Vol. 18 no. 2 2002
Pages 327–328

CGO: utilizing and integrating gene expression microarray data in clinical research and data management

Klaus Bumm, Mingzhong Zheng, Clyde Bailey, Fenghuang Zhan, M. Chiriva-Internati, Paul Eddlemon, Julian Terry, Bart Barlogie and John D. Shaughnessy, Jr*

Donna D. and Donald M. Lambert Laboratory of Myeloma Genetics and Myeloma & Transplantation Research Center, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA

Received on July 26, 2001; revised on September 9, 2001; accepted on September 18, 2001

ABSTRACT

Summary: *Clinical GeneOrganizer* (CGO) is a novel windows-based archiving, organization and data mining software for the integration of gene expression profiling in clinical medicine. The program implements various user-friendly tools and extracts data for further statistical analysis. This software was written for Affymetrix GeneChip© *.txt files, but can also be used for any other microarray-derived data. The MS-SQL server version acts as a data mart and links microarray data with clinical parameters of any other existing database and therefore represents a valuable tool for combining gene expression analysis and clinical disease characteristics.

Availability: The software is available free of charge at http://lambertlab.uams.edu

Contact: ShaughnessyJohn@uams.edu

Supplementary information: A snapshot of detailed program-specific table relationships and information on other programs including various cancer gene expression profiles are available on above mentioned web site.

SOFTWARE DESIGN

Gene expression microarrays are high throughput tools for understanding the biology and genetics of many diseases. However, the seemingly unlimited capacity of microarrays has not yet been fully exploited, hampered by the fact that genetic analysis is typically limited to basic research facilities with the need of highly specialized computational biologists to analyzing and mining microarray-derived data (Ermolaeva *et al.*, 1998; Quackenbush, 2001). Microarrays also generate vast amounts of data (12 000 data points × 200–300 patients) that, without prior modification is not in suitable formats for clinical use. In addition microarray data has not been integrated into existing clinical databases. Tools for primary data handling such as data storage and processing (Liao *et al.*, 2000) facilitate the high throughput usage of microarrays. The next step is to bring this data into clinics and include it in a well-rounded patient evaluation. At the University of Arkansas for Medical Sciences we established an Affymetrix GeneChip© based facility in 1999 as part the Myeloma & Transplantation Research Center. Bone marrow samples from patients are routinely processed for gene expression analysis and the need for tools to visualize and extract clinically relevant data became imminent. We developed this software for users that intend to utilize microarray derived gene expression data in clinical settings. CGO differs from already existing data mining databases/analysis tools in its focus on only clinically relevant features and it's capability to interface with other clinical hospital databases.

For developing this software the following requirements had to be taken into consideration:

- due to the accelerated accumulation of microarray files in rather unordered fashion, files had to be stored efficiently on a server and be easily retrievable;
- microarray files had to be linked with patients medical record numbers and be integrated into patients history as a diagnostic event;
- data extraction tools should be on a program surface that is familiar to clinicians and requires little computer knowledge.

SOFTWARE SPECIFICATIONS

CGO, as distributed is a semi-compiled (*.mde) standalone MS-Access 97/2000 database. The interface is a slightly modified version of the standard user interface

*To whom correspondence should be addressed.

for the UAMS 'Multiple Myeloma DataBase' (MMDB) and acts as a front end to a data mart consisting of several MS-SSQL databases and proprietary format databases available on the University of Arkansas for Medical Sciences (UAMS) campus through ODBC (Open DataBase Connectivity). The MS-SSQL server based data mart version including CGO is available through UAMS.

CLINICAL DATA MINING TOOLS

Virtual chip

The 'virtual chip' tool allows the user to create named groups of gene probes (e.g. oncogenes, apoptosis genes, etc.). Using these predefined groups of genes and the sample selector (described below) allows only those genes included in the virtual chip to be reported, graphed, or extracted for analysis. This tool includes a 'text search' function, which facilitates the creation of virtual chips based on descriptions of the genes in the template described below.

Edit template

All genes are represented in the software with a unique identifier number (Array ID). We created a template that linked every gene to all descriptions edited for it such as: chromosome location, symbol, description, activity, function, potential drug target, enzyme nomenclature, etc. The template can be edited and additional information can be added or deleted.

Sample selector

Groups of samples/patients can be selected, either manually or automatically by various criteria (gender, race, age, etc.). Using this sample selector and one of the virtual chips, specific subsets of the data in the system can be analyzed or extracted. Up to ten groups of patients can be defined using this system and summary data for these can be extracted.

Chromosome selector

The chromosome selector allows a focus on specific chromosomes. Genes are extracted by the information edited in the template under G-band position and can be differentiated by chromosome and p- or q-arm location.

Single gene analyzer

This function includes several options. Selected groups of patients can be graphed from highest to lowest gene expression according to their quantitative average difference call (AvgDiff) for any one gene in the template. We integrated an option to flag samples as a normal or control group. This allows evaluating gene expression in contrast to the average as well as highest and lowest normal expression for the gene of interest. All data can be exported to an MS-Excel spreadsheet for further analysis.

Drug target screening

This function was created to screen and rank patients for expression of genes known to be targets of specific drugs (e.g. Campath-1 against CDw52) prior to enrollment on phase I/II clinical trials. Drug target genes are grouped in a virtual chip and selected groups of patients are screened for expression of these genes. The data is presented in an MS-Excel spreadsheet and can be further sorted by expression for each gene.

Statistical query

Every piece of information created by a microarray can be extracted through this program in a predefined fashion. We also integrated statistical tests into this program. In order to create dendrograms with Treeview (Eisen *et al.*, 1998) on a regular basis we wrote a routine that extracts gene expression values on a selected group of patients and presents the data in a format that can be simultaneously imported by GeneCluster.

FUTURE ASPECTS

Given that alterations in gene expression portray a fundamental mechanism of cancer cell growth, expression profiling represents an exciting new frontier in the realm of clinical cancer research. It is anticipated that the identification of novel pathways will lead to the formation of new clinical trials that exploit this knowledge. The application of CGO with its 20–40 already established clinical features in carefully controlled clinical trials may have important implications in predictive medicine and in pharmacogenomic based therapy design.

ACKNOWLEDGEMENTS

Research conducted through private funding provided by Donna D. and Donald M. Lambert Laboratory of Myeloma Genetics and CA55819 from National Cancer Institute, Bethesda, MD.

REFERENCES

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

Ermolaeva,O., Rastogi,M. *et al.* (1998) Data management and analysis for gene expression arrays. *Nature Genet.*, **20**, 19–23.

Liao,B., Hale,W. *et al.* (2000) MAD: a suite of tools for microarray data management and processing. *Bioinformatics*, **16**, 946–947.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nature Rev. Genet.*, **2**, 418–427.