BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 18 2004, pages 3636–3637
doi:10.1093/bioinformatics/bth355

Breakpoint identification and smoothing of array comparative genomic hybridization data

Kees Jong[1,*], Elena Marchiori[1], Gerrit Meijer[2], A. v. d. Vaart[1] and Bauke Ylstra[2]

[1]Faculty of Sciences and [2]VU University Medical Center, Vrije Universiteit, De Boelelaan 1117, Amsterdam 1081HV, The Netherlands

Received on October 10, 2003; revised on May 11, 2004; accepted on June 7, 2004
Advance Access publication June 17, 2004

ABSTRACT

Summary: We describe a tool, called aCGH-Smooth, for the automated identification of breakpoints and smoothing of microarray comparative genomic hybridization (array CGH) data. aCGH-Smooth is written in visual C++, has a user-friendly interface including a visualization of the results and user-defined parameters adapting the performance of data smoothing and breakpoint recognition. aCGH-Smooth can handle array-CGH data generated by all array-CGH platforms: BAC, PAC, cosmid, cDNA and oligo CGH arrays. The tool has been successfully applied to real-life data.

Availability: aCGH-Smooth is free for researchers at academic and non-profit institutions at http://www.few.vu.nl/~vumarray/

Contact: cjong@few.vu.nl

INTRODUCTION

Array comparative genomic hybridization (array CGH) is a laboratory approach for genome-wide measurement of aberrations in chromosomal copy numbers. The goal of the array CGH technique is the detection of DNA sequence copy number changes and determination of the associated breakpoints along the chromosomes.

DNA copy number aberrations are used, for instance, to search for genes located in regions of recurrent chromosomal gains, amplification or deletions. It is therefore of fundamental relevance to identify as precisely as possible the boundaries of chromosomal regions with aberrant copy numbers.

Because chromosomal DNA copy numbers for technical reasons cannot be measured directly, DNA from test cells is directly compared to DNA from normal cells, using several thousand small DNA fragments, with known identity and genomic position (frequently referred to as clones or BACs), as probes. Every single experiment yields tumor to normal ratios for each clone on the array, and thus each chromosomal location (see Pinkel *et al*., 1998).

*To whom correspondence should be addressed.

The ratios found in the experiment have some noise generated by polymorphic sites (sequence variation between individuals), some experimental noise as well as compression of the ratios due to aneuploidy and admixed non-tumor cells. This noise renders the identification of breakpoints and the determination of the true copy number values problematic. To facilitate and standardize this process, aCGH-Smooth has been developed.

The core of aCGH-Smooth is a heuristic algorithm, originally introduced in Jong *et al*. (2003). It identifies potential breakpoints and smooths the observed array CGH values between consecutive breakpoints to a suitable common value. The output of aCGH-Smooth is a mapping which associates to each clone a new value. For every batch of experiments, aCGH-Smooth allows to adapt the settings for smoothing and breakpoint recognition to the requirements of the raw data, depending on the biological and technical quality of the samples analysed.

Other methods/tools for array-CGH analysis based on different computational approaches (Fridlyand *et al*., 2004), which uses HMCM for modelling the possible dependence of a clone with its near neighbour, and Olshen and Venkatraman (2002), which uses a variant of the binary segmentation approach.

METHODS

aCGH-Smooth takes as input Excel files, e.g. the ones produced by, the GenePix Pro software (Axon inc, CA), the 'UCSF Spot' software (Jain *et al*., 2002) or Imagene (Biodiscovery Inc., El Segundo, CA). A file describes one array CGH experiment, which includes a sequence of clone positions in the genome and clone values measured in the experiment.

aCGH-Smooth uses a heuristic algorithm for identifying breakpoints in such a sequence and for smoothing its values. This algorithm is the best performing of the three memetic algorithms (population-based stochastic iterative algorithms incorporating local search) introduced in Jong *et al*. (2003).

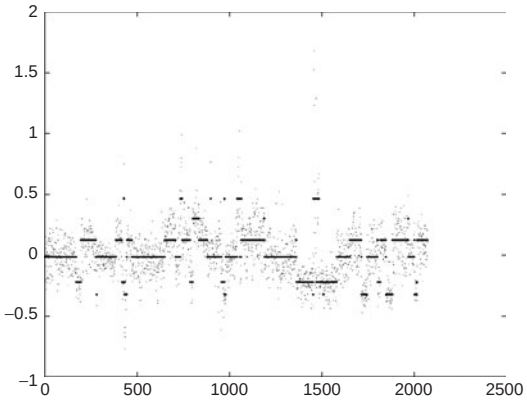

Fig. 1. aCGH-Smooth applied to BAC data.

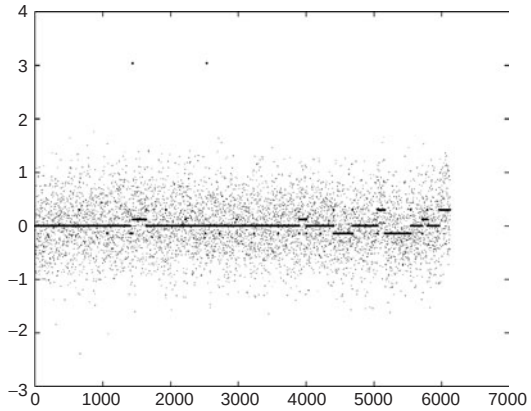

Fig. 2. aCGH-Smooth applied to oligo data.

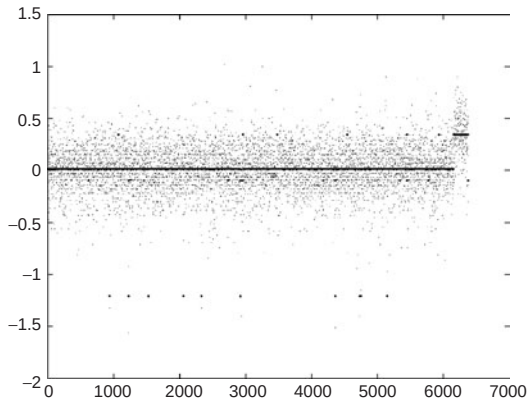

Fig. 3. aCGH-Smooth applied to cDNA data.

The algorithm is based on the assumption that the experimental noise in the data is generated by a Gaussian process. The algorithm uses a regularized maximum likelihood criterion for measuring the goodness of a given placement of breakpoints.

A novel option of aCGH-Smooth allows the user to set the value of a threshold used for detecting potential amplicons and outliers. The corresponding clones are not considered by the smoothing algorithm.

The tool, a demo input file, a user guide and the paper describing the core algorithm can be downloaded at http://www.few.vu.nl/~vumarray/.

RESULTS

We show the results of aCGH-Smooth with default parameters when applied to three types of array-CGH experiments.

Figure 1 shows a gastric tumor experiment, performed at the UCSF Cancer Center, with a CGH array that has PCR representations of BAC and PAC clones as a probe spotted on the array. Each spot on the array covers ∼100–200 kb of the human genome.

Figure 2 is a 'normal to normal' experiment, except that there is an extra copy of chromosome 18. It was performed at the VU Medical Center of Amsterdam. It uses CGH arrays that have synthetically synthesized oligos as a probe spotted on the array. Each spot on the array covers 60 bp of the human genome.

Figure 3 shows a 'normal to normal' experiment, except that there are 3 copies of the X chromosome. It was performed at Stanford, and uses CGH arrays that have cDNAs as a probe spotted on the array. Each spot on the array covers between 200 and 1500 bp of the human genome.

We found that only in few points there is a clear difference between the results given by aCGH-Smooth and by the expert. aCGH-Smooth can thus easily and effectively be applied for data generated by different analysis programs and different array-CGH platforms.

REFERENCES

Fridlyand,J., Snijders,A., Pinkel,D., Albertson,D., and Jain,A. (2004) Understanding array CGH data. *J. Multivariate Anal.*, in press.

Jain,A.N., Tokuyasu,T.A., Snijders,A.M., Segraves,R., Albertson,D.G., and Pinkel,D. (2002) Fully automatic quantification of microarray image data. *Genome Res.*, **12**, 325–332.

Jong,K., Marchiori,E., van der Vaart,A., Ylstra,B., Meijer,G. and Weiss,M. (2003) Chromosomal breakpoint detection in human cancer. In LNCS, vol. 2611, Springer.

Olshen,A. and Venkatraman,E. (2002) Change-point analysis of array-based comparative genomic hybridization data. In *Proceedings of Joint Statistical Meetings*, pp. 2530–2535.

Pinkel,D., Segraves,R., Sudar,D., Clark,S., Poole,I., Kowbel,D., Collins,C., Kuo,W., 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.