Gene expression

Analyzing microarray data using CLANS

Tancred Frickey and Georg Weiller*

ARC Centre of Excellence for Interactive Legume Research and Bioinformatics Laboratory, Genomic Interactions Group, Research School of Biological Sciences, Australian National University, GPO Box 475, Canberra, ACT 2601, Australia

Received on October 23, 2006; revised on February 7, 2007; accepted on February 27, 2007

Advance Access publication March 7, 2007

Associate Editor: Thomas Lengauer

ABSTRACT
Summary: Analysis of microarray experiments is complicated by the huge amount of data involved. Searching for groups of co-expressed genes is akin to searching for protein families in a database as, in both cases, small subsets of genes with similar features are to be found within vast quantities of data. CLANS was originally developed to find protein families in large sets of amino acid sequences where the amount of data involved made phylogenetic approaches overly cumbersome. We present a number of improvements that greatly extend the previous version of CLANS and show its application to microarray data as well as its ability of incorporating additional information to facilitate interactive analysis.

Availability: The program is available for download from: http://bioinfoserver.rsbs.anu.edu.au/downloads/clans/

Contact: Georg.Weiller@anu.edu.au

Supplementary information: http://bioinfoserver.rsbs.anu.edu.au/programs/clans

1 INTRODUCTION

Gaining useful insights from microarray experiments is frequently hampered by the amount of data the experiments generate as well as the difficulty of relating changes in gene expression to observable effects in a cell or organism. Transforming expression data into useful biological hypotheses requires both a reduction in the amount of data to analyze and taking into account additional information that provides the background on which to base such a hypothesis.

The method of choice to reduce the amount of data is often to disregarding all data except for sets of co-expressing genes or genes behaving according to a specific expectation. The hypotheses derived from the experiments generally arise from a synthesis between expression data and a combination of the evolutionary history of genes, annotation, known interactions, metabolic pathways and cellular localization.

The problem of finding groups of genes co-expressed across experimental conditions is comparable to the problem of finding groups of similar proteins in a large sequence database. In both cases, finding groups is a matter of using similarities in the features of the genes or sequences to group them into sets, maximizing the amount of information a set provides and minimizing the amount of conflicting information the grouping generates. In one case the features of interest are the expression values, in the other the nucleic or amino acid sequences.

CLANS (Frickey and Lupas, 2004) was developed to facilitate detection of protein families within large and diverse sequence datasets. Due to the similarity of the tasks, we decided to extend CLANS to microarray data analysis. The examples below present some of the new features of the program and how these can be used to facilitate analysis of microarray experiments.

A tutorial and further information on CLANS are available as part of the Supplementary Material.

2 IMPLEMENTATION

CLANS provides an interactive analysis environment using self-organizing maps to visualize datasets of pairwise similarities. In the case of microarray data, these similarities are based on how the expression of one gene correlates with that of another. Linear correlation was used in the provided example, but many other correlation measures are available. Positive correlation values provide ‘attractive’ forces between the genes represented in the graph. Negative values can either be disregarded or used to provide ‘attractive’ or ‘repulsive’ forces, depending on what the analyst requires.

Annotation and pathway data, such as MAPMAN-bins (Thimm et al., 2004), GeneBins (Goffard and Weiller, 2006) or Gene Ontology (Ashburner et al., 2000), can be integrated in the map to facilitate analysis. Graphs depicting hypothetical expression levels for the various experiments can be drawn to find genes showing similar behavior. Combined with the ability to exclude any number of experiments, this allows recovery of groups of genes that were defined based on subsets of the currently available data.

Finally, three automated cluster-detection methods and extensive selection and coloring features were added to facilitate visualization and analysis.

3 APPLICATION

In a first step, microarray expression data is converted to a CLANS file using the program ‘expr2clans.jar’, available as part of the package. CLANS derives 2D or 3D maps of the
pairwise similarities to allow interactive detection of groups of co-expressed genes (Fig. 1). Varying the minimum correlation cutoff can reveal the sub- or super-structure of any cluster by causing large clusters to dissociate into smaller clusters of higher correlation or vice versa. The resulting map is used to focus on specific groups or clusters of genes and various colors, shapes and sizes can be used to track these throughout the analysis. Graphs showing the expression responses of the genes in any group can be used to visualize the differences between groups.

The program provides an environment in which gene expression can be tightly coupled with annotation data. The bidirectional lookup between expression and annotation data, using GeneBins files, for example, provides an easy way to see which of the KEGG (Kanehisa et al., 2004) pathways contain co-expressing genes as well as which of the groups of co-expressing genes share a common pathway, functional annotation or cellular localization.

ACKNOWLEDGEMENTS

This research was funded by an Australian Research Council Centre of Excellence grant. Funding to pay for the Open Access publication charges was provided by the same grant. We would like to thank Christine Beveridge and Julia Cremer for extensive testing and feedback on how to make the program more useful and user friendly.

Conflict of Interest: none declared.

REFERENCES


