BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 1 2002
Pages 207–208

Genesis: cluster analysis of microarray data

Alexander Sturn[1, 2], John Quackenbush[2] and Zlatko Trajanoski[1,*]

[1]Institute of Biomedical Engineering, Graz University of Technology, Krenngasse 37, 8010 Graz, Austria and [2]The Institute for Genomic Research, Rockville, MD 20850, USA

Received on June 23, 2000; revised on August 21, 2000; accepted on September 5, 2000

ABSTRACT

Summary: A versatile, platform independent and easy to use Java suite for large-scale gene expression analysis was developed. Genesis integrates various tools for microarray data analysis such as filters, normalization and visualization tools, distance measures as well as common clustering algorithms including hierarchical clustering, self-organizing maps, k-means, principal component analysis, and support vector machines. The results of the clustering are transparent across all implemented methods and enable the analysis of the outcome of different algorithms and parameters. Additionally, mapping of gene expression data onto chromosomal sequences was implemented to enhance promoter analysis and investigation of transcriptional control mechanisms.

Availability: http://genome.tugraz.at

Contact: zlatko.trajanoski@tugraz.at

INTRODUCTION

High throughput techniques are becoming more and more important in many areas of basic and applied biomedical research. Microarray techniques using cDNAs or oligonucleotides are such high throughput approaches for large-scale gene expression analysis and enable the investigation of mechanisms of fundamental processes and the molecular basis of diseases on a genomic scale.

Microarray experiments have been used to identify differentially expressed genes in a highly parallel manner. Beyond simple discrimination of differentially expressed genes, functional annotation of coexpressed genes (guilt-by-association), diagnostic classification, and investigation of regulatory mechanisms (coregulation from coexpression) require clustering of genes into sets with similar expression patterns. Several clustering techniques have been recently developed and applied to analyze microarray data and tools for exploring expression data using single or multiple clustering algorithms were subsequently reported (Dysvik and Jonassen, 2001). However, to the best of our knowledge, there is no single tool which integrates the common clustering and visualization methods and enables the comparison of the results obtained using different clustering methods.

*To whom correspondence should be addressed.

PROGRAM OVERVIEW

A platform independent Java suite was developed to visualize and analyze a whole set of gene expression experiments. After reading the data from tab-delimited flat files, several graphical representations of the measured signal intensities can be generated showing a matrix of genes (rows) and experiments (columns) (Figure 1). At each point of the analysis, the information for a certain gene can be retrieved from the GenBank by pointing an element in the matrix. Various filters can be applied to the dataset to extract genes of interest for the specific question. Fluorescence ratios can be normalized in several ways to choose the appropiate representation of the data for further analysis. Eleven different similarity distance measurements have been implemented, ranging from simple Pearson correlation or Euclidean distance to more sophisticated approaches like mutual information or Spearman's rank correlation coefficients. To enable thorough cluster analysis and data mining the common clustering algorithms were implemented: hierarchical clustering (Eisen *et al.*, 1998), k-means (Tavazoie *et al.*, 1999), self-organizing maps (Tamayo *et al.*, 1999), principal component analysis (Raychaudhuri *et al.*, 2000), and support vector machines (Brown *et al.*, 2000). Detailed information on the supported data formats and used methods is available on the web site and in the accompanying documentation.

An important and valuable feature of this tool is the transparency of the clustering analysis across all implemented methods (Figure 1). Due to the underlying assumptions in each clustering technique and the necessity to adjust various parameters, the manner in which the data are normalized within and across experiments, and the used similarity measure, the outcome of the clustering can

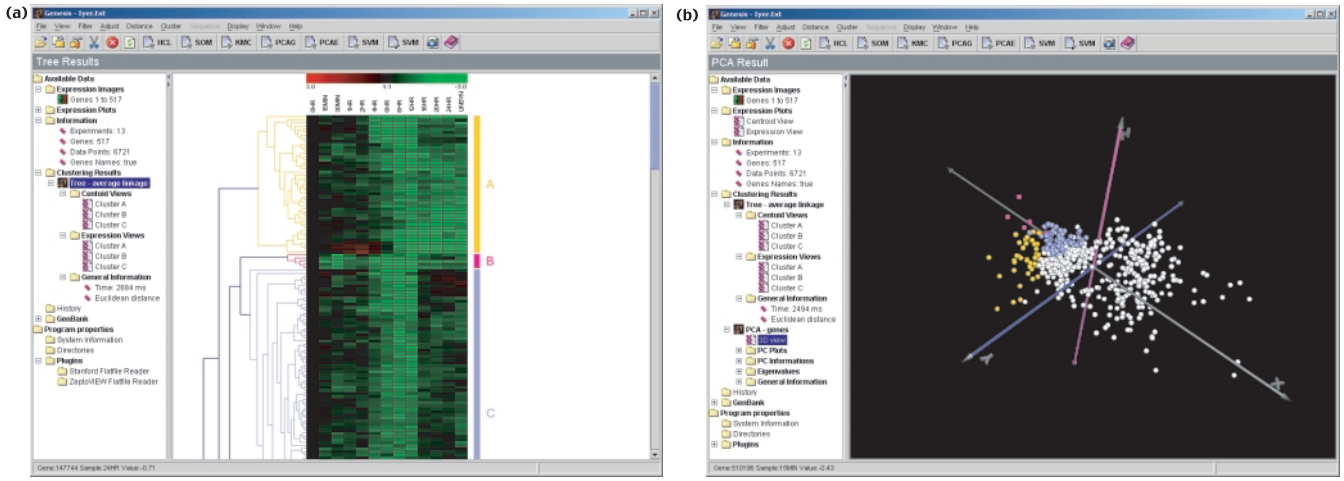

Fig. 1. Program window of the Genesis suite. (a) Hierarchical clustering result. (b) Principal component analysis. The pink, orange and blue balls correspond to the genes with the colored bars A, B, and C shown in (a).

differ substantially. Thus, it is imperative to apply several clustering techniques and parameter values on the same dataset and illuminate different relationships between the data. The comparison of clusters obtained using several techniques can provide the researcher with additional information compared to a single method approach. For instance, one can begin with hierarchical clustering to get a first impression on the number of patterns hidden in the dataset and then use this information to set the number of clusters required for k-means and self-organizing map techniques. Principal component analysis can be then used to visualize these clusters in 3-dimensional space and get an impression on cluster size, integrity, and distribution.

Additionally, Genesis provides the ability to map gene expression data onto chromosomal sequences to enhance the investigation of regulatory mechanisms. Genes on consecutive chromosomal locations are often coexpressed and can be easily identified by this method. Finally, extensive work has been undertaken to accomplish visualization of the gene expression data and clustering results in a user friendly and intuitive way. After designing and coding, the software was extensively tested at several institutions which produce microarray data and the user feedback was taken into consideration. Up to date, the software was tested on Windows 2000, LINUX, Thru64 Unix, Solaris, and Irix platforms.

ACKNOWLEDGEMENTS

This work was supported by the Austrian Ministry for Transport, Innovation and Technology, the Jubiläumsfonds der Österreichischen Nationalbank (8773), and the Austrian Science Fund (P14298,F718).

REFERENCES

Brown,M.P., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M. Jr. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Dysvik,B. and Jonassen,I. (2001) J-express: exploring gene expression data using Java. *Bioinformatics*, **17**, 369–370.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Raychaudhuri,S., Stuart,J.M. and Altman,R.B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.*, 455–466.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.