

Microarrays and Statistics

Microarray e statistica

Mauro Gasparini
Dipartimento di Matematica
Politecnico di Torino
gasparin@calvino.polito.it

Annibale Biggeri ¹
Dipartimento di Statistica
Università di Firenze
abiggeri@unifi.it

Riassunto: L'importante tecnologia post-genomica dei microarrays è brevemente presentata e quindi discussa nei suoi aspetti statistici. In particolare, viene posto l'accento sulle tecniche riassuntive, inferenziali e sperimentali con cui uno statistico può contribuire all'analisi di questi nuovi tipi di dati.

Keywords: microarrays, design of experiments, gene expression, Bayesian networks.

1. A short description of the biology

The past decade has seen the accomplishment of the decoding of the human genome, so that now we know the anatomy of the genes, but we do not know their physiology or pathology. Here lies the great challenge of the years ahead, as we are entering the postgenomic era of functional genomics.

The genome is a set of instructions for the production of, mainly, proteins with different biological functions. The translation of instructions from the nuclear DNA to the protein producers (the ribosomes) happens via the messenger RNA, mRNA. In the lack of knowledge of the gene-protein correspondences, biologists are measuring the quantity of mRNA expressed by genes, since a primary mechanism by which proteins are regulated is through variation of the amount of mRNA present in the cell.

Microarrays are a tool to measure mRNA. DNA clones from gene data-banks are selected, amplified by PCR and mechanically printed on coated microscopic slides. Total RNA (or mRNA) from test and reference cell samples are extracted and, through reverse transcription complementary DNA (cDNA) fragments (between 600-2400 bases) or oligonucleotide sequences (of exactly 70 bases, by industrial kits), are allowed to hybridize on the slides (microarrays), then measured. For example, cDNA or oligonucleotide may be labelled by fluorescent dyes and the corresponding intensities measured under a laser fluoroscope. A microarray provides simultaneously several thousands of such pairs of values, one for each of several thousand genes, resulting in a computerized image of relative intensities.

2. Bioinformatics and Statistics

Even from the very short description contained in the previous section, it should be clear that constructing very large databases and computer images based on microscopic

¹Lavoro svolto con i finanziamenti della Cassa di Risparmio di Firenze, 2001 Centro di Genomica Funzionale

biological entities involves a large effort in Information Technology. The production of data from microarrays makes for a big chunk of a new emerging science called Bioinformatics.

When it comes to the analysis of such data and their use in drawing scientific results though, microarrays are planned experiments and as such they require DOE (Design of Experiments) techniques and inferential statistical concepts rather than Informatics. The statistics deals here with a very large amount of data, a feature it shares with other current applications, such as data-mining in marketing or the analysis of web traffic. This paper is a discussion of some crucial aspects of the statistical analysis of microarrays. The discussion is done with reference to some of the relevant literature, but one has to keep in mind that the literature in the field is currently exploding and any attempt to reviewing it would be vain.

3. Data normalization and data reduction

Raw data from microarray experiments require careful quality control to avoid spurious results. For example, one particular type of dye may be more fluorescent; the background intensity may vary from spot to spot due to substrate dye fixation or glass natural fluorescence; there could be variations between print tip dimensions or deformations after many hours of printing. Normalization is the first step in the analysis of gene expression data, and it is usually treated as a descriptive issue, preceding statistical data analysis (see for example Yang *et al.* (2001)).

It is of utmost importance to separate the issues of data normalization and quality control from the subsequent task of data reduction. Reducing high dimensional data to a lower dimensional summary is of main interest in order for the researcher to be able to manage and interpret the evidence. It is very important that the reduction be done not just by an apparently reasonable algorithm but, as much as possible, by referring to the statistical principle of sufficiency (i.e. the reduction of data be done without losing useful information as far as the inferential conclusions are concerned).

For example, cDNA microarrays provide raw data which are, for each gene, fluorescence intensities for the red and the green dyes associated to two experimental conditions for each of many pixels in a single spot associated with the gene. Two data reduction steps are usually performed. First, the distribution over the pixels is summarized by a single number like the mean or the median of red and green, say R_i and G_i , $i = 1, \dots, N$, N being the number of genes. Next, a second data reduction using the ratio R_i/G_i is ordinarily used to check in which of the two experimental conditions the mRNA has revealed more abundant. A k -fold rule is usually applied on the ratio to decide whether the gene was over- or under-expressed.

It has been observed that the variability associated with the absolute intensity may be increasing with the intensity, something which would make the k -fold rule senseless. In other words, the ratio may not sufficient in capturing the information in the data. Therefore, a statistical model build on the original quantities may be more appropriate than a statistical model build on the ratios. This argument is the main thesis of the paper by Newton *et al.* (2001), where a Bayesian hierarchical model is built for the absolute intensities using formulae

$$R|\theta_R \sim \text{Gamma}(a, \theta_R) \quad , \quad G|\theta_G \sim \text{Gamma}(a, \theta_G)$$

as building blocks to express the R_i and the G_i as exchangeable quantities and to expand into a three-layer hierarchy for inferential purposes. The bottom line is that expression parameters are estimated simultaneously and the decision on the over- or under-expression of each gene is based on a statistical model rather than *ad hoc* algorithms.

Another example is Efron *et al.* (2000), where oligonucleotide microarrays, not cDNA, are discussed. There, the choice between a mean difference and a less intuitive function to measure differential expression is made based on exploratory statistical analysis.

4. Design of experiments

Following the simple k -fold rule would have the consequence to consider just the genes with large change in expression that are obvious to detect, wasting most of the information collected. In order to benefit from the new technology, scientists need statistically designed experiments and data analyses that not only produce estimates of relative expression but, in addition, an evaluation of the size of the possible errors of those estimates.

Following the DOE ideas expressed in Kerr and Churchill (2001) for example, the main sources of variability in a microarray experiment are "arrays" (A) "dyes" (D), "varieties" (V) (or experimental conditions) and "genes" (G), which can result in the following simplified linear model:

$$Y = A + D + V + G + AD + AV + AG + DV + DG + VG + \text{error}$$

or in a model with higher order interactions, if appropriate. The VG interactions are the effect of interest: identifying genes whose expressions change in different varieties. If the number of varieties is greater than two, the reader can recognize here an example of incomplete block design since a microarray can only accommodate for two varieties at a time.

The widely used reference design (one dye labels a reference variety and the other dye labels other varieties considered) has the major drawback that the VG effects are completely confounded with DG interactions. Moreover, having v varieties and n genes, the reference design gives $2vn$ observations: after accounting for main effects, VG and AG effects, no degrees-of-freedom are left to estimate error.

Biologists adopted the reference design since they recognized that the amount of cDNA varies from spot to spot and therefore fluorescent intensities are meaningful only in a relative sense. Statisticians can do better designs, for example the loop design proposed in Kerr and Churchill (2001).

5. Data classification and functional relationships

Identification of unknown classes using gene expression profiles (unsupervised learning, cluster analysis) or classification into known classes (supervised learning, discriminant analysis) are two common techniques used in gene expression experiments. The goal is to identify clusters of genes that behave in similar ways under different experimental conditions.

Heuristic algorithms have been proposed so far. Clustering algorithm based on probability models offer an alternative whereby the data is assumed to be generated by a

finite mixture of underlying probability distributions. Specifically, following Yeung *et al.* (2001), let the data $\mathbf{X}$ consist of independent multivariate observations, $X_i, i = 1, \dots, n$ and let the component (group) be denoted by $k = 1, \dots, G$. The likelihood is:

$$L = \prod_{i=1}^n \sum_{k=1}^K f(X_i | \theta_k) \pi_k$$

where π_k is the probability to belong to the k -th component. One important feature of mixture modelling is that posterior probabilities of class membership are obtained.

However, data classification does not directly address the very important point of functional relationships among genes, the interest here being in which gene influence which physiological or pathological processes. The study of the observed conditional associations between genes via Bayesian Networks appears to be promising, even in the ill-defined framework of microarray experiments.

A Bayesian Network consists of two distinct parts: a directed acyclic graph (DAG) and a set of parameters for the DAG. From a given sample we can perform a series of tests of conditional independence or assign a score to each DAG and select the one with the highest one. Friedman *et al.* (2000) discusses the long-term prospects in discovering causal patterns, but warns that Bayesian Networks require larger sample sizes, are sensitive to the choice of local model and discretization method, and do not allow for non-linear dependencies.

6. Momentum

In February 2001 a group of biologists, bioinformaticians and statisticians gathered at the Advanced Biotechnology Center in Genoa, Italy, to discuss the multidisciplinary aspects of the analysis of the booming technology of microarrays, and finalized the first comprehensive Italian workshop on the field, "Tecnologia degli Oligo- e DNA-microarray e analisi statistica dei dati" held in Genoa, December 14-th 2001.

References

- Efron B., Tibshirani R., Goss V. and Chu G. (2000) Microarrays and their use in a comparative experiment, Technical report, Department of Statistics, Stanford University.
- Friedman N., Linial M., Nachman I. and Pe'er D. (2000) Using Bayesian networks to analyze expression data, *Journal of Computational Biology*, 7, 601–620.
- Kerr M. and Churchill G. (2001) Experimental design for gene expression microarrays, *Biostatistics*, 2, 183–201.
- Newton M., Kendzioriski C., Richmond C., Blattner F. and Tsui K. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data, *Journal of Computational Biology*, 8, 37–52.
- Yang Y., Dudoit S., Luu P. and Speed T. (2001) Normalization for cDNA microarray data, Technical report, Department of Statistics, UC Berkeley.
- Yeung K., Fraley C., Murua A., Raftery A. and Ruzzo W. (2001) Model-based clustering and data transformations for gene expression data, Technical report, University of Washington Statistics Department.