APPLICATION NOTE

Genomics & Informatics Vol. 5(3) 137-139, September 2007

RAN-aCGH: R GUI Tools for Analysis and Visualization of an Array-CGH Experiment

Sangcheol Kim[1] and Byung-Soo Kim[1]*

[1]Department of Applied Statistics, Yonsei University, Seoul 102-749, Korea

Abstract

RAN-aCGH is an R GUI tool for the analysis and visualization of array comparative genomic hybridization (array-CGH) experiments. The tool consists of data-loading, preprocessing for missing data, several methods for statistical identification of DNA copy number aberration, and visualization of the copy number change. RAN-aCGH requires a single input format, provides various visualizations, and allows the addition of a new statistical method, all in a user-friendly graphic user interface (GUI).

Availability: RAN-aCGH (Windows, Mac OS) is freely available for academic researchers. http://web.yonsei.ac.kr/BSKimLab/Public_paper.htm

Supplementary information: Source program, data files, and technical documentation are available at http://web.yonsei.ac.kr/BSKimLab/Public_paper.htm

Keywords: array-CGH, R GUI, copy number aberration

Introduction

DNA copy number aberrations are key genetic events in the development and progression of various human cancers (Pollack *et al.*, 2002). Array comparative genomic hybridization (array-CGH) technology has been recently developed, and it allows us to measure aberrations pertaining to DNA copy number changes at tens of thousands of chromosomal loci simultaneously. The array-CGH is used to search for chromosomal regions of gain, loss, amplification, or deletion related to cancer development and recurrence. The purpose of array-CGH analysis is to divide the whole genome into segments such that the copy number differs between contiguous segments and subsequently to quantify the copy number in each segment (Pinkel *et al.*, 1998; Jong *et al.*, 2004).

*Corresponding author: E-mail bskim@yonsei.ac.kr
Tel +82-2-2123-4541, Fax +82-2-313-5331
Accepted 4 August 2007

Recently, several statistical methods have been developed for the analysis of array-CGH experiments. These are aCGH-Smooth (Jong *et al.*, 2004), CBS (Olshen *et al.*, 2004), GLAD (Hupé *et al.*, 2004), HMM (Fridlyand *et al.*, 2004), CGHseg (Picard *et al.*, 2005), CLAC (Wang *et al.*, 2005), Lasso-based penalized least squares regression (Huang *et al.*, 2005), Wavelet (Hus *et al.*, 2005), and a spatially-correlated mixture model (Broët and Richardson, 2006). Many analysis tools are available in the MATLAB toolbox or in R packages in *Bioconductor* (http://www.bioconductor.org) or the R-Project website (http://www.r-project.org). However, those are provided only with simple visualization and a few statistical methods for identifying aberrations of chromosomal regions. Biologists may find it difficult to employ a subset of these methods due to their inconsistent input format and different preprocessing steps.

In this study, RAN-aCGH (Fig. 1) has been developed for the analysis of array-CGH experiments. RAN-aCGH is built on the open-source R software and provides a user-friendly interface. The RAN-aCGH interface consists of data loading, preprocessing, choosing methods, and a window with a scrollbar showing software parameters that a user can choose. The advantage of RAN-aCGH is its single input format through which one can process the preprocessing steps and run several statistical analyses to completion with just a few clicks. It provides various plots of DNA copy number change and one can add a new statistical method to it.

Components of RAN-aCGH

Data Loading

The input dataset to be processed should be given in tab-delimited text files. Chromosome mapping information of genes and postnormalization data of array-CGH experiments can be uploaded as two independent files. The data file consists of a header line explaining information and sample name by column. Detailed information on the data.txt and info.txt is available in the manual at http://web.yonsei.ac.kr/BSKimLab/Public_paper.htm.

Preprocessing

For the imputation of missing values, we implement the

preprocessing steps that consist of no missing proportion (NMP) and imputation. The NMP of a gene is defined as the proportion of valid observations out of the total number of samples (Kim *et al.*, 2005). We can impute missing values for each chromosome using the k-nearest neighbors (K-NN) algorithm (Troyanskaya *et al.*, 2001). The preprocessing has default values of 80% for NMP and k=10 for K-NN. The R package *impute* is used for K-NN.

Methods

The RAN-aCGH includes almost all the methods that have been proposed in R packages. These include a circular binary segmentation (CBS) method (Olshen *et al.*, 2004), unsupervised hidden Markov model (HMM) approach (Fridlyand *et al.*, 2004), gain and loss analysis of DNA (GLAD) using adaptive weights smoothing (AWS) (Hupe *et al.*, 2004), cluster along chromosomes (CLAC) method using the hierarchical clustering algorithm (Wang *et al.*, 2005), a penalized least squares regression (Huang *et al.*, 2005), and the wavelet approach (Hsu *et al.*, 2005). For implementing these statistical methods in RAN-aCGH, the *DNAcopy*, *GLAD*, *clac*, *lars*, *waveslim*, *cluster*, *splines*, *multtest*, *Biobase*, and *tools* packages of R and *Bioconductor* are used.

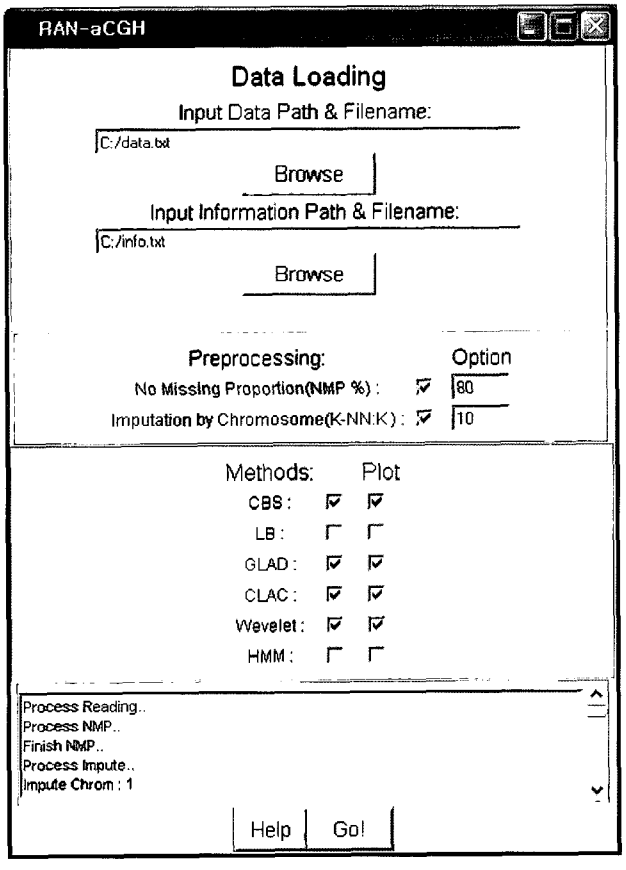

Fig. 1. Graphic User Interface (GUI) of RAN-aCGH.

Results and Discussion

One can click the "Go!" button after loading the data set, select preprocessing parameters such as NMP proportion and a k value for K-NN, and finally choose statistical methods (Fig. 1). RAN-aCGH returns preprocessing and statistical methods that are chosen by the user in a small window (Fig. 1). Output files are automatically saved as text format files.

We noted that most of the statistical methods proposed for detecting copy number aberrations were run as web-based tools, such as MATLAB toolbox or R packages. However, these tools have their own input formats, which are quite different from one another. Furthermore, one must learn MATLAB or R to run these tools. Biologists may find it difficult to employ a subset of these methods due to their inconsistent input format and different preprocessing steps. RAN-aCGH features single input format of input data files and user-friendly GUI with a few mouse clicks. Furthermore, users can add new statistical methods into RAN-aCGH.

Acknowledgements

This work was supported by a Korean Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MOST) (No. R01-2004-000-10057-0).

References

Broët, P., and Richardson, S. (2006). Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics* 22, 911-918.

Fridlyand, J., Snijder, A.M., Pinkel, D., Albertson, D.G., and Jain, A.N. (2004). Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis* 90, 132-153.

Hsu, L., Self, S.G., Grove, D., Randolph, T., Wang, K., Delrow, J.J., Loo, L., and Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* 6, 211-226.

Huang, T., Wu, B., Lizardi, P., and Zhao, H. (2005). Detection of DNA copy number alterations using pealized least squares regression. *Bioinformatics* 21, 3811-3817.

Hupé, P., Stransky, N., Thiery, J.P., Radvanyi, F., and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* 20,

3413-3422.

Jong, K., Marchiori, E., Meijer, G., Vaart, A., and Ylstra, B. (2004). Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics* 20, 3636-3637.

Kim, B.S., Kim, I., Lee, S., Kim, S., Rha, S.Y., and Chung, H.C. (2005). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics* 21, 517-528.

Olshen, A.B., Venkatraman, E.S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5, 557-572.

Picard, F., Robin, S., Livielle, M., Vaisse, C., and Daudin, J.J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics* 6, 27.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B., Gray, J.W., and Albertson, D.G. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* 20, 207–211.

Pollack, J.R., Sørlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Børresen-Dale, A.L., and Brown, P.O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences* 99, 12963-12968.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* 17, 520-525.

Venkatraman, E.S., and Olshen, A.B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 23, 657-663.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B., and Tibshirani, R. (2005). A method for calling gains and losses in array CGH data. *Biostatistics* 6, 45-58.