Proceedings of the 18th IEEE International Symposium on Computer-Based Medical Systems (CBMS'2005) 2005:389-394.

An Ontology-Driven Clustering Method for Supporting Gene Expression Analysis

Haiying Wang[1], Francisco Azuaje[1], Olivier Bodenreider[2]
[1] School of Computing and Mathematics, University of Ulster, Jordanstown, UK
[2] National Library of Medicine, National Institutes of Health, Bethesda, U.S.A
Emails : hy.wang@ulster.ac.uk, fj.azuaje@ieee.org, , olivier@nlm.nih.gov

Abstract

The Gene Ontology (GO) is an important knowledge resource for biologists and bioinformaticians. This paper explores the integration of similarity information derived from GO into clustering-based gene expression analysis. A system that integrates GO annotations, similarity patterns and expression data in yeast is assessed. In comparison with a clustering model based only on expression data correlation, the proposed framework not only produces consistent results, but also it offers alternative, potentially meaningful views of the biological problem under study. Moreover, it provides the basis for developing other automated, knowledge-driven data mining systems in this and related application areas.

1. Introduction

A modern approach to systems biology integrates different knowledge sources to make large-scale datasets, such as gene expression data, meaningful. Expression data clustering is a fundamental tool to support functional predictions. Based on the assumption that genes exhibiting similar expression patterns should be co-regulated, and therefore contained in the same functional pathway, data clustering techniques have fueled several potential applications to disease diagnosis and therapy design [1]. However, due to the complexity of the biological problems under study and the lack of complete experimental and analytical models, there is a need to design automated, knowledge-driven techniques to assist in the explanation and validation of predictive outcomes.

It has been shown that traditional, data-driven clustering approaches lack the ability to automatically describe the biological meaning of similarity relationships represented in the clusters [2]. These methods mainly generate lists of similar genes with respective to expression levels, which may not necessarily reflect prior knowledge. Thus, biologists apply semi-automated procedures to describe clusters in terms of their functional composition using existing knowledge bases (e.g. annotations), which may be a complex and time-consuming task [3].

The *Gene Ontology™* (GO) is one such important functional knowledge source [4]. This paper focuses on the integration of similarity information derived from GO to support clustering-based gene expression analysis. The remainder of this paper is organized as follows. Section 2 introduces GO and relevant applications. A framework that incorporates GO-driven similarity information into a clustering process is proposed in Section 3, followed by results obtained from the analysis of a gene expression dataset in *S. cerevisiae* (yeast). The paper concludes with a discussion of limitations and potential applications of the methods studied.

2. Gene Ontology and its applications to clustering–based analysis

GO [4] provides a set of controlled, structured vocabularies to describe key functional aspects in different organisms. It comprises three independent *hierarchies* that define functional attributes of gene products: *Molecular function* (MF), *biological process* (BP), and *cellular component* (CC). Each hierarchy consists of *directed acyclic graphs* (DAGs) of terms, in which each term may be linked to more than one parent term. For example, the GO term *regulation of development* is a child of both *development* and *regulation of biological process* in the BP hierarchy (Figure 1). There are two types of child-to-parent relationships in GO: “is a” and “part of” types. A child term more specialized than its parent term (is_a relationship) or a component of its parent term (part_of relationship). From the BP ontology, for example, the term *regulation of cellular process* is a child of *regulation of biological process* and part of *cellular process* (Figure 1).

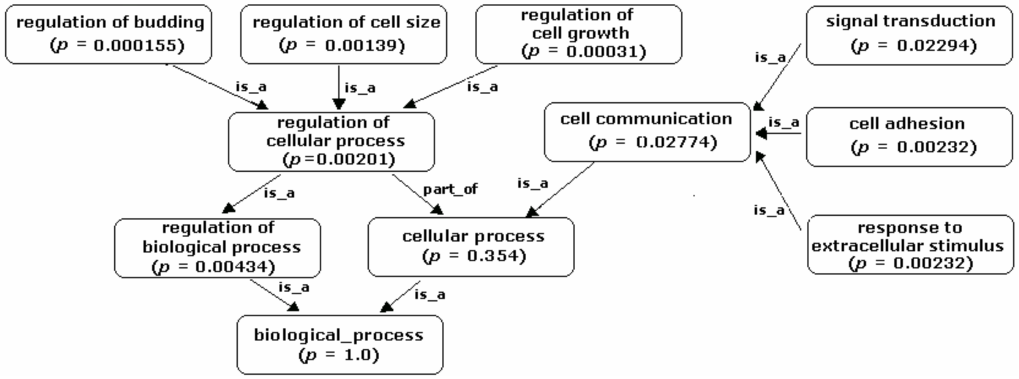
regulation of budding (p = 0.000155)
regulation of cell size (p = 0.00139)
regulation of cell growth (p = 0.00031)
signal transduction (p = 0.02294)
is_a
is_a
is_a
is_a
regulation of cellular process (p=0.00201)
cell communication (p = 0.02774)
is_a
cell adhesion (p = 0.00232)
is_a
part_of
is_a
is_a
regulation of biological process (p = 0.00434)
cellular process (p = 0.354)
response to extracellular stimulus (p = 0.00232)
is_a
is_a
biological_process (p = 1.0)

Figure 1 Partial view of the BP hierarchy in GO. Rounded rectangles represent terms and arrows stand for edges indicating the relationships between two terms. *p* represents the probability of finding a GO term in the Saccharomyces Genome Database (SGD) (February 2004 release).

GO is becoming the *de facto* standard for annotating gene products. The widespread adoption of GO to annotate genes facilitates cross-species/cross-database queries. However, its significance is not limited to annotation applications. GO may facilitate large-scale predictive applications in functional genomics. The inclusion of GO annotations in gene expression studies may help to explain why a particular group of genes share similar expression patterns. It also helps to identify functionally-enriched clusters of genes. *FatiGO* [5], for example, extracts GO terms that are significantly over- or under-represented in clusters of genes. Adryan and Schuh [6] 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. However, these approaches do not fully exploit the knowledge that can be extracted from analyzing functional relations of GO terms and their information content in different annotation databases. Moreover, there is a need to offer alternative GO-driven clustering methods to improve the predictive accuracy and biological relevance.

3. A GO-Driven Approach to Hierarchical Clustering

3.1 A GO-based distance function

In order to incorporate GO knowledge into a clustering algorithm, we first implemented similarity/distance measures that take into account topological and information content features encoded in the GO hierarchies. Such techniques are referred to as *semantic similarity assessment* approaches, which have been previously investigated by the authors [7].

Based on the assumption that the more information two terms share in common, the more similar they are, three semantic similarity measures: Resnik's, Lin's and Jiang's metrics, have been studied as possible approaches to GO-driven clustering analysis [7]. Lin's similarity model has shown to produce both biologically meaningful and consistent similarity predictions [7]. Given terms, c_i and c_j , their Lin's similarity is defined as:

$$sim(c_i, c_j) = \frac{2 \times \max_{c \in S(c_i, c_j)} [\log(p(c))]}{\log(p(c_i)) + \log(p(c_j))} \qquad (1)$$

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. The similarity of a pair of genes is computed as the average similarity between terms from the two genes (as described in [7]).

3.2 A GO-driven, hierarchical clustering framework

The incorporation of GO-driven similarity information into a clustering algorithm is summarized in Figure 2. For a given gene pair, data-driven similarity values were calculated with the *Pearson correlation coefficient* and GO-driven similarity values were calculated using Lin's semantic similarity model. Thus, both data- and GO-driven similarity matrices and different types of hierarchical clustering schemes were implemented.

3.3 Gene expression and GO annotation datasets

GO annotations derived from the *Saccharomyces Genome Database* (SGD), February 2004 release, were analyzed to calculate similarity using Lin's model. Experiments ignored *IEA* annotations (Inferred from Electronic Annotation) due to their lack of reliability. The expression data originated from a study by Eisen *et al.* [8], which contains responses to several perturbations in yeast. Each gene is described by 79 expression values that are associated with 79 time points during several important conditions [8]. Eisen *et al.* systematically analyzed 2467 genes and identified 10 relevant groups of co-expressed genes. Table 1 shows the distribution of genes over these 10 groups.

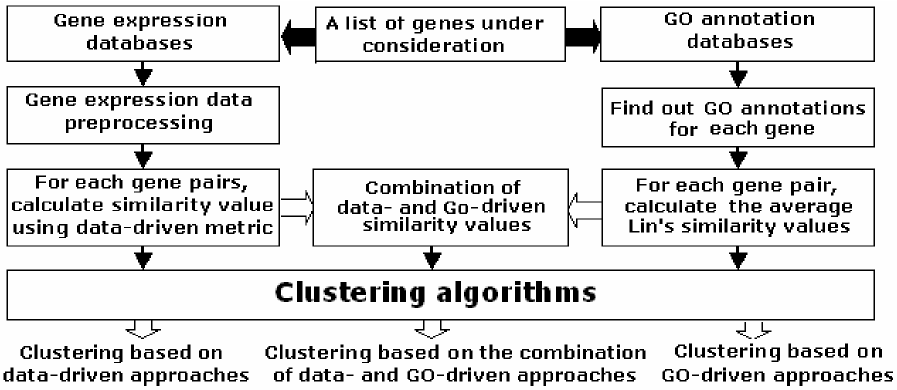
Gene expression databases
A list of genes under consideration
GO annotation databases
Gene expression data preprocessing
Find out GO annotations for each gene
For each gene pairs, calculate similarity value using data-driven metric
Combination of data- and Go-driven similarity values
For each gene pair, calculate the average Lin's similarity values
Clustering algorithms
Clustering based on data-driven approaches
Clustering based on the combination of data- and GO-driven approaches
Clustering based on GO-driven approaches

Figure 2 A framework for incorporating GO-driven similarity information into clustering.

Table 1 Distribution of Genes over ten clusters identified by Eisen *et al*.

Cluster	B	C	D	E	F	G	H	I	J	K
Number of genes	11	27	14	17	22	15	8	126	5	16

4. Results

4.1 GO-Driven Cluster Interpretation

GO-driven similarity information was first generated to assess clusters initially obtained only with expression correlation (data-driven clustering). The distribution of Lin's similarity values over the 10 clusters analyzed by Eisen *et al.* [8] is shown in Figure 3. The significance of the differences between these clusters in terms of their GO-driven similarity was established by a one-way ANOVA. The results shown in Table 2 confirm that the functional differences between these clusters are significant (p < 0.0005).

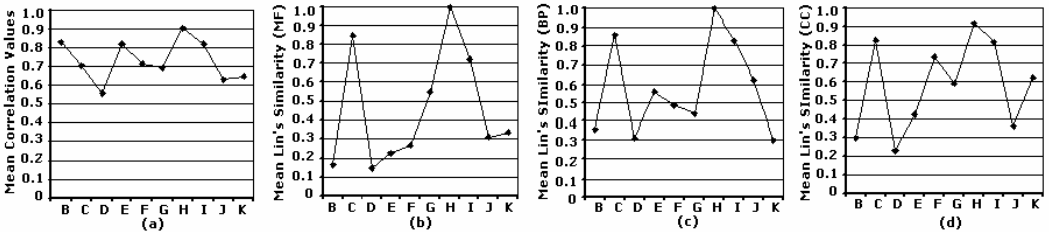

Figure 3 The distribution of (a) Pearson Correlation; (b) Lin's Similarity (MF); (c) Lin's Similarity (BP); (d) Lin's Similarity (CC) over the ten clusters analyzed by Eisen *et al.*

Table 2 *F* values and significance levels across the GO hierarchies using one-way ANNOVA analysis

Ontology	*F* value	Significance
MF	132.03	*p* < 0.0005
BP	324.46	*p* <0.0005
CC	180.05	*p* <0.0005

For most of the clusters, the results shown in Figure 3 are consistent with previous research on the relationship between GO-driven similarity and expression correlation: Clusters exhibiting stronger expression correlation tend to have higher GO-based similarity values. For example, the average expression correlation and Lin's similarity values in Cluster H, which includes eight histone genes, are all greater than 0.90. This is also consistent with results obtained by Hereford *et al.*[9] that indicated that these genes are co-regulated. Similar trends can be observed from Clusters C and I. An inconsistency was found in Cluster B, which shows a relatively high mean expression correlation (0.83) and a low mean Lin's similarity across all the GO hierarchies (Figure 3). In the case of the MF hierarchy, for instance, the mean Lin's similarity value for Cluster B was equal to 0.16 with more than half of its gene pairs showing similarity values equal to zero. It might highlight the functional diversity exhibited by this cluster. Further analyses with the FatiGO system confirm this hypothesis. Eleven genes from this cluster are significantly associated with six molecular functions (at the MF level 3): *structural constituent of cytoskeleton*, *protein binding*, *lipid binding*, *hydrolase activity*, *liqase activity* and *kinase regulator activity*.

4.2 GO-Driven Hierarchical Clustering

Average-linkage hierarchical clustering using Lin's similarity model was implemented on 261 genes with GO annotations obtained from SGD, which were included in the 10 groups identified by Eisen *et al.* [8]. The results are shown in Figure 4. The 79-dimensional gene expression vectors associated with 79 separate time courses are visualized as a heatmap, in which red, black and green in the original pcture represent up-regulated, unchanged and down-regulated genes respectively.

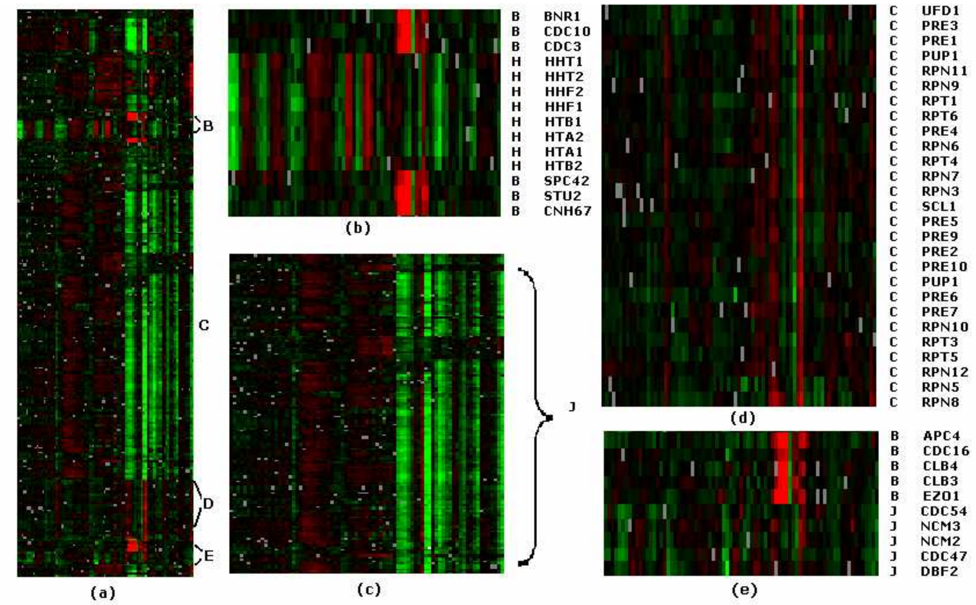

Figure 4 GO-driven hierarchical clustering of 261 genes included in the 10 groups analyzed by Eisen *et al.* based on Lin's BP similarity values. (b), (c), (d) and (e) are the zoomed images of four marked areas B, C, D and E in (a). The cluster labels used by Eisen *et al.* and the list of gene symbols for each cluster are included next to the zoomed images.

In general these results are consistent with the clusters generated by Eisen *et al.* using only gene expression correlation [8]. For example, 5 genes in Cluster J, 8 genes in Cluster H, 27 genes in Cluster C and 126 genes in Cluster I are grouped together by the GO-driven clustering (Figures 4 (b) to (e)). This confirms that genes belonging to the same cluster participate in common biological processes. A FatiGO analysis further supports this observation. For example, Cluster C contains a significantly higher percentage of genes involved with protein catabolism than other clusters (100% of genes). A similar observation applies to Clusters H, J and I. A closer examination of Cluster H shows that its 8 histone genes were also assigned to the same cluster using the GO-driven approach for all the hierarchies. These results are also consistent with the findings shown in the Section 4.1. A closer look at the genes assigned to each cluster additionally stresses the advantages of GO-driven clustering methods. For example, 11 genes from Cluster B were separated into two groups using the GO-driven clustering. As illustrated in Figure 4, 6 genes involved in *cell organisation and biogenesis* (BNR1, CDC10, CDC3, SPC42, STU2, CNH6) were clustered with 8 histone genes. The other 5 genes (APC4, CDC16, CLB4, CLB3, EXO1) are involved in *cell proliferation*, and they were grouped with the genes belonging to Cluster J, which is also associated with *cell proliferation*. These results illustrate the capacity of a GO-driven clustering to detect relevant functional relationships that may not be represented by a data-driven clustering algorithm. Similar results were observed when using similarity information from the MF and CC hierarchies.

5. Discussions and Conclusion

This paper presented a clustering strategy that incorporates similarity information extracted from GO. Its results were compared with the clusters obtained from a data-driven clustering method, which was solely based on gene expression correlation. The results were in general consistent. However, the GO-driven method may be able to identify functional relationships and differences, which may not be identified by traditional data-driven clustering. Moreover, similarity information derived from GO can be used to interpret data-driven clustering results in a more biologically meaningful way. It may provide indicators to detect irrelevant expression correlations between pairs of genes within a cluster. This investigation suggests that these approaches may lead to more biologically meaningful clusters. Genes with similar functions tend to be clustered together. Additionally, it might support the identification of genes with similar expression patterns that may actually be involved in different biological pathways.

Speer *et al.* [10] incorporated Lin's similarity metric into a *Memetic Clustering Algorithm (MCA)* to study human fibroblasts expression data. Their method may also detect clusters of functionally related genes. Unlike our study, Speer *et al.* adopted maximum similarities/minimum distances in their clustering analysis. They assumed that only single term-term similarity is required to measure gene-gene similarity. However, Lord *et al.* [11] have indicated that this may not always be an accurate assumption. Future research will include a comparison between our approach and the MCA. These techniques should be tested on data from other organisms. We plan to continue studying relationships between expression correlation, gene co-regulation and GO-driven similarity. Analyses on recent releases of GO and SGD are being conducted.

6. References

[1]. F. Azuaje, “Clustering-based approaches to discovering and visualizing expression patterns”, *Briefings in Bioinformatics*, 4 (1), pp. 31-42, 2003.

[2]. A. Lagreid., T. R. Hvidsten, H. Midelfart, J. Komorowski, and A. K. Sandvik, “Predicting gene ontology biological process from temporal gene expression patterns”, *Genome Res*., 13(5), pp. 965-79, 2003.

[3]. D. A. Hosack, G. Dennis, B. T. Sherman, H. C. Lane, and R. A. Lempicki, “Identifying biological themes within lists of genes with EASE”, *Genome Biology*, 4:R70, 2003.

[4]. The Gene Ontology Consortium, “Gene ontology: Tool for the unification of biology.” *Nat. Genet*. 25, pp. 25-29, 2000.

[5]. F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo, “FatiGo: a web tool for finding significant associations of Gene Ontology terms with groups of genes”, *Bioinformatics*, 20(4), pp. 578-580, 2004.

[6]. B. Adryan and R. Schuh, "Gene-Ontology-based clustering of gene expression data", *Bioinformatics,* 20(16), pp. 2851-2852, 2004.

[7]. H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo, “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, 2004, pp. 25-31.

[8]. 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*, 95, 14863—14868, 1995.

[9]. L. M. Hereford, M. A. Osley, T. R. 2nd. Ludwig, and C. S. mcLaughin, “Cell-cycle regulation of yeast histone mRNA.” *Cell*, 24(2), pp. 367-75, 1981.

[10]. N. Speer, C. Spieth, and A. Zell, “A memetic clustering algorithm for the functional partition of genes based on the gene ontology.” In *Proc. of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, La Jolla, CA, USA, 2004, pp. 252-259.

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