Assessing differential gene expression using two-component microarray mixture models.

Geoff Laslett*, Phil Brown† and Albert Trajstman
Victorian Bioinformatics Consortium
Monash University, Clayton, Victoria 3800, Australia
and CSIRO Mathematical and Information Sciences

Microarray experiments are today's challenge for statisticians. Already numerous formal and informal methods of analysis have been proposed, and it is unclear which are best suited. We are concerned with replicated microarray studies in which cDNA from a tissue of interest is labelled red, say, and from a control tissue is labelled green, and there are no external covariates or outcome variables. A number of papers tackle this problem using either a Gamma observational model on the two channels or a log-normal distribution of expressions, see Kendziorski *et al* (2003) for recent extensions. We follow the model-based log-normal method of Lönnstedt and Speed (2002) for ranking genes according to their differential expression levels. This is an empirical Bayes style of hierarchical mixture model in which a proportion p (typically 1%) of genes is differentially expressed. The variances of the residual log differential expression values vary between genes according to an inverse Gamma distribution, and the means of the differentially expressed genes also vary in magnitude according to a natural conjugate model. Lönnstedt and Speed (2002) recommend that p be fixed *a priori*, and suggest an informal method for estimating the other three global parameters in the model.

We have fitted this model to some replicated microarray data generated from studies of the fowl cholera bacteria in chickens. For our data the Lönnstedt and Speed (2002) model seems to miss some variation. In our presentation we indicate how to extend the model so that it fits these data better, and we also recommend fitting all the parameters formally using likelihood or Bayesian methods. The ranking of genes is not as straightforward under the extension as it is under the original model, so we propose a plot that summarises the relevant information in an easy-to-understand way. Our extension also takes account of dyeswap, and can incorporate other random effects.

REFERENCES

Kendziorski, C. M., Newton, M. A., Lan, H. and Gould, M. N. (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Statistics in Medicine, to appear.

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

*E-mail: geoff.laslett@csiro.au

†On leave from University of Kent