APPLIED STOCHASTIC MODELS IN BUSINESS AND INDUSTRY
Appl. Stochastic Models Bus. Ind., 2006; **22**:631–638
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asmb.654

Experimental designs for 2-colour cDNA microarray experiments

Nam-Ky Nguyen[1,‡] and E. R. Williams[2,*,†]

[1] *School of Mathematics, Statistics and Computer Science, University of New England, Armidale, NSW 2351, Australia*
[2] *Statistical Consulting Unit, The Australian National University, Canberra, Australia*

SUMMARY

Kerr and Churchill (*Biostatistics* 2001; **2**:183–201) showed how varieties (e.g. type of tissues, drug treatments, etc.) are paired onto arrays by a catalogue of A-optimal incomplete block designs (IBDs) for 6–10 varieties (v), and number of blocks of size 2 between v and $\binom{v}{2}$. These A-optimal IBDs were obtained by (i) generating all non-isomorphic connected graphs on v vertices using Brendan McKay's, MAKEG program (http://cs.anu.edu.au/people/bdm/nauty/) and (ii) comparing all designs of the same size on the basis of A-optimality to obtain the best ones.

In this paper we will give a quick overview on IBDs and describe an algorithmic approach to extend the mentioned catalogue. We aim at IBDs with up to 100 varieties with equal as well as unequal replications. A catalogue of 2007 IBDs is given. We will also extend the concept of *even* designs in Kerr and Churchill (*Biostatistics* 2001; **2**:183–201) to *row-orthogonal* designs. Copyright © 2006 John Wiley & Sons, Ltd.

Received 30 June 2005; Revised 1 May 2006; Accepted 4 July 2006

KEY WORDS: A-optimality; even designs; row-orthogonal designs; incomplete block designs; row–column designs; optimal designs

1. INTRODUCTION

Microarrays are a powerful tool for the interrogation of gene function. In cDNA microarrays, single-stranded DNA of known sequence content (e.g. from a library) is spotted out onto a glass slide. There can be many thousands of spots (genes) on each slide (array). Then mRNA from varieties (e.g. cell populations) of interest is reverse transcribed into cDNA and at the same time labelled with red or green fluorescent dye. Differentially-labelled cDNA from two varieties is then applied to the microarray. The single strands of cDNA hybridize to their complementary sequences on the array and this process is measured as a digital signal. The intensity of the red

*Correspondence to: E. R. Williams, Statistical Consulting Unit, The Australian National University, Canberra, Australia.
† E-mail: emlyn.williams@anu.edu.au
‡ Currently at SAS Institute Inc., SAS Campus Drive, Cary, NC 27513, U.S.A.

and green signals is recorded for each spot and indicates which genes are being used by the varieties. For more background on microarray technology, see Nguyen *et al.* [1]. In microarray experiments, as with other experiments such as field variety trials, we want to find the best separation of the varietal variation (the signal) from other sources of variation, such as differences between slides and/or dye labels (the noise). An optimal experimental design provides the greatest separation between signal and noise; this is the case regardless of the methods employed to correct for background noise in the image analysis and whether single or multiple channel normalization is carried out [2].

In microarray experimental design there are two main considerations: (i) the efficient allocation of pairs of varieties (one for each dye colour) to slides and (ii) the arrangement of spots on each slide. In this paper we will concentrate on (i). Previous work on the construction of suitable designs for microarray experiments (see for example Kerr and Churchill [3], hereafter called KC; [4–6]) has not fully taken into account the development of software to generate efficient optimal or near optimal designs and the facility to offer these designs for a wide range of parameter values. A function of this paper is to emphasize the algorithmic approach to the generation of designs suitable for microarray experiments [7]. We will think of the experimental design problem as one where there is an array of two rows (dyes) by b columns (slides) to which v varieties must be allocated; in other words, the construction of an efficient $2 \times b$ row–column design for v varieties.

In this paper we will first present some commonly used microarray designs. We will then discuss the use of IBDs of block size 2 as microarray designs and introduce an efficient algorithm which allocate varieties to slides. These varieties do not necessarily have the same number of replications. This is a common situation in microarray experiments. We will also extend the concept of *even* designs to *row-orthogonal* designs in which we attempt to optimize the allocation of varieties to both slides and dye colours.

2. COMMONLY USED MICROARRAY DESIGNS

Figure 1 gives an example of a *reference* design. This design has five (test) varieties (1–5), one reference variety (0) in five arrays. In this design, one dye is used to label the reference variety (1st row) and the other dye is used to label the test varieties (2nd row).

Each block design is characterized by a $v \times v$ concurrence matrix. The diagonal element λ_{ii} of this matrix gives the number of replications of variety i and off-diagonal element λ_{ij} $(i<j)$ gives the number of arrays (i.e. blocks) in which varieties i and j both appear. A formal definition of this matrix will be given in the next section. The concurrence matrix of the design in Figure 1 is

$$\begin{pmatrix} 5 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{1}$$

It can be seen that most information is collected on the reference variety (i.e. the least interesting variety) and the variety effects are completely confounded with dye effects (see KC).

DOI: 10.1002/asmb

0	0	0	0	0
1	2	3	4	5

Figure 1. A reference design.

0	1	2	3	4
1	2	3	4	0

Figure 2. A loop design.

1	2	0	0	2	3	1	3	4	4
0	0	3	4	1	1	4	2	2	3

Figure 3. A balanced incomplete block design.

Figure 2 gives an example of a *loop* design for five varieties (0–4) in five arrays.
The concurrence matrix of this design is

$$\begin{pmatrix} 2 & 1 & 0 & 0 & 1 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 1 & 0 & 0 & 1 & 2 \end{pmatrix} \tag{2}$$

This design is an example of an IBD for v varieties in v blocks of size 2. It collects twice as much data on the varieties of interest. The varieties are also row-orthogonal (each variety is labelled once with the red and green dyes). However, it does not provide direct comparisons between all variety pairs.

Figure 3 gives an example of a balanced IBD (BIBD) for five varieties (0–4) in 10 arrays.
The concurrence matrix of this design is

$$\begin{pmatrix} 4 & 1 & 1 & 1 & 1 \\ 1 & 4 & 1 & 1 & 1 \\ 1 & 1 & 4 & 1 & 1 \\ 1 & 1 & 1 & 4 & 1 \\ 1 & 1 & 1 & 1 & 4 \end{pmatrix} \tag{3}$$

This typical BIBD has v varieties in $\binom{v}{2}$ arrays of size 2. Unlike previous designs, a BIBD provides direct comparisons for all variety pairs. With large v, however, a BIBD constructed with $\binom{v}{2}$ arrays requires a very large number of arrays. As an economic alternative to BIBDs, A-optimal IBDs for $v \leqslant 10$ and $v \leqslant b \leqslant \binom{v}{2}$ have been catalogued by KC (http://www.jax.org/staff/churchill/labsite/research/expression/design.html). In the next sections, we will give a quick overview on IBDs and describe an efficient algorithm to extend the mentioned catalogue.

3. IBDS AS MICROARRAY DESIGNS

An IBD of size (v, k, b) has v varieties set out in b blocks of size k $(k < v)$ such that each variety is replicated r_i $(i = 1, \ldots, v)$ times. We assume that no variety occurs more than once in a block. A common criterion for comparing designs of the same size is $\sum \mu_i^{-1}$ where the μ_i's are the $v - 1$ non-zero roots of the information matrix for the adjusted variety effects is

$$C = \mathbf{r}^{\delta} - k^{-1}NN' \tag{4}$$

Here $\mathbf{r}^{\delta} = \text{diag}(\mathbf{r})$, where $\mathbf{r} = (r_1, r_2, \ldots, r_v)'$. $NN' = \{\lambda_{ij}\}$ is the $v \times v$ *concurrence* matrix in which $\lambda_{ii} = r_i$ $(i = 1, \ldots, v)$ and λ_{ij} $(i<j)$ is the number of blocks in which varieties i and j both appear. An IBD which minimizes $\sum \mu_i^{-1}$ (i.e. minimizes the average pairwise variance) is said to be A-optimal.

IBDs used in 2-colour cDNA microarray experiments have block size 2. IBDs whose r_i's take the same value and λ_{ij}'s take the same value are called BIBDs. IBDs whose r_i's take the same value and λ_{ij}'s differ by at most 1 are called regular graph designs [8]. When r_i's differ by at most 1 and λ_{ij}'s differ by at most 1, the IBD is called a near-BIBD [9].

In the following, we outline the steps of an algorithm for constructing IBDs of size (v, k, b). This algorithm is an adaptation of the algorithm of Nguyen [10] for constructing IBDs with equal replications:

1. Construct a starting design D of size (v, k, b).
2. Calculate $-kC$ $(= NN' + A$, where $A = -k\mathbf{r}^{\delta}$ is constant) and f (the sum of squares of the upper-diagonal elements of $-kC$). Find a pair of treatments in two different blocks such that the swap of these two treatments results in the biggest reduction in f. If the search is successful, update f, $-kC$ and D. Repeat the search process until $f = b\binom{k}{2}$ (for block size 2, it can be proved that this lower bound is b).
3. If D is not a BIBD, calculate $\sum \mu_i^{-1}$. Find a pair of treatments in two different arrays such that the swap of these two treatments does not alter f but does result in the biggest reduction in $\sum \mu_i^{-1}$. Repeat the search until $\sum \mu_i^{-1}$ cannot be reduced further.

The basic algorithm (steps 1–3) is repeated a number of times in an attempt to avoid local optima. Each repeat is called a *try*.

Note.

1. Step 2 makes use of the (M, S)-optimality criterion (see [7, Section 2.5]) to quickly filter good designs.
2. Let i and t be two varieties in block I, and m and t' be two varieties in block M (assuming t is not in M and t' is not in I). The swapping of i and m increases λ_{tm} and $\lambda_{t'i}$ by 1 and decreases λ_{ti} and $\lambda_{t'm}$ by 1. This observation is used to quickly update f and $-kC$ in Step 2.
3. The formula by John [11] can be used to speed up the update of $\sum \mu_i^{-1}$ in Step 3.

DOI: 10.1002/asmb

1	1	5	3	1	3	5	3
4	0	2	0	2	0	2	4

(a)

$\underline{2}$	1	5	3	1	3	5	3
4	0	$\underline{1}$	0	2	0	2	4

(b)

$\underline{0}$	1	5	3	1	3	5	3
4	0	1	$\underline{2}$	2	0	2	4

(c)

$\underline{1}$	1	5	3	1	3	5	3
4	0	$\underline{0}$	2	2	0	2	4

(d)

Figure 4. Three iterations in the construction of an IBD of size $(v, k, b) = (6, 2, 8)$.

In Figure 4, we illustrate the steps in constructing an IBD of size $(v, k, b) = (6, 2, 8)$.

In Step 1, a starting design (a) is constructed by allocating the replications of the varieties to the spots of the arrays row wise and randomizing the positions of the varieties within each array (column) and within each row. Step 2 consists of (b) where f decreases from 12 to 10 and (c) where f decreases from 10 to 8. Step 3 consists of (d) where $\sum \mu_i^{-1}$ decreases from 4.2333 to 3.75 (f remains 8).

4. DISCUSSION

KC took advantage of the fact that every IBD for v varieties in $b < \binom{v}{2}$ blocks of size 2 corresponds to a simple graph on v nodes to do a search for all possible designs when $v \leqslant 10$. They used Brendan McKay's, MAKEG program (http://cs.anu.edu.au/people/bdm/nauty/) to generate full sets of non-isomorphic-connected designs and searched the A-optimal designs among these designs. Their search is restricted to IBDs in which each pair of varieties appears at most in one array. They found 110 A-optimal IBDs for $v = 6$–10 in $v \leqslant b \leqslant \binom{v}{2}$ blocks. Additional even solutions (solutions in which the number of replications for each variety is even) have been found for 69 (v, b) combinations. Eight solutions for $(v, b) = (11, 13), (12, 14), (13, 14)$ and $(13, 15)$ are also given. They reported that there are 11 716 571, 1006 700 565 and 164 059 830 476 non-isomorphic connected graphs in the searches for 10, 11 and 12 varieties,

DOI: 10.1002/asmb

respectively. As a result, this approach becomes computationally infeasible for large v and the algorithmic approach is suggested as a more feasible alternative.

Our algorithmic approach obtained all 179 IBDs in the KC catalogue in only 1 min on our 2 MHz laptop PC (30 tries are used for each (v, b) combination). The computer time increases as v increases. The 170 designs for $v = 20$, for example, consume $3\frac{3}{4}$ h on the same laptop. Like KC, we restrict our search to IBDs in which each pair of varieties appears at most in one array (IBDs with the objective function f reaching its lower bound). All found designs for $v = 6$–20 in $v \leqslant b \leqslant \binom{v}{2}$ blocks are listed at http://designcomputing.net/mad/. Our IBDs and the ones catalogued by KC are either A- or near A-optimal and are more useful and flexible for microarray experiments than combinatorial designs in the published literature.

Microarray experimenters have special interest in *even* designs (KC Section 4.6). This is because the varieties within each array can be rearranged such that the design becomes row orthogonal. A row-orthogonal design is the one where for each variety, the replication of the variety is the same in each row. The concurrence matrix of the row component will be $k^{-1}\mathbf{rr}'$. As such a row-orthogonal design which is A-optimal with respect to the column component will also be A-optimal with respect to both rows and columns. This definition is an unequal replication extension of that in Section 5.7 of John and Williams [7]. Figure 5 is an example of an A-optimal row-orthogonal IBD with $(v, b) = (6, 8)$ and $\sum \mu_i^{-1} = 3.8333$.

When v is large or when the design is not even, arranging the varieties manually within each row so that the resulting design is optimal with respect to both dye colour and array is not easy. More sophisticated algorithms for row–column designs such as those described by Nguyen and Williams [12] and Nguyen [13] are required for this purpose. Their adaptation for microarray designs will be described elsewhere. The website http://designcomputing.net/mad/ also provides listing of 1010 even designs for $v = 6$–20 when even solutions are available. These designs are row orthogonal. The listed uneven designs cannot be row orthogonal. However, we ensure that the numbers of replications for each variety coloured in the red and green dyes differ by at most 1.

Note that for $(v, b) = (13, 14)$, our design (http://designcomputing.net/mad/v13.txt) has $\sum \mu_i^{-1} = 20.31$. KC design (http://www.jax.org/staff/churchill/labsite/research/expression/v13k14-15results.txt) has $\sum \mu_i^{-1} = 19.19$ and is thus more A-optimal than our design. However, the range of variety replications of our design is 2–3 and of KC design is 1–9.

Figure 6 displays the relationship between the square root of the average variance of all estimated pairwise comparisons $\sqrt{\bar{V}}$ $(= \{2/(v-1) \sum \mu_i^{-1}\}^{1/2})$ and the number of arrays for A-optimal row-orthogonal IBDs with $v = 6$–20. It can be seen that there is substantial reduction in $\sqrt{\bar{V}}$ if few additional arrays are added to an IBD with v arrays. However, this reduction is fairly small if additional arrays are added to an IBD with $2v$ or more arrays.

Note that when the varieties are equi-replicated, the design generation package CycDesigN [14] can also be used to generate A-optimal IBDs and row–column designs.

0	0	1	5	2	3	1	4
4	2	3	0	1	0	5	1

Figure 5. A row-orthogonal IBD for six varieties in eight arrays.

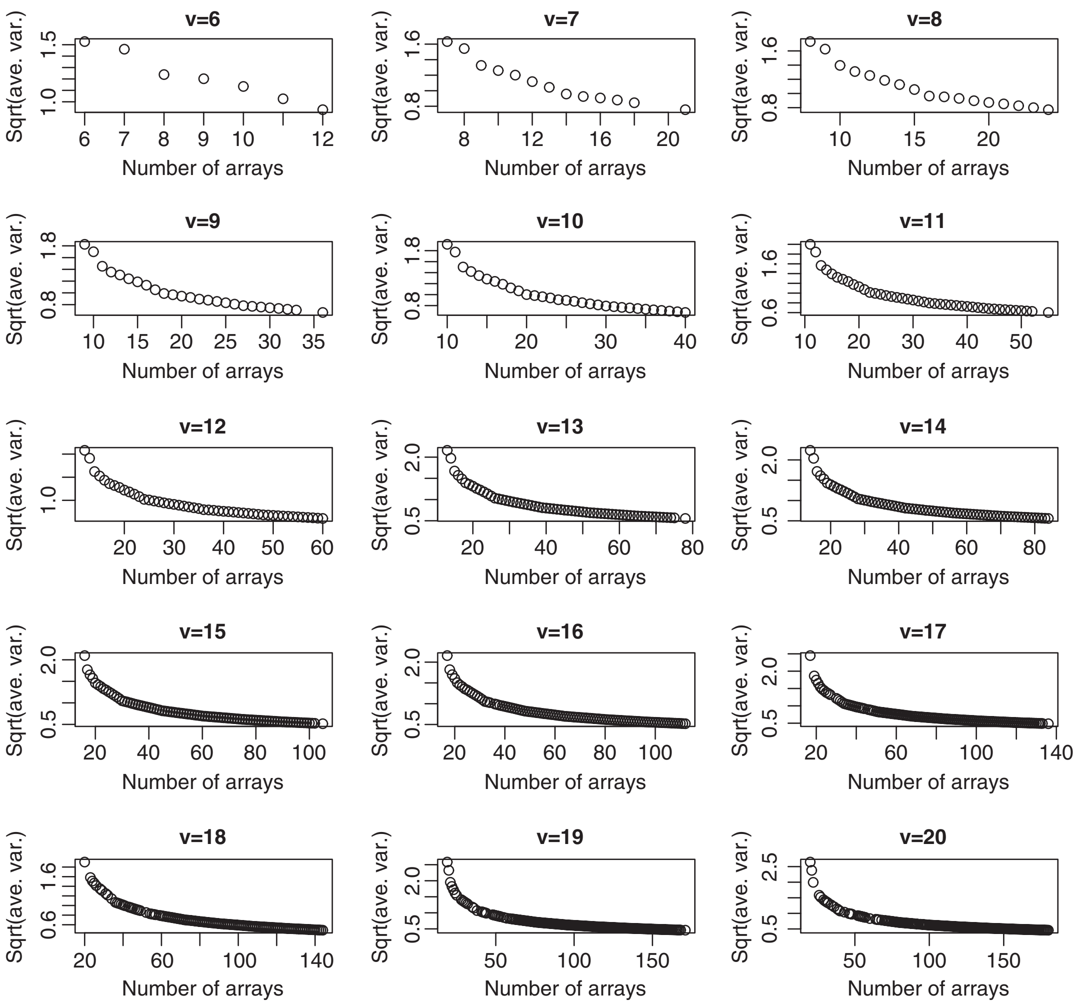

Figure 6. Graphs showing the relationship between $\sqrt{\bar{V}}$ and the number of arrays for 1010 A-optimal row-orthogonal IBDs with $v = 6$–20.

5. CONCLUDING REMARKS

In this paper we provide an alternative approach to construct IBDs of block size 2 for microarray experiments. This approach enables us to extend the useful catalogue of KC. We also go a further step by making the even designs row orthogonal and uneven design near row orthogonal. Researchers wishing to use designs with $v > 20$ can have access to our JAVA program MAD (microarray designs), available from the first author.

Our discussion in this paper restricts to block size 2 as in most cDNA microarray experiments, this block size corresponds to the 2-colour microarray system. However, the algorithm described in Section 3 is general and can be used for block sizes greater than 2 when three or more colour microarray systems are used [15–17].

DOI: 10.1002/asmb

REFERENCES

1. Nguyen DV, Arpat AB, Wang N, Carroll RJ. DNA microarray experiments: Biological and technological aspects. *Biometrics* 2002; **58**:701–717.
2. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 2004; **3**(1).
3. Kerr MK, Churchill GA. Experimental design for gene expression microarrays. *Biostatistics* 2001; **2**:183–201.
4. Yang YH, Speed TP. Design issues for cDNA microarray experiments. *Nature Reviews* 2002; **3**:579–588.
5. Churchill GA. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics Supplement* 2002; **32**:490–495.
6. Kerr MK. Design considerations for efficient and effective microarray studies. *Biometrics* 2003; **59**:822–828.
7. John JA, Williams ER. *Cyclic and Computer Generated Designs*. Chapman & Hall: London, 1995.
8. John JA, Mitchell TJ. Optimal incomplete block designs. *Journal of the Royal Statistical Society*, *Series B* 1977; **39**:39–43.
9. Cheng C-S, Wu C-F. Nearly balanced incomplete block designs. *Biometrika* 1981; **68**:493–500.
10. Nguyen N-K. Construction of optimal incomplete block designs by computer. *Technometrics* 1994; **36**:300–307.
11. John JA. Updating formula in an analysis of variance model. *Biometrika* 2001; **88**:1175–1178.
12. Nguyen N-K, Williams ER. An algorithm for constructing optimal resolvable row–column designs. *Australian Journal of Statistics* 1993; **35**:363–370.
13. Nguyen N-K. Construction of optimal row–column designs by computer. *Computing Science and Statistics* 1997; **28**:471–475.
14. Whitaker D, Williams ER, John JA. *CycDesigN*: *A Package for the Computer Generation of Experimental Designs*. CSIRO Forestry and Forest Products: Canberra, 2002.
15. Hessner MJ, Wang X, Hulse K, Meyer L, Wu Y, Nye S, Guo SW, Ghosh S. Three color cDNA microarrays: quantitative assessment through the use of fluorescein-labeled probes. *Nucleic Acids Research* 2003; **31**:e14.
16. Hessner MJ, Wang X, Khan S, Meyer L, Schlicht M, Tackes J, Datta MW, Jacob HJ, Ghosh S. Use of a three-color cDNA microarray platform to measure and control support-bound probe for improved data quality and reproducibility. *Nucleic Acids Research* 2003; **31**:e60.
17. Woo Y, Krueger W, Kaur A, Churchill G. Experimental design for three-color and four-color gene expression microarrays. *Bioinformatics* 2005; **21**(Suppl. 1):i459–i467.

DOI: 10.1002/asmb