A database-based application for management and statistical analysis of high-throughput gene expression data

[*] Marco Masseroli, PhD, [*] Pietro Cerveri, PhD, [†] Pier Giuseppe Pelicci, MD, [†] Myriam Alcalay, MD, PhD

[*] Dipartimento di Bioingegneria, Politecnico di Milano, Milan – Italy
[†] Istituto Europeo di Oncologia, Milan - Italy

Background

Gene DNA sequences identified by many genome projects and improvements of nanotechnology have made high-throughput experiments a powerful tool to study the differential expression of thousands of genes at once. Nevertheless, these experiments produce a huge amount of data, quantifying the expression level of each of thousand genes in a number of different tissue types and conditions, presenting variability of gene expression levels and noise, which require specific data analysis. The different types of arrays available for high-throughput gene expression experiments, the large amount of spotted genes in a single array experiment, the number of arrays used in replicate experiments[1], and the analyses required[2] for the massive data produced, demand an adequate software framework helping investigators uncovering new biological information.

Implementation Methods and Results

An integrated database-based application, called Gene Array Analyzer Software (GAAS) (http://www.medinfopoli.polimi.it), has been implemented in MS-Visual C++ programming language and interconnected to a MS-Access 2000 database whose schema enables flexible data management and multi-user analyses. GAAS is characterized by management, analysis and visualization frameworks.

The *management framework* allows processing virtually any kind of gene expression dataset format, storing - separately for each user - all parameter values used during data analysis, and defining custom suitable data visualization and storage configurations of analysis results.

The *analysis framework* enables flexible parametric gene expression data analyses according to the following sequential steps.

1) Automatic background and spot quality evaluation, and background correction. 2) Data normalization by median or mean intensity, either of all clone intensities in the array or of a subset of clone intensities assumed not to vary among the evaluated conditions. 3) Differential gene expression (gene regulation) in a single experiment (i.e. test vs. control condition) evaluated on statistically significant expression intensity ratios with respect to confidence intervals automatically determined on log-ratio distribution, or manually defined folding thresholds[2]. 4) Gene regulation in multiple replica experiments according to regulation reproducibility cut-off or conditional probability calculated for each gene according to its regulation probability in each single experiment.

The *visualization framework* provides powerful user interfaces to interactively navigate input data and analysis results of each evaluation step, both in tabular and graphical format (i.e. histogram and scatter plots of expression level distributions).

Discussion and Conclusions

GAAS enables flexible management and fast differential expression analyses of high-throughput gene expression data also from many replica experiments at once. Background, spot, and clone quality analyses, windowing of noise-affected low and high expression levels, and normalization enable excluding data altered by topological hybridization differences or spotting errors and standardizing data from distinct experiments to avoid bias in analysis results. Statistical analysis of differential expression in a single experiment (test vs. control condition) provides for each clone probability of regulation (up, down, not) according to a selected significance level. Evaluation of clone differential expression in multiple replica experiments on the basis of single experiment expression ratios allows comparison of results also from repetitions of a whole biological experiment in the same experimental conditions. These replicate expression data present much higher variability than those from experiments starting from the same RNA pool and replicating only hybridization and quantification processes.

References

1. Lee M-LT, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA 2000; 97, 9834-9.
2. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J Biomed Optics 1997; 2, 364-74.