Bioinformatics, 33(3), 2017, 458–460
doi: 10.1093/bioinformatics/btw639
Advance Access Publication Date: 14 October 2016
Applications Note

Systems biology

LDAP: a web server for lncRNA-disease association prediction

Wei Lan[1], Min Li[1], Kaijie Zhao[1], Jin Liu[1], Fang-Xiang Wu[2], Yi Pan[3] and Jianxin Wang[1,*]

[1]School of Information Science and Engineering, Central South University, Changsha 410083, China, [2]Division of Biomedical Engineering, University of Saskatchewan, Saskatoon SKS7N5A9, Canada and [3]Department of Computer Science, Georgia State University, Atlanta, GA 30302, USA

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on June 1, 2016; revised on September 21, 2016; accepted on October 3, 2016

Abstract

Motivation: Increasing evidences have demonstrated that long noncoding RNAs (lncRNAs) play important roles in many human diseases. Therefore, predicting novel lncRNA-disease associations would contribute to dissect the complex mechanisms of disease pathogenesis. Some computational methods have been developed to infer lncRNA-disease associations. However, most of these methods infer lncRNA-disease associations only based on single data resource.
Results: In this paper, we propose a new computational method to predict lncRNA-disease associations by integrating multiple biological data resources. Then, we implement this method as a web server for lncRNA-disease association prediction (LDAP). The input of the LDAP server is the lncRNA sequence. The LDAP predicts potential lncRNA-disease associations by using a bagging SVM classifier based on lncRNA similarity and disease similarity.
Availability and Implementation: The web server is available at http://bioinformatics.csu.edu.cn/ldap
Contact: jxwang@mail.csu.edu.cn.
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Long noncoding RNAs (lncRNAs) are the biggest class of noncoding RNAs with greater than 200nt in length which can regulate gene expression at different levels including transcriptional, post-transcriptional and epigenetic regulation (Ponting *et al.*, 2009). Similar to mRNA transcripts, lncRNAs are transcribed by RNA polymerase II and have 3′ polyadenylation, 5′ cap characteristic. LncRNAs exhibit less conservative than other small noncoding RNAs such as microRNAs or snoRNAs in sequence and lower than protein coding genes in expression. LncRNAs play important roles in many biological processes such as chromosome dosage compensation, genomic imprinting, epigenetic regulation, nuclear and cytoplasmic trafficking, cell proliferation, cell differentiation, cell growth, metabolism, apoptosis and diseases. Some lncRNA-disease databases have been developed to store lncRNA-disease association dataset such as LncRNADisease (Chen *et al.*, 2013a) and Lnc2Cancer (Ning *et al.*, 2015).

Accumulating evidences indicate that lncRNAs have close associations with many human diseases (Wapinski and Chang, 2011). Identifying potential associations between lncRNAs and diseases contributes to exploring the complex pathogenesis and etiology of diseases. In recent years, several computational methods have been proposed to predict lncRNA-disease associations based on assumption that functionally similar lncRNAs tend to be related with phenotypically similar diseases (Jalali *et al.*, 2015). Chen *et al.* (2013b) developed a semi-supervised learning method, RLSLDA (Laplacian Regularized Least Squares for LncRNA-Disease Association), to infer potential lncRNA-disease associations based on lncRNA expression profiles and lncRNA-disease associations. Yang *et al.* (2013) presented a network-based method to predict

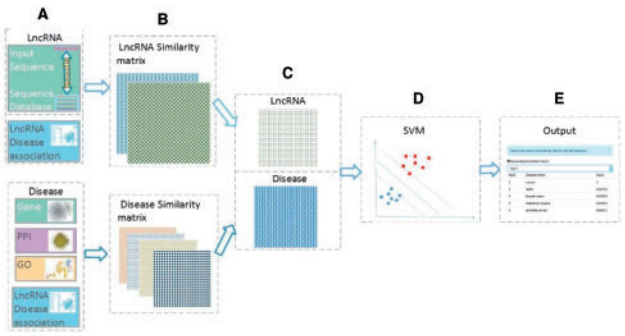

Fig. 1. The pipeline of LDAP. (**A**) Calculating similarities of lncRNAs and diseases, respectively. (**B**) Constructing similarity matrices of lncRNAs and diseases, respectively. (**C**) Fusing similarity matrices of lncRNAs and diseases in term of Karcher mean, respectively. (**D**) Predicting potential lncRNA-disease associations by using bagging SVM. (**E**) The output of predicted result (Color version of this figure is available at *Bioinformatics* online.)

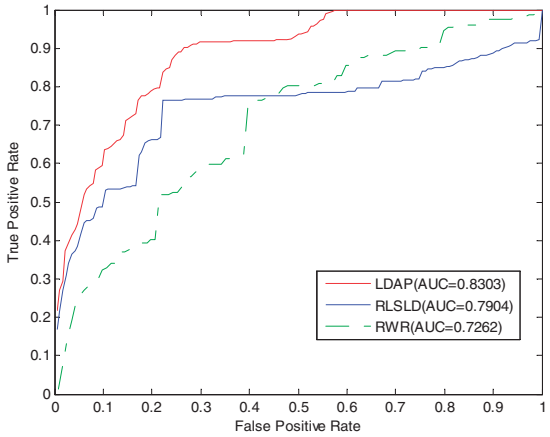

Fig. 2. ROC curves of different methods for predicting lncRNA-disease association (Color version of this figure is available at *Bioinformatics* online.)

lncRNA-disease associations by utilizing information propagation algorithm. Sun *et al.* (2014) constructed lncRNA functional similarity network and applied random walk with restart (RWR) to infer potential lncRNA-disease associations.

In this article, we present the LDAP web server for lncRNA-disease discovery by integrating multiple biological data resources. The geometric mean of matrix is employed to fuse different data resources while the bagging SVM is used to predict potential lncRNA-disease associations. The experimental results show that the performance of our method is superior to state-of-the-art methods for lncRNA-disease association prediction.

2 Methods

Figure 1 shows the pipeline of LDAP for predicting lncRNA-disease associations based on the integration of different data resources. In this pipeline, we use two methods (*LncRNA_Seq* and *LncRNA_Gip*) for lncRNA similarity measurement and five methods (*Dis_Icod*, *Dis_Top*, *Dis_Gf*, *Dis_GO* and *Dis_Gip*) for disease similarity measurement based on different data resources (for more details see the Supplementary Material). The Karcher mean of matrices is employed to fuse similarity matrices of lncRNAs and diseases, respectively (Zakeri *et al.*, 2014). For a set of similarity matrices:$K_1 \cdots K_m$, their Karcher mean (K) is defined as:

$$K = \underset{X \in S}{argmin} \sum_{i=1}^{m} \| \log (K_i^{-1/2} X K_i^{-1/2}) \|_F^2$$

where S denotes the set of all semi-positive matrices and F denotes the Frobenius norm. The bagging SVM is used to identify potential lncRNA-disease associations (Mordelet *et al.*, 2014). It assumes that positive unlabeled learning problems have a particular structure that leads to instability of classifiers while bagging can be used to enhance the performance of instable classifiers (Claesen *et al.*, 2015). In practice, the resampling method can be used to obtain multiple sub-samples. The weighted SVM is employed to build classifiers to discriminate positive samples from each sub-sample. At the last, the predicted lncRNA-disease associations are outputted.

3 Implementation and experiment

The input of LDAP web server is a lncRNA sequence or a txt file with multiple sequences with FASTA format. The sequence should be greater than 200nt in length. The user can upload a sequence or text with multiple sequences each time. It should be noted that the user is required to provide the email address when submitting text with multiple sequences. In each submission, a job ID will be assigned. When the job is finished, the result will be displayed. In case that a user submits text with multiple sequences, the link of result page will be sent to the user by email.

In order to test the performance of the proposed LDAP, we implemented the leave-one-out-cross validation. We compare our LDAP with two state-of- the-art methods: RWR (Sun *et al.*, 2014) and RLSLDA (Chen *et al.*, 2013b). The area under the curve (AUC) is used to evaluate the performance of methods. Figure 2 shows the prediction performance of three methods. Comparing with RWR (0.7262) and RLSLDA (0.7904), the LDAP obtains the highest value (0.8303) of AUC. It demonstrates that LDAP is an accurate and effective method for recovering the known lncRNA-disease associations.

According to the statistical data of American cancer society, over 200 000 women and 2000 men are diagnosed with invasive breast cancer each year in the United States. The pathogenic mechanism of breast cancer is viewed as the result of interaction between the environmental factor and the genetically susceptible host (Friedenson *et al.*, 2007). Many lncRNAs have association with breast cancer via up-regulating or down-regulating of breast cancer genes. For example, long non-coding RNA UCA1 promotes breast tumor growth through interaction with the 5′-untranslated region (5′-UTR) of p27 mRNAs to suppress of p27 (Kip1) gene expression (Huang *et al.*, 2014). Table 1 shows the top 10 potential breast cancer related lncRNAs which are predicted by LDAP. Amount the top 10 predicted breast cancer related lncRNAs, 6 lncRNAs are validated by recent publications. WRAP53 ranked at top 2 is related with breast cancer. According to the recent researches (Garcia-Closas *et al.*, 2007, Mahmoudi *et al.*, 2011), single-nucleotide polymorphisms (SNPs) in WRAP53 were found to be overrepresented in women with breast cancer, in particular estrogen receptor-negative breast cancer. Ube3a-as ranked at top 3 is associated with breast cancer. It is verified that Ube3a-as can encode the genen E6AP and several potential substrates of E6AP have been reported, including human homolog of Rad23 A (HHR23A) and HHR23B, amplified in breast cancer 1 protein (AIB1) (Kühnle *et al.*, 2013, Mani *et al.*, 2006). The literature (Ahmed *et al.*, 2010) shows DAPK1 ranked at top 4 is found to be related with breast cancer or benign breast diseases. The SNP in human RRP1B ranked at top 5 has been proved to associate with breast cancer development and progression (Nanchari *et al.*,

Table 1. The top ten predicted lncRNA of breast cancer

Rank	lncRNA	References
1	PDZRN3-AS1	Unknown
2	WRAP53	Garcia-Closas et al. (2007) and Mahmoudi et al. (2011)
3	Ube3a-as	Kühnle et al. (2013) and Mani et al. (2006)
4	DAPK1	Ahmed et al. (2010)
5	RRP1B	Nanchari et al. (2015)
6	DLX6-AS1	Unknown
7	SCAANT1	Unknown
8	HAR1A	Gumireddy et al. (2013)
9	PCGEM1	Enciso-Mora et al. (2010a,b)
10	DAOA-AS1	Unknown

2015). The HAR1A ranked at top 8 is associated with SF3B3 in breast cancer cells (Gumireddy *et al.*, 2013). The PCGEM1 ranked at top 9 facilitates the transcription of ERα target genes in the absence of estrogen in breast cancer cells (Enciso-Mora *et al.* 2010a,b). In addition, 4 lncRNAs are not found in literature. The functions of these lncRNAs are still unknown which is deserved for biologists to validated their functions via biological experiments.

4 Conclusion

In this article, we have presented the LDAP web server for discovering lncRNA–disease associations based on multiple data resources. Two lncRNA similarity and five disease similarity methods are employed to calculate similarities between lncRNA–lncRNA and disease–disease, respectively. We use the geometric mean of matrix to fuse lncRNA and disease similarities, respectively. The bagging SVM is employed to identify potential lncRNA–disease associations. The experimental results have shown that our approach is able to identify known and potential new lncRNA–disease associations.

Acknowledgements

The authors would like to thanks anonymous reviews for their useful comments to improve the quality of this paper.

Funding

This work was supported in part by the National Natural Science Foundation of China No. 61232001, No. 61420106009, No. 61428209, No. 61622213 and No. 61370712.

Conflict of Interest: none declared.

References

Ahmed,I.A. *et al.* (2010) Epigenetic alterations by methylation of RASSF1A and DAPK1 promoter sequences in mammary carcinoma detected in extracellular tumor DNA. *Cancer Genet. Cytogenet.*, **199**, 96–100.

Chen,G. *et al.* (2013a) LncRNADisease: a database for long noncoding RNA associated diseases. *Nucleic Acids Res.*, **41**, 983–986.

Chen,X. *et al.* (2013b) Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics*, **29**, 2617–2624.

Claesen,M. *et al.* (2015) A robust ensemble approach to learn from positive and unlabeled data using SVM base models. *Neurocomputing*, **160**, 73–84.

Enciso-Mora,V. *et al.* (2010a) Risk of breast and prostate cancer is not associated with increased homozygosity in outbred populations. *Eur. J. Hum. Genet.*, **18**, 909–914.

Enciso-Mora,V. *et al.* (2010b) Mechanisms of endocrine resistance in breast cancer: an overview of the proposed roles of noncoding RNA. *Breast Cancer Res.*, **17**, 40.

Friedenson,B. *et al.* (2007) The BRCA1/2 pathway prevents hematologic cancers in addition to breast and ovarian cancers. *BMC Cancer*, **7**, 152–162.

Garcia-Closas,M. *et al.* (2007) Common genetic variation in TP53 and its flanking genes, WDR79 and ATP1B2, and susceptibility to breast cancer. *Int. J. Cancer*, **121**, 2532–2538.

Gumireddy,K. *et al.* (2013) Identification of a long non-coding RNA-associated RNP complex regulating metastasis at the translational step. *EMBO J.*, 2013, **32**, 2672–2684.

Huang,J. *et al.* (2014) Long non-coding RNA UCA1 promotes breast tumor growth by suppression of p27 (Kip1). *Cell Death Dis.*, **5**, e1008.

Jalali,S. *et al.* (2015) Computational approaches towards understanding human long non-coding RNA biology. *Bioinformatics*, **31**, 2241–2251.

Kühnle,S. *et al.* (2013) Role of the ubiquitin ligase E6AP/UBE3A in controlling levels of the synaptic protein Arc. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 8888–8893.

Mahmoudi,S. *et al.* (2011) WRAP53 promotes cancer cell survival and is a potential target for cancer therapy. *Cell Death Dis.*, **2**, e114.

Mani,A. *et al.* (2006) E6AP mediates regulated proteasomal degradation of the nuclear receptor coactivator amplified in breast cancer 1 in immortalized cells. *Cancer Res.*, **66**, 8680–8686.

Mordelet,F. *et al.* (2014) A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognit. Lett.*, **37**, 201–209.

Nanchari,S.R. *et al.* (2015) Rrp1B gene polymorphism (1307T>C) in metastatic progression of breast cancer. *Tumour Biol.*, **36**, 615–621.

Ning,S.W. *et al.* (2015) Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.*, **44**, D980–D985.

Ponting,C.P. *et al.* (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.

Sun,J. *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.

Wapinski,O. and Chang,H.Y. (2011) Long noncoding RNAs and human disease. *Trends Cell Biol.*, **21**, 354–361.

Yang,X.F. *et al.* (2013) A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLoS One*, **9**, e87797.

Zakeri,P. *et al.* (2014) Protein fold recognition using geometric kernel data fusion. *Bioinformatics*, **30**, 1850–1857.