[UCI]I804:24011-200000233948

*2007*年 *2*月

博士學位論文

균형요인실험계획법을 이용한 *cDNA Microarray Design*에 관한 연구

朝鮮大學校 大學院

電算統計學科

李 正 和

균형요인실험계획법을 이용한 *cDNA Microarray Design*에 관한 연구

Analysis of cDNA Microarray Design using the Balanced Factorial Designs

*2007*年 *2*月 日

朝鮮大學校 大學院

電算統計學科

李 正 和

균형요인실험계획법을 이용한 *cDNA Microarray Design*에 관한 연구

指導敎授 崔 圭 丁

이 **論文**을 **博士學位申請 論文**으로 **提出**함

*2006*年 *10*月 日

朝鮮大學校 大學院

電算統計學科

李 正 和

李正和의 博士學位論文을 認准함

委 員 長　翰林大學校 敎授 _____ 印

委　　員　서울大學校 敎授 _____ 印

委　　員　朝鮮大學校 敎授 _____ 印

委　　員　朝鮮大學校 敎授 _____ 印

委　　員　朝鮮大學校 敎授 _____ 印

*2006*年　*12*月　　日

朝鮮大學校 大學院

목 차

목차 ····· i
표목차 ····· iv
그림목차 ····· v
ABSTRACT ····· vi

Ⅰ. 서 론 ····· *1*
1.1 연구배경 ····· 1
1.2 연구목적 ····· 2

Ⅱ. 완전랜덤화 설계 ····· *4*
2.1 실험계획 ····· 4
2.1.1. 실험계획의 원리 ····· 4
2.1.2. 실험계획법의 분류 ····· 5
2.1.3. 분산분석법 ····· 7
2.2 완전랜덤화 설계 ····· 8
2.2.1. 일원배치법 ····· 9
2.2.2. 이원배치법 ····· 9

Ⅲ. 요인실험설계 ····· *11*
3.1 요인실험법 ····· 11

3.1.1. 다요인실험법 ····· 11
3.1.2. 2^k-요인 실험 ····· 14
(1) 2^2-요인 실험 ····· 14
(2) 2^3-요인 실험 ····· 16
3.1.3. 3^k-요인 실험법 ····· 18
(1) 3^2-요인 실험 ····· 18

Ⅳ. 교락법과 일부실시법 ····· *20*

4.1 교락법 ····· 20
4.1.1. 교락의 의미 ····· 20
4.1.2. 정의대비(defining contrast) ····· 22
4.1.3. 2^k-요인실험에 대한 교락법 ····· 24
(1) 2^2-요인실험을 2블록(크기=2)에 실험하는 경우 ····· 24
(2) 2^2-요인실험을 2블록(크기=4)에 실험하는 경우 ····· 26
4.2 일부실시법 ····· 27
4.2.1. 부분요인실시법 ····· 27
4.2.2. (0, 1) 직교배열표를 이용한 부분실시 ····· 30
4.2.3. 선명도(resolution) ····· 31

Ⅴ. *Microarray* 실험계획 ····· *34*

5.1 Microarray ····· 34
5.2 Microarray 실험계획의 목적 ····· 38
5.3 Microarray 실험계획법의 종류 ····· 39
5.3.1. 그래프 표현법 ····· 39

5.2.2. 기준계획(reference design) ····· 41
5.2.3. 고리계획(loop design) ····· 43

Ⅵ. 요인 *Microarrays* ····· ***45***
6.1 2수준 요인 Microarrays ····· 45
6.1.1. 2^2-요인실험에서의 설계 ····· 45
6.1.2. 2^3-요인실험에서의 설계 ····· 52
6.1.3. 3×2 실험에서의 설계 ····· 54

Ⅶ. 결론 및 토론 ····· ***57***

참고문헌 ····· ***58***

부록 ····· ***66***

표 차 례

[표 2-1] 분산분석표 ····· 7
[표 2-2] 반복이 있는 이원배치법의 분산분석표 ····· 10
[표 3-1] 3요인의 제곱합과 자유도 ····· 13
[표 3-2] 반복이 없는 2^2-요인실험의 자료구조 ····· 15
[표 3-3] 2^2-요인실험의 요인효과 계수표 ····· 16
[표 3-4] 2^3-요인실험의 요인효과 계수표 ····· 17
[표 4-1] 2^3-요인실험에서 $I=ABC$인 경우 ····· 29
[표 4-2] 2^2형 직교배열표 ····· 30
[표 4-5] 2^3형 직교배열표 ····· 31
[표 6-1] 2^2형 요인배치법에서 유전자의 표현법 ····· 46
[표 6-2] 2^2형 요인배치 실험디자인 ····· 47
[표 6-3] 2^2요인 모델 ····· 47
[표 6-4] 2^2요인 실험설계 ····· 51
[표 6-5] 3×2요인을 위한 모델 ····· 52
[표 6-6] 2^3요인을 위한 모델 ····· 54

그 림 차 례

[그림 3-1] 직교다항식의 이해 ····· 18
[그림 5-1] cDNA가 만들어지는 과정 ····· 35
[그림 5-2] 유전자 발현 연구에 사용되는 DNA 어레이의 general overview ····· 36
[그림 5-3] cDNA 칩의 이미지 ····· 37
[그림 5-4] cDNA microarray실험 ····· 37
[그림 5-5] 대조군(C)을 Cy5로 처리군(T)을 Cy3로 부착 후 합성한 실험 ····· 39
[그림 5-6] 실험을 4회 반복한 경우 ····· 40
[그림 5-7] dye swap 디자인의 예 ····· 41
[그림 5-8] 기준계획의 예 ····· 42
[그림 5-9] 처리군이 5개인 고리계획 ····· 44
[그림 6-1] 2^2형 요인배치법에서 가능한 실험조합 ····· 46
[그림 6-2] 실험횟수가 4일 때의 실험 ····· 50
[그림 6-3] 6개의 슬라이드에 배치된 2^2 요인 실험 ····· 51

ABSTRACT

Analysis of cDNA Microarray Design using the Balanced Factorial Designs

By Lee, Jung-Hwa
Advisor : Prof. Choi, Kuey-Chung
Department of Computer Science and Statistics,
Graduate School of Chosun University

In our view, the main design issue with cDNA microarray experiments is the determination of which mRNA samples should be hybridized together on the same slide, bearing in mind the objectives of the experiment and the constraints on reagents and materials.

Usually, can define that is action that raises change under several conditions about study subject who is fixed as important terminology that is used in society diapason such as experiment(experiment) Iran physiotherapy, chemistry, biomedical engineering and observes and observe the phenomenon. Therefore, experimenter shall observes reason which make change in result through experiment and examines closely. Causes that can explain what present state actually exists much innumerably, and also, various constraints with expense, time make perfect cause

searching examination difficult. Accordingly, an actuality experiment between a lot of causes empirically or work which do to arrive in target that want illuminating attribute of leading persons handled in an experiment thus directly selecting some by dictionary information in several angles as purpose of an experiment speak can.

A.Fisher is scholar that is lead self-satisfaction which draft importance of experimental design in role and , adversely, statistics of statistics in experimental design while extends in the 1920s and the 30s and works in Rothamsted Experimental Station. He branch of advocated important principle those randomization(randomization), block anger (blocking), number of replication (replicates) be. And in formality of experimental design realization of problem and purpose establishment of an experiment, comprehension of constraint, suitable specification of the model etc. (that any leading person amounted interest when consider whether covariance(covariance) confirmed Itneun and model is crossing (crossed) model or is interest (nested) model because includes necessary leading persons out of leading persons and those leading persons reciprocal action in matrix to be included in matrix, and is interest model consider.

Also, model whether fixed effect(fixed effect) model or probability (random) effect whether search and in each model whether normality despotic rule about error is sound process of consideration), decision (decision of number of replication In Accordance With path of confidence interval or power of test required under constraint) of number of replication, randomization, selection of experimental design, a reserve experiment, source collection and analysis of variance follow.

In this paper, I studied desirable and reasonable balance block design for cDNA microarray design by an experiment once for all using traditional confounding method. Confirmed that traditional factor experimental design that is gotten from this paper is more useful than other what confounding method. Specially, suggested balanced block design for microarray experiment design than method that is suitable according to algorithm and circumstance, and generalized various microarray designs by latest literature.

In a sense, the tools and techniques of statistical design must continue to develop as more and more imaginative experiments are devised by biologists, but they must always be based on the sound principles of analysis of these experiments.

Ⅰ 서 론

1.1 연구배경

DNA Microarray 기술은 수천 개 또는 수만 개의 유전자의 발현을 동시에 탐색할 수 있는 새로운 과학기술이다. Human Genome Project(HGP)의 연구 결과로 우리는 인간이 지니고 있는 30억 개의 DNA 염기서열을 모두 해독하게 되었다. HGP는 생명공학의 급속한 발전과 함께 수많은 종류의 새로운 첨단 기술을 탄생시켰으며 이중에는 DNA chip 기술도 포함되어 있다. DNA chip에서 얻어진 자료를 간단히 Microarray라고 한다. Microarray에 대한 연구가 활발해 지면서 이에 Microarray의 설계에 대한 중요성도 인식되었다.

Landgreber 외 (2005)는 Yang 외 (2002)는 표준 데이터 정상화 절차가 단지 dyes 때문에 전체적인 효력을 제거한다고 지적했다. 그리고 특별한 방법을 사용하여 2^2, 2^3 그리고 3×2 실험을 설계했다. Wit 외 (2004)는 simulated annealing 알고리즘을 사용하여 효율적인 설계법을 찾아냈고, dye-swap 설계법은 연구에 속하는 조건의 수에 증가로써 매우 비효율적이 되는 것을 보여줬다. 그들은 또한 simulated annealing 알고리즘을 사용하여 얻는 설계법과 비교할 때 확실히 잘 실행하는 혼합된 루프(loop) 설계법을 정의했다.

Shah (1960)와 Kshirsagar (1966)에 의해 특성을 보기위한 균형요인배치법(balanced factorial design)은 농업 실험의 환경에서 주요하게 개발되어졌고, 균형 디자인(balanced design)은 다른 환경에서도 유용하게 사용하기 위하여 발견되었다. 예를 들어, Gupta 와 Mukerjee (1996), Kshirsagar 와 Wang (1996), 및 거기에 인용되는 것을 보면, 균형 설계법(balanced design)은 생물검정(bioassay) 실험을 위해 또한 광범위하게 연구되었다. Kerr와 Churchill (2001a,b), 및 Churchill (2002) 은 ANOVA 모형을 사용하여 microarray 실험에 관련시켜 몇몇 설계법에 세부적인 논의를 준비했다. 그들은 common reference microarray를 위해 루프설계법을 소개하고, 루프(loop) 설계법이 상당한 개선 이상의 효율적인 점에서 “common reference" 설계법에 대하여 제시한다.

Kerr와 Churchill (2001a)은 설계를 만들기 위해 A-optimality criterion을 이용하였다. Glonek과 Solomon (2003)은 역시 효율적인 설계법을 만들기 위한 방법을 제시하였다. 그들은 중요성의 특별한 대비의 추정을 하는 한은 종합적인 능률에 설계법이 대부분의 효율적인 설계법이 아닐지도 모른다는 것을 주장했다. 본 논문에서는 허용할 수 있는 설계법을 고려하고, 허용할 수 있는 설계법의 종류에 대해 그들 컴퓨터의 설계법의 조사를 제안했다.

Yang과 Speed (2002) 에 의해 제안된 모든 비교 설계법은 실험자가 연구에 속하는 조건사이에 모든 짝을 이룬 비교에서 동등하게 중요성을 가질 때 유용하다고 주장하였고 특히 Glonek과 Solomon (2003)은 고전적인 설계법은 단지 모든 추정가능한 대비의 변동을 최소화하는 것을 찾는 것이고 중요성의 대비는 아니라고 명확히 제시했다. 그런데 요인 실험의 고전적인 설계법에서 모든 추정 가능한 대비는 동등한 중요성이 드물다는 것을 유념해야한다. 그러므로 추정가능 할 수도 불가능 할 수도 있는 다른 대비와 비교하여 더 높고 정확한 중요성에 따라 대비를 추정하는 균형요인배치법(balanced factorial design)을 선택하는 것은 당연한 것이다. 실제로 고전적인 교락법은 블록에 완전 또는 부분 교락을 이용하여 중요성이 적은 것을 희생시켜 중요성의 대비에 더 높고 정확하게 달성시키는 것을 확실하게 한다. 고전적인 교락법은 균형요인배치법(balanced factorial designs)을 위해 매우 유용하여 요인 microarray 실험을 위해 적당한 설계법을 작성하는데 도움이 된다.

1.2 연구목적

이 논문의 목적은 cDNA microarray 실험을 위해 균형요인설계(balanced factorial design)를 제공하기 위함이다. 2개의 실험 조건을 각각의 microarray 슬라이드에 교배시키기 때문에, 배열은 크기 2의 블록을 형성한다. 따라서 블록 크기 2인 균형요인배치법(balanced factorial designs)과 요인 microarray 설계법에 대해서 연구하고자 한다. microarray 실험을 위한 균형설계법(balanced designs)은 연구에 유효한 균형요인설계법(balanced factorial designs)을 이용함으로서 쉽게 얻을 수 있음을 보인다. 또한 그룹분할가능하고 일반화된 사이클(cyclic) 디자인과 같은 블록 크기 2인 균형요인설계법

(balanced factorial designs) 역시 유용하다는 것을 보인다. 본 논문에서는 고차교호작용을 희생시키지 않고 main block을 만들어서 실험함으로서 그 효과를 구할 수 있는 방법을 제시하고자 한다.

일반적인 실험계획법에 관한 내용을 2장에서 소개하고 3장에서는 요인실험설계에 대하여 더 자세히 알아보고 4장에서는 교락법과 일부실시법에 대하여 논한다. 그리고 5장에서는 microarray에 대해서 알아본다. 그리고 고전적인 교락된 설계법과 요인 microarray을 이용한 cDNA microarray design에 대하여 논하기로 한다. 본 논문에서는 2^m에 대해서만 다루기로 한다. 또한 부록에 R-프로그램을 이용한 2요인에 대한 교락에 대한 내용과 design을 제시한다.

II 완전랜덤화 설계

2.1 실험계획

2.1.1 실험계획의 원리

실험설계의 기본원리는 수십 년 동안 비교실험을 위한 실험설계의 기본원칙으로 지켜지고 있다. 그 내용은 다음과 같다. 첫 번째 랜덤화(randomization)는 실험해야할 처리들을 실험단위에 동일한 확률로 배치하는 방법을 말한다. 이것을 실시하는 이유는 비교분석하고자 하는 처리효과가 다른 잡음 요인들의 효과에 고루 영향을 받게 함으로써 실제로 처리간의 비교에는 잡음 요인들의 효과들이 상쇄되도록 하기 위해서이다. 이것이 바로 랜덤화를 이용한 평준화 효과이다. 또한 랜덤화는 통계적 분석을 용이하게 해준다. 실험 자료에 대한 분산분석을 실시할 경우 여기에서 주로 사용된 F-검정은 정규분포로부터 뽑은 랜덤표본을 가정으로 한 것에서부터 시작된 것이다. 이 랜덤화 방법이 이러한 가정의 타당성을 제공하는 것이다.

실험환경의 동질성은 단순한 랜덤화의 방법으로는 한계가 있을 수 있다. 실험단위들의 차이를 단순한 랜덤화에 의하여 평균화시켰을 경우 오차제곱합이 커져 분석력이 떨어질 수 있다. 따라서 동질의 실험단위들을 모아 블록을 만들어 각각의 블록에 모든 처리들을 실험하는 것이 합리적이라 하겠다. 이 블록화방법은 오차제곱합을 크게 줄여 실험의 정도(precision)를 높일 수 있다. 모든 실험설계마다 반드시 블록화를 해야 하는 것은 아니지만, 실험의 성격상 필요하다고 판단되면 블록화는 하는 편이 유리하다. 그리고 여러 집단의 특성치를 비교할 때 각 집단으로부터 여러 번의 독립적 무작위 추출을 실시한다.

이 여러 번의 독립적 무작위 추출이 바로 반복실험의 개념이다. 반복(replication)은 각각의 처리를 둘 이상의 실험단위에 가함을 의미한다. 실험 결과의 정밀성을 유지하려면 반복 실험을 하여야 한다. 다른 한편으로는 같은 자극에 대하여 반복실험을 함으로써 오차제곱합을 얻을 수 있어 분산분석을 통한 통계분석이 용이하게 된다. 반복화

가 이루어지지 않는 실험의 경우 대개 오차에 대한 추정이 불가능하여 분산분석을 하기 어렵다. 이러한 경우를 우리는 포화된 모형(saturated model)이라 한다.

그리고 교락의 원리와 직교화의 원리도 있다. 교락법은 구할 필요가 없는 2인자 교호작용이나 고차의 교호작용을 블록과 교락시키는 방법으로 검출할 필요가 없는 요인이 블록의 효과와 교락하게 함으로써 실험의 효율을 높일 수 있다. 그리고 직교화의 원리는 요인간에 직교성을 갖도록 실험계획하여 데이터를 구하면 같은 실험횟수라도 검출력(power of test)이 더 좋은 검정을 할 수 있고, 정도가 더 높은 추정을 할 수 있는 것으로 밝혀져 있다. 교호작용 가운데 기술적으로 생각하여 무시될 수 있는 것을 주효과(main effect)와 교락시켜 실험의 크기를 줄일 수 있는 실험계획을 짜도록 고안된 표를 직교배열표(orthogonal array table)라고 하며, 직교배열표는 요인간의 직교성을 이용하여 만들어 놓은 표이다.

2.1.2 실험계획법의 분류

실험계획법의 종류는 무수히 많지만 비교적 많이 사용되는 것들을 실험을 배치하는 형에 의하여 다음과 같이 분류할 수 있다.

*(1)*요인배치법*(factorial design)*

인자의 각 수준의 모든 조합에 대하여 실험을 행하는 것으로 실험순서는 랜덤하게 정한다. 인자의 수에 따라서 인자가 하나인 경우는 일원배치법(one-way factorial design), 인자가 둘인 경우는 이원배치법(two-way factorial design), 그리고 인자가 세 개 이상인 경우는 다원배치법(multi-way factorial design)이라고 부른다.

인자수가 n개 있고 모든 인자의 수준이 2인 경우는 2^n형 요인배치법(2^n factorial design), 3수준인 경우는 3^n형 요인배치법이라고 부른다. 만약 p개의 인자는 2수준이고 q개의 인자는 3수준이면 2^p3^q형 요인배치법이라고 부른다.

이처럼 인자 각 수준의 모든 조합에서 실험이 행해지며, 실험순서가 완전히 랜덤하게 행해지는 실험계획법을 완비형배치법이라고 부르기도 한다.

(2) 분할법(split-plot design)

요인배치법에서 실험순서가 완전히 랜덤하게 정해지지 않고, 실험전체를 몇 단계로 나누어서 단계별로 랜덤화하는 실험계획법을 분할법이라고 부른다. 따라서 분할법은 실험순서를 정하는 방법에 있어서 요인배치법과 다르다. 예를 들어 A(l수준), B(m수준) 2인자의 실험을 이원배치법으로 한다면 lm개의 조합 전체를 랜덤화하여 주어야한다. 그런데 실험에 취급되는 인자의 종류에 따라 완전 랜덤화가 극히 곤란한 경우가 있다. 인자 A가 극히 높은 수준의 가열온도라고 하고 B는 원료의 종류라고 하자. 높은 가열온도는 수시로 바꾸기 곤란하므로 우선 A에 대하여 랜덤화하고 다음에 B에 대하여 랜덤화하여, 하나의 가열온도 수준에서 모든 원료에 대하여 실험을 끝마치고 다른 가열온도 수준으로 옮기면 편리할 것이다. 이것은 A, B의 두 단계로 나누어 랜덤화하였다고 볼 수 있다. 이와 같은 분할법의 경우는 요인배치법과 데이터의 나열은 동일하나 분석방법에 차이가 있다.

(3) 교락법(confounding method)

요인배치법의 실험에서는 인자간의 모든 교호작용이 검출되지만 고차의 교호작용은 구하지 않아도 좋을 때가 많다. 검출할 필요가 없는 교호작용을 다른 요인(특히 블록을 많이 사용함)과 교락하도록 배치하는 방법을 교락법이라고 한다. 이 실험은 실험횟수를 늘리지 않고도 전체 실험을 블록화하여 간편하게 실험할 수 있는 장점이 있다.

(4) 일부실시법(fractional factorial design)

불필요한 교호작용이나 고차의 교호작용은 구하지 않고 각 인자의 조합 중에서 일부만 선택하여 실험을 실시하는 방법을 일부실시법이라고 한다. 교락법에서는 블록과 교호작용을 교락시키나 일부실시법에서는 불필요한 교호작용을 중요한 요인들과 교락시켜 주는 것이 보통이다. 이 실험계획법은 실험횟수를 가능한 한 작게 하고자 할 때 사용된다. 일부실시법을 계획할 때에는 직교배열표(orthogonal array table)를 사용하는 것이 편리하다.

위에서 설명한 것 외에도 불완비블록계획법이나 반응표면계획법이 있다. 모든 실험계획법이 위의 어느 하나에만 반드시 속하는 것은 아니다. 일부실시법을 짜면서 교락

의 원리를 적용시키면 일부실시법도 되고 교락법도 될 수 있다.

2.1.3 분산분석법

분산분석은 1923년 피셔에 의하여 처음 소개된 것으로 알려져 있으나 실제로는 1885년 에지워스(Edgeworth)가 분산분석표와 비슷한 형태의 계산을 했다는 사실이 스네데커(Snedecor)에 의하여 밝혀졌다.

피셔는 처음 멘델의 유전(genetics)을 상관관계의 일종으로 설명하면서 이것을 유전될 수 있는 부분과 유전될 수 없는 부분으로 분할하여 설명하였다. 또 유전될 수 있는 부분을 상가적(additive) 요인과 우성적(dominance) 요인 그리고 유전인자 사이의 교호작용(interaction)으로 분할하여 설명하였다. 이것이 분산분석의 시초이다.

분산분석은 실험이나 표본조사를 통하여 수집된 값에 대하여 총평균을 중심으로 한 자료들의 총변동(total variation)을 계산하고, 이 총변동을 다시 각 변인(source)들의 변동으로 분류한 후 이것을 통하여 그 변인이 반응변수에 영향을 주고 있는지를 분석하는 방법이다. 따라서 분산분석은 모분산 혹은 어떤 종류의 분산에 대하여 분석이 목적이 아니고 변동의 원인이 무엇인지를 밝히는 것이 목적이라 하겠다.

$$\text{모형 : } y_{ij}=\mu+a_i+e_{ij} \qquad i=1,\cdots,l;\ j=1,\cdots,m$$

일 때 분산분석표는 다음과 같다.

[표 2-1] 분산분석표

변인	제곱합	자유도	평균제곱	평균제곱의 기대치	F값
처리	SS_{TR}	$l-1$	$SS_{TR}/(l-1)$	$\sigma^2_E+m\sigma^2_{TR}$	$\frac{MS_{TR}}{MS_E}$
오차	SS_E	$l(m-1)$	$SS_E/(l(m-1))$	σ^2_E	
전체	SS_T	$lm-1$			

위의 [표 2-1]을 보면 처리의 효과가 각각 통계적으로 유의할 경우 평균제곱합은 모분산 σ^2의 불편추정치가 될 수 없다. 그러나 여기에서는 모분산 σ^2을 구하는 것이 목적이 아니고 처리라는 변인이 총변동에 어느 정도 비중을 차지하고 있는가에 관심이 있다고 볼 수 있다. 즉 각 변인의 통계적 유의성을 검정함으로써 자료의 변동원인이 어디에 있는지를 분석하는 것이 목적이라 하겠다.

2.2 완전랜덤화 설계

완전 랜덤화 설계(CRD : completely randomized design)는 비교실험에 대한 오차조절 설계 중 가장 단순한 형태이다. 그리고 독립 이표본에서 데이터를 얻는 상황이 완전 랜덤화 설계의 가장 초보적인 경우이다. 완전 랜덤화 설계의 실용성은 낮은 편이지만 이 설계방식의 개념과 대응되는 모형, 모형의 분석 방법 등의 기본 아이디어는 더 복잡한 여타의 설계에도 그대로 적용된다.

완전 랜덤화 설계는 동질적인 실험단위들이 가용한 경우에 사용한다. 이때 동질성을 너무 좁은 의미로 해석할 필요는 없다. 물론 동질적인 실험단위들보다는 완벽하게 동일한 실험단위들을 실험에 동원할 수 있다면 더 바랄 것이 없겠으나, 자연 상태에서 완벽한 동일성을 보여주는 개체는 존재하지 않는다. 따라서 이상적인 동질의 실험단위란 제반 성질이 최대한 흡사한 실험단위를 의미한다. 하지만 실험에 따라 동질성 의미부여는 달라지는 것이 정상이다. 예를 들어 한국인 성인 남자가 실험 대상이라면, 한국인으로서 성인이며 남자인 점에서 동질성이 있는 것이지, 얼굴이 닮았다든가 체중이 비슷하다는 의미는 아니다. 이런 경우에는 동질적인 실험단위라도 변동이 클 것이다. 하지만 공장에서 일정하게 조절해 놓은 공정 조건 하에서 생산되는 카세트테이프들을 실험단위로 간주하면, 테이프들의 측정 특성의 변동 폭은 매우 작으리라고 짐작된다. 결국 동질성의 확보란 실험에 따라 상대적이고 가변적일 수밖에 없다.

완전 랜덤화 설계는 정의는 다음과 같다.

$N = rt$개의 동질적인 실험단위가 가용하고, t개의 처리를 비교하려한다. N개의 실험단위를 t개의 그룹으로 등확률(equal probability)로 랜덤 분할하는데, 각 그룹마다 r개

의 실험단위를 갖도록 한다. 그 다음 l개의 그룹에 랜덤 배정한다.

이렇게 완전 랜덤화 실험 결과 얻어지는 데이터의 분석을 분산분석의 원리에 의하여 실시하게 된다.

2.1.1 일원배치법

실험설계의 기본원리인 랜덤화(randomization), 블록화(blocking), 반복(replication) 중 랜덤화의 한 방법이다. 이것은 랜덤화의 원리를 충실히 따른 원리이다. 선택된 변인 이외의 환경을 랜덤화의 방법을 통하여 평준화시킴으로써 실험의 분석력을 높이는데 목적이 있다.

모형 : $y_{ij}=\mu+a_i+e_{ij} \qquad i=1,\cdots,l,\ j=1,\cdots,m$ (2.1)

이처럼 어떤 반응 값에 대하여 하나의 변인의 영향만을 살피고자 할 때 이러한 실험설계를 일원배치법(one-way layout)이라 한다. 분산분석의 결과는 앞에서 제시한 [표 2-1]과 같다.

2.2.2 이원배치법

모형 : $y_{ijk}=\mu+a_i+b_j+ab_{ij}+e_{ijk} \quad i=1\cdots l,\ \ j=1,\cdots m,\ k=1,\cdots,r$ (2.2)

제시된 모형처럼 변인이 둘인 경우를 이원배치법(two-way layout)이라 한다. 이때 변인을 어떻게 취급하느냐에 따라 난괴법과 요인실험으로 생각할 수 있다. 첫 번째 요인의 수준 하나를 선택한 다음 다른 요인을 임의로 선택하는 경우이다. 이 방법은 선택된 한 요인의 수준이라는 균등한 조건하에서 다른 요인을 실험하는 것으로서 이것이

바로 실험설계의 원리 가운데 하나인 블록화이다. 이를 흔히 난괴법(randomized complete block design : RCBD)이라 한다. 이 난괴법은 이원배치법임에는 틀림없지만 블록과 처리라는 두 변인의 교호작용이 없다는 전제가 필요하다. 다음으로 두 변인에 대한 이원배치법으로 이른바 요인실험(factorial experiment)을 생각할 수 있다.

[표 2-2] 반복이 있는 이원배치법의 분산분석표

변인	제곱합	자유도	평균제곱	평균제곱의 기대치	F값
A	SS_A	$l-1$	$\frac{SS_A}{l-1}$	$\sigma^2_E + m\sigma^2_A$	$\frac{MS_A}{MS_E}$
B	SS_B	$m-1$	$\frac{SS_B}{m-1}$	$\sigma^2_E + l\sigma^2_A$	$\frac{MS_B}{MS_E}$
$A \times B$	$SS_{A\times B}$	$(l-1)(m-1)$	$\frac{SS_{A\times B}}{(l-1)(m-1)}$	$\sigma^2_E + r\sigma^2_{A\times B}$	
오차	SS_E	$lm(r-1)$	$\frac{SS_E}{lm(r-1)}$	σ^2_E	
전체	SS_T	$lm-1$			

Ⅲ 요인실험설계

실험설계의 가장 중요한 목적은 최소의 실험횟수(비용)로 최대의 실험효과를 얻는데 있다. 실험의 과정으로부터 한 요인을 고정시키고 다른 요인의 수준을 변화시켜가면서 실험하는 방법으로는 최적 값을 찾기 힘들 뿐 아니라 실험비용도 많이 든다는 사실을 알 수 있다. 따라서 이러한 문제를 효과적으로 해결하는 방법으로 이른바 요인실험(factorial experiment)을 생각할 수 있다. 요인실험에서는 처리들(treatments)이 각 요인들의 수준별 처리조합(treatment combination)으로 구성된다. 이러한 처리들을 실험함으로써 각 요인의 효과는 물론 요인들의 비교식 등을 추정할 수 있다.

3.1 요인실험법

3.1.1 다요인실험법

요인은 보통 질적 요인(qualitative factor)과 양적 요인(quantitative factor)으로 구분된다. 질적 요인이란 요인의 성질이 질적인 것이어서 수준이 종류로 표시되는 요인을 말한다. 예로서 남녀의 성별, 촉매의 종류 등과 같이 수량적으로 측정할 수 없는 요인을 의미한다. 반면에 양적 요인이란 요인의 수준이 수량적으로 표시되는 요인으로서 온도, 습도 등의 요인을 일컫는다. 보통 각 처리의 수준은 0,1,…로 표시하는데 양적 요인인 경우 숫자가 적을수록 낮은 수치를 갖는 수준을 의미한다. 예로서 양적 요인이 온도이고 수준의 수가 3으로서 0,1,2로 표시하면 0의 수준은 가장 낮은 온도를, 1은 중간 온도를, 2는 가장 높은 온도를 나타낸다. 두 개의 수준을 갖는 또 하나의 양적 요인 습도를 택하여 0,1로 표시한 후 두 요인의 처리조합을 고려할 때 예로서 00이란 가장 낮은 온도와 가장 낮은 습도에서 처리되는 처리조합을 의미한다.

다요인실험법의 모형은 다음과 같이 쓸 수 있다. 여기서는 모든 처리조합이 n번 반복되는 균형자료라 가정하자.

$$y_{ijk..p} = \mu + a_i + b_j + (\text{나머지 주효과}(main\ \ effect))$$

$$+ (ab)_{ij} + (\text{나머지 2차 교호작용}(two - factor\ \ actions)) \qquad (3.1)$$

$$+ (\text{모든 고차 교호작용}(high\ \ order\ \ actions) + e_{ij\cdots p}$$

$$i = 1,\cdots,l\ ;\ \ j = 1,\cdots,m\ \ ;\ k = 1,\cdots,n\ \ ;\ \ \cdots\ \ ;\ p = 1,\cdots,r$$

예로서 요인의 수가 셋일 때 $y_{ijk.} = \sum_{p=1}^{r} y_{ijk}$은 처리조합 ijk에서의 총 n개의 관찰치의 합을 나타낸다. 이때의 모형은 식 (3.1)을 이용하여 다음과 같이 나타낼 수 있다.

$$y_{ijkp} = \mu + a_i + b_i + c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ijk} + e_{ijkp} \qquad (3.2)$$

$$i = 1,\cdots,l\ ; j = 1,\cdots,m\ ; k = 1,\cdots,n\ ;\ p = 1,\cdots,r$$

그리고 각 제곱합과 자유도는 다음 [표 3-1]과 같이 나타난다.

[표 3-1] 3요인의 제곱합과 자유도

요인	제곱합	자유도
A	$SS_A = mnr\sum_{i=1}^{l}(\bar{y}_{i\cdots} - \bar{y}_{\cdots\cdot})^2$	$l-1$
B	$SS_B = lnr\sum_{j=1}^{m}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots\cdot})^2$	$m-1$
C	$SS_C = lmr\sum_{k=1}^{n}(\bar{y}_{\cdot\cdot k} - \bar{y}_{\cdots\cdot})^2$	$n-1$
$A\times B$	$S_{AB} - S_A - S_B$	$(l-1)(m-1)$
$A\times C$	$S_{AC} - S_A - S_C$	$(l-1)(n-1)$
$B\times C$	$S_{BC} - S_B - S_C$	$(m-1)(n-1)$
$A\times B\times C$	$S_{ABC} - (S_A + S_B + S_C + S_{A\times B} + S_{A\times C} + S_{B\times C})$	$(l-1)(m-1)(n-1)$
E	$SSE = \sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{l=1}^{r}(y_{ijkl} - \bar{y}_{ijk\cdot})^2$	$lmn(r-1)$
전체	$SST = \sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{l=1}^{r}(y_{ijkl} - \bar{y}_{\cdots\cdot})^2$	$lmnr-1$

$$SS_{ABC} = r\sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{k=1}^{n}(\bar{y}_{ijk\cdot} - \bar{y}_{ij\cdot\cdot} - \bar{y}_{i\cdot k\cdot} - \bar{y}_{\cdot jk\cdot} + \bar{y}_{i\cdots} + \bar{y}_{\cdot j\cdot\cdot} + \bar{y}_{\cdot\cdot k\cdot} - \bar{y}_{\cdots})^2$$

$$SS_{AB} = nr\sum_{i=1}^{l}\sum_{j=1}^{m}(\bar{y}_{ij\cdot\cdot} - \bar{y}_{i\cdots} - \bar{y}_{\cdot j\cdot\cdot} + \bar{y}_{\cdots\cdot})^2$$

$$SS_{AC} = mr\sum_{i=1}^{l}\sum_{k=1}^{n}(\bar{y}_{i\cdot k\cdot} - \bar{y}_{i\cdots} - \bar{y}_{\cdot\cdot k\cdot} + \bar{y}_{\cdots\cdot})^2$$

$$SS_{BC} = lr\sum_{j=1}^{m}\sum_{k=1}^{n}(\bar{y}_{\cdot jk\cdot} - \bar{y}_{\cdot j\cdot\cdot} - \bar{y}_{\cdot\cdot k\cdot} + \bar{y}_{\cdots\cdot})^2$$

3.1.2 2^k-요인 실험

다요인실험 중 가장 널리 이용되는 실험은 모든 k개의 요인이 각각 수준이 2이거나 3인 2^k- 또는 3^k-요인실험이다. 본 논문에서는 2^k-요인실험(2^k-factorial experiment)에 대해서 알아본다. 2^k-요인설계는 실험점의 수효가 가장 적은 설계이기 때문에 특히 많은 요인들이 개재된 실험의 초기 단계에서 유용하다.

(1) 2^2- 요인실험

$k=2$인 경우를 생각해 보자. 이것은 A, B 두 요인의 수준이 각각 2인 경우로 A요인의 수준을 0,1 B 요인의 수준을 0,1이라 하면 이들의 처리조합은 00, 10, 01, 11 (실제로는 A_0B_0, A_0B_1, A_1B_0, A_1B_1)의 4개가 가능하다. 반복수가 각 처리조합 당 하나인 가장 단순한 형태의 계획 아래서 얻어지는 모형은 다음과 같다. 단, 분석을 간단히 하기 위하여 실험오차 e_{ij}는 없는 것으로 하고 데이터의 구조식을 표현하였다.

$$y_{ij}=\mu+a_i+b_i+(ab)_{ij} \quad i,j=0,1$$

위의 모형에 $i,j=0,1$을 대입하여 계산하면

$$\mu=\frac{1}{4}(y_{00}+y_{01}+y_{10}+y_{11})$$

$$a_0=\frac{1}{4}(y_{00}+y_{01}-y_{10}-y_{11})$$

$$b_0=\frac{1}{4}(y_{00}-y_{01}+y_{10}-y_{11})$$

$$(ab)_{00}=\frac{1}{4}(y_{00}-y_{01}-y_{10}+y_{11})$$

[표 3-2] 반복이 없는 2^2-요인실험의 자료구조

		요인 B		합계
		B_0	B_1	
요인 A	A_0	y_{00}	y_{01}	$T_{0\cdot}$
	A_1	y_{10}	y_{11}	$T_{1\cdot}$
합 계		$T_{\cdot 0}$	$T_{\cdot 1}$	T

인자 A의 두 수준간의 효과의 차이 $a_1 - a_0$을 A로 표시하면

$$\begin{aligned} A &= a_1 - a_0 \\ &= \frac{1}{2}(y_{11} + y_{10} - y_{01} - y_{00}) \\ &= \frac{1}{2}[ab + a - b - (1)] \end{aligned} \tag{3.2}$$

B의 주효과는

$$B = \frac{1}{2}[ab + b - a - (1)] \tag{3.3}$$

이고, 두 인자 A, B간의 교호작용(interaction)의 효과는

$$AB = \frac{1}{2}[ab + (1) - a - b] \tag{3.4}$$

로 정의한다.

위에서 구한 주효과와 교호작용은 다음과 같이 해석할 수 있다.

$$A = \frac{1}{2}[(A\text{인자의 수준 } 1\text{의 데이터의 합}) - (A\text{인자의 수준 } 0\text{의 데이터의 합})]$$

$$B = \frac{1}{2}[(B\text{인자의 수준 } 1\text{의 데이터의 합}) - (B\text{인자의 수준 } 0\text{의 데이터의 합})]$$

$$AB = \frac{1}{2}[(A,B\text{인자의 수준합이 짝수인 데이터의 합}) - (A,B\text{인자의 수준합이 홀수인 데이터의 합})]$$

위의 내용을 요인 효과 계수표를 다시 정리하였다.

[표 3-3] 2^2-형 요인효과 계수표

처 리 조 합	요 인 효 과		
	A	B	AB
00 (1)	−	−	−
10 a	+	−	+
01 b	−	+	−
11 ab	+	+	+

(2) 2^3-요인 실험

$k=3$인 경우를 나타낸다. 이것은 A,B,C의 세 요인의 수준이 각각 2인 경우로 A 요인의 수준을 0,1 B 요인의 수준을 0,1 C 요인의 수준을 0,1이라 하면 이들의 처리조합은 000, 001, 010, 011, 100, 101, 110, 111 (실제로는 $A_0B_0C_0$, $A_0B_0C_1$, $A_0B_1C_0$, $A_0B_1C_1$, $A_1B_0C_0$, $A_1B_0C_1$, $A_1B_1C_0$, $A_1B_1C_1$)의 8개가 가능하다. 그리고 2^2-요인 실험과 동일한 절차를 통해 2^3-요인 실험의 요인효과도 다음과 같이 구할 수 있다.

[표 3-4] 2^3-요인실험의 요인효과 계수표

처 리 조 합		요 인 효 과						
		A	B	AB	C	AC	BC	ABC
000	(1)	−	−	+	−	+	+	−
100	a	+	−	−	−	−	+	+
010	b	−	+	−	−	+	−	+
110	ab	+	+	+	−	−	−	−
001	c	−	−	+	+	−	−	+
101	ac	+	−	−	+	+	−	−
011	bc	−	+	−	+	−	+	−
111	abc	+	+	+	+	+	+	+

위의 표는 몇 가지 중요한 특징을 지닌다. 모든 계수열의 "+", "-" 수는 같으며. 모든 계수열은 서로 직교하며, 두 열의 곱은 다른 열을 생성한다.

위의 [표 3-4]로부터 주효과 A는

$$\begin{aligned} A &= \frac{1}{4}[a+ac+ab+abc-(1)-c-b-bc] \\ &= \frac{1}{4}[(A\text{인자의 }1\text{수준의 데이터의 합}) \\ &\qquad -(A\text{인자의 }0\text{수준의 데이터의 합})] \\ &= \frac{1}{4}(T_{1..}-T_{0..}) \end{aligned}$$

가 된다. 같은 방법으로 주효과 B, C와 교호작용 AB, AC, BC를 구하면

$$B=\frac{1}{4}(T_{.1.}-T_{.0.})$$

$$C=\frac{1}{4}(T_{..1}-T_{..0})$$

$$AB = \frac{1}{4}[(1) + c + ab + abc - b - bc - a - ac]$$
$$= \frac{1}{4}[(A, B\text{인자의 수준합이 짝수인 데이터의 합}) - (A, B\text{인자의 수준합이 홀수인 데이터의 합})]$$
$$= \frac{1}{4}(T_{11.} + T_{00.} - T_{01.} - T_{10.})$$

$$AC = \frac{1}{4}(T_{1.1} + T_{0.0} - T_{0.1} - T_{1.0})$$

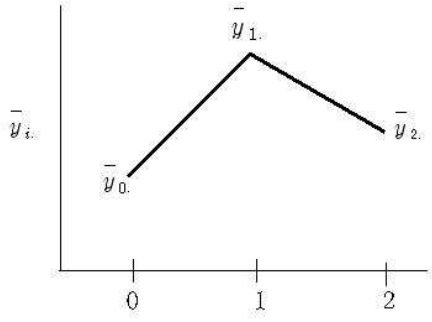
$$BC = \frac{1}{4}(T_{.11} + T_{.00} - T_{.01} - T_{.10})$$

3.1.3 3^k-요인 실험법

(1) 3^2-요인실험

각 요인의 수준이 3인 3^k-요인실험법을 생각해 보자. 그 중 $k=2$인 3^2-실험의 경우 수준을 0,1,2라 놓으면 가능한 처리조합은 00, 01, 02, 10, 11, 12, 20, 21, 22가 된다. 2^2-요인실험에서는 각 제곱합의 자유도가 1이어서 더 이상의 분해는 불가능하였다. 하지만 3^2-요인실험의 경우 주효과의 자유도는 2, 2차 교호작용의 자유도는 2^2, 3차 교호작용의 자유도는 2^3이 된다.

[그림 3-1] 직교다항식의 이해



만약 A가 양적인 관계를 나타내는 계량인자(quantitative factor)이고, 수준간의 간격이 일정할 때에는 인자 A의 일차효과(linear effect)와 이차효과(quadratic effect)의 존재여부를 찾아볼 수 있다. [그림 3-1]을 고려하여 보자. 만약 인자 A가 직선적은 반응관계를 나타낸다면, 0수준에서 1수준으로 변하는 동안 A의 일차효과는 $\bar{y}_{1.}-\bar{y}_{0.}$이고, 수준이 1에서 2로 변하는 동안의 일차효과는 $\bar{y}_{2.}-\bar{y}_{1.}$가 된다. 따라서 요인 A의 일차효과 C_L은 두 직선의 평균기울기로서 나타낼 수 있어

$$\begin{aligned} C_L &= (T_{1.}-T_{0.})+(T_{2.}-T_{0.}) \\ &= T_{2.}-T_{0.} \end{aligned} \tag{3.5}$$

로 쓸 수 있는데. 하나의 대비(contrast)이다. 여기서 대비 [-1 0 1]을 선형(1차)다항계수 (linear polynomial coefficient vector)라 한다.

반면에 2차 효과(quadratic effect)가 있다면, 2차 효과 C_Q는 두 기울기의 차이로 표현할 수 있다.

$$\begin{aligned} C_Q &= (T_{2.}-T_{1.})-(T_{1.}-T_{0.}) \\ &= T_{2.}-2T_{1.}+T_{0.} \end{aligned} \tag{3.6}$$

를 2차효과라고 부르고, 이것도 하나의 대비이다. 여기서 대비 [1 -2 1]을 2차 다항계수(quadratic polynomial coefficient vector)라 한다.

Ⅳ 교락법과 일부실시법

지금까지 처리조합이 완전확률화계획에 의해 실험되는 상황을 연구의 대상으로 삼았었다. 그러나 요인의 수가 많은 경우 처리조합의 수가 과다하게 커져서 실제로 실험하는데 어려움이 따른다. 예로서 2^7-요인실험의 경우 각 처리조합의 반복수를 하나로 국한하여도 128번의 실험을 필요로 한다. 이와 같은 경우 확률화에 의한 실험을 하는데도 어려움이 있고, 비용 또한 너무 많이 들어 부득이 실험횟수를 제한할 경우가 발생한다. 이러한 문제를 해결하기 위해 실험횟수를 늘리지 않고 블록으로 나누어 동질적인 환경 내에서 처리조합을 배치하는 경우와 처리조합 중 부분만을 선택하여 실험을 행하는 부분요인실시법 등이 연구의 대상이 된다.

4.1 교락법

4.1.1 교락의 의미

교락(confounding)이란 블록 내에서 처리조합을 효율적으로 배치하도록 이용되는 요인실험계획의 한 기법이라 할 수 있다. 개념을 이해하기 위해 반복수가 하나인 2^2-요인실험을 고려하여 보자. 4가지 가능한 처리조합을 동일한 실험조건 아래서 실험할 수 있다면 실험의 형태와 결과는 앞에서 다룬 이원배치법이나 3장에서 언급한 2^2-요인실험으로 분석을 행할 수 있다. 그렇지 않은 경우는 블록의 사용을 고려하여야 한다.

우리는 여기서 이 실험이 반드시 블록 크기가 2인 두 블록에서 행해져야 한다고 가정하자. 앞에서처럼 y_{ij}을 요인 A의 i번째 수준과 요인 B의 j번째 수준에서 관찰된 반응치라 하고 아래와 같이 배치되었다고 가정하자.

d_1 블록 1 : $(y_{00} \quad y_{11})$ 혹은 $(00 \;\; 11)$로 표시
블록 2 : $(y_{01} \quad y_{10})$ 혹은 $(01 \;\; 10)$로 표시

가능한 모형은

$$y_{ijk} = \mu + a_i + b_j + ab_{ij} + \beta_k + \varepsilon_{ijk} \qquad (i,j=0,1, k=1,2) \tag{4.1}$$

이고, 행렬의 형태로 고쳐 쓰면 다음과 같다.

$$\begin{pmatrix} y_{001} \\ y_{012} \\ y_{102} \\ y_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ a_0 \\ a_1 \\ b_0 \\ b_1 \\ ab_{00} \\ ab_{01} \\ ab_{10} \\ ab_{11} \\ \beta_1 \\ \beta_2 \end{pmatrix} + \boldsymbol{\epsilon} \tag{4.2}$$

여기서

추가제약식 $\sum_i a_i = 0$, $\sum_j b_i = 0$, $\sum_i ab_{ij} = 0$(모든 j에 대해서), $\sum_j ab_{ij} = 0$(모든 i에 대해서), $\sum_k \beta_k = 0$을 주면 $a_0 = -a_1$, $b_0 = -b_1$, $ab_{01} = ab_{10} = -ab_{11} = -ab_{00}$, $\beta_1 = -\beta_2$가 되어 (식 4.2)를 다시 쓰면

$$\begin{pmatrix} y_{001} \\ y_{012} \\ y_{102} \\ y_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ a_0 \\ b_0 \\ ab_{00} \\ \beta_1 \end{pmatrix} + \boldsymbol{\epsilon} \tag{4.3}$$

가 된다. 계획행렬의 계수는 4이고 4번째 열과 5번째 열은 같아 ab_{00}과 β_1은 분리해서

추정할 수 없음을 알 수 있다. 정리하면 2^2-요인실험을 d_1과 같이 두 블록에 배치하면 주효과는 추정가능하다. 이와 같이 두(혹은 그 이상)요인의 효과를 분리해서 추정할 수 없을 때 두 효과는 교락(confounded)되었다고 한다. 교락법을 논하기 이전에 정의대비에 대하여 알아보도록 한다.

4.1.2 정의대비(defining contrast)

만약 2^3형 요인설계에서 교호작용 ABC를 블록과 교락시킨다면 이것을 $I=ABC$ 로 표현하고, 정의대비(defining contrast)라고 부른다. 정의 대비란 단순히 어떤 요인이 블록과 교락되어 있는가를 나타내는 표현식이다. 정의대비를 이용하여 다음과 같은 선형 표현식

$$L = x_1 + x_2 + x_3 \ (\mathrm{mod} 2)$$

를 만든다. 여기서 x_i는 i번째 인자의 수준으로 0, 1만을 취하는 값이다. 만약 정의대비에 i번째 인자가 나타나 있지 않으면 선형표현식에 x_i는 나타나지 않는다. 위의 식은 L은 0, 1만이 될 수 있으며 2개의 블록을 $L=0$이 되는 것과 $L=1$이 되는 조합으로 나누면 된다.

블록 1	블록 2
(1)	a
ab	b
ac	c
bc	abc

일반적으로 2^k-요인실험의 경우 이를 2^s블록에 배치하고자 할 때는 다음과 같은 s개의 정의대비가 필요하게 된다.

$$\begin{cases} L_1 = 0,1 \\ L_2 = 0,1 \\ \vdots \\ L_s = 0,1 \end{cases}$$

위의 정의대비에 의하며 배치하면 다음과 같은 결과를 얻는다.

	블록	1	2		b
L_1		0	0		1
L_2	=	0	0	⋯	1
⋮		⋮	⋮		⋮
L_S		0	1		1

2^3-요인실험을 2^2블록(블록의 크기=2)에 배치하는 경우를 생각해 보자. 이때 교호작용 AB와 ABC을 블록과 교락시키려고 한다. 따라서 독립정의대비로 다음을 선택하였다고 하자.

$$\begin{cases} L_1 = x_1 + x_2 = 0 \\ L_2 = x_1 + x_2 = 1 \end{cases} \quad : \; AB\text{가 블록과 교락}$$

$$\begin{cases} L_1 = x_1 + x_2 + x_3 = 0 \\ L_2 = x_1 + x_2 + x_3 = 1 \end{cases} \quad : \; ABC\text{가 블록과 교락}$$

이 정의대비에 의한 배치결과는 다음과 같다.

	블록	1	2	3	4
L_1		0	0	1	1
L_2	=	0	1	0	1

4.1.2 2^k-요인실험에 대한 교락법

(1) 2^2-요인실험을 *2*블록*(*크기*=2)*에 실험하는 경우

2^2-요인 실험에 대한 블록이 2개이고 요인이 2개인 불완비블록설계를 생각해 보자. x_1, x_2을 각각 요인 A와 B의 수준이라 하면 다음과 같은 방법에 의하여 실험을 실시할 수 있다.

반복 1		반복 2		반복 3	
$x_1=0$	$x_1=1$	$x_2=0$	$x_2=1$	$x_1+x_2=0$	$x_1+x_2=1$
블록 1	블록 2	블록 3	블록 4	블록 5	블록 6
0 0 (1)	1 0 a	0 0 (1)	0 1 b	0 0 (1)	0 1 b
0 1 b	1 1 ab	1 0 a	1 1 ab	1 1 ab	1 0 a

여기에서 각 반복 내 처리들의 관측치를 y_{ijk}라 하면 y_{ijk}는 다음과 같다.

$$y_{ijk} = \mu + \alpha x_{1i} + \beta x_{2j} + \alpha\beta x_{1i} x_{2j} + \gamma_k + e_{ijk} \qquad (4.4)$$

단 $x_{1i} = 0,1,\ \ x_{2j} = 0,1,\ \ i,j,k = 0,1,\ \ \gamma_k$는 블록의 효과

반복 1의 경우는 주효과 A의 블록효과가 교락된 경우이다.

$E(y_{000}) = \mu + \gamma_1$

$E(y_{011}) = \mu + \beta + \gamma_1$

$E(y_{102}) = \mu + \alpha + \gamma_2$

$E(y_{112}) = \mu + \alpha + \beta + \alpha\beta + \gamma_2$

$E(y_{000} + y_{011} - y_{102} - y_{112}) = 2\alpha + \alpha\beta + 2(\gamma_2 - \gamma_1)$

이때 $\alpha\beta = 0$인 경우에도 $E(y_{000} + y_{011} - y_{102} - y_{112}) = 2\alpha + 2(\gamma_2 - \gamma_1)$이므로 주효과 A와 블록효과가 서로 교락된 것을 알 수 있다. 따라서 블록의 효과가 존재할 경우 주효과 A는 추정불가능하게 된다.

한편 대비$_A$=대비$_{block}$=$-y_{001} - y_{011} + y_{102} + y_{112}$로 주효과 A와 블록효과에 대한 대비가 같음을 알 수 있다. 따라서 주효과 A는 반복 1의 실험결과를 이용할 수 없고 반복 2와 3의 실험 자료를 이용하여 구해야 한다.

반복 2의 경우는 주효과 B의 블록효과가 교락된 경우이다.

$E(y_{001}) = \mu + \gamma_1$

$E(y_{101}) = \mu + \alpha + \gamma_1$

$E(y_{012}) = \mu + \beta + \gamma_2$

$E(y_{112}) = \mu + \alpha + \beta + \alpha\beta + \gamma_2$

$E(y_{012} + y_{112} - y_{001} - y_{101}) = 2\beta + \alpha\beta + 2(\gamma_2 - \gamma_1)$

이때 $\alpha\beta = 0$인 경우에도 $\frac{1}{2}(y_{012} + y_{112} - y_{001} - y_{101})$이므로 주효과 β와 블록효과가 서로 교락된 것으로 주관심사인 주효과 β가 추정불가능하게 된다. 이 경우에도 역시 주효과 β와 블록효과를 나타내는 대비는 다음과 같다.

대비$_B$=대비$_{block}$=$-y_{001} - y_{012} + y_{012} + y_{112}$

주효과 B는 반복 2의 실험결과를 이용할 수 없고 반복 1과 3의 실험 자료를 이용하

여 구해야 한다.

마지막으로 반복 3의 경우는 교호작용 AB가 블록효과와 교락된 경우이다.

$E(y_{001}) = \mu + \gamma_1$

$E(y_{111}) = \mu + \alpha + \beta + \alpha\beta + \gamma_1$

$E(y_{012}) = \mu + \beta + \gamma_2$

$E(y_{102}) = \mu + \alpha + \gamma_2$

위 식들의 관계로부터 다음과 같은 사실을 얻을 수 있다.

$$E(y_{001} + y_{111} - y_{012} - y_{102}) = \alpha\beta + 2(\gamma_1 - \gamma_2)$$

이 경우 블록효과와 교호작용 $\alpha\beta$가 서로 교락된 것으로, 블록효과가 없을 경우 교호작용 $\alpha\beta$는 $(y_{011} + y_{111} - y_{012} - y_{102})$로 추정된다. 그러나 블록효과가 있을 경우 $(\gamma_1 \neq \gamma_2)$ 교호작용 $\alpha\beta$는 추정불가능하게 된다. 이 경우 역시 교호작용과 블록효과에 대한 대비는 서로 같음을 알 수 있다. 즉 대비$_{AB}$=대비$_{block}$=$y_{001} - y_{012} - y_{012} + y_{111}$이다. 따라서 교호작용 AB의 경우에도 반복 1과 2의 실험결과를 이용하여 구할 수 있다.

이렇게 한 반복에 한 요인이 교락되게 실험하는 방법을 부분적으로 교락된 실험이라고 한다.

(2) 2^3-요인실험을 *2*블록*(*크기*=4)*에 실험하는 경우

2^3-요인실험을 실시하고자한다. 단, 교호작용 ABC을 블록과 교락시키기로 한다.

$$L_1 = x_1 + x_2 + x_3 \pmod 2$$

| $L_1=0$ |

블록 1

0	0	0	(1)
0	1	1	bc
1	0	1	ac
1	1	0	ab

주블록

| $L_1=1$ |

블록 2

0	0	1	c
0	1	0	b
1	0	0	a
1	1	1	abc

여기에서 블록 1을 구성한 다음 블록 1에 포함되지 않은 다른 처리를 하나 골라 (예를 들면 (1 1 1)) 블록 1의 각 처리들에 더하면 블록 2의 결과를 얻을 수 있음을 알 수 있다. 이때 블록 1은 처리 (0 0 0)를 포함하며 블록 1을 이용하여 블록 2를 생성할 수 있다는 점에서 블록 1을 주블록(principal block 혹은 key block)이라 한다.

4.2 일부실시법

4.2.1 부분요인실시법

요인실험법에서 요인의 수 k의 증가는 많은 실험횟수를 요구한다. 예로서 2^7-요인실험의 경우 처리조합 당 반복수를 한 번으로 제한하여도 128번의 실험을 필요로 한다. 7개의 요인이 동시에 연구 대상이 되는 것은 기업체 등 실제의 적용의 경우 매우 흔히 일어날 수 있는 일이다. 2^7-요인실험의 경우 주효과에 대한 자유도는 불과 7이고, 2차 교호작용에 대한 자유도는 $\binom{7}{2}=21$인데 반해 3차 이상의 교호작용에 대한 자유도는 99나 된다. 따라서 의미가 상대적으로 적은 고차의 교호작용을 무시할 수 있다면, 주효과나 2차 교호작용 등 저차의 교호작용에 관한 정보를 실험횟수를 줄이고도 얻을 수 있을 것이다. 이와 같이 모든 처리조합을 다 실험하는 완전요인실험(complete factorial design)과는 다르게 그 부분만을 택하여 실험하는 요인실험을 부분요인실시법(fractio

-nal factorial design)이라 한다.

요인과 수준 그리고 실험횟수가 정해졌을 때 처리조합을 선택하는 기준을 필요로 한다. 그 기준을 이해하기 위해 8개의 처리조합을 가진 2^3-요인실험 중 4개의 처리조합을 택하여 실험하는 (즉 반을 택하는) 부분요인실험 (2^{3-1} - 부분요인실시라 함)을 생각하여 보자. 2^3-요인실험의 요인효과 계수표는 앞에서 제시한 [표 3-4]와 같다.

그 표를 이용하여 네 처리조합 100, 010, 001, 111을 택하여 실험했다고 가정하자. A, B, C 주효과와 AB, AC, BC, ABC 교호작용 효과의 추정량은 선택된 네 처리조합의 선형결합으로서 다음 식과 같은 방법으로 해석하면 (예로서 A의 효과의 추정량은 A의 수준 1에서의 처리합 - A의 수준 0에서의 처리합)

$$A\text{의 효과} = \frac{1}{2}[a + abc - (b + c)]$$

$$B\text{의 효과} = \frac{1}{2}[b + abc - (a + c)]$$

$$C\text{의 효과} = \frac{1}{2}[c + abc - (a + b)]$$

$$AB\text{의 효과} = \frac{1}{2}[c + abc - (a + b)]$$

$$BC\text{의 효과} = \frac{1}{2}[a + abc - (b + c)]$$

$$AC\text{의 효과} = \frac{1}{2}[b + abc - (a + c)]$$

$$ABC\text{의 효과} = \frac{1}{2}[b + abc - (a + c)]$$

가 되어 A의 효과 = BC의 효과, B의 효과 = AC의 효과, C의 효과 = AB의 효과가 됨을 알 수 있다.

다시 말하면 2^{3-1}-부분요인실험은 요인효과 계수표에서 요인 효과 ABC열에 '+'를 갖는 처리조합을 선택함으로서 얻어지고 그 결과 A와 BC, B와 AC, C와 AB는 분리하여 추정할 수 없다. 이와 같이 분리하여 추정할 수 없는 두 요인은 교락(혹은 별명 ; aliased) 되었다고 한다. ABC의 효과를 보면 선택된 모든 처리조합에 대한 계수

가 '+'임을 볼 수 있어서 이에 대응하는 정의관계식은

$$I = ABC$$

로 표시한다. 한 요인효과와 교락된 요인효과를 얻기 위해 정의관계식에 요인효과를 곱하면 된다. 위의 경우에서 보면

$$A\text{의 별명} = A \times I = A(ABC) = A^2BC = BC$$
$$B\text{의 별명} = B \times I = B(ABC) = AB^2C = AC$$
$$C\text{의 별명} = C \times I = C(ABC) = ABC^2 = AB$$

가 되어 같은 결과가 나옴을 알 수 있다.

위의 결과를 보면 부분요인실험의 경우 주효과도 부득이 어느 요인과는 교락되나 될 수 있으면 고차의 교호작용과 교락되도록 처리조합을 선택하여야 할 것이다.

[표 4-1] 2^3-요인실험에서 $I = ABC$인 경우

처 리 조 합	요 인 효 과						
	A	B	AB	C	AC	BC	ABC
100 a	+	−	−	−	−	+	+
010 b	−	+	−	−	+	−	+
001 c	−	−	+	+	−	−	+
111 abc	+	+	+	+	+	+	+

이와 같이 2^3-요인실험의 $\frac{1}{2}$실시를 $\left(\frac{1}{2}\right)2^3$-요인실험 혹은 2^{3-1}요인실험이라고 한다. 이것은 2^3-요인실험에서 2^3개의 처리조합 중 2^{3-1}개만을 선택하여 실험하는 방법을 의미하고 이것을 일부실시법(fractional replication)이라 한다.

이 결과로 주효과 A와 교호작용 BC가 서로 교락되어 BC의 효과를 무시할 수 있을 때 주효과 A의 추정이 가능하다. 이때 두 요인 A와 BC는 서로 별명관계(alias)에 있

다고 하거나 교락(confounding)되어 있다고 한다. 주효과 B는 교호작용 AC와, 주효과 C는 교호작용 AB와 서로 별명관계에 있음을 알 수 있다. 이것은 일부실시에서 비롯된 결과이다.

4.2.2 (0, 1) 직교배열표를 이용한 일부실시

요인실험에서 처리조합을 표시할 때 요인의 수준을 나타내는 숫자로 표시하는 경우가 흔하다 예를 들면 2^2-요인실험의 경우 4종류의 처리 (0, 0), (0, 1), (1, 0), (1, 1)으로 표시한다. 이것을 일반적으로 다음과 같이 나타낸다.

$$(i, j),\ \text{단}\ i = 0, 1,\ j = 0, 1$$

또 $i + j = h\,(mod\ 2)$와 같은 연산법을 이용하면 각 요인에 대한 효과를 계산하는데 필요한 계수를 얻을 수 있다. 이러한 결과를 직교배열표(orthogonal array)라 한다.

[표 4-2] 2^2-형 직교배열표

처 리	A	B	AB
	i	j	$i+j$
0 0	0	0	0
0 1	0	1	1
1 0	1	0	1
1 1	1	1	0

이때 주효과 A에 대한 대비$_A$는 첫째 열인 i열을 이용하여 1에 대응하는 측정치에서 0에 대응하는 측정치를 빼면 된다. 주효과 B는 둘째 열인 j열을, 마지막으로 교호작용 AB는 셋째 열인 $(i+j)$열을 이용하면 된다. 2^3-형 직교배열표는 [표 4-3]과 같다.

[표 4-3] 2^3-형 직교배열표

처 리	A	B	AB	C	AC	BC	ABC
	i	j	k	$i+j$	$i+k$	$j+k$	$i+j+k$
000	0	0	0	0	0	0	0
100	0	0	1	0	1	1	1
010	0	1	0	1	0	1	1
110	0	1	1	1	1	0	0
001	1	0	0	1	1	0	1
101	1	0	1	1	0	1	0
011	1	1	0	0	1	1	0
111	1	1	1	0	0	0	1

일반적으로 2^n-요인실험에 대한 2^n-형 직교배열표 역시 이러한 방법으로 얻을 수 있다.

4.2.3 선명도(resolution)

2^{k-p}-요인실험에서 $(R-t)$차 이상의 교호작용이 모두 0이라 가정하면 $t < \frac{R}{2}$차 교호작용까지 추정이 가능할 때 이러한 요인실험을 결정지표 R인 일부실시법이라 한다. R는 $R = \text{Ⅲ}, \text{Ⅳ}, \text{Ⅴ}$ 인 경우의 세 종류가 있다.

선명도 Ⅲ : 선명도 Ⅲ을 갖는 실시법은 어떤 주효과도 다른 주효과와는 교락되지 않으나, 주효과는 2차 요인효과와는 교락되며, 2차 요인효과는 다른 2차 요인효과와 교락된다. 다음과 같이 정리할 수 있다.

1. 주효과끼리는 서로 별명관계에 있지 않으며, 주효과는 2차 교호작용과 2차 교호작용 일부는 서로 자기들끼리 별명관계에 있다.
2. 따라서 2차 교호작용 이상을 무시하면 모든 주효과를 구할 수 있다.

예를 들어 2^{3-1}Ⅲ-요인 실험의 $\frac{1}{2}$실시를 한 경우를 보면 3요인에 대하여 실험횟수가 8번이며 정의대비가 $I=ABC$이면 별명관계는 $A=BC,\ \ B=AC,\ \ C=AB$가 된다.

선명도 Ⅳ : 선명도 Ⅳ를 갖는 실시법은 어떤 주효과도 다른 주효과나 2차 요인효과와는 교락되지 않으나, 2차 요인효과는 다른 2차 요인효과와는 교락된다. 이 경우는 짝수 결정지표로서 R가 홀수인 경우와 다소 다른 특징이 있다. $R=4$인 경우 $t<\frac{R}{2}$로부터 $t=1$이고 $(R-t)=3$이 되어 3차 이상의 교호작용을 무시하면 모든 주효과가 독립적으로 구해진다. 이때 2요인 교호작용은 주효과와 별명관계에 있지 않으며, 2요인 교호작용끼리 서로 별명관계에 있다. 따라서 다음과 같이 정리할 수 있다.

1. 주효과는 3요인 교호작용과 별명관계에 있고 주효과 및 2요인 교호작용과는 별명관계에 있지 않다.
2. 따라서 3요인 이상 교호작용을 무시하면 주효과를 독립적으로 구할 수 있으며 2요인 교호작용의 일부를 무시하면 2요인 교호작용의 나머지 일부를 구할 수 있다.

예를 들어 2^{4-1}Ⅳ-요인실험에 대하여 정의대비가 $I=ABCD$인 경우의 별명관계는 다음과 같다.

$A=BCD,\ \ B=ACD,\ \ C=ABD,\ \ D=ABC,$

$AB=CD,\ \ AC=BD,\ AD=BC$

선명도 Ⅴ : 선명도 Ⅴ를 갖는 실시법은 어떤 주효과와 2차 요인효과도 다른 주효과나 2차 요인 효과와는 교락되지 않으나, 2차 요인효과는 다른 2차 요인효과와는 교락된다. $R=5$이면 $t<\frac{R}{2}$에서 $t=2$이고 $(R-t)=3$이므로 3요인 이상의 교호작용을 무시하면 주효과 및 2요인 교호작용을 구할 수 있다.

1. 주효과 및 2요인 교호작용은 3요인 이상의 교호작용과 별명관계에 있으며 주효과 및 2요인 교호작용과는 별명관계에 있지 않다.

2. 따라서 3요인 이상의 교호작용을 무시하면 주효과 및 2요인 교호작용을 구할 수 있다.

예를 들어 2^{5-1} V-요인실험에서 정의대비 $I = ABCDE$이면 별명관계는 다음과 같다.

$AB = CDE$, $BD = ACE$, $AC = BDE$, $BE = ACD$, $CD = ABE$, $AE = BCD$,
$BD = ACE$, $DE = ABC$, $A = BCDE$, $B = ACDE$, $C = ABDE$, $D = ABCE$,
$E = ABCD$

Ⅴ *Microarray* 실험계획

5.1 Microarray

세포내에서 기능을 하는 단백질을 구성하는 아미노산들을 결정짓는 염기들이 있는 DNA부분을 유전자라고 한다. genomic DNA는 이중나선구조로 핵 안에서 단단히 꼬여져 있다. DNA의 염기 배열 중 단백질로 변화되는 부분은 핵 안에서 mRNA로 전사된 다음 핵공을 통해 빠져나온 후, 세포질의 리보솜에서 단백질 합성을 시작한다. 핵공이란 핵막에 있는 작은 구멍으로 mRNA를 비롯한 여러 가지 물질의 이동통로이다. 즉, DNA의 염기 배열 중 단백질로 변환되는 부분은 전사되어 세포 안에서 mRNA로 존재한다.

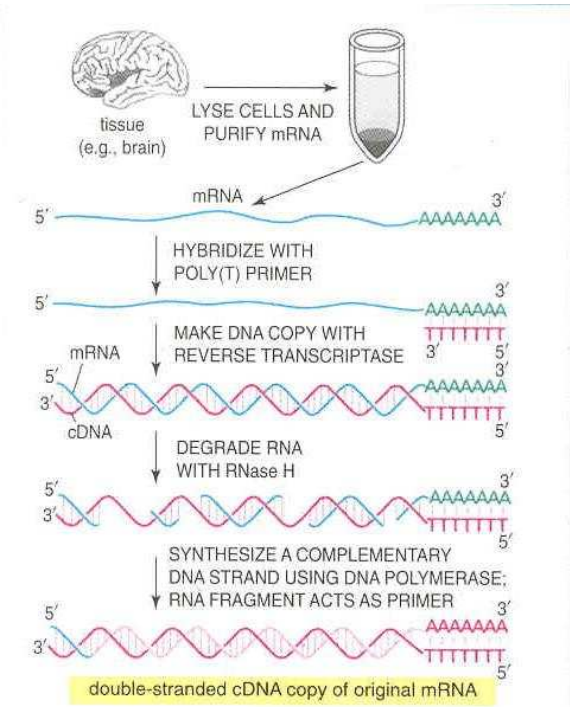
세포에 존재하는 mRNA의 양이 그대로 단백질의 양과 반드시 동일한 것은 아니지만, 앞으로 만들어질 가능성 있는 단백질의 양을 나타내고 있는 것이기에 이를 통해서 세포의 활동상과 상태를 알 수 있다. 대체로 특정 유전자를 전사(transcription)하여 만들어진 mRNA는 추후 번역(translation)과정을 통해 단백질로 생성하는데 사용하므로 mRNA의 양은 해당 유전자의 발편정도를 나타내는 척도로 삼을 수 있다. 즉, 측정된 mRNA의 양이 많을 때에는 해당 유전자가 활성화되었다는 것을 뜻하며, 적을 때에는 유전자가 비활성화 되었다는 것을 뜻한다. 따라서 mRNA를 통해서 유전자에 대한 연구를 할 수 있다. 그러나 RNA는 DNA와 달리 단일 나선으로 불안정하며, 또한 수명이 짧기 때문에 실험실에서 다루기가 힘들다. 그래서 보통 유전자 발굴에는 RNA를 인공적으로 이중나선의 DNA로 전환시키는데, 이것이 바로 cDNA(complementary DNA)다. cDNA는 DNA안에 mRNA로 전사되는 부분만을 인공적으로 만들어 낸 것이다.

[그림 5-1]은 cDNA를 만드는 실험 과정이다. DNA에서 전사된 mRNA는 핵에서 빠져나올 때 3‘ 부분에 아데니(A)를 달고 나온다. 이것을 poly A tail 이라고 하며, 세포질에 있는 mRNA가 가지고 있는 특성 중의 하나이다. mRNA가 있는 곳에 티민(T)으로 구성된 poly T primer와 RNA를 DNA로 역전사 시키는 역전사효소(reverse transcriptase)를 넣어주면, DNA가 합성되면서 RNA-DNA 결합체가 생긴다. 여기에 다시 RNA를 분해하는 RNase H 효소를 넣어주면, RNA가 분해되는데, 이때 모든

RNA가 분해되지는 않는다. 분해되지 않고 남아있는 RNA 조각들을 primer로 사용, DNA 중합효소(DHA polymerase)를 이용하여 상보적인 DNA 가닥을 만들게 되는데. 이것이 cDNA이다.

[그림 5-1] cDNA가 만들어지는 과정

(Alberts et al., 2002)

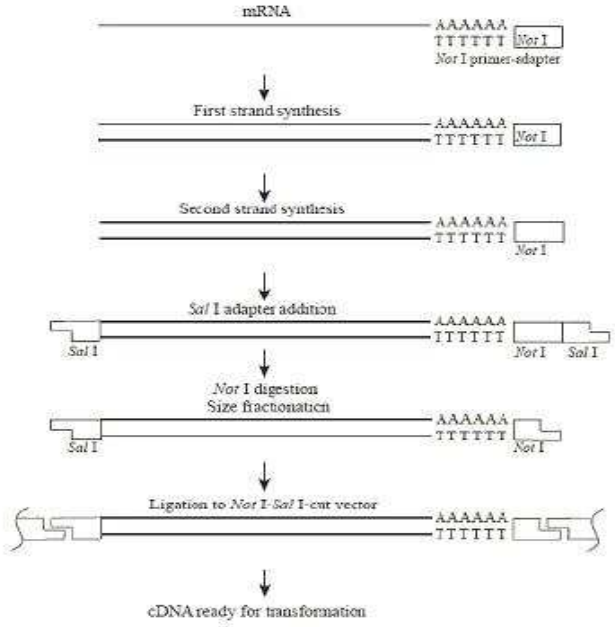


cDNA 자체가 반드시 하나의 완전한 유전자를 의미하지는 않는다. cDNA는 생체내에서 발현되는 유전정보의 일부이다. 생체 내에서 발현되는 유전정보들을 모아놓은 것을 cDNA library라 한다. 보통 cDNA microarray에서는 소형 기판 위에 고밀도로 미리 올려놓은 프로브(probe)로 cDNA를 주로 쓴다. 프로브는 어레이 위에 미리 심어 놓는 DNA 서열로, 이를 통해 특정 조건에서 어느 서열 정보가 발현되었는지를 알 수 있다.

DNA microarray는 DNA가 상보적으로 결합하는 성질을 이용하는 것이다. [그림 5-2]에 유전자 발현 연구에 사용되는 DNA microarray의 핵심 원리가 잘 나타나 있다.

관심 있는 조직에서 mRNA를 추출하여 유전자 발현을 탐지할 수 있는 염료(dye)를 붙인 cDNA를 만들고, 이를 미리 만들어진 DNA microarray에 혼성화(hybridization) 시키면 서로 상보적인 서열이 있는 부분에서는 형광이 나타난다. 형광의 강도를 측정해서 유전자의 발현량을 알아보는 것이다.

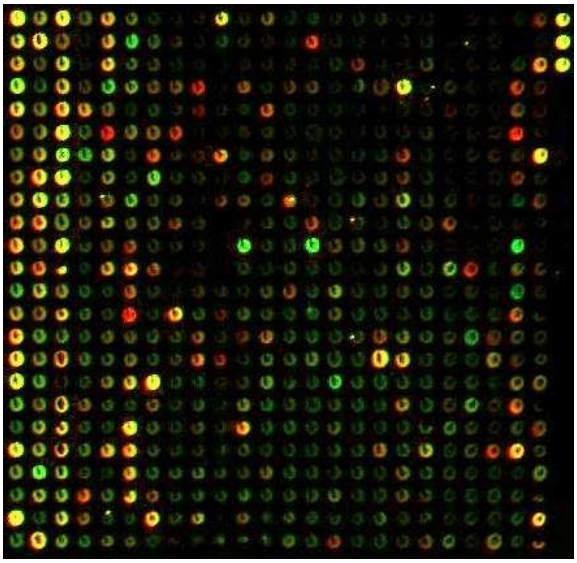
[그림 5-2] 유전자 발현 연구에 사용되는 DNA 어레이의 general overview



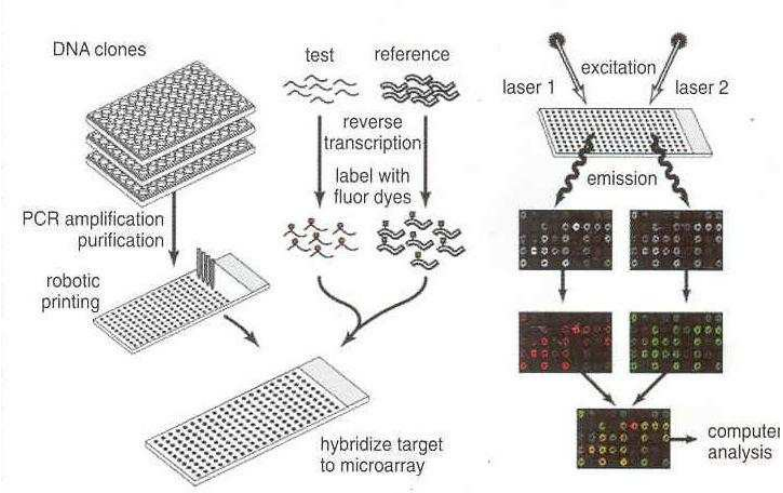
microarray 실험 과정은 크게 세 부분으로 나눌 수 있다. 첫 번째 microarray를 만드는 과정, 두 번째 표본들로부터 RNA를 추출하여 microarray와 혼성화시키는 과정, 그리고 스캐닝을 하여 정량화하고 분석을 하는 과정이다.

[그림 5-3]은 유전자 발현연구에 관한 microarray 실험을 통해 얻어진 cDNA칩 이미지이다. 녹색이나 적색은 특정 조건(예를 들면, 암 또는 정상세포)에 치우친 유전자임을 뜻하고, 노란색은 동일하게 발현되는 유전자임을 나타낸다.

[그림 5-3] cDNA 칩의 이미지



[그림 5-4] cDNA microarray실험(Duggan et al., 1999)



cDNA microarray 실험에 대한 그림은 [그림 5-4]에 나타나 있다. cDNA microarray 슬라이드 한 장에 기록되는 cDNA 수는 적게는 수 백개에서 많게는 수 만개에 달한

다. 즉 cDNA 실험을 통해서 동시에 수 백 개에서 수 만개에 달하는 유전자의 발현양상을 한 번에 살펴 볼 수 있음을 뜻한다.

5.2 Microarray 실험계획의 목적

마이크로어레이 실험을 통해서 나온 자료를 이용하여, 각 유전자의 발현정도를 객관적으로 판단할 수 있어야 한다. 또한, 각 처리의 효과 차이를 검출할 수 있는 자료구조가 되어야 한다. 유의한 유전자의 발굴작업은 microarray 실험을 통해서 얻을 수 있는 가장 기본적인 결과이다. 유의성 검정에는 여러 처리간의 비교를 통하여 유전자의 발현 정도를 판별하는 유의성 검정을 할 수 있다. 서로 다른 조직에서 추출한 유전자의 비교, 정상 세포에서 얻은 조직과 암 세포에서 얻은 조직의 비교, 서로 다른 시점에서 얻은 조직의 비교 등을 생각해 볼 수 있다. 관측된 유전자들의 발현정도를 이용하여 반응 값에 대한 예측을 하거나, 비슷한 성질을 갖는 유전자들을 그룹화하여 묶을 수 있다. 이러한 실험을 통하여 이전에 밝혀지지 않은 새로운 형태의 질병을 발견할 수도 있다. 이와 같이 microarray 실험을 통하여 다양한 실험결과를 도출할 수 있다. 따라서 microarray 실험은 실험자가 원하는 목적을 가장 잘 설명할 수 있도록 계획되어야 한다.

실험의 대상은 처리를 받는 사람, 동물 및 세포 등이 되며, 이 때 처리라 함은 실험에서 비교하고자 하는 인자의 수준이 된다. 세포의 분화과정을 연구하는 경우에는 실험을 관측하게 되는 시점이 중요한 처리가 된다. 이러한 실험에서는 염료(dye)의 효과, 슬라이드 내에서 동일 유전자를 반복하여 프린트 한 경우에 반응값의 변동 정도 및 슬라이드 간의 반복실험에 있어서 동일 유전자의 슬라이간의 변동 정도를 고려한 실험계획이 필요하다.

변동요인으로는 생물학적 변동요인, 기술적 변동요인, 측정오차를 생각해 볼 수 있으며 실험의 정도를 높여서 분석의 타당성을 제공할 수 있는 반복실험도 microarray 실험에서 고려하여한다.

5.3 Microarray 실험계획법의 종류

5.3.1 그래프 표현법

Microarray 실험은 하나의 슬라이드에 두 개의 처리가 실시되는 형태를 띠게 된다. 이 경우는 슬라이드가 블록이 되고 블록내의 처리의 수준은 2가 된다. (Kerr and Churchill, 2001b). [그림 5-3]은 이러한 특성을 쉽게 표시해줄 수 있는 표현법이다. 이러한 그래프 표현법은 디자인의 설정을 손쉽게 도와줄 수 있으며, 디자인의 성질을 쉽게 평가할 수 있다. 또한 부수적으로 디자인의 이름을 결정짓기도 한다. (Kerr and Churchill, 2001b)

[그림 5-5] 대조군(C)을 Cy5로 처리군(T)을 Cy3로 부착 후 합성한 실험

상자로 표현된 것은 처리를 나타낸다. 화살표는 한 번의 실험을 나타내는데 한 장의 슬라이드라고 생각하면 된다. 화살표의 시작점은 Cy5가 되고 화살표가 가리키는 쪽은 Cy3가 된다. [그림 5-5]는 대조그룹에 Cy5를 처리그룹에 Cy3를 처리한 실험을 표시한 것이다.

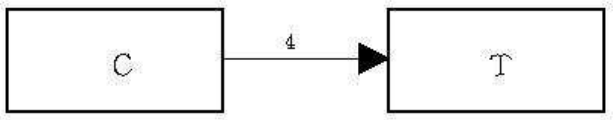
대조군(C)과 처리군(T)의 발현강도를 각각 $log(C)$와 $log(T)$로 표현하면, 이것은 다음과 같이 표현 할 수 있다.

$$log(C) = \mu + D_R + V_C + \epsilon^C \qquad (5.1)$$
$$log(T) = \mu + D_G + V_T + \epsilon^T$$

여기서, μ는 전체평균을, D_R과 D_G는 각각 Cy_5와 Cy_3의 염료 효과를, 그리고 V_C와 V_T는 대조군과 처리군의 효과를 나타낸다. 이때 ϵ^C와 ϵ^T는 서로 독립인 오차항이다.

이 때 특정유전자의 상대적인 발현정도는 $log(T/C)$와 같이 나타낼 수 있으며, 위의 식 (5.1)을 이용해서 살펴보면, $log(T/C) = V_T - V_C + D_G - D_R + \epsilon$과 같이 표현될 수 있다. 여기서, $\epsilon = \epsilon^T - \epsilon^C$이라고 나타낸다. 이 때, 각 처리군과 대조군사이의 효과의 차이와 염료의 효과차가 서로 교락되어 있는 것을 볼 수 있다.

[그림 5-6] 실험을 4회 반복한 경우

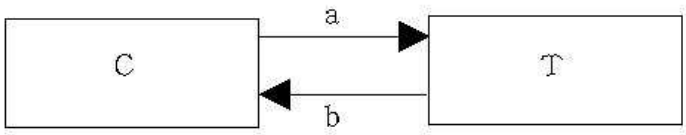


한편 동일한 조건하에서 실험을 여러 번 반복하는 경우도 생각해 볼 수 있다. [그림 5-6]을 보면, 화살표 위에 숫자 4가 표시되어 있다. 이는 반복회수를 나타내며, 동일한 실험 조건하에서 실험이 4번 반복되었다는 것을 표현한다. 따라서 [그림 5-5]에서 설명한 실험을 4번 반복한 것이며, 슬라이드는 4장이 된다. 이 실험에서의 상대적인 발현정도는

$$log(T/C) = [log(T^1/C^1) + log(T^2/C^2) + log(T^3/C^3) + log(T^4/C^4)]/4 \quad (5.2)$$

와 같이 측정될 수 있으며, [그림 5-5]의 실험보다는 효과의 차이를 보다 더 정밀하게 추정하리라 기대할 수 있다. 그러나 [그림 5-6]에서의 실험도 [그림 5-5]의 실험처럼 처리효과와 염료효과가 교락되어 있다. 여기서, 위 첨자는 실험의 순서를 나타낸다.

[그림 5-7] dye swap 디자인의 예



이러한 문제점들을 해결하기 위하여 [그림 5-7]과 같은 실험을 생각해 볼 수 있다. 이와 같이 화살이 반대 방향으로 가리키는 실험을 dye swap 디자인이라고 부르며, 실험할 때 염료를 바꾸어서 혼성화시킨다. 화살표 위의 알파벳은 실험의 이름을 나타내기 위하여 사용하였다. a 슬라이드는 [그림 5-5]의 실험처럼 대조군에 Cy5를 처리군에 Cy3를 처리한 실험이며, b 슬라이드는 반대로 대조군에 Cy3을 처리군에 Cy5를 실험하였다. 실험은 각 유전자의 발현 정도에서 염료의 효과를 제거할 수 있게 되리라 기대할 수 있다.

[그림 5-7]의 실험에서는 발현정도를 슬라이드에 따라서 각각 $log(T^a/C^a)$와 $log(T^b/C^b)$처럼 측정하였다고 생각했을 때, $log(T/C) = (log(T^a/C^a) - log(T^b/C^b))/2$로 사용하게 되면, 처리효과와 교락되어 있는 염료의 효과를 분리해 낼 수 있다.

[그림 5-6]과 [그림 5-7]의 실험을 할 때 주의할 점은, 앞에서도 설명하였듯이 반복실험이 되기 때문에 실험에 사용하는 표본을 생물학적으로 동일한 것을 사용할 것인지 여부를 결정해야한다. Kerr et al.(2002)은 [그림 5-7]과 같은 dye swap 디자인을 세 번 반복하여 분산분석을 실시하였다.

5.3.2 기준계획(reference design)

[그림 5-8]에서 표현된 실험과 같이 하나의 기준을 이용하여 여러 가지 처리간의 실험을 비교하는 실험계획을 기준계획(reference design)이라고 한다. 여러 가지 처리군을 비교해야하는 경우나, 기준시점에서 변화해가는 과정을 살펴보고자 할 때 생각해 볼 수 있는 방법이다. 위의 방법은 해석이 간편하고 실험과 분석 방법이 쉬우며, 새로운 표본이나 처리를 추가하여 기존의 실험들과 비교할 수 있기 때문에 실험의 확장성

측면에서 사용이 용이하다. 다만, 기준계획은 한 슬라이드로부터 기준채널 이외의 실제 표본에 대해서도 한 채널값만 보기 때문에 실험회수가 많아져 비효율적이다.

실험을 통해서 발현정도는 $log(T1/R)$, $log(T2/R)$, $log(T3/R)$ 및 $log(T4/R)$로 측정될 수 있다. 여기서 우리가 처리 1(T1)과 처리 2(T2)의 효과의 차이를 알고 싶으면, $log(T2/R) = log(T2/R) - log(T1/R)$을 계산하여 분석하면 된다.

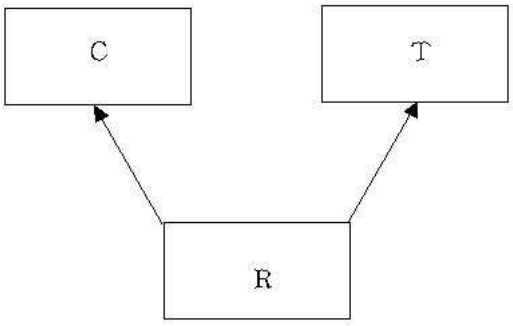
$$\begin{aligned} \log(T2/T1) &= D_G + V_{T2} - D_R - V_R - (D_G + V_{T1} - D_R - V_R) + \epsilon \\ &= V_{T2} - V_{T1} + \epsilon \end{aligned} \quad (5.3)$$

와 같이 표현된다. 그러나 실험(슬라이드)마다 표본을 어디서 추출했느냐에 따라서 꼭 식(5.3)과 같이 표현되는 것은 아니다.

[예제 5.1] dye swap 계획과 기준계획

처리군과 비교군의 효과를 비교하고자 어떤 실험을 디자인하려 한다. 실험의 제약조건으로 인하여 총 실험회수는 2회로 한정한다. 이 문제에서는 [그림 5-7]의 실험과 [그림 5-8]의 실험을 생각해 볼 수 있다.

[그림 5-8] 기준계획의 예



[그림 5-8]과 같은 실험에서 기준으로 사용되는 표본은 생물학적으로 동일한 재료를 사용하여 대조군과 처리군에 대하여 microarray 실험을 한 것이다. 만약 실험에서 기

준으로 사용된 표본이 서로 다른 재료일 수 있다. 이런 경우에는 각 슬라이드마다 서로 구분해 줄 필요가 있으며, Yang and Speed (2002b)는 이 경우에 R과 R'과 같이 표현하였으며, 이 경우에는 처리군과 대조군의 효과의 차이 이외에도 기준으로 사용한 표본의 차이도 포함된다. 따라서 기준디자인을 사용할 때는 기술적인 반복이 되도록 실험을 계획하는 것이 바림직하다. 만약 서로 다른 생물학적 재료에서 추출하였다면 풀링(pooling)을 이용하면 된다.

[그림 5-7]과 [그림 5-8]의 두 실험계획을 비교하기 위하여, Yang and Speed (2002a)는 모든 로그비의 분산이 σ^2으로 동일하며, 서로 다른 슬라이드 사이는 독립이라고 가정하였다.

[그림 5-7]의 dye swap 디자인은 $log(T/C) = (log(T^a/C^a) - log(C^b/T^b))/2$이므로, $Var(log(T/C)) = \sigma^2/2$이 된다. [그림 5-8]은 $log(T/C) = (log(T/R) - log(C/R)$이므로 $Var(log(T/C) = 2\sigma^2$이 된다. 따라서 이 경우에는 dye swap 디자인이 선호된다.

[그림 5-5]의 실험처럼 처리효과를 하나의 슬라이드에서 비교하기 위해 설계하는 방법을 직접 설계법(direct design)이라고 하며, 공통인 기준에서의 차이를 이용하여 처리효과를 다른 슬라이드를 사용하며 비교하는 방법을 간접 설계법(indirect design)이라고 한다. (Yang and Speed, 2002b)

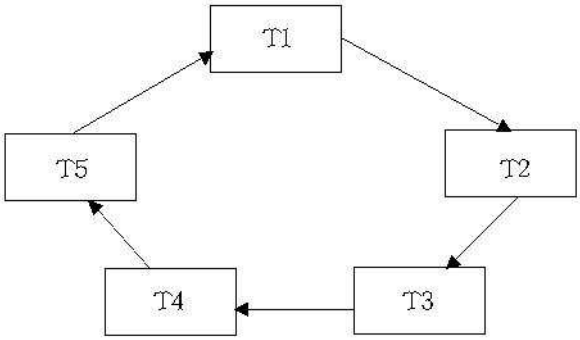
5.3.3 고리계획(loop design)

각기 서로 다른 표본들과 염료를 바꿔가면서 실험하는 방법을 고리계획(loop design)이라고 한다. 비교하고자 하는 대상이 적은 경우에 효율적인 방법이다. 다만, 하나라도 결측치가 생기는 경우에 다른 실험에서 얻은 결과들만 가지고는 분석이 어려워지게 된다. 이때 발현정도의 측정은 $log(T2/T1)$, $log(T3/T2)$, $log(T4/T3)$ 및 $log(T5/T4)$의 형태로 관측된다. 여기서 처리 1(T1)과 처리 2(T2)의 효과의 차이를 알고 싶은 경우에는 $log(T2/T1)$을 분석에 사용하면 된다.

그러나 $log(T2/T1) = -(log(T3/T2) + log(T4/T3) + log(T5/T4) + log(T1/T5))$을 이용하

여 간접적으로 계산될 수 있으며 이를 분석에 고려하여야 한다.

[그림 5-9] 처리군이 5개인 고리계획



Ⅵ 요인 *Microarrays*

*6.1 2*수준 요인 *Microarrays*

6.1.1 2^2 실험에서의 설계

Glonek와 Solomon(2002)은 microarray 실험에서 요인 배치법을 고려하였다. 2^2형 요인배치법의 실험에 대해서 보면 이 실험은 인자가 두 개이고 각 인자의 수준수는 2가 된다. 인자를 A, B로 나타내고 A의 수준을 0, 1로 B의 수준을 0과 1로 나타낸다. 즉 두 개 수준의 색깔을 0($Cy5$:빨강)과 1($Cy3$:녹색)로 된 코드로 표시할 것이다. 데이터의 모델은 아래와 같이 쓰인다.

$$(y_{i1} - y_{i2}) = (\delta_1 - \delta_2) + (\tau_{t_{i1}} - \tau_{t_{i2}}) + \epsilon_{i1i2}$$

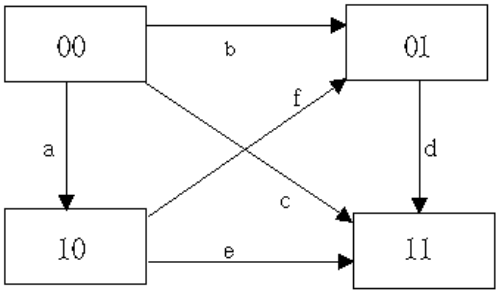
i번째 블록으로부터의 블록내 비교 $y_{i1} - y_{i2}$는 i번째 배열에서 교배시킨 두 처리조합 t_{i1}과 t_{i2}의 $\log_2$ 표현비와 일치한다. $\delta_1 - \delta_2$는 염료 효과, τ_j는 j번째 처리조합의 효과, 그리고 ϵ_{i1i2}은 평균이 0, 분산이 σ^2을 갖는 독립적으로 분포된 확률오차이다. $\log_2$ 표현비는 표준화와 원인수정방법에 종속된다고 추정된다. 블록효과는 블록내 비교 $y_{i1} - y_{i2}$로 배출되어진다. 색깔 효과 $\delta_1 - \delta_2$는 처리들이 색깔에 직교한다면 주효과와 교호작용의 요인 추정으로부터 자동적으로 배출되어진다. 다시 말하면, 색깔의 두 수준이 각각의 처리조합과 함께 동등하게 나타난다면, 위의 모델은 각 유전자가 개별적으로 적용할 수 있다.

[표 6-1] 2^2형 요인배치법에서 유전자의 표현법

실험의 조건	로그 발현정도
00 (c)	$E(log(C)) = \mu$
01 (b)	$E(log(A)) = \mu + \alpha$
10 (a)	$E(log(B)) = \mu + \beta$
11 (ab)	$E(log(AB)) = \mu + \alpha + \beta + (\alpha\beta)$

[표 6-1]은 이 실험에서 발현강도의 표현식으로 나타내었다. μ는 대조집단인 00(C)의 모수를 나타낸다. α는 A의 주효과라 불리며, 실험조건 10(a)과 00(c)사이의 차이를 나타낸다. β는 B의 주효과라 하며, $(\alpha\beta)$는 인자 A와 B의 교호작용이라 불린다. 식으로는 $(\alpha\beta) = (11-01)-(10-00)$으로 표현될 수 있다.

[그림 6-1] 2^2형 요인배치법에서 가능한 실험조합



[그림 6-1]은 2^2형 요인배치법에서 생각할 수 있는 실험을 그림으로 표현한 것이다. 이 실험의 내용이 [표 6-2]에 정리되어 있다.

[표 6-2] 2^2형 요인배치 실험디자인

실험번호	Cy5	Cy3	로그비
a	00	10	α
b	00	01	β
c	00	11	$\alpha+\beta+(\alpha\beta)$
d	01	11	$\alpha+(\alpha\beta)$
e	10	11	$\beta+(\alpha\beta)$
f	10	01	$\beta-\alpha$

각 반복에서 교락된 효과에 따라 교락의 방법을 사용하여 얻어진 단일 반복 모형은 아래의 [표 6-3]에 주어졌다.

[표 6-3] 2^2요인 모델

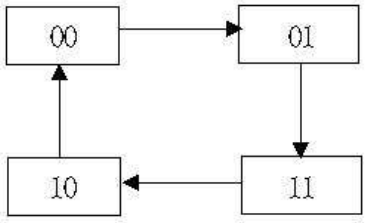
모델	블록	교락된 효과
D_{11}	[00, 01],[10, 11]	A
D_{12}	[00, 10][01, 11]	B
D_{13}	[00, 11],[01, 10]	AB

microarray 모델은 행요인 색깔과 열요인 블록(슬라이드)을 갖는 행렬 모델이다. 수준 0(빨강)과 1(녹색)은 첫 번째와 두 번째 열에 각각 해당한다. D_{11}과 D_{12}을 사용함으로써 우리는 열에 4개의 슬라이드가 해당하는 뒤따라오는 microarray 모델을 얻는다.

00	11	10	01
01	10	00	11

첫 번째 행에서의 모든 처리조합은 cy5로 분류하고 두 번째 행은 cy3로 분류한다.

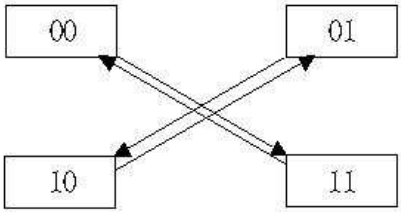
예를 들면, 첫 번째 열 [00, 01]은 각각 빨강과 녹색색깔로 구분된 처리조합 00과 01로 교배된 하나의 슬라이드를 나타낸다. 열에서의 처리조합은 행들의 처리조합에 직교로 배열되어졌다. 오직 사용된 슬라이드의 수는 처리조합수가 복합적일 때 행과 처리의 직교는 얻어졌다. 위의 microarray 모델은 50%의 효율을 가진 주효과 A과 B을 추정한다. 그에 반해, 교호작용 AB는 D_{11}과 D_{12}에서 동시에 교락되어지지 않았기 때문에 완전한 효율을 가지고 추정한다. 사실 이 모델은 Kerr and Churchill(2001a)에 의해서 주어진 common loop(CL)모델이다.



슬라이드 수	A	B	AB
4	0.50	0.50	1

뒤따라오는 모델은 D_{13}의 두 번의 반복을 취함으로써 Landgrebe et al.(2005)에 의해서 언급된 cross-swap(CS)모델이다.

00	01	11	10
11	10	00	01



Landgrebe et al.(2005)는 16개의 슬라이드에서 두개의 모델을 고려했다. 한 모델은 CL의 4개의 반복을 사용하는 것과 또 하나는 각각 CL과 CS를 두 번씩 반복해 사용하는 것이다. 사실, D_{11}, D_{12}, D_{13}의 적절한 결합은 A, B 그리고 AB의 추정의 원하는 효율과 함께 추가적인 모형을 얻을 수 있다.

예를 들면, D_{11}의 3번 반복, D_{12}의 2번 반복 그리고 D_{13}의 3번 반복을 취함으로써 16개의 슬라이드를 갖는 뒤따라오는 microarray 모델을 낳는다.

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

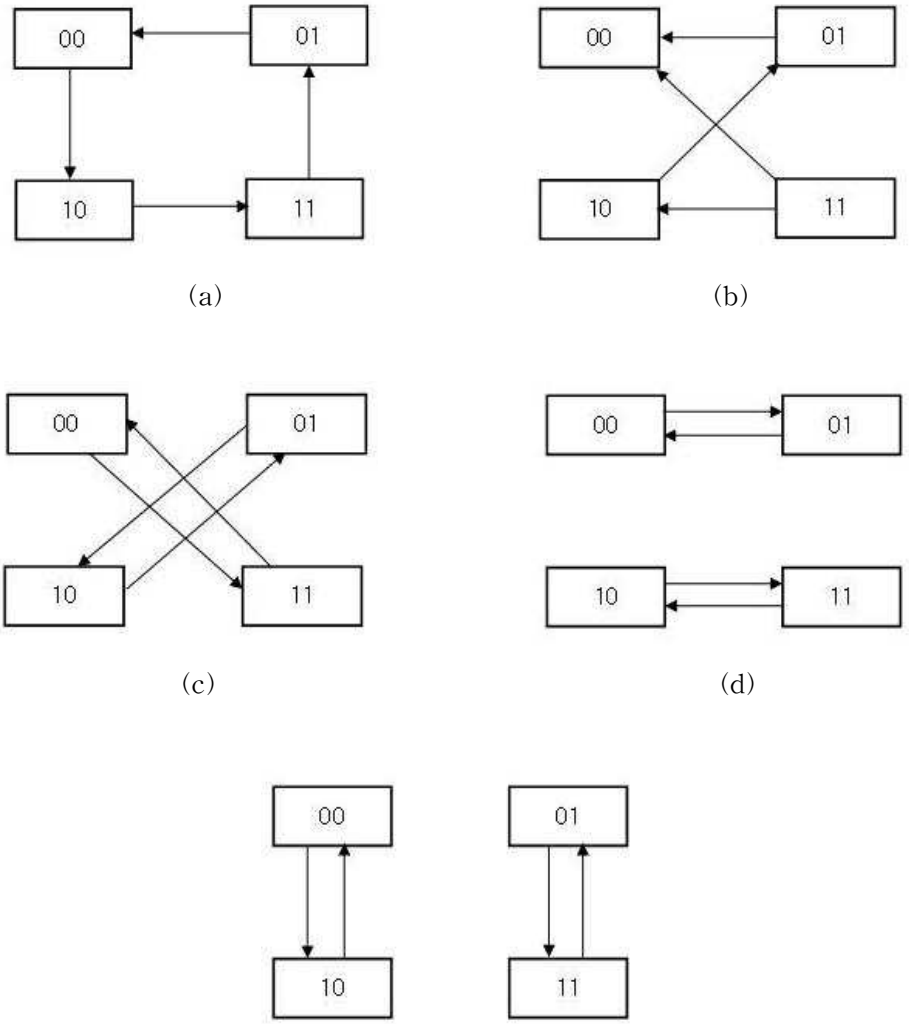
이 모델은 각각 62.5%, 75% 그리고 62.5% 효율을 가진 A, B 그리고 AB을 추정한다. 모델의 선택은 바라는 효율 또는 대상의 대조 추정의 정확에 의존할 것이다.

D_{11}, D_{12}, D_{13}을 반복하여 효율성을 추정함을 알 수 있다. 유전자의 발현강도를 알기 위해서 실시하는 단일 실험은 많은 변동요인을 포함하고 있다. 유의한 유전자를 찾기 위한 여러 가지 검정법이 개발되었지만, 한 번의 실험으로는 불충분하다. 따라서 신뢰할 수 있는 결과를 얻기 위해서는 반복 실험이 필요하다. Lee et al.(2000)은 3번의 반복실험을 하면 충분하다고 보았지만, microarray 실험에서 얼마나 많은 반복 실험을 해야 하는지 결정하는 것은 무척 어려운 일이다. 또한, 한 슬라이드에서 동일한 유전자를 몇 번 반복하여 스팟팅 할 것인가를 결정하는 것도 실험자의 관심사가 될 수 있다.

슬라이드에서 기술적인 반복은 동시에 추출된 mRNA에 대한 반복으로 언급된다. 그

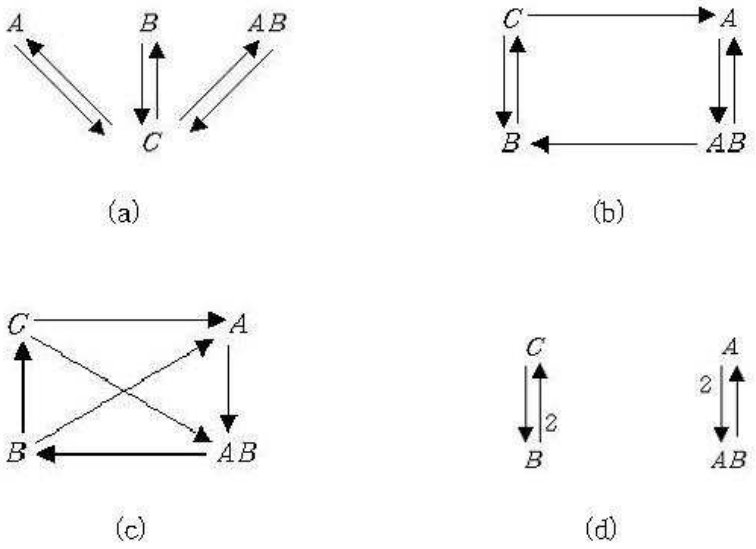
것으로부터 특성, 추출의 반복적인 특징을 얻을 수 있다. 즉, 기술적인 반복을 통해 생물학적인 반복보다 분산의 자유도가 더 작게 나타난다. 만약에 위의 실험에서 여러 가지 제약조건 때문에 실험의 횟수를 4회로 고정하였을 때는 다음과 같은 실험들이 제시된다.

[그림 6-2] 실험횟수가 4일 때의 실험



위의 모형을 이용하여 주효과 및 교호작용 효과의 효율성을 추정할 수 있다. 본 논문에서는 6개의 슬라이드에 배치하는 design에 대한 효율성을 제시하였다.

[그림 6-3] 6개의 슬라이드에 배치된 2^2 요인 실험



[표 6-4] 2^2 요인 실험 설계

Design choices	Main effect A	Main effect B	Interaction A,B
	Indirect design		
design Ⅰ (a)	0.50	0.50	1.50
	A balance of direct and indirect design		
design Ⅱ (b)	0.67	0.43	0.67
design Ⅲ (c)	0.50	0.50	1.00
design Ⅳ (d)	N/A	0.30	0.67

또한 이 논문에서는 [그림 6-3]에 제시된 것처럼 6개의 슬라이드에 포함되는 여러 디자인에 대하여서 나타내었다. 처리할 cell line 2요인인 경우를 생각해 보자. C(00)을 control cells이라고 하고, A(10)과 B(01)을 첫 번째와 두 번째 방법으로 분리된 처리

된 cell로부터 파생된 mRNA라 하고 AB(11)은 동시에 두 요인을 처리한 cells이라 정의하자. 디자인 II와 IV는 $\log(AB \times C / AB)$로 정의되는 즉 00에 대비한 11의 상호작용의 분산이 가장 작다. 하지만 디자인 IV의 경우는 A(10)에 결과가 나타나지 않았다. 또한 디자인 I은 상호작용의 추정치가 가장 안 좋다. 따라서 균형요인설계를 이용한 디자인의 분산이 더 작음을 알 수 있다. 이에 본 논문에서는 균형요인설계를 이용한 microarray design을 제공하고자 한다. 일반적으로 디자인 II는 모든 조합을 고려할 때 좋은 예측을 제공하는 적절한 디자인이다.

6.1.2 3×2 실험에서의 설계

각 수준이 다른 요인설계에 대해서는 3×2실험을 대해서는 언급하기로 한다. 이에 대한 요인 모델은 Lewis and Tuck(1985)에 의해서 작성되어졌다. 일반화된 순환 방법의 구성에 기본을 둔 이러한 모델들은 [표 6-5]에 표현됐다.

[표 6-5] 3×2요인을 위한 모델

모델	블록	추정의 효율		
		A	B	AB
D_{41}	[00,01],[10,11],[20,21]	0	1	1
D_{42}	[00,11],[10,21],[20,01],[01,10],[11,20],[21,00]	0.75	1	0.25
D_{43}	[00,10],[10,20],[20,00],[01,11],[11,21],[21,01]	0.75	0	0.75

일반화된 순환의 방법은 다음과 같다.

10씩 합하여 처음으로 순환될 때까지 계속 실시한다. 10 10 11씩 순서대로 합을 순환될 때까지 계속 시행하는데 10을 20, 11을 21로 하여도 상관없다.

[예 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) 순환됨

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

[예 2] D_{42} (00, 11) ⊕ 10(mod(3, 2)) = (10, 21) ⊕ 10 mod(3, 2)
= (20, 01) ⊕ 11 mod(3, 2) = (01, 10) ⊕ 10 mod(3, 2)
= (11, 20) ⊕ 10 mod(3, 2) = (21, 00) ⊕ 11 mod(3, 2)
= (00, 11) : 순환
main block

(2) 다른 main design에 대하여 살펴보면 다음과 같다.

[예 3] D_{44} (00, 20) (10,00) (20,10)
[처음부터 ⊕ 10 mod(3, 2) ,⊕ 10 mod(3, 2),⊕ 11 mod(3, 2) 순환]
(01, 21) (11,01) (21,11) (00,20) [순환]
[처음부터 ⊕ 10 mod(3, 2), ⊕ 10 mod(3, 2), ⊕ 11 mod(3, 2)]

[예 4] D_{45} (00, 21) (10, 01) (20, 11) (01, 20) (11, 00) (21, 10) (00,21) [순환]
[처음부터 ⊕10, ⊕10, ⊕11, ⊕10, ⊕10, ⊕11 순환]
여기서 $D_{44} \cong D_{43}$ $D_{45} \cong D_{42}$이다.

따라서 3 block으로 된 BS design은 5개가 아니고 3개이며 6 block으로 된 XL, AL design은 5개가 아니고 3개임을 알 수 있다.

D_{42}는 6개의 슬라이드를 가진 모델이다.

00	10	20	01	11	21
11	21	01	10	20	00

Landgrebe et al.(2005)은 D_{41}는 BS모델과 같이 그리고 D_{42}, D_{43}는 각각 XL, AL과 같은 모델의 두 번의 반복을 언급했다. 게다가 모델들은 D_{41}, D_{42}, D_{43}의 적절한 결합

에 의해서 얻어질 수 있다.

6.1.3 2^3실험에서의 설계

교락의 방법을 사용하여 얻어진 2^3실험을 위한 모든 가능한 단일 반복 모델은 부록에 제시한 프로그램을 실행하여 얻어졌으며 그 결과는 다음 [표 6-6]에 나타내었다.

[표 6-6] 2^3요인을 위한 모델

모델	블록	교락된 효과
D_{21}	[000,100],[001,101],[010,110],[011,111]	B, C, BC
D_{22}	[000,010],[001,011],[100,110],[101,111]	A, C, AC
D_{23}	[000,001],[010,011],[100,101],[110,111]	A, B, AB
D_{24}	[000,111],[001,110],[010,100],[011,100]	AB, AC, BC
D_{25}	[000,011],[001,010],[101,110],[100,111]	A, BC, ABC
D_{26}	[000,101],[001,100],[011,110],[010,111]	B, AC, ABC
D_{27}	[000,110],[010,100],[011,101],[001,111]	C, AB, ABC

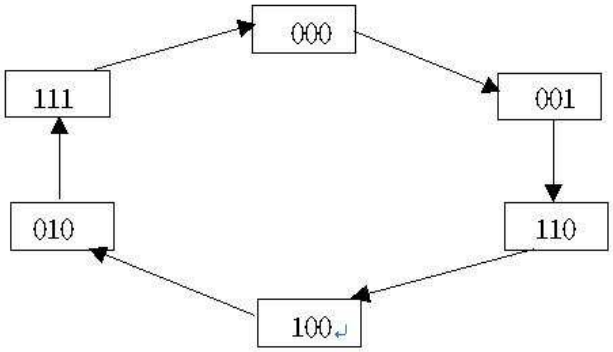
2^3실험 역시 반복모형 모델의 표현을 다른 방법으로 접근해보면 main 블록을 000과 나머지 7개중에서 1개를 선택하고 나머지 블록은 main 블록의 요인합을 $\mod 2$로 하여 main 블록과 동치가 되도록 나머지 6개를 2개씩 조합한다. 예를 들어 D_{23}의 main 블록은 [000, 001]이고 나머지 블록은 (000 ⊕ 001)(mod 2) = 001 되게 2개씩 조합하여 [010, 011] [100, 101] [110, 111] 3개의 블록을 얻을 수 있다. 이것을 [표 6-3]에서 예를 들었던 것처럼 곱하게 되면 [표 6-6]과 같은 결과를 얻을 수 있다. 예를 들어 D_{21}, D_{24}, D_{27}에 대한 교락여부를 살펴보면 다음과 같다.

D_{21} : (000⊕100) = (---⊕+--) = (- + +) B, C, BC

D_{24} : (000⊕111) = (---⊕+++) = (- - -) AB, AC, BC

D_{27} : (000⊕110) = (---⊕++-) = (- - +) C, AB, ABC

위의 모델에서 D_{21}을 이용하여 B, C, BC에 대한 효과를 추정할 수가 없다. 따라서 이들의 효과를 추정하기 위해선 다른 모델을 함께 고려하여야 한다. 만약 위의 모델을 이용하지 않고 단순히 Loop 모형을 고려하여 설계하면 다음과 같다.



위의 그래프 표현법을 슬라이드에 배치하면 다음과 같이 7개의 슬라이드에 배치된다.

000	001	110	100	010	111
001	110	100	010	111	000

효과를 추정하면 AB, AC, BC의 분산이 동일하게 나타나지 않는다. 그래서 [표 6-6]에서 D_{24}, D_{25}, D_{26}그리고 D_{27}을 사용하였다. 16개의 슬라이드를 가진 모델을 고려하면 그 결과는 다음과 같다.

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

슬라이드 수	A	B	C	AB	AC	BC
16	0.75	0.75	0.75	0.5	0.5	0.5

세 개의 주효과들은 75% 효율성으로 추정되어졌고, 세 개의 2인자 교호작용들은 50% 효율성으로 추정되어졌다.

따라서 적절한 design의 선택을 필요함을 알 수 있다. 이처럼 [표 6-6]에서 제시한 design을 이용하여 적절한 모델을 택하면 주효과 뿐만 아니라 교호작용의 효과도 추정할 수 있음을 알 수 있다.

이로서 2수준 실험에 있어서 교락법을 이용하면 교락여부에 따른 모델을 제시함으로서 원하는 효과 추정이 가능하다. 또한 최고차항을 무시하지 않고 교호작용의 효과 추정도 가능함을 알 수 있다.

Ⅶ 결론 및 토론

cDNA microarray 실험의 많은 디자인들의 목적은 같은 슬라이드에서 함께 혼성화시킬 샘플을 결정하는 것이다. 본 논문에서는 microarray 디자인 블록 구성을 위하여 이용 가능한 균형요인실험설계에 대하여 논하였다. 특히, 2수준 요인에 대한 고전적인 교락 디자인들이 microarray 디자인으로는 효율적이라는 사실을 쉽게 알 수 있다.

Lewis와 Tuck(1985) 그리고 Gupta(1987)에 의해 블록크기 2인 블록 모델들이 효과적이라는 것을 입증하였다. 이에 요인의 수와 수준이 많아지면 아주 복잡해지므로 단일 실험에 의한 블록을 만드는 방법과 고전적인 교락법에 의한 단일실험 요인모델에 의한 계획과 주로 고전적인 디자인과 연결하여 그의 관계를 나타내어 수준과 요인이 동시에 많거나 수준과 요인이 독립적으로 많을 때를 고려하여 균형 요인실험설계를 사용하였다. 향후 본 디자인에 의한 실험분석은 4개의 카테고리(classical methods, tree methods, machine learning method, 그리고 최근 일반화된 LDA(Linear discriminant analysis) 및 PAM(Predictive Analysis of Microarrays) method를 연구하여 비교분석하는데 역점을 두어야 할 것이다. 또한 프로그램을 보완하여 많은 수의 요인과 수준에 대한 교락 여부 및 계획의 효율성을 제시하는데 역점을 두어야 할 것이다.

참고 문헌

[1] Abraham, B., H. Chipman, and K. Vijayan. (1999). "Some Risks in the Construction and Analysis of Supersaturated Designs." Technometrics, Vol. 41, pp. 135-141.

[2] Barnett, V., and T. Lewis (1994). *Outliers in Statistical Data, 3rd edition. Wiley*, New York.

[3] Bingham, D., and R. R. Sitter. (1999). "Minimum Aberration Two-Level Fractional Factorial Split-Plot Designs." Technometrics, vol. 41, pp. 62-70.

[4] Black, M. A. and Doerge, R. W. (2002). "Calculation of the minimum of repli -cate spots required for detection of significant gene expression fold change in microarray experiments." Bioinformatics, vol. 18, pp. 1609-1616.

[5] Box, G.E.P. and Wilson, K.G. (1951). On the Experimental Attainment of Optimum Conditions, Journal of Royal Statistical Society, B, 13, 1-45

[6] Box, G.E.P. (1992-1993). "Sequential Experimentation and Sequential Assembl -y of Designs." Quality Engineering, Vol. 5, No. 2, pp. 321-330.

[7] Box, G.E.P. (1999). "Statistics as a Catalyst to Learning by Scientific Metho -d Part Ⅱ－A Discussion" (with Discussion), Journal of Quality Technology, Vol. 31, pp. 16-29.

[8] Burdick, R. K., C. M. Borror, and D. C. Montgomery. (2003). "A Review of Methods for Measurement Systems Capability Analysis." Journal of Quality

Technology, Vol. 35, No. 4, pp. 342-354.

[9] Burdick, R. K., and F. A. Graybill (1992). Confidence Intervals on Variance Components. *Dekker*, New York.

[10] Coleman, D. E., and D. C. Montgomery (1993). "A Systematic Approach to Planning for a Designed Industrial Experiment" (with Discussion). Technom -etrics, Vol. 35, pp. 1-27.

[11] Cornell, J. A. (1990). Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data. *2nd edition. Wiley*, New York.

[12] Chen, J. and Wu, C.F.J. (1991). Some Results on s^{n-k} Fractional Factorial Designs with Minimum Aberration or Optimal Moments, Annals of Statistics, Vol. 19, pp. 1028-1041

[13] Cornfield, J. and Tukey, J.W. (1956). Average Values of Mean Squares in Factorials, Annals of Mathematical Statistics, Vol. 27, pp. 907-949.

[14] Daniel, C. and Wilcoxon, F. (1966). Factorial 2^{p-q} Plans Robust Against Lin -ear and Quadratic Trends, Technometrics, Vol. 8, pp. 259-278.

[15] Fries, A. and Hunter, W.G. (1980). Minimum Aberration 2^{k-p} Designs, Technometics, Vol. 22, pp. 601-608.

[16] Glonek, G. F. V. and solomon, P. J. (2002). Facrorial and time course design -s for cDNA microarray experiments. *Technical Report, Department of Appl -ied Mathematics, University of Adelaide*, 10/2002.

[17] Hamada, M., and N. Balakrishnan (1998). "Analyzing Unreplicated Factorial Experiments: A Review with Some New Proposals" (with Discussion). Statis -tica Sinisa, Vol. 8, pp. 1-41.

[18] Hamada, M., and C. F. J. Wu (1992). "Analysis of Designed Experiments with Complex Aliasing." Journal of Quality Technology, Vol. 24, No. 3, pp. 130-137.

[19] Hines, W. W., D. C. Montgomery, D. M. Goldsman, and C. M. Borror.(2003). Probability and Statistics in Engineering. *4th edition. Wiley*, New York.

[20] Holcomb, D. R., D. C. Montgomery, and W. M. Carlyle. (2003). "Analysis of Supersaturated Designs." Journal of Quality Technology, Vol. 35, No. 1, pp. 13-27.

[21] Huang, P., D. Chen, and J. Voelkel. (1999), "Minimum Aberration Two-level Split-Plot Designs." Technometrics, Vol. 41, pp. 314-326.

[22] Jones, B. (1980). Algorithm AS 156: Combining two component designs form a row-and-column design. Applied Statistics, Vol. 29, pp. 334-337.

[23] Khuri, A. I., and J. A. Cornell (1996). Response Surfaces: Designs and Analyses. *2nd edition. Dekker*, New York.

[24] Kshirsagar, A.M. (1966). Balanced factorial designs. Journal of the Royal Statistical Society, B28, pp. 59-67

[25] Montgomery, D.C. (1976). Design and Analysis of Experiments, *John Wiley & Sons*, New York.

[26] Larntz, K., and P. Whitcomb (1998). "Use of Replication in Almost Unreplicated Factorials." *Presented at the Fall Technical Conference, Cornin -g*, New York.

[27] Li, W. W., and C. F. J. Wu. (1997). "Columnwise-Pairwise Algorithms with Applications to the Construction of Supersaturated Designs." Technometrics, Vol. 39, pp. 171-179.

[28] Lin, D. K. J. (1993). "A New Class of Supersaturated Designs." Technometri -cs, Vol. 35, pp. 28-31.

[29] Lin, D. K. J. (1995). "Generating Systematic Supersaturated Designs." Technometrics, Vol. 37, pp. 213-223.

[30] Lin, D. K. J. (2000). "Recent Developments in Supersaturated Designs." Chapter 18 in Statistical Process Monitoring and Optimization," *S. H. Park and G. G. Vining, eds. Marcel Dekker*, New York, pp. 305-319.

[31] Loughin, T. M. (1998). "Calibration of the Length Test for Unreplicated Factorial Designs." Journal of Quality Technology, Vol. 30, pp. 171-175.

[32] Loughin, T. M., and W. Noble (1997). "A Permutation Test for Effects in an Unre plicated Factorial Design." Technometrics, Vol. 39, pp. 180-190.

[33] Mee, R. W.(2004). "Efficient Two-Level Designs for Estimating All Main Effects and Two-Factor Interactions." Journal of Quality Technology, Vol. 36,pp. 400-412.

[34] Mee, R. W., and Peralta, M. (2000). "Semifolding 2^{k-p} Designs."

Technometrics, Vol. 42, No. 2, pp.122-143

[35] Miller, R. G.(1991). Simultaneous Statistical Inference. *Springer-Verlag*, New York.

[36] Montgomery, D. C.(2001). Introduction to Statistical Quality Control. *4th edition. Wiley*, New York.

[37] Montgomery, D. C.(1999). "Experimental Design for Product and Process Design and Development." Journal of the Royal Statistical Society, D, Vol. 48, pp. 159-177

[38] Montgomery, D. C., C. M. Borror, and J. D. Stanley (1997-1998). "Some Cautions in the Use of Plackett-Burman Designs." Quality Engineering, Vol. 10, pp. 371-381

[39] Montgomery, D. C., E.A. Peck, and G. G. Vining(2001). Introduction to Line -ar Regression Analysis. *3rd edition. Wiley*, New York.

[40] Montgomery, D. C., and G. C. Runger (1996). "Foldovers of 2^{k-p} Resolution IV Experimental Designs." Journal of Quality Technology, Vol. 28, pp. 446-450.

[41] Montgomery, D. C., and G. C. Runger (2003). Applied Statistics and Probab -ility for Engineers. *3rd edition. Wiley*, New York.

[42] Myers, R. H. (1990). Classical and Modern Regression with Applications. *2nd edition. PNS-Kent*, Boston.

[43] Myers, R. H., and D. C. Montgomery (2002). Response Surface Methodology:

Process and Product Optimization Using Designed Experiments. *2nd edition. Wiley*, New York.

[44] Myers, R. H., and D. C. Montgomery (1997). "A Tutorial on Generalized lin -ear Models." Journal of Quality technology, Vol. 29, pp. 274-291.

[45] Myers, R. H., and D. C. Montgomery, and G. G. Vining. (2002). Generalized Linear Models with Applications in Engineering and the Sciences. *John Wile -y & Sons*, New York.

[46] Myers, R. H., and D. C. Montgomery, and G. G. Vining, C. M. Borror, and S.M.Kowalski. (2004). "Response Surface Methodology: A Retrospective and Literature Survey." Journal of Quality Technology, Vol. 36, pp. 53-77.

[47] Nair, V. N., et al., (eds.) (1992). "Taguchi's Parameter Design: A Panel Discussion." Technometrics, Vol. 34, pp. 127-161.

[48] Nelson, L. S. (1995a). "Using Nested Designs Ⅰ: Estimation of Stansard Deviations." Journal of Quality Technology, Vol. 27, No. 2, pp. 169-171.

[49] Nelson, L. S. (1995b). "Using Nested Designs Ⅱ: Confidence Limits for Standard Deviations." Journal of Quality Technology, Vol. 27, No. 3, pp. 265-267.

[50] Raghavarao, D. (1971). Construcions and Combinatorial Problems in Design of Experiments. *Wiley,* New York.

[51] Searle, S. R., G. Casella, and G. E. McCulloch (1992). Variance components. *Wiley*, New York.

[52] Two-Level Factorial Experiments." Quality Engineering, Vol. 11, pp. 267-271.

[53] Taguchi, G. (1991). Introduction to Quality Engineering. *Asian Productivity Organization, UNIPUB, White Plains*, New York.

[54] Ting, N., R. K. Burdick, F. A. Graybill, S. Jeyaratnam, and T.-F. C. Lu (1990). "Confidence Intervals on Linear Combinations of Variance Components That Are Unrestricted in Sign." Journal of Statistical Computation and Simulatio -n, Vol. 35, pp. 135-143.

[55] Wu, C. F. J.(1993). "Construction of Supersaturated Designs Through Partia -lly Aliased Interactions." Biometrika, Vol. 80, pp. 661-669.

[56] Ye, K., and M. Hamada (2000). "Critical Values of the Lenth Method for Unreplicated Factorial Designs." Journal of Quality Technology, Vol. 32, pp. 57-66.

[57] Yang, Y. H., and Speed, T. P. (2002a). "Designs issues for cDNA microarray experiments" Nature Reviews Genetics, Vol. 3, pp.579-588.

[58] Yang, Y. H., and Speed, T. P. (2002b). "Direct and indirect hybridizations for cDNA microarray experiments" Sankhya Sereis A, Vol. 643, pp.707-721

[59] Yang, Y. H., and Speed, T. P. (2002b). "Direct and indirect hybridizations for cDNA microarray experiments" Sankhya Sereis A, Vol. 643, pp.707-721

[60] 박태성 외 5인 (2005). microarray 자료의 통계적 분석. *자유아카데미*, 서울.

[61] 박성현 (1984). 현대실험계획법. *민영사*, 서울.

[61] 박홍선 (2001). SAS를 이용한 실험계획과 분산분석. *자유아카데미*, 서울.

[62] 이우선 (1998). 최신실험설계. *영풍문고*, 서울.

[63] 박동권 (1995). 실험계획법. *자유아카데미*, 서울.

부록

```
 f=function(i)
{
a<-rep(c(0,1),c(2^(i-1),2^(i-1)))
b<-rep(c(0,1),c(2^(i-2),2^(i-2)))
bb<-rep(b,2^(i-(i-1)))
c<-rep(c(0,1),c(2^(i-3),2^(i-3)))
cc<-rep(c,2^(i-(i-2)))
d<-rep(c(0,1),c(2^(i-4),2^(i-4)))
dd<-rep(d,2^(i-(i-3)))
e<-rep(c(0,1),2^(i-1))
D<-matrix(c(a,bb,cc,dd,e),2^i,i)
print(D)

j<-1
while(j<i^i)
{
j<-j+1
D21<-rbind(D[1,], D[j,])
print(D21)
mod1=(D21[2,])%%2
M21=matrix(c(mod1),1,i)
}

if(i==2)
{
 if(M21[1,1]==1)
 {
```

```
      if(M21[1,2]==1) "AB" else " A"
     }
   else "B"
 }

 else if(i==3)
 {
 if(M21[1,1]==1)
 {
  if(M21[1,2]==1)
 {
    if(M21[1,3]==1) " AB AC BC"
   else " C AB ABC"
   }
}
 else if(M21[1,2]==0)
{
 if(M21[1,3]==1) "B AC ABC"  else " B C BC"
}
 else if(M21[1,1]==0)
{
 if(M21[1,2]==1)
{
 if(M21[1,3]==1) "A BC ABC" else " A C AC"
}
}
else "A B AB "
}
}
```

감사의 글

대학원 생활하는 동안 그리고 논문을 쓰기까지 도와주신 많은 분들이 없었다면 저 혼자의 힘으로는 결코 오늘의 결과를 얻지 못했을 것입니다. 도움을 주신 모든 분들께 이 글을 빌어 감사의 마음을 전합니다.

학부때부터 논문을 쓰기까지 부족한 저를 항상 챙겨주시며 지도해주신 최규정 교수님께 감사드립니다. 큰 관심과 사랑으로 대해주신 조병엽 교수님께도 감사의 말씀을 올립니다. 항상 웃는 얼굴로 대해주시며 격려해주신 배상현 교수님, 늘 인자한 미소로 대해주신 이응기 교수님께도 감사드립니다. 논문을 성심을 다해 심사해주셨던 정채영 교수님, 조용한 목소리로 배우는 사람 입장에서 먼저 생각해주신 권용만 교수님, 편안함과 친근함으로 대해주시고 늘 챙겨주신 장인홍 교수님께도 감사드립니다.

먼 거리도 마다하지 않고 시간을 내 주시고 저의 미흡한 논문을 자상하게 심사해 주시고 학문적 조언을 해주신 박동호 교수님, 전종우 교수님께 머리 숙여 감사의 말씀을 드립니다.

오랜 시간 함께 동문수학했던 이장재 선배님, 송병호 선배님, 늘 친동생처럼 따르며 함께 어울릴 수 있었던 오지애, 이승우, 전준훈 등 통계실습실 선·후배님들에게도 감사의 마음을 전하며 건강과 발전을 기원합니다.

교수님들은 물론 선후배들과 인간적인 교류를 하면서 돈독한 유대감을 형성하도록 도와주신 양성기 선배님, 박병전 선배님, 이광옥 선배님, 이대영 선배님 그리고 대학원 생활동안 고민도 함께 하며 즐거움을 함께했던 대학원 선후배님들 또한 잊지 못할 고마우신 분들입니다. 또한 변함없는 우정으로 대해준 친구들에게도 감사의 말을 전합니다.

지금까지 헌신적인 사랑으로 키워주시고 공부할 수 있도록 도와주시고 항상 곁에서 응원해주신 어머님, 늘 도와주고 사랑을 베풀어준 때론 오빠 같은 존재 동생 창훈에게도 고맙다는 말을 전합니다. 그리고 언제나 자상하게 대해주시고 부족한 며느리 늘 챙겨주시는 시어머님과 시댁 식구들께도 감사드립니다.

마지막으로 늘 사랑으로 대해주고 논문과 결혼 준비로 힘들어할 때마다 격려해주고 이해해 준 세상에서 가장 소중하고 사랑하는 남편에게도 고마움과 사랑의 마음을 전합니다.

오늘의 결과에 자만하지 않고 계속 정진하면서 살아가도록 하겠습니다. 다시 한번 도움을 주신 모든 분들에게 감사드립니다.

*2006*年 *12*月

李 正 和

저작물 이용 허락서

학 과	전산통계	학 번	10141065	과 정	박사
성 명	한글: 이 정 화 한문 : 李 正 和 영문 : Lee Jung-Hwa				
주 소	경기도 성남시 분당구 구미동 무지개마을 주공12단지1207-1103				
연락처	016-606-2789	E-MAIL	cssdream@hanmail.net		
논문제목	한글 : 균형요인실험계획법을 이용한 cDNA Microarray Designs에 관한 연구				
	영어 : Analysis of cDNA Microarray Design using the Balanced Factorial Designs				

본인이 저작한 위의 저작물에 대하여 다음과 같은 조건아래 조선대학교가 저작물을 이용할 수 있도록 허락하고 동의합니다.

- 다 음 -

1. 저작물의 DB구축 및 인터넷을 포함한 정보통신망에의 공개를 위한 저작물의 복제, 기억장치에의 저장, 전송 등을 허락함
2. 위의 목적을 위하여 필요한 범위 내에서의 편집·형식상의 변경을 허락함. 다만, 저작물의 내용변경은 금지함.
3. 배포·전송된 저작물의 영리적 목적을 위한 복제, 저장, 전송 등은 금지함.
4. 저작물에 대한 이용기간은 5년으로 하고, 기간종료 3개월 이내에 별도의 의사 표시가 없을 경우에는 저작물의 이용기간을 계속 연장함.
5. 해당 저작물의 저작권을 타인에게 양도하거나 또는 출판을 허락을 하였을 경우에는 1개월 이내에 대학에 이를 통보함.
6. 조선대학교는 저작물의 이용허락 이후 해당 저작물로 인하여 발생하는 타인에 권리 침해에 대하여 일체의 법적 책임을 지지 않음
7. 소속대학의 협정기관에 저작물의 제공 및 인터넷 등 정보통신망을 이용한 저작물의 전송·출력을 허락함.

동의여부 : 동의(○) 반대( )

2007년 2 월 일

저작자: 이 정 화 (서명 또는 인)

조선대학교 총장 귀하