Ontology Learning and Knowledge Discovery Using the Web:

Challenges and Recent Advances

Wilson Wong
The University of Western Australia, Australia

Wei Liu
The University of Western Australia, Australia

Mohammed Bennamoun
The University of Western Australia, Australia

Information Science
REFERENCE

Senior Editorial Director: Kristin Klinger
Director of Book Publications: Julia Mosemann
Editorial Director: Lindsay Johnston
Acquisitions Editor: Erika Carter
Development Editor: Joel Gamon
Production Editor: Sean Woznicki
Typesetters: Jennifer Romanchak and Mike Brehm
Print Coordinator: Jamie Snavely
Cover Design: Nick Newcomer

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

Library of Congress Cataloging-in-Publication Data

Ontology learning and knowledge discovery using the Web: challenges and recent advances / Wilson Wong, Wei Liu and Mohammed Bennamoun, editors.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-60960-625-1 (hardcover) -- ISBN 978-1-60960-626-8 (ebook) 1. Ontologies (Information retrieval) 2. Data mining. I. Wong, Wilson, 1981- II. Liu, Wei, 1972 July 14- III. Bennamoun, M. (Mohammed)
TK5105.88815.O587 2011
006.3’12--dc22
2010043008

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 9
GO-Based Term Semantic Similarity

Marco A. Alvarez
Utah State University, USA

Xiaojun Qi
Utah State University, USA

Changhui Yan
North Dakota State University, USA

ABSTRACT

As the Gene Ontology (GO) plays more and more important roles in bioinformatics research, there has been great interest in developing objective and accurate methods for calculating semantic similarity between GO terms. In this chapter, the authors first introduce the basic concepts related to the GO and then briefly review the current advances and challenges in the development of methods for calculating semantic similarity between GO terms. Then, the authors introduce a semantic similarity method that does not rely on external data sources. Using this method as an example, the authors show how different properties of the GO can be explored to calculate semantic similarities between pairs of GO terms. The authors conclude the chapter by presenting some thoughts on the directions for future research in this field.

GENE ONTOLOGY AND GENE ONTOLOGY ANNOTATION

The most successful effort for systematically describing current biological knowledge is the GO project (Ashburner et al., 2000), which maintains a dynamic, structured, precisely defined, and controlled vocabulary of terms for expressing the roles of genes and gene products. The GO is dynamic in the sense that its structure changes as more information is available. The GO consists of three different ontologies describing: 1) biological processes (BP), where a process often involves a chemical or physical transformation (e.g. cell growth); 2) molecular functions (MF), where functions are defined as the biochemical activity of gene products (e.g. enzyme); and 3) cellular components (CC), which refers to places in the cell where gene products are active (e.g. nuclear membrane). Each ontology contains nodes (GO

DOI: 10.4018/978-1-60960-625-1.ch009

terms) linked to each other through *"is-a"* or *"part-of"* relationships forming a directed acyclic graph. Such organization enables the retrieval and visualization of biological knowledge at different levels.

The Gene Ontology Annotation (GOA) project (Barrell et al., 2009) at the European Bioinformatics Institute (EBI) is a project that aims to provide high-quality electronic and manual associations (annotations) between GO terms and UniProt KnowledgeBase (UniProtKB) entries (Consortium, 2009). Crucial to this project is the integration of different databases, a problem that has been addressed by the GO project. The GO maintains a common vocabulary of terms that can be applied to all organisms enabling the annotation across species and databases. The GOA project associates GO terms to UniProtKB entries using strictly controlled manual and electronic methods where every association is supported by a distinct evidence source. A protein can be annotated to multiple GO terms from any of the three ontologies in GO. Functional annotations of UniProtKB proteins currently consists of over 32 million annotations to more than 4 million proteins (Barrell et al., 2009).

SEMANTIC SIMILARITY BETWEEN GENE ONTOLOGY TERMS

The calculation of semantic similarity between pairs of ontology terms aims to capture the relatedness between the semantic content of the terms. Researchers have made great efforts to develop objective and accurate methods to calculate term semantic similarity. For example, semantic similarity between concepts has been a central topic in natural language processing where several robust methods have been proposed based on the WordNet ontology (Budanitsky & Hirst, 2006). In recent years, ontologies have grown to be a popular topic in the biomedical research community creating a demand for computational methods that can exploit their hierarchical structure, in particular, methods for calculating semantic similarity between terms in the GO. Such methods are designed to reflect the closeness or distance between the semantic content of the terms, in other words, their biological relationships.

Additionally, semantic similarity methods can easily be extended to infer higher level semantic relationships. For example, at the protein level, scores for a given protein pair can be calculated by combining the pairwise semantic similarities for the GO terms associated with the proteins. These scores can be used in a broad range of applications such as clustering of genes in pathways (Wang, Du, Payattakool, Yu, & Chen, 2007, Sheehan, Quigley, Gaudin, & Dobson, 2008, Nagar & Al-Mubaid, 2008, Du, Li, Chen, Yu, & Wang, 2009), protein-protein interaction (Xu, Du, & Zhou, 2008), expression profiles of gene products (Sevilla et al., 2005), protein sequence similarity (Pesquita et al., 2008, Mistry & Pavlidis, 2008, Lord, Stevens, Brass, & Goble, 2003), protein function prediction (Fontana, Cestaro, Velasco, Formentin, & Toppo, 2009), and protein family similarity (Couto, Silva, & Coutinho, 2007). An armada of semantic similarity measures using the GO are available in the biomedical literature. A representative collection of available methods have been reviewed and categorized by (Pesquita, Faria, Falcão, Lord, & Couto, 2009).

SEMANTIC SIMILARITY BETWEEN GENE PRODUCTS

In the research related to biological ontologies, great interest has been seen in exploiting ontological annotations to estimate the relationship between gene products, particularly proteins. The use of ontological annotations to measure the similarities between gene products was first introduced in (Lord et al., 2003), where three different methods (Jiang & Conrath, 1997, Lin, 1998, Resnik, 1995) originally designed for the

WordNet ontology were evaluated under the biological context.

Bearing in mind that each gene product is annotated by multiple GO terms, we focus our attention on pairwise approaches that measure semantic similarities between gene products by combining the pairwise semantic similarities between their terms. Three methods have been widely used for this purpose: average (Lord et al., 2003), maximum (Sevilla et al., 2005), and the best match average (Couto et al., 2007, Schlicker, Domingues, Rahnenfuhrer, & Lengauer, 2006, Wang et al., 2007). Such methods can be explained using the following example.

Let P_k be a protein, and $A(P_k) = \{t_{k1}, t_{k2}, \ldots\}$ be the set of non-redundant GO terms that annotate P_k. Then, given two input proteins P_i and P_j with annotation sets $A(P_i) = \{t_{i1}, t_{i2}, \ldots, t_{im}\}$ and $A(P_j) = \{t_{j1}, t_{j2}, \ldots, t_{jn}\}$, we obtain the similarity matrix $M_{m \times n}$ where every $M(a,b)$ is the semantic similarity between GO terms t_{ia} and t_{jb}. The similarity matrix $M_{m \times n}$ is not necessarily square or symmetric since proteins may be annotated by any number of GO terms. Let's assume that $SSA(t_1, t_2)$ denotes the semantic similarity between GO terms t_1 and t_2 calculated by our method.

In the maximum method, the semantic similarity between proteins P_i and P_j is the largest term similarity found in the similarity matrix $M_{m \times n}$. This approach can be expressed as:

$$P_{max}(P_i, P_j) = \max_{1 \le a \le m, 1 \le b \le n} SSA(t_{ia}, t_{jb})$$

Similarly, the average method defines the semantic similarity between proteins P_i and P_j as the average over all term similarities in the matrix. This approach can be expressed as:

$$P_{avg}(P_i, P_j) = \frac{\sum_{a=1}^{m} \sum_{b=1}^{n} SSA(t_{ia}, t_{jb})}{m + n}$$

The best match average method is based on the different meanings of rows and columns in the similarity matrix $M_{m \times n}$. Note that the values in row a represent the similarities between term t_{ia} and all the terms in $A(P_j)$, while column b contains the similarities between term t_{jb} and all the terms in $A(P_i)$. Let's define the row maxima of a row of matrix $M_{m \times n}$ as the maximum value in that row and column maxima of a column of $M_{m \times n}$ as the maximum value in that column. Then, the vector consisting of all row maxima represents the best hits when comparing one protein with the other, and the vector consisting of all column maxima represents the best hits when comparisons are made in the other direction. The best match average method calculates the averages of both vectors and then takes the average of them. This method is summarized below:

$$row_{max}(P_i, P_j) = \frac{1}{m} \cdot \sum_{a=1}^{m} \max_{1 \le b \le n} SSA(t_{ia}, t_{jb})$$

$$col_{max}(P_i, P_j) = \frac{1}{n} \cdot \sum_{b=1}^{n} \max_{1 \le a \le m} SSA(t_{ia}, t_{jb})$$

$$P_{bma}(P_i, P_j) = \frac{row_{max}(P_i, P_j) + col_{max}(P_i, P_j)}{2}$$

According to (Pesquita et al., 2008), from a biological point of view there are limitations to the average and maximum approaches. Imagine two functionally identical proteins with more than one annotation. The average method will yield a similarity below 1.0 because the average is calculated over all the pairwise combinations. On

the other hand, the maximum approach can yield similarities of 1.0 even when the proteins are not functionally identical, because it ignores unrelated terms. In (Pesquita et al., 2008) the authors claimed that the best match average approach does not suffer from the above limitations, and accounts for both similar and dissimilar terms as expected biologically. In our experiments, we tested these three methods with our proposed semantic similarity algorithm. The results confirm that the best match average is a better way to combine GO term similarities to obtain semantic similarities between proteins.

EVALUATION OF METHODS FOR COMPUTING SEMANTIC SIMILARITY BETWEEN GO TERMS

Although there exist a few methods for calculating the semantic similarity between GO terms, the fair evaluation and comparison of these methods has proven very difficult. The main challenge is that there is no gold standard to compare with, more specifically, there is no well-accepted quantitative definition for semantic similarities between GO terms. An approach used in several studies is to calculate semantic similarities for a set of proteins and correlate the resulting semantic similarities with sequence similarity (Lord et al., 2003), Pfam similarity (Couto et al., 2007), protein interactions (Guo, Liu, Shriver, Hu, & Liebman, 2006), among others. In general, correlations are determined by the Pearson Correlation Coefficient, which is in the range of $[-1,1]$. Then the correlation is used as a measure to evaluate the performance of the proposed method. Better methods are expected to achieve higher correlation. Two different types of functional similarities have been used in our study for this purpose. In the first type, sequence similarity between proteins is used to estimate functional similarity. The foundation of this approach is that similar sequence leads to similar function. The second type of function similarity is based on the Pfam (Finn et al., 2008) annotations of proteins. Let families $F(P_i) = \{f_{i1}, f_{i2}, \ldots, f_{im}\}$ and $F(P_j) = \{f_{j1}, f_{j2}, \ldots, f_{jn}\}$ be the Pfam families that protein P_i and P_j are associated with respectively. Then the functional similarity between the two proteins is calculated similarly to (Couto et al., 2007), where the Jaccard coefficient between the two sets is defined as shown below:

$$P_{pfam}(P_i, P_j) = \frac{|F(P_i) \cap F(P_j)|}{|F(P_i) \cup F(P_j)|}$$

PREVIOUS METHODS FOR COMPUTING SEMANTIC SIMILARITY BETWEEN GO TERMS

Given two input terms, a semantic similarity algorithm returns a numerical score that quantifies the relatedness of the input terms. Based primarily on the structure of the GO and GOA annotations, several algorithms for estimating the semantic similarity of GO terms have been proposed in the biomedical literature. In one dimension, they can be roughly classified as node-based methods (Couto et al., 2007, Lord et al., 2003, Schlicker et al., 2006), edge-based methods (Cheng et al., 2004, Wu, Zhu, Guo, Zhang, & Lin, 2006), and hybrid methods (Wang et al., 2007, Othman, Deris, & Illias, 2008). In edge-based methods, the semantic similarity varies according to the shortest distance connecting the input terms, while in node-based methods the similarity is evaluated by comparing specific properties of the input terms, and optionally their ancestors. In a second dimension, they can be classified as intrinsic methods that only rely on the ontologies and external methods that depend on additional information from external data sources. For example, annotation databases

Table 1. Statistics for the Gene Ontology. For each of the ontologies we show respectively the number of GO terms, the number of "is-a" links, the number of "part-of" links, and the maximum depth.

Ontology	Terms	*"is-a"* links	*"part-of"* links	*max-depth*
Biological Process	16,819	27,532	3,446	14
Molecular Function	8,628	10,079	3	14
Cellular Location	2,416	3,670	941	10

like GOA can be used to calculate the frequency of annotation and/or information content scores for any node. Several existing semantic similarity algorithms include in their calculations information content measures determined from an external corpus. For instance, algorithms reported in (Pesquita et al., 2009) use, direct or indirectly, information content calculations proposed by (Resnik, 1995, Jiang & Conrath, 1997, Lin, 1998).

A limitation of methods that rely on external data sources is their sensitivity to changes in the involved corpus. If the lexical corpus is changed, the semantic similarity values will also change. In our context, GOA annotations are commonly used as a corpus. As the GOA is updated, the semantic similarity between the same pair of GO terms will also change. However, an ideal semantic similarity method should only rely on the ontologies and should not be affected by the change of the external corpus. In this regard, semantic similarity methods that are intrinsic to the ontologies should be the target of future research. In the following section, we will use an example to show how to develop a semantic similarity method that only relies on the ontologies.

A SEMANTIC SIMILARITY METHOD THAT DOES NOT RELY ON EXTERNAL DATA SOURCES

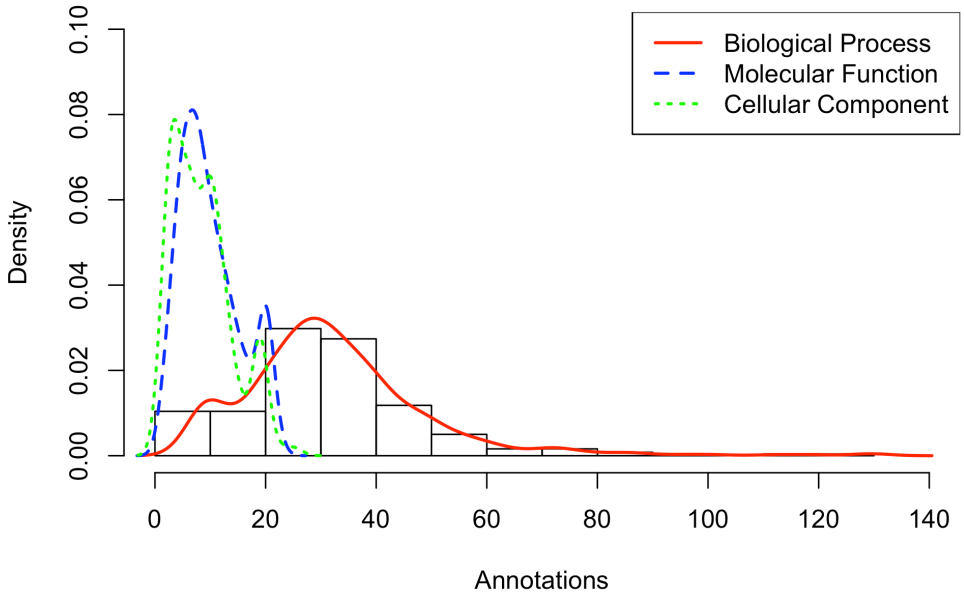
Dataset

For our experiments we downloaded the revision 1.723 of the GO. Table 1 shows the number of terms, *"is-a"* links, *"part-of"* links, and maximum depth of the three ontologies in such version of the GO.

Evaluation

For the purpose of evaluating our method, we downloaded the release 15.6 of the UniProtKB database, which is the most comprehensive and highly annotated publicly accessible protein sequence database, having recorded more than 6.6 million proteins through a combination of manual and electronic techniques. Then the annotations of the proteins in UniProtKB were extracted from the release 74.0 of GOA-UniProt, which provides the mapping of GO terms to UniProtKB entries. The mapping is done in the GOA project using both manual and electronic methods, both of which are strictly controlled to produce high-quality GO annotation and both require the involvement of biologists and software engineers (Barrell et al., 2009). Bear in mind that our semantic similarity method does not use the information in GOA-UniProt to calculate the semantic similarity. From the UniProtKB protein database, we selected the top 500 proteins with the highest number of annotations. We ensure that all the selected proteins have at least one annotation in GOA-UniProt for each of the three GO ontologies. In Figure 1 we can observe the density function for the number of annotations existing in our selected dataset of the top 500 most annotated proteins. In total we count 16,248 unique annotations for Biological Process, 5,029 for Molecular Function, and 4,298 for Cellular Component ontologies.

Figure 1. Histogram for the number of BP annotations considering the dataset of the top



500 most annotated proteins. The figure also shows the probability density estimate for the number of annotations in the BP, MF, and CC ontologies in the same dataset. Note how the proteins have considerably fewer MF and CC annotations than BP ones.

To evaluate our method, we calculated the pairwise semantic similarities between the top 500 proteins. For every pair of proteins we used our method to calculate pairwise semantic similarities between the GO terms annotating them and combined those similarities using the best match average method. Then, the resulting semantic similarities between the top 500 proteins were compared with the functional similarities between these proteins using the Pearson Correlation Coefficient (PCC). Higher PCC values imply a better semantic similarity method. The pairwise functional similarities between proteins were calculated based on the Pfam annotations associated with the proteins. For this purpose, when we selected the top 500 proteins, we also guaranteed that the selected proteins have at least one Pfam-A annotation by checking the online service available at http://pfam.sanger.ac.uk/protein/ provided by the Pfam database. According to recommendations found in (Yon Rhee, Wood, Dolinski, & Draghici, 2008), we excluded all annotations containing the *not*, *contributes_to*, and *colocalizes_with* qualifiers.

Semantic Similarity Method that only Considers the Shortest Path Between GO Terms on the Ontology

The first information that we explored to calculate semantic similarity between two GO terms was the shortest path that connects them on the ontology following either *"is-a"* and *"part-of"* links. We used the length of the path as a measure for the relatedness of the terms. The hypothesis is that if two GO terms are semantically similar, they should be close to each other in the ontology. The edges at different depths of the GO imply different distances in the biological setting, with the edges closer to the root implying longer distances than edges farther from the root. Thus, we assigned weights to edges based on the depths of

their endpoints in the ontology. For example an edge with endpoints t_1 and t_2 is given a weight as follows:

$$weight(t_i, t_j) = 1 - \frac{depth(t_i) + depth(t_j)}{2 \cdot max}$$

where $depth(t_i)$ and $depth(t_j)$ are their corresponding depths in the graph, and max is the maximum depth in the respective ontology. Due to multiple inheritance there are special cases where a given node can have different paths from the root with different lengths. We choose to consider the maximum depth possible which indicates the higher degree of specialization of the node.

The length of the path between two GO terms is defined as the sum of edge weights on the path. The shortest path is then the path with the smallest length. We transform the length of the path into a similarity measure using the following quadratic function so that shorter distances imply higher similarities:

$$spsim(t_i, t_j) = \left(\frac{sp(t_i, t_j)}{max} - 1 \right)^2$$

where $sp(t_i, t_j)$ is the length of the shortest path between node t_i and node t_j, and max is the maximum depth in the ontology. It can be easily proven that the similarity values are in the range of $[0,1]$. We use this function to express a similarity score between two GO terms as follows:

$$SSA(t_1, t_2) = spsim(t_1, t_2)$$

When this method was used to calculate semantic similarity between GO terms, the correlation between semantic similarities and functional similarities for the top 500 proteins was 0.668, 0.600 and 0.444 respectively when BP, MF and CC ontologies were considered.

Add the Depth of the Nearest Common Ancestor into the Semantic Similarity Method

The second property that we explored to calculate semantic similarity between two GO terms was the depth of the nearest common ancestor (NCA) for a pair of nodes in the ontologies. The hypothesis for this is that two semantically similar GO terms should share a long common path from the root and only branch at a place close to the bottom of the ontology. Based on this assumption, the deeper the NCA, the more similar the terms are. We used the following function to convert the depth of NCA into a similarity score:

$$nca(t_i, t_j) = \frac{dnca(t_i, t_j)}{max}$$

where $dnca(t_i, t_j)$ simply returns the depth of the NCA between terms t_i and t_j, and max is the maximum depth of the ontology. The output of $nca(t_i, t_j)$ is also in the interval $[0,1]$. Then, we combined this function with the length of the shortest path to develop a similarity score between two GO terms as follows:

$$SSA(t_1, t_2) = \frac{spsim(t_1, t_2) + nca(t_1, t_2)}{2}$$

When this function was used to calculated semantic similarity between GO terms, the correlation between semantic similarities and functional similarities for the top 500 proteins was 0.787, 0.765 and 0.510 respectively when BP, MF and CC ontologies were considered. Compared these results with those in the previous section, we can see that adding depth of NCA into the semantic similarity method significantly improve

the performance and the improvement is consist across the three ontologies.

CONSIDER THE SIMILARITY BETWEEN THE DEFINITIONS OF GO TERMS

On the GO, each term is associated with a definition which is a textual description of the term. If two GO terms are similar in semantics, they are very likely to share some common words in their definitions. Thus, we also explored the definition of GO terms to calculate semantic similarity between GO terms. First we defined the long definition of a term as the union of the terms' name and definition. We refined every long definition by removing common words (e.g. of, a, the) and applied the Porter algorithm (Porter, 1980) for stemming. Then, we created long definition vectors in a n dimensional ontological space. The value of n is the total number of unique stemmed words found in all long definitions for the respective ontology, which is $10,130$ for the biological process ontology, $12,979$ for the molecular function ontology and $5,884$ for the cellular component ontology. Every value in the long definition vector represents the *tf-idf* weight (term frequency-inverse document frequency) for the corresponding word. This weight evaluates how important a word is to its long definition. A high *tf-idf* weight is reached by words with high frequency in the long definition and with low frequency in the corpus (i.e. the collection of all long definitions in the respective ontology), therefore *tf-idf* weights tend to filter out common words. The similarity score is the cosine similarity defined by:

$$ld(t_i, t_j) = \frac{\overrightarrow{ld_i} \cdot \overrightarrow{ld_j}}{\| \overrightarrow{ld_i} \| \cdot \| \overrightarrow{ld_j} \|}$$

where $\overrightarrow{ld_i}$ and $\overrightarrow{ld_j}$ are the long definition vectors for terms t_i and t_j respectively.

We then combined all the three functions to obtain a score for the semantic similarity between two GO terms:

$$SSA(t_1, t_2) = \frac{spsim(t_1, t_2) + nca(t_1, t_2) + ld(t_1, t_2)}{3}$$

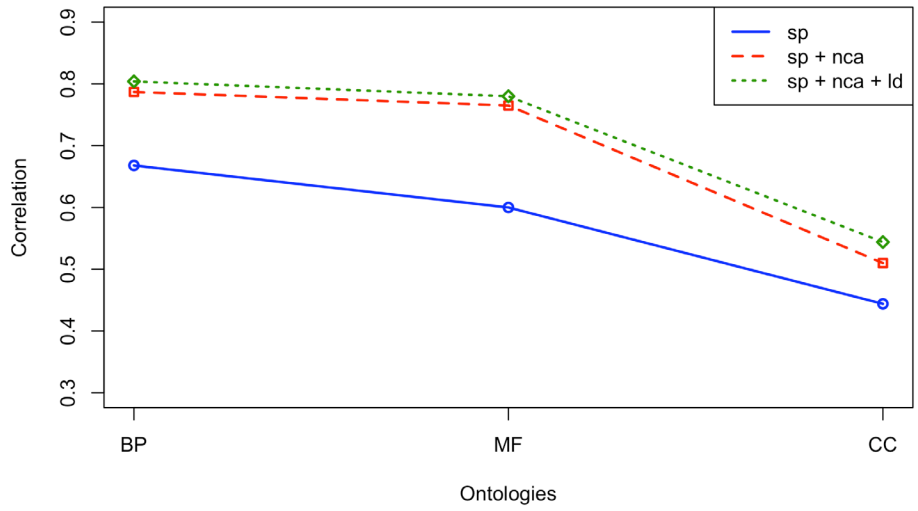
When this function was used to calculate semantic similarity, the correlation between semantic similarities and functional similarities for the top 500 proteins was increased to 0.804, 0.780 and 0.544 respectively for BP, MF and CC ontologies.

DISCUSSION AND FUTURE DIRECTIONS

Till this point, we have shown how to use different properties of the ontologies to calculate semantic similarity between GO terms. Figure 2, shows how the performance of the method is improved gradually by including more properties from the GO.

We also compared our method with other semantic similarity methods using the CESSM Web server available at http://xldb.fc.ul.pt/biotools/cessm developed by the XLDB research group at the University of Lisbon. CESSM currently implements 11 semantic similarity measures, all of which rely on external data sources. CESSM allows users to evaluate their semantic similarity algorithms using sequence similarities, which are calculated by means of RRBS (Pesquita et al., 2008), a relative measure of sequence similarity based on BLAST bitscores. Instead of using PCC, CESSM provides a resolution score which represents how well semantic similarities match sequence similarities. According to CESSM, resolution is the relative intensity where variations in

Figure 2. The performance of the semantic similarity method improves gradually as more properties of the ontologies are included. BP: biological process ontology; MF: molecular function ontology; CC; cellular component ontology; sp : the length of the shortest path between GO terms; nca : the depth of the nearest common ancestor; and ld : the similarity between long definitions.



the sequence similarity scale are translated into the semantic similarity scale. Higher resolution values mean that the semantic similarity method has higher capability to distinguish between different levels of protein function.

In the comparison, we only use the MF ontology because BP and CC ontologies present poor correlation with sequence similarity, as stated in (Lord et al., 2003). In this context, the best result achieved by the 11 methods in CESSM is a resolution score of 0.967 . In contrast, our method achieves a resolution score of 0.972 . In addition to the high performance, the key advantage of our method is that it is intrinsic to the ontology, that is, it does not rely on the external data sources in the calculation of semantic similarity.

Future research can be performed in the following directions in the search of better semantic similarity algorithms.

Different Weightings for Terms

In the method we developed, when we combined different properties of the ontologies to obtain a similarity score, equal weights are given to the three different terms: sp , nca , and ld . But it is possible to obtain a better combination by assigning different weights to the three terms. As a simple exploration, we tried every combination of the weights in the range of $[0,2]$ (with increments of 1) for the three terms. No combination can improve the performance consistently across the three ontologies. One possibility is that the relationships among the three terms are not linear. Thus, to find the optimal weighting, systematic research will be needed to explore the individual contribution of each term and the mutual dependency among these properties.

More Properties of the Ontology

Exploring other properties of the ontology may also help improve the performance in calculating

semantic similarity. For example, there are two different types of links between GO terms: *"is-a"* and *"part-of"*. In different studies, these two types of links have been used indiscriminately and represented by the same kind of edges. But since these links characterize different relationships, modeling separate types of links using different types of edges in the graph may better explore the knowledge in the ontologies.

Weighting Edges in the Ontology

Our results have shown that the length of the shortest path between two GO terms is a good indication of the semantic similarity between them. We assign weights to the edges based on the depth of their endpoints. Other strategies may be explored to assign biological weights to the edges of the ontologies. For example, weights corresponding to evolutionary distances may be considered. One simple way to do that is to assign weight to an edge by averaging the evolution distances between proteins associated with the two endpoints of the edge.

Evaluation Benchmarks

A very important but still unsolved problem in the development of semantic similarity measures is how to evaluate a semantic similarity method. Most researchers do this by comparing the semantic similarities between proteins given by a method with functional similarities between proteins. The functional similarities between proteins may be estimated using sequence similarity or annotations of the proteins in some functional databases. However, for a protein, there are many aspects that biologists are interested in. For example, the amino acid sequence, the 3D structure, and the evolutionary history of the protein. The semantic similarities between proteins given by a method should reflect all these aspects of a protein. Thus, an ideal evaluation should also compare the semantic similarity with the structural similarity, evolutionary distance, and other biological aspects between proteins. In addition to that, the correlation between semantic similarity and functional similarity (or other aspects) may not be linear. So Pearson Correlation Coefficient may not be the best way to compare them. Other methods should also be explored. For example, one can compare the rankings of all pairwise similarities based on semantic similarity with that based on functional similarity.

ACKNOWLEDGMENT

The authors would like to thank the XLDB Research Team of the University of Lisbon for providing an online tool for the evaluation of GO-based semantic similarity measures. In particular, we thank Catia Pesquita for all the kind support given for using their tool.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., & Cherry, J. M. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. doi:10.1038/75556

Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C., & Apweiler, R. (2009). The GOA database in 2009–an integrated gene ontology annotation resource. *Nucleic Acids Research*, *37*(1), D396–D403. doi:10.1093/nar/gkn803

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, *32*(1), 13–47. doi:10.1162/coli.2006.32.1.13

Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., & Kulp, D. (2004). A knowledge-based clustering algorithm driven by gene ontology. *Journal of Biopharmaceutical Statistics*, *14*(3), 687–700. doi:10.1081/BIP-200025659

Consortium, T. U. (2009). The universal protein resource (uniprot) 2009. *Nucleic Acids Research*, *37*(1), D169–D174. doi:10.1093/nar/gkn664

Couto, F. M., Silva, M. J., & Coutinho, P. M. (2007). Measuring semantic similarity between gene ontology terms. *Data & Knowledge Engineering*, *61*(1), 137–152. doi:10.1016/j.datak.2006.05.003

Du, Z., Li, L., Chen, C. F., Yu, P. S., & Wang, J. Z. (2009). G-sesame: Web tools for go-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Research*, *37*(2), W345–349.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., & Hotz, H. R. (2008). The pfam protein families database. *Nucleic Acids Research*, *36*(1), D281–D288.

Fontana, P., Cestaro, A., Velasco, R., Formentin, E., & Toppo, S. (2009). Rapid annotation of anonymous sequences from genome projects using semantic similarities and a weighting scheme in gene ontology. *PLoS ONE*, *4*(2), e4619. doi:10.1371/journal.pone.0004619

Guo, X., Liu, R., Shriver, C. D., Hu, H., & Liebman, M. N. (2006). Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics (Oxford, England)*, *22*(8), 967–973. doi:10.1093/bioinformatics/btl042

Jiang, J. J., & Conrath, D. W. (1997). *Semantic similarity based on corpus statistics and lexical taxonomy*. In International Conference Research on Computational Linguistics.

Lin, D. (1998). *An information-theoretic definition of similarity*. In International Conference on Machine Learning (pp. 296–304).

Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation. *Bioinformatics (Oxford, England)*, *19*(10), 1275–1283. doi:10.1093/bioinformatics/btg153

Mistry, M., & Pavlidis, P. (2008). Gene ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, *9*(1), 327. doi:10.1186/1471-2105-9-327

Nagar, A., & Al-Mubaid, H. (2008). *A new path length measure based on GO for gene similarity with evaluation using sgd pathways*. In IEEE International Symposium on Computer-Based Medical Systems (pp. 590–595).

Othman, R., Deris, S., & Illias, R. (2008). A genetic similarity algorithm for searching the gene ontology terms and annotating anonymous protein sequences. *Journal of Biomedical Informatics*, *41*(1), 65–81. doi:10.1016/j.jbi.2007.05.010

Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E., Falcão, A. O., & Couto, F. M. (2008). Metrics for GO based protein semantic similarity: A systematic evaluation. *BMC Bioinformatics*, *9*(5). doi:10.1186/1471-2105-9-S5-S4

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, *5*(7), e1000443. doi:10.1371/journal.pcbi.1000443

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Resnik, P. (1995). *Using information content to evaluate semantic similarity in a taxonomy*. In International Joint Conference on Artificial Intelligence (vol. 1, pp. 448–453).

Schlicker, A., Domingues, F. S., Rahnenfuhrer, J., & Lengauer, T. (2006). A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, *7*, 302. doi:10.1186/1471-2105-7-302

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., & Martinez-Cruz, L. A. (2005). Correlation between gene expression and GO semantic similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *2*(4), 330–338. doi:10.1109/TCBB.2005.50

Sheehan, B., Quigley, A., Gaudin, B., & Dobson, S. (2008). A relation based measure of semantic similarity for gene ontology annotations. *BMC Bioinformatics*, *9*(1), 468. doi:10.1186/1471-2105-9-468

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C. F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics (Oxford, England)*, *23*(10), 1274–1281. doi:10.1093/bioinformatics/btm087

Wu, X., Zhu, L., Guo, J., Zhang, D. Y., & Lin, K. (2006). Prediction of yeast protein-protein interaction network: Insights from the gene ontology and annotations. *Nucleic Acids Research*, *34*(7), 2137–2150. doi:10.1093/nar/gkl219

Xu, T., Du, L., & Zhou, Y. (2008). Evaluation of go-based functional similarity measures using s. cerevisiae protein interaction and expression profile data. *BMC Bioinformatics*, *9*(1), 472. doi:10.1186/1471-2105-9-472

Yon Rhee, S., Wood, V., Dolinski, K., & Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nature Reviews. Genetics*, *9*(7), 509–515. doi:10.1038/nrg2363

KEY TERMS AND DEFINITIONS

Ontology.: The formal representation of knowledge for a given domain by a hierarchical organization of concepts and relationships between them.

Gene Ontology.: A project that provides a controlled vocabulary of terms describing gene product characteristics and relationships across all species.

Semantic Similarity.: A measure of how related two or more concepts are.

Gene Product.: Biochemical material, either RNA or protein, resulting from gene expression.