Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Statistical Planning and Inference 129 (2005) 19–37

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Sequential prediction bounds for identifying differentially expressed genes in replicated microarray experiments

Robert D. Gibbons[a,*], Dulal K. Bhaumik[a], David R. Cox[b], Dennis R. Grayson[c], John M. Davis[c], Rajiv P. Sharma[c]

[a] *Departments of Biostatistics, Psychiatry, and Center for Health Statistics, University of Illinois at Chicago, 1601 W. Taylor, Chicago, IL 60612, USA*

[b] *Department of Genetics, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, USA*

[c] *The Psychiatric Institute, University of Illinois at Chicago,1601 W. Taylor, Chicago, IL 60612, USA*

Available online 21 August 2004

Abstract

Microarrays are new biotechnological devices that permit the simultaneous evaluation of expression levels of thousands of genes in one or more tissue samples. We develop a new method for identifying differentially expressed genes in replicated cDNA and oligonucleotide microarray experiments. The method is based on a nonparametric prediction interval which is computed as an order statistic of n control measurements and is applied sequentially to a series of p replicate sets of experimental measurements, each of size n_i. We illustrate how reasonable experiment-wise false positive and false negative rates can be attained for any practical number of genes based on manipulating the order statistics, n, p and n_i. The method is used to identify gene expression levels that are associated with a pathological condition beyond chance expectations given the large number of genes tested. We illustrate use of the method on replicated gene expression data in tumor and normal colon tissues, and compare it to an alternative approach based on permutation tests.

Keywords: Microarray; Simultaneous prediction intervals; Molecular genetics; Statistical genetics; Multiple comparisons; Nonparametric statistics

[*] Corresponding author. Tel.: +1-312-413-7755; fax: +1-312-996-2113.
E-mail address: rdgib@uic.edu (R.D. Gibbons).

0378-3758/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.jspi.2004.06.037

1. Introduction

With the advent of methods for large-scale gene expression studies, such as microarray technology (Chee et al., 1996), evaluation of large numbers of gene mRNA levels amongst different individuals is now possible. An immediate consequence of this new technology is the ability to differentiate gene expression among normal and diseased states, a problem of central importance to modern biology and medical research and has wide applicability to high throughput pharmaceutical screening. From a statistical perspective, the problem is complex for three primary reasons. First, determining whether an observed difference between two sources (e.g., normal children versus children with Down's syndrome) is due to chance, is an almost insurmountable problem when testing several thousand genes. Second, little is known about the distributional form of intensity data of this type. Often, the data vary over several orders of magnitude with a proportion of the distribution censored below a limit of detection (Audic and Claverie, 1998). In many cases (e.g., cDNA microarrays), the data for each gene are expressed as a ratio of two fluorescence intensity measures, leading to both left and right censoring problems. Third, the number of available measurements is typically small (i.e., 10 or 20 at most), at least in part, due to the high expense of these new technologies, and for some applications due to the limited availability of postmortem human tissue of sufficient quality for the analysis.

Note that the problem addressed in this paper (i.e., identifying differential expression levels in replicated microarray experiments) is only one of several important statistical problems associated with microarrays (see Claverie, 1999 for a review of the earlier work). Some of the earliest approaches to this problem involved the identification of groups of genes with similar function using cluster analysis (Eisen et al., 1998). Such methods are often referred to as "unsupervised" in that they do not consider auxiliary information regarding gene function or their relation to relevant outcomes of interest. Alternative "supervised" methods have been introduced to take advantage of auxiliary information by incorporating disease classes or related outcomes. For example, some investigators have focused on the identification of genes of similar function (e.g., classification of various tumor types) using various forms of discriminant analyses applied to the expression level profiles (Slonim et al., 2000; Dudoit et al., 2000; Brown et al., 1999; Ben-Dor et al., 2000). Hastie et al. (2000) have combined supervised and unsupervised approaches by developed "gene shaving" and "tree harvesting"' algorithms in which hierarchical clustering is used to reduce the dimensionality of the problem and the resulting gene clusters are then related to outcomes of interest (e.g., survival time). A similar approach using linear modeling of "genetic profiles" (i.e., identified using unsupervised hierarchical clustering) has been suggested by van Someren et al. (2000).

Methods for analysis of variance components in a single microarray slide have been developed by Newton et al. (2001); Kerr et al. (2000) and Sapir and Churchill (2000). In such experiments, the data for each gene consist of two fluorescence intensity measures, (R, G), representing the expression level of the gene in the red (Cy5) and green (Cy3) labeled mRNA samples. Typically, one of the two dyes corresponds to a pooled set of control tissues (e.g., normal colon tissue samples) and the other to an experimental tissue of interest (e.g., tissue taken from a malignant tumor in the colon). Dudoit et al. (2002) classify various "single-slide" methods in terms of whether they are based solely on the