Bethesda, 7th October 2005

From: Francisco Azuaje
To: May Cheh
Cc: Olivier Bodenreider

Report of research visit, August 29th to 7th October 2005.

During this visit we have expanded our investigations on the application of Gene Ontology-driven similarity (GOS) to functional genomics. We focused on two specific problems:

- Study of significant associations between GOS and tissue-specific gene co-expression in a multi-cellular organism (mouse).
- Study of a method for reconstructing functional networks of genes and proteins using GOS in yeast.

The direct outcomes of this research are reflected in two publications:

1) H. Wang, H. Zheng, F. Azuaje, O. Bodenreider, A. Chesneau, "Linking Gene Ontology-Driven Similarity and Gene Co-Expression in Mouse", submitted to the conference on *Research in Computational Molecular Biology* (*RECOMB* 2006).

2) F. Azuaje1, O. Bodenreider, H. Wang, H. Zheng, "Gene Ontology-Driven Similarity for Supporting the Prediction of Integrated Functional Networks", to be submitted to *BMC Bioinformatics*.

This visit also allowed us to revise a paper previously submitted:

3) F. Azuaje, H.Wang, H. Zheng, O. Bodenreider and A. Chesneau, "Predictive integration of Gene Ontology-driven similarity and functional interactions", to be submitted to *Bioinformatics*.

Furthermore, the following paper was completed and submitted during this visit:

4) H. Wang, F. Azuaje, H. Zheng and O. Bodenreider, "seGOsa: Software environment for Gene Ontology-driven similarity assessment", to be submitted to *Bioinformatics*.

Please find attached papers 1 and 3 as part of this report.

Yours sincerely,

Francisco Azuaje, PhD.

Linking Gene Ontology-Driven Similarity and Gene Co-Expression in Mouse

Haiying Wang[1], Huiru Zheng[1], Francisco Azuaje[1], Olivier Bodenreider[2], Alban Chesneau[3]

[1]University of Ulster, Jordanstown, UK
[2] National Library of Medicine, National Institutes of Health, Bethesda, USA
[3] EMBL Grenoble, France,

Abstract. The integration of multiple sources of information is becoming prevalent in post-genome biology. An important task toward that goal is the quantitative assessment of relationships between relevant predictive resources. Here we assess key associations between the functional similarity of pairs of gene products and their expression correlation. Two approaches to computing similarity among genes based on knowledge extracted from the Gene Ontology are compared: *Average* and *highest average*. We investigate integrative properties of a tissue-specific data set from mouse, which describes expression profiles during retinal development, and annotations derived from the *Mouse Genome Informatics* database. We show that highly co-expressed genes are more likely to display a high degree of functional similarity for all GO hierarchies than other pairs of genes. GO-driven similarity information is best used in combination with gene co-expression as it helps narrow down the space of biological significance in terms of the number of genes and gene pairs. In practice, this integration may be useful to support the selection of relevant genes and pairs of co-expressed genes. The highest average similarity assessment method appears to be more suitable to stress functional differences between co-expressed genes. This investigation suggests that GO-driven similarity is a valuable complementary predictive resource for interactome prediction problems in multi-cellular organisms.

1 Introduction

In recent years, a number of public efforts have been focusing on the annotation and curation of gene-specific functional data. The outcome originating from these efforts can now be accessed through several databases, which provide exceptional depth and coverage of the functional data available for gene products [1]. The combination of biological knowledge extracted from diverse resources has become a fundamental goal to achieve comprehensive, large-scale functional predictions, such as the inference of networks of interactions in different model organisms. *The Gene Ontology™* (GO) [1] is one of such important functional knowledge resources, which has been designed to offer controlled, structured vocabularies to describe key domains of molecular biology across model organisms. It has traditionally facilitated the development of several organism-specific databases and enabled the implementation of cross-database queries. Moreover, its ability to provide detailed classification, controlled vocabulary and organized terminology has made it relevant to support functional predictive applications. GO has been used as a gold standard for functional prediction applications and to estimate the biological significance of gene expression analyses. *FatiGO*, for example, may be applied to identify GO terms that are significantly over- or under-represented in clusters of genes [2].

Recent research has found significant relationships between different types of functional datasets. For example, the correlation between gene co-expression, protein complex membership and gene regulatory interactions has been reported. Previous research has shown significant relationships between functional and sequence-based similarities of pairs of genes [3]. We have also demonstrated significant relationships between gene co-expression and GO-driven similarity in *Sacharomices cerevisiae* [4]. More recently GO annotations have been used to support the prediction of protein-protein interactions in *S. cerevisiae* [5]. However, significant relationships between GO-driven similarity and other functional properties have not been adequately studied in multi-cellular model organisms.

Here we investigate relationships between GO-driven similarity, which takes into account information associated with both the structure of the GO and the information content of its terms, and gene expression correlation in *Mus musculus*. The primary objective of this study is to expand our understanding of the relationships between GO-driven gene similarity and expression correlation in multi-cellar organisms. Our hypothesis is that combining GO-driven similarity and expression correlation results in better prediction of functional association than any method used in isolation. A secondary objective is to compare two approaches to aggregating between-term similarity for computing between-gene similarity.

The remainder of this paper is organized as follows. Section 2 introduces the GO and some of its applications in functional genomics, followed by a description of GO-driven similarity assessment techniques in Section 3. We implemented two approached to computing between-gene similarity, both based on an information-theoretic approach. Section 4 describes the datasets under study. Section 5 presents the results. This paper concludes with a discussion of the relevance of the results and possible applications.

2 The Gene Ontology and its Applications

2.1 The Gene Ontology

Starting in 1998 as a collaboration between three model organism databases: *FlyBase* [6], the *Saccharomyces Genome database* (SGD) [7] and the *Mouse Genome Informatics* (MGI) database [8], the GO project endeavors to provide a set of structured, controlled vocabularies and classifications for key biological domains that can be used to describe gene products in any organism [1]. The GO consists of three hierarchies that describe attributes of gene products in three non-overlapping domains of molecular biology: *Molecular function* (MF), *biological process* (BP), and *cellular component* (CC). MF represents information on the role played by individual gene products, for example *G-protein coupled receptor activity*. BP refers to a biological objective accomplished by one or more ordered assemblies of molecular function such as *signal transduction*. CC represents the cellular localization of the gene product, including cellular structure and complexes, for example *nucleus* or *anaphase-promoting complex.*

GO terms and their relationships within each hierarchy form of *directed acyclic graph* (DAG) that represents a network in which each term – except for the root – has one or more parent terms. For example, the GO term *negative regulation of cellular process* is a child of both *regulation of cellular process* and *negative regulation of biological process* in the BP hierarchy as illustrated in Fig. 1. The relationship between a child and its parent can be either "is a" (is a kind of) or "part of". The former is used when the child is more specific than its parent term (is_a relationship); the latter when the child term refers to a part while the parent term refers to the whole of which the child is a component (part_of relationship). From the BP hierarchy, for example, the term *regulation of cellular process* is a kind of *regulation of biological process* and a component of *cellular process*. The majority of GO links are "is a" links.

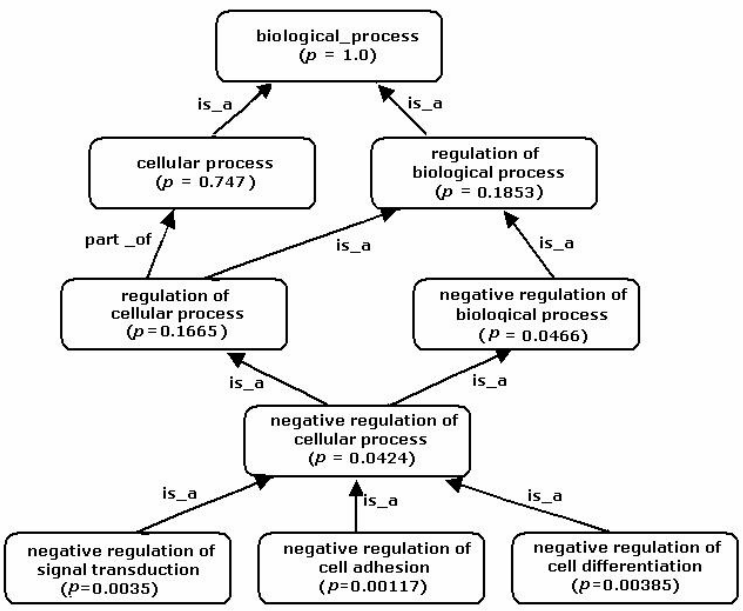

Fig. 1 Partial view of the BP hierarchy in the GO. Rounded rectangles represent terms and arrows stand for edges indicating the relationships between two terms. *p* represents the probability of finding a gene annotated to this GO term in the MGI (August 2005 release).

The terms defined by the GO have been used to annotate in a consistent way the genes and gene products described in multiple model organism databases. The source of each annotation is recorded, e.g., a literature reference, another database or a computational analysis. A standard set of *evidence codes* is used to indicate the nature of evidence on which a particular annotation is based. For example, if an annotation is inferred from the timing or location of expression of a gene, the evidence code associated with this annotation will be IEP (*inferred from expression pattern*). IEA (*inferred from electronic annotation*) is used to denote annotations that depend directly on computation or automated transfer of annotations from a database, the accuracy of which has not been verified by curators. Understandably, such annotations tend to be less reliable. The products of GO-driven projects, including vocabularies, annotations, databases and accompanying tools, are freely available from the GO website: http://www.geneontology.org/.

2.2 Overview of Applications of the Gene Ontology to Functional Genomics

It has been demonstrated that the GO may facilitate large-scale applications for functional geneomics. One such application is the integration of GO annotations into gene expression data clustering tasks [2]. Such an application is now referred to as *ontological analysis* of gene expression data and is becoming the *de facto* standard for post-processing high throughput experiments [9]. Examples of ontological analysis tools include FatiGO [2] and GOToolBox [10]. A comprehensive review of these systems can be found in [9].

In addition to providing gene annotations, the GO also provides a structure for organizing genes into biologically relevant groupings. Such information can be used as an important *prior* biological knowledge base to facilitate functional prediction, hypothesis generation and validation studies. For example, based on the analysis of phenotypic annotations extracted from the *Munich Information Center for Protein Sequences* (MIPS) and GO annotations, King *et al.* [11] inferred gene-phenotype associations in yeast using decision trees. Lægreid *et al.* [12] used supervised learning methods to predict GO biological process annotation terms from microarray-derived time-series gene expression data. Adryan and Schuh [13] recently developed a clustering system

that incorporates GO information for selecting subsets of gene expression data. Hierarchical clustering based on the Pearson's correlation coefficient was applied to those genes with GO terms defined by the user.
Unlike most methods which rely solely on the hierarchical organization of GO terms, our approach to computing gene similarity takes advantage of the information content of GO terms. The following section introduces the problem of measuring between-term and between-gene similarity in the GO.

3 Gene Product Similarity Measurement Using GO Annotations

The similarity between two genes g_1 and g_2 is computed by aggregating at the gene level the similarity values computed at the term level between the GO terms to which these genes have been annotated. The methods used for computing between-term similarity and aggregating schemes at the gene level are presented below.

Between-term similarity

The first step towards measuring similarity among gene products using GO annotations is to establish between-term similarity within each hierarchy. Given a pair of terms, c_1 and c_2, traditional methods for measuring their similarity are based on an edge counting approach, i.e. counting the number of edges between the nodes associated with these two terms in the ontology. A small number of edges between two terms corresponds to highly similar terms. Conversely, terms farther apart tend to be less similar. One of the main limitations of this approach is that it assumes that nodes and edges are uniformly distributed in an ontology, which is not an accurate assumption in the GO because it exhibits variable link densities. For example, the pair "*cellular process*" and "*regulation of biological process*" has the same similarity as the pair "*negative regulation of cell adhesion*" and "*negative regulation of cell differentiation*" by using this method. This is because these two pairs have an immediate common parent term. However, terms in the latter pair appear to be *semantically* more closely related than in the former.
An alternative method to measure similarity between two terms is based on the assessment of their *information content*, which exploits *information-theoretic* principles to reflect semantic similarity between two terms. Let C be the set of terms in the GO. The *information content* (*IC*) of a term, $c \in C$, can be quantified as follows:

$$IC(c) = -\log(p(c)) \tag{1}$$

Where *p(c)* is the probability of finding term *c* and a child of *c* in the annotation database under analysis. Based on the assumption that the more information two terms share in common, the more similar they are, three semantic similarity measures have been developed. Resnik's [14], Lin's [15] and Jiang's [16] metrics have been studied elsewhere as possible approaches to calculating GO-driven similarity [3], [4]. Resnik's method does not differentiate the similarity of any pair of terms in a sub-hierarchy as long as they have the same lowest common ancestor. The Resnik's values can vary between 0 and infinity, which is not a straightforward way to reflect similarity. Jiang's method deals with similar issues, but high values reflect dissimilarity rather than similarity. Lin's similarity model has been shown to produce both biologically meaningful and consistent similarity predictions. Given terms, c_i and c_j , their Lin's similarity is defined as:

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

Where $S(c_i, c_j)$ represents the set of parent terms shared by both c_i and c_j, 'max' represents the maximum operator, and $p(c)$ is the probability of finding c or one of its children in the annotation database being analyzed. It generates normalized similarity values between 0 and 1.

Between-gene similarity

After calculating the similarity between GO terms describing two gene products, it is then possible to establish the similarity between these two gene products. The basic idea is to combine the calculated similarities from the sets of GO terms used to describe the gene products. 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, Lord *et al*. [3] used average values determined as the average inter-set similarity between terms from A_i and A_i, as shown in Equation (3).

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

where m and n are the number of terms included in A_i and A_j respectively, and $sim(c_k, c_p)$ can be calculated using Lin's model. This approach has been previously applied to structural and functional genomics. Nevertheless, this method does not always produce meaningful results. For example, intuitively, the similarity between two genes having the same sets of annotation terms is expected to be equal to 1. However, this is not true when several annotations within a hierarchy are assigned to the genes. It will define, for instance, $SIM(g_i, g_j) = 0.5$, for $g_i = g_j$ when A_i and A_j are described by the same set of annotations with more than one GO term within a hierarchy. In order to address such a limitation we have introduced an alternative approach that selectively aggregates maximum inter-set similarity values [17] as follows:

$$SIM(g_i, g_j) = \frac{1}{m+n} \times (\sum_k \max_p(sim(c_k, c_p)) + \sum_p \max_k(sim(c_k, c_p))) \tag{4}$$

From now on we will refer to the aggregation schemes based on (3) and (4) as the *average* and *highest average* similarity methods respectively. The procedure to establish gene-gene similarity using GO annotations in this study is summarized in Table 1.

Table 1 GO-Based Similarity Measure

1: **Initialization**: Download the latest version of GO database Find out the total number of gene products, N, associated with the root node of each ontology.
2: **Repeat** establishing the information content for each GO term
3: Find out the number of gene products associated with each GO term, c, and its child terms, $freq(c)$.
4: Calculate the probability of finding a child of c in the annotation database being analyzed as follows: $p(c) = \frac{freq(c)}{N}$
5: Compute the information content for each GO term using Equation (1)
6: Fill in GO information content table
7: **Until** there is no term left
8: **Repeat** estimating similarity value for each pair of genes, g_i and g_j.
9: Find a set of GO terms associated with each gene, A_i and A_j.
10: Calculate the Lin's similarity value between terms from A_i to A_j using Equation (2).

11: Compute the average and maximum inter-set similarity values for each gene pair using Equations (3) and (4).

12: **Until** There is no more gene pair left

4 Data and Methods

The GO annotations derived from the MGI (August 2005 release of the GO database) were analyzed to calculate the functional similarity of mouse gene products. Experiments ignored IEA annotations due to their lack of reliability. We concentrated on a mouse retina dataset from Dorrell *et al.* study [18], which contains gene expression profiles of thousands of genes in eight different time points (P0, P4, P8, P10, P12, P14, P21, and adult (P42)) during postnatal mouse retinal development. This is the first dataset describing global gene expression profiles in the developing postnatal mouse retina. It reflects different expression patterns during postnatal retinal development such as glial and neuronal differentiation, vascularization, and the onset of vision. A detailed description of this dataset can be found in [18].

Our analysis includes about one million gene pairs derived from this dataset for which GO annotations are available. For each pair of genes, the GO-based similarity in each hierarchy was compared to the absolute expression correlation value. Expression correlation was calculated using the *Pearson correlation coefficient*. The comparison was done separately for the three hierarchies of the GO.

5 Results

Fig. 2 summarizes the relationship between GO-driven highest average similarity (under BP) and the absolute expression correlation between pairs of gene products. For these and all subsequent figures, 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. Gene pairs exhibiting absolute expression correlation close to 1 exhibit the highest similarity values. By excluding weakly correlated genes (absolute correlation values lower than 0.5) such a difference is even more significant (right panel of Fig. 2). Similar trends were obtained from the average similarity method (Fig. 3)

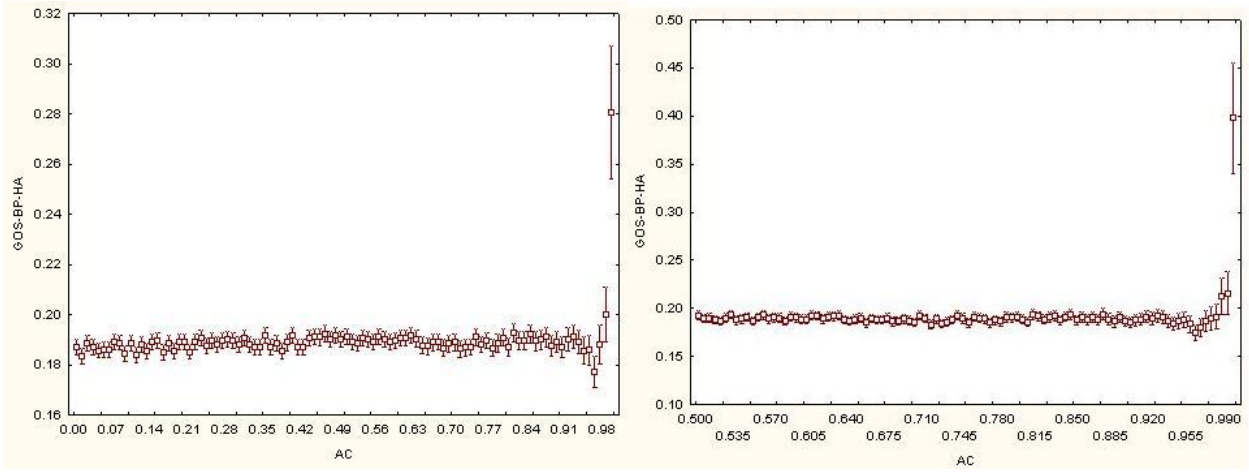

Fig. 2 GO-driven similarity (GOS) and absolute expression correlation (AC) for **BP hierarchy**. Mean similarity values for each correlation interval and their 95% confidence intervals. Between-

gene similarity was calculated using the **highest average similarity method**. Right panel depicts relationships after excluding a set of weakly correlated genes (AC values lower than 0.5).

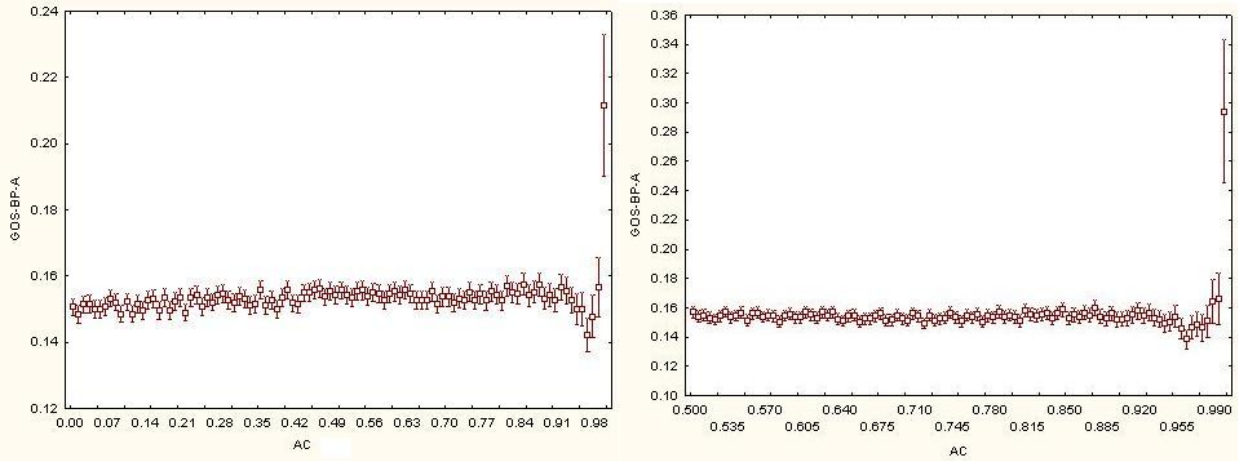

Fig. 3 GO-driven similarity (GOS) and absolute expression correlation (AC) for **BP hierarchy**. Mean similarity values for each correlation interval and their 95% confidence intervals. Between-gene similarity was calculated using the **average similarity method**. Right panel depicts relationships after excluding a set of weakly correlated genes (AC values lower than 0.5).

Figs. 4 and 5 show the results obtained from the application of the highest average and average similarity methods under the CC hierarchy respectively. Figs. 6 and 7 summarize the relationships with regard to highest average and average similarity methods under the MF hierarchy respectively. Like in the BP hierarchy, in the CC and MF hierarchies only the most highly correlated gene pairs tend to exhibit the highest similarity values. Moreover, the average and highest average similarity methods are able to capture similar relationships. They differ in the sense that in the latter method the distinction between the highest and lower correlation genes is more pronounced.

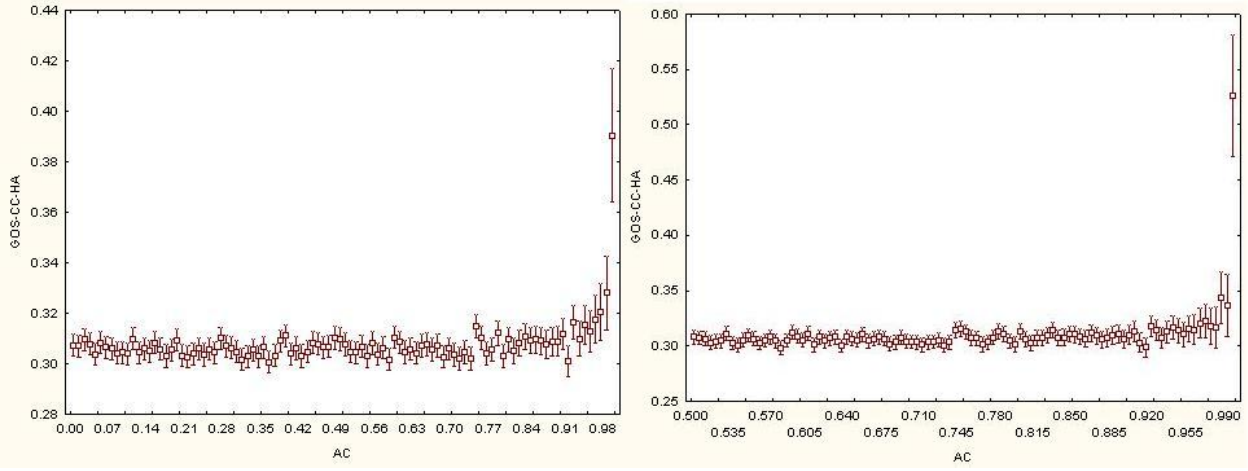

Fig. 4 GO-driven similarity (GOS) and absolute expression correlation (AC) for **CC hierarchy**. Mean similarity values for each correlation interval and their 95% confidence intervals. Between-gene similarity was calculated using the **highest average similarity method**. Right panel depicts relationships after excluding a set of weakly correlated genes (AC values lower than 0.5).

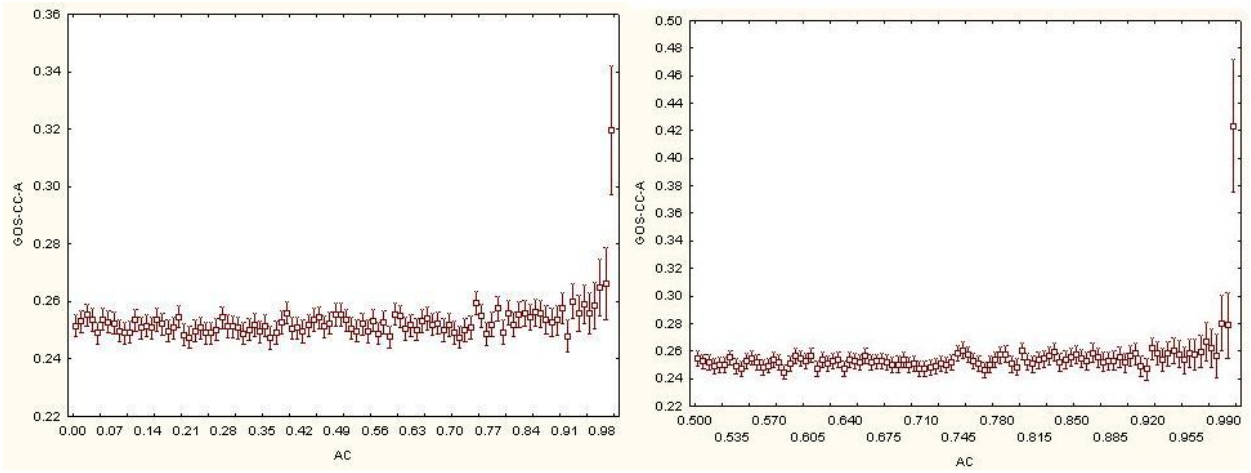

Fig. 5 GO-driven similarity (GOS) and absolute expression correlation (AC) for **CC hierarchy**. Mean similarity values for each correlation interval and their 95% confidence intervals. Between-gene similarity was calculated using the **average similarity method**. Right panel depicts relationships after excluding a set of weakly correlated genes (AC values lower than 0.5).

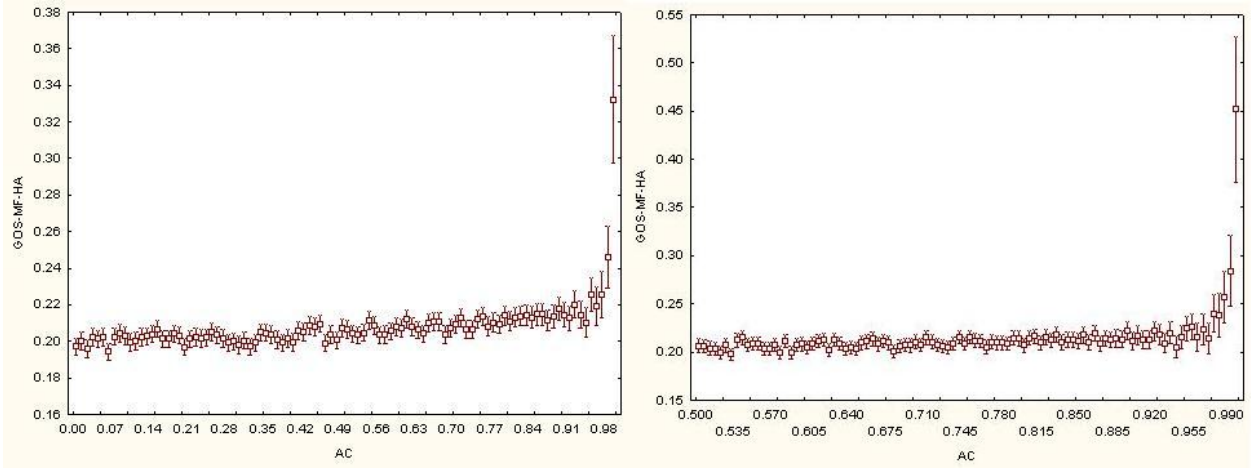

Fig. 6 GO-driven similarity (GOS) and absolute expression correlation (AC) for **MF hierarchy**. Mean similarity values for each correlation interval and their 95% confidence intervals. Between-gene similarity was calculated using the **highest average similarity method**. Right panel depicts relationships after excluding a set of weakly correlated genes (AC values lower than 0.5).

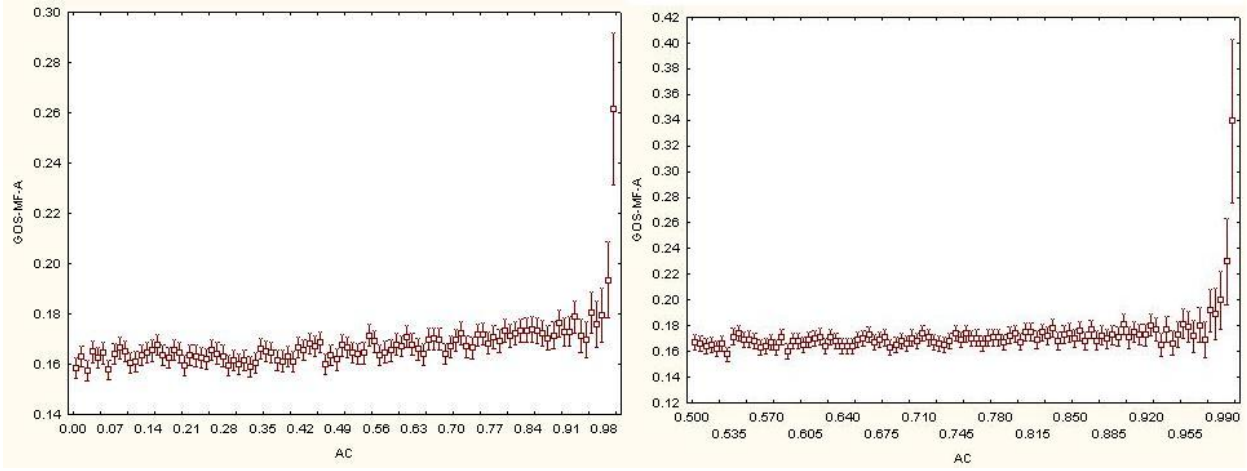

Fig. 7 GO-driven similarity (GOS) and absolute expression correlation (AC) for **MF hierarchy**. Mean similarity values for each correlation interval and their 95% confidence intervals. Between-gene similarity was calculated using the **average similarity method**. Right panel depicts relationships after excluding a set of weakly correlated genes (AC values lower than 0.5).

6. Discussion

Functional similarity and co-expression. This investigation demonstrated significant relationships between gene co-expression patterns and GO-driven similarity of pairs of genes for all three GO hierarchies. Overall, highly co-expressed genes tend to exhibit higher GO-driven similarity values than weakly co-expressed genes. However, a large number of highly co-expressed genes also have relatively small GO-driven similarity values. These results further illustrate the limitations of functional prediction models solely based on gene expression data due to the presence of multiple spurious functional associations. The results also stress the complexity of the relationship between gene co-expression and functional associations in multi-cellular organisms.

Assessment of GO-driven similarity. This research has also allowed us to assess two methods for computing GO-driven similarity: average and highest average. Overall, both methods are able to represent similar patterns. However, the results produced by the highest average method tend to be more easily interpretable ([0-1] range for similarity) and more suitable to quantitatively highlight functional differences between co-expressed genes (larger differences between groups of co-expressed genes).

Biological interpretation. In terms of protein interactions, Figures 2 to 7 suggest the involvement of two main groups of genes. The first group is characterized by highly co-expressed genes with a high degree of GO-driven similarity, which may reflect the presence of complexes in this dataset. The second group (corresponding to the majority of the genes in the dataset) exhibits relatively low levels of co-expression and GO-driven similarity. This group may reflect – since a majority of these pair of genes are not coding for interacting proteins – transient interactions at particular intervals of the retinal development, such as transcription factor-cofactors interactions required to activate a developmental gene. In this case we may expect pairs of genes showing inconsistent relations between co-expression and GO-driven similarity, even when both products are linked into a pairwise interaction.

Applications. This study confirms the feasibility of applying GO-driven similarity approaches to support the prediction of significant functional associations in multi-cellular organisms. More precisely, it suggests that GO-driven similarity and co-expression data may be best used in

combination. Practically, it is difficult to establish a biological threshold for co-expression correlation: 0.9 or 0.8 would seem like high correlation values, but were shown to reflect limited functional similarity. In contrast, the use of GO-driven similarity information for these genes helps narrow down the space of biological significance. On most figures above, there is a breaking point around the value of 0.98 for co-expression correlation. In other words, the slope of the curve GO driven similarity/co-expression correlation differs dramatically on both sides of this value. This finding provides *biological motivation* for selecting the value of 0.98 (in this case) as the threshold for highly co-expressed genes. As shown in Fig. 8, threshold selection has important consequences in terms of the amount of human resources required for interpreting these data. In this example, setting the threshold to T=0.98, only 1440 pairs of genes would require further examination, compared to 8648 for T=0.95 and 104566 for T=0.80. Without the supporting evidence provided by GO-driven similarity, the threshold for co-expression correlation would have to be set to an arbitrary value. Setting this threshold too low would result in high costs for interpreting the data. Conversely, potentially relevant associations would be missed if the threshold is set too high.

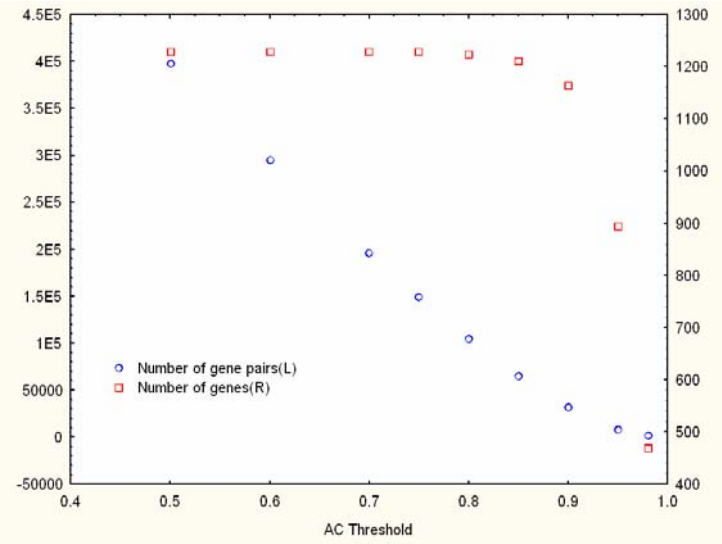

Fig. 8. Relationships between the number of genes, gene pairs and co-expression thresholds for BP hierarchy using the highest average similarity method.

The results also suggest that, in the case of highly correlated genes, the GO-driven similarity approach may be used to predict GO terms for partially characterized gene products. One of the key applications may be, for example, to assign potentially novel GO terms to pairs of gene products based on their co-expression patterns. In order to support this task another important goal is to define in more detail which GO hierarchy encodes the most significant relationship with gene co-expression for specific groups of genes.

Future work. To the best of our knowledge, this study is the first integrative analysis of GO-driven similarity and gene co-expression reported for this *M. musculus*. We will expand this research for other multi-cellular organisms including *C. elegans* and *Homo sapiens*, as well as different tissue-specific expression datasets. This will allow us to gain a more complete view of the overlapping predictive properties between these types of data, which may further justify their incorporation, for instance, into large-scale interactome inference models.

Acknowledgement

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).This work was done while Francisco

Azuaje was a visiting scholar at the Lister Hill National Center for Biomedical Communications, NLM, NIH.

References

1. The Gene Ontology Consortium: Creating the gene ontology resource: Design and implementation. Genome Research, 11 (2001) 1425-1433.
2. Al-Shahrour, F., Díaz-Uriarte, R. and Dopazo, J.: FatiGO: a web tool for finding significant associations of Gene Ontology terms to groups of genes. Bioinformatics. 20 (2004) 578–580.
3. Lord, P., Stevens, R., Brass, A. and Goble, C.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics, 19 (2003) 1275-1283.
4. Wang, H., Azuaje, A., Bodenreider, O. and Dopazo, J.: Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In the Proc. of IEEE 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology, La Jolla, CA, USA, (2004) 25-31.
5. Lu, L.J., Xia, Y., Paccanaro, A., Yu, H., Gerstein, M.: Assessing the limits of genomic data integration for predicting protein networks. Genome Research, **15** (2005) 945-53.
6. The FlyBase Consortium: The FlyBase database of the Drosophila genome projects and community literature. Nucleic Acids Research, **31** (2003) 172-175.
7. Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, Schroeder, L., M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D. and Cherry, J. M.: Saccharomyces genome database (SGD) provides secondary gene annotation using the gene ontology (GO). Nucleic Acids Research, 30(1), (2002) 69-72.
8. Blake, J. A., Richardson, J. E., Bult, C. J., Kadin, J. A., Eppig, J. T., and the members of the Mouse Genome Database Group: MGD: The Mouse Genome Database. Nucleic Acids Res, **31** (2003) 193-195.
9. Khatri, P. and Drăghici, S.: Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics, 21(18), (2005)3587-3595.
10. Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., and Jacq, B.: GOToolBox : functional investigation of gene datasets based on Gene Ontology. Genome Biology 2004; 5(12):R101.
11. King, O. D., Foulger, R. E., Dwight, S. S., White, J. V., and Roth, F. P.: Predicting gene function from patterns of annotation. Genome Research, **13** (2003) 896-904.
12. Lægreid, A., Hvidsten, T. R., Midelfart, H., Komorowski, J. and Sandvik, A. K.: Predicting gene ontology biological process from temporal gene expression patterns. Genome Research, 13 (2003) 965-979.
13. Adryan, B. and Schuh, R.: Gene-Ontology-based clustering of gene expression data. Bioinformatics, 20(16), (2004) 2851-2852.
14. Resnik, P. and Diab, M.: Measuring verb similarity. In Proc. of Twenty Second Annual Meeting of the Cognitive Science Society (COGSCI2000), Philadelphia, August 2000.
15. Lin, D.: An information-theoretic definition of similarity. In Proc. of 15th International Conference on Machine Learning, San Francisco, (1998) 296-304.
16. Jiang, J. J. and Conrath, D. W.: Semantic similarity based on corpus statistics and lexical taxonomy. In Proc. of International Conference on research in Computational Linguistics, Taiwan (1998).
17. Azuaje, F., Wang, H., and Bodenreider, O.: Ontology-driven similarity approaches to supporting gene functional assessment. In *Proc. Of The Eighth Annual Bio-Ontologies Meeting*, Michigan, 25 June 2005, http://bio-ontologies.man.ac.uk/.

18. Dorrell, M., Aguilar, E., Weber, C., and Friedlander, M.: Global gene expression analysis of the developing postnatal mouse retina. Investigative Ophthalmology & Visual Science, 45(3), (2004) 1009-1019.