



PreP: gene expression data pre-processing

Jorge García de la Nava¹, Sacha van Hijum² and
Oswaldo Trelles^{1,*}

¹Computer Architecture Department, Universidad de Málaga, 29080, Málaga, Spain
and ²Department of Molecular Genetics, University of Groningen, The Netherlands

Received on February 27, 2003; revised on May 13, 2003; accepted on May 31, 2003

ABSTRACT

Summary: PreP is a versatile, powerful, standalone application that aims at pre-processing gene expression data.

Availability: Documentation and executable file for MS-Windows are available at <http://chirimoyo.ac.uma.es/bitlab/services/index.htm>

Contact: ots@ac.uma.es

INTRODUCTION

Technological breakthroughs, such as gene expression monitoring technology have enabled a dynamic view of biological processes by studying expression patterns of thousands of genes in a variety of experimental conditions. The massive generation of data in these experiments demands sensible automatic methods for their analyses.

As with any other computational-based methodology, results are strongly related to the quality of input data. There are many sources of systematic and random variations introduced along the various steps in measuring gene expression levels. These variations in expression levels might lead to a false understanding on gene expressions under certain changing experimental conditions. From this perspective, the application of data pre-processing techniques will produce a significant improvement in the quality of results.

Great efforts have been done in the design and development of methodologies in the process of removing undesirable variation in data. Diverse proposals have been made with the aim to tackle problems, such as (i) sample preparation, fluctuation in efficiency of reverse transcription and labelling (Dudoit *et al.*, 2001); (ii) differences in labelling efficiency between the two fluorescence dyes (Finkelstein *et al.*, 2000); (iii) background interference and noise problems in low expression level data (Yang *et al.*, 2002); (iv) intensity and spatial dependence in measurements, variability within and between slides in the amount of DNA spotted, i.e. pin geometry, fluctuations in volume spotted, fixation to slide, or efficiency of hybridization, differences in spotted arrays (Ideker *et al.*, 2000).

Nevertheless, none of the current available pre-processing software tools cover the following combined features: (a) an integrated and broad gallery of techniques to deal with the many sources of measurement errors; (b) an interactive user friendly interface for visualization of data in an appropriate representation; (c) a standalone application for data privacy and (d) free availability. The PreP application aims to contribute in closing this gap. Furthermore, PreP will enable the user to standardize data handling procedures.

PreP is a standalone interactive graphical suite for the pre-processing of gene expression data that aim to minimize sources of systematic and random variation in the measured data, other than differential expression. PreP integrates a variety of analytical tools. In some cases these can be applied in any context (such as the normalization, adjusting and ratio scaling). In other cases, some specific conditions have to be met (e.g. gene replication). Once the error have been minimized, PreP allows to extract the individual channel signals and ratio's between both channels. Table 1 summarizes the different procedures supplied in PreP for handling specific sources of error.

The core of PreP is a Visual C++ Studio library of algorithms and data handling routines. A group of control functions provides extra functionality in PreP. A gallery of visualization methods increase the level of user friendliness and interactive analysis in PreP (Table 2). PreP input and output files are tab-delimited text files, which can be readily imported into for instance Microsoft Excel. In addition, PreP has its own proprietary data format (*engene compatible*) (García de la Nava *et al.*, 2003). The general set-up of PreP allows the user a high freedom in the (order of) approaches to correct data.

ACKNOWLEDGEMENTS

This work has been partially supported by grant QLK3-2001-01473 under the EU sub-programme area 'Quality of Life and Management of Living Resources—The Cell factory'.

*To whom correspondence should be addressed.

Table 1. Possible sources of experimental errors, the causes and the solutions provided by PreP

Error type	Causes	Correction procedure in PreP
Systematic		
Ratio shift	Different dye or label incorporation efficiencies; different scanning sensitivities or laser power for different colours, non-linear transfer functions in photodetector	Ratio correction via fitting curve, dye-swap
Contrast variations	Nonlinearities and different conditions when measuring	Scaling
Spatial effects	Non-uniform solution spreading, irregular lighting, differences in print-tips; tilted slide	Division of slide in sectors
Random		
Data acquisition (Saturation and quantization error)	Finite dynamic range Translating continuous values into discrete ones	Intensity thresholding and filtering
Multiple Generalized	Imprecision's due to parameters associated to the experiment Intrinsic errors of the measure processes	Value estimation via replication

Table 2. Visualization methods and their uses employed by PreP

Name	Method	Use
Slide view	A synthetic reproduction of the scanned image from the available data	Comparison with the scanned image, identifying single spots, splitting the slide in blocks and manual testing
Slide view of coherent spots	A synthetic reproduction of the scanned image only for coherent data	Evaluation of the quality of the slide and zones poorly scanned (negative or null intensity values are not shown)
Slide view with quality	Uses the blue channel for displaying the quality of the measure	Combined with algorithms that provide a quality value for each spot
MA and RG plots	Logarithmic plot of ratio versus intensity and logarithmic plot of red versus green channel	The MA-plot displays the dependencies of the ratio on the intensity (for ratio correction and filtering); in the RG-plot case the two colour channels are emphasized separately
Box graph	Box graph of each block of the slide	Classical statistical graph for detecting outliers and comparing the distribution of diverse data sets (useful tool for detecting contrast variations interslide or intraslide)
Density graph and Density graph per block	This graph estimates the density distribution of ratios (per slide and per grid block)	Preliminary test on the distribution of the ratios. The expected density graph is a normal distribution (per block, helps detecting spatial errors)
Intensity-Intensity graph	A scatter plot showing the intensity values of one scan versus the same values of another scan	This is a first step for comparing two slides. The data should be near the diagonal if the slides are good replicates of each other
Dispersion, deviation and correlation of replicates	The intensity values of the individual spots of a group, its deviation or its correlation versus the mean of all the spots from the same replication group	Quality estimation of the replication. For dispersion graph, the data points should be along the diagonal, and the more noise, the more blurred they will be. If the deviation is high the quality will decrease
Normality of replications	Applies the inverse of the normal distribution function to the distribution function of each replication group	One typical assumption is that the noise is normally distributed. This graph will test that hypothesis. If the data points lie along the diagonal, the noise is very close to be normal

SUPPLEMENTARY DATA

For supplementary data, please refer to *Bioinformatics* online.

REFERENCES

- Dudoit,S., Yang,Y.H., Luu,P. and Speed,T.P. (2001) Normalization for cDNA microarray data. In Bittner, M.L., Chen, Y., Dorsel,A.N. and Dougherty,E.R. (eds), *Microarrays: Optical Technologies and Informatics*. Proceedings of SPIE, Vol. 4266, 141–152.
- Finkelstein,D.B., Gollub,J. and Cherry,J.M. (2000) Normalization and systematic measurement error in cDNA microarray data. *Joint Statistical Meeting 2000*.
- García de la Nava,J., Franco-Santaella,D., Cuenca,J., Carazo,J.M., Trelles,O. and Pascual-Montano,A. (2003) Engene: the processing and exploratory analysis of gene expression data. *Bioinformatics*, **19**, 657–658.
- Ideker,T., Thorsson,V., Siegel,A.F. and Hood,L.E. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.*, **7**, 805–817.
- 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.