1 Comparing Gene Ontology Semantic Similarity in a Systems Discovery Workbench

1.1 Introduction

Gene Ontology (GO) is a controlled, dynamic vocabulary used to contextualize eukaryotic genes in a broad and intuitive manner (Ashburner et al., 2000). GO contains three independent ontologies: Cellular Component, Biological Process, and Molecular Function. Each particular gene contains a set of GO annotations, which can range from general terms that are widely applicable to more specific terms that succinctly describe the location, pathway, and role of the gene of interest. Due to its hierarchical nature, researchers have used GO terms to explore relationships between genes of interest (Lord et al., 2003).

Semantic similarity as described by Resnik (1999), was the first measure used to examine relationship of GO terms sets between genes. The principle behind the Resnik (1999) measure is that terms that are rarely used, or deeper in the GO tree hierarchy, have a lower probability of being selected. Therefore, these terms will have higher Information Content (IC), since these terms likely describe more specific concepts. IC is calculated by taking the negative logarithm of the probability of the usage of the term where rarely used terms will have a higher IC score. The IC measure is used to calculate semantic similarity between two genes by considering the IC score of the lowest common ancestor of the two GO sets of interest.

The Resnik measure remains widely implemented when comparing GO semantic similarity. However, the original method does not take into account the amount of detail a GO term may actually possess since some parts of the GO tree are not deeply annotated, nor does it normalize the distance of two GO terms to their lowest common ancestor. In addition, the Resnik measure does not have an upper limit on scores, making it difficult to interpret the meaning of scores with a high magnitude. One measure that takes both of these factors into account and limits scores between 0 and 1 is the Relevance similarity score proposed by Schlicker and colleagues (2006). This measure is a combination of the original Resnik measure as well as the measure proposed by Lin (1998) and has shown to estimate semantic similarity between two genes at the same sensitivity as other standard measures (Schlicker et al., 2006).

Based on these previous findings, it is conceivable that two genes can be inferred to be a part of the same biological system based on GO semantic similarity. Therefore, we have decided to implement GO semantic similarity as a track in our systems discovery workbench. Our GO semantic similarity score is based on the average of Cellular Component and Biological Process scores, as Molecular Function scores are likely to be uninformative when searching for novel interactions. High Molecular Function scores will indicate functional similarity but will hinder our search, as biological systems likely require genes with vastly different functions rather than clusters of genes with similar functions. The measures and cutoff from this particular track will provide additional information to systems discovery and potentially aid in tentatively naming clusters created through the integration of all information tracks.

1.2 Materials and Methods

GO annotations as compiled by the GO Consortium were imported into R using GO.db (version 3.0). Entrez IDs were obtained for the gene pairs and semantic similarity was determined using a modified Information Content score proposed by Schlicker et al. (2006). Information Content (IC) as defined by Resnik is calculated as:

$$p(t) = \frac{n_{t'}}{N} | t' \epsilon \{t, childrenoft\}$$

$$IC(t) = -log(p(t))$$

where $n_{t'}$ represents the number of term t' and N represents the total number of GO terms in the GO tree. IC is then calculated as the negative log of the probability of term t. Schlicker et al. normalized Information Content scores by the depth and number of connections with the following equation:

$$sim_{Rel}(t_1, t_2) = \frac{2IC(MICA)(1 - p(MICA))}{IC(t_1) + IC(t_2)}$$

where MICA is the Most Informative Information Ancestor (ie. deepest common ancestor) between the two terms.

Semantic Similarity scores of gene pairs are calculated using the GOSemSim R package (version 3.0), available through Bioconductor, by Yu et al. (2010). The package requires Entrez Gene IDs in order to determine the GO semantic similarity between two genes of interest. The semantic similarity in Cellular Component (CC) and Biological Process (BP) GO categories will be determined. CC and BP scores will be determined using the best-match average strategy (average of maximum similarity scores on each row and column). The average of the two scores will determine the overall Semantic Similarity scores. Gene pairs that have scores greater than 0.2 (as recommended by Schlicker et al. (2006)) will be considered to be semantically similar. This track of the analysis pipeline will output gene pairs that are considered to be semantically similar alongside their average semantic similarity score.

1.3 References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G.

Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.

Lin D. An information-theoretic definition of similarity. ICML. 1998.

Lord PW, Stevens RD, Brass A, Goble CA. Semantic similarity measures as tools for exploring the gene ontology. Pac Symp Biocomput. 2003:601-12.

Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J Artif Intell Res. 1999:95-130.

Schlicker A, Domingues FS, Rahnenfuhrer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics. 2006 Jun 15;7:302.

Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics. 2010 Apr 1;26(7):976-8.