Gene-Ontology-based clustering of gene expression data
Boris Adryan* and Reinhard Schuh*

Department of Molecular Developmental Biology, Max-Planck-Institute for Biophysical Chemistry, 37077 Göttingen, Germany

Received on September 20, 2003; revised on March 9, 2004; accepted on April 19, 2004
Advance Access publication April 29, 2004

ABSTRACT
Summary: The expected correlation between genetic co-regulation and affiliation to a common biological process is not necessarily the case when numerical cluster algorithms are applied to gene expression data. GO-Cluster uses the tree structure of the Gene Ontology database as a framework for numerical clustering, and thus allowing a simple visualization of gene expression data at various levels of the ontology tree.
Availability: The 32-bit Windows application is freely available at http://www.mpibpc.mpg.de/go-cluster/
Contact: boris.adryan@mpi-bpc.mpg.de; rschuh@gwdg.de
Supplementary information: A comprehensive software manual, hints for troubleshooting and commented examples are available from the website.

INTRODUCTION
Gene expression profiling by microarrays results in a large amount of numerical tabular data. Many statistical methods for normalization and data comparison between experiments, such as analysis of variance (ANOVA) or principal component analysis (PCA) are available and belong to a set of procedures that can be easily automated. However, the essential step in the analysis of those experiments remains to be the biological interpretation by manual inspection.

One strategy for the analysis of gene expression data is hierarchical clustering, which creates a binary tree of genes ordered by their similarity of regulation. The measure for this similarity is usually a correlation coefficient that compares the regulation of one gene with the regulation of another one at various experimental conditions or time points. Also, SOM (self organizing maps) and K-means clustering are widely used for the grouping of microarray datasets. All the algorithms mentioned are standard and are implemented in various programs, i.e. Cluster (Eisen et al., 1998) or dChip (Li and Wong, 2001).

Clustering aims for the identification of regulated biological processes through the evaluation of co-regulated genes, based on the assumption that a cellular response is mainly reflected in transcriptional levels. Unfortunately, assignment to a certain cluster, genetic co-regulation and biological function do not necessarily coincide. The reasons for this are manifold and lie mainly in the biological response. First, cellular processes are affected by both up- and down-regulation. Therefore, genes involved in a common pathway can end up in completely different clusters. Second, many processes are only regulated by post-translational modification. Notably, the cluster algorithms themselves are very sensitive to statistical variation. The clustering becomes more robust as the variation becomes minimized and therefore a certain minimal number of experiments and replicates are necessary to build meaningful clusters. The potential to assign genes to reasonable groups of biological function was shown for various cluster algorithms, and it becomes clear that these approaches perform differently on distinct datasets (Gibbons and Roth, 2002).

FRAMEWORK
The Gene Ontology (GO) database holds functional gene annotation in a hierarchical structure that reflects the relationship between the biological terms and associated genes (The Gene Ontology Consortium, 2000). The three ontologies are biological process, molecular function and cellular component. Within an ontology, the terms are organized in parent–child relationships, so that the GO database can be pictured as a tree structure where each node can give rise to various numbers of experiments and terms in turn give rise to other terms or to annotated genes as their leaves, respectively. A term and its leaves represent a ‘functional cluster’, but recursively, also its parent term and all child terms of this parent form a cluster.

We and others (e.g. Cheng et al., 2002; Hanisch et al., 2002) believe that these functional groups can be used as a framework for numerical clustering in order to create more meaningful clusters. Purely mathematical approaches are underway to incorporate the GO structure as an additional distance measure. Mathematically speaking, the GO is a directed acyclic graph, and uneven granularity and

*To whom correspondence should be addressed.
CONCLUSION

We have developed the GO-Cluster application that incorporates the structure of the GO database as a framework for selecting subsets of gene expression data that is then subject to cluster analysis. We found the program very useful for rapid visual identification of regulated biological processes. The cellular response can be evaluated on virtually every level of the GO tree, thus allowing quick browsing through all of the ontologies. We believe that our approach is not only good for evaluating the transcriptional regulation of biological functions of interest, but may also help to identify common regulation of processes that otherwise may have escaped identification. GO-Cluster was successfully applied to Affymetrix *Drosophila* GeneChip data in order to assess FGF/Branchless signaling in the fruit fly (B. Adryan and R. Schuh, manuscript in preparation).

ACKNOWLEDGEMENTS

B.A. is supported by stipends from Fonds der Chemischen Industrie and Lutz Adolf-Stiftung für Hochbegabte.

REFERENCES


INTERPRETATION

The visual output obtained from GO-Cluster requires some interpretation. In contrast to the traditional cluster analysis where users observe certain clusters for genes that play roles in a common pathway, in GO-Cluster one has to decide whether the clustering result for a certain GO term is subjectively acceptable or not (for a commented example, see the website). The greatest advantage of this approach is that in complex regulatory networks one can see antagonistically regulated genes as two child branches originating from one common root whereas in other clustering approaches they would end up in completely distinct branches of the cluster tree. The former cluster may appear out of the ordinary for researchers used to pure statistical clustering, but in the biological sense it is true.

biological relevance of certain terms can be considered when evaluating the shortest distance between two distinct GO terms (Cheng et al., 2002). However, this approach is not transparent and one might wonder why two differently regulated genes end up in one cluster.

Our approach does not qualify for the mathematical solution of the problem but aims for the visual identification of differentially regulated functional clusters. The 32 bit Windows program GO-Cluster is designed for the biological researcher, who wants convenient access to GO and wants to visualize gene regulation on various levels of the GO tree. In contrast to other known visualization tools, i.e. MAPPFinder (Doniger et al., 2003) or GoMiner (Zeeberg et al., 2003), our software does not judge statistically the regulation of a GO term, but performs hierarchical average-distance clustering by applying Pearson’s correlation coefficient to the genes that are allocated to the corresponding term. The advantage of clustering is that no ‘rules’ have to be predefined and essentially all of the available datasets are informative. Also, the color-coded display of clustered gene expression data is more revealing and easier to assess for the biologist. The program is easy to use, since the GO tree can be accessed in a file manager-like manner and tabular expression data can quickly be loaded into it. In GO-Cluster virtually every node of the GO can be selected for cluster analysis, the corresponding tree of clusters can be visualized in real time and simultaneously displayed as a left-to-right tree structure (see the website for an exemplary screenshot). Here, the expression level is not only displayed color-coded in the context of the gene vector, but also in another panel in the context of its expression level in comparison to the other genes in the respective GO term for one experimental condition.