BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 20 2010, pages 2643–2644
doi:10.1093/bioinformatics/btq477

Data and text mining

Advance Access publication August 27, 2010

Ontology- and graph-based similarity assessment in biological networks

Haiying Wang[1], Huiru Zheng[1] and Francisco Azuaje[2,*]

[1]Computer Science Research Institute, School of Computing and Mathematics, University of Ulster and [2]Laboratory of Cardiovascular Research, Public Research Centre for Health (CRP-Santé), L-1150, Luxembourg

Associate Editor: John Quackenbush

ABSTRACT

Summary: A standard systems-based approach to biomarker and drug target discovery consists of placing putative biomarkers in the context of a network of biological interactions, followed by different 'guilt-by-association' analyses. The latter is typically done based on network structural features. Here, an alternative analysis approach in which the networks are analyzed on a 'semantic similarity' space is reported. Such information is extracted from ontology-based functional annotations. We present *SimTrek*, a Cytoscape plugin for ontology-based similarity assessment in biological networks.

Availability: http://rosalind.infj.ulst.ac.uk/SimTrek.html

Contact: francisco.azuaje@crp-sante.lu

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

Received on May 19, 2010; revised on August 4, 2010; accepted on August 13, 2010

1 INTRODUCTION

A standard systems-based approach to biomarker and drug target discovery consists of placing putative or known biomarkers in the context of a network of biological interactions, followed by different 'guilt-by-association' analyses (Merico *et al.*, 2009). Putative biomarkers may be derived, for example, from standard differential expression analysis, including those from large-scale gene expression and proteomics experiments. Networks may encode gene–gene, gene–protein or protein–protein interactions. They are typically inferred from published interactions or from computational prediction models. Guilt-by-association approaches include graph theoretic techniques, such as network clustering algorithms. Thus, standard network-driven approaches tend to be based on the analysis of biological network structures.

Here, we report an alternative analysis approach in which the user-defined networks are mapped onto a 'semantic similarity' space. In such a space, between-protein relationships are represented by estimates of ontology-based functional similarity. Semantic similarity is computed by using statistical information encoded in functional databases annotated to the Gene Ontology (GO). We and others have previously investigated semantic similarity assessment in the context of functional genomics and network-based biology (Azuaje *et al.*, 2006; Pesquita *et al.*, 2009).

*To whom correspondence should be addressed.

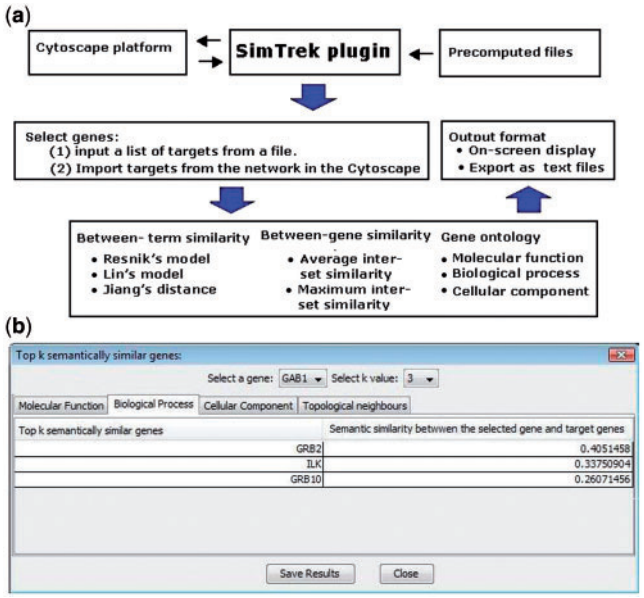

Fig. 1. Overview of *SimTrek*. (**a**) Software architecture. (**b**) Output panel screenshot.

2 THE *SimTrek* SYSTEM

We implemented a semantic similarity assessment system, *SimTrek*, which can be integrated with different types of biological networks under Cytoscape (Shannon *et al.*, 2003). Figure 1a depicts the main components and processes of *SimTrek*.

SimTrek allows the interrogation of networks based on user-defined queries and the retrieval of their *k*-nearest neighbors in the network and semantic similarity spaces. The user-defined inputs are: the network to be analyzed and a list of gene/protein queries. The network can represent different types of functional relationships, e.g. regulatory or signaling interaction networks. The queries can represent putative biomarkers or targets obtained from previous experimental or computational analyses, such as differentially expressed genes. The semantic similarity between gene products is calculated using different information theoretic techniques (Pesquita *et al.*, 2009). Moreover, different numbers, *k*, of 'semantically similar' neighbors can be retrieved for further analyses. Figure 1 shows a screenshot of *SimTrek* under Cytoscape.

Based on the assumption that the more information two terms share in common, the more similar they are, three semantic similarity measures (Jiang and Conrath, 1997; Lin, 1998; Resnik, 1995) are

implemented to measure between-term similarity within each of the GO hierarchies. In addition, *SimTrek* offers two approaches to assessing similarity between gene products based on the aggregation of between-term similarities. The first method is defined as the *average inter-set similarity* between the sets of GO terms associated with two gene products. The second approach is the *highest between-term similarity* approach, which selectively aggregates maximum inter-set similarity values. A more detailed description of these techniques can be found in the Supplementary Section.

Semantic similarity assessment results are displayed in different panels (Figure 1). The first includes two combo boxes which allow users to select the target genes and k-values. The second is a tabbed pane that shows the results from the GO hierarchies and the topology-based neighborhood analysis methods. The corresponding tables under the first three tabs show the top k genes and their similarity values. In the case where the number of gene pairs having semantic similarity values is lesser than k, only those gene pairs with values are displayed. The last table shows the immediate network (graph-based) neighbors of the selected queries. To minimize installation effort and overcome potential technical constraints posed by firewalls, the current system provides compact pre-computed files required to calculate semantic similarity. Regular updates of these files may be obtained from our website.

3 APPLICATION EXAMPLE

Semantic similarity can provide information that cannot be extracted from the input network topology alone. Unlike traditional graph-based methods, *SimTrek* incorporates annotated knowledge that can be exploited to discover systems-level, functional relationships.

Figure 1 illustrates the application of *SimTrek* using a network that regulates the activity of p53 in humans (Abdi *et al.*, 2008; Ma'ayan *et al.*, 2005;). After defining a target query, e.g. GAB1 (Figure 1), the system enables the user to implement different types of semantic similarity estimation techniques. This can be done according to species, GO hierarchy and evidence criteria. The retrieval of the most semantically-similar (k-nearest neighbors) gene products can be visualized and saved in a user-specified file. The user can either interactively select queries on the network or provide lists of targets previously stored in a text file. *SimTrek* also automatically detects network topology-based nearest neighbors.

Different analyses of the k-most semantically similar genes/proteins to the query sets can be implemented, for different values of k. The retrieved neighbors can represent the inputs to subsequent tasks for exploring the predictive potential of these genes/proteins. For example, they can represent functionally specialized clusters of genes (modules) or pathways connecting phenotype-relevant processes or putative biomarkers. The example shown in Figure 1 was obtained with $k=3$, Lin's semantic similarity, and concentrated on human GO annotations to the Biological Process hierarchy, excluding those with IEA evidence code. Note that the resulting set of semantic nearest-neighbors (GRB2, ILK, GRB10) includes proteins that were not included in the nearest neighborhood defined by the input network topology alone (SHP2, GRB2, PI3K). The ontology-based predictions are consistent with results from other studies. For example, GAB1, a known cell death mediator, encodes a GRB2-associated-binding protein (Holgado-Madruga *et al.*, 1996). A direct functional association between GRB10 and GAB1 through mitogenic signaling has been reported (Deng *et al.*, 2008). Currently, there are no GO annotation terms from the Biological Process hierarchy assigned to genes SHP2 and PI3K. Other results may suggest the existence of novel protein–protein interactions or key functional non-physical associations.

Unlike other tools, such as (standalone application) DynGO (http://gauss.dbb.georgetown.edu/liblab/dyngo.html) and the R package GOstats (http://bioconductor.org/packages/2.3/bioc/html/GOstats.html), *SimTrek* was implemented as a Cytoscape plugin (Shannon *et al.*, 2003) to facilitate its integration within a widely used software framework, which also offers powerful visualization functionality. *SimTrek* also goes beyond the calculation of semantic similarity to implement a network-centric analytical approach. It can be used to detect potentially novel associations, which may not be explicitly represented in the input network under analysis. This allows users to move beyond the analysis of graph-derived clusters of gene products. Functionally relevant associations across the network or clusters are detected to aid in the interpretation of network analysis or to guide the prediction of potential novel targets or biomarkers. Furthermore, *SimTrek* contributes to the incorporation of prior biological knowledge into network-based analyses.

ACKNOWLEDGEMENT

We thank Jaine Blayney for supporting testing of SimTrek.

Conflict of Interest: none declared.

REFERENCES

Abdi,A. *et al.* (2008) Fault diagnosis engineering of digital circuits can identify vulnerable molecules in complex cellular pathways. *Sci. Signal.*, **1**, ra10.

Azuaje,F. *et al.* (2006) Predictive integration of gene ontology-driven similarity and functional interactions. In *Proceedings of the 6th IEEE International Conference on Data Mining, Hong Kong*, IEEE Computer Science, pp. 114–119.

Deng,Y. *et al.* (2008) Mitogenic roles of Gab1 and Grb10 as direct cellular partners in the regulation of MAP kinase signaling. *J. Cell Biochem.*, **105**, 1172–1182.

Jiang,J.J. and Conrath,D.W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*, Academia Sinica, Taiwan, pp. 19–33.

Holgado-Madruga,M. *et al.* (1996). A Grb2-associated docking protein in EGF- and insulin-receptor signaling. *Nature*, **379**, 560–564.

Lin,D. (1998) An information-theoretic definition of similarity. In *Proceedings of 15th International Conference on Machine Learning*, Morgan Kaufmann, Madison, Wisconsin, pp. 296–304.

Ma'ayan,A. *et al.* (2005) Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science*, **309**, 1078–1083.

Merico,D. *et al.* (2009) How to visually interpret biological data using networks. *Nat. Biotechnol.*, **27**, 921–924.

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. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, Montreal, pp. 448–453.

Shannon,P. *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.