Biostatistics (2004), **5**, 1, *pp*. 89–111
Printed in Great Britain

Factorial and time course designs for cDNA microarray experiments

G. F. V. GLONEK
School of Applied Mathematics, The University of Adelaide, Adelaide, SA 5005, Australia
P. J. SOLOMON*
School of Applied Mathematics, The University of Adelaide, Adelaide, SA 5005, Australia
patty.solomon@adelaide.edu.au

SUMMARY

Microarrays are powerful tools for surveying the expression levels of many thousands of genes simultaneously. They belong to the new genomics technologies which have important applications in the biological, agricultural and pharmaceutical sciences. There are myriad sources of uncertainty in microarray experiments, and rigorous experimental design is essential for fully realizing the potential of these valuable resources. Two questions frequently asked by biologists on the brink of conducting cDNA or two-colour, spotted microarray experiments are 'Which mRNA samples should be competitively hybridized together on the same slide?' and 'How many times should each slide be replicated?' Early experience has shown that whilst the field of classical experimental design has much to offer this emerging multi-disciplinary area, new approaches which accommodate features specific to the microarray context are needed. In this paper, we propose optimal designs for factorial and time course experiments, which are special designs arising quite frequently in microarray experimentation. Our criterion for optimality is statistical efficiency based on a new notion of admissible designs; our approach enables efficient designs to be selected subject to the information available on the effects of most interest to biologists, the number of arrays available for the experiment, and other resource or practical constraints, including limitations on the amount of mRNA probe. We show that our designs are superior to both the popular reference designs, which are highly inefficient, and to designs incorporating all possible direct pairwise comparisons. Moreover, our proposed designs represent a substantial practical improvement over classical experimental designs which work in terms of standard interactions and main effects. The latter do not provide a basis for meaningful inference on the effects of most interest to biologists, nor make the most efficient use of valuable and limited resources.

Keywords: cDNA microarrays; Factorial experiments; Optimal experimental design; Time course experiments.

1. INTRODUCTION

Microarrays are powerful tools for surveying the expression levels of many thousands of genes simultaneously. They belong to the new genomics technologies which are rapidly transforming molecular biology from its historical paradigm of the identification, cloning and analysis of specific gene products. There are many different microarray technologies, ranging from the high-density nylon membrane arrays

*To whom correspondence should be addressed.

popular amongst medical and agricultural scientists, to the short oligonucleotide (Affymetrix) arrays which are more accurate, but proprietry and expensive. Our experience has been with the class of spotted complementary DNA (cDNA) microarrays (Brown and Botstein, 1999; Eisen and Brown, 2000) and recently also with spotted long oligonucleotide arrays; the latter and cDNA arrays are often collectively referred to as 'two-colour' or spotted microarrays.

Microarray experiments are conducted in many different contexts with applications ranging from analysing cellular responses to biological and environmental stimuli, genetic mapping studies and diagnosing disease states, to understanding gene regulation and interactions; the key motivation for research in these and related fields is the expectation of rapid advance in understanding the genetic basis of disease, and hopefully, of finding cures. The interface of biology, medicine, computer science and statistics which used to be data-poor is now data-mega-rich, and statistics has a central role to play in producing and processing that information, and to making it intelligible.

Our paper is concerned with planning microarray experiments, a topic on which there are still relatively few published papers, despite the fact that rigorous experimental design is essential for accurately measuring the effects of most interest to biologists. Kerr and Churchill (2001) is dedicated to a discussion of classical experimental designs for microarray experiments and primarily considers A-optimality as the efficiency criterion for choosing a design. More recently, Yang and Speed (2002, 2003) and Churchill (2002) provide suggestions for planning microarray experiments and overview the major design issues involving cDNA microarrays. Design issues have also been discussed by other researchers, including Jin *et al*. (2001), who demonstrate the importance of experimental design and replication in a study of sex, age and genotype in *Drosophila melanogaster*, Wolfinger *et al*. (2001) and Pan *et al*. (2002).

In this paper, we address issues of experimental design which entail important statistical and practical considerations specific to the microarray context. We contend, as do Yang and Speed (2002, 2003), that whatever the primary aim of the experiment, be it to identify a list of candidate genes for differential expression or to discriminate between different tissue types, the optimal design should estimate the effects of interest to biologists with maximum precision, subject to resource and any other practical constraints. The major practical constraint is effectively the number of slides which can be hybridized in any given experiment which in turn may be due to the limited availability of the requisite mRNA probes, or due to cost considerations. For readers unfamiliar with microarrays, it is perhaps helpful to emphasize that the experimental design chosen applies simultaneously to all genes on the array. However, for the purposes of statistical analysis, the genes are treated more or less separately and for this reason it is necessary only to consider the question of design for a single gene.

Thus for a given amount of experimental effort and any practical constraints on the problem, we seek to optimize the information on the key biological effects of interest. A key premise is that it is possible to define *a priori* a number of contrasts that are of specific interest. The approach is then to design experiments that provide maximal information for these contrasts. Obtaining such designs is not straightforward, and this paper is devoted to describing our approach to the problem. We demonstrate the utility of our approach by determining optimal designs for factorial experiments with a small number of factors, as well as for time course experiments with a relatively small number of time points. A more conventional approach is to select designs on the basis of the usual orthogonal contrasts and standard optimality criteria such as A-optimality. However, we argue that such designs typically lead to improved efficiency for certain contrasts that are not relevant at the expense of those that are. We also demonstrate how to improve upon the widely-used reference designs usually favoured by biologists, and upon the all-pairwise comparison factorial designs recently proposed by Speed (2001). We place some simple results established by Yang and Speed (2002, 2003) in a broad conceptual and formal framework for the design of two-colour spotted microarray experiments. Our primary criterion for optimality is statistical efficiency, and using a new notion of *admissibility*, we propose classes of designs which accommodate the special features of microarray experiments. The efficiency gains over commonly used designs can be substantial.

Section 2 of the paper provides a brief background to the cDNA microarray process: some basic knowledge of the process itself is essential for understanding the statistical issues involved. Section 3 motivates what we mean by 'design' for microarrays and presents a motivating biological example. In Section 4, we describe the underlying conceptual and mathematical framework of our approach to designing microarray experiments, introduce the notion of admissibility, and derive efficient designs for 2×2 factorial experiments. More complete elucidation of efficient designs for 2×2 factorial experiments is given in the Appendix. Section 5 describes how admissible designs can be extended to search for optimal designs when there are limitations on the amount of mRNA available. In Section 6, we extend the applications to higher-order factorial designs. Efficient designs for time course experiments are set out and discussed in Section 7. Some further issues in the planning of experiments are discussed in Section 8, where we also briefly summarize our key findings and compare our designs with classical experimental designs.

2. THE cDNA MICROARRAY TECHNIQUE

In cDNA microarrays, known single-stranded DNA clones are robotically spotted out and fixed onto a glass slide. At the same time, two mRNA samples from cell populations to be compared are reversed transcribed into cDNA and separately labelled with dyes, usually red (Cy5) and green (Cy3). The two labelled probes are then mixed together and applied to the microarray. During hybridization, single strands in the probe solution competitively combine with their complementary base-pair nucleotide sequences spotted on the slide. The motivation behind the technique is that the mRNA in the original cell sample reflects which genes are being used by the cell, and the intensity ratio at a spot is thus a measure of the relative abundance of the gene in the two samples. We refer readers to Nguyen *et al*. (2002) and Schena (2003) who provide detailed accounts of the relevant biological and technical background.

The relative intensities of red and green at a spot are extracted by image processing the scanned microarray slides. Yang *et al*. (2002a) give a comprehensive discussion of the statistical issues involved. The intensity ratios are usually adjusted for background noise, then normalized to remove systematic sources of variation. These steps are motivated by the fact that a substantial proportion of the observed variation in cDNA microarray data is due to systematic biases, which are described in detail by Dudoit *et al*. (2002) and Yang *et al*. (2002b). The raw red and green intensities are usually transformed to the log base 2 scale, being a natural scale of measurement for multiplicative (i.e. fold) changes and inducing effective additivity of effects. As the starting point, we assume that the requisite processing of the image has been conducted, and that appropriate data pre-processing steps have been performed to produce data in the form of a single ratio of a red and green intensity for each gene, and that these values are reasonably assumed to be relatively free of systematic bias.

3. DESIGN: MOTIVATION AND ILLUSTRATION

3.1 *What is meant by 'design' for microarrays?*

Experimental design for microarrays entails numerous statistical and practical considerations. Some of the questions most frequently asked by biologists include: which mRNA samples should be competitively hybridized together on the same slide, and how many times should each slide be replicated? Other important questions arise in considering the use of pooled samples when several individuals are sampled from each of the populations under study, including whether pooling improves precision and what would be the optimal number of pools. Even the definition of replication is not straightforward in microarray experimentation, and correlations between observations from different slides used in a single experiment can occur for various reasons. These and related issues have been studied by Speed and Yang (2002), who

distinguish 'technical replicates' from true biological replicates in experiments in which the same purified mRNA sample is applied to several arrays.

For the purpose of our discussion, we assume replicated hybridizations are statistically independent in the sense of representing either true biological variability between individuals, or variability between extractions within an individual. Practice varies, and the nature of the replication involved in a particular experiment will obviously determine the scope and extent of biological inference that may be drawn from that experiment, and this should always be made explicit.

To date, there is no formal basis for conducting so-called *dye-swapped* experiments. The motivation for repeating hybridizations with the dye-assignment reversed is an empirical one in that it allows a direct measure of the extent of bias due to the physical and other properties of the dyes in the normalization step. In many experiments, there are biases arising from sources not related to the dyes and often these will be substantial. Although the use of dye-swapped replication is not necessary or sufficient for the elimination of such biases, we take the view that if hybridizations are to be replicated, then they should be performed as dye-swapped replicates.

In this paper, we address the questions of which samples should be hybridized together and which hybridizations should be performed when a factorial design with a small number of factors, or a simple time course design, is appropriate. The goal of a microarray experiment may be to identify candidate genes for differential expression, or it may be to distinguish between different tissue types, or to classify tissues. Our premise is that the appropriate way to achieve such goals is to prescribe a design that is best able to identify differential expression subject to the practical constraints of the problem.

The ability to identify differential expression is expressed most naturally in terms of statistical power against a suitable alternative hypothesis. This can then be optimized by choosing a statistically efficient design. We take this approach in Section 4, but first motivate our development with an illustrative case study.

3.2 *Case study: a cDNA experiment in leukaemic mice*

We are collaborating with researchers from Adelaide's Child Health Research Institute and Hanson Institute on a study to identify genes that play an important role in receptor signalling and leukaemogenesis. The experiment described here is part of a broader research programme focusing on signalling pathways activated by the granulocyte/macrophage colony stimulating factor (GM-CSF) receptor. Several approaches are being taken to investigate the nature of differential signalling that occurs in activated mutants of the GM-CSF receptor, and to relate this to the wild-type GM-CSF receptor. Two classes of activated mutants (extracellular and transmembrane mutants) display contrasting biological effects, especially in relation to leukaemogenic potential. One cell line under study, V449E, proliferates into leukaemia, and another cell line, FIΔ, undergoes differentiation to macrophages and neutrophils. The hypothesis is that there is a set of genes induced specifically in response to expression of V449E that results in its leukaemic effects.

A 2×2 factorial experiment was conducted to compare the two mutants at times zero hours and 24 hours; it was anticipated that measuring changes over time would distinguish genes involved in promoting or blocking differentiation, or that suppress or enhance growth, as genes potentially involved in leukaemia. We are interested in genes differentially expressed between the two samples i.e. in the sample *main effect*, but more particularly, in those genes which are differentially expressed in the two samples at time 24 hours but not at time zero hours. This is the *interaction* of sample and time.

From the perspective of designing a suitable experiment, the key points to observe are the following. The primary objective is to detect non-zero sample by time interactions and therefore the design should be efficient with respect to the estimation of that parameter. The time and sample main effects are also of some interest and should be estimable. In terms of constraints, a total of eight slides printed with the

Table 1. *Expression of a given gene in the* 2×2 *factorial experiment*

Experimental condition	Log intensity
00	μ
$a0$	$\mu + \alpha$
$0b$	$\mu + \beta$
ab	$\mu + \alpha + \beta + (\alpha\beta)$

15 K mouse cDNA library were available, and since adequate mRNA probe was available, there were no further constraints on the possible hybridizations. Note that in contrast to most statistical work, where interactions are often thought of as a nuisance, the interaction parameter in a two-factor gene expression experiment is frequently the parameter of prime importance.

4. Admissible designs

4.1 *Notation and parametrization*

We now introduce the general notation for 2×2 factorial designs and describe the usual types of experiments conducted to measure the interaction parameter. The discussion will be given in terms of a single gene and it is intended that the same parametrization be applied separately for every gene on a slide.

Consider two *factors*, A and B having levels 0, a and 0, b respectively. For example, in the leukaemic mice experiment, factor A represents sample with the levels 0 and a indicating the two cell lines, and factor B represents time with the levels 0 and b indicating the times zero and 24 hours, respectively. Where applicable, the value '0' will represent the baseline level of a factor. In the present example, it is natural to take the non-leukaemic line, FIΔ, as the baseline level for A and time zero hours as the baseline level for B.

In the 2×2 factorial experiment, there are four possible experimental conditions, and in the context of a single hybridization, the expected log intensities can be described by the parameters μ, α, β, $(\alpha\beta)$ as shown in Table 1. It should be noted that the description of the expected intensities shown in Table 1 is completely general in the sense that the possible values for the intensities are not constrained by the parametrization. It is also worth noting that this parametrization is not unique and that other formulations of the main effect and interaction parameters are commonly used. However, all such parametrizations lead to identical conclusions for any specific contrast. The present choice is motivated by the fact that, in our application, the parameters correspond directly to the contrasts of interest to biologists.

The parameter μ may be thought of as the baseline intensity under the control condition 00, i.e. with each factor at its lower level. In the context of cDNA microarrays experiments, μ typically does not have a useful interpretation. The parameter α is often called a main effect parameter and represents the difference in intensities between the two experimental conditions $a0$ and 00. In the context of our example, the difference $\alpha = a0 - 00$ is the difference between V449E and FIΔ observed at time zero. This parameter can be estimated directly from a single slide on which the two cell lines taken at time zero have been hybridized. Similarly, the parameter $\beta = 0b - 00$ is the main effect for B. In the present example, it represents the change in intensity that occurs in FIΔ between zero and 24 hours. As with the main effect α, it can also be estimated directly from a single slide on which the FIΔ cell lines at times zero and 24 hours have been hybridized.

Finally, the parameter $(\alpha\beta)$ is called the AB interaction and is typically the parameter of primary

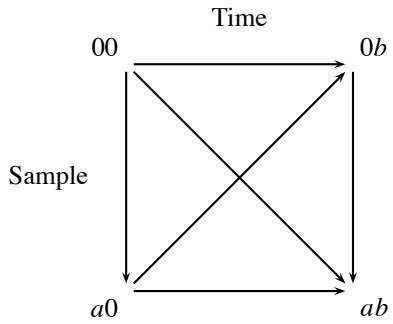

Fig. 1. The four sample–time combinations and the six possible pairwise hybridizations in a 2×2 factorial design of block size two.

interest in a 2×2 factorial microarray experiment. Again for our example, the purpose of the experiment is to identify genes that display a different pattern of expression in the two cell lines over time. Now consider the difference $(ab - 0b) - (a0 - 00)$. Observe that the first term, $ab - 0b$, is the difference between V449E and FIΔ measured at time 24 hours, and the second term, $a0 - 00$, is the same quantity measured at zero hours. Hence the difference of the two represents the differential expression between the two cell lines that exists at time 24 hours beyond what was present at time zero. In terms of the parametrization in Table 1, we find

$$(\alpha\beta) = (ab - 0b) - (a0 - 00)$$

so that, in this case, $(\alpha\beta)$ is the parameter of interest. Unlike the main effects parameters, the interaction cannot be estimated directly from a single slide but can be obtained in various ways from two or more slides. For example, experimenters could perform the following two hybridizations: $0b$ versus ab and 00 versus $a0$ or, alternatively, they could measure the interaction effect by performing the pair of hybridizations $0b$ versus 00 and ab versus $a0$. The interaction can also be measured in less direct ways. For example, from the three hybridizations $a0$ versus 00, $0b$ versus 00 and ab versus 00, we obtain

$$(\alpha\beta) = (ab - 00) - (a0 - 00) - (0b - 00).$$

Having established that to estimate an interaction requires an experiment with multiple slides and that this can be done in several different ways, we now consider the question of which particular hybridizations should be used. To begin, observe that for the four sample–time combinations there are six possible pairs of sample–time combinations that can be hybridized on a single slide. The four sample–time combinations and six possible hybridizations are represented in Figure 1. In fact, there are 12 possible types of slides since the dye allocation can also be reversed for each pair of sample–time combinations, but these need not be considered separately. In Figure 1 we adopt the convention that the arrow-head sample is labelled with the red dye, and the arrow-tail sample with the green dye. Note that in practice, it is desirable to balance the red and green labellings of a probe as much as possible within a given experiment, but for the purposes of describing the parametrization in Table 2 and Figure 1, we have used hybridizations which give a convenient representation of the parameters. The expected log ratio, $M = \log(R/G)$, for each pair of sample–time combinations can be calculated from Table 1 and these are shown in Table 2.

An experimental design is specified by the number of slides of each configuration to be made and, for a fixed total number of slides, a number of different designs are possible. For example, if a total of six

Table 2. *Expected log ratio* $M = \log(R/G)$

	Configuration		Expected
	Green	Red	log ratio
1	00	$a0$	α
2	00	$0b$	β
3	00	ab	$\alpha + \beta + (\alpha\beta)$
4	$0b$	ab	$\alpha + (\alpha\beta)$
5	$a0$	ab	$\beta + (\alpha\beta)$
6	$a0$	$0b$	$\beta - \alpha$

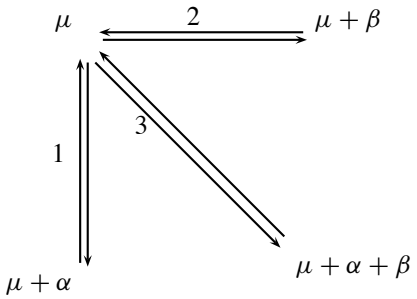

Fig. 2. The usual reference design for six slides allocated as three dye-swapped pairs of hybridizations; the combination with both factors at their lower level is the reference sample, represented by the parameter μ.

slides were available, a reference design comprising two replicates of each of configurations 1, 2 and 3 of Table 2 could be used, as illustrated in Figure 2. The reference design allows for the estimation of all three parameters of interest and has been used extensively in practice. An alternative design considered by Speed (2001), is to use a single replicate of each of the six possible configurations as shown in Figure 1. As with the reference design, the all-pairwise comparison design allows for the estimation of the three parameters of interest but it can be shown to have superior properties to the reference design. However, in the analyses that follow, we will demonstrate designs that are superior to both the reference and all-pairwise comparison designs. In particular, we show that despite its popularity and widespread acceptance, the reference design is very inefficient. We establish also that designs incorporating all possible direct pairwise comparisons are rarely optimal by the criterion of statistical efficiency, regardless of whether interest centres on the main effects and interaction equally, or on the interaction effects alone; any benefit appears to lie solely in estimation of the main effects.

4.2 *Statistical power and standard errors*

The question of design can now be stated as: How many replicates of each configuration should be produced? The standard way to answer this question would be to prescribe a suitable threshold value for M, say 4, and then require that the experiment have a pre-determined level of *power*, say 80%, against any such alternatives. Such an experiment should then have an 80% chance of detecting any gene that is

log four fold over- or under-expressed. It is well known that such a requirement can be met by choosing a design such that the standard error for each parameter of interest falls below a certain value. However, in practice, the situation for microarrays is complicated. In particular, it can be shown that the standard error of a given parameter estimate is given by $\sigma\sqrt{c}$, where σ is the standard deviation between slides for a particular gene and c is a number derived from the design. However, a single experiment typically involves anything between 10 000 and 20 000 genes in which σ varies greatly from gene to gene, and is usually unknown. Therefore, the power typically cannot be determined in advance and, in a single experiment, we should not expect to attain the same level of power for every gene. Nevertheless, the design with the smallest standard error and thus the highest power will be that which has the smallest value of c and this does apply equally to every gene.

4.3 *Least squares estimates*

A major step in the statistical analysis of a factorial microarray experiment is to obtain estimates of the parameters of interest and their standard errors. Both of these quantities can be obtained from the well-known theory of least squares estimation; see, for example, Searle (1971). For illustration, consider again the 2×2 factorial case and the reference design with two slides allocated to each of configurations 1, 2 and 3 from Table 2. To calculate the least squares estimates, the design matrix X must be formed to reflect the expected log intensity ratio for each slide, as specified in Table 2. In this case, the parameter vector is taken to be $\gamma = (\alpha, \beta, (\alpha\beta))^T$, so that the product

$$X\gamma = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ (\alpha\beta) \end{pmatrix} = \begin{pmatrix} \alpha \\ \alpha \\ \beta \\ \beta \\ \alpha + \beta + (\alpha\beta) \\ \alpha + \beta + (\alpha\beta) \end{pmatrix}$$

gives the expected log intensity ratios for the design. If the observed log intensity ratios are given by $\boldsymbol{m} = (m_1, m_2, m_3, m_4, m_5, m_6)^T$, then the least squares estimates of the parameters are given in vector form by $(X^T X)^{-1} X^T \boldsymbol{m}$, and the standard error of the ith parameter estimate is given by $\sigma\sqrt{c_i}$, where c_i is the ith diagonal element of the matrix $(X^T X)^{-1}$.

4.4 *Admissible designs*

It is reasonable, all other things being equal, that we should choose a design that makes each of the c_i as small as possible. Unfortunately this criterion is not straightforward. If the total number of slides is fixed and a certain pair of designs is to be compared, it could be expected that some of the c_i will be smaller for the first design and some will be smaller for the second design. To illustrate, consider two experiments in the 2×2 case, one comprising three replicates of each of the configurations 1, 2, 4 and 5 from Table 2 and the other having four replicates of configurations 1, 2 and two of 4, 5. The design matrices are, respectively,

$$X_1 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}^T$$

and

$$X_2 = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}^T.$$

The diagonal elements of $(X_1^T X_1)^{-1}$ are $1/4$, $1/4$ and $1/3$ and the diagonal elements of $(X_2^T X_2)^{-1}$ are $5/24$, $5/24$ and $3/8$. Hence, for the same total number of slides, the first design provides slightly better estimates for the interaction parameter $(\alpha\beta)$ and the second provides slightly better estimates of the main effects α and β.

On the other hand, it can happen that one design is better than another. For example, consider the reference design with four replicates of each of the configurations 1, 2, 3, so that the design matrix is

$$X_3 = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}^T$$

and the diagonal elements of $(X_3^T X_3)^{-1}$ are given by $1/4$, $1/4$ and $3/4$. The conclusion is that the design X_3 is inferior to both X_1 and X_2 in terms of statistical efficiency. That is, for the same number of slides, the design X_3 provides less accurate estimates of *all* parameters. This is especially the case for the interaction parameter $(\alpha\beta)$ which, as previously discussed, is often the parameter of primary importance. These considerations motivate the following definition.

DEFINITION 1 A design with a total of n slides and design matrix X is said to be *admissible* if there exists no other design with n slides and design matrix X_* such that

$$c_i \geqslant c_i^*$$

for all i *with strict inequality for at least one* i, where c_i, c_i^* are respectively the diagonal elements of $(X^T X)^{-1}$ and $(X_*^T X_*)^{-1}$. A design that is not admissible is said to be *inadmissible*.

According to this definition, the design X_3 is inadmissible since X_1 and X_2 are examples of X_* that violate the conditions for admissibility.

To illustrate the importance of choosing an efficient design, it is useful to compare the performance of admissible designs and some commonly used inadmissible alternatives. For simplicity we will consider the 2×2 factorial experiment with six slides. In this case there are 462 possible designs of which 21 are admissible; these designs are shown in Table 3. Figure 3 presents diagrams of the three admissible designs which estimate the interaction parameter most efficiently. The design in the third row of Table 3 corresponds to the third diagram in Figure 3 and Admissible Design 1 of Table 4, and is subject to the constraint that $c_\alpha = c_\beta$.

We now consider the performance of some inadmissible designs. To simplify the comparison, we compare them only to the admissible designs which satisfy the additional constraint $c_\alpha = c_\beta$. There are three such designs and these are shown in Table 4. The first inadmissible design we consider is the reference design with two slides allocated to each of configurations 1, 2 and 3, as shown previously in Figure 2. Although this design has been widely used, the results of Table 4 show it to be very inefficient under our formulation, especially with respect to the interaction parameter. Based on the comparison of $c_{(\alpha\beta)}$, the Admissible Design 1 is clearly far superior to the reference design, and improves the efficiency of estimation by 100%. In fact, it can be shown that the reference design would require 12 slides to achieve the same precision in estimating the crucial parameter $(\alpha\beta)$.

The design considered recently by Speed (2001), comprising six slides corresponding to all six possible comparisons as shown in Figure 1, is also analysed in Table 4. Although superior to the reference design, it is nevertheless inadmissible and provides a substantially less precise estimate of $(\alpha\beta)$ than Admissible Design 1, which here provides an efficiency gain of 33%. The reference design and the all-pairwise comparison design do provide for more efficient estimates of certain other contrasts, such as $\alpha - \beta$ and $\alpha + \beta + \gamma$. However, a key element of our approach is to identify explicitly the contrasts that

Table 3. *Admissible designs with six slides*

Replication configuration									Replication configuration								
1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$	1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$
2	1	0	2	1	0	0.42	0.67	0.67	2	1	1	1	0	1	0.38	0.46	1.15
1	2	0	1	2	0	0.67	0.42	0.67	1	2	1	0	1	1	0.46	0.38	1.15
2	2	0	1	1	0	0.42	0.42	0.75	4	1	0	1	0	0	0.25	1.00	1.25
3	1	0	1	1	0	0.30	0.70	0.80	1	4	0	0	1	0	1.00	0.25	1.25
1	3	0	1	1	0	0.70	0.30	0.80	3	1	0	1	0	1	0.29	0.57	1.29
2	1	0	1	1	1	0.38	0.54	0.85	1	3	0	0	1	1	0.57	0.29	1.29
1	2	0	1	1	1	0.54	0.38	0.85	3	2	0	1	0	0	0.33	0.50	1.33
3	1	1	1	0	0	0.33	0.67	1.00	2	3	0	0	1	0	0.50	0.33	1.33
1	3	1	0	1	0	0.67	0.33	1.00	2	2	0	1	0	1	0.38	0.38	1.38
2	2	1	1	0	0	0.50	0.40	1.10	2	2	0	0	1	1	0.38	0.38	1.38
2	2	1	0	1	0	0.40	0.50	1.10									

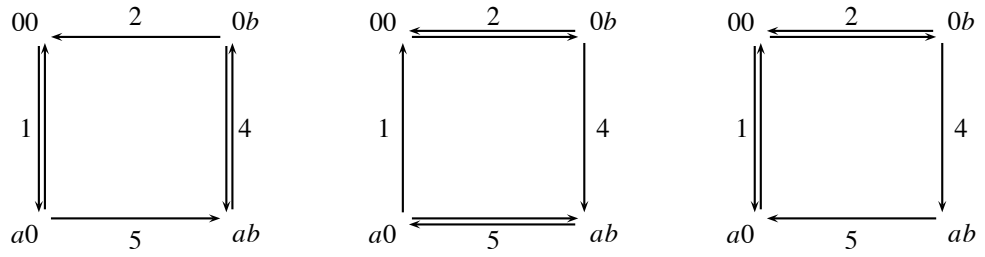

Fig. 3. Three optimal admissible designs for 2×2 factorial experiments with six slides: the three designs correspond to those with the smallest $c_{(\alpha\beta)}$ for estimation of the interaction parameter as set out in Table 3. The third design is subject to the constraint that the main effects c's are equal, i.e. $c_\alpha = c_\beta$, and corresponds to Admissible Design 1 in Table 4.

Table 4. *Designs with six slides*

	Replication configuration								
Design	1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$
Admissible 1	2	2	0	1	1	0	0.42	0.42	0.75
Admissible 2	2	2	0	1	0	1	0.38	0.38	1.38
Admissible 3	2	2	0	0	1	1	0.38	0.38	1.38
Reference	2	2	2	0	0	0	0.50	0.50	1.50
All comparisons	1	1	1	1	1	1	0.50	0.50	1.00

are of interest and optimize with respect to those. We argue that improved estimation of other contrasts is an unwarranted diversion of experimental effort rather than a virtue.

These examples show that large gains in efficiency can be obtained by using admissible designs. The improved efficiency ultimately translates into an enhanced ability to detect differential expression for a given amount of experimental effort. For this reason, it is recommended that only admissible designs be

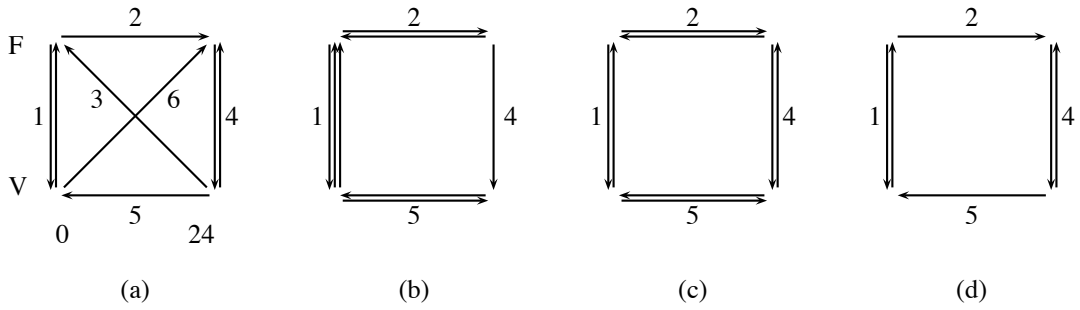

Fig. 4. Four designs for the study of mutant leukaemic mice in which cell lines FIΔ (F) and V449E (V) were compared at times zero and 24 hours. Each arrow represents a hybridization, and cell line comparisons at zero and 24 hours were replicated with the dye assignment reversed. Eight slides in total were available in this experiment: (a) the design that was used; (b) an admissible design; (c) our recommended admissible design; (d) the design used but with both diagonal comparisons omitted.

considered.

In a given problem, that is, a set of possible configurations and total number of slides, there is no simple way to identify the set of admissible designs. It transpires that, even for relatively small experiments, there are a very large number of designs to choose from. For example, if a total of 24 slides are available, then there are 118 755 possible ways to allocate them amongst the six slide types shown in Table 2. However, for relatively small problems they can be identified by simple enumeration of all possibilities. In the Appendix, admissible designs for the 2×2 case are listed for experiments with up to 18 slides and subject to the additional constraint that $c_\alpha = c_\beta$. In situations where the total number of slides is the only constraint, it is our recommendation that only admissible designs be used.

4.5 *The leukaemic mice case study revisited*

The study of leukaemic mice motivated our consideration of the design issues discussed in this paper, but the experiment itself was conducted prior to our elucidation of these issues. At the time, the best design appeared to be the all-pairwise comparison design—this design provides a robust and comprehensive basis for estimation and statistical inference. Using the eight slides available, the six possible pairwise hybridizations were conducted and the cell line (i.e. sample) comparisons at times zero and 24 hours were replicated since they represented the direct comparisons of most biological interest. The two replicated hybridizations were performed as dye-swapped replicates.

The (inadmissible) experimental design we actually employed is shown in Figure 4(a). For this design, it can be checked that $c_\alpha = 1/3$, $c_\beta = 5/12$ and $c_{(\alpha\beta)} = 2/3$. The fact that this design is inadmissible is demonstrated by the design shown in Figure 4(b). For that design, we have $c_\alpha = 0.29$, $c_\beta = 0.39$ and $c_{(\alpha\beta)} = 0.54$. However, since estimation of the interaction parameter $(\alpha\beta)$ is of primary interest, and for reasons of balance, we would recommend in practice that the admissible design shown in Figure 4(c) be used. For that design, we have $c_\alpha = c_\beta = 3/8$ and $c_{(\alpha\beta)} = 1/2$. The simplest case of the 2×2 factorial experiment with eight slides corresponds to the 'loop' design advocated by Kerr and Churchill (2001).

It is of interest to observe that the diagonal comparisons used in the (inadmissible) case study design

Table 5. *Two designs with six slides derived from the leukaemic mice experiment*

	Replication configuration								
Design	1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$
A	2	1	0	2	1	0	0.42	0.67	0.67
B	1	1	1	1	1	1	0.5	0.5	1.0

do not contribute to the estimation of the interaction parameter. In particular, the design matrix is given by

$$X = \begin{pmatrix} 1 & -1 & 0 & 1 & 1 & -1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & -1 & 1 & 0 \end{pmatrix}^T$$

where the second and sixth rows of X represent the dye-swapped replicates of configurations 1 and 4. The coefficients for estimation of the interaction $(\alpha\beta)$ are given by the third row of $(X^TX)^{-1}X^T$ and are in this case $\frac{1}{3}(-1, 1, -1, 0, 1, -1, 1, 0)$. This shows that $(\alpha\beta)$ could have been estimated with the same precision if the experiment had used only the six slides of configurations 1, 2, 4 and 5, as shown in Figure 4(d). In other words, two out of the eight slides do not contribute to the estimation of the key parameter of interest.

It is important to demonstrate that the differences in efficiency shown theoretically for different designs is observable in experimental data. A complete comparison of the data from the original eight-slide leukaemic mice experiment to the reduced version with six slides would need to take account of the corresponding reduction in degrees of freedom. To avoid this complication, we compared the data obtained from two six-slide experiments. In particular, we compared the six-slide experiment shown in Figure 4(d) to the all-pairs design obtained by removing one slide from of each of the dye-swapped replicate pairs. In what follows, we refer to the two designs as A and B respectively. Table 5 shows the calculated variances for designs A and B. We observe that design A is admissible but, as shown previously, design B is not.

To demonstrate the apparent improvement likely to be realized in practice, the data corresponding to designs A and B were analysed as separate experiments and the variance estimates for $\widehat{(\alpha\beta)}$ calculated. Each slide consists of 16 128 spots so that there are 16 128 pairs of variances to be compared. The mean of these variances was 0.068 for design A and 0.088 for design B. There is also considerable variability from gene to gene arising from the variability in the mean squared error, and this is shown in Figure 5. The black points correspond to genes where the variance estimate from design A was lower and the grey points are those for which design B was lower. We see that there is a modest but nevertheless clearly discernible benefit associated with the use of admissible design A. We would expect that a comparison with the reference design shown in Figure 2 would demonstrate an even more marked effect. It should be noted that the high degree of variability apparent in Figure 5 is due to the very small residual degrees of freedom for these designs and this aspect of the problem is quite separate from the issues considered in this paper. From the perspective of designing an experiment, subject to a constraint on the total number of slides, the most relevant comparison is that given in Table 5.

5. ADDITIONAL CONSTRAINTS

In practice, it sometimes happens that in addition to constraints on the total number of slides available, the amount of mRNA from the different sources is also limited. In this situation, the principle of selecting

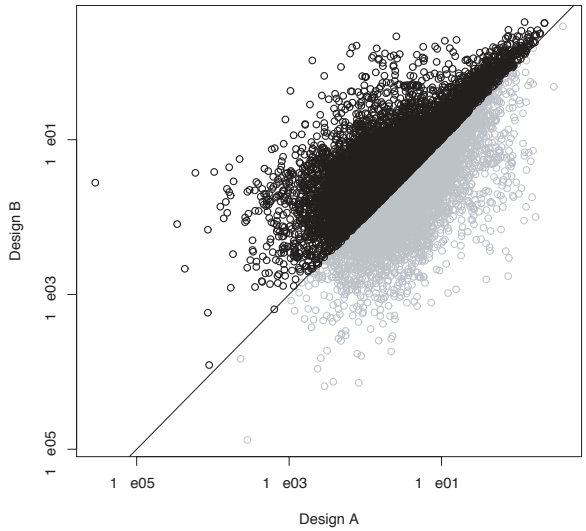

Fig. 5. Variance estimates for $\widehat{(\alpha\beta)}$ from admissible design A and the all-pairs comparison design B, both with six slides, for the leukaemic mice experiment.

admissible designs can still be applied. The only practical consideration is that the search must be constrained to those designs compatible with the limitations on the available mRNA.

Consider the 2×2 factorial design with a total of 18 slides available. With no constraints on the available mRNA, there are a total of 33 649 possible designs; the 16 admissible designs with $c_\alpha = c_\beta$ are shown in Table 17 in the Appendix. For illustration, suppose that for each of the combinations $a0$, $0b$ and ab, there is only sufficient mRNA to produce m slides, but that there is no limit on the baseline combination 00. If $m = 6$, there is only one possible design with 18 slides, namely the reference design with six replicates of each of the configurations 1, 2 and 3. At the other extreme, if $m = 18$, then there are no restrictions and the number of possible designs is 33 649. Table 6 shows the admissible designs with $c_\alpha = c_\beta$ for $m = 6, 7, 8$ and 9. When $m = 9$, the number of possible designs is restricted to 2002 and the design comprising five replicates of configurations 1 and 2 and four replicates of configurations 4 and 5 appears as an admissible design within this restricted subset. It is of interest to observe that the same design is also admissible when no restrictions are imposed, and produces the lowest possible value for $c_{(\alpha\beta)}$.

Another situation in which constraints may arise is in enabling comparability for multiple experiments. Suppose a certain treatment combination is likely to be used as the baseline treatment in several different experiments. We may then require that each treatment combination be hybridized with the baseline a prescribed minimum number of times. For example, consider again the 2×2 factorial experiment with a total of 18 slides and suppose that each of the combinations $a0$, $0b$ and ab are to be hybridized with the baseline 00 at least twice. Our approach is then to consider admissible designs from within the constrained subset which, in this example, comprises 6188 possible designs. There are eight admissible designs with $c_\alpha = c_\beta$, and these are shown in Table 7.

The admissible designs obtained without constraint (see Tables 13–17 in the Appendix) generally include only a very small number of slides of the configurations 3 and 6. The designs that give the overall minimum value for $c_{(\alpha\beta)}$ contain no slides of either configuration. When additional constraints are introduced, similar patterns are observed. In particular, the numbers of slides of the configurations 3

Table 6. *Designs with 18 slides when available mRNA is restricted to m slides for each of a*0, 0*b and ab*

Replication configuration									Replication configuration								
1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$	1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$
$m = 6$									$m = 9$								
6	6	6	0	0	0	0.17	0.17	0.50	5	5	0	4	4	0	0.16	0.16	0.23
$m = 7$									6	6	0	3	3	0	0.14	0.14	0.25
6	6	4	1	1	0	0.15	0.15	0.38	6	6	2	2	2	0	0.14	0.14	0.30
6	6	5	0	0	1	0.15	0.15	0.53	6	6	1	2	2	1	0.13	0.13	0.31
7	7	4	0	0	0	0.14	0.14	0.54	7	7	0	2	2	0	0.13	0.13	0.32
$m = 8$									6	6	2	1	1	2	0.12	0.12	0.42
5	5	2	3	3	0	0.15	0.15	0.26	7	7	1	1	1	1	0.12	0.12	0.45
6	6	2	2	2	0	0.14	0.14	0.30	8	8	0	1	1	0	0.12	0.12	0.56
6	6	3	1	1	1	0.13	0.13	0.39	7	7	2	0	0	2	0.12	0.12	0.79
7	7	2	1	1	0	0.13	0.13	0.40	8	8	1	0	0	1	0.11	0.11	1.25
7	7	3	0	0	1	0.13	0.13	0.62									
8	8	2	0	0	0	0.12	0.12	0.75									

Table 7. *Designs with 18 slides when at least two replicates of configurations 1, 2 and 3 are prescribed*

Replication configuration									Replication configuration								
1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$	1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$
4	4	2	4	4	0	0.17	0.17	0.25	6	6	3	1	1	1	0.13	0.13	0.39
5	5	2	3	3	0	0.15	0.15	0.26	7	7	2	1	1	0	0.13	0.13	0.40
6	6	2	2	2	0	0.14	0.14	0.30	6	6	2	1	1	2	0.12	0.12	0.42
5	5	2	2	2	2	0.13	0.13	0.32	7	7	2	0	0	2	0.12	0.12	0.79

and 6 tend to be small and the designs which minimize $c_{(\alpha\beta)}$ prescribe the smallest possible numbers for both configurations. It is interesting to observe that, excluding the trivial case where it was the only possible design, the reference design was not admissible in any context.

6. FACTORIAL EXPERIMENTS WITH MORE THAN TWO FACTORS OR TWO LEVELS

Our approach to the design of factorial microarray experiments can be used for experiments with more than two factors and/or more than two levels per factor, and we now demonstrate the utility of admissible designs on a 2×3 factorial experiment.

Illustration: a 2×3 *experiment of GM-CSF*. We are collaborating with researchers from the Institute of Medical and Veterinary Science, Adelaide, on investigating the role of the cell-surface receptors for a group of signalling molecules in human disease. The signalling molecules are the cytokines GM-CSF, interleukin 5 and interleukin 3. The aim is to understand how these receptors are activated normally and what goes wrong with them in diseases such as leukaemia, certain solid cancers, asthma and rheumatoid arthritis. The experiment involves knocking out the GM-CSF receptor in a cell line and comparing it to the parental 'normal' cells. Two cell lines are compared in this experiment, mutant (M) and normal wild-type (W), at three time points: zero, six and 12 hours. The biologists are particularly interested in the cell line

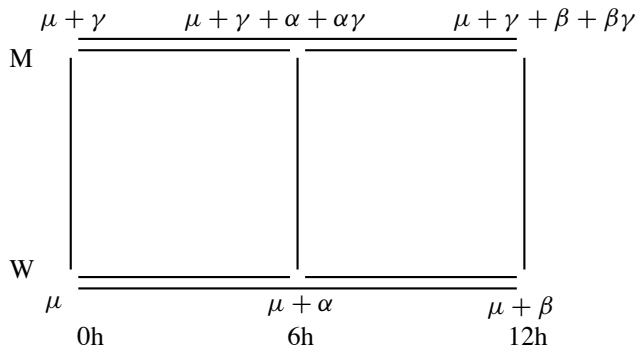

Fig. 6. Possible hybridizations for a 2×3 factorial experiment. The interaction parameters $(\alpha\gamma)$ and $(\beta\gamma)$ are of most interest in this experiment, representing the changes in gene expression at six and 12 hours respectively.

comparisons at six and 12 hours. In what follows, we will suppose that 10 slides are available for the experiment and that there are no other restrictions.

We decided in advance to consider only hybridizations for which the samples differ on one factor. That is, we consider comparisons of the two cell lines at the same time or comparisons of the same cell line at two different times, but exclude the case of comparing the two different cell lines at two different times. The nine hybridizations we consider are shown in Figure 6. The parameter μ represents the baseline (wild-type W at time zero), γ is the cell line (sample) parameter at time zero, α the difference in W between zero and six hours, and β the change in W from six to 12 hours. The parameters of primary interest in this experiment are $(\alpha\gamma)$ and $(\beta\gamma)$, the changes in expression between the cell lines at six hours compared to time zero, and 12 hours compared to time zero, respectively. The main effect parameters α, β and γ are of secondary but nevertheless significant interest. The *a priori* exclusion of the other six comparisons was largely for practical reasons. However, in the light of our experience with the 2×2 case, we considered it very unlikely that any such slides would feature in an optimal design. To choose an optimal design, all possible designs using 10 slides were enumerated and, subject to the constraint $c_{(\alpha\gamma)} = c_{(\beta\gamma)}$, 55 admissible designs were found. Of these, the design shown in Figure 7 provides the smallest smallest value for $c_{(\alpha\gamma)}$. In particular, for this design we have $c_\alpha = c_\beta = 0.66$, $c_\gamma = 0.36$ and $c_{(\alpha\gamma)} = c_{(\beta\gamma)} = 0.63$.

7. TIME COURSE EXPERIMENTS

7.1 *Parametrizations for time course experiments*

The approach illustrated for factorial designs may also be applied in other situations such as time course experiments. Consider a simple time course experiment in which a single sample is to be analysed at times $0, 1, 2, \ldots, n$, and for a single gene, let μ_t denote the level of expression at time t. A key step is to first identify the effects of interest. In contrast to the 2×2 factorial experiment, where it is frequently the case that the interaction parameter is unambiguously of prime importance, the situation for a time course will depend more specifically on the particular context. In what follows, we will consider three approaches that may be applicable in certain practical situations.

In the first, we assume that time zero represents a meaningful baseline and that the purpose is simply to detect differential expression relative to this baseline at any time. In this case, it would appear reasonable to define the parameters to be $\alpha_t = \mu_t - \mu_0$ for $t = 1, 2, \ldots, n$. The second situation we consider is

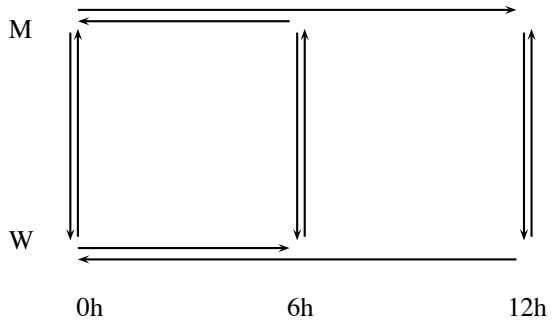

Fig. 7. Best admissible design for a 2×3 factorial experiment with 10 slides when interest is in the interaction parameters $(\alpha\gamma)$ and $(\beta\gamma)$.

to define the parameters of interest as the differences between adjacent time points, $\delta_t = \mu_t - \mu_{t-1}$ for $t = 1, 2, \ldots, r$. This approach is relevant when the time scale is such that changes in expression are likely to be observed as an abrupt step from one time point to the next and the purpose of the experiment is to identify, for each gene, the time points at which such changes occur.

The third situation relevant to time course experiments and where our methods can be applied is when certain time-profiles are specified in advance as being of interest. Suppose $r < n$ particular profiles are defined in advance to be of interest. The n-dimensional space of all possible time profiles can then be parametrized by constructing a basis comprising the r profiles (vectors) of interest and another $n - r$ vectors that span the complementary space. The concept of admissibility can then be applied to identify those designs that provide for the most efficient estimation of the coefficients associated with the profiles of interest. Suppose, for example, four equally-spaced time points are taken and let the space of all four-dimensional profiles, $M = \{(\mu_1, \mu_2, \mu_3, \mu_4)\}$, be parametrized in terms of orthogonal polynomials. The four basis vectors are

$$\boldsymbol{v}_0 = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix}, \ \boldsymbol{v}_1 = \begin{pmatrix} -0.6708 \\ -0.2236 \\ 0.2236 \\ 0.6708 \end{pmatrix}, \ \boldsymbol{v}_2 = \begin{pmatrix} 0.5 \\ -0.5 \\ -0.5 \\ 0.5 \end{pmatrix} \text{ and } \boldsymbol{v}_3 = \begin{pmatrix} -0.2236 \\ 0.6708 \\ -0.6708 \\ 0.2236 \end{pmatrix}.$$

For the purposes of illustration, we will consider designs that are optimal for the estimation of the linear and quadratic terms. The same approach can also be used for any other specified profiles of interest. However, it would seem that the designs tailored for the estimation of the linear and quadratic terms would also be well suited to the estimation of any other profile that is well approximated by a quadratic function of time.

7.2 *Designs*

To demonstrate the application of our methods, the admissible designs are calculated under each of the three situations discussed above for time course experiments with four time points and six or 12 slides. In a time course experiment with n time points, there are $n(n + 1)/2$ possible slides and the question of design again amounts to that of how many slides of each type should be made. In the case $n = 4$,

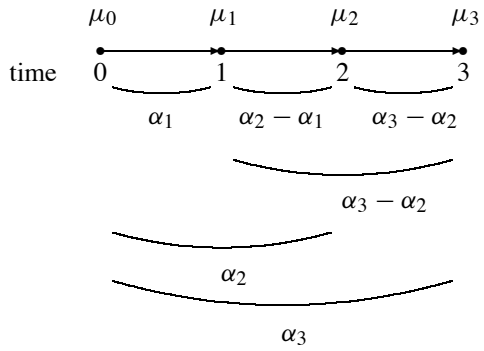

Fig. 8. Possible hybridizations for a time course experiment with four time points.

Table 8. *Expected log ratio M for the time course experiment, for the three parametrizations*

	Configuration				Expected log ratio
1	Time 0	Time 1	α_1	δ_1	$0.4472\beta_1 - \beta_2 + 0.8944\beta_3$
2	Time 0	Time 2	α_2	$\delta_1 + \delta_2$	$0.8944\beta_1 - \beta_2 - 0.4472\beta_3$
3	Time 0	Time 3	α_3	$\delta_1 + \delta_2 + \delta_3$	$1.3416\beta_1 + 0.4472\beta_3$
4	Time 1	Time 2	$\alpha_2 - \alpha_1$	δ_2	$0.4472\beta_1 - 1.3416\beta_3$
5	Time 1	Time 3	$\alpha_3 - \alpha_1$	$\delta_2 + \delta_3$	$0.8944\beta_1 + \beta_2 - 0.4472\beta_3$
6	Time 2	Time 3	$\alpha_3 - \alpha_2$	δ_3	$0.4472\beta_1 + \beta_2 + 0.8944\beta_3$

Table 9. *Time course experiments: α parameters*

Replication configuration								
1	2	3	4	5	6	c_{α_1}	c_{α_2}	c_{α_3}
2	2	2	0	0	0	0.50	0.50	0.50
1	1	1	1	1	1	0.50	0.50	0.50
3	3	3	1	1	1	0.22	0.22	0.22

there are six possible types of slides as illustrated in Figure 8. The expected log ratio for each of the six types of slides are shown in Table 8 using each of the three different parametrizations. It should be emphasized that the three different parametrizations are equivalent representations of the same model. As previously explained, in any single application it is likely that one particular set of parameters would be most relevant to the question at hand. In our approach, the appropriate parametrization is first chosen and then designs admissible with respect to that parametrization considered. When six slides are available, and the α parameters are of interest, there are 462 possible designs of which 44 are admissible. Of these, two provide equal variances for all three parameter estimates and are shown in Table 9. When 12 slides are available, there is only one admissible design that provides equal variances for the three parameter estimates and this is also shown in Table 9.

The two designs in Table 9 for estimation of the α parameters involving six slides give identical

Table 10. *Time course experiments: δ parameters*

Replication configuration								
1	2	3	4	5	6	c_{δ_1}	c_{δ_2}	c_{δ_3}
2	0	0	2	0	2	0.50	0.50	0.50
1	1	1	1	1	1	0.50	0.50	0.50

Table 11. *Time course experiments: δ parameters with 12 slides*

Replication configuration									Replication configuration								
1	2	3	4	5	6	c_{δ_1}	c_{δ_2}	c_{δ_3}	1	2	3	4	5	6	c_{δ_1}	c_{δ_2}	c_{δ_3}
3	2	1	2	1	3	0.22	0.25	0.23	3	1	1	2	2	3	0.23	0.25	0.22
4	0	1	3	1	3	0.21	0.24	0.24	3	1	1	3	0	4	0.24	0.24	0.21
3	1	1	3	1	3	0.23	0.21	0.23	3	2	0	2	1	4	0.24	0.25	0.21
4	1	0	2	2	3	0.21	0.25	0.24									

variances for the parameter estimates and it is therefore of interest to compare the covariance matrices. These are given respectively by

$$\begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}.$$

The positive covariances in the second design indicate, as would be expected, that the second design provides superior estimates of differences $\alpha_{t_1} - \alpha_{t_2}$.

When six slides are available, and the δ parameters are of interest, there are 462 possible designs of which 36 are admissible. Of these, two provide equal variances for all three parameter estimates and are shown in Table 10. However, if a total of 12 slides are available, it transpires that of the 352 admissible designs, none provide equal variance for each of the three parameter estimates. An obvious choice for the design with 12 slides might then be to simply double the numbers of slides from the two admissible designs for the six slides. Clearly both of those designs will give $c_{\delta_1} = c_{\delta_2} = c_{\delta_3} = 0.25$. In this case, there are seven admissible designs that provide lower variances for all three parameter estimates and these are listed in Table 11.

Finally, we consider admissible designs for the parametrization based on the orthogonal polynomials. When six slides are available, there are 462 possible designs of which 71 are admissible. Of these, one provides equal variances for the parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ and these are shown in Table 12. Similarly, when 12 slides are available, there are two admissible designs that provide equal variances and these are also given in Table 12.

In the light of these considerations, designs that allocate equal numbers of each of the six possible types of slides would seem to be well suited to the time course experiment with four time points. Although they are not necessarily optimal, the preceding calculations show that they are quite efficient in all three situations. Moreover, the balance in these designs is an attractive property that may well out-weigh the minor losses in efficiency.

Table 12. *Time course experiments: β parameters*

Replication configuration								
1	2	3	4	5	6	c_{β_1}	c_{β_2}	c_{β_3}
1	1	1	1	1	1	0.25	0.25	0.25
2	2	2	2	2	2	0.13	0.13	0.13
1	4	2	0	4	1	0.10	0.10	0.29

8. Discussion and further design issues

8.1 *The role of parameters*

One of the key steps in the development of this paper is the identification of the parameters of interest. Although this may be an unfamiliar step for many experimentalists, it is necessary for any formulation of good design and in many contexts should be straightforward. It is important to note that, for a given set of possible hybridizations, the corresponding levels of expression can be described using appropriate parameters in several different, but equivalent, ways. See, for example, Tables 9 and 10 in the context of the time course experiment. Moreover, the experimental designs that are optimal for one particular parametrization may not be optimal for a different parametrization of the same experiment. The key point to be made here is that the parameters must be formulated to correspond directly to the underlying questions of substantive interest. In non-technical terms, this amounts to the fact that experiments can be designed to answer specific questions in a given context. One should not expect an experiment designed for a particular question to be optimal for answering some other question. Further issues arise when parameters of subsidiary interest, such as main effect parameters, are present or when a question involving several parameters simultaneously is of interest. The treatment of those issues is somewhat technical and beyond the scope of the present paper.

8.2 *Additional contrasts*

In this paper, we began with the parametrization given in Table 1 and then studied the question of how best to design an experiment to estimate efficiently the parameters of interest. In the 2×2 experiment, it is frequently the case that the interaction parameter is the sole parameter of interest. However, in different experiments it may happen that both the original parameters and some additional derived contrasts are of equal interest. For example, in the simple time course experiment, it may be of equal interest to estimate both the α and δ parameters. In algebraic terms, this leads to a certain redundancy in the sense that if we know the values of the α parameters we can use that information to deduce the values of the δ parameters. As was previously discussed, it is not the case that the designs which give the best estimate of the α parameters are also optimal for estimating δ. In this case, the definition of admissibility could be extended to find designs that best accommodate both requirements. This extension will be the subject of future work.

8.3 *Larger scale studies*

The examples considered in this paper have been small in terms of both the number of parameters involved and the number of slides available. In situations involving a larger number of parameters, the same arguments for considering only admissible designs can be made. However, in such cases it may happen that the number of admissible designs is so large that it is not useful simply to examine the list. In such cases, additional criteria for selecting a design are needed. If the number of parameters is small

but a large number of slides are available, then a different problem arises. Namely, the total number of configurations rapidly becomes too large for the enumeration methods used here to be feasible. Although the problems outlined above are yet to be resolved, it has been our experience that our methods are useful for many experiments currently being considered in practice.

8.4 *Robustness*

In this paper, we have been primarily concerned with finding admissible designs subject to a single constraint on the total number of slides. In Section 5, we considered further contraints owing to limitations on the available mRNA. In practice, it may be necessary to introduce constraints for other reasons. For example, we might require a design with even numbers of each type of slide so that dye-swapping can be used or, more importantly, the requirement may be for a design in which all parameters can be estimated even if one slide fails completely. The latter is only likely to be a problem in small experiments with a very small number of replicated slides, but it raises the general issue of robustness of admissible designs. One approach would be to consider only designs with the property that all parameters remain estimable when any single slide is removed, and then choose a design admissible within the restricted subset. However, most of the admissible designs in Tables 13–17 already have this property so, in practice, it would appear that conducting a restricted search to identify robust, admissible designs may not be necessary.

8.5 *Classical designs*

In this paper, we have described in detail simple notional applications of factorial and time course designs to microarray experiments. It is important to recognize that classical designs and standard approaches to estimation seek to minimize the standard error of all estimable treatment contrasts, whereas we are interested in particular contrasts, frequently although not exclusively the interaction parameter. Moreover, owing to the practical constraints often arising in microarray experimentation due to limited numbers of slides, limitations on the available mRNA probes, uncertainty about the actual experimental process, and so on, each complex experiment needs its own tailor-made design. In other words, although it is possible to generate banks of admissible designs, it is very useful to have a way of treating each experiment on a case-by-case basis to accommodate features particular to that experiment. Furthermore, classical experimental design does not offer optimal designs of direct practical utility to the microarray context, and although there is a large and established literature on classical experimental design, it has not been able to offer definitive answers for even the simplest microarray experiments.

In summary, we have proposed classes of *admissible designs*, for factorial and time course microarray experiments with a fixed number of arrays available for experimentation and information on the effects of primary interest to biologists. For relatively small problems, this may be done simply by enumerating the possibilities. For larger problems, where the number of possible configurations is so large that enumeration is not feasible, we are presently exploring approximate methods of optimisation. The anticipated result of this research will be a computational tool that enables experimentalists and other researchers to identify good designs in problems too large to be analysed by enumeration.

ACKNOWLEDGEMENTS

This work was supported in part by the Australian Research Council. We are grateful to Terry Speed for advice and inspiration, and to Anna Tsykin for helpful discussions on the biological aspects of microarray experiments. We acknowledge our biological collaborators Richard D'Andrea, Brenton Reynolds, Tom Gonda, Mark Guthridge and Greg Goodall.

The calculations for this paper were performed using a computer program written in C++ by the authors. A copy of the program will be provided on request.

Table 13. *Designs with eight slides*

Replication configuration									Replication configuration								
1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$	1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$
2	2	0	2	2	0	0.38	0.38	0.50	3	3	0	1	1	0	0.29	0.29	0.67
2	2	0	2	1	1	0.34	0.34	0.59	3	3	0	1	0	1	0.27	0.27	1.27
2	2	0	1	2	1	0.34	0.34	0.59	3	3	0	0	1	1	0.27	0.27	1.27

Table 14. *Designs with 10 slides*

Replication configuration									Replication configuration								
1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$	1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$
3	3	0	2	2	0	0.27	0.27	0.42	4	4	0	1	1	0	0.23	0.23	0.63
3	3	0	2	1	1	0.25	0.25	0.51	4	4	0	1	0	1	0.21	0.21	1.21
3	3	0	1	2	1	0.25	0.25	0.51	4	4	0	0	1	1	0.21	0.21	1.21
3	3	1	1	1	1	0.23	0.23	0.60									

Table 15. *Designs with 12 slides*

Replication configuration									Replication configuration								
1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$	1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$
3	3	0	3	3	0	0.25	0.25	0.33	4	4	1	1	1	1	0.19	0.19	0.54
3	3	0	3	2	1	0.23	0.23	0.37	5	5	0	1	1	0	0.18	0.18	0.60
3	3	0	2	3	1	0.23	0.23	0.37	4	4	0	1	1	2	0.18	0.18	0.63
4	4	0	2	2	0	0.21	0.21	0.38	5	5	0	1	0	1	0.17	0.17	1.17
4	4	0	2	1	1	0.19	0.19	0.47	5	5	0	0	1	1	0.17	0.17	1.17
4	4	0	1	2	1	0.19	0.19	0.47									

APPENDIX A

Further results for 2×2 *experiments*

In this section, we present the key admissible designs for 2×2 factorial experiments on 8, 10, 12, 16 and 18 slides. The results presented generalize those in Table 3 for six slides, but assume that the main effects are to have equal variances. These tables were produced using a C++ program that identifies the admissible designs in a given situation by enumeration. We offer these designs here in the hope that they may be informative to researchers on the brink of conducting 2×2 factorial microarray experiments. Note that it is not necessary to assume an even number of slides; we have done so for convenience and to enable dye-swapped replication when feasible. It is clear that the 'cross-hybridizations' rarely enter the optimal admissible designs.

Table 16. *Designs with 16 slides*

Replication configuration									Replication configuration								
1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$	1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$
4	4	0	4	4	0	0.19	0.19	0.25	6	6	0	1	2	1	0.14	0.14	0.42
5	5	0	3	3	0	0.16	0.16	0.27	6	6	1	1	1	1	0.13	0.13	0.47
5	5	0	3	2	1	0.15	0.15	0.30	7	7	0	1	1	0	0.13	0.13	0.57
5	5	0	2	3	1	0.15	0.15	0.30	6	6	0	1	1	2	0.13	0.13	0.58
6	6	0	2	2	0	0.15	0.15	0.33	6	6	0	1	0	3	0.13	0.13	1.13
5	5	0	2	2	2	0.15	0.15	0.35	6	6	0	0	1	3	0.13	0.13	1.13
6	6	0	2	1	1	0.14	0.14	0.43									

Table 17. *Designs with 18 slides*

Replication configuration									Replication configuration								
1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$	1	2	3	4	5	6	c_α	c_β	$c_{(\alpha\beta)}$
5	5	0	4	4	0	0.16	0.16	0.22	6	6	0	2	2	2	0.13	0.13	0.33
5	5	0	4	3	1	0.15	0.15	0.24	7	7	0	2	1	1	0.12	0.12	0.41
5	5	0	3	4	1	0.15	0.15	0.24	7	7	0	1	2	1	0.12	0.12	0.41
6	6	0	3	3	0	0.14	0.14	0.25	7	7	1	1	1	1	0.12	0.12	0.45
6	6	0	3	2	1	0.13	0.13	0.29	6	6	1	1	1	3	0.12	0.12	0.47
6	6	0	2	3	1	0.13	0.13	0.29	7	7	0	1	1	2	0.11	0.11	0.57
6	6	1	2	2	1	0.13	0.13	0.31	7	7	0	1	0	3	0.11	0.11	1.11
7	7	0	2	2	0	0.13	0.13	0.32	7	7	0	0	1	3	0.11	0.11	1.11

REFERENCES

BROWN, P. AND BOTSTEIN, D. (1999). Exploring the new world of the genome with DNA microarrrays. *Nature Genetics (Suppl.)* **21**, 33–37.

CHURCHILL, G. A (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics (Suppl.)* **32**, 490–495.

DUDOIT, S., YANG, Y., CALLOW, M. AND SPEED, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–139.

EISEN, M. AND BROWN, P. O. (2000). DNA arrays for analysis of gene expression. *Methods in Enzymology* **303**, 179–205.

JIN, W., RILEY, R., WOLFINGER, R., WHITE, K., PASSADOR-GURGEL, G. AND GIBSON, G. (2001). The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genetics* **29**, 389–395.

KERR, M. AND CHURCHILL, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201.

NGUYEN, D., ARPAT, A., WANG, N. AND CARROLL, R. (2002). DNA microarray experiments: biological and technical aspects. *Biometrics* **58**, 701–717.

PAN, W., LIN, J. AND LE, C. (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biology* **3**, research0022.1–0022.10.

SCHENA, M. (2003). *Microarray Analysis*. Hoboken: Wiley.

SEARLE, S. (1971). *Linear Models*. New York: Wiley.

SPEED, T. P. (2001). *Gene Expression*, Wald Lecture III, Joint Statistical Meetings, August 2001. `http://www.stats.berkeley.edu/users/terry`

SPEED, T. P AND YANG, Y. H. (2002). Direct versus indirect hybridizations for cDNA microarrayexperiments. *Sankhya A Series A* **64**, 707–721.

WOLFINGER, R., GIBSON, G., WOLFINGER, E., BENNETT, L., HAMADEH, H., BUSHEL, P., AFSHARI, C. AND PAULES, R. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**, 625–638.

YANG, Y. H., BUCKLEY, M., DUDOIT, S. AND SPEED, T. P. (2002a). Comparison of methods for image analysis on cDNA micrarray data. *Journal of Computational and Graphical Statistics* **11**, 108–136.

YANG, Y., DUDOIT, S., LUU, P., LIN, D., PENG, V., NGAI, J. AND SPEED, T. P. (2002b). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**, e15.

YANG, Y. H. AND SPEED, T. P. (2002). Design issues for cDNA microarray experiments. *Nature Reviews Genetics* **3**, 579–588.

YANG, Y. H. AND SPEED, T. P. (2003). Design and analysis of comparative microarray experiments. In Speed, T. P. (ed.), *Statistical Analysis of Gene Expression Microarray Data*, Boca Raton, FL: CRC Press.

[*Received December, 9 2002; first revision April, 2 2003; second revision June, 29 2003; accepted for publication August, 7 2003*]