Biomathematical information compression and signal extraction for gene expression microarray data analysis

Sofiane Lariani[1], Elena Comelli[1], Patrick Descombes[2] and Martin Grigorov[1]
[1]Nestlé Research Center, CH-1000 Lausanne 26 – [2]NCCR, Geneva (Switzerland)

Nestlé Good Food, Good Life

Introduction

Gene expression Microarray data covers both signal (active key genes) and noise. *A tentative to recognize any genes expression pattern without clearing the data is like recognizing faces and details on a jammed TV screen: it is hard and we have a high chance to be mistaken.*

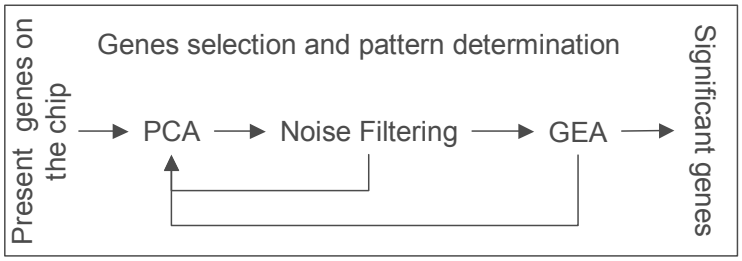
Noise in this context can be related to fluctuation due to the experimental condition, to outliers, and also genes presenting no modulation between groups. In this work we present a biostatistical method for noise filtering and statistical significance assessment. This method is based on combination of:

❑ Multivariate approach, Principle Component Analysis (PCA) for a first data compression and signal extraction and

❑ Univariate approach using robust statistics for the estimation of statistical significance of the selected genes.

The built workflow provides the biologist a flexible but robust way of knowledge extraction to recognise expression pattern with high confidence in the identified modulated genes.

Methodology



❑ **Class determination and genes selection:**

✓ We use the PCA on the data to generate a series of uncorrelated variables by looking for linear combinations that create the largest spread among the values. The axes of the new space are obtained by maximizing the variance within the data. No particular assumption (as equal variance) will be made for the PCA except that the mean vector and the covariance exist.

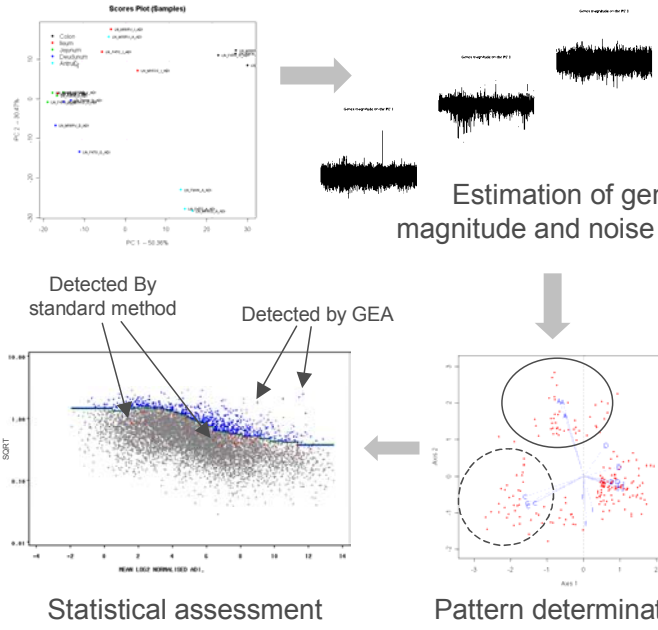
✓ After which, based on loading distribution, we select genes that contribute highly in the observed energy between clusters.

✓ Then we map score (chips representation) and loading (genes representation) on the same plot (called biplot); As a consequence of the matrix factorisation the modulation of any gene in a chip can be geometrically inferred from the biplot.

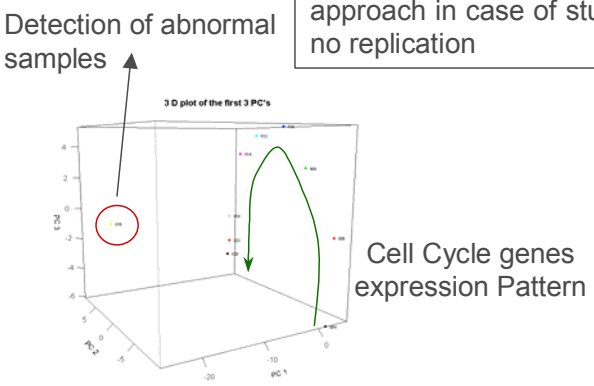
❑ **Statistical significance:**

In order to mitigate the effects of a low number of replicates and to precisely estimate the error, we use Global Error Assessment to determine statistically significant genes difference. Using this method we have higher confidence in the results of the test.

Results



Efficiency of the multivariate approach in case of study with no replication



Conclusions

In order to increase the confidence of researchers in the gene expression data that identify a real biological change, the noise filtering at an early stage seems primordial. The combinations of classical multivariate methods, such as PCA with robust statistics strengthen the credibility of the results by decreasing the number of false discoveries. This approach was tested on different studies and gave promising results. For the purpose we used the publicly distributed statistical language R (www.r-project.org) and Bioconductor the open source and open development software project for the analysis and comprehension of genomic data (www.bioconductor.org)

References

1. R. Mansourian, The Global Error Assessment (GEA) model for the selection of differentially expressed genes in microarray data. Bioinformatics 20(16): 2726-2737 (2004)
2. E. Marshall, Getting the Noise Out of Gene Arrays Science, Vol 306, 630-631 , 22 October 2004
3. Hardle, "Applied multivariate statistical analysis". Springer-Verlag, 2003
4. Y. E. Pittelkow, Visualisation of Gene Expression Data – the GE-biplot, the Chip-plot and the Gene-plot, Statistical Applications in Genetics and Molecular Biology: Vol. 2: No. 1, Article 6. 2003.

Contact: sofiane.lariani@rdls.nestle.com

Nestlé Research Center