Measure the Semantic Similarity of GO Terms Using Aggregate Information Content

Xuebo Song, Lin Li, Pradip K. Srimani, Philip S. Yu, and James Z. Wang

Abstract—The rapid development of gene ontology (GO) and huge amount of biomedical data annotated by GO terms necessitate computation of semantic similarity of GO terms and, in turn, measurement of functional similarity of genes based on their annotations. In this paper we propose a novel and efficient method to measure the semantic similarity of GO terms. The proposed method addresses the limitations in existing GO term similarity measurement techniques; it computes the semantic content of a GO term by considering the information content of all of its ancestor terms in the graph. The aggregate information content (AIC) of all ancestor terms of a GO term implicitly reflects the GO term's location in the GO graph and also represents how human beings use this GO term and all its ancestor terms to annotate genes. We show that semantic similarity of GO terms obtained by our method closely matches the human perception. Extensive experimental studies show that this novel method also outperforms all existing methods in terms of the correlation with gene expression data. We have developed web services for measuring semantic similarity of GO terms and functional similarity of genes using the proposed AIC method and other popular methods. These web services are available at http://bioinformatics.clemson.edu/G-SESAME.

Index Terms—Gene ontology, GO similarity, gene expression, G-SESAME

1 Introduction

Gene ontology (GO) [1] describes the attributes of genes and gene products (either RNA or protein, resulting from expression of a gene) using a structured and controlled vocabulary. GO consists of three ontologies: biological process (BP), cellular component (CC) and molecular function (MF), each of which is modeled as a directed acyclic graph. In recent past, many biomedical databases, such as model organism databases (MODs) [2], UniProt [3], SwissProt [4], have been annotated by GO terms to help researchers understand the semantic meanings of biomedical entities. With such a large diverse biomedical data set annotated by GO terms, computing functional or structural similarity of biomedical entities has become a very important research topic. Many researchers have tried to measure the functional similarity of genes or proteins based on their GO annotations [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. Since different biomedical researchers may annotate the same or similar gene function with different but semantically similar GO terms based on their research findings, an accurate measure of semantic similarity of GO terms is critical for accurate measurement of gene functional similarities.

While those existing studies have proposed different methods to measure the semantic similarity of GO terms, they all have their limitations. In general, there are three types of methods for measuring the semantic similarity of GO terms: node-based [9], [20], [21], [22], edge-based [10], [17], [23], [24], and hybrid [6], [11], [18], [19] methods. See Section 2 for a brief discussion of some most representative methods and their limitations.

In this paper, we propose a novel method to measure the semantic similarity of GO terms. This method is based on two major observations: (1) In general, the dissimilarity of GO terms near the root (more general terms) of GO graph should be larger than that of the terms at a lower level (more specific terms); (2) the semantic meaning of one GO term should be the aggregation of all semantic values (SVs) of its ancestor terms (including the term itself). The first observation follows the human perception of term semantic similarity at different specialization levels of the ontology. The second observation agrees with how human beings use the term to annotate genes.

The rest of the paper is organized as follows. We review existing most representative methods for semantic similarity measurement of GO terms in Section 2. We introduce our proposed aggregate information content (AIC) based approach in Section 3. We present details of experimental evaluation of AIC in Section 4 and discuss the advantages of our proposed AIC based algorithms in Section 5. After briefly introducing the implemented web services in Section 6, we conclude this paper and discuss the future studies in Section 7.

2 Related Prior Work

A large number of studies [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [17], [18], [19] have appeared in the literature in the last 15 years to measure the semantic similarity of GO

- *X. Song, P.K. Srimani, and J.Z. Wang are with the School of Computing, Clemson University, Clemson, SC 29634. E-mail: {xuebos, srimani, jzwang}@clemson.edu.*
- *L. Lin is with the Department of Computer Science and Information Systems, Murray State University, Murray, KY 42071. E-mail: lli6@murraystate.edu.*
- *P.S. Yu is with the Department of Computer Science, University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607. E-mail: psyu@uic.edu.*

Manuscript received 14 July 2013; revised 5 Oct. 2013; accepted 26 Nov. 2013. Date of publication 11 Dec. 2013; date of current version 5 June 2014.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2013.176

terms. All of these methods can be broadly classified into three categories: node-based, edge-based, and hybrid methods. The three most cited representative methods [20], [21], [22] were originally designed to measure the semantic similarity of natural language terms. While they have been widely adopted by bioinformatics researchers to measure the semantic similarity of GO terms, each of them has its own limitations. In 2007, Wang et al. [6] proposed a new measure of the semantic similarity of GO terms: this new hybrid method considers both the GO structure and the semantic content (biological meaning) of the GO terms in measuring the semantic similarity of GO terms, and many studies [5], [11], [15], [16] have shown the superiority of this hybrid method. It has been widely accepted by biomedical researchers [11] since it was published.

2.1 Limitations of Current Methods

Node-based measures (e.g., Resnik's [20], Lin's [21], Jiang and Conrath's [22], Schlicker's [9]) rely mainly on information content (IC) of the GO terms to represent their semantic values; IC of a GO term is derived from the frequency of its presence (including the presence of its children terms) in a certain corpus (e.g., SGD database, GO database). Resnik's [20] method concentrates only on the maximum information contained in ancestors (MICA) of the compared GO terms, but ignores the locations of these terms in the GO graph, e.g., a GO term's distance from the root of the ontology, and the semantic impact of other ancestor terms. A term's distance to the root of the ontology shows the specialization level of this term in human perception. If a term is far from the root in the ontology, it means biomedical researchers know more details about this term and the meaning of the term is more specific. On the other hand, if a term is closer to the root of the ontology, it means the term is a more general term, such as cellular process or metabolic process, which does not provide too much details about the related biomedical entities. Ignoring the specialization level of a term in the ontology is the principal reason that the semantic similarity obtained by these methods is inconsistent with human perception; they suffer from "shallow annotation" problem [6], [8], [13] in which the semantic similarity of GO terms near the root of the ontology are sometimes measured very high.

Edge-based approaches [10], [17], [23], [24] are based on the length of graph paths connecting the terms being compared. Some edge-based approaches [23] treat all edges equally, ignoring the levels of edges in the ontology. This simple equal-edge-based approach also suffers from "shallow annotation" because based on this approach, the semantic similarity of two terms with a certain graph distance near the root would be equal to the semantic similarity of two terms with the same graph distance but away from the root. To address the "shallow annotation" problem, other edge-based methods [10], [17], [24] assign different weights to the edges at the different levels of the ontology, assuming that the edges at the same level of the ontology have the same weight. However, the terms at the same level of the GO graph do not always have the same specificity because different gene properties demand different levels of detailed studies. It means the edges at the same level of the GO graph but in different GO branches do not necessarily have the same weights.

The hybrid method [6] considers both the GO structure and the semantics (biological meanings) of GO terms at different ontological levels. However, this method uses two semantic contribution factors, obtained from empirical study of gene classification of certain species, to calculate the semantic values of GO terms. Semantic contribution factors obtained by empirical studies on genes from certain species may not be optimal for measuring the functional similarity of genes in other species. A recent study [19] has proposed to consider the interaction between the descendants of the GO terms in computing semantic similarity between them. However, this is predominantly an add-on to other existing methods and thus still inherits their drawbacks.

2.2 Review of Existing Representative Methods

We provide a brief overview of the four most representative methods for GO term semantic similarity measure: Method A by Resnik [20], Method B by Lin [21], Method C by Jiang and Conrath [22], and Method D by Wang et al. [6]. We use these four methods as benchmarks to evaluate the relative performance of our proposed AIC method in the next sections.

Method A: The frequency of a GO term is recursively defined as

$$freq(t) = annotation(t) + \sum_{i \in child(t)} freq(i), \tag{1}$$

where $annotation(t)$ is the number of gene products annotated with term t in the GO database. $child(t)$ is the set of children of term t. For each term t, $p(t)$ denotes the probability that term t occurs in the GO database,

$$p(t) = freq(t)/freq(root). \tag{2}$$

Information Content of term t is defined as

$$IC(t) = -\log p(t). \tag{3}$$

Method A uses maximum information contained in ancestorsof two terms to measure the semantic similarity between them

$$sim_{GO}(a, b) = \max_{c \in P(a,b)} IC(c), \tag{4}$$

where $P(a, b)$ denotes the set of common ancestor terms of term a and term b in the ontology graph. Based on the definition of IC in Method A (Equations (1), (2), (3)), MICA often happens to be the IC value of the least common ancestor (LCA) of terms a and b.

The principal limitation of method A derives from the fact that it considers only MICA of two terms while ignoring the distances of the two terms to their LCA and the semantic contribution of other ancestor terms. For example, terms a and b have the same LCA with terms c and b in the partial GO graph shown in Fig. 1. Using method A, the semantic similarity between term a and b would be equal to the semantic similarity between term c and d, inconsistent with human perception, which suggests that the semantic

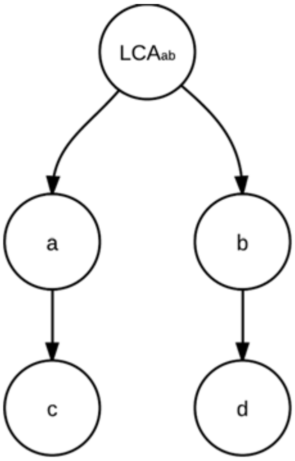


Fig. 1. GO terms at different ontology levels sharing the same LCA.

similarity between term c and d should be less than the one between term a and b.

Method B: It is based on the ratio between IC values of two terms and that of their MICA; the semantic similarity between two terms a and b is defined as

$$sim_{GO}(a, b) = \frac{2 * max_{c \in P(a,b)}\, IC(c)}{IC(a) + IC(b)}. \tag{5}$$

Method C: It introduces the concept of term distance into the semantic similarity calculation. The intuition is that two terms closer in the GO graph should be more similar than two terms farther in the GO graph. The distance between two terms a and b is defined as

$$Dis_{GO}(a, b) = IC(a) + IC(b) - 2 * \max_{c \in P(a,b)} IC(c). \tag{6}$$

The semantic similarity of two terms a and b are then defined as

$$sim_{GO}(a, b) = \frac{1}{1 + Dis_{GO}(a, b)}. \tag{7}$$

Note: Methods B and C ameliorated the principal limitation of Method A by implicitly considering the graph distance of the two terms in the semantic similarity measure. Consider the example in Fig. 1; $sim_{GO}(c, d)$ should be less than $sim_{GO}(a, b)$ according to human perception because the graph distance between c and d is greater than the graph distance between a and b. Since term a is a parent of term c, we have $freq(a) > freq(c)$ and $p(a) > p(c)$ (Equations (1) and (2)). According to the definition of IC in Equation (3), we have $IC(c) > IC(a)$. Similarly, we have $IC(d) > IC(b)$. Therefore, the semantic similarity values obtained by both methods B and C are consistent with human perception in this aspect.

However, it is possible that a GO term has multiple parent terms with different semantic relations; using MICA alone does not account for multiple parents. Also, two terms at a higher level (more general terms) of GO graph should be, as is perceived by humans, semantically more dissimilar than two terms with the same graph distance at a lower level (more specific terms). Since neither methods B nor C factor in the specialization level of the LCA of the two terms in their semantic similarity measure, the semantic similarity values obtained by these two methods may still be inconsistent with human perception as demonstrated in our experiment in Section 4.

Method D: Method D attempts to address the shortcomings of the existing methods by aggregating the semantic contributions of ancestor terms in the GO graph. The S-value of GO term t related to term x (where term t is an ancestor of term x, including term x itself) is defined as,

$$S_x(t) = \begin{cases} 1, & \text{if } t = x, \\ \max\{w_e * S_x(t') | t' \in \text{children of } t\}, & \text{if } t \neq x \end{cases} \tag{8}$$

where w_e is the semantic contribution factor of an edge (weight of the edge in the GO graph). Then the semantic value of a GO term x is defined as

$$SV(x) = \sum_{t \in T_x} S_x(t), \tag{9}$$

where T_x is the set of GO terms in DAG_x (directed acyclic graph consisting all ancestors of the term x, including term x). Finally, the semantic similarity between two GO terms a, b is defined as

$$sim_{GO}(a, b) = \frac{\sum_{t \in T_a \cap T_b} (S_a(t) + S_b(t))}{SV(a) + SV(b)}, \tag{10}$$

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. While this method combines both the semantic and the topological information of GO terms to address weaknesses of methods A, B and C, it still suffers from two disadvantages. First, it needs to use semantic contribution factor values (weight) empirically obtained from gene classification to calculate the semantic values of GO terms. Using semantic contribution factors obtained from the classification of genes from certain species may not be suitable for measuring the functional similarity of genes in other species. Second, some biomedical studies need to obtain the similarity matrix for a large group of GO terms or genes. Dynamically calculating the semantic values of GO terms is time consuming and may result in a long user response time, which will be shown in our experimental studies.

3 Aggregate Information Content Based Method

We address the limitations of the existing methods using an aggregate information content approach.

3.1 GO Similarity

This *aggregate information content* based similarity measurement method (Method AIC) considers the aggregate contribution of the ancestors of a GO term (including this GO term) to the semantics of this GO term, and takes into account how human beings use the terms to annotate genes. We use a term's IC value, as defined before (Equations (1), (2), (3)), to represent their semantic contribution values. Given the fact that terms at upper levels (more general

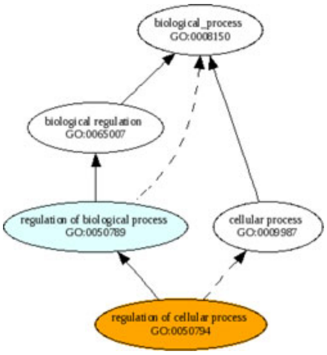


Fig. 2. GO Graph containing terms GO:0050794 and GO:0050789.

terms) of ontology graph are less specific than those at lower levels, we define the knowledge of a term t as

$$K(t) = 1/IC(t). \tag{11}$$

Unlike the weight in Method D, which represents the semantic contributions of ancestor terms using contribution factors obtained from empirical study, this $K(t)$ incorporates the statistical distribution of GO terms in the entire gene ontology. The deeper a GO term dwells in the ontology, the more we know about this term (with more knowledge, i.e., K(t)). Thus, we would say this newly defined $K(t)$ represents how much people have studied term t, thus the knowledge of term t. We further propose a logarithmic model to normalize $K(t)$ into a semantic weight $SW(t)$:

$$SW(t) = \frac{1}{1 + e^{-K(t)}}. \tag{12}$$

We then compute semantic value $SV(x)$ of the GO term x by adding the semantic weights of all its ancestors (i.e., aggregating semantic contribution of the ancestors)

$$SV(x) = \sum_{t \in T_x} SW(t), \tag{13}$$

where T_x is the set of all of its ancestors including x itself. We define the semantic similarity between GO terms a and b, based on their aggregate information content, as follows:

$$sim_{GO}(a, b) = \frac{\sum_{t \in T_a \cap T_b} 2 * SW(t)}{SV(a) + SV(b)}, \tag{14}$$

where $SW(t)$ is the semantic weight of term t defined in Equation (12), and $SV(t)$ is the semantic value of term t defined in Equation (13). Aggregating the semantic contribution of all ancestor terms implicitly factors in the position of the term in the GO graph, and overcomes the weakness of the MICA based approaches.

We demonstrate how to use the AIC method to compute the semantic similarity between two terms, GO:0050794 and GO:0050789, shown in Fig. 2. (All GO DAGs used in this paper are obtained from the web tools in the popular G-SESAME Website [25].) First, we obtain the IC values of all related GO terms from the GO database released in June 2013. The results are shown in Table 1. We note that many studies [26], [27], [28] used GOSim R package [29] to obtain the IC information for all related GO terms. However, IC values in GOSim R package are not always derived from the latest GO database. Due to continuous evolution of the GO database, IC values and semantic similarities of GO terms may change over time with the change of the GO database content. In addition, GOSim R package is hard to be integrated with the popular G-SESAME Website. Another widely used R package called GOSemSim [30] also suffers those issues. Therefore, we chose, in this paper, to calculate the IC values of GO terms directly from the latest GO database release. Second, we calculate the semantic weight for each GO term using Equation (12). Finally, we use Equation (13) and Equation (14) to get the semantic similarity of GO terms GO:0050794 and GO:0050789 as $sim_{GO}(0050794, 0050789) = 0.748$.

TABLE 1
IC Values and Semantic Weights of GO Terms

Go Terms	IC value	SW value
0050794	1.2461	0.6905
0050789	0.9906	0.7329
0065007	0.9403	0.7434
0009987	0.2610	0.9788
0008150	0	1

3.2 Gene Similarity

There are several methods [6], [8], [12] to measure the functional similarity of gene products based on the semantic similarity of GO terms. The common methods are: MAX [6], [8] and AVE [12] methods; they define functional similarity between gene products as the maximum or average semantic similarity values over the GO terms annotating the genes respectively. In this paper, we use AVE method as follows:

$$sim_{AVE}(g_1, g_2) = \underset{\substack{t_1 \in annotation(g_1) \\ t_2 \in annotation(g_2)}}{\text{average}} sim(t_1, t_2), \tag{15}$$

where $annotation(g)$ is the set of GO terms that annotates gene g. Although some studies [6], [8] use the MAX method to compute the functional similarity of genes, people [5] found that the AVE method is more stable and less sensitive to outliers. In addition, the AVE method is more compatible with our original objective of capturing all available information while the MAX method often ignores the contribution of other GO terms.

4 Experimental Evaluation of AIC

It is well known, as demonstrated in [5], [7], [8], that there is a high correlation between gene expression data and the gene functional similarity obtained from GO term similarities, i.e., genes with similar expression patterns should have high similarity in GO based measures because they should be annotated with semantically similar GO terms. We use the correlation of genes obtained from gene expression data to validate the gene functional similarities obtained by GO based similarity measures. As in many existing studies [13], [31], [32], [33], we use gene expression data from Spellman data set [34], which comprises of 6,178 genes, to obtain the

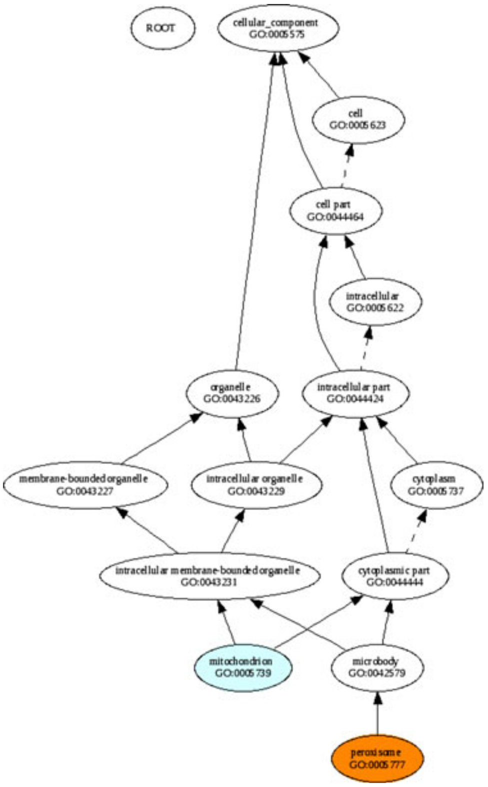


Fig. 3. GO graph of terms GO:0005739 and GO:0005777.

gene correlation patterns. The gene annotation data used to calculate the gene functional similarity is obtained from the GO database released in June 2013. In the next two subsections, we provide comparison of our method (AIC) with the state-of-the-art current methods: Method A [20], Method B [21], Method C [22], and Method D [6] in terms of GO term semantic similarity and gene functional similarity. The reason we compare our new AIC method with these four existing methods is that Method D is one of the most used methods for measuring the semantic similarity of GO terms since 2007 (72.5 million times by researchers from 71 countries according to G-SESAME statistics), while recently proposed methods have not yet been widely adopted by biomedical researchers. In addition, most newly proposed methods are different variants of methods A, B, and C which are still widely used as benchmark methods for measuring the semantic similarity of GO terms.

4.1 Comparison Analysis Based on Correlation with Human Perception

From human perspective, we know that two GO terms at higher levels of the gene ontology should have larger dissimilarity than two GO terms with the same graph distance at lower levels. Our AIC method is compatible with this observation in that two GO terms with the same graph distance at the lower levels of the gene ontology usually share more common ancestors. Therefore, the semantic similarity of GO terms obtained by our AIC method is consistent with human perception as shown in an illustrative example from our experimental results in Fig. 3 and Table 2.

TABLE 2
Semantic Similarity Values of GO Term Pairs Obtained by Different Methods

Dataset	Method	**Similarity**
SW(GO:0005739, GO:0005777)	A	1.047
	B	0.424
	C	0.260
	D	0.797
	AIC	0.902
SW(GO:0044424, GO:0005622)	A	0.430
	B	0.918
	C	0.928
	D	0.845
	AIC	0.898
SW(GO:0044444, GO:0005737)	A	0.821
	B	0.879
	C	0.815
	D	0.879
	AIC	0.939

Consider the two GO terms GO:0005739 and GO:0005777 as shown in Fig. 3. The semantic similarity values obtained by Methods A, B, C, D and AIC are shown in Table 2. These two very specific GO terms have only one different ancestor term GO:0042579; the semantic similarity between them should be very high. However, the semantic similarity values obtained by Method B [21] and Method C [22] fail to exhibit this expected behavior while Method D [6] and the proposed AIC method correctly exhibit this expected behavior. This observation reinforces our previous contention that use of MICA alone in computing similarity is not sufficient because of loss of important information. The semantic similarity values obtained by Method A are not normalized; hence, it is hard to determine the relative similarity levels without reviewing all pair-wise semantic similarity values in the GO database. Accordingly, we have excluded Method A from this comparison study.

Now, we check whether all these semantic similarity measurement methods agree with the human perspective aforementioned: two GO terms at higher levels of the gene ontology should have larger dissimilarity than two GO terms with the same graph distance at lower levels. We calculate the semantic similarity between GO:0044424 and GO:0005622 (Group 1) and the semantic similarity between GO:0044444 and GO:0005737 (Group 2). The semantic similarity values are shown in Table 2. These two groups of GO terms have similar structure in the GO graph except group 1 is closer to the root of the GO graph. Based on human perception, the semantic similarity of GO terms in group 1 should be less than that in group 2 since GO terms in group 2 are at a lower level of the GO graph. However,

TABLE 3
Pearson's Correlation Coefficients between Gene Expression Data and Gene Functional Similarities Obtained by Different Semantic Similarity Measurement Methods

Groups	Method A [20]	Method B [21]	Method C [22]	Method D [6]	**Proposed AIC**
4	0.614	0.789	0.930	0.929	**0.966**
5	0.561	0.717	**0.889**	0.802	0.850
6	0.413	0.569	0.700	0.745	**0.774**
7	0.519	0.622	**0.761**	0.725	0.733
8	0.496	0.597	0.675	0.706	**0.714**
9	0.417	0.659	0.664	0.745	**0.778**
10	0.403	0.620	0.730	0.733	**0.772**
11	0.419	0.665	0.691	0.725	**0.761**
12	0.246	0.485	0.722	0.716	**0.782**
13	0.321	0.525	0.715	0.709	**0.791**

only methods A, D and our AIC method satisfy this property. The semantic similarity values obtained by methods B and C are inconsistent with the human perception because these two methods do not consider the specialization level of two terms' LCA in the semantic similarity measure. The "shallow annotation" problem is clearly shown in these experiments.

4.2 Comparison Analysis Based on Correlation with Gene Expression Data

In our next set of experiments, we first use Pearson's correlation to compute the gene expression similarity with the Spellman data set [34]. Then, we calculate the correlation between the functional similarity of these genes obtained from BP ontology and the gene expression similarity. The objective is, as stated in [7], to test the hypothesis that pairs of genes exhibiting similar expression levels which are measured by the absolute correlation values in gene expression data tend to have high functional similarities between each other. The average of correlation coefficients between genes within an expression similarity interval estimates the mean of the statistical distribution of correlations; and it shows the underlying trend that relates expression similarity and functional similarity. We split the gene pairs into groups with equal intervals according to the absolute gene expression correlation values between gene pairs, as in previous studies [5], [7], [8], [13], and then compute Pearson's correlation coefficient between the mean of gene functional similarities and the mean of gene expression correlation values in each group. We split gene pairs into 4-13 groups respectively to avoid under-fitting and over-fitting problems [35]. We again compare the results obtained using four existing methods (Methods A, B, C and D) and those obtained using our AIC method, as shown in Table 3. The experimental results show that our AIC method generally outperforms other four methods with the highest correlation coefficients between gene functional similarity and gene expression similarity in most cases. Method D and C also showed excellent correlation between the GO based functional similarity and the gene expression similarity. Methods A and B did not perform well in this experimental study, with Method A being the worst.

4.3 Computational Efficiency of the AIC Method

While methods D and AIC show superiority to other three methods in agreement with human perception and in correlation with gene expression data, Method D requires computation of the S-value of a node in a DAG by doing a breadth first search starting from the node and exhausting the subtree of the node in the DAG (S-values are not stored at the node); this is a major computation cost of method D. In the proposed AIC method, the similarity values are precomputed using the DAG (the GO graph and its relationships are static and stored as the aggregate information content of the node) and are stored at the node; thus this method does not need to do the expensive traversal of the graph during runtime. This is the primary reason for the computational efficiency of AIC method over Method D. We use the execution time of computing the functional similarities of a large number of gene pairs to evaluate the computation efficiency of our proposed AIC method. In this experiment, we use methods D and AIC to compute the functional similarities of three sets of gene pairs. The numbers of genes in these sets are 200, 500 and 2,000 respectively. The experiment was conducted on a Linux box with a i7-2600K CPU @ 3.40 GHz, 8G memory. The execution time are shown in Table 4. As demonstrated by the experimental results, method AIC is considerably faster than method D.

TABLE 4
Computation Efficiency of Methods D and AIC

	Execution Time (seconds)		
# of Gene Pairs	200	500	2000
Method D	173	3506	36123
Method AIC	56	261	7632

5 Advantages of AIC Method

Experimental results in Section 4 demonstrate the superiority of the proposed AIC method over the representative ones, Method A [20], Method B [21], Method C [22] and Method D [6]. Method AIC is characterized with the following unique features:

- AIC shows advantages over Method A by taking into account the structural difference as stated in Section 2.2.
- AIC does not suffer from "shallow annotation" as in Methods B and C. Note that, in Equation (14) the denominator is smaller when terms are annotated at the top levels, i.e., the equal difference on the numerator will result in a larger difference in the semantic similarity value. Thus, the semantic similarity value of two terms at top levels is less than that of two terms with the same graph distance at lower levels. This is consistent with human perspectives.
- AIC exhibits high correlation coefficient between the gene expression similarity and the GO based functional similarity.
- AIC is computationally significantly faster than the popular hybrid Method D since the information content values can be precomputed. It does not use the empirically determined semantic contribution factors in semantic similarity computation.

In summary, the proposed AIC method is very promising in that it outperforms all existing state-of-the-art methods in terms of consistency with human perception, correlation with gene expression data and computational efficiency.

6 Integration with G-SESAME Web Services

Due to the high demand for computing GO semantic similarity and gene functional similarity by biomedical researchers, a number of web services, such as ProteInOn (http://lasige.di.fc.ul.pt/webtools/proteinon) [36], FunSimMat (http://www.funsimmat.de) [37], GOToolBox (http://genome.crg.ex/GOToolBox) [38], and G-SESAME (http://bioinformatics.clemson.edu/G-SESAME) [6] have been developed. All of them are very convenient and easy to use. However, none of these tools, except G-SESAME, provides the visualization, batch-mode support, and web-based APIs simultaneously. These enhanced services are very important and useful for biomedical researchers to conveniently and efficiently run their applications. We augmented and extended the original G-SESAME website by incorporating our proposed AIC method, and also implemented the other three popular methods, A, B and C, to allow users to select the appropriate method at their own interests. The redesigned G-SESAME Website has the following characteristic features:

1. It provides a list of user-friendly, easy-to-use web services for researchers to use.
2. It provides several state-of-the-art semantic measurement methods at the same place. Users can select different methods to measure the semantic similarity of GO terms and functional similarity of genes, and compare their measurement results.
3. It provides web-based visualization to allow users to inspect the locations of the GO terms within the GO graph and visually determine their semantic similarity.
4. It provides batch mode support to allow users to measure the semantic similarities of a group of GO terms or functional similarities of a group of genes.
5. It provides a list of web-based APIs to allow users to easily integrate these web services into their own applications.

7 Conclusion and Future Work

In this paper, we proposed a novel aggregate information content method to measure the semantic similarity of GO terms accurately and efficiently. This AIC approach aggregates the information content of all ancestor terms of a particular GO term while the computation of GO term's information content implicitly considers the semantic contribution of its descendant terms. Thus, this approach ensures the completeness of the semantic information in the semantic similarity measure. Our analysis and experimental results show the superiority of the proposed AIC method over the state-of-the-art methods [6], [20], [21], [22]. We further enhance the popular G-SESAME Website [6] http://bioinformatics.clemson.edu/G-SESAMEby providing web services for GO term semantic similarity measure and gene functional similarity measure using different methods, including the proposed AIC method, Resnik's, Lin's and Jiang's methods (Methods A, B, and C, respectively). We are currently implementing better visualization tools to allow users to easily compare the measurement results obtained by different methods using various data sets.

Acknowledgments

This work was supported in part by US National Science Foundation (NSF) through grants DBI-0960586, DBI-0960443, CNS-1115234, OISE-1129076, CCF 0832582, NIH grants 1 R15 CA131808-01 and 1 R01 HD069374-01 A1.

References

[1] The Gene Ontology Consortium, "Gene Ontology: Tool for the Unification of Biology," *Nature Genetics*, vol. 25, pp. 25-29, 2000.

[2] L.D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J.E. Stajich, T.W. Harris, A. Arva, and S. Lewis, "The Generic Genome Browser: A Building Block for a Model Organism System Database," *Genome Research*, vol. 12, pp. 1599-1610, 2002.

[3] The UniProt Consortium, "The Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 36, pp. D190-D195, 2008.

[4] E.V. Kriventseva, W. Fleischmann, E.M. Zdobnov, and R. Apweiler, "Clustr: A Database of Clusters of Swiss-Prot+Trembl Proteins," *Nucleic Acids Research*, vol. 29, pp. 33-36, 2001.

[5] T. Xu, L. Du, and Y. Zhou, "Evaluation of Go-Based Functional Similarity Measures Using s.cerevisiae Protein Interaction and Expression Profile Data," *BMC Bioinformatics*, vol. 9, article 472, 2008.

[6] J.Z. Wang, Z. Du, R. Payattakool, P.S. Yu, and C.-F. Chen, "A New Method to Measure the Semantic Similarity of GO Terms," *Bioinformatics*, vol. 23, pp. 1274-1281, 2007.

[7] H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo, "Gene Expression Correlation and Gene Ontology-Based Similarity: An Assessment of Quantitative Relationships," *Proc. IEEE Symp. Computational Intelligence in Bioinformatics and Computational Biology*, pp. 25-31, 2004.

[8] J.L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J.M. Mato, L.A. Martinez-Cruz, F.J. Corrales, and A. Rubio, "Correlation Between Gene Expression and GO Semantic Similarity," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 2, no. 4, pp. 330-338, Oct.-Dec. 2005.

[9] A. Schlicker, F.S. Domingues, J. Rahnenfuhrer, and T. Lengauer, "A New Measure for Functional Similarity Functional Similarity of Gene Products Based on Gene Ontology," *BMC Bioinformatics*, vol. 7, article 302, 2006.

[10] J. Cheng, M. Cline, J. Martin, D. Finkelstein, T. Awad, D. Kulp, and M.A. Siani-Rose, "A Knowledge-Based Clustering Algorithm Driven by Gene Ontology," *J. Biopharmaceutical Statistics*, vol. 14, no. 3, pp. 687-700, 2004.

[11] C. Pesquita, D. Faria, A.O. Falcao, P. Lord, and F.M. Couto, "Semantic Similarity in Biomedical Ontologies," *PLoS Computational Biology*, vol. 5, no. 7, e1000443, 2009.

[12] F. Azuaje, H. Wang, and O. Bodenreider, "Ontology-Driven Similarity Approaches to Supporting Gene Functional Assessment," *Proc. ISMB'2005 SIG Meeting on Bio-ontologies*, pp. 9-10, 2005.

[13] B. Li, J.Z. Wang, F. Luo, F.A. Feltus, and J. Zhou, "Effectively Integrating Information Content and Structural Relationship to Improve the Gene Ontology Similarity Measure between Proteins," *Proc. Int'l Conf. Bioinformatics and Computational Biology (BioComp '10)*, pp. 166-172, 2010.

[14] C. Pesquita, D. Faria, H. Bastos, A.O. Falcao, and F.M. Couto, "Evaluating GO-Based Semantic Similarity Measures," *Proc. 10th Annual Bio-Ontologies Meeting*, pp. 37-40, 2007.

[15] T. Ravasi et al., "An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man," *Cell*, vol. 140, no. 5, pp. 744-752, 2010.

[16] N.L. Washington, M.A. Haendel, C.J. Mungall, M. Ashburner, M. Westerfield, and S.E. Lewis, "Linking Human Diseases to Animal Models Using Ontology-Based Phenotype Annotation," *PLoS Biology*, vol. 7, no. 11, p. e1000247, 2009.

[17] M. Li, X. Wu, Yi. Pan, and J. Wang, "HF-measure: A New Measurement for Evaluating Clusters in Proteinprotein Interaction Networks," *PROTEOMICS*, vol. 13, no. 2, pp. 291-300, 2013.

[18] Z. Teng, M. Guo, X. Liu, Q. Dai, C. Wang, and P. Xuan, "Measuring Gene Functional Similarity Based on Group-Wise Comparison of GO Terms," *Bioinformatics*, vol. 29, no. 11, pp. 1424-1432, 2013.

[19] H. Yang, T. Nepusz, and A. Paccanaro, "Improving GO Semantic Similarity Measures by Exploring the Ontology beneath the Terms and Modelling Uncertainty," *Bioinformatics*, vol. 28, no. 10, pp. 1383-1389, May 2012.

[20] P. Resnik, "Semantic Similarity in Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language," *J. Artificial Intelligence Research*, vol. 11, pp. 95-130, 1999.

[21] D. Lin, "An Information-Theoretic Definition of Similarity," *Proc. Int'l Conf. Machine Learning*, pp. 296-304, 1998.

[22] J.J. Jiang and D.W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," *Proc. Int'l Conf. Research in Computational Linguistics*, pp. 19-33, 1997.

[23] V. Pekar and S. Staab, "Taxonomy learning: Factoring the Structure of a Taxonomy into a Semantic Classification Decision," *Proc. Int'l Conf. Computational Linguistics*, vol. 2, pp. 786-792, 2002.

[24] H. Wu, Z. Su, F. Mao, V. Olman, and Y. Xu, "Prediction of Functional Modules Based on Comparative Genome Analysis and Gene Ontology Application," *Nucleic Acids Research*, vol. 33, no. 9, pp. 2822-2837, 2005.

[25] Z. Du, L. Li, C.-F. Chen, P.S. Yu, and J.Z. Wang, "G-sesame: Web Tools for GO-Term-Based Gene Similarity Analysis and Knowledge Discovery," *Nucleic Acids Research*, vol. 37, pp. W345-W349, 2009.

[26] K. Ovaska, M. Laakso, and S. Hautaniemi, "Fast Gene Ontology Based Clustering for Microarray Experiments," *BioData Mining*, vol. 1, no. 1, p. 11+, Nov. 2008.

[27] Y. Li and B.-L. Lu, "Semantic Similarity Definition over Gene Ontology by Further Mining of the Information Content," *Proc. Sixth Asia-Pacific Bioinformatics Conf. (APBC)*, vol. 6, pp. 155-164, 2008.

[28] P. Smialowski et al., "The Negatome Database: A Reference Set of Non-Interacting Protein Pairs," *Nucleic Acids Research*, vol. 38, Database Issue, pp. D540-D544, 2010.

[29] H. Froehlich, N. Speer, A. Poustka, and T. Beissbarth, "GOSim—An R-Package for Computation of Information Theoretic GO Similarities between Terms and Gene Products," *BMC Bioinformatics*, vol. 8, article 166, 2007.

[30] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: An R Package for Measuring Semantic Similarity Among GO Terms and Gene Products," *Bioinformatics*, vol. 26, no. 7, pp. 976-978, 2010.

[31] L.J. Heyer, S. Kruglyak, and S. Yooseph, "Exploring Expression Data: Identification and Analysis of Coexpressed Genes," *Genome Research*, vol. 9, pp. 1106-1115, 1999.

[32] D. Jiang, C. Tang, and A. Zhang, "Cluster Analysis for Gene Expression Data: A Survey," *IEEE Trans. Knowledge and Data Eng.*, vol. 16, no. 11, pp. 1370-1386, Nov. 2004.

[33] F.D. Gibbons and F.P. Roth, "Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation," *Genome Research*, vol. 12, pp. 1574-1581, 2002.

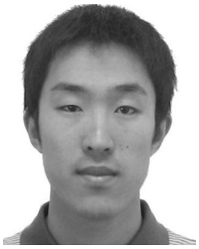
[34] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273-3297, 1998.

[35] P. Courrieu, M. Brand-D'abrescia, R. Peereman, D. Spielerand, and A. Rey, "Validated Intraclass Correlation Statistics to Test Item Performance Models," *Behavior Research Methods*, vol. 43, pp. 37-55, 2010.

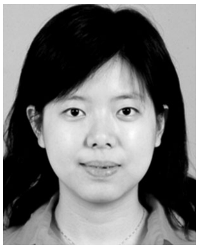
[36] D. Faria, C. Pesquita, F. Couto, and A. Falcao, "A Web Tool for Protein Semantic Similarity," Technical Report DI-FCUL-TR-07-6, Dept. of Informatics, Univ. of Lisbon, 2007.

[37] A. Schlicker and M. Albrecht, "FunSimMat: A Comprehensive Functional Similarity Database," *Nucleic Acids Research*, vol. 36, Database Issue, pp. 434-439, 2008.

[38] D. Martin, C. Brun, E. Remy, P. Mouren, D. Thieffry, and B. Jacq, "GOToolBox: Functional Analysis of Gene Datasets Based on Gene Ontology," *Genome Biology*, vol. 5, no. 12, p. R101, 2004.



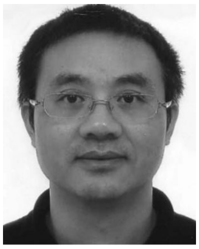
Xuebo Song received the BS degree in electronic engineering and information science from the University of Science and Technology of China. He is currently working toward the PhD degree in computer science at Clemson University. His research interests include data mining, information retrieval, and storage systems.



Lin Li received the BS and MS degrees in computer science from the Beijing University of Posts and Telecommunications in China, and the PhD degree in computer science from Clemson University in May 2012. She is currently an assistant professor in the Department of Computer Science and Information Systems at Murray State University. Her research strives to bring principles and techniques from computer science to solve problems in other disciplines, with a particular interest in health informatics and bioinformatics.

Pradip K. Srimani received the BTech, MTech, and PhD degrees from the University of Calcutta, India, in 1973, 1975, and 1978, respectively. He is currently a professor of computer science at Clemson University, South Carolina. His research interests include heuristic search, distributed computing, mobile computing, and graph theory applications. He has published more than 250 papers in journals, conference proceedings, and books. He co-edited two books for the Computer Society Press. He has served on editorial boards and as special issue guest editor for a number of journals including IEEE Transactions and Magazines. He is a fellow of the IEEE and a distinguished scientist of the ACM.

Philip S. Yu received the BS degree in electrical engineering from National Taiwan University, the MS and PhD degrees in electrical engineering from Stanford University, and the MBA degree from New York University. He is a distinguished professor in computer science at the University of Illinois at Chicago and also holds the Wexler Chair in Information Technology. He spent most of his career at IBM, where he was the manager of the Software Tools and Techniques group at the Watson Research Center. His research interests include big data, including data mining, data stream, database, and privacy. He has published more than 780 papers in refereed journals and conferences. He holds or has applied for more than 250 US patents. He is the editor-in-chief of the *ACM Transactions on Knowledge Discovery from Data*. He is on the steering committee of the IEEE Conference on Data Mining and ACM Conference on Information and Knowledge Management and was a member of the IEEE Data Engineering Steering committee. He was the editor-in-chief of the *IEEE Transactions on Knowledge and Data Engineering* (2001-2004). He received the IEEE Computer Society 2013 Technical Achievement Award for *"pioneering and fundamentally innovative contributions to the scalable indexing, querying, searching, mining and anonymization of big data"*, the EDBT Test of Time Award (2014), and the Research Contributions Award from IEEE International Conference on Data Mining (2003). He had received several IBM honors including two IBM Outstanding Innovation Awards, an Outstanding Technical Achievement Award, two Research Division Awards, and the 94th plateau of Invention Achievement Awards. He was an IBM Master Inventor. He is a fellow of the ACM and the IEEE.



James Z. Wang received the BS and MS degrees from the University of Science and Technology of China and the PhD degree from the University of Central Florida, Orlando, both in computer science. He is currently a professor in the School of Computing, Clemson University, South Carolina. His research interests include multimedia systems, database, distributed computing, information retrieval, data mining, and bioinformatics. He has published more than 70 papers in international journals and conference proceedings. He is an associate editor of the *International Journal of Data Mining and Bioinformatics*. He has served as a paper reviewer, program committee member, session chair, or program vice chair for several international conferences. He is a senior member of the IEEE and the ACM.