False Discovery Rate-Controlled Gene Selection under Variance Heterogeneity and Correlated Expression Measurements

Michael G. Schimek[1] and Tomáš Pavlík[2]

[1] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria
[2] Institute of Biostatistics and Analyses, Masaryk University, Czech Republic

Statistical procedures for the identification of differentially expressed genes involve a serious multiple comparison problem. The false discovery rate (abb. *FDR*) is a key tool for type I error control. *FDR* estimation and the obtained selection results can suffer from high variability of the gene expression measurements, especially when low variance genes are involved. This problem is usually tackled by a 'correcting' constant applied to the pooled variance in the parametric test statistic. But the results are also influenced by correlated measurements because of co-expressed genes.

In this presentation our main interest is the popular SAM approach ('Significance Analysis of Microarrays', Chu et al. (2005) Tech. Rep., Stanford Univ.), frequently applied in molecular biology and medicine. It comprises a *FDR* estimation concept (Tusher, Tibshirani and Chu (2001) PNAS, 98, 5116–21). Moreover we study two alternative procedures (Schwender, Krause and Ickstadt (2003) Tech. Rep. SFB 475, Univ. of Dortmund; Grant, Liu and Stoeckert (2005) Bioinformatics 21, 2684–90).

Via extensive simulations we compared these procedures applying the ordinary as well as the modified t-test statistic under various choices of the 'correcting' constant for typical data models of varying complexity. The widely discussed correlation model proposed by Qiu, Xiao, Gordon and Yakovlev (2006, BMC Bioinformatics, 7:50) was studied too. We show that the gene selection results do not differ much between the above *FDR* procedures, but rather depend on the complexity of the adopted data model and not so much on the extent of correlation. Finally we draw conclusions for the practical use of SAM and its variants, also with respect to the appropriate choice of the 'correcting' constant. Last but not least we consider the potential of an 'automatic' empirical Bayes procedure avoiding *FDR* estimation at all.