Poster L-29

Identifying functional gene sets from hierarchically clustered expression data: map of abiotic stress regulated genes in Arabidopsis thaliana.

Authors:
Matti Kankainen (*University of Helsinki, Institute of Biotechnology*)
Günter Brader (*University of Helsinki, Department of Biological and Environmental Sciences, Division of Genetics*)
Petri Törönen (*University of Helsinki, Institute of Biotechnology*)
E. Tapio Palva (*University of Helsinki, Department of Biological and Environmental Sciences, Division of Genetics*)
Liisa Holm (*University of Helsinki, Institute of Biotechnology*)

Short Abstract: MultiGO is a tool to automatically identify biologically relevant gene sets from hierarchically clustered gene trees (http://ekhidna.biocenter.helsinki.fi/poxo/multigo). Since the entire tree is analyzed, all gene clusters sharing a common biological function, as defined by GO, are reported. The tool also identifies a cluster set representing the experimental key functions.

Long Abstract:

Typical expression data analysis encompasses the pre-processing of the data and the use of statistical tests to detect genes with altered expression (1). Genes with similar expression profiles can then be clustered together, to present the data in a more comprehensible form, using different clustering algorithms (1). Hierarchical algorithms classify genes into nested clusters and the result resembles phylogenetic classification (1). In this approach, e.g. in agglomerative algorithms, single expression profiles are joined to form groups, which are further joined until the process has been carried to completion, forming a single hierarchical expression tree (1-2). Partitional data clustering algorithms, e.g. k-means clustering, separate distinctly expressed genes into a predetermined number of clusters (1, 3). Despite the different clustering approaches, their final aim is to create gene sets that could be used as indicators of the status of the cellular functions (1-2). However this is not always achieved. The generated clusters can be biologically irrelevant and can contain genes involved in incoherent functions. For example, improper parameter settings may result in spurious clusters. Thus, it is essential to evaluate the clusters.

A reasonable approach to evaluate co-expressed gene clusters is to explore the gene functions of the genes in the cluster and to calculate the statistical significance of the found functions (4-9). Gene functions can be explored using text mining tools for biological and medical literature (7-9). Another approach to explore gene functions is to use controlled vocabularies, such as Gene Ontology (GO) (10). The difference between these two is that in the first, numerous alternative wordings may describe the same function, whereas in the latter this is eliminated. After associating gene functions with clusters, the statistical significance of these associated functions can be calculated, e.g. using basic statistical methods that compare the frequency of the term within a set of query genes against the

frequency of the term in the transcriptome of the organism. There are numerous tools, which can perform both the exploration of the gene functions and the statistical calculus with respect to GO-terms (11-13). These tools enable the rapid creation of overviews of the stimulated functions and therefore they have facilitated the interpretation of expression data. However, since existing tools have mainly been designed to analyze single gene lists, they unnecessarily complicate the analysis of large expression data sets. For example, tools designed to analyze single gene lists ignore the fact that the analyzed genes can as well be a part of a data set containing several other clusters. The systematic evaluation of all clusters can also be labour-intensive using these tools, which complicates the listing of all significant GO-term of the data set. For example, finding the affected key functions from expression data set can be difficult, although there might be a preliminary hypothesis about them. There are some tools that estimate the optimal set of clusters using the gene function information (5-6). Although these tools enumerate all clusters in the analysis, they still suffer from the same deficiency and do not report all significant GO-terms to the user.

To address the above problems, we developed a web-enabled tool, MultiGO. Our tool analyzes every gene cluster, assigns a representative function to each of them and then reports to the user all clusters that have an enrichment of genes involved in coherent functions, with respect to the GO classification. Since every cluster is analyzed, the tool identifies all significant GO-terms of the entire data set. The systematic analysis can also be used to discover aspects and functions that would have been otherwise missed. For example, MultiGO can highlight similar functions that are located in distinct branches of the hierarchically clustered expression tree, e.g. two clusters that are expressed differently but are involved in the same function. The tool also estimates the best vertical cutting point of the hierarchically clustered expression trees using Fisher's combined p-value test (14). This cutting point yields an illustrative set of clusters, which corresponds to the key stimulated functions of the experiment. Selecting fewer clusters of larger size would decrease the functional coherence of the clusters whereas selecting a larger number of smaller size clusters would separate functionally related genes from each others.

We illustrate the utility of the tool by analyzing a gene expression data set from Arabidopsis thaliana under abiotic stress conditions.

REFERENCES

1. Quackenbush,J. (2001) Nature Rev. Genet., 2, 418-427.
2. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Proc. Natl. Acad. Sci. U S A., 95, 14863-14868.
3. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Nature Genet., 22, 281-285.
4. Gibbons,F.D. and Roth,F.P. (2002) Genome Res. 10, 1574-1581.
5. Törönen,P. (2004) BMC Bioinformatics, 5.
6. Raychaudhuri,S., Chang,J.T., Imam,F. and Altman,R.B. (2003)
Nucleic Acids Res. 31, 4553-4560.
7. Raychaudhuri,S., Schutze,H. and Altman,R.B. (2002) Genome Res., 12, 1582-1590.
8. Masys,D.R, Welsh,J.B., Lynn,F.J., Gribskov,M., Klacansky,I. and Corbeil,J. (2001) Bioinformatics, 17, 319-326.
9. Jenssen,T.K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) Nature Genet., 28, 21-28.
10 Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P.,

Dolinski,K., Dwight,S.S., Eppig,J.T., et al. (2000) Nature Genet., 25, 25-29.
11. Beissbarth,T. and Speed,T.P.(2004) Bioinformatics, 20, 1464-1465.
12. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) Bioinformatics, 20, 578-580.
13. Boyle,E.I., Weng,S., Gollub,J., Jin,H., Botstein,D., Cherry,J.M. and Sherlock,G. (2004) Bioinformatics, 20, 3710-3715.
14. Fisher,R.A (1932) Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh