文章编号:0253-9861(1999)02-0327-09

运用混合线性模型定位复杂数量性状基因的方法

朱 军
(浙江大学 生物数学研究中心,浙江 杭州 310029)

摘 要:运用混合线性模型分析原理,提出了复杂数量性状基因定位的方法,可以分析QTL复杂的遗传效应及QTL×环境互作效应. 采用混合线性模型随机效应的无偏预测方法,可以预测基因型值和基因型×环境互作效应值,再运用区间作图法或复合区间作图法间接分析QTL的加性、显性遗传主效应及其与环境的互作效应,还能定位在特定发育阶段表达的QTL. 基于混合模型的复合区间作图法(MCIM法)可以分析多环境的遗传实验资料,直接分析包括上位性效应的遗传主效应及其与环境的互作效应. Markov链蒙特卡罗(MCMC)分析方法可用于推断QTL的统计特征.

关键词:数量性状基因定位;QTL定位方法;上位性效应;基因型×环境互作;发育性状基因定位
中图分类号:O242.29;Q348 **文献标识码**:A

作物产量、品质和抗逆性等重要的农艺性状大多为数量性状基因座位(Quantitative Trait Loci,简称QTL)所控制. 采用数量性状基因的定位方法如区间作图法[1]、复合区间作图法[2]、标记回归法[3]等已能将众多的数量基因定位在相应的分子标记连锁图上.

许多重要的农艺性状都是复杂的数量性状. 它们除了受简单的加性、显性效应控制以外,还可能受上位性效应、母体遗传效应及其与环境的互作效应等控制. 国内外现有的QTL分析方法大都运用方差分析、简单回归或多元回归等统计方法,只能分析简单的数量遗传模型,尚不能分析复杂的遗传现象及在不同时空下表达的基因效应,QTL分析方法的滞后,使作物遗传育种工作者不能有效地对一些重要的农艺性状进行精细定位及遗传效应分析.

70年代以来发展的混合线性模型分析方法,可以同时分析固定效应和多项随机效应,已成为研究数量性状遗传的重要统计方法[4]. 近年来,我们在发展混合线性模型统计方法、创立分析复杂数量性状的遗传模型等领域开展了系统的基础理论研究[5],为经典数量遗传分析提供了一些稳健的遗传模型和无偏的统计方法.

1 回归模型的QTL定位方法

1.1 加性-显性效应的QTL定位

近十年来,在植物中已先后发表了番茄、玉米、水稻等30多种分子标记连锁图谱[6]. 这些连锁图谱可用于对控制质量性状或数量性状的目标基因进行定位,并检测基因与分子标记的连锁关系.

利用方差分析的方法,可以检测出数量性状与单一分子标记之间的相互关联. 但是这种方法

收稿日期:1998-11-16
基金项目:国家自然科学基金重大项目资助(39893350)
作者简介:朱 军(1949—),男,上海市人,博士,浙江大学教授.

无法对数量性状基因的位置及其效应进行有效的定位和估计,因而不能满足 QTL 精细作图的研究需要. Lander 和 Botstein(1989)提出了基于最大似然分析原理的区间作图(Interval Mapping,简称 IM)QTL 分析方法[1],可用于推断相邻分子标记 (M_{i-} 和 M_{i+}) 之间的某个 QTL (Q_i) 的位置及其遗传效应. 如果假定所分析的数量性状只受一对基因控制,并且不存在基因型与环境的互作效应,个体 j 的表现型值(y_j) 可以用以下简单回归模型表示:

$$y_j = b_0 + b^* X_j^* + \varepsilon_j,$$

式中 b_0 是群体均值,b^* 是 QTL 遗传效应,X_j^* 是遗传效应的系数,ε_j 是随机机误.

数量性状一般都受多基因控制. 当搜索某个 Q_i 时,其它QTL 的影响可能会干扰区间作图分析的结果. Zeng (1994) 提出的复合区间作图(Composite Interval Mapping,简称 CIM) 分析方法[2],在回归模型中包括了与其它 QTL 紧密连锁的分子标记

$$y_j = b_0 + b^* X_j^* + \sum_f b_f X_{fj} + \varepsilon_j,$$

式中 b_0 是群体均值,b^* 是 QTL 遗传效应,X_j^* 是遗传效应的系数,b_f 是第 f 个分子标记基因型的效应,X_{fj} 是个体 j 的 M_f 分子标记效应的系数. CIM 方法可以在一定程度上消除背景遗传变异的干扰.

区间作图法和复合区间作图法都基于回归模型分析原理,并建立在一个简单的遗传假设上:数量性状的表现型变异受遗传效应(固定效应) 和残差机误(随机效应) 控制. 因此这两种方法都不能分析基因型 × 环境的互作效应.

1.2 遗传主效应及 GE 互作效应的 QTL 定位

在多环境下实施的遗传实验,其遗传群体的表现型变异除了受遗传效应(G) 和残差机误(ε) 控制以外,还会受到环境效应(E) 和基因型 × 环境互作效应(GE) 的控制. 第 j 种基因型在环境 h 中第 k 次重复中的表现型值可以用以下线性模型表示:

$$y_{hjk} = \mu + G_j + E_h + GE_{hj} + \varepsilon_{hjk} \tag{1}$$

模型(1) 中,y_{hjk} 是第 j 种基因型在第 h 个环境内第 k 次重复中的观察值;μ 是群体均值,固定效应;G_j 是基因型效应,$G_j \sim N(0, \sigma_G^2)$;$E_h$ 是环境效应,$E_h \sim N(0,\sigma_E^2)$;GE_{hj} 是基因型 × 环境互作效应,$GE_{hj} \sim N(0,\sigma_{GE}^2)$;$\varepsilon_{hjk}$ 是残差效应,$\varepsilon_{hjk} \sim N(0, \sigma_\varepsilon^2)$. 以上公式可改写为矩阵形式的混合线性模型,

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{U}_G\boldsymbol{e}_G + \boldsymbol{U}_E\boldsymbol{e}_E + \boldsymbol{U}_{GE}\boldsymbol{e}_{GE} + \boldsymbol{e}_\varepsilon = \mathbf{1}\mu + \sum_{u=1}^{4}\boldsymbol{U}_u\boldsymbol{e}_u \sim N(\mathbf{1}\mu, \boldsymbol{V} = \sum_{u=1}^{4}\sigma_u^2\boldsymbol{U}_u\boldsymbol{U}_u^T) \tag{2}$$

式(2) 中 $\boldsymbol{y}$ 是表现型值向量;μ 是固定效应的群体均值;$\boldsymbol{I}$ 是系数为 1 的向量;$\boldsymbol{e}_1 = \boldsymbol{e}_G \sim N(0,\sigma_G^2\boldsymbol{I})$ 是基因型效应的随机向量;$\boldsymbol{e}_2 = \boldsymbol{e}_E \sim N(0,\sigma_E^2\boldsymbol{I})$ 是环境效应的随机向量;$\boldsymbol{e}_3 = \boldsymbol{e}_{GE} \sim N(0,\sigma_{GE}^2\boldsymbol{I})$ 是基因型 × 环境互作效应的随机向量;$\boldsymbol{e}_4 = \boldsymbol{e}_\varepsilon \sim N(0,\sigma_\varepsilon^2\boldsymbol{I})$ 是残差效应随机向量.

采用随机效应的预测方法(如 BLUP 法、LUP 法、AUP 法)[7] 可以获得基因型效应值 G_j 及基因型 × 环境互作效应值 GE_{hj} 的无偏预测,然后可进一步计算 $y_{j(G)} = \mu + G_j$ 和 $y_{hj(GE)} = \mu + E_h + GE_{hj}$ 的预测值.

如果采用区间作图法或复合区间作图法,分析基因型效应值的预测数据 $\hat{y}_{j(G)}$,可以定位具有遗传主效应(加性主效应、显性主效应) 的 QTL.

区间作图法:$\hat{y}_{j(G)} = b_{0(G)} + b_{(G)}^* X_j^* + \varepsilon_{j(G)}$;

复合区间作图法:$\hat{y}_{j(G)} = b_{0(G)} + b_{(G)}^* X_j^* + \sum_f b_{f(G)} X_{fj} + \varepsilon_{j(G)}$.

以上两个方程式中,$b_{0(G)}$ 是群体均值,$b_{(G)}^*$ 是QTL 的遗传主效应,$b_{f(G)}$ 是第 f 个分子标记基因型的主

效应.

如果采用区间作图法或复合区间作图法,分析基因型与第 h 个环境的互作效应的预测数据 $\hat{y}_{hj(GE)}$,则可定位具有基因型 × 环境互作效应(加性 × 环境互作效应、显性 × 环境互作效应)的 QTL. 即

区间作图法:$\hat{y}_{hj(GE)} = b_{h0(GE)} + b^{*}_{h(GE)}X^{*}_{hj} + \varepsilon_{hj(GE)}$;

复合区间作图法:$\hat{y}_{hj(GE)} = b_{h0(GE)} + b^{*}_{h(GE)}X^{*}_{hj} + \sum_{f} b_{hf(GE)}X_{fhj} + \varepsilon_{hj(GE)}$.

以上两个方程式中,$b_{h0(GE)}$ 是在第 h 个环境中的群体均值,$b^{*}_{h(GE)}$ 是 QTL × 环境 h 的互作效应,$b_{hf(GE)}$ 是第 f 个分子标记基因型 × 环境 h 的互作效应.

1.3　发育数量遗传的 QTL 定位

数量性状的最终表现是生物体不同发育阶段基因表达的综合结果. 研究基因在特定时刻的表达及其对数量性状的影响,是发育数量遗传学的一项重要研究内容.

环境 h 中第 k 次重复的第 j 种基因型在 t 时刻($t = 1,2,\cdots$)的表现型值可以用以下线性模型表示:

$$y_{hjk(t)} = \mu_{(t)} + G_{j(t)} + E_{h(t)} + GE_{hj(t)} + \varepsilon_{hjk(t)}. \tag{3}$$

模型(3)可以改写为矩阵形式的混合线性模型,

$$\boldsymbol{y}_{(t)} = \boldsymbol{I}\mu_{(t)} + \boldsymbol{U}_G\boldsymbol{e}_{G(t)} + \boldsymbol{U}_E\boldsymbol{e}_{E(t)} + \boldsymbol{U}_{GE}\boldsymbol{e}_{GE(t)} + \boldsymbol{e}_{\varepsilon(t)} = \boldsymbol{I}\mu_{(t)} + \sum_{u=1}^{4}\boldsymbol{U}_u\boldsymbol{e}_{u(t)} \tag{4}$$

$$\sim N(\boldsymbol{I}\mu_{(t)}, \boldsymbol{V}_{(t)} = \sum_{u=1}^{4}\sigma^2_{u(t)}\boldsymbol{U}_u\boldsymbol{U}_u^T).$$

模型(4)中 $\boldsymbol{y}_{(t)}$ 是在 t 时刻的表现型值向量;$\mu_{(t)}$ 是 t 时刻的群体均值;$\boldsymbol{I}$ 是系数为 1 的向量;$\boldsymbol{e}_{1(t)} = \boldsymbol{e}_{G(t)} \sim N(0, \sigma^2_{G(t)}\boldsymbol{I})$ 是 t 时刻的基因型主效应的随机向量;$\boldsymbol{e}_{2(t)} = \boldsymbol{e}_{E(t)} \sim N(0, \sigma^2_{E(t)}\boldsymbol{I})$ 是 t 时刻的环境效应的随机向量;$\boldsymbol{e}_{3(t)} = \boldsymbol{e}_{GE(t)} \sim N(0, \sigma^2_{GE(t)}\boldsymbol{I})$ 是 t 时刻的基因型 × 环境互作效应的随机向量;$\boldsymbol{e}_{4(t)} = \boldsymbol{e}_{\varepsilon(t)} \sim N(0, \sigma^2_{\varepsilon(t)}\boldsymbol{I})$ 是 t 时刻的残差效应随机向量.

采用随机效应的预测方法[7]可以获得 $E_{h(t)}$,$G_{j(t)}$ 及 $GE_{hj(t)}$ 的无偏预测,然后进一步计算 $y_{j(G)(t)} = \mu_{(t)} + G_{j(t)}$ 和 $y_{hj(GE)(t)} = \mu_{(t)} + E_{h(t)} + GE_{hj(t)}$ 的预测值. 用区间作图法或复合区间作图法分析 t 时刻基因型主效应的预测值数据 $\hat{y}_{j(G)(t)}$,定位的 QTL 具有初始时刻至 t 时刻($0 \rightarrow t$)的遗传主效应. 采用区间作图法或复合区间作图法,分析基因型与第 h 个环境在 t 时刻的互作效应的预测值数据 $\hat{y}_{hj(GE)(t)}$,可分析在发育阶段($0 \rightarrow t$)具有基因型 × 环境互作效应的 QTL.

给定 $t-1$ 时刻的表现型值,t 时刻的条件表现型值是条件随机变量:

$$y_{hjk(t|t-1)} = \boldsymbol{I}\mu_{(t|t-1)} + G_{j(t|t-1)} + E_{h(t|t-1)} + GE_{hj(t|t-1)} + \varepsilon_{hjk(t|t-1)}, \tag{5}$$

并且具有条件变量分布

$$\boldsymbol{y}_{(t|t-1)} = \boldsymbol{I}\mu_{(t|t-1)} + \boldsymbol{U}_G\boldsymbol{e}_{G(t|t-1)} + \boldsymbol{U}_E\boldsymbol{e}_{E(t|t-1)} + \boldsymbol{U}_{GE}\boldsymbol{e}_{GE(t|t-1)} + \boldsymbol{e}_{\varepsilon(t|t-1)} =$$

$$\boldsymbol{I}\mu_{(t|t-1)} + \sum_{u=1}^{4}\boldsymbol{U}_u\boldsymbol{e}_{u(t|t-1)} \sim N(\boldsymbol{I}\mu_{(t|t-1)}, \boldsymbol{V}_{(t|t-1)} = \sum_{u=1}^{4}\sigma^2_{u(t|t-1)}\boldsymbol{U}_u\boldsymbol{U}'_u. \tag{6}$$

采用条件随机效应的预测方法[8],可以获得条件遗传效应 $G_{j(t|t-1)}$ 及条件互作效应 $GE_{hj(t|t-1)}$ 的无偏预测.

采用区间作图法或复合区间作图法分析条件变量 $y_{j(G)(t|t-1)} = \mu_{(t|t-1)} + G_{j(t|t-1)}$ 的预测值,所定位的 QTL 可以推断特定发育阶段($t-1 \rightarrow t$)的净遗传主效应. 采用区间作图法或复合区间作图法分析条件变量 $y_{hj(GE)(t|t-1)} = \mu_{(t|t-1)} + E_{h(t|t-1)} + GE_{hj(t|t-1)}$ 的预测值,则可以分析($t-1 \rightarrow t$)阶段

的 QTL 及净 QTL × 环境互作效应.

2 混合线性模型的 QTL 定位方法

2.1 包括 QTL × 环境互作的遗传模型

如果对多环境下实施的遗传试验资料进行 QTL 定位分析,基因型 j 在环境 h 中的表现型值可用以下混合线性模型表示,

$$y_{hj}=\mu+a\chi_{A_j}+d\chi_{D_j}+u_{E_{hj}}e_{E_h}+u_{AE_{hj}}e_{AE_h}+u_{DE_{hj}}e_{DE_h}+\sum_f u\,M_{fj}e\,M_f+\sum_l u\,ME_{lhj}e\,ME_{lh}+\varepsilon_{hj} \tag{7}$$

式(7)中,μ 是群体均值;a 和 d 分别是 QTL 的加性主效应和显性主效应,固定效应;χ_{A_j} 和 χ_{D_j} 是遗传主效应的系数;e_{E_h} 是环境 h 的随机效应,其系数为 $u_{E_{hj}}$;e_{AE_h} 是加性 × 环境互作随机效应,其系数为 $u_{AE_{hj}}$;e_{DE_h} 是显性 × 环境互作随机效应,其系数为 $u_{DE_{hj}}$;e_{M_f} 是标记基因型 f 的主效应,$u_{M_{fj}}$ 是标记主效应的系数;$e_{ME_{eh}}$ 是标记 l × 环境 h 的互作随机效应,其系数为 $u_{ME_{ehj}}$;ε_{hj} 是随机的残差效应.

模型(7)可用以下混合线性模型的矩阵形式表示:

$$\boldsymbol{y}=\boldsymbol{Xb}+\boldsymbol{U}_E\boldsymbol{e}_E+\boldsymbol{U}_{AE}\boldsymbol{e}_{AE}+\boldsymbol{U}_{DE}\boldsymbol{e}_{DE}+\boldsymbol{U}_M\boldsymbol{e}_M+\boldsymbol{U}_{ME}\boldsymbol{e}_{ME}+\boldsymbol{e}_\varepsilon=\boldsymbol{Xb}+\sum_{u=1}^{6}\boldsymbol{U}_u\boldsymbol{e}_u$$

$$\sim N(\boldsymbol{Xb},\boldsymbol{V}=\sum_{u=1}^{6}\sigma_u^2\boldsymbol{U}_u\boldsymbol{R}_u\boldsymbol{U}_u^T). \tag{8}$$

模型(8)中,$\boldsymbol{y}$ 是表现型值向量;$\boldsymbol{b}$ 是固定效应的参数向量;$\boldsymbol{X}$ 是固定效应的系数矩阵;$\boldsymbol{e}_1=\boldsymbol{e}_E\sim N(0,\sigma_E^2\boldsymbol{I})$ 是环境效应的随机向量;$\boldsymbol{e}_2=\boldsymbol{e}_{AE}\sim N(0,\sigma_{AE}^2\boldsymbol{I})$ 是加性 × 环境互作效应的随机向量;$\boldsymbol{e}_3=\boldsymbol{e}_{DE}\sim N(0,\sigma_{DE}^2\boldsymbol{I})$ 是显性 × 环境互作效应的随机向量;$\boldsymbol{e}_4=\boldsymbol{e}_M\sim N(0,\sigma_M^2\boldsymbol{R}_M)$ 是分子标记基因型主效应的随机向量;$\boldsymbol{e}_5=\boldsymbol{e}_{ME}\sim N(0,\sigma_{ME}^2\boldsymbol{R}_{ME})$ 是分子标记 × 环境互作效应的随机向量;$\boldsymbol{e}_6=\boldsymbol{e}_\varepsilon\sim N(0,\sigma_\varepsilon^2\boldsymbol{I})$ 是残差效应随机向量.

采用以下 2.2 一节中介绍的混合线性模型分析方法,可以直接定位 QTL,并估算其遗传主效应(加性 a、显性 d),还可无偏预测 QTL 与环境的互作效应(加性 × 环境互作 e_{AE}、显性 × 环境互作 e_{DE}).

数量遗传研究结果表明,非等位基因之间的互作效应(上位性效应)是不可忽略的遗传效应组成部分[9,10]. 目前,通常采用方差分析方法分析分子标记之间的互作效应,然后推断 QTL 的上位性效应[10]. 这种分析方法不能精细定位具有上位性效应的 QTL,也无法估算上位性效应值.

利用 DH 群体或 RIL 群体,可以分析加性和加 × 加上位性效应. DH 群体或 RIL 群体的个体 j 在环境 h 中的表现型值可用以下混合线性模型表示:

$$y_{hk}=\mu+a_i\chi_{A_{ik}}+a_j\chi_{A_{jk}}+aa_{ij}\chi_{AA_{ijk}}+u_{E_{hk}}e_{E_h}+u_{A_iE_{hk}}e_{A_iE_h}+u_{A_jE_{hk}}e_{A_jE_h}+u_{AA_{ij}E_{hk}}e_{AA_{ij}E_h}+$$

$$\sum_f u_{M_{fk}}e_{M_f}+\sum_l u_{MM_{lk}}e_{MM_l}+\sum_p u_{ME_{phk}}e_{ME_{ph}}+\sum_q u_{MME_{qhk}}e_{MME_{qh}}+\varepsilon_{hk}. \tag{9}$$

模型(9)中,μ 是群体均值;a_i 和 a_j 分别是两个基因位点 Q_i 和 Q_j 的加性主效应,aa_{ij} 是 Q_i 和 Q_j 的加 × 加上位性主效应;$\chi_{A_{ik}}$,$\chi_{A_{jk}}$ 和 $\chi_{AA_{ijk}}$ 分别是加性和上位性主效应的系数;e_{E_h} 是环境 h 的随机效应,其系数为 $u_{E_{hk}}$;$e_{A_iE_h}$ 是位点 Q_i 的加性 × 环境互作随机效应,其系数为 $u_{A_iE_{hk}}$;$e_{A_jE_h}$ 是位点 Q_j 的加性 × 环境互作随机效应,其系数为 $u_{A_jE_{hk}}$;$e_{AA_{ij}E_h}$ 是双位点 Q_i 和 Q_j 的上位性 × 环境互作随机效应,

其系数为 $u_{AA_{ij}E_{hk}}$; e_{M_f} 是标记基因型的随机效应,其系数是 $u_{M_{fk}}$; e_{MM_l} 是互作分子标记的随机效应,其系数是 $u_{MM_{lk}}$; $e_{ME_{ph}}$ 是分子标记 × 环境 h 的互作效应,其系数为 $u_{ME_{hpk}}$; $e_{MME_{qh}}$ 是互作分子标记 × 环境 h 的互作效应,其系数为 $u_{MME_{qhk}}$; ε_{hk} 是随机的残差效应.

基于模型(9)的所有个体,其表现型值可以用以下混合线性模型的矩阵形式表示:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{U}_E\boldsymbol{e}_E + \boldsymbol{U}_{A_iE}\boldsymbol{e}_{A_iE} + \boldsymbol{U}_{A_iE}\boldsymbol{e}_{A_jE} + \boldsymbol{U}_{AAE}\boldsymbol{e}_{AAE} + \boldsymbol{U}_M\boldsymbol{e}_M + \boldsymbol{U}_{MM}\boldsymbol{e}_{MM} +$$

$$\boldsymbol{U}_{ME}\boldsymbol{e}_{ME} + \boldsymbol{U}_{MME}\boldsymbol{e}_{MME} + e_\varepsilon = \boldsymbol{Xb} + \sum_{u=1}^{9}\boldsymbol{U}_u\boldsymbol{e}_u \sim N(\boldsymbol{Xb}, \boldsymbol{V} = \sum_{u=1}^{9}\sigma_u^2\boldsymbol{U}_u\boldsymbol{R}_u\boldsymbol{U}_u^T). \tag{10}$$

模型(10)中,$\boldsymbol{y}$ 是表现型值向量; $\boldsymbol{b}$ 是固定效应的参数向量; $\boldsymbol{X}$ 是固定效应的系数矩阵; $\boldsymbol{e}_1 = \boldsymbol{e}_E \sim N(0,\sigma_E^2\boldsymbol{I})$ 是环境效应的随机向量; $\boldsymbol{e}_2 = \boldsymbol{e}_{A_iE} \sim N(0,\sigma_{A_iE}^2\boldsymbol{I})$ 是位点 Q_i 的加性 × 环境互作效应的随机向量; $\boldsymbol{e}_3 = \boldsymbol{e}_{A_jE} \sim N(0,\sigma_{A_jE}^2\boldsymbol{I})$ 是位点 Q_j 的加性 × 环境互作效应的随机向量; $\boldsymbol{e}_4 = \boldsymbol{e}_{AAE} \sim N(0,\sigma_{AAE}^2\boldsymbol{R}_{AAE})$ 是双位点 Q_1 和 Q_2 的加加上位性 × 环境互作效应的随机向量; $\boldsymbol{e}_5 = \boldsymbol{e}_M \sim N(0,\sigma_M^2\boldsymbol{R}_M)$ 是分子标记主效应的随机向量; $\boldsymbol{e}_6 = \boldsymbol{e}_{MM} \sim N(0,\sigma_{MM}^2\boldsymbol{R}_{MM})$ 是互作分子标记效应的随机向量; $\boldsymbol{e}_7 = \boldsymbol{e}_{ME} \sim N(0,\sigma_{ME}^2\boldsymbol{R}_{ME})$ 是分子标记 × 环境互作效应的随机向量; $\boldsymbol{e}_8 = \boldsymbol{e}_{MME} \sim N(0,\sigma_{MME}^2\boldsymbol{R}_{MME})$ 是互作分子标记 × 环境互作效应的随机向量; $\boldsymbol{e}_9 = \boldsymbol{e}_\varepsilon \sim N(0,\sigma_\varepsilon^2\boldsymbol{I})$ 是残差效应随机向量.

采用下面 2.2 节介绍的混合线性模型的分析方法,可以定位具有上位性的 QTL,并估算其遗传主效应(加性 a_i、加性 a_j 及其加性与加性的上位性 aa_{ij}),还能预测 QTL 与环境的互作效应(加性 × 环境互作 e_{AE}、加加上位性 × 环境互作 e_{AAE}).

2.2　混合线性模型的定位方法

在运用分子标记定位数量性状基因的分析中,确定 QTL 的位置和基因的遗传效应是分析的主要目的,为固定效应所采用的分子标记可以视为该基因组所具有的分子标记的一个随机样本,记可以作为随机效应。基于这个原理建立的定位基因的遗传模型,可用混合线性模型的通式表示为:

$$\boldsymbol{y} = \boldsymbol{Xb} + \sum_{u=1}^{m}\boldsymbol{U}_u\boldsymbol{e}_u \tag{11}$$

该模型是具有均值 $\boldsymbol{Xb}$ 和方差 $\boldsymbol{V} = \sum_{u=1}^{m}\sigma_u^2\boldsymbol{U}_u\boldsymbol{R}_u\boldsymbol{U}_u^{\mathrm{T}}$ 的混合多变量正态分布,$\boldsymbol{b}$ 是固定效应向量($p \times 1$),$\boldsymbol{e}_u \sim N(0,\sigma_u^2\boldsymbol{R}_u)(u = 1,2,\cdots,m-1)$ 是随机效应向量($q_u \times 1$),$\boldsymbol{e}_m \sim N(0,\sigma_m^2\boldsymbol{I})$ 是残差随机效应向量($n \times 1$). 均值和方差的似然函数 (L) 为

$$L(\boldsymbol{b},\boldsymbol{V}) = (2\pi)^{-\frac{n}{2}}|\boldsymbol{V}|^{-\frac{1}{2}}\exp[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{Xb})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{Xb})]. \tag{12}$$

似然函数的对数(l) 为

$$l(\boldsymbol{b},\boldsymbol{V}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{Xb})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{Xb})]. \tag{13}$$

如果固定效应 $\boldsymbol{b}$ 是可估计的,其最大似然估计值可以由下式算得

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{y}. \tag{14}$$

并具有抽样方差

$$\mathrm{Var}(\hat{\boldsymbol{b}}) = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}. \tag{15}$$

如果固定效应的系数矩阵 $\boldsymbol{X}$ 奇异,且 $\boldsymbol{b}$ 不可无偏估计时,仍能获得固定效应的可估算函数 $\boldsymbol{c}'\boldsymbol{b}$ 的最大似然估计值

$$\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{b}} = \boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{+}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{y},$$

其抽样方差矩阵为

$$\mathrm{Var}(\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{b}}) = \boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{+}\boldsymbol{c}.$$

当我们在相邻的两个分子标记 M_{i-} 和 M_{i+} 之间搜索 QTL,可以设置分子标记 M_{i-} 与 QTL 之间交换值 $r_{M_{i-}Q}$ 的先验值 $\hat{r}_{M_{i-}Q}$,然后计算似然比(LR) 统计量。

$$LR = 2l_1(\hat{\boldsymbol{b}},\hat{\boldsymbol{V}},r_{M_{i-}Q}) - 2l_0(\hat{\boldsymbol{b}},\hat{\boldsymbol{V}},r_{M_{i-}Q} = 0.5); \tag{16}$$

$\boldsymbol{V}$ 中的方差分量可用其无偏估计值替代:

$$\hat{\boldsymbol{V}} = \sum_u \hat{\sigma}^2_{U_u}\boldsymbol{R}_u\boldsymbol{U}_u^{\mathrm{T}}$$

对于无效假设 $H_0: r_{M_{i-}Q} = 0.5$ (QTL 与分子标记 M_{i-} 相互独立) 及其相应的备择假设 H_1: $r_{M_{i-}Q} < 0.5$ (QTL 与分子标记 M_{i-} 相连锁),可以采用似然比(LR) 统计量检验 QTL 是否与分子标记 M_{i-} 相连锁. LR 近似地具有 χ^2 分布. 卡方分布的自由度,对于模型(4) 为 $\mathrm{d}f = 4$,对于模型(8) 为 $\mathrm{d}f = 6$.

当 $LR \geqslant \chi^2$ 时拒绝无效假设,可推断 QTL 与相邻的分子标记 M_{i-} 和 M_{i+} 连锁. QTL 的遗传距离可由 $\hat{r}_{M_{i-}Q}$ 计算,QTL 遗传主效应则由 $\hat{\boldsymbol{b}}$ 算得. 可以采用 t 测验对遗传效应进行检验. 如果设置的无效假设与备择假设分别是

$$\mathrm{H}_0: \boldsymbol{c}'\boldsymbol{b} = m \quad vs. \quad \mathrm{H}_1: \boldsymbol{c}'\boldsymbol{b} \neq m.$$

当统计量 $|\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{b}} - m/\sqrt{\boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})\boldsymbol{c}}| \geqslant t_\alpha$ 时,拒绝无效假设,接受备择假设.

混合线性模型(8) 和(10) 中的 QTL × 环境互作效应(加性 × 环境互作向量 $\boldsymbol{e}_{AE}$,显性 × 环境互作向量 $\boldsymbol{e}_{AE}$,上位性 × 环境互作向量 $\boldsymbol{e}_{AAE}$),可采用 BLUP、LUP 或 AUP 等随机效应预测方法预测[7].

也可采用 Henderson (1984) 提出的以下方法[11],同时估计 QTL 主效应及预测 QTL × 环境互作效应:

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{U}_1 & \boldsymbol{X}^T\boldsymbol{U}_2 & \cdots \\ \boldsymbol{U}_1^T\boldsymbol{X} & \boldsymbol{U}_1^T\boldsymbol{U}_1 + \boldsymbol{R}_1^{-1}\dfrac{\sigma_m^2}{\sigma_1^2} & \boldsymbol{U}_1^T\boldsymbol{U}_2 & \cdots \\ \boldsymbol{U}_2^T\boldsymbol{X} & \boldsymbol{U}_2^T\boldsymbol{U}_1 & \boldsymbol{U}_2^T\boldsymbol{U}_2 + \boldsymbol{R}_1^{-1}\dfrac{\sigma_m^2}{\sigma_2^2} & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{e}}_1 \\ \hat{\boldsymbol{e}}_2 \\ \hat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{U}_1^T\boldsymbol{y} \\ \boldsymbol{U}_2^T\boldsymbol{y} \\ \hat{\boldsymbol{u}} \end{bmatrix} \tag{17}$$

方程式(17) 可以简单表示为 $\boldsymbol{W}\boldsymbol{\beta} = \boldsymbol{d}$.

2.3 Markov 链蒙特卡罗分析方法

近年来,基于 Bayesian 推断原理的 Markov 链蒙特卡罗(MCMC) 分析方法已应用于数量性状的基因定位[12]. 对于混合线性模型(11) 中未知的参数,Bayesian 先验分布或先验值可确定为[13]:

$f(\boldsymbol{b}) \propto$ 某常数;

$\boldsymbol{e}_u|\boldsymbol{R}_u,\sigma_u^2 \sim N(0,\sigma_u^2\boldsymbol{R}_u), u = 1,2,\cdots,m-1;$

$f(\sigma_u^2|\lambda_u,\alpha_u) \propto (\sigma_u^2)^{-\lambda_u/(2-1)}\exp(-\dfrac{1}{2}\lambda_u\alpha_u/\sigma_u^2), u = 1,2,\cdots,m-1;$

$f(\sigma_m^2|\lambda_m,\alpha_m) \propto (\sigma_m^2)^{-\lambda_m/(2-1)}\exp(-\dfrac{1}{2}\lambda_m\alpha_m/\sigma_m^2).$

在以上式中,λ_m 或 λ_u 是确信度(degree of believe),而 α_m 或 α_u 是方差的先验值.

设定

$\boldsymbol{\beta}^{\mathrm{T}}=[\boldsymbol{b}^{\mathrm{T}},\boldsymbol{e}_1^{\mathrm{T}},\boldsymbol{e}_2^{\mathrm{T}},\cdots,\boldsymbol{e}_{m-1}^{\mathrm{T}}]=[\beta_1,\beta_2,\cdots,\beta_N]$;

$\boldsymbol{\beta}_{-i}^{\mathrm{T}}=[\beta_1,\beta_2,\cdots,\beta_{i-1},\beta_{i+1},\cdots,\beta_N]$.

式中,$N=p+\sum_{u=1}^{m-1}q_u=p+q$.

$\boldsymbol{v}^{\mathrm{T}}=[\sigma_1^2,\sigma_2^2,\cdots,\sigma_{m-1}^2]$;

$\boldsymbol{v}_{-u}^{\mathrm{T}}=[\sigma_1^2,\sigma_2^2,\cdots,\sigma_{u-1}^2,\sigma_{u+1}^2,\cdots,\sigma_{m-1}^2]$;

$\boldsymbol{\alpha}^{\mathrm{T}}=[\alpha_1,\alpha_2,\cdots,\alpha_{m-1},\alpha_m]$;

$\boldsymbol{\lambda}^{\mathrm{T}}=[\lambda_1,\lambda_2,\cdots,\lambda_{m-1},\lambda_m]$.

因此,所求参数的联合后验密度函数是[14]:

$$f(\boldsymbol{\beta},\boldsymbol{v},\sigma_m^2|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\lambda})\propto$$

$$(\sigma_m^2)^{-(n+\lambda_m+2)/2}\exp\{-\frac{1}{2\sigma_m^2}[(\boldsymbol{y}-\boldsymbol{Xb}-\sum_{u=1}^{m-1}\boldsymbol{U}_u\boldsymbol{e}_u)^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{Xb}-\sum_{u=1}^{m-1}\boldsymbol{U}_u\boldsymbol{e}_u)+\lambda_m\alpha_m]\}\times$$

$$\prod_{u=1}^{m-1}\left\{(\sigma_u^2)^{-(q_u+\lambda_u+2)/2}\exp[-\frac{1}{2\sigma_u^2}(\boldsymbol{e}_u^{\mathrm{T}}\boldsymbol{R}_u^{-1}\boldsymbol{e}_u+\lambda_u\alpha_u)]\right\}. \tag{18}$$

对于未知参数$(\boldsymbol{\beta},\boldsymbol{v},\sigma_m^2)$可以通过其相应的边缘密度作推断. $\boldsymbol{\beta}$的每一个位置参数(β_i)的条件后验密度函数是正态的,具有均值$\tilde{\beta}_i$和方差$\tilde{s}_i^2$:

$$\beta_i|\boldsymbol{y},\boldsymbol{\beta}_{-i},\boldsymbol{v},\sigma_m^2,\boldsymbol{\alpha},\boldsymbol{\lambda}\sim N(\tilde{\beta}_i,\tilde{s}_i^2),i=1,2,\cdots,N. \tag{19}$$

式(19)中,$\tilde{\beta}_i=(d_i-\sum_{j=1,j\neq i}^{N}w_{ij}\beta_j)/w_{ii}$,$\tilde{s}_i^2=\sigma_m^2/w_{ii}$,$w_{ij}$是方程组(17)的矩阵$\boldsymbol{W}$中第$i$行、第$j$列的元素,$d_i$是方程组(17)中$\boldsymbol{d}$向量的第$i$项元素.

残差方差σ_m^2的条件后验密度函数是标量化的逆卡方:

$$\sigma_m^2|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{v},\boldsymbol{\alpha},\boldsymbol{\lambda}\sim\tilde{\lambda}_m\tilde{\alpha}_m\chi^{-2}_{(df=\tilde{\lambda}_m)}. \tag{20}$$

式(20)中$\tilde{\lambda}_m=n+\lambda_m$,$\tilde{\alpha}_m=[(\boldsymbol{y}-\boldsymbol{Xb}-\sum_{u=1}^{m-1}\boldsymbol{U}_u\boldsymbol{e}_u)^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{Xb}-\sum_{u=1}^{m-1}\boldsymbol{U}_u\boldsymbol{e}_u)+\lambda_m\alpha_m]/\tilde{\lambda}_m$.

第u项方差分量σ_u^2的条件后验密度函数也是标量化的逆卡方:

$$\sigma_u^2|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{v}_{-u},\sigma_m^2,\boldsymbol{\alpha},\boldsymbol{\lambda}\sim\tilde{\lambda}_u\tilde{\alpha}_u\chi^{-2}_{(df=\tilde{\lambda}_u)}. \tag{21}$$

式(21)中,$\tilde{\lambda}_u=q_u+\lambda_u$,$\tilde{\alpha}_u=\boldsymbol{e}_u^{\mathrm{T}}\boldsymbol{R}_u^{-1}\boldsymbol{e}_u+\lambda_u\alpha_u)/\tilde{\lambda}_u$.

通常可以设定所有的先验值为$\lambda_u=-2$,$\alpha_u=0$ $(u=1,2,\cdots,m)$. 采用以下 Gibbs 抽样技术,可以对未知参数$(\boldsymbol{\beta},\boldsymbol{v},\sigma_m^2)$的边缘分布作 Bayesian 推断:

[i] 以混合线性模型的无偏估计值或预测值[7],作为参数$\boldsymbol{\beta},\boldsymbol{v}$和$\sigma_m^2$的初始值;

[ii] 根据式(19)的概率分布生成并更新$\beta_i,i=1,2,\cdots,N$;

[iii] 根据式(20)的概率分布生成并更新σ_m^2;

[iv] 根据式(21)的概率分布生成并更新$\sigma_u^2,u=1,2,\cdots,m-1$;

重复步骤[ii]～[iv]k次$(k\to\infty)$,便可获得的 Markov 链长为k的平衡分布. 这些k个样本是从联合后验分布(18)生成的随机样本. 第i个样本

$$\{\boldsymbol{\beta}_i,\boldsymbol{v}_i\text{和}(\sigma_m^2)_i\},i=1,2,\cdots,k, \tag{22}$$

是一个$N+m$的向量,其中的每一个元素都是根据相应的边缘分布生成的随机样本,即称为供统计推断的 Gibbs 样本.

如果$x_i(i=1,2,\cdots,k)$是式(22)的一个成分,后验分布$F(\chi)$的特征可以由下式估算:

$$\hat{c}=\frac{1}{k}\sum_{i=1}^{k}g(\chi_i). \tag{23}$$

式(23) 中,$g(\chi_i)$ 可以是 $F(\chi)$ 的任何特征,如均值或方差.

采用Markov链蒙特卡罗(MCMC)分析方法,还可以对遗传参数的各种函数作Bayesian推断.

3 讨 论

区间作图法和复合区间作图法都基于回归模型分析原理,并建立在一个简单的遗传假设上:数量性状的表现型变异受遗传效应(固定效应)和残差机误(随机效应)控制,不存在基因型 × 环境的互作效应。可以用简单的数量遗传模型表示为:

$$y = \mu + G + \varepsilon. \tag{24}$$

式(24) 中 y 是个体的表现型值,μ 是群体均值,G 是遗传效应值,ε 是随机机误.

区间作图法假定群体的遗传变异只受一对基因控制,因此遗传效应就是 QTL 的效应($G_Q = b^* X$).考虑到数量性状实际上受多基因控制,复合区间作图法把总的遗传效应分解为被搜索的 QTL 效应($G_Q = b^* X$) 及与其它 QTL 连锁的分子标记效应($G_M = \sum b_f X_{fj}$) 两个分量.

但是在多环境下实施的遗传实验,其遗传群体的表现型变异除了受遗传主效应(G) 和残差机误(ε) 控制以外,还受到环境效应(E) 和基因型 × 环境互作效应(GE) 的控制.包括环境及基因型 × 环境互作效应的遗传模型可简单表示为

$$y = \mu + G + E + GE + \varepsilon \tag{25}$$

区间作图法和复合区间作图法是应用回归模型,分析 QTL 与分子标记的连锁关系及遗传效应.除了残差机误以外,所有回归效应只能设为固定效应.因此,这两种方法不能直接分析环境及环境互作等随机效应.采用 1.2 节和 1.3 节中提出的 QTL 定位间接分析方法,可以有效地分析 QTL 主效应和 QTL × 环境互作效应,并定位在特定发育阶段($t \to t-1$) 表达的、具有净遗传主效应以及净环境互作效应的发育特异性 QTL[15, 16].

本文 2.2 节中所提出的 QTL 分析方法是基于混合模型的复合区间作图方法.该方法把控制背景遗传变异的分子标记效应($G_M = \sum u_{fj} e_f$) 归为随机变量,使它们不会影响对 QTL 位置及遗传效应的无偏估算.这种分析方法还可以在模型中包括环境效应及环境互作效应,

$$y = \mu + G_Q + E + G_Q E + G_M + G_M E + \varepsilon \tag{26}$$

可有效地直接分析 QTL 的遗传主效应以及 QTL×环境互作效应.

采用 2.3 节中所介绍的 Markov 链蒙特卡罗(MCMC)分析方法定位 QTL,不但可以无偏估算基因的各种效应,还能对遗传参数的统计特征作 Bayesian 推断,因而具有更广的应用前景.

参 考 文 献

[1] Lander E S, Botstein D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps [J]. Genetics, 1989,121: 185~199.

[2] Zeng Z B. Precision mapping of quantitative trait loci[J]. Genetics, 1994, 136: 1457~1468.

[3] Moreno-Gonzalez J. Estimates of marker-associated QTL effects in Monte Carlo backcross generations using multiple regression[J]. Theor Appl Genet, 1992, 85:423~434.

[4] 朱 军. 广义遗传模型与数量遗传分析新方法[J]. 浙江农业大学学报, 1994, 20: 551~559.

[5] 朱 军. 遗传模型分析方法[M]. 北京: 中国农业出版社, 1997.

[6] 陆朝福, 朱立煌. 植物育种中的分子标记辅助选择[J]. 生物工程进展, 1995, 15(4): 11~17.

[7] Zhu J, Weir B S. Diallel analysis for sex-linked and maternal effects[J]. Theor Appl Genet, 1996, 92: 1~

9.

[8] Zhu J. Analysis of conditional genetic effects and variance components in developmental genetics[J]. Genetics, 1995, 141: 1633～1639.

[9] Li Z K, Pinson S R M, Park W D, *et al*. Epistasis for three grain components in rice (Oryza sativa L.)[J]. Genetics, 1997, 145: 453～465.

[10] Yu S B, Li J X, Xu C G, *et al*. Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid[J]. Proc Natl Acad Sci, USA, 1997, 94: 9226～9231.

[11] Henderson C R. Application of Linear Models in Animal Breeding[M]. Canada: University of Guelph, 1984.

[12] Bink M C A M, Quass Q L, Van Arendonk J A M. Bayesian estimation of dispersion parameters with a reduced animal model including polygenic and QTL effects[J]. Genet Sel Evol, 1998, 30: 103～125.

[13] Wang C S, Rutledge J J, Gianola D. Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs[J]. Genet Sel Evol, 1994, 26: 91～115.

[14] Macedo F W M, Gianola D. Bayesian analysis of univariate mixed models with informative priors[A]. Eur Assoc Anim Prod 38th Annu Meet[C]. Lisbon: Prortugal, 1987.35.

[15] Yan J Q, Zhu J, He C X, *et al*. Quantitative trait loci analysis for developmental behavior of tiller number in rice (Oryza sativa L.) [J]. Theor Appl Genet, 1998, 97: 267～274.

[16] Yan J Q, Zhu J, He C X, *et al*. Molecular dissection of developmental behavior of plant height in rice (Oryza sativa L.)[J]. Genetics, 1998, 150: 1257～1265.

Mixed model approaches of mapping genes for complex quantitative traits

ZHU Jun

(Research Center of Biomathematics, Zhejiang Univ., Hangzhou 310029, China)

Abstract: New QTL mapping methods based on mixed linear model approaches were proposed for analyzing complex genetic effects and their interaction with environments. Unbiased prediction can be applied for predicting genotype effects and genotype × environment interaction effects, which can then be further used for mapping QTL or developmental QTL with genetic main effects and GE interaction effects by interval mapping or composite interval mapping approaches. Mixed-model-based composite interval mapping approaches are capable of handling genetic data derived from multiple environments and directly analyzing genetic main effects (including epistasis) and GE interaction effects. Markov chain Monte Carlo (MCMC) methods can be applied to make inference for the statistical properties of QTL.

Key words: quantitative trait loci; QTL mapping methods; epistasis; genotype × environment interaction; developmental QTL mapping

(责任编辑:张　明)