LDAP: a web server for IncRNA-disease association prediction

Wei Lan¹, Min Li¹, Kaijie Zhao¹, Jin Liu¹, Fang-Xiang Wu², Yi Pan³ and Jianxin Wang¹,*

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

*To whom correspondence should be addressed.

Abstract

Motivation: Increasing evidences have demonstrated that long noncoding RNAs (lncRNAs) play important roles in many human diseases. Therefore, predicting novel IncRNA-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 IncRNA-disease associations by integrating multiple biological data resources. Then, we implement this method as a web server for IncRNA-disease association prediction (LDAP). The input of the LDAP server is the IncRNA sequence. The LDAP predicts potential IncRNA-disease associations by using a bagging SVM classifier based on IncRNA 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 non-coding 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
In 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 method, we use two methods (LncRNA_Seq and LncRNA_Gimp) for lncRNA similarity measurement and five methods (Disicod, Dis_Top, Dis_Gf, Dis_GO and Dis_Gimp) 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 (Zakerti et al., 2014). For a set of similarity matrices $K_1,\ldots,K_m$, their Karcher mean (K) is defined as:

$$K = \arg \min_{S \in S} \sum_{s=1}^{m} \| \log (K_s^{-1/2}K^{-1/2}) \|_F^2$$

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., 2019). 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., 2010).

---

**Figure 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.)

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

Table 1. The top ten predicted lncRNA of breast cancer

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

References


