A Laplace Mixture Model for Identification of Differential Expression in Microarray Experiments

Debjani Bhowmick, A. C. Davison*, and Darlene R. Goldstein

Ecole Polytechnique Fédérale de Lausanne

Institute of Mathematics

EPFL-FSB-IMA, Station 8

CH-1015 Lausanne, Switzerland

* corresponding author

email: anthony.davison@epfl.ch

Tel: + 41 21 693 5502

Fax: + 41 21 693 4250

April 20, 2006

Abstract

Microarrays have become an important tool for studying the molecular basis of complex disease traits and fundamental biological processes. A common purpose of microarray experiments is the detection of genes that are differentially expressed under two conditions, such as treatment versus control, or wild-type versus knock-out.

We introduce a Laplace mixture model as a long-tailed alternative to the normal distribution when identifying differentially expressed genes in microarray experiments, and provide an extension to asymmetric over- or under- expression. This model permits greater flexibility than models in current use as it has the potential, at least with sufficient data, to accommodate both whole genome and restricted coverage arrays.

We also propose a REML-type approach to hyperparameter estimation which is equally applicable in the Normal mixture case.

The Laplace model appears to give some improvement in fit to data, although simulation studies show that our method performs similarly to several other statistical approaches to the problem of identification of differential expression.

Keywords: Laplace distribution; Marginal likelihood; Microarray experiment; Mixture model; REML.

1 Introduction

Microarrays have become an important tool for studying the molecular basis of complex disease traits and fundamental biological processes. Two-channel microarrays, such as spotted cDNA or long oligonucleotide arrays, measure relative gene expression in two samples. Once preprocessed, the data from such arrays take the form of normalized base 2 logarithm of the expression ratios. A common purpose of microarray experiments is the detection of genes that are differentially expressed under two conditions, such as treatment versus control, or wild-type versus knock-out. Numerous statistical methods have been proposed for identification of differential expression, with new ones continuing to be introduced.

In early analyses such as (1; 2), fold change between conditions exceeding a constant was used to identify differentially expressed genes. This method performs poorly, however, because it ignores the different variability of expression across genes. Such variability can be taken into account by using a t-test on the average log fold change, but variation in gene expression is poorly estimated with small numbers of replicates, so genes with artificially low variance may be selected even if they are not truly differentially expressed. Numerous refinements have been proposed in order to reduce the numbers of false positive and false negative results. (3) and (4) suggested adding a constant to the t denominator so that it does not become too small. (5) proposed a normal mixture model for the gene expression data and defined a log posterior odds statistic, their B-statistic, for

3

ranking genes. Their approach was extended to linear models for more general designs by (6); these methods are implemented in the `R` package `limma` as part of the BioConductor project (`www.bioconductor.org`). (7) use a similar approach but also include a heuristic, iterative method for estimating the proportion of differentially expressed genes.

The present paper makes three main contributions: we consider robust and asymmetric variants of the mixture modelling approach, we propose a new approach to parameter estimation, and we perform a comparative study intended to elucidate key features of such methods. The methods studied include those listed above, along with our own proposal of a Laplace mixture as a long-tailed alternative to the normal distribution for mean gene expression across replicate arrays proposed by (5).

This paper is organised as follows: Section 2 describes the model, with estimation of the hyperparameters discussed in Section3. Simulation studies outlined in Section 4 show that this method performs similarly to several other statistical approaches to the problem of identification of differential expression. In Section 5 we apply our method to a published dataset and compare its results with other methods. The paper ends with a brief discussion.

References

[1] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science,* **270**, 467–470.

[2] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences, USA,* **93**, 10614–10619.

[3] Tusher, V. G., Tibshirani, R. J., and Chu, G. (2001) Significance analysis of microarrays applied to the ionising radiation response. *Proceedings of the National Academy of Sciences, USA,* **98**, 5116–5124.

[4] Efron, B., Tibshirani, R. J., Goss, V., and Chu, G. 'Microarrays and their use in a comparative experiment'. Technical report Stanford University (October, 2000) Department of Health Research and Policy.

[5] Lönnstedt, I. and Speed, T. P. (2002) Replicated microarray data. *Statistica Sinica,* **12**, 31–46.

[6] Smyth, G. K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology 3, No. 1, Article 3,* **3**(1), Article 3.

[7] Gottardo, R., Pannucci, J. A., Kuske, C. R., and Brettin, T. (2003) Statistical analysis of microarray data: a Bayesian approach. *Biostatistics,* **4**, 597–620.