BIOINFORMATICS APPLICATIONS NOTE
Vol. 18 no. 8 2002
Pages 1139–1140

MArray: analysing single, replicated or reversed microarray experiments

Junbai Wang*, Vigdis Nygaard, Birgitte Smith-Sørensen, Eivind Hovig and Ola Myklebost

Department of Tumour Biology, Norwegian Radium Hospital, N0310 Oslo, Norway

Received on December 21, 2001; revised on February 14, 2002; accepted on March 1, 2002

ABSTRACT

Summary: MArray is a Matlab toolbox with a graphical user interface that allows the user to analyse single or paired microarray datasets by direct input of the raw data output file from image analysis packages, such as QuantArray or GenePiX. The application provides simple procedures to manually evaluate the quality of each measurement, multiple approaches to both ratio normalization (simple normalization, intensity dependent normalization) and evaluation of the reproducibility of paired experiments (using the techniques ‘simple statistical method’ and ‘quality control ellipse’ and ‘significance analysis of microarrays’). Specifically, interactive spot evaluation functions are available in MArray and an online gene information database (NCBI UniGene) is linked. The application may provide a valuable aid in selecting and optimizing experimental procedures, as well as serving as an analytical tool for two-state biological comparisons, such as a study of single-dose activation. It is entirely platform independent, and only requires Matlab installed.

Availability: http://matrise.uio.no/marray/marray.html

Contact: junbaiw@radium.uio.no

INTRODUCTION

DNA microarray analysis has attained wide interest from researchers in the study of gene expression. Recently, reports have increasingly focused on the importance of quality control (Lee *et al.*, 2000; Wang *et al.*, 2001) and statistical design (Kerr and Churchill, 2001) of microarray experiments. Some publicly available software, i.e. J-express (Dysvik and Jonassen, 2001) and MaxdView (Manchester, 2001) provide simple functions for filtering and normalization of raw microarray data, but the quality control and significance analysis of paired experiments (like either identical or reciprocal fluorochrome labelling of sample and reference) are not available. Although ‘significance analysis of microarray’ (SAM; Tusher and Chu, 2001) is a statistical package for finding significant genes in a set of microarray experiments, it does not provide any function to evaluate the quality of input data. However, there is a need for tools for experimental quality control and identification of systematic errors, as well as relatively simple tools for ‘two-state’ comparisons of biological models.

We here present an application of this type, MArray, which encompasses the basic tools needed for processing of raw data from microarray experiments. Most subroutines of MArray can be used in other Matlab scripts, which enables advanced users to build their own functions or add new functions.

DESCRIPTION

With a graphical user interface, MArray enables the user to easily access the various analysis tools. QuantArray or GenePix data export files in tab delimited text format that can be directly loaded. As noise in a dataset will directly affect later studies (Lee *et al.*, 2000), the raw intensity measurements should be pre-processed before any statistical procedures are performed. This data cleaning procedure excludes all empty and manually flagged spots. Secondly, the intensity values for each channel are calibrated relative to the median of the channel background. This prevents a bias in the scale of intensity measurements obtained from the two channels and provides a reasonable ratio comparison between their expression levels. By small modifications of the source code, advanced users may add extra functions to MArray, such as here done with F-Scan (Munson and Young, 2001), which shows the composite microarray image in an interactive window. Finally, two manually adjustable parameters, minimum signal intensity of each channel and the signal to noise ratio [(Signal $-$ Background)/(Background $+$ $2 \times$ Background standard deviation)] are applied to assess the quality of the filtering (Wang *et al.*, 2001).

After pre-processing, there are a number of ways to view the data. These include histograms of spot ratio or intensity distributions, scatter plots of log2, linear or cubic root transformed data, where the gene name will be displayed when a data point is clicked, while more detailed gene in-

*To whom correspondence should be addressed.

formation can be obtained from NCBI's UniGene database (NIH, 2000) by pressing the Clone2Web button. A log2-ratio versus overall intensity plot may reveal a dye bias resulting from an uneven labelling efficiency and background level for each fluorochrome. Due to the non-linear behaviour of this effect, a simple normalization (assuming both channels have equal total signal intensities) is not suited to correct such a bias. However, intensity dependent normalization (lowest regression method) can compensate these effects through robust local linear fits (Yang *et al.*, 2001).

For the analysis of paired experiments, MArray provides interactive tools to investigate the reproducibility of experiments. In the scatter plot view, a quality control ellipse is centered around the intersection of the means in each experiment. This represents the T^2 by limit and is a solution of $T^2 = y'y$ ($y'y$ being the sum of squares of the principal components). T^2 is a quantity indicating the overall conformance of an individual observation vector to its mean and show the level of reproducibility (Jackson, 1980). When reproducibility is high the T^2 ellipse is a narrow ellipse, and the data can be further analysed. If the T^2 ellipse is wide, then one needs to examine the trends that lead to the poor reproducibility.

Correlation coefficients measure the strength of a relationship between two variables, not the agreement between them and data that seem to be in poor agreement can produce quite high correlations (Bland and Altman, 1986). Therefore, we have implemented another simple statistical method; a plot of the differences between paired measurements against their mean intensity, as this may be more informative (Altman, 1991). This analysis is done in an interactive plot where the blue line is the mean of differences and two red lines represent the 95% confidence interval for the mean differences. Gene information and the significance score $d(i)$ of each spot will be displayed in the analysis window when the data point is clicked. The $d(i)$ score is a solution of $d(i) = r(i)/(s(i) + s0)$, where $r(i)$ is the difference between paired observations, $s(i)$ is a standard deviation and $s0$ is a fudge factor. The value of $s0$ was chosen to minimize the coefficient of variation of $d(i)$ (Tusher and Chu, 2001).

Due to Matlab's built-in functions, MArray allows all plots to be saved as JPEG format image files, and further allows the export of the filtered dataset with normalized ratios, as well as all analysis parameters as text files. There is no real limit to the size of dataset to be loaded in MArray. A paired experiment with 12 000 genes in each, the data loading was finished in approximately 120 s and the rest of computations were completed in about 60 s on a desktop workstation (733 MHz, Pentium III, Windows NT 4.0). MArray is a work in progress and future work will include applying visualization methods capable of assessing the repeatability of more than three microarray experiments, as well as filtering procedures based on repeatedly spotted clones (Jenssen *et al.*, 2002) and a web application of MArray.

ACKNOWLEDGEMENTS

This work was supported by the Norwegian Cancer Society (http://www.kreft.no).

SUPPLEMENTARY MATERIAL

For Supplementary Material, please refer to *Bioinformatics* Online.

REFERENCES

Altman,D.G. (1991) *Practical Statistics for Medical Research*. Chapman & Hall, London.

Bland,J.M. and Altman,D.G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 307–310.

Brown,C.S., Goodwin,P.C. and Sorger,P.K. (2001) Image metrics in the statistical analysis of DNA microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 8944–8949.

Dysvik,B. and Jonassen,I. (2001) J-Express: exploring gene expression data using Java. *Bioinformatics*, **17**, 369–370.

Jackson,J.E. (1980) *A User's Guide to Principal Components*. John Wiley, New York.

Jenssen,T.K., Langaas,M., Winston,P.Kuo, Brigitte,Smith-Sorensen, Ola,Myklebost and Eivind,Hovig (2002) Analysis of repeatability in spotted cDNA microarrays. *Nucleic Acids Res.*, submitted

Kerr,M.K. and Churchill,G.A. (2001) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.

Lee,M.L., Frank,C.Kuo, Whitemore,G.A. and Sklar,Jeffery (2000) Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9839.

Manchester, (2001) Bioinformatis, MaxdView. http://bioinf.man.ac.uk/microarray/maxd/maxdView/index.html

Munson,P.J., Prabhu,V.V. and Young,L. (2001) F-scan: fluorescently probed cDNA microarray analysis. http://abs.cit.nih.gov/fscan.

NIH, (2000) BioInformatics BIMAS/CBEL/CIT, mAdb UniGene search. http://nciarray.nci.nih.gov/search_UG.shtml

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wang,X.J., Ghosh,S. and Guo,S.W. (2001) Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res.*, **29**, 75e.

Yang,Y.H., Dudoit,S., Luu,S. and Terry,S. (2001) *Normalization for cDNA Microarray Data*, SPIE BIOS 2001, San Jose, California.