G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery

Zhidian Du1, Lin Li1, Chin-Fu Chen2, Philip S. Yu3 and James Z. Wang1,∗

1School of Computing, 2Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634 and 3Department of Computer Science, University of Illinois at Chicago, 851 S. Morgan St., Rm 1138 SEO, Chicago, IL 60607, USA

Received February 21, 2009; Revised April 28, 2009; Accepted May 16, 2009

ABSTRACT
We have developed a set of online tools for measuring the semantic similarities of Gene Ontology (GO) terms and the functional similarities of gene products, and for further discovering biomedical knowledge from the GO database. The tools have been used for about 6.9 million times by 417 institutions from 43 countries since October 2006. The online tools are available at: http://bioinformatics.clemson.edu/G-SESAME.

INTRODUCTION
We provide and maintain a set of online tools to measure the semantic similarities of Gene Ontology (GO) terms and the functional similarities of gene products, and to discover biomedical knowledge through GO database. These tools are developed based on methods and algorithms proposed in our G-SESAME article (1) using MySQL 5.0.45 and PHP 5.1.6 and hosted by an Apache Web server (version 2.2.3) running on a Linux operating system (CentOS 5).

MOTIVATIONS
Although the GO project (2) has provided us the organized GO terms, one challenge is to accurately measure the semantic similarities of two GO terms, from which we can determine the functional similarities of genes. Some methods, such as Resnik’s, Jiang’s, Lin’s (3–5) and their variants (6,7), have been used to measure the semantic similarities of GO terms. However, their accuracies for measuring GO terms were the main issues that we addressed in our previous article (1) as these methods were originally proposed for other specific taxonomies and were adapted to measure the semantic similarities of GO terms (8).

*To whom correspondence should be addressed. Tel: +1 864 506 0283; Fax: +1 858 658 2100; Email: jzwang@cs.clemson.edu

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
path rule, when calculating a GO term’s semantic value, it must include the biological meanings of all its ancestor terms. Therefore, when measuring the semantic similarities of GO terms, we must consider not only the number of the common ancestor terms but also the locations of these ancestor terms related to the two specific terms in the GO graph.

Semantic similarities of two GO terms

According to our above analysis, we proposed a method to measure the semantic similarities of two GO terms in (1). We first decode the semantics of a GO term into a numeric value. Since the semantics (biological meanings) of a GO term are determined by its location in the entire GO hierarchy and its semantic relations with all of its ancestor terms, we use the DAG which is a subgraph of an ontology starting from this GO term and ending at any of the root term (biological process, cellular component, or molecular function) to represent this term. The detailed method has been introduced in Wang’s article (1).

Our semantic similarities measurement algorithm has two advantages. First, it relies only on the relationships of the GO terms within a specific ontology (biological process, cellular component, or molecular function) to determine their semantic similarities. Therefore, it provides a consistent measurement for the semantic similarities between two GO terms, independent of the annotation statistics. Second, our algorithm is designed to decode the human perception of the semantic relationships between child and parent terms. Thus, the semantic similarities of GO terms obtained by our algorithm can reflect the closeness of their biological meanings in human perspectives.

In our previous study (1), we used many cases to evaluate the ‘is-a’ and ‘part-of’ values. We found that the gene clustering results by choosing 0.8 as the contribution factor for ‘is-a’ relations would be most consistent with the manual classification. In (8), the authors compared the covariance of the semantic similarities between GO terms in all three ontologies with the results of the BLAST. The covariance is not significantly impacted by using ‘part-of’ relations or not. We also know that, compared with the number of ‘part-of’ links, the number of ‘is-a’ links dominates the whole corpus in the GO database. Therefore, the semantic value of ‘is-a’ links should be bigger than that of ‘part-of’. Based on the evaluation of different values, we suggest to use the value 0.8 for ‘is-a’ and 0.6 for ‘part-of’.

Semantic similarities of two genes

Given the semantic similarities of two GO terms, a method to compare the similarities of two genes which are annotated by two sets of GO terms was also proposed by us in (1). When we compare two sets of GO terms, we let each pair of GO terms with higher semantic similarities dominate the relationship while neglecting other insignificant values. All the assumptions are based on the monotonic properties of the information content (IC) of a term with those of its ancestors and descendants. Each term in one GO term set should pair up its counterpart in the other set with close ‘distance’. The distance can be the one defined in Jiang’s method; however, it is also feasible to use other definitions which should be in compliance with their own monotonic properties of IC. For example, in our previous article (1), we give our own distance definition.

Evaluation of our method and Resnik’s

We use the gene annotation and classification information in pathways at the Saccharomyces genome database (SGD) as the reference for our evaluation. There are 152 biological pathways in the SGD database. Most of these pathways contain at least three genes annotated by both GO molecular function terms and EC numbers (10). These genes are also manually clustered by their molecular functions. We compared the results of our method with Resnik’s method on 111 pathways listed at SGD, which are well accepted pathways and the related genes are completely annotated in the GO database. From the pairs of clustering patterns, it is easy to tell that our method generated more similar pathways to those in SGD. The detailed comparisons in all cases between our method and Resnik’s can be found in the supplemental materials of (1). The summary of evaluation results are shown in Table 1.

### Table 1. Comparison results of our method and Resnik’s based on SGD pathways

| Cases                          | Numbers |
|-------------------------------|---------|
| Total of pathways in the SGD website | 152     |
| Number of pathways having at least three | 111     |
| Number of pathways used in evaluation | 111     |
| Evaluations showing our method is better | 66      |
| Evaluations showing that both methods are equal | 45      |
| Evaluations showing that our method is worse | 0       |

ONLINE TOOLS

The following sections introduce the set of our online tools, which include the tool for measuring the semantic similarities of two and multiple GO terms, the tool for the semantic comparison of two gene products from two different species, and multiple gene comparison and clustering tools. All the interfaces of our tools are consistent and user friendly. Users need to specify the values of ‘is-a’ and ‘part-of’ relations for all the tools. Our tools can run in an interactive mode or in a batch mode.

Tool for measuring the semantic similarities of two GO terms

This tool is the simplest tool in our family of tools, with which users need to provide two GO terms and the ‘is-a’ and ‘part-of’ values. After receiving two GO terms online and the related ‘is-a’ and ‘part-of’ values, the program first searches the GO database to check the existence of the two GO terms. If both of the two GO terms exist, it retrieves the ancestors of these two GO terms. Figure 1 shows the output of comparing semantic similarities of two GO terms, GO: 0005739 and GO: 0005777. The DAGs of
these two GO terms which are shown in Figure 2 are displayed as a hyperlink in the output webpage. In the DAG, solid arrows represent the ‘is-a’ relation and dashed arrows represent the ‘part-of’ relation.

**Tool for measuring the semantic similarities of two GO term sets**

Based on the two GO term similarities comparison function, it is easy to implement the tool to compare the semantic similarities of two GO term sets. This tool allows users to upload two files which contain multiple GO terms in each file. Figure 3 shows the output of the comparison of two GO term sets.

**Tool for measuring the semantic similarities of two genes from two species**

Based on the semantic correlation of GO terms used to annotate genes and gene products, we implemented the gene functional similarities comparison tool which can be used to compare the functional similarities of two genes. One unique feature of this tool is that it can compare two different genes from two different species. Users can specify one of the ontologies, two different species, and other criteria, such as data bases and evidence code. If multiple genes with the same symbols exist in the database, users have to choose one gene among these different species.

Figure 4 illustrates the output of measuring the semantic similarity of two genes, adh1 and adh4, from two different species. Not only the similarity value of these two genes is displayed at the top of the output webpage, but also the detail information of GO terms which annotate these two genes and other information, such as data sources and evidence codes, are displayed in two tables at the center of the page. The similarity table related with each pair of GO terms is also displayed at the bottom of the web page.

**Tool for multiple genes comparison and gene clustering**

Based on the gene functional similarity measurement, we implemented the multiple gene comparison and clustering tool for gene functionality analyses and knowledge discovery using our method, Jiang’s, Lin’s and Resnik’s methods.

This tool needs to utilize a file which contains gene symbols with one symbol in each line. Users can specify different ontologies, species and other criteria. The interface of the tool is shown in Figure 5.

After users submit the query online, the tool not only outputs the similarities of these genes, but also displays the clustering results of these genes. It is easy to read the
different patterns from these clustering results. Figures 6 and 7 show the similarity values and clustering results of genes in tryptophan-degradation pathway obtained by the tool based on Resnik’s method.

Our tools can be used not only in an interactive mode, but they are also designed and can be used in a batch mode for massive data analysis and data mining via the public Application Programming Interfaces (APIs). All our programs are listed at http://bioinformatics.clemson.edu/G-SESAME/Program/ and all of them are well organized. We provide the public APIs, so it is very easy to let the third parties’ programs call the public APIs via the standard Hyper Text Transfer Protocol (HTTP).

Figure 5. Interface of the gene clustering tool.

Figure 4. Output of measuring the semantic similarity of two genes, adh1 and adh4, from two different species.

Figure 6. Semantic similarity table of genes in tryptophan degradation pathway obtained by Resnik’s method.

Figure 7. Clustering results of genes in tryptophan degradation pathway based on Resnik’s method.

Users can use their preferred languages to call our public APIs when they need to run our tools in a batch mode. Their programs need to provide the necessary query information which is needed by our tools. After sending the query information to our tools via the HTTP protocol, their programs will receive the results sent back by our tools and further perform their analysis. The details of how to use our tools in both modes have been described on our web pages.

Our tools have been optimized for efficiency and robustness. They have been used extensively and intensively by many institutions.

CONCLUSION AND FUTURE STUDIES

In this article, after introducing the background of the GO project, we introduce our solution for how to compare two different GO terms and genes. We then introduce a set of online tools based on our methods proposed in (1). These tools include the GO term comparison tool, the GO term set comparison tool, the comparison tool of two genes from two different species, the multiple gene comparison
tool, and the clustering tool. We not only provide these online tools using our method, but also provide the clustering tools using Resnik’s, Jiang’s and Lin’s methods to satisfy different user requirements.

All these tools can be used interactively or in a batch mode. They have been maintained and optimized by our team. As shown by our web log records, these tools have been widely used by researchers in biomedical research community.

Currently, we find that most of our knowledge discovery tools are related with one species scope. However, we believe that we need to do more research for multiple gene products among two or more species. Besides, due to the intensive usage of our tools, we may consider the feasibility to parallelize the implementation.

FUNDING

Funding for open access charge: National Institutes of Health grant 1R15CA131808-01.

Conflict of interest statement. None declared.

REFERENCES

1. Wang,J.Z., Du,Z., Payattakool,R., Yu,P.S. and Chen,C.-F. (2007) A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23, 1274–1281.

2. Consortium,G.O. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25, 25–29.

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

4. Lin,D. (1998) An information-theoretic definition of similarity, semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA, pp. 296–304.

5. Jiang,J.J. and Conrath,D.W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research on Computational Linguistics (ROCLING X)*. Taiwan.

6. Couto,F.M., Silva,M.J. and Coutinho,P. (2003) Implementation of a functional semantic similarity measure between gene-products. In *DI/FCUL TR 03-29*. Department of Informatics, University of Lisbon.

7. Lee,S.G., Hur,J.U. and Kim,Y.S. (2004) A graph-theoretic modeling on go space for biological interpretation of gene clusters. *Bioinformatics*, 20, 381–388.

8. Lord,P.W., Stevens,R.D., Brass,A. and Goble,C.A. (2003) Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19, 1275–1283.

9. Sevilla,J.L., Segura,V., Podhorski,A., Guruceaga,E., Mato,J.M., Martinez-Cruz,L.A., Corrales,F.J. and Rubio,A. (2005) Correlation between gene expression and go semantic similarity. *IEEE/ACM Transact. Comput. Biol. Bioinformatics*, 2, 330–338.

10. Ball,C.A., Dolinski,K., Dwight,S.S., Harris,M.A., Issel-Tarver,L., Kasarskis,A., Scafe,C.R., Sherlock,G., Binkley,G., Jin,H. et al. (2000) Integrating functional genomic information into the sac-charomyces genome database. *Nucleic Acids Res.*, 28, 77–80.