BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 7 2003, pages 893–894
DOI: 10.1093/bioinformatics/btg085

MARAN: normalizing micro-array data

Kristof Engelen, Bert Coessens, Kathleen Marchal and Bart De Moor*

ESAT-SCD, KULeuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

Received on September 3, 2002; revised on November 8, 2002; accepted on November 22, 2002

ABSTRACT

Summary: MARAN is a web-based application for normalizing microarray data. MARAN comprises a generic ANOVA model, an option for Loess fitting prior to ANOVA analysis, and a module for selecting genes with significantly changing expression.

Availability: http://www.esat.kuleuven.ac.be/maran/

Contact: kristof.engelen@esat.kuleuven.ac.be

INTRODUCTION

The use of ANOVA (ANalysis Of VAriance) for normalizing microarray data is increasingly gaining interest (Jin *et al.*, 2001; Kerr *et al.*, 2000). A major advantage of this approach is that different sources of variation are assessed at once across the entire experiment. Moreover, residuals obtained from fitting the model provide a means for statistical analysis of the results, e.g. selecting genes with significantly changing expression. MARAN is a user-friendly web-based application for normalizing microarray data. The normalization procedure implemented in MARAN consists of a generic ANOVA model; it is readily applicable to any type of experimental design. Additional functionalities are plots for assessing the appropriateness of the model, an option for Loess fitting (Yang *et al.*, 2002) the data in case of severe non-linearities in the data set, and a module for selecting genes with significantly changing expression.

PREPROCESSING A DATA SET

Modeling the data

Use of ANOVA for microarray normalization basically comes down to modeling the measured expression level of each gene as a linear combination of the major sources of variation (i.e. explanatory variables or effects), such as the array or dye for which the measurement was taken. An advantage of the model implemented in MARAN is its generic nature with respect to the experimental design, i.e. it can be used to normalize any type of microarray design in a single run. The major effects included in the model are *batch*, *dye*, *array*, *pin* and *array* × *dye*. A *batch* is a collection of slides which contain the same set of genes, representative for part of the genome. This effect needs to be taken into account when the entire set of genes is too large to be spotted on a single array. The collection of arrays on which the same set of genes was spotted constitutes a 'batch'. The *dye* effect models the difference in measured intensities between the red and green dye; the *array* effect compensates for global intensity differences between arrays. Likewise, a set of measurements that share the same *pin* effect, were spotted by the same spotting pin. The *array* × *dye* interaction effect is used, instead of a *condition* effect, for alleviating any condition-dependent variations in the measured intensities. Both effects are confounded and using a *condition* effect would render the analytical solutions of the model fit dependent on the experimental design.

Apart from these global effects, a *gene* and *expression* effect have been included as well. The *gene* effect normalizes each gene with respect to its basal expression level; *expression* refers to the effect of interest, i.e. the condition-affected change in intensity for each gene.

Modeling the data is fairly straightforward and user-friendly. Any explanatory variable, that is not relevant for a specific experimental design, is automatically discarded. All other effects (except *expression*) can be included or excluded by clicking the corresponding checkboxes on the 'Modeling' page.

While it is technically possible to normalize any type of experiment design with our model, we would suggest to make sure each gene is measured at least twice in every condition, regardless of the other parameters. The underlying reasons are explained in detail in the 'User Guide' section of the website.

Interpretation of the results

After completion of the analysis, normalized expression values (and all parameters and residuals of the fitted model) can be downloaded from the 'Results' page. Also on this page, an ANOVA-table, for interpreting the different effects and their contribution to the total amount of variation, and several plots for analyzing the ANOVA

*To whom correspondence should be addressed.

modeling assumptions, are included. These assumptions are 2-fold: firstly, the data should be adequately described by a linear model. Secondly, the error terms are assumed to be normally distributed with mean zero and constant variance. Information about the heteroscedasticity (non-constant error variance) and normality of the residual distribution can be obtained from the 'Global residual plot' and the 'NQ plot' respectively. Serious heteroscedastic features should be avoided when using the residual distribution for selecting genes with significantly changing expression. It should be noted, however, that deviations from normality, in the form of widened tails, can often be acceptable due to the small amount of data points compared to the number of parameters to be estimated. As explained later on, bootstrap methods are advisable for selecting genes with significantly changing expression when serious heteroscedasticity or non-normality occurs in the residual distribution.

More problematic, however, is an apparent heteroscedasticity caused by a superposition of non-linear trends in the residuals for each combination of major effects, indicating that a linear model is not adequate for describing the data (i.e. the first assumption is not satisfied). All other plots are residual plots for each specific array–dye combination. When obvious curvilinear trends are observed on these plots (Marchal *et al.*, 2002), remedial measures should be taken, as described below.

Remedial measure for non-linear residual trends

A well-established remedial measure for removing non-linear effects in the data set has previously been described by Yang *et al.* (2002). This Loess-based method has been made available in the MARAN web application ('Loess' page). The 'Loess' page can be accessed directly or after inspecting the results of an initial fit. It is important to keep in mind that a Loess-fit based correction for non-linearities is performed in the dye direction, separately for each array. This implies that, for complex experimental designs, non-linearities across arrays or conditions cannot be completely alleviated.

Filtering the results

As mentioned above, the obtained estimates of the error terms can be used for various statistical analysis concerning the ANOVA parameters. Two different methods for selecting genes with significantly changing expression have been made available on the website.

The first method is valid under the assumption of normally distributed error terms, with mean zero and constant error variance. A selection of genes can be obtained by entering a preferred significance or, when desired, *p*-values for all genes can be downloaded.

The alternative method is based on a bootstrap procedure (Efron, 1979). It is a 'fixed predictor sampling method', similar to the one described by Kerr *et al.* (2000) and is appropriate when the residuals show serious deviations from normality, but no apparent heteroscedasticity is present. A selection of genes can be obtained by entering a preferred significance.

Further analysis

The complete results file, or a selection of genes obtained after filtering, can be used for further analysis using a direct link to the INCLUSive website (INCLUSive is a web-based application for integrated clustering, upstream sequence retrieval and regulatory motif detection; Thijs *et al.*, 2002).

ACKNOWLEDGEMENTS

K. E. is a research assistant of the IWT; B. C. is a research assistant at the KULeuven; K.M. is a post-doctoral researcher of the FWO; Professor B.D.M. is full professor at the KULeuven. We would like to thank G. Thijs, F. De Smet, J. De Brabanter and P. Van Hummelen for their useful comments. This work is partially supported by: (1) IWT projects: STWW-Genprom 980396, GBOU SQUAD; (2) Research Council KULeuven: GOA Mefisto-666; (3) FWO projects: G.60413.03, G.0115.01; (4) DWTC (IUAP IV-02 (1996-2001) and IUAP V-22 (2002-2006); (5) IDO (IOTA Oncology, Genetic networks); (6) Flanders Interuniversity Institute of Biotechnology (VIB).

REFERENCES

Efron,B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1–26.

Jin,W., Riley,R.M., Wolfinger,R.D., White,K.P., Passador-Gurgel,G. and Gibson,G. (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genet.*, **29**, 389–395.

Kerr,M.K., Martin,M. and Churchill,G.A. (2000) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *J. Comput. Biol.*, **7**, 819–837.

Marchal,K., Engelen,K., DeBrabanter,J., Aerts,S., DeMoor,B., Ayoubi,T. and Van Hummelen,P. (2002) Comparison of different methodologies to identify differentially expressed genes in two-sample cDNA microarrays. *J. Biol. Systems*, in press.

Thijs,G., Moreau,Y., DeSmet,F., Mathys,J., Lescot,M., Rombauts,S., Rouze,P., DeMoor,B. and Marchal,K. (2002) Inclusive: integrated clustering, upstream sequence retrieval and motif sampling. *Bioinformatics*, **18**, 331–332.

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