Significance analysis of microarrays outcomes may be altered by data structures

Fabien Valet[1−3], Eléonore Gravier[1−3], Yann de-Rycke[1−3], and Bernard Asselain[1−3]

[1] Institut Curie, Paris, F-75248 France
[2] INSERM, U900, Paris, F-75248 France
[3] Ecole des Mines de Paris, ParisTech, Fontainebleau, F-77300 France

The Significance Analysis of Microarrays (SAM) is one of the most popular method [1] to identify genes that are differentially expressed under different experimental conditions, or among different types of tissue sample (for example breast cancer which developped metastasis and breast cancer which did not). A modified T-test is computed for each gene to identify genes with significant changes in expression between the samples. In order to take into account the multiplicity of tests, the proportion of genes that are identified by chance (ie identified genes that are false discoveries) is estimated using the FDR (False Discovery Rate) .
As pointed out by several authors [2, 3], SAM outcomes may be altered by both pre-SAM data selection (for example arbitrary filtering), and within-SAM data filtering (for example, there is no particular recommendation concerning the choice of the FDR threshold, nor on the number of necessary permutations to calculate this FDR). In addition to these possible alterations, the structure of the data, as for example the distribution of gene expressions and the correlation between these genes, may also influence SAM outcomes.
In this study, we proposed to investigate the influence of the data structure in SAM outcomes. With this aim, we performed simulations with different particular data structures, in particular correlation structures. We showed that the SAM ability to detect the true significant genes varied under different conditions of data structures.

References

[1] Goss Tusher V, Tisbshirani R, Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA, 98:5116-5121

[2] Larssin O, Wahlestedt C, Timmons JA (2005). Considerations when using the significance analysis of microarrays (SAM) algorithm. BMC Bioinformatics, 6:129

[3] Xie Y, Pan W, Khodursky AB (2005). A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. Bioinformatics, 21:4280-8