Proceedings of the IEEE Fourth Symposium on Bioinformatics and Bioengineering (BIBE-2004) 2004:317-324.

Incorporating Ontology-Driven Similarity Knowledge into Functional Genomics: An Exploratory Study

Francisco Azuaje
University of Ulster at Jordanstown, UK
fj.azuaje@ieee.org

Olivier Bodenreider
National Library of Medicine, Bethesda, USA
olivier@nlm.nih.gov

Abstract

This research explores the feasibility of semantic similarity approaches to supporting predictive tasks in functional genomics. It aims to establish potential relationships between ontology-based similarity of gene products and important functional properties, such as gene expression correlation. Similarity measures based on the information content of the Gene Ontology (GO) were analyzed. Models have been implemented using data obtained from well-known studies in S. cerevisiae. Results suggest that there may exist significant relationships between gene expression correlation and semantic similarity. Analyses of protein complex data show that, in general, there is a significant correlation between the semantic similarity exhibited by a pair of genes and the probability of finding them in the same complex. These results can also be interpreted as an assessment of the quality and consistency of the information represented in the GO.

1. Introduction

One of the main goals of the post-genome era is to integrate relevant data sources, which may be useful for implementing comprehensive, large-scale protein characterization studies. Outcomes originating from biological research can now be accessed through data repositories, which provide specialized information to describe the structure and function of genes and their products. This type of information becomes an important source of background knowledge, which may be exploited to facilitate cross-database queries and to support validation studies. The Gene Ontology™ (GO) [1] is one such resource, which has been designed to offer controlled vocabularies and shared hierarchies for aiding in the annotation of molecular attributes across model organisms. Moreover, it has been shown that the GO can be used for describing gene expression clustering results and for implementing advanced gene querying systems [1],[2].

Incorporating background knowledge, such as that represented in the GO, is crucial for generating or testing novel hypotheses in functional genomics. The information derived from this process may be used to develop new predictive systems, which may be integrated with other models in large-scale genomic research. Moreover, it can be used to assess the consistency and validity of emerging knowledge.

One strategy to exploit the information encoded in the GO may consist of processing it to measure similarity between gene products. Similarity assessment is at centre of important tasks in bioinformatics. This is fundamental to implement predictive models for large-scale functional genomics. This type of information is known as *semantic similarity*, because it takes into account information relevant to the definition of concepts and their inter-relationships within a specific problem domain.

There are two major similarity assessment schemes for studying proteins [3]. *Structural classification* measures similarity based on protein sequence or tertiary structure. Lord *et al.* [4] have investigated relationships between semantic and sequence similarity. They applied different semantic similarity measures and the BLAST tool on the Swiss-Prot database. *Functional classification* assesses similarity in terms of functional features such as biochemical pathways. It does not comprise structural similarity features. This paper places emphasis on the incorporation of ontology-based similarity for functional classification problems. It aims to study potential relationships between the semantic similarity of gene products and key functional properties: Gene expression correlation and protein complex membership. The results are based on the GO annotations from the *Saccharomyces Genome Database* (SGD). Section 2 overviews the GO and

applications. Section 3 introduces the problem of measuring semantic similarity in the GO. Section 4 describes datasets and problems under consideration. Section 5 presents results. Section 6 discusses methods and future research.

2. The Gene Ontology and its applications

2.1 The Gene Ontology

The primary goal of the GO is to define a shared, structured and controlled vocabulary to annotate molecular attributes across models organisms [1]. Moreover, the GO project allows users to access annotation information resulting from different model organisms. For instance, the databases SGD and FlyBase use terms defined by this ontology. The GO annotation files also provide useful information about the evidence for the knowledge represented in its taxonomies. This information is stored in the form of evidence codes, which is associated with each gene annotated using the GO. There are different types of evidence codes supported by the GO. For example, the evidence codes *TAS* (Traceable Author Statement) and *IEA* (Inferred from Electronic Annotation). The evidence code TAS refers to annotations supported by articles or books. In contrast, IEA annotations are based on results automatically derived from sequence similarity searches, which have not been reviewed by curators. The reader is referred to the GO website to obtain additional information on databases and evidence codes supported (www.geneontology.org).

The GO consists of three ontologies, sometimes referred to as taxonomies: Molecular function (MF), biological process (BP), and cellular component (CC). The first ontology refers to information on what a gene product does. BP is related to a biological objective to which a gene product contributes. CC refers the cellular location of the gene product, including cellular structures and complexes. Figure 1 depicts the general organization of the GO and a partial view of the first level of terms included under MF. The reader is referred to [1] for further information on the GO design principles. These vocabularies (one for each ontology) and their relationships are represented in the form of *directed acyclic graphs* (DAGs). Thus, a hierarchy in the GO may be seen as a network in which each term may represent a "child node" of one or more "parent nodes". There are two types of child-to-parent relationships: "is a" and "part of" types. The first type is defined when a child class is a subclass of a parent class. For example, from the BP ontology, cell proliferation is a child of cell growth. The second type is used to describe when a child node is a component of a parent. For example, from the same ontology, DNA replication is part of the cell cycle. Figure 1.a illustrates a partial view of the type of DAGs found in the GO.

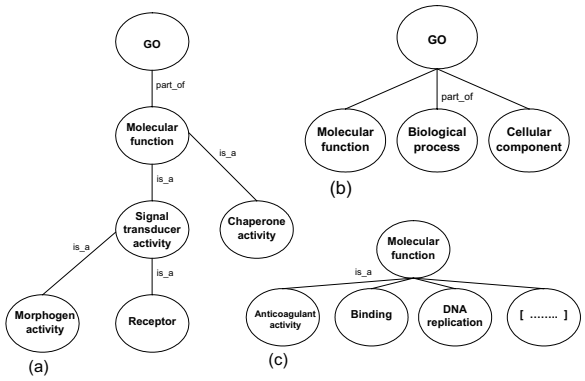

Figure 1. Different views of the GO. (a) Typical example of a DAG. (b) GO taxonomies. (c) Partial view of the first level of MF. Dashed lines indicate the presence of several terms not included here.

2.2 Gene Ontology-based applications

The GO may facilitate information search tasks across databases, because it offers a framework to store different repositories using the same query terms. However, the relevance of the GO goes beyond information retrieval applications. It has been demonstrated that the GO may facilitate large-scale applications for functional genomics. One such application is the *FatiGO* system, which is a Web-based interface for analyzing groups of genes and their associations with GO terms [2]. It allows the user to analyze the differential distributions of GO terms for two sets of genes. Based on statistical tests, it assigns the most representative GO terms associated with a group of genes. King *et al.* [5] predicted known and novel gene-phenotype associations in yeast. Their model uses phenotypic annotations extracted from the Munich Information Center for Protein Sequences (MIPS) database and gene annotations based on more than 3000 GO terms. Decision trees were implemented to infer associations. Hvidsten *et al.* [6] combined gene expression data with annotations originating from the GO biological process taxonomy. They propose a supervised classification system, based on *rough set theory*, to predict biological processes linked to expression patterns. Although these methods process GO terms, they do not fully exploit the information associated with the structure of the GO

and its information content. The following section introduces the problem of measuring semantic similarity in the GO.

3. Semantic similarity methods and the GO

The semantic similarity between terms represented in an ontology may be defined, for example, in terms of topological and statistical patterns. In order to understand the problem of measuring semantic similarity between gene products based on their annotations, it is first necessary to describe approaches to calculating the similarity between annotation terms.

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 ontology. The shorter this distance, the higher the similarity. If there are multiple paths, one may use the shortest or the average distance. This approach is commonly referred to as the *edge counting* method. A variation of this method defines weights for the links according to their position in the taxonomy [7]. Constraints exhibited by this type of models have been previously studied [7]. One of its limitations is that it heavily relies on the idea that nodes and links in an ontology are uniformly distributed. This is not an accurate assumption in taxonomies exhibiting variable link densities. An alternative approach to measuring semantic similarity exploits *information-theoretic* principles [8]. It has been demonstrated that this type of approaches is less sensitive, and in some cases not sensitive, to the problem of link density variability [9]. These methods traditionally consider only the "is a" links in a taxonomy. However, it has been shown that other types of links may also be processed to perform similarity assessment [9]. Moreover, it has been suggested that processing all types of links as equals may be an effective strategy for dealing with the problem of orphan nodes [4],[9]. It is also important to stress that the majority of the GO links are "is a" links [4]. Thus, this bias regarding link type usage supports the application of this approach. This research considers the two types of links as equally relevant to the similarity assessment process.

Let C be the set of terms in the GO. One key approach to assessing the similarity between terms, $c \in C$, is to analyze the amount of information they share in common. In the GO this information may be represented by the set of parent nodes, which subsume the terms under consideration. For example, in Figure 1.a the terms "morphogen activity" and "receptor" are subsumed by the terms "signal transducer activity" and "molecular function". Thus, one may say that the terms "morphogen activity" and "receptor" shared those attributes (parents) in common. For each term, $c \in C$, *p(c)* is the probability of finding a child of *c* in the taxonomy. Thus, as one moves up to the root node of the GO (i.e. terms "molecular function", "biological process" and "cellular component") *p(c)* monotonically approaches a value equal to 1. This together with the principle of information theory allows the quantification of the information content of a term as equal to –log(*p(c)*). It allows to measure similarity between terms based on the assumption that the more information two terms share in common, the more similar they are. In this situation the information shared by two terms may be calculated using the information content of the terms subsuming them in the ontology. Such a semantic similarity model was proposed by Resnik [9]:

$$sim(c_i, c_j) = \max_{c \in S(c_i, c_j)} [-\log(p(c))] \qquad (1)$$

where $S(c_i, c_j)$ comprises the set of parent terms shared by both terms c_i and c_j, and 'max' represents the maximum operator. The value of this metric can vary between 0 and infinity. For example, in Figure 1.a "signal transducer activity" and "molecular function" belong to $S(c_1, c_2)$, where c_1 and c_2 are "morphogen activity" and "receptor" respectively. Nevertheless, "signal transducer activity", which provides the minimum *p(c)* and the maximum $-\log(p(c))$, also represents the most informative term. Thus, equation (1) provides the information content of the lowest common ancestor of two terms.

Lin [10] proposed an alternative information-theoretic approach. It takes into account not only the parent commonality of two query terms, but also the information content associated with the query terms. Thus, given terms, c_i and c_j, their similarity may be defined as:

$$sim(c_i, c_j) = \frac{2 \times \max_{c \in S(c_i, c_j)} [\log(p(c))]}{\log(p(c_i)) + \log(p(c_j))} \qquad (2)$$

where $p(c_i)$ and $p(c_j)$ are defined as above. The values generated by equation (2) vary between 0 and 1. This technique may be seen as a normalized version of (1). Lin's values also increase in relation to the degree of similarity shown by two terms, and decreases with their difference. For additional information on these and related techniques the reader is referred to [8], [10]. Based on equations (1) or (2) it is then possible to calculate the similarity between gene products based

on their annotations. 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 . Thus, this method aggregates similarity contributions originating from all of the terms used to describe g_i and g_j. This is formally defined as:

$$SIM(g_i, g_j) = \frac{1}{m \times n} \times \sum_{c_k \in A_i, c_p \in A_j} sim(c_k, c_p) \quad (3)$$

P.W Lord and colleagues [4] have investigated the relationship between semantic and sequence similarities. They suggested that semantic similarity metrics, such as those based on equations (1) to (3), are correlated with gene sequence similarity. Their results are based on the analysis of the Swiss-Prot-Human database, and they indicate that semantic similarity may support more powerful gene sequence search tasks. The contribution of our study is to explore the potential relationships between the semantic similarity of gene products, their gene expression correlation and their membership in the same complex.

4. Datasets and methods

Results are based on the analysis of the GO database available on April 7th, 2003. This research comprises associations between GO terms and gene products included in the SGD. The analyses considered only non-IEA annotations due to their reliability (Section 2). Quantitative relationships between the semantic similarity of pairs of gene products and functional genomics data were studied. Two types of functional data were analyzed: Gene expression correlation and complex membership in *S. cerevisiae*. The integration of gene expression correlation and semantic similarity is based on data that characterize mRNA transcript levels during the cell cycle of *S. cerevisiae.* These data were obtained from [11]. This dataset was selected because of its scientific relevance, which has been demonstrated in previous research [12], and because it reflects fundamental cellular states of this organism. Similarity analyses were performed on 225 genes that show significant and periodic transcriptional fluctuations during the five cell cycle phases: early G1, late G1, S, G2 and M phases [11]. Each gene is described by 17 expression values. The total number of gene pairs generated by this dataset is 25200. Thus, 25200 pairs of similarity values and 25200 absolute expression correlation values were calculated. Expression correlation was calculated using the well-known *Pearson correlation* coefficient. Associations between semantic similarity and protein complex membership were studied using a dataset consisting of 83 pairs of proteins, which were characterized by Jansen *et al.* [12]. Each pair is labelled on the basis of their membership (or non-membership) in the same protein complex in *S. cerevisiae*. There are 46 pairs categorized as belonging to the same complex (true positives). Jansen *et al.* showed that these predictions exhibited low error rates in a study that integrated several whole-genome data resources, using the MIPS complexes catalogue as the gold standard. In our study similarity values were calculated for each pair. Graphical analyses and *ANOVA* were performed to visualize potential correlations between semantic similarity of a pair of genes and the probability of finding them in the same complex.

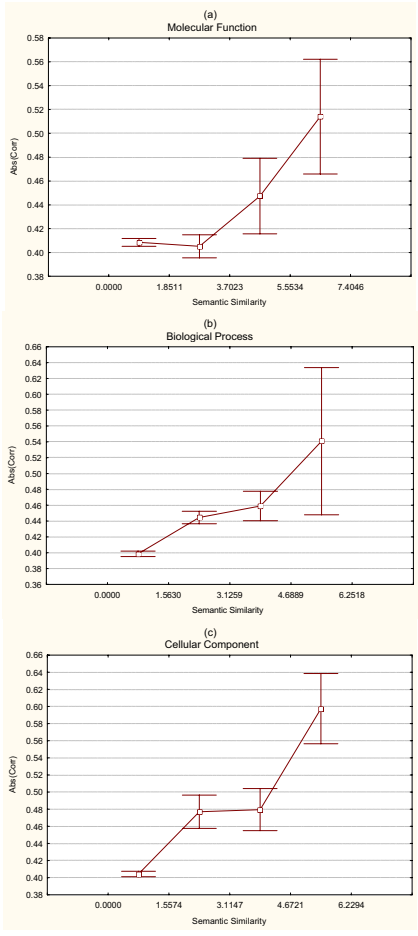

Figure 2. Expression correlation and GO-based similarity. (a) MF, (b) BP and (b) CC taxonomies. Mean absolute expression correlation values for each similarity interval and their 95% confidence intervals. Similarity based on equation (1).

5. Results

5.1 GO-based similarity and expression data

Figure 2 shows expression correlation values between pairs of gene products against semantic similarity. The axis of abscissas is divided into a number of similarity intervals, and the axis of ordinates shows the absolute mean expression correlation values for these intervals and their 95% confidence intervals. Each panel in Figure 2 summarizes information from the MF, BP and CC taxonomies.

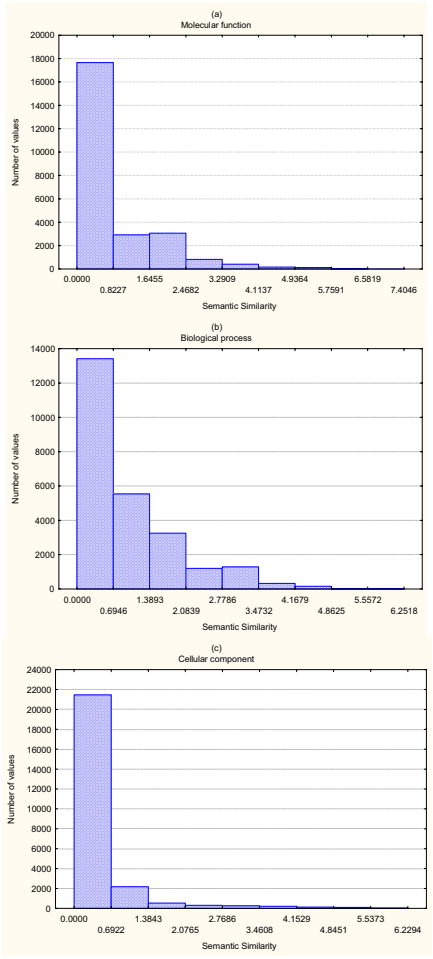

Figure 3. Distribution of semantic similarity values for GO taxonomies: (a) MF, (b) BP, (c) CC, based on equation (1).

These analyses are restricted to non-IEA annotations and based on equation (1). In general, high similarity values are associated with high expression correlation values, and weak semantic similarity is associated with low expression correlation. By augmenting or reducing the number of similarity intervals this trend is observed for extreme values: lowest/highest similarity/correlation values for all ontologies. This response is significantly stronger in the case of the lowest expression correlation values. One factor that should be considered to evaluate the relevance or reliability of these outcomes is the relatively small number of high similarity values generated by this dataset and the important variability of these values. Figure 3 depicts the distribution of semantic similarity values generated by the GO taxonomies. Figure 4 illustrates results based on equation (2) obtained for all the taxonomies. They are in general consistent with the results obtained with equation (1). These figures also suggest that the strongest and weakest semantic similarities may be linked to the highest and lowest expression correlation values respectively.

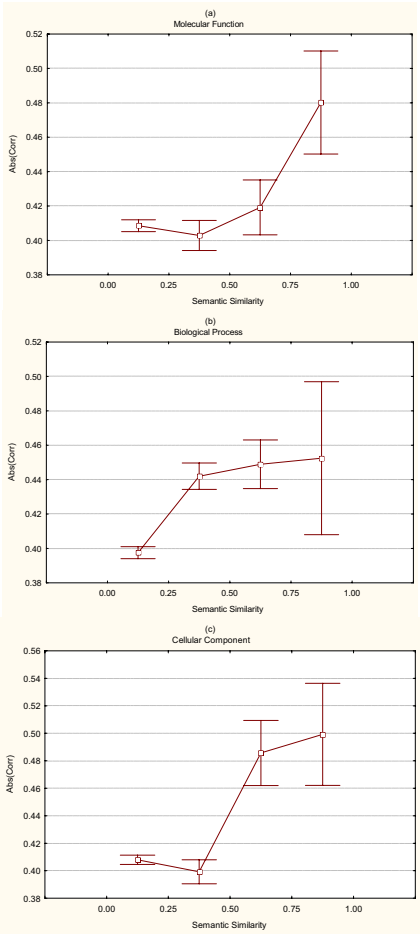

Figure 4. Expression correlation vs. GO-based similarity using (a) MF, (b) BP and (c) CC taxonomies. Similarity based on equation (2).

5.2 GO-based similarity and protein complex membership

Figure 5 summarizes the relationship between equation (1) and protein complex membership for each GO hierarchy. The dataset consisted of 83 pairs of proteins characterized by Jansen *et al.* [12], in which each pair is labeled on the basis of their membership (or non-membership) in the same complex. Thus, Figure 5 portrays the probability of finding in the same complex two proteins exhibiting a similarity value within an interval, *SS*. This probability is referred to as *P*(*same_complex*|*SS*).

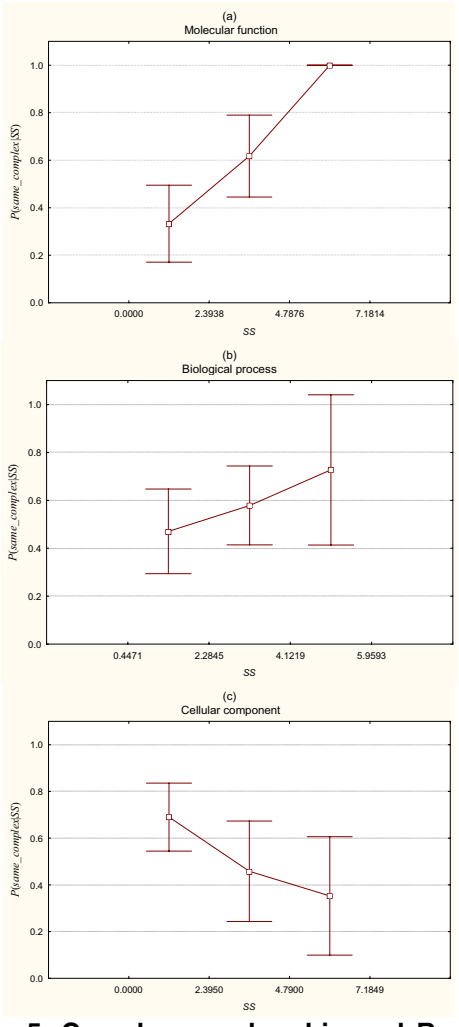

Figure 5. Complex membership and Resnik's similarity. Probability, *P*(*same_complex*|*SS*), of finding in the same complex two proteins exhibiting a similarity value within *SS*. Data obtained from [12]. Results based on (a) MF, (b) BP and (c) CC taxonomies.

Figures 5.a and 5.b suggest that if two proteins belong to the same complex it is also likely that such proteins are highly similar on the basis of their function and involvement in biological processes. This relationship is significantly stronger in the case of function. All of the protein pairs exhibiting the highest similarity values can also be found in the same complexes (Figure 5.a). However, the CC taxonomy produced results inconsistent with these observations. Figure 5.c indicates an inverse relationship between these two properties. These results are based on three SS intervals, but similar results were obtained for two and four intervals.

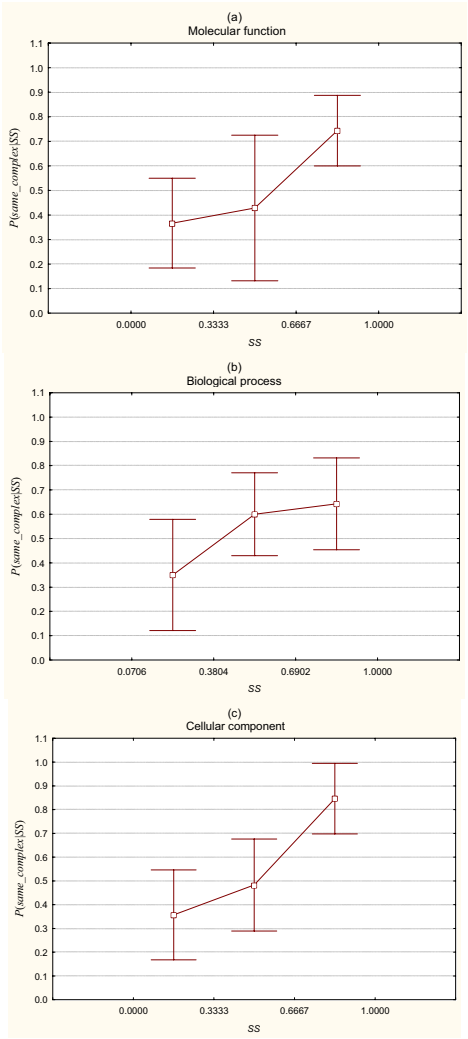

Figure 6. Complex membership and Lin's semantic similarity based on the GO. Probability, *P(same_complex|SS)*, of finding in

the same complex two proteins exhibiting a similarity value within *SS*. Results based on (a) MF, (b) BP and (c) CC taxonomies.

Figure 6 illustrates the results obtained from the similarity model proposed by Lin. These experiments also indicate a significant relationship: The stronger the semantic similarity between two proteins, the more likely the possibility of finding them in the same complex. Nevertheless, unlike the results shown in Figure 5, the information represented in all of the GO taxonomies supports this hypothesis. Further research involving larger data sets is required to confirm it.

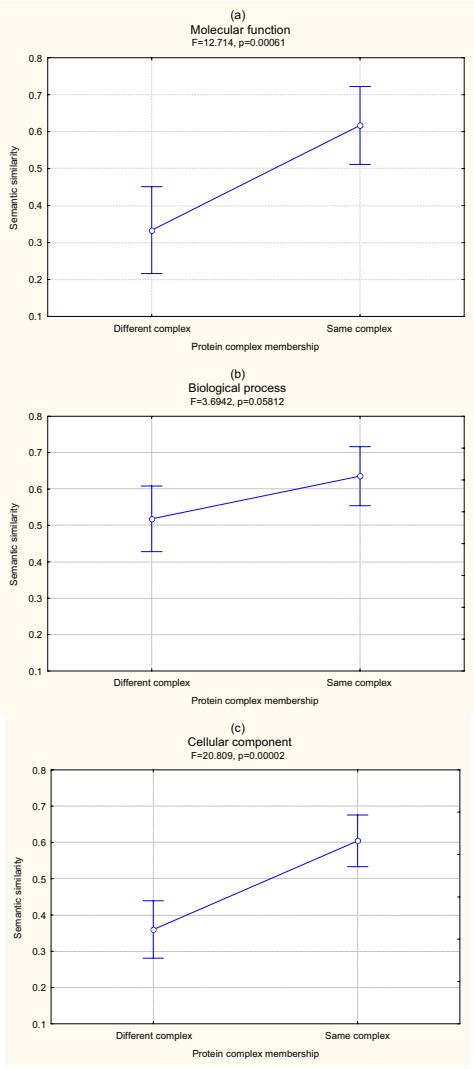

Figure 7. ANOVA: Semantic similarity values between pairs of proteins assigned to the same complex vs. those belonging to different complexes. Results based on (a) MF, b) BP and (c) CC, using equation (2).

On the basis of their semantic similarity, Figure 7 compares the differences between protein pairs belonging to the same complex and those assigned to different complexes. Similarity values are considered for each GO taxonomy using Lin's metric. It also presents F and p values produced by this ANOVA procedure. It shows that there may be significant differences between these two categories of protein pairs in terms of their semantic similarity. Protein pairs belonging to the same complex generally exhibit strong similarities. These differences are significantly stronger in the case of similarity originating from the CC taxonomy (Figure 7.c). There was no significant difference in the results produced by the BP ontology (Figure 7.b). Regarding MF and BP, similar results were obtained with Resnik's method.

6. Discussion and conclusions

Similarity models. Semantic similarity models have been implemented using information stored in the GO. Previous research in natural language processing has shown that information theoretic approaches may outperform traditional taxonomic-based similarity assessment methods, such as edge counting approaches [8],[10]. However, semantic similarity methods such as those proposed by Resnik and Lin may also be susceptible to the limitations observed in traditional methods. One problem that requires further investigation is the effect of taxonomy link variability on the similarity calculation process. This may represent a difficult problem to control using the GO taxonomies because of the presence of term sub-taxonomies that may be denser than others. Resnik has shown that his similarity model is not sensitive to this problem in the case of the WordNet's taxonomy of concepts [9].

Contributions and limitations. This research studied potential significant relationships between GO-based similarity and functional genomics data. In general one may expect a strong connection between the degree of GO-based similarity and the absolute expression correlation of two gene products. This is perhaps more clearly illustrated in the case of pairs of genes showing a low expression correlation and weak semantic similarity. One factor that needs to be considered to interpret these results is the relatively small number of gene pairs producing high semantic similarity values. Based on the analysis of a relatively small number of protein pairs, this study suggests that pairs assigned to the same complex may exhibit strong semantic similarity. Such similarity levels may be

significantly stronger than the similarity shown by protein pairs belonging to different complexes.

Applications. The relationship between GO-based similarity and gene expression correlation may be applied to support functional prediction applications together with other genomic resources. It may be used, for example, to evaluate or validate clusters of genes or similarity-based predictions using different types of data [13]. The quality of expression clusters may be assessed on the basis of the GO-based similarity exhibited by these clusters. Therefore, it would be useful to study methods for representing cluster coherence and isolation. Moreover, semantic similarity models may be used to automatically label gene clusters in terms of their coherence. Semantic similarity could also be applied to support annotation tasks. For instance, groups of gene products could be annotated using their lowest common ancestor rather than multiple annotations. These models may also contribute to assess differences in annotations across genes, within a database or across multiple organisms.

Future work. Current studies include replication of experiments using recent releases of the SGD, and implementation of other similarity measures. Future research will provide stronger evidence to support these claims. It will be necessary to incorporate different and larger sets of expression data in yeast and other organisms. We also aim to analyze larger protein complex datasets. This will allow us to confirm the feasibility of using these models for supporting the prediction of complex membership. Another crucial problem is the integration of similarity information originating from the GO hierarchies. One basic approach is to calculate the average of the similarity values obtained from each hierarchy. Preliminary results are in general consistent with the findings of this paper.

7. References

[1] The Gene Ontology Consortium. "Creating the gene ontology resource: Design and implementation", *Genome Research*, vol. 11, pp. 1425-33, 2001.

[2] J. Herrero, F. Al-Shahrour, R. Diaz-Uriarte, et al, "GEPAS: A web-based resource for microarray gene expression data analysis", *Nucleic Acids Research*, vol. 31, pp. 3461-3467, 2003.

[3] C. Ouzounis, R. Coulson, A. Enright, V. Kunin, J. Pereira-Leal, "Classification schemes for protein structure and function", *Nature Reviews Genetics*, vol. 4, pp. 508-19, 2003.

[4] P. Lord, R. Stevens, A. Brass, C. Goble, "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation", *Bioinformatics*, vol. 19, pp. 1275-83, 2003.

[5] O. King, J. Lee, A. Dudley, D. Janse, G. Church, F. Roth. "Predicting phenotype from patterns of annotation", *Bioinformatics*, vol. 19 (Suppl. 1), pp. 183-189, 2003.

[6] T. Hvidsten, A. Laegreid, J. Komorowski, "Learning rule-based models of biological process from gene expression time profiles using Gene Ontology", *Bioinformatics*, vol. 19, pp. 1116-23, 2003.

[7] J. Zhong, H. Zhu, Y. Li, Y. Yu, "Conceptual graph matching for semantic search", in *Proc. of Conceptual Structures: Integration and Interfaces* (ICCS-2002), Priss, U., Corbett, D., Angelova, G. (eds). Springer Verlag: London; pp. 92-106. 2002.

[8] A. Budanitsky, G. Hirst, "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures", in *Proc. of Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA. 2001.

[9] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy", in *Proc. of the 14th International Joint Conference on Artificial Intelligence*, Montreal, pp. 448-453, 1995.

[10] D. Lin, "An information-theoretic definition of similarity", in *Proc. 15th International Conference on Machine Learning,* pp. 296-304, San Francisco, 1998.

[11] R. Cho, M. Campbell, E. Winzeler et al., "A genome-wide transcriptional analysis of the mitotic cell cycle", *Molecular Cell* , vol. 2, pp. 65-73, 1998.

[12] R. Jansen, N. Lan, J. Qian, M. Gerstein, "Integration of genomic datasets to predict protein complexes in yeast", *Journal of Structural and Functional Genomics*, vol. 2, pp. 71-81, 2002.

[13] S. Raychaudhuri, J.T. Chang, F. Imam, R.B. Altman, "The computational analysis of scientific literature to define and recognize gene expression clusters", *Nucleic Acids Res.*, vol. 31, pp. 4553-60, 2003.

Acknowledgements

This research was partly supported by a visiting fellowship from the U.S-NLM. We thank Joaquin Dopazo (CNIO) and Mark Gerstein (Yale) for useful comments.