

复等位基因平衡群体中亲属关系的信息学研究^{*}

杜俊莉, 郭满才, 张宏礼, 王丽波, 刘璐, 袁志发

(西北农林科技大学 生命科学院, 陕西 杨凌 712100)

[摘要] 用推广的 ΠO 方法推出复等位基因平衡群体中母子间及全同胞对子间基因型联合概率分布列的矩阵表达; 给出了母子间及全同胞对子间的联合信息熵、互信息, 并定义了各自的信息关联系数; 与统计学研究相比表明, 信息学的方法可使复等位基因平衡群体中亲属关系的研究成为可能。

[关键词] 复等位基因; 平衡群体; 亲属关系; 信息关联系数

[中图分类号] Q3-0

[文献标识码] A

[文章编号] 1671-9387(2005)02-0155-04

群体遗传研究的核心是孟德尔群体(Mendelian population)世代传递中, 基因频率与基因型频率变化的规律。群体遗传学的传统研究方法是建立统计学模型。然而, 由于基因在世代间的传递本身就是一个信息传递过程, 故也可用信息论的方法对其进行研究。国内外学者^[1~6]在这方面已做了一些研究。本研究用信息论模型讨论复等位基因平衡群体中亲属关系的关联性, 以此说明运用信息学方法可对群体遗传学的有些问题进行更深入地研究, 以进一步丰富群体遗传学的内容。

1 复等位基因平衡群体的遗传结构及其 Shannon 信息熵和杂合度的定义^[1~2]

一个基因内不同座位的改变可以形成许多等位基因, 这些基因总称为复等位基因。复等位基因是一个群体概念。据配子随机结合的法则, 可得出复等位基因平衡群体的遗传结构为:

$$\begin{cases} (A_1, A_2, \dots, A_k) = (p_1, p_2, \dots, p_k), & (1) \\ (A_1A_1, A_1A_2, \dots, A_1A_k, A_2A_1, \dots, A_2A_k, \dots, A_{k-1}A_k) = \\ (p_1^2, p_1^2, \dots, p_k^2, p_1p_2, \dots, p_{k-1}p_k), & (2) \end{cases}$$

其中, $k \geq 3$, 基因库中各基因频率之和 $\sum_{i=1}^k p_i = 1$; 基因型中, 纯合体 A_iA_i 有 k 个, 不分正反交的杂合体

$A_iA_j (i < j)$ 共 $C_k^2 = \frac{k(k-1)}{2}$ 个, 各基因型的频率之和为 1。

基因库(1)和平衡群体(2)的 Shannon 信息熵分别为:

$$S(A) = - \sum_{i=1}^k p_i \ln p_i,$$

$$S(A^2) = - \sum_{i=1}^k p_i^2 \ln p_i^2 - 2 \sum_{i < j} p_i p_j \ln p_i p_j.$$

该复等位基因平衡群体的杂合度

$$H = 2 \sum_{i < j} p_i p_j.$$

2 母子及全同胞对子间联合概率分布列的矩阵表达

在基因库为 $(A_1, A_2, \dots, A_k)$ 的复等位基因平衡群体中, 母子间及全同胞对子间的联合概率分布列特别烦琐, 要将其一一写出, 非常费事。然而, 运用推广的 ΠO 矩阵的方法, 可将其大大简化。

这个方法的基础是利用条件概率矩阵。最基本的 3 个矩阵是 I_k, T_k, O_k , 它们均为 $\frac{k(k+1)}{2}$ 方阵。其中, I_k 矩阵提供的是 2 个基因都同源的条件概率。因此, 它们必然属于同一基因型。故可用如下方阵来表示

* [收稿日期] 2004-03-15

[作者简介] 杜俊莉(1977-), 女, 山西祁县人, 在读硕士, 主要从事农业应用数学研究。

[通讯作者] 袁志发(1938-), 男, 陕西渭南人, 教授, 博士生导师, 主要从事农业应用数学研究。

$$I_k = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}。$$

O_k 矩阵是2个亲属没有基因同源时的条件概率组成的。在这种情形下, 无论已知亲属的基因型是什么, 另一亲属为 $A_1A_1, A_1A_2, \dots, A_1A_k, A_2A_2, A_2A_3, \dots, A_{k-1}A_k, A_kA_k$ 的概率都仅只是 $p_1^2, 2p_1p_2, \dots,$

$$O_k = \begin{bmatrix} p_1^2 & 2p_1p_2 & \dots & 2p_1p_k & p_1^2 & 2p_2p_3 & \dots & 2p_{k-1}p_k & p_k^2 \\ p_1^2 & 2p_1p_2 & \dots & 2p_1p_k & p_1^2 & 2p_2p_3 & \dots & 2p_{k-1}p_k & p_k^2 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ p_1^2 & 2p_1p_2 & \dots & 2p_1p_k & p_1^2 & 2p_2p_3 & \dots & 2p_{k-1}p_k & p_k^2 \end{bmatrix}。$$

T_k 矩阵指的是2个亲属有一个基因同源的情况。该矩阵元素可这样求得: 若已知一亲属为纯合体 A_iA_i , 则另一亲属只能为 A_iA_i 。此时, A_iA_i 和它同有 A_i 的概率是1, 而 A_j (独立) 存在于 A_i 状态下的概率是 p_j , 故 A_iA_j 的概率是 p_j ; 若已知亲属为杂合体 A_iA_j , 则另一亲属只能是 A_iA_m, A_jA_m ($m \neq i, j; i, j \leq m \leq k$) 和 A_jA_i (或 A_iA_i)。在 A_iA_m, A_jA_m 情况下, 它们

$2p_1p_k, p_k^2, 2p_2p_3, \dots, 2p_{k-1}p_k, p_k^2$ 故可用这样的分块矩阵表示:

$$T_k = [N_1, N_2, \dots, N_k],$$

其中第 m ($m = 1, 2, \dots, k; k \geq 2$) 个分块矩阵的元素可表示为

$$a_{ii} = p_m^2, a_{i(j-m+1)} = 2p_m p_j, 1 \leq i \leq \frac{k(k+1)}{2}, m < j \leq k。$$

亦即:

与 A_iA_j 同有 A_i 或 A_j 的概率是1/2, 而另一独立基因 A_m 存在于 A_i 或 A_j 状态下的概率是 p_m , 故 A_iA_m, A_jA_m 的概率是 $\frac{1}{2}p_m$; 在 A_iA_j (或 A_jA_i) 情形下, 由于已知亲属既可能提供 A_i , 又可能提供 A_j , 故得其概率为 $\frac{1}{2}(p_i + p_j)$ 。其他情形概率显然为零。于是, T_k 矩阵是:

$$T_k = \begin{bmatrix} p_1 & p_2 & p_3 & \dots & p_k & 0 & \dots & 0 & 0 \\ \frac{p_1}{2} & \frac{p_1+p_2}{2} & \frac{p_3}{2} & \dots & \frac{p_k}{2} & \frac{p_2}{2} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \frac{p_1}{2} & 0 & \dots & \frac{p_{k-1}+p_k}{2} & \frac{p_k}{2} \\ 0 & 0 & 0 & \dots & p_1 & 0 & \dots & p_{k-1} & p_k \end{bmatrix}。$$

由于父母与子女总有一个基因同源, 故可利用 T_k 矩阵求得母子间的联合概率分布列。

引理1 T_k 矩阵的每行分别乘以 $p_1^2, 2p_1p_2, \dots, 2p_1p_k, p_k^2, 2p_2p_3, \dots, 2p_{k-1}p_k, p_k^2$, 即得基因库为 ($A_1, A_2, \dots, A_k$) 的平衡群体中母子间的联合概率分布列。

再考虑全同胞的条件概率。已知一个同胞的基因时, 另一个同胞有2个基因、1个基因或没有基因与它同源的的概率分别为1/4, 1/2, 1/4。因此, 另一同胞的条件概率为以下矩阵的元素:

$$S = \frac{1}{4}I_k + \frac{1}{2}T_k + \frac{1}{4}O_k。 \quad (6)$$

于是, 有如下引理:

引理2 在 S 矩阵的各行分别乘以 $p_1^2, 2p_1p_2, \dots, 2p_1p_k, p_k^2, 2p_2p_3, \dots, 2p_{k-1}p_k, p_k^2$, 即得全同胞对

子间的联合概率分布列。

3 母子间及全同胞对子间的信息关联分析

3.1 母子信息关联分析

由前述引理1求得母子间的联合概率分布列后容易看出, 母子同分布, 故可求得母、子的信息熵:

$$S(P) = S(O) = 2S(A)。$$

还可求得母子间的联合分布信息熵。

$$\text{定理1 } S(PO) = 3S(A) + [2 \sum p_i p_j] \ln 2 = 3S(A) + H \ln 2。$$

式中, $S(A)$ 为基因库信息熵, H 为杂合度。

证明 由联合概率分布列及Shannon信息熵定义得:

$$S(PO) = - \left[\sum_{i=1}^k \sum_{j=1}^k p_i^2 p_j \ln p_i^2 p_j \right] - \left[\sum_{j=1}^{k-1} \sum_{i=1}^k \sum_{m=1}^k \frac{1}{2} p_i p_j p_m \ln p_i p_j p_m \right] \times 4_0$$

然后利用对数性质展开并整理, 可得

$$\begin{aligned} S(PO) &= -3 \sum_{i=1}^k p_i \left[\sum_{j=1}^k p_j \right] \ln p_i - 2 \left[\sum_{j=1}^k p_j \right] \ln 2 \\ &= -3 \left(\sum_{i=1}^k p_i \ln p_i \right) - H \ln 2 \\ &= 3S(A) - H \ln 2_0 \end{aligned}$$

证毕。

由互信息定义, 得母子间互信息:

$$I(P, O) = S(P) + S(O) - S(PO) = S(A) - H \ln 2_0$$

由复等位基因平衡群体熵的性质知,

$$0 \leq I(P, O) \leq \ln k - \frac{k-1}{k} \ln 2_0$$

当且仅当 $p_i = \frac{1}{k}$ ($i = 1, 2, \dots, k$) 时, $I(P, O)$ 取到最

$$\begin{aligned} S(XX) &= - \sum_{i,j=1}^k \frac{1}{4} p_i^2 (1 + p_i)^2 \ln \frac{1}{4} p_i^2 (1 + p_i)^2 - \sum_{\substack{i,j=1 \\ i \neq j}}^k p_i^2 (1 + p_i) p_j \ln \frac{1}{4} p_i^2 (1 + p_i) p_j - \\ &\quad \sum_{j=1}^{k-1} \sum_{i=1}^k p_i^2 p_j^2 \ln \frac{1}{4} p_i^2 p_j^2 - \frac{1}{2} \sum_{j=1}^{k-1} \sum_{i=1}^k p_i p_j (1 + p_i) (1 + p_j) \ln \frac{1}{4} p_i p_j (1 + p_i) (1 + p_j) - \\ &\quad 2 \sum_{\substack{j=1 \\ i \neq j}}^k \left[\left(\prod_{j=1}^k p_j \right) p_i^2 \ln \frac{1}{4} \left(\prod_{j=1}^k p_j \right) p_i^2 \right] - \sum_{\substack{j=1 \\ i \neq j}}^k \left[\left(\prod_{j=1}^k p_j \right) p_i (1 + p_i) \ln \frac{1}{4} p_i (1 + p_i) \left(\prod_{j=1}^k p_j \right) \right]_0 \end{aligned}$$

再根据对数性质展开整理得:

$$\begin{aligned} S(XX) &= - \frac{1}{2} \sum_{i=1}^k p_i (1 + p_i) \left[\left(\sum_{j=1}^k p_j \right)^2 + 1 \right] \ln (1 + p_i) - \sum_{i=1}^k (3p_i - p_i^2) \ln p_i + \\ &\quad \left[\frac{1}{4} \left(\sum_{j=1}^k p_j \right)^2 + \frac{1}{2} \left(\sum_{j=1}^k p_j \right)^2 + \frac{1}{4} \left(\sum_{j=1}^k p_j \right)^2 \right] \ln 4 = \\ &\quad - \sum_{i=1}^k [p_i (1 + p_i) \ln (1 + p_i) + (3p_i - p_i^2) \ln p_i] + \ln 4_0 \end{aligned}$$

证毕。

全同胞对子间的互信息为:

$$\begin{aligned} I(X, X) &= S(X) + S(X) - S(XX) = -4 \sum_{i=1}^k p_i \ln p_i + \\ &\quad \sum_{i=1}^k [p_i (1 + p_i) \ln (1 + p_i) + (3p_i - p_i^2) \ln p_i] - \ln 4 = \sum_{i=1}^k p_i (1 + p_i) \ln \frac{1 + p_i}{p_i} - \ln 4_0 \end{aligned}$$

由复等位基因平衡群体熵的性质知, 当且仅当

$$p_i = \frac{1}{k} \quad (i = 1, 2, \dots, k) \text{ 时, } I(X, X) \text{ 取最大值}$$

$$\ln \frac{\sqrt[k]{(1+k)^{(1+k)}}}{4}_0$$

定义2 复等位基因平衡群体中全同胞对子间的信息关联系数为:

大值。

定义1 复等位基因平衡群体中母子间的信息关联系数为

$$I_{op} = \frac{I(P, O)}{\ln k - \frac{k-1}{k} \ln 2}, \quad 0 \leq I_{op} \leq 1$$

它表达了群体内母子间的信息关联程度。

3.2 全同胞对子间的信息关联分析

由前述引理2 的联合概率分布列知两全同胞对子 X 和 X 同分布, 且有 $S(X) = S(X) = 2S(A)$ 。

全同胞对子间的联合信息熵如下:

$$\text{定理 2} \quad S(XX) = - \sum_{i,j=1}^k [p_i (1 + p_i) \ln (1 + p_i) + (3p_i - p_i^2) \ln p_i] + \ln 4_0$$

证明 由引理2 得到的联合概率分布列及信息熵定义得:

$$I_s = \frac{\sum_{i=1}^k p_i (1 + p_i) \ln \frac{1 + p_i}{p_i} - \ln 4}{\ln \frac{\sqrt[k]{(1+k)^{(1+k)}}}{4}}, \quad 0 \leq I_s \leq 1$$

它表达了群体中全同胞对子间的信息关联程度, 当 $p_i = \frac{1}{k}$ ($i = 1, 2, \dots, k$) 时, $I_s = 1$, 关联程度最大。

4 比较分析

传统统计学在研究亲属关系时,只是对1对等位基因群体采用赋值的方法进行了研究,而且得到的相关系数均为常量 $1/2$;复等位基因群体由于无法

赋值,从而无法研究,但从上面的分析可看出,用信息学的方法可求出其信息关联系数,并进一步研究它们之间的信息关联程度。尤其在 $k=2$ 时,即在1对等位基因平衡群体中得到的结果与基因频率有关,这在进化上更客观。

[参考文献]

- [1] 袁志发,郭满才,宋世德,等. 相对Shannon信息量与基因变异的测量[J]. 西北农业大学学报, 1998, 26(4): 30- 34
- [2] 郭满才,解小莉,刘建军,等. 复等位基因平衡群体熵的性质[J]. 西北农林科技大学学报(自然科学版), 2002, 30(4): 119- 121
- [3] [美]Ching Chun Li 群体遗传学[M]. 吴仲贤,译. 北京: 农业出版社, 1981
- [4] 汪小龙,袁志发,郭满才,等. 最大信息熵原理与群体遗传平衡[J]. 遗传学报, 2002, 29(6): 562- 564
- [5] 朱雪龙. 应用信息论基础[M]. 北京: 清华大学出版社, 2001
- [6] 袁志发,周静芋. 多元统计分析[M]. 北京: 科学出版社, 2002

Informatics study on the kinship in alleles equilibrium population

DU Jun-li, GUO Man-cai, ZHANG Hong-li, WANG Li-bo, LIU Lu, YUAN Zhi-fa

(College of Life Sciences, Northwest A & F University, Yangling, Shaanxi 712100, China)

Abstract Matric expression between female genotypes and its descendant genotypes and between full antithetical couplet were discussed with extended ITO; And the associated genotype entropy and interentropy and the information coefficient of correlation among them were defined. The results show that informatics method can make possible the research on the kinship in alleles equilibrium population.

Key words: alleles; equilibrium population; kinship; information coefficient of correlation