Design and Analysis of Screening Experiments with Microarrays

Paola Sebastiani[1], Joanna Jeneralczuk[2], and Marco F. Ramoni[3]

[1] Department of Biostatistics, Boston University School of Public Health, Boston
sebas@bu.edu
[2] Department of Mathematics and Statistics, University of Massachusetts, Amherst jeneral@math.umass.edu
[3] Children's Hospital Informatics Program, Harvard Medical School, Boston
marco_ramoni@harvard.edu

Summary. Microarrays are nowadays used as exploratory tool in many screening experiments. The objectives of these experiments are several and include the identification of the genes that change expression in two or more biological conditions, the discovery of new cellular or molecular functions of genes, or the definition of a molecular profile that characterizes different biological conditions underlying for example normal or tumor cells. A very important question arising in the design of screening experiments with microarrays is the choice of the sample size. In this chapter, we first review the technology of microarrays and then describe some simple comparative experiments and some of the statistical techniques that are used for their analysis. We then discuss the issue of sample size and describe two approaches to sample size determination. The first approach is based on the concept of reproducibility, while the second approach uses a Bayesian decision theoretic criterion to trade off information gain and experimental costs. We conclude with a discussion of some of the open problems in the design and analysis of microarray experiments that need further research.

1 Introduction

One of the results of the Human Genome project is that the human DNA comprises between 30,000 and 35,000 genes. Only about 50% of these genes have known functions and several projects around the world are currently under way to characterize these newly discovered genes and to understand their role in cellular processes or in mechanisms leading to disease.

An avenue of research focuses on gene expression: the process by which a gene transcribes the genetic code stored in the DNA into molecules of mRNA that are used for producing proteins. The measurement of the expression levels of all the genes in a cell is nowadays made possible by the technology of microarrays (Lockhart and Winzeler, 2000). The basic intuition underlying

the technology of microarrays is that the genes responsible for different biological conditions may have different expression and hence produce molecules of mRNA in different amount. Microarray technology allows the measurement of the expression levels of all the genes in a cell, thus producing its *molecular profile*. By measuring the molecular profiles of cells in different conditions, researchers can identify the genes responsible for the different biological conditions as those with different expression level, or *differential expression*.

One important use of the microarray technology is the generation of scientific hypotheses: many microarray experiments are conducted to discover new genes that may have a role in particular biological process or may be responsible for disease. Because of their high costs, however, microarray experiments are often limited in sample size. From the experimental design point of view, the use of microarray technology as a hypothesis generator tool opens novel design and methodology issues. Even the design of a simple experiment conducted to discover the molecular profiles of two biological conditions opens basic issues such us the choice of the minimum sample size required to stake a reliable claim.

In this chapter, we review the technology of synthetic oligonucleotide microarrays and describe some of the popular statistical methods that are used to discover genes with differential expression in simple comparative experiments. We introduce a novel Bayesian procedure to analyze differential expression that addresses some of the limitations of current procedures. We proceed by discussing the issue of sample size and describe two approaches to sample size determination in screening experiments with microarrays. The first approach is based on the concept of reproducibility, while the second approach uses a Bayesian decision theoretic criterion to trade off information gain and experimental costs. We conclude with a discussion of some of the open problems in the design and analysis of microarray experiments that need further research.

2 Synthetic Oligonucleotide Microarrays

The modern concept of gene expression dates back to the seminal work of Jacob and Monod (1961) and their fundamental discovery that differential gene expression — when and in what quantities a gene is expressed — determines different protein abundance that induces different cell functions. During its expression, a gene transcribes its DNA sequence combining the nucleotides A, T, C and G into molecules of mRNA (messenger ribonucleic acid) that are then transported out of the cell nucleus and used as a template for making a protein. This two-step representation of the protein-synthesis process constitutes the *central dogma of molecular biology* (Crick, 1970).

Because the first step of a gene expression consists of copying its DNA sequence into mRNA molecules, the amount of mRNA molecules provides a quantitative measure of the gene expression level. The basic idea behind microarray technology is to measure the expression level of all genes in a cell by

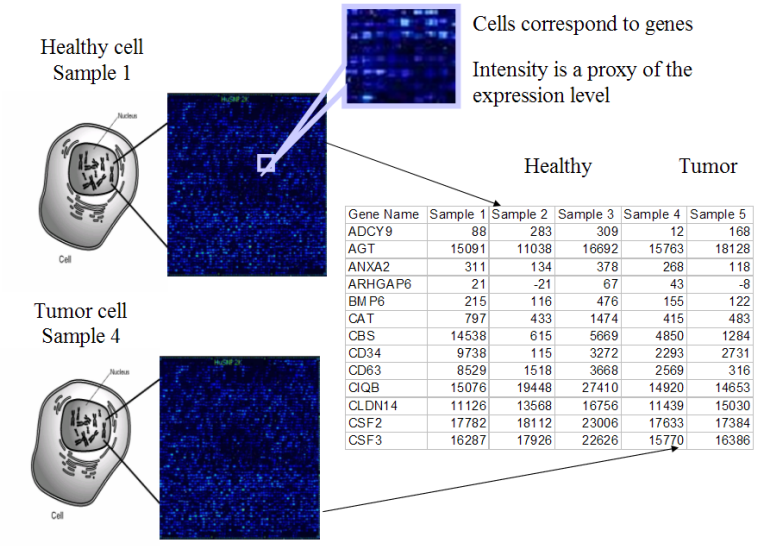
measuring the mRNA abundance of each gene. This is achieved by exploiting one property of the DNA sequence and the mRNA molecule produced during the gene expression: the two molecules bind together at a particular temperature. This fact is known as *hybridization* (Lennon and Lehrach, 1991).

There are different technologies for microarrays and we remind to Chapter **??** and the review in Sebastiani et al. (2003a) for a description of cDNA microarrays. Here, we focus on synthetic oligonucleotide microarrays. Technically, a synthetic oligonucleotide microarray is a platform gridded in such a way that each location of the grid corresponds to a gene and contains several copies of a short specific DNA segment that is characteristic of the gene (Duggan et al., 1999). The short specific segments are known as *synthetic oligonucleotides* and the copies of synthetic oligonucleotides that are fixed on the platform are called the *probes*.

The rationale behind synthetic oligonucleotide microarrays is based on the concept of probe redundancy: a set of well-chosen probes is sufficient to uniquely identify a gene. Therefore, synthetic oligonucleotide microarrays represent each gene by a set of probes unique to the DNA of the gene. On the GeneChip® platform, each probe consists of a segment of DNA, and each gene is represented by a number of *probe pairs* ranging from 11 in the Human Genome U133 set, to 16 in the Murine Genome U74v2 set and the Human Genome U95v2. A probe pair consists of a perfect match probe and a mismatch probe. Each perfect match probe is chosen on the basis of uniqueness criteria and proprietary, empirical rules designed to improve the odds that probes will hybridize to mRNA molecules with high specificity. The mismatch probe is identical to the corresponding perfect match probe except for the nucleotide in the central position, which is replaced with its complementary nucleotide, so A is replaced by T and viceversa, and C is replaced by G and viceversa. The inversion of the central nucleotide makes the mismatch probe a further specificity control because, by design, hybridization of the mismatch probe can be attributed to either non specific hybridization or background signal caused by the hybridization of cell debris and salts to the probes (Lockhart et al., 1996). Each cell of an Affymetrix oligonucleotide microarray consists of millions of samples of a perfect match or mismatch probe, and the probes are scattered across the microarray in a random order to avoid systematic bias.

To measure the expression level of the genes in a cell, investigators prepare the *target* by extracting the mRNA from the cell and making a fluorescence-tagged copy. This tagged copy is then hybridized to the probes in the microarray. During the hybridization, if a gene is expressed in the target cells, its mRNA representation will bind to the probes on the microarray, and its fluorescence tagging will make the corresponding probe brighter. Studies have demonstrated that the brightness of a probe is correlated with the amount of mRNA in the original sample. Therefore, the measure of each probe intensity is taken as a proxy of the mRNA abundance for the corresponding gene in the sample, and a robust average of the intensities of the probe set determines a relative expression for the corresponding gene. Full details are in

the Affymetrix document describing the statistical algorithm that is available from www.affymetrix.com/support/technical/whitepapers, and a summary is in Sebastiani et al. (2003a). Figure 1 sketches the three steps of a microarray experiment.



Gene Name	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
ADCY9	88	283	309	12	168
AGT	15091	11038	16692	15763	18128
ANXA2	311	134	378	268	118
ARHGAP6	21	-21	67	43	-8
BMP6	215	116	476	155	122
CAT	797	433	1474	415	483
CBS	14538	615	5669	4850	1284
CD34	9738	115	3272	2293	2731
CD63	8529	1518	3668	2569	316
CIQB	15076	19448	27410	14920	14653
CLDN14	11126	13568	16756	11439	15030
CSF2	17782	18112	23006	17633	17384
CSF3	16287	17926	22626	15770	16386

Fig. 1. A sketch of a microarray experiment. The mRNA in a cell is fluorescently labelled and hybridized to the microarray. After the hybridization, the intensity of each probe is captured into an image that is then processed to produce a proxy of the expression level of each gene in the target. Each microarray measures the molecular profile of a cell, and several microarray samples are needed to be able to detect the genes that have differential expression. In this figure, five microarrays were used to measure the molecular profiles of three healthy cells (Samples 1–3) and two tumor cells (Samples 4 and 5).

3 Design of Comparative Experiments

A typical microarray experiment produces the expression level of thousands of genes in two or more biological conditions. We denote the expression levels measured with microarrays by $y = \{y_{kji}\}$, where the index k specifies the kth gene in the microarray, $(k = 1, \ldots, p)$, and the index i denotes the ith sample measured in condition j. Because of technical and biological variability that are due to difficulties in the execution of the experiment and variability between different tissues used to extract the mRNA, more than one sample in

each biological condition is usually measured. We denote by n_j the number of samples measured in condition j so that $i = 1, \ldots, n_j$. Note that samples of the same biological condition may be pure replications or biological replications. In the first case, the target hybridized to the microarrays is made of mRNA extracted from the same cell while, in the second case, the target hybridized to the microarrays is made of mRNA extracted from different cells.

We term the set of expression levels measured for a gene across different conditions its *expression profile*, and we use the term *sample molecular profile* (or simply sample) to denote the expression level of the genes measured with one microarray, in a particular condition. Formally, the expression profile of a gene k in condition j will be the set of measurements $y_{kj} = \{y_{kj1}, \ldots, y_{kjn_j}\}$, the overall expression profile of the same gene across all conditions will be the set $y_k = \{y_{kj}\}_j$, and the ith sample profile of condition j will be the set of measurements $y_{ji} = \{y_{1ji}, \ldots, y_{pji}\}$.

Common experimental objectives are the identification of the genes with significant differential expression in two or more conditions, and the development of models that can classify new samples on the basis of their molecular profiles. In some experiments, the conditions may be controllable experimental factors such as doses of a drug or the time point at which to conduct the experiment. In general observational studies, which amount to a large proportion of microarray studies, the experimenter defines the conditions of interest (often disease and normal tissues) and measures the molecular profile of samples that are randomly selected. The study design are typically *case-control* (Schildkraut, 1998) with subjects selected according to their disease status: cases are subjects affected with the particular disease of interest, while controls are unaffected with the disease. For example, in an experiment conducted to identify the genes that are differentially expressed between normal lung cells and tumor lung cells, tissues from unaffected and affected patients are randomly chosen and each tissue provides the mRNA sample that is hybridized to the microarray.

In observational studies the main design issue is the choice of the sample size, while sample size determination and treatment choice are the primary design issues in factorial experiments. Sample size determination depends on the analytical method used to identify the genes with different expression and the optimality criterion. These topics will be examined in the next two sections

4 Analysis of Comparative Experiments

Popular techniques for identifying the genes with different expression in two biological conditions 1 and 2 are based on the t-statistic:

$$t_k = \frac{\bar{y}_{k1} - \bar{y}_{k2}}{SE(\bar{y}_{k1} - \bar{y}_{k2})},$$

where $\bar{y}_{kj}$ is the mean expression level of gene k in condition j, and the standard error of the sample mean difference, $SE(\bar{y}_{k1} - \bar{y}_{k2})$, is computed assuming different variances in the two conditions. Because of the large variability of gene expression data measured with microarrays, authors have suggested some forms of penalization for the denominator of the t-statistic. For example, Golub et al. (1999) suggest to compute the standard error $SE(\bar{y}_{k1} - \bar{y}_{k2})$ by the quantity

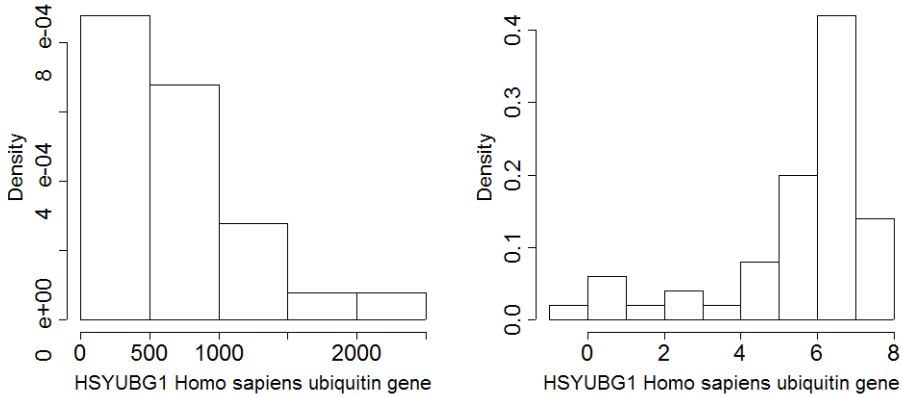
$$s_{S2Nk} = \frac{s_{k1}}{\sqrt{n_1}} + \frac{s_{k2}}{\sqrt{n_2}},$$

where s_{kj} is the sample standard deviation of condition j. The ratio $|\bar{y}_{k1} - \bar{y}_{k2}|/s_{S2Nk}$ is termed the *signal-to-noise ratio*. Other forms of penalization are justified by the fact that the standard error may be very small for genes with small expression values, thus inflating the value of the t-statistic. Based on this intuition, Tusher et al. (2000) suggest to adjust the standard error by $a + SE(\bar{y}_{k1} - \bar{y}_{k2})$ where the constant a is chosen to minimize the coefficient of variation of the t-statistic of all the genes. More recently, Efron et al. (2001) suggest to replace a by the 90th percentile of the standard error of all the genes.

The choice of the threshold to select the genes with a statistically significant change of expression is often distribution free. The main idea is to compute the value of a statistic from the data in which the sample labels that represent the conditions are randomly reshuffled. By repeating this process a large number of times, it is possible to construct the empirical distribution of a statistic under the null hypothesis of no differential expression. From this distribution one can select a gene specific threshold to reject the null hypothesis with a particular significance. Authors have also developed algorithms for multiple comparison adjusted p-values (Dudoit et al., 2001).

Distribution free methods tend to be widely used in practice, but they often require a large sample size to detect the genes with different expression and a small false positive rate (Zien et al., 2003). Some authors have suggested making distribution assumptions on the gene expression data, and the most popular choice is to assume that gene expression data follow a Lognormal distribution (Baldi and Long, 2001; Ibrahim et al., 2002). Another stream of work focuses on the estimation of the fold change of expression, that is, the ratio of the sample means assuming Gamma distribution for the gene expression data (Chen et al., 1997; Newton et al., 2001). We investigated the adequacy of these distributional assumptions on some large data sets available from http://www-genome.wi.mit.edu/cancer and none of these distributions appear to be, by themselves, appropriate for all genes.

An example is in Figure 2, which depicts the histogram of one sample of size 50 of the probe set corresponding to the "HSYUBG1 Homo sapiens ubiquitin" gene in the U95Av2 Affymetrix microarray. The distribution in panel (a) has an exponential decay, with a long right tail. The histogram in panel (b) displays the distribution of the log-transformed data and shows



(a) (b)

Fig. 2. Distribution of gene expression data from 50 prostatectomy samples measured with the U95Av2 Affymetrix microarray. (a): histogram of the expression level of the "HSYUBG1 Homo sapiens ubiquitin gene". (b): histogram of the same gene expression level after the log-transformation was used.

the fact that log-transforming the original data removes the left skewness by introducing a right skewness. This phenomenon is typically observed when log-transforming data that follow a Gamma distribution, with consequent bias induced to estimate the mean (McCullagh and Nelder, 1989).

This probe set was selected from a publicly available data set of expression profiles comprising 50 normal prostatectomy samples and 52 tumor prostatectomy samples (Singh et al., 2002). We tested the distribution assumption on each of the 12,625 probe sets using the likelihood ratio test described in Jackson (1969), with 5% significance. About 50% of gene expression data appeared to be better described by Lognormal distributions, whereas the remaining 50% were better described by Gamma distributions. This finding opens a serious issue because discriminating between Lognormal and Gamma distributions is notoriously difficult, particularly in small samples (Jackson, 1969). To overcome this issue, we developed a methodology for differential analysis that uses model averaging to account for model uncertainty.

5 Bayesian Analysis of Differential Gene Expression

BADGE (Bayesian analysis of differential gene expression) is a program for Bayesian analysis of differential gene expression that uses model averaging to solve the problem of model uncertainty in gene expression data. BADGE measures the differential expression by the fold change θ_k. Formally, if we let

μ_{kj} denote the average expression level for the gene k in condition j, the fold change is the ratio

$$\theta_k = \frac{\mu_{k1}}{\mu_{k2}}, \quad k = 1, \ldots, p$$

where p is the number of genes. No change of expression is represented by $\theta_k = 1$, and changes of expression are represented by a fold change $\theta_k < 1$ and $\theta_k > 1$.

The method implemented in BADGE is Bayesian and regards the fold change θ_k as a random variable so that the differential expression of each gene is measured by the posterior probability $p(\theta_k > 1|y_k)$. Clearly, values of $p(\theta_k > 1|y_k)$ near 0.5 identify the genes that do not change expression across the two conditions while values of $p(\theta_k > 1|y_k)$ near 1 identify the genes that have larger expression in condition 1 than in condition 2, and values of $p(\theta_k > 1|y_k)$ near 0 identify the genes that have smaller expression in condition 1 than in condition 2. The posterior probability of differential expression of a gene k is independent of the measurements of the other genes, because we assume that the expression values of different genes are independent, given the parameter values. This assumption may not be realistic, because genes are known to interact with each other, but it allows to screen for genes with differential expression. More advanced methods to take gene-gene dependence into account are described in Sebastiani et al. (2004).

BADGE computes the posterior probability of differential expression of each gene by assuming Gamma and Lognormal distributions, and then averages the results of each analysis. This technique is known as *Bayesian model averaging* and is described in Hoeting et al. (1999). If we let M_{lk} and M_{gk} denote the model assumptions that the expression data of gene k follow either a Lognormal or a Gamma distribution, the posterior probability $p(\theta_k > 1|y_k)$ can be computed as:

$$p(\theta_k > 1|y_k) = p(\theta_k > 1|M_{lk}, y_k)p(M_{lk}|y_k) + p(\theta_k > 1|M_{gk}, y_k)p(M_{gk}|y_k) \quad (1)$$

where $p(\theta_k > 1|M_{lk}, y_k)$ and $p(\theta_k > 1|M_{gk}, y_k)$ are the posterior probabilities of differential expression assuming a Lognormal and a Gamma model. The weights $p(M_{lk}|y_k)$ and $p(M_{gk}|y_k) = 1 - p(M_{lk}|y_k)$ are the posterior probabilities of the two models. Because a Bayesian point estimate of the fold change is the expected value of the posterior distribution of θ_k, say $E(\theta_k|y_k)$, the point estimate of the fold-change θ_k is computed by averaging the point estimates conditional on the two models

$$E(\theta_k|y_k) = E(\theta_k|M_{lk}, y_k)p(M_{lk}|y_k) + E(\theta_k|M_{gk}, y_k)p(M_{gk}|y_k). \quad (2)$$

Similarly, an approximate $(1-\alpha)\%$ credible interval is computed by averaging the credible limits computed under the two models. Particularly, if (l_{kl}, u_{kl}) and (l_{kg}, u_{kg}) are the $(1-\alpha)\%$ credible limits conditional on the two models, an approximate $(1-\alpha)\%$ credible interval for θ_k is $(\theta_{kl}, \theta_{ku})$ where

$$\theta_{kl} = l_{kl}p(M_{lk}|y_k) + l_{kg}p(M_{gk}|y_k)$$
$$\theta_{ku} = u_{kl}p(M_{lk}|y_k) + u_{kg}p(M_{gk}|y_k)$$

Details of the calculations are reported in Appendix A. To select the subset of genes characterizing the molecular profile of the two experimental conditions, we proceed as follows. The posterior probability of differential expression $p(\theta_k > 1|y_k)$ is the probability that the gene k has larger expression in condition 1 than in condition 2, given the available data. If we fix a threshold s to select as differentially expressed the genes with $p(\theta_k > 1|y_k) < s$ and $p(\theta_k < 1|y_k) < 1 - s$, then the expected number of genes selected by chance would be $2(p \times s)$, where p is the number of genes in the microarray. By fixing this number to be f, then the threshold s is $f/(2p)$, that can be interpreted as the expected error rate in the detection of the genes with differential expression.

6 Sample Size Determination

A crucial question in the design of comparative experiments is the determination of the sample size sufficient to analyze the data with some level of confidence. The traditional approach to sample size determination is power-based and leads to choose the sample size to achieve a desired power for a particular alternative hypothesis. Dow (2003) and Zien et al. (2003) have investigated this approach in simulation studies, and their results show that the sample size depends on the minimum fold change to be detected, the statistical method used for the estimation of the fold change and the trade off between false positive and false negative rates. So, for example, Zien et al. (2003) identify a minimum of 25 samples per condition to detect genes that change by more than 2 folds with a false positive rates of 0.1% and a power of 80% using the standard t-test. However, this approach appears to be too restrictive for an essentially screening experiment, and it is also strongly dependent on debatable assumptions about the distribution of gene expression data. Therefore, we introduce two different criteria based on the concept of reproducibility and information gain.

6.1 Reproducibility

The first approach to sample size determination that we investigate is based on the concept of reproducibility. The intuition is to identify the minimum sample size that is needed to reproduce the same results with high probability in other experiments. To investigate this issue, we need a large database of microarray experiments from which we can select non-overlapping subsets that are analyzed with some statistic. The reproducibility is then measured

by computing the agreement between the statistics in the different subsets. A measure of agrement is the rescaled correlation $(1 + \rho_i)/2$, where ρ_i is the average correlation between statistics in samples of size i. Suppose, for example, the differential expression of a gene k in two biological conditions is measured by the t-statistics $t_k(D_{1i})$, where D_{1i} is the data set of size i used in the comparison. As we repeat the analysis in non-overlapping data sets of the same size i, we derive the set of values $t(D_{1i}) = \{t_k(D_{1i})\}, \ldots, t(D_{mi}) = \{t_k(D_{mi})\}$, and we can measure the pairwise agreement by the $m(m-1)/2$ correlations

$$\rho_{rs,i} = cor(t(D_{ri}), t(D_{si})).$$

The average correlation ρ_i is then computed by averaging the $m(m-1)/2$ pairwise correlations.

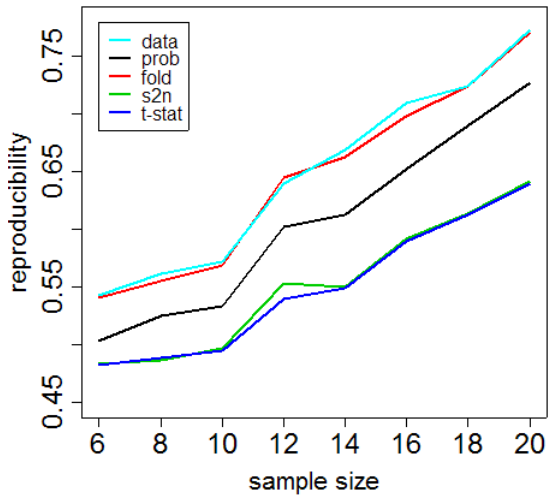


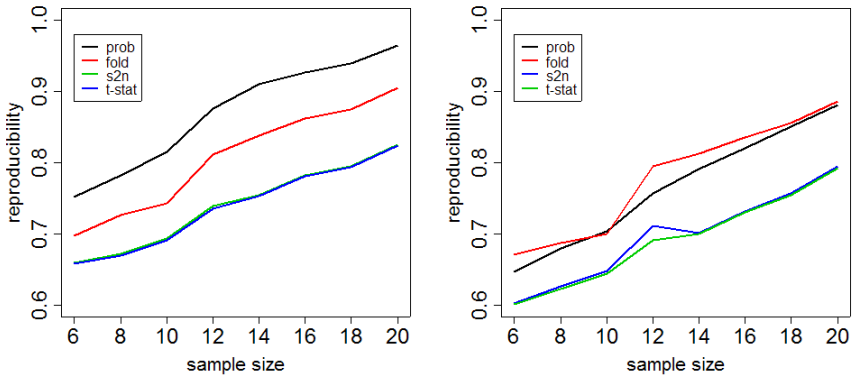
Fig. 3. Reproducibility of the posterior probability (black line); of the estimate of the fold change (red line), of the t (dark blue) and of the signal-to-noise ratio (green) statistics scores compared to the data reproducibility (pale blue), for different sample sizes. The data reproducibility is measured by the empirical fold change: the ratio between sample means. The x-axis reports the number of samples per group, and the y-axis reports the reproducibility measured by $(1+\rho_i)/2$, where ρ_i is the average correlation between statistics in samples of size i.

As an example, Figure 3 plots the reproducibility of the posterior probability and the estimate of the fold change (black and red lines) computed

by BADGE together with the reproducibility of the t-statistic (dark blue) and of the signal-to-noise ratio statistic (green) implemented in GeneCluster. The line in pale blue reports the data reproducibility that was measured by the rescaled correlation between the ratio of sample means. To measure the reproducibility, we selected 32 non-overlapping subsets from the large data set of 102 expression profiles of prostatectomy samples described in Section 4. Specifically, we chose eight different sample sizes ($n_j = 6, 8, 10, 12, 14, 16, 18, 20$) and, for each of the eight sample sizes n_j, we created four data sets by selecting n_j normal samples and n_j tumor samples from the original database. This procedure generated 32 data sets, and then we used BADGE to compute the posterior probability of differential expression and the estimate of the fold change $\hat{\theta}_k$ in each data set. We also analyzed the data sets with GeneCluster using the standard t and signal-to-noise ratio statistics.

The plot in Figure 3 shows a substantially larger reproducibility of the fold change and posterior probability computed by BADGE compared to the t and signal-to-noise ratio statistics. Furthermore, the reproducibility of the estimated fold change is virtually undistinguishable from the data reproducibility. Compared to the estimated fold change, the reproducibility of the posterior probability is about 5% less than the reproducibility of the data, whereas both the t and signal-to-noise ratio statistics are on average 10% less reproducible than the data.

However, we also notice the very low data reproducibility — below 60% — of experiments with less than 10 samples per group, and the fact that a reproducibility higher than 70% requires at least 20 samples per group. To further investigate the effect of sample size on the reproducibility of detecting differential expression, we examined the reproducibility of the analysis with 1329 genes that were selected by BADGE with probability smaller than 0.01 or larger than 0.99 in the whole data set comprising 102 samples. The objective of this comparison was to investigate whether these genes would be detected as differentially expressed in experiments with smaller sample sizes. Figure 4, panel (a), summarizes the results and we notice the large reproducibility of the analysis for small sample sizes: the reproducibility is above 70% even in experiments with only 6 samples per group, and above 80% when the number of samples per group is at least 12. Once again, the reproducibility of the fold analysis conducted by BADGE is consistently larger than that of the analysis conducted with the t or signal-to-noise ratio statistics. We also repeated the analysis using about 1300 genes that were selected by values of the t-statistic smaller than -2 or larger than 2 in the whole data set. The results are summarized in the plot in panel (b) of Figure 4, and show that the selection of the gene by the t-statistic is 5% less reproducible compared to the selection based on BADGE. These results suggest the need for at least 12 samples per conditions, to have substantial reproducibility with BADGE, whereas the analysis based on the t or signal-to-noise ratio statistics would require more than 20 samples per conditions.



(a) (b)

Fig. 4. (a) Reproducibility of the posterior probability (black line); of the estimate of the fold change (red line), of the t (dark blue) and of the signal-to-noise ratio (green) statistics for different sample sizes, for the 1329 genes selected as most differentially expressed by BADGE on the whole data set. (b) same analysis for the 1329 genes selected as most differentially expressed by the t-statistic. The x-axis reports the number of samples per group, and the y-axis reports the reproducibility measured by $(1 + \rho_i)/2$, where ρ_i is the average correlation between statistics in samples of size i.

6.2 Average Entropy

Although suggestive, sample size determination based on reproducibility does not take into account the experimental costs. In this section we introduce a formal decision theoretic approach that allows us to choose the sample size by trading off the gain of information provided by the experiment and the experimental costs.

The decision problem is represented by the decision tree in Figure 5, in which circles represent chance nodes, squares represent decision nodes, and leaves (black circles) are value nodes. The first decision node is the selection of the sample size n used in the experiment, and c represents the cost of one sample. The experiment will generate random data y that have to be analyzed by an inference method a, and the difference between the true state of nature, represented in this case by the fold changes $\theta = (\theta_1, \ldots, \theta_p)$, and the inference will determine a loss $L(\cdot)$ that is a function of the two actions n and a, the data, and the experimental costs. In this decision problem, there are two actions to choose: the optimal sample size and the optimal inference.

The solutions are found by "averaging out" and "folding back" (Raiffa and Schlaifer, 1961), so that, starting from the terminal node, we compute the expected loss at the chance nodes, given everything on the left of the

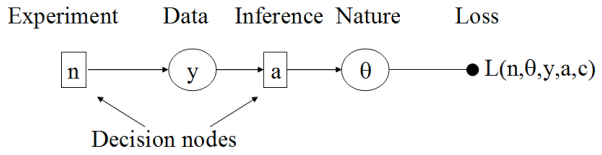


Fig. 5. A decision tree describing the choice of sample size. The first decision node represents the choice of the sample size. After this decision, the experiment is conducted and generates the data y that are assumed to follow a distribution with parameter θ. The data are used to make an inference on the parameter θ, and the second decision node a represents the statistical procedure that is used to make this inference. The last node represents the loss induced by choosing an experiment with sample size n and an inference a, when the true parameter value is θ. The loss is a function of the sample size n and the inference a, the data y, the true parameter value θ and the experimental cost c.

node, and we determine the best actions by minimizing the expected loss at the decision nodes. The first decision is the choice of the inference method a and the optimal decision a^*, or Bayes action, is found by minimizing the expected loss $E\{L(n, \theta, y, a, c)\}$, where the expectation is with respect to the conditional distribution of θ, given n and y. The expected loss evaluated in the Bayes action a^* is called the *Bayes risk* and we denote it by $R(n, y, a^*, c) = E\{L(e, \theta, y, a^*, c)\}$. This quantity is also a function of the data y, so that the optimal sample size is chosen by minimizing the expected Bayes risk $E\{R(n, y, a^*, c)\}$, where the expectation is with respect to the marginal distribution of the data.

A popular choice for the loss function $L(\cdot)$ is the log-score that is defined as

$$L(n, \theta, y, a, c) = -\log a(\theta|n, y) + nc \tag{3}$$

in which $a(\theta|y, n)$ is a distribution for the parameter θ, given the data and the sample size n. This loss function was originally advocated by Good (1952) as a proper measure of uncertainty conveyed by a probability distribution. Lindley (1956) proposed the use of this loss function to measure the information gain provided by an experiment and to determine the optimal sample size of an experiment (Lindley, 1997). With this choice of loss function, the Bayes action a^* is the posterior distribution of θ, given n, y, say $p(\theta|n, y)$, and the Bayes risk is given by:

$$\begin{aligned} R(n, y, a^*, c) &= -\int \log p(\theta|n, y) p(\theta|n, y) d\theta + nc \\ &\equiv Ent(\theta|n, y) + nc. \end{aligned}$$

The quantity $Ent(\theta|n, y) = -\int \log p(\theta|n, y) p(\theta|n, y) d\theta$ is known as the Shannon entropy, or entropy, and the negative Shannon entropy represents the

amount of information about θ contained in the posterior distribution. Therefore the negative Bayes risk represents the trade off between information and experimental costs.

To choose the optimal sample size $n = (n_1 + n_2)$, we need to minimize the expected Bayes risk

$$\min_{n_j} E\{R(n, y, a^*, c)\} = \min_{n_j} \left\{ \int Ent(\theta|n, y) p(y|n) dy + nc \right\}.$$

Because we assume that expression data are independent, given the parameters, the joint posterior density of the parameter vector θ is $p(\theta|n, y) = \prod_k p(\theta_k|n, y_k)$. This independence implies that the overall entropy $Ent(\theta|n, y)$ is the sum of the entropies $\sum_k Ent(\theta_k|n, y_k)$, and the expected Bayes risk is

$$E\{R(n, y, a^*, c)\} = \sum_k \int Ent(\theta_k|n, y_k) p(y_k|n) dy_k + nc.$$

In BADGE we account for model uncertainty by averaging the results of the posterior inference, conditional on the Gamma and Lognormal distribution for the gene expression data. To parallel the sample size determination with the inference process based on model averaging, we therefore introduce the *Average Entropy* $Ent_a(.)$ that we define as

$$\begin{aligned} &Ent_a(\theta_k|y_k, e) \\ &= p(M_{lk}|n, y_k) Ent(\theta_k|n, y_k, M_{lk}) + p(M_{gk}|n, y_k) Ent(\theta_k|n, y_k, M_{gk}). \end{aligned}$$

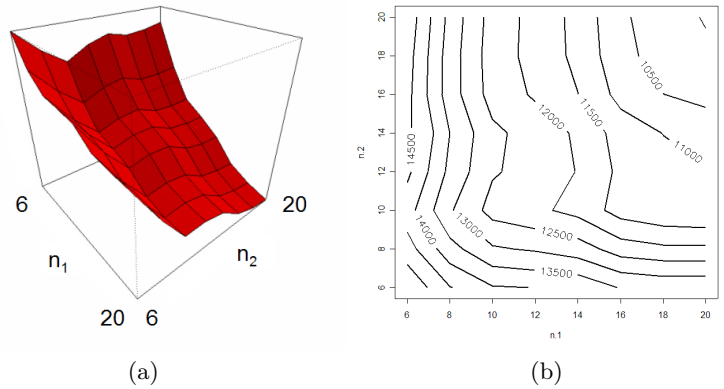
This quantity averages the Shannon entropies conditional on the Gamma and Lognormal models with weights given by their posterior probabilities. In the Appendix we show that the average entropy is a concave function on the space of probability distributions, it is monotone under contractive maps and has some nice decomposition properties. These properties ensure that

$$Ent_a(\theta|n, y) = \sum_k Ent_a(\theta_k|n, y_k).$$

This last simplification allows us to simplify the calculation of the expected Bayes risk $E\{R(n, y, a^*, c)\}$ as

$$\begin{aligned} E\{R(n, y, a^*, c)\} &= \sum_k E\{Ent_a(\theta_k|n, y_k)\} + nc \\ &= \sum_k p(M_{lk}) \int p(y_k|n, M_{lk}) Ent(\theta_k|n, y_k, M_{lk}) dy_k \\ &+ \sum_k p(M_{gk}) \int p(y_k|n, M_{gk}) Ent(\theta_k|n, y_k, M_{gk}) dy_k + nc. \end{aligned}$$

The last formula describes the expected Bayes risk as an average of Bayes risks conditional on the Gamma and Lognormal models, with weights given by their prior probabilities. The importance of this result is an overall objective criterion for sample size determination that averages criteria based on specific model assumptions, thus providing a solution that is robust to model uncertainty. Because computations in closed form are intractable, we have developed numerical approximations to the conditional entropies $Ent(\theta_k|n, y_k, M_{lk})$ and $Ent(\theta_k|n, y_k, M_{gk})$. The calculations of the integrated risk is performed via stochastic simulations and the exact objective function is estimated by curve fitting as suggested in Müller and Parmigiani (1995). These details will be published elsewhere, but are available upon request.



(a) (b)

Fig. 6. Estimate of the expected Bayes risk. The surface in panel (a) shows the estimated Bayes risk (z-axis) as a function of the number of samples n_1 (x-axis) and n_2 (y-axis) per condition. Panel (b) shows the contour plot of the surface displayed in panel (a).

One example is in Figure 6 that plots the stochastic estimation of the Bayes risk as a function of the sample sizes n_1 and n_2. In this example, the data were resampled from the data set of 102 prostatectomy samples described in Section 6.1. From the results on the reproducibility, we estimated that a sample of size n induces a reproducibility $(22.5\log(n) - 4)\%$, so that we used as loss function $-\log(p(\theta_k|n, y_k) + .22 * \log(n) - .04$. An interesting fact is the evident decrease of the estimated Bayes risk when the sample size increases from six to ten samples per condition, whereas the reduction in risk is less effective for larger sample sizes. This result agrees with the findings in Section 6.1 about the reproducibility of the analysis. Furthermore, the effect of changing the number of samples in the two conditions is not symmetrical. This finding is more intriguing and would suggest that, at least in microarray experiments

comparing normal versus tumor samples, it is best to have a larger number of normal samples than tumor samples. An intuitive explanation of this finding is that tumor samples are less variable because the individuals are all affected by the disease.

7 Discussion

Although this chapter has focused on the design of comparative experiments conducted to identify genes with differential expression, microarrays are used for broader experimental objectives and challenge statisticians with novel design questions. In comparative experiments, an important question is whether it is best to make pure replications of the expression measurements of the same cell. Arguments have been made to show that a single replication is not sufficient to achieve reproducible results and authors have suggested to use at least three pure replications of each measurement (Lee et al., 2000). The costs of microarray experiments still impose serious sample size limitations, and the designer of the experiment needs to trade off the number of biological replications with the number of pure replications. The best solution depends of course on the objective of the analysis: if the interest is to have an accurate estimate of the *technical variability* of the microarray measurements, then an experiment with a large number of replications and a small number of biological replications will be preferable to an experiment with one replication of each biological replications. However, in experiments in which the *biological variability* between samples is expected to be large, such as in clinical studies involving human subjects, investing resources in biological replications rather than pure replications is intuitively the best strategy. This dilemma in the design of the experiments and the lack for an "out-of-the-box" answer shows the needs for researching this area further.

Sample size and treatment choice are the design questions for general multifactor experiments. Authors have proposed the use of standard factorial experiments in completely randomized designs, block designs or Latin squares, see (?; Churchill, 2003). However, the unusual distribution of gene expression data questions the relevance of standard orthogonal factorial experiments in this context. Another important problem that has not received large attention in the design community is the development of design criteria for experiments that are not limited to the estimation of particular parameters. For example, data from comparative experiments are often used to define classification models able to predict a clinical feature by using the molecular profile of cells in a tissue. This objective is particularly important for cancer classification (Golub et al., 1999), when subtypes of cancer are difficult to discriminate. The typical approach is to select the genes with differential expression and use them to build a classification model. Several models have been proposed in the literature and an overview is in Sebastiani et al. (2003a). Validation of the classification accuracy is carried out by using a training set to build the

model and a test set to assess its classification accuracy. In this context, an important design question is the sample size needed to determine a classification model that is sufficiently accurate, and an interesting approach based on learning curves is described in Mukherjee et al. (2003).

More complex are design issues for microarray experiments conducted to identify gene functions or their network of interaction. The assumption that genes with similar functions have similar expression patterns underlies the popular approach of clustering gene expression profiles and sample molecular profiles to identify subgroups of genes with similar expression patterns in a subset of the samples (Eisen et al., 1998). Design issues are the sample size determination, and also the selection of the time points at which to make the measurements in temporal experiments. When the experimental goal is to model the network of gene interactions, we move into the area of experimental design for causal inference. Popular knowledge representation formalisms such as Bayesian networks (Cowell et al., 1999) and dynamic Bayesian networks seem to be the ideal tool for capturing the dependency structure among genes (Friedman et al., 2000; Segal et al., 2001; Yoo et al., 2002; Sebastiani et al., 2004). Proper experiments to learn Bayesian networks from data are unknown and, besides preliminary work in Pearl (1999), Spirtes et al. (1999), experimental design to enable causal inference with Bayesian networks is an unexplored research area.

8 Acknowledgments

This research was supported by the NSF program in Bioengineering and Environmental Systems Division/Biotechnology under Contract ECS-0120309. The authors are grateful to the editors and anonymous referees for their help to improve the initial version of the chapter.

A Details of Computations

In this section, we describe briefly the details of some numerical approximations used to compute the posterior distribution of the fold change θ_k, for $k = 1, \ldots, p$. We assume that, given the model parameters, the expression data y_{kji} are independent between genes and samples.

Computation details: Lognormal distribution

Suppose the expression data y_{kji} are generated from a variable Y_{kj} that follows a Lognormal distribution with parameters η_{ki} and σ^2_{kj} defining the mean $\mu_{kj} = e^{\eta_{kj}+\sigma^2_{kj}/2} =$ and the variance $\mu^2_{kj}(e^{\sigma^2_{kj}} - 1)$. Particularly, $X_{kj} = \log(Y_{kj})$ is normally distributed with mean η_{kj} and variance σ^2_{kj}. Because

$$p(\theta_k > 1|M_{lk}, y_k) = p(\log(\mu_{k1}) - \log(\mu_{k2}) > 0|M_{lk}, y_k)$$
$$= p(\eta_{k1} - \eta_{k2} + (\sigma^2_{k1} - \sigma^2_{k2})/2 > 0|M_{lk}, y_k)$$

any inferences about θ_k can be done equivalently on the parameters η_{kj}, σ^2_{kj} of the log-transformed variables. The posterior probability $p(\theta_k > 1|M_{lk}, y_k)$ can be computed as

$$p(\theta_k > 1|M_{lk}, y_k) = \int p(\eta_{k1} - \eta_{k2} > (\sigma^2_{k2} - \sigma^2_{k1})/2|\sigma^2_{k1}, \sigma^2_{k2}, M_{lk}, y_k) \times f(\sigma^2_{k1}, \sigma^2_{k2}|M_{lk}, y_k) d\sigma^2_{k1} d\sigma^2_{k2} \quad (4)$$

where $f(\sigma^2_{k1}, \sigma^2_{k2}|M_{lk}, y_k)$ denotes the posterior density of the parameters $\sigma^2_{k1}, \sigma^2_{k2}$. We assume a standard uniform prior on η_{kj} and $\log(\sigma^2_{kj})$ and prior independence of $(\eta_{k1}, \sigma^2_{k1})$ from $(\eta_{k2}, \sigma^2_{k2})$. Then, it is well known that, given the data, the parameters $\sigma^2_{k2}, \sigma^2_{k1}$ are independent and distributed as $s^2_{kj}/\sigma^2_{kj} \sim \chi^2_{n_i-1}$, $i = 1, 2$, where χ^2_n denotes a χ^2 distribution on n degrees of freedom, and $s^2_{kj} = \sum_j (x_{kji} - \bar{x}_{kj})^2/(n_i - 1)$ is the sample variance of the log-transformed data $x_{kji} = \log(y_{kji})$ in condition i. Similarly, $\eta_{kj}|\sigma^2_{kj}$ is normally distributed with mean $\bar{x}_{kj}$ and variance σ^2_{kj}/n_i, and the marginal distribution of η_{kj} is $\eta_{kj} \sim (s^2_i/n_i)^{1/2} t_{n_i-1} + \bar{x}_{kj}$, where t_n is a Student's t distribution on n degrees of freedom, (Box and Tiao, 1973).

To compute the integral in (4), we notice that, for fixed $\sigma^2_{k2}, \sigma^2_{k1}$, the quantity $p(\eta_{k1} - \eta_{k2} > (\sigma^2_{k2} - \sigma^2_{k1})/2)$ is the cumulative distribution function of a standard normal distribution evaluated in $-\{(\sigma^2_{k2} - \sigma^2_{k1})/2 - (\bar{x}_{k1} - \bar{x}_{k2})\}/\sqrt{\sigma^2_{k1}/n_1 + \sigma^2_{k2}/n_2}$, and then this quantity should be averaged with respect to the joint posterior distribution of $\sigma^2_{k2}, \sigma^2_{k1}$. Because there does not seem to be a closed form solution, we use a two-step numerical approximation. First we approximate the integral in (4) by the first order approximation

$$p(\eta_{k1} - \eta_{k2} > (s^2_{k2} - s^2_{k1})/2|M_{lk}, y_k),$$

and then we use the numerical approximation to the Behrens-Fisher distribution described by Box and Tiao (1973), to approximate the posterior probability by

$$p(\theta_k > 1|M_{lk}, y_k) \approx p\left(t_b > -\frac{\bar{x}_{k1} - \bar{x}_{k2} + s^2_{k1}/2 - s^2_{k2}/2}{a(s^2_{k1}/n_1 + s^2_{k2}/n_2)^{1/2}}\right).$$

The scaling factor a and the adjusted degrees of freedom b are given in Box and Tiao (1973). For large n_1, n_2 the scaling factor a approaches 1 and the degrees of freedom b approach $n_1 + n_2 - 2$ so that the posterior distribution of $\eta_{k1} - \eta_{k2}$ is approximately the non central Student's t $(s^2_{k1}/n_1 + s^2_{k2}/n_2)^{1/2} t_{n_1+n_2-2} + \bar{x}_{k1} - \bar{x}_{k2}$. The approximation is applicable for n_1, n_2 greater than 5, and comparisons we have conducted against inference based on MCMC methods have shown that this approximation works well for samples of size 6 or more.

An approximate estimate of the fold change θ_k is

$$\hat{\theta}_k = e^{\bar{x}_{k1} - \bar{x}_{k2} + s_{k1}^2/2 - s_{k2}^2/2}$$

and approximate credible limits are given by

$$l_{kl} = e^{(\bar{x}_{k1} - \bar{x}_{k2} + s_{k1}^2/2 - s_{k2}^2/2) - t_{1-\alpha/2,b} a (s_{k1}^2/n_1 + s_{k2}^2/n_2)^{1/2}}$$
$$u_{kl} = e^{(\bar{x}_{k1} - \bar{x}_{k2} + s_{k1}^2/2 - s_{k2}^2/2) + t_{1-\alpha/2,b} a (s_{k1}^2/n_1 + s_{k2}^2/n_2)^{1/2}}$$

where $t_{1-\alpha/2,b}$ is the $1-\alpha/2$ quantile of a Student's t distribution on b degrees of freedom.

Computation details: Gamma distribution

Suppose now that the gene expression data follow a Gamma distribution with parameters α_{kj}, β_{kj} that specify the mean and the variance of the distribution as $\mu_{kj} = \alpha_{kj}/\beta_{kj}$ and $V(Y_{kj}|\alpha_{kj}, \beta_{kj}) = \mu_{kj}^2/\alpha_{kj}$. We wish to compute the posterior distribution of $\theta_k = \mu_{k1}/\mu_{k2}$, or equivalently

$$\theta_k = \frac{\alpha_{k1}}{\alpha_{k2}} \frac{\beta_{k2}}{\beta_{k1}}.$$

If α_{kj} is known, say $\alpha_{kj} = \hat{\alpha}_{kj}$, using a uniform prior for β_{kj} determines the posterior distribution for $\beta_{kj}|y_k \sim \text{Gamma}(n_i\hat{\alpha}_{kj} + 1, n_i\bar{y}_{kj})$. The value $\hat{\alpha}_{kj}$ can be for example the maximum likelihood estimate of α_{kj}, which is the solution of the equation:

$$f(\alpha_{kj}) = \log(\alpha_{kj}) - \psi(\alpha_{kj}) - \log(\bar{y}_{kj}) + \sum_j \log(y_{kji})/n_i = 0$$

where $\psi(\alpha) = d\log(\Gamma(\alpha))/d\alpha$ is the digamma function. Then it is easily shown that $2n_i\bar{y}_{kj}\beta_{kj} \sim \chi^2_{2(n_i\hat{\alpha}_{kj}+1)}$ (Casella and Berger, 1990). Furthermore, $\beta_{k1}|y_k$ and $\beta_{k2}|y_k$ are independent and, because the ratio of two independent random variables that are distributed as χ^2 distribution is proportional to an F distribution (Box and Tiao, 1973), the distribution of the ratio β_{k2}/β_{k1} is easily found to be

$$\frac{\beta_{k2}}{\beta_{k1}} \sim \frac{n_1\bar{y}_{k1}}{n_2\bar{y}_{k2}} \frac{n_2\hat{\alpha}_{k2} + 1}{n_1\hat{\alpha}_{k1} + 1} F_{2(n_2\hat{\alpha}_{k2}+1), 2(n_1\hat{\alpha}_{k1}+1)}$$

and an approximation to the probability $p(\theta_k > 1|M_{gk}, y_k)$ is

$$p(\theta_k > 1|M_{gk}, y_k) = p\left(F_{2(n_2\hat{\alpha}_{k2}+1), 2(n_1\hat{\alpha}_{k1}+1)} > \frac{\bar{y}_{k2}}{\bar{y}_{k1}} \frac{\hat{\alpha}_{k2}}{\hat{\alpha}_{k1}} \frac{\hat{\alpha}_{k1} + 1/n_1}{\hat{\alpha}_{k2} + 1/n_2}\right)$$

The point estimate for θ_k is given by $\hat{\theta}_k = \bar{y}_{k1}/\bar{y}_{k2}$, and $(1-\alpha)\%$ credible limits are

$$l_{kg} = \frac{\bar{y}_{k1}}{\bar{y}_{k2}} \frac{\hat{\alpha}_{k1}}{\hat{\alpha}_{k2}} \frac{\hat{\alpha}_{k2} + 1/n_2}{\hat{\alpha}_{k1} + 1/n_1} f_{\alpha/2, 2(n_2\hat{\alpha}_{k2}+1), 2(n_1\hat{\alpha}_{k1}+1)}$$
$$u_{kg} = \frac{\bar{y}_{k1}}{\bar{y}_{k2}} \frac{\hat{\alpha}_{k1}}{\hat{\alpha}_{k2}} \frac{\hat{\alpha}_{k2} + 1/n_2}{\hat{\alpha}_{k1} + 1/n_1} f_{1-\alpha/2, 2(n_2\hat{\alpha}_{k2}+1), 2(n_1\hat{\alpha}_{k1}+1)}.$$

The assessment of the error of the approximation depends on the posterior variance of α_{kj} of which we do not have a closed form expression. Empirical comparisons we conducted on gene expression data sets suggest that the results based on our numerical approximation are virtually indistinguishable from those obtained by Markov Chain Monte Carlo methods when $n_1, n_2 > 10$. Details are described in the report Sebastiani et al. (2003b).

Computation details: mixing weights

To compute the mixing weights in equations (1) and (2), we assume that changes in the average expression levels between the two experimental conditions can at most affect the parameter values but not the distribution membership. Therefore, the mixing weights are the posterior probabilities $p(M_{lk}|y_k)$ and $p(M_{gk}|y_k)$, computed by disregarding the distinction between the two conditions $j = 1, 2$. We use the approximation to the posterior odds $B_k = p(M_{lk}|y_k)/p(M_{gk}|y_k)$ given by the Bayesian information criterion to make the choice independent of the prior probabilities (Kass and Raftery, 1995). In this way, the posterior probability $p(M_{lk}|y_k)$ is $B_k/(1 + B_k)$ and $p(M_{gk}|y_k) = 1/(1 + B_k)$. The Bayesian information criterion is essentially the likelihood ratio:

$$B_k = \frac{p(M_{lk}|y_k)}{p(M_{gk}|y_k)} = \frac{f_l(y_k|\hat{\eta}_k, \hat{\sigma}_k^2)}{f_g(y_k|\hat{\alpha}_k, \hat{\beta}_k)} \tag{5}$$

where $f_l(y_k|\hat{\eta}_k, \hat{\sigma}_k^2)$ and $f_g(y_k|\hat{\alpha}_k, \hat{\beta}_k)$ are the likelihood functions for the Lognormal and Gamma models evaluated in the maximum likelihood estimates $\hat{\eta}_k, \hat{\sigma}_k^2, \hat{\alpha}_k, \hat{\beta}_k$ of the parameters. See Sebastiani et al. (2003b) for further details.

B Properties of the Average Entropy

In this appendix, we prove some general properties of the average entropy in the context of gene expression analysis. We denote by θ the change of expression of a generic gene in two conditions, and we suppose that the expression values follow either a Gamma distribution, M_g, or a Lognormal distribution, M_l. In this case, the average entropy becomes:

$$Ent_a(\theta) = w_1 Ent(\theta|M_l) + (1 - w_1) Ent(\theta|M_g)$$

where, for simplicity of notation, w_1 denotes the probability of the model M_l, and $1 - w_1$ is the probability of the model M_g. The quantities $Ent(\theta|M_l)$ and

$Ent(\theta|M_g)$ denote, respectively, the Shannon entropy of θ computed under the assumption that the gene expression data follow a Lognormal and a Gamma distribution.

Theorem 1 (Concavity). *The average entropy $Ent_a(\theta)$ is a concave function of the set of probability distributions for θ.*

Proof. The result follows by the fact that Shannon Entropy is concave in the space of probability distribution (DeGroot, 1970), and the average entropy is a convex combination of Shannon entropies.

Theorem 2 (Monotonicity). *Let $\eta = g(\theta)$ be a smooth transformation of θ, such that g^{-1} exists, and let J be the Jacobian of the transformation g^{-1}. Then*

$$\begin{cases} Ent_a(\eta) > Ent_a(\theta), \text{ if } |J| < 1; \\ Ent_a(\eta) < Ent_a(\theta), \text{ if } |J| > 1. \end{cases}$$

Proof. The result follows by the monotony of Shannon Entropy (Sebastiani and Wynn, 2000).

Theorem 3 (Decomposability). *The average entropy of the random vector $\theta = \{\theta_1, \theta_2\}$ can be decomposed as*

$$Ent_a(\theta_1, \theta_2) = Ent_a(\theta_1) + E_{\theta_1}\{Ent_a(\theta_2|\theta_1)\}.$$

Proof. Let M_{l1} and M_{l2} denote Lognormal distributions for the expression values of two genes, and let w_1 and w_2 be the posterior probability assigned to the models M_{l1} and M_{l2}. When we decompose the average entropy of θ_1 and θ_2 we need to consider the space of model combinations

$$\mathcal{M} = \{(M_{1l}, M_{2l}), (M_{1l}, M_{2g}), (M_{1g}, M_{2l}), (M_{1g}, M_{2g}).$$

If we assume that the model specifications are unrelated, and that expression values of different genes are independent given the parameter values, then the probability distribution over the model space $\mathcal{M}$ is $w_1w_2, w_1(1-w_2), (1-w_1)w_2, (1-w_1)(1-w_2)$. Then we have

$$\begin{aligned} Ent_a(\theta_1, \theta_2|\mathcal{M}) = {} & w_1w_2 Ent(\theta_1, \theta_2|M_{1l}, M_{2l}) \\ & + w_1(1-w_2) Ent(\theta_1, \theta_2|M_{1l}, M_{2g}) \\ & + (1-w_1)w_2 Ent(\theta_1, \theta_2|M_{1g}, M_{2l}) \\ & + (1-w_1)(1-w_2) Ent(\theta_1, \theta_2|M_{1g}, M_{2g}) \end{aligned}$$

By the property of Shannon entropy $Ent(\theta_1, \theta_2) = Ent(\theta_1) + E_{\theta_1}\{Ent(\theta_2|\theta_1)\}$, where $E_\theta(\cdot)$ denotes expectation with respect to the distribution of θ, there follows that

$$w_1 w_2 Ent(\theta_1, \theta_2 | M_{1l}, M_{2l}) \\ = w_1 w_2 Ent(\theta_1 | M_{1l}) + w_1 w_2 E_{\theta_1|M_{1l}}\{Ent(\theta_2|\theta_1, M_{2l})\}$$

and similarly

$$w_1(1-w_2) Ent(\theta_1, \theta_2 | M_{1l}, M_{2g}) \\ = w_1(1-w_2) Ent(\theta_1 | M_{1l}) + w_1(1-w_2) E_{\theta_1|M_{1l}}\{Ent(\theta_2|\theta_1, M_{2g})\};$$

$$(1-w_1)w_2 Ent(\theta_1, \theta_2 | M_{1g}, M_{2l}) \\ = (1-w_1)w_2 Ent(\theta_1 | M_{1g}) + (1-w_1)w_2 E_{\theta_1|M_{1g}}\{Ent(\theta_2|\theta_1, M_{2l})\};$$

$$(1-w_1)(1-w_2) Ent(\theta_1, \theta_2 | M_{1g}, M_{2g}) \\ = (1-w_1)(1-w_2) Ent(\theta_1 | M_{1g}) + (1-w_1)(1-w_2) E_{\theta_1|M_{1g}}\{Ent(\theta_2|\theta_1, M_{2g})\}.$$

Now group the terms

$$w_1 w_2 Ent(\theta_1|M_{1l}) + w_1(1-w_2) Ent(\theta_1|M_{1l}) = w_1 Ent(\theta_1|M_{1l})$$

and

$$(1-w_1)w_2 Ent(\theta_1|M_{1g}) + (1-w_1)(1-w_2) Ent(\theta_1|M_{1g}) = (1-w_1) Ent(\theta_1|M_{1g})$$

to derive

$$w_1 Ent(\theta_1|M_{1l}) + (1-w_1) Ent(\theta_1|M_{1g}) = Ent_a(\theta_1).$$

Similarly, we can group the terms

$$w_1 E_{\theta_1|M_{1l}}\{w_2 Ent(\theta_2|\theta_1, M_{2l}) + (1-w_2) Ent(\theta_2|\theta_1, M_{2g})\} = w_1 E_{\theta_1|M_{1l}}\{Ent_a(\theta_2|\theta_1)\}$$

and

$$(1-w_1) E_{\theta_1|M_{1g}}\{w_2 Ent(\theta_2|\theta_1, M_{2l}) + (1-w_2) Ent(\theta_2|\theta_1, M_{2g})\} = (1-w_1) E_{\theta_1|M_{1g}}\{Ent_a(\theta_2|\theta_1)\},$$

to derive

$$w_1 E_{\theta_1|M_{1l}}\{Ent_a(\theta_2|\theta_1)\} + (1-w_1) E_{\theta_1|M_{1g}}\{Ent_a(\theta_2|\theta_1)\} = E_{\theta_1}\{Ent_a(\theta_2|\theta_1)\}$$

that concludes the proof.

Theorem 4 (Additivity). *If θ_1, θ_2 are independent , then*

$$Ent_a(\theta_1, \theta_2) = Ent_a(\theta_1) + Ent_a(\theta_2).$$

Proof. The result follows from the previous theorem.

Index

Bibliography

Baldi, P., and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics*, *17*, 509–519.

Box, G. E. P., and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Wiley, New York, NY.

Casella, G., and Berger, R. L. (1990). *Statistical Inference*. Duxbury Press, Belmont, Ca.

Chen, Y., Dougherty, E., and Bittner, M. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomed. Optics*, *2*, 364–374.

Churchill, G. (2003). Comment to "statistical challenges in functional genomics". *Statist. Sci.*, *18*, 64–69.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York, NY.

Crick, F. H. C. (1970). Central dogma of molecular biology. *Nature*, *227*, 561–563.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York, NY.

Dow, G. S. (2003). Effect of sample size and P-value filtering techniques on the detection of transcriptional changes induced in rat neuroblastoma (NG108) cells by mefloquine. *Malaria Journal*, *2*. Available from http://www.malariajournal.com/content/2/1/4.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2001). Statistical methods for identifying genes with differential expression in replicated cDNA microarrays experiments.. *Statistica Sinica*, *12*, 111–139.

Duggan, J. D., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nat. Genet.*, *21*, 10–14.

Efron, B., Storey, J. D., and Tibshirani, R. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, *96*, 1151–1160.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, *95*, 14863–14868.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using bayesian network to analyze expression data. *J. Comput. Biol.*, *7*, 601–620.

Golub, R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., H. Coller, . M. L. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, *286*, 531–537.

Good, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. B*, *14*, 107–114.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.*, *14*, 382–417. With discussion.

Ibrahim, J. G., Chen, M. H., and Gray, R. J. (2002). Bayesian models for gene expression with DNA microarray data. *J. Amer. Statist. Assoc.*, *97*, 88–99.

Jackson, O. A. Y. (1969). Fitting a Gamma or Log-normal distribution in fibre-diameter measurements on wool tops. *Appl. Statist.*, *18*, 70–75.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, *3*, 318–356.

Kass, R. E., and Raftery, A. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, *90*, 773–795.

Lee, M. T., Kuo, F. C., Whitmorei, G. A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA*, *18*, 9834–9839.

Lennon, G. G., and Lehrach, H. (1991). Hybridization analyses of arrayed cDNA libraries. *Trends Genet*, *7*, 314–317.

Lindley, D. V. (1956). On a measure of information provided by an experiment. *Ann. Math. Statist.*, *27*, 986–1005.

Lindley, D. V. (1997). The choice of sample size. *J. Roy. Statist. Soc. C*, *46*, 129–138.

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, *14*, 1675–1680.

Lockhart, D., and Winzeler, E. (2000). Genomics, gene expression and DNA arrays. *Nature*, *405*, 827–836.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models* (2nd edition). Chapman and Hall, London.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., and Mesirov, T. R. G. A. J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *J. Comput. Biol.*, *10*, 119–142.

Müller, P., and Parmigiani, G. (1995). Optimal design via curve fitting of Monte Carlo experiments. *J. Amer. Statist. Assoc.*, *90*.

Newton, M., Kendziorski, 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. *J. Comput. Biol.*

Pearl, J. (1999). Graphs, structural models, and causality. In *Computation, Causation, and Discovery*, pp. 95–140. The MIT Press, Menlo Park, CA.

Raiffa, H. A., and Schlaifer, R. S. (1961). *Applied Statistical Decision Theory*. MIT Press, Cambridge, MA.

Schildkraut, J. M. (1998). Examining complex genetic interactions. In *Gene Mapping in Complex Human Diseases*, pp. 379–410. John Wiley & Sons, New York.

Sebastiani, P., Abad, M., and Ramoni, M. F. (2004). Bayesian networks for genomic analysis. In *EURASIP: Book Series on Signal Processing and Communications.* To appear.

Sebastiani, P., Gussoni, E., Kohane, I. S., and Ramoni, M. (2003a). Statistical challenges in functional genomics (with discussion). *Statist. Sci.*, *18*, 33–70.

Sebastiani, P., Ramoni, M., and Kohane, I. (2003b). BADGE: Technical notes.. Tech. rep., Department of Mathematics and Statistics, University of Massachusetts at Amherst.

Sebastiani, P., and Wynn, H. P. (2000). Maximum entropy sampling and optimal Bayesian experimental design. *J. Roy. Statist. Soc. B*, *62*, 145–157.

Segal, E., Taskar, B., Gasch, A., Friedman, N., and Koller, D. (2001). Rich probabilistic models for gene expression. *Bioinformatics*, *1*, 1–9.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., DAmico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, *1*, 203–209.

Spirtes, P., Glymour, C., Scheines, R., Meek, C., Fienberg, S., and Slate, E. (1999). Prediction and experimental design with graphical causal models. In *Computation, Causation, and Discovery*, pp. 65–94. The MIT Press, Menlo Park, CA.

Tusher, V. G., Tibshirani, R., and Chu, G. (2000). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, *98*, 5116–5121.

Yoo, C., Thorsson, V., and Cooper, G. (2002). Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. In *Proceedings of the Pacific Symposium on Biocomputing*. Available from http://psb.stanford.edu.

Zien, A., Fluck, J., Zimmer, R., and Lengauer, T. (2003). Microarrays: How many do you need?. *J. Comput. Biol.* In press.