CGH-Plotter: MATLAB toolbox for CGH-data analysis

Reija Autio1, Sampsa Hautaniemi1,∗, Päivikki Kauraniemi3, Olli Yli-Harja1, Jaakko Astola1, Maija Wolf2 and Anne Kallioniemi3

1Institute of Signal Processing, Tampere University of Technology, PO Box 553, 33101 Tampere, Finland, 2Medical Biotechnology Group, VTT Technical Research Centre of Finland, PO Box 106, FIN-20521 Turku, Finland and 3Laboratory of Cancer Genetics, Institute of Medical Technology, University of Tampere and Tampere University Hospital FIN-33520 Tampere, Finland

Received on November 8, 2002; revised on February 14, 2003; accepted on March 4, 2003

ABSTRACT
Summary: CGH-Plotter is a MATLAB toolbox with a graphical user interface for the analysis of comparative genomic hybridization (CGH) microarray data. CGH-Plotter provides a tool for rapid visualization of CGH-data according to the locations of the genes along the genome. In addition, the CGH-Plotter identifies regions of amplifications and deletions, using k-means clustering and dynamic programming. The application offers a convenient way to analyze CGH-data and can also be applied for the analysis of cDNA microarray expression data. CGH-Plotter toolbox is platform independent and requires MATLAB 6.1 or higher to operate.

Availability: http://sigwww.cs.tut.fi/TICSP/CGH-Plotter
Contact: sampsa.hautaniemi@tut.fi

INTRODUCTION
Copy number changes, such as deletions and amplifications, are common aberrations in cancer and are known to involve genes that play a crucial role in the development and progression of this malignant disease (Gray and Collins, 2000). Identification of such gene copy number changes is extremely important as it offers a basis for understanding the pathogenesis of cancer and provides essential tools for improved clinical management of cancer.

The copy number changes in cancer usually span large regions of the genome and therefore influence several genes at the same time. Comparative genomic hybridization (CGH) on DNA microarray allows simultaneous monitoring of copy numbers of thousands of genes throughout the genome (Monni et al., 2001; Pollack et al., 1999). In this technique, differentially labeled tumor and normal reference DNA are hybridized on DNA microarrays. Differences in the tumor to normal fluorescence intensity are quantitated and the resulting copy number ratios reflect the relative gene copy number changes in the tumor sample.

Acquisition of copy number information on thousands of genes creates problems for the analysis and interpretation of the data. Here, we have developed a software tool (CGH-Plotter) for automated high-throughput analysis of CGH microarray data. CGH-Plotter allows researchers (i) to plot CGH copy number data as a function of the position of the genes along the human genome, and (ii) to rapidly determine the exact locations of copy number changes, such as amplicons and deletions.

ASSUMPTIONS AND METHODS
We assume that CGH-data and indices to chromosome boundaries are stored in Matlab format. Detailed instructions for data formats and usage of the CGH-Plotter are given in User Manual, which is available at our website. Currently, the CGH-Plotter utilizes three algorithms: moving median and mean filtering, k-means clustering, and dynamic programming. As the copy number changes in cancer usually span large regions of the genome, CGH copy number ratios from one sample can be considered as a signal consisting of constant levels, which are to be identified. Constant levels that are above the baseline are referred to as amplicons, while constant levels below the baseline are called deletions. We have employed a three stage procedure for detecting constant levels:

1. Filtering. To reduce the effect of noise the data are filtered with a moving median or mean filter. Preferred filtering method and the size of the moving window can be defined at the graphical user interface.

2. k-means clustering. The k-means clustering is applied to estimate the maximum number of changes in each chromosome. Here k is three denoting amplified, deleted and baseline clusters.
CGH-plotter: MATLAB toolbox

3. Dynamic programming. At this stage actual constant levels (i.e. amplicons, deletions, and baselines) in the CGH-data are determined.

The core of the analysis procedure is dynamic programming, which is based on the assumption that a measured CGH-ratio is the sum of underlying ‘true’ value and Gaussian noise. The dynamic program utilizes the Markov property and identifies change points of the constant levels by minimizing sum of mean square errors for all combinations of CGH ratios.

RESULTS

The main purpose of the CGH-Plotter is to display CGH-data and the boundaries computed in the analysis phase. The data can be plotted against genomic index or base-pairs. The genomic index reflects the order of genes along the genome whereas base-pairs represents the genuine locations of the genes (Figure 1A). In addition to being more informative and easily accessible from public databases (e.g. UCSC/Santa Cruz, http://genome.ucsc.edu), the cumulative base-pair plots permit the user to evaluate whether a specific region of genome has a large enough representation of genes in the array in order to make reliable conclusions. Accordingly, the user may utilize the CGH-Plotter for designing new microarray experiments.

CGH-ratios can be plotted either from one chromosome across multiple samples or from a single sample genomewide. Also, the constant levels representing amplicons and deletions can be displayed with or without the data. The CGH-Plotter allows results to be saved easily to a tab delimited text file (Figure 1B). The output file includes information on the type of the change (amplicon or deletion), start and end points using genomic indices and cumulative base-pairs, the height of the amplicon or deletion, and the maximum or minimum ratio within the amplicon or deletion.

In summary, CGH-Plotter is freely available toolbox that enables rapid analysis and illustration of CGH microarray data. The graphical interface is user-friendly and functions can be modified easily. Even though CGH-Plotter is designed for the analysis of CGH-data, it is also applicable for the analysis of cDNA microarray expression data.

ACKNOWLEDGEMENTS

The authors are grateful to David Venet whose insightful suggestions for improving the code speeded up the CGH-Plotter considerably.

REFERENCES

