BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 14 2007, pages 1843–1845
doi:10.1093/bioinformatics/btm238

Gene expression

CRCView: a web server for analyzing and visualizing microarray gene expression data using model-based clustering

Zuoshuang Xiang[1], Zhaohui S. Qin[2,3] and Yongqun He[1,3,4,*]

[1]Unit for Laboratory Animal Medicine, [2]Department of Biostatistics, Center for Statistical Genetics, [3]Bioinformatics Program, and [4]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI

Received on March 15, 2007; revised on April 22, 2007; accepted on April 26, 2007

Advance Access publication May 7, 2007

Associate Editor: Olga Troyanskaya

ABSTRACT

Summary: CRCView is a user-friendly point-and-click web server for analyzing and visualizing microarray gene expression data using a Dirichlet process mixture model-based clustering algorithm. CRCView is designed to clustering genes based on their expression profiles. It allows flexible input data format, rich graphical illustration as well as integrated GO term based annotation/interpretation of clustering results.

Availability: http://helab.bioinformatics.med.umich.edu/crcview/

Contact: yongqunh@umich.edu

1 INTRODUCTION

Microarray gene expression technology is a powerful tool and has been widely used in modern biomedical research to monitor gene expression levels on a global scale. Although substantial noise is contained in microarray data, functional related genes tend to show detectable correlated expression patterns across multiple experiments. This property can be exploited by unsupervised clustering analysis to reveal underlying functional relationships among genes. Many different clustering techniques (e.g. hierarchical clustering, k-means, self-organizing map) have been developed (Gollub and Sherlock, 2006). These techniques have been widely used in downloadable software programs such as GenePattern (Reich *et al.*, 2006) and MeV (http://www.tm4.org/mev.html), and web-based tools such as GEPAS (http://www.gepas.org). These clustering approaches are easy to use and visually appealing. However, since they are distance-based, the results are sensitive to noise and distance definitions. Recently, state-of-the-art model-based clustering approaches have been proposed to analyze microarray data, and favorable performances have been reported (Medvedovic *et al.*, 2004; Qin, 2006; Yeung *et al.*, 2001). More importantly, model-based clustering approaches are built under a sound probability framework, thus offering crucial benefits such as incorporation of uncertainties and formal statistical inference. Model-based clustering software programs including MCLUST (Fraley and Raftery, 1999) and GIMM (Medvedovic *et al.*, 2004) are primarily command line-based and all require software installation. They also do not directly provide graphical outputs whereas visualization is an important part of clustering analysis. To fill the vacancy, we developed CRCView, an easy to use point-and-click web interface for users to run model-based clustering analysis without software download and installation.

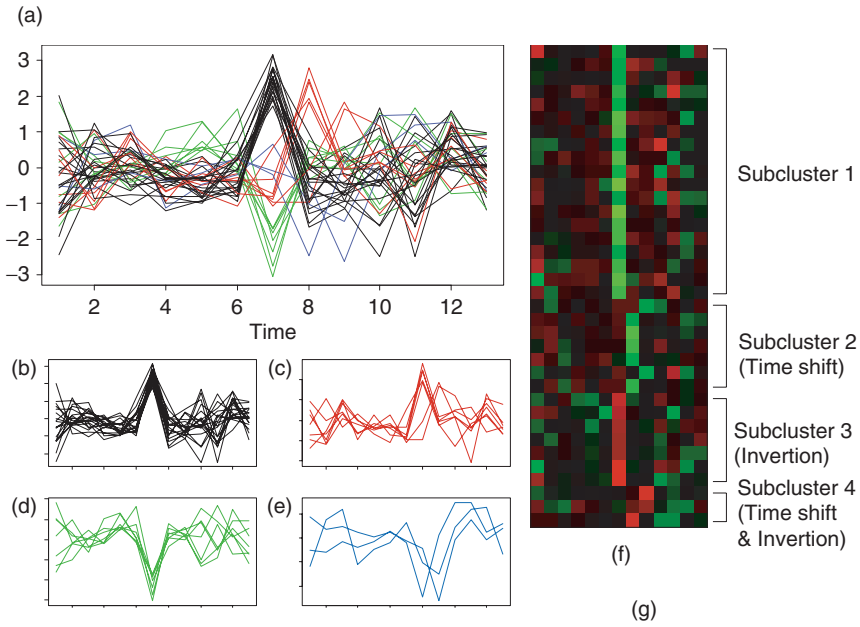
*To whom correspondence should be addressed.

2 FEATURES AND USAGE

CRCView is based on our recently developed Dirichlet process mixture model-based clustering algorithm—CRC, or Chinese restaurant cluster (Qin, 2006). This algorithm is designed for clustering genes based on their expression profiles across multiple experiments to infer functional relationships. CRC extends existing model-based clustering algorithms by offering the following additional features: (1) cluster genes showing non-synexpression correlation patterns (time-shifted and/or inverted) together in the same group. (2) Provide accurate estimation of number of clusters. (3) Handle missing data automatically during clustering. No preprocessing steps such as elimination or imputation are needed. (4) Provide multiple strength measurements for each cluster produced, including tightness measure and stability measure such that resulting clusters can be sorted by various criteria to prioritize for follow-up verification.

Significant improvements have been made in CRCView in addition to CRC functionalities. The improvements are 3-fold: user-friendliness, rich visualization options and automatic annotation/interpretation of clustering results. To illustrate these useful features, an example is demonstrated using the sample data extracted from a published human cell cycle study http://www-sequence.stanford.edu/human_cell_cycle/ (Cho *et al.*, 2001). Some sample graphical outputs are illustrated in Figure 1.

CRCView provides an easy-to-use, point-and-click web interface accessible through the internet. No software installation is needed. Due to differences in design, pre-processing and summarization steps, microarray gene expression data formats can be quite different. Users often need to make tedious adjustments to their original data before any analysis can be done. CRCView is designed to allow users to bypass this inconvenient step. In order to handle diverse input data formats, we create an interactive module which prompts the user to supply information on any data format different from a



GO Hierarchy	GO Term	Description	p-values	No.of genes
Molecular function	GO:0019992	Diacylglycerol binding	0.007	2
	GO:0005529	Sugar binding	0.034	2
	GO:0004872	Receptor activity	0.039	7
	GO:0004871	Signal transducer activity	0.044	10

Fig. 1. CRCView analysis of human cell cycle microarray data. Eight clusters are obtained by CRC analysis. (**a**) Trace plot of cluster 7 of 36 genes. Four subclusters are found. The first subcluster is shown in black (**b**), the red genes form the second cluster with time shift pattern (**c**), the third cluster (green) represents inverted gene expression pattern (**d**) and the fourth cluster (blue) includes both time shift and gene expression inversion (**e**). (**f**) Eisen plot (Eisen *et al.*, 1999) of all genes in cluster 7. The color scheme represents the gene expression level instead of subcluster patterns. (**g**) Twenty one genes from four GO Molecular function categories are found significantly enriched based on the GOStats function available in CRCView. Each GO term and gene name are also linked to public GO and Entrez Gene databases.

basic template. Further, missing data is allowed and handled automatically. We also provide automated filtering functionalities including minimum coefficient of variation (CV) based on a Bioconductor (Gentleman *et al.*, 2005) package named genefilter as well as minimum expression level and minimum fold change filtering.

CRCView provides multiple graphical display options of the clustering results including trace plots (Fig. 1a–e) and heatmap (Fig. 1f). The thumb views of trace plots of the all clusters are shown in the results page with basic properties of each cluster shown on top. The clusters can be rearranged by sorting according to criteria such as cluster size. Users can also select a subset of the clusters for further analyses or display (Fig. 1a–e). A Heatplus Bioconductor package is customized to generate modified Heatmap of selected clusters (Fig. 1f). The color matrix is sorted by subclusters in the rows and displayed based on order in the columns. The size and ratio of each heat spot are fixed so that the column and row labels are always readable. Selected cluster images can be exported as one zip file.

An important step in clustering analysis of microarray gene expression data is to rationalizing, verifying and interpreting the clustering results. However, most of the current clustering tools do not provide such functionalities. GO enrichment analysis is a powerful tool for understanding the common features shared by a group of genes, therefore we incorporated it into our CRCView system to enable easy interpretation of the clustering results. A customized GOstats Bioconductor package is used to run GO analysis. Two functions in GOstats, hyperGTest and probeSetSummary, are run sequentially to analyze data and print out the GO term table and the associated significant genes simultaneously (Fig. 1g). To facilitate analysis process, the CRCView server installs from Bioconductor. Seventy eight annotation libraries of microarray chips for human (31), mouse (24), rat (14), zebrafish (1), chicken (1), *Drosophila* (3), *Arabidopsis* (2), *Caenorhabditis elegans* (1) and *Xenopus Laevis* (1). More species will be added and the library will be updated frequently.

The CRCView system is implemented using a three-tier architecture. Users' access and data analysis in CRCView through front-end web browsers are processed using PHP (middle-tier, application server based on Apache) against a MySQL (version 5.0) relational database (back-end,

database server). The analysis results are then presented to the user in the web browser. The two servers also backup each others' data and systems regularly. Any user can register for a CRCView account enabling secure login, data input and analysis and storage of analysis parameters and results. For each project, users can run multiple CRC analyses, and all analysis records are stored in the CRCView database. All functionalities of CRC are available in CRCView. CRCView also provides a user-friendly interactive web interface to implement all the other features discussed in this report including visualization and GO enrichment analysis.

3 SUMMARY

Model-based clustering analysis is a powerful method for microarray gene expression data analysis. Its usage is currently restricted mainly due to lacking of visualization and follow-up analysis tools. The web-based CRCView system fills this vacancy by providing user-friendly CRC analysis pipeline, visualization of clustering results and GO enriched interpretation. We believe this program will make the powerful model-based clustering tool more accessible to researchers conducting data analysis on complex microarray experiment data. We will keep updating the CRCView site by adding more features and functions to make it even more useful.

ACKNOWLEDGEMENT

We thank James MacDonald for valuable suggestions and comments on an earlier version of the website and manuscript.

Conflict of Interest: none declared.

REFERENCES

Cho,R.J. *et al.* (2001) Transcriptional regulation and function during the human cell cycle. *Nat. Genet.*, **27**, 48–54.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Fraley,C. and Raftery,A.E. (1999) MCLUST: software for model-based cluster analysis. *J. Classif.*, **16**, 297–306.

Gentleman,R. *et al.* (2005) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York.

Gollub,J. and Sherlock,G. (2006) Clustering microarray data. *Meth. enzymol.*, **411**, 194–213.

Medvedovic,M. *et al.* (2004) Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, **20**, 1222–1232.

Qin,Z.S. (2006) Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics*, **22**, 1988–1997.

Reich,M. *et al.* (2006) GenePattern 2.0. *Nat. Genet.*, **38**, 500–501.

Yeung,K.Y. *et al.* (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.