Bioinformatics Advance Access published May 7, 2007

Application Note

CALIB: a Bioconductor package for estimating absolute expression levels from two-color microarray data

Hui Zhao [1], Kristof Engelen [1], Bart De Moor [2] and Kathleen Marchal [1,2 *]

[1] CMPG-BioI, Department of Microbial and Molecular Systems, K.U.Leuven, Kasteelpark Arenberg 20, B-3001 Leuven-Hevelee, Belgium

[2] BIOI@SCD, Department of Electrical Engineering (ESAT), K.U.Leuven, Kasteelpark Arenberg 10, B-3001 Leuven-Hevelee, Belgium

Associate Editor: Dr. Joaquin Dopazo

ABSTRACT

In this paper we describe a new Bioconductor package "CALIB" for normalization of two-color microarray data. This approach is based on the measurements of external controls and estimates an absolute target level for each gene and condition pair, as opposed to working with log-ratios as a relative measure of expression. Moreover, this method makes no assumptions regarding the distribution of gene expression divergence.

Availability: http://bioconductor.org/packages/2.0/bioc Open Source

1 INTRODUCTION

Normalization of microarray measurements is the first step in a microarray analysis flow. It aims at removing consistent sources of variations to make measurements mutually comparable. Reliable normalization is essential since the results of all subsequent analyses, such as clustering, might largely be influenced by the normalization procedure. For normalization of two-color arrays different methods have been described. Although some approaches inherently work with absolute intensities (e.g. ANOVA (Kerr et al., 2000)), in general, preprocessing of two-color microarrays largely depends on the calculation of the log-ratios of the measured intensities. A common normalization step is the removal of the nonlinear intensity-dependent discrepancy between Cy3 and Cy5 intensities (e.g. loess (Yang et al., 2002)). This normalization step assumes the distribution of gene expression to be balanced and showing little change between the biological samples tested, an assumption which is referred to as Global Normalization Assumption or GNA. Global mRNA changes that result in an uneven distribution of expression changes, however, have been shown to occur more frequently than currently believed (van Bakel and Holstege, 2004; van de Peppel et al., 2003), and could have a significant impact on the interpretation of data normalized according to the Global Normalization Assumption.

Recently, a different way of normalizing two-color microarray data was proposed that avoids the GNA and poses several advantages over ratio based approaches (Engelen et al., 2006). Briefly, the normalization is based on a physically motivated model, explicitly modeling the hybridization of transcript targets to their corresponding DNA probes, and the relation between the measured fluorescence and the amount of hybridized, labeled target. The parameters of this model and incorporated error distributions are estimated from external control spikes: targets that are added to the hybridization solution in known concentrations. This, together with the inherent nonlinearity of the model, allows for normalizing the data without making any assumptions on the distribution of gene expression, as opposed to procedures relying on the GNA. More importantly, since the model links target concentration to measured intensity, estimating absolute expression levels of transcript targets in the hybridization solution becomes possible.

To increase this method's usability and accessibility, we implemented it as a user-friendly Bioconductor package (Gentleman et al., 2004), CALIB.

2 DESCRIPTION

Below we give a short description of the data structures adopted by the CALIB package, the implemented normalization approach and the corresponding visualization tools.

2.1 Data Structure

In the CALIB package, microarray data are stored in a structure called *RGList_CALIB*, an extension of the *limma::RGList* (Smyth, 2004). The advantage of constructing *RGList_CALIB*, as an extension of the *limma::RGList* enhances its usability for users already familiar with normalization methods such as *normalizeWithinArrays()* in *limma*, allowing for maximal flexibility in using CALIB alongside other packages available within Bioconductor. The CALIB microarray data structure *RGList_CALIB*, as well as other CALIB specific data structures, such as *SpikeList* (to store spike related data) and *ParameterList* (to store estimated calibration parameters), are all inherited directly from the R data type *LIST*. Therefore, any method that works on *LIST*, will also work on these CALIB defined classes.

2.2 Normalization

The normalization method implemented in the CALIB package consists of two major steps: 1. estimation of the calibration parameters: *estimateParameter()* and 2. normalizing the data accordingly: *normalizeData()*.

The function *estimateParameter()* takes the objects of *RGList_CALIB* and *SpikeList* as input arguments and uses spike intensity measurements and corresponding concentrations to estimate the parameters of the calibration models. Since each array has a different set of parameters, the estimation is performed separately

[*] To whom correspondence should be addressed.

for each array. The output of this function is an object of the *ParameterList* class, containing all the model parameters (representing the intensity saturation characteristics of both dyes, and the hybridization of target to complementary probes) of the specified arrays.

The function *normalizeData()* takes the raw microarray data (*RGList_CALIB* object) and the estimated model parameters (*ParameterList* object) as input arguments. The parameters of the spike based calibration model are used to estimate the absolute expression levels for each combination of a gene and condition in the design of the microarray experiment, irrespective of the number of microarray slides or replicate spots on one slide. When required, this function can be used on individual genes, or on a selected group of genes instead of on the entire gene set.

These functions are written in C++. A dynamic link library, compiled from the C++ code is used as a plug-in for the R wrapper functions.

2.3 Visualization

The CALIB package provides different visualization functions that allow quality control and data exploration. As the estimation of model parameters depends to some extent on the quality of the external control spikes, it is advisable to check the quality of the external controls and the estimated parameters prior to running the normalization function by using these visualization functions. Additional functions are also provided to view the final results of the normalization procedure.

The functions *plotSpikeCI()* and *plotSpikeRG()* can be used to check the quality of external controls before running the estimation function, as well as to evaluate the model fit after parameter estimation. *plotSpikeCI()* compares measured intensities to actual spike concentrations. *plotSpikeRG()* compares Cy3 and Cy5 measurements from the same spot. Other functions can only be called after estimating the model parameters: *plotSpikeHI()* plots the estimated amount of hybridized target against corresponding intensities of calibration controls (i.e. with a 1:1 ratio) and serves to evaluate the calibration model, and *plotSpikeSpotError()* plots the distribution of estimated spot capacity errors for all spikes of an array (as a histogram, box plot or density plot).

The functions *plotSpikeACEC()* and *plotNormalizedData()* can be called after data normalization. *plotSpikeACEC()* compares the actual concentrations of calibration controls (i.e. with a 1:1 ratio) with the estimated concentrations. Ratio controls are not included by default, as they do not necessarily reflect the design of the experiment. *plotNormalizedData()* allows comparing the estimated expression levels of two selected conditions.

3 USAGE

The CALIB package has a manual and vignette describing its use. In order to illustrate the workings and principles of this method, and the usage of functions in the package, we include a test example derived from the study of Hilson et al., 2004 consisting of a color-flip of two conditions.

4 DISCUSSION

We implemented CALIB, a user-friendly Bioconductor package for the normalization of two-color microarray data. The underlying method relies on the presence of external control spikes to estimate the parameters of a calibration model, which are then used to obtain absolute expression levels for all genes. This spike based normalization procedure provides an alternative solution to the standard ratio based normalization, which is particularly applicable in cases where, either the global normalization assumption is violated and no alternative solutions exist, or for applications where absolute expression levels are more convenient than ratios'. Besides the normalization procedure, CALIB provides some convenient visualization tools for quality control of the experimental protocol based on externally added control spikes.

ACKNOWLEDGEMENTS

K. Engelen is a postdoctoral fellow of the K.U.Leuven Research Council. This work is partially supported by: 1. IWT projects: GBOU-SQUAD-20160; SBO-BioFrame; SBO-060045, 2. Research Council K.U.Leuven: GOA-Ambiorics, IDO genetic networks; EF/05/007 SymBioSys, 3. FWO projects: G.0413.03, and G.0241.04. We would like to thank Hervé Pagès and Wolfgang Huber for their advice and useful comments on our package.

REFERENCES

Engelen,K. et al. (2006) A calibration method for estimating absolute expression levels from microarray data. *Bioinformatics*, 22, 1251-1258.

Gentleman,R.C. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5, R80.1-R80.16.

Hilson,P. et al. (2004) Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications. *Genome Res.*, 14, 2176-2189.

Kerr,M.K., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, 7, 819-837.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3, Article3.

van Bakel,H. and Holstege,F.C. (2004) In control: systematic assessment of microarray performance. *EMBO Rep.*, 5, 964-969.

van de Peppel,J. et al. (2003) Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Rep.*, 4, 387-393.

Yang,Y.H. et al. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, 30, e15.