Fall 12-2-2011

A Study of Correlations between the Definition and Application of the Gene Ontology

Yuji Mo
University of Nebraska-Lincoln, ymo@cse.unl.edu

A STUDY OF CORRELATIONS BETWEEN THE DEFINITION AND APPLICATION OF THE GENE ONTOLOGY

by

Yuji Mo

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Stephen Scott

Lincoln, Nebraska

December, 2011

A STUDY OF CORRELATIONS BETWEEN THE DEFINITION AND APPLICATION OF THE GENE ONTOLOGY

Yuji Mo, M. S.

University of Nebraska, 2011

Adviser: Stephen Scott

When using the Gene Ontology (GO), nucleotide and amino acid sequences are annotated by terms in a structured and controlled vocabulary organized into relational graphs. 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.

We also study how terms from different ontologies in GO relate to each other, Such relationships can be helpful in refining term definitions. In order to identify such "cross-terms", we propose a generalized semantic measure which can be used to identify related terms across GO ontologies. Results based on Saccharomyces Genome Database show that the measure is correlated with the degree of co-occurrence for term pairs. By thresholding the level of similarity, we found a list of highly correlated cross ontology term pairs. These term pairs show a high level of biological correlation.

Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Dr. Stephen Scott. His broad knowledge and constructive ideas have been of great value for me. It is an enjoyable experience for me to work with him, a very encouraging and enthusiastic professor.

I am grateful to all the professors on my committee for their insightful comments and suggestions throughout my research.

I wish to thank Catherine Anderson. Those brainstorming talks for new ideas with her have always greatly motivated me in my research. I would like also thank her invaluable help in interpreting results.

I owe my thanks to my family and all my friends for their constant encouragement and support.

This research was supported by National Science Foundation grant number 0743783.

Chapter 1

Introduction

The recent increase of fast affordable genomic sequencing technology has accelerated the availability of molecular sequences [2]. The enormous amount of sequence data transformed how biologists describe and characterize gene products. Biologists' progress in conceptualizing biological terms can no longer keep pace with the volume of sequence data. Thus Gene Ontology (GO) Consortium was formed to address this issue. The initiative's aim is to push the genomic community towards standardizing the representation of gene and gene product attributes across species and databases. During the past decade, GO became widely accepted in this community as a concise means of annotating gene products for machine translation [30].

The wide acknowledgement of GO has brought in concerns as well. Because the wide scope of the community GO targeted, curators may not be able to reach a consensus in the vocabulary's (usually referred to as *terms*) definition and usage. Recent research [25, 17, 9, 26, 23] in semantic similarity measures allow us to quantify relationships between terms. These measures reflect the terms' biological relation [18] based on their usage. Our research uses the semantic similarity measures as tools to examine the correlation between term definition and usage.

GO consists of three orthogonal ontologies which are controlled vocabularies describing the domain of gene products, i.e., enzymes and other proteins encoded in DNA. 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 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.

Since the GO project is a collaborative effort between groups sharing their vocabularies, terms can be added, deleted or edited when its usage came into doubt [7].

Sometimes, controversial terms get split into two child terms. Meanwhile gene product are usually annotated by a series of terms. Correlated terms in these cases, especially those from different aspects of an ontology, can be of interest to biologists as well. To distinguish such related terms, a unified similarity measure is proposed to evaluate the level of similarity between any two terms across ontologies.

In order to justify that our unified similarity measure reflects the correlation between terms, we compared the measure with the Resnik measure, which we found to be the most stable measure. Results show that our measure behaves similarly to the Resnik measure evaluating terms from the same ontology. We further verify our measure by investigating term pairs that annotate the same gene product. Results show that term pairs more frequently co-occurred together have higher similarity values. Thus we believe our measures can effectively quantify the similarity between terms.

By thresholding the unified semantic similarity values, we determined a list of highly correlated cross ontology term pairs. After examining some of them, we found high levels of biological correlation between these terms. We also identified several pair patterns by aggregating the relation between the terms.

In summary, this research presents a quantitative analysis of the correlation between GO's definition and application. The correlation is analyzed based on well annotated synthetic data with similar configuration as GO. We also found that one measure ("Resnik") is robust against various changes in annotation statistics. In addition, we proposed a unified similarity measure which can be used to quantify relationships between terms across ontologies. We found a high biological correlation between highly correlated term pairs identified using this measure.

Chapter 2

Background

2.1 Ontology

The value of any kind of knowledge can be greatly enhanced when it is allowed to be integrated with other data. For instance, bridging biodiversity data with genomics data enables reasoning from morphology to gene sequence. Such integration enables systems which exploit computational possibilities in multiple domains. To facilitate such sharing, reuse and integration of knowledge among systems, it is useful to define a common vocabulary in which shared knowledge is formally represented [5]. The common controlled vocabulary is usually referred to as an ontology.

2.1.1 Concept of Ontology

The word ontology is borrowed from philosophy, which studies the existence of objects, basic categories and their relations. In the fields of AI and information systems, the ontology is usually referring to a vocabulary which consists of a set of objects and describable relationships among them. Formally, an ontology is a description of the concepts and relationships that can exist for an agent or a community of agents [28].

Common ontologies guarantee consistency, but not completeness, with respect to queries and assertions using the vocabulary defined in the ontology [6].

Constructing an ontology usually requires integrating information from different sources. For instance, building an inventory management ontology across many warehouses requires records for all items, customer records and depository information. A specimen cataloguing ontology for biologists uses geographic coordinates, species identification numbers or even gene sequences collected to bridge knowledge from different fields. Ontologies are designed to be shared among a community. This always involves related concepts and ontologies: for instance, an inventory management ontology has to be related customer reviews to increase its usefulness. A specimen cataloguing ontology has to be related to literature published and ontologies from other laboratories. The need for sharing and integration requires ontologies to use a common vocabulary.

Maintaining an ontology is expensive and requires a lot of effort. It is natural to see that in the domains of interest, application requirements change over time. This change is often brought in by a distributed and collaborative manner. Developers from different communities may not share identical understandings of the concepts defined in an ontology. Therefore, modifications of the Gene Ontology come out every week. As an ontology grows to be larger and more popular, maintaining an ontology becomes a problem.

2.1.2 Ontology in the Fields of Biological and Biomedical Research

An ontology is usually designed to meet the interests of a domain. Advancement in biological and clinical research generates swarms of data. Organizing this infor-

mation involves the creation and analysis of annotations which link data collected to controlled vocabularies. This approach improves human readability, facilitates searching and makes data available to algorithmic processing [31]. Gene Ontology [2] is the most successful collaborative effort towards this goal, integrating millions of annotations across thousands of species.

Compared with molecular biology where data is publicly available and well defined, the biodiversity and biomedical domains only have limited amounts of data for research purposes. Due to the nature of these data, knowledge is mostly defined in natural language in the literature. Even in the field of clinical research where systematic data are available, the use of local schemas prevent data to accumulate [13]. To face this problem, the Open Biological and Biomedical Ontologies (OBO) [29] initiative provides a lightweight solution. Rather than defining an integrated ontology like GO, it approaches consensus by developing a set of expanding orthogonal life science ontologies where ontologies are managed by individual interest groups. OBO has been widely accepted in the biodiversity side and gained lots of interests.

2.1.3 Gene Ontology

The Gene Ontology project is a collaborative effort aiming to standardize the representation of gene and gene product attributes across species and databases. GO is made up of three independent, orthogonal ontologies:

- Cellular Component (CC) ontology, which describes where a gene product is located at a sub-cellular level;
- Molecular Function (MF) ontology, which describes the function a gene product can perform;

- Biological Process (BP) ontology, which describes series of events and molecular functions.

Terms in GO can have any number and type of relationships to other terms. The relations are endowed with descriptive logic so that inferences can be made between terms. There are three types of relationships defined in go: "is a", "part of" and "regulate". When we say A "is a" B, A is a subtype of B. For example, "lyase activity" is a subtype of "catalytic activity". The "part of" relationship represents a whole-part relation. When A is "part of" B, then B is a necessarily part of A. For instance, "replication fork" is a part of "chromosome" but not all instances of "chromosome" have "replication fork". Within biological process ontology, "regulate" relation describes one process's direct effect on the other process, and be either positive or negative.

"Regulate" relations only appear in biological process ontology and can possibly forms cycles. For example, a series of functions which one promotes another in a cycle. Since this type of relation does not reflect any hierarchical order between terms, we are only interested in "is a" and "part of" relations in this thesis. In addition Lord [18] mentioned that the "part of" and "is a" relations are usually exclusive. He found that the semantic meaning of the two relation types varies between different ontologies. Thus we consider the two type of relations equally and do not distinguish between types of relations. Thus GO can be structured as a directed acyclic graph (DAG) using these relations. Each node of each DAG is a term with a distinct name and description. The edges of a DAG represent relations between the connected 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 earned popularity among the genomics community. 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 [7]. Not all gene products are well annotated in GO. Quite a few of them are annotated by only one or two terms. Since these terms may biologically correlated with other terms, the gene products should likely to be annotated by others as well. These biological correlation could also provides insights for biologist to refine term.

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 and to identify biologically correlated term pairs. We found that GO annotations might not be applied in a manner consistent with their definition. We also found a list of term pairs with high levels of biological correlation.

2.2 Semantic Similarity Measures

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.

2.2.1 Similarity between Terms

There are many different functions for calculating semantic similarity between terms. We consider the following five measures because these widely used measures quantify the relation between terms based on their annotations. Thus they are suitable to represent the applications in GO.

Resnik [25] proposed that the amount of information provided by the common ancestors of the two terms may be used as a measure:

$$Sim_{Resnik}(c_i, c_j) = \max_{c_k \in S(c_i, c_j)} -\log P(c_k) \ , \tag{2.1}$$

where $S(c_i, c_i)$ 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 [17] extended Resnik's measure by modifying the information content of a term to take both descendants into consideration:

$$Sim_{Lin}(c_i, c_j) = \max_{c_k \in S(c_i, c_j)} \left(\frac{2 \log P(c_k)}{\log P(c_i) + \log P(c_j)} \right) \ . \tag{2.2}$$

Generic terms do not have a high relevance for the comparison of different gene products. Andreas's [26] 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:

$$Sim_{Rel}(c_i, c_j) = \max_{c_k \in S(c_i, c_j)} \left(\frac{2(1 - P(c_k)) \log P(c_k)}{\log P(c_i) + \log P(c_j)} \right) \ . \tag{2.3}$$

Jiang [9] proposed a similarity measure as the reciprocal of semantic distance:

$$Sim_{Jiang}(c_i, c_j) = \max_{c_k \in S(c_i, c_j)} \left(\frac{1}{-\log P(c_i) - \log P(c_j) + 2 \log P(c_k)} \right) . \tag{2.4}$$

The Cosine similarity [23] 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 $v_i = (v_{i1}, v_{i2}, \ldots, v_{in})$, in which $v_{ij} = 1$ if c_i annotates e_j, and 0 otherwise. The measure is then defined as

$$Sim_{\cos}(c_i, c_j) = \frac{\langle v_i, v_j \rangle}{\|v_i\| \|v_j\|} , \tag{2.5}$$

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

2.2.2 Similarity between Gene Products

In addition to the similarity between GO terms, we may want to compare gene products as well. Since correlated gene products are more likely have similar descriptions in the literature or experimental results, they may have sets of terms that are related to each other. Thus we need a method to quantitate the similarity between two sets of terms. Common approaches take the maximum similarity between every pair of terms from each set. For example, the similarity between two sentences is determined by the closest pair of words from each sentence because the words in the sentences have only one sense at a time. However, the gene product will have plenty of the annotations contributed by many curators at the same time. Thus Lord [18] points out that the maximum similarity method is not suitable in the case of GO after in-

vestigating SWISS-PROT-Human. He further suggests using the average similarity between terms instead.

$$Corr^G(e_i, e_j) = \frac{1}{|C_i||C_j|} \sum_{c_k \in C_i, c_l \in C_j} Sim^G_{measure}(c_k, c_l) \ , \tag{2.6}$$

where C_i and C_j are the sets of terms that annotate gene products e_i and e_j (respectively), $|C_i|$ is the size of the set C_i, and $Sim^G_{measure}$ can be any of the term similarity measures described in Section 2.2 on ontology G.

2.3 Related Work

The Gene Ontology has established itself as one of the most important source for computational knowledges in the field of gene products. After Lord [18] investigated the correlation between semantic similarity and gene product sequence similarity, the semantic similarity on GO arouses great interest from many aspects.

Lee [16] presented a graph theoretic algorithm to extract common biological attributes of the genes within a cluster from gene ontology (GO). With the information extracted, they were able to perform various microarray analyses. Al-Shahrour [1] presented a tool to extract GO terms that are significantly popular over sets of genes within the context of a genome-scale experiment like DNA microarray. Kohler [14] introduced a different database integration method. This method can be used to integrate life science databases with the help of GO. Schlicker [27] compared the human protein interaction network derived from experiment with the one predicted by similarity measure.

There are also researches on the quality between a term and its definitions. The algorithm published by Kohler [14] is able to identify terms and definitions which

are defined in a problematic way, where reasoning over relations shows contradiction. Using ontology alignment, their algorithm can propose alternative synonyms and definition for those problematic terms.

Other research focuses on the similarity measures themselves. Mistry [19] proposed similarity measure that can avoid some problems that affect the probability-based measures. They showed that their measure is significantly faster than information content based measures. There are also improvements [19] proposed to overcome difficulties in the existing similarity measures. Pesquita [22] systematically evaluated all these measures and their variations using the relationship with sequence similarity.

The demand to identify related ontology terms across ontologies has been address in general ontology research. They are mostly aimed for matching ontology schemas from different perspectives like databases [15], information systems [3] and web services [11]. They attempt to solve this problem using fuzzy logic or information theory.

Compared with research in general ontology, the need to search for cross terms in GO has not received enough attention. This motivates our study to identify biologically related term pairs. Also, all the literature described above interprets each ontology individually. This is another motivation of our study, which is to use a generalized method to assess *semantic similarity* using all three ontologies.

2.4 Summary of the Literature

As more biology experts come to understand the convenience of storing data in the interchangeable format, we can expect more structural knowledge to become available to computational analysis. Plenty of semantic similarity measures are designed for their applications. This includes similarities between primary data being annotated

and between vocabulary terms used to annotate them. Related work has been done to predict experimental results. In contrast, we use some of the same measures they do, but for the purposes of measuring the consistency of the use of GO.

2.5 Objectives of our Study

From the literature, we understand that term usage in GO is not necessary consistent. Also we know that ambiguous terms and related terms exist in GO. Our study has two objectives. First, we measure the consistency of the use of GO terms by comparing GO's defined structure to the terms' application. Second, we design a generalized semantic measure which can be used to identify a correlated vocabulary across GO.

Chapter 3

Methodology

In order to achieve the objectives described in Section 2.5, we will utilize both synthetic data and data from real ontologies. We will use synthetic data to characterize the sensitivity of several similarity measures to various properties of the data. We then interpret the correlation between the definition and the usage of real ontologies using results obtained from synthetic data. At last, we will present a unified method to compute the similarity between any two terms, which could be coming from either the same or different ontologies.

3.1 Characterize Similarity Measures with Annotation of Various Properties

Before we apply our correlation technique to real ontological data, we must first determine what similarity 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.

3.1.1 Ontology Formalization

The gene ontology $G = (V, E)$ is organized as 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 both "is a" and "part of" relations are transitive, c_j "is a" or "part of" 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 [20], 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: $E_{root} = \bigcup_i E_i$.

3.1.2 Synthetic Data: Generating Ontology Annotations

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.

Formally, let $G = (V, E)$ be the ontology DAG and $m = |V|$. 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 ω_0. The annotation data set is then generated using three parameters n, r, and γ as follows.

1. Choose an initial distribution $\omega_0 = \{P_0(c_1), P_0(c_2), P_0(c_3), ..., P_0(c_m)\}$ over terms

$C = \{c_1, c_2, c_3, ..., c_m\}$. We will examine the distribution ω_0 in Section 3.1.4.

2. Randomly choose a starting term $s_i \in C$ according to ω_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.

3.1.3 Quantitative Analysis of Similarity Measures

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 τ coefficient [12].

The basic τ method requires all values in the ranked lists to be unique, which cannot be guaranteed in our problem setting. Therefore, we make a common modifi-

cation [24] 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 ith element in L_1, and $\ell_2^i \in L_2$ be the ith element in L_2. Similarly define ℓ_1^j and ℓ_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 τ coefficient is defined as:

$$\tau(L_1, L_2) = \frac{n_c - n_d}{\sqrt{(n_c + n_d + n_1)(n_c + n_d + n_2)}} \ . \tag{3.1}$$

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

3.1.4 Synthetic Data: Parameter Sensitivity Analysis

To observe how the parameters of Section 3.1.2 influence correlation, we start by choosing ω_0 to be the uniform distribution. We evaluated the mean values of the correlation between L_{DAG} defined in Section 3.1.3 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.

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 3.1, 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. We then modify the synthetic data generation model to be more realistic by taking two variations.

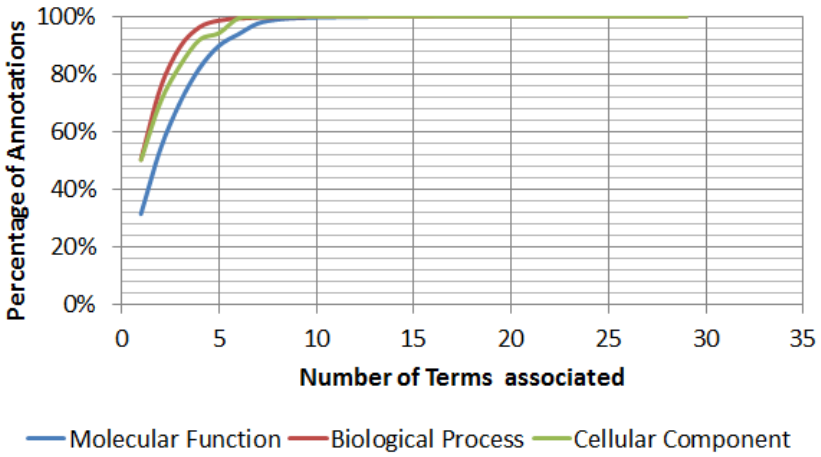

Figure 3.1: Percentage of gene products annotated in GO versus number of terms used to annotate them.

In our updated model, we let r (the number of terms annotating a gene product) vary among the gene products. Based on Figure 3.1, 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 3.1 suggests a value of p between 0.35 and 0.50.

The second variation we made over the experiments of Section 4.1 is in the distribution ω_0. Our results in Section 4.1 used a uniform distribution for initial distribution ω_0. We now examine the effect of non-uniformity of the ω_0 on the τ correlation coefficient for each similarity measure using skewed ω_0, where non-uniformity is measured by the normalized entropy H_0:

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

To sum up, similarity measures are characterized by five annotation parameters:

- n: the number of annotations
- r: the number of terms annotating a gene product
- γ: the sparseness the annotation of a gene product is distributed in the ontology
- p: the probability that a randomly selected gene product is annotated by a single term
- H_0: the entropy of ω_0, the initial distribution of choosing starting terms.

3.2 Evaluating the Correlation Between Ontology Usage and Definitions

3.2.1 Real Data: Partial Ontology

We empirically compare Rel, Cosine, Resnik's, Lin's, and Jiang's similarity measures using annotations from UniProtKB [8] with a corresponding sub-ontology from GO. 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 compute correlations using only UniProtKB/Swiss_Prot, then using the entire set (UniProtKB).

3.2.2 Real Data: Full Gene Ontology

We studied each of GO's three ontologies by computing the Kendall τ rank correlation coefficient for every pair of measures in Section 2.2 as well as the ontology DAG distance D. In order to compute τ 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 τ rank correlation coefficient in our case has a complexity of $\Theta(m^4 \log(m))$ [4]. Given that the number of terms ranges from 1653 to 9497 (Table 4.2), it is infeasible to evaluate τ directly. Instead, we estimate τ 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 τ_i, and then repeat this sampling process 50 times. We estimate τ as the mean of $\tau_1, \ldots, \tau_{50}$. Since the standard deviation of $\tau_1, \ldots, \tau_{50}$ between each measure was < 0.01, we consider the mean to be a good estimate.

3.3 Identifying Similar Terms Across Ontologies

Semantically similar terms may occur in different ontologies. For example "zinc ion transmembrane transporter activity" (GO:0005385) from molecular function ontology and "zinc ion transmembrane transport" (GO:0071577) from biological process ontology are highly related with each other. Even though the two terms come from different ontologies, 98% of gene products they annotated are the same. Correlated pairs may be of great interest to biologists to refine terms within the ontologies. For instance if term A from CC ontology correlates very highly with terms B and C from MF ontology but only a weak correlation exists between term B and C, then it could be argued to redefine term A from CC, A1 and A2 as two terms in CC, such that all A1 would correlate with term B and A2 would correlate with term C.

3.3.1 Unified Semantic Similarity

High semantic similarity values usually imply similar terms. Section 2.2 lists measures that quantify the semantic similarities between terms within the same ontology. These measures exploit both statistics in the term usage and the term to term relations defined in the ontologies.

There are problems in these approaches to measuring semantic similarity. First, terms not sharing a set of identical gene products are considered to be uncorrelated. Even though there is no overlapping gene product in the two sets, gene products in one set might have a synonym with almost identical gene sequence, function and hence annotation in the other set. Second, these measures rely on the ontology structure to compute IC (information content). If there is no direct path in the ontology between two terms (like terms each from a different ontology), the measures cannot be applied.

The term frequency–inverse document frequency weight (tf-idf) [10] is a weight often used to quantify term's importance against a document (a gene product in our scenario):

$$tfidf(c, e) = tf(c, e) * idf(c) \ , \tag{3.2}$$

where $tf(c, e)$ is the posterior probability to see term c given gene product e and $idf(c)$ is the inverse of the probability to see term c in any gene product. The method is not suitable in our study because of two reasons. First the weight considers the statistic in the occurrence of terms alone and unable to exploit the relations between terms. For example, two synonymous terms are treated as two distinctive terms in tf-idf. Second there are only unique annotations in Gene Ontology. Gene product e can be annotated by c at most once. Thus it is not possible to compute $tf(c, e)$ since we cannot compute probability solely based on annotations. Thus instead of using tf-idf to search for correlated terms, we use a semantic similarity measure we

proposed.

We look into using the measures described in Section 2.2 as the basis to construct new measures. Under the same concept, we may arguably believe that two terms are more similar if the two sets of gene products being annotated have higher semantic similarity values. These values can be computed by extending Equation (2.6) from Section 2.2.2.

Let e_i and e_j by any two arbitrary gene products. Equation (2.6) computes the similarity between e_i and e_j with respect to each ontology. Measures in Equation (2.6) consider only terms annotating e_i and e_j from a same ontology. We combine these measures to utilize terms from all three ontologies. These measures reflect the similarity between gene product in different aspect. Thus the weight between these measures can be further fine tuned to match the application of the unified measure. For simplicity, they are equally weighted in this thesis. We define the new similarity between two gene products e_i and e_j as the arithmetic mean of the similarities from the three ontologies:

$$Corr(e_i, e_j) = \frac{1}{3}[Corr^{MF}(e_i, e_j) + Corr^{BP}(e_i, e_j) + Corr^{CC}(e_i, e_j)] \ , \qquad (3.3)$$

where $Corr^{MF}(e_i, e_j)$, $Corr^{BP}(e_i, e_j)$ and $Corr^{CC}(e_i, e_j)$ correspond to the semantic similarity values between gene products e_i and e_j in Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). In this measure, the similarity between e_i and e_j is determined by terms annotating them from all three ontologies.

Given two terms c_i and c_j, each annotating a set of gene products E_i, E_j, the similarity between c_i and c_j is determined by the pairwise similarity between E_i and E_j. Since we already have a measure defined for every pair of gene products, we define the similarity between two sets of gene products as the average similarity between

every pair of gene product from E_i and E_j. Formally, our unified term similarity measure is defined as:

$$Sim_{Uni}(c_i, c_j) = \frac{1}{|E_i||E_j|} \sum_{e_k \in E_i, e_l \in E_j} Corr(e_k, e_l) \ , \tag{3.4}$$

where E_i is the set of gene products annotated by term c_i and $|E_i|$ is the size of set E_i.

For example, we consider the unified similarity between c_i and c_j, which come from two different ontologies. Methods in Section 2.6 cannot be used to evaluate the semantic similarity because the two terms do not come from the same ontology. Alternatively, we first compute the similarity $Corr^G(e_k, e_l)$ between each pair of gene products $e_k \in E_i$ and $e_l \in E_j$ for each ontology G, where G can be either MF, BP or CC. The similarity $Corr(e_k, e_j)$ between e_k and e_l is the mean of $Corr^{MF}(e_i, e_j)$, $Corr^{BP}(e_i, e_j)$ and $Corr^{CC}(e_i, e_j)$. Now we have a semantic similarity $Corr(e_k, e_j)$ between every pair of gene products annotated by c_i and c_j. Our method $Sim_{Uni}(c_i, c_j)$ takes the average similarity $Corr(e_k, e_j)$ between every pair of gene products as the unified similarity. Since the similarity between gene products ranges in $[0, +\infty)$, the unified similarity, which is the average similarity of a gene product pairs, ranges in $[0, +\infty)$ as well. We can measure Sim_{uni} for every pair of terms in GO. In this thesis we define term pairs with similarity value one standard deviation higher than the average similarity value between every term pair as *correlated terms* and those below as *uncorrelated terms*. When two correlated terms come from different ontologies, we say they are *correlated cross ontology terms*. We will further differentiate highly correlated terms from the others in Section 4.3.4.

Here we have a generalized semantic similarity between two terms. It differs from previous methods by two aspects. First, the similarity of the two terms considers

not only annotation from identical gene products but also those from similar gene products. Second, the two terms are no longer restricted to coming from the same ontology.

3.3.2 Validating Semantic Similarity

In Section 3.3.1, we created a method to measure the similarity between terms from different ontologies. In order to justify the unified method we proposed, we validate it against the Saccharomyces Genome Database (SGD). SGD project maintains a database of genomic and biological information. Compared with general gene product in GO, these yeast genomes, shown in Table 3.1, are better annotated with a smaller set of terms, see Table 3.1. Since these annotations are maintained and updated only by SGD curators, they may show a higher level of consistency. It is suitable for our purpose of validation.

Table 3.1: Number of terms and annotation in Saccharomyces Genome Database.

Aspect	Terms	Annotations
Cellular Component	525	10516
Molecular Function	1346	13933
Biological Process	1682	16835
Total	3553	41284

We expect our new measure to correlate with similarities computed via the methods of Section 2.2. So highly similar terms within an ontology should also be highly similar using the unified method. We will use an approach similar to that used in Section 3.1.3 to compare two measures. Formally, we compute the similarity via the Resnik method and our method for each pair of terms (c_i, c_j) from the same ontology, organize them into sorted lists L_{Resnik} and L_{Uni}, and then measure the τ coefficient

$\tau(L_{Resnik}, L_{Uni})$. Since there are three ontologies in GO, we will have three coefficients $\tau(L^{MF}_{Resnik}, L^{MF}_{Uni})$, $\tau(L^{CC}_{Resnik}, L^{CC}_{Uni})$ and $\tau(L^{BP}_{Resnik}, L^{BP}_{Uni})$.

We also expect that term pairs used together to annotate the same gene products to have high semantic similarity. In order to show this, we first need to identify these pairs. There is already plenty of research [21] invested in this topic. For simplicity, we consider all term pairs appeared in SGD since there are a significant number of term pairs that only appear once. For instance, for a gene product annotated by k terms, $\binom{k}{2}$ pairs of terms will be extracted pairwise. The $\binom{k}{2}$ co-occurred term pairs annotate the same gene product. Terms in these pairs appear together could either do so by true biological correlation or just by accident. To demonstrate that these co-occurred term pairs have higher semantic similarity values, average similarity values are computed based on randomly selected term pairs. We would like to see the difference between the value obtained from pairs occurred together and pairs in random cases.

Our third expectation is that term pairs more heavily used tend to have higher similarity values. In SGD, term usage is non-uniformly distributed. A few terms prevail among hundreds or even thousands of gene products. This results in heavily used term pairs from these terms. To deal with this fact, we define a degree of co-occurrence. The degree measures the co-occurrence of two terms in a value between 0 and 1, 0 as never seen together and 1 as alway appear together. For term pair (c_i, c_j), our degree of co-occurrence is defined as:

$$Freq(c_i, c_j) = \sqrt{\frac{|E_i \bigcap E_j|^2}{|E_i||E_j|}} \ , \tag{3.5}$$

where E_i is the set of gene products annotated by term c_i and $|E_i|$ is the size of set E_i. If $E_i = E_j$ which means the two terms annotate identical sets of gene products,

Equation (3.5) gives the highest degree. If the two terms are never used together in a gene product, the equation gives a degree of 0. Also, we will organize this measure into a list L_{Freq} and compare it with L_{Uni} using τ coefficient. The τ coefficient shows how the measure correlates term pair usage. Thus it can be another touchstone for the unified measure.

Chapter 4

Result and Discussion

In this chapter, we follow the steps described in Chapter 3 to analyze semantic similarity. We will use synthetic data as a tool to interpret results from real ontologies. Also we will validate the cross ontology semantic similarity we proposed by comparing it with existing measures and the degree of co-occurrence in term pairs using Kendall's tau coefficient and Pearson's linear coefficient. After that we will apply the measure to search for cross ontology term pairs by thresholding the minimal similarity values.

4.1 Analysis on Synthetic Data

4.1.1 Uniformly Distributed Number of Annotations

Before starting to choose parameters for synthetic data, we first need to understand the structure of GO. Because of the size of GO, it is infeasible to analyze it as a whole. Instead, we computed the number of descendants under each term in GO and randomly picked 30 terms which have around 90–110 children each. Three were from cellular component ontology, 9 were from molecular function and 18 were from biological process. By visualizing the DAG under these terms (Figure 4.1.1), we found

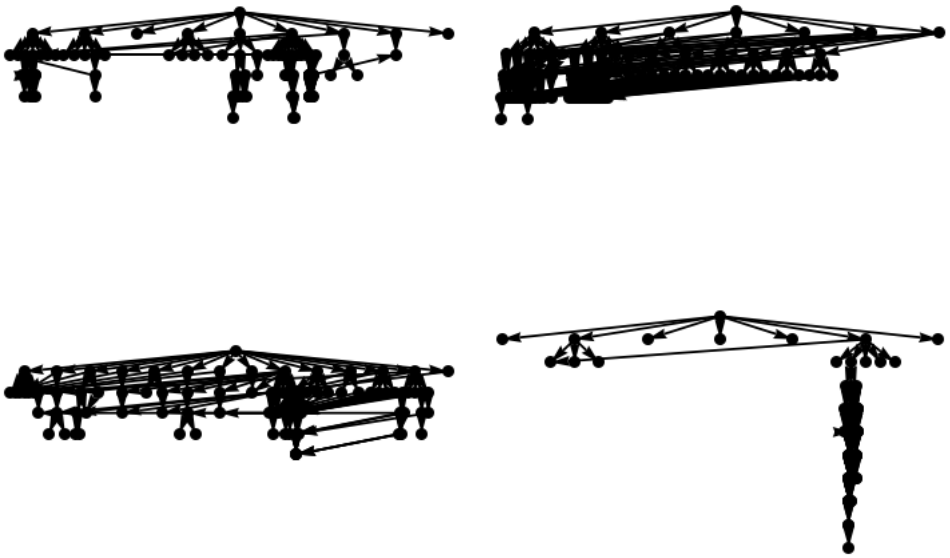

Figure 4.1: Four of the sampled ontologies from GO with approximately 100 terms each.

these 30 DAGs to be reasonably well balanced with only one exception (shown in the bottom right figure) which has a long chain in its branch.

We generated twenty sets of annotations with three configurations ($n = 120, r = 15, \gamma = 0.6$), ($n = 50, r = 5, \gamma = 0.6$), ($n = 50, r = 15, \gamma = 0.3$) following the procedures in Section 3.1.2 for the 30 DAGs. The average τ values in each measure for the DAGs in each configuration are close to each other (less than 0.15 maximum difference) except for the skewed DAG, which correlates much lower than the others. Because the 29 DAGs are more balanced than the skewed one, this indicates that the structure of the ontology does not have a significant impact on the similarity values as long as it is reasonably well balanced like the 29 DAGs. In order to extensively test the annotation parameters, we need a representative DAG to avoid testing on all possible DAGs. Thus for simplicity, we choose a complete binary tree of depth 7 as

the DAG for our synthetic data since the effective branching factor over the 29 DAGs is 1.76.

Twenty sets of annotations were generated on a complete binary tree of depth 7 for each configuration of (n, r, γ), where n, r and γ range from $[40, 200]$, $[2, 20]$, $[0.2, 0.9]$ respectively with a uniformly distributed ω_0. We evaluated the mean values of the correlation by changing one parameters while fixing the other two.

Figure 4.2 shows the the average τ for a variable number n of gene products using $r = 15$ and $\gamma = 0.6$. In Figure 4.2, the average correlation for Cosine increases with the number of annotations n, while the four other measures are not affected by n. Also, we notice that when $n > 170$, further increase of n will not increase τ for any measure very much.

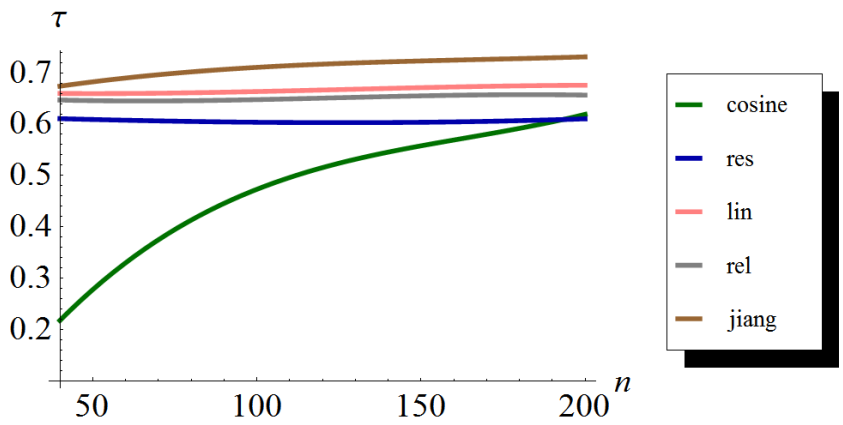

Figure 4.2: Average τ of each similarity measure with respect to n the number of distinct gene product when fixing r and γ ($n \in [40, 200], r = 15, \gamma = 0.6$).

Figure 4.3 shows the results for variable γ when $n = 200$ and $r = 8$. For $\gamma < 0.65$, the correlation for Jiang's measure decreases with growing γ. In contrast, τ for Cosine increases with growing γ. Also, the change of γ does not influence the correlation for other three measures. When $\gamma > 0.65$, τ for every measure begins to decrease with increasing γ, especially for Cosine, which decreases dramatically.

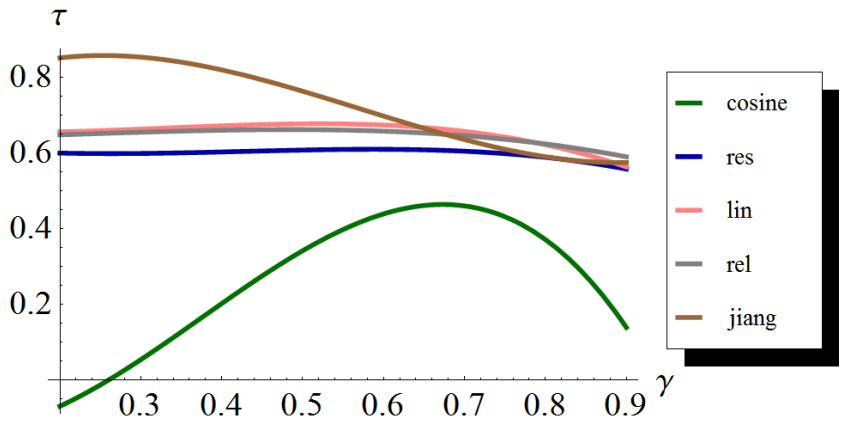

Figure 4.3: Average τ of each similarity measure with respect to γ when fixing n and r ($n = 200, r = 8, \gamma \in [0.2, 0.9]$).

In Figure 4.4, 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 4.2, correlations for Resnik's, Lin's, and Rel change little with increasing r, Jiang's decreases slightly, and the correlation for Cosine increases significantly.

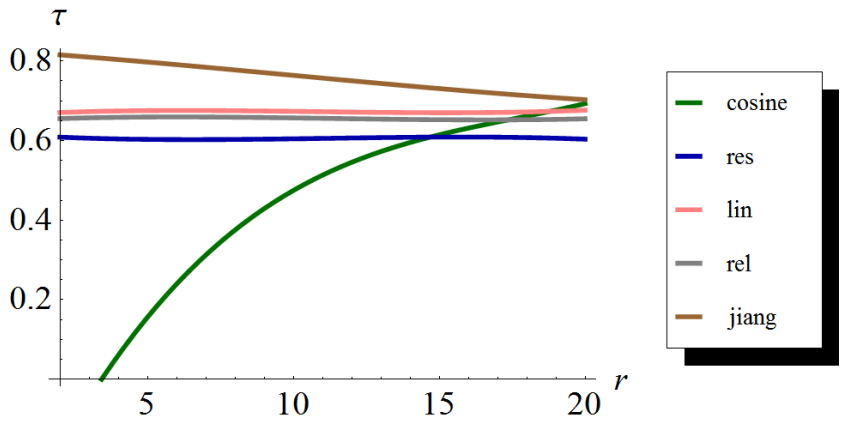

Figure 4.4: Average τ of each similarity measure with respect to r the number of terms associated with each gene product when fixing n and γ ($n = 200, r \in [2, 20], \gamma = 0.6$).

From the three figures, we can see that γ affects τ of all similarity measures, though less so for Lin's, Rel, and Resnik's. A gene product can be associated with a

number of distinct terms, and γ defines how sparse the annotation of a gene product is distributed in the ontology. A small γ indicates that the gene product has been annotated by several terms close to each other. Results show that Cosine correlates more when $\gamma \approx 0.65$ while the correlation for the other four increases when γ 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 γ and r, the number of gene products n has limited influence on the correlation. Generally, higher τ 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 γ 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.

4.1.2 Non-Uniformly Distributed Number of Annotations

To understand how skewed popularity in gene product impacts the semantic similarity measures, ten sets of annotations were generated on each configuration of $n = 100$, $\gamma = 0.3$ and p (see Section 3.1.4), whose values ranged from 0.1 to 0.9, on a complete binary tree of depth 7. In Figure 4.5, we show the average value of τ 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.

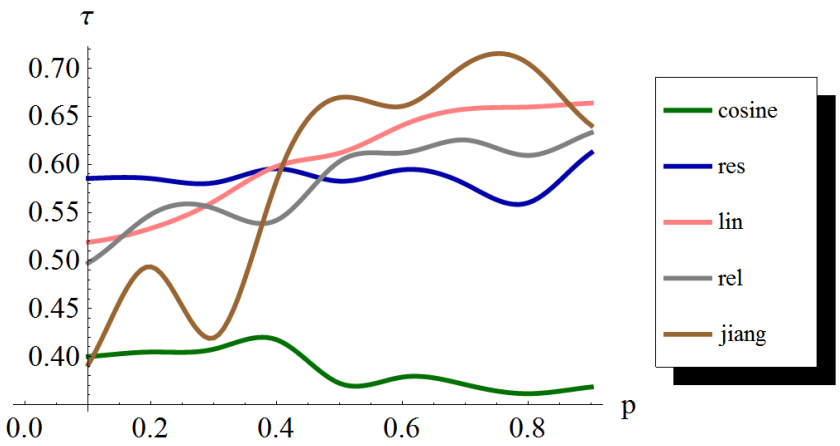

Figure 4.5: Average value of τ based on variable number of annotations r geometrically distributed with parameter p ($n = 100$, $\gamma = 0.3$).

To understand how heavily used terms impacts the semantic similarity measures, 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 ω_0. Figure 4.6 shows the impact of ω_0's normalized entropy on τ. We can see that increasing H_0 (making ω_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.

From these results we can see that Cosine and Jiang's are not well suited for skewed data (with a low-entropy ω_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

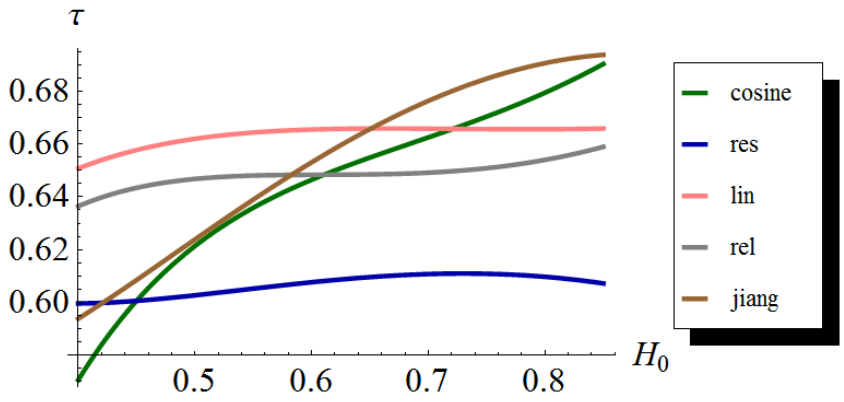

Figure 4.6: Average value of τ versus the normalized entropy H_0 of the starting distribution ω_0 ($n = 200$, $\gamma = 0.6$, $r = 5$).

Resnik's, Lin's and Rel (especially Resnik's) are more stable across many parameter values.

4.2 Analysis on Real Ontologies

4.2.1 Analysis on Partial Ontology

We used a subset of 25593 annotations from UniProtKB 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.

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 4.1 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 Sec-

tion 4.1.2), we can use their correlations from our tests on synthetic data as a baseline for our experiments here. The results ($\tau \approx 0.6$) for these three measures from Table 4.1 are very close to the baseline suggested by Figures 4.2–4.4. This leads us to believe that this partial ontology correlates well to its usage.

Table 4.1: Comparison of τ 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

4.2.2 Analysis on Full Gene Ontology

Our experiment on the full ontology was performed on a copy of GO annotations dated April 2010, which consisted of $32,651,844$ annotations of $6,729,320$ gene products using terms from three ontologies (see Table 4.2). There are $43,645$ relations defined over the $26,664$ 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.

Tables 4.3–4.5 present the τ 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

Table 4.2: 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

vs. Lin and Res vs. Rel, which correlate near 0.99. This is unsurprising given the relationships among the definitions of these measures.

Table 4.3: Estimated τ 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 4.4: Estimated τ 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				

From Section 4.1, we understand how values for n, r, γ, 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 γ 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

Table 4.5: Estimated τ 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				

of the initial distribution) and γ are generally low in the real data. Specifically, we use $H_0(\omega)$ as an upper bound of $H_0(\omega_0)$. Table 4.6 shows values of the relevant parameters in GO; γ is omitted and instead is qualitatively estimated as "low", since Table 4.6 gives $H_0(\omega)$ as relatively low, ranging from 0.44 to 0.58.

Table 4.6: 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 4.4 and 4.5 for synthetic data of similar characteristics.

Table 4.6 gives low $H_0(\omega)$ from 0.44 to 0.58, which suggests that both γ 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 γ in Figure 4.3, 0.45 given by $p \approx 0.25$ in Figure 4.5 or 0.6 given by $H_0(\omega_0)$ around 0.4 in Figure 4.6.

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.

4.3 Result on Unified Similarity Measure

Based on results in Section 4.1, the Resnik measure is the most stable measure of those we tested according to parameter sensitivity analysis. However, one of the limitations of the Resnik measure is that it only works with term pairs within the same ontology. That is the motivation for our unified similarity measure from Section 2.1.3. But before we can apply our unified similarity measure to search for cross terms, we should validate that the unified measure behaves in a way similar to that of Resnik. In Section 4.3.1, we compare the Resnik measure with our unified similarity measure to see how well it performs within the same ontology.

The degree of co-occurrence between term pairs reflects how terms are used together with each other. Term pairs frequently used together could indicate a relationship. In Section 4.3.2, we compare the degree of co-occurrence with our similarity measure to find out how the measure correlates with term usage.

4.3.1 Comparison with Direct Measure

Since SGD is a relatively small database compared to full GO, we can directly compare our unified measure to Resnik without sampling. For each ontology in SGD, we compute the similarity using both Resnik method and our method for each pair of terms in Table 3.1. That is, we compute the similarity via Resnik method and our method for each pair of terms from BP, CC and MF and organize them into two lists L_{Resnik} and L_{Uni}. We then measured both rank correlation and linear correlation between the two lists.

The results, shown in Figure 4.7, show that there is a good correlation. We can see that when the similarity for unified measure increases the direct measure increases, especially when the similarity value is below 4. Since there are very few term pairs in the biological process ontology with similarity values over 5, we see a slower increase beyond value 5.

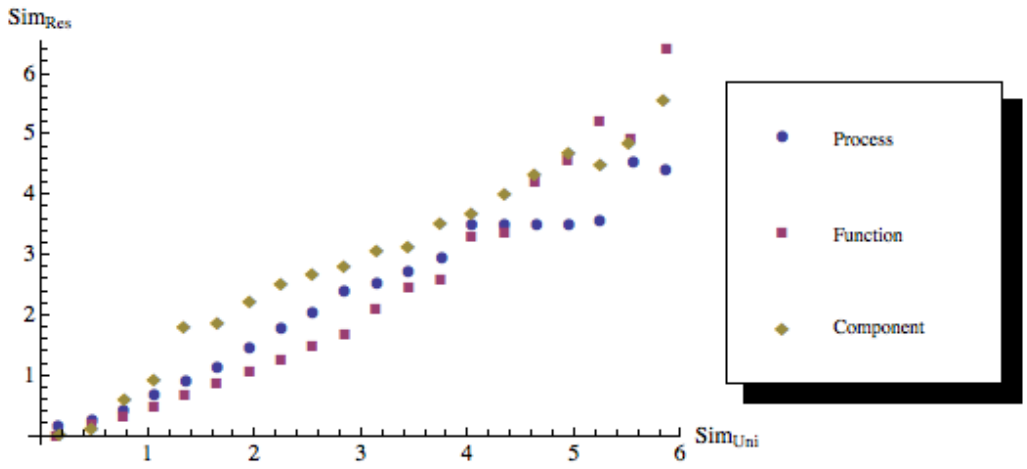

Figure 4.7: Comparing semantic similarity between Resnik's measure and unified measure in SGD.

Since the unified measure is in the same scale as Resnik's direct measure, in addition to Kendall's τ, we compute the linear correlation coefficient ρ between the two as well. Table 4.7 shows the correlation values between the two measures. From the table we can see that $\rho > 0.5$ in all three ontologies. Thus we can see that our unified measure correlates with Resnik, which evaluates semantic similarity between terms within the same ontology.

4.3.2 Correlation in the Co-occurred Term Pairs

In SGD the 6353 gene products have 41284 annotations, and gene products are usually labeled with multiple terms. The terms appearing together could be correlated. If

Table 4.7: Correlation value between Resnik's measure and unified measure in each ontology of SGD using Kendall's τ and Pearson's linear correlation coefficient ρ

Ontology	τ	ρ
Molecular Function	0.4452	0.6326
Cellular Component	0.3915	0.5801
Biological Process	0.3517	0.5093

they are correlated, we expect them to have higher semantic similarity values. Since there are not very many of them, we measure the degree of co-occurrence between every pair of term terms using the method described in Section 3.3.2 and examine only those with degree greater than zero. Table 4.8 shows the average semantic similarity values for these term pairs. In order to get a sense on how large these similarity values can be, we use the similarity between two uniformly randomly chosen term pairs selected from all term in the three oncologies as a baseline. Average similarity over 100000 arbitrary chosen term pairs shows a value of 1.8312. Table 4.8 shows that the average similarity values are much larger than 1.8312.

Table 4.8: Average semantic similarity for co-occurred term pairs grouped by source ontology.

Ontology	Number of pairs	Average Similarity
Biological Process	20008	4.2284
Cellular Component	5137	3.3822
Molecular Function	2707	5.7382
BP and CC	23211	3.4174
BP and MF	18719	4.515
CC and MF	11725	3.250
Total	81507	3.922

When evaluating co-occurrence results, we consider the results involving pairs from the same ontology separately from pairs from different ontologies. The first three rows in Table 4.8 show results for pairs from the same ontology. We can read that term pairs from molecular function have significantly higher similarity values. This could

mean that terms from molecular function have a higher tendency to be used in groups. The second three rows show results for pairs from different ontologies, which we refer to as *cross ontology term pairs*. We can see that term pairs between biological process and molecular function have higher correlation values. Quick examination on these term pairs shows that they describe related activities from different perspectives.

For example GO:0006864 "pyrimidine nucleotide transport" from biological process and GO:0015218 "pyrimidine nucleotide transmembrane transporter activity" from molecular function both describe the transfer of a pyrimidine nucleotide. The former refers to the directed movement process itself while the latter focuses on the catalysis function from labelled gene products in the process. Similarly, we have GO:0008277 "regulation of G-protein coupled receptor protein signaling pathway" and GO:0005057 "receptor signaling protein activity" from molecular function. The two terms both describe the process that proteins pass signals. GO:0005057 mainly refers to the gene product's function that can convey a signal and trigger another state change or activity, while GO:0008277 focuses on the protein's role in the modulation process of such signaling pathway.

4.3.3 Correlation with Degree of Co-occurrence

In the previous section, we have shown that co-occurred term pairs usually have high semantic similarity compared with uniformly randomly selected term pairs. Term pairs that appear more frequently should have higher semantic similarity. To test this, we compare the degree of co-occurrence with unified similarity for all 81507 pairs in Table 4.8, where the degree is in the interval of $(0, 1]$ by Equation 3.5. Figure 4.8 plots the degree of co-occurrence against the unified similarity measure. From this figure, we can see that the average degree of co-occurrence grows with semantic similarity

when the similarity is over 4. When the similarity is lower than 4, the degree of co-occurrence is 0, which means such term pairs have been never be used together.

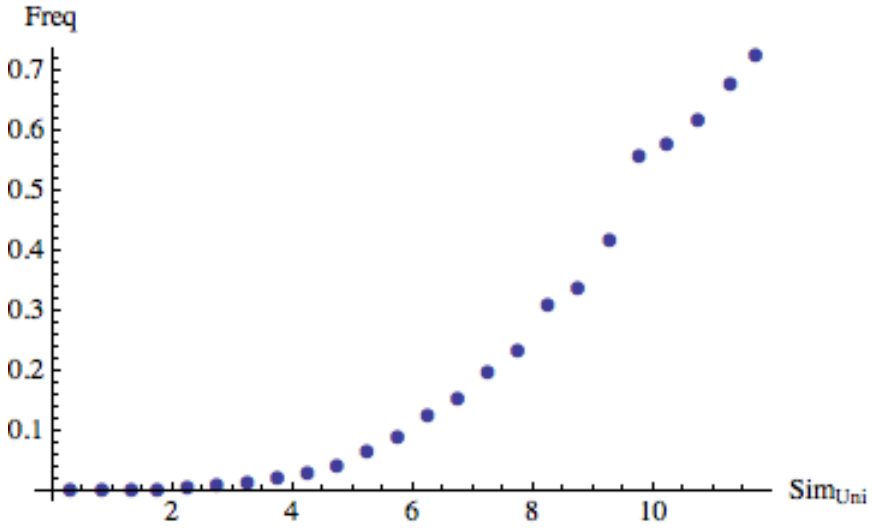

Figure 4.8: Comparing degree of co-occurrence and unified semantic similarity. ($\tau = 0.6603$, $\rho = 0.6919$)

Kendall's τ and Pearson's correlation coefficients between degree of co-occurrence and semantic similarity show values of $\tau = 0.6603$ and $\rho = 0.6919$ respectively. This indicates a strong correlation between degree of co-occurrence and our semantic similarity measure. The similarity given by our measure for three ontologies is correlated with their direct measure. Terms used together have higher average semantic similarity, which grows with their degree. These correlations serve as validation of our unified similarity measures.

4.3.4 Cross Ontology Term Pairs

Using the unified similarity measure, we can evaluate the semantic similarity for term pairs from different ontologies. Term pairs with low similarity are more likely to be correlated by accident than the others. From Section 4.3.2, we see there are

linear correlations between similarity values and degrees of co-occurrence. The linear correlation should increase if we choose only term pairs with higher similarity values. In order to identify those highly correlated cross ontology term pairs, we need to set a minimum similarity threshold. The threshold can filter out low similarity term pairs. For each of the 11 threshold values between 0 and 10, we plot the histogram of $Sim_{Uni}/Freq$ for term pairs above it. Figure 4.9 shows the distribution of the term pairs by choosing different similarity thresholds. The X-axis shows the ratio between similarity value and the degree of co-occurrence. The Y-axis shows the number of cross ontology term pairs for a given ratio. The number of pairs in the peak values can be an order of magnitude higher than other X values, and the ratio can go very high when the degree of co-occurrence is low. Thus, in order to better visualize the data, we use log scale in both axes.

The wider the range of the ratio is, the lower the linear correlation between the similarity measure and degree of co-occurrence. If the minimal similarity threshold does not matter, we expect to see an identical distribution interval for all threshold values. However, from the figure we can see that by increasing the threshold the range of the distribution decreases, which means fewer accidental correlated term pairs. When the threshold is lower than 6, the increase in threshold decreases the range dramatically. In contrast, the range does not change significantly with the threshold when the threshold is beyond 6. Thus, to ensure enough term pairs while also maintaining the quality, we choose a threshold of 6 to filter cross ontology term pairs. We consider these terms pairs to be *highly correlated cross ontology terms*.

Now we have identified a set of cross ontology term pairs[1] which have high semantic similarity (a similarity of 6.0 which is much higher than the 1.8312 average). We also

[1]The complete list of cross ontology term pairs (all pairs under curve threshold 0) can be downloaded from `http://cse.unl.edu/~ymo/thesis/yeast-pair.zip`

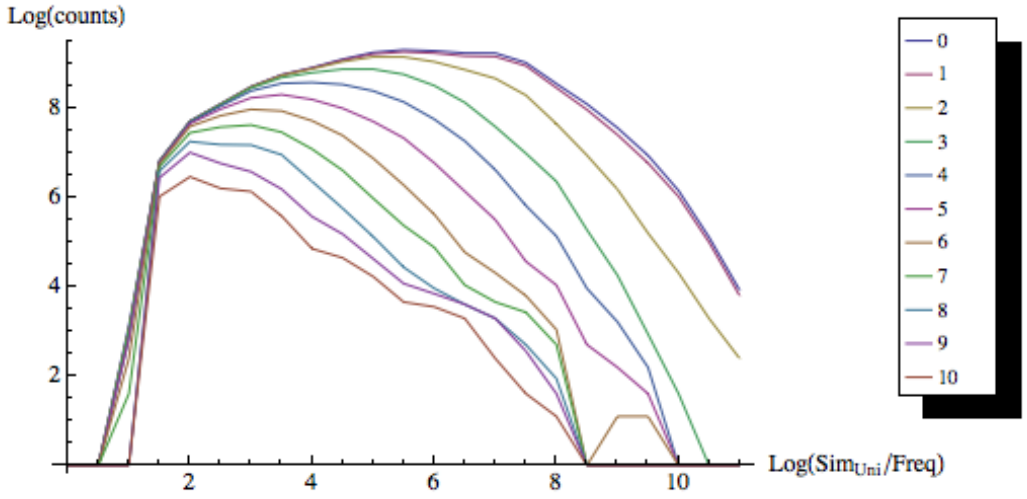

Figure 4.9: Distribution of the ratio of unified similarity and degree of co-occurrence by choosing different minimal similarity threshold (from 0 to 10).

would like to review a few of them case by case. Term pairs on the two sides of the curves correspond to pairs which have either high similarity low occurrence or low similarity high occurrence. These types of terms pairs occurred much less often than pairs near the peak of the distribution. In this study, we want to investigate typical term pairs only. Figure 4.9 suggests that a majority of term pairs has ratio around 2 in log scale. We consider term pairs around this ratio as typical.

Appendix A lists a few highly correlated pairs around the peak. After examining these pairs, we found it makes good sense biologically for them to achieve a high correlation values. The pairs in the list can be generally classified into three categories: 1) cellular component term that supports a specific molecular function correlates with the term for that molecular function; 2) biological process term that consists of a series of molecular function correlates each individual molecular function terms; 3) biological process term that takes place in specific cellular component correlates with

term for the cellular component. Results show a very high degree of actual correlation between the term.

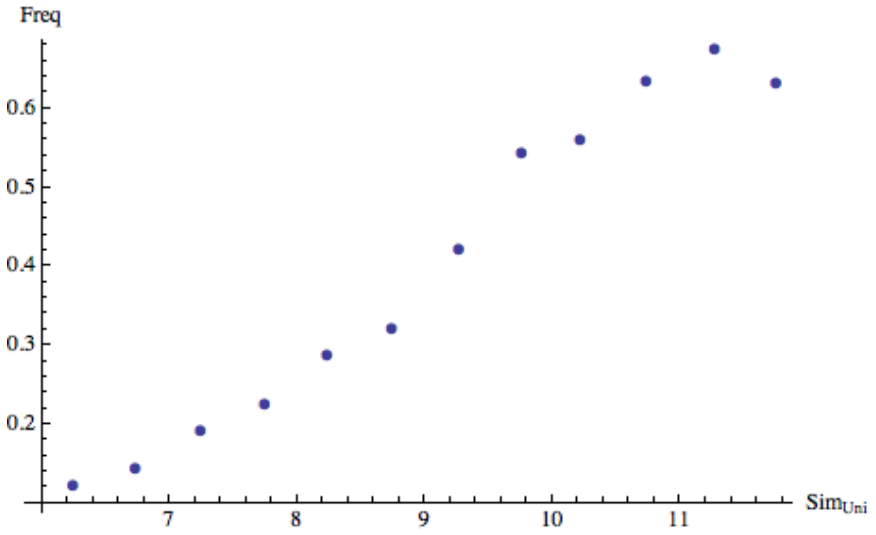

Figure 4.10: Comparing degree of co-occurrence and unified semantic similarity for identified cross term pairs. ($\tau = 0.6661$, $\rho = 0.6862$)

Similarly, we can examine how well of our unified measure correlates the degree of co-occurrence on this portion of cross ontology term pairs. Figure 4.10 demonstrates the correlation between degree of co-occurrence and unified semantic similarity for these cross term pairs. From the figure, we can read that cross ontology term pairs, which are identified as high semantic similarities using our method, are co-occurred more frequently than others. This shows the similarity for highly correlated term pairs also correlated with the degree of co-occurrence.

By examining the list of highly correlated term pairs, we found that there are still many term pairs with value zero in the degree of co-occurrence because they are not co-occurred in SGD. Admittedly, these pairs need to be verified against other metrics beyond SGD. Also, the term pairs we identified are domain-specific because of the SGD database we are using. Thus, these terms should be considered as similar only within the scope of the yeast genome. In order to overcome these difficulties,

additional databases need to be introduced. We will investigate them in the future research.

Chapter 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 ω_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.

From our 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.

It is also unknown whether terms from different aspects of GO correlate with each other. Although the Resnik measure is the most stable measure to identify similar term pairs according to parameter sensitivity analysis, the Resnik measure is limited only to measure term pairs within the same ontology. Hence we proposed a unified similarity measure which can compute semantic similarity between any two terms either within the same ontology or across ontologies.

By comparing with direct measure which evaluates semantic similarity between terms within the same ontology, we can see that our unified measure correlates with the existing direct measure. Then we compare it with the degree of co-occurrence, which reflects the likelihood of two terms to annotate a same gene product, for pairs in SGD. Results show that term pairs with degree greater than zero have much higher similarity values than others. We noticed that the level of similarity correlates well with the degree of co-occurrence. This evidence shows our unified similarity measures are suitable to evaluate term similarity.

We further identified a list of highly correlated cross ontology term pairs by thresholding the unified semantic similarity values. After examining a few of them with high similarity values (see Appendix A), we found high level of biological correlation between these terms. For instance, when a term from cellular component ontology correlates another term from biological process ontology, the former term defines physical cell structure which supports the chemical reaction defined in the latter term. These relations can be useful to refine term definition.

Appendix A

List of a Few Highly Correlated Cross Ontology Term Pairs

The correlation value between cross ontology term pairs are above 6 according to Section 4.3.4 using the unified similarity measure proposed in Section 3.3.1. This list shows a few top hits the ratio of which between the semantic similarity and degree of co-occurrence (defined in Section 3.3.2) is around 7.38 suggest by Figure 4.9. The label BP, CC and MF in term ID represent term's source biological process, cellular component and molecular function respectively.

Term1 ID	Term2 ID	Term1 Name	Term2 Name
BP GO:0051123	CC GO:0005669	RNA polymerase II transcriptional preinitiation complex assembly	transcription factor TFIID complex
BP GO:0032568	CC GO:0005669	general transcription from RNA polymerase II promoter	transcription factor TFIID complex

CC GO:0070860	MF GO:0010843	RNA polymerase I core factor complex	promoter binding
CC GO:0017053	MF GO:0016565	transcriptional repressor complex	general transcriptional repressor activity
MF GO:0034246	BP GO:0006391	mitochondrial transcription initiation factor activity	transcription initiation from mitochondrial promoter
MF GO:0003840	BP GO:0042908	gamma-glutamyltransferase activity	xenobiotic transport
BP GO:0006550	MF GO:0004148	isoleucine catabolic process	dihydrolipoyl dehydrogenase activity
BP GO:0006550	MF GO:0004738	isoleucine catabolic process	pyruvate dehydrogenase activity
BP GO:0006574	MF GO:0004148	valine catabolic process	dihydrolipoyl dehydrogenase activity
BP GO:0006574	MF GO:0004738	valine catabolic process	pyruvate dehydrogenase activity
BP GO:0042743	MF GO:0004148	hydrogen peroxide metabolic process	dihydrolipoyl dehydrogenase activity
BP GO:0042743	MF GO:0004738	hydrogen peroxide metabolic process	pyruvate dehydrogenase activity
MF GO:0070463	BP GO:0070462	tubulin-dependent ATPase activity	plus-end specific microtubule depolymerization
MF GO:0015129	BP GO:0015727	lactate transmembrane transporter activity	lactate transport

MF GO:0034202	BP GO:0034203	glycolipid-translocating activity	glycolipid translocation
CC GO:0043626	MF GO:0030337	PCNA complex	DNA polymerase processivity factor activity
MF GO:0051575	BP GO:0071047	5'-deoxyribose-5-phosphate lyase activity	polyadenylation-dependent mRNA catabolic process
CC GO:0030678	BP GO:0001682	mitochondrial ribonuclease P complex	tRNA 5'-leader removal
MF GO:0034084	BP GO:0034210	steryl deacetylase activity	sterol deacetylation
CC GO:0000802	MF GO:0032184	transverse filament	SUMO polymer binding
MF GO:0015505	BP GO:0015857	uracil:cation symporter activity	uracil transport
CC GO:0005962	MF GO:0004449	mitochondrial isocitrate dehydrogenase complex (NAD+)	isocitrate dehydrogenase (NAD+) activity
CC GO:0005950	MF GO:0004049	anthranilate synthase complex	anthranilate synthase activity
CC GO:0009328	BP GO:0006432	phenylalanine-tRNA ligase complex	phenylalanyl-tRNA aminoacylation
BP GO:0006830	MF GO:0000006	high-affinity zinc ion transport	high affinity zinc uptake transmembrane transporter activity

BP GO:0015879	MF GO:0015226	carnitine transport	carnitine transporter activity
BP GO:0015890	MF GO:0015663	nicotinamide mononucleotide transport	nicotinamide mononucleotide transmembrane transporter activity

Bibliography

[1] F Al-Shahrour, R Diaz-Uriarte, and J Dopazo. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, February 2004.

[2] M Ashburner, C A Ball, and J A Blake. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25:25–29, 2000.

[3] C Batini and M Lenzerini. A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys (CSUR)*, 1986.

[4] D. Christensen. Fast algorithms for the calculation of Kendall's . *Computational Statistics*, 20:51–62, 2005.

[5] TR Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 1993.

[6] TR Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies*, 1995.

[7] David P Hill, Barry Smith, Monica S McAndrews-Hill, and Judith A Blake. Gene Ontology annotations: what they mean and where they come from. *BMC Bioinformatics*, 9(Suppl 5):S2, 2008.

[8] E Jain, A Bairoch, and S Duvaud. Infrastructure for the life sciences: Design and implementation of the UniProt website. *BMC Bioinformatics*, 10:136, 2009.

[9] J Jiang and D Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of ROCLING X*, page 9008, 1997.

[10] KS Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

[11] Y Kalfoglou. IF-Map: An ontology-mapping method based on information-flow theory. *Journal on data semantics I*, 2003.

[12] MG Kendall. *Rank Correlation Methods*. Griffin, 1970.

[13] IS Kohane, P Greenspun, and J Fackler. Building national electronic medical record systems via the World Wide Web. *... the American Medical ...*, 1996.

[14] J Köhler, K Munn, A Rüegg, and A Skusa. Quality control for terms and definitions in ontologies and taxonomies. *BMC ...*, 2006.

[15] JA Larson and SB Navathe. A theory of attributed equivalence in databases with application to schema integration. ..., page 243, 1989.

[16] SG Lee and JU Hur. A graph-theoretic modeling on GO space for biological interpretation of gene clusters. *Bioinformatics*, 2004.

[17] D Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, 1998.

[18] P W Lord, R D Stevens, A Brass, and C A Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, July 2003.

[19] Meeta Mistry and Paul Pavlidis. Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, 9(1):327, 2008.

[20] Gene Ontology. Go annotation policies and guidelines. `http://www.geneontology.org/GO.annotation.shtml`, 2011.

[21] T Ozaki and T Ohkawa. Efficient Discovery of Closed Hyperclique Patterns in Multidimensional Structured Databases. In *Data Mining Workshops, 2009. ICDMW '09. IEEE International Conference on*, pages 533–538. IEEE Computer Society, 2009.

[22] Catia Pesquita, Daniel Faria, Hugo Bastos, Antonio Ferreira, Andre Falcao, and Francisco Couto. Metrics for GO based protein semantic similarity: A systematic evaluation. *BMC Bioinformatics*, 9(Suppl 5):S4, 2008.

[23] 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*, 3:263–274, 2006.

[24] Wolfram Research. Kendall rank correlation. `http://reference.wolfram.com/mathematica/MultivariateStatistics/ref/KendallRankCorrelation.html`, 2011.

[25] P Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

[26] 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*, 7:302, 2006.

[27] Andreas Schlicker, Carola Huthmacher, Fidel Ramirez, Thomas Lengauer, and Mario Albrecht. Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics*, 23(7):859–865, 2007.

[28] B Smith. Ontology: Towards a new synthesis. *Formal Ontology in Information Systems*, 2001.

[29] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, November 2007.

[30] L. Stein. Genome annotation: from sequence to biology. *Nature reviews genetics*, 2001.

[31] L Yue. Pathway and ontology analysis: emerging approaches connecting transcriptome data and clinical endpoints. *Current molecular medicine*, 2005.