doi:10.1093/jmcb/mjab005
Published online January 25, 2021

Application Note

Random distributed logistic regression framework for predicting potential lncRNA–disease association

Predicting potential lncRNA–disease association pairs is an important issue in the field of biomedicine. Traditional lncRNA–disease association prediction algorithms are mainly based on biological network models. For instance, RWRlncD is developed for predicting the potential lncRNA–disease association by performing random walk with restart on lncRNA functionally similar networks (Sun et al., 2014). However, biological graph network algorithms often have certain limitations and low accuracy. With the development of machine learning, new ideas have been brought to the construction of the association prediction algorithm (Liu et al., 2016, 2019a, b; Wan et al., 2019). Recently simboost algorithm based on matrix decomposition is developed for predicting drug–target and miRNA–disease potential associations (He et al., 2017; Chen et al., 2019). Previous studies have shown that simboost can extract relevant features from known association pairs and use machine learning algorithms to evaluate the possibility of association pairs with unknown labels. We propose that this algorithm framework can be improved and applied to lncRNA–disease association prediction.

The proposed algorithm that combines simboost feature extraction and logistic regression is named as random distributed logistic regression framework (RDLRF). The flow chart of the algorithm is shown in Figure 1A. To test the feasibility of RDLRF, we apply it to a dataset with 656 lncRNAs and 119 diseases. Based on known lncRNA–disease associations downloaded from lncRNA–disease database and lnc2cancer database, we implement leave-one-out cross-validation (LOOCV) to evaluate the performance of RDLRF. Since receiver operating characteristic (ROC) curves are widely used to evaluate model performance in previous literature of predicting lncRNA–disease associations, it is employed in this work to compare the performance of several models. RDLRF scores of lncRNA–disease pairs without association evidences can be obtained after implementing RDLRF. Area under the ROC curve (AUC) is calculated to quantitatively evaluate model performance. Under the same data conditions (i.e. the lncRNA similarity, disease similarity, and the known association between lncRNAs and diseases), we draw ROC curves and calculate AUC values for RDLRF algorithm, RDLRF algorithm with removing topological similarity steps (annotated with WTS, without topological similarity), IRWRLDA algorithm (Chen et al., 2016), HGLDA algorithm (Chen, 2015b), and KATZLDA algorithm (Chen, 2015a), respectively, based on the LOOCV. As shown in Figure 1B, the ROC curve of RDLRF algorithm almost contains the curves of other algorithms, which means that under the same conditions, the discrimination performance of RDLRF algorithm is optimal. From the perspective of AUC value, we can also see that the AUC value of RDLRF algorithm has reached 0.9429, far more than those of other algorithms. In addition, we remove the step of integrating topological similarity in the framework and calculate the performance of the algorithm separately, annotated with RDLRF (WTS), and the AUC value slightly decreases, but it still exceeds other algorithms. This shows that the integration of topological similarity steps can improve the performance of this framework, but compared with other classical algorithms, the biggest improvement comes from feature engineering and distributed logistic regression.

Since there is no definite conclusion about lncRNA similarity, it is necessary to discuss the parameter β of lncRNA integration similarity (see Supplementary Methods and materials). We divide the threshold from 0.1 to 0.9 in steps of 0.1 and then calculate the AUC values of RDLRF framework under different thresholds. The performance of RDLRF is not particularly sensitive to parameter changes, i.e. AUC fluctuates in a small range, and the best performance is achieved when the value is 0.6 (Figure 1C).

To demonstrate the ability of RDLRF to recover unknown lncRNA–disease association pairs, we implement case studies. We select the top 20 lncRNAs based on RDLRF score predicted for each disease and search the database for evidence supporting that the lncRNA–disease association did exist based on PubMed database (Supplementary Tables S1–S3). Note that the known association pairs are eliminated in this step. For breast cancer, recent evidence supports that 12 of the top 20 potential lncRNAs exist. In particular, for the top 13 potential lncRNAs, the evidence supports 76.9%

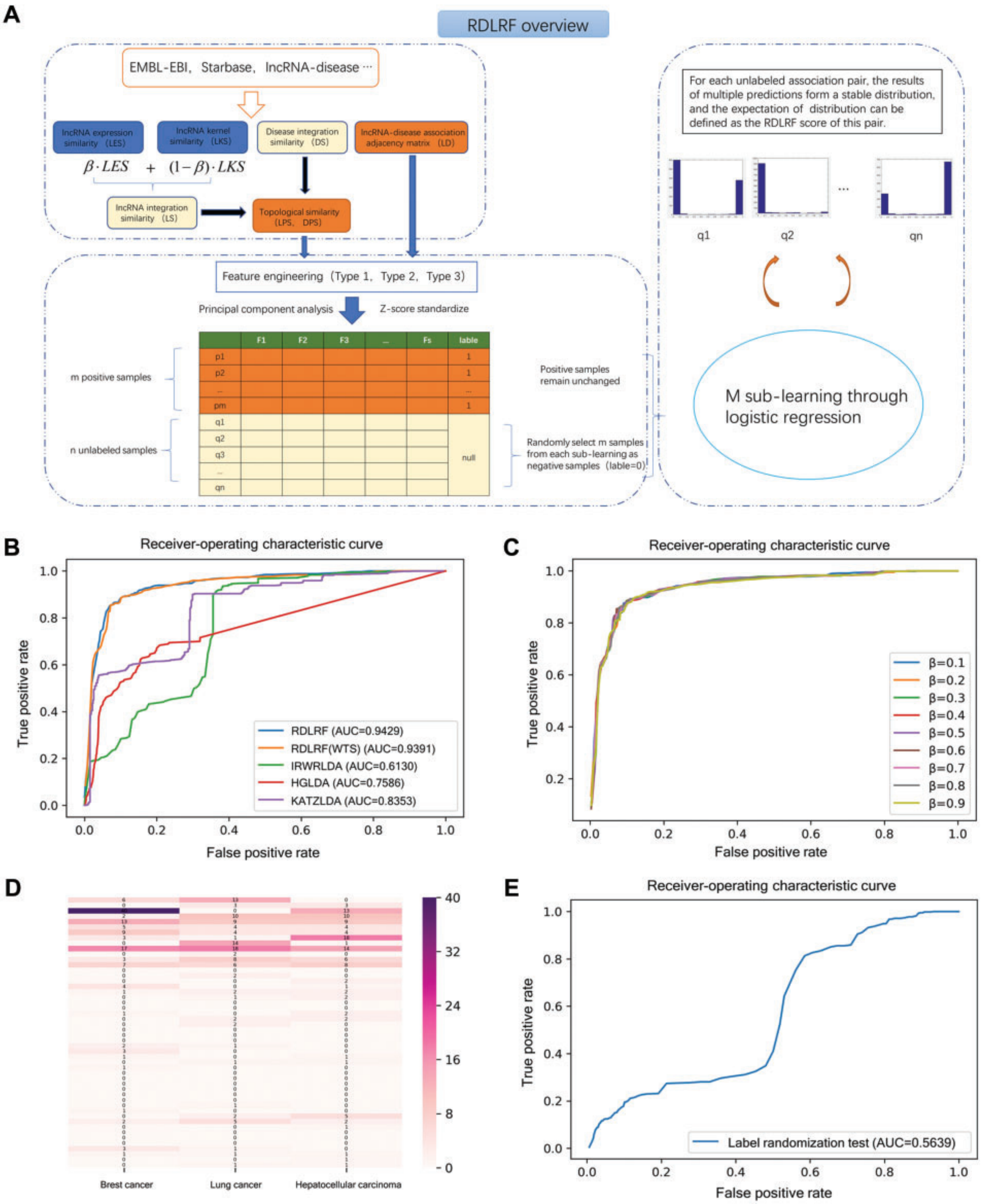

Figure 1 RDLRF for predicting potential lncRNA–disease association. (**A**) The proposed algorithm framework can be roughly divided into three parts. The first part is to calculate the similarity. The second part is feature engineering. The third part is integrated learning based on logistic regression. (**B**) Performance comparisons between RDLRF, RDLRF (WTS), IRWRLDA, HGLDA, and KATZLDA. RDLRF obtains an AUC value of 0.9429, ranking first among all selected algorithms. This shows that RDLRF has the best computing performance. (**C**) The effect of parameter β on RDLRF is not great, and when β is equal to 0.6, RDLRF can obtain the best performance. (**D**) Heat map of top 50 potential lncRNAs' PubMed hits. The higher the potential lncRNA ranking, the darker the color, which means it is supported by more PubMed literature. (**E**) The ROC curve of label randomization test. Contrary to the performance test, in the label randomization test, the ROC curve is close to the linear function (or the value of AUC is close to 0.5), indicating that the algorithm is not suffered from overfitting.

existing. Research by Tripathi et al. (2016) shows overexpression of lnc-MTAP (CDKN2B-AS1) and lnc-FAM (H19) in breast cells, which suggests that these lncRNAs may have significant roles to play in breast cancer. The RDLRF scores of CDKN2B-AS1 and H19 rank first and fifth, respectively (Supplementary Table S1). For lung cancer, 75% of the top 20 potential lncRNAs have been confirmed base on PubMed. According to the research of Loewen et al. (2014), lncRNA HOX transcript antisense RNA (HOTAIR) represses gene expression through recruitment of chromatin modifiers. The expression of HOTAIR is elevated in lung cancer and correlates with metastasis and poor prognosis. Moreover, HOTAIR promotes proliferation, survival, invasion, metastasis, and drug resistance in lung cancer cells. HOTAIR's RDLRF score ranks first among all potential lncRNAs associated with lung cancer (Supplementary Table S2). Hepatocellular carcinoma is one of the leading causes of cancer-related death and the mechanism of its progression remains poorly understood. Research by Huang et al. (2015) confirms that lncRNA ANRIL, as a growth regulator, can be used as a new biomarker and a therapeutic target for hepatocellular carcinoma. ANRIL's RDLRF score ranks first among all potential lncRNAs associated with hepatocellular carcinoma (Supplementary Table S3). In addition, we draw a heat map of the top 50 potential lncRNAs with highest RDLRF scores for three cancers (Figure 1D). If the color is darker, it means the corresponding lncRNA is supported by more PubMed literature.

In order to test whether RDLRF suffers from overfitting, we randomly mix '0' and '1' elements in the known association matrix of lncRNA and disease. As shown in Figure 1E, the AUC value is 0.5639 under the verification of LOOCV, and the curve displaying discrimination function is close to the random state, which shows that RDLRF effectively avoids overfitting. Comparing the different results of the original label and the mixed label, we can conclude that RDLRF is an effective tool to reveal more potential lncRNAs related to diseases.

Studying the key role of lncRNA in the biological process based on priori information is conducive to promoting the study of disease pathogenesis and treatment. We propose a computational model, RDLRF, which uses machine learning algorithms to rank the likelihood of potential disease-related lncRNAs while avoiding the limitations of traditional graph models. Specifically, in order to transform the association prediction problem into a binary classification problem that can be solved by machine learning algorithms, this study focuses on solving the two problems of feature extraction and negative samples missing. First of all, we define the new lncRNA similarity and disease similarity and introduce the restart random walk algorithm to fully consider the topological properties of each node in the network. Relying on techniques such as non-negative matrix factorization, we then extract effective features from similarity networks and known-related networks. Principal component analysis is used to remove redundancy and noise in high-dimensional features. Random sampling is used to obtain negative samples, the logistic regression is repeated multiple times to obtain the distribution formed by the probability of each candidate lncRNA, and the expectation of the distribution is used as the prediction result of RDLRF. Compared with traditional graph-based algorithms, RDLRF greatly improves its prediction performance and is more compatible with new lncRNAs or new diseases.

[Supplementary material is available at Journal of Molecular Cell Biology online. This research was supported by the Key Projects of National Natural Science Foundation of China (11831015), the Major Research Plan of National Natural Science Foundation of China (91730301), and Postgraduate Research & Practice Innovation Program of Jiangnan University (JNKY19_051).]

Yichen Sun, Hongqian Zhao, Gang Zhou, Tianhao Guan, Yujie Wang, and Jie Gao*

School of Science, Jiangnan University, Wuxi 214122, China
*Correspondence to: Jie Gao,
E-mail: gaojie@jiangnan.edu.cn

Edited by Zefeng Wang

References

Chen, X. (2015a). KATZLDA: KATZ measure for the lncRNA–disease association prediction. Sci. Rep. *5*, 16840.

Chen, X. (2015b). Predicting lncRNA–disease associations and constructing lncRNA functional similarity network based on the information of miRNA. Sci. Rep. *5*, 13186.

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

Chen, X., Zhu, C.C., and Yin, J. (2019). Ensemble of decision tree reveals potential miRNA–disease associations. PLoS Comput. Biol. *15*, e1007209.

He, T., Heidemeyer, M., Ban, F.Q., et al. (2017). SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. J. Cheminformatics *9*, 24.

Huang, M.D., Chen, W.M., Qi, F.Z., et al. (2015). Long non-coding RNA ANRIL is upregulated in hepatocellular carcinoma and regulates cell apoptosis by epigenetic silencing of KLF2. J. Hematol. Oncol. *8*, 50.

Liu, R., Wang, H.Y., Aihara, K., et al. (2019a). Hunt for the tipping point during endocrine resistance process in breast cancer by dynamic network biomarkers. J. Mol. Cell Biol. *11*, 649–664.

Liu, X.P., Chang, X., Leng, S.Y., et al. (2019b). Detection for disease tipping points by landscape dynamic network biomarkers. Natl Sci. Rev. *6*, 775–785.

Liu, X.P., Wang, Y.T., Ji, H.B., et al. (2016). Personalized characterization of diseases using sample-specific networks. Nucleic Acids Res. *44*, e164.

Loewen, G., Jayawickramarajah, J., Zhuo, Y., et al. (2014). Functions of lncRNA HOTAIR in lung cancer. J. Hematol. Oncol. *7*, 90.

Sun, J., Shi, H.B., Wang, Z.Z., et al. (2014). Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network. Mol. Biosyst. *10*, 2074–2081.

Tripathi, R., Soni, A., and Varadwaj, P. (2016). Integrated analysis of dysregulated lncRNA expression in breast cancer cell identified by RNA-seq study. Noncoding RNA Res. *1*, 35–42.

Wan, F.P., Hong, L.X., Xiao, A., et al. (2019). NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. Bioinformatics *35*, 104–111.