MulCom: a novel program for the statistical analysis of genomic data obtained on multiple microarray platforms

Riccardo Roasio[1], Li-Min Fu[1], Marco Botta[2] and Enzo Medico[1]

[1] Institute for Cancer Research and Treatment, University of Torino School of Medicine, 10060 Candiolo, Italy.
[2] Department of Computer Science, University of Torino, 10149 Torino, Italy.

Keywords. Microarray, statistics, permutations, new software.

Introduction

The increasing pace at which DNA microarray-based genomic expression profiles are generated and published poses the issue of efficient and reliable comparison between datasets obtained by different laboratories and on different microarray platforms. Statistical analysis of microarray data is in continuous evolution, and several procedures have been described for detection and weighing of systematic and random errors coming from the highly parallel -but poorly replicated- microarray expression data [1]. However, data obtained from different microarray platforms may be of substantially different nature. This is particularly evident when comparing two commonly used platforms, spotted cDNA microarrays and High-Density Oligonucleotide (HDO) microarrays of the Affymetrix type. cDNA microarrays yield a reproducible ratio between two signals, deriving respectively from the reference and from the sample. Conversely, absolute signals tend to vary across microarrays. Therefore, cDNA microarray data have to be analyzed with statistics handling repeated measurements or paired data, such as paired T-test. In the case of HDO microarrays, an absolute signal level is obtained from each single mRNA sample. As a consequence, non-paired statistics have to be applied to this type of data.

Given the intrinsic differences between cDNA microarrays, data analysis procedures have generally been developed on one of the two platforms and only in some cases adapted to the other, however without a specific focus on systematic comparison and validation across platforms. It is still unclear whether data obtained in the two systems can be treated, compared and eventually merged under a common analysis framework. We addressed these issues by generating expression profiles from the same RNAs with both microarray platforms and by developing an analysis procedure in which inter-platform differences in data treatment are reduced to the minimum essential. We then developed a novel statistical test specifically designed to handle multiple comparisons against the same reference condition (eg many points of stimulation against one unstimulated control). In the Multiple Comparison (MulCom) test, regulated genes are identified by a ‘tunable’ statistic test weighing expression change in each stimulation point against replicate variability calculated across the whole set of stimulation points.

MulCom

1. Preliminary data treatment

Data obtained from cDNA and HDO microarrays have a different format and a different nature; therefore, data from the two platforms had to undergo independent preliminary treatments aimed at generating a common format for homogeneous statistical analysis and comparison. Data from each cDNA microarray, obtained as tab-delimited text files from Incyte, had first to undergo Cy3/Cy5 normalization to eliminate the non-linearity in fluorochrome ratios. Quantile normalization of Cy5 values against Cy3 values allowed efficient non-linearity correction. Comparison between stimulated samples and the control involved calculation and averaging of the two Cy5/Cy3 $\log_2$ratios, and calculation of Standard Deviation (SD) for each ratio pair. Data from GeneChip microarrays, were obtained as tab-delimited text files through the Bioconductor R package, using quantile normalization (www.bioconductor.org). Stimulation points were compared to the controls through pair-wise $\log_2$ ratio calculation and averaging, and calculation of SDs.

2. The statistical test

At the end of the preliminary data treatment, the following data were obtained for each gene, in both platforms: 1) Average $\log_2$ratio between stimulations and control; 2) SD for each average $\log_2$ratio. In our experimental case of stimulated/control comparisons, both platforms provide the average ratios of six different stimulation points versus the control, and the SD of each ratio (each point done in duplicate). To obtain a more reliable estimate of variability, we calculated the Root Mean

Square of the six SDs (RMS SD) for each explored sequence. We hypothesized that while the SD of a single duplicate can easily be aberrantly high or low by chance, the RMS of SD from six duplicates is a more stable and reliable parameter.

These parameters were included in a statistical test aimed at identifying significantly regulated genes. The test is named MulCom as it is specifically designed for multiple comparisons against the same reference/control. The general structure of MulCom is that of a t-test:

$$\frac{M-T}{MAXs_M} \geq m$$

where M = Mean, average log2 ratio (absolute); T = threshold log2ratio; $MAXs_M$ = the higher between the SD of the duplicate and the RMS SD. The test is passed if the average $\log_2$ratio is greater than a certain threshold T and greater than the SD. Test tuning consists in modifying the T value and/or the m value.

3. The MulCom Program

MulCom is written in c++. It reads tab-delimited text files with normalized microarray data, formatted according to precise indications. The program checks the format and generates tables containing, for each gene, (1) Average log2 ratio; (2) Standard deviation of the AVG log2Ratio; (3) RMS SD. Starting from these three tables the MulCom test is calculated for each gene, according to the algebraically equivalent formula:

$$\text{Absolute(Average Log2 Ratio)} - \text{m*[MAX(SD,RMSSD)]} \geq \text{T}$$

where m and T are taken as input from the user. The user can also decide if the program uses the max between SD and RMS-SD or use only one of the two.

The program can also make n permutations of the columns' order to generate randomized datasets, and estimate the percentage of false-positives (False Discovery Rate, FDR, [2]). The user may decide the randomization seed, which allows re-generating the same permutations in subsequent or parallel sessions. The MulCom program has also implemented the classical t-test, for comparisons.

Conclusions

Data permutations and extensive false discovery analysis allowed accurate, platform-specific optimization of the MulCom test. Comparison analysis indicated that the MulCom approach, by weighing the overall SD (RMSSD), and by introducing a threshold value for the fold-change, is much more sensitive and specific than traditional t-tests. Moreover, it is equally applicable to microarray data from different platforms.

At this time MulCom is a standalone program, but it will be modified to use the computer grid architecture for parallelization, to reduce the computational time necessary to re-run the test on a high number of permutations, which is necessary for the correct estimate of the FDR.

References

[1] Nadon, R., & Shoemaker, J. (2002). Statistical issues with microarrays - processing and analysis. Trends in Genetics, 18, 265-271.

[2] Tusher VG, Tibshirani R, Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response.