Proceedings of the second conference on Standards and Ontologies for Functional Genomics (SOFG 2) 2004:53.

Standards and Ontologies for Functional Genomics

FROM FUNCTIONAL SIMILARITY AMONG GENE PRODUCTS TO DEPENDENCE RELATIONS AMONG GENE ONTOLOGY TERMS

Bodenreider O. [1], Aubry M. [2], Burgun A. [3]
[1] U.S. National Library of Medicine, National Institutes of Health, Rockville, Maryland, U.S.A
[2] Unité de Génétique Humaine, UMR 6061 CNRS, Avenue du Pr Léon Bernard 35043 Rennes Cedex, France
[3] Laboratoire d'Informatique Médicale, Université de Rennes I, Av. Pr Léon Bernard 35043 Rennes Cedex, France

The Gene Ontology (GO) is a controlled vocabulary widely used for the annotation of gene products. GO is organized in three hierarchies for molecular functions, cellular components, and biological processes but no relations are provided among terms across hierarchies. More generally, dependence relations both within and across the three hierarchies are not recorded in GO. Methods based on lexical similarity have been used to identify such relations [1]. However, lexically similar words can have different meanings and the same meaning can be expressed by lexically dissimilar words. These limitations which are inherent to lexical methods led us to explore different approaches. In the annotation databases, each gene product is described a vector of GO terms. A vector space model (VSM) can thus be used to compute similarity among gene products, based on their annotations. Analogously, similarity can be computed among GO terms, based on the gene products with which these terms are associated. Figure 1 summarizes our approach. The vectors of gene products for each GO term are obtained by transposing the original matrix of gene products by GO terms. As it is usual with vector space models (e.g., in information retrieval applications), the similarity between two GO terms is computed as the dot product of the corresponding vectors of genes, after normalization of these vectors [2]. The dot product of two vectors varies between 0 (no similarity) and 1 (perfect similarity). We applied this method to five annotation databases (FlyBase, the Human subset of GOA, MGI, SGD, and WormBase). Term-term similarity was computed pairwise for all GO terms present, resulting in a half-matrix for each model organism database. Relations with a similarity lower than .5 were ignored. As shown in Table 1, a total of 4,316 relations among GO terms were identified by this method, restricted to relations across hierarchies, in at least one annotation database. Examples of pairs of related terms include *potassium channel activity* (MF) / *potassium ion transport* (BP) and *hemoglobin complex* (CC) / *oxygen transport* (BP). Both lexical similarity and VSM similarity identified large numbers of dependence relations. However, only a few percent of these relations are common to both methods. Further validation (manual or against other methods) is needed to assess the validity of the relations identified. For further information on this research the reader is referred to [3].

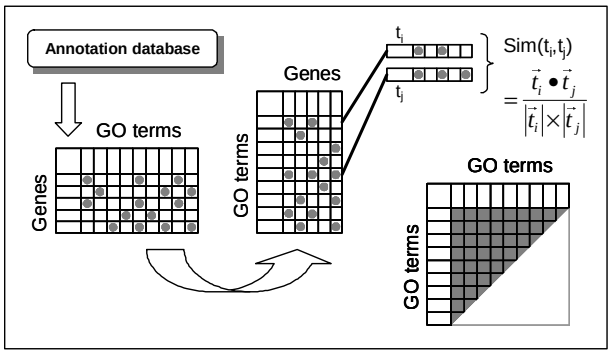

Figure 1. Similarity between GO terms in the vector space model from a given annotation database.

Table 1 – Number of associations identified by the vector space model (VSM), by lexical similarity (LEX), and by both method (Both) for each category of association (MF: molecular function; CC: cellular component; BP: biological process)

	VSM	LEX	Both
MF-CC	499	917	81
MF-BP	3057	2523	33
CC-BP	760	2053	23
Total	4316	5493	137

[1] P Ogren, P.V., Cohen, K.B., Acquaah-Mensah, G.K., Eberlein, J. & Hunter, L. The compositional structure of Gene Ontology terms. *Pac Symp Biocomput*, 214-25 (2004)
[2] Baeza-Yates, R. & Ribeiro-Neto, B. *Modern information retrieval*, 513 p. (ACM Press ; Addison-Wesley, New York; Harlow, England, 1999).
[3] Bodenreider O, Aubry M, Burgun A. Non-lexical approaches to identifying associative relations in the Gene Ontology. In: Pacific Symposium on Biocomputing 2005; (in press).