基于性状 – 标记回归的 QTL 区间测验方法

吴为人,李维明
(福建农业大学作物科学学院,福州 350002)

摘 要 本文提出了两种基于性状 – 标记回归的 QTL 区间测验方法,分别称为 TMRIT – I 和 TMRIT – II。前者采用似然比统计量进行显著性测验,与基于最小二乘的简化复合区间定位法(sCIM)等价,但计算上明显简单快捷;后者则采用一种"伪似然比"统计量进行显著性测验,不仅进一步简化计算,而且明显提高统计功效。二者皆可通过排列测验估计显著阈值。给出了一个模拟例子。

关键词:QTL;定位;性状 – 标记回归;区间测验
中图分类号:Q348 文献标识码:A 文章编号:0253 – 9772(2001)02 – 0143 – 04

Methods of QTL Interval Test Based on Trait-Marker Regression

WU Wei-ren, LI Wei-ming
(*College of Crop Science, Fujian Agricultural University, Fuzhou* 350002, *China*)

Abstract: Two methods of interval test of quantitative trait loci (QTLs) based on trait – marker regression are proposed, named as TMRIT – I and TMRIT – II, respectively. The former uses likelihood ratio statistic for significance test, equivalent to the method of simplified composite interval mapping (sCIM) based on least squares, but much simpler and quicker in calculation. The latter uses a 'pseudo – likelihood ratio' statistic for significance test, not only simplifying calculation further, but also significantly increasing statistical power. For both methods, significance threshold can be estimated by permutation tests. A simulated example is given.

Key words: QTL; mapping; trait – marker regression; interval test

统计分析方法的好坏对数量性状基因座(QTL)的定位效果影响很大。在已有的 QTL 定位方法中,复合区间定位法(composite interval mapping, CIM)[1]被认为是最有效的。但最初提出的 CIM 是基于最大似然法的,计算上复杂费时。后来发展出了基于最小二乘的简化(simplified)CIM 方法(sCIM)[2],使计算速度大为提高。但由于 CIM 必须以一定步长(通常取 1 cM)对整个基因组逐点进行扫描,因此即使采用了最小二乘法,速度仍显较慢。Whittaker 等[3]证明,在回交一代(BC)、加倍单倍体(DH)、重组自交系(RI)等每个基因座只存在两种基因型的群体中,性状 – 标记回归(trait – marker regression, TMR)分析与 sCIM 是等价的,当相邻两个标记的偏回归系数同号(即同为正或负)时,可以由它们计算出该标记区间内的 QTL 位置和效应。这使得 QTL 定位变得极为简单和快捷。但他们建议以标记偏回归系数的大小以及相邻标记的偏回归系数是否同号来判断标记区间内 QTL 的存在与否,这显得不够直观和可靠。

另外,TMR 在统计上的一个重要特性是,一个 QTL 的效应只被与它直接相邻的左右侧两个标记(的偏回归系数)所吸收,而与相邻标记之外的标记无关。CIM 正是利用 TMR 的这个特性来提高 QTL 定位的准确性的,但它同时也带来了降低QTL检测

收稿日期:2000 – 04 – 11 修回日期:2000 – 06 – 30
基金项目:国家自然科学基金重大项目资助(资助号 39893350)
作者简介:吴为人(1960 –),男(汉族),博士,教授,目前主要研究方向:QTL 定位,分子标记应用,数量遗传学。E-mail:wuwr@fjau.edu.cn

统计功效的负作用。这是因为,在零假设条件下,被检标记区间内的 QTL 的效应会被区间外侧相邻标记部分吸收。当外侧相邻标记与被检区间靠得很近时,会大大降低 QTL 检测的统计功效。为解决这个问题,目前只好在被检区间两侧设置一定宽度的窗口,只有在窗口之外的标记才允许用作余因子(cofactor)。但这使得 CIM 不能鉴别两个较紧密连锁的 QTL。

本文旨在给出一种类似于 sCIM,以似然比统计量为依据的基于 TMR 的区间测验方法(TMR – based interval test, TMRIT),记为 TMRIT – I,并提出一种以"伪似然比"统计量为依据的 TMRIT 方法,记为 TMRIT – II,以提高 QTL 检测的统计功效。

1 TMRIT – I

考虑一个 DH 群体。若没有上位性,则性状 – 标记回归模型为:

$$y_i = b_0 + \sum_{j=1}^{m} b_j x_{ij} + e_i \qquad (i = 1,2,\cdots,n) \quad (1)$$

式中 y_i 为第 i 个体的性状值,b_j 为第 j 标记的偏回归系数,x_{ij} 为第 i 个体中第 j 标记的基因型指示变量(当基因型为 +/+ 时取值为 1,–/– 时为 –1),e_i 为误差。当某个标记区间$(k,k+1)$内存在 QTL 时,两端标记的偏回归系数(b_k 和 b_{k+1})将同号,并且有[3]:

$$r_k = \frac{1}{2}\left[1 - \sqrt{(1-2r)\,\frac{b_k + b_{k+1}(1-2r)}{b_{k+1} + b_k(1-2r)}}\right] \quad (2)$$

$$a^2 = \frac{[b_k + b_{k+1}(1-2r)][b_{k+1} + b_k(1-2r)]}{1-2r} \quad (3)$$

式中 r_k 为第 k(亦即左侧)标记与 QTL 间的重组率,r 为两标记间的重组率,a 为 QTL 的加性效应。因此,从相邻标记的偏回归系数,就足以求出区间内 QTL 的位置和效应。

在 CIM 中,对 QTL 进行统计测验的假设为 $H_0: a = 0$ 和 $H_1: a \neq 0$。由式(3)知,这相当于在 TMRIM 中对 $H_0: b_k = b_{k+1} = 0$ 和 $H_1: b_k \neq 0$ 和(或) $b_{k+1} \neq 0$ 进行测验。对应于 H_1,模型(1)是一个完全模型。当 H_0 成立时,模型(1)约简为

$$y_i = b_0 + \sum_{\substack{j=1 \\ j \neq k,k+1}}^{m} b_j x_{ij} + e_i \qquad (i = 1,2,\cdots,n) \quad (4)$$

将完全模型和约简模型的最小剩余平方和分别记为 RSS_1 和 RSS_0,则似然比统计量为

$$LR = -p \ln \frac{RSS_1}{RSS_0} \quad (5)$$

式中,根据 Box 近似[4],$p = n - m - 1$。在 H_0 下,LR 近似服从自由度为 2 的卡方分布。对 LR(或 LOD≈0.217 LR)的显著性进行测验,就能对 QTL 的存在性进行推断。

但是,由于必须对基因组中所有标记区间依次进行测验,这就遇到对同一组数据重复测验的问题,其后果是使假阳性机会增高。为此,就必须提高各次测验的显著水平。但到底应该用多高的显著水平才合适,理论上很难确定。目前流行的做法是,从整个基因组水平上控制总的错误率,即零假设为 H_0:基因组中不存在 QTL。于是,对 LR 显著阈值的确定依赖于整个基因组上各单次测验中最大 LR 值的概率分布。不幸的是,虽然已知 LR 近似服从卡方分布,但对最大 LR 值的理论分布却是未知的。因此通常只能借助模拟抽样的方法来估计 LR 的显著阈值。目前一种较好的数值计算方法是排列测验法(permutation test)[5],其最大优点是只依赖于当前样本,不必知道性状的总体分布。

2 TMRIT – II

Wu 等[6]曾提出利用 TMR 来滤除相对于单条染色体的遗传背景噪音,以提高在每条染色体上检测 QTL 的灵敏度。这一思想可以推广到单个标记区间的情形。先用模型(1)(即完全模型)进行回归分析。然后利用估得的回归系数,针对要检验的标记区间(假定为 k, $k+1$)计算矫正性状值:

$$y_i^* = y_i - \sum_{\substack{j=1 \\ j \neq k,k+1}}^{m} \hat{b}_j x_{ij} \qquad (i = 1,2,\cdots,n) \quad (6)$$

根据 TMR 的性质可知,y_i^* 基本不包含被检标记区间之外的 QTL(亦即遗传背景)的效应。因此,对于矫正性状值,TMR 模型便简化为

$$y_i^* = b_0 + b_k x_{ik} + b_{k+1} x_{ik+1} + e_i \quad (i = 1,2,\cdots,n) \quad (7)$$

对 QTL 的统计测验与前面相似。记在 H_0 和 H_1 下的剩余平方和分别为 RSS_0^* 和 RSS_1^*,则似然比为

$$LR^* = -p^* \ln \frac{RSS_1^*}{RSS_0^*} \quad (8)$$

为与式(5)区别起见,这里将 LR 和 p 皆加上星号。如果 Box 近似仍然适用的话,则 $p^* = n - 3$。这样,通过检验 LR^* 的显著性,就能推断 QTL 存在与否。然而,LR^* 并不服从卡方分布。根据式(6)和(7)容

易看出,RSS_1^* 与 RSS_1 是等价的,但 RSS_0^* 与 RSS_0 却不相等(事实上,$RSS_0^* > RSS_0$)。这是因为,

$$RSS_0^* = \sum_{i=1}^{n}(y_i^* - y^*)^2 = \sum_{i=1}^{n}(y_i - \hat{b}_0 - \sum_{\substack{j=1 \\ j\neq k,k+1}}^{m}\hat{b}_j x_{ij})^2 \quad (9)$$

虽然形式上与 RSS_0 相似,但式(9)中的回归系数是从模型(1)而非模型(4)估得的,RSS_0^* 并不是最小剩余平方和。因此,LR^* 并不是一个真正的似然比,称为"伪似然比"(pseudo – likelihood ratio)。

尽管 LR^* 的概率分布是未知的,但并不妨碍用它进行显著性测验。事实上,只要它是一个统计量,就能够用排列测验的方法来估计它的显著阈值。由于 LR^* 的计算非常简单,因此排列测验并不费时,是现实可行的。顺便一提,在式(8)中,p^* 是一个常数。既然 Box 近似对计算 LR^* 没有什么实际意义,那么 p^* 取何值也就无关紧要(当然必须 $p^* > 0$)。在下面的例子里,为了比较 TMRIT – I 和 TMRIT – II,我们取 $p^* = p$。

3　模拟例子

假设某 DH 群体由 80 个株系组成,有 1 对同源染色体分离,该染色体长 150 cM,含 16 个标记座位,均匀分布在染色体上,间距为 10 cM;染色体上有 4 个 QTL,依次位于 16、48、90 和 125 cM 处,其加性效应分别为 1.2、1.0、– 1.2 和 0.9;随机误差的标准离差为 1。在计算机上产生样本,用 TMRIT – I 和 TMRIT – II 分别进行 QTL 定位分析,并用排列测验方法(共重复 10,000 次)估计所需的显著阈值。另外,为了比较,还进行了 CIM 分析(采用完全模型,亦即对余因子未作筛选)。结果如表 1 和图 1 所示。

从表 1 和图 1 可以看出:(1)LOD^* 要比 LOD 大得多,显著阈值也相应大大提高。这是预料中的事,因为 $RSS_0^* > SSR_0$。(2)在 5% 总体显著水平上,1 号和 2 号 QTL 的 LOD 值均不显著(阈值为 2.24),而 LOD^* 则皆达到显著(阈值为 7.88),说明正如所期望的那样,TMRIT – II 的统计功效确实比 TMRIT – I 高。(3)TMRIT – I 求得的某个标记区间的 LOD 值非常接近于由 sCIM 算得的该区间中的最大 LOD 值。所以,可以预计,二者的显著阈值也应非常接近。这说明,TMRIT – I 和 sCIM 不仅在 QTL 的位置和效应估计上是等价的,而且在 QTL 检测的统计功效上也是相当的。

表 1　QTL 定位结果

Table 1　QTL mapping results

标记号	标记位置 (cM)	偏回归系数	LOD 值 TMRIT-I	LOD 值 TMRIT-II	QTL 位置 (cM)	QTL 效应
1	0	0.207				
2	10	0.494	1.20	4.76		
3	20	0.498	1.42	8.08*	15.0	0.902
4	30	– 0.057	0.48	2.29		
5	40	0.326	0.29	0.93		
6	50	0.686	2.21	8.32*	46.8	0.919
7	60	0.248	2.16	7.35		
8	70	0.641	1.61	6.66		
9	80	– 0.674	1.06	1.75		
10	90	– 1.106	5.92*	17.28*	86.2	– 1.618
11	100	– 0.197	3.38*	12.04*		
12	110	– 0.133	0.20	1.06		
13	120	1.648	5.61*	15.61*		
14	130	– 0.369	5.22*	13.37*	120.0	1.648
15	140	0.110	0.29	1.03		
16	150	– 0.280	0.38	0.53		

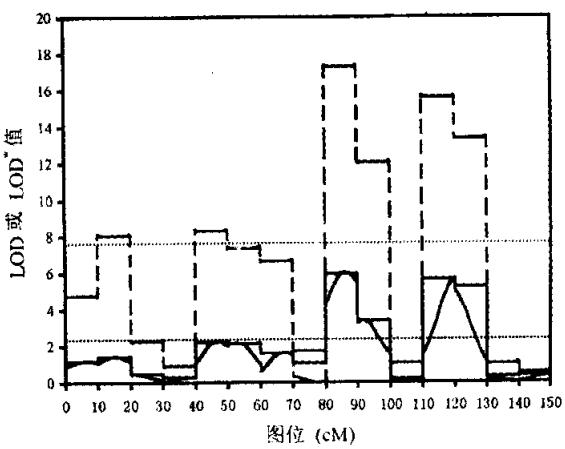

图 1　TMRIT-I、TMRIT-Ⅱ和 sCIM 得到的 LOD 或 LOD^* 轮廓线

实直方线:TMRIT-I;虚直方线;TMRIT-Ⅱ:实曲线;sCIM:下水平点线:TMRIT-I 的 LOD 阈值;上水平点线:TMRIT-I 的 LOD^* 阈值。

Fig.1　LOD and LOD^* profiles obtained by TMRIT – Ⅰ、TMRIT – Ⅱ and sCIM

另外,值得注意的是 4 号 QTL。在标记区间(12,13)和(13,14)上的 LOD 和 LOD^* 值均极显著,表明在那里应存在 QTL。但 b_{12} 与 b_{13} 及 b_{13} 与 b_{14} 之间却是异号的。这说明,4 号 QTL 不可能位于 12 和 13 号标记之间或 13 和 14 号标记之间,唯一的可能就是正好位于 13 号标记上。sCIM 的结果也证明了这一点。然而,如果按照 Whittaker 等[3]所建议的那样,仅根据相邻两个标记的偏回归系数是否同号来

判断 QTL 的存在与否,则无法得出 4 号 QTL 存在的结论。由此可以看出依据标记偏回归系数来判断 QTL 的方法的局限性,同时也反映了 TMRIT 方法的优越性。

4 讨 论

本文提出了两种基于 TMR 的 QTL 区间测验方法。其中 TMRIT-I 与 sCIM 基本上是等价的,但计算上却简单快捷得多。在 CIM 中,要从理论上确定显著阈值是困难的,因而最好采用排列测验等模拟抽样方法。但模拟抽样计算量大,使排列测验变得不太实用。在已报道的不少实际研究中,都没有采用排列测验的方法来确定显著阈值,而是人为地选定一个大致的阈值。这样必然会影响对结果的正确判断。TMRIT-I 大大简化了计算,使利用排列测验估计显著阈值的方法变得切实可行。

TMRIT-II 进一步简化了计算过程,但更重要的是它提高了 QTL 检测的统计功效。不过,这里需要考察的是,"伪似然比"是否确实具有统计量的性质。图 2 给出了前面模拟例子 10000 次排列测验所得到的最大 LOD 值和最大 LOD* 值的频率分布。可以看出,两个分布的形状是比较相似的。由此看来,LOD*(或 LR*)确实具有统计量的性质。因此,用它来进行 QTL 的显著性测验应是可行的。

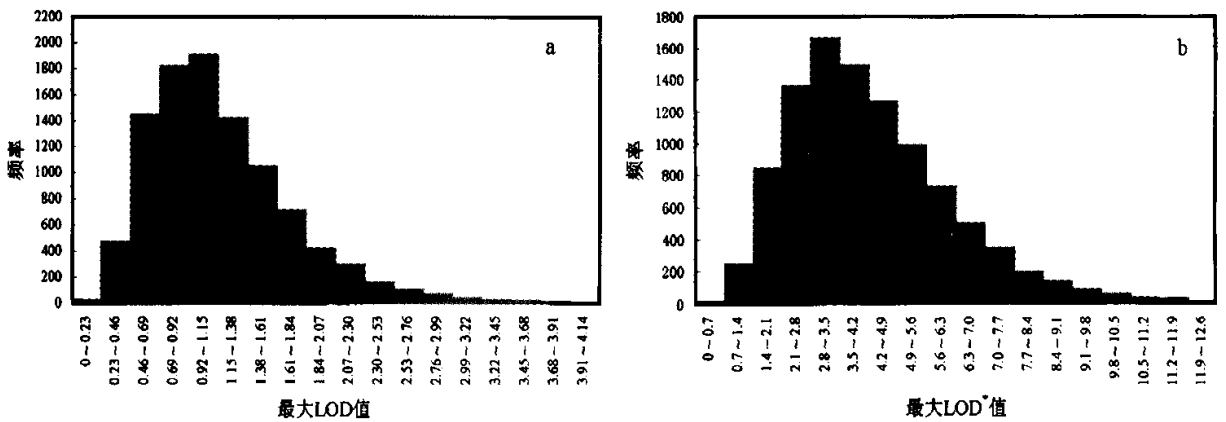

图 2 TMRIT-I 和 TMRIT-II 的最大 LOD 值(a)和最大 LOD* 值(b)的频率分布

Fig.2 Frequency distributions of maximum LOD (a) and LOD* (b) scores of TMRIT-I and TMRIT-II

TMRIT 的基础是对模型(1)进行回归分析。但在实际研究中,由于样本容量不会很大,常常会遇到标记数多于个体数的情况,因而无法同时对所有标记进行回归分析。在这种情况下,考虑到不同染色体上的标记是相互独立的,因此可以采取对各条染色体分别进行 TMRIT 分析的策略。当然,由于抽样误差,不同染色体上的标记可能会存在一定的相关。因此,在定位出 QTL 之后,最好再用多元回归分析对这些 QTL 的真实性进行重新评价,以消除假阳性[7]。另外,在 CIM 中,通过筛选余因子,从回归模型中排除多余的标记,可以显著提高 QTL 测验的统计功效。这一点对 TMRIT 而言也应是成立的。

参考文献(References):

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

[2] Wu Weiren, Li Weiming, Lu Haoran. Composite interval mapping of quantitative trait loci based on least squares estimation [J]. 福建农业大学学报, 1996, 25(4):394~399.

[3] Whittaker J C, Thompson R, Visscher P M. On the mapping of QTL by regression of phenotype on marker-type [J]. Heredity, 1996, 77:23~32.

[4] Press S J. Applied Multivariate Analysis [M]. New York:Holt Rinehart & Winston, 1972.

[5] Churchill G A, Doerge R W. Empirical threshold values for quantitative trait mapping [J]. Genetics, 1994, 138:963~971.

[6] Wu Weiren, Li Weiming, Lu Haoran. A general approach for filtrating genetic background noise in QTL mapping [J]. 生物数学学报, 1998, 13(5):592~595.

[7] Wu W-R, Li W-M, Tang D-Z, *et al*. Time-related mapping of quantitative trait loci underlying tiller number in rice [J]. Genetics, 1999, 151:297~303.