Biometrical Analysis and Estimation of the Number of Genes for Seed Protein Content of Soybean, *Glycine max* (L.) Merrill.

By TERUO ISHIGE

Department of Physiology and Genetics, National Institute of Agricultural Sciences
(Yatabe, Ibaraki, 305 Japan)

Soybean is one of the crops used as a principal source of vegetable protein and oil in human foods and animal feeds. Especially in Asian countries including Japan, soybean is the most important resource of vegetable protein for human foods. An efficient method to increase the yield of soybean protein is to improve protein content of soybean seeds. The study reviewed here deals with genetic analysis of the protein content of soybean seeds by means of biometrical methods,[3] and estimates the number of genes controlling the seed protein content by the maximum likelihood estimation method.

Chemical analysis of soybean seed protein content

For the study on genetics of protein content in soybean, it was necessary to measure protein contents of a large number of seed samples by the Kjeldahl method. However, the analysis by this method requires a lot of time and labor. Therefore, it is desired to develop a rapid, simple, and inexpensive method of chemical analysis. We attempted to modify the biuret method,[4] and the result of measurement by the modified method was compared with the protein content analysed by the Kjeldahl method.

* Present address:
Department of Cell Biology, National Institute of Agrobiological Resources (Yatabe, Ibaraki, 305 Japan)

As shown in Table 1, there was no significant differences in means and variances obtained from these two methods. Therefore, chemical analysis in this study was conducted by the modified biuret method.

1) Modified biuret method

Biuret solution: To 920 m*l* of distilled water, add 10 m*l* of 10 N potassium hydroxide solution and 20 m*l* of 25% sodium potassium tartrate hydroxide solution. Add slowly with vigorous stirring 50 m*l* of 4% $CuSO_4 \cdot 5H_2O$ solution.

Procedure: About 80 mg of sample (ground meal) filtered through 80 mesh screen was dried for one hr at 100°C and weighed. Mix 1 m*l* of chloroform with the sample, add 30 m*l* of biuret solution, and shake vigorously for 15 min. Centrifuge at 8,000 rpm for 15 min. Determine the absorbance intensity of supernatant at 550 nm by a spectrophotometer.

Calibration: In establishing a calibration line to convert biuret to protein values, a regression equation should be derived from the biuret: protein values of Bovine Serum Albumin (BSA) of 10 samples selected to represent the normal protein content range.

Variations of protein content

1) Variation of protein content among seeds within a single plant

Genetic segregation of protein content among seeds within a single plant was esti-

Table 1. Protein content of soybean seeds analyzed by Kjeldahl method and modified biuret method

Materials	Kjeldahl method			Modified biuret method		
	Replication	Mean(%)	Variance	Replication	Mean(%)	Variance
Bonminori	6	47.3	0.057	6	47.4	0.102
Norin 2	6	44.2	0.019	6	43.8	1.090
Sakagami 2	6	35.9	0.569	5	36.4	0.173
F_5 9	12	39.4	0.332	12	39.4	0.173
F_5 21	10	43.0	0.868	9	43.0	0.342
Albumin (cont.)	10	97.8	0.213	12	97.8	0.573
Mean		51.27	0.384		51.30	0.463

Table 2. Variation of protein content among F_2 seeds

Materials	No. of plants	No. of seeds	Protein content (%)	Variance
Norin 2	5	100	43.9	2.29
Sakagami 2	5	100	40.8	4.51
Tachisuzunari	1	30	43.5	1.81
Shinmejiro	1	30	39.2	5.64
Tamahikari	1	30	35.2	3.78
Sayohime	1	30	46.4	4.18
Mean				3.70
F_2 seeds				
Norin 2 × Sakagami 2	1	29	35.2	2.80
Sakagami 2 × Norin 2	2	71	39.3	4.43
Tachisuzunari × Shinmejiro	1	30	41.4	3.17
Tachisuzunari × Tamahikari	1	30	39.5	2.93
Tachisuzunari × Sayohime	1	30	44.8	2.58
Shinmejiro × Tachisuzunari	1	30	40.2	2.71
Shinmejiro × Tamahikari	1	30	41.5	4.08
Shinmejiro × Sayohime	1	30	39.7	4.51
Tamahikari × Tachisuzunari	1	30	39.3	4.51
Tamahikari × Shinmejiro	1	30	36.6	3.21
Tamahikari × Sayohime	1	30	41.4	2.25
Sayohime × Tachisuzunari	1	30	42.6	4.10
Sayohime × Shinmejiro	1	30	41.2	9.94
Sayohime × Tamahikari	1	30	40.2	3.98
Mean				3.94

mated by measuring the protein content of individual seeds of a plant, with Norin 2, Sakagami 2, Tachisuzunari, Shinmejiro, Tamahikari, and Sayohime, and with F_1 seeds and F_2 seeds produced by crossing them. As shown in Table 2, variance of protein content among individual seeds of a plant ranged from 1.81 to 5.64 with an average of 3.70, while that of F_2 seeds was not larger than that of parental lines or F_1 seeds. Therefore, it was concluded that the selection based on the protein content of a single seed in F_1 is not effective.

2) Variation of seed protein content among plants within a variety and variation of seed protein content among varieties

Seed protein content of 20 plants for each of 22 varieties was examined. Variation of

seed protein content among plants within a variety ranged from 0.86 to 2.66. Seed protein content of 22 varieties ranged from 33.5% to 45.0%, showing a normal distribution.[3]

Diallel analysis for seed protein content

Protein content of soybean seeds from F_1's, F_2's and parental populations in the diallel crosses among varieties was measured by the modified biuret method using seeds without seedcoats (mostly cotyledon part of seeds). Hayman's model[1] was adopted to this diallel analysis.

Table 3. Protein content of F_1 seeds (%)

♀ \ ♂	Tachi-suzunari	Shin-mejiro	Tama-hikari	Sayo-hime
Tachi.	41.70	42.30	42.90	43.26
Shin.	42.49	41.19	42.92	43.04
Tama.	37.88	38.70	37.78	37.89
Sayo.	42.72	45.21	44.21	43.73

Table 4. Seed-protein content of F_1 plants (%)

♀ \ ♂	Tachi-suzunari	Shin mejiro	Tama-hikari	Sayo-hime
Tachi.	43.55	41.06	40.10	43.56
Shin.	40.42	39.09	39.28	41.50
Tama.	40.73	37.51	36.64	40.71
Sayo.	42.11	41.28	41.05	44.60

The data of protein content in F_1 seeds presented in Table 3 indicate that there exists a significant difference in maternal effects, but not in paternal effects. This finding suggests that protein content in soybean seeds was determined by the genotypes of mother plants rather than by those of the seeds themselves after fertilization. On the other hand, seed protein content in F_1 plants showed no different maternal effects (Table 4). The additive genetic effects on seed protein content in F_1 plants were statistically significant, but dominance, and cytoplasmic and any interaction effects were not significant. Since the additive genetic effects alone were significant, protein content of soybean seeds was regarded as a highly heritable character.

Estimation of the number of genes

1) Genetic model

Distribution theories of segregating F_2 populations in self-fertilized populations were reported by Tan and Chang.[5] They assumed that there were n loci segregating independently and that there were only two alleles at each locus, say A_i and a_i for the ith locus, and that genotypic value of A_iA_i was d_i, that of A_ia_i was h_i, that of a_ia_i was $-d_i$. If Y_2 is a random variable for the distribution of F_2 individuals, P_1 for a parental line, P_2 for another one, and Y_1 for F_1, then

$$P_1=d_1+d_2+\ldots+d_n+E,$$
$$P_2=-d_1-d_2\ldots-d_n+E,$$
$$Y_1=h_1+h_2+\ldots+h_n+E,$$
$$Y_2=X_1+X_2+\ldots+X_n+E,$$

where $X_1, X_2, \ldots, X_n$ are discrete random variables associated with the segregation of genes, and E is a continuous random variable associated with random disturbance. To estimate number of genes, it was supposed that $d_i=d$ and $h_i=h$. Then $P_i=n\cdot d+E$, $P_2=-n\cdot d+E$, $Y_1=n\cdot h+E$, and $Y_2=(2d)Z_1+(d+h)Z_2-n\cdot d$. Where $(Z_1, Z_2)\sim$Mult$(n;\ 1/4, 1/2)$ and $E\sim N(0, \sigma^2)$. Under this genetic model, the segregation of quantitative traits and qualitative ones which are not associated with random disturbance was as shown in Fig. 1. The computer program to estimate number of genes using maximum likelihood method (ML) and moment method[6] were developed under this model.[2]

2) Monte Carlo simulation

In the estimation of the number of genes, the random disturbance, E, affected the variance of estimators. To study the effect of random disturbance, Monte Carlo simulation method was applied and many genotypic values were generated by electric computer (NEC ACOS-6). An example, in which the number of genes is one, s=0.3, d=1, h=0, is indicated in Fig. 2. In the simulation experiment with 20 replications, the number of genes was esti-

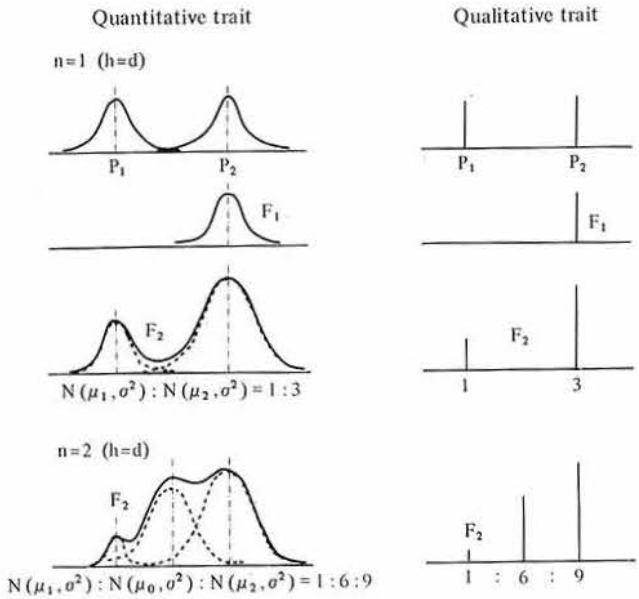

Fig. 1. Genetic models for quantitative and qualitative trait

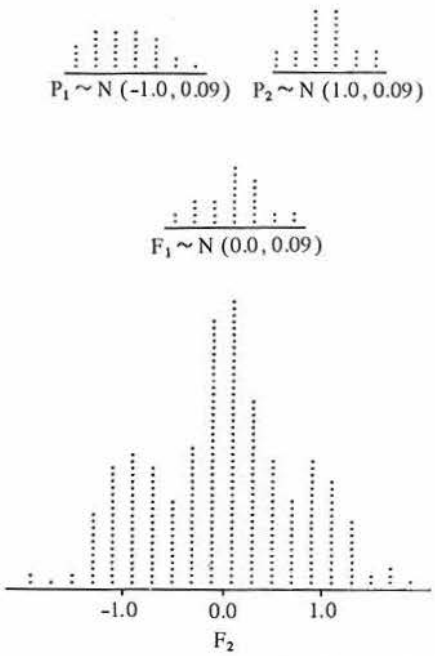

Fig. 2. Test data derived by Monte Carlo simulation

mated by the maximum likelihood method (Table 5) and moment method (Table 6) for two cases, h=d and h=0. In Table 5, the ratio of likelihood shows the significant levels. If the value of the ratio was smaller than 0.05, we concluded that the number of genes was estimated exactly. The simulation test showed that the ratio of likelihood took large values under large environmental variance and that variances of estimates took large values too (Table 5 and 6). The variance of estimates by ML method was smaller than that of moment method, indicating that the moment method was likely to express a smaller number of genes than ML method because of a wide F_2 variation.

Number of genes controlling protein content of soybean seeds

The results obtained in the progenies from six crosses among four varieties are shown in Table 7. For references, these data were analyzed by the Castle-Wright formula which infers genetic parameter separately. The estimates of number of genes by the ML method were two or three, and additive effects per each gene were decided to be 0.01–1.32(%). The ratio of likelihood (R) was calculated by the ratio between the first maximum likeli-

Table 5. Monte Carlo simulation for environmental variance in estimating number of genes by ML method

Given no. of genes	σ					
	0.1	0.3	0.5	0.7	0.9	1.1
$h=d$						
1	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.020)
2	2.00(.000)	2.00(.000)	2.00(.007)	2.05(.204)	0.95(.763)	2.10(.931)
3	3.00(.000)	3.00(.002)	3.05(.314)	3.10(.686)	3.15(.951)	3.10(.842)
4	4.00(.000)	3.95(.456)	3.90(.756)	4.15(.880)	3.95(.988)	4.25(.966)
5	5.00(.003)	5.15(.846)	5.05(.903)	4.85(.857)	4.85(.971)	5.20(.959)
$h=0$						
1	1.00(.000)	1.00(.000)	1.00(.000)	1.00(.040)	1.00(.083)	2.05(.753)
2	2.00(.000)	2.00(.062)	2.05(.522)	2.10(.773)	2.10(.907)	2.25(.955)
3	3.00(.000)	3.95(.209)	2.90(.700)	3.10(.900)	3.35(.803)	3.40(.974)
4	4.00(.808)	4.05(.895)	4.15(.974)	4.20(.998)	3.80(.963)	4.15(.924)
5	5.05(.929)	5.10(.957)	5.20(.988)	4.85(.991)	4.80(.994)	5.15(.986)

( ): ratio of likelihood h: dominance effect
d: additive effect σ^2: environmental variance

Table 6. Monte Carlo simulation for environmental variance in estimating number of genes by moment method

Given no. of genes	σ					
	0.1	0.3	0.5	0.7	0.9	1.1
$h=d$						
1 α	1.04(0.02)	1.04(0.03)	1.08(0.04)	1.10(0.15)	1.27(0.22)	1.40(0.53)
β	0.69(0.01)	0.68(0.00)	0.69(0.01)	0.71(0.04)	0.82(0.11)	1.00(0.36)
2 α	1.93(0.03)	1.98(0.10)	2.04(0.22)	2.54(0.92)	1.57(0.40)	2.49(1.73)
β	1.31(0.01)	1.28(0.04)	1.38(0.09)	1.74(0.39)	1.14(0.25)	1.51(0.64)
3 α	3.06(0.11)	3.13(0.16)	3.43(0.31)	3.30(0.41)	3.88(18.9)	3.31(10.1)
β	2.02(0.02)	2.05(0.01)	2.31(0.11)	2.31(0.34)	3.08(16.2)	2.52(8.23)
4 α	4.18(0.29)	4.00(0.60)	3.47(0.50)	5.22(10.2)	3.08(16.2)	4.97(60.1)
β	2.77(0.07)	2.74(0.23)	2.37(0.20)	4.15(10.7)	2.25(3.07)	4.25(974.)
5 α	4.78(0.31)	5.40(0.86)	5.01(3.87)	5.44(16.3)	4.45(8.66)	5.86(25.9)
β	3.18(0.09)	3.53(0.32)	3.42(1.68)	4.53(23.6)	3.26(3.83)	5.04(91.5)
$h=0$						
1 α	1.03(0.00)	1.01(0.01)	1.08(0.03)	1.26(0.23)	1.08(0.05)	1.15(0.53)
β	1.03(0.00)	1.00(0.01)	1.08(0.03)	1.23(0.20)	1.08(0.06)	1.13(0.46)
2 α	2.06(0.02)	2.13(0.08)	2.24(0.28)	2.77(2.68)	3.30(7.37)	7.57(196.)
β	2.06(0.02)	2.12(0.07)	2.20(0.26)	2.55(1.83)	2.90(4.98)	3.58(20.9)
3 α	3.18(0.14)	2.68(0.16)	2.61(0.61)	3.53(2.22)	6.21(104.)	7.19(119.)
β	3.18(0.14)	2.66(0.16)	2.57(0.54)	3.36(1.89)	4.04(18.9)	4.45(19.5)
4 α	4.07(0.20)	4.25(0.58)	5.21(3.88)	6.44(33.9)	4.46(3.71)	4.41(38.5)
β	4.07(1.00)	4.19(0.54)	4.95(3.27)	5.35(18.8)	2.24(2.34)	3.17(14.6)
5 α	5.32(0.10)	5.36(0.49)	7.21(19.7)	3.99(5.60)	3.54(4.97)	5.36(27.2)
β	5.31(0.10)	5.26(0.45)	6.44(10.6)	3.71(4.58)	3.06(3.01)	3.82(11.0)

$$\alpha: \frac{\frac{1}{2}\left[\frac{\bar{P}_1-\bar{P}_2}{2}\right]^2+\frac{1}{4}\left[\bar{F}_1-\frac{\bar{P}_1+\bar{P}_2}{2}\right]^2}{V_{F_2}-\frac{V_{P_1}+V_{P_2}+V_{F_1}}{3}} \qquad \beta: \frac{\frac{1}{2}\left[\frac{\bar{P}_1-\bar{P}_2}{2}\right]^2}{V_{F_2}-\frac{V_{P_1}+V_{P_2}}{2}}$$

( ): variance of estimates h: dominance effect
d: additive effect σ^2: environmental variance

Table 7. Estimates of number of genes for protein content

Cross	d	h	N	R	N_1	N_2	E*
Tachisuzunari × Shinmejiro	1.32	-1.78	2	0.75	1.46	2.78	0.49
Tachisuzunari × Tamahikari	0.91	-0.43	3	0.90	1.35	1.60	0.62
Tachisuzunari × Sayohime	0.44	-0.13	3	0.99	1.06	1.11	2.72
Shinmejiro × Tamahikari	0.01	0.59	1	0.89	0.01	0.01	72.28
Shinmejiro × Sayohime	0.45	0.24	3	0.99	1.33	1.45	1.21
Tamahikari × Sayohime	0.68	0.06	2	0.97	1.66	1.72	2.91

d: additive effect, h: dominance effect, N: estimated number of genes by ML method, N_1, N_2: estimated number of genes by the Castle-Wright formula

$$N_1=\frac{\frac{1}{2}\left(\frac{\bar{P}_1}{2}-\frac{\bar{P}_2}{2}\right)^2}{V_{F_2}-\hat{\sigma}^2} \qquad N_2=\frac{\frac{1}{2}\left(\frac{\bar{P}_1}{2}-\frac{\bar{P}_2}{2}\right)^2+\frac{1}{4}\left(\frac{\bar{F}_2}{2}-\frac{\bar{P}_1+\bar{P}_2}{2}\right)^2}{V_{F_2}-\hat{\sigma}^2}$$

R: ratio of likelihood

E*: environmental variance $E^*=\frac{2\hat{\sigma}}{|\bar{P}_1-\bar{P}_2|}$

hood and the second maximum likelihood estimated from several number of genes. As none of the ratios of likelihood for each cross was statistically significant, we could not definitely decide the number of genes, although the result suggests that the number of genes may be 2 or 3, as mentioned above. The estimated number of genes by the Castle-Wright formula was smaller than the estimate by ML method. To estimate the number of genes more precisely, it may be necessary to use materials with wider genetic variation in the future.

References

1) Hayman, B. I.: The analysis of variance of diallel tables. *Biometrics*, **10**, 235–244 (1954).
2) Ishige, T.: Estimation of number of genes by maximum likelihood method. *Bull. Computing Center for Research in Agriculture, Forestry and Fisheries*, A15, 231–262 (1979) [In Japanese with computer programs written by FORTRAN VI].
3) Ishige, T.: Genetic variation and biometrical analysis of seed protein content in soybean, *Glycine Max* (L.) Merrill. *Bull. Nat. Inst. Agr. Sci.*, Ser. D, **32**, 45–92 (1981) [In Japanese with English summary].
4) Pinkney, A. J.: The biuret test as applied to the estimation of wheat protein. *Cereal Chem.*, **38**, 501–506 (1961).
5) Tan, W. Y. & Chang, W. C.: Convolution approach to the genetic analysis of quantitative characters of self-fertilized populations. *Biometrics*, **10**, 1073–1090 (1954).
6) Wright, S.: Evolution and the genetics of population. I. Genetic and biometric foundation. University of Chicago Press, Chicago, 381–391 (1968).

(Received for publication, April 4, 1983)