

Pooling microarray data for assessing treatment/exposure effects: a bayesian approach

Microarray con miscele di materiale genico per valutare effetti di trattamento/esposizione: un approccio Bayesiano

Annibale Biggeri, Marta Blangiardo¹

Department of Statistics "G. Parenti"

University of Florence

abiggeri@ds.unifi.it, blangiar@ds.unifi.it

Summary: In microarray experiments we need to pool mRNA from different samples by technical reasons before the array production phase. This study design permits to save materials and to reduce costs of analysis, with the drawback of making impossible to estimate the biological variability. We propose a full range of study designs in which several independent pools of individual samples are randomly selected. This strategy allows to reduce costs and to estimate the biological variability. Inference from these alternative designs can be done using a Bayesian approach. Hierarchical modelling could be a powerful instrument to distinguish variance components and to identify differentially expressed genes.

Keywords: Microarray analysis, Bayesian hierarchical model, Design of experiment.

1. Introduction

Microarray technology is a powerful tool to analyze and classify thousands of genes at the same time. Currently gene expression analysis is used heuristically to gain insight into protein production and regulation. Moreover, it promises to revolutionize the traditional approach to the analysis of gene function, allowing to study thousands of genes at the same time.

This new technology has influenced many research areas and it has involved also the statistical field. Actually, the analysis of microarray implies both the development of new statistical methods, and the adapting of already known procedures. The main steps in microarray data analysis start from image analysis (Chen *et al.* (1997), Yang (2002a)), normalization (Yang *et al.* (2002), Tseng *et al.* (2001)) and inference of differentially expressed genes. The validity of the analysis relies, however, on the validity of the experimental design, which determines how samples have been selected and how biological specimens have been processed. Microarray experiments at the same time present peculiarities and questions to be tackled by study design options from both statistical and biological point of view.

In the present work we review the different aspects concerning the design of experiment in microarray analysis, revising the more popular approaches in the literature and then focussing on the analysis of pooled data from different statistical units

Finally, we discuss advantages and drawbacks of pooling data and formulate a Bayesian hierarchical model to analyze such experiments.

¹59, viale Morgagni, 50124 Florence (Italy)

2. Microarray technologies

There are many microarray technologies and the more used are two: cDNA microarray and oligonucleotide arrays. They differ in how the sequences of DNA are processed to produce the array and in the length of the sequences used. We treat mostly the first technology, but in this section we wish to introduce both briefly. In cDNA microarray mRNA from two different samples is reverse-transcribed to form cDNA, labelled with two different dyes and hybridized on a glass slide. After hybridization the image is scanned and the fluorescence of the two dyes are calculated. Quantitative evaluation of gene expression is conducted in a comparative way, that is comparing, for a given gene, level of expression under the two different experimental conditions.

The second technology uses high-density oligonucleotide arrays (mostly Affymetrix). In such array the expression level of each gene is measured by hybridization reaction of the sample of mRNA to a set of probes, composed of 11-20 pairs of oligonucleotides, each with a predefined base pairs. The first type of probe in each pair is known as Perfect Match (PM) and is taken from the gene sequence, while the second is known as Mismatch (MM) and is created changing the middle base of the PM sequence to control the rate of unspecific binding. A single RNA sample is prepared, labelled with a dye and hybridized on the array. Therefore, differently from cDNA microarray a single sample is hybridized onto the array and, after scanning, an image is produced and analyzed to obtain a fluorescence intensity for each probe. These intensities represent how much hybridization occurred for each oligonucleotide.

During the complex procedure of microarray fabrication, many sources of variability arise and can obscure the biological signal of interest. In particular, it is convenient to classify them into five phases of data acquisition: microarray manufacturing, preparation of mRNA from biological samples, hybridization, scanning and imaging. Each of these phases can introduce an amount of systematic variation and artifactual that should be taken under control. To this aim, quality controls and normalization procedures are used to detect and eliminate the artifacts and the systematic variations, both within a single array and across arrays. However, even after normalization some sources of variations can remain and these can be classified as biological variation and technical variation. Speed (Speed (2003)) distinguishes two version of biological variability: one if mRNA is extracted from different samples from a given cell line or tissue, and the other if mRNA samples come from different individuals in the same species: the variability associated to the first case is lower than the second one. Technical variability arises when different hybridizations are performed using mRNA extracted from the same biological sample.

3. Design of experiment

We concentrate on design of experiment for cDNA microarray; however all the considerations can be easily extended to oligonucleotide arrays. For the array based on cDNA the elementary design is a (incomplete) block design of size two: in each spot gene material from treatment and reference is hybridized in the same conditions. The measured intensities on two channels of fluorescence are compared to each other. This design is originated from the knowledge of the principal sources of variability in this type of measurement and try to control them by means of comparisons on the same spot (gene, print-tip and array being equal). The important measure for the biologist is a relative measure,

the fluorescence intensity (proportional to the level of gene expression) for the channel related to the treatment with respect to the reference: the fold change in a logarithmic scale. Normalization procedures are used to remove the source of variability; alternatively it has been proposed to control the dye effect by design: the dye swap design provides for replicates with inverse assignment of fluorochrome to the two conditions.

In applied research literature there are two types of designs most frequently used: the reference design and the loop design. The first is used to compare each variety of treatment or exposure to the reference (i.e. for three categories: T1 vs ref, T2 vs ref, T3 vs ref) while the second compares each pairs of treatment changing the assignment of treated and non treated (i.e. for the same three categories: T1-green vs Ref-red; T2-green vs T1-red; Ref-green vs T2-red), eventually balancing the assignment of fluorochrome. Each design has advantages and drawbacks: if, on one hand using the first the researcher collects more material for the reference category, that could be the less interesting from a scientific point of view, on the other hand the second one lacks of robustness (Kerr (2003)) and should be avoided in experimental situations where array hybridizations could fail and cannot be replaced. The underlying framework of both of these designs is a balanced incomplete block design (BIB). It has a long history in statistical experimental design (Cochran and Cox (1957)) and seems very natural in cDNA microarray analysis where each spot is a block.

Moreover, to choose the appropriate design for a specific experiment, it is important to understand the goals of microarray experiments. To this aim, we can recognize two principal goals: one is the identification of differentially expressed genes among several varieties (class comparisons, Bittner *et al.* (2000)), while the other is the discovery of clusters within a collection of samples (class discovery, Alizadeh *et al.* (2000)). Class comparison is related to exposure or treatment (i.e. comparison of gene expression for different tumors, for a population of smokers and non smokers) and the comparison between two varieties are performed directly (i.e. loop design) or indirectly (i.e. reference design). Viceversa, class discovery is based on distances between gene profiles of pairs of samples. Dobbin and Simon (Dobbin and Simon (2002)) show that if the main goal of experiment is class discovery, then a reference design is preferable to a loop design, while if the class comparison is the main goal and there is enough material available, a BIB design could be the most efficient.

Nevertheless, if the material is not sufficient, other strategies are needed.

4. Pooling data: composite sampling

Applied researchers propose to pool RNA sample from different units (Jin *et al.* (2001)); this strategy is motivated by lack of material from a single unit to produce an array. Eventually the only things to do are either pooling samples or proceed to an amplification of RNA. There are two strategies to amplify RNA, based on PCR or on T7 system. Both amplifications are questionable when the amount of material is really small and the procedure could introduce source of bias that is not clearly identifiable. For this reason, very often the strategy followed is to performing a pool of individual samples. The design of experiment based on pools is known in literature as "Composite sampling" (Elder (1977), Boswell and Patil (1987)) and began to be used during the World War II, when researchers wanted to analyze soldiers with respect to syphilitic antigens. Dorfman (Dorfman (1943)) showed that the number of tests could had been reduced if blood from

different subjects were mixed before testing ; then only groups in which presence of anti-gens was detected were retested: if they were few enough each subjects was then tested; otherwise they split groups in subgroups and performed again the analysis. This strategy has been used since in many applications on blood banks and in environmental studies to monitor waste disposal sites.

Moreover, if it is possible to classify individuals on the basis of some relevant features (i.e. healthy vs diseased), a composite strategy could be to form pools from the different sub-populations and to compare them.

This type of composite seems to be useful in microarray analysis whenever a specific treatment or exposure needs be evaluated in terms of expressed genes.

Many authors have contributed on estimation from composite sampling data: Gibbs and Gower (Gibbs and Gower (1960)), Kerr (Kerr (1971)). In particular, the two arguments treated are related to the following points:

- how does composite compare with traditional uncomposite sampling?
- how many samples and which size should be used?

5. Estimation of mean and variance

In order to answer to these questions let us start with a very simple situation: suppose we wish to compare mRNA from population A and mRNA from population B; from each of these, a sample of n individual is selected. A way to conduct the analysis is to pool all RNA from population A and to compare with RNA from population B (Figure 1); this design makes impossible any type of inference on variance components, because both biological variability (between subjects) and technical variability (within subjects) are not estimable; the only (limited) simple-array inference is on differentially expressed genes under a single class comparison design. However, if replicates are performed for each pool(Figure 2), it becomes possible to estimate technical variability, while, using a global pool for each population, the biological one remains reset. In this case, it is possible to apply a reference design and compute the average expression estimation $\bar{x}_1 = \sum_{j=1}^J x_{1j}$ and $\bar{x}_2 = \sum_{j=1}^J x_{2j}$ where x_{1j} is the expression value for the first pool and j -th replication, the sum is over all the J replications. Then it is possible to compute a t statistic:

$$\frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{2}/\sqrt{J}}$$

where the estimation of s is based on the J replicates.

If enough material would have been collected for each individual unit, we could have been able to perform a number of arrays equal to the number of individual units (Figure 3). Following this strategy, it would have been possible to estimate the biological variability, but the technical component of variability would have been confused with the error component. Inference on differentially expressed genes would follow the t-statistic strategy as before.

If replicates would have been performed, estimation of variance components, biological and technical, would be possible.

A feasible compromise between the two extremes could be to divide randomly the n subjects selected from each population in two subgroups (Figure 4). The comparison between the subgroups allows estimation of biological variability. Technical variability becomes estimable if replicates are performed.

Therefore we propose to use replicates and several independent pools from a population to estimate both the variance components even after normalization procedure.

6. Bayesian Linear model

We use the Bayesian approach to inference. We model the unnormalised log gene expression $y_{gs} \sim N(\mu_{gs}, \sigma_g^2)$ as a gaussian for gene g and channel $s = 1, \dots, 2 \cdot n_a$ (Lewin *et al.* (2003)).

So, for each gene, we have the normalization procedure modelled as follow:

$$\begin{aligned}\mu_{gs} &= m_g + \alpha_g + \delta + \tau_g + \nu_{1g} + \nu_{2g} \\ \nu_{1g} &\sim N(0, \sigma_\epsilon^2) \\ \nu_{2g} &\sim N(0, \sigma_\eta^2)\end{aligned}$$

where α_g is the array effect, δ is the dye effect, τ_g is the treatment effect, m_g is the expression value for the g gene after normalization and ν_{1g}, ν_{2g} are the variance components (biological variability and technical variability).

Different prior distribution will be used for the parameters of the model and a sensitivity analysis will be performed.

7. Figures

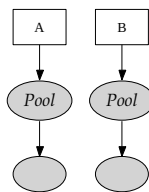


Figure 1: Global pool without replication

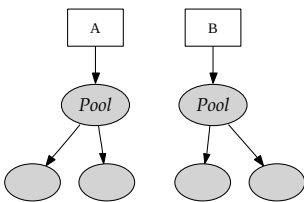


Figure 2: Global pool with 2 replications

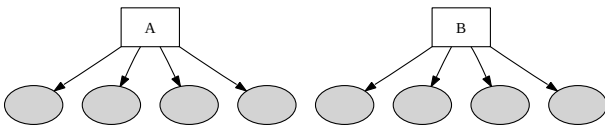


Figure 3: Single individual comparison without replications

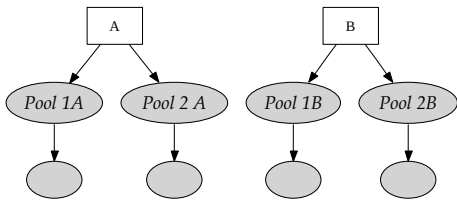


Figure 4: Independent pools without replications

References

- Alizadeh A., Eisen M., Davis R., Ma C., Lossos I., Rosenwald A., Boldrick J., Sabet H., Tran T. and Yu X. (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling, *Nature*, 403, 503–511.
- Bittner M., Meltzer P., Chen Y., Jiang Y. and Seftor E., Hendrix M., Radmacher M., Simon R., Yakhini Z. and Ben-Dor A. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling, *Nature*, 406, 536–540.
- Boswell M. and Patil G. (1987) A perspective of composite sampling., *Communications in Statistics, Part A Theory and Methods*, 16, 3069–3093.
- Chen Y., Dougherty E. and Bittner M. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images, *Journal of Biomedical Optics*, 2, 364–374.
- Cochran W. and Cox G. (1957) *Experimental designs*, Wiley.
- Dobbin K. and Simon R. (2002) Comparison of microarray designs for class comparison and class discovery, *Bioinformatics*, 18, 1438–1445.
- Dorfman R. (1943) The detection of defective members of large populations, *Annals of mathematical Statistics*, 14, 436–440.
- Elder R. (1977) *Properties of composite sampling procedures.*, Phd dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.
- Gibbs A. and Gower J. (1960) The use of a multiple-transfer method in plant virus transmission studies some statistical points arising in the analysis of results, *Annals of Applied Biology*, 48, 75–83.
- Jin W., Riley R., Wolfinger R., White K., Passador-Gurgel G. and Gibson G. (2001) The contribution of sex, genotype and age to transcriptional variance in drosophila melanogaster, *Nature Genetics*, 29, 389–395.
- Kerr J. (1971) The probability of disease transmission, *Biometrics*, 27, 219–222.
- Kerr M. (2003) Design considerations for efficient and effective microarray studies, *Berkeley electronic Press*.
- Lewin A., Richardson S., Marshall C. and Glazier A. and Aitman T. (2003) Bayesian modelling of differential gene expression, *submitted*.
- Speed T. (Ed.) (2003) *Statistical analysis of gene expression microarray data*, Chapman and Hall/CRC.
- Tseng G., Oh M., Rohlin L., Liao J. and Wong W. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects, *Nucleic Acids Research*, 29, 2549–2557.
- Yang Y. (2002a) Comparison of methods for image analysis on cDNA microarray data, *Journal of Computational Graphical Statistics*, 11, 1, 108–136.
- Yang Y., Dudoit S., Luu P., Lin D., Peng V., Ngai J. and Speed T. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research*, 30, e15.