ClutrFree: cluster tree visualization and interpretation

Ghislain Bidaut\textsuperscript{1,2} and Michael F. Ochs\textsuperscript{1,*}

\textsuperscript{1}Division of Population Science, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, USA and \textsuperscript{2}Structural and Genomic Information Laboratory, CNRS-UPR 2589, 13402 Marseille, France

Received on September 29, 2003; revised on March 24, 2004; accepted on April 28, 2004
Advance Access publication May 14, 2004

ABSTRACT

Summary: ClutrFree facilitates the visualization and interpretation of clusters or patterns computed from microarray data through a graphical user interface that displays patterns, membership information of the genes and annotation statistics simultaneously. ClutrFree creates a tree linking the patterns based on similarity, permitting the navigation among patterns identified by different algorithms or by the same algorithm with different parameters, and aids the inferring of conclusions from a microarray experiment.

Availability: The ClutrFree Java source code and compiled bytecode are available as a package under the GNU General Public License at http://bioinformatics.fccc.edu

Contact: Ghislain.Bidaut@fccc.edu; Michael.Ochs@fccc.edu

Supplementary information: Extended documentation on the software usage, as well as supplemental figures and an appropriate test dataset are provided on the website.

Microarrays have become a standard tool for exploration of biological processes in numerous areas, including cancer research (Ochs and Godwin, 2003). Data analysis often includes use of clustering algorithms or machine learning techniques, however, no de facto standards have emerged. Therefore, for many microarray experiments, it is useful to apply multiple algorithms and explore the consistency of the results. For example, Yeung \textit{et al.} (2001) assessed the differences between clustering procedures by deriving figures of merit for clusters, while Datta and Datta (2003) applied multiple analysis methods and measured the consistency by gene membership lists for fixed numbers of clusters. The approach here, ClutrFree, aims primarily to visualize the similarities (measured as correlations) between the results of multiple pattern recognition algorithms regardless of the number of patterns, including cases where the same algorithm is used with different parameters producing different numbers of patterns. Since there are usually no strong a priori reasons for choosing parameter values, this permits the user to explore the sensitivity of the analysis to such choices. Although we focus on pattern recognition in the gene space here, patterns in condition space could be visualized as well, with the input files comprising clusters or patterns on conditions. The similarities between patterns comprise both measures of robustness (how often do genes or conditions appear in related patterns) and correlations (how similar are pattern shapes). When exploring gene space, the gene lists provide a possible way to enhance interpretation through the use of gene ontology, so tools for exploring ontological assignments in the patterns are included. Although genes perform multiple functions and hence belong to multiple ontological branches, by including many genes in a group, ontological enhancement can still provide an indication of function in some cases.

ClutrFree provides a graphical user interface (GUI) for visualization and a simple method for constructing meaningful trees across multiple results from pattern recognition. The tree (a directed graph) contains levels, such that each level summarizes the result of a single application of a pattern recognition algorithm, with nodes representing individual patterns or clusters (such as three nodes from a single run of a self-organizing map). Branches that are most stable (e.g. arrow A in Fig. 1) regarding the variation of certain parameters can be easily visualized. In some cases, we have seen from gene ontology that clusters that branch (e.g. arrow B in Fig. 1) identify sets of genes constituting subprocesses of a single biological process, such as splitting of routine housekeeping processes into carbohydrate metabolism and biosynthesis processes (Ochs \textit{et al.}, 2004).

The ClutrFree GUI contains three windows: the main window displaying graphs of the patterns linked by similarity measure and permitting navigation in the tree, the tree window and the gene table (Fig. 1). The tree of patterns is created by an iterative method beginning with the results comprising the minimum number of clusters and following these steps:

1. Each analysis is represented as a tree level, beginning with the experiment having the fewest patterns...
Fig. 1. The major ClutrFree GUI components: clusters are displayed in the main window (a) in the form of a stem plot. The bottom arrows (b) permit the navigation within the data tree. The upper slider (c) adjusts the value/uncertainty ratio used to binarize the data. Genes and their membership to each cluster (d) are simultaneously displayed in the gene window along with gene ontology and annotations (e), and persistence measurement (f). Genes can be filtered out on the basis of their persistence using the upper group of checkboxes and sliders (g). When hierarchical ontologies are used, the system attempts to use a color scheme to aid the display readability. The gene ontology counting list (h) can also be filtered using the two bottom sliders (i). The first one defining the highlighting level, and the second one adjusting the cutoff used to filter the list by removing ontologies with a low number of instances.


(n patterns, the \( n \) level). For analyses having the same number of patterns, the order is not important.

(2) The \( n + 1 \) level of the tree is compared with the \( n \) level by calculating the Pearson correlation for all nodes at \( n \) with all nodes at \( n + 1 \), with each node representing a pattern. The connections are made from highest to lowest correlation, with nodes removed from the process as they are connected. Each remaining node at level \( n + 1 \) is connected to the node at level \( n \) that gives the highest Pearson correlation coefficient. This creates a tree where all nodes have at least one connection between levels. The correlation is calculated on the pattern shapes input into the algorithm, so that only one shape is allowed for each pattern at this time, requiring that the same shape is used from each pattern in all similarity measures (e.g. average linkage, centroids, etc.). The number of nodes must be a monotonically increasing function with level (although two neighboring levels may have the same number of nodes), so the data must be ordered prior to the use of ClutrFree.

(3) Step 2 is repeated for each additional level, until the full tree has been constructed.

Patterns themselves are displayed in the main window as shown in Figure 1a, and buttons permit the user to switch rapidly from one pattern to the next or to navigate through the tree to infer the change in a pattern as parameters increase or decrease. Genes related to each cluster (or node) are displayed in a separate window as a data matrix, shown in Figure 1d, along with their annotation and ontology, each row being a gene and each column quantifying the membership of each gene to the corresponding patterns, which can be a binary
value, as in hierarchical clustering, or a continuous value, as in Bayesian Decomposition (Moloshok et al., 2002). If the clustering method does not produce binary cluster membership, ClutrFree binarizes each data point based on its corresponding uncertainty: if the data value is greater, that is, \( m \sigma > 0 \) (\( m \) being a user selectable parameter), the data point is assigned as 1 (member), otherwise it is assigned as 0 (non-member). From the binary values, the system has the ability to quantify gene stability by measuring the persistence, defined as the number of consecutive times that a gene remains a member of a given branch as we move down the tree. ClutrFree also permits filtering of genes that do not meet a defined level of persistence. To infer biological functions for a cluster, enhancement of annotation categories (such as gene ontology) of the member genes is calculated and displayed on a separate table, as shown in Figure 1h. Each single annotation description for genes belonging to a cluster is counted, normalized by the number of genes present in the cluster, and compared to the total number of ontology matches in genes in the experiment to see enhancements of biological functions (Ochs et al., 2004). Users have the possibility to filter out ontology classes with a low number of genes to avoid accidental enhancement in a small cluster. ClutrFree is written in Java and can be used on all platforms that support Java 1.4. It currently permits data importation from tab-delimited files and Bayesian Decomposition output files (Moloshok et al., 2002). The gene annotation can be encoded under the form of gene names, accession numbers or informal string descriptions. The use of the standardized gene ontology format defined by the Gene Ontology Consortium (Ashburner et al., 2000) has been implemented, permitting ClutrFree to import annotations retrieved by tools, such as the Automated Sequence Annotation Pipeline (Kossenkov et al., 2003). When genes are annotated with hierarchical ontologies (such as the one from the Sanger Institute), a color-highlighting scheme is provided for enhanced visualization. Clusters and trees are exportable as image files (TIFF, PNG, JPG or SVG) and the tree is also exportable in the Dot graph language (Graphviz package, AT&T Laboratories). Tables can be exported in the tab-delimited format after sorting or filtering for further processing. Future enhancement will include modularization of import routines, facilitating the extension of the file formats currently supported by ClutrFree, as well as incorporating visualization of biological information on the tree itself.

ACKNOWLEDGEMENTS

We thank the National Institutes of Health, National Cancer Institute (CCCG CA06927 to R. Young), Pennsylvania Department of Health (grant to M.F.O.) and the Pew Foundation for support.

REFERENCES


