Measure the Semantic Similarity of GO Terms Using Aggregate Information Content

Xuebo Song[1], Lin Li[1], Pradip K. Srimani[1], Philip S. Yu[2], and James Z. Wang[1,⋆]

[1] School of Computing, Clemson University, Clemson, SC 29634–0974
{xuebos,ll,srimani,jzwang}@clemson.edu
[2] Department of Computer Science, University of Illinois, Chicago, IL 60607
psyu@uic.edu

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. This paper proposes a novel and efficient method to measure the semantic similarity of GO terms. This method addresses the limitations in existing GO term similarity measurement methods by using the information content of all ancestor terms of a GO term to determine the GO term's semantic content. The aggregate information content 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 outperforms all existing methods in terms of the correlation with gene expression data.

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 [5–16]. Since different biomedical researchers may annotate the same or similar gene function with different but semantically similar GO terms based on

⋆ Corresponding author.

Z. Cai et al. (Eds.): ISBRA 2013, LNBI 7875, pp. 224–236, 2013.

their research findings, an accurate measure of semantic similarity of GO terms is critical to 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, 17–19], edge-based [10, 20, 21], and hybrid [6, 11] 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 of its ancestor terms (including the term itself). The first observation follows the human perception of term semantic similarity at different ontology levels. 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 based approach (AIC) in section 3. Section 4 provides details of experimental evaluation of AIC, while section 5 concludes the paper with a summary of unique charateristics of AIC.

2 Related Prior Work

A large number of studies [5–14] have appeared in the literature in the last 15 years to measure the semantic similarity of GO 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 [17–19] were originally designed to measure the semantic similarity of natural language terms. While each of them has its limitations they have been widely adopted by bioinformatics researchers to measure the semantic similarity of GO terms. In 2007, Wang [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. Besides, 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 [17], Lin's [18], Jiang and Conrath's [19], 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 [17] 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 is the principal reason that the semantic similarity obtained by these methods is inconsistent with human perception; they suffer from "shallow annotation" problem [8, 13, 6] in which the semantic similarity of GO terms near the root of the ontology are sometimes measured very high.

Edge-based approaches [10, 20, 21] are based on the length of graph paths connecting the terms being compared. Some edge-based approaches [20] treat all edges equally, ignoring the levels of edges in the ontology. This simple 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, 21] 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 suitable for genes of other species.

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 [17], Method B by Lin [18], Method C by Jiang and Conrath [19], 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 this paper in the next sections.

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

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

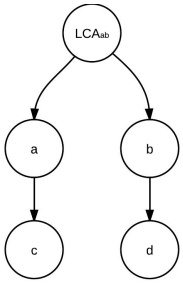

Fig. 1. GO terms at different ontology levels sharing the same LCA

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 $t \in T$, $p(t)$ denotes the probability that term t occurs in the GO database,

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

Information Content(IC) of term t is defined as

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

Method A uses Maximum Information Contained in Ancestors (MICA) of 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 Figure 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.

Methods B & C. Method B 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 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 Figure 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). Because methods B and C do not consider the specialization level of two terms' LCA in the semantic similarity measure, the semantic similarity values obtained by these two methods may still be inconsistent with the human perception as demonstrated in our experiment in Section 4.

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

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

where w_e is the semantic contribution factor of an edge. Then the semantic value (SV) of a GO term a is,

$$SV(a) = \sum_{t \in T_a} S_a(t) \tag{9}$$

where T_a is the set of GO terms in DAG_a (Directed Acyclic Graph consisting all the ancestors of the term including the term itself). 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 a semantic contribution factor value (weight) empirically obtained from gene classification to calculate the semantic values of GO terms. Using a semantic contribution factor 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 Proposed Aggregate Information Content Based Method (AIC)

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

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

We further propose a logarithmic model to normalize W(t) into a semantic weight $SW(t)$:

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

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

$$SV(a) = \sum_{t \in T_a} SW(t) \tag{13}$$

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

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

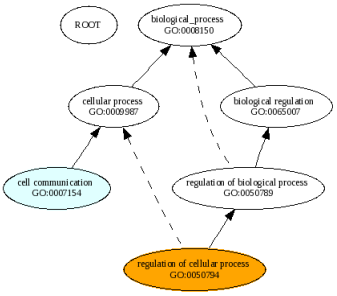

Fig. 2. GO Graph containing terms GO:0050794 and GO:0007154

Table 1. IC values & Semantic Weights of GO terms

Go Terms	IC value	SW value
0050794	1.2931	0.6842
0007154	2.0939	0.6172
0050789	1.1339	0.7072
0065007	1.0343	0.7245
0009987	0.4346	0.9090
0008150	0	1

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:0007154, shown in Figure 2.(All the similarity comparison figures showed in this paper are retrieved from the tools in [22].) First, we use the GOSim R package [23] to retrieve the IC information for all related GO terms, shown in Table 1. 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:0007154 as $sim_{GO}(0050794, 0007154) = 0.5828$.

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 [7, 5, 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, 24–26], we use gene expression data from Spellman dataset [27], which comprises of 6178 genes, to obtain the gene correlation patterns. The gene annotation data used to calculate the gene functional similarity is obtained from the GO database (2012-07). In the next two subsections, we provide comparison of our method (AIC) with the state-of-the-art current methods: Method A [17], Method B [18], Method C [19], and Method D [6] in terms of GO term semantic similarity and gene functional similarity.

4.1 Evaluating AIC Method Using GO Term Semantic Similarity

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 Figure 3 and Table 2.

Consider the two GO terms GO:0005739 and GO:0005777 as shown in Figure 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 A [17], Method B [18], and Method C [19] 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.

Now, we check whether all these semantic similarity measurement methods agree with the human perspective: 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, only methods A, D and our AIC method satisfy this property. The

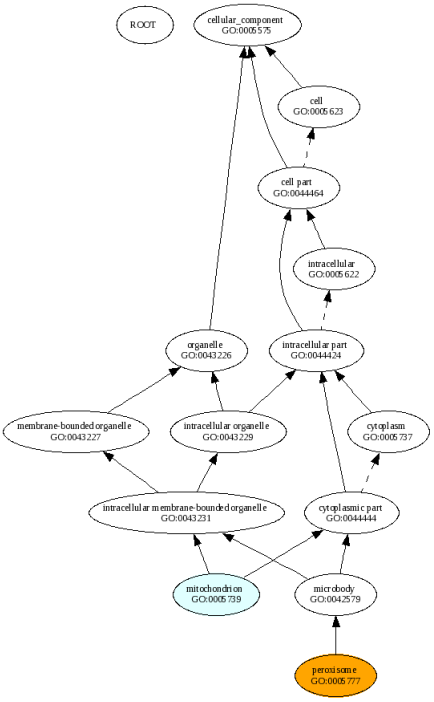

Fig. 3. GO graph of terms GO:0005739 and GO:0005777

Table 2. Semantic similarity values of GO term pairs obtained by different methods

Dataset	Method	**Similarity**
SW(GO:0005739, GO:0005777)	A	0.135
	B	0.335
	C	0.464
	D	0.797
	AIC	0.915
SW(GO:0044424, GO:0005622)	A	0.049
	B	0.948
	C	0.990
	D	0.845
	AIC	0.902
SW(GO:0044444, GO:0005737)	A	0.104
	B	0.872
	C	0.960
	D	0.879
	AIC	0.942

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 Evaluating AIC Using 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 dataset [27]. 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

Table 3. Pearson's correlation coefficients between gene expression data and gene functional similarities obtained by different semantic similarity measurement methods

Groups	Method B [18]	Method C [19]	Method A [17]	Method D [6]	**Proposed AIC**
4	0.789	0.930	0.614	0.929	**0.966**
5	0.717	**0.889**	0.561	0.802	0.850
6	0.569	0.700	0.413	0.745	**0.774**
7	0.622	**0.761**	0.519	0.725	0.733
8	0.597	0.675	0.496	0.706	**0.714**
9	0.659	0.664	0.417	0.745	**0.778**
10	0.620	0.730	0.403	0.733	**0.772**
11	0.665	0.691	0.419	0.725	**0.761**
12	0.485	0.722	0.246	0.716	**0.782**
13	0.525	0.715	0.321	0.709	**0.791**

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

gene expression correlation values between gene pairs, as in previous studies [13, 5, 7, 8], 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. 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 higher correlation coefficients between gene functional similarity and gene expression similarity.

4.3 Evaluating the Computation 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 is computationally expensive due to the recursive computation of semantic values of GO terms. On the other hand, our proposed AIC method uses the aggregate IC value, which can be precomputed, to represent the semantic value of a GO term. Thus, method AIC should be computationally more effective. 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 2000 respectively. The experiment was conducted on a Linux box with a i7-2600K CPU @ 3.40GHz, 8G memory. The execution time are shown in Table 4. As demonstrated by the experimental results, method AIC is considerably faster than method D.

5 Conclusion

Experimental results in Section 4 demonstrate the superiority of the proposed AIC method over the current ones. Method AIC is characterized with the following unique features:

- It does not suffer from "shallow annotation". 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.
- It exhibits high correlation coefficient between the gene expression similarity and the GO based functional similarity.
- It is computationally much faster than the popular hybrid method [6].

In summary, the proposed method AIC 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.

Acknowledgement. The work was partially supported by NSF Awards DBI-0960586, DBI-0960443 and CCF 0832582, and NIH award 1 R15 CA131808-01.

References

1. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nature Genetics 25, 25–29 (2000)
2. Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., Lewis, S.: The generic genome browser: A building block for a model organism system database. Genome Research 12, 1599–1610 (2002)
3. The UniProt Consortium. The uniprot consortium: The universal protein resource (uniprot). Nucleic Acids Research, pp. 190–195 (2008)
4. Kriventseva, E.V., Fleischmann, W., Zdobnov, E.M., Apweiler, R.: Clustr: a database of clusters of swiss-prot+trembl proteins. Nucleic Acids Research 29, 33–36 (2001)
5. Xu, T., Du, L., Zhou, Y.: Evaluation of go-based functional similarity measures using s.cerevisiae protein interaction and expression profile data. BMC Bioinformatics 9, 472 (2008)

6. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.-F.: A new method to measure the semantic similarity of go terms. Bioinformatics 23, 1274–1281 (2007)
7. Wang, H., Azuaje, F., Bodenreider, O., Dopazo, J.: Gene expression correlation and gene ontology-based similarity: An assessment of quantitative relationships. In: Proc. of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp. 25–31 (2004)
8. Sevilla, J.L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J.M., Martinez-Cruz, L.A., Corrales, F.J., Rubio, A.: Correlation between gene expression and go semantic similarity. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2, 330–338 (2005)
9. Schlicker, A., Domingues, F.S., Rahnenfuhrer, J., Lengauer, T.: A new measure for functional similarity functional similarity of gene products based on gene ontology. BMC Bioinformatics 7, 302 (2006)
10. Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., Kulp, D., Siani-Rose, M.A.: A knowledge-based clustering algorithm driven by gene ontology. Journal of Biopharmaceutical Statistics 14(3), 687–700 (2004)
11. Pesquita, C., Faria, D., Falcao, A.O., Lord, P., Couto, F.M.: Semantic similarity in biomedical ontologies. PLoS Computational Biology 5(7), e1000443 (2009)
12. Azuaje, F., Wang, H., Bodenreider, O.: Ontology-driven similarity approaches to supporting gene functional assessment. In: Proc. of the ISMB 2005 SIG Meeting on Bio-ontologies, pp. 9–10 (2005)
13. Li, B., Wang, J.Z., Luo, F., Feltus, F.A., Zhou, J.: Effectively integrating information content and structural relationship to improve the gene ontology similarity measure between proteins. In: The 2010 International Conference on Bioinformatics & Computational Biology (BioComp 2010), pp. 166–172 (2010)
14. Pesquita, C., Faria, D., Bastos, H., Falcao, A.O., Couto, F.M.: Evaluating go-based semantic similarity measures. In: Proc. of the 10th Annual Bio-Ontologies Meeting 2007, pp. 37–40 (2007)
15. Ravasi, T., et al.: An atlas of combinatorial transcriptional regulation in mouse and man. Cell 140(5), 744–752 (2010)
16. Washington, N.L., Haendel, M.A., Mungall, C.J., Ashburner, M., Westerfield, M., Lewis, S.E.: Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biology 7(11), e1000247 (2009)
17. Resnik, P.: Semantic similarity in taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11, 95–130 (1999)
18. Lin, D.: An information-theoretic definition of similarity. In: Proc. Int. Conf. on Machine Learning, pp. 296–304 (1998)
19. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. Int. Conf. on Research in Computational Linguistics, pp. 19–33 (1997)
20. Pekar, V., Staab, S.: Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. In: Proc. Int. Conf. on Computational Linguistics, vol. 2, pp. 786–792 (2002)
21. Wu, H., Su, Z., Mao, F., Olman, V., Xu, Y.: Prediction of functional modules based on comparative genome analysis and gene ontology application. Nucleic Acids Research 33(9), 2822–2837 (2005)
22. Du, Z., Li, L., Chen, C.-F., Yu, P.S., Wang, J.Z.: G-sesame: web tools for go-term-based gene similarity analysis and knowledge discovery. Nucleic Acids Research 37, W345–W349 (2009)

23. Froehlich, H., Speer, N., Poustka, A., Beissbarth, T.: Gosim - an r-package for computation of information theoretic go similarities between terms and gene products. BMC Bioinformatics 8, 166 (2007)
24. Heyer, L.J., Kruglyak, S., Yooseph, S.: Exploring expression data: Identification and analysis of coexpressed genes. Genome Research 9, 1106–1115 (1999)
25. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: A survey. IEEE Transactions on Knowledge and Data Engineering 16, 1370–1386 (2004)
26. Gibbons, F.D., Roth, F.P.: Judging the quality of gene expression-based clustering methods using gene annotation. Genome Research 12, 1574–1581 (2002)
27. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell 9, 3273–3297 (1998)