BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 8 2005, pages 1724–1726
doi:10.1093/bioinformatics/bti199

Gene expression

OLIN: optimized normalization, visualization and quality testing of two-channel microarray data

Matthias E. Futschik[1,2,*] and Toni Crompton[3]

[1]Institute for Theoretical Biology, Humboldt-Universität, Invalidenstrasse 43, 10115 Berlin, Germany, [2]Department of Information Science, PO Box 56 and [3]Otago School of Medical Sciences, Division of Health Science, PO Box 913, University of Otago, Dunedin, New Zealand

Received on July 31, 2004; revised on November 8, 2004; accepted on November 30, 2004
Advance Access publication December 7, 2004

ABSTRACT

Summary: Microarray data are generated in complex experiments and frequently compromised by a variety of systematic errors. Subsequent data normalization aims to correct these errors. Although several normalization methods have recently been proposed, they frequently fail to account for the variability of systematic errors within and between microarray experiments. However, optimal adjustment of normalization procedures to the underlying data structure is crucial for the efficiency of normalization. To overcome this restriction of current methods, we have developed two normalization schemes based on iterative local regression combined with model selection. The schemes have been demonstrated to improve considerably the quality of normalization. They are implemented in a freely available R package. Additionally, functions for visualization and detection of systematic errors in microarray data have been incorporated in the software package. A graphical user interface is also available.

Availability: The R package can be downloaded from http://itb.biologie.hu-berlin.de/~futschik/software/R/OLIN. It underlies the GPL version 2.

Contact: m.futschik@biologie.hu-berlin.de

Supplementary information: Further information about the methods used in the OLIN software package can be found at http://itb.biologie.hu-berlin.de/~futschik/software/R/OLIN

INTRODUCTION

Microarray measurements are affected by a variety of systematic experimental errors limiting the accuracy of data produced. Such errors have to be identified and removed before further data analysis is conducted. Although this so-called normalization procedure is only an intermediate step in the analysis, it has considerable impact on the results of follow-up analysis (Hoffmann *et al*., 2002). Assessment of the efficiency of a chosen normalization method should therefore be an integral part of every normalization procedure.

A popular class of normalization methods for two-channel microarrays is based on local regression. They have become the standard approach for many researchers, as they are flexible and easy to use, and have been implemented in numerous freely available or commercial microarray data-analysis systems (Holloway *et al*., 2002; Quackenbush, 2002; Yang *et al*., 2002). However, one unresolved challenge in using local regression methods has been the choice of regression parameters. This has commonly been left to the user with only default values given. Instructions on how to adjust the parameters to the underlying data structure are generally not given. In a previous analysis, however, we have shown that the use of such default parameters can severely compromise the quality of normalized data (Futschik and Crompton, 2004). Therefore, an optimization of the model parameters is required to ensure high efficiency of normalization.

In order to overcome the limitations of current methods, we introduced two normalization schemes based on iterative local regression and model selection. Both normalization schemes aim to correct intensity- and location-dependent dye bias in the two-channel microarray data. Extensive comparison of normalization efficiencies have shown that these schemes can considerably reduce systematic errors in microarray data, increase the overall consistency within experiments and also improve the correlation of microarray measurements with actual biological changes in the expression. For the detailed comparison and discussion of the normalization schemes and the underlying concepts, the reader is referred to the study by Futschik and Crompton (2004). Here, we present the software package including these schemes and further functions for analyzing microarray data.

ALGORITHM

The main algorithm developed for optimized normalization is Optimized Local Intensity-dependent Normalization (OLIN). It is based on iterative local regression of spots' logged intensity ratios regarding spot intensity and location with a subsequent correction of the dye bias. For local regression, the LOCFIT algorithm is used, as it offers more flexibility than the commonly used lowess method (Loader, 1999). Model parameters are optimized in each regression step by generalized cross-validation, which considerably reduces the computational costs compared to standard cross-validation. This is crucial, as the optimization task would not be feasible applying conventional cross-validation procedures due to the large number of spots on the arrays. A detailed description of the OLIN algorithm can be found in the study by Futschik and Crompton (2004) or on the OLIN Web page.

The second normalization algorithm developed is Optimized Scaled Local Intensity-dependent Normalization (OSLIN). Basically, it comprises the OLIN procedure with a subsequent optimized scaling of the range of logged intensity ratios across the spatial array

*To whom correspondence should be addressed.

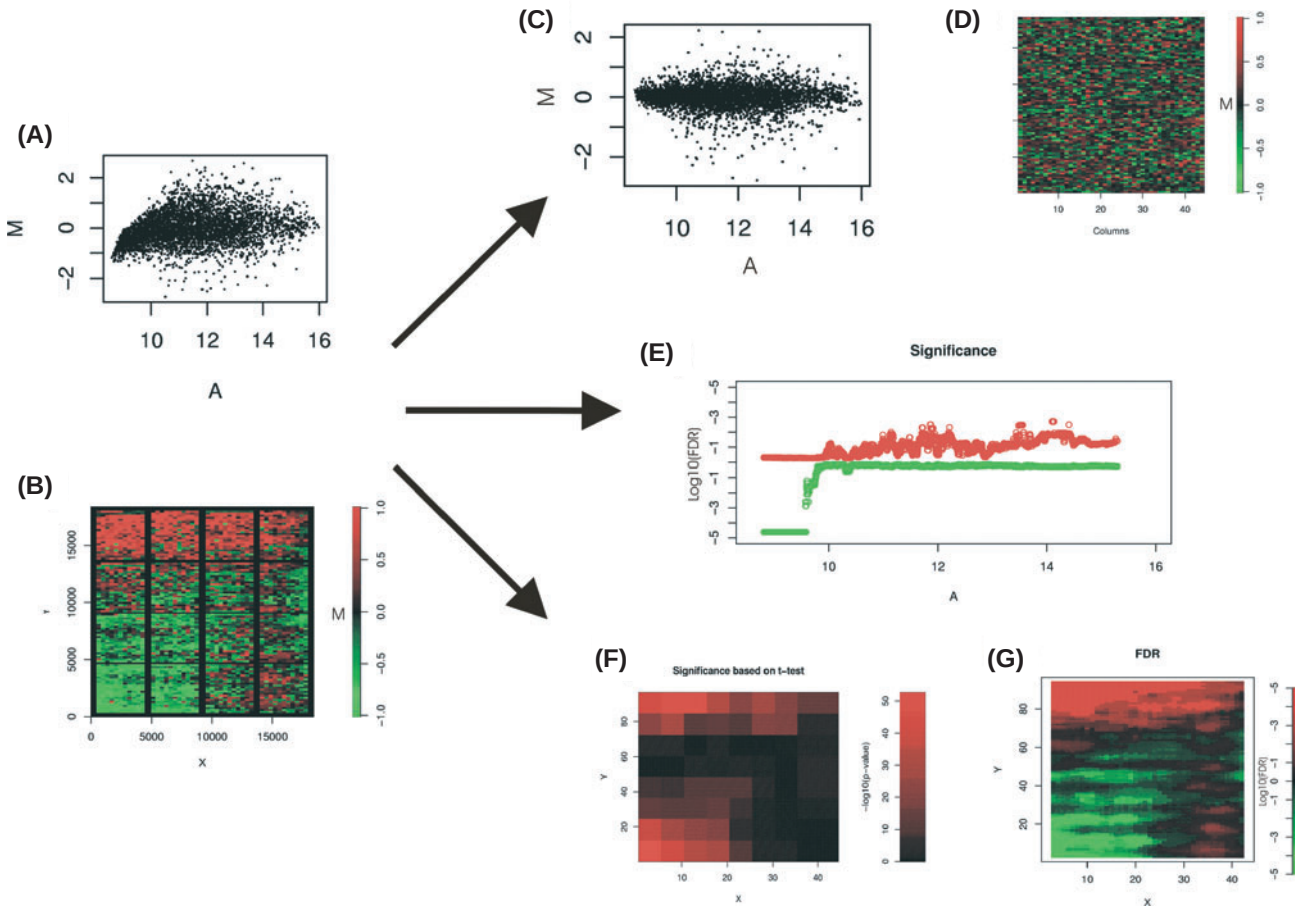

Fig. 1. Optimized normalization, visualization and detection of systematic errors using the OLIN package. Illustrated for the array 1 of the SW480/620 microarray experiment (Futschik *et al*., 2002). (**A**) MA-plot of raw data: logged intensity ratios $M = (\log_2 \text{Cy5} - \log_2 \text{Cy3})$ with respect to average logged spot intensity $A = 0.5 * (\log_2 \text{Cy5} + \log_2 \text{Cy3})$. (**B**) MXY-plot (produced by OLIN function *mxy2.plot*) of raw data: M with respect to X- and Y-coordinate of spots. (**C** and **D**) MA- and MXY-plot (*mxy.plot*) of normalized data by OSLIN. (**E**) Significance of intensity-dependent dye bias (indicated by false discovery rate) for spot neighborhoods as detected by *fdr.int* and visualized by *sigint.plot.* (**F**) Visualization of one-factorial ANOVA model assessing spatial bias (*anovaspatial*). (**G**) Two-dimensional plot of significance of spatial dye bias detected by one-sided random permutation test (*fdr.spatial*, *sigxy.plot*).

dimensions. Both OLIN and OSLIN assume that most genes on the array are not differentially expressed or that up- and down-regulation is balanced across the spot intensity range. Additionally, random spotting is required. Although these assumptions are frequently fulfilled in genome-wide microarray experiments, they should be carefully checked.

IMPLEMENTATION

The OLIN/OSLIN algorithm is implemented in the R language (Ihaka and Gentleman, 1996). Together with various other R functions, they form the OLIN package. It is placed within the Bioconductor framework (Gentleman *et al*., 2004) allowing a convenient integration with other data analysis tools. To increase its user-friendliness, most functions of the OLIN package can be accessed via a graphical user interface (*OLINgui*) based on TclTk widgets.

Additional functions in the OLIN package serve for visualization and quality testing of (normalized) microarray data. These functions can be used to assess stringently the efficiency in removing systematic errors as well as to identify possible problems in the experimental protocol. Several distinct statistical tests were implemented to detect and localize plate-, pin-, intensity- and location-dependent systematic errors: (1) F-tests based on one-factorial ANOVA models, (2) one-sided random permutation tests (based on sampling with replacement) and (3) one-sided random permutation tests (based on sampling without replacement) detecting spot neighborhoods affected by experimental bias.

Significance of (local) systematic errors is indicated either by (adjusted) P-values or by false discovery rates. In contrast to global assessment methods such as correlation measures, the implemented tests allow a stringent identification of regions of significant bias in microarray data. This feature can be especially valuable for a rapid detection of artifacts and may assist in the improvement of experimental protocols. Examples for optimized normalization, visualization and error detection are shown in Figure 1.

Besides within-array normalization performed by the OLIN algorithm, the R package also includes methods for between-array normalization.

CONCLUSIONS

We developed an R package for optimized normalization and quality testing of microarray data. Optimized normalization was demonstrated to significantly improve the data quality and, thus, to support the validity of results derived in follow-up gene expression analysis (Futschik and Crompton, 2004). Emphasis was put on the integration of methods for the detection and visualization of systematic errors.

Whereas the OLIN procedure implemented here aims primarily to correct intensity- and location-dependent dye bias, the basic iterative procedure incorporating model selection should be easily adapted to the correction of other types of systematic errors. To support the implementation of optimized normalization in other software, the complete source code is included in the freely available R package.

Finally, the methods implemented in the OLIN package are not restricted to two-channel microarrays, but can be applied to other array platforms as well. We believe therefore that they will be of general use to many researchers using array technologies.

ACKNOWLEDGEMENTS

We thank Simone Monreal for critical reading of the manuscript and Anna Tschaut for constructive support of the research project. We also like to thank Aaron Jeffs and Sharon Pattison for providing the data that the original studies were based on, Koen Bossers for suggestions helping to improve the software and the reviewer for constructive comments on the manuscript. M.F. would like to thank Hanspeter Herzel, for current support and his former PhD supervisors Nik Kasabov, Michael Sullivan, Parry Guilforld and Anthony Reeve for previous support of the research project. M.F. was supported by a PhD scholarship from the University of Otago.

REFERENCES

Futschik,M. and Crompton,T. (2004) Model selection and efficiency testing for normalisation of cDNA microarray data. *Genome Biol.*, **5**, R60.

Futschik,M., Jeffs,A., Pattison,S., Kasabov,N., Sullivan,M. and Reeve,A. (2002) Gene expression profiling of metastatic and nonmetastatic colorectal cancer cell lines. *Genome Lett.*, **1**, 26–33.

Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Hoffmann,R., Seidl,T. and Dugas,M. (2002) Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol.*, **3**, research0033.1–research0033.11.

Holloway,A.J., van Laar,R.K., Tothill,R.W. and Bowtell,D. (2002) Options available-from start to finish-for obtaining data from DNA microarrays II. *Nat. Genet.*, **32** (Suppl. 2), 481–489.

Ihaka,R. and Gentleman,R. (1996) R: a language of data analysis and graphics. *J. Comput. Graphic Stat.*, **5**, 299–314.

Loader,C. (1999) *Local Regression and Likelihood*. Springer, NY.

Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32** (Suppl. 2), 496–501.

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