Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 891842, 9 pages
http://dx.doi.org/10.1155/2014/891842

Research Article

Correlating Information Contents of Gene Ontology Terms to Infer Semantic Similarity of Gene Products

Mingxin Gan

Dongling School of Economics and Management, University of Science and Technology Beijing, Beijing 100083, China

Correspondence should be addressed to Mingxin Gan; ganmx@ustb.edu.cn

Received 29 January 2014; Accepted 29 April 2014; Published 22 May 2014

Academic Editor: Huiru Zheng

Successful applications of the gene ontology to the inference of functional relationships between gene products in recent years have raised the need for computational methods to automatically calculate semantic similarity between gene products based on semantic similarity of gene ontology terms. Nevertheless, existing methods, though having been widely used in a variety of applications, may significantly overestimate semantic similarity between genes that are actually not functionally related, thereby yielding misleading results in applications. To overcome this limitation, we propose to represent a gene product as a vector that is composed of information contents of gene ontology terms annotated for the gene product, and we suggest calculating similarity between two gene products as the relatedness of their corresponding vectors using three measures: Pearson's correlation coefficient, cosine similarity, and the Jaccard index. We focus on the biological process domain of the gene ontology and annotations of yeast proteins to study the effectiveness of the proposed measures. Results show that semantic similarity scores calculated using the proposed measures are more consistent with known biological knowledge than those derived using a list of existing methods, suggesting the effectiveness of our method in characterizing functional relationships between gene products.

1. Introduction

Over the last few years, domain ontologies have been successfully applied to describe entities within a variety of biological domains, with examples including the derivation of functional relationships between gene products based on the gene ontology (GO) [1–3], the inference of phenotype similarity between human diseases based on the human phenotype ontology (HPO) [4, 5], the modeling of general computational tasks in systems biology based on the systems biology ontology (SBO) [6], and many others [7–9]. With an ontology to provide controlled and structured vocabularies in a specific biological domain and annotations to characterize entities in the domain with the vocabularies, relationships between the entities can be quantified by their semantic similarities in the ontology, thereby providing a convenient yet powerful means of profiling the entities and their semantic relationships [1]. Nevertheless, the automated derivation of semantic similarity between entities based on their annotations in a domain specific ontology still remains a great challenge, appealing for the development of effective and convenient computational methods [10].

In general, a domain ontology provides a set of controlled and relational vocabularies for describing domain specific knowledge. The vocabularies, also referred to as concepts or terms, are often organized as a directed acyclic graph (DAG), in which vertices denote terms and edges represent semantic relationships between the terms. It is also common that an ontology has more than one semantic relationship. For example, in the gene ontology, there are multiple types of semantic relationships such as "A is_a B" (any instance of A is also an instance of B) and "A part_of B" (an instance of A is a component of some instances of B) [1]. Given such a domain specific ontology and annotations that map entities onto the terms, most existing methods first calculate pairwise semantic similarity between the terms using the structure

of the ontology and annotations of entities and then derive similarity between the entities based on similarity between the terms [10–14].

Taking the gene ontology as an example, in order to achieve the former objective, Resnik proposed to use the information content (the negative logarithm of the relative frequency of occurrence of a term in annotations for a set of gene products) of the lowest common ancestor of two query terms to measure their semantic similarity [11]. Lin modified this measure by taking information contents of the query terms into consideration [12]. Schlicker et al. further incorporated the relative frequency of occurrence of the lowest common ancestor into the measure of Lin [14]. Jiang and Conrath proposed to incorporate the information contents of the query terms by using a formula different from that of Lin [13]. As another branch, Wang et al. proposed to calculate semantic similarity between GO terms using only the structural information of the underlying gene ontology, with the consideration of two types of semantic relationships: is_a and part_of [10].

With similarities between GO terms calculated, the semantic similarity between two query gene products was often calculated using a mean-max rule [10]. More specifically, given a single GO term and a collection of GO terms, the similarity between the term and the collection was defined as the maximum similarity between the term and every term in the collection. Furthermore, the similarity between two collections of GO terms was defined as the average of similarity between every term in a collection and the other collections. Finally, since a gene product was annotated by a collection of GO terms, semantic similarity between two gene products was defined as the similarity between the corresponding two sets of GO terms.

The above methods have been successfully applied to a variety of fields, with examples including the calculation of functional similarity between proteins based on the gene ontology (GO) for the inference of disease genes [2], the characterization of phenotype similarity between human diseases based on the human phenotype ontology (HPO) [5], and many others [7]. Software packages implementing these methods have also been released and publically available in the community of bioinformatics and computational biology, with examples including GOSemSim [15], FuSSiMeG [16], and OWLSim [4]. However, disadvantages of these methods are also obvious. For example, although methods such as those in [12–14] took efforts to modify the method of Resnik [11], their methods often performed worse than that of Resnik in real applications [10], suggesting that the revision of information contents can hardly be effective. Also, although Wang et al. systematically considered the structure and multiple semantic relationships of the gene ontology [10], they discarded the valuable resource of information contents of GO terms, resulting in a method performing worse than that of Resnik in many applications such as the prioritization of candidate genes [2]. In addition, as we shall see in the Results section, all of these methods tend to overestimate similarity between proteins that are actually not similar in their functions, thereby yielding misleading results in applications.

With these understandings, we propose in this paper to represent a gene product using a vector that is composed of information contents of GO terms annotated for the product in the gene ontology. Based on this notion, we suggest calculating semantic similarity between gene products as the relatedness of their corresponding vectors using three measures: Pearson's correlation coefficient, cosine similarity, and the Jaccard index. We focus on the biological process namespace of the gene ontology and annotations of proteins of the budding yeast *Saccharomyces cerevisiae* to perform a series of comprehensive studies on the effectiveness of the proposed measures. We calculate semantic similarity scores between yeast genes relying on the biological process domain of the gene ontology, use the resulting semantic similarity scores to measure functional relationships between the proteins, and study the consistency between such relationships and known biological knowledge. Results on 141 yeast biochemical pathways, 1,022 protein families, and two large-scale yeast protein-protein interaction networks show that semantic similarity scores calculated using the proposed measures are more consistent with biological knowledge than those derived using a list of existing methods, suggesting the effectiveness of our method in characterizing semantic similarity between gene products.

2. Methods

2.1. The Gene Ontology and Species Specific Annotations. The gene ontology (GO) provides a controlled vocabulary of terms for describing characteristics of gene products. This ontology covers three domains: biological process (BP), molecular function (MF), and cellular component (CC). The biological process domain defines operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of living cells, tissues, organs, and organisms. The molecular function domain represents the elemental activities of a gene product at the molecular level, such as binding or catalysis. The cellular component domain describes the parts of a cell or its extracellular environment [1]. Each of these three domains is organized according to a directed acyclic graph (DAG) structure, represented as $G = (V, E)$, where V is a set of vertices denoting concepts and E is a set of edges denoting semantic relationships between the terms. In such a graph, we use P_t and C_t to denote the sets of parents and children of term t, including t itself, respectively, and we use A_t and D_t to denote ancestors and descendants of term t, including t itself, respectively. Note that in the gene ontology, there are multiple types of semantic relationships such as "A is_a B" (any instance of A is also an instance of B) and "A part_of B" (an instance of A is a component of some instance of B).

A species specified annotation provides a mapping from a gene product of the species to a term in a domain (BP, MF, or CC) of the gene ontology. Following common specifications, the annotation of a gene product with term t implies the annotation of the gene product with all ancestors of t. With this notion, we represent annotations of gene product g using a binary annotation vector $\mathbf{a}_g = (a_{gi})_{|V|\times 1}$, where $a_{gi} = 1$ if g

is annotated by the term indexed by i or its descendants and $|V|$ the total number of terms in a domain.

3. Semantic Similarity as Correlation of Information Contents

Given a domain of the gene ontology and annotations for a set of gene products, the probability that a product annotated by term t or its descendants is estimated using the relative frequency of occurrence of term t and its descendants in the annotations is calculated by

$$\Pr(t) = \frac{1}{N}\sum_{i\in D_t} n_i, \tag{1}$$

where n_i is the number of annotations with term i and N the total number of annotations. The information content of term t is then calculated as

$$\mathrm{IC}(t) = -\log \Pr(t). \tag{2}$$

Moreover, information contents of all terms in the domain can be represented as a vector $\mathbf{q} = (q_i)_{|V|\times 1}$ with q_i being the information content of the term indexed by i. Calculating the Hadamard (entrywise) product of $\mathbf{a}_g$ and $\mathbf{q}$, we obtain the vector of information contents for gene product g as $\mathbf{x}_g = \mathbf{q} \circ \mathbf{a}_g = (x_{gi})_{|V|\times 1}$, where $x_{gi} = q_i \times a_{gi}$ for $i = 1, \ldots, |V|$. With such a vector calculated for every gene product, we propose the following three measures to quantify semantic similarity between two entities.

First, we propose to calculate the similarity as the absolute value of Pearson's correlation coefficient between the two vectors $\mathbf{x}_g$ and $\mathbf{x}_h$ for two gene products g and h as

$$S_{gh}^{(\mathrm{correlation})} = \left| \frac{\sum_{1\le i\le |V|} \left(x_{gi} - \overline{x}_g\right)\left(x_{hi} - \overline{x}_g\right)}{\sqrt{\sum_{1\le i\le |V|} \left(x_{gi} - \overline{x}_g\right)^2}\sqrt{\sum_{1\le i\le |V|} \left(x_{hi} - \overline{x}_h\right)^2}} \right|. \tag{3}$$

In this measure, we assume that information contents for the two gene products, $\mathbf{x}_g$ and $\mathbf{x}_h$, have a linear relationship, say,

$$\mathbf{x}_g = \alpha + \beta \mathbf{x}_h. \tag{4}$$

Hence, it is natural to use the coefficient of determination (r^2) that measures how good the observations fit this linear model to quantify the similarity between the two vectors. To ease the computation, we simply calculate the absolute value of the correlation coefficient instead of r^2. Note that exchanging $\mathbf{x}_g$ and $\mathbf{x}_h$ in the linear model yields the same r^2.

Second, we calculate the similarity as the cosine of the angle between the two vectors $\mathbf{x}_g$ and $\mathbf{x}_h$ for two gene products g and h as

$$S_{gh}^{(\mathrm{cosine})} = \frac{\sum_{1\le i\le |V|} x_{gi}x_{hi}}{\sqrt{\sum_{1\le i\le |V|} x_{gi}^2}\sqrt{\sum_{1\le i\le |V|} x_{hi}^2}}. \tag{5}$$

This is equivalent to calculating the uncentered correlation coefficient of the two vectors. It is evident that the cosine measure will yield similar results as those of the correlation measure when the means of $\mathbf{x}_g$ and $\mathbf{x}_h$ are small.

Third, we calculate the similarity as the Jaccard index of the two annotation vectors $\mathbf{a}_g$ and $\mathbf{a}_h$ for two gene products g and h as

$$S_{gh}^{(\mathrm{Jaccard})} = \frac{\sum_{1\le i\le |V|} \left(a_{gi} \wedge a_{hi}\right)}{\sum_{1\le i\le |V|} \left(a_{gi} \vee a_{hi}\right)}. \tag{6}$$

This is equivalent to calculating the ratio of the number of elements in the intersection and union of the two annotation sets for gene products g and h.

4. Existing Methods for Calculating Semantic Similarity

Most existing methods first derive similarity scores between terms and then calculate semantic similarity scores between gene products as similarity scores between collections of annotated terms for the products. More precisely, there have been two main categories of methods for calculating pairwise concept similarity scores: (1) approaches based on information contents of terms in the gene ontology and (2) methods based on the structure of the gent ontology.

The first group of approaches calculates similarity between two terms u and v relying on the information content of the most specific term m_{uv} in their common ancestors. Generally, a term with more specific meaning tends to have a higher information content and hence

$$m_{uv} = \arg \max_{w\in A_u\cap A_v} \mathrm{IC}(w). \tag{7}$$

With this notion, Resnik [11] defined the similarity between u and v as

$$T_{uv}^{(\mathrm{Resnik})} = \mathrm{IC}(m_{uv}) = -\log \Pr(m_{uv}). \tag{8}$$

Lin [12] defined the similarity as

$$T_{uv}^{(\mathrm{Lin})} = \frac{2\log \Pr(m_{uv})}{\log \Pr(u) + \log \Pr(u)}. \tag{9}$$

Schlicker et al. [14] define the similarity as

$$T_{uv}^{(\mathrm{Schlicker})} = \frac{2\log \Pr(m_{uv})}{\log \Pr(u) + \log \Pr(u)}\left(1 - \Pr(m_{uv})\right). \tag{10}$$

Jiang and Conrath [13] define the dissimilarity between two terms as

$$D_{uv}^{(\mathrm{Jiang})} = \log \Pr(u) + \log \Pr(u) - 2\log \Pr(m_{uv}). \tag{11}$$

This is equivalent to defining its reciprocal as the similarity as

$$T_{uv}^{(\mathrm{Jiang})} = \frac{1}{\log \Pr(u) + \log \Pr(u) - 2\log \Pr(m_{uv})}. \tag{12}$$

The second group of approaches calculates similarity between GO terms depending on the structure of the gene ontology. Briefly, given a term indexed by t, Wang et al. iteratively calculate an s-value for every ancestor $a \in A_t$ to measure the contribution of a to the semantic of t as

$$s_t(a) = \begin{cases} 1 & \text{if } a = t, \\ \max_{x \in C_a} w_e s_t(x) & \text{if } a \neq t, \end{cases} \quad (13)$$

where the weight $w_e = 0.8$ if x and t have the is_a relationship and $w_e = 0.6$ if x and t have the part_of relationship [10]. Then, a semantic value for term t is defined as $s(t) = \Sigma_{x \in A_t} s_t(x)$. Finally, the semantic similarity score between two terms u and v is defined as

$$T_{uv}^{(\text{Wang})} = \sum_{x \in A_u \cap A_v} \frac{s_u(x) + s_v(x)}{s(u) + s(v)}. \quad (14)$$

With pairwise semantic similarity scores between GO terms being ready, the similarity between term t and a set of terms T is defined as

$$\text{Sim}(t, T) = \max_{t' \in T} T_{tt'}, \quad (15)$$

where $T_{tt'}$ is calculated using either of the above methods. The similarity between two sets of terms S and T can then be calculated as

$$\text{Sim}(S, T) = \frac{1}{|S| + |T|} \left(\sum_{s \in S} \text{Sim}(s, T) + \sum_{t \in T} \text{Sim}(t, S) \right). \quad (16)$$

Finally, for two gene products g and h annotated by two sets of terms G and H, respectively, the semantic similarity between the two objects is then defined as

$$S_{gh} = \text{Sim}(G, H). \quad (17)$$

5. Results

5.1. Data Sources. There have been quite a few domain specific ontologies available for characterizing entities in a variety of biological domains. Particularly, the OBO (open biological and biomedical ontologies) Foundry has released eight ontologies to provide standard descriptions of entities in biological domains [14]. Among these ontologies, biological process (BP), molecular function (MF), and cellular component (CC) are typically referred to as the gene ontology (GO), which has been widely used to describe functions of genes. The gene ontology also provides annotations of gene products for several well-studied model organisms, including yeast, fruit fly, and mouse [1]. In this paper, we focus on the biological process domain of GO and annotations of the budding yeast *Saccharomyces cerevisiae* to validate the effectiveness of the proposed measures. We extract 22,688 terms from the biological process domain of the gene ontology (released on April 27, 2012) and obtain 22,798 annotations of 6,383 yeast genes (released on April 28, 2012).

5.2. Distribution of Semantic Similarity Scores of Random Gene Pairs. It is evident that a pair of genes selected at random can hardly have similar functions, and thus the semantic similarity score between such a pair of genes should be close to zero. To validate this argument, we calculate semantic similarity scores of 100,000 pairs of yeast genes selected at random, and we summarize the distribution of the scores in Figure 1. We can clearly see from the figure that the median similarity score of the correlation measure (0.004894) is almost 0 so is that of the cosine measure (0.003196). The median similarity score of the Jaccard measure (0.03846) is higher than those for both the correlation and the cosine measures but still lower than those for all the five existing methods. The method of Resnik generates the smallest median similarity score (0.04395) among the existing methods, followed by the methods of Schlicker et al. (0.04810), Lin (0.09115), and Wang et al. (0.2138). The method of Jiang et al. generates the largest median similarity score (0.3460). From these observations, we conclude that the existing methods tend to overestimate semantic similarity between genes that are actually not related in their functions. On the other hand, the proposed measures, though much simpler than the existing methods, do not have such a drawback and thus yield much more reasonable results in assessing semantic similarity between randomly selected gene pairs.

5.3. Consistency between Gene Semantic Similarity and Pathway Data. It is known that most biological functions rise from collaborative effects of several proteins that usually involve in the same biological process and form a pathway [17]. Hence, gene products (proteins) in the same pathway should have similar annotations in the biological process ontology and in turn own high semantic similarity scores according this ontology. On the contrary, gene products belonging to different pathways should own relatively low semantic similarity scores. To assess whether the proposed similarity measures are consistent with this knowledge, we compare semantic similarity scores between proteins within a pathway and those between proteins involved in different pathways as follows.

We download from the *Saccharomyces* Genome database (SGD) [18] 141 pathways, each including at least two proteins. For each of these pathways, we calculate pairwise semantic similarity scores of proteins involved in the pathway, and we average these scores over all pairs of proteins to obtain the mean semantic similarity score within the pathway (μ_{in}). Meanwhile, for each pathway, we further select at random 10 times the number of proteins as those in the pathway, calculate semantic similarity scores between these proteins and those in the pathway, and average over these scores to obtain the mean semantic similarity score outside the pathway (μ_{out}). Then, we plot the distribution of mean similarity scores within and outside all pathways in Figure 2. From the figure, we observe that the mean similarity scores within pathways are in general large, while those outside pathways are typically small. Particularly, for all of the three proposed measures (correlation, cosine, and the Jaccard), the differences between the medians of the mean similarity

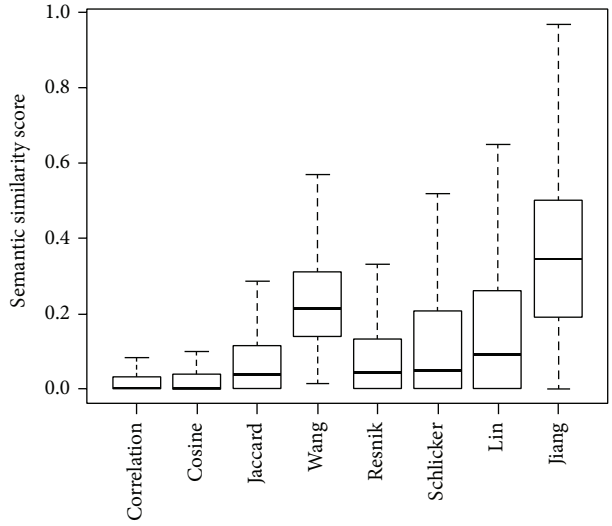

Figure 1: Distributions of semantic similarity scores of 100,000 randomly selected pairs of yeast genes.

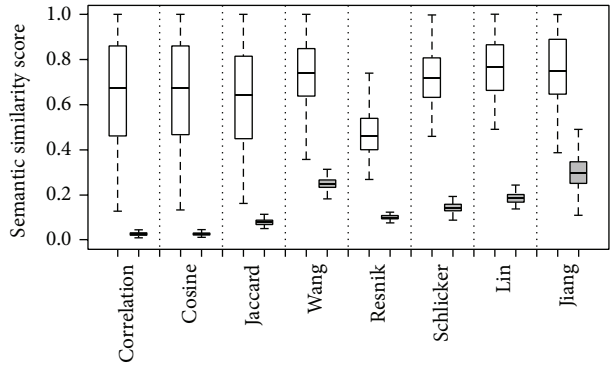

Figure 2: Distributions of mean semantic similarity scores within pathways (white) and outside pathways (gray).

scores within and outside pathways are much more obvious than those of the five existing methods. For example, using the correlation measure, we obtain the median μ_{in} over all pathways as 0.6578 and the median μ_{out} as 0.02564. Using the cosine measure, we obtain a median μ_{in} of 0.6600 and a median μ_{out} of 0.02733. In contrast, the method of Wang produces a median μ_{in} of 0.7405 and a median μ_{out} of 0.2489, and the method of Resnik produces a median μ_{in} of 0.4662 and a median μ_{out} of 0.09956.

We further calculate for each pathway the ratio of the mean semantic similarity scores within the pathway over that outside the pathway ($\mu_{\text{in}}/\mu_{\text{out}}$), and we average such ratios over all 141 pathways to obtain a criterion called fold change of semantic similarity scores within pathways against those outside pathways. We summarize the fold changes in Figure 3, from which we can clearly see the effectiveness of the proposed measures. For example, using the correlation measure, we obtain a fold enhancement of 29.93. Using the cosine measure, we obtain a fold change of 26.65. In contrast, the method of Wang only produces a fold change of 3.03, and

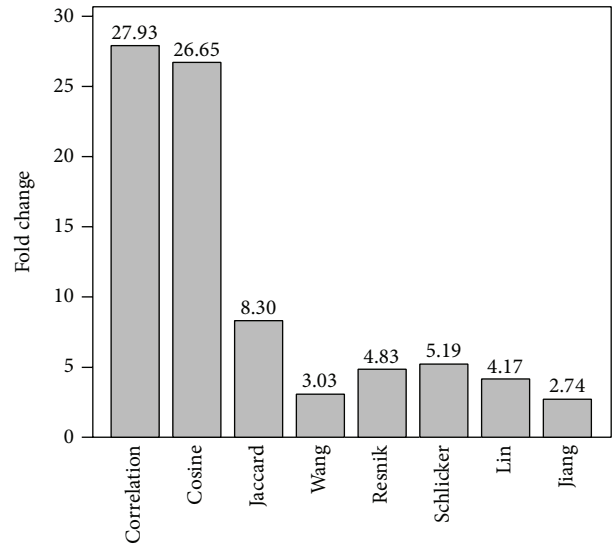

Figure 3: Fold change of semantic similarity scores within pathways against those outside pathways.

the method of Resnik produces a slightly larger fold change of 4.83.

These observations support the conclusion that the proposed measures yield much more reasonable results in assessing functional relationships between proteins within pathways, and thus these measures are more consistent with biological knowledge than existing methods.

5.4. Consistency between Gene Semantic Similarity and Protein Domain Data. Proteins are often composed of one or more functional regions, commonly referred to as protein domains [19]. Different domains typically account for different functions of proteins containing them, and thus different combinations of protein domains give rise to the diverse range of proteins found in nature. Hence, proteins can be classified into different families according to the domains that the proteins contain. Moreover, proteins containing the same domain, or say belonging to the same family, should have some similar functions and thus share some similar annotations in the biological domain of the gene ontology. Consequently, proteins belonging to the same family should have high semantic similarity scores according to the gene ontology. On the contrary, proteins belonging to different familiess should own relatively low semantic similarity scores. To assess whether the proposed similarity measures are consistent with this knowledge, we compare semantic similarity scores between proteins within a protein family and those between proteins belonging to different families as follows.

The Pfam database [20] provides a large collection of both high quality protein families (Pfam-A) and low quality protein families (Pfam-B). In version 26.0 of the Pfam-A collection (released in November 2011), 13,672 protein families are collected. From this data source, we extract 1,022 protein families, each including at least two yeast proteins. For each of these families, we calculate pairwise semantic similarity scores of proteins belonging to the family, and

we average these scores over all pairs of proteins to obtain the mean semantic similarity score within the family (ν_{in}). Meanwhile, for each protein family, we further select at random 10 times the number of proteins as those in the family, calculate semantic similarity scores between these proteins and those belonging to the family, and average over these scores to obtain the mean semantic similarity score outside the family (ν_{out}). Then, we calculate for each protein family the ratio of the mean semantic similarity scores within the family over that outside the family (ν_{in}/ν_{out}), and we average such ratios over all 1,022 protein families to obtain a criterion called fold change of semantic similarity scores within protein families against those outside families. We summarize the fold changes in Figure 4, from which we can clearly see the effectiveness of the proposed measures. For example, using the correlation measure, we obtain a fold change of 6.915. Using the cosine measure, we obtain a fold change of 6.511. Using the Jaccard measure, we obtain a fold change of 3.267. In contrast, the method of Wang only produces a fold change of 1.856, and the method of Resnik produces a slightly larger fold change of 2.370.

We further change the minimum number proteins belonging to a protein family from 2 to 10, calculate the fold change in each situation, and present the results in Table 1. Briefly, the fold change varies with the minimum number of proteins in a protein family, but the observation that the fold changes of the proposed measures are greater than those of the existing methods remains unchanged. For example, when considering protein families containing at least 10 proteins, we obtain fold changes of 9.273, 9.814, and 4.516 for the correlation, cosine, and the Jaccard measures, respectively. In contrast, the fold change for the measures of Wang, Resnik, and Schlicker are 2.090, 2.846, and 3.430, respectively. From these results, we make the conjecture that the proposed measures yield much more reasonable results in assessing functional relationships between proteins that belong to the same protein family. Hence, we conclude that the proposed measures are more consistent with biological knowledgethan existing methods.

5.5. Consistency between Gene Semantic Similarity and PPI Data. Biological knowledge suggests that proteins often interact with each other in the collaborative generation of biological functions [21]. The collection of all physical interactions in a living organism is typically referred to as the protein-protein interaction (PPI) network, in which nodes are proteins and edges are physical interactions between the proteins. Interacting proteins are usually involved in similar biological process and thus have similar annotations in the biological process domain of the gene ontology and high semantic similarity scores. To assess whether our similarity measures are consistent with this knowledge, we assess relationships between interacting proteins and their semantic similarity scores as follows.

We download two manually curated PPI networks of *Saccharomyces cerevisiae*. From BioGrid (biological generic repository for interaction datasets) [22, 23], we extract a PPI network composed of 3,529 nodes and 16,285 edges. From

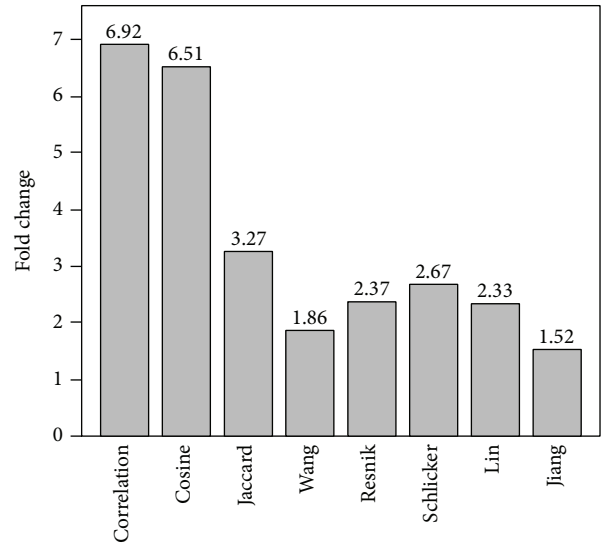

Figure 4: Fold change of semantic similarity scores within protein families against those outside protein families.

DIP (database of interacting proteins) [24, 25], we extract a relative small PPI network including 2,902 nodes and 7,005 edges. For each of these networks, we calculate semantic similarity scores for interacting proteins and those for the same number of randomly selected noninteracting pairs of proteins, and we plot the distribution of these scores in Figures 5(a) and 5(b). From the figure, we obviously see that the semantic similarity scores for interacting proteins are in general larger than those for noninteracting proteins, and this observation exists for both the BioGrid and the DIP networks.

Then, for each of these networks, we average over semantic similarity scores between interacting proteins to obtain the mean semantic similarity score of interacting proteins (τ_{int}). Meanwhile, we average over semantic similarity scores of noninteracting pairs of proteins to obtain the mean semantic similarity score of noninteracting proteins (τ_{non}). Finally, we calculate the fold change as τ_{int}/τ_{non} to measure the effectiveness of a method in distinguishing the functional relationship between interacting proteins. We present the results summarized in Figure 6, from which we can see the effectiveness of the proposed measures. For example, for the BioGrid network, we obtain a fold change of 6.15 when using the correlation measure. For the DIP network, the fold change is 5.44 for the correlation measure. For the cosine and the Jaccard measures, we observe similar results. From these observations, we make the conjecture that the semantic similarity scores calculated by the proposed measures are consistent with biological knowledge about interacting proteins.

It has also been shown that proteins closer in a PPI network tend to have more similar functions [4]. With this understanding, we use the length of the shortest path between two proteins in a PPI network to measure the network proximity of the proteins, use the semantic similarity score of the two proteins to measure their functional similarity, and

Table 1: Fold changes of semantic similarity scores within protein families against those outside families.

m	n	Semantic similarity measures							
		Correlation	Cosine	Jaccard	Wang	Resnik	Schlicker	Lin	Jiang
2	1022	6.915	6.511	3.267	1.856	2.370	2.669	2.331	1.524
3	562	8.986	8.446	3.827	1.988	2.680	3.100	2.641	1.629
4	360	9.608	8.760	4.027	2.037	2.799	3.247	2.761	1.656
5	240	9.359	9.135	4.131	2.065	2.843	3.324	2.827	1.662
6	182	9.997	9.214	4.224	2.105	2.901	3.410	2.888	1.692
7	141	10.10	9.741	4.363	2.106	2.952	3.476	2.918	1.690
8	110	9.921	9.409	4.432	2.101	2.853	3.409	2.895	1.661
9	89	9.880	9.321	4.445	2.094	2.857	3.419	2.908	1.643
10	75	9.814	9.273	4.516	2.090	2.846	3.430	2.898	1.644

m: minimum number of proteins in a family. n: number of protein families, each containing at least m proteins.

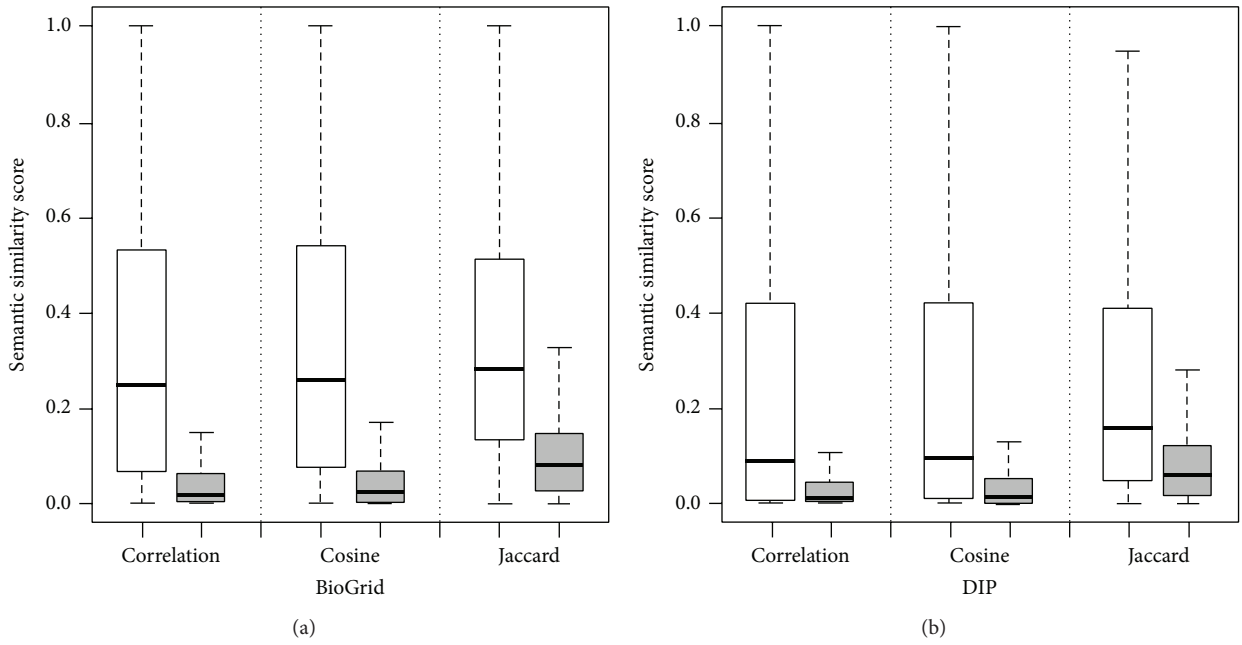

Figure 5: Relationships between semantic similarity scores and protein-protein interaction data. (a) Distributions of similarity scores of interacting proteins (white) against noninteracting proteins (gray) for the BioGrid dataset. (b) Distributions of similarity scores of interacting proteins (white) against noninteracting proteins (gray) for the DIP dataset.

plot the change of the similarity score with the closeness of proteins in Figure 6. From the figure, we can see that protein pairs tend to have higher semantic similarity scores if they are closer in the PPI network. For example, for the BioGrid network and the cosine measure, the median semantic similarity score is 0.2590 for direct interacting protein pairs, 0.0720 for protein pairs intermediated by another protein, 0.0372 for protein pairs intermediated by two other proteins, and so forth. Similar results are observed for the other two measures. These results suggest that protein similarity scores are correlated with protein closeness in a PPI network, again consistent with biological knowledge.

6. Conclusions and Discussion

In this paper, we have proposed an approach to represent annotations of a gene product in the gene ontology using vectors that are composed of information contents of terms in the ontology. Based on this notion, we have proposed to calculate pairwise semantic similarity between gene products by using three measures (Pearson's correlation coefficient, cosine similarity, and the Jaccard index) to quantify the relatedness of the corresponding vectors. We have performed a series of comprehensive studies on the effectiveness of the proposed measures using the ontology of biological process and annotations of the budding yeast

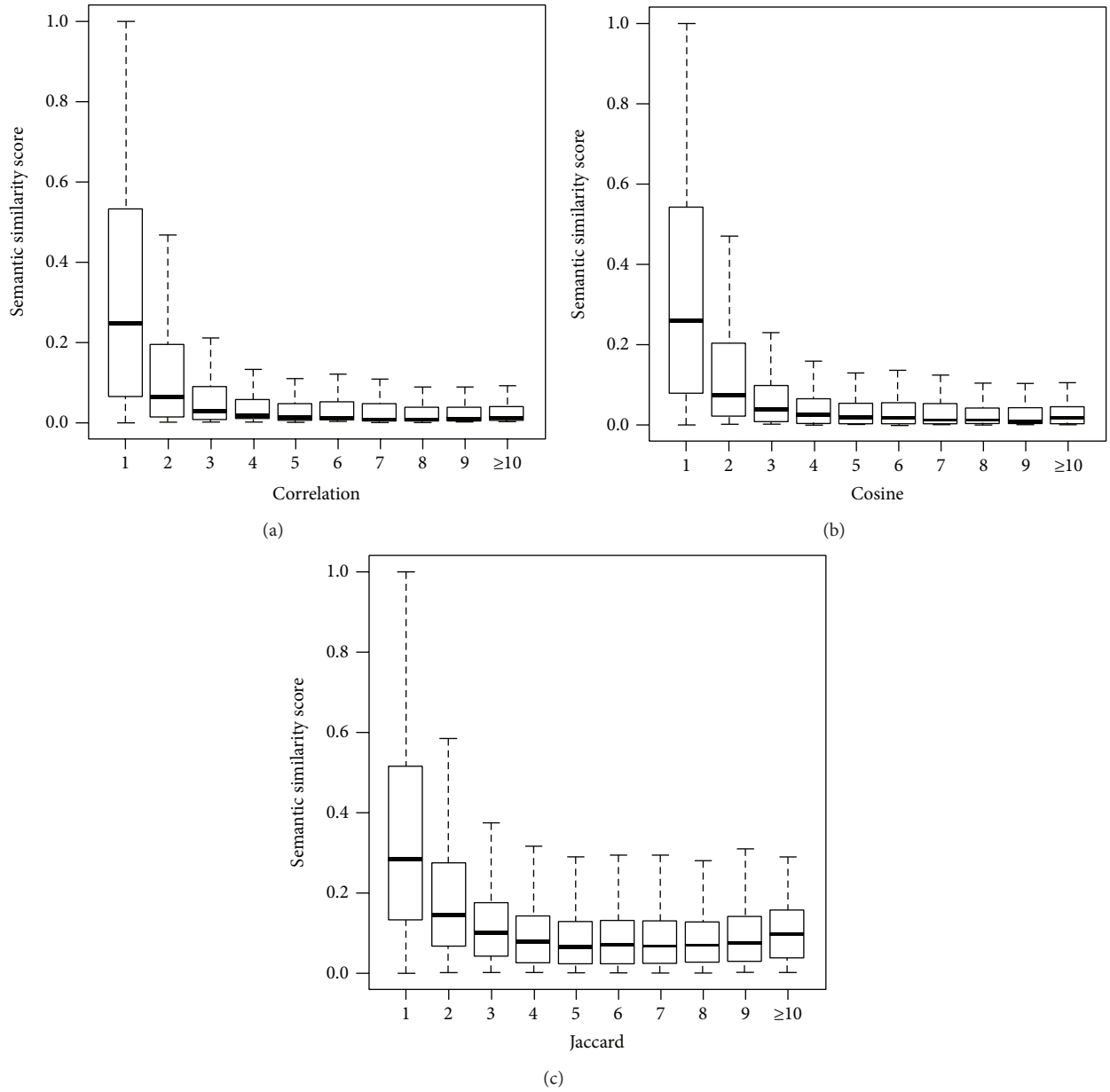

Figure 6: Distributions of semantic similarity scores against the shortest path distance of interacting proteins for the BioGrid dataset. (a) Results for the measure of correlation. (b) Results for the measure of cosine. (c) Results for the measure of Jaccard.

Saccharomyces cerevisiae. Comprehensive studies on the relationships between semantic similarity of gene products and biochemical pathways, protein families, and protein-protein interaction networks show that semantic similarity scores calculated using the proposed measures are more consistent with biological knowledge than those derived using a list of five existing methods, suggesting the effectiveness of our method in characterizing functional similarity between gene products based on the gene ontology.

The main advantage of the proposed measures is the simplicity in calculation and the effectiveness in characterizing semantic similarity between gene products. The representation of gene products as vectors of information contents of ontology terms is straightforward, making the followed computation easy to understand. The simplicity in presentation also benefits the computation with a low time complexity, thereby making our method suitable for large scale calculation of semantic similarity for not only applications based on the gene ontology but also those using other ontologies.

Certainly, the proposed measures can be further improved from the following aspects. First, although the contribution of a term in a domain ontology has been characterized by its information content, it is possible to further refine such contribution by adjusting the information contents with prior knowledge. For example, it is not hard

to combine annotations of different organisms to achieve a more precise estimation of information contents for concepts in the gene ontology. Another possibility is to develop a Bayesian method to estimate the information contents, using existing annotations to derive the prior distribution.

Second, although the presentation of domain entities as vectors of concepts is simple yet effective, the incorporation of the structure of the concepts in the underlying ontology may further improve the performance of the proposed method. Existing algorithms for calculating similarity between two tree structures [26] might be a potential candidate along this direction.

Conflict of Interests

The author does not have any conflict of interests.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (no. 71101010).

References

[1] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[2] R. Jiang, M. Gan, and P. He, "Constructing a gene semantic similarity network for the inference of disease genes," *BMC Systems Biology*, vol. 5, supplement 2, article S2, 2011.

[3] J. Wu, Y. Li, and R. Jiang, "Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies," *PLoS Genetics*, vol. 10, no. 3, Article ID e1004237, 2014.

[4] 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, Article ID e1000247, 2009.

[5] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The human phenotype ontology: a tool for annotating and analyzing human hereditary disease," *American Journal of Human Genetics*, vol. 83, no. 5, pp. 610–615, 2008.

[6] M. Courtot, N. Juty, C. Knupfer et al., "Controlled vocabularies and semantics in systems biology," *Molecular Systems Biology*, vol. 7, p. 543, 2011.

[7] M. Gan, X. Dou, and R. Jiang, "From ontology to semantic similarity: calculation of ontology-based semantic similarity," *The Scientific World Journal*, vol. 2013, Article ID 793091, 11 pages, 2013.

[8] Y. Chen, J. Hao, W. Jiang et al., "Identifying potential cancer driver genes by genomic data integration," *Scientific Reports*, vol. 3, article 3538, 2013.

[9] Y. Chen, X. Wu, and R. Jiang, "Integrating human omics data to prioritize candidate genes," *BMC Medical Genomics*, vol. 6, no. 1, article 57, 2013.

[10] 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, no. 10, pp. 1274–1281, 2007.

[11] P. Resnik, "Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language," *Journal of Artificial Intelligence Research*, vol. 11, pp. 95–130, 1999.

[12] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the 15th International Conference on Machine Learning*, pp. 296–304, Morgan Kaufmann, 1998.

[13] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of the International Conference on Research in Computational Linguistics*, pp. 19–33, 1997.

[14] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer, "A new measure for functional similarity of gene products based on gene ontology," *BMC Bioinformatics*, vol. 7, article 302, 2006.

[15] 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, Article ID btq064, pp. 976–978, 2010.

[16] F. M. Couto, M. J. Silva, and P. M. Coutinho, "Measuring semantic similarity between Gene Ontology terms," *Data and Knowledge Engineering*, vol. 61, no. 1, pp. 137–152, 2007.

[17] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[18] J. M. Cherry, E. L. Hong, C. Amundsen et al., "Saccharomyces genome database: the genomics resource of budding yeast," *Nucleic Acids Research*, vol. 40, pp. D700–D705, 2012.

[19] A. Bateman, L. Coin, R. Durbin et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 32, pp. D138–D141, 2004.

[20] M. Punta, P. C. Coggill, R. Y. Eberhardt et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 40, pp. D290–D301, 2012.

[21] R. Jiang, Z. Tu, T. Chen, and F. Sun, "Network motif identification in stochastic networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 25, pp. 9404–9409, 2006.

[22] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri et al., "The BioGRID interaction database: 2011 update," *Nucleic Acids Research*, vol. 39, no. 1, pp. D698–D704, 2011.

[23] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, pp. D535–D539, 2006.

[24] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic Acids Research*, vol. 32, pp. D449–D451, 2004.

[25] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "DIP: the database of interacting proteins," *Nucleic Acids Research*, vol. 28, no. 1, pp. 289–291, 2000.

[26] Y. Zhong, C. A. Meacham, and S. Pramanik, "A general method for tree-comparison based on subtree similarity and its use in a taxonomic database," *Biosystems*, vol. 42, no. 1, pp. 1–8, 1997.