Bioinformatics, 34(11), 2018, 1953–1956
doi: 10.1093/bioinformatics/bty002
Advance Access Publication Date: 22 January 2018
Applications Note

Systems biology

DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function

Liang Cheng[1,†], Yang Hu[2,†], Jie Sun[1], Meng Zhou[1,*] and Qinghua Jiang[2,*]

[1]Department of College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang Sheng 150081, China and [2]Department of Bioinformatics, School of Life Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang Sheng 150001, China

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Cenk Sahinalp

Received on May 22, 2017; revised on November 11, 2017; editorial decision on January 2, 2018; accepted on January 22, 2018

Abstract

Summary: DincRNA aims to provide a comprehensive web-based bioinformatics toolkit to elucidate the entangled relationships among diseases and non-coding RNAs (ncRNAs) from the perspective of disease similarity. The quantitative way to illustrate relationships of pair-wise diseases always depends on their molecular mechanisms, and structures of the directed acyclic graph of Disease Ontology (DO). Corresponding methods for calculating similarity of pair-wise diseases involve Resnik's, Lin's, Wang's, PSB and SemFunSim methods. Recently, disease similarity was validated suitable for calculating functional similarities of ncRNAs and prioritizing ncRNA–disease pairs, and it has been widely applied for predicting the ncRNA function due to the limited biological knowledge from wet lab experiments of these RNAs. For this purpose, a large number of algorithms and priori knowledge need to be integrated. e.g. 'pair-wise best, pairs-average' (PBPA) and 'pair-wise all, pairs-maximum' (PAPM) methods for calculating functional similarities of ncRNAs, and random walk with restart (RWR) method for prioritizing ncRNA–disease pairs. To facilitate the exploration of disease associations and ncRNA function, DincRNA implemented all of the above eight algorithms based on DO and disease-related genes. Currently, it provides the function to query disease similarity scores, miRNA and lncRNA functional similarity scores, and the prioritization scores of lncRNA–disease and miRNA–disease pairs.

Availability and implementation: http://bio-annotation.cn:18080/DincRNAClient/

Contact: biofomeng@hotmail.com or qhjiang@hit.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Disease similarity researches include the designment of the quantitatively measure to calculate the similarity of pair-wise diseases (SPWD) and similarity of pair-wise disease sets (SPWDS). This domain attracted much more attention since its wide application in establishing functional similarity network (FSN) of non-coding RNAs (ncRNAs) (Sun *et al.*, 2014; Wang *et al.*, 2010), and predicting ncRNA–disease associations (Sun *et al.*, 2014) and so on.

The development of Disease Ontology (DO) (Kibbe *et al.*, 2015) provides a way to investigate disease similarity using system biology methods, because it is the first vocabulary established around disease names. The state-of-art methods for calculating the similarity of pair-wise DO terms involve Wang's, Resnik's, Lin's process-similarity based (PSB) and SemFunSim methods (Cheng *et al.*, 2014; Lin, 1998; Mathur and Dinakarpandian, 2012; Resnik, 1995; Wang *et al.*, 2007). In addition, the commonly used method for calculating SPWDS based on SPWD is 'pair-wise best, pairs-average' (PBPA) (Sun *et al.*, 2014), and an alternative method is 'pair-wise all, pairs-maximum' (PAPM) (Pesquita *et al.*, 2009). Recently, disease similarity was widely used for predicting the ncRNA function and prioritizing ncRNA–disease pairs based on random walk with restart (RWR) method as the lack of proteins from these RNAs. First, the functional similarity of pair-wise ncRNAs is converted to the similarity of their inducing disease sets, which can be calculated based on PBPA and PAPM methods. Then the similarities of all the pair-wise ncRNAs are used to construct ncRNA FSN, where each RNA is deemed as a node and the functional similarity score as the weight of edge. Finally, the network is utilized to prioritize ncRNA–disease pairs using a RWR method. The details of these methods were described in the 'Supplementary Methods' section of Supplementary Material.

Although SPWDS is widely applied in computing functional similarity of ncRNAs and predicting novel ncRNA–disease association, no tools were designed for the purpose nowadays. In this paper, we presented a comprehensive toolkit to explore disease associations and ncRNA function from the perspective of disease similarity. The toolkit provides tools for retrieving SPWD, SPWDS, lncRNA functional similarity (LFS), miRNA functional similarity (MFS), prioritization of lncRNA–disease pair and prioritization of miRNA–disease pair. Our toolkit is freely available at http://bio-annotation.cn:18080/DincRNAClient/.

2 Materials and methods

2.1 Priori semantic associations between diseases

Semantic associations between diseases were extracted from DO. Currently, DO contains 7124 'IS_A' relationships between 6920 disease terms.

2.2 Datasets for calculating disease similarity

Disease-related genes, gene-related biological processes (BPs) and the human gene functional network are utilized for calculating disease similarity. Priori disease–gene associations are scattered in multiple manual databases. Main of these databases include Gene Reference into Function (GeneRIF) (Mitchell *et al.*, 2003), Online Mendelian Inheritance in Man (OMIM) (Amberger *et al.*, 2011), Genetic Association Database (GAD) (Becker *et al.*, 2004) and Comparative Toxicogenomics Database (CTD) (Davis *et al.*, 2013). After mapping disease names in these four databases to DO terms based on SIDD (Cheng *et al.*, 2013), the integrated disease–gene associations are obtained. In addition, gene-related BP is from GO Annotation (GOA) (Camon, 2004), and the human gene functional network are from HumanNet (Lee *et al.*, 2011).

2.3 Priori ncRNA–disease associations

Here ncRNA–disease associations contain lncRNA–disease associations and miRNA–disease associations. lncRNA–disease associations are from LncRNADisease (Chen *et al.*, 2013) and Lnc2Cancer (Ning *et al.*, 2016). miRNA–disease associations are from HMDD v2.0 (Li *et al.*, 2014). All of the disease names in these databases are manually mapped to DO terms. As a result, we get 5710 associations between 265 diseases and 556 miRNAs, and 596 associations between 161 diseases and 343 lncRNAs.

2.4 Designment and implementation of the system

The three-layer architecture involving DATABASE, ALGORITHM and TOOLS layer is designed in the Figure 1. All the datasets are stored in the relational database management system MySQL 5.5 as DATABASE layer. Eight methods involving Wang's, Resnik's, Lin's, PSB, SemFunSim, PBPA, PAPM and RWR methods are implemented in ALGORITHM layer. DincRNA has been implemented on a JavaEE framework and run on the web server [2-core (2.26 GHz) processors] of UCloud (Sqalli *et al.*, 2012). It provides web application by a typical browser/server model in Apache Tomcat container. The querying results are packed into JSON objects for transferring and displaying. In the front end of the DincRNA, the Asynchronous JavaScript and XML (AJAX) technique is adopted for exchanging data asynchronously between the browser and the server to avoid full page reloads.

3 Results

3.1 Web interface

The DincRNA provides web pages for users to query disease similarity score, ncRNA functional similarity score, and ncRNA–disease prioritization score. All of these scores can be downloaded from 'Resources' page or result page.

The DincRNA provides a search engine to query entities involving disease names, DOIDs, lncRNA symbols and miRNA symbols in the input page. The autocomplete function of the page can help users to input the interested diseases and ncRNA names easily. To support the calculation of the similarities of multiple pairs of disease sets online, the DincRNA also provides the batch processing function which allows users to input disease sets by files (Supplementary Fig. S2).

Each entry of query results by DincRNA contains names of the pair-wise entities and their similarity score or prioritization score. The result page provides the paging and sorting functions to show all of the entries. Each page can show 20, 50 or 100 entries as required based on paging function (Supplementary Fig. S2). Entries can be displayed in descending or ascending order of similarity score or prioritization score by sorting function (Supplementary Fig. S2). In addition, all the query results can be downloaded in JSON and CSV formats (Supplementary Fig. S2) for easing to browse the querying results locally. Detailed usage of the DincRNA is described in the Supplementary Material.

3.2 Performance evaluation of the prioritization of ncRNA–disease pairs

Since SemFunSim, PSB, Wang's, Lin's and Resnik's methods can be combined with PBPA or PAPM methods for prioritizing

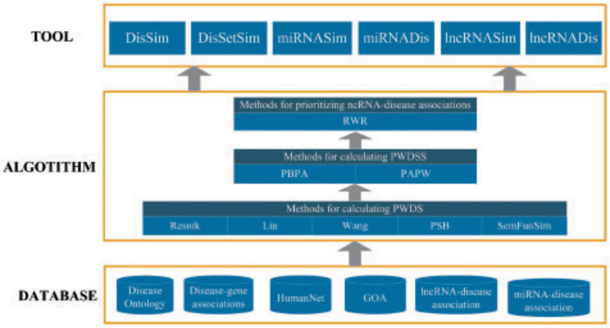
TOOL
DisSim
DisSetSim
miRNASim
miRNADis
lncRNASim
lncRNADis
ALGOTITHM
Methods for prioritizing ncRNA-disease associations
RWR
Methods for calculating PWDSS
PBPA
PAPW
Methods for calculating PWDS
Resnik
Lin
Wang
PSB
SemFunSim
DATABASE
Disease Ontology
Disease-gene associations
HumanNet
GOA
lncRNA-disease association
miRNA-disease association

Fig. 1. System overview of DincRNA

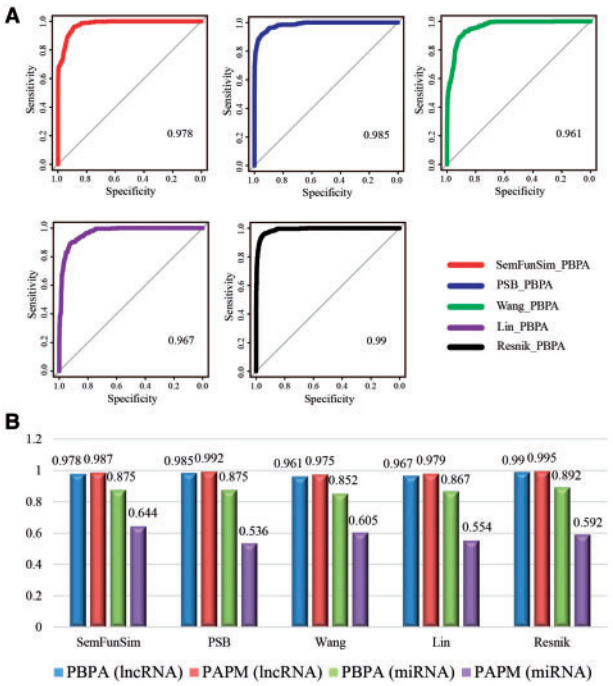

Fig. 2. Performance evaluation results. (**A**) The ROC curve of prioritizing lncRNA–disease pairs based on PBPA method. (**B**) The performance evaluation of prioritizing ncRNA–disease pairs based on different methods

Table 1. Accuracy of prioritizing lncRNA–disease pairs based on PBPA method for specified false positive rate

Method	Accuracy (FPR = 0.05)	Accuracy (FPR = 0.10)	Accuracy (FPR = 0.15)
Lin	0.831	0.903	0.939
PSB	0.913	0.958	0.973
Resnik	0.953	0.977	0.996
SemFunSim	0.825	0.941	0.975
Wang	0.729	0.896	0.934

ncRNA–disease pairs using RWR method, it is not easy for users to choose the most suitable ones. Hence, we evaluated the performance of each type of combinations in prioritizing lncRNA–disease and miRNA–disease pairs by leave-one-out cross validation.

Totally 525 priori lncRNA–disease associations, including 88 diseases with at least two lncRNAs, were used for this assessment. The receiver operating characteristic (ROC) curves of prioritizing lncRNA–disease pairs based on lncRNA FSN (LFSN) constructed by PBPA method with different algorithms for calculating SPWD were shown in Figure 2A. The areas under ROC curve (AUCs) based on the combinations between SemFunSim, PSB, Wang's, Lin's, Resnik's methods and PBPA method are 0.978, 0.985, 0.961, 0.967 and 0.990 respectively. The further evaluation for accuracy under lower false positive rate (FPR) is shown in Table 1. When the FPR was set to 15%, the accuracy based on the combinations between SemFunSim, PSB, Wang's, Lin's, Resnik's methods and PBPA method are 0.975, 0.973, 0.934, 0.939 and 0.996 respectively. All of the evaluation results based on different combinations are very well and close. The similar evaluation results occur based on the combinations between SemFunSim, PSB, Wang's, Lin's, Resnik's methods and PAPM method, which is shown in Figure 2B. These results show that PBPA and PAPM methods combined with other algorithms are very suitable for prioritizing lncRNA–disease pairs.

Analogously 5661 priori miRNA–disease associations of 261 diseases were utilized for leave-one-out cross validation. The AUCs based on PBPA and PAPM methods are shown in Figure 2B. Overall, the evaluation results of miRNAs are lower than those of lncRNAs using PBPA method. For example, The AUCs of lncRNA–disease and miRNA–disease pairs based on the combination of SemFunSim and PBPA methods are 0.978 and 0.875 respectively. This may be caused by the difference of the size of the sample in lncRNAs and miRNAs. Intuitively the assessment should be more comprehensive for larger size of sample. And corresponding the performance should be more accurate. Although the AUCs based on PBPA method decline slightly with the increasement of size of the sample, its performance is still very well due to the high AUC. In comparison with the prioritization of lncRNA–disease pairs, the AUCs of prioritizing miRNA–disease pairs using PAPW declines dramatically. The decline of the performance with the larger size of sample may be caused by its more noise. These results also show that the performance of PAPW is unstable. That may be the reason why PBPA is applied more widely than PAPW (Chen *et al.*, 2015; Sun *et al.*, 2014; Wang *et al.*, 2010).

The AUCs based on SemFunSim, PSB, Wang's, Lin's and Resnik's methods are close to each other according to Figure 2B. For example, the maximum and minimum AUCs of prioritizing miRNA–disease pairs based on PBPA are 0.892 and 0.852 respectively. These stable results validate that existing methods for calculating SPWD are mature.

Acknowledgements

We thank Dr. Jiajie Peng for discussion of the system design. We also thank Dr. Hengqianq Zhao for DincRNA testing.

Funding

This work was supported by the National Natural Science Foundation of China [Grant No. 61502125], Heilongjiang Postdoctoral Fund [LBH-Z15179] and China Postdoctoral Science Foundation [2016M590291].

Conflict of Interest: none declared.

References

Amberger,J. *et al.* (2011) A new face and new challenges for Online Mendelian Inheritance in Man (OMIM$^{®}$). *Hum. Mutat.*, **32**, 564–567.

Becker,K.G. *et al.* (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.

Camon,E. (2004) The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Res.*, **32**, D262–D266.

Chen,G. *et al.* (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983–D986.

Chen,X. *et al.* (2015) Constructing lncRNA functional similarity network based on lncRNA–disease associations and disease semantic similarity. *Sci. Rep.*, **5**, 11338.

Cheng,L. *et al.* (2014) SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PLoS One*, **9**, e99415.

Cheng,L. *et al*. (2013) SIDD: a semantically integrated database towards a global view of human disease. *PLoS One*, **8**, e75504.

Davis,A.P. *et al*. (2013) The comparative toxicogenomics database: update 2013. *Nucleic Acids Res*., **41**, D1104–D1114.

Kibbe,W.A. *et al*. (2015) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res*., **43**, D1071–D1078.

Lee,I. *et al*. (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res*., **21**, 1109–1121.

Li,Y. *et al*. (2014) HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res*., **42**, D1070–D1074.

Lin,D. (1998) An information-theoretic definition of similarity. *ICML*, p 296–304.

Mathur,S. and Dinakarpandian,D. (2012) Finding disease similarity based on implicit semantic similarity. *J. Biomed. Inf*., **45**, 363–371.

Mitchell,J.A. *et al*. (2003) Gene indexing: characterization and analysis of NLM's GeneRIFs. In: *AMIA … Annual Symposium proceedings/AMIA Symposium*. AMIA Symposium, pp. 460–464.

Ning,S. *et al*. (2016) Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res*., **44**, D980–D985.

Pesquita,C. *et al*. (2009) Semantic similarity in biomedical ontologies. *PLoS Comput. Biol*., **5**, e1000443.

Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy, arXiv preprint cmp-lg/9511007.

Sqalli,M.H. *et al*. (2012) UCloud: a simulated Hybrid Cloud for a university environment. In: *2012 IEEE 1st International Conference on Cloud Networking (CLOUDNET)*, IEEE, pp. 170–172.

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

Wang,D. *et al*. (2010) Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics*, **26**, 1644–1650.

Wang,J.Z. *et al*. (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.