An empirical Bayes method for gene expression analysis in R

Michael G. Schimek and Wolfgang Schmidt

Keywords: Empirical Bayes, Bioconductor project, microarray, multiple testing, R, sparse sequence.

Revised abstract for oral presentation

In recent years the new technology of microarrays has made it feasible to measure expression of thousands of genes to identify changes between different biological states. In such biological experiments we are confronted with the problem of high-dimensionality because of thousands of genes involved and at the same time with small sample sizes (due to limited availability of cases). The set of differentially expressed genes is unknown and the number of its elements relatively small. Due to a lack of biological background information this is a statistically and computationally demanding task.

The fundamental question we wish to address is differential gene expression. The standard statistical approach is significance testing. The null hypothesis for each gene is that the data we observe have some common distributional parameter among the conditions, usually the mean of the expression levels. Taking this approach, for each gene a statistic is calculated that is a function of the data. Apart from the type I error (false positive) and the type II error (false negative) there is the complication of testing multiple hypotheses simultaneously. Each gene has individual type I and II errors. Hence compound error measures are required. Recently several measures have been suggested ([1]). Their selection is far from trivial and their calculation computationally expensive.

As an alternative to testing we propose an empirical Bayes thresholding (EBT) approach for the estimation of possibly sparse sequences observed with white noise (modest correlation is tolerable). A sparse sequence consists of a relatively small number of informative measurements (in which the signal component is dominating) and a very large number of noisy zero measurements. Gene expression analysis fits into this concept. For that purpose we apply a new method outlined in [5]. It circumvents the complication of multiple testing. More than that, user-specified parameters are not needed, apart from distributional assumptions. This automatic and computationally efficient thresholding technique is implemented in R.

The practical relevance of EBT is demonstrated for cDNA measurements. The preprocessing steps and the identification of differentially expressed genes is performed using R functions ([4]) and Bioconductor libraries ([3]). Finally comparisons with selected testing approaches based on compound error measures available in `multtest` ([2]) are shown.

References

[1] Benjamini, Y. and Hochberg, Y (1995). *Controlling the false discovery rate: A practical and powerful approach to multiple testing.* J. Royal Statist. Soc.,

B 85, 289 – 300.

[2] Dudoit, S. and Ge, Y. (2003). *Bioconductor's multtest package.* Report, `http://www.stat.berkeley.edu/∼sandrine`.

[3] Dudoit, S. and Hornik, K. (2003) The Bioconductor FAQ. `http://www.bioconductor.org/`.

[4] Ihaka, R. and Gentleman, R. (1996). *R: A language for data analysis and graphics.* J. Computat. Graph. Statist., **5**, 299 – 314.

[5] Johnstone, I. M. and Silverman, B. W. (2004). *Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences.* To appear in Annal. Statist.

Acknowledgement: The second author acknowledges funding from Gen-AU project "A Comprehensive Disease Bank for Functional Genomics".

Affiliation: Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, Auenbruggerplatz 2, A-8036 Graz, Austria, michael.schimek@meduni-graz.at.