Proceedings of the second conference on Standards and Ontologies for Functional Genomics (SOFG 2) 2004:47.

GENE ONTOLOGY-DRIVEN SIMILARITY FOR GENE EXPRESSION CORRELATION ANALYSIS

Azuaje F [1], Wang H [1], Bodenreider O [2], Dopazo J [3]
[1]School of Computing and Mathematics, University of Ulster, Jordanstown, Co. Antrim, UK
[2]U.S. National Library of Medicine, National Institutes of Health, Rockville, Maryland, U.S.A
[3] Bioinformatics Unit, Spanish National Cancer Centre, Madrid, Spain

The Gene Ontology (GO) and its related annotation databases provide information for measuring similarity between gene products. The topological features of GO terms (i.e., their inter-relationships in the ontology) and the statistical features of the terms in annotation databases (i.e., frequency) are both exploited by information-theoretic approaches to measuring functional similarity among gene products. Previous research has shown that GO-driven, functional similarity of pairs of genes correlates with sequence similarity [1]. This study aims to support the integration of GO-driven similarity for functional prediction problems. It focuses on the quantitative assessment of relationships between GO-driven similarity and expression correlation. It also offers insights into the consistency of the functional information represented in the GO and resulting databases. The GO and annotations derived from the *S. cerivisiae* Genome Database (SGD) were analyzed to calculate functional similarity of gene products. Three methods for measuring similarity were implemented: Resnik's, Lin's and Jiang's metrics. Using a known gene expression dataset in yeast [2], several million pairs of gene products were compared on the basis of these properties. This analysis was performed separately on the three hierarchies of the GO. It confirms that highly correlated genes exhibit strong similarity based on information originating from the GO hierarchies. Such a similarity is significantly stronger than that observed between weakly correlated genes. This observation holds for the three GO hierarchies and for the three metrics under investigation.

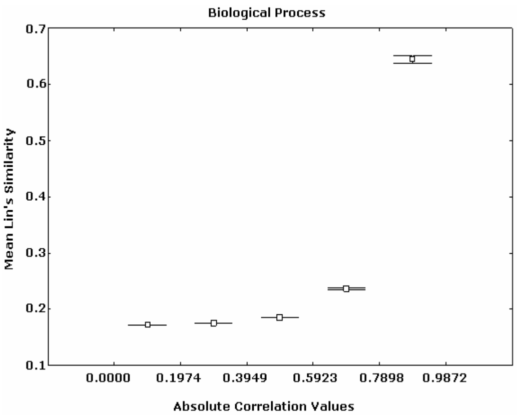

Figure 1. Expression correlation and GO-driven similarity based on Lin's technique for the BP hierarchy.

Figure 1 illustrates the relation between one of the similarity methods and the absolute expression correlation values between pairs of genes. In this case similarity was calculated for the Biological Process (BP) hierarchy using Lin's model. The axis of abscissas is divided into a number of absolute correlation intervals, and the axis of ordinates shows the mean similarity values detected in these intervals and their 95% confidence intervals. This study further demonstrates the relevance of applying GO-driven similarity assessment techniques for validating gene expression correlation. Similarity values may provide indicators to detect irrelevant expression correlations. It may also support the detection of false-positives interactions by indicating when two potentially-interacting proteins are not functionally associated.

We are investigating the application of these methods for defining *cluster validity indices*, which may aid in the identification of significant gene clusters. We are also designing clustering strategies that combine expression correlation and similarity information. For further information on this research the reader is referred to [3].

[1] P. Lord, R. Stevens, A. Brass, and C. Goble, "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation," *Bioinformatics*, vol. 19, pp. 1275-1283, 2003.

[2] M. Eisen, P. L. Spellman, P. O. Brown, and D. Botsein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 14863-14868, 1998.

[3] H Wang, F Azuaje, O Bodenreider, J Dopazo, "Gene expression correlation and gene ontology-driven similarity: An assessment of quantitative relationships", *Proc. of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2004.