Large-scale significance testing of high-throughput data with FAMT

Magalie Houée-Bigot[1,2], Chloé Friguet[3], Sandrine Lagarrigue[2], Yuna Blum[2], and David Causeur[1]

[1] Agrocampus Ouest- Applied Mathematics Department, 65 rue de St Brieuc 35042 RENNES Cedex, France.
[2] INRA- UMR598, Animal Genetics, 65 rue de St Brieuc 35042 RENNES, France.
[3] Lab-STICC- University of South Brittany, 8 rue Montaigne, 56000 VANNES, France.

Abstract. Analysis of complex systems using high-throughput technologies offers new challenges for statistics. In systems biology for example, microarray technology gives access to whole-genome transcription datasets. It has therefore turned out to be a powerful tool to find out genes which expression variations are significantly related to a given trait using large-scale significance testing. Since Benjamini and Hochberg (1995)'s procedure to control the False Discovery Rate (FDR), the multiple testing theory has been deeply renewed. But the heterogeneity of microarray data has long been ignored in statistical models. However, some recent papers (see Friguet *et al.*, 2009) suggest that unmodeled heterogeneity factors may generate some dependence across gene expressions and affect consistency of the multiple testing results.

Friguet *et al.* (2009) propose a supervised factor model to identify the latent heterogeneity components, method implemented in the R package FAMT (Factor Analysis for Multiple Testing). The talk aims both at presenting the statistical handling of multiple testing dependence as proposed in Friguet *et al.* (2009) and at illustrating the performance of the method by a microarray data analysis using the `R` package FAMT. As described in Blum *et al.* (2010), this microarray study analyses the relationships between the abdominal fatness of chickens and hepatic transcriptome profiles. Some heterogeneity components are extracted from the data by an EM algorithm. Additional functionalities optimize the procedure, such as the estimation of the proportion of true null hypotheses or the optimal number of factors.

Keywords: factor analysis, multiple testing, dependence, high dimension, R.

References

1.Y. Benjamini, and Y. Hochberg. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 289–300, 1995

2.Y. Blum, G. Le Mignon, S. Lagarrigue and D. Causeur. A Factor Model to Analyze Heterogeneity in Gene Expression. *BMC Bioinformatics*, 11, 368, 2010.

3.C. Friguet, M. Kloareg and D. Causeur. A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104, 1406–1415, 2009.