

近亲繁殖下一对等位基因群体的熵性质*

张宏礼^{1,2}, 郭满才¹, 解小莉¹, 刘建军¹, 周静芋¹, 袁志发¹

(1 西北农林科技大学 生命科学学院, 陕西 杨陵 712100; 2 黑龙江八一农垦大学 基础部, 黑龙江 密山 158308)

[摘 要] 研究了在近亲繁殖下一对等位基因群体的基因库信息熵和基因型信息熵, 并讨论了在世代交替中基因型信息熵的性质、配子间的互信息性质及其与近交系数的关系, 并定义了配子间的近交关联信息系数。结果表明, 应用信息论模型研究群体遗传学与统计学方法具有一致性, 而且还具有信息学含义。

[关键词] 等位基因; Shannon 信息熵; 基因型信息熵; 互信息

[中图分类号] S11⁺ 4; S813.1

[文献标识码] A

[文章编号] 1671-9387(2003)01-0148-03

群体遗传学是研究孟德尔群体世代传递中基因频率变化规律的科学。世代传递的过程本身是一个信息传递的过程, 因此其研究的数学模型不应仅限于统计学模型, 信息论模型也应是其重要的研究内容。国内学者^[1,2]在用信息论研究群体遗传学方面已作了一些有意义的工作。本研究以群体遗传学的信息论模型为基础, 研究了一对等位基因的群体在近亲繁殖制度下, 达到平衡时群体的熵性质, 以此丰富群体遗传学的内容, 并为以信息论为工具研究遗传学问题打下基础。

1 基因型信息熵

设一对等位基因位点基因库信源为:

$$(A, P): (A, a) = (p, q) \quad (1)$$

对应的群体若为平衡群体, 则基因型(正反交分开)信源为:

$$(G_0, P): (AA, Aa, aA, aa) = (p^2, pq, pq, q^2) \quad (2)$$

式中, $0 \leq p, q \leq 1, p + q = 1$ 。

从群体(2)开始, 群体中出现了某种程度的近亲繁殖, 即出现了配子间相关, 则在配子随机结合中, 使纯合体基因型频率增加, 杂合体基因型频率减少, 偏离了随机交配^[3], 其基因型信源为:

$$(G_F, P): (AA, Aa, aA, aa) =$$

$$(p^2 + \epsilon pq - \epsilon pq - \epsilon q^2 + \epsilon) \quad (3)$$

式中, ϵ 为两性配子的协方差, $0 \leq \epsilon \leq pq$ 。若 F 为两性配子的相关系数, 即为近交系数, 则 $\epsilon = Fpq, 0 \leq F \leq 1$ 。相应地(3)式又可写成:

$$(C_F, P): (AA, Aa, aA, aa) = (p^2 +$$

$$Fpq, pq - Fpq, pq - Fpq, q^2 + Fpq) \quad (4)$$

这种规律称为 Wright 平衡定律。根据 Wright 平衡定律, 从群体(3)或(4)开始, 在施以随机交配的世代交替中, 若 F 保持不变, 则群体(3)或群体(4)保持不变。

从群体(2)开始, 在遗传同型交配制度下, 随世代交替, 群体最终趋向于:

$$(G_1, P): (AA, Aa, aA, aa) = (p, 0, 0, q) \quad (5)$$

根据 Shannon 信息熵的定义, (1)的基因库信息熵为:

$$S(A) = p \ln p - q \ln q \quad (6)$$

群体(2)的基因型信息熵(正反交分开)为:

$$S(G_0) = -p^2 \ln p^2 - 2pq \ln pq - q^2 \ln q^2 \quad (7)$$

群体(3)或(4)的基因型信息熵(正反交分开)为:

$$S(G_F) = - (p^2 + \epsilon) \ln (p^2 + \epsilon) - 2(pq - \epsilon) \ln (pq - \epsilon) - (q^2 + \epsilon) \ln (q^2 + \epsilon) \quad (8)$$

或

$$S(G_F) = - (p^2 + Fpq) \ln (p^2 + Fpq) - 2(pq - Fpq) \ln (pq - Fpq) - (q^2 + Fpq) \ln (q^2 + Fpq) \quad (9)$$

群体(5)的基因型信息熵为:

$$S(G_1) = -p \ln p - q \ln q \quad (10)$$

2 基因型信息熵的性质

根据国内学者的研究^[1,2], 在随机交配制度下, 基因型信息熵有如下性质:

性质 1 在随机交配制度下, 具有基因库(1)的群体达到平衡群体(2)时, 基因型信息熵最大, 为

* [收稿日期] 2002-04-16

[基金项目] 西北农林科技大学重点科研基金项目(0808)

[作者简介] 张宏礼(1970-), 男, 黑龙江鸡西人, 讲师, 主要从事农业应用数学研究。

$S(G_0)$ 。

性质2 在随机交配制度下, 具有基因库(1)的群体达到平衡群体(2)时, 基因型信息熵和基因库信息熵满足:

$$S(G_0) = 2S(A)。$$

性质3 在随机交配制度下, 当 $p = q = 1/2$ 时, 群体(1)的基因库信息熵达到最大值, 且最大值为:

$$S(A)_{\max} = \ln 2。$$

性质4 在随机交配制度下, 当 $p = q = 1/2$ 时, 平衡群体(2)的基因型信息熵达到最大值, 且最大值为:

$$S(G_0)_{\max} = 2\ln 2。$$

性质5 在遗传同型交配制度下, 平衡群体(2)最终趋向于群体(5), 群体(5)基因型信息熵和基因库信息熵相等, 即

$$S(G_1) = S(A) = S(G_0)/2。$$

从群体(2)开始, 群体中出现了某种程度的近亲繁殖, 则在世代交替中其基因型信息熵具有如下性质:

性质6 从平衡群体(2)开始, 在近亲繁殖制度下, 其基因型信息熵逐代减少。

证明: 从平衡群体(2)开始之后的每一代, 在近亲繁殖制度下, 基因型频率为(3)式, 基因型信息熵为:

$$S(G_F) = - (p^2 + \epsilon) \ln(p^2 + \epsilon) - 2(pq - \epsilon) \ln(pq - \epsilon) - (q^2 + \epsilon) \ln(q^2 + \epsilon)$$

将 ϵ 看作是在 $[0, pq]$ 内连续取值的变量, 在开区间 $(0, pq)$ 内对 ϵ 求导数

$$\begin{aligned} \frac{dS}{d\epsilon} &= - [1 + \ln(p^2 + \epsilon)] + 2[1 + \ln(pq - \epsilon)] - \\ &\quad [1 + \ln(q^2 + \epsilon)] = - \ln(p^2 + \epsilon) + \\ &\quad 2\ln(pq - \epsilon) - \ln(q^2 + \epsilon) = \\ &\quad \ln \frac{(pq - \epsilon)^2}{(p^2 + \epsilon)(q^2 + \epsilon)} < \ln \frac{(pq)^2}{p^2 q^2} = 0 \end{aligned}$$

这说明 $S(G_F)$ 是关于 ϵ 的减函数, 从而在世代交替中, 随 ϵ 的不断增大, 其基因型信息熵逐代减少。

性质7 从平衡群体(2)开始, 在某种固定的近亲繁殖制度下, 其基因型信息熵满足:

$$S(G_1) = S(G_0)/2 \leq S(G_F) \leq S(G_0)。$$

证明: 由性质6知, 从平衡群体(2)开始之后的每一代, 在固定的近亲繁殖制度下, 基因型信息熵 $S(G_F)$ 逐代减少。而当 $\epsilon = 0$ 时, $S(G_F) = S(G_0)$; 当 $\epsilon = pq$ 时, $S(G_F) = S(G_0)/2$ 。所以有 $S(G_1) = S(G_0)/$

$$2 \leq S(G_F) \leq S(G_0)。$$

性质8 从平衡群体(2)开始, 在近亲繁殖制度下, 其基因型信息熵逐代减少的速度不断减小。

证明: 由性质6的证明知, 群体(2)的基因型信息熵变化率为:

$$\frac{dS}{d\epsilon} = - \ln(p^2 + \epsilon) + 2\ln(pq - \epsilon) - \ln(q^2 + \epsilon)$$

再对 ϵ 求导数

$$\frac{d^2 S}{d\epsilon^2} = \frac{d}{d\epsilon} \left(\frac{dS}{d\epsilon} \right) = - \frac{1}{p^2 + \epsilon} - \frac{2}{pq - \epsilon} - \frac{1}{q^2 + \epsilon} < 0$$

这说明 $\frac{dS}{d\epsilon}$ 是关于 ϵ 的减函数, 从而在世代交替中, 随 ϵ 的不断增大, 其基因型信息熵减少的速度逐代减慢。

性质9 当 $p = q = \frac{1}{2}$ 时, $\frac{dS}{d\epsilon}$ 是关于 ϵ 递减的上

凸函数; 当 $p \neq q$ 时, 存在 $\epsilon_0 \in (0, pq)$, 使得 $\frac{dS}{d\epsilon}$ 在 $[0, \epsilon_0]$ 上是关于 ϵ 递减的上凹函数, 在 $[\epsilon_0, pq]$ 上是关于 ϵ 递减的上凸函数。

证明: 由性质8知, $\frac{d^2 S}{d\epsilon^2} = \frac{d}{d\epsilon} \left(\frac{dS}{d\epsilon} \right) < 0$, 则 $\frac{dS}{d\epsilon}$ 是关于 ϵ 的减函数。更进一步, 在开区间 $(0, pq)$ 内, 将 $\frac{d^2 S}{d\epsilon^2}$ 再对 ϵ 求导数

$$\begin{aligned} \frac{d^3 S}{d\epsilon^3} &= \frac{d}{d\epsilon} \left(\frac{d^2 S}{d\epsilon^2} \right) = \frac{1}{(p^2 + \epsilon)^2} - \frac{2}{(pq - \epsilon)^2} + \frac{1}{(q^2 + \epsilon)^2} \\ \text{由于 } \frac{d^3 S}{d\epsilon^3} &= \frac{d^2}{d\epsilon^2} \left(\frac{dS}{d\epsilon} \right)。 \text{ 于是, 当 } p = q = \frac{1}{2} \text{ 时, } \frac{d^3 S}{d\epsilon^3} = \\ &\quad \frac{2}{(1/4 + \epsilon)^2} - \frac{2}{(1/4 - \epsilon)^2} < 0, \text{ 即 } \frac{dS}{d\epsilon} \text{ 是关于 } \epsilon \text{ 的上凸函} \end{aligned}$$

数; 当 $p \neq q$ 时, $\lim_{\epsilon \rightarrow 0} \frac{d^3 S}{d\epsilon^3} = \left(\frac{1}{p^2} - \frac{1}{q^2} \right)^2 > 0$, $\lim_{\epsilon \rightarrow pq} \frac{d^3 S}{d\epsilon^3} = -\infty$, 则存在 $\epsilon_0 \in (0, pq)$, 使得 $\left. \frac{d^3 S}{d\epsilon^3} \right|_{\epsilon = \epsilon_0} = 0$ 。在 $(0, \epsilon_0)$ 内 $\frac{d^3 S}{d\epsilon^3} > 0$, 则 $\frac{dS}{d\epsilon}$ 在 $[0, \epsilon_0]$ 上是关于 ϵ 递减的上凹

函数; 在 (ϵ_0, pq) 内 $\frac{d^3 S}{d\epsilon^3} < 0$, 则 $\frac{dS}{d\epsilon}$ 在 $[\epsilon_0, pq]$ 上是关于 ϵ 递减的上凸函数。

基因型信息熵的大小反映了群体多型性的丰富程度。以上性质恰好说明: 从平衡群体(2)开始, 在固定的近亲繁殖制度下, 群体的基因型多态性递减的过程和基因型信息熵递减的过程一致。因此, 其纯合度逐代增加而杂合度逐代减少, 可以通过基因型信息熵指标反映。从随机交配的平衡状态(2)达到近亲繁殖制度下的极限状态, 是群体从多型共存到群体纯合化的过程, 这也正好从基因型信息熵的变化过

程得到反映。经过无穷代后,群体达到极限状态,基因型信息熵达到最小值,群体完全纯合。但是,随世代数的增加,这种递减趋势越来越慢。

3 配子间信息熵的性质

群体(3)配子间联合频率分布为:

(X, Y)	A	a
A	$p^2 + \epsilon p q$	$p q - \epsilon p q$
a	$p q - \epsilon p q$	$q^2 + \epsilon p q$

其中 X, Y 分别对应着雄、雌配子。

在近亲繁殖制度下,群体(3)基因频率世代不变,所以配子库信源和基因库信源(1)相同,配子库信息熵为

$$S(X) = S(Y) = S(A) \quad (11)$$

因为配子间联合频率分布信源和基因型信源一致,所以配子间联合信息熵为

$$S(XY) = S(G_F) \quad (12)$$

再根据 Shannon 平均互信息定义,配子间平均互信息为:

$$I(X, Y) = S(X) + S(Y) - S(XY) \quad (13)$$

把(11), (12)式代入(13)式中,得

$$I(X, Y) = S(G_0) - S(G_F) \quad (14)$$

由群体(3)基因型信息熵 $S(G_F)$ 性质 6, 7, 8 可得配子间平均互信息性质:

性质 10 从平衡群体(2)开始,在近亲繁殖制度下,配子间平均互信息 $I(X, Y)$ 是关于 ϵ 单调递增

的上凹函数,且 $0 \leq I(X, Y) \leq S(X) = S(Y) = S(A) = S(G_1) = S(G_0)/2$ 。

平均互信息反映的是配子间的信息交流程度。性质 10 说明,在近亲繁殖制度下,随世代交替,配子间的信息交流逐代增加,直到极限状态时达最大。此时群体完全纯合,基因库信源和基因型信源一致,亲代向子代传递的信息是全部基因库信源(或基因型信源)的信息。

性质 10 给出了配子间平均互信息和近交系数 F 之间的关系式,即 $0 \leq I(X, Y) \leq S(X)$ 。

定义 在近亲繁殖下,称 $I(X, Y)/S(X)$ 或 $I(X, Y)/S(Y)$ 为配子间的近交关联信息系数,记作

$$I_F = I(X, Y)/S(X) = I(X, Y)/S(Y)。$$

由性质 10 及 $\epsilon = F/pq$ 可证:

性质 11 I_F 是关于 F 在 $[0, 1]$ 上的增函数,且 $0 \leq I_F \leq 1$ 。当 $F = 0$ 时, $I_F = 0$; 当 $F = 1$ 时, $I_F = 1$ 。

性质 11 说明,配子间的近交关联信息系数 I_F 随近交系数 F 的增大而增大,二者反映的问题是一致的,且取值范围相同。 F 是统计意义下配子间的相关系数,而 I_F 则从信息学的角度反映了配子间的相对关联程度。

以上引入了信息论模型,研究在一定近亲繁殖制度下群体基因库的变化规律,并得到了一些信息熵性质。这些结果不仅和统计模型一致,而且有其信息论意义。由于熵是描述发展和进化的,因而用信息论方法研究群体遗传学具有其重要意义。

[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, 宋世德, 周静芋, 等. 非平衡群体基因变异测量的 Shannon 信息量方法[J]. 生物数学学报, 2001, 16(3): 341- 347
- [3] [美] Li C. 群体遗传学[M]. 吴仲贤译. 北京: 中国农业出版社, 1981

The entropy character of a pair alleles population in inbreeding

ZHANG Hong-li^{1,2}, GUO Man-cai¹, XIE Xiao-li¹, LIU Jian-jun¹, ZHOU Jing-yu¹, YUAN Zhi-fa¹

(1 College of Life Science, Northwest Sci-Tech University of Agriculture and Forestry, Yangling, Shaanxi 712100, China;

2 Heilongjiang August First Land Reclamation University, Mishan, Heilongjiang 158308, China)

Abstract: This paper introduces the entropy of the genepool and the genotype of a pair alleles population in inbreeding, and demonstrates the character of genotype entropy, the character of inter-zygote's interentropy, and the relation between the latter and coefficient of inbreeding in the generation change. The results show that: the utilization of information model on population genetics not only has good uniformity with statistical method, but also has the informational significance.

Key words: a pair alleles; Shannon entropy; genotype entropy; interentropy