Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society, April 28 - 29, 2006, pp. 121-129

Balanced Experimental Designs for cDNA Microarray data

Kuey Chung Choi[1]

Abstract

Two color or cDNA microarrays are extensively used to study relative expression levels of thousands of genes simultaneously. Only two tissue samples can be hybridized on a single microarray slide. Thus, a microarray slide necessarily forms an incomplete block design with block size two when more than two tissue samples are under study.

We also need to control for variability in gene expression values due to the two dyes. Thus, red and green dyes form the second blocking factor in addition to slides. General design problem for these microarray experiments is discussed in this paper. Designs for factorial cDNA microarrays are also discussed.

Keywords : Confounding, cDNA microarrays, Efficient design, Factorial experiments.

1. Introduction

Two color or cDNA microarrays are now extensively used to study relative expression levels of thousands of genes simultaneously across biological samples.

Although experimental designs were developed by Shah (1960) and Kshirsagar (1966) mainly in the context of agricultural experiments, factorial designs have been found to be useful in other settings as well. The purpose of this paper is to provide designs for cDNA microarray experiments. Since two experimental conditions are hybridized on each microarray slide, the arrays form blocks of size two. However, loop designs become inefficient for larger number of treatment combinations.

Landgreber et al. (2005) pointed out that designs for 2^2, 2^3 and 3×2 experiments using ad-hoc methods. Wit et al. (2004) searched for efficient designs using the simulated annealing algorithm, and showed that dye-swap designs become very inefficient with the increase in the number of conditions under investigation. They also defined interwoven loop designs, which perform quite well when compared with designs obtained using the simulated annealing algorithm.

[1]Dept.of Statistics, Chosun University. E-mail : kjchoi@chosun.ac.kr

In fact, the classical method of confounding does exactly that it achieves higher precision on contrasts of interest at the expense of less important ones that are completely or partially confounded with blocks.

The connection between designs with block size two and microarray designs is discussed in this paper. It is shown that balanced designs for microarray experiments can be readily obtained using balanced designs available in literature. In particular, designs obtained using the classical method of confounding are discussed in some detail. Other balanced factorial designs with block size two, such as group divisible and generalized cyclic designs, are also shown to be equally useful.

Microarrays are presented in section 2 for 2^2, 2^3 and 3×2 experiments are discussed in detail. Microarray designs for other experiments can be obtained similarly. Finally, some concluding remarks are made in section 3.

2. Confounded Designs with Block Size Two

Consider a factorial experiment involving m factors $F_1, F_2, \cdots, F_m$ at two levels each. Treatment combinations will be denoted by m-tuples $a_1 a_2 \cdots a_m$ where a_i=0 or 1 denote the two coded levels of the ith factor, $i=1,2,\cdots m$. Let D be a single replicate design having 2^{m-1}blocks of size two for the m-factor experiment obtained using the classical method of confounding, see e. g. Raghavarao (1971). Let the 2^{m-1}factorial effects (main effects and/or interactions) confounded between the blocks of D be denoted by B_i, $i=1,2,\cdots,2^{m-1}$, and let A_i, $i=1,2,\cdots,2^{m-1}$ denote the factorial effects that are unconfounded, and hence estimable in D.

Let (t_{i1}, t_{i2}) denote the two treatment combinations in the ith block of D, and let (y_{i1}, y_{i2}) be the observations corresponding to them, $i=1,2,\cdots,2^{m-1}$. Then, the contrasts of block totals $y_{i1}+y_{i2}$, $i=1,2,\cdots,2^{m-1}$, estimate the block effects, the factorial effects B_i, $i=1,2,\cdots,2^{m-1}$, are confounded with. Whereas, the unconfounded factorial effects A_i, $i=1,2,\cdots,2^{m-1}$, are estimated using the within block comparisons $y_{i1}-y_{i2}$, $i=1,2,\cdots,2^{m-1}$. Let

$$y_d=[(y_{11}-y_{12}),(y_{21}-y_{22}),\cdots,(y_{b1}-y_{b2})]'$$

where $b=2^{m-1}$ denotes the number of blocks. Let $h_j=(h_{j1}, h_{j2}, \cdots, h_{jb})'$ be such that $E(h'_j y_d)=A_j$, $j=1,2,\cdots,2^{m-1}$. Then, clearly h_j, $j=1,2,\cdots,2^{m-1}$, form a complete set of

mutually orthogonal column vectors of size 2^{m-1}. Also, $Var(h'_j y_d) = h'_j h_j \sigma^2$ where σ^2 denotes the constant variance of the difference $y_{i1} - y_{i2}$. In the case of a microarray experiment involving 2^{m-1} arrays, the within block comparison $y_{i1} - y_{i2}$ corresponds to the $\log_2$ expression ratio for the two treatment combinations hybridized on the th array.

A partially confounded design in which all the 2^{m-1} factorial effects are estimable may be obtained by adding further replicates that confound a different set of factorial effects. Although we have illustrated the method of confounding using 2^m experiments, the method is applicable in general to factors with number of levels a prime or power of a prime number. In fact, similar conclusions also hold for balanced factorial designs with block size two obtained using methods other than the classical method of confounding.

3. Application to Factorial Microarrays

For an m factor microarray experiment, the linear model for these gene expression values is given by

$$y_{ijk} = \mu_j + \alpha_{ij} + \delta_{jk} + \tau_{ijk} + \varepsilon_{ijk}, \quad i = 1, 2, \ldots, b; j = 1, 2, \ldots, g; k = 1, 2$$

where for the jth gene, μ_j is the overall mean, α_{ij} is the ith array effect, δ_{j_1} and δ_{j_2} are the effects of levels 1 and 2 respectively of the dye factor F_g, τ_{ijk} is the effect of the treatment combination i_k, and ε_{ijk} are random errors independently distributed with mean 0 and variance σ_j^2. Because treatment contrasts are estimated from within block comparisons $y_{ij} - y_{ij'}$, the model for these differences is then given by

$$(y_{i1} - y_{i2}) = (\delta_1 - \delta_2) + (\tau_{i1} - \tau_{i2}) + (\varepsilon_{i1} - \varepsilon_{i2}).$$

3.1 2^2 experiments

3.1.1 Main designs of factorial Experiment for 2^2

(1) The number of blocks with block size 2 : $\frac{2^2}{2} = 2^{2-1}$

Table 1) 2^2 experiments

d_{11} : [00, 01], [10, 11]
d_{12} : [00, 10], [01, 11]
d_{13} : [00, 11], [01, 10]

(2) Effected confounded

The method on main design of confounded factors are handled as following and be given above table1.

ⓧ : product of signs for each factor, 0 and 1 are denoted by - and +, respectively.

1) main effect : those which give "+" after multiplying block signs

2) interaction effect: those which give "+" after multiplying signs for at least two factors from multiplication of block signs

3) Example of the main design for confounded factors

D_{11} : $00ⓧ01 = (--ⓧ-+) = (+,-) \Rightarrow F_1$

D_{12} : $00ⓧ10 = (--ⓧ+-) = (-,+) \Rightarrow F_2$

D_{13} : $00ⓧ11 = (--ⓧ++) = (-,-) \Rightarrow F_1F_2$

(3) microarray design

1) common loop design(CL)

Main design consists of D_{11} and D_{12} without replication (4 slides).

$$D_1 = \begin{pmatrix} \underline{00} & \underline{11} & \underline{10} & \underline{01} \\ 01 & 10 & 00 & 11 \end{pmatrix}$$

D_{11} D_{12}

2) cross swap design(CS)

Main design consists of two replications of D_{13} (4 slides).

$$D_2 = \begin{pmatrix} 00 & 01 & 11 & 10 \\ 11 & 10 & 00 & 01 \end{pmatrix}$$

The above design configuration was obtained after adjusting and rearranging the main design so that two rows have the same number of treatment combinations and the factor sums for the rows are the same.

In D_2, $(00 \oplus 01 \oplus 11 \oplus 10)(\mathrm{mod}\,2) = 00$: *row*1

$(11 \oplus 10 \oplus 00 \oplus 01)(\mathrm{mod}\,2) = 00$: *row*2

where the symbol $\oplus$ is defined as factor sum.

3) 4 replications of CL with 16slides

$$D_3 = \begin{pmatrix} 00 & 11 & 10 & 01 : & 01 & 10 & 00 & 11 : & 00 & 11 & 10 & 01 : & 01 & 10 & 00 & 11 \\ 01 & 10 & 00 & 11 : & 00 & 11 & 10 & 01 : & 01 & 10 & 00 & 11 : & 00 & 11 & 10 & 01 \end{pmatrix}$$

4) 2 replications each of CL and CS with 16 slides

$$D_4 = \begin{pmatrix} 00 & 11 & 10 & 01 : & 01 & 10 & 00 & 11 : & 00 & 01 & 11 & 10 : & 11 & 10 & 00 & 01 \\ 01 & 10 & 00 & 11 : & 00 & 11 & 10 & 01 : & 11 & 10 & 00 & 01 : & 00 & 01 & 11 & 10 \end{pmatrix}$$

5) 3 replication of D_{11}, 2 replication of D_{12} and 3 replication of D_{13} with 16 slides

$$D_5=\begin{pmatrix} 00 & 10 & 01 & 11 & 00 & 10 & : & 10 & 01 & 00 & 11 & : & 01 & 11 & 00 & 01 & 11 & 10 \\ 01 & 11 & 00 & 10 & 01 & 11 & : & 00 & 11 & 10 & 01 & : & 10 & 00 & 11 & 10 & 00 & 01 \end{pmatrix}$$
$$\underbrace{\qquad\qquad\qquad}_{D_{11}}\ \underbrace{\qquad\qquad\qquad}_{D_{11}}\ \underbrace{\qquad\qquad\qquad}_{D_{11}}\quad \underbrace{\qquad\qquad\qquad}_{D_{12}}\ \underbrace{\qquad\qquad\qquad}_{D_{12}}\quad \underbrace{\qquad\qquad\qquad}_{D_{13}}\ \underbrace{\qquad\qquad\qquad}_{D_{13}}\ \underbrace{\qquad\qquad\qquad}_{D_{13}}$$

3.2 2^3 experiments

3.2.1 Main designs of factorial experiment for 2^3

(1) number of blocks of size 2 : $\frac{2^3}{2}=2^{3-1}=4$

(2) block construction of D_{21}

1) main block : 000, and select 1 out of the remaining 7 treatment combinations

2) other blocks : Make pairs from the remaining six so as to be equivalent to factor sum for main block by the mod (2)

Table 2) 2^3 experiments

D_{21}	(000, 100)(001, 101)(010, 110)(011, 111)	F_2, F_3, F_2F_3
D_{22}	(000, 010)(001, 011)(100, 110)(101, 111)	F_1, F_3, F_1F_3
D_{23}	(000, 001)(010, 011)(100, 101)(110, 111)	F_1, F_2, F_1F_2
D_{24}	(000, 111)(001, 110)(010, 101)(110, 111)	F_1F_2, F_1F_3, F_2F_3
D_{25}	(000, 011)(001, 010)(101, 111)(100, 111)	$F_1, F_2F_3, F_1F_2F_3$
D_{26}	(000, 101)(001, 100)(100, 110)(010, 111)	$F_2, F_1F_3, F_1F_2F_3$
D_{27}	(000, 110)(010, 100)(011, 101)(001, 111)	$F_3, F_1F_2, F_1F_2F_3$

(3) Effect confounded

The method on main design of confounded factors are handled as following and be given also above table2.

1) D_{21}: 000ⓧ100 = (---ⓧ+--) = (- + +) [F_2, F_3, F_2F_3]

2) D_{24}: 000ⓧ111 = (---ⓧ+++) = (- - -) [F_1F_2, F_1F_3, F_2F_3]

3) D_{27}: (000ⓧ110) = (---ⓧ++-) = (- - +) [F_3, F_1F_2, $F_1F_2F_3$]

(4) Microarray design with 16 slides

1) Select 4 main designs randomly from Table 2, and adjust and rearrange each block as row 1 and row 2 so as to have the same number of each treatment combination.

example : Use $D_{24}D_{25}D_{26}D_{27}$

$$D_1 = \begin{pmatrix} 111 & 110 & 010 & 011 : & 000 & 001 & 101 & 100 : & 000 & 001 & 011 & 010 : & 110 & 100 & 101 & 111 \\ 000 & 001 & 101 & 100 : & 011 & 010 & 110 & 111 : & 101 & 100 & 110 & 111 : & 000 & 010 & 011 & 001 \end{pmatrix}$$

D_{24} D_{25} D_{26} D_{27}

2) Constructed by properly combining the main design in Table 2

i) Constructed by combining two replications of two arbitrary main designs and another main design

ii) Constructed by combining two replications of two arbitrary main designs Especially in case 2^4 experiments, We can be constructed by table3.

Table3) 2^4 experiments

D_{31}	(0000, 0001) (0010, 0011) (0100, 0101) (0110, 0111) (0100, 1100) (0101, 1101) (0110, 1110) (0111, 1111)	$F_1.F_2.F_3.F_1F_2.$ $F_2F_4.F_3F_4.F_2F_3F_4$
D_{32}	(0000, 0010) (0001, 0011) (0100, 0110) (0101, 0111) (1000, 1001) (1010, 1011) (1100, 1101) (1110, 1111)	$F_1.F_2.F_4.F_1F_2$ $F_1F_4.F_2F_4.F_1F_2F_4$
D_{33}	(0000, 0011) (0001, 0010) (0100, 0111) (0101, 0110) (1000, 1011) (1001, 1010) (1100, 1111) (1101, 1110)	$F_1.F_2.F_1F_2.F_3F_4$ $F_1F_3F_4.F_2F_3F_4.F_1F_2F_3F_4$
D_{34}	(0000, 0100) (0001, 0101) (0010, 0110) (0011, 0111) (1000, 1100) (1001, 1101) (1010, 1110) (1011, 1111)	$F_1.F_3.F_4.F_1F_3$ $F_1F_4.F_3F_4.F_1F_3F_4$
D_{35}	(0000, 0101) (0001, 0100) (0010, 0111) (0011, 0110) (1000, 1101) (1001, 1100) (1010, 1111) (1011, 1110)	$F_1.F_3.F_1F_3.F_2F_4$ $F_1F_2F_4.F_2F_3F_4.F_1F_2F_3F_4$
D_{36}	(0000, 0111) (0001, 0110) (0010, 0101) (0011, 0100) (1000, 1110) (1001, 1111) (1010, 1100) (1011, 1101)	$F_1.F_4.F_1F_4.F_2F_3$ $F_1F_2F_3.F_2F_3F_4.F_1F_2F_3F_4$
D_{37}	(0000, 0111) (0001, 0110) (0010, 0101) (0011, 0100) (1000, 1111) (1001, 1110) (1010, 1101) (1011, 1100)	$F_1.F_2F_3.F_2F_4.F_3F_4$ $F_1F_2F_3.F_1F_2F_4.F_1F_3F_4$
D_{38}	(0000, 1000) (0001, 1001) (0010, 1010) (0011, 1011) (0100, 1100) (0101, 1101) (0110, 1110) (0111, 1111)	$F_2.F_3.F_4.F_2F_3$ $F_2F_4.F_3F_4.F_2F_3F_4$
D_{39}	(0000, 1001) (0001, 1000) (0010, 1011) (0011, 1010) (0100, 1101) (0101, 1101) (0110, 1111) (0111, 1110)	$F_2.F_3.F_1F_4.F_2F_3$ $F_1F_2F_4.F_1F_3F_4.F_1F_2F_3F_4$
D_{310}	(0000, 1010) (0001, 1011) (0010, 1000) (0011, 1001) (0100, 1110) (0101, 1111) (0110, 1100) (0111, 1101)	$F_2.F_4.F_1F_3.F_2F_4$ $F_1F_2F_3.F_1F_3F_4.F_1F_2F_3F_4$
D_{311}	(0000, 1011) (0001, 1010) (0010, 1001) (0011, 1000) (0100, 1111) (0101, 1110) (0110, 1101) (0111, 1100)	$F_2.F_1F_3.F_1F_4.F_3F_4$ $F_1F_2F_3.F_1F_2F_4.F_2F_3F_4$
D_{312}	(0000, 1100) (0001, 1101) (0010, 1110) (0011, 1111) (0100, 1000) (0101, 1001) (0110, 1010) (0111, 1011)	$F_3.F_4.F_1F_2.F_3F_4$ $F_1F_2F_3.F_1F_2F_4.F_1F_2F_3F_4$
D_{313}	(0000, 1101) (0001, 1100) (0010, 1111) (0011, 1110) (0100, 1001) (0101, 1000) (0110, 1011) (0111, 1010)	$F_3.F_1F_2.F_1F_4.F_2F_4$ $F_1F_2F_3.F_1F_3F_4.F_2F_3F_4$
D_{14}	(0000, 1110) (0001, 1111) (0010, 1100) (0011, 1101) (0100, 1010) (0101, 1011) (0110, 1000) (0111, 1001)	$F_4.F_1F_2.F_1F_3.F_2F_3$ $F_1F_2F_4.F_1F_3F_4.F_2F_3F_4$
D_{15}	(0000, 1111) (0001, 1110) (0010, 1101) (0011, 1100) (0100, 1011) (0101, 1010) (0110, 1001) (0111, 1000)	$F_1F_2.F_1F_3.F_1F_4.F_2F_3$ $F_2F_4.F_3F_4.F_1F_2F_3F_4$

3.4 3×2 experiments

3.4.1 Main designs of factorial design for 3×2

Table 4) 3×2 experiments

D_{41}	(00, 01) (10, 11) 20,21)
D_{42}	(00, 11) (10, 21) (20, 01) (01, 10) (11, 20) (21, 00)
D_{43}	(00, 10) (10, 20) (20, 00) (01, 11) (11, 21) (21, 01)

(1) method of construction : Use gerneralized cyclic method

i) Add by 10 and continue until first circulation.

ii) Add by 10, 10, 11 by order and continue until circulation.

In the above, 20 and 21 may be used instead of 10 and 11, respectively.

(example)

1) D_{41} (00, 01) $\oplus$ 10(mod(3, 2)) = (10, 11) $\oplus$ 10 mod(3, 2)

= (20, 21) $\oplus$ 10 mod(3, 2) = (00, 01) : main block

D'_{41} (00, 01) $\oplus$ 20 mod(3, 2) = (20, 21) $\oplus$ 20 mod(3, 2)

= (10, 11) $\oplus$ 20 mod(3, 2) = (00, 01) circulation

$\therefore$ $D_{41} \cong D'_{41}$

2) D_{42} (00, 11) $\oplus$ 10(mod(3, 2)) = (10, 21) $\oplus$ 10 mod(3, 2)

= (20, 01) $\oplus$ 11 mod(3, 2) = (01, 10) $\oplus$ 10 mod(3, 2)

= (11, 20) $\oplus$ 10 mod(3, 2) = (21, 00) $\oplus$ 11 mod(3, 2)

= (00, 11) : circulation of main block

(2) other main design

i) D_{44} (00, 20) (10,00) (20,10)

[From the beginning, $\oplus$ 10 mod(3, 2) ,$\oplus$ 10 mod(3, 2),$\oplus$ 11 mod(3, 2)]

(01, 21) (11,01) (21,11) (00,20) [circulation of main block]

[From the beginning, $\oplus$ 10 mod(3, 2), $\oplus$ 10 mod(3, 2), $\oplus$ 11 mod(3, 2)]

ii) D_{45} (00, 21) (10, 01) (20, 11) (01, 20) (11, 00) (21, 10) (00,21) [circulation of main block]

[From the beginning, $\oplus$10, $\oplus$10, $\oplus$11, $\oplus$10, $\oplus$10, $\oplus$11 circulation]

Here $D_{44} \cong D_{43}$, $D_{45} \cong D_{42}$.

So we can see that the number of BS designs with 3 blocks is 3, not 5, and the number of XL and AL designs with 6 blocks is 3, not 5.

4. Microarry design

(1) BS : 2-replicates of D_{41} with 6 slides

$$D_1\begin{pmatrix} 00\,10\,20 : 01\,11\,21 \\ 01\,11\,21 : 00\,10\,20 \end{pmatrix}$$

(2) 2 replicates of D_{41} and D_{42} with 12 slides

$$D_2\begin{pmatrix} 00\,10\,20\,01\,11\,21 : 00\,10\,20\,01\,11\,21 \\ 01\,11\,21\,00\,10\,20 : 11\,21\,01\,10\,20\,00 \end{pmatrix}$$

5. Conclusion

Balanced factorial designs available in literature provide building blocks for constructing microarray designs with desired statistical properties. In particular, classical confounded designs for two level factors readily provide efficient designs for microarrays. Designs with block size two tabulated by Lewis and Tuck(1985) and Gupta(1987) can also be used with advantage. Fractions of full-factorials may also be used as building blocks, especially when the number of factors is large. We have used the row-column structure to display designs mainly to show their close connection with classical factorial designs. Also, directed graph displays of microarray designs become quite cumbersome with an increase in the number of levels and/or the number of factors.

6. Reference

[1] Eun Soon Shin, Jong Eun Lee, Kwan Jeh Lee (2005), A study on the item analysis of korea Developmental Inventory for preschoolers, *Journal of the Korean Data Analysis Society,* Vol. 7, No. 4, pp. 1213-1222.

[2] Hyun-Cheol Kang, Sang-Tae Han, Byoung-Cheol Jung, Youn-Ju Shin (2004). On the Algorithm of Recommendation System for Personalization, *Journal of the Korean Data Analysis Society,* Vol. 6, No. 4, pp. 1043-1050.

[3] So-Young Lee, Seung-Bae Choi, Kyu-Kon Kim, Hyung-Do Kim, Chang-Wan Kang (2005). Study on Development the Profit Model using Customer Segmentation in Medical Field - using Data Mining Technique, *Journal of the Korean Data Analysis Society*, Vol. 7, No. 2, pp. 523-532.

[4] Glonek, G., Solomon, P. (2004). Factorial and time course designs for cDNA microarray experiments, *Biostatistics,* 5, 89-111.

[5] Gupta, S. (1987). Generating generalized cyclic designs with factorial balance, *Commun. Statist. - Theor. Meth.*, 16, 1885-1900.

[6] Gupta, S., Mukerjee, R. (1996). Developments in incomplete block designs for parallel line assays, In: Ghosh, S., Rao, C. R. ed., *Handbook of Statistics,* 13, 875-902, Amsterdam: North-Holland.

[7] Hoil Kim (2005), A Note of Randem Effect in case of Two-level Data, *Journal of the Korean Data Analysis Society,* Vol. 7, No. 3, pp. 797-807.

[8] Kerr, M., Churchill, G. (2001a). Experimental designs for gene expression microarrays, *Bio statistics,* 2, 183-201.

[9] Kerr, M., Churchill, G. (2001b). Statistical design and analysis of gene expression microarray data, *Genet. Res.,* 77, 123-128.

[10] Kshirsagar, A. M. (1966). Balanced factorial designs, *J. R. Statist. Soc.,* B, 28, 59-67.

[11] Kshirsagar, A. M., Wang, W. (1996). Use of Lattice square designs in bioassays, *J. Biophar. Statist.,* 6, 185-199.

[12] Kuey Chung Choi, Byung Jun Park, Sung Ki Yang, Jung Hwa Lee (2005). Optimality Criteria in Experimental Designs, *Journal of the Korean Data Analysis Society,* Vol. 7, No. 2, pp. 441-452.

[13] Landgrebe, J., Bretz, F., Brunner, E. (2005). Efficient design and analysis of two colour factorial microarray experiments, *Comp. Statist. and Data Analysis* (In press)

[14] Lee, M. L. T. (2004). *Analysis of Microarray Gene Expression Data,* New York: Springer.