BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 17 2005, pages 3575–3577
doi:10.1093/bioinformatics/bti574

Gene expression

goCluster integrates statistical analysis and functional interpretation of microarray expression data

Gunnar Wrobel, Frédéric Chalmel and Michael Primig*

Biozentrum and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, CH-4056 Basel, Switzerland

Received on February 17, 2005; revised on May 9, 2005; accepted on July 5, 2005

Advance Access publication July 14, 2005

ABSTRACT

Motivation: Several tools that facilitate the interpretation of transcriptional profiles using gene annotation data are available but most of them combine a particular statistical analysis strategy with functional information. goCluster extends this concept by providing a modular framework that facilitates integration of statistical and functional microarray data analysis with data interpretation.

Results: goCluster enables scientists to employ annotation information, clustering algorithms and visualization tools in their array data analysis and interpretation strategy. The package provides four clustering algorithms and GeneOntology terms as prototype annotation data. The functional analysis is based on the hypergeometric distribution whereby the Bonferroni correction or the false discovery rate can be used to correct for multiple testing. The approach implemented in goCluster was successfully applied to interpret the results of complex mammalian and yeast expression data obtained with high density oligonucleotide microarrays (GeneChips).

Availability: goCluster is available via the BioConductor portal at www.bioconductor.org. The software package, detailed documentation, user- and developer guides as well as other background information are also accessible via a web portal at http://www.bioz.unibas.ch/gocluster.

Contact: michael.primig@unibas.ch

INTRODUCTION

Microarray-based expression profiling helps identify genes involved in processes such as growth, development, the response to external stimuli (Futcher, 2000; Schlecht and Primig, 2003; Stoughton, 2004; Wrobel and Primig, 2005) and drug target discovery (Sausville and Holbeck, 2004). This approach is based on the assumption that functionally related genes are transcriptionally coregulated during the biological process for which they are important. Such gene groups, until recently, had to be identified by manually searching for recurring functional references among the gene lists obtained by statistical analysis. Several tools are available that employ the hypergeometric distribution to determine if loci with similar GeneOntology (GO) annotation (Harris *et al*., 2004) are significantly enriched in a given group of genes (Al-Shahrour *et al*., 2004; Hosack *et al*., 2003; Martin *et al*., 2004). Other packages use GO data to generate a global overview of an expression dataset instead of relying on single gene lists (Buehler *et al*., 2004). These tools are important because they help correlate functional annotation information with the outcome of expression profiling experiments. However, each of the currently available solutions has certain limitations. For example, FatiGO and EASE allow analysis of only one single gene list at a time. Therefore, the ability of the algorithm to identify functional groups in a given list is largely dependent upon the filtering strategy used. CRASSS, a package that generates a global overview of the distribution of functions within array data, employs only one clustering method and relies on commercial microarray data analysis software. Moreover, these tools are often provided as web services and, therefore, they cannot be integrated into an automated analysis procedure.

In this study, we describe goCluster, a solution that implements a statistical analysis procedure yielding gene lists that are subsequently searched for non-random enrichment of related GO annotation terms from all three categories (biological process, molecular function and cellular component) (Harris *et al*., 2004). This innovative and flexible tool enables scientists to employ custom tailored approaches to microarray data analysis and interpretation, thereby accelerating and facilitating the process of hypothesis building.

APPROACH

The software goCluster is written in the statistical programming language R (http://www.r-project.org) as a part of the BioConductor project that provides freely available open source software solutions (Carey *et al*., 2005; Gentleman *et al*., 2004). Note that running the current release of the package requires R installed in a Linux or Windows environment (see System Requirements section and online help files for more details). goCluster is based on an object-oriented framework designed to conduct searches for functionally related genes in microarray data. It is possible to import the complete dataset for all genes represented on a given microarray or to employ filtering steps to remove loci for which no reliable data are available (e.g. those that are flagged with an 'absence' or 'marginal' call in the case of GeneChips). The approach is based on two assumptions. First, particular expression patterns displayed by groups of genes are related to the phenomenon studied (cell-cycle progression, meiotic development, stress response). Second, a non-randomly enriched GO term associated with an expression group (cluster) pinpoints a

*To whom correspondence should be addressed.

highly relevant biological process, molecular function or cellular component associated with the phenomenon studied.

The program combines annotation data with statistical and functional analysis algorithms without relying on a particular clustering algorithm, annotation type or statistical test. The software is organized into six functionally distinct modules. The *Data* module contains the expression signals to be analysed and links it with the Annotation module that can hold a variety of different types of information including chromosomal localization or functional annotation data. The *Filtration and Clustering* module can contain gene filtering and clustering algorithms used to select and group genes. These groups are processed by the *Functional Analysis* module that calculates a statistical value (e.g. *P*-value) for every annotation term found within a group. These values are corrected for multiple testing by the algorithm provided in the *Significance Analysis* module that also selects interesting annotation terms. Finally, a graphical display of the results by a so-called heat map (a false colour bar diagram of expression signal intensities) is generated in the *Visualization* module.

The current version of goCluster provides efficient means to analyse a dataset by hierarchical clustering, CLARA, partitioning around medoids (PAM) or *k*-means clustering. The resulting groups are subsequently searched for enrichment of GO annotation terms using the hypergeometric distribution. Multiple testing errors are accounted for using the Bonferroni correction or by employing the false discovery rate (FDR) based on randomized clustering results. goCluster provides several methods to visualize results obtained using GO information. For example, expression values of clustered genes can be displayed as a heat map to visualize expression signal intensities (Wrobel and Primig, 2005). It is also possible to produce a graphical display of expression data for individual genes (see Hochwagen *et al*., Figure 6, panels D and E). The information content of hierarchical clustering is increased by displaying the annotation term with the smallest *P*-value right next to the genes in a particular cluster. This makes it straightforward to verify the functional relevance of each cluster.

IMPLEMENTATION

Each of the six sections mentioned above is represented by its own class within goCluster. A module meant to provide one of the sections with new functionality needs to be derived from the corresponding class (e.g. a module that contributes a new type of clustering algorithm needs to extend the class of the *Filtration and Clustering* module). All six basic classes are derived from the clusterModule class that provides generic features common to all modules of goCluster. An important advantage of this structure is a common front-end that can be used to configure all modules. It also reduces the effort necessary to extend the framework since it provides basic functionality for the derived classes. A new clustering algorithm can thus be added by embedding the corresponding clustering function into a short wrapper class that extends the *Filtration and Clustering* module. The new algorithm will be immediately available within goCluster and can thus be combined with the statistical tests provided by the framework.

It is also possible to add new ontologies to the framework by introducing a two-column table that associates genes with annotation terms. The annotation data need a simple wrapper class that loads the table and stores it in the clusterAnnotation class. In this case, the input table should only list the genes present on the microarray to be analysed. More complicated annotation classes are needed in order to reduce generic lists to the annotation terms relevant for a specific microarray or for direct web access to annotation information available online.

SYSTEM REQUIREMENTS

The hardware requirements for running goCluster depend primarily on the size of the datasets to be processed and on the type of clustering algorithms to be employed. Most analysis procedures can be run on a standard desktop computer. Note that it might take several hours to compute the results for large batches of more than 100 hybridization experiments. Hierarchical clustering is more expensive in terms of computing power because it requires an exponentially increasing amount of memory with an increasing number of features on the microarray. In these cases it is recommended to filter for differentially expressed transcripts prior to the current goCluster analysis. The current goCluster Release 1.0.3. features a simple Tk graphical user interface (GUI; using the R package *tkWidgets*) functional under all operating systems capable of running R.

APPLICATION

We have employed goCluster to help interpret previously published expression profiles of rat transcripts that are upregulated in male germ cells (Schlecht *et al*., 2004). Indeed, the tool identified GO terms that were particularly relevant for the process of male meiosis and spermatogenesis on the basis of germ-cell enriched transcription (Wrobel and Primig, 2005). Furthermore, exploiting the capabilities of goCluster we analysed a complex experiment using the budding yeast *Saccharomyces cerevisiae*. The approach involved several parameters including vegetative growth in rich medium, meiotic development (sporulation), cold-shock treatment and cell culturing in the presence of a toxic substance that inhibits growth and spore formation. Again, the tool correctly correlated GO annotation and expression patterns of genes involved in detoxification as well as temperature stress. Moreover, it helped identify loci involved in meiosis and spore development that were negatively regulated by low temperature and treatment with a toxic compound that disrupts the process (Hochwagen *et al*., 2005); see also the web portal at http://www.bioz.unibas.ch/primig/benomyl/). Data interpretation was supported by combining GO annotation with hierarchical clustering and identification of enriched terms using the hypergeometric distribution. The resulting *P*-values were corrected using the FDR. Note that the visualization method is equivalent to the overview produced by CRASSS. It is, however, important to note that our approach just shows a subset of the results since only the most significant and non-overlapping GO terms are displayed.

DISCUSSION

goCluster demonstrates how different approaches to functional data analysis can be integrated into one tool without the need to develop a novel application for each new dataset or annotation type. This is especially important considering that GO is not the only source of functional annotation data that can (or indeed should) be used to interpret the results of profiling experiments. goCluster functionalities may include protein–protein interaction data or information on protein localization (Ghaemmaghami *et al*., 2003; Huh *et al*., 2003).

Since the program is developed in R it can exploit all algorithms available via BioConductor that filter or cluster differentially expressed genes. Although the current version of goCluster focuses on clustering to generate a global overview of the dataset, it is straightforward to include other algorithms (e.g. ANOVA models, self-organizing maps) and to combine them with functional analysis.

The first release of goCluster is restricted to countable types of annotation (e.g. GO) amenable to analysis by employing the hypergeometric distribution. It would be interesting to include algorithms that explore more complex patterns than enrichment of terms, e.g. the distances of terms in the GO hierarchy or physical distances of genes according to their chromosomal localization. Another potentially useful addition would be an approach that exploits information encoded in tree structures, such as those obtained from a hierarchical clustering approach (Barriot *et al*., 2004). Thereby, the hierarchical clustering result of one dataset can be used as an annotation for a second experiment in order to rapidly identify common pathways between both datasets. Currently, goCluster does not provide a file loader covering formats, such as XML, RDF, OWL or CLIPS. We intend to provide such generic import modules in the next goCluster release.

CONCLUSION

The goCluster software has been successfully employed for the interpretation of complex microarray expression profiling data from rat and budding yeast (Hochwagen *et al*., 2005; Wrobel and Primig, 2005). The approach implemented in the package lends itself to a very broad range of experimental conditions used to study the rapidly growing list of organisms for which a critical amount of annotation and functional genomics data is available. Therefore, we suggest that tool described here facilitates and accelerates the process of microarray data based hypothesis building. To carry out a goCluster analysis in R, users simply have to set up configuration parameters and execute the analysis function. To further increase the versatility of our approach, we intend to provide a web-based goCluster service in the not too distant future. The package is freely available under the GNU public license and can be downloaded from the BioConductor website at http://www.bioconductor.org (Carey *et al*., 2005) and the goCluster portal at http://www.bioz.unibas.ch/gocluster. This website also provides detailed information on how to install and run the software.

ACKNOWLEDGEMENTS

We thank L. Hermida and C. Niederhauser-Wiederkehr for the critical reading of the manuscript. F.C. and G.W. are supported by the Biozentrum. Funding to pay the Open Access publication charges for this article was provided by the Swiss Institute of Bioinformatics.

Conflict of Interest: none declared.

REFERENCES

Al-Shahrour,F. *et al*. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.

Barriot,R. *et al*. (2004) New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. *Nucleic Acids Res.*, **32**, 3581–3589.

Buehler,E.C. *et al*. (2004) The CRASSS plug-in for integrating annotation data with hierarchical clustering results. *Bioinformatics*, **20**, 3266–3269.

Carey,V.J. *et al*. (2005) Network structures and algorithms in Bioconductor. *Bioinformatics*, **21**, 135–136.

Futcher,B. (2000) Microarrays and cell cycle transcription in yeast. *Curr. Opin. Cell. Biol.*, **12**, 710–715.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Ghaemmaghami,S. *et al*. (2003) GIobal analysis of protein expression in yeast. *Nature*, **425**, 737–741.

Harris,M.A. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.

Hochwagen,A. *et al*. (2005) A novel response to microtubule perturbation in meiosis. *Mol. Cell. Biol.*, **25**, 4767–4781.

Hosack,D.A. *et al*. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.

Huh,W.K. *et al*. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.

Martin,D. *et al*. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol.*, **5**, R101.

Sausville,E.A. and Holbeck,S.L. (2004) Transcription profiling of gene expression in drug discovery and development: the NCI experience. *Eur. J. Cancer*, **40**, 2544–2549.

Schlecht,U. *et al*. (2004) Expression profiling of mammalian male meiosis and gametogenesis identifies novel candidate genes for roles in the regulation of fertility. *Mol. Biol. Cell*, **15**, 1031–1043.

Schlecht,U. and Primig,M. (2003) Mining meiosis and gametogenesis with DNA microarrays. *Reproduction*, **125**, 447–456.

Stoughton,R.B. (2004) Applications of DNA Microarrays in Biology. *Annu. Rev. Biochem.*

Wrobel,G. and Primig,M. (2005) Mammalian male germ cells are fertile ground for expression profiling of sexual reproduction. *Reproduction*, **129**, 1–7.