRNA-Disease Association Identification Combing Similarity Information and Rotation Forest

Zhen-Hao Guo,[1,2,3] Zhu-Hong You,[1,3,4,*] Yan-Bin Wang,[1,3,*] Hai-Cheng Yi,[1] and Zhan-Heng Chen[1]

SUMMARY

Long non-coding RNA (lncRNA) play critical roles in the occurrence and development of various diseases. The determination of the lncRNA-disease associations thus would contribute to provide new insights into the pathogenesis of the disease, the diagnosis, and the gene treatments. Considering that traditional experimental approaches are difficult to detect potential human lncRNA-disease associations from the vast amount of biological data, developing computational method could be of significant value. In this paper, we proposed a novel computational method named LDASR to identify associations between lncRNA and disease by analyzing known lncRNA-disease associations. First, the feature vectors of the lncRNA-disease pairs were obtained by integrating lncRNA Gaussian interaction profile kernel similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity. Second, autoencoder neural network was employed to reduce the feature dimension and get the optimal feature subspace from the original feature set. Finally, Rotating Forest was used to carry out prediction of lncRNA-disease association. The proposed method achieves an excellent preference with 0.9502 AUC in leave-one-out cross-validations (LOOCV) and 0.9428 AUC in 5-fold cross-validation, which significantly outperformed previous methods. Moreover, two kinds of case studies on identifying lncRNAs associated with colorectal cancer and glioma further proves the capability of LDASR in identifying novel lncRNA-disease associations. The promising experimental results show that the LDASR can be an excellent addition to the biomedical research in the future.

INTRODUCTION

Long non-coding RNAs (lncRNAs) are an important class of transcripts, with the length longer than 200 nt, which participates in various physiological processes, such as immune surveillance, post-translational regulation, cell differentiation, proliferation, apoptosis, and epigenetic regulation. Especially, accumulating studies have indicated that a large number of lncRNAs are involved in numerous complex human diseases, such as various cancers (Chung et al., 2011; Zhang et al., 2012), blood diseases (Congrains et al., 2012; Alvarez-Dominguez and Lodish, 2017; Sallam et al., 2018), and neurodegeneration diseases (Johnson, 2012). Therefore, inferring the potential association between lncRNA and disease is helpful to understand the pathogenesis of complex diseases at the molecular level and provide new insights into the diagnosis, treatment, and prognosis of diseases.

Profit from the development of high-throughput experimental techniques, such as Microarray, Northern blots and qPCR, Fluorescence *in situ* hybridization, RNA interference, and RNA immunoprecipitation (Yan et al., 2012), a large amount of data about lncRNAs-disease associations have been determined and distributed in different public databases, such as lncRNAdb (Amaral et al., 2010), NRED (Dinger et al., 2008), and NONCODE (Xie et al., 2013). However, although experimentally validated lncRNA-disease associations drive research and development of medical molecular biology, they often have high false positives and false negatives. Moreover, many experimental methods are expensive and time-consuming. Consequently, it is essential to develop a computational prediction approach based on the accumulated biological data to accurately and rapidly find potential lncRNAs-disease associations. Computational method can quantitatively describe the associations between lncRNAs and diseases and efficiently screen out the most promising lncRNA-disease association pairs for further biological experimental validation.

The proposed computational method for predicting lncRNA-disease association can be roughly divided into three categories. Methods in the first category uncover ncRNA-disease associations based on the idea of network or link prediction. The underlying assumption is that lncRNAs associated with the same

[1]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China

[2]University of Chinese Academy of Sciences, Beijing 100049, China

[3]These authors contributed equally

[4]Lead Contact

*Correspondence: zhuhongyou@ms.xjb.ac.cn (Z.-H.Y.), wangyanbin15@mails.ucas.ac.cn (Y.-B.W.)

https://doi.org/10.1016/j.isci.2019.08.030

or similar diseases are more likely to have similar functions. Liao et al. constructed a coding-non-coding gene co-expression network based on public microarray expression profiles to discover the potential functions of lncRNA (Liao et al., 2011). Yang et al. applied a propagation algorithm to predict lncRNA-disease associations by constructing a coding-non-coding gene-disease bipartite network based on known associations between diseases and disease-causing genes (Yang et al., 2014). Chen et al. came up with the model called IRWRLDA to identify potential associations by integrating known lncRNA-disease associations, disease semantic similarity, and various lncRNA similarity measures (Chen et al., 2016). Huang et al. proposed a model called PBMDA to predict microRNA (miRNA)-disease associations by integrating known human miRNA-disease associations, miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity (You et al., 2017). Methods in the second category utilize matrix factorization to identify potential lncRNA-disease associations. The basic assumption is that unknown association information can be derived from other known association information. Fu et al. predicted lncRNA-disease associations by decomposing data matrices of heterogeneous data sources into low-rank matrices (Fu et al., 2017). Lu et al. developed a method called SIMCLDA for potential lncRNA-disease association prediction based on inductive matrix completion (Lu et al., 2018). These two types of methods are based on specific assumptions, but these assumptions are not unanimously accepted. Relevant studies have shown that in many cases bio macromolecules with similar structures or ligands do not have the same functions. Matrix factorization approaches will experience dramatic performance degradation when the known associated information is insufficient. In addition, these methods both cannot mine the similarity feature of lncRNA and disease, and consider the inherent logic of the association between lncRNA and disease from the perspective of data-driven. Machine learning models are used in the third category to discover the unknown lncRNA-disease associations. Lan et al. proposed a method called LDAP to identify latent associations between lncRNAs and diseases by using a bagging support vector machine (SVM) classifier based on lncRNA similarity and disease similarity (Lan et al., 2016). Since these methods are the beginning of machine learning application for lncRNA-disease association prediction, there is still much room for improvement in the prediction performance, prediction accuracy of such methods can be still greatly improved by increasing training samples and using more appropriate and advanced learning algorithms. Recently, the accumulation of association data between lncRNA and disease and the development of machine learning technology provide a better opportunity for predicting the association between lncRNA and disease using supervised learning model.

Instead of using network-based and matrix factorization-based methods to compute association scores directly, we explored to extract association features from lncRNA-disease pairs by multiple similarity matrices and trained machine learning models in a supervised manner to predict their association. In this study, we proposed a novel supervised computational method named (LDASR) for large-scale lncRNA-disease association prediction based on collaborative filtering and machine learning technologies. First, the feature vectors of the lncRNA-disease pairs were obtained by integrating lncRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity. Second, autoencoder neural network was employed to low the feature dimension and get the optimal feature subspace from the original feature set. Finally, considering the size of training samples and the possible non-linear relationship in input, we trained rotating forest to carry out prediction of LncRNA-Disease Association. The flow of LDASR is represented in Figure 1. In leave-one-out cross-validation (LOOCV) and five cross-validation to evaluate test data, the proposed LDASR model achieved better results than some previous methods, with AUC of 0.9502 and 0.9428, respectively. The test results show that supervised learning model can achieve better performance.

RESULTS

Leave-One-Out Cross-Validation

For LOOCV, each sample in the dataset is selected for testing in turn, and the remaining samples are used as the training set to construct the prediction model. As we have mentioned, 1,765 lncRNA-disease associations, which have been experimental verified, were regarded as positive samples. Then we randomly picked 1,765 lncRNA-disease associations in the remaining associations as negative samples. The total number of datasets was 3,530, so we trained and tested 3,530 times according to the LOOCV method to get the final experimental result. At the same time, we drew ROC (receiver operating characteristic curve) and calculated AUC (area under curve) under LOOCV as shown in Figure 2 to quantify the prediction results and facilitate comparison with other methods. For LOOCV, LDASR obtained AUCs of 0.9502, indicating

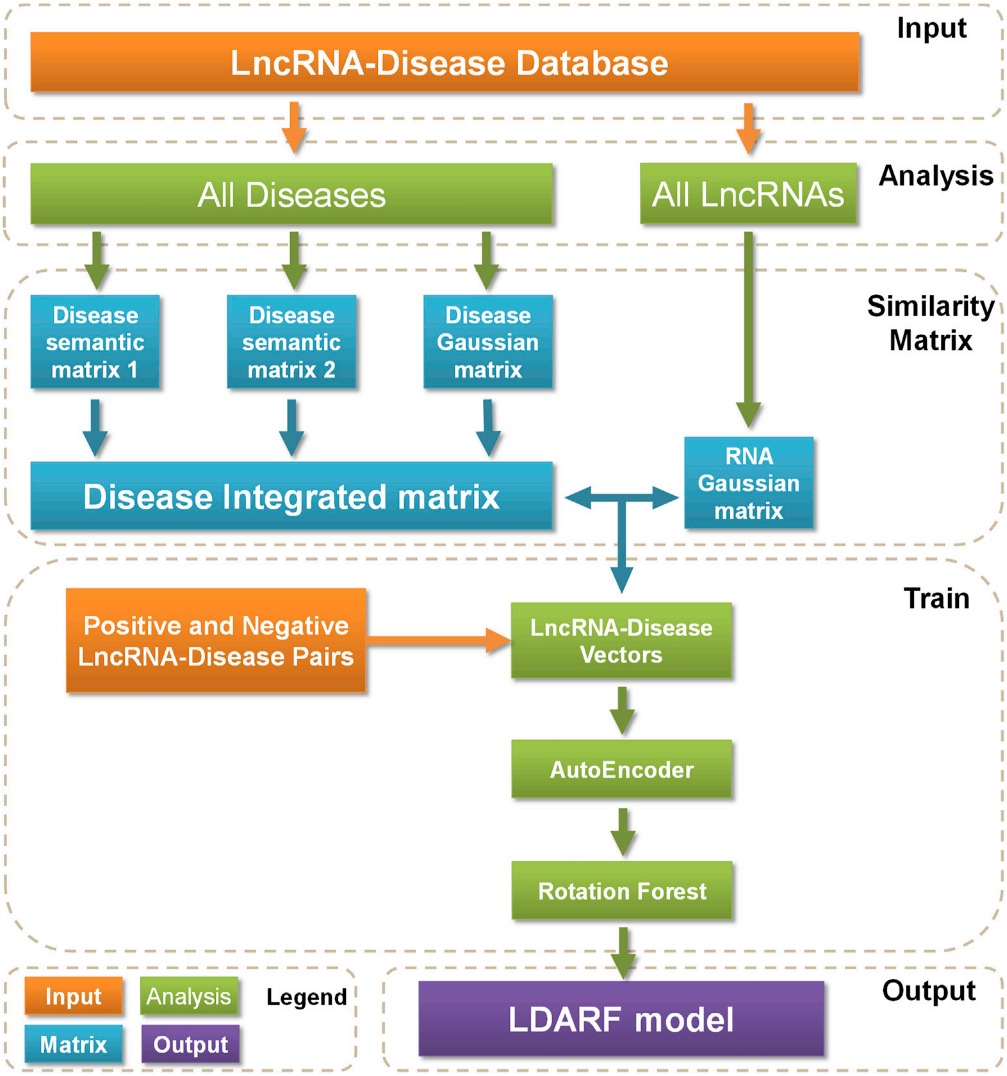


Figure 1. Flowchart of LDASR

Step 1: Building three similarity matrices for disease by combining semantic information and Gaussian kernel information. Step 2: Building 1 similarity matrix for lncRNA. Step 3: Extraction of similarity feature vectors for disease and lncRNA from disease similarity matrix and lncRNA similarity matrix. Step 4: Extracting the same number of positive and negative samples from the adjacency matrix to construct the dataset used in this paper. Step 5: Selecting the most valuable features and reducing feature noise by using autoencoder. Step 6: more discriminant feature vectors were put into Rotation Forest ensemble classifier for training, verification, and prediction. The construction of disease semantic matrix can see also Figure S1.

that the model combining various similarities and rotation forest had a strong ability to distinguish the difference between positive and negative samples.

Five-fold Cross-Validation

For 5-fold cross-validation, the entire dataset is randomly divided into five mutually exclusive subsets of roughly equal size, each of which is used in turn as a test set for evaluation, and the remaining four subsets served as training sets to build the model. To better verify the performance of our method and save computing resources, LDASR was further evaluated by 5-fold cross-validation. For 5-fold cross-validation, LDASR obtained mean AUC of 0.9428 in the end as shown in Figure 3.

To more comprehensively evaluate our model, we used a broader range of evaluation criteria, including accuracy (Acc.), sensitivity (Sen.), specificity (Spec.), precision (Prec.), and MCC. The prediction

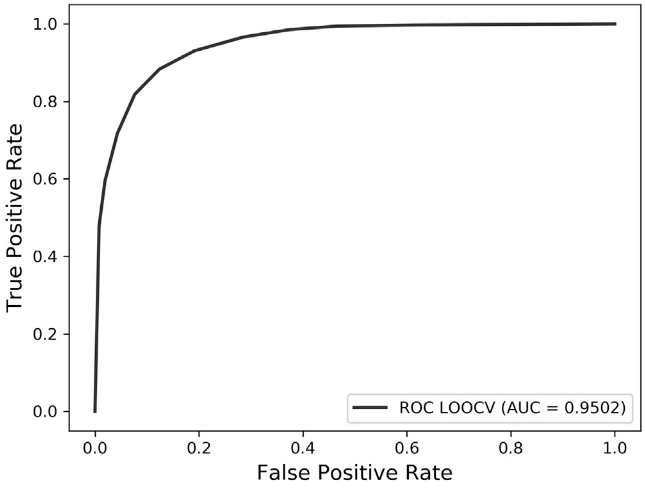


Figure 2. The ROC and AUC of LDASR in LOOCV Based on the v2017 Dataset (3,530 lncRNA-Disease Associations)

performance is listed in Table 1. The results of average Acc., Sen., Spec., Prec., MCC, and AUC were 85.72%, 90.14%, 81.31%, 82.86%, 71.74%, and 94.28% when using the proposed method to predict lncRNA-disease associations. The standard deviations of these values were 1.59%, 0.77%, 2.69%, 2.11%, 3.05%, and 0.94%, respectively. For 5-fold cross-validation shown in Figure 5, LDASR obtained high mean AUC of 0.9428. The high AUCs showed that LDASR combining multiple similarities and rotation forest was feasible and effective to predict lncRNA disease associations. At the same time, the lower standard deviation of these standards implied that the proposed model was robust and stable.

Compared with Other Classifiers

To assess the performance of Rotation Forest: In this section, we compared Rotation Forest with several common classifiers in 5-fold cross-validation, including Random Forest, Logistic Regression, Naive Bayes, and SVM. To be fair, all settings except classifiers are default and the same dataset is used. The ROC curves

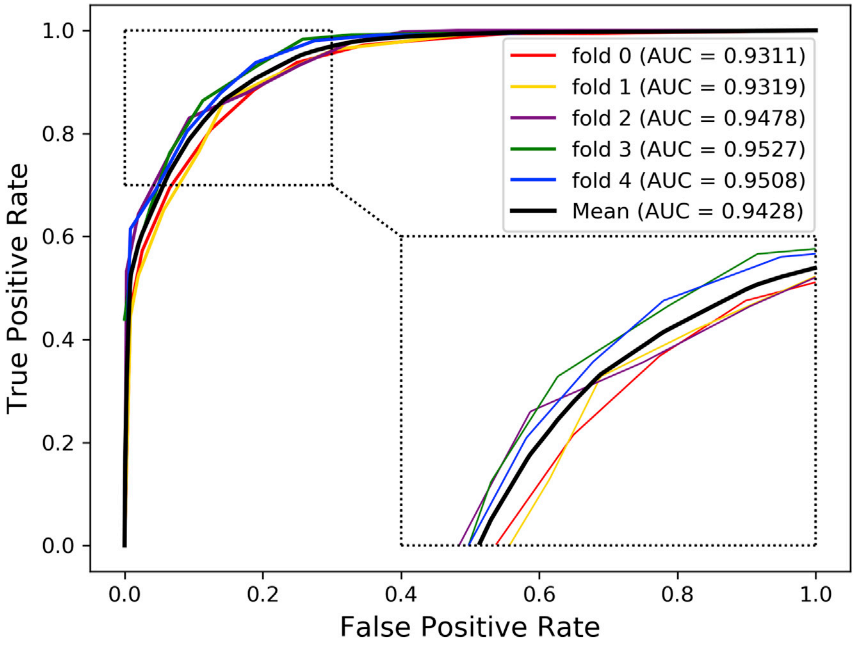


Figure 3. The ROCs and AUCs of LDASR in 5-Fold Cross-validation Based on the v2017 Dataset (3,530 lncRNA-Disease Associations)

Fold	Acc. (%)	Sen. (%)	Spec. (%)	Prec. (%)	MCC (%)	AUC (%)
0	83.85	90.08	77.62	80.10	68.24	93.11
1	85.27	88.95	81.59	82.85	70.73	93.19
2	84.42	89.80	79.04	81.07	69.24	94.78
3	88.10	91.22	84.99	85.87	76.35	95.27
4	86.97	90.65	83.29	84.43	74.14	95.08
Average	85.72 ± 1.59	90.14 ± 0.77	81.31 ± 2.69	82.86 ± 2.11	71.74 ± 3.05	94.28 ± 0.94

Table 1. Five-fold Cross-validation Results Performed by LDASR on the v2017 Dataset (3,530 lncRNA-Disease Associations)

implemented by five classifiers are summarized in Figure 4. As seen in Figure 4, it is obvious that the Rotation Forest achieves the best results. The effectiveness mainly has the following factors: (1) Based on the idea of collaborative filtering, there might be cases of non-independence between feature attributes generated by the similarity between lncRNA and disease. This affects the predictive capability of Naive Bayes. (2) The performance of the SVM classifier is more sensitive to data. Proper selection of the parameters and the correct choice of the kernel function will result in a large training cost. (3) The prediction performance of the logistic regression is limited by the assumption that the feature and the target must be linearly separable, so the lower AUC was obtained. (4) Tree-based assemble algorithms such as Random Forest and Rotating Forest are not affected by the nonlinear relationships in the data, so they have achieved excellent results in the five-fold cross-validation. Compared with the Random Forest, Rotating forest randomly combines the sample attribute sets before each subsample is extracted, and Principal Component Analysis (PCA) is utilized to transform the data between the divided sets of sub-attributes. This operation not only makes each sub-sample different, but also plays a certain role in data pre-processing, thereby improving the accuracy and difference of each base classifier to obtain excellent assemble effect.

Compared with Other Methods

The LDASR was further tested by comparing it with other three state-of-the-art methods involving LRLSLDA (Chen and Yan, 2013), LRLSLDA-LNCSIM1 (Chen et al., 2015), and LRLSLDA-LNCSIM2 (Chen et al., 2015).

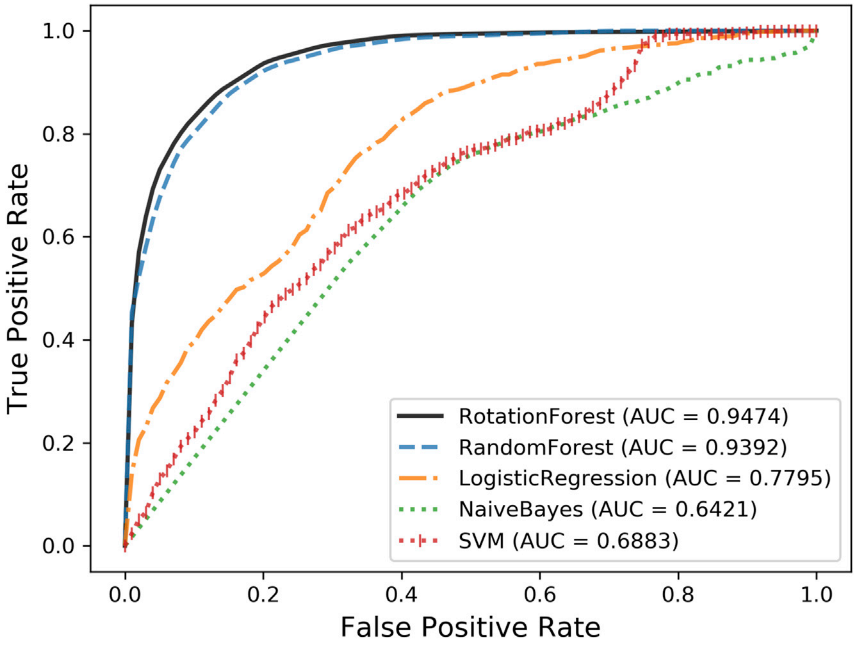


Figure 4. Comparison with Random Forest, Logistic Regression, Naive Bayes, and SVM in 5-Fold Cross-validation Based on the v2017 Dataset (3,530 lncRNA-Disease Associations)

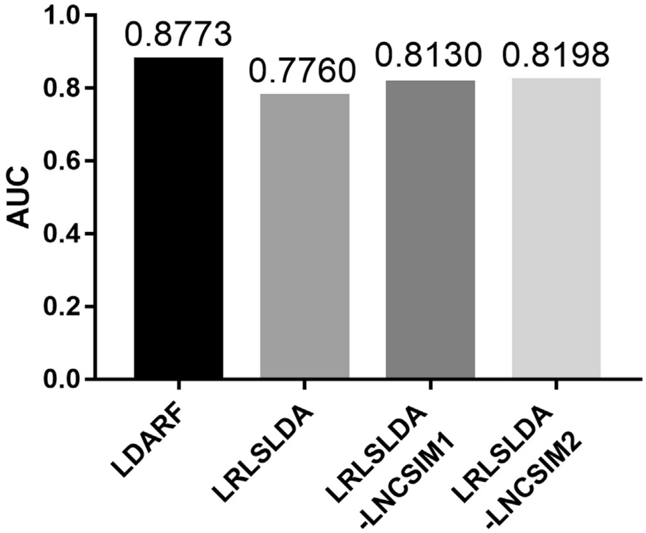


Figure 5. Under the v2012 Dataset (586 lncRNA-Disease Associations), LDASR and LRLSLDA, LRLSLDA-LNCSIM1, LRLSLDA-LNCSIM2 Were Compared between the AUCs Obtained under LOOCV

The comparison of the obtained AUC between LDASR and previous methods in LOOCV is shown in Figure 5. The results on 2012v dataset showed obviously that our proposed method made some progress. As a result, the proposed method achieved 0.1013, 0.0643, and 0.0575 improvements in terms of AUC compared with other three approaches in terms of AUC under LOOCV. Because the Leave-one-out cross-validation uses one observation as the validation set and the remaining observations as the training set, the model under LOOCV will always be a very stable solution. Therefore, it was difficult to add error bars to Figure 5. Compared with traditional computational methods for the prediction of lncRNAs-diseases associations, machine learning can consider similarity information from the perspectives of probability, statistics, approximation, and convex analysis and iteratively and optimally grasp the essential associations rule between RNA and disease. And this data-driven approach will show a stronger advantage as data are accumulated. The improved prediction performance produced by this method further validates the potential of machine learning algorithms on such problems. Although 5% does not seem to be a considerable improvement, we hope to draw the attention of relevant researchers and open a novel perspective on solving problems by machine learning strategies.

Case Study

To evaluate the capability of the model in practical application, we applied LDASR to predict Colorectal Cancer, Glioma, and Prostate cancer as two kinds of case studies. For the purpose of simulating the real environment and ensuring the fairness of case studies, the associations of LncRNADisease database (v2017) were used to train the model and the remaining four additional databases, including Lnc2Cancer (Ning et al., 2015), MNDR (Cui et al., 2017), CRlncRNA (Wang et al., 2018), and LncRNAWiki (Ma et al., 2014), were used to verify the results.

For the first kind of case study, Colorectal Cancer was chosen as the research subject. In this case study, all 1,765 associations in LncRNADisease database (v2017) were used as positive samples. The negative samples of the same size as positive samples are generated by random selection in the rest. Therefore, the test set was constructed by connecting colorectal cancer diseases to all lncRNAs in the other three databases, respectively. As a result, a total of 881 lncRNA-colorectal cancer pairs were verified as test samples. Finally, samples with a predicted probability greater than 0.5 are screened out and sorted according to the probability values from large to small. Recent results in biological experiments confirmed that Colorectal Cancer was related to gene XIST (Lassmann et al., 2007), AB073614 (Xue et al., 2018; Wang et al., 2017), and SNHG3 (Huang et al., 2017). They were all in the top of the list, but they were not included in the LncRNADisease database. Then, we ranked all the 881 lncRNAs based on their predicted association scores and validated the top 10 lncRNAs in LncRNADisease, CRlncRNA, MNDR 2.0, and Lnc2Cancer. The results are shown in the Table 2.

Num	lncRNA	Confirmed Database
1	snhg3	LncRNAWiki
2	linc00237	Unconfirmed
3	kcna2	Unconfirmed
4	xist	LncRNADisease/MNDR 2.0
5	cahm	LncRNADisease/CRlncRNA
6	bx649059	LncRNADisease/CRlncRNA/MNDR 2.0
7	ab073614	Lnc2Cancer
8	bx648207	LncRNADisease/CRlncRNA/MNDR 2.0
9	ak123657	LncRNADisease/CRlncRNA/MNDR 2.0
10	fas-as1	Unconfirmed

Table 2. Top 10 Colorectal Cancer-Associated lncRNAs Predicted by LDASR

In the second kind of case study, Glioma and Prostate Cancer were the subjects of the study. A glioma is a tumor that begins with glial cells in the brain or spine (Mamelak and Jacoby, 2007). The experimental design for case study 2 is as follows: For positive sample set, we removed all Glioma-related associations in the positive sample set. There were 42 positive samples related to Glioma here, so the number of positive samples was 1,723. Like case study 1, we used the same method to select 1,723 negative samples and 881 test samples. The test sample set was put into the classifier after training with positive and negative samples. In the end, we found that XIST and CYTOR were both at the top of the list, but they were not included in the LncRNADisease database. Recent results in biological experiments confirmed that Glioma was related to XIST (Yao et al., 2015) and CYTOR (Yu et al., 2017). Then, we ranked all the 881 lncRNAs based on their predicted association scores and validated the top 10 lncRNAs in LncRNADisease, CRlncRNA, MNDR 2.0, and Lnc2Cancer. The results can be seen in Table 3.

For Prostate Cancer, all the experimental steps are the same as Glioma. For the positive sample set, we removed all Prostate Cancer-related associations in the positive sample set. There were 55 positive samples related to Glioma here, so the number of positive samples was 1,710. Like case study Glioma, we used the same method to select 1,710 negative samples and 881 test samples. The list of the validated top 10 lncRNAs are listed in Table 4.

DISCUSSION

Accumulating evidences have highlighted the positive role of developing a powerful machine-learning-based method to predict potential associations between lncRNAs and diseases, which could significantly help people to understand the pathogenesis of complex diseases at the molecular level and provide new insights into the diagnosis, treatment, and prognosis of diseases. In this paper, we proposed a novel computational method, LDASR, to predict the unknown lncRNA-disease associations by integrating multiple similarity information. Compared with previous methods, we embed this task into a machine learning framework to better understand the essential law of the association between lncRNA and diseases. First, we extracted feature vectors for lncRNA and disease from multiple similarity matrices and constructed the feature vector of RNA-disease pairs by connecting features of lncRNA to that of disease. Then, autoencoder neural network was employed to reduce the feature dimension and improve the efficiency and accuracy of classifier. Finally, we applied rotation forest to carry out prediction. LDASR first shows its good performance by LOOCV and 5-fold cross-validation experiments. Furthermore, the comparison test shows that LDASR has a powerful prediction ability to distinguish positive and negative samples, which is obviously better than the other state-of-the-art methods. Finally, the analyses of case studies further prove that LDASR holds significant value in inferring potential lncRNA-disease associations in practice. As a novel computational method, it is anticipated that LDASR has potential value in biomedical research for comprehending the pathogenesis of diseases, which can further advance the quality of disease diagnostics, therapy, prognosis, and prevention. As a novel computational approach, LDASR could not only play a positive

Num	lncRNA	Confirmed Database
1	zfat-as1	Unconfirmed
2	xist	CRlncRNA/MNDR 2.0
3	spry4-it1	LncRNADisease/CRlncRNA/MNDR 2.0
4	cytor	MNDR 2.0
5	neat1	LncRNADisease/CRlncRNA
6	meg3	LncRNADisease/CRlncRNA
7	malat1	LncRNADisease/CRlncRNA/MNDR 2.0
8	cdkn2b-as1	LncRNADisease
9	h19	LncRNADisease/CRlncRNA
10	hotair	LncRNADisease/CRlncRNA/MNDR 2.0

Table 3. Top 10 Glioma-Associated lncRNAs Predicted by LDASR

role in rapidly understanding the pathogenesis of disease and improving the quality of disease diagnosis, treatment, prognosis, and prevention but also confirms the great potential of machine learning in predicting the relationship between RNA and disease.

Limitations of the Study

There are several limitations in the current model. First, in the stage of characterizing lncRNA, we hope to fully combine and utilize a variety of information such as the sequence of lncRNA in the future instead of using only the Gaussian Interaction Profile Kernel Similarity. Second, Gaussian Interaction Profile Kernel Similarity is a traditional network representation learning method widely used in the embedding of bipartite graph nodes. With the rise of deep learning, novel network representation learning methods emerge in an endless stream, which can more effectively characterize the behavior of nodes and the structure of the entire network. We hope to take advantage of deep learning to improve prediction ability in the future. Third, all parameters are default in the process of constructing the model, and we believe that the performance of models can achieve a visible progress through the adjustment of the parameters.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

Num	lncRNA	Confirmed Database
1	pcat29	LncRNADisease/CRlncRNA/MNDR 2.0
2	tug1	Unconfirmed
3	malat1	LncRNADisease/CRlncRNA/MNDR 2.0
4	hif1a-as2	Unconfirmed
5	h19	LncRNADisease/CRlncRNA
6	dleu1	LncRNADisease/MNDR 2.0
7	dgcr5	Unconfirmed
8	cytor	Unconfirmed
9	cdkn2b-as3	Unconfirmed
10	cdkn2b-as11	LncRNADisease/CRlncRNA/MNDR 2.0/Lnc2Cancer

Table 4. Top 10 Prostate Cancer-Associated lncRNAs Predicted by LDASR

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.08.030.

ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China (Grant nos. 61722212 and 61572506) and in part by the Pioneer Hundred Talents Program of the Chinese Academy of Sciences. The authors would like to thank the editors and anonymous reviewers for their constructive advice.

AUTHOR CONTRIBUTIONS

Z.-H.G., Z.-H.Y., and Y.-B.W. considered the algorithm, arranged the datasets, and performed the analyses. H.-C.Y. and Z.-H.C. wrote the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 19, 2019
Revised: July 16, 2019
Accepted: August 19, 2019
Published: September 27, 2019

REFERENCES

Alvarez-Dominguez, J.R., and Lodish, H.F. (2017). Emerging mechanisms of long noncoding RNA function during normal and malignant hematopoiesis. Blood *130*, 1965–1975.

Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E., and Mattick, J.S. (2010). lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res. *39*, D146–D151.

Chen, X., Yan, C.C., Luo, C., Ji, W., Zhang, Y., and Dai, Q. (2015). Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. Sci. Rep. *5*, 11338.

Chen, X., and 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., and Gong, D.-W. (2016). IRWRLDA: improved random walk with restart for lncRNA-disease association prediction. Oncotarget *7*, 57919.

Chung, S., Nakagawa, H., Uemura, M., Piao, L., Ashikawa, K., Hosono, N., Takata, R., Akamatsu, S., Kawaguchi, T., and Morizono, T. (2011). Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. Cancer Sci. *102*, 245–252.

Congrains, A., Kamide, K., Oguro, R., Yasuda, O., Miyata, K., Yamamoto, E., Kawai, T., Kusunoki, H., Yamamoto, H., and Takeya, Y. (2012). Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B. Atherosclerosis *220*, 449–455.

Cui, T., Zhang, L., Huang, Y., Yi, Y., Tan, P., Zhao, Y., Hu, Y., Xu, L., Li, E., and Wang, D. (2017). MNDR v2. 0: an updated resource of ncRNA–disease associations in mammals. Nucleic Acids Res. *46*, D371–D374.

Dinger, M.E., Pang, K.C., Mercer, T.R., Crowe, M.L., Grimmond, S.M., and Mattick, J.S. (2008). NRED: a database of long noncoding RNA expression. Nucleic Acids Res. *37*, D122–D126.

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2017). Matrix factorization-based data fusion for the prediction of lncRNA–disease associations. Bioinformatics *34*, 1529–1537.

Huang, W., Tian, Y., Dong, S., Cha, Y., Li, J., Guo, X., and Yuan, X. (2017). The long non-coding RNA SNHG3 functions as a competing endogenous RNA to promote malignant development of colorectal cancer. Oncol. Rep. *38*, 1402–1410.

Johnson, R. (2012). Long non-coding RNAs in Huntington's disease neurodegeneration. Neurobiol. Dis. *46*, 245–254.

Lan, W., Li, M., Zhao, K., Liu, J., Wu, F.-X., Pan, Y., and Wang, J. (2016). LDAP: a web server for lncRNA-disease association prediction. Bioinformatics *33*, 458–460.

Lassmann, S., Weis, R., Makowiec, F., Roth, J., Danciu, M., Hopt, U., and Werner, M. (2007). Array CGH identifies distinct DNA copy number profiles of oncogenes and tumor suppressor genes in chromosomal-and microsatellite-unstable sporadic colorectal carcinomas. J. Mol. Med. *85*, 293–304.

Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., Zhao, G., Luo, H., Bu, D., and Zhao, H. (2011). Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. Nucleic Acids Res. *39*, 3864–3878.

Lu, C., Yang, M., Luo, F., Wu, F.-X., Li, M., Pan, Y., Li, Y., and Wang, J. (2018). Prediction of lncRNA–disease associations based on inductive matrix completion. Bioinformatics *34*, 3357–3364.

Ma, L., Li, A., Zou, D., Xu, X., Xia, L., Yu, J., Bajic, V.B., and Zhang, Z. (2014). LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs. Nucleic Acids Res. *43*, D187–D192.

Mamelak, A.N., and Jacoby, D.B. (2007). Targeted delivery of antitumoral therapy to glioma and other malignancies with synthetic chlorotoxin (TM-601). Expert Opin. Drug Deliv. *4*, 175–186.

Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., Gao, Y., Guo, M., Yue, M., and Wang, L. (2015). Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. Nucleic Acids Res. *44*, D980–D985.

Sallam, T., Sandhu, J., and Tontonoz, P. (2018). Long noncoding RNA discovery in cardiovascular disease: decoding form to function. Circ. Res. *122*, 155–166.

Wang, J., Zhang, X., Chen, W., Li, J., and Liu, C. (2018). CRlncRNA: a manually curated database of cancer-related long non-coding RNAs with experimental proof of functions on clinicopathological and molecular features. BMC Med. Genomics *11*, 114.

Wang, Y., Kuang, H., Xue, J., Liao, L., Yin, F., and Zhou, X. (2017). LncRNA AB073614 regulates proliferation and metastasis of colorectal cancer cells via the PI3K/AKT signaling pathway. Biomed. Pharmacother. *93*, 1230–1237.

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R., and Zhao, Y. (2013). NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. *42*, D98–D103.

Xue, J., Liao, L., Yin, F., Kuang, H., Zhou, X., and Wang, Y. (2018). LncRNA AB073614 induces epithelial-mesenchymal transition of colorectal cancer cells via regulating the JAK/STAT3 pathway. Cancer Biomarkers *21*, 1–10.

Yan, B., Wang, Z.-H., and Guo, J.-T. (2012). The research strategies for probing the function of long noncoding RNAs. Genomics *99*, 76–80.

Yang, X., Gao, L., Guo, X., Shi, X., Wu, H., Song, F., and Wang, B. (2014). A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. PLoS One *9*, e87797.

Yao, Y., Ma, J., Xue, Y., Wang, P., Li, Z., Liu, J., Chen, L., Xi, Z., Teng, H., and Wang, Z. (2015). Knockdown of long non-coding RNA XIST exerts tumor-suppressive functions in human glioblastoma stem cells by up-regulating miR-152. Cancer Lett. *359*, 75–86.

You, Z.-H., Huang, Z.-A., Zhu, Z., Yan, G.-Y., Li, Z.-W., Wen, Z., and Chen, X. (2017). PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. PLoS Comput. Biol. *13*, e1005455.

Yu, M., Xue, Y., Zheng, J., Liu, X., Yu, H., Liu, L., Li, Z., and Liu, Y. (2017). Linc00152 promotes malignant progression of glioma stem cells by regulating miR-103a-3p/FEZF1/CDC25A pathway. Mol. Cancer *16*, 110.

Zhang, Z., Hao, H., Zhang, C., Yang, X., He, Q., and 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.

ISCI, Volume *19*

Supplemental Information

A Learning-Based Method for LncRNA-Disease Association Identification Combing Similarity Information and Rotation Forest

Zhen-Hao Guo, Zhu-Hong You, Yan-Bin Wang, Hai-Cheng Yi, and Zhan-Heng Chen

Supplemental Figures

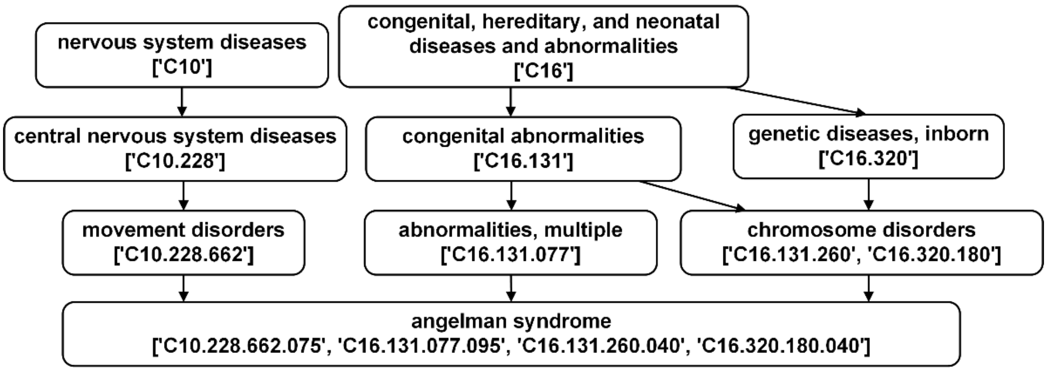


Figure S1. Construction of a disease's DAG. Related to Figure 1.

Transparent Methods

Data Collection

Known lncRNA-disease associations were downloaded from the LncRNADisease database (v2017) (Geng Chen *et al.*, 2012). which contained 2947 experimentally validated lncRNA–disease associations between 914 lncRNAs and 329 diseases. After deleting duplicate data caused by multiple experiment validations, we selected 1765 associations involving 881 lncRNAs and 328 diseases. The lncRNA–disease associations can be visualized as a network, the nodes represent specific lncRNA or disease, the edges connect a lncRNA to a disease. To extract positive and negative samples from this network, all experimentally validated lncRNA-disease pairs (i.e. 1765 lncRNA-disease pairs) constitute the golden standard positive dataset. The remaining edges of this network can be considered as nonassociation, and the corresponding lncRNA and disease can be collected as negative samples. In this paper, we followed previous method collect negative samples with the same size as positive samples using random selection (Ben-Hur and Noble, 2005). Although false negative samples may be included in the negative dataset, considering that the size of data collected only accounts for a small part of the whole network, the impact can be neglected. This can be treated as an issue with unbalanced data set processing, i.e. the process of down-sampling from negative sample (unlabeled sample). The picked negative samples are a very small percentage which only accounts for 0.61% (1765/ (881*328)-1765) and then a total of 3530 lncRNA–disease pairs were collected.

LncRNADisease v2017 and LncRNADisease v2012 are 2 different versions of the same database, of which v2012 is a true subset of v2017. The previous proposed by Chen *et al.* is to train and test based on lncRNADisease v2012, in order to ensure the fairness of the experiment, the 293 lncRNA–disease associations in version 2012 involving 118 lncRNAs and 167 diseases were also collected to constitute positive set. The negative set was constituted by the method mentioned above. As a result, the entire dataset consists of 586 lncRNA–disease pairs, of which half is from the positive samples and the other is from the negative samples.

Disease MeSH Descriptors And Directed Acyclic Graph

Medical Subject Headings (MeSH) is an authoritative subject vocabulary compiled by the National Library of Medicine, which provide a hierarchically-organized terminology for indexing and cataloging of various diseases. Each disease can be represented as a Directed Acyclic Graph (DAG) by the information provided by MeSH, which is described as follows: *DAG(D)= (D,* N_D, E_D*)*. Here, *D* represents specific diseases, N_D is node set that contains all disease in *D*'s DAG. E_D represents the relationship between the nodes in *D*'s DAG. Specific examples are shown in Figure S1.

Disease Semantic Similarity Matrix 1

We computed disease semantic similarity based on DAG. the contribution of disease *t* to the semantic value of disease *D* is defined as:

$$\begin{cases} D1_D(t) = 1 & if\ t = D \\ D1_D(t) = \max\{\Delta * D1_D(t') | t' \in children\ of\ t\} & if\ t \neq D \end{cases} \tag{1}$$

Where Δ denotes the semantic contribution decay factor and equals to 0.5. In the DAG on disease *D*, disease *D* is at the top, and its contribution to its semantic value is defined as 1. The semantic contribution of the next layer to disease *D* is equal to the contribution of the layer disease to itself multiplied by the semantic contribution attenuation factor. Therefore, the semantic value of disease A can be defined as follows:

$$D1(D) = \Sigma_{t \in N_D} D1_D(t) \tag{2}$$

The measure of disease similarity can be derived from set theory. The similarity between two diseases is calculated by the following：

$$DS1(i,j) = \frac{\Sigma_{t \in N_i \cap N_j}(D1_i(t) + D1_j(t))}{DV1(i) + DV1(j)} \tag{3}$$

Disease Semantic Similarity Matrix 2

The above disease similarity measure only considers local information and the intersection between two sets. Some scholars considered that it was one-sided and incomplete. Another semantic similarity measure method is used to complement the previous one. Inspired by information theory, the method suggests that diseases that often occur in DAGs should have a higher status and contribute more to other diseases (Xing Chen *et al.*, 2015, Xuan *et al.*, 2013). The new disease contribution values are measured as follows:

$$D2_D(t) = -\log\left(\frac{the\ number\ of\ DAGs\ including\ t}{the\ number\ of\ disease}\right) \tag{4}$$

The sum of the contributions of all nodes in the DAG of disease *D* is as follows:

$$DV2(\mathrm{D}) = \Sigma_{t \in N_D} D2_D(t) \tag{5}$$

The semantic similarity value could be calculated just like *DS1*:

$$DS2(i,j) = \frac{\Sigma_{t \in N_i \cap N_j}(D2_i(t) + D2_j(t))}{DV2(i) + DV2(j)} \tag{6}$$

Gaussian Interaction Profile Kernel Similarity For Diseases And LncRNA

In order to overcome the gap caused by the lack of MeSH information, the idea of collaborative filtering is employed to construct the third similarity matrix. In this paper, we first construct an adjacency matrix using the association data of lncRNA and disease. The columns of the matrix represent lncRNA and the rows represent diseases. Then, the Radial Basis Function (RBF) Gaussian kernel function was applied to adjacency matrix to obtain similarity matrix of disease (van Laarhoven,Nabuurs and Marchiori, 2011, Xing Chen *et al.*, 2018). The similarity defined by the Gaussian interaction profile kernel is as follows:

$$DG(i,j) = exp(-\alpha_d \left\| d_i - d_j \right\|^2) \tag{7}$$

Where d_i and d_j are *i*-th row and the *j*-th row of the adjacency matrix, respectively. α_d that is the weight factor used to regulate the kernel bandwidth, can be defined as follows:

$$\alpha_d = \alpha'_d \left(\frac{1}{nd} \sum_{i=1}^{nd} \| d_i \|^2\right) \tag{8}$$

Here, *nd* is the number of the diseases, the parameter α_d' is set to 0.5 empirically.
Analogous to the Gaussian similarity calculation method of disease, the Gaussian similarity of RNA is calculated by the same method. Formula 7 is replaced by Formula 9:

$$RS(i,j) = RG(i,j) = exp(-\alpha_r \left\| r_i - r_j \right\|^2) \quad (9)$$

Where r_i and r_j are *i*-th column and the *j*-th column of the adjacency matrix, respectively. α_r is the weight factor used to regulate the kernel bandwidth, defined by Formula (10):

$$\alpha_r = \alpha_r'(\frac{1}{nr}\sum_{i=1}^{nr}\|r_i\|^2) \quad (10)$$

Here, *nr* is the number of the diseases, the parameter α_r' is set to 0.5 empirically. After constructing the similarity matrix based on adjacency matric *A*, the representation vector of each lncRNA or disease will not change with cross-validation. The impact of this on the results will be discussed in a follow-up work.

Construction of Feature Vectors for Disease and lncRNA

Disease Semantic Similarity Matrix and Disease Gaussian Interaction Profile Kernel Similarity are two different types of information so neither is redundant. One of the above is often imperfect, to get a complete disease similarity matrix *DS*, we integrated disease semantic similarity matrix 1, disease semantic similarity matrix 2 and disease Gaussian interaction profile kernel similarity matrix by formula 11.

$$DS(i,j) = \begin{cases} \frac{DS1(i,j)+DS2(i,j)}{2} & if\ i\ and\ j\ have\ semantic\ similarity \\ DG(i,j) & otherwise \end{cases} \quad (11)$$

The row or column of matrix *DS* is regarded as the feature vectors of disease. Similarly, the row or column of matrix RS is regarded as the feature vectors of lncRNA. It's remarkable that all similarity matrices are symmetric matrices.
The *i*-th disease can be represented by the *i*-th row of the matrix *DS*:

$$DS_{i*} = (DS_{i1}, DS_{i2}, \dots, DS_{i328}) \quad (12)$$

The *j*-th disease can be represented by the *j*-th row of the similarity matrix RS:

$$RS_{j*} = (DS_{j1}, DS_{j2}, \dots, DS_{j881}) \quad (13)$$

The association consists of the *i*-th disease and the *j*-th lncRNA can be represented by the following vector:

$$Pair_{ij} = (DS_{i*}, RS_{j*}) = (DS_{i1}, DS_{i2}, \dots, DS_{i328}, RS_{j1}, RS_{j2}, \dots, RS_{j881}) \quad (14)$$

Each positive sample is given a label 1 and each negative sample is given a label 0.

AutoEncoder

Each association can be abstracted into a 1209-dimensional vector through the above step. Training set and test set consisting of thousands of such vectors take up a lot of storage space, which is not conducive to the training of classifiers. In order to reduce noise and improve feature quality, the autoencoder was used to obtain the optimal feature space from the original feature (Yi *et al.*, 2018). The autoencoder consists of an encoder and decoder. The coding part is responsible for compressing input data and the decoding part is responsible for restoring initial input. The main steps are as follows:
$f(x)$ is the activation function of the encoder, $g(h)$ is the activation function of the decoder. It will generally do this using a sigmoid function:

$$h = f(x) \coloneqq S_f(Wx + p) \quad (15)$$

$$y = g(h) \coloneqq S_g(W'x + q) \quad (16)$$

Here, we choose the sigmoid function as the activation function:

$$S_f(t) = S_g(t) = \frac{1}{1+e^{-t}} \tag{17}$$

The difference between *x* and *y* can be described by a reconstruction error function which is defined as follows:

$$L(x, y) = -\sum_{i=1}^{n}[x_i \, log(y_i) + (1 - x_i)log(1 - y_i)] \tag{18}$$

Through the above the loss function can be defined as follows:

$$Loss = \sum_{i=1}^{n} L(x_i, g(f(x_i))) \tag{19}$$

Therefore, the most suitable argument was obtained by minimizing the loss function. We can use *h* instead of *x* to represent the original vector. In this study, we used the keras library to implement the autoencoder and set the parameters batch size and epoch to 128 and 100, respectively.

RotationForest

Building an integrated learning algorithm by merging multiple models helps to achieve better prediction effects (Wang *et al.*, 2017, Li *et al.*, 2017). The idea of ensemble learning is to solve the defects and limitations inherent in the model of a single model by integrating more models. In 1990, Schapire analyzed and proved the equivalence between the weak learning algorithm and the strong learning algorithm based on the PAC (Probably Approximately Correct) learning model (Schapire, 1990). Since then, it has gradually attracted the focus of a wide range of scholars and shown outstanding effects on many classification or regression tasks. Assemble learning classifiers have stronger generalization capabilities and simpler parameter adjustments than traditional single models. Rotation Forest here was chosen to carry out the prediction. The Rotation Forest algorithm is based on the idea of feature transformation and focuses on improving the variability and accuracy of the base classifier (Rodriguez,Kuncheva and Alonso, 2006). Suppose $x = [x_1, x_2, \dots, x_n]^T$ represents the sample with *n* features. A matrix *X* of *N*n* to represent a training sample set with *N* data records. $y = [y_1, y_2, \dots, y_n]^T$ represents the corresponding sample class label in the training sample set *X. F* represents the attribute set, and *D1*, *D2*, …, *DL* represent *L* base classifiers. The main steps are as follows:

(1) The attribute set *F* is randomly divided into *K* sub-attribute sets, and each sub-attribute set contains about *M* = *n/K* attributes.

(2) Denote by $F_{i,j}$ the *j*-th subset of features for the training set of classifier *Di*. Then a bootstrap subset of objects is drawn with the size of 75% of the dataset to form a new training set, which is denoted by X'_{ij}. Using the selected subset of samples to transform the sub-attribute set in $F_{i,j}$, the principal component analysis (PCA) is used to obtain *Mj* principal components: $a_{ij}^1, a_{ij}^2, \dots, a_{ij}^{M_j}$.

(3) Repeat step 2 to store the obtained *K* principal component coefficients into a coefficient matrix *Ri.* According to the order of the original data attribute set, rearrange the matrix *Ri* to obtain R'_i, then the training set will be transformed into XR'_i. The base classifier *Di* will be trained on the new training set.

$$R_i = \begin{bmatrix} \left[a_{ij}^1, a_{ij}^2, \dots, a_{ij}^{M_1}\right] & [0] & \dots & [0] \\ [0] & \left[a_{ij}^1, a_{ij}^2, \dots, a_{ij}^{M_2}\right] & \dots & [0] \\ \vdots & \vdots & \vdots & \vdots \\ [0] & [0] & \dots \left[a_{ij}^1, a_{ij}^2, \dots, a_{ij}^{M_K}\right] & \end{bmatrix} \tag{20}$$

(4) After the above steps, *L* base classifiers can be obtained. The final prediction category is determined with maximum confidence.

Supplemental References

Ben-Hur, A. and Noble, W.S. (2005) 'Kernel methods for predicting protein–protein interactions'. *Bioinformatics,* 21 (suppl_1), pp. i38-i46.

Chen, G. *et al.* (2012) 'LncRNADisease: a database for long-non-coding RNA-associated diseases'. *Nucleic acids research,* 41 (D1), pp. D983-D986.

Chen, X. *et al.* (2018) 'Novel Human miRNA-Disease Association Inference Based on Random Forest'. *Molecular Therapy-Nucleic Acids,* 13 568-579.

Chen, X. *et al.* (2015) 'Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity'. *Scientific reports,* 5 11338.

Li, J.-Q. *et al.* (2017) 'PSPEL: in silico prediction of self-interacting proteins from amino acids sequences using ensemble learning'. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),* 14 (5), pp. 1165-1172.

Rodriguez, J.J., Kuncheva, L.I. and Alonso, C.J. (2006) 'Rotation forest: A new classifier ensemble method'. *IEEE transactions on pattern analysis and machine intelligence,* 28 (10), pp. 1619-1630.

Schapire, R.E. (1990) 'The strength of weak learnability'. *Machine learning,* 5 (2), pp. 197-227.

van Laarhoven, T., Nabuurs, S.B. and Marchiori, E. (2011) 'Gaussian interaction profile kernels for predicting drug–target interaction'. *Bioinformatics,* 27 (21), pp. 3036-3043.

Wang, L. *et al.* (2017) 'An ensemble approach for large-scale identification of protein-protein interactions using the alignments of multiple sequences'. *Oncotarget,* 8 (3), pp. 5149.

Xuan, P. *et al.* (2013) 'Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors'. *PloS one,* 8 (8), pp. e70204.

Yi, H.-C. *et al.* (2018) 'A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information'. *Molecular Therapy-Nucleic Acids,* 11 337-344.