复等位基因平衡群体熵的性质*

郭满才, 解小莉, 刘建军, 张宏礼, 宋世德, 周静芋, 袁志发
(西北农林科技大学 生命科学学院, 陕西 杨陵 712100)

[摘 要] 研究了复等位基因平衡群体熵的性质, 并与基因多样度 D、相对信息多样度 $S'(A)$、相对纯合度信息多样度 $S'_J(A)$、相对杂合度信息多样度 $S'_H(A)$ 及多态信息含量 PIC 进行了比较研究。结果表明, 在表征基因变异与遗传变异上, 信息论方法与统计学方法具有一致性, 而且信息论方法还具有信息学含义。

[关键词] 复等位基因; 群体熵; 遗传平衡; 多样度; 多态信息含量
[中图分类号] S813.1 [文献标识码] A [文章编号] 1000-2782(2002)04-0119-04

群体遗传学是研究孟德尔(Mendelian population)群体世代传递中基因频率变化规律的科学。英国数学家哈代(Hardy)和德国医生温伯格(Weinberg)经过各自独立的研究, 于 1908 年分别发表了群体遗传学的基本规律, 即 Hardy-Weinberg 定律。群体遗传学与数量遗传学的很多结论都是以它为出发点展开的。基因在世代之间的传递过程本身是一个信息传递过程, 因而, 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[1~5]在这方面做了一些有意义的工作。本研究拟以群体遗传学的信息论模型为基础, 研究复等位基因平衡群体熵的性质, 以丰富群体遗传学内容, 并为以信息论为工具研究遗传学问题打下坚实的基础。

1 复等位基因平衡群体熵的定义

设复等位基因位点 A 为:

$$(A_1, A_2, \cdots\cdots, A_k) = (p_1, p_2, \cdots\cdots, p_k) \quad (1)$$

其中, p_i 为 A_i 的频率, $p_i \geq 0$, 且 $\sum_{i=1}^{k} p_i = 1$。群体平衡时, 各基因型的频率为:

$$(p_1A_1 + p_2A_2 + \cdots + p_kA_k)^2 = \sum_{i=1}^{k} p_i^2 A_iA_i + 2\sum_{i<j} p_ip_jA_iA_j \quad (2)$$

其中, p_i^2 为 A_iA_i 的频率, A_iA_j 与 A_jA_i 为正反交, 其频率均为 p_ip_j。

据 Shannon 信息熵的定义, 复等位基因库(1)的信息熵为:

$$S(A) = -\sum_{i=1}^{k} p_i \ln p_i \quad (3)$$

平衡群体(2)的信息熵为:

$$S(A^2) = -\left\{\sum_{i=1}^{k} p_i^2 \ln p_i^2 + 2\sum_{i<j} p_ip_j \ln p_ip_j\right\} = -2\left\{\sum_{i=1}^{k} p_i^2 \ln p_i + \sum_{i<j} p_ip_j \ln p_ip_j\right\} \quad (4)$$

2 基因一致度 J、基因多样度 D、多态信息含量 PIC[6] 及其性质

基因库(1)或群体(2)的基因一致度 J 和基因多样度 D 及多态信息含量 PIC 分别定义为:

$$J = \sum_{i=1}^{k} p_i^2, D = 1 - \sum_{i=1}^{k} p_i^2,$$

$$PIC = 1 - \sum_{i=1}^{k} p_i^2 - 2\sum_{i<j} p_i^2 p_j^2 \quad (5)$$

其性质为:

$$\frac{1}{k} \leq J \leq 1, 0 \leq D \leq \frac{k-1}{k},$$

$$0 \leq PIC \leq \frac{(k-1)^2(k+1)}{k^3} \quad (6)$$

当 p_i 之一为 1 而其余为 0 时, 有 $J_{max} = 1, D_{min} = 0$, $PIC_{min} = 0$; 当 p_i 均等于 $\frac{1}{k}$ 时, 有 $J_{min} = \frac{1}{k}, D_{max} = \frac{k-1}{k}, PIC_{max} = \frac{(k-1)^2(k+1)}{k^3}$。显然, 基因一致度 J 与多样度 D 分别是以平衡群体的纯合体频率与杂合体频率为代表, 来度量具有同一基因库的任一群体的基因变异。

* [收稿日期] 2002-02-26
[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] 郭满才(1963-), 男, 陕西宝鸡人, 副教授, 在读博士, 主要从事农业应用数学的研究。

3 复等位基因平衡群体熵的性质

对于给定的基因库(1),其对应的群体有无数多个,不失一般性,设其 k^2 个基因型(正反交分开)分布为:

$$(A_1A_1, \cdots\cdots, A_kA_k, A_1A_2, \cdots\cdots, A_{k-1}A_k) = (p_{11}, \cdots\cdots, p_{kk}, p_{12}, \cdots\cdots, p_{k-1,k}) \quad (7)$$

根据最大信息熵原理,可以证明:

性质 1: 对于给定的基因库,当群体平衡时其基因型信息熵最大。

证明: 对群体(7),其信息熵为

$$S = -\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\ln p_{ij} \quad (8)$$

其中,

$$\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij} = 1 \quad (9)$$

$$\frac{1}{2}\sum_{j=1}^{k}(p_{ij} + p_{ji}) = p_i \quad (i = 1, 2, \cdots, k) \quad (10)$$

应用拉格朗日乘数法,引入 $\lambda_0, \lambda_1, \cdots, \lambda_k$ 构造目标函数:

$$= -\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\ln p_{ij} + (\ln\lambda_0 + 1)\left(\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij} - 1\right) + \sum_{i=1}^{k}\ln\lambda_i\left(\frac{1}{2}\sum_{j=1}^{k}(p_{ij} + p_{ji}) - p_i\right) \quad (11)$$

令各一阶偏导数为零: $\frac{\partial G}{\partial p_{ij}} = 0$, 即:

$$\ln p_{ij} - \ln\lambda_0 - \frac{1}{2}(\ln\lambda_i + \ln\lambda_j) = 0$$

可得:

$$p_{ij} = \lambda_0\sqrt{\lambda_i\lambda_j} \quad (i, j = 1, 2, \cdots, k) \quad (12)$$

将(12)代入方程(9),(10)可以解出 $\lambda_0, \lambda_1, \cdots, \lambda_k$, 进而求出最大熵分布概率 p_{ij}。当 $k = 2$ 时, $p_{11} = \lambda_0\lambda_1$, $p_{22} = \lambda_0\lambda_2$, $p_{12} = p_{21} = \lambda_0\sqrt{\lambda_1\lambda_2}$, 代入(9),(10):

$$\begin{cases} \lambda_0(\lambda_1 + \lambda_2 + 2\sqrt{\lambda_1\lambda_2}) = 1 \\ \lambda_0(\lambda_1 + \sqrt{\lambda_1\lambda_2}) = p_1 \\ \lambda_0(\lambda_2 + \sqrt{\lambda_1\lambda_2}) = p_2 \end{cases}$$

解之可得: $\lambda_1 = \frac{p_1^2}{\lambda_0}$, $\lambda_2 = \frac{p_2^2}{\lambda_0}$, 代回(12)即得:

$$p_{11} = \lambda_0\lambda_1 = p_1^2,\ p_{22} = \lambda_0\lambda_2 = p_2^2,\ p_{12} = p_{21} = \lambda_0\sqrt{\lambda_1\lambda_2} = p_1p_2,$$ 证毕。

这正是 Hardy-Weinberg 定律所给出的平衡群体的基因型频率,由此证明了最大熵分布就是 Hardy-Weinberg 平衡分布,而且可以证明对于任意 k 都是如此。

Hardy-Weinberg 平衡定律与最大信息熵原理的内在一致性说明,随机交配导致群体基因型信息熵增大,当群体基因型信息熵达到最大时,其基因型频率不再变化,即达到"平衡",因而,随机交配过程是一个不可逆过程,而选择和近亲交配等育种手段则使群体的信息熵降低,有序性增加。育种的实质是通过施行选择和近亲交配等手段,来增加群体的一致性和有序性,或通过杂交和随机交配等手段增加群体的杂合性和多态性,以调节群体的信息熵。

性质 2: 对某一等位基因个数确定的位点,当各等位基因频率相等时,该位点的信息熵最大。

证明: 设位点 A 如(1)所示,则该位点的信息熵为

$$S(A) = -\sum_{i=1}^{k} p_i\ln p_i \quad (13)$$

其中

$$\sum_{i=1}^{k} p_i = 1$$

应用拉格朗日乘数法,引入 $\ln\lambda + 1$ 构造目标函数:

$$G = -\sum_{i=1}^{k} p_i\ln p_i + (\ln\lambda + 1)\left(\sum_{i=1}^{k} p_i - 1\right) \quad (15)$$

令各一阶导数为零: $\frac{\partial G}{\partial p_i} = 0$, 即:

$$\ln p_i - \ln\lambda = 0 \quad (i = 1, 2, \cdots, k) \quad (16)$$

可得:

$$p_i = \lambda \quad (i = 1, 2, \cdots, k) \quad (17)$$

将(17)代入(14),可求出当 $p_i = \frac{1}{k}$ $(i = 1, 2, \cdots\cdots, k)$ 时,基因库(1)的信息熵最大,最大值为 $\ln k$。即 $0 \le S(A) \le \ln k$。

由性质 1 与性质 2 容易得到,对于等位基因数目固定的基因库,当 $p_1 = p_2 = \cdots = p_k = \frac{1}{k}$ 时,不同基因库所对应的群体(2)的最大信息熵达到最大值,最大值为 $2\ln k$,即 $0 \le S(A^2) \le 2\ln k$,且 $S(A^2) = 2S(A)$。

由于 $S(A)$ 和 $S(A^2)$ 与等位基因数 k 有关,引入基因库(1)和群体(2)的相对信息熵 $S'(A)$ 和 $S'(A^2)$ 为:

$$S'(A) = \frac{1}{\ln k}S(A),\ S'(A^2) = \frac{1}{2\ln k}S(A^2) \quad (18)$$

显然有 $S'(A^2) = \frac{1}{2\ln k}S(A^2) = \frac{1}{2\ln k}2S(A) = S'(A)$。对于相对信息熵有:

$$0 \le S'(A^2) = S'(A) \le 1$$

$S(A^2)$ 为平衡群体和位点的多样性的度量, $S'(A^2)$ 的意义为平衡群体和位点的多样性程度, 即它占最大可能多样性的比例。$S'(A^2)$ 可以作为群体中基因变异与基因型变异的测度, 称为基因相对信息量多样度。

将相对纯合度信息熵 $S_J'(A^2)$ 和相对杂合度信息熵 $S_H'(A^2)$ 分别定义为:

$$\begin{cases} S_J'(A^2) = -\dfrac{1}{\ln k}\sum_{i=1}^{k} p_i^2 \ln p_i \\ S_H'(A^2) = -\dfrac{1}{\ln k}\sum_{i<j} p_i p_j \ln p_i p_j \end{cases}$$

$S_J'(A^2)$ 的意义是指群体中纯合体的多样性的比例。$S_H'(A^2)$ 的意义是指群体中杂合体的多样性的比例。显然

$$S'(A^2) = S_J'(A^2) + S_H'(A^2)$$

$S_J'(A^2)$ 和 $S_H'(A^2)$ 又分别称为纯合相对信息量多样度和杂合相对信息量多样度。

性质 3: D、$S(A)$ 和 $S'(A)$ 是关于各 p_i 的凸函数; $S(A^2)$、$S'(A^2)$、$S_J'(A^2)$、$S_H'(A^2)$、PIC 是关于各 p_{ij} 的凸函数。

证明: 当 $k=2$ 时, 由于

$$D = 1 - p_1^2 - (1-p_1)^2, \frac{d^2D}{dp_i^2} = -4 < 0 \quad (i=1,2);$$

$$S(A) = -(p_1\ln p_1 + p_2\ln p_2), \frac{d^2S(A)}{dp_i^2} = -\left(1+\frac{1}{1-p_i}\right) < 0 \quad (i=1,2);$$

$$S'(A) = -\frac{1}{\ln 2}(p_1\ln p_1 + p_2\ln p_2), \frac{d^2S'(A)}{dp_i^2} = -\frac{1}{\ln 2}\times\left(1+\frac{1}{1-p_i}\right) < 0 \quad (i=1,2)。$$

$S(A^2)$、$S'(A^2)$、$S_J'(A^2)$、$S_H'(A^2)$、PIC 是关于各 p_{ij} 的凸函数, 以及 k 等于其他值的情形时的证明与之类似。

通过模拟数据还可证明:

性质 4: 多样度 D、基因库相对信息多样度 $S'(A^2)$、相对纯合信息量多样度 $S_J'(A^2)$、相对杂合信息量多样度 $S_H'(A^2)$、多态信息含量 PIC 呈正相关。

事实上, 当 $k=2$, p_1 的初始频率为 0, 步长为 0.01, 终止频率为 1.0 时, 模拟了 101 个群体, 可得出表 1。

表 1　D、$S'(A^2)$、$S_J'(A^2)$、$S_H'(A^2)$、PIC 间的相关系数

Table 1　The correlation coefficient among D、$S'(A^2)$、$S_J'(A^2)$、$S_H'(A^2)$ and PIC

	D	$S'(A^2)$	$S_J'(A^2)$	$S_H'(A^2)$	PIC
D	1.000 000	0.995 533	0.998 988	0.962 631	0.996 006
$S'(A^2)$		1.000 000	0.990 358	0.983 879	0.999 629
$S_J'(A^2)$			1.000 000	0.949 617	0.991 004
$S_H'(A^2)$				1.000 000	0.982 204
PIC					1.000 000

该性质说明, 在表征基因变异与遗传变异上, 各个指标体系是相当的, 但相对信息量多样度除具有遗传学意义外, 还具有其信息学内涵。并且还有如下特点:

(1) 各多样性指标中, 相对信息量多样度变化范围最大, 介于 0～1, 因而对变异的描述最清晰, 并且可以对变异来源进行剖分。PIC 取值范围最小。

(2) 随着位点等位基因数目的增加, D 与 PIC 的取值趋于 0～1, 表明各指标在描述基因变异与基因型变异上的作用是类似的, 但各有其不同的遗传学含义。

[参考文献]

[1] 王身立. 生物物理遗传学[M]. 长沙: 湖南科学技术出版社, 1992.

[2] 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.

[3] 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.

[4] 周士谔, 衡红刚, 张国庆. 数量遗传学中一种新的求综合性状的方法[J]. 遗传学报, 1989, 16(4): 269-275.

[5] 柯卫东, 刘采芹. 数量遗传学中的信息问题初探[J]. 自然杂志, 1998, 13(4): 210-212.

[6] Botstein David, White R L, Skolnick M, et al. Construction of a genetic linkage map in man using restriction fragment length polymorphisms[J]. Am J Hum Genet, 1980, 32: 314-331.

The entropy characters of alleles equilibrium population

GUO Man-cai, XIE Xiao-li, LIU Jian-jun, ZHANG Hong-li,
SONG Shi-de, ZHOU Jing-yu, YUAN Zhi-fa
(*College of Life Sciences, Northwest Sci-Tech University of Agriculture and Forestry, Yangling, Shaanxi* 712100, *China*)

Abstract: The entropy characters of alleles equilibrium population were discussed and the comparison among gene diversity D, relative information diversity $S'(A)$, relative homozygosity information diversity $S_J'(A)$, relative heterozygosity information diversity $S_H'(A)$ and polymorphism information content (PIC) were studied. The results show that the statistic method and the informative method have very good uniformity in measuring the gene variation and the genetic variation, but the information method has the informatic means.

Key words: alleles; population entropy; genetic equilibrium; diversity; polymorphism information content

(上接第 118 页)

Studies on breed resource of Langkazi sheep in Tibet

REN Zhan-jun[1], CHANG Hong[2], Mimaciren[3], LI Qiu-yan[4]
(1 *College of Animal Sciences and Technology, Northwest Sci-Tech University of Agriculture and Forestry, Yangling, Shaanxi* 712100, *China*;
2 *College of Animal Science and Veterinary Medicine, Yangzhou University, Yangzhou, Jiangsu* 225009, *China*;
3 *Animal Science and Veterinary Medicine Service Center in Tibet, Lasa, Tibet* 850000, *China*;
4 *Deer General Farm of thirty-three rigiment second Agriculture division, Weili, Xinjiang* 841505, *China*)

Abstract: Studies were carried out on Langkazi sheep with regard to its origion, development, ecological conditions, external physical characteristics, population structure, production performance, raising and management, etc. The results showed that: the Langkazi sheep has a long history in its development and can be widely used for the production of hair, milk and meat. It has poor production performance, but its heredity quality is steady. It can be raised with crude forage and can survive in vile environment such as in the cold moutain areas. Langkazi sheep is a precious local breed resource that cannot be replaced by other sheep varieties living in high and cold areas.

Key words: Langkazi sheep; breed recource; ecological characteristic; performance of production