第 24 卷 第 6 期
1998 年 11 月

作 物 学 报
ACTA AGRONOMICA SINICA

Vol. 24, No. 6
Nov., 1998

数量性状主-多基因混合遗传的 P_1、P_2、F_1、F_2 和 $F_{2:3}$ 联合分析方法*

王建康** 盖钧镒***

(国家大豆改良中心 南京农业大学大豆研究所，江苏南京，210095)

提 要 本文提出利用亲本 P_1 和 P_2、杂种 F_1、F_2 和 $F_{2:3}$ 五世代联合分析数量性状主基因和多基因遗传的统计方法，共建立可供选择的单基因遗传、多基因遗传以及一个主基因＋多基因混合遗传三类 11 个遗传模型；*AIC* 信息准则用于选择最适遗传模型，通过适合性检验对所选择的遗传模型做进一步检验；以 D 类模型为例，给出参数估计 EM 过程的一般步骤。以邳县天鹅蛋(P_1)和 1138－2(P_2)杂交组合为例，分析了大豆抗豆秆黑潜蝇性状的遗传，发现该性状符合主基因＋多基因混合遗传模式，主基因的加显性效应分别为－1.86 和－1.64，F_2 世代主基因的遗传率为 43.84%，$F_{2:3}$家系世代主基因的遗传率为 88.59%，F_2 群体的抗感分界线为 $11 \leqslant x \leqslant 12$，$F_{2:3}$群体的抗感分界线为 $10.6 \leqslant x \leqslant 11.0$。

关键词 数量性状；主基因-多基因混合遗传模型；极大似然估计；EM 算法；多世代联合分析

数量性状基因在遗传效应上有很大差异，效应较大的可以表现出主效基因的特征，效应较小者表现为微效多基因，这种遗传现象称为主基因-多基因混合遗传(major gene and polygene mixed inheritance)[1~5]。王建康和盖钧镒(1996)[4]系统研究了数量性状主基因-多基因混合遗传模型的鉴别和遗传参数估计问题，本文报道利用 P_1、F_1、P_2、F_2、$F_{2:3}$(F_2 单株衍生的 F_3 家系)世代的联合分析方法，此法将适用于不易得到回交种子的作物，各群体中个体的表型值分别用 X_{1i}，X_{2i}，X_{3i}，X_{4i}和 X_{5i}表示，样本量分别用 n_1，n_2，n_3，n_4 和 n_5 表示，模型中的一些基本假定是：所考虑的性状是二倍体核遗传，不存在细胞质或母体效应，主基因和多基因之间不存在互作和连锁(对多基因之间有无互作没有限制)，所有配子有相同的生活力，分离世代中个体的多基因效应是正态随机变量。所考虑的遗传模型包括单基因遗传、多基因遗传以及一个主基因＋多基因的混合遗传。

1 遗传模型的构造

1.1 一个主基因位点的遗传模型(以 A 表示)

假定某一性状在两个亲本中的差异是由一个基因位点的分离所引起的，亲本的基因型用 AA 和 aa 表示，设 σ^2 为环境方差，P_1、F_1 和 P_2 群体的分布为：

$$X_{1i} \sim N(\mu_1, \sigma^2),\ X_{2i} \sim N(\mu_2, \sigma^2),\ X_{3i} \sim N(\mu_3, \sigma^2)$$

* 国家自然科学基金项目。 ** 现在河南省农业科学院科学实验中心工作，郑州 450002
*** 联系作者，E-mall:sri@njau.edu.cn
收稿日期：1997-05-06，收到修改稿日期：1998-01-23

F_2 和 $F_{2,3}$为三个正态分布的混合，表示为：

$$X_{4i} \sim (1/4)N(\mu_1, \sigma^2) + (1/2)N(\mu_2, \sigma^2) + (1/4)N(\mu_3, \sigma^2)$$

$$X_{5i} \sim (1/4)N(\mu_{51}, \sigma^2_{51}) + (1/2)N(\mu_{52}, \sigma^2_{52}) + (1/4)N(\mu_{53}, \sigma^2_{53})$$

其中，

$\mu_{51} = \mu_1$，$\mu_{53} = \mu_3$，分别为 AA 家系和 aa 家系平均数的平均，

$\mu_{52} = (1/4)\mu_1 + (1/2)\mu_2 + (1/4)\mu_3$，为 Aa 家系平均数的平均，

$\sigma^2_{51} = \sigma^2_{53} = \sigma^2/n$，为 AA 家系和 aa 家系平均数的方差，$n$ 为家系内个体数，

$\sigma^2_{52} = (1/8)(\mu_1 - \mu_3)^2 + (1/16)(\mu_1 - 2\mu_2 + \mu_3)^2 + \sigma^2/n$ 为 *Aa* 家系平均数的方差；Aa 家系平均数的分布只是一个近似的正态分布，Aa 家系中的个体应服从一个正态混合分布，根据概率论的大数定理，当家系内个体数较多(经计算机模拟，当家系内个体数 $n>5$)时，家系平均数构成的分布就很接近于正态分布。由此构造样本似然函数，似然函数中所包含的参数有 μ_1，μ_2，μ_3 和 σ^2 共四个，参数的估计及最大似然函数值的确定采用 EM 算法，以 P_1、F_1 和 P_2 群体的样本平均数和三个同质群体样本方差的加权平均作为 EM 算法中的初始值。

群体平均数(m)、基因的加性效应(d)和显性效应(h)分别为：

$$m = (1/2)(\mu_1 + \mu_3), \quad d = (1/2)(\mu_1 - \mu_3), \quad h = \mu_2 - (1/2)(\mu_1 + \mu_3)$$

称这时的遗传模型为 A－1，在一定的约束条件下还可建立一个基因位点的加性模型(A－2)、完全显性模型(A－3)和负向完全显性模型(A－4)。

1.2 多基因遗传模型(以 C 表示)

由于多基因的遗传效应近似，数量较多，而且多基因间又可能存在连锁和互作，性状的表型在分离世代也表现为一正态或近似正态分布，在所考虑的群体中，亲本和 F_1 的分布同上，F_2 和 $F_{2,3}$家系平均数的分布为：

$$X_{4i} \sim N(\mu_4, \sigma^2_4), \quad X_{5i} \sim N(\mu_5, \sigma^2_5)。$$

如果多基因服从加-显性模型，加性和显性效应分别用$[d]$和$[h]$表示[10]，m 表示群体平均数，那么，

$$\mu_1 = m + [d], \quad \mu_2 = m + [h], \quad \mu_3 = m - [d],$$

$$\mu_4 = m + (1/2)[h], \qquad \mu_5 = m + (1/4)[h]$$

则对 μ_1、μ_2、μ_3、μ_4 和 μ_5 所施加的约束条件为：

$$\mu_1 + 2\mu_2 + \mu_3 - 4\mu_4 = 0 \text{ 和}$$

$$\mu_2 - 3\mu_4 + 2\mu_5 = 0$$

称这时的模型为 C-1，群体平均数(m)、多基因的加性效应($[d]$)和显性效应($[h]$)分别表示为：

$$m = (1/2)(\mu_1 + \mu_3), \quad [d] = (1/2)(\mu_1 - \mu_3),$$

$$[h] = \mu_2 - (1/2)(\mu_1 + \mu_3)。$$

1.3 一个主基因＋多基因混合遗传模型(以 D 代表)

P_1、F_1 和 P_2 群体的分布为单一正态分布，F_2 和 F_3 是三个正态分布的混合，但是，混合分布中所包含的成分分布与 P_1、F_1 和 P_2 的分布不同，用 μ_{41}、μ_{42}、μ_{43}和 μ_{51}、μ_{52}、μ_{53}分别表示 F_2 和 $F_{2,3}$家系中各个主基因基因型的均值，F_2 群体中的成分分布有相同的方差，用 σ^2_4 表示，

$F_{2,3}$家系群体中各个成分分布的方差分别用 σ^2_{51}、σ^2_{52}和 σ^2_{53}表示，则：

$$X_{4i} \sim (1/4)N(\mu_{41},\sigma^2_4) + (1/2)N(\mu_{42},\sigma^2_4) + (1/4)N(\mu_{43},\sigma^2_4),$$

$$X_{5i} \sim (1/4)N(\mu_{51},\sigma^2_{51}) + (1/2)N(\mu_{52},\sigma^2_{52}) + (1/4)N(\mu_{53},\sigma^2_{53})。$$

如果多基因符合加显性遗传模型，沿用上面的符号，则有以下的关系式：

$\mu_1 = m + d + [d]$，$\mu_2 = m + h + [h]$，$\mu_3 = m - d - [d]$，

$\mu_{41} = m + d + (1/2)[h]$，$\mu_{42} = m + h + (1/2)[h]$，$\mu_{43} = m - d + (1/2)[h]$，

$\mu_{51} = m + d + (1/4)[h]$，$\mu_{52} = m + (1/2)h + (1/4)[h]$，$\mu_{53} = m - d + (1/4)[h]$，

$\sigma^2_4 = (1/2)D + (1/4)H + \sigma^2$，$D$、$H$ 分别为多基因的加性方差和显性方差，

$\sigma^2_{51} = \sigma^2_{53} = (1/2)D + (1/16)H + [(1/4)D + (1/8)H + \sigma^2]/n$，

$\sigma^2_{52} = (1/2)d^2 + (1/4)h^2 + \sigma^2_{51}$

对一阶统计量应施加的约束条件为：

$$\mu_1 + 2\mu_2 + \mu_3 - \mu_{41} - 2\mu_{42} - \mu_{43} = 0$$

$$\mu_2 - 2\mu_{41} - \mu_{42} + 2\mu_{51} = 0$$

$$\mu_1 + \mu_3 + 2\mu_{42} - 4\mu_{52} = 0$$

$$2\mu_1 + 3\mu_2 + 2\mu_3 - 2\mu_{41} - 3\mu_{42} - 2\mu_{53} = 0$$

遗传参数的估计为：

$$m = (1/2)(\mu_1 + \mu_3),$$

$$d = \mu_{41} - (1/2)\mu_1 - (1/2)\mu_3 - \mu_2 + \mu_{42},$$

$$h = 2\mu_{42} - \mu_2 - (1/2)\mu_1 - (1/2)\mu_3,$$

$$[d] = \mu_1 - m - d,$$

$$[h] = 2\mu_2 - 2\mu_{42}$$

称这时的模型为 D-1，在一定的约束条件下，还可建立加性主基因和加显性多基因模型(D-2)、完全显性主基因和加显性多基因模型(D-3)以及负向完全显性主基因和加显性多基因模型(D-4)。

2　合适模型的选择与检验

2.1　*AIC*(Akaike's information criterion)准则

Akaike(1977)[6]认为根据熵最大原理，可将统计推断视为观测值概率分布的估计，通过比较估计分布与真实分布之间的适合程度进行推断，适合度用概率熵来度量，概率熵定义为两个分布的期望对数似然值之差，用函数 $B(f; g)$表示，其中 f 为真实分布，g 为估计分布，一个优良的推断应使熵取得最大值，即熵最大原理。*AIC* 值与熵之间有以下关系：

$$AIC = -2B(f;g) + C = -2L(\hat{\varphi}) + 2N$$

C 是只与真实分布有关的常数，$\hat{\varphi}$ 是模型中参数的极大似然估计，$L(\cdot)$是对数似然函数，N 为模型中独立参数的个数。熵最大原理在此处的应用是选择 *AIC* 值最小的模型为最适模型。由于似然比统计量在混合分布的假设测验中不具有一般情形下的渐进性质，也就是不服从渐进卡方分布，其真实渐进分布通常难以确定，因此本文考虑应用 *AIC* 准则首先在不同类遗传模型中选择最适合的一类。Akaike(1977)还证明了在一定条件下，*AIC* 准则与似然比检验是等价的，但 *AIC* 可用于无法应用似然比检验的场合，如混合分布中的一些假设检验问题。

2.2 似然比检验(LRT,likelihood ratio test)

如果一个模型是另一模型的特殊情形，则可应用似然比检验来比较这个模型是否显著地优于另一模型。假定从某一模型(用 H_1 表示)到它的特例(用 H_0 表示)包含 γ 个独立的限制方程，那么似然比统计量 λ 渐进服从于自由度为 γ 的 χ^2 分布，即：

$$\lambda = 2[L(\hat{\varphi}_1) - L(\hat{\varphi}_0)] \sim \chi^2(\gamma)$$

$\hat{\varphi}_1$ 和 $\hat{\varphi}_0$ 分别是模型 H_1 和 H_0 下参数的极大似然估计，$L(\cdot)$是对数似然函数。在利用 *AIC* 值选择到某一类模型后，可以利用似然比检验比较同一类模型中不同模型的优劣，例如在利用 *AIC* 值选择到 D 类模型后，可用似然比检验比较 D 与 D−1 的优劣，如果二者没有显著差异，则选择包含较少参数、较简单的 D−1 模型为最适遗传模型，反之，如果二者有显著差异，则不能选择 D−1 而只选择模型 D。

2.3 适合性检验

通过比较期望分布与观测分布的拟合程度[9]，可以确定所选择的合适模型是否和观测数据相一致，一致的话，则说明可用该模型解释这一组数据，否则，可能所考察的遗传模型都不合适，应该考虑采用其它模型分析这组数据。本文考虑采用均匀性检验、Smirnov 检验和 Kolmogorov 检验确定期望分布与样本分布间的适合性。

设 $F(x)$为概率分布函数，X_1，X_2，…，X_n 为样本观测值，则 $F(X_i)$是[0,1]上的均匀分布，利用以下三个自由度为 1 的 χ^2 统计量，可以检验 $F(X_i)$是否是[0,1]上的均匀分布：

$$U_1^2 = 12[\sum F(X_i) - n/2]^2/n \sim \chi^2(1)$$

$$U_2^2 = (45/4)[\sum F(X_i)^2 - n/3]^2/n \sim \chi^2(1)$$

$$U_3^2 = 180[\sum (F(X_i) - 0.5)^2 - n/12]^2/n \sim \chi^2(1)$$

设 $F_n^*(x)$为经验分布函数，$X_{(1)}$，$X_{(2)}$，…，$X_{(n)}$为顺序统计量，$F_0(x)$为期望分布(通过模型得到的总体的分布)，Smirnov 在 1936 年提出利用统计量

$${}_nW^2 = n\int_{-\infty}^{+\infty}[F_n(X) - F_0(X)]^2 dF_0(X) = \frac{1}{12n} + \sum[F(X_{(r)}) - \frac{r - 0.5}{n}]^2$$

做适合性检验，并证明了${}_nW^2$ 的极限分布，Marshall 在 1958 年证明了${}_nW^2$ 达到它的极限分布的速度非常之快，当 $n=3$ 时，${}_nW^2$ 已很接近它的极限分布。

Kolmogorov 在 1933 年提出适合性检验的另一统计量 D_n：

$$D_n = \mathrm{Sup}|F_n^*(x) - F_0(x)|$$

以上五个检验适合性的统计量都是分布自由的，即检验统计量的分布与总体的分布类型无关。在一定的遗传模型下，可以获得各个世代群体的期望分布，然后利用以上统计量做适合性检验，利用 *AIC* 选择的最适模型应该通过所有的适合性检验。

3 极大似然估计的计算

对于 D 类模型，多世代联合估计的 EM(expectation and maximization)[2,5,11]算法中，E 步骤完全数据对数似然函数的期望函数为：

$$L_c(\Phi) = \sum \log f(X_{1i};\mu_1,\sigma^2) + \sum \log f(X_{2i};\mu_2,\sigma^2) + \sum \log f(X_{3i};\mu_3,\sigma^2)$$
$$+ \sum[W_{4i1}\log f(X_{4i};\mu_{41},\sigma_4^2) + W_{4i2}\log f(X_{4i};\mu_{42},\sigma_4^2) + W_{4i3}\log f(X_{4i};\mu_{43},\sigma_4^2)]$$

$$+\sum[W_{5i1}\log f(X_{5i};\mu_{51},\sigma_{51}^2)+W_{5i2}\log f(X_{5i};\mu_{52},\sigma_{52}^2)+W_{5i3}\log f(X_{5i};\mu_{53},\sigma_{53}^2)]$$

式中省去求和符上的求和范围，$f(X;\mu,\sigma^2)$表示正态分布 $N(\mu,\sigma^2)$的密度函数，W_{4i1}、W_{4i2}、W_{4i3}、W_{5i1}、W_{5i2}和 W_{5i3}是样本属于混合总体中各成分分布的后验概率。M 步骤在计算 $L_c(\Phi)$ 的极大值和极大值点时，D 为全效应模型，无约束条件，可直接对 $L_c(\Phi)$求偏导数得到新一轮的参数估计值；在 D－1 等非全效应模型时，还各有一组约束条件，可利用 Lagrange 函数确定 M 步骤中 $L_c(\Phi)$的条件极值，其推演过程此处省略。

按上列似然函数，利用 EM 算法进行分布参数估计的过程为：

① 根据样本观测值选择适当参数的初始值；

② 计算混合群体中样本观测值的后验概率 W_{4i1}、W_{4i2}、W_{4i3}、W_{5i1}、W_{5i2}和 W_{5i3}，从而得到完全分类数据的似然函数 $L_c(\Phi)$(E 步骤)；

③ 对 $L_c(\Phi)$求极值或条件极值得到各成分分布的均值和方差的估计(M 步骤)；

④ 将③得到的估计值作为下一轮 EM 迭代的初始值，重复以上过程，直到预定的精度为止。

4　实例分析

4.1　数据描述

杂交组合的亲本为邳县天鹅旦(P_1)和 1138－2(P_2)，经鉴定表明前者是感豆秆蝇材料，后者是抗豆秆蝇材料，所调查的性状为单株虫量。F_1 代平均表型与 P_2 亲本接近，F_2 代表现为偏态分布，峰不明显，从 127 棵 F_2 单株获得 $F_{2,3}$家系世代，每个家系观测 5 株，各个世代的次数分布如表 1。

表 1　大豆邳县天鹅旦×1138－2 各世代茎秆虫量的分布

Table 1　Frequency distributions of the number of beanfly in stem in various generations of a cross between Pixiantianedan(P_1) and 1138－2(P_2)

	4	5	6	7	8	9	10	11	12	13	14	15	16	n
P_1						1	2	7	2	3	3	0	2	20
F_1		1	3	3	3	6	1	3						20
P_2	1	2	2	7	4	3	1							20
F_2	2	11	17	29	33	34	30	18	11	10	4	1		200
$F_{2,3}$				4	17	50	18	4	14	19	1			127

4.2　统计分析结果与讨论

表 2 给出了不同模型的极大似然函数值、*AIC* 值及极大似然估计，表 3 给出 *AIC* 值较小的部分模型的适合性检验的结果，从 C-模型的适合性检验可以看出所有检验统计量在 P_1、F_1 和 P_2 群体中均未达到显著水平，说明这三个群体可以视为单一正态分布，因此认为所研究的性状在纯合群体中符合正态分布，而在分离群体中则表现为正态分布的混合，且不需考虑数据转换问题。在所有模型中，D-模型有最小的 *AIC* 值(表 2)，适合性检验中所有统计量都没达到显著水平(表 3)，因此 D-模型应是最适合模型，图 1 给出各世代的次数分布图及由 D 模型得到的拟合曲线图。利用以下关系可以进一步估计主基因的加性效应 d 和显性效应 h：

$$\mu_1=m_1+d=7.20,\ \mu_2=m_2+h=8.25,\ \mu_3=m_3-d=12.15,$$

$$\mu_{41}=m_4+d=8.56,\ \mu_{42}=m_4+h=9.23,\ \mu_{43}=m_4-d=12.39,$$

$\mu_{51} = m_5 + d = 8.54$, $\mu_{52} = m_5 + (1/2)h = 8.73$, $\mu_{53} = m_5 - d = 12.14$

表 2 不同模型中成分分布参数的估计值

Table 2 Maximum likelihood estimates of component parameters in various genetic models

模型 Model	对数似然值 Logarithm likelihood	*AIC*	μ_1	μ_2	μ_3	μ_{41}	μ_{42}	μ_{43}	μ_{51}	μ_{52}	μ_{53}	σ^2	σ_4^2	σ_{51}^2	σ_{52}^2
A−1	−796.45	1600.89	8.08	8.64	12.18	—	—	—	—	—	—	2.10	—	—	—
A−2	−824.73	1655.45	7.97	10.06	12.16	—	—	—	—	—	—	2.30	—	—	—
A−3	−797.98	1601.97	8.30	8.30	12.16		—	—	—	—	—	2.14	—	—	
A−4	−868.01	1742.02	8.13	10.90	10.90	—	—	—	—	—		3.50	—	—	—
C	−813.39	1642.78	7.20	8.25	12.15	9.86	—	—	9.60	—		2.94	5.11	2.98	—
C−1	−818.87	1649.75	7.44	9.12	12.39	9.52			9.72	—	—	3.18	5.23	2.99	—
D	−775.63	1575.26	7.20	8.25	12.15	8.56	9.23	12.39	8.54	8.73	12.14	2.94	2.87	0.34	0.89
D−1	−785.30	1586.60	7.30	9.32	12.25	9.03	8.22	12.19	8.48	9.00	12.17	3.28	2.61	0.31	0.98
D−2	−797.28	1608.56	7.07	9.17	12.02	8.07	9.36	10.65	8.02	9.45	12.08	3.19	4.53	0.36	0.57
D−3	−786.53	1587.07	7.45	9.15	12.40	8.57	8.57	12.43	8.29	9.25	12.14	3.20	2.65	0.33	1.02
D−4	−819.22	1652.44	7.44	9.12	12.39	8.76	9.77	9.77	9.08	9.84	10.09	3.18	5.05	2.62	2.67

注：—表示模型中无此参数 Note: —represents no such parameters in this model

表 3 部分模型的适合性检验(括号内为概率值)

Table 3 Tests for goodness of fit(probability in paratheses)

模型 Model	群体 Population	统计量 Statistic				
		U_1	U_2	U_3	$_nW^2$	D_n
D	P_1	0.01(0.93)	0.01(0.91)	0.67(0.41)	0.17(>0.10)	0.20(>0.10)
	F_1	0.00(0.98)	0.02(0.88)	0.23(0.63)	0.13(>0.10)	0.17(>0.10)
	P_2	0.12(0.73)	0.02(0.90)	0.66(0.42)	0.15(>0.10)	0.25(>0.10)
	F_2	0.01(0.94)	0.03(0.87)	0.14(0.70)	0.32(>0.10)	0.10(>0.10)
	$F_{2,3}$	0.15(0.70)	0.14(0.71)	0.00(0.99)	0.11(>0.10)	0.06(>0.10)
D−1	P_1	0.03(0.86)	0.20(0.65)	1.22(0.27)	0.14(>0.10)	0.18(>0.10)
	F_1	6.13 *	4.77 *	0.74(0.39)	0.38 *	0.37 *
	P_2	0.29(0.59)	0.14(0.70)	0.32(0.57)	0.14(>0.10)	0.25(>0.05)
	F_2	8.17 * *	7.39 * *	0.04(0.84)	1.26 * *	0.18 * *
	$F_{2,3}$	0.44(0.51)	0.22(0.64)	0.50(0.48)	0.13(>0.10)	0.11(>0.05)
D−3	P_1	0.33(0.56)	0.66(0.42)	1.05(0.31)	0.12(>0.10)	0.20(>0.10)
	F_1	4.44 *	3.47(0.06)	0.49(0.48)	0.27(>0.10)	0.33 *
	P_2	0.79(0.37)	0.43(0.51)	0.66(0.42)	0.15(>0.10)	0.28(>0.05)
	F_2	4.97 *	4.23 *	0.14(0.71)	0.96 * *	0.16 * *
	$F_{2,3}$	1.19(0.28)	0.77(0.38)	0.52(0.47)	0.28(>0.10)	0.13 *
D−2	P_1	0.18(0.67)	0.03(0.86)	0.93(0.34)	0.24(>0.10)	0.23(>0.10)
	F_1	4.65 *	3.62(0.06)	0.55(0.46)	0.28(>0.10)	0.34 *
	P_2	0.00(0.99)	0.01(0.92)	0.20(0.65)	0.17(>0.10)	0.22(>0.10)
	F_2	6.95 * *	6.54 * *	0.00(0.98)	1.01 * *	0.15 * *
	$F_{2,3}$	2.61(0.11)	1.40(0.24)	2.30(0.13)	0.63 *	0.18 * *
C	P_1	0.01(0.93)	0.01(0.91)	0.67(0.41)	0.17(>0.10)	0.20(>0.10)
	F_1	0.00(0.98)	0.02(0.88)	0.23(0.63)	0.13(>0.10)	0.17(>0.10)
	P_2	0.12(0.73)	0.02(0.90)	0.66(0.42)	0.15(>0.10)	0.25(>0.10)
	F_2	0.16(0.70)	0.07(0.79)	0.24(0.63)	0.34(>0.10)	0.11 *
	$F_{2,3}$	1.02(0.31)	0.45(0.50)	1.53(0.21)	0.98 * *	0.20 * *

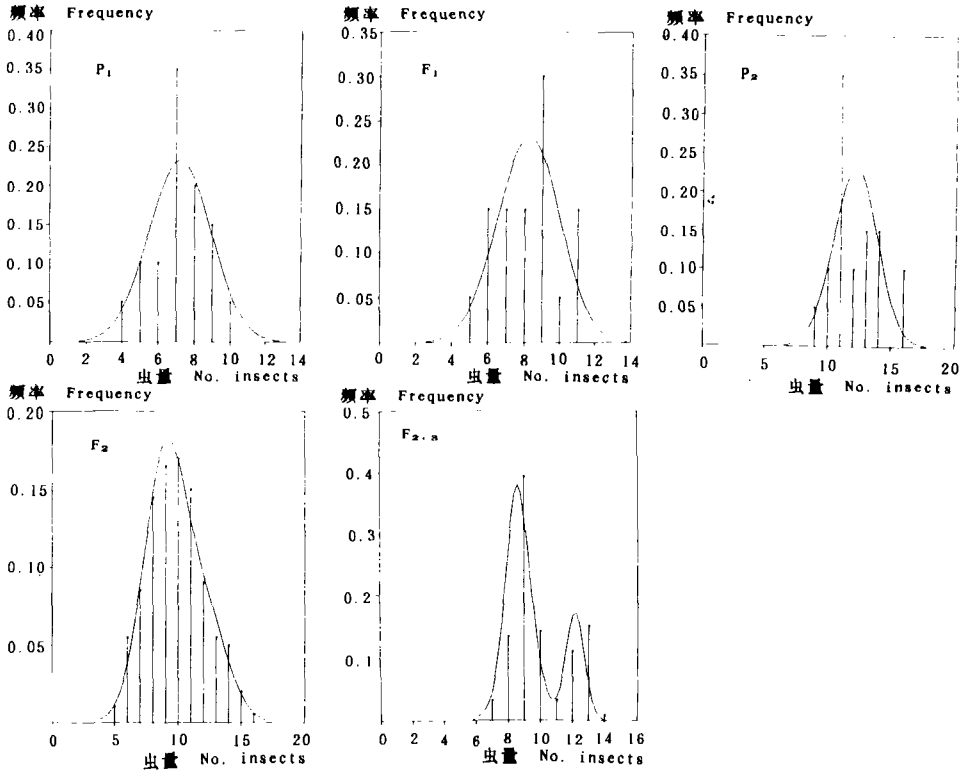

图 1　大豆邳县天鹅蛋×1138-2 各世代豆秆黑潜蝇全株虫量的次数分布及其适合模型的拟合曲线

Fig. 1　Frequency distributions and fitted curves of the number of insects/plant in various generations derived from Pixian tianedan×1138−2

利用最小二乘算法可以得到：$m_1=9.06$，$m_2=9.89$，$m_3=10.29$，$m_4=10.61$，$m_5=10.08$，$d=-1.86$，$h=-1.64$；除 D-模型外，模型 D-1 和模型 D-3 的 *AIC* 值也较小，在适合性检验中只有少数统计量达到显著水平，D-1 与 D-3 的似然比检验差异不显著，若认为模型 D-3 是合适模型，则可得到 $m=9.93$，$d=h=-1.93$，$[d]=-0.56$，$[h]=1.14$。此例中，模型 D 与 D-1 的似然比检验有显著差异，以下的结果均来自 D 模型。

在 D 模型中，如果不考虑主基因与多基因之间的互作，那么 $\sigma_4^2=2.87$ 为 F_2 群体中多基因与环境之和的估计，$\sigma_{51}^2=\sigma_{53}^2=0.34$ 为 $F_{2,3}$ 群体中多基因与环境之和的估计，F_2 和 $F_{2,3}$ 群体总的表型变异分别为 5.11 和 2.98(见表 2 中 C 模型方差的估计)，这样可以得到 F_2 和 $F_{2,3}$ 群体中主基因的遗传方差分别为 2.24 和 2.64，因此得到主基因遗传率的估计分别为43.84% 和 88.59%。表 2 中 σ^2 为 P_1、F_1 和 P_2 群体的合并方差，一般可以作为环境方差的估计。在 D 模型中 $\sigma^2=2.94$，此处若将 σ^2 视为 F_2 群体的环境变异、σ^2/n 视为 $F_{2,3}$ 群体的环境变异，则在这两个分离群体中多基因和环境变异之和将小于环境的变异，得到多基因的变异为 0，而模型检验又存在多基因变异，其原因可能有二，一是不宜将 P_1、F_1 和 P_2 群体方差的估计作为分离群体中环境变异部分的估计，二是可能存在有主基因与多基因间以及多基因间的互作。

在模型分析的基础上按 Bayes 方法计算分离世代 F_2 植株及其 $F_{2,3}$ 家系各种主基因型的

后验概率，以0.05为小概率事件对F_2植株和$F_{2,3}$家系的主基因型进行分类，其结果成表4。由表4可以看出，F_2全株虫量为5头/株的个体可能归属于AA及Aa，两个成分分布相重叠，

表4 由F_2植株及$F_{2,3}$家系茎秆虫量的后验概率推测的F_2个体基因型归属

Table 4 The estimated genotype of F_2 individuals according to thein respective posterior probability in F_2 and $F_{2,3}$

虫量 Number of insects	次数 Frequency	后验概率 Posterior probability			F_2基因型归属 Estimated genotype of F_2
		AA	Aa	aa	
F_2					
5	2	0.55	0.45	0.00	AA+Aa
6～9	90	0.50－0.31	0.50－0.64	0.00－0.04	Aa+AA
10	34	0.24	0.63	0.13	Aa+AA+aa
11	30	0.16	0.52	0.32	Aa+aa+AA
12	18	0.08	0.32	0.60	aa+Aa+AA
13～14	21	0.03－0.01	0.15－0.06	0.82－0.93	aa+Aa
15～16	5	0.00	0.02－0.01	0.98－0.99	aa
$F_{2,3}$					
6.8～10.0	89	0.07－0.46－0.07	0.93－0.54－0.92	0.00	Aa+AA
10.4	2	0.02	0.94	0.04	Aa
10.6	1	0.00	0.75	0.25	Aa+aa
11.0～11.4	7	0.00	0.32－0.05	0.68－0.95	aa+Aa
11.6～13.2	28	0.00	0.02－0.00	0.98－1.00	aa

归于AA的可能性比Aa稍大；6～9头/株的个体可能归属于Aa及AA，归于Aa的可能性比AA稍大；10头/株的个体归属于三种基因型的可能性并存，三个成分分布相重叠，但归于Aa的可能性最大，归于AA的可能性次之，归于aa的可能性最小；11头/株的个体归属于三种基因型的可能性并存，但归于Aa的可能性最大，归于aa的可能性次之，归于AA的可能性最小；12头/株的个体归属于三种基因型的可能性并存，但归于aa的可能性最大，归于Aa的可能性次之，归于AA的可能性最小；13～14头/株的个体归属于aa及Aa，但归于aa的可能性比Aa大；15～16头/株的个体归属于aa基因型。$F_{2,3}$家系平均数为6.8～10.0头/株时有89个家系，其上代可能为Aa或AA，但Aa的可能性较大些；10.4头/株时有2个家系，其上代可能为Aa；10.6头/株时有1个家系，其上代可能为Aa或aa，但Aa的可能性大些；11.0～11.4头/株时有7个家系，其上代可能为aa或Aa，但aa的可能性大些；11.6～13.2头/株时有28个家系，其上代可能为aa。

参 考 文 献

1 姜长鉴、徐辰武、惠大丰、邵元建，1995，作物学报，21(5)，632～636

2 姜长鉴、莫惠栋，1995，作物学报，21(6)，641～648

3 姜长鉴、刘学锋，1995，遗传学报，22(1)，59～64

4 王建康、盖钧镒(导师)，1996，数量性状主基因和多基因混合遗传的鉴别和遗传参数的估计(博士学位论文)，南京农业大学

5 王建康、盖钧镒，1997，遗传学报，24(5)，432～440

6 Akaike，H.，1977，Applications of Statistics，P. R. Krishnaiah(ed.)，North Holland Publishing Company，Amsterdam，27～41

7 Dempster, A. P., N. M. Laird and D. B. Robin, 1977, J. R. Statist. Soc. B., 39, 1～38

8 Elston, R. C., 1984, Genetics, 108, 733～744

9 Kendall, M. G. and A. Stuart, 1979, The Advanced Theory of Statistics, Volume 2 Inference and Relationship, Charles Griffin & Company Limited

10 Mather, K. and J. L. Jinks, 1982, Biometrical Genetics 2nd edn., Chapman and Hall

11 McLachlan, G. J., 1988, Mixture Models: Inference and Applications to Clustering, Marcel Dekker, Inc

Identification of Major Gene and Polygene Mixed Inheritance Model of Quantitative Traits by Using Joint Analysis of P_1, F_1, P_2, F_2 and $F_{2,3}$ Generations

Wang Jiankang Gai Junyi

(National Center of Soyhean Improvement Soybean Research Institute of Nanjing Agricultural University, Nanjing, Jiangsu 210095)

Abstract The statistical method for identification of major gene and polygene mixed inheritance of quantitative traits by using joint analysis of P_1, F_1, P_2, F_2 and $F_{2,3}$ generations was proposed in this paper. Eleven genetic models were established, which could be classified into three types: single major gene inheritance, polygene inheritance and one major gene plus polygene mixed inheritance. The most suitable genetic model could be selected by using Akaike's Information Criterion and could be further tested by using a set of tests of fitness. The EM algorithm was derived for computing maximum likelihood estimates in D－type model, as an example. Inheritance of resistance of soybean to the Agromyzid fly was analyzed by using the above method. Major gene and polygene mixed model was the most fitted genetic model for this trait. The additive and dominance effects of the major gene were estimated as －1.86 and －1.64, respectively. The major gene heritability values of F_2 and $F_{2,3}$ were 43.84% and 88.59%, respectively. The F_2 individuals were classified into appropriate major gene genotypes according to the posterior probability values of F_2 plants and $F_{2,3}$ familits. The critical values of F_2 and $F_{2,3}$ to distinguish resistant and susceptable ones were drawn as $11 \leqslant x \leqslant 12$ and $10.6 \leqslant x \leqslant 11.0$.

Key words Quantitative trait; Major gene and polygene mixed inheritance; Maximum likelihood estimate; EM algorithm; Joint analysis of multiple generations