Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies 2005:9-10.

Ontology-driven similarity approaches to supporting gene functional assessment

Francisco Azuaje[1,], Haiying Wang[1] and Olivier Bodenreider[2]*

[1]*University of Ulster, Northern Ireland,, UK*
[2]*National Library of Medicine, Bethesda, USA*

ABSTRACT

Motivation: Bio-ontologies, such as the Gene Ontology, represent important sources of prior knowledge that may be automatically integrated to support predictive data analysis tasks. The assessment of similarity of gene products provides the basis for the implementation of classification tools and the automated validation of functional associations. This study discusses alternative techniques for measuring ontology-driven similarity of gene products. Relationships between these types of similarity information and key functional properties, such as gene co-expression, are discussed.

1 INTRODUCTION

Bio-ontologies represent important knowledge bases, which have traditionally been applied to enhance database annotation and interoperation as well as cross-database information retrieval tasks. The *Gene Ontology™* (GO) (The Gene Ontology Consortium, 2001) is one such resource that is becoming the *de facto* standard for annotating gene products.

The relevance of the GO goes beyond annotation and information retrieval applications. It has been shown that GO may facilitate large-scale predictive applications in functional genomics. The analysis of GO annotations in gene expression analysis may help to explain why a particular group of genes share similar expression patterns. Several tools have been proposed to identify functionally-enriched clusters of genes. *FatiGO* (Al-Shahrour *et al.*, 2004), for example, extracts GO terms that are significantly over- or under-represented in clusters of genes. GO-based annotations have been incorporated to construct functional predictors that in combination with other information resources have shown to improve functional association prediction (e.g. protein-protein interactions) (Jansen *et al.*, 2003). Hvidsten *et al.* (2003) combined gene expression data with annotations originating from the GO biological process taxonomy. They applied *rough set theory* to assign biological process terms to genes represented by expression patterns. King *et al.* (2003) implemented *decision trees* and *Bayesian networks* to predict new GO terms-gene associations based on existing annotations from the SGD and FlyBase. Although these functional prediction tools process GO annotations they do not fully exploit the knowledge that can be extracted from analyzing relations of GO terms and their information content in different annotation databases. For instance, traditional functional prediction support or cluster analysis tools mainly process information about the frequency of individual annotation terms associated with a list of genes. Furthermore, such applications may be improved by explicitly considering similarity relationships between the genes, which may be estimated by analyzing both the information content and structure of the GO. It has been suggested that by ignoring such *semantic similarity* between closely related GO terms (e.g., between a parent and a child), traditional methods may fail to identify the functional similarity between genes annotated with these closely related yet distinct terms.

Thus, the GO has been proposed as a tool for measuring similarity between genes. Previous research showed significant relationships between semantic similarity of pairs of genes and their sequence-based similarity (Lord *et al.*, 2003). Also we have evaluated relevant quantitative relationships between GO-driven similarity and gene expression correlation (Wang *et al.*, 2004). GO-driven clustering algorithms based on such approaches have been recently reported (Wang *et al.*, 2005, Speers *et al.*, 2004). Moreover, they have provided the basis for developing tools that may facilitate the identification of relevant partitions from clustering, using, for example, GO-driven cluster validity indices (Bolshakova *et al.*, 2005)

This paper discusses our current research on the design of GO-driven similarity assessment techniques. It aims to compare two approaches to estimating between-gene similarity, which may be implemented using different schemes for measuring between-term similarity. Relationships between semantic similarity and gene co-expression are further investigated taking into account both approaches.

2 SEMANTIC SIMILARITY APPROACHES TO ASSESSING GENE SIMILARITY

Given a pair of terms, c_1 and c_2, a traditional method for measuring their similarity consists of calculating the distance between the nodes associated with these terms in the

* To whom correspondence should be addressed.

ontology, whose limitations have been discussed elsewhere (Zhong *et al.*, 2002). *Information-theoretic* models have been studied as alternative approaches to measuring similarity in an ontology. Let *C* be the set of terms in the GO. Information-theoretic approaches to measuring similarity between terms, $c \in C$, may be based on the *amount of information* associated with them or shared by them in common. Several techniques may be implemented using this principle, such as those proposed by Lin, Resnik and Jiang (Lord *et al.*, 2003, Wang *et al.*, 2004). Similarity (or distance) values for a pair of gene products described by GO terms may be calculated based on such techniques (Lord *et al.*, 2003, Wang *et al.*, 2004). Given a pair of gene products, g_i and g_j, which are annotated by a set of terms A_i and A_j respectively, where A_i and A_j comprise *m* and *n* terms respectively, the semantic similarity, $SIM(g_i, g_j)$, may be defined as the average inter-set similarity between terms from A_i and A_j:

$$SIM(g_i, g_j) = \frac{1}{m \times n} \times \sum_{c_k \in A_i, c_p \in A_j} sim(c_k, c_p) \qquad (1)$$

where $sim(c_k,c_p)$ represent the similarity between terms. This approach does not always meaningfully estimate similarity. For example, similarity is expected to be equal to 1 when the gene pair has the same set of annotation terms. However, this is not true when several annotations within a hierarchy are assigned to the genes. In order to address such a limitation we are currently evaluating an alternative approach that selectively aggregates maximum inter-set similarity values as follows:

$$SIM(g_i, g_j) = \frac{1}{m+n} \times \left(\sum_k \max_p (sim(c_k, c_p)) + \sum_p \max_k (sim(c_k, c_p))\right) \qquad (2)$$

2.1 Linking semantic similarity and other functional properties

The analysis of quantitative relationships between semantic similarity and other functional information resources is important to allow the identification of novel integrative prediction strategies. Such relationships may indicate whether semantic similarity may be combined with other large-scale predictive resources (e.g. gene expression correlation, sequence binding patterns, etc.) to improve key functional prediction factors, such as accuracy and coverage. Based on (1) previous research has confirmed that GO-driven similarity and expression correlation of pairs of gene products in *S. cerevisiae* are significantly interrelated (Wang *et al.*, 2004). This property has shown to be consistently valid for similarity information originating from all of the GO hierarchies. We are currently analyzing these relationships using (1) and (2) on the latest GO annotation release for *S. cerevisiae*.

We are assessing relationships between semantic similarity and other functional properties such as gene co-regulation and protein-protein interactions in *S. cerevisiae* and *C. elegans.* One of our hypotheses is that the GO-driven similarity of a pair of genes may be used as an indicator of regulatory and protein-protein interactions.

Furthermore, we are investigating how GO-driven semantic similarity may be applied to support the detection of spurious (co-regulation or protein-protein) interaction predictions. After studying this, one could in principle justify the design of prediction support tools for co-regulation and protein-protein interactions, which in combination with other resources, e.g. co-expression, may support a more accurate and biologically meaningful identification of functional networks.

REFERENCES

The Gene Ontology Consortium (2001) Creating the gene ontology resource: Design and implementation. *Genome Research*, **11**, 1425-1433.

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

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302** (17), 449-453.

Hvidsten, T., Lægreid, A. and Komorowski, J. (2003) Learning rule-based models of biological process from gene expression time profiles using Gene Ontology. *Bioinformatics*, **19**, 1116-1123.

King, O. D., Foulger, R. E., Dwight, S. S., White, J. V., and Roth, F. P. (2003) Predicting gene function from patterns of annotation. *Genome Research*, **13**, 896--904.

Lord, P., Stevens, R., Brass, A. and Goble, C. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275--1283.

Wang,H., Azuaje,F., Bodenreider,O., and Dopazo,J. (2004) Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In *Proc. of IEEE 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, La Jolla, CA, USA, 25-31.

Wang,H., Azuaje,F., and Bodenreider,O. (2005) An ontology-driven clustering method for supporting gene expression analysis. In *Proc. of the 18th IEEE International Symposium on Computer-Based Medical Systems*, in press.

Speer, N., Spieth, C. and Zell, A. (2004) A memetic clustering algorithm for the functional partition of genes based on the gene ontology. In the *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, San Diego, USA, 252-259.

Bolshakova,N., Azuaje, F., and Cunningham,P. (2005) A knowledge-driven approach to cluster validity assessment. *Bioinformatics*, Advance Access published on February 15, 2005.

Zhong, J., Zhu, H., Li, Y. and Yu, Y. (2002) Conceptual graph matching for semantic search. In *Proc. of Conceptual Structures: Integration and Interfaces*, 92-106.