Hierarchical modelling to handle heteroscedasticity in microarray data

Renée X. de Menezes & Hans c. van Houwelingen

Medical Statistics Department, Leiden University Medical Centre, Netherlands[1]

One of the main challenges in microarray data analysis is to develop statistical procedures which can, based upon a small number of observations per gene but a large number of genes, be powerful enough to detect effects of interest. To compare gene expression levels from two groups, such as *treated* and *untreated*, several approaches have been proposed: some suggest handling the problem using permutation tests (Efron *et al.* , 2001, Tusher *et al.* , 2001), whilst others do so via hierarchical models (Baldi & Long, 2001, Lonnsted & Speed, 2002). Merits of these approaches have been illustrated using empirical arguments, which allow for comparisons between approaches, but not much is learnt about the approach's optimality or power.

We propose a more structured approach: construct a likelihood-ratio test statistic which makes use of a hierarchical structure in a way similar to Baldi & Long's approach, but purely frequentist, so that estimation is easily implemented. This test statistic can then be re-written as a generalized Student's t statistic which, under the null hypothesis of no treatment effect, has a Student t distribution with augmented degrees of freedom.

Sets of genes appointed as affected by the new test showed agreement with sets appointed by Significance Analysis of Microarrays (Tusher *et al.* , 2001), with the new test showing more power in situations where the proportion of affected genes was small, and the sample size in each group was small, while false discovery rates yielded by the new test and SAM were of comparable sizes.

Extensions to other statistical tests and methods to improve their power in microarray data analysis, such as the F-test and ANOVA, will be discussed.

References

Baldi, P. & Long, A. D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17** (6), 509-519.

Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151-1160.

Lonnstedt, I. & Speed, T. (2002). Replicated microarray data. *Statistica Sinica* **12** (1): pp.31-46.

Tusher, V. G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* **98**, 5116-5121.

[1]Address: P.O. Box 9604, 2300 RC, Leiden, Netherlands. E-mail: `r.x.menezes@lumc.nl`