Genomics Proteomics Bioinformatics 14 (2016) 62–71

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

A Bipartite Network-based Method for Prediction of Long Non-coding RNA–protein Interactions

Mengqu Ge [1,a], Ao Li [1,2,*,b], Minghui Wang [1,2,c]

[1] *School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China*
[2] *Centers for Biomedical Engineering, University of Science and Technology of China, Hefei 230027, China*

Received 12 November 2015; revised 4 January 2016; accepted 6 January 2016
Available online 22 February 2016

Handled by Zhihua Zhang

KEYWORDS

lncRNA;
Protein;
Interaction;
Bipartite network;
Propagation

Abstract As one large class of non-coding RNAs (ncRNAs), long ncRNAs (**lncRNAs**) have gained considerable attention in recent years. Mutations and dysfunction of **lncRNAs** have been implicated in human disorders. Many **lncRNAs** exert their effects through **interactions** with the corresponding RNA-binding **proteins**. Several computational approaches have been developed, but only few are able to perform the prediction of these **interactions** from a network-based point of view. Here, we introduce a computational method named lncRNA–protein **bipartite network** inference (LPBNI). LPBNI aims to identify potential lncRNA–interacting **proteins**, by making full use of the known lncRNA–protein **interactions**. Leave-one-out cross validation (LOOCV) test shows that LPBNI significantly outperforms other network-based methods, including random walk (RWR) and protein-based collaborative filtering (ProCF). Furthermore, a case study was performed to demonstrate the performance of LPBNI using real data in predicting potential lncRNA–interacting proteins.

Introduction

An increasing number of studies show that approximately 2% of the whole mammalian genome represents protein-coding genes, whereas the majority of the genome consists of non-coding RNA (ncRNA) genes. ncRNAs had long been regarded as transcriptional noise, but recent investigations demonstrate that ncRNAs play an important role in the regulation of diverse biological processes [1–5]. Long ncRNAs (lncRNAs), which consist of more than 200 nucleotides, constitute a large class of ncRNAs [6–7]. In the past several years, the number of identified lncRNAs has been increasing sharply because of the development of both bioinformatics tools and experimental techniques. Functional studies of lncRNAs show that mutated and dysfunctional lncRNAs are implicated in a range of cellular processes [8–12] and human diseases, ranging from neurodegeneration to cancer [13–18]. Although some lncRNAs, *e.g.*, Xist [19] and

* Corresponding author.
E-mail: aoli@ustc.edu.cn (Li A).
[a] ORCID: 0000-0001-5185-7045.
[b] ORCID: 0000-0001-9910-8967.
[c] ORCID: 0000-0002-5788-894X.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2016.01.004

MALAT1 [20], have been well studied, the functions of most lncRNAs remain unclear. Usually lncRNAs function through interacting with RNA-binding proteins (RBPs) [21–24]. Therefore, it is important to predict the potential lncRNA–protein interactions, in order to study the complex function of lncRNAs.

Since the experimental identification of lncRNA–protein interactions remains costly, developing effective predictive approaches becomes essential. Recently, several computational methods have been reported for predicting potential lncRNA–protein interactions. For instance, Bellucci et al. developed catRAPID in 2011 [25] by taking into account secondary structure, hydrogen bonds, and van der Waals forces between lncRNAs and proteins. Next, Muppirala et al. [26] introduced a method named RPISeq, using only sequence information of lncRNAs and proteins. Support vector machine (SVM) classifiers [27] and random forest (RF) [28] are used to predict RBPs. In 2013, Lu et al. [29] developed a novel approach, named lncPro, which uses secondary structure, hydrogen bond, van der Waals force features, and yields the prediction score using Fisher's linear discriminate method. Later on, an approach named RPI-Pred was developed by Suresh et al. [30], they trained SVM-based approach, by extracting sequence and high-order 3D structure features of lncRNAs and proteins.

All the aforementioned methods are based on the biological characteristics of ncRNAs and proteins. CatRAPID and lncPro combined sequence and structural features of lncRNAs and proteins. RPISeq was based on sequence features. RPI-Pred used the high-order structure features of lncRNAs and proteins. However some studies show that lncRNAs generally exhibit low sequence conservation [1], which may make it difficult to predict interactions based on the intrinsic properties of lncRNAs. Biological network-based methods are widely used in many types of studies, such as disease gene prioritization [31] and drug-target interaction prediction [32]. The development of bioinformatics technologies such as CLIP-seq and cross-linking immunoprecipitation, has enabled us to construct lncRNA–protein interaction networks. We introduce here a novel computational method, lncRNA–protein bipartite network inference (LPBNI), for the prediction of lncRNA–protein interactions. LPBNI identifies novel lncRNA–protein pairs by efficiently using the lncRNA–protein bipartite network. In order to evaluate the performance of the proposed method, we compared LPBNI with other network-based methods, including random walk (RWR) [31] and protein-based collaborative filtering (ProCF) [33]. RWR [31] has been used to predict genes associated with potential diseases. ProCF is derived from the recommendation algorithms, similar to the item-based collaborative filtering method [33]. The performance evaluation is based on leave-one-out cross validation (LOOCV) of the known lncRNA–protein interactions extracted from NPInter [34]. To further demonstrate the effectiveness of lncRNA–protein bipartite network, six lncRNAs were used to evaluate the performance of LPBNI in comparison with the existing methods, lncPro [29] and RPISeq [26]. These evaluation tests demonstrated that LPBNI outperforms the other methods significantly. In a case study, several potential interactions between lncRNAs and proteins identified by LPBNI were well supported by starBase [35], indicating the superior predictive ability of LPBNI.

Results

Performance comparison with other network-based methods on lncRNA–protein interactions prediction

We compared the performance of LPBNI with RWR [31] and ProCF [33]. ProCF is based on the idea that if a protein interacts with an lncRNA, similar proteins will be recommended as interacting with this lncRNA. The linkage between p_i and l_j can be defined as: $score_{ij} = \frac{\sum_{k=1,k\neq i}^{m} S_P(p_i,p_k)a_{kj}}{\sum_{k-1,k\neq i}^{m} S_P(p_i,p_k)}$, where $S_P(p_i,p_k)$ is the similarity between proteins p_i and p_k. Here, we used cosine vector similarity to measure the similarity of proteins: $S_P(p_i,p_k) = \frac{|d(i)\cap d(k)|}{\sqrt{|d(i)||d(k)|}}$, where $d(i)$ and $d(k)$ are the degrees of proteins i and k, respectively.

We extracted 4870 lncRNA–protein interactions from NPInter 2.0 [34] (see "*Data collection and preprocessing*" for detail). In LPBNI, for one node, at least two interactions are required to perform LOOCV. Therefore, the nodes that have only one link are not considered in the performance evaluation, so we further get 4796 lncRNA–protein interactions which match that condition, and this dataset is taken as 'gold standard' data in the LOOCV test. The receiver operating characteristic (ROC) curves and the area under the curve (AUC) obtained using these methods are shown in **Figure 1**. It is obvious that LPBNI shows the highest true positive rate (TPR) at each false positive rate (FPR). In addition, the AUC value of LPBNI is 0.878 (**Table 1**), which is higher than that obtained using RWR (0.765) and ProCF (0.738), respectively. These data suggest that LPBNI has a better predictive ability compared with RWR and ProCF. To validate the reliability of LPBNI, we compared the sensitivity, accuracy, precision, and Matthew's correlation coefficient (MCC) of LPBNI, RWR, and ProCF with specificities of 99.0% and 95.0%, respectively. As shown in Table 1, with specificity of 99.0%, sensitivity, accuracy, precision, and MCC of LPBNI are all higher than that with RWR and ProCF. When specificity was reduced to 95.0%, sensitivity and MCC increased for all three methods, with decreased precision, although the accuracy remained comparable. However, LPBNI still showed a higher performance in terms of sensitivity, accuracy, precision, and MCC, compared to RWR and ProCF.

The fold enrichment is also used to evaluate the performance of the proposed method, which can be defined as: $N/2/n$ [37], where N represents the number of candidate proteins, and n is the ranking of the tested protein among the candidate proteins for the evaluation. Based on the formula, the average fold enrichments are 4.007, 3.590, and 1.653 for LPBNI, RWR, and ProCF, respectively. These data suggest that LPBNI outperforms the other methods in identifying lncRNA–related proteins with a higher rank. **Table 2** shows the number of lncRNA–protein interactions that were correctly retrieved at 5%, 10%, 15%, 20%, and 50% of all the prediction results, respectively. Among 4796 true interactions between lncRNAs and proteins, LPBNI achieves a higher retrieval compared with RWR and ProCF, at each of the investigated percentiles. The biggest difference was observed for 5%, where LPBNI recovered 579 interactions successfully, and only 410 and 116 interactions were retrieved using RWR and ProCF, respectively.