

# A Study of Correlations Between the Definition and Application of the Gene Ontology

Yuji Mo, Catherine Anderson and Stephen D. Scott  
Dept. of Computer Science  
University of Nebraska  
Lincoln, NE 68588-0115  
{ymo, anderson, sscott}@cse.unl.edu

## Abstract

When using the Gene Ontology (GO), nucleotide and amino acid sequences are annotated by terms in a structured and controlled vocabulary organized into a relational graph. The usage of the vocabulary (GO terms) in the annotation of these sequences may diverge from the relations defined in the ontology. We measure the consistency of the use of GO terms by comparing GO's defined structure to the terms' application. To do this, we first use synthetic data with different characteristics to understand how these characteristics influence the correlation values determined by various similarity measures. Using these results as a baseline, we found that the correlation between GO's definition and its application to real data is relatively low, suggesting that GO annotations might not be applied in a manner consistent with its definition. In contrast, we found a sub-ontology of GO that correlates well with its usage in UniProtKB.

## 1. Introduction

The Gene Ontology (GO) [1] is a controlled vocabulary describing the domain of gene products, i.e., enzymes and other proteins encoded in DNA. GO is made up of three independent, orthogonal ontologies: (1) the Cellular Component ontology, which describes where a gene product is located at a subcellular level; (2) the Molecular Function ontology, which describes the function a gene product can perform; and (3) the Biological Process ontology, which describes series of events and molecular functions. Each ontology is structured as a directed acyclic graph (DAG). Each node of each DAG is a term with a distinct name and description. The edges of a DAG represent the relations between the connected nodes. The relations are endowed with descriptive logic so that inferences can be made between parent and child nodes. A gene product can be annotated by assigning GO terms to the description of the gene product. This assignment is also referred to as an *association* between a term and a gene product.

GO has become widely accepted in the genomics community as a concise means of annotating gene products for machine translation [2]. However, due to the wide scope of the genomics community, ambiguities in term usage exist.

The GO project is a collaborative effort between groups sharing their vocabularies. Group members participate on a self-interested, best-effort basis to reach consensus on the addition, deletion or editing of terms within the three ontologies. However, individual curators from different communities may interpret the definitions differently, resulting in inconsistent usage, and thus it is necessary to continually refine terms. With the large increase of gene products that are annotated with GO, methods to evaluate semantic similarity based on annotations are critical in evaluating the consistency of usage. This motivates our study, which is to apply measures of *semantic similarity* to estimate the consistency between how GO is defined and how it is used in practice.

The notion of semantic similarity is frequently used in information retrieval, where terms are indexed by similar meaning rather than similar words. This concept was used in early research with natural language processing techniques: associating descriptive language with terms and quantifying this similarity. The ontology terms in GO may be examined by clustering terms together with similar semantics [3] using these techniques.

Earlier work done [4], [5] to determine semantic similarity of terms using the annotation they have been associated with were designed for specific applications: malapropism correction (the correction of outliers in the annotation), assessing functional similarity of gene products [6], predicting protein interaction [7], assessing the influence of electronic annotations [8] and assisting in the annotation of new sequences [4]. In contrast, we use some of the same measures they do, but for the purposes of measuring the consistency of the use of GO.

All three ontologies within GO contain many biologically/biochemically descriptive terms that have not been used (not applied to any annotation). A large number of terms are used only once or not at all. This creates a usage pattern where a large percent of GO terms fall in the tail of the distribution, (called the *long tail phenomenon*). Because of this phenomenon, certain types of similarity measures may be preferable to others in evaluating ontology usage. Thus, one of our results is a test using synthetic data with different characteristics to understand how various similarity measures

In Proceedings of BIOCOMP '11: The 2011 International Conference on Bioinformatics and Computational Biology, to appear.

measure correlation, and how these measures are influenced by various properties of the data. We then describe how the synthetic data parameters imply properties of real data. Our results show that one measure (called ‘‘Cosine’’) is only useful in recognizing correlations when the gene product usage comes with a long tail and each term is annotated by many moderately concentrated terms in the ontology. Another measure (‘‘Jiang’s’’) is not well suited for unbalanced usage of terms in the ontology. The remaining measures (‘‘Resnik’s,’’ ‘‘Lin’s,’’ and ‘‘Rel’’) are almost independent of the data characteristics that we varied, especially Resnik’s.

Using our results on synthetic data as a baseline, we then sampled partial ontologies from GO and measured correlations between their definitions and their usage. Relative to correlation results found in synthetic data with similar configurations to the real data, we found that the average correlation is low. This might suggest that GO annotations are not applied in a manner consistent with their definition. In contrast, we found that the sub-ontology rooted at the term ‘‘GO:0005275: amine transmembrane transporter activity’’ correlates well with its usage in UniProtKB.

## 2. Method

### 2.1 Problem Formalization

An ontology  $G = (V, E)$  is a directed acyclic graph (DAG), where each vertex corresponds to a term  $c_i$ . There is an edge from  $c_i$  to  $c_j$  if and only if  $c_j$  is explicitly a  $c_i$ . Since this ‘‘is\_a’’ relation is transitive,  $c_j$  is\_a  $c_i$  if and only if there is a path from  $c_i$  to  $c_j$ . We consider  $c_j$  to be a descendant of  $c_i$  if a path from  $c_i$  to  $c_j$  exists.

According to the gene product annotation guidelines [9], a gene product can be annotated by zero or more nodes of each ontology. Let  $C_i$  be the set of terms used to annotate gene product  $e_i$ . Similarly, we can define  $E_j$  as the set of gene products annotated by term  $c_j$ . By definition,  $c_j \in C_i \Leftrightarrow e_i \in E_j$ . In addition, annotating a gene product with a term implies that the gene product is also annotated by all ancestors of the term. Thus,  $c_i$  is a descendant of  $c_j$  implies  $E_i \subseteq E_j$ . The ancestor term inherits all annotations from its descendant, so the root term has all annotations:  $E_{root} = \bigcup_i E_i$ .

### 2.2 Similarity Measures

There are many different functions for calculating semantic similarity between terms. We consider the following five measures.

Resnik [10] proposed that the amount of information provided by the common ancestors of the two terms may be used as a measure:

$$Sim_{Resnik}(c_i, c_j) = \max_{c_k \in S(c_i, c_j)} -\log P(c_k) , \quad (1)$$

where  $S(c_i, c_j)$  is the set of ancestors shared by both  $c_i$  and  $c_j$  and  $P(c_k)$  is the probability that a randomly selected gene product is annotated by term  $c_k$ :  $P(c_k) = |E_k|/|E_{root}|$ .

Lin [11] extended Resnik’s measure by modifying the information content of a term to take both descendants into consideration:

$$Sim_{Lin}(c_i, c_j) = \max_{c_k \in S(c_i, c_j)} \left( \frac{2 \log P(c_k)}{\log P(c_i) + \log P(c_j)} \right) . \quad (2)$$

Generic terms do not have a high relevance for the comparison of different gene products. Andreas’s [5] relevance measure combined both Lin’s and Resnik’s measure by weighting Lin’s similarity measure with  $1 - P(c_k)$ . For a detailed term  $c_k$ ,  $P(c_k)$  becomes relatively very small and makes  $1 - P(c_k)$  close to 1 and negligible:

$$Sim_{Rel}(c_i, c_j) = \max_{c_k \in S(c_i, c_j)} \left( \frac{2(1 - P(c_k)) \log P(c_k)}{\log P(c_i) + \log P(c_j)} \right) . \quad (3)$$

Jiang [12] proposed a similarity measure as the reciprocal of semantic distance:

$$Sim_{Jiang}(c_i, c_j) = \max_{c_k \in S(c_i, c_j)} \left( \frac{1}{-\log P(c_i) - \log P(c_j) + 2 \log P(c_k)} \right) . \quad (4)$$

The Cosine similarity [13] is a measure frequently used in data mining. It is defined as the cosine of the angle between two vectors in a hyperspace. We model each term  $c_i$  as a vector  $v_i = (v_{i1}, v_{i2}, \dots, v_{im})$ , in which  $v_{ij} = 1$  if  $c_i$  annotates  $e_j$ , and 0 otherwise. The measure is then defined as

$$Sim_{cos}(c_i, c_k) = \frac{\langle v_i, v_k \rangle}{\|v_i\| \|v_k\|} , \quad (5)$$

where  $\langle v_i, v_k \rangle$  is the dot product of vectors  $v_i$  and  $v_k$  and  $\|v_i\|$  is the length of  $v_i$ .

### 2.3 Evaluation

In order to measure how well an ontology’s usage correlates with its definition, we measure the correlation between how the gene products are annotated with terms (via the similarity measures in Section 2.2) and the terms as they are defined in the ontology. Formally, for each pair of terms  $(c_i, c_j)$ , we measure their distance in the ontology DAG. We then sort all term pairs in descending order (greatest distance first) and put them into a sorted list  $L_{DAG}$ . We then measure the similarity between each pair of terms via the similarity measures in Section 2.2, sort the term pairs in ascending order (lowest similarity first) and put them into a sorted list  $L_{measure}$ , where the measure is Resnik’s, Lin’s, Jiang’s, Rel or Cosine. Finally, we measure the correlation between

the two sorted lists  $L_{DAG}$  and  $L_{measure}$  using Kendall's  $\tau$  coefficient [14].

The basic  $\tau$  method requires all values in the ranked lists to be unique, which cannot be guaranteed in our problem setting. Therefore, we make a common modification [15] to the basic method as follows. Let  $L_1$  and  $L_2$  be the two (equal-length) lists that we are comparing. Let  $\ell_1^i \in L_1$  be the  $i$ th element in  $L_1$ , and  $\ell_2^i \in L_2$  be the  $i$ th element in  $L_2$ . Similarly define  $\ell_1^j$  and  $\ell_2^j$  for  $j \neq i$ . Now consider each pair of pairs  $((\ell_1^i, \ell_2^i), (\ell_1^j, \ell_2^j))$  for  $i \neq j$ . We say that this pair is *concordant* if  $\ell_1^i > \ell_1^j$  and  $\ell_2^i > \ell_2^j$  or  $\ell_1^i < \ell_1^j$  and  $\ell_2^i < \ell_2^j$ . The pair is *discordant* if  $\ell_1^i > \ell_1^j$  and  $\ell_2^i < \ell_2^j$  or  $\ell_1^i < \ell_1^j$  and  $\ell_2^i > \ell_2^j$ . (Note that all inequalities are strict.) Now let  $n_c$  be the number of concordant pairs, and  $n_d$  be the number of discordant pairs. Finally, let  $n_1$  be the number of ties among elements of  $L_1$  and  $n_2$  be the number of ties among elements of  $L_2$ . Then the  $\tau$  coefficient is defined as:

$$\tau(L_1, L_2) = \frac{n_c - n_d}{\sqrt{(n_c + n_d + n_1)(n_c + n_d + n_2)}}. \quad (6)$$

The  $\tau$  coefficient ranges from  $-1$  (perfect negative correlation) to  $+1$  (perfect positive correlation).

### 3. Generating Synthetic Data

Before we apply our correlation technique to real ontological data, we must first determine what  $\tau$  values we should expect if an ontology's application to annotating gene products in fact does reflect its definition, under each similarity measure of Section 2.2. Thus we generated pairs  $(e_i, C_i)$ , where  $e_i$  is a synthetic gene product and  $C_i$  is its simulated annotation set, i.e. each term  $c_j \in C_i$  annotates gene product  $e_i$ . The synthetic data has various properties, which we use to characterize the similarity measures.

Let  $G = (V, E)$  be the ontology DAG and  $m = |V|$ . For simplicity, we assume  $G$  to be a complete tree of depth  $d$  and branching factor  $k$ . The synthetic annotation data was generated using the following randomized process on  $G$ . For each of the  $n$  distinct gene products, we select one term as the first term according to a predetermined initial distribution  $\omega_0$ . The annotation data set is then generated using three parameters  $n$ ,  $r$ , and  $\gamma$  as follows.

- 1) Choose a initial distribution  $\omega_0 = \{P_0(c_1), P_0(c_2), P_0(c_3), \dots, P_0(c_m)\}$  over terms  $C = \{c_1, c_2, c_3, \dots, c_m\}$ . We will examine the distribution  $\omega_0$  in Section 4.
- 2) Randomly choose a starting term  $s_i \in C$  according to  $\omega_0$  for each of the  $n$  synthesized gene products  $e_i$ .
- 3) Let  $D$  be the all-pairs shortest path matrix on the ontology DAG  $G$ , where  $D_{ij}$  is the number of steps needed to reach  $c_j$  from  $c_i$ . For each  $s_i$ , generate a distribution  $Q_i$  over  $C$ , where the probability for each term decreases exponentially with its distance to  $s_i$ , i.e.  $Q_i(c_j) = \gamma^{D_{ij}}$ .

- 4) Choose  $r$  terms from  $C$  according to  $Q_i$ , and add them to  $C_i$ . For each  $c_j$  chosen, add all of its ancestors to  $C_i$ .

## 4. Result and Discussion

### 4.1 Synthetic Data: Parameter Sensitivity Analysis

To observe how the parameters of Section 3 influence correlation, we start by choosing  $\omega_0$  to be the uniform distribution. Thus each starting term was chosen uniformly from the ontology DAG. Twenty sets of annotations were generated for each configuration of  $(n, r, \gamma)$  on a complete binary tree of depth 7. We evaluated the mean values of the correlation between  $L_{DAG}$  defined in Section 2.3 and the sorted list for each measure, which are  $\tau(L_{DAG}, L_{Lin})$ ,  $\tau(L_{DAG}, L_{Resnik})$ ,  $\tau(L_{DAG}, L_{Rel})$ ,  $\tau(L_{DAG}, L_{Jiang})$  and  $\tau(L_{DAG}, L_{Cos})$  on various configurations of parameter values.

Figure 1 shows the the average  $\tau$  for a variable number  $n$  of gene products using  $r = 15$  and  $\gamma = 0.6$ . In Figure 1, the average correlation for Cosine increases with increasing  $n$  (the number of annotations), while the four other measures are not affected by  $n$ . Also, we notice that when  $n > 170$ , further increase of  $n$  will not increase  $\tau$  for any measure very much.

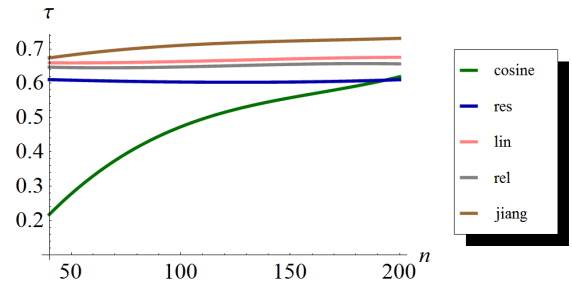


Fig. 1: Average  $\tau$  of each similarity measure with respect to  $n$  the number of distinct gene product when fixing  $r$  and  $\gamma$  ( $n \in [40, 200]$ ,  $r = 15$ ,  $\gamma = 0.6$ ).

Figure 2 shows the results for variable  $\gamma$  when  $n = 200$  and  $r = 8$ . For  $\gamma < 0.65$ , the correlation for Jiang's measure decreases with growing  $\gamma$ . In contrast,  $\tau$  for Cosine increases with growing  $\gamma$ . Also, the change of  $\gamma$  does not influence the correlation for other three measures. When  $\gamma > 0.65$ ,  $\tau$  for every measure begins to decrease with increasing  $\gamma$ , especially for Cosine, which decreases dramatically.

In Figure 3, we chose a moderate  $\gamma = 0.6$  and sufficiently large  $n = 200$  to examine the trend in the values of  $r$ . Similar to the results in Figure 1, correlations for Resnik's, Lin's, and Rel change little with increasing  $r$ , Jiang's decreases slightly, and the correlation for Cosine increases significantly.

From the three figures, we can see that  $\gamma$  affects  $\tau$  of all similarity measures, though less so for Lin's, Rel, and

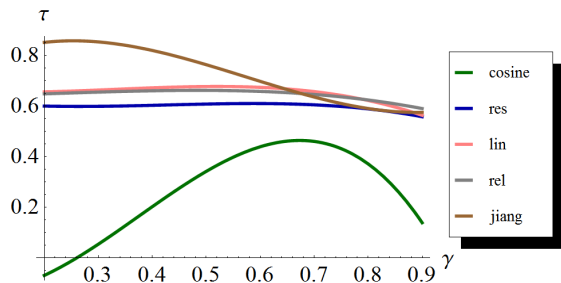


Fig. 2: Average  $\tau$  of each similarity measure with respect to  $\gamma$  when fixing  $n$  and  $r$  ( $n = 200, r = 8, \gamma \in [0.2, 0.9]$ ).

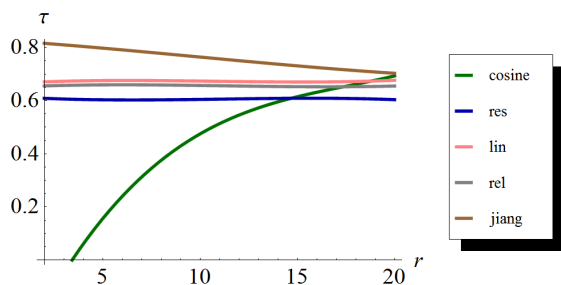


Fig. 3: Average  $\tau$  of each similarity measure with respect to  $r$  the number of terms associated with each gene product when fixing  $n$  and  $\gamma$  ( $n = 200, r \in [2, 20], \gamma = 0.6$ ).

Resnik’s. A gene product can be associated with a number of distinct terms, and  $\gamma$  defines how sparse the annotation of a gene product is distributed in the ontology. A small  $\gamma$  indicates that the gene product has been annotated by several terms close each other. Results show that Cosine correlates more when  $\gamma \approx 0.65$  while the correlation for the other four increases when  $\gamma$  is low.

The parameter  $r$  defines the number of terms assigned to a gene product. Higher  $r$  indicates that an individual gene product receives more annotations. This parameter affects Cosine significantly: its correlation goes high with increasing  $r$ . In contrast, Resnik’s, Lin’s and Rel show a very slight decrease when  $r$  increases, though they are still quite stable.

In contrast to  $\gamma$  and  $r$ , the number of gene products  $n$  has limited influence on the correlation. Generally, higher  $\tau$  can be obtained for all measures when more annotations are made. However, as long as there is a sufficient number of annotation records ( $n > 170$ ), further increase brings only a slight increase to the correlation.

From these results we see that Cosine is only suited for evenly annotated data with moderate  $\gamma \approx 0.65$  and high  $r$ , which means each gene product is annotated by many moderately concentrated terms in the ontology. Jiang’s measure is best suited for data with low  $\gamma$  and  $r$ , which means each gene product is annotated by very few closely related terms in the ontology. Also, we found that Resnik’s,

Lin’s and Rel are almost independent of the three parameters.

## 4.2 Synthetic Data: Geometrically Distributed Number of Annotations

We now modify the synthetic data generation model to be more realistic. When an ontology is used in practice, the terms commonly used often come from a relatively small subset of the entire set of terms. As an example, refer to Figure 4, which shows that in the database UniProtKB/Swiss\_Prot, 40% of the gene products are annotated by at most two GO terms, and less than 10% of gene products receive annotation from more than 5 terms. On average, there are five terms used to annotate each gene product. Thus, in our updated model, we let  $r$  (the number of terms annotating a gene product) vary among the gene products. Based on Figure 4, we assume the number of terms follows a geometric distribution with parameter  $p$ , which is the probability that a randomly selected gene product is annotated by a single term. (So a smaller value of  $p$  results in a longer tail.) Figure 4 suggests a value of  $p$  between 0.35 and 0.50.

Ten sets of annotations were generated on each configuration of  $n = 100, \gamma = 0.3$  and  $p$ , whose values ranged from 0.1 to 0.9, on a complete binary tree of depth 7. In Figure 5, we show the average value of  $\tau$  that resulted from running our experiments for variable values of  $p$ . The figure suggests that larger values of  $p$  tend to increase the correlation for all measures, except for Cosine (which decreases) and Resnik’s (which is the most stable of all). The correlation of Jiang’s increases dramatically with  $p$ .

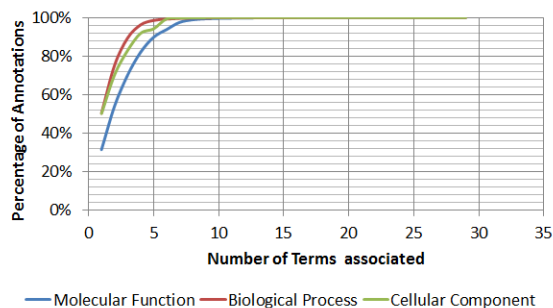


Fig. 4: Percentage of gene products annotated in GO versus number of terms used to annotate them.

The second variation we made over the experiments of Section 4.1 is in the distribution  $\omega_0$ . Our results in Section 4.1 used a uniform distribution for initial distribution  $\omega_0$ . We now examine the effect of nonuniformity of the  $\omega_0$  on the  $\tau$  correlation coefficient for each similarity measure using skewed  $\omega_0$ , where nonuniformity is measured by the

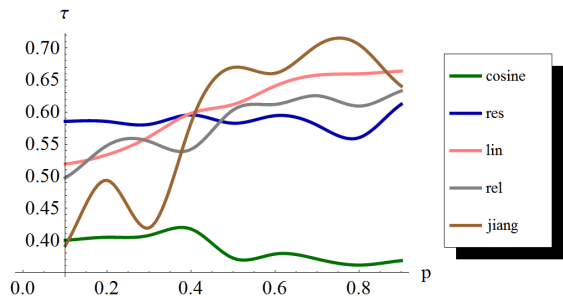


Fig. 5: Average value of  $\tau$  based on variable number of annotations  $r$  geometrically distributed with parameter  $p$  ( $n = 100$ ,  $\gamma = 0.3$ ).

normalized entropy  $H_0$ :

$$H_0(\omega_0) = \frac{H(\omega_0)}{H_{max}} = \frac{-\sum_{i=1}^m P(c_i) \log_2 P(c_i)}{\log_2 m}.$$

Two hundred sets of annotations were generated from the configuration  $n = 200$ ,  $\gamma = 0.6$  and  $r = 2$ . In each set, we chose  $m$  values at random from  $[0, 1]$  according to an exponential distribution with parameter  $\lambda \in [0.5, 10]$  and then normalized them to get  $\omega_0$ . Figure 6 shows the impact of  $\omega_0$ 's normalized entropy on  $\tau$ . We can see that increasing  $H_0$  (making  $\omega_0$  more uniform) generally increases the correlation of all five measures, though Resnik's and Lin's are fairly stable. In particular, Cosine and Jiang's increase dramatically with increasing  $H_0$ .

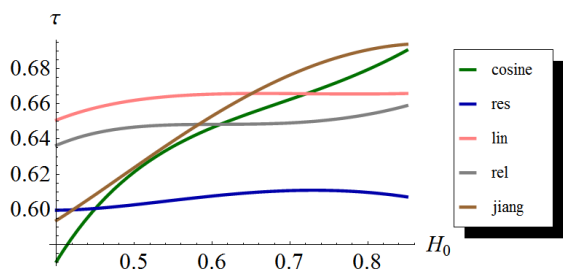


Fig. 6: Average value of  $\tau$  versus the normalized entropy  $H_0$  of the starting distribution  $\omega_0$  ( $n = 200$ ,  $\gamma = 0.6$ ,  $r = 5$ ).

From these results we can see that Cosine and Jiang's are not well suited for skewed data (with a low-entropy  $\omega_0$ ), and Cosine is not well suited for data with a short tail (high  $p$  value). Also, unlike Cosine and Jiang's, the correlation values of Resnik's, Lin's and Rel (especially Resnik's) are more stable across many parameter values.

### 4.3 Real Data: Partial Ontology

We empirically compared Rel, Cosine, Resnik's, Lin's, and Jiang's similarity measures using annotations from UniProtKB [16] with a corresponding sub-ontology from

Table 1: Comparison of  $\tau$  on "GO:0005275"

Measure	UniProtKB/Prot	UniProtKB
Cos	0.424	0.319
Resnik	0.596	0.576
Lin	0.621	0.602
Rel	0.618	0.630
Jiang	0.441	0.480
Terms	17	25
Genes	895	25105
Annotations	907	25593

GO. We used a subset of 25593 annotations along with the subtree from GO, rooted at the term "GO:0005275: amine transmembrane transporter activity." This annotation set consists of 25105 identified genes and contains 25 unique terms. UniProtKB is comprised of two sections, UniProtKB/Swiss\_Prot and UniProtKB/TrEMBL. UniProtKB/Swiss\_Prot contains curated annotations while UniProtKB/TrEMBL contains entries with computationally analyzed annotations generated by automatic procedures. These are not reviewed and curated by an author. Thus, UniProtKB/Swiss\_Prot may have data of higher quality than UniProtKB/TrEMBL. Note that 98% of the records are electronically annotated. We first computed correlations using only UniProtKB/Swiss\_Prot, then using the entire set (UniProtKB).

The electronic annotations in UniProtKB/TrEMBL have many gene products that are each annotated by a single term. Further, the annotation in UniProtKB/TrEMBL contains only a subset of GO terms and is significantly larger than UniProtKB/Swiss\_Prot. Thus, in Table 1 we see that Cosine's correlation decreased dramatically while only Rel and Jiang's have slightly improved correlation when switching from UniProtKB/Swiss\_Prot to UniProtKB. Since Resnik's, Lin's and Jiang's are almost immune to changes in parameter values (according to Section 4.2), we can use their correlations from our tests on synthetic data as a baseline for our experiments here. The  $\tau \approx 0.6$  for these three measures from Table 1 is very close to the baseline suggested by Figures 1–3. This leads us to believe that this partial ontology correlates well to its usage.

### 4.4 Real Data: Full Ontology

Our experiment on the full ontology was performed on a copy of GO annotations dated April 2010, which consisted of 32651844 annotations of 6729320 gene products using terms from three ontologies (see Table 2). There are 43645  $is\_a$  relations defined over the 26664 terms. From the table we see that the three ontologies differ in size. The Biological Process ontology is much larger than the other two. Also, the table shows that more than one third of the terms are defined but have never been used. For Biological Process, almost half are unused.

We studied each of GO's three ontologies by computing

Table 2: Number of terms and relations for each GO ontology. Numbers exclude obsolete terms. “Active” refers to terms that have been used at least once. “Relations” refers to is\_a relations.

Ontology	Terms		Relations
	Total	Active	
Cellular Component	2626	1653	3992
Molecular Function	8659	5885	10132
Biological Process	18005	9497	29521

the Kendall  $\tau$  rank correlation coefficient for every pair of measures in Section 2.2 as well as the ontology DAG distance  $D$ . In order to compute  $\tau$  for  $m$  terms, we would need to compute the sorted similarity measure list on all  $\binom{m}{2}$  term pairs. Thus the algorithm for computing the Kendall  $\tau$  rank correlation coefficient in our case has a complexity of  $\Theta(m^4 \log(m))$  [17]. Given that the number of terms ranges from 1653 to 9497 (Table 2), it is infeasible to evaluate  $\tau$  directly. Instead, we estimate  $\tau$  by uniformly randomly sampling term pairs from the list. In order to do so, each time we sample 1000 term pairs from the list and compute  $\tau_i$ , and then repeat this sampling process 50 times. We estimate  $\tau$  as the mean of  $\tau_1, \dots, \tau_{50}$ . Since the standard deviation of  $\tau_1, \dots, \tau_{50}$  between each measure was  $< 0.01$ , we consider the mean to be a good estimate.

Tables 3–5 present the  $\tau$  values for each pair of similarity measures for each of the three ontologies. The first column of each table shows the correlations between DAG distance and the five measures. Res, Lin, Rel and Jiang each correlate with DAG at about the same values, while Cosine only shows a weak correlation. Also, we noticed that the first four are highly correlated with each other, especially Jiang vs. Lin and Res vs. Rel, which correlate near 0.99. This is unsurprising given the relationships among the definitions of these measures.

Table 3: Estimated  $\tau$  between similarity measures on Cellular Component.

	DAG	Cos	Jiang	Rel	Lin
Res	0.44	0.25	0.85	0.99	0.83
Lin	0.40	0.45	0.98	0.83	
Rel	0.44	0.25	0.84		
Jiang	0.40	0.43			
Cos	0.23				

Table 4: Estimated  $\tau$  between similarity measures on Molecular Function.

	DAG	Cos	Jiang	Rel	Lin
Res	0.40	0.20	0.90	0.99	0.89
Lin	0.37	0.33	0.99	0.89	
Rel	0.40	0.20	0.90		
Jiang	0.38	0.32			
Cos	0.19				

Table 5: Estimated  $\tau$  between similarity measures on Biological Process.

	DAG	Cos	Jiang	Rel	Lin
Res	0.37	0.25	0.96	0.99	0.96
Lin	0.37	0.29	0.99	0.95	
Rel	0.37	0.25	0.96		
Jiang	0.37	0.29			
Cos	0.24				

From Section 4.1, we understand how values for  $n$ ,  $r$ ,  $\gamma$ ,  $p$ , and  $H_0(\omega_0)$  for an ontology and its annotations affect correlation values for the similarity measures we use. The values of  $n$ ,  $r$ , and  $p$  are directly estimated from the data. However, it is not obvious how to directly estimate  $\gamma$  and  $H_0(\omega_0)$  from the data. But if we look at  $H_0(\omega)$  (the normalized entropy of the final distribution over the terms), we find that it is generally low. From this we estimate that both  $H_0(\omega_0)$  (the normalized entropy of the initial distribution) and  $\gamma$  are generally low in the real data. Specifically, we use  $H_0(\omega)$  as an upper bound of  $H_0(\omega_0)$ . Table 6 shows values of the relevant parameters in GO;  $\gamma$  is omitted and instead is qualitatively estimated as “low”, since Table 6 gives  $H_0(\omega)$  as relatively low, ranging from 0.44 to 0.58.

Table 6: Corresponding parameters for each ontology.

Ontology	$n$	$r$	$p$	$H_0(\omega)$
Molecular Function	5860336	2.85	0.35	0.58
Cellular Component	3217382	2.13	0.47	0.44
Biological Process	5127003	1.94	0.52	0.55

Since increasing  $n$  beyond a sufficient number (170 in synthetic data) brings only minimal changes in correlation, we expect  $n$  will have little effect on correlation values even though it is four orders of magnitude higher than the values used in our synthetic data. The  $\tau \approx 0.2$  for Cosine in GO lies in the interval  $[0.1, 0.4]$  that is suggested by Figures 3 and 5 for synthetic data of similar characteristics.

Table 6 gives low  $H_0(\omega)$  from 0.44 to 0.58, which suggests that both  $\gamma$  and  $H_0(\omega_0)$  are low. The  $\tau \approx 0.39$  for Jiang’s is low compared to either 0.8 given by low  $\gamma$  in Figure 2, 0.45 given by  $p \approx 0.25$  in Figure 5 or 0.6 given by  $H_0(\omega_0)$  around 0.4 in Figure 6.

In addition, the average  $\tau \in [0.37, 0.44]$  for Resnik’s, Lin’s and Rel are low compared with those from the synthetic data and GO:0005275, where similar configurations show that correlations around 0.6 are possible (and very stable in the case of Resnik’s). All these results suggest that GO’s use correlates less with its definition compared to GO:0005275, though more experimentation should be performed to confirm this.

## 5. Conclusion

The Gene Ontology (GO) terms are widely used to annotate gene products. However, it is unknown whether

the terms defined in GO are used to label gene products in a manner consistent with their definition. Since there are many ways to measure semantic similarity, we first used various synthetic data models to study several similarity measures to characterize their sensitivity to various properties of the data. We found that Cosine is only suitable for annotation sets that have with long tails (low  $p$  values) and in which each term is annotated by many moderately concentrated terms in the ontology. Jiang's measure is not well suited for skewed data (with a low-entropy  $\omega_0$ ) and in which each gene product is annotated by very few closely related terms in the ontology. Also, we found that Resnik's, Lin's and Rel are almost independent of the these parameters, especially Resnik's.

Then we investigated a small sub-ontology and its annotations of data from UniProtKB and found that Rel, Resnik's and Jiang's measures indicate correlations between the DAG and its application relative to what seems to be the best possible based on tests on synthetic data. Thus we conclude that this partial ontology's definition relates well to its usage.

Finally, from our preliminary result on the full GO ontologies, we found that correlation results using the more stable measures (especially Resnik's) seem to indicate that the correlation between GO's use and its definition is low, especially when compared to the correlation between GO:0005275 and UniProtKB. More experimentation should be performed to confirm this.

In addition to a more detailed analysis, future work includes examining other measures that evaluate semantic similarity, and characterizing them based on synthetic data parameters as we did with those of this paper. This might reveal measures that are even less sensitive to the parameter values and might in turn be even more useful for studying real data.

Our synthetic data model was based on complete binary trees that were not similar to the DAGs in GO. Thus it is possible that the trends observed in our synthetic data results might not reflect what we would see in a full ontology. Therefore, in our ongoing work, we randomly selected 100 terms from GO, each with around 100 child terms, yielding 100 subDAGs, each of size approximately 100. We then measured the sensitivity of each similarity measure's  $\tau$  value to the five parameters by repeating the tests of Section 4.1 on each of the 100 subDAGs. Our preliminary results show that Resnik's measure remained almost invariant to changes in parameter values when the subDAG remains unchanged. However, Resnik's  $\tau$  value was sensitive to the topology of the subDAG. In our continued research, we will further investigate this, attempting to correlate the similarity measures'  $\tau$  values to properties of the subDAGs, such as branching factor, depth, diameter, and skewness.

## Acknowledgments

The authors thank the anonymous reviewers for their comments. This research was supported by National Science Foundation grant number 0743783.

## References

- [1] M. Ashburner, C. A. Ball, and J. A. Blake, "Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25–29, 2000.
- [2] L. Stein, "Genome annotation: From sequence to biology," vol. 2, pp. 493–503, 2001.
- [3] K. Verspoor, K. B. C. D. Dvorkin, and L. Hunter, "Ontology quality assurance through analysis of term transformations," *Bioinformatics*, vol. 25, pp. 77–84, 2009.
- [4] P. W. Lord, "Investigating semantic similarity measures across the Gene Ontology: The relationship between sequence and annotation," *Bioinformatics*, vol. 19, pp. 1275–1283, 2003.
- [5] A. Schlicker, F. S. Domingues, J. Rahnenfuhrer, and T. Lengauer, "A new measure for functional similarity of gene products based on Gene Ontology, including indels," *BMC Bioinformatics*, vol. 7, p. 302, 2006.
- [6] M. Mistry and P. Pavlidis, "Gene Ontology term overlap as a measure of gene functional similarity," *BMC Bioinformatics*, vol. 9, p. 327, 2008.
- [7] A. Schlicker, C. Huthmacher, F. Ramirez, T. Lengauer, and M. Albrecht, "Functional evaluation of domain-domain interactions and human protein interaction networks," *Bioinformatics*, vol. 23, no. 7, pp. 859–865, 2007.
- [8] C. Pesquita, D. Faria, H. Bastos, A. Ferreira, A. Falcao, and F. Couto, "Metrics for GO based protein semantic similarity: A systematic evaluation," *BMC Bioinformatics*, vol. 9, no. Suppl 5, p. S4, 2008.
- [9] "GO annotation policies and guidelines." [Online]. Available: <http://www.geneontology.org/GO.annotation.shtml>
- [10] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 448–453.
- [11] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the 15th International Conference on Machine Learning*, 1998, pp. 296–304.
- [12] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of ROCLING X*, 1997, p. 9008.
- [13] M. Popescu, J. M. Keller, and J. A. Mitchell, "Fuzzy measures on the Gene Ontology for gene product similarity," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 3, pp. 263–274, 2006.
- [14] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, pp. 81–93, 1938.
- [15] "Kendall Rank Correlation. Wolfram Mathematica." [Online]. Available: <http://reference.wolfram.com/mathematica/MultivariateStatistics/ref/KendallRankCorrelation.html>
- [16] E. Jain, A. Bairoch, and S. Duvaud, "Infrastructure for the life sciences: Design and implementation of the UniProt website," *BMC Bioinformatics*, vol. 10, p. 136, 2009.
- [17] D. Christensen, "Fast algorithms for the calculation of Kendall's  $\tau$ ," *Computational Statistics*, vol. 20, pp. 51–62, 2005.