Infection, Genetics and Evolution xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

A python module to normalize microarray data by the quantile adjustment method

Ibrahima Baber [a,*], Jean Philippe Tamby [b], Nicholas C. Manoukis [c], Djibril Sangaré [a], Seydou Doumbia [a], Sekou F. Traoré [a], Mohamed S. Maiga [d], Doulaye Dembélé [e]

[a] Malaria Research and Training Center (MRTC), Faculté de Médecine, de Pharmacie et d'Odontostomatologie, Université de Bamako, Mali
[b] Unité de Recherche en Génomique Végétale (URGV)-UMR INRA 1165-UEVE, ERL CNRS 8196, 2 Rue Gaston Crémieux, F-91057 Evry Cedex, France
[c] Section of Vector Biology, Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892-8132, USA
[d] Faculté des Sciences et Techniques, Université de Bamako, Mali
[e] Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS, UdS, 67404 Illkirch Cedex, France

ARTICLE INFO

Article history:
Received 30 April 2010
Received in revised form 17 September 2010
Accepted 10 October 2010
Available online xxx

Keywords:
Module
Quantile method
Python
Microarray
Normalization

ABSTRACT

Microarray technology is widely used for gene expression research targeting the development of new drug treatments. In the case of a two-color microarray, the process starts with labeling DNA samples with fluorescent markers (cyanine 635 or Cy5 and cyanine 532 or Cy3), then mixing and hybridizing them on a chemically treated glass printed with probes, or fragments of genes. The level of hybridization between a strand of labeled DNA and a probe present on the array is measured by scanning the fluorescence of spots in order to quantify the expression based on the quality and number of pixels for each spot. The intensity data generated from these scans are subject to errors due to differences in fluorescence efficiency between Cy5 and Cy3, as well as variation in human handling and quality of the sample. Consequently, data have to be normalized to correct for variations which are not related to the biological phenomena under investigation. Among many existing normalization procedures, we have implemented the quantile adjustment method using the python computer language, and produced a module which can be run via an HTML dynamic form. This module is composed of different functions for data files reading, intensity and ratio computations and visualization. The current version of the HTML form allows the user to visualize the data before and after normalization. It also gives the option to subtract background noise before normalizing the data. The output results of this module are in agreement with the results of other normalization tools.

Published by Elsevier B.V.

1. Introduction

The two-color microarray technique involves the immobilization of thousands of DNA fragments (probes) on a substrate, generally a ~6 cm^2 glass cover slip. Sample RNAs extracted from test and control tissue sources are converted to cDNA and labelled with two fluorescent markers, a green Cy3 and a red Cy5 cyanine. Then the labelled cDNAs are hybridized with the probes on the microarray (Schena et al., 1995). Gene expression levels are quantified by a scanner which estimates the fluorescence intensity in two wavelengths for each spot. The scanner computes the signal and background values for each spot via a dedicated algorithm (Yang et al., 2002). The reliability of these values can be affected by technical factors. The aim of data normalization is to adjust variations due to these factors in order to obtain reliable results that reflect real gene expression (Smyth and Speed, 2003).

There are two types of normalization methods: linear and nonlinear. Each of these methods may be applied at local or global level. In the linear method, the normalization consists of finding a proportional constant between data from the Cy5 and Cy3 dyes. This can be done using a set of probes for known or housekeeping genes on the array or by using data from all probes. Nonlinear approaches are more general and considered to be more powerful. LOWESS, based on local regression (Cleveland, 1979), and quantile adjustment methods are the

Abbreviations: B, background; F, foreground; CGI, Common Gateway Interface; MySQL, my structured query language; MIDAS, microarray data analysis system; LOWESS, local weighted Scatter plot smoothing.

* Corresponding author at: Malaria Research and Training Center (MRTC), Faculté de Médecine, de Pharmacie et d'Odontostomatologie, Université de Bamako, Mali B.P.1805, Bamako, Mali (West Africa). Tel.: +223 20 22 52 77; fax: +223 20 22 49 87.
E-mail addresses: baber@icermali.org (I. Baber), tamby@evry.inra.fr (J.P. Tamby), manoukisn@niaid.nih.gov (N.C. Manoukis), dsangare@icermali.org (D. Sangaré), sdoumbi@icermali.org (S. Doumbia), cheick@icermali.org (S.F. Traoré), mohmaiga@ml.refer.org (M.S. Maiga), doulaye@igbmc.fr (D. Dembélé).

1567-1348/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.meegid.2010.10.008

most widely used statistical methods for microarray data normalization (Fujita et al., 2006; Saviozzi et al., 2006; Dabney and Storey, 2007; Oshlack et al., 2007). However, the quantile method has the advantage to be quicker and simpler than the LOWESS. The algorithm of the quantile adjustment method was described by Bolstad et al. (2003).

Our module implements the quantile algorithm (*read the manual for more details*). It involves sorting intensity values before doing the calculation of their mean for each rank. After that, means are redistributed according the ranks of the initial list of values. An important advantage of the quantile method is that it does not require parameter tuning like more complex methods such as LOWESS.

There are several existing tools for microarray analysis, some of which are commercial. Most of these do not have available the source codes for modification and customization for special situations. Due to this limitation, it is important for biologists to contribute to the development of more flexible tools. In this spirit, we decided to develop a normalization module with source code freely available to anyone who wants to modify it in order to fit his/her own needs. Specifically, we have implemented a Python (Guido van Rossum, 2005) module for microarray data normalization using the quantile adjustment method which can be run via a web interface. As far as we know, there is no module for quantile adjustment normalization available in the biopython library; our attempt tries to fill this lack. We chose Python due to its clean syntax, easy typing and many available libraries which are well adapted to the problem we are addressing.

2. Method

The module is composed of several functions which are processing all calculations for data normalization and for graphical view. There is also a CGI (Common Gateway Interface) script which provides access to the normalization calculation via a web server where the module is installed. An HTML form that allows the user to submit data and the server sends back the result to the user is also included. In summary, the implemented module is capable of:

- read a data file with .GPR extension → input
- compute normalization according to quantile method
- view results as a graphically and as text → output
- present web interface for option choice

2.1. Input file

Currently accepted input file of our implementation is the .GPR (GenePix Results) (in Molecular Devices, 2010). This kind of file has a header comment which includes experiment date, description of the scanner parameters and the type of experiment. Our program analyzes only the data of signal and background.

2.2. Web interface form

Fig. 1 shows the form which allows a user to submit a data file (.gpr extension, GenePix) to the program. A user can select the signal of his preference (Median or Mean) before he or she proceeds with the normalization by checking a button in the second column. The box corresponding to subtract background can be checked if the user decides to subtract the background value from the signal value (Fig. 1). A CGI script is included for interaction between the module source code and the web interface.

From this form, the input file is uploaded from the field File Name, and then a user can select median or mean data. He can check one of these boxes (before or after normalization) to view data plots.

2.3. Output file

The output data of the program has four columns. It can contain from 20 to about 50,000 lines. A graphical review of the data is produced by plotting ratios against intensities.

In the output file, the first and second columns are normalized signal values (Cy5 and Cy3). Those data have been used to get ratio and intensity values. The prefix (Norm_) of columns names means that Ratio and intensity values (log 2) have been computed after data normalization. In the case where the program is executed with normalization option checked, in the output file, the prefix preceding intensity and ratio.

2.4. Plot view of ratio against intensity

Fig. 2 shows the plot of ratio and intensity before and after normalization.

Fig. 2(a) is graphical view of data (median values) before normalization. The ratio values are not symmetric to the zero ratio line while this is the case for the second plot (Fig. 2(b)). From Fig. 2(b): (i) if the ratio log value is greater than a given threshold,

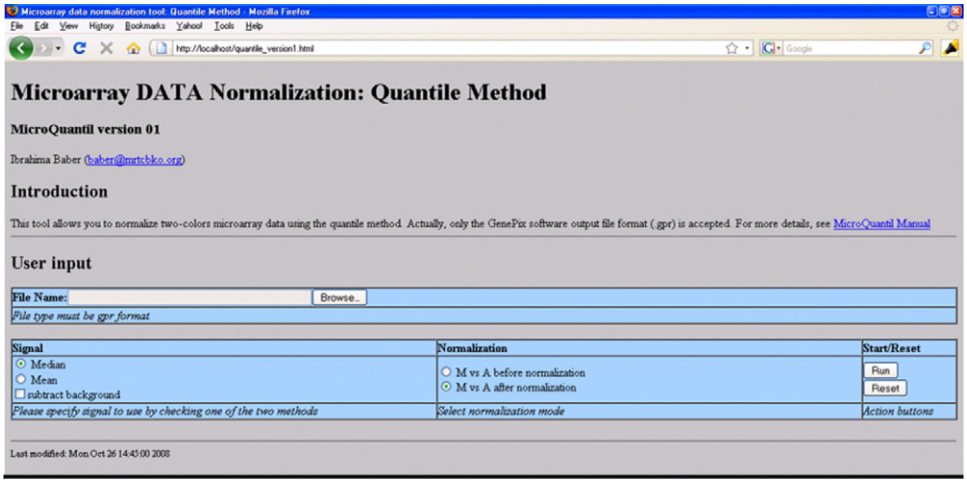

Fig. 1. Data submission form.

Please cite this article in press as: Baber, I., et al., A python module to normalize microarray data by the quantile adjustment method. Infect. Genet. Evol. (2010), doi:10.1016/j.meegid.2010.10.008

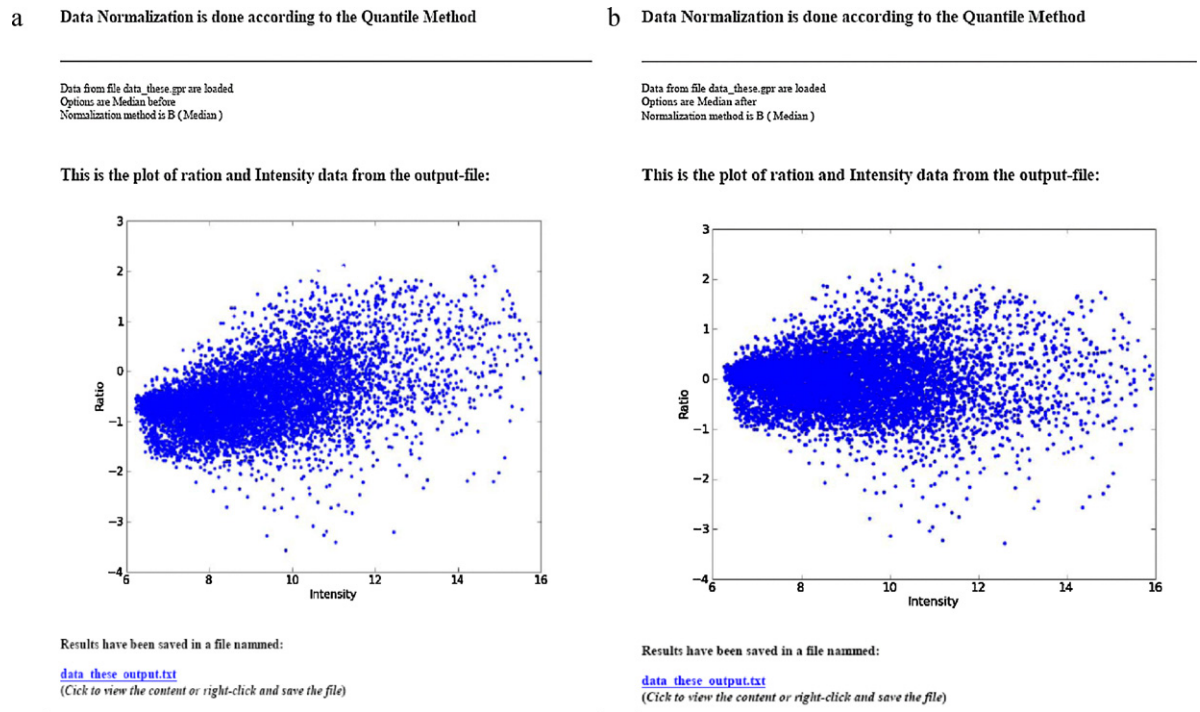

Fig. 2. Graphical view of genes repartition according to ratio values.

e.g. 1, corresponding genes are up-regulated. (ii) If ratio log value is less than a given threshold, e.g. −1, corresponding genes are down-regulated. (iii) If the log ratio is equal to 0, the expression level of genes did not change. All genes between the upper and lower threshold (around the $M = 0$ line) are not changed ones.

Data generated by the module are similar to those obtained with standard software named Elea (Microarray & Sequencing Platform, 2009). For error quantification between values obtained from the two methods (our module versus Elea), we have calculated Mean Square Error (MSE) of couple values. The results are respectively 1.9985E−5 and 8.2341E−6 for ratios and mean intensities (Fig. 3). These error values are quasi null, validating the results of our module.

Fig. 3(right) shows that most of the ratio data points are along the diagonal of the scatterplot, again illustrating that the results obtained from both two tools are in good agreement. All intensity data points are likewise along the diagonal of the scatterplot (Fig. 3(left)).

3. Discussion

Normalizing microarray data is required to adjust for variations which are not related to the biological phenomena under study (Knudsen, 2004). There are many tools for microarray data normalization based on the two principal methods, the quantile method and LOWESS. One good example comes from a web site named MIDAS (microarray data analysis system) accessible via its URL (TM4-MIDAS, 2005), where there are efficient tools for microarray data analysis, including normalization, clustering and functional database search modules.

In spite of the availability of those tools which include graphical interfaces, and executables run on command line such as the

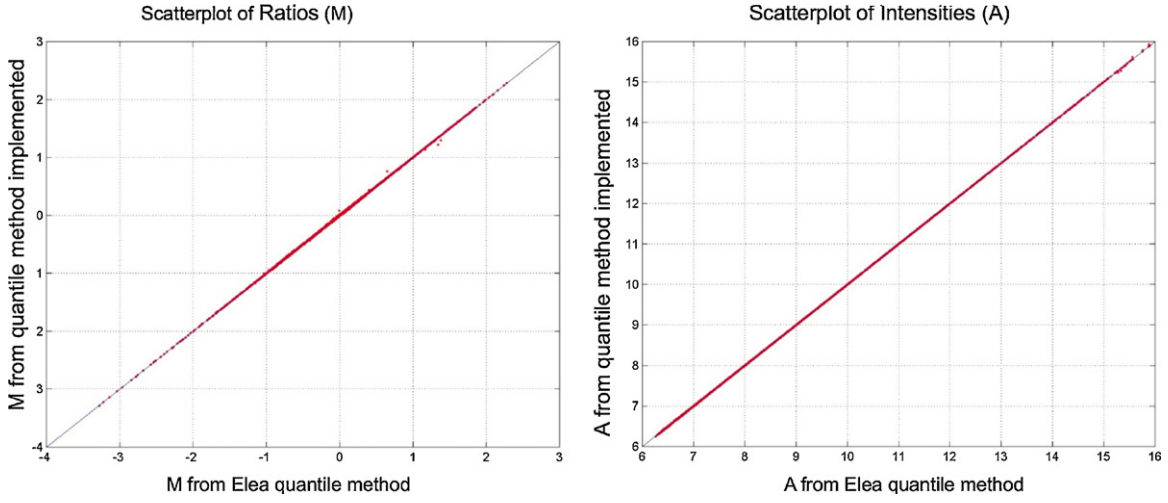

Fig. 3. Scatterplot of ratio M made with results of our module and with Elea quantile method result (left). Scatterplot of intensity A made with results of our module and with Elea quantile method result (right).

Please cite this article in press as: Baber, I., et al., A python module to normalize microarray data by the quantile adjustment method. Infect. Genet. Evol. (2010), doi:10.1016/j.meegid.2010.10.008

limma library in Bioconductor (Gentleman et al., 2004), there is a real need for software which may be adapted to users's options. Tools which can be run via Internet are largely used nowadays (Oliveros, 2008). For these reasons we developed this normalization module with an HTML interface module.

The main limit of online normalization tools is network bandwidth. Uploading microarray data files, which are often large, requires significant bandwidth to be time-efficient.

In future versions of our module a MySQL database will be included to improve its handling of submitted data management. In this case, if the user sends consecutively the same data file to a server, the file name will be saved, and the module will be able send back to the user the result of his first submission.

Authors' contributions

IB implemented the module and wrote the manuscript. DD has designed the project and reviewed the manuscript. JPT participated in implementing the quantile module and in reviewing the manuscript. NCM participated in web dynamic form design and in editing the manuscript. DS, SD, SFT, MSM participated to the manuscript review.

Availability and requirements

The code source of this module may be obtained from: http://taylor0.biology.ucla.edu/~manoukis/Quantile.tar.gz.

The installation of this module requires having python and several libraries (Numpy, matplotlib) available, in addition to a web server such as Karrigell and Apache (more details are provided in the program download).

Acknowledgements

This research was supported in part by the Intramural Research Program of the NIH, NIAID. We thanks Dr. Olivier Gascuel and Dr. Chantal Pacteau from CNRS, France for their help in training grant funding, and Dr Cathrine Letondal from Pasteur Institute, Paris, France for training in python. I would like to thank the NIH, USA and TOKTEN (Transfer of Knowledge through Expatriate Nationals), Mali for technical support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2010.10.008.

References

Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19 (2), 185–193.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. 74, 829–836.

Dabney, A.R., Storey, J.D., 2007. Normalization of two-channel microarrays accounting for experimental design and intensity-dependent relationships. Genome Biol. 8 (3), R44.

Fujita, A., Sato, J.R., Rodrigues Lde, O., Ferreira, C.E., Sogayar, M.C., 2006. Evaluating different methods of microarray data normalization. BMC Bioinform. 7 (October 23), 469.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., Zhang, J., 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5 (10), R80.

Guido van Rossum: Tutoriel Python, Release 2.4.1.: Traduction française dirigée par Olivier Berger, Mise à jour par Henri Garreta. Python Software Foundation, <docs@python.org>, September 24, 2005. <ftp://ftp-developpez.com/python/cours/TutoVanRossum/fichiers/TutorielPython.pdf> (15.07.06).

Knudsen, S., 2004. Guide to Analysis of DNA Microarray Data, 2nd edition. Wiley Liss, ISBN: 184 pp.

Microarray & Sequencing Platform: «IGBMC, Strasbourg, France. IGBMC Microarray and Sequencing Platform».<http://www-microarrays.u-strasbg.fr/base.php?page=analysisExpressionNormEleaE.php> (10.06.09).

Molecular Devices, Inc.©2010. Produced in the US. <http://www.moleculardevices.com/pages/software/gn_genepix_file_formats.html#gpr> (10.05.10).

Oliveros, J.C., 2008. GPR Normalizer. An Interactive Server for Normalizing Standard GenePix GPR files. <http://bioinfogp.cnb.csic.es/tools/normalize_gpr> (6.01.06).

Oshlack, A., Emslie, D., Corcoran, L.M., Smyth, G.K., 2007. Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. Genome Biol. 8 (1), R2.

Saviozzi, S., Cordero, F., Lo Iacono, M., Novello, S., Scagliotti, G.V., Calogero, R.A., 2006. Selection of suitable reference genes for accurate normalization of gene expression profile studies in non-small cell lung cancer. BMC Cancer 6 (July 26), 200.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270 (5235 (October 20)), 467–470.

Smyth, G.K., Speed, T.P., 2003. Normalization of cDNA microarray data. Methods 31, 265–273.

TM4-MIDAS 2005, Dana-Farber Cancer Institute, 44 Binney St., Boston, MA, USA. <http://www.tm4.org/midas.html> (17.05.10).

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., Speed, T.P., 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30 (4), e15.