# LETS ROC ON MICROARRAYS

Carina Silva-Fortes[1,3], Maria Antónia Amaral Turkman[2,3], Lisete Sousa[2,3]

[1] Higher School of Health Technologies of Lisbon-Portugal
[2]Faculty of Sciences of University of Lisbon-Portugal
[3]Center of Statistics and Applications of University of Lisbon-Portugal

**Abstract**
There are new statistical challenges posed by data from microarray experiments, due to the exploratory nature of experiments and the huge number of genes under investigation. There are many statistical techniques to analyze those data but sometimes they are too difficult to implement. We present the advantages of the application of receiver operating characteristic (ROC) analysis in microarray data analysis, in particular on selection of genes that are differentially expressed (DE) in different know classes of tissue. We also present one example of application of ROC analysis to select genes that are differentially expressed and compare the results with dChip analysis.
**Key-words:** ROC curves, microarrays, multiple testing, optimal cut-off, differential expression.

## Introduction

Microarrays are a new technology where entire genomes can be accessed at once. Besides of being more efficient than the classical gene-by-gene approach, this opens up entirely new avenues for research, on particulary in statistics (Goor, 2005).

A microarray measurement is the result of a multi-step experiment: manufacturing the arrays, acquiring and preparing the mRNA from the cells, hybridizing them to the arrays, and finally, scanning the hybridized arrays. Each of these steps contributes to variability of the data, some of which is caused by biological differences, by technology, environment, etc.

To minimize those variations preprocessing steps such as image analysis, background correction, normalization and summarization are needed before DE genes selection. These steps strongly depend on the platform used. In this work we use Affymetrix microarrays, where multiple distinct oligonucleotide probes on each chip represent every gene (Affymetrix, 1999).

All those steps involve distinction or discrimination between two alternative states, like spots classification in faulty and good spots on image analysis. Figure 1 shows where a dichotomy decision is needed on microarray data analysis.

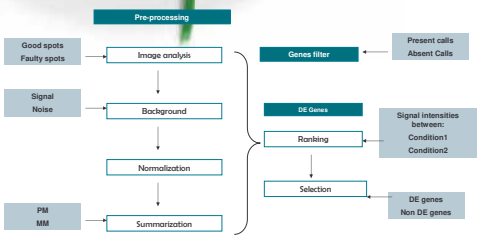


Figure 1: Dichotomy decisions on microarray data analysis

There are several methodologies for all those steps, although a gold standard is not yet availabe.

Receiver operating characteristic (ROC) methodology is appropriate in situations where there are two possible true states (Zweig and Campbell, 1993; Baker, 2003). In assessing the performance of a test, the question is whether the test result distributiond from the two states differ? If they do not differ, obviously the test results cannot discriminate between the two groups. If the distributions do not overlap at all, then there is a perfect discrimination. Most often the distributions of the test results are partially overlapped (figure 2). We can use ROC methologies in each step of microarrays analysis where a dichotomy decision is needed (figure 1).

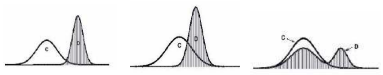

Figure 2: Overlap between two distributions

ROC analysis provides a numerical and graphical representation of the tradeoff between false-positive rates and true-positive rates that are produced by any given test, and allows estimation of optimal cutoff levels for discrimination and filtration of test results. We can use ROC analysis on microarray data sets for two approaches: to compare different methodologies for microarray data analysis and select the best one; to select between two truth states.

No statistical test has the ideal combination of 100% specificity and 100% sensitivity, and the ROC curve documents the combination of specificity and sensitivity achieved in a given test. The area under the curve (AUC) (1) is as an index of accuracy. The AUC ranges between (0.5;1) and the accuracy of a given test is higher for values of AUC closer to 1. In general a gold standard is needed, to construct a ROC curve, but Affymetrix has spike-in data sets (McGee and Chen, 2006) where the 'spiked-in' experiment provides a controlled dataset with known sequence and known concentration.

ROC curves can play an important role in identifying DE genes. Suppose a gene expression microarray experiment compares specimens from subjects with two different phenotypes (e.g. Control (C) and Disease (D)). In this case, AUC represents the amount of overlapping of the two samples distribution. When AUC is near to 1 we have a DE candidate gene. If the sample size is small, as is often the analysis is generaly limited to ranking genes by differential expression. For ranking genes, Pepe (2003) proposed computing the partial area under the ROC curve (pAUC) (2) near a low false-positive rate ($t_0$) of interest.

To compare the distributions of two populations ($C$ = control, $D$ = disease) let $X^i g$ denote the expression level of gene $g$ in sample $i$ = C,D after normalization. Each point of the ROC curve ($t$, $ROC(t)$), corresponds to a different expression level $u$, where:

$$t = 1 - P(X_g^C < u) \quad \text{1-specificity}$$
$$ROC(t) = P(X_g^D \geq u) \quad \text{sensitivity}$$

$$AUC = \int_0^1 ROC(t)dt \quad (1) \qquad pAUC(t_0) = \int_0^{t_0} ROC(t)dt \quad (2)$$

Several statistical methods for microarray gene selection have been explored, although all of them have a weakness, namely the choice of the threshold for the decision. Fold-change thresholds has been the most commonly used method for filtering false positives and declaring significant changes, usually varying from 2 to 6 fold. Such constant thresholds tend to produce false positives when signal intensities are low and false negatives when signal intensities are high (Li *et al.*, 2005). However it remains an open question, how threshlds for significant changes should be determined.

ROC analysis can provide an optimal selecting threshold. There are several approaches for selecting an optimal cutoff value (Greiner *et al.*, 2000; Bilban, 2002). The cutoff selection criteria should take into consideration the maximization of sensititivity and specificity, minimizing costs associated with bad decisions, etc.

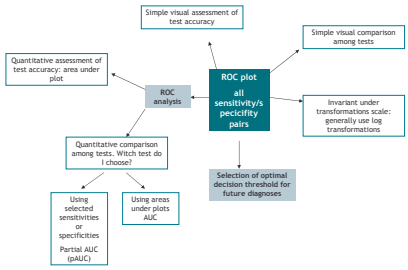


Figure 3: Properties of ROC analysis

## Application

The expression set to be used in this application is Huang.RE which is discussed in Huang *et al.*,(2003). The data contains microarrays of 52 women with breast cancer of whom 34 did not experience a recurrence of the tumour during a 3 years time period.
The scientific objective in Huang *et al.* (2003) study is to find gene expression as predictors on breast cancer outcome. Does the data help us to find new biomarker which can be used in the follow-up of cancer patients, for example to diagnose recurrence of the tumour?

This application follows the arguments given by Pepe *et al.* (2003). They argue as follows: in general, scientists are more interested in identifying genes that are over expressed, rather than under-expressed, in cancer diagnostic research.

The data consists of VSN normalised expression measures which is summarised by median polish. The array contains 12625 probe sets. We are interested in looking for important genes which have a good chance to be differentially expressed between both groups.

We say that gene $g$ is differentially expressed if the distribution of the gene expression in the two groups is different.

Figure 4 shows the boxplot for both distributions "Recurrence" and "No recurrence" for gene "1454_at" expression values and figure 5 shows a ROC curve for the same gene.

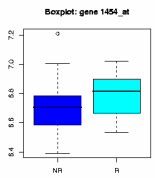


Figure 4: Boxplot for gene "1454_at"

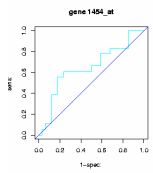


Figure 5: ROC curve for gene "1454_at"

An important measure for the quality of separation is the area under the ROC curve, the AUC. The AUC for the probe set "1454_at" is 0.6471. This value indicates that gene is not a good candidate to be a biomarker, so this gene wouldn't be analyzed further.
We calculated the AUC for all 12625 genes and we selected those for which the AUC is higher than 0.9. Hence we selected 7 biomarkers candidates. We also calculated the partial area above the ROC curve (pAUC) for genes who had AUC higher than 0.9, with false positive rate (2), t0=0.1. Finally we ranked those genes and selected those for which the pAUC was higher than 0.05. The number of genes selected was 4 (Table 1).

| Gene | AUC>0.9 | pAUC(0.1) |
|---|---|---|
| 32625_at | 0.93 | 0.0565 |
| 33706_at | 0.90 | 0.0261 |
| 35222_at | 0.90 | 0.0382 |
| 38795_at | 0.91 | 0.0339 |
| 38895_at | 0.94 | 0.05621 |
| 39280_at | 0.91 | 0.0588 |
| 965_at | 0.90 | 0.0513 |

Table 1: DE genes

The R packages with the functions used for the data analysis in this work are part of the Bioconductor project (Gentleman, 2004).

## Acknowledgments

This research was partially funded by the project FCT/POCI 2010.

## References

•Affymetrix (1999). *Gene Chip Analysis Suite User Guide*. Affymetrix, Santa Clara, CA.

•Baker, S.G. (2003). The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *Journal of the National Cancer Institute*, Vol 95, nº. 7.

•Bilban, M.; Buehler, L.K.; Head, S.; Desoye, G. And Quaranta, V. (2002). Defining signal thresholds in DNA microarrays: exemplary application for invasive cancer. *BMC Genomics*, 3:19.

•Gentleman, R.C. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5, R80.

•Goor, T.A. (2005). A History of DNA Microarrays. *Pharmaceutical Discovery*. [www.pharmadd.com].

•Greiner, M.; Pfeiffer, D. and Smith, R.D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*, 45, 23-41.

• Huang, E.; Cheng, S.H.; Dressman, H.; Pittman, J.; Tsou, M-H; Horng, C-F; Bild, A.; Iversen, E.I.; Liao, M.; Chen, C-M; West, M.; Nevins, J.R. and Huang, A.T. (2003). Gene expression predictors of breast cancer outcomes. *The Lancet*, *361:159-1596.*

•Irizarry, R. A.; Hobbs, B.; Collin, F.; Beazer-Barclay, Y. D.; Antonellis, K. J.; Scherf, U. and Speed, T.P. (2003) Exploration ; Normalization and Summaries of High Density gonucleotide array probe level data *Biostatistics* 4, 249-64.

•Li, X.; Kim, J.; Zhou, J.; Gu, W. and Quigg, R. (2005). Use of signal thresholds to determine significant changes in microarray data analyses. *Genetics and Molecular Biology*, 28, 2,191-200.

•McGee, M. and Chen, Z. (2006).New Spiked-in Probe sets for the Affymetrix HGU-133A latin-square experiment. COBRA,Paper 4.

•Pepe, M.S. . S.; Longton, G. M.; Anderson, G. L. and Schummer, *M.* (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics*, 59:133-42.

•Saviozzi, S. and Calogero, R.A. (2003). Microarray probe expression measures, data normalization and statistical validation. *Comparative and Functional Genomics*. 4:442-446.

•Zweig, M. H. and Campbell, G. (1993). Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. *Clinical Chemistry*, vol 39, n.°4.