

# Reproducible Statistical Analysis in Microarray Studies

Ulrich Mansmann  
Department of Medical Biometry and Informatics  
University of Heidelberg

Reproducibility of calculations is a longstanding issue within the statistical community (Leisch and Rossini, 2003). Due to the complexity of the algorithms, the size of the data sets, and the limitations of the medium printed paper it is usually not possible to report all the minutiae of the data processing and statistical computations. Like the critical assessment of a mathematical proof it should be possible to check the software behind a complex data analysis.

Microarrays are a recent biotechnology that offer the hope of improved cancer classification, providing clinicians with the information to choose the most appropriate form of treatment (van't Veer et al., 2002; van de Vijver et al., 2002; Huang et al., 2003). A number of publications presented clinically promising results by combining this new kind of biological data with specifically designed algorithmic approaches. But, reproducing published results in this domain is harder than it may seem. For example, Tibshirani and Efron (2002) report: "We reanalysed the breast cancer data from van't Veer et al. (2002). ... Even with some help of the authors, we were unable to exactly reproduce this analysis."

To achieve reproducible calculations and to offer an extensible computational framework the tool of a compendium (Sawitzki, 1999; Leisch, 2002; Gentleman and Temple Lang, 2004; Sawitzki, 2002) is discussed. A compendium is a document that bundles primary data, processing methods (computational code), derived data, and statistical output with the textual documentation and conclusions. It is interactive in the sense that it allows to modify the processing options, plug in new data, or insert further algorithms and visualisations.

This talk presents examples, discusses the problems hidden in the published analyses and demonstrates a strategy to improve the situation which is based on the vignette technology available from the R and Bioconductor projects (Ihaka and Gentleman, 1996; Gentleman and Carey, 2002).

## References

- Gentleman, R. and Carey, V. (2002). Bioconductor. R News 2(1) 11-16.
- Gentleman, R. and Temple Lang, D. (2004). Statistical analyses and reproducible research. unpublished manuscript
- Huang, E., Cheng, S.H., Dressman, H., Pittman, J., Tsou, M.H., Horng, C.F., Bild, A., Iversen, E.S., Liao, M., Chen, C.M., West, M., Nevins, J.R. and Huang, A.T. (2003). Gene expression predictors of breast cancer outcomes. The Lancet 361:1590-1596.
- Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 5: 299-314.
- Leisch, F. (2002). Dynamic generation of statistical reports using literate data analysis. Compstat 2002 - Proceedings in Computational Statistics 575-580.
- Leisch, F. and Rossini, A.J. (2003). Reproducible Statistical research. Chance 16(2):41-45.
- Sawitzki, G. (1999) Software components and document integration for statistical computing. Proceedings ISI Helsinki 1999 (52nd session) Bulletin of the International Statistical Institute Tome LVIII 2 117-120.
- Sawitzki, G. (2002). Keeping Statistics Alive in Documents. Computational Statistics 17: 65-88.
- Tibshirani, R.J. and Efron, B. (2002). Pre-validation and inference in microarrays Statistical Applications in Genetics and Molecular Biology 1(1).
- van de Vijver, M.J., He, Y.D., van't Veer, L.J. et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347: 1999-2009.
- van't Veer, L., Dai H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van de Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend S.H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530-536.