The Korean Communications
in Statistics Vol. 9, No. 3, 2002
pp. 675-682

Program Development of Genetic Analysis for Diallel Cross Experiment

Seo Young Kim[1], Jong Sung Bae[2]

Abstract

In this study, we develop the statistical analysis program for genetic analysis of diallel crosses data by SAS/MACRO, SAS/IML. Genetic analysis is to estimate of genetics parameters and heredity with reciprocal cross and without reciprocal cross. Statistical analysis program solve the problem of the difficulties on the data analysis in field denetics and breeding. Therefore the user whoever want to analysis of data on genetics and breeding easily conduct the work saving time and suffering.

Keywords : Genetic Analysis, Diallel Cross, Expeimental Design.

1. 서론

국내외 농사시험연구를 포함한 대부분의 분야에서 연구목적 달성을 위해 얻어지는 시험자료는 통계분석을 거쳐 연구결과를 검증하고 있다. 따라서 연구자들은 보다 정확하고 사용하기 편리한 분석도구를 필요로 하고 있다. 그러나 사회과학 및 일반적인 자연과학 자료를 분석할 수 있는 프로그램은 다양하게 개발되어 있지만, 유전 분석을 위한 전문 통계분석 프로그램의 개발은 상당히 미비한 것으로 조사되었다. 특히, 작물육종연구에서 널리 사용되고 있는 이면교배 실험은 현재 국내에서 농촌진흥청을 중심으로 한 국가 연구기관에서 많이 사용되고 있다. 이면교배는 주로 식물의 육종실험(breeding experiment)에서 각 계통(inbred line)의 유전적인 성질을 연구하기 위해 사용되는 짝짓기 계획이다. 이면교배는 식물육종에서도 자식성 식물에 주로 사용되는 방법으로 현재 국내에서는 벼, 보리, 밀, 담배와 같은 자식성식물의 육종에 주로 사용되고 있다.

육종가들은 시험자료에 대해서 유전분석을 통해 최종적인 작물에 대한 유전정보를 얻게 된다. 따라서 실험자료에 대한 정확한 분석은 연구자로 하여금 정확한 유전적 해석을 가능하게 한다. 현재 사용되고 있는 이면교배자료의 유전분석 프로그램은 AGRISP이 대표적이고 Genstat 이나 Mystat, APL등이 있다. 그러나 이러한 프로그램은 여러 가지 단점을 지니고 있다. 첫째, 대부분 Dos용 프로그램으로 자료의 호환이 어렵다. 둘째, 자료입력이 대화식(interactive)으로 Dos용 화면에 직접 입력하기 때문에 수정이 번거럽다. 셋째, 잘못된 대화로 인해 처음부터 프로그램 수행을

1) Lecturer, Department of statistics, Chonnam National University, Gwangju, 500-757, Korea.
E-mail : gong@chonnam.ac.kr
2) Professor, Information and Telecommunication Reasearch Institute, Department of statistics, Chonnam National University, Gwangju, 500-757, Korea.
E-mail : jsbae@chonnam.ac.kr

다시 해야 하는 경우가 발생한다. 넷째, 통계분석 프로그램이지만 통계분석 전문 프로그램이 발달하지 않았던 70, 80년대 초에 개발된 관계로 결과에 대한 정확한 검증을 할 필요가 있다.

현재 사용되고 있는 통계분석 전용 프로그램은 대부분의 연구기관에서 보유하고 있지만, 유전분석을 위한 분석기법은 포함하고 있지 않다. 예를 들어 SAS를 이용하여 유전분석을 하고자 하는 경우 우선, 분석을 위해서는 복잡한 자료 핸들링이 필수적이다. 또한 한가지 분석 결과를 얻기 위해 여러 단계의 프로시져를 단계적으로 수행해야 하기 때문에 통계적 알고리즘 및 유전분석 이론을 정확하게 이해할 필요가 있다. 따라서 이러한 문제점을 보완하고 사용자 위주의 유전분석 프로그램을 개발할 필요가 있다.

본 연구는 통계전문가가 아닌 육종연구자들이 유전분석을 보다 용이하게 할 수 있도록 하는 분석 프로그램을 개발하는데 그 목적을 두었다. 2장에서 유전분석에 필요한 개념 및 분석방법을 살펴보고, 3장에서 SAS/IML과 MACRO기능을 이용하여 실제자료에 대한 유전분석 결과를 도출을 위한 구체적인 프로그램 설계과정을 살펴본다. 4장에서 결론 및 논의사항을 언급하였다. 참고로 본문에서 사용되는 유전·육종 용어와 관련된 유전 및 육종학적 해석은 그 내용이 어렵고 본 연구의 범위를 벗어나므로 언급하지 않았음을 미리 밝혀둔다. 이에 관련된 내용은 Prem et al(1979), 안장순 등(2000), 채영암 등(1993)을 참고할 수 있다.

2. 유전분석 내용 및 방법

2.1 유전분석 내용

이면교배의 유전분석은 역교배(reciprocal cross)가 있는 경우와 없는 경우로 나누어 수행하였다. 대부분의 육종가는 이면교배 자료를 행렬형태로 보관한다. 따라서 본 연구에서는 분석에 사용될 자료는 사용자가 이해하기 쉽게 행렬형태로 입력하도록 하였다. 하나의 유전자 A가 만드는 두 유전자형 AA와 aa의 이면교배에서 얻어진 F1에서 각 열(array)의 평균, 분산 및 공분산을 구하고, V_p은 교배친 분산, V_r은 각 열의 분산, W_r은 각 열의 유전자형과 대각선 상의 교배친 사이의 공분산을 나타낸다. 이때 유전분석에 사용되는 표기법은 Hayman의 표기법(채영암, 이영만,1993)에 따른다.

1) 분산과 공분산

V_{0L0} : 교배친간의 분산(= V_P)

V_{0L1} : 각 행(array)의 평균의 분산

V_{1L1} : 각 행의 분산의 평균

W_{0L01} : 교배친과 행 사이의 공분산 평균

$M_{L1} - M_{L0}$: 교배친 평균과 전체 1 세대의 평균 사이의 차

2) 유전모수 추정

- $D = V_{0L0} - E$: 유전자의 상가성 효과에 의한 분산성분
- $F = 2V_{0L0} - 4W_{0L01} - \frac{2(n-2)}{n} E$: 행에 대한 상가성 및 우성효과에 대한 공분산의 평균
- $H_1 = V_{0L0} - 4W_{0L01} + 4V_{1L1} - \frac{3n-2}{n} E$: 유전자의 우성효과에 의한 분산성분
- $H_2 = V_{1L1} - 4V_{0L1} - 2E$: 부모세대에서의 유전자 비율
- E : 환경분산 성분

3) 환경분산 E를 추정

① 분산분석결과 집구간에 유의성이 없을 때에는 오차의 평균제곱(MS)를 이용한다.

$$E = \frac{\text{오차 MS}}{\text{집구수}}$$

② 집구간에 유의성이 있을 때에는 집구와 오차를 합해 추정한다.

$$E = \frac{(\text{오차 제곱합(SS)} + \text{집구SS})/(\text{오차자유도} + \text{집구 자유도})}{\text{집구수}}$$

4) 유전력 추정

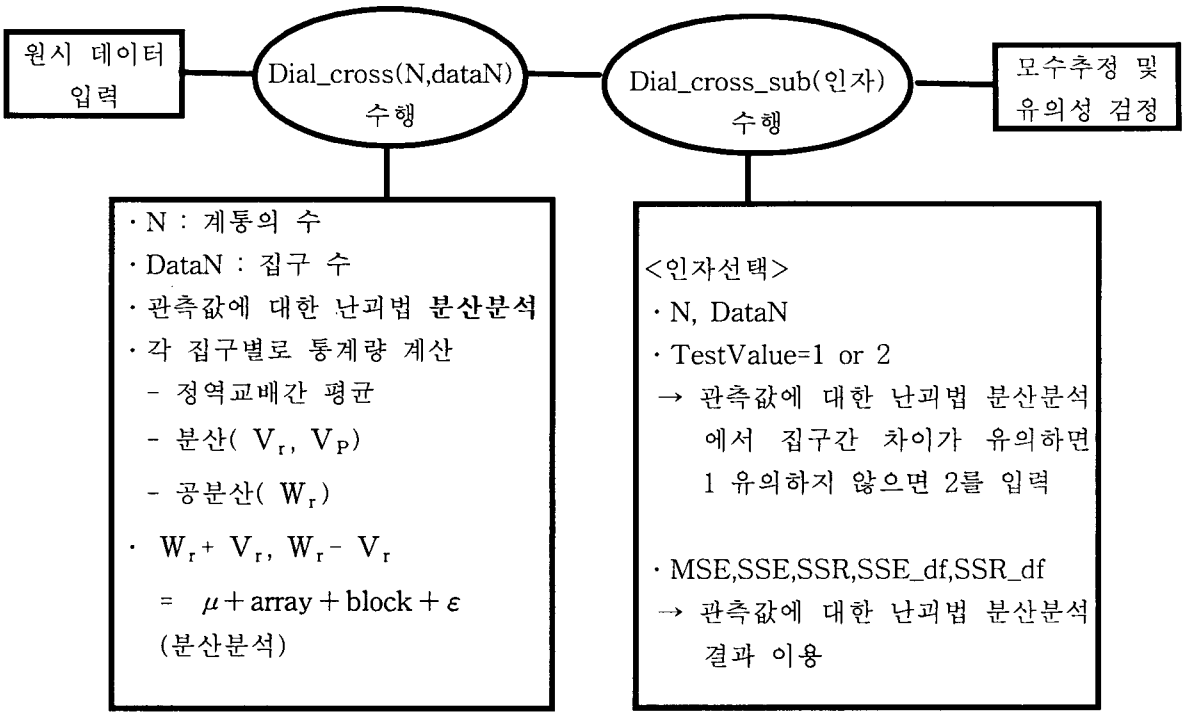
- 상가성 분산(Additive variance) : $V_A = \frac{1}{2} D + \frac{1}{2} H_1 - \frac{1}{2} H_2 - \frac{1}{2} F$
- 우성분산(Dominance variance) : $V_D = \frac{1}{4} H_2$, $V_E = E$
- 협의의 유전력(Heritability in narrow-sence) : $h_N^2 = \frac{V_A}{V_A + V_D + V_E}$
- 광의의 유전력(Heritability in broad-sence) : $h_B^2 = \frac{V_A + V_D}{V_A + V_D + V_E}$

5) 유전모수 및 유전력의 유의성 검정

유의성 검정에서는 신뢰구간을 이용하여 유의수준 5%와 1%에서의 유의성 결과를 도출한다.

2.2 프로그램 설계

연구자는 다음 그림과 같이 처음 3단계에 몇 개의 인자를 입력함으로써 유전분석을 쉽게 수행할 수 있다.



[그림1] 프로그램 구성도

[그림1]과 같이 SAS/MACRO가 수행되는 과정에서 계산되는 결과들을 구체적으로 살펴보면 다음과 같다.

① 원시자료에 대한 난괴법 분산분석

② 집구별 원시 데이터와 교배친 분산(V_r)과 공분산(W_r), Wr + Vr, Wr − Vr을 계산

③ Wr + Vr=집구+행+오차, Wr − Vr=집구+행+오차의 분산분석

④ 반복별 성적의 평균값, V_r, W_r계산

⑤ V_{0L0}, V_{0L1} , V_{1L1} , W_{0L0i} , $M_{L1} - M_{L0}$ 계산

⑥ 유전모수 추정

⑦ 각 분산 및 유전력 추정

⑧ 각 유전모수 추정치의 표준오차 계산을 위한 곱의 수(multiplier)

⑨ 유전모수의 분산과 표준오차

⑩ 유전모수의 유의성 검정(추정치의 신뢰구간 이용)

3. 분석결과 출력

프로그램 사용방법을 설명하기 위해 채영암, 이영만(1993)의 역교배가 없는 반이면교배(half diallel cross) 자료를 이용하였다. 대부분 육종 연구가들은 다음과 같이 행렬형태로 데이터를 보관하는 것이 일반적이다. 다음의 예는 보리의 6개 품종간 이면교배에서 6개 교배친과 30개의 F1을 난괴법 2반복으로 조사된 수당입수에 관한 자료이다.

집구2						
품종	1	2	3	4	5	6
1	45.9	48.7	38.2	30.8	33.7	55.4
2	37.4	39.5	35.2	31.1	42.2	52.5
3	35.1	34.0	24.8	29.8	31.2	40.9
4	29.3	29.1	30.1	25.1	35.8	32.8
5	34.7	44.0	30.7	26.5	41.0	43.6
6	56.4	51.6	38.9	34.5	42.5	53.4

집구1						
품종	1	2	3	4	5	6
1	47.9	47.6	40.5	31.3	39.4	52.5
2	48.8	41.9	39.0	30.5	45.1	53.0
3	41.6	38.4	31.8	30.5	35.7	37.9
4	32.2	30.0	29.6	22.6	34.8	30.2
5	38.4	46.2	34.8	33.9	42.3	43.0
6	51.9	52.2	38.6	29.3	42.4	53.5

위의 이면교배 자료에 대해 개발된 프로그램을 이용하여 유전분석을 수행하는 과정을 살펴보면 다음과 같다.

1) 원시자료를 입력한다.

난괴법 2 반복이므로 두 개의 DATA set을 구성하고, 품종 수가 6이므로 INPUT문에 X1-X6이라는 6개의 변수 명을 사용한다. 만약 반복이 3이상이면 이와 똑같은 방법으로 DATA SET만 더 추가하면 된다.

```
DATA _class_1;                          DATA _class_2;
      input x1-x6 @@;                         input x1-x6 @@;
cards;                                  cards;
47.9  47.6  40.5  31.3  39.4  52.5      45.9  48.7  38.2  30.8  33.7  55.4
...                                       ...
51.9  52.2  38.6  29.3  42.4  53.5      56.4  51.6  38.9  34.5  42.5  53.4
; RUN;                                  ; RUN;
```

2) dial_cross(N, dataN)을 수행한다.

이때 N : 품종의 수, dataN : 집구수(반복수)를 나타낸다. 위 예의 경우 품종이 6이고, 반복이 2이므로, dial_cross(6, 2)를 입력하고 수행하면, 다음과 같은 결과들이 출력된다. [표1]에서 환경분산 E를 추정하기 위해 먼저 집구간에 유의성을 판단한다. 유의확률 p값이 0.0174로 유의수준 5%에서 유의하다. 이 단계의 분석결과로부터 dial_cross_sub(인자들) 수행에 필요한 다음과 같은 인자들을 넘겨준다. SSE=218.418194, MSE = 6.240520, SSR=38.866806, SSE_DF=35, SSR_DF=1.

[표1]. 관측값에 대한 분산분석

Dependent Variable: _VALUE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	36	4762.205000	132.283472	21.20	<.0001
Error	35	218.418194	6.240520		
	↓	↓	↓		
	sse_df	sse	mse		
Corrected Total	71	4980.623194			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
cross(유전자형)	35	4723.338194	134.952520	21.63	<.0001
Group(집구)	1	38.866806	38.866806	6.23	0.0174
	↓	↓			↓
	ssr_df	ssr			집구간 차에 대한 유의성 검정

[표2]는 정교배와 역교배에 대한 집구1의 평균과 분산, 공분산, W_r+V_r, W_r-V_r의 결과를 나타낸 것이다. 집구2에 대한 결과도 동일한 형태로 출력된다.

[표2]. 정역교배간 평균, V_r, W_r, W_r+V_r, W_r-V_r,

_VR1	_WR1	SUMVRWR1	DIFVRWR1
56.439667	75.02	131.45967	18.580333
61.680667	86.359	148.03967	24.678333
18.265	42.329	60.594	24.064
15.322417	29.856	45.178417	14.533583
20.015417	36.209	56.224417	16.193583
93.019667	100.296	193.31567	7.2763333

47.9	48.2	41.05	31.75	38.9	52.2
48.2	41.9	38.7	30.25	45.65	52.6
41.05	38.7	31.8	30.0	35.25	38.25
31.75	30.25	30.05	22.5	34.35	29.75
38.9	45.65	35.25	34.35	42.3	42.7
52.2	52.6	38.25	29.75	42.7	53.5

[표3]. Wr+Vr의 array와 집구에 대한 이원분산분석

Dependent Variable: sumVrWr(Wr+Vr)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	32080.10226	5346.68371	13.14	0.0063
Error	5	2034.25631	406.85126		
Corrected Total	11	34114.35857			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
array	5	31891.25852	6378.25170	15.68	0.0045
Group	1	188.84375	188.84375	0.46	0.5260

[표3]의 분석결과는 Wr+Vr의 행(array)과 집구에 대한 분산분석을 수행한 것이다. Wr-Vr에 대한 분석결과도 동시에 출력된다. 마지막으로 Dial_cross(인자들) 수행결과에서 넘겨받은 인자들을 이용하여 다음 단계를 수행한다.

3) dial_cross_sub(N, dataN, TestValue, MSE, SSE, SSR, SSE_Df, SSR_Df)를 수행한다.

여기서 N : 품종수, dataN: 반복수, TestValue: 원시자료 분산분석에서 집구간에 유의하면 1, 유의하지 않으면 2를 입력하고, 2반복 성적의 분산분석 결과에서 MSE : 오차 MS, SSE : 오차 SS, SSR : 집구 SS, SSE_Df : 오차 자유도, SSR_Df : 집구 자유도 값을 사용하면 된다. 따라서 위 예제의 경우, 다음과 같이 입력하여 수행하면 된다.

***%dial_cross_sub*(6, 2, 1, 6.2405, 218.418, 38.866, 35, 1, 1.96);**

%dial_cross_sub를 수행함으로써 [표4]와 같이 각종 유전모수와 유전분산, 유전력, 유전모수의 표준편차 및 신뢰구간을 유의성 검정결과 등 이면교배의 유전분석에서 필요한 대부분의 결과들을 출력할 수 있다.

[표4]. 유전모수 추정 및 유의성 검정

유전모수 추정치

E_HAT	D_HAT
3.5733889	121.43003

H1_HAT	H2_HAT	F_HAT	D_H1
45.679324	40.577454	-3.86863	75.750704

F_P	ROOTH1_D	H2_4H1	KD_KR
-9.86863	0.6133336	0.2220778	0.9998257

유전모수 추정치의 분산과 표준오차

VAR_D	STD_D	VAR_F	STD_F
27.206435	5.2159788	162.37491	12.742642

VAR_H1	STD_H1	VAR_H2	STD_H2	VAR_E	STD_E
175.33036	13.241237	139.91881	11.828728	3.8866336	1.9714547

유전모수 추정치의 신뢰구간

신뢰구간이 0을 포함하면 주어진 유의수준하에서 유의하지 않다.

LD_INT	HD_INT
111.20671	131.65335

LF_INT	HF_INT
-28.84421	21.106948

LH1_INT	HH1_INT
19.726499	71.632149

LH2_INT	HH2_INT
17.393147	63.761761

LE_INT	HE_INT
-0.290662	7.4374401

[표4]의 결과에서 LD_INT는 유전모수 D에 대한 신뢰하한, HD_INT는 신뢰상한을 나타내고, 다른 모수에 대해서도 동일한 방법으로 해석할 수 있다. 분석에 필요한 이론적 배경에 대해서는 통계유전육종학(1993, 채영암, 이영만)을 참고하기 바란다.

4. 결론 및 의견

본 연구는 작물육종분야에서 주로 사용되는 이면교배실험 자료에 대한 유전분석 프로그램을 육종 연구자가 보다 쉽게 이용할 수 있도록 기존에 사용되고 있는 프로그램들을 보완하여 작성하고자 하였다. 실제로 연구자가 2001년 한해동안 농촌진흥청에서 경험한 결과에 따르면, 이면교배 방법은 현재 육종연구분야에서 유용하게 널리 이용되고 있는 실험방법은 아닌 듯 하였다. 그러나 농촌진흥청과 같은 주요 국가연구기관에서는 작물육종에서 이면교배 방법은 중요하게 사용되고 있는 것 또한 사실이다. 그럼에도 불구하고 대다수의 이면교배 육종연구자들은 적절한 분석프로그램이 없어 많은 어려움에 직면하고 있음을 경험할 수 있었다. 이미 언급하였듯이 대부분의 프로그램들이 사용방법이 까다롭고, 그 프로그램들을 유능하게 다룰 수 있는 연구자마저 거의 없는 실정이다. 농촌진흥청을 비롯한 대부분의 농사시험 연구기관에서는 통계분석용 SAS를 보유하고 있기 때문에 본 연구는 SAS를 이용하여 사용자가 보다 쉽게 이용할 수 있도록 하는데 중점을 두었다.

본 연구에서 개발된 이면교배자료의 유전분석 프로그램은 완전한 사용자 위주는 아니지만, 육종연구자들에 상당한 도움이 될 것이다. 실제로 농촌진흥청에서 몇 건의 시험자료에 대한 분석을 시행한 결과 연구자로 하여금 상당히 만족한 분석결과를 얻을 수 있었다. 좀더 비쥬얼한 메뉴방식의 분석 시스템 개발이 절실한 시점에서 앞으로 SAS/AF와 같은 소프트웨어 개발 툴을 이용하여 보다 사용하기 쉬운 전문 유전분석 소프트웨어를 개발을 추진하고 있다.

프로그램은 gong@chonnam.ac.kr에서 이용할 수 있다.

5. 참고문헌

[1] 채영암, 이영만 (1993). 통계유전육종학, 향문사

[2] 안장순, 이영만, 민경수 (2000). 작물육종학, 전남대학교 출판부

[3] CRISP 해설집 (1983). 농촌진흥청

[4] Manugistics Institute.(1993). *Apl*Plus PC Reference mannual, ver.11.* Manugistics, Inc.

[5] SAS Institute.(1997). *SAS/STAT software:Changes and enchancements tgrough release 6.12.* SAS Inst., Cary,NC.

[6] SAS Institute.(1997). *SAS/IML software:Changes and enchancements through release 6.12.* SAS Inst., Cary,NC.

[7] Statistics Department, Rorhamsted Experimental Station.(1987). *Genstat 5 Refernece mannual.* Clarendon Press · Oxford.

[8] Prem Narain, V.K.Bhatia and P.K.Malhotra(1979). Handbook of statistical genetics, ICAR in New Delhi.

[2002년 6월 접수, 2002년 10월 채택]