BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 4 2015, pages 608–609
doi:10.1093/bioinformatics/btu684

Systems biology

Advance Access publication October 17, 2014

DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis

Guangchuang Yu[1,2,*], Li-Gen Wang[3], Guang-Rong Yan[2] and Qing-Yu He[2,*]

[1]State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The University of Hong Kong, Hong Kong SAR, [2]Key Laboratory of Functional Protein Research of Guangdong Higher Education Institutes, College of Life Science and Technology, Jinan University, Guangzhou 510632 and [3]Guangdong Information Center, Guangzhou 510031, China

Associate Editor: Igor Jurisica

ABSTRACT

Summary: Disease ontology (DO) annotates human genes in the context of disease. DO is important annotation in translating molecular findings from high-throughput data to clinical relevance. DOSE is an R package providing semantic similarity computations among DO terms and genes which allows biologists to explore the similarities of diseases and of gene functions in disease perspective. Enrichment analyses including hypergeometric model and gene set enrichment analysis are also implemented to support discovering disease associations of high-throughput biological data. This allows biologists to verify disease relevance in a biological experiment and identify unexpected disease associations. Comparison among gene clusters is also supported.

Availability and implementation: DOSE is released under Artistic-2.0 License. The source code and documents are freely available through Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/DOSE.html).

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

Contact: gcyu@connect.hku.hk or tqyhe@jnu.edu.cn

Received on May 7, 2014; revised on October 9, 2014; accepted on October 14, 2014

1 INTRODUCTION

Characterizing disease-disease relationships and mining gene-disease associations provides insights in analyzing high-throughput data to elucidate molecular mechanisms of complex diseases. Understanding similarities among diseases and among genes in disease context helps in early diagnosis, drug repurposing, and new drug development. Investigating gene-disease associations with gene lists obtained by high-throughput experiments helps exploring biological questions in disease context and discovering unanticipated functions.

Disease ontology (DO) provides a consistent description of genes in disease perspectives. To provide researchers with more accessible of disease knowledge, the DO database (Schriml *et al.*, 2012) supplies a web browser for users to explore DO vocabularies while disease and gene annotations database (Peng *et al.*, 2013) supplies a web interface for mapping genes and diseases. DO is organized as a directed acyclic graph, laying the foundation for computation of disease knowledge using semantic similarity algorithms. There are many generic quality tools for computation of semantic measures including SML, SimPack, SemMF, OWLSim and Similarity Library (http://goo.gl/3xCuJ6). These generic libraries can be employed to analyze DO semantic similarities. DOSim (Li *et al.*, 2011) was designed specific for DO, but the authors fail to maintain the package. Functional DO (FunDO) (Osborne *et al.*, 2009) implemented hypergeometric test to assess significant of DO associations with a gene list. However, FunDO doesn't allow users to customize the background set of genes and thus may introduce biases in the results.

To address the shortcoming of lack of R/Bioconductor package that designed for computation of semantic and enrichment analyses based on DO, we present DOSE, that allows measuring semantic similarity among DO terms and genes using several information-content and graph-structure based algorithms. For evaluating functional associations with gene lists of high-throughput genomic and proteomic studies, DOSE supports hypergeometric test and gene set enrichment analysis (GSEA), which incorporate expression level measurements to extract disease relevance of biological experiments. More importantly, DOSE provides several DO-specific visualization functions to produce highly customizable, publication-quality figures of similarity and enrichment analyses that are not available elsewhere. With these visualization tools, the results obtained by DOSE are more interpretable.

2 IMPLEMENTATION

DOSE provides *doSim* function to compute semantic similarity among DO terms. DOSE implemented four information content based algorithms proposed by Resnik (Resnik, 1999), Lin (Lin, 1998), Jiang and Conrath (Jiang and Conrath, 1997) and Schlicker (Schlicker *et al.*, 2006), respectively, and one graph based algorithm proposed by Wang (Wang *et al.*, 2007) to measure the semantic similarity among DO terms.

These algorithms were extended from in-house developed R package GOSemSim (Yu *et al.*, 2010). By mapping genes to DO annotations, *geneSim* function measures the semantic similarities among genes based on their annotated DO terms. Four combine strategies were implemented in DOSE for aggregating semantic similarity scores of multiple DO terms associated with genes,

*To whom correspondence should be addressed.

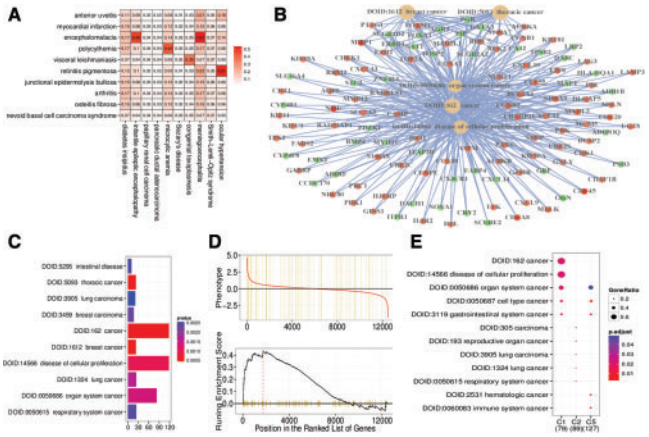

Fig. 1. Five graphs produced by DOSE. (A) Heatmap of semantic similarity matrix; (B) Disease and gene association network; (C) Barplot of enrichment result; (D) Plot of running sum of enrichment scores and its association with phenotype and (E) Comparison of disease associations among different gene sets

including *max* which calculates the maximum similarity score over all pairs of DO terms, *avg* which uses the average of similarity scores over all pairs of DO terms, *rcmax* which measures the maximum of RowScore and ColumnScore, where RowScore (ColumnScore) is the average of maximum similarity on each row (column) and best-match average which measures the average of maximum similarity scores on each row and column. The semantic similarity results obtained from *doSim* and *geneSim* can be visualized by *simplot* function.

DOSE provides hypergeometric model to assess disease associations of differential express genes. The *enrichDO* function allows users to select an appropriate background of genes as baseline. The *gseAnalyzer* function supports GSEA to evaluate disease relevance of high-throughput data. These approaches can be used to verify whether the genes implicated in biological experiment are disease associated and to identify unexpected disease associations. Multiple comparison corrections including Bonferroni, Benjamini, False Discovery Rate and *q*-values are also incorporated. Disease associations among different gene clusters or gene lists from different conditions can be compared using in-house developed R package clusterProfiler (Yu *et al.*, 2012). Several visualization functions including *barplot* and *cnetplot* are implemented for visualizing significant disease associations and gene-disease association network respectively. Running sum of enrichment scores and its association with phenotype can be visualized using *gseaplot* function.

3 RESULTS AND CONCLUSION

DOSE was developed using the R statistical computing language and is released within Bioconductor project. It provides five algorithms for DO and gene semantic similarity measurements (Fig. 1A); hypergeometric test for identifying significant disease association of gene list (Fig. 1B and C); GSEA for interpreting genome wide expression profiles in disease context (Fig. 1D) and comparison of significant disease associations among different gene sets (Fig. 1E). R scripts to generate Figure 1 are presented in Supplemental File.

The DOSE package presented here makes use of semantic similarity approaches and enrichment analyses to facilitate users to investigate large gene sets. Moreover, DOSE provides users the abilities to visualize semantic similarities, significant gene-disease associations, and gene set comparison.

Funding: This work was supported by the National Natural Science Foundation of China (21271086 to Q.-Y.H.) and Fundamental Research Funds for the Central Universities (21613414 to G.Y.).

REFERENCES

Jiang,J. and Conrath,D. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: *International Conference Research on Computational Linguistics (ROCLING X), Taiwan*. p. 9008.

Li,J. *et al.* (2011) DOSim: an R package for similarity between diseases based on disease ontology. *BMC Bioinformatics*, **12**, 266.

Lin,D. (1998) An information-theoretic definition of similarity. In: *Proceedings of the Fifteenth International Conference on Machine Learning*. pp. 296–304.

Osborne,J. *et al.* (2009) Annotating the human genome with disease ontology. *BMC Genomics*, **10**, S6.

Peng,K. *et al.* (2013) The disease and gene annotations (DGA): an annotation resource for human disease. *Nucleic Acids Res.*, **41**, D553–D560.

Resnik,P. (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artifical Intell. Res.*, **11**, 95–130.

Schlicker,A. *et al.* (2006) A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, **7**, 302.

Schriml,L.M. *et al.* (2012) Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.

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

Yu,G. *et al.* (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**, 976–978.

Yu,G. *et al.* (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS J. Integr. Biol.*, **16**, 284–287.