Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 2602–2620
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Identifying differentially expressed genes in dye-swapped microarray experiments of small sample size

I.B. Lian[a,c,*], C.J. Chang[b], Y.J. Liang[c], M.J. Yang[a], C.S.J. Fann[c,**]

[a] *Department of Mathematics, National Changhua University of Education, Changhua 50058, Taiwan*
[b] *Graduate Institute of Medical Science, Chang Gung University, Taipei, Taiwan*
[c] *Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan*

Received 11 July 2005; received in revised form 16 November 2005; accepted 7 January 2006
Available online 26 January 2006

Abstract

When using microarray analysis to determine gene dependence, one of the goals is to identify differentially expressed genes. However, the inherent variations make analysis challenging. We propose a statistical method (SRA, swapped and regression analysis) especially for dye-swapped design and small sample size. Under general assumptions about the structure of the channels, scanner, and target effects from the experiment, we prove that SRA removes bias caused by these effects. We compare our method with ANOVA, using both simulated and real data. The results show that SRA has consistent sensitivity for the identification of differentially expressed genes in dye-swapped microarrays, particularly when the sample size is small. The program for the proposed method is available at http://www.ibms.sinica.edu.tw/~csjfann/firstflow/program.htm.

Keywords: Dye-swapped design; Robust regression; ANOVA; Two color microarray

1. Introduction

Microarray experiments allow simultaneous monitoring of the expression levels of numerous genes. The primary purpose of microarray analyses is to characterize the numerical difference between expression profiles from two distinct cell samples, which may reveal clinical subtypes (Van't Veer et al., 2002) or otherwise clarify biological pathways involving certain genes (Roberts et al., 2000). Since this technology generates large and complex multivariate data sets, one of the major challenges lies in the development of computational and statistical approaches as analytical tools.

Microarray data contains inherent "noise" such as between-slide bias, differential dye efficiency, and background noise, among other factors. In addition, large-scale microarray experiments can be expensive in terms of equipment, consumables, and time. As such, Yang and Speed (2002) have advocated the importance of careful design in order to extract the maximum amount of information.

* Corresponding author. Department of Mathematics, National Changhua University of Education, Changhua 50058, Taiwan. Tel.: +886 47232105; fax: +886 47211192.
E-mail addresses: maiblian@cc.ncue.edu.tw (I.B. Lian), csjfann@ibms.sinica.edu.tw (C.S.J. Fann).
** Also for correspondence.

0167-9473/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.csda.2006.01.003

Statistical approaches to the analysis of microarray data fall into two major groups, namely methods that identify differentially expressed genes (Jain et al., 2003; Dudoit et al., 2002) and those that classify the functional dependency of genes (e.g., hierarchical clustering; Eisen et al., 1998). In the present study, we focused on identifying differentially expressed genes. The first method, an intuitive approach, identifies differentially expressed genes based on fold change, i.e., any gene for which the ratio of two-color channel intensities exceeds a pre-set value is said to be differentially expressed. A few methods have used the conventional *t*-test or similar statistical tests (Roberts et al., 2000; Tusher et al., 2001), and a categorical summary measurement of differential expression based on ranking has also been proposed (Tsodikov et al., 2002). These approaches have contributed to the development of even more accurate methods. When prior information is available, some Bayesian approaches (e.g., Baldi and Long, 2001) can further improve the inference. Here we assumed that prior information is unavailable.

Most cDNA microarray designs assess gene expression in reference cells using one channel (green or red fluorescent dye), whereas that in target cells is assessed by the other channel. Among these designs, ratio-based and intensity-based analyses are frequently used. The latter design uses separate intensities in the analysis and has been claimed to be more efficient and flexible ('t Hoen et al., 2004). However, two-channel designs often have systematic variation due to different efficiencies of fluorescent dyes. Dye biases can be caused by various factors, and they may be confounding when searching for subtle biological differences (Yang et al., 2002). To circumvent this type of variation, it is recommended that data be obtained by swapping pairs of dyes between the channels (Jin et al., 2001; Yang et al., 2002).

Data reproducibility and between-slide variations have been major concerns in microarray technology. That is, relative gene expression levels from replicate experiments might not be the same due to errors introduced from diverse experimental conditions. Therefore, Quackenbush (2001) has advocated the concept of experimental replication. The present study utilized a method that identifies differentially expressed genes in experiments that incorporate a dye-swapping design and replications with respect to channel, scanner and target effects. Using this statistical framework, the dye-swapping and summation steps canceled out any nebulous relationships that may have otherwise affected readings between two cell types for most genes that were, in fact, not differentially expressed. Therefore, the corresponding data pairs clustered around the origin or along a diagonal line in the x–y scatter plot, where x and y represent the dye-swapped readings from two cell types. For genes that are expressed differentially between cell types, the corresponding points of the paired data lay away from the diagonal in the scatter plot. Aggregated statistics for standardized residuals from the fitting of a robust regression method was used to identify genes that were differentially expressed in a statistically significant manner.

Using simulated and real data sets, we compared our method (swapped regression analysis, SRA) with an ANOVA-type test and a non-parametric version of *t*-test. Our simulations show that the dye-swapping design enhances the sensitivity for identifying differentially expressed genes. In this respect, SRA is generally robust in sensitivity compared with other methods, at a cost of modest higher false discovery rate (FDR) than ANOVA method. We conclude that SRA is a good complement to ANOVA, particularly when the sample size is small. Both SRA and ANOVA have better overall performance than the *t*-test, therefore only the comparisons between the first two were shown.

2. Dye-swapping design

In a dye-swapping design, the hybridization step is performed twice, with the dye assignment reversed in the second hybridization. Let g_i and r_i be the true intensities for the green and red channels for the ith gene, $i = 1, \ldots, n$, and G_{ij} and R_{ij} be the observed readings of g_i and r_i from the jth subject, $j = 1, \ldots, m$. Assume that the first cell type (e.g., reference cells) on the first slide is labeled with a green fluorescent dye and the other cell type (e.g., target cells) with a red dye. For the second slide, the labeling is reversed. A Latin square can be designed, as shown in Table 1.

ANOVA, an intuitive approach, requires linear assumptions for all effects. To avoid such strict assumptions, we considered a different approach. To explore the properties analytically, we made general assumptions for the majority of the non-differentially expressed genes with regard to the channel effect in Eq. (1) as well as the scanner effect for green and red dyes in Eqs. (2) and (3), respectively. Assuming that there is an unknown relationship, (f_i), between green and red intensities, for gene i:

$$r_i = f_i(g_i). \tag{1}$$