BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 6 2005, pages 821–822
doi:10.1093/bioinformatics/bti113

Genetics and population analysis

CGH-Explorer: a program for analysis of array-CGH data

Ole Christian Lingjærde[1,*], Lars O. Baumbusch[2], Knut Liestøl[1], Ingrid K. Glad[3] and Anne-Lise Børresen-Dale[2]

[1]Department of Informatics, University of Oslo, PO Box 1080 Blindern, N-0316 Oslo, Norway, [2]Department of Genetics, Institute for Cancer Research, Norwegian Radium Hospital, Montebello, N-0310 Oslo, Norway and [3]Department of Mathematics, University of Oslo, PO Box 1053 Blindern, N-0316 Oslo, Norway

Received on April 14, 2004; revised on October 24, 2004; accepted on October 25, 2004
Advance Access publication November 5, 2004

ABSTRACT

Summary: CGH-Explorer is a program for visualization and statistical analysis of microarray-based comparative genomic hybridization (array-CGH) data. The program has preprocessing facilities, tools for graphical exploration of individual arrays or groups of arrays, and tools for statistical identification of regions of amplification and deletion.

Availability: The program is available as Java class files that runs on any platform with the Java 2 runtime environment (J2SE JRE) installed, and as a Windows executable. Java source files are also available. See http://www.ifi.uio.no/bioinf/ Papers/CGH/

Contact: ole@ifi.uio.no

INTRODUCTION

Genomic DNA copy number changes are key genetic events in the development of many cancers. Copy number changes vary widely with respect to genomic location, magnitude of change and size of the affected genomic region. Changes may cause dysregulation of gene expression for genes that are critical to tumorigenesis and the progression of cancer. The detection and mapping of these genes is important for understanding cancer progression and for associating aberrations with cancer phenotypes (Lengauer *et al*., 1998).

In array-CGH, differentially labeled tumor and normal DNA are cohybridized to a microarray of mapped sequences, and fluorescence intensities are measured. Intensity ratios are roughly proportional to the copy number ratios of the corresponding DNA samples, providing a means to quantitatively measure genomic copy number for thousands of genes at a time and to map these changes to chromosomal regions (Solinas-Toldo *et al*., 1997; Pinkel *et al*., 1998).

Earlier reportings of software for analysis of array-CGH data include a MATLAB toolbox for plotting the data genome-wide or chromosome-by-chromosome and for detecting change points (Autio *et al*., 2003). Their approach does not include any assessment of the statistical significance of the reported copy number changes. Also, Frankenberger *et al*. (2003) reported on the development of a data analysis tool for array-CGH data.

In this paper, we report on a comprehensive menu-driven standalone program for graphical exploration and statistical analysis of array-CGH data. The program is open-source and portable. The program supports import of text files and Excel files, data preprocessing (including missing data imputation, centering, various data transformations and smoothing), graphical exploration of individual arrays or groups of arrays on arbitrary genomic scales, and detection of statistically significant changes in copy number ratio. The program offers the user many choices to modify settings and customize views. Microarray data measuring mRNA expression levels may also be imported and displayed, either separately or together with copy number data.

GRAPHICAL ANALYSIS

Basic plotting functionality includes point plots, frequency plots and checkerboard plots of copy numbers as a function of genomic position. The program supports data views on arbitrary genomic scales, ranging from the whole genome to part of a chromosome. The user can zoom in on any part of the present view using the mouse to indicate the region of interest. Data from different arrays can be plotted together in the same coordinate system, with layout (color and plotting symbol) chosen independently for each array (Fig. 1).

Smoothing filters may be applied and the effect instantly seen on a plot. The user may choose between a symmetric moving average filter and a filter specifically designed for discontinuous signals that identifies change points (i.e. starts and ends of all amplified and deleted regions) and thus provides an estimate of the copy number ratio between each pair of consecutive change points.

Information about individual genes in a plot can be obtained in several ways. One may search for genes with names that contain a particular phrase and highlight these genes in a plot. Alternatively, pointing with the mouse at a measurement in a plot shows the name of the gene. Finally, using the mouse the user can select a collection of measurements and obtain a table with textual information about the corresponding genes and links for each gene to a genomic gateway such as *Entrez* or *Ensembl*.

DETECTION OF SIGNIFICANT COPY NUMBER CHANGES

Several methods for detecting copy number changes are implemented in CGH-Explorer. The most basic functionality is to manually specify an upper and lower threshold and the number of hits required within a region of specified width in order to count the region as an amplification or deletion. Thresholds may also be found automatically, using either a normal DNA reference sample or a novel

*To whom correspondence should be addressed.

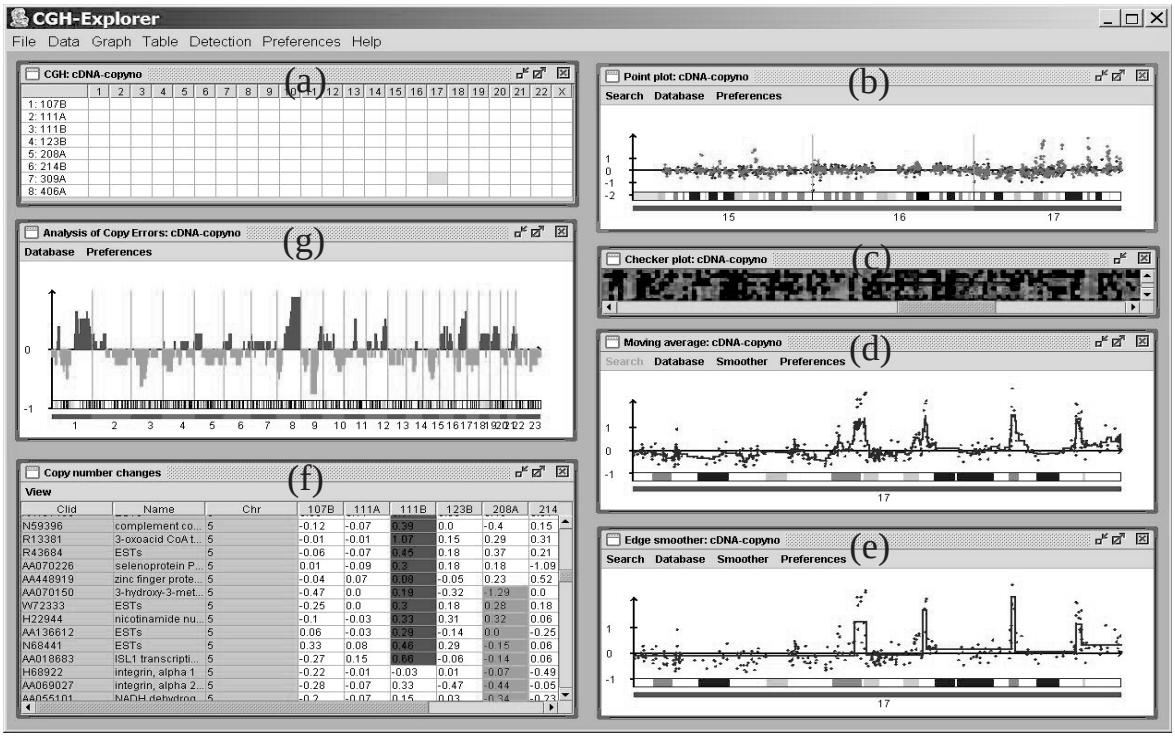

Fig. 1. CGH-Explorer Screenshot. Clockwise from top left corner: (a) window for selecting subset of data to plot or analyze; (b) scatterplot of data over three chromosomes where different colors correspond to different arrays (the bars at the bottom in this and other plots show cytoband locations and chromosomes); (c) heat map of subset of the data; (d) estimation of copy number ratios using a moving average filter; (e) estimation of copy number ratios using an edge-preserving filter; (f) table showing the result of using ACE to find copy number errors; (g) frequency plot showing for each gene the percentage of samples being altered.

bootstrap-based method implemented in CGH-Explorer. In addition, a novel method called analysis of copy errors (ACE), not based on thresholding the copy number ratios, is implemented. See the paper's website for information about the above methods.

REFERENCES

Autio,R., Hautaniemi,S., Kauraniemi,P., Yli-Harja,O., Astola,J., Wolf,M. and Kallioniemi,A. (2003) CGH-Plotter: MATLAB toolbox for CGH-data analysis. *Bioinformatics*, **19**, 1714–1715.

Frankenberger,C., Urzúa,U., Church,D., Powell,J., Burgess,T., Tawady,T., Gangi,L. and Munroe,D. (2003) aCGH explorer: a tool for high-resolution analysis of chromosomal imbalances using cDNA and oligo DNA arrays (poster abstract). *Biomedical Information Science and Technology Initiative (BISTI) 2003 Symposium, NIH*, Bethesda, MD, November 6–7.

Lengauer,C., Kinzler,K.W. and Vogelstein,B. (1998) Genetic instabilities in human cancers. *Nature*, **396**, 643–649.

Pinkel,D., Segraves,R., Sudar,D., Clark,S., Poole,I., Kowbel,D., Collins,C., Kuo,W.L., Chen,C., Zhai,Y., *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Solinas-Toldo,S., Lampel,S., Stilgenbauer,S., Nickolenko,J., Benner,A., Dohner,H., Cremer,T. and Lichter,P. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.