특집: 생물공정 기술의 최적화(II-I)

통계학적방법을 이용한 배지조성의 최적화

이희찬 · 김병기[1]
선문대학교 공과대학 화학공학과 및 연세대학교 생물산업소재센터
[1]서울대학교 유전공학연구소 및 공과대학 공업화학과

미생물이 인간생활에 이용된 것은 인류문화의 발전과 그 역사적 궤도를 함께 하고 있지만, 산업에 체계적으로 이용된 것은 Pasteur가 1850년대에 미생물과 효모를 생리학적으로 정의하고 무균기법, 최소배지(minimal media) 등을 도입한 이후라고 할 수 있다. 1869년에는 Raulin에 의해 *Aspergillus*의 배양을 위한 최초의 completely defined medium이 보고되었고, 1870년에는 Koch에 의해 순수배양법이 소개되었다. 1933년에는 Kluyver와 Perquin에 의해 플라스크를 사용한 진탕배양법이 개발되어 미생물배양에 기구의 이용이 본격화되었고, 1950년대에 들어서면서 Monod, Hinshelwood 등에 의한 세포성장속도와 배양조건과의 관계가 본격적으로 연구되기 시작하였다. 이와 같이 다른 학문영역에 비하여 상당히 느린속도로 발전하던 중, 2차 세계대전으로 인하여 페니실린의 대량생산이 요구되면서 생물공정 기술들이 급격히 개발된 초기에는 경험에 의하여 얻어진 것들이 대부분이었다. 최근에 들어 NMR, Flowcytometer 등의 분석능이 뛰어난 분석기기와 유전공학적 방법을 이용한 정량적인 대사기능의 해석이 시도되고 있지만 현재까지도 생물공정의 많은 부분이 경험에 의해 축적된 know-how에 의존하는 경우가 많은데, 그 이유는 체계적인 실험에 의한 분석이 미비한 점도 있지만 미생물이 갖는 변이성에도 기인하는 바가 크다고 할 수 있다. 이와 같은 변이성은 배지조성 등의 중요한 배양조건을 결정하는 데 크게 영향을 미치고 새로운 균주가 지속적으로 개발됨에 따라 계속 새로운 배지를 최적화 하여야하는 어려움이 있다. 따라서, 본고에서는 계속되는 새로운 배지의 최적화를 잘 알려진 통계적 방법에 의한 적은 수의 실험으로 노력을 극소화 할 수 있는 방안에 대하여 논하고자 한다. 궁극적으로는 주어진 조건하에서 목적하는 바(예: 세포질량, 재조합단백질 등)를 극대화 할 수 있는 최적의 배지조성을 얻는 것이 목표이고, 이와 더불어 모든 실험에 기본적으로 수반되는 오차를 줄이는 방안도 함께 논의 될 것이다.

실험디자인방법

실험의 기본적인 목적은 data를 수득하여 필요한 결정을 내리기 위함이다. 이와 같은 data를 수득하기 위한 계획을 실험디자인이라 할 수 있다. 실험디자인의 기본적인 목표는 짧은 시간 안에 신뢰성 있고 유용한 data의 생산에 있고, 효율적인 실험계획의 수립 및 그에 따른 실천은 많은 경우에 시간과 경비를 상당히 절감하게 한다. 실험디자인이 성공적이지 못하면 경우에 따라 이럴 수도 저럴 수도 있는 애매한 결론을 내리게 하는 데 시간과 자본을 투자하게 되는 결과를 얻게 될 것이다. 따라서, 원하는 결과를 도출하는데 급급하기보다는 실험의 계획단계에서 충분한 시간을 투자하여 신중한 결정을 내릴 수 있도록 하는 것이 효율적일 것이다. 본고에서는, 현재 일반적으로 사용되고 있는 대표적인 실험디자인법중 가장 간단한 몇 가지를 그 개념과 함께 간단한 예를 소개하고자 한다. 우선, factorial design을 통해 기본적인 개념파악을 하고, random screening에 사용되는 fractional factorial design의 일종인 Plackett-Burman design에 관하여 간단히 설명할 것이다.

One Factor At a Time

전통적인 실험법으로 "한번에 한가지씩 (OFAT : one factor at a time)" 변화시키는 방법이 많이 사용되고 있다. 이 방법은 다른 모든 조건들을 고정시키고, 한 가지 요소만을 변화시키기 때문에, 주어진 조건에서 그 요소의 영향을 심도 있게 파악할 수 있는 것이 장점이지만, 다른 요소(factor)들의 조건이 변하면, 그 결과를 예측하기 어려운 문제점을 갖고 있다. 특히, 요소들간에 상호작용이 존재하는 경우에는 얻어진 결과가 실험대상의 요소만의 효과인지, 다른 요소와의 상승효과에 의한 것인지를 구분할 수 없게 된다. 그러나, 이 방법의 가장 큰 단점은 최대(또는 극대) 값을 찾기가 상당히 어렵다는 것이다. 예를 들면, 미생물 셀룰로오스의 생산에 영향을 미치는 두 가지 배지 성분인 구연산과 에탄올의 영향을 조사하는 실험을 수행하는 경우 에탄올과 구연산의 효과를 개별적으로 조사하는 OFAT의 전략으로 실험을 한다고 하자. 다음 페이지의 그림 1의 예에서 나타난 결과는, 에탄올이 존재하지 않는 조건하에서 구연산의 농도가 0에서 5 g/L로 증가하면 셀룰로오스의 생산은 점진적으로 증가하고, 구연산이 존재하지 않는 조건하에서 배지 1리터당 0에서 20 ml의 에탄올을 가입하는 것은 셀룰로오스의 생산에 좋지 않은 영향을 미침을 보여준다. 한번

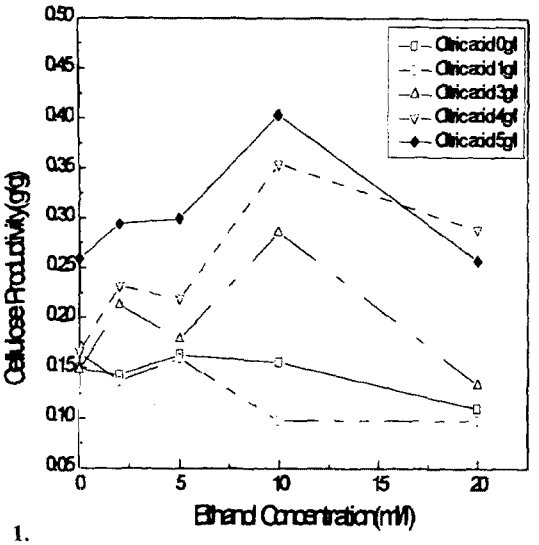

그림 1.

에 한가지씩 조사하는 전략에 의하여 얻은 결과는 에탄올 0 m*l*와 구연산 5 g을 사용하도록 권장하게 될 것이다. 에탄올의 영향을 먼저 조사하고 구연산의 효과를 조사하는 순차적인 조사방법에서도 에탄올의 효과가 없다고 단정하여 에탄올의 농도는 0 m*l*로 지정되고 그 조건하에서 구연산의 농도는 5 g으로 결정될 것이다. 운이 좋아서, 구연산의 영향을 먼저 조사하고 그 후에 에탄올의 영향을 조사한다면 최대수율을 얻을 수 있는 조건이 구연산 5 g과 에탄올 10 m*l*임을 알아낼 수 있을 것이다.

이러한 불확실한 방법에 의존하기가 싫을 경우는 구연산 및 에탄올의 농도를 5가지씩 25번 또는 그 이상의 실험을 수행해야 할 것이다. OFAT의 전략은 요인들의 상호작용을 무시할 수 있다고 가정하여 위와 같이 낮은 성공률을 갖게 되는 것이다. 한가지 더 고려해야 할 사항은 서로 다른 조건에서 얻어진 결과들의 차이가 실험오차의 크기보다 커야만 그 결과가 유효한 것이라는 것이다. 실험오차는 표준편차(s)에 의해 표시가능하고 결과의 유효성은 정상분포(normal distribution)이나 t-분포(normal distribution)를 가정하여 평균(Y)으로부터 표준편차의 몇 배수 이상이나 이하인 경우에 각 지점에 해당되는 확률로서 인정될 수 있다 (아래의 식 참조).

$$t=\frac{Y-\mu}{s/\sqrt{n}} \geq Y+(t_{(1-\alpha),d.f.})\cdot\frac{s}{\sqrt{n}},\ \text{또는}\ t \leq Y-(t_{(1-\alpha),d.f.})\cdot\frac{s}{\sqrt{n}}$$

where, t=t value of the sample, Y=sample mean value
α=probability value d.f.=degree of freedom
s=standard deviation μ=population sample mean
n=number of sample

표준편차의 일반적인 특성은 실험의 실행 횟수가 커질 수록 작아지는 것이다. 표준편차의 값이 작아지면 실험군들간의 아주 작은 차이들을 구분해 낼 수 있게 되는데, 이는 역설적으로 말하면 적은 수의 반복 실험으로는 작은 효과의 차이는 구분할 수 없게 된다. 따라서, 작은 차이를 구분하는 실험을 수행하려면 수 많은 실험을 수행해야 하고 이는 엄청난 시간과 경비를 요구하게 될 것이다. 실험오차는 실험과정에서 실수가 발생하지 않아도 생성되는 것으로 주로, 생물학적 변이성, 실험기술의 차이 등 실험계획에 의해 제거될 수 없는 것들이다. 이를 최소화하기 위해서 사용되는 기법으로 앞에서 거론된 반복실험 (replication), 그리고 실험의 순서를 무작위로 선택하는 무차별화 (randomization), 비교적 동일한 특성을 갖는 군끼리 묶는 구획화 (blocking) 등이 중요한 것이다. 원료물질의 lot no. 등과 같은 비슷한 성질을 갖는 것을 하나의 군으로 묶어주는 것 (Blocking)으로 block내의 오차를 줄여 주고, 실험순서에 따라 조직적으로 따라서 변하는 특성에 의한 오차를 줄이기 위해서 randomization을 사용하게 된다. Randomization과 blocking을 실험디자인에 도입하기 위해 Randomized Block Design(RBD)이 일반적으로 많이 사용된다. 조사대상 A, B, C, D를 3번 반복해야 한다면 전부 12번의 실험을 수행해야 하는데 A, B, C, D를 block단위로 하여 각 block내에서 A, B, C, D의 순서와 block간의 실험순서를 Random Number Table을 이용하거나, 뽑기 등의 방법으로 정해진 무작위순서에 의해 수행하는 것이다. 위에서 서술한 문제점들을 적은 수의 실험으로 완벽하게 해결할 수 있는 방안은 없다. 단지, 그러한 문제점의 발생 가능성을 최소의 노력으로 극소화하는데 그 의의를 찾아야 할 것이다.

Factorial Design

개의 변수(factor)를 가진 실험을 factorial design에 의해 수행하는 과정을 설명하기 위해 그림 2에 나타 낸 정육면체를 이용하자. 정육면체의 각 변은 각각 의 변수 x1, x2, x3를 나타내고 각 변수를 두 개의 단계(*l*evel)-높음, 낮음-으로 나누면 이와 같은 형태의 디자인을 2^3 (l^P: *l*-number of level, P-number of factor) Factorial이라 한다. 변수 x1은 정육면체의 왼쪽에 있는 4개의 꼭지점에 낮음, 그리고 오른쪽에 있는 4개의 꼭지점에 높음이 할당되고, x2, x3의 경우에는 전면과 밑면에 낮음과 후면과 윗면에 높음이 배정된다. 결과적으로 정육면체와 그의 내부공간은 실험에 사용될 수 있는 변수의 변화공간이 된다. 이와 같은 2단계(two-level) factorial은 각 변수의 주요효과(main effect)와 변수간의 상호작용을 결정할 수 있다. x1에 의한 주요효과는 정육면체의 왼쪽 면과 오른쪽 면의 결과를 비교하여 구할 수 있다. 정육면체의 전면아래의 모서리에서는 x2와 x3의 값이 일정하기 때문에 x1값의 변화에 따른 OFAT에 의한 실험치를 구할 수 있고, 이와 평행을 이루는 나머지 3개의 모서리에서도 마찬가지로 x1의 주요효과를 구할 수 있게 된다. 결과적으로, x1의 주요효과는 4개의 모서리에서 구한 효

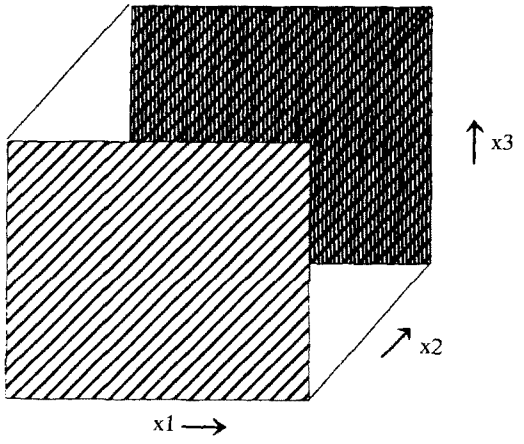

그림 2.

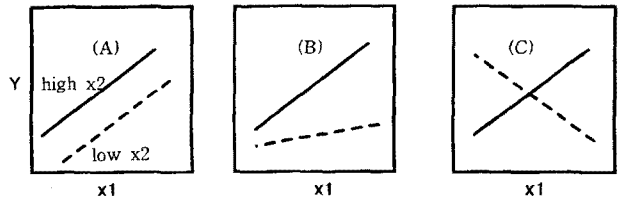

그림 3. 변수간의 상호작용

과의 평균치가 될 것이다. 이러한 결과는 고정된 x2, x3의 값에서만 결과를 얻게되는 OFAT의 접근방법의 단점을 보완하게 되고, 동시에 x1의 주요효과를 측정하기 위해 8개의 실험치 모두가 사용되어 자연스러운 반복효과(hidden replication)를 갖게 되는 것이다.

Factorial design에서 각 변수간의 상호작용(interaction)의 정도 또한 확인 할 수 있다. 두 개의 변수간의 상호작용이란, 하나의 변수에 의한 효과가 다른 변수의 높고 낮음에 따라 다르게 나타남을 말한다. 그림 3은 두 개의 변수 x1과 x2사이의 상호작용의 정도에 따른 변화를 보여준다. (A)와 같이 x2의 높고 낮음에 관계없이 x1의 변화에 대한 효과 Y가 동일한 기울기를 갖는 경우는 상호작용이 없고, (C)와 같이 x2의 높고 낮음에 의한 효과 Y의 변화가 극명하게 나타나는 경우는 상호작용이 강한 경우고 (B)는 그들의 중간의 경우이다.

이러한 상호작용의 정도는 2단계 실험디자인으로는 그 관계가 선형일 것으로 가정하고 분석할 수밖에 없으므로, 선형이 아닌 관계를 탐색하기 위해 정육면체 디자인의 중앙에 하나의 실험을 할당하여 총 9개의 실험으로 2단계 3변수의 실험디자인을 구성할 수 있다. 표 1의 예로서 보여진 실험디자인에서 높음을 +로 낮음을 -로 나타내고 중앙 점을 0으로 표시하고 있다. 1번부터 8번까지의 시도를 2반복하고 9번의 중앙점을 4반복하여 총 20회의 실험을 수행하며 그 순서는 무작위로 배열하되(randomization) 중앙점을 균등하게 배분하였다.

표 1. 2단계 3변수 Factorial Design

실험번호	실험순서	x1	x2	x3	Y	Y 평균	분산
1	2	-	-	-	9	10	2
	4				11		
2	6	+	-	-	30	33	18
	13				36		
3	5	-	+	-	14	17	18
	10				20		
4	15	+	+	-	38	43	50
	18				48		
5	3	-	-	+	91	95	32
	12				99		
6	9	+	-	+	125	127	8
	11				129		
7	8	-	+	+	155	155	0
	17				155		
8	16	+	+	+	177	178	2
	19				199		
9	1	0	0	0	63	65.5	27
	7				50		
	14				60		
	20				72		

(예를들면, x1=glucose conc., x2=pH, x3=phosphate conc., Y=cell conc.)

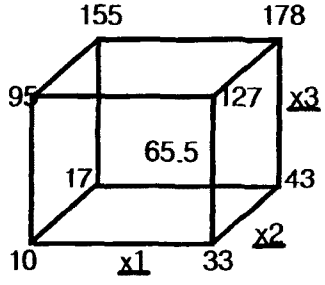

그림 4. 2단계 Factorial과 중심점

실험결과를 원래의 정육면체 디자인에 도시한 것이 그림 4이다. 앞에서 설명되었듯이 정육면체의 왼쪽과 오른쪽의 결과를 평균하여 비교하면 x1의 효과를 분석할 수 있고, x2, x3의 효과도 동일한 방법으로 쉽게 분석할 수 있다. 상호작용의 정도를 분석하기 위해, 정육면체의 윗면과 아랫면을 포개어 중첩되는 점을 평균한 값을 x1과 x2의 관계로 도시한 것이 그림 5a이고, 같은 방법으로 정육면체의 왼쪽과 오른쪽을 포개서 얻은 것이 그림 5b이다. x1과 x2의 관계는 상호작용이 무시할 만하고, x2와 x3의 사이에는 상호작용이 존재함을 알 수 있다.

이들 효과에 대한 유효성과 중심점을 이용한 상호작용의 곡률의 유효성은 Student's "t" test를 이용하여 계산할 수 있다.

Screening Design

Factorial design은 변수(factor)의 숫자가 늘어남에 따라 실험의 횟수가 기하급수적으로 증가하기 때문에 실용성이 낮아

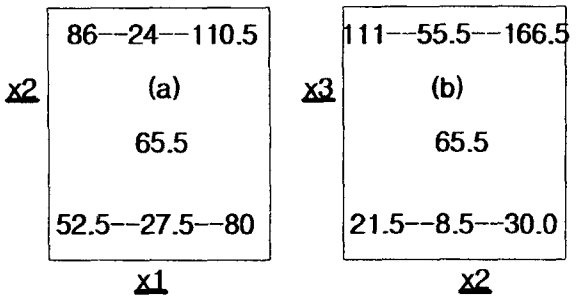

그림 5. 상호작용의 분석

지게 된다. 변수가 10개가 되면 1024회의 실험을 중복해서 수행해야 하기 때문이다. 이러한 경우에는 factorial Design의 일부를 사용하는 fractional factorial design이나 fractional factorial design의 특별한 경우인 Plackett-Burman design을 사용하여 중요한 변수를 선별하는 screening 작업이 필요하다. 이러한 실험 디자인방법은 여러 참고도서에서 자세히 다루고 있고, 종류도 상당히 많고 다양하고 상세히 소개되어 있다. 대표적인 것으로 Plackett-Burman design의 경우 변수가 11개[n개]인 경우 6개의 변수까지가 중요하다고 가정하고 12번[(n+1)번, 단 (n+1)은 4의 배수]의 실험을 요구하는 경우를 표 2에 나타내었다(반복실험이 요구되면 그 횟수는 24회). 표를 자세히 분석해보면 12회의 실험 중에서 각각의 변수들을 6회씩 포함시켰고, 각각의 변수들이 3회씩 중복되는 수학적 균형을 이루고 있는 것을 알 수 있다. 변수 1(x1)의 높은 값이 포함되는 1, 2, 4, 5, 6, 10번의 실험 중에 1, 4, 5번은 x2의 높은 값과 2, 6, 10번은 x2의 낮은 값과 중복되고, 2, 4, 10번은 x3의 높은 값과 1, 5, 6번은 x3의 낮은 값과 중복되고, 1, 5, 10번은 x6의 높은 값과, 2, 4, 6번은 x6의 낮은 값과 중복되는 등이 그 예이다. Factorial design에 의하면 2^{11}(=2048)회의 실험을 수행해야 하므로 12회의 실험횟수는 1%가 채 못되는 숫자이므로 시간과 실험경비를 상당히 절약할 수 있음을 알 수 있다. 적은 숫자는 그 만큼의 희생을 요구하는데 이 경우 변수들 간의 상호작용에 대한 정보를 얻을 수가 없게된다. 따라서, Plackett-Burman design에서는 단지 중요한 변수만을 선별하는데 중점을 두고 있는 것이다. 7번째 칼럼부터 11번째 칼럼은 실험오차를 계산하는데 사용하여 변수에 의한 효과의 유효성을 계측할 수 있게 한다. 일차적으로 11개의 변수 중에서 효과가 탁월한 것 3~4개를 선별하면 2단계로서 factorial design에서 이들 주요변수들의 상호작용 여부를 조사하기 위해 정육면체를 구성하였던 것과 같은 방법으로 실험을 수행하면 된다. Screening 과정에서 나타난 효과가 factorial design에서도 명백히 확인되면 더 이상의 실험을 진행하지 않고도 바로 실전에 이용할 수도 있을 것이다.

위에서는 단지 12회의 실험으로 11개의 변수로부터 중요한 변수 3~4개를 효과적으로 선별할 수 있는 Plackett-Burman design을 보았다. 실험의 회수를 16, 24회로 증가시킴에 따라 15 또는 23개의 변수에서 각 변수의 주요효과와 일부의 상호작용에 대한 자료도 얻을 수 있는 design이 가능하다. 이와 같이 10개 이상의 변수에 대한 design이 다수가 보고되어 있으므로, 초기의 screening 단계에서는 가능한 한 변수의 수를 줄이려고 노력할 것이 아니라, 가능한 변수를 최대한으로 포용하려고 노력해야 할 것이다. 원하는 결과를 도출하는데 급급하기보다는 실험의 초기단계에서 충분한 시간을 투자하여 중요한 변수를 누락시키지 않는 신중한 결정을 내릴 수 있도록 하는 것이 중요하다. 중요한 변수를 나중에 발견하여 모든 실험을 다시 하여야 하는 일이 발생하지 않도록 하여야 할 것이다.

표 2. 12-Run Plackett-Burman Design

	A	B	C	D	E	F	G	H	I	J	K
1	+	+	-	+	+	+	-	-	-	+	-
2	+	-	+	+	+	-	-	-	+	-	+
3	-	+	+	+	-	-	-	+	-	+	+
4	+	+	+	-	-	-	+	-	+	+	-
5	+	+	-	-	-	+	-	+	+	-	+
6	+	-	-	-	+	-	+	+	-	+	+
7	-	-	-	+	-	+	+	-	+	+	+
8	-	-	+	-	+	+	-	+	+	+	-
9	-	+	-	+	+	-	+	+	+	-	-
10	+	-	+	+	-	+	+	+	-	-	-
11	-	+	+	-	+	+	+	-	-	-	+
12	-	-	-	-	-	-	-	-	-	-	-

Latin Square의 응용

10개의 변수에 대한 factorial design이 1024회의 실험을 요구하는 것은 상호작용에 관한 효과를 검증하기 위한 것으로, 예를 들면 2개의 변수간의 상호작용을 위해서 ${}_{10}C_2$=45회, 3개의 변수간에는 ${}_{10}C_3$=120회 등의 실험이 요구되는 것이다. 실제의 경우에는 4개 이상의 변수간의 상호작용은 거의 존재하지 않는다고 가정해도 큰 무리가 없고, 간혹 상호작용이 중요한 경우라도 그것을 찾아내기란 쉬운일이 아니기 때문에, 일부의 상호작용이 존재하지 않는다고 가정할 수 있으면 Latin Square 등을 이용한 방법을 응용할 수 있을 것이다.

결 론

발효배지의 조성을 최적화하는데 필요한 통계학적 방법으로 factorial design의 기본개념을 설명하였고 이 방법이 전통적으로 사용되어 온 OFAT의 기법보다 효율적임을 보여주었다. 변수간의 상호작용을 조사하기 위한 factorial design의 중심점이 있는 예로서 주어졌다. 변수의 수가 많아지는 경우 중요한 변수를 선별하기 위해 factorial design의 일부로 구성되

어 있는 Plackett-Burman design등을 사용할 수 있음을 알았다. 발효배지의 최적화에는 우선 product의 생산에 영향을 미칠 수 있는 가능한 모든 변수를 나열하고 이들 중에서 문헌조사, 연구진간의 토의, 실험운영조건 등의 제반조건을 고려하여 screening 대상변수를 선정한 후 중요한 요소의 선별을 위해 fractional factorial design이나 Plackett-Burman design 등을 사용할 것이고, 선별된 중요변수들로 factorial design이나 2차적인 fractional factorial design 등으로 최종 조성을 결정하여야 할 것이다. 실험디자인에는 오차를 최소화하기 위한 중복실험(replication), 실험순서의 randomization, 그리고 유사한 부류를 군으로 통합하는 blocking을 항상 고려하여야 한다. 실험오차의 극소화는 적은 수의 반복실험으로도 대조군과 차별이 되는 유효변수를 선별해 낼 수 있게 하는 기본적인 요소이기 때문이다. Blocking을 하는 방법은 매우 다양하고 통계학의 이론적인 이해도 곁들여야하기 때문에 지면관계상 이곳에서는 다루지 못하지만 관심 있는 독자들은 제시된 참고문헌들을 참조할 수 있을 것이다. 본고에서 다루지 못한 많은 부분에 대해서도 제시된 참고문헌을 이용할 수 있을 것이다.

참고문헌

1. Haaland, P. D. 1989. Experimental Design in Biotechnology, Marcel Dekker, Inc., New York and Basel.
2. Ostle, B. and Linda C. Malone 1988. Statistics in Research, 4th ed., Iowa State University Press/AMES.
3. Petersen, R. G. 1985. Design and Analysis of Experiments, Marcel Dekker, Inc. New York and Basel.
4. Box, G. E. P., W. G. Hunter, and J. S. Hunter 1978. Statistics for Experimenters, Wiley, New York.
5. Plackett, R. L. and J. P. Burman 1946. The design of optimum multifactorial experiments. *Biometrika,* **33**, 305-325.