BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 10 2007, pages 1274–1281
doi:10.1093/bioinformatics/btm087

Data and text mining

A new method to measure the semantic similarity of GO terms

James Z. Wang[1,*], Zhidian Du[1], Rapeeporn Payattakool[1], Philip S. Yu[2] and Chin-Fu Chen[3]

[1]School of Computing, Clemson University, Clemson, SC 29634, USA, [2]IBM T. J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532, USA and [3]Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA

Received on October 6, 2006; revised and accepted on March 1, 2007

Advance Access publication March 7, 2007

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Although controlled biochemical or biological vocabularies, such as Gene Ontology (GO) (http://www.geneontology.org), address the need for consistent descriptions of genes in different data sources, there is still no effective method to determine the functional similarities of genes based on gene annotation information from heterogeneous data sources.

Results: To address this critical need, we proposed a novel method to encode a GO term's semantics (biological meanings) into a numeric value by aggregating the semantic contributions of their ancestor terms (including this specific term) in the GO graph and, in turn, designed an algorithm to measure the semantic similarity of GO terms. Based on the semantic similarities of GO terms used for gene annotation, we designed a new algorithm to measure the functional similarity of genes. The results of using our algorithm to measure the functional similarities of genes in pathways retrieved from the saccharomyces genome database (SGD), and the outcomes of clustering these genes based on the similarity values obtained by our algorithm are shown to be consistent with human perspectives. Furthermore, we developed a set of online tools for gene similarity measurement and knowledge discovery.

Availability: The online tools are available at: http://bioinformatics.clemson.edu/G-SESAME

Contact: jzwang@cs.clemson.edu

Supplementary information: http://bioinformatics.clemson.edu/Publication/Supplement/gsp.htm

1 INTRODUCTION

Although controlled biochemical or biological vocabularies, such as Gene Ontology (GO) (http://www.geneontology.org), address the need for consistent descriptions of genes in different data sources, automatically measuring the functional similarities of genes based on these annotation data remains a challenge. Currently, researchers use online information retrieval tools, such as AmiGO (http://www.godatabase.org) and QuickGO (http://www.ebi.ac.uk/ego/), to collect gene annotation data from various databases and manually discover the correlations or similarities of gene products by visually examining their biological functions. However, because the manual discovery of this important knowledge requires significant time and effort, there is a critical need to build automated tools to measure and visualize the functional similarities of gene products based on existing annotation information from heterogeneous data sources.

In past years, some online tools such as eGOn (Langaas *et al*., 2005), FuSSiMeG (*http://xldb.fc.ul.pt/rebil/tools/ssm/*) and DAVID (*http://david.abcc.ncifcrf.gov/*) were developed to measure the functional similarity of genes. However, their similarity measurement methods have drawbacks. Some approaches (Langaas *et al*., 2005; *http://david.abcc.ncifcrf.gov/*) measure gene functional similarities based on the probability of the appearance of GO terms or the kappa statistics of similar annotation terms correlated with different genes, and ignore the semantic relations ('is-a' and 'part-of') among these terms in the GO graph. Although other methods (Jiang and Conrath, 1997; Lin, 1998; Resnik, 1999) were proposed to measure the semantic similarity of terms in a specific taxonomy, these methods were originally developed for the natural language taxonomies and it is unclear whether they are suitable for measuring the semantic similarity of GO terms.

These existing methods (Jiang and Conrath, 1997; Lin, 1998; Resnik, 1999) and their variants (Coute *et al*., 2003; Kriventseva *et al*., 2001; Lee *et al*., 2004) determine the similarity of two GO terms based on their distances to the closest common ancestor term and/or the annotation statistics of their common ancestor terms. Although recent studies (Guo *et al*., 2006; Sevilla *et al*., 2005; Wang *et al*., 2004) evaluating these methods showed that Resnik's method is better than other methods in terms of the correlation with gene sequence similarities and gene expression profiles, none of these evaluation studies provided direct evidences on how well these methods measure the functional similarity of genes. Instead, they pointed out some drawbacks in these existing similarity measurement methods that hinder their ability of determining the functional similarity of genes.

A drawback of Resnik's method is that it ignores the information contained in the structure of the ontology by only concentrating on the information content of a term derived from the corpus statistics. However, the specificity of a GO term is usually determined by its location in the GO graph and a GO term's semantics (biological meanings) are inherited from

*To whom correspondence should be addressed.

all its ancestor terms. Therefore, using the information content as the sole determination factor for the semantic similarity of GO terms is inappropriate. On the other hand, based on human perspectives, if two terms sharing the same parent are near the root of the ontology (terms are more general), they should have larger semantic difference than two terms having the same parent and being far away from the root of the ontology because the later are more specific terms. However, using Jiang's or Lin's method, as pointed out by Sevilla *et al.*, if two gene products are well annotated near the root of the ontology (shallow annotation), their semantic similarity will always be measured at very high (close to 1) and their semantic distance will always be computed close to nil, thus providing a misleading result. The effect of shallow annotation is a serious drawback of both Jiang and Lin's methods.

Besides their individual drawbacks, a common problem with these methods is that they depend on the gene annotation statistics to measure the semantic similarity of GO terms. Hence, people may get different semantic similarity values for the same two GO terms if they use different gene annotation data. However, the purpose of having a set of controlled vocabularies (an ontology) is that the biological terms in an ontology should have a fixed semantics (biological or biochemical meanings) when it is used to annotate genes. Therefore, it is desirable to determine the semantic similarity of GO terms based only on the structure and annotation specification of GO ontologies. Unfortunately, most existing ontology-structure-based methods (Langaas *et al.*, 2005; Wang *et al.*, 2004) also have their drawbacks in that they determine the semantic similarity of two GO terms either based on their distances to the closest common ancestor term or based on the number of their common ancestor terms.

First, the distances to the closest common ancestor term cannot accurately represent the semantic difference of two GO terms. As discussed previously, if two terms sharing the same parent are near the root of the ontology, they should have larger semantic difference than two terms having the same parent and being far away from the root of the ontology. In addition, one GO term may have multiple parent terms with different semantic relations. A GO term's semantics (biological meanings) must be the aggregate semantic contributions from all ancestor terms (including this specific term). Second, measuring the semantic similarity of two GO terms based only on the number of common ancestor terms cannot discern the semantic contributions of the ancestor terms to these two specific terms. In fact, a common ancestor of two GO terms may have different contributions to the semantics of these specific terms because their distances to this common ancestor in the GO graph may differ and the semantic relations (edges in the GO graph) leading to this common ancestor may vary as well. Based on human perspectives, an ancestor term farther from a descendant term in the GO graph contributes less to the semantics of the descendant term while an ancestor term closer to a descendant term in the GO graph contributes more to the semantics of this descendant term.

As the GO website states, 'The ontologies are structured as directed acyclic graphs, which are similar to hierarchies but differ in that a child, or more specialized, term can have many parent, or less specialized, terms. For example, the biological process term *hexose biosynthesis* has two parents, *hexose metabolism* and *monosaccharide biosynthesis*. This is because *biosynthesis* is a subtype of *metabolism*, and a *hexose* is a type of *monosaccharide*. When any gene involved in *hexose biosynthesis* is annotated to this term, it is automatically annotated to both *hexose metabolism* and *monosaccharide biosynthesis*, because every GO term must obey the true path rule: if the child term describes the gene product, then all its parent terms must also apply to that gene product'. This statement clearly indicates that a GO term's semantics (biological meanings) must include the biological meanings of all its ancestor terms. Therefore, measuring the semantic similarity of GO terms 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.

2 GENE FUNCTIONAL SIMILARITY

To address the drawbacks in existing methods, we propose a new method to measure the semantic similarity of GO terms and, in turn, devise an algorithm to determine the functional similarity of genes based on the semantic similarities of GO terms used to annotate these genes.

2.1 Semantic values of GO terms

GO is a large collaborative public database that provides a set of controlled vocabularies (biological or biochemical terms) describing gene products based upon their functions in the cell. Three ontologies, *biological process*, *cellular component* and *molecular function*, are defined in this database. They are presented as directed acyclic graphs (DAGs) in which the terms form nodes and the two kinds of semantic relations ('is-a' and 'part-of') form edges. 'is-a' is a simple class-subclass relation, where A *is-a* B means that A is a subclass of B. 'part-of' is a partial ownership relation; C *part-of* D means that whenever C is present, it is always a part of D, but C need not always be present.

To measure the semantic similarity of GO terms, we first encode the semantics of a GO term into a numeric format. Since the semantics (biological meanings) of a GO term are determined by its location in the entire GO graph and its semantic relations with all of its ancestor terms, we use the DAG (a subgraph of an ontology) starting from the specific GO term and ending at any of the root term (biological process, cellular component or molecular function) to represent this term. For example, Figure 1 depicts the DAG for GO term *Intracellular Membrane-bound Organelle*: 0043231. This DAG has seven GO terms connected by eight relations. A dotted arrow represents the 'part-of' relation and a solid arrow shows the 'is-a' relation. For instance, GO term *Intracellular Organelle*: 0043229 is a subclass of GO term *Organelle*: 0043226 and also a part of GO term *Intracellular*: 0005622.

We note that the DAGs and gene annotation information used in this article were obtained from the GO database in May 2006. Due to the daily evolution of the GO database, a GO term's DAG may change due to the addition of new terms or removal of obsolete terms, and gene annotation data may

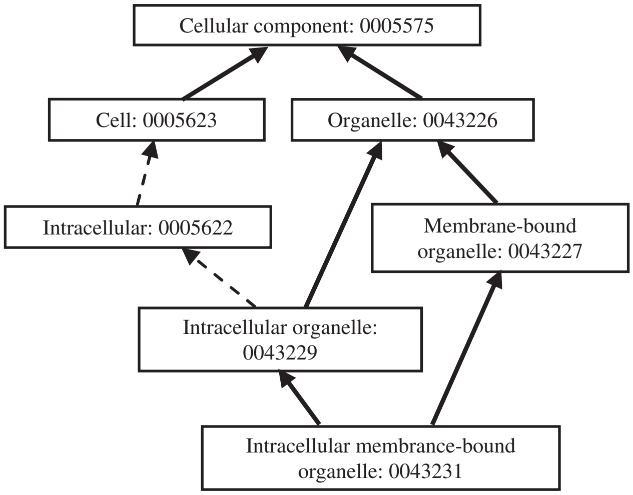

Cellular component: 0005575
Cell: 0005623
Organelle: 0043226
Intracellular: 0005622
Membrane-bound organelle: 0043227
Intracellular organelle: 0043229
Intracellular membrance-bound organelle: 0043231

Fig. 1. DAG for GO term *Intracellular Membrane-bound Organelle:* 0043231.

change as new information about genes is found and added into the database.

Formally, a GO term A can be represented as $DAG_A = (A, T_A, E_A)$ where T_A is the set of GO terms in DAG_A, including term A and all of its ancestor terms in the GO graph, and E_A is the set of edges (semantic relations) connecting the GO terms in DAG_A. To encode the semantics of a GO term in a measurable format to enable a quantitative comparison of two term's semantics, we define the semantic value of term A as the aggregate contribution of all terms in DAG_A to the semantics of term A. Terms closer to term A in DAG_A contribute more to its semantics, while terms farther from term A in DAG_A contribute less as they are more general terms. Therefore, we define the contribution of a GO term t to the semantics of GO term A as the S-value of GO term t related to term A. For any term t in $DAG_A = (A, T_A, E_A)$, its S-value related to term A, $S_A(t)$, is defined as:

$$\begin{cases} S_A(A) = 1 \\ S_A(t) = \max\{w_e * S_A(t') | t' \in \text{childrenof}(t)\} \text{ if } t \neq A \end{cases} \tag{1}$$

where w_e is the semantic contribution factor for edge $e \in E_A$ linking term t with its child term t'. In DAG_A, GO term A is the most specific term and we define its contribution to its own semantics as one. Other terms in DAG_A are more general and, hence, contribute less to the semantics of GO term A. Therefore, we have $0 < w_e < 1$. After obtaining the S-values for all terms in DAG_A, we calculate the semantic value of GO term A, $SV(A)$, as:

$$SV(A) = \sum_{t \in T_A} S_A(t) \tag{2}$$

For instance, if we assume that the semantic contribution factors for 'is-a' and 'part-of' relations are 0.8 and 0.6 respectively, we use Equation (1) to calculate the S-values of the GO terms in the DAG representing GO term *Intracellular Membrane-bound Organelle*: 0043231 and list the results in Table 1. We note that the semantic value of a GO term differs from its S-value. The semantic value of a GO term is the aggregate semantic contribution of all terms in a DAG representing this GO term (For GO term 0043231, it is the summation of values in Table 1). The S-value of a GO term related to one of its descendant terms is its contribution to the semantics of this descendant term.

Table 1. S-values for GO terms in DAG for term *Intracellular Membrane-bound Organelle*: 0043231

GO terms	0043231	0043229	0043227	0005622
S-value	1.0	0.8	0.8	0.48
GO terms	0005623	0043226	0005575	
S-value	0.288	0.64	0.512	

2.2 Semantic similarity of GO terms

Given $DAG_A = (A, T_A, E_A)$ and $DAG_B = (B, T_B, E_B)$ for GO terms A and B respectively, the semantic similarity between these two terms, $S_{GO}(A, B)$, is defined as

$$S_{GO}(A,B) = \frac{\sum\limits_{t \in T_A \cap T_B} (S_A(t) + S_B(t))}{SV(A) + SV(B)} \tag{3}$$

where $S_A(t)$ is the S-value of GO term t related to term A and $S_B(t)$ is the S-value of GO term t related to term B.

This formula determines the semantic similarity of two GO terms based on both the locations of these terms in the GO graph and their semantic relations with their ancestor terms, addressing the drawbacks in the existing approaches. For any term $t \in T_A \cap T_B$, $S_A(t)$ may differ from $S_B(t)$ even if term t is a common term in both DAG_A and DAG_B. This is because the locations of term A and B are different in the entire GO graph. For example, Figures 1 and 2 depict the DAGs for GO terms *Intracellular Membrane-bound Organelle*: 0043231 and *Intracellular Organelle*: 0043229, respectively. Although there are five common GO terms in these two DAGs, the same term in different DAGs has different S-values related to these two specific GO terms. These values differ because the locations of the DAGs in the entire GO graph differ, and the contributions of these common terms to the semantics of the two specific GO terms also differ.

Assuming that the semantic contribution factors for 'is-a' and 'part-of' relations are 0.8 and 0.6, respectively, Table 1 contains the S-values for all terms in the DAG for GO term *Intracellular Membrane-bound Organelle*: 0043231 and Table 2 contains the S-values for all terms in the DAG for GO term *Intracellular Organelle*: 0043229. Using Equation (3), the semantic similarity between GO terms 0043231 and 0043229 is calculated as:

$$S_{GO}(0043231{,}0043229) = 0.7727.$$

2.3 Functional similarity of genes

Usually one gene is annotated by many GO terms. For instance, genes ADh4 and Ldb3 are annotated by molecular function term sets {*0004022, 0004024, 0004174,*

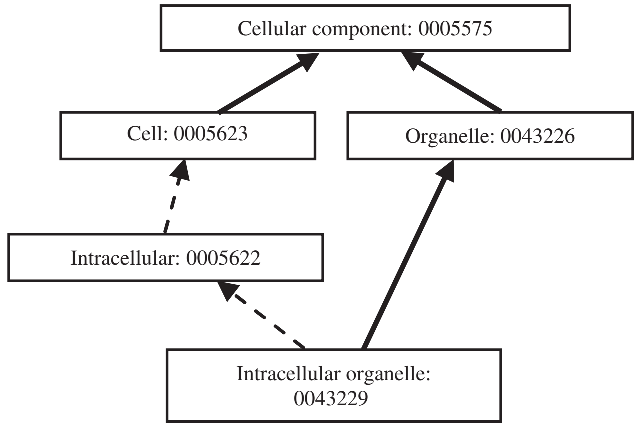


Fig. 2. DAG for GO term *Intracellular Organelle*: 0043229.

Table 2. S-values for GO terms in DAG for term *Intracellular Organelle*: 0043229

GO terms	0043229	0005622	0005623	0043226	0005575
S-value	1.0	0.6	0.36	0.8	0.64

0046872, 0008270, 0004023} and {*0009055, 0005515, 0046872, 0008270, 0020037*}, respectively in the GO database. Therefore, the similarity between the molecular functions of these two genes can be determined by comparing the semantic similarities of GO terms in these two sets.

Assuming $GO_1 = \{go_{11}, go_{12}, \ldots, go_{1m}\}$ and $GO_2 = \{go_{21}, go_{22}, \ldots, go_{2n}\}$ are two sets of GO terms that annotate genes G_1 and G_2 respectively, the following simple method was used in many studies (Langaas *et al.*, 2005; Wu *et al.*, 2005) to determine the functional similarity of genes G_1 and G_2:

$$\mathrm{Sim}(G_1, G_2) = \frac{|GO_1 \cap GO_2|}{|GO_1 \cup GO_2|} \tag{4}$$

This simple approach only considers the contribution from the exactly matched GO terms when calculating the functional similarity of two genes, ignoring the impact of different yet semantically similar GO terms. For instance, using Equation (4), the similarity between the molecular functions of genes Adh4 and Ldb3 is calculated as Sim(ADh4, Ldb3) = 0.22, since there are only two out of nine terms matched exactly in their associated molecular function term sets. This result suggests that the molecular functions of ADh4 and Ldb3 are not similar since the range of the similarity value obtained by Equation (4) is between 0 and 1, and value 0.22 is at the lower end of this range.

However, if we look into the detailed information about the molecular function terms annotating these two genes in Table 3, we can tell that besides the two exactly matched terms, 'metal ion binding' and 'zinc ion binding', other two terms associated with gene Ldb3, 'protein binding' and 'heme binding', are semantically similar to these two matched terms because they are very close to these terms in the GO graph. On the other hand, term 'electron carrier activity' (associated with gene Ldb3) is a child of the term 'oxidoreductase activity' (associated with gene ADh4) in the molecular function ontology. Therefore, the semantics of these two terms should be very similar. Furthermore, terms 'alcohol dehydrogenase activity', 'alcohol dehydrogenase activity, zinc-dependent', 'alcohol dehydrogenase activity, metal ion-independent' and 'electron-transferring-flavoprotein dehydrogenase activity' are great great grand children of the term 'oxidoreductase activity' in the molecular function ontology. Thus, they should have some similarities with the term 'oxidoreductase activity'. Based on these facts, the molecular functions of genes ADh4 and Ldb3 should be similar. Therefore, using Equation (4) to measure the functional similarity of genes is not practical because the results are not consistent with human perspectives.

Table 3. Information on GO terms associated to genes ADh4 and Ldb3

ADh4	
GO:0004022	Alcohol dehydrogenase activity
GO:0004024	Alcohol dehydrogenase activity, zinc-dependent
GO:0004174	Electron-transfer-flavoprotein dehydrogenase activity
GO:0046872	Metal ion binding
GO:0008270	Zinc ion binding
GO:0004023	Alcohol dehydrogenase activity, metal ion-independent
Ldb3	
GO:0009055	Electron carrier activity
GO:0005515	Protein binding
GO:0046872	Metal ion binding
GO:0008270	Zinc ion binding
GO:0020037	Heme binding

To accurately measure the functional similarity between two genes, we must also consider the contributions from the semantically similar terms that annotate these two genes respectively. Thus, we first define the semantic similarity between one GO term and a set of GO terms. The semantic similarity between one term *go* and a GO term set $GO = \{go_1, go_2, \ldots, go_k\}$, Sim(*go*, *GO*), is defined as the maximum semantic similarity between term *go* and any of the terms in set *GO*. That is,

$$\mathrm{Sim}(go, GO) = \max_{1 \le i \le k}(S_{GO}(go, go_i)) \tag{5}$$

Therefore, given two genes G_1 and G_2 annotated by GO term sets $GO_1 = \{go_{11}, go_{12}, \ldots, go_{1m}\}$ and $GO_2 = \{go_{21}, go_{22}, \ldots, go_{2n}\}$ respectively, we define their functional similarity as,

$$\mathrm{Sim}(G1, G2) = \frac{\sum_{1 \le i \le m} \mathrm{Sim}(go_{1i}, GO_2) + \sum_{1 \le j \le n} \mathrm{Sim}\big(go_{2j}, GO_1\big)}{m + n} \tag{6}$$

Now we measure the molecular function similarity between genes ADh4 and Ldb3 using Equations (5) and (6). Assuming the semantic contribution factors for 'is-a' and 'part-of' relations to be 0.8 and 0.6 respectively, we obtain the semantic similarities between the molecular function terms that annotate genes ADh4 and Ldb3 respectively, and list the results in

Table 4. Similarities between molecular function terms that annotate genes ADh4 and Ldb3, respectively

	GO:0046872	GO:0008270	GO:0009055	GO:0020037	GO:0005515
GO:0004023	0.112	0.071	0.427	0.112	0.141
GO:0004024	0.112	0.071	0.427	0.112	0.141
GO:0046872	1	0.664	0.173	0.390	0.480
GO:0016491	0.213	0.142	0.814	0.213	0.262
GO:0004022	0.126	0.081	0.482	0.126	0.157
GO:0008270	0.664	1	0.115	0.259	0.321
GO:0004174	0.126	0.081	0.482	0.126	0.157

Table 4. Using Equations (5) and (6), we get Sim(ADh4, Ldb3) = 0.693.

This similarity value is in the high end of the range from 0 to 1, indicating that ADh4 and Ldb3 are analogous in terms of their molecular functions. Therefore, it confirms that the functional similarity obtained by our algorithm matches the human perception.

3 VALIDATION OF OUR APPROACH

The evaluation of semantic similarity measurement methods is a challenging task, because it usually requires human involvement. In natural language domain, most studies collect a small set of term pairs and let people rank their semantic similarities. Then, the correlations between the measured semantic similarity values and the human similarity rankings are used to evaluate the semantic similarity measurement method. In this article, we use a similar approach to evaluate our similarity measurement algorithm. We use the gene annotation and classification information in pathways manually curated by researchers at the Saccharomyces genome database(SGD) (http://pathway.yeastgenome.org/biocyc/) as the reference for our similarity measurement. Although recent studies (Guo *et al*., 2006; Sevilla *et al*., 2005; Wang *et al*., 2004) used the correlation with gene sequence or gene expression similarities to evaluate the semantic similarity measurement methods, the feasibility of these evaluation methods is still debatable because there is not always correlation between the gene functional similarities and the gene sequence or gene expression similarities.

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 (Ball *et al*., 2000). These genes are also manually clustered by their molecular functions. For instance, there are five genes, ZWF1, GND1, GND2, SOL3 and SOL4, in the *oxidative branch of the pentose phosphate pathway*. Among these genes, SOL3 and SOL4 are annotated by the same GO term and the same EC number, and GND1 and GND2 are also annotated by the same GO term and the same EC number. Conversely, the GO term and EC number used to annotate gene ZWF1 are very similar to those annotating genes GND1 and GND2.

To demonstrate the advantages of our similarity measurement algorithm over the existing methods, we implemented two online gene-clustering tools (http://bioinformatics.clemson.edu/G-SESAME/knowledge Discovery.html) based on our algorithm and Resnik's method, respectively. These tools first measure the functional similarities between the input genes and, then, cluster these genes based on the obtained similarity values. We also implement a visualization tool to display the annotation information for a pair of genes on the molecular function ontology to demonstrate the functional similarity of these two genes. With the annotation information in the pathway and GO database (visualized by our visualization tool if necessary), we can evaluate whether the similarity values obtained by a similarity measurement method are consistent with human perspectives by visual examination.

Since Resnik's method is the best among these existing methods according to recent evaluation studies (Guo *et al*., 2006; Sevilla *et al*., 2005; Wang *et al*., 2004), comparing with this method will further validate our algorithm. Using these tools to cluster genes in pathways containing at least three genes in the SGD database, we found that similarity values and clustering results obtained by our algorithm are consistent with human perspectives while similarity values and clustering results obtained by Resnik's method are often inconsistent with the human perception.

Due to the space limitation, we cannot present all evaluation results in this article. Therefore, we only use one example to explain why our similarity measurement algorithm is better than Resnik's method. The rest of the evaluation results can be found at http://bioinformatics.clemson.edu/Publication/Supplement/gsp.htm. In this example, we use gene-clustering tools to cluster genes in *tryptophan degradation pathway* depicted in Figure 3. In our gene-clustering tool, we choose the molecular function ontology for measuring the semantic similarity of GO terms because genes on these pathways are manually annotated and clustered by their molecular functions. We assign 0.8 to be the semantic contribution factor for the 'is-a' relation. The semantic contribution factor for the 'part-of' relation does not affect the similarities between the molecular functions of genes because no 'part-of' relation exists in the molecular function ontology. Since we use the annotation information from the SGD database as the reference, we select this database as the data source.

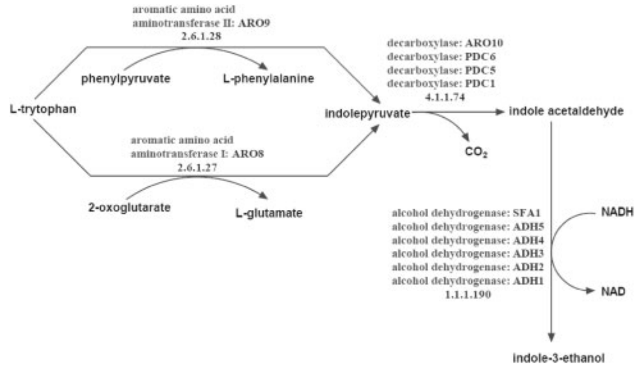
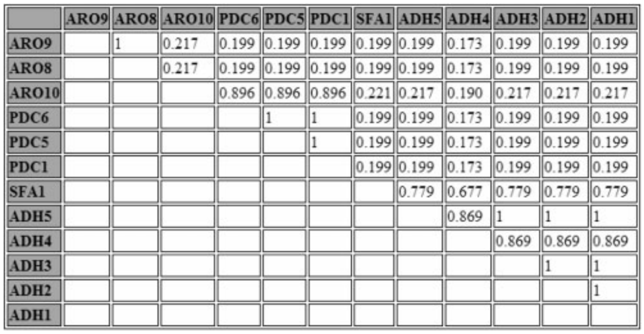


Fig. 3. Functions of genes in a *S.cerevisiae* pathway: tryptophan degradation from the SGD.



	ARO9	ARO8	ARO10	PDC6	PDC5	PDC1	SFA1	ADH5	ADH4	ADH3	ADH2	ADH1
ARO9		1	0.217	0.199	0.199	0.199	0.199	0.199	0.173	0.199	0.199	0.199
ARO8			0.217	0.199	0.199	0.199	0.199	0.199	0.173	0.199	0.199	0.199
ARO10				0.896	0.896	0.896	0.221	0.217	0.190	0.217	0.217	0.217
PDC6					1	1	0.199	0.199	0.173	0.199	0.199	0.199
PDC5						1	0.199	0.199	0.173	0.199	0.199	0.199
PDC1							0.199	0.199	0.173	0.199	0.199	0.199
SFA1								0.779	0.677	0.779	0.779	0.779
ADH5									0.869	1	1	1
ADH4										0.869	0.869	0.869
ADH3											1	1
ADH2												1
ADH1												

Fig. 4. Similarity values among genes in tryptophan degradation pathway obtained by our algorithm.

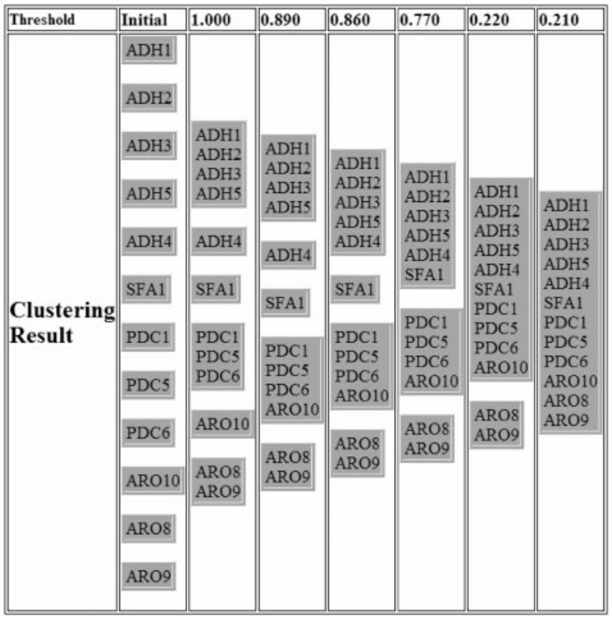
The similarity values among these genes obtained by our algorithm are depicted in Figure 4. The similarity between genes ARO8 and ARO9 is measured to be 1 by our algorithm. Because both of these genes are annotated by only one molecular function term 'aromatic-amino-acid transaminase activity' in the SGD database, they have the same molecular function based on the annotation data. Therefore, the similarity value obtained by our algorithm is consistent with the human perception. Our algorithm also determines that the molecular function similarity between genes ADH1, ADH2, ADH3 and ADH5 is 1. This is again consistent with the human perception since these genes are annotated by only one molecular function term 'alcohol dehydrogenase activity' in the SGD database. Meanwhile, the similarity between PDC1, PDC5 and PDC6 are measured as 1 by our algorithm since they are annotated by only one molecular function term 'pyruvate decarboxylase activity' in the SGD database. Besides these genes that have the same molecular function, gene ADH4 is annotated by the molecular function term 'alcohol dehydrogenase activity, zinc-dependent' which is a child term of 'alcohol dehydrogenase activity'.

Thus, the molecular function of gene ADH4 should be very similar to the molecular function of gene ADH1, ADH2, ADH3 and ADH5. Our algorithm determines that the similarity between ADH4 and the other four genes is 0.869, consistent with the human perspective. Gene ARO10 is annotated by three GO terms 'carboxylase activity', 'pyruvate decarboxylase activity', and 'phenylpyruvate decarboxylase activity'. Besides sharing the same term 'pyruvate decarboxylase activity' with genes PDC1, PDC5 and PDC6, the other two terms that annotate ARO10 are also very similar to term 'pyruvate decarboxylase activity' because they are closely located in the GO graph. Based on our algorithm, the similarity between genes ARO10 and PDC1 is 0.896, again consistent with the human perception. Gene SFA1 is annotated by two GO terms 'formaldehyde dehydrogenase (glutathione) activity' and 'alcohol dehydrogenase activity'. Because SFA1 shares one term 'alcohol dehydrogenase activity' with gene ADH1 and has another term similar to term 'alcohol dehydrogenase activity', our algorithm determines that the functional similarity between SFA1 and ADH1 is 0.779. This similarity value is also consistent with the human perception. As shown in Figure 5, when the similarity threshold is at 0.77, our gene-clustering tool groups these genes into three clusters which are associated with three different stages in *tryptophan degradation* pathway.

Resnik's method relies on the corpus statistics to determine the information content of the GO terms. We use the gene annotation data in the GO database to calculate the corpus statistics. We note that using different gene annotation data to obtain the corpus statistics may yield different semantic similarity values for the same GO term pair based on Resnik's method. Figures 6 and 7 show the similarity values and clustering results of genes in *tryptophan degradation* pathway obtained by the gene-clustering tool based on Resnik's method. These results demonstrate the weakness of Resnik's method. That is, it ignores the information contained in the structure of the ontology by only concentrating on the information content of a term.

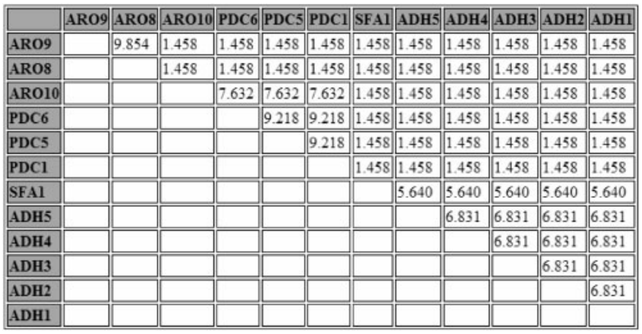
For instance, genes ARO8 and ARO9 are annotated by only one molecular function term 'aromatic-amino-acid transaminase activity' in the SGD database, and genes ADH1, ADH2, ADH3 and ADH5 are also annotated by only one molecular function term 'alcohol dehydrogenase activity'. Furthermore, the distances from these two GO terms to the root term 'molecular function' are equal in the GO graph. Therefore, the similarity between genes ARO8 and ARO9 should be equal to the similarity among genes ADH1, ADH2, ADH3 and ADH5 in terms of their molecular functions. However, Resnik's method measured the similarity between genes ARO8 and ARO9 to be 9.854, and the similarity among genes ADH1, ADH2, ADH3 and ADH5 to be 6.831. Similarly, genes PDC1, PDC5 and PDC6 are annotated by only one molecular function term 'pyruvate decarboxylase activity' in the SGD database. However, the similarity value among PDC1, PDC5 and PDC6 does not equal to the similarity value between genes ARO8 and ARO9 based on Resnik's method. On the other hand, gene ADH4 is annotated by a molecular function term 'alcohol dehydrogenase activity, zinc-dependent' which is a child term of 'alcohol dehydrogenase activity'. It means the molecular function of gene ADH4 is different from the molecular function of genes ADH1, ADH2, ADH3 and ADH5 based on the gene annotation information.

However, Resnik's method cannot tell this difference because the similarity value between ADH4 and ADH1 is



Threshold	Initial	1.000	0.890	0.860	0.770	0.220	0.210
Clustering Result	ADH1 ADH2 ADH3 ADH5 ADH4 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9	ADH1 ADH2 ADH3 ADH5 ADH4 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9	ADH1 ADH2 ADH3 ADH5 ADH4 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9	ADH1 ADH2 ADH3 ADH5 ADH4 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9	ADH1 ADH2 ADH3 ADH5 ADH4 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9	ADH1 ADH2 ADH3 ADH5 ADH4 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9	ADH1 ADH2 ADH3 ADH5 ADH4 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9

Fig. 5. Clustering results of genes in tryptophan degradation pathway based on the similarity values obtained by our algorithm.



	ARO9	ARO8	ARO10	PDC6	PDC5	PDC1	SFA1	ADH5	ADH4	ADH3	ADH2	ADH1
ARO9		9.854	1.458	1.458	1.458	1.458	1.458	1.458	1.458	1.458	1.458	1.458
ARO8			1.458	1.458	1.458	1.458	1.458	1.458	1.458	1.458	1.458	1.458
ARO10				7.632	7.632	7.632	1.458	1.458	1.458	1.458	1.458	1.458
PDC6					9.218	9.218	1.458	1.458	1.458	1.458	1.458	1.458
PDC5						9.218	1.458	1.458	1.458	1.458	1.458	1.458
PDC1							1.458	1.458	1.458	1.458	1.458	1.458
SFA1								5.640	5.640	5.640	5.640	5.640
ADH5									6.831	6.831	6.831	6.831
ADH4										6.831	6.831	6.831
ADH3											6.831	6.831
ADH2												6.831
ADH1												

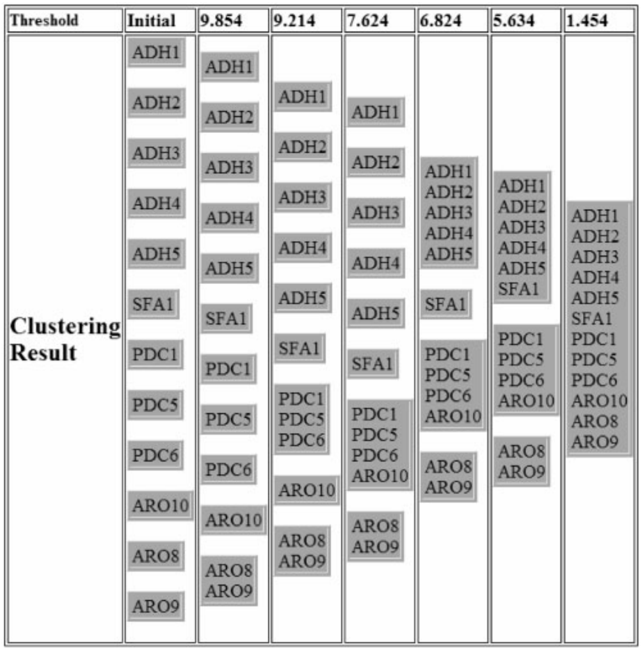
Fig. 6. Similarity values among genes in tryptophan degradation pathway obtained by Resnik's method.

the same as the similarity value between ADH2 and ADH1 based on this method. These results show that the similarity values and clustering results obtained by Resnik's method are inconsistent with human perspectives, indicating using only information content derived from annotation statistics is not suitable for measuring the semantic similarity of GO terms.

4 ADVANTAGES OF OUR ALGORITHMS

Our evaluation in Section 3 demonstrates the advantages of our algorithm over the Resnik's algorithm. As we discussed in Section 1, Lin and Jiang's methods have a serious drawback in dealing with shadow annotation. Our method does not have such a problem because the denominator in Equation (3) is smaller when terms are near the root of the ontology, and the same amount of difference on the numerator will cause a larger difference in the semantic similarity value. Therefore, using our method, if two terms sharing the same parent are near the root of the ontology (terms are more general), their semantic similarity value is less than that of two terms having the same parent and being far away from the root of the ontology. This is consistent with human perspectives.

In summary, our semantic similarity 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 similarity. Therefore, it provides a consistent measurement for the semantic similarity between two GO terms, independent of the annotation statistics. Second, our algorithm is designed to encode the human perception of the semantic relationships between child and parent terms. Thus, the semantic similarity of GO terms obtained by our algorithm can reflect the closeness of their biological meanings in human perspectives.



Threshold	Initial	9.854	9.214	7.624	6.824	5.634	1.454
Clustering Result	ADH1 ADH2 ADH3 ADH4 ADH5 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9	ADH1 ADH2 ADH3 ADH4 ADH5 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9	ADH1 ADH2 ADH3 ADH4 ADH5 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9	ADH1 ADH2 ADH3 ADH4 ADH5 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9	ADH1 ADH2 ADH3 ADH4 ADH5 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9	ADH1 ADH2 ADH3 ADH4 ADH5 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9	ADH1 ADH2 ADH3 ADH4 ADH5 SFA1 PDC1 PDC5 PDC6 ARO10 ARO8 ARO9

Fig. 7. Clustering results of genes in tryptophan degradation pathway based on the similarity values obtained by Resnik's method.

5 DETERMINING SEMANTIC CONTRIBUTION FACTORS

The semantic similarity value of GO terms measured by our method relates to the semantic contribution factors for 'is-a' and 'part-of' relations. These two parameters determine how much a parent term contributes to the semantics of its child term. If the semantic contribution factor is set to 0, it means no contribution from the ancestor terms is considered for the semantics of a specific term, conflicting with the rule on how to use GO terms to annotate genes discussed in the GO website. If the semantic contribution factor is set to 1, it means all ancestor terms equally contribute to the semantics of a specific term, inconsistent with the human perception. Therefore, the semantic contribution factor should be greater than 0 and less than 1. We suggest selecting the semantic contribution factor to be at least 0.5 to allow an ancestor term farther from a specific term to have a meaningful impact on its semantics. By measuring the similarities between the molecular functions of genes in all pathways retrieved from the SGD database under various semantic factors (varying from 0.5 to 0.9) and clustering these genes based on the obtained similarity values, we found that selecting 0.8 as the contribution factor for 'is-a' relation produced similarity values that are most consistent with human perspectives and the gene clustering results are most consistent with their manual classification. However, the similarity values between the molecular functions of these genes were not changed when the semantic contribution factor for 'part-of' relation varies. This is because no 'part-of' relation exists in the molecular function ontology. However, there are many 'part-of' relationship edges in the biological process and cellular component ontologies although 'is-a' relations are dominant ones. Based on our experimental studies with these two ontologies, the semantic contribution factor for the 'part-of' relation should be 0.6 or 0.7.

6 CONCLUSION AND FUTURE STUDIES

In this article, we proposed a novel method to encode a GO term's semantics into a numeric value by aggregating the semantic contributions of their ancestor terms in the GO graph and, in turn, devised an algorithm to measure the semantic similarity of two GO terms. Then, we designed an algorithm to measure the functional similarity of two genes based on the semantic similarities among the GO terms annotating these genes. With these algorithms, we implemented a set of online tools to measure the semantic similarities of GO terms and the functional similarities of genes, and to cluster genes based on their functional similarity values (http://bioinformatics.clemson.edu/G-SESAME). Although we are currently using fixed semantic contribution factors for all 'is-a' or 'part-of' edges respectively, we will study whether we can improve our similarity measurement by varying the semantic contribution factors for relationship edges based on their distance to the root of the ontology. We will also extend our algorithms to handle the gene annotation data using other ontologies.

ACKNOWLEDGEMENTS

We would like to thank Dr Wayne Madison, Dr Pradip Srimani and Clemson DCIT for providing computing facilities to host our online tools. We would also thank the associate editor Dr Jonathan Wren and all reviewers for their valuable comments and suggestions.

Conflict of Interest: None declared.

REFERENCES

Ball,C.A. *et al.* (2000) Integrating functional genomic information into the Saccharomyces Genome Database. *Nucleic Acids Res.*, **28**, 77–80.

Couto,F. *et al.* (2003) Implementation of a Functional Semantic Similarity Measure between Gene-Products, Department of Informatics, University of Lisbon, *DI/FCUL TR 03-29*, November, 2003.

Guo,X. *et al.* (2006) Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, **22**, 967–973.

Jiang,J. and Conrath,D. (1997) Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *tenth International Conference on Research on Computational Linguistics (ROCLING X)*, Taiwan.

Kriventseva,E.V. *et al.* (2001) CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Res*, **29**, 33–36.

Langaas,M. *et al.* (2005) Statistical hypothesis testing of associa-tion between two reporter lists within the GO-hierarchy, *Technical report*. Department of Mathematical Sciences, Norwegian University of Science and Technology.

Lee,S.G. *et al.* (2004) A graph-theoretic modeling on GO space for biological interpretation of gene clusters. *Bioinformatics*, **20**, 381–388.

Lin,D. (1998) An information-theoretic definition of similarity, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *fifteenth International Conference on Machine Learning*. pp. 296–304.

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

Sevilla,J.L. *et al.* (2005) Correlation between Gene Expression and GO Semantic Similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **2**, pp. 330–338.

Wang,H. *et al.* (2004) Gene Expression Correlation and Gene Ontology-Based Similarity: An Assessment of Quantitative Relationships. In *Proceedings of the 2004 IEEE Symposium on Computational In-telligence in Bioinformatics and Computational Biology (CIBCB'2004)*, pp. 25–31.

Wu,H. *et al.* (2005) Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Res.*, **33**, 2822–2837.