

Identification of differential gene expression from DNA microarrays

C.A. Glasbey and C.D. Mayer

Biomathematics and Statistics Scotland

JCMB, King's Buildings, Edinburgh, EH9 3JZ, Scotland

1 Introduction

Microarray (or DNA chip)-based hybridisation analyses using high density DNA probe arrays is a powerful tool for a broad and diverse set of genetic applications, including gene expression monitoring, sequence information, and genotype analysis (see, for example, Chipping Forecast, 1999). It allows new biomolecular approaches that promise to revolutionise our understanding of physiology and disease. A good starting point for finding out more about this rapidly developing field is the web site: www.gene-chips.com

To study the difference in gene expression between two samples, each gene is placed as a spot on a glass slide, and the slide is then hybridized with the two samples, each labelled with a fluorescent marker. Finally, the microarray is scanned at high spatial resolution at the two wavelengths. For example, Fig 1 shows an array comparing two strains of Human Cytomegalo Virus (HCMV). Red spots reveal genes only expressed by virus strain one, green spots show genes only expressed by strain two, yellow spots show genes expressed by both strains and dark spots by neither. The first stages in the analysis of such data are estimation of expression of each gene, and identification of differential gene expression. We consider them in the following two sections.

2 Image analysis

Image analysis is the first step in analysing microarray data. Methods of noise reduction, background correction and segmentation are needed before the integrated intensity of individual spots can be obtained. Work is reported at the US National Human Genome Research Institute web page www.nhgri.nih.gov/DIR/LCG/15K/HTML/img_analysis.html and in Yang et al. (2000). However, there remain needs and opportunities for exploration of alternative, improved methods and for extensive, empirical and theoretical comparisons between methods.

In the talk we illustrate the use of median filters to reduce the effects of speckle noise. We also use morphological operators such as the top-hat filter to correct the images for background trend, as proposed by Yang et al. (2000). Then we consider alternative segmentation methods to isolate the spots in the images.

To estimate the difference in spot intensity between the two samples, we model pixel values by bivariate log-normal distributions. The maximum likelihood estimator for this model, and

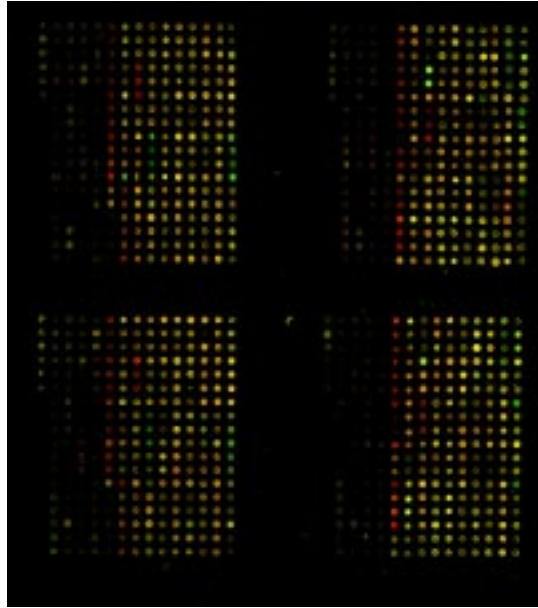


Figure 1: DNA microarray comparing two strains of Human Cytomegalo Virus (HCMV).

the estimator used in most applications, of the differential rate of expression are given by

$$\exp \left[\frac{1}{n} \sum_{i,j} \log \left(y_{i,j}^{(2)} / y_{i,j}^{(1)} \right) \right] \quad \text{and} \quad \sum_{i,j} y_{i,j}^{(2)} / \sum_{i,j} y_{i,j}^{(1)}.$$

Here $y_{i,j}^{(k)}$ denotes the pixel value in row i , column j , for sample k , and summations are over the n pixels in a segmented spot. However, simulations show the difference in efficiency between the estimators to be small for typical values of parameters.

3 Identification of differential gene expression

To facilitate comparison between two samples on a single array, statistical models are needed, building on the work of (Chen et al., 1997; Newton et al., 2001). The appropriateness of several mathematical assumptions need to be assessed under experimental conditions, such as that of a latent binary response. Resampling methods, such as those in Dudoit et al. (2000), can be used to indicate whether the differential expression of a gene is statistically significant. Robust, nonparametric methods are also of relevance, to reduce the influence of contaminated data, which are typical for microarray data.

In the talk, we show how mixture distributions can be use to identify gene expression and differential gene expression for data such as in Fig 2. For a single sample, we assume that

$$\log(y^{(k)}) \sim pN(\mu_1, \sigma^2) + (1 - p)N(\mu_2, \sigma^2),$$

whereas for the difference between samples, we assume that

$$\log(y^{(2)} / y^{(1)}) \sim pN(\mu, \sigma_1^2) + (1 - p)N(\mu, \sigma_2^2).$$

In both cases, we can then estimate the probability of either mixture component, given the data.

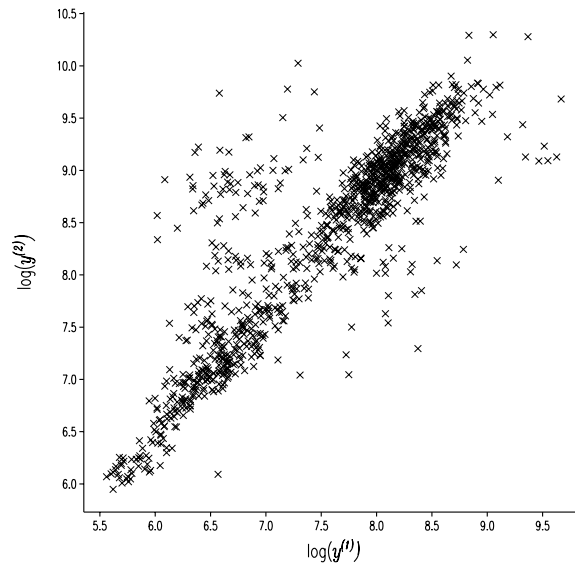


Figure 2: Spot intensities from sample 2 plotted against those from sample 1, on log-scale, for data in Fig 1.

Acknowledgements

The work was supported by funds from the Scottish Executive Rural Affairs Department, and we thank the Scottish Centre for Genome Technology and Informatics for the data.

References

- Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997). Ratio based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, 2:364–374.
- Chipping Forecast (1999). *The Chipping Forecast*, volume 21 of *Supplement to Nature Genetics*.
- Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. (Available at www.stat.Berkeley.EDU/users/terry/zarray/Html/matt.html).
- Newton, M. A., Kenzioriski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*. (In press).
- Yang, Y. H., Buckley, M. J., Dudoit, S., and Speed, T. P. (2000). Comparison of methods for image analysis on cDNA microarray data. (Available at www.stat.Berkeley.EDU/users/terry/zarray/Html/image.html).