BIOINFORMATICS

Vol. 19 no. 13 2003, pages 1714–1715
DOI: 10.1093/bioinformatics/btg230

CGH-Plotter: MATLAB toolbox for CGH-data analysis

Reija Autio[1], Sampsa Hautaniemi[1,], Päivikki Kauraniemi[3], Olli Yli-Harja[1], Jaakko Astola[1], Maija Wolf[2] and Anne Kallioniemi[3]*

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

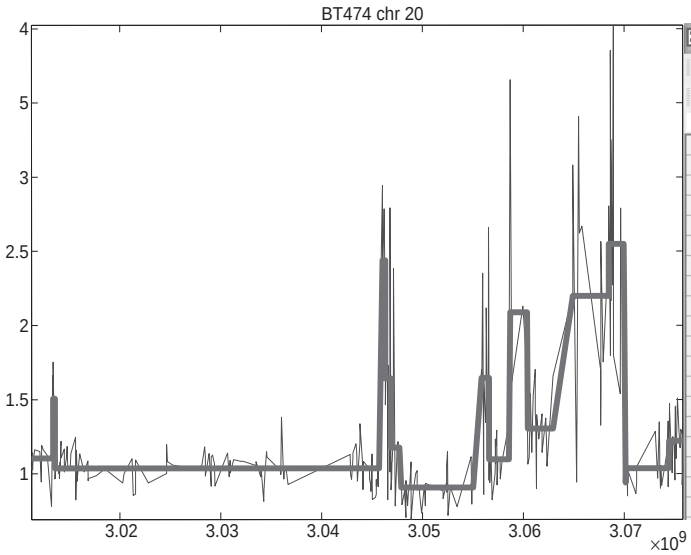
ASSUMPTIONS AND METHODS

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.

*To whom correspondence should be addressed.

A



B

Microsoft Excel - boundaries

File Edit View Insert Format Tools Data Window Help Acrobat

	A	B	C	D	E	F
1	CHROMOSOME20	SKBR3	MCF7	BT474	BT20	MDA361
2						
3	1 Type	Amplicon	Amplicon	Amplicon	Deletion	Amplicon
4	1 Start	10713	10712	10713	10696	10713
5	1 End	10721	10721	10721	10704	10721
6	1 Start Basepair	3013370320	3013370320	3013370320	3011267311	3013370320
7	1 End Basepair	3013567065	3013567065	3013567065	3012060376	3013567065
8	1 Height	1.7515	1.3385	1.506	0.75281	1.3293
9	1 Max/Min	2.2184	1.506	1.7524	0.9545	1.5851
10						
11	2 Type	Amplicon	Amplicon	Amplicon	Amplicon	
12	2 Start	10840	10892	10840	10713	
13	2 End	10888	10904	10845	10720	
14	2 Start Basepair	3046048414	3052383073	3046048414	3013370320	
15	2 End Basepair	3052040082	3054882912	3046336026	3013559603	
16	2 Height	1.4106	1.2839	2.4384	1.7382	
17	2 Max/Min	1.9364	1.8971	2.9431	1.9778	
18						

Fig. 1. (**A**) Copy number ratio profile and constant levels from the analysis of chromosome 20 in BT474 breast cancer cell line. *X*-axis depicts the cumulative base-pair location from the short arm to the long arm of the chromosome and *Y*-axis depicts the copy number ratios between the cell line and the reference. (**B**) Partial results from the analysis of five breast cancer cell line samples.

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 genome-wide. 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

Gray,J. and Collins,C. (2000) Genome changes and gene expression in human solid tumors. *Carcinogenesis*, **21**, 443–452.

Monni,O., Bärlund,M., Mousses,S., Kononen,J., Sauter,G., Heiskanen,M., Chen,Y., Bittner,M. and Kallioniemi, A. (2001) Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. *Proc. Natl Acad. Sci. USA*, **98**, 5711–5716.

Pollack,J., Perou,C., Alizadeh,A., Eisen,M., Pergamenschikov,A., Williams,C., Jeffrey,S., Botstein,D. and Brown,P. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, **23**, 41–46.