

非平衡群体基因变异 Shannon
信息量的测量方法

Q343.13

Q141

郭满才, 宋世德[✓], 周静芋, 袁志发

(西北农林科技大学 基础科学系, 陕西 杨陵 712100)

[摘要] 在 Shannon 信息量的基础上, 对非平衡群体建立了群体基因型相对信息量 $S'(G)$, 纯合体相对信息量 $S_J(G)$ 和杂合体相对信息量 $S_H(G)$ 的概念, 并赋予遗传学意义。与基因一致度 J 和基因多样性 D 的理论比较表明, 二者在数量规律上有很好的 consistency, 但又是相对独立的指标体系, 且相对信息量还有新的内涵。 $S'(G)$ 既能表征基因变异, 又能反映基因型水平上的遗传变异, $S_J(G)$ 主要反映纯合体的遗传变异, $S_H(G)$ 主要反映杂合体的遗传变异。各相对信息量既可反映群体的遗传变异程度, 又能比较不同位点间的遗传变异程度。

[关键词] 非平衡群体; Shannon 信息量; 基因一致度; 基因多样性; 群体基因型信息量

基因变异, 多样性

[中图分类号] S813.1

[文献标识码] A

多样性

对如何定量地描述非平衡群体的基因变异, Nei^[1] 认为在非随机交配群体中, 虽然杂合度与群体内杂合体的频率无关, 但它仍然是一个群体内基因变异的理想尺度, 可以适用于任何生物, 无论是自体受精生物还是异体受精生物, 是单倍体还是多倍体。并建议把杂合度 H 改称为“基因多样性”, 纯合度 J 改称为“基因一致度”。Weir^[2] 在论述基因多样性时也沿用同样的指标。但实际上, 对于非平衡群体, 这种描述只利用了基因频率而没有利用基因型频率。袁志发等^[3] 提出用相对 Shannon 信息量描述平衡群体的基因变异, 并赋予遗传学意义。本文拟研究非平衡群体基因变异的 Shannon 信息量测量方法, 并与基因多样性进行比较, 为定量描述群体的基因变异提供一种新的方法, 以使 Shannon 信息量方法得到更好的应用。

1 基因多样性 $D^{[1]}$ 及其性质

设复等位基因位点 A 为:

$$(A_1, A_2, \dots, A_K) = (P_1, P_2, \dots, P_K) \quad (1)$$

其中, P_i 为 A_i 的频率, $P_i \geq 0$, 且 $\sum_{i=1}^K P_i = 1$ 。其对应的群体在理论上有无穷多个, 不失一般性, 设它们为:

$$(A_1A_1, A_2A_2, \dots, A_KA_K, A_1A_2, A_1A_3, \dots, A_{K-1}A_K) = (D_1, D_2, \dots, D_K, H_{12}, H_{13}, \dots, H_{K-1}H_K) \quad (2)$$

其中, 纯合基因型 A_iA_i 有 K 个, 频率为 D_i , 杂合基因型 $A_iA_j (i < j)$ 有 $K(K-1)/2$ 个, 频

[收稿日期] 2000-01-20

[作者简介] 郭满才(1963—), 男, 讲师, 博士。

率为 H_{ij} , 显然 $\sum_{i=1}^K D_i + \sum_{i < j} H_{ij} = 1$ 。共有基因型 $K(K+1)/2$ 个。

基因位点或基因型的一致度 J 和多样性 D 分别定义为:

$$J = \sum_{i=1}^K P_i^2, D = 1 - \sum_{i=1}^K P_i^2 \quad (3)$$

其性质为:

$$\frac{1}{K} \leq J \leq 1, 0 \leq D \leq \frac{K-1}{K} \quad (4)$$

当 P_i 之一等于 1 而其余均为 0 时, 有 $J_{\max} = 1, D_{\min} = 0$; 当 P_i 均等于 $\frac{1}{K}$ 时, 有 $J_{\min} = \frac{1}{K}, D_{\max} = \frac{K-1}{K}$ 。显然, 基因一致度 J 与基因多样性 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。

2 非平衡群体相对信息量及其性质

当群体不平衡时, 其 Shannon 信息量定义为:

$$S(G) = - \left(\sum_{i=1}^K D_i \ln D_i + \sum_{i < j} H_{ij} \ln H_{ij} \right) \quad (5)$$

群体的纯合体信息量 $S_J(G)$ 和杂合体信息量 $S_H(G)$ 分别定义为:

$$S_J(G) = - \sum_{i=1}^K D_i \ln D_i, S_H(G) = - \sum_{i < j} H_{ij} \ln H_{ij} \quad (6)$$

则有

$$0 \leq S(G) = S_J(G) + S_H(G) \leq \ln \frac{K(K+1)}{2} \quad (7)$$

当群体中只有一种基因型时 $S(G)_{\min} = 0$; 当 $\frac{K(K+1)}{2}$ 种基因型频率相等, 且均为 $\frac{2}{K(K+1)}$ 时, $S(G)_{\max} = \ln \frac{K(K+1)}{2}$, 此时 $S_J(G) = \frac{2}{K+1} \ln \frac{K(K+1)}{2}, S_H(G) = \frac{K-1}{K+1} \ln \frac{K(K+1)}{2}$ 。由于群体不平衡, $S_J(G)$ 与 $S_H(G)$ 以及 $S(G)$ 的最大值不在同一点取得, 结论如下(证明略)。

当 $K \geq 3$ 时:

$$0 \leq S_J(G) \leq \ln K, 0 \leq S_H(G) \leq \ln \frac{K(K-1)}{2} \quad (8)$$

其中当 D_i 皆等于 $\frac{1}{K}$ 时, $S_J(G)_{\max} = \ln K$; 当 H_{ij} 皆等于 $\frac{2}{K(K-1)}$ 时, $S_H(G)_{\max} = \ln \frac{K(K-1)}{2}$ 。 D_i 或 H_{ij} 之一为 1, 其余为 0 时, 它们均取最小值 0。

当 $K=2$ 时:

$$0 \leq S_J(G) \leq \frac{2}{e}, 0 \leq S_H(G) \leq \frac{1}{e} \quad (9)$$

其中当 $D_1 = D_2 = \frac{1}{e}$ 时, $S_J(G)_{\max} = \frac{2}{e}$; 当 $H_{12} = \frac{1}{e}$ 时, $S_H(G)_{\max} = \frac{1}{e}$ 。当 D_1, D_2 之一为 1, 另一个为 0 时 $S_J(G)_{\min} = 0$; 当 $H_{12} = 1$ 时, $S_H(G)_{\min} = 0$ 。

定义 $S'(G)$, $S'_J(G)$ 和 $S'_H(G)$ 分别为群体的相对信息量、纯合体相对信息量和杂合体相对信息量, 则 $K \geq 3$ 时

$$0 \leq S'_J(G) = \frac{S_J(G)}{\ln \frac{K(K+1)}{2}} \leq \frac{\ln K}{\ln \frac{K(K+1)}{2}} \quad (10)$$

$$0 \leq S'_H(G) = \frac{S_H(G)}{\ln \frac{K(K+1)}{2}} \leq \frac{\ln \frac{K(K-1)}{2}}{\ln \frac{K(K+1)}{2}} \quad (11)$$

$$0 \leq S'(G) = \frac{S(G)}{\ln \frac{K(K+1)}{2}} = S'_J(G) + S'_H(G) \leq 1 \quad (12)$$

当 $K=2$ 时, 由 (7)、(9) 可得与 (10)~(12) 类似的式子。

$S(G)$ 的遗传学意义为群体的遗传变异大小, $S_J(G)$ 表示纯合体的遗传变异大小, $S'(G)$ 为群体的遗传变异程度, $S'_J(G)$ 为纯合体的遗传变异程度, 即占最大可能遗传变异的比例。 $S_H(G)$ 与 $S'_H(G)$ 可仿比解释。

对于具有多个位点的非平衡群体, 可建立相应的平均信息量与平均相对信息量公式:

$$\bar{S}(G) = \frac{1}{r} \sum_{i=1}^r S_i(G), \bar{S}_J(G) = \frac{1}{r} \sum_{i=1}^r S_{J_i}(G), \bar{S}_H(G) = \frac{1}{r} \sum_{i=1}^r S_{H_i}(G) \quad (13)$$

$$\bar{S}'(G) = \frac{1}{r} \sum_{i=1}^r S'_i(G), \bar{S}'_J(G) = \frac{1}{r} \sum_{i=1}^r S'_{J_i}(G), \bar{S}'_H(G) = \frac{1}{r} \sum_{i=1}^r S'_{H_i}(G) \quad (14)$$

其中, r 为位点个数。

3 模拟结果与分析

对上述理论结果, 在 $K=2$ 时进行计算机模拟。当 A_1 的初始频率为 0.0, 步长 0.1, 终止频率为 1.0, 基因型频率步长为 0.005 时, 模拟了 511 个群体。当 A_1 的初始频率为 0.00, 步长 0.01, 终止频率为 0.50, 基因型频率步长为 0.005 时, 模拟了 2 601 个群体。用 $\bar{H}$, $\bar{S}'_J(G)$, $\bar{S}'_H(G)$ 与 $\bar{S}'(G)$ 分别表示具有同一基因库的所有模拟群体杂合体频率 H_{12} 的平均值、纯合体相对信息量 $S'_J(G)$ 的平均值、杂合体相对信息量 $S'_H(G)$ 的平均值以及群体的相对信息量 $S'(G)$ 的平均值。模拟结果与理论结果完全相符, 如最大值、最小值、最大值点、最小值点等。对模拟结果进行统计分析(部分数据见表 1~3, 表 2, 3 为表 1 数据计算结果, 其余略), 得出如下结果。

1) 同一基因库的每一模拟群体, 只要 3 种基因型的频率均不小于 0.03, 则其 $S'(G)$, $S'_J(G)$ 和 $S'_H(G)$ 分别与该库的 $\bar{S}'(G)$, $\bar{S}'_J(G)$ 和 $\bar{S}'_H(G)$ 差异不显著。

2) D 与 $\bar{H}$ 的关系为: 当 A_1 的频率为 $[0.00, 0.06]$, $[0.47, 0.53]$, $[0.94, 1.00]$ 时, 二者差异不显著。对其他基因库, 二者有显著或极显著的差异, 且 D 与 $\bar{H}$ 相比偏大。

3) D 与 $\bar{S}'(G)$ 的关系为: 当 A_1 为 $[0.00, 0.03]$ 及 $[0.97, 1.00]$ 时, 二者差异不显著。对其他基因库, 二者差异显著或极显著, 且 D 与 $\bar{S}'(G)$ 相比偏小。 $\bar{H}$ 与 $\bar{S}'(G)$ 有与之完全类似的性质。

4) 基因库相对信息量 $S'(A)^{[3]}$ 与每个库所对应模拟群体的最大相对信息量 $S'(G)_{\max}$

数值几乎处处相等。

5) 从简单相关看 $\bar{S}_J(G)$, $\bar{S}_H(G)$ 和 $\bar{S}(G)$ 均与 J 有强的负相关, 与 D 有强的正相关; $\bar{S}_J(G)$, $\bar{S}_H(G)$, $\bar{S}(G)$ 间有强的正相关。从偏相关看, $\bar{S}_J(G)$ 与 $\bar{S}_H(G)$ 间有强的负偏相关; $\bar{S}_J(G)$, $\bar{S}_H(G)$ 与 $\bar{S}(G)$ 间有强的正偏相关; $\bar{S}_J(G)$, $\bar{S}_H(G)$ 与 J 的偏相关为负, 与 D 的偏相关为正; $\bar{S}(G)$ 与 J 的偏相关为正, 与 D 的偏相关为负。

表1 模拟群体的 D , H 和 $\bar{S}(G)$ 分析

基因库	模拟组数	D	H	$\bar{S}(G)$	D, H 间 t 检验	$D, \bar{S}(G)$ 间 t 检验	$H, \bar{S}(G)$ 间 t 检验
(0.0, 1.0)	1	0.00	0.0	0.000 0	0.000 0	0.000 0	0.000 0
(0.1, 0.9)	21	0.18	0.1	0.441 2	1.061 3	-2.374 1	-3.175 9
(0.2, 0.8)	41	0.32	0.2	0.671 3	1.592 0	-3.192 7	-4.386 7
(0.3, 0.7)	61	0.42	0.3	0.798 0	1.592 0	-3.438 3	-4.638 2
(0.4, 0.6)	81	0.48	0.4	0.835 3	1.061 3	-3.229 1	-4.051 4
(0.5, 0.5)	101	0.5	0.5	0.766 0	0.000 0	-2.419 6	-2.477 4
(0.6, 0.4)	81	0.48	0.4	0.835 3	1.061 3	-3.229 1	-4.051 4
(0.7, 0.3)	61	0.42	0.3	0.798 0	1.592 0	-3.438 3	-4.638 2
(0.8, 0.2)	41	0.32	0.2	0.671 3	1.592 0	-3.192 7	-4.386 7
(0.9, 0.1)	21	0.18	0.1	0.441 2	1.061 3	-2.374 1	-3.157 9
(1.0, 0.0)	1	0.00	0.0	0.000 0	0.000 0	0.000 0	0.000 0
(0.00, 1.00)	1	0.000 0	0.00	0.000 0	0.000 0	0.000 0	0.000 0
(0.01, 0.99)	3	0.019 8	0.01	0.073 5	0.475 8	-1.144 7	-1.429 8
(0.02, 0.98)	5	0.039 2	0.02	0.131 0	0.932 1	-1.964 1	-2.507 2
(0.03, 0.97)	7	0.058 2	0.03	0.181 5	1.369 0	-2.638 1	-3.421 9
(0.04, 0.96)	9	0.076 8	0.04	0.227 1	1.786 5	-3.215 8	-4.226 0
(0.05, 0.95)	11	0.095 0	0.05	0.269 0	2.184 6	-3.722 9	-4.946 6
(0.06, 0.94)	13	0.112 8	0.06	0.307 9	2.563 3	-4.174 3	-5.599 3
(0.07, 0.93)	15	0.130 2	0.07	0.344 3	2.922 5	-4.580 9	-6.195 6
(0.08, 0.92)	17	0.147 2	0.08	0.378 5	3.262 3	-4.948 9	-6.742 2
(0.09, 0.91)	19	0.163 8	0.09	0.410 8	3.582 7	-5.284 8	-7.245 9
(0.10, 0.90)	21	0.180 0	0.10	0.441 2	3.883 7	-5.588 6	-7.706 7
(0.40, 0.60)	81	0.480 0	0.40	0.835 3	7.832 9	-48.418 1	-35.606 2
(0.41, 0.59)	83	0.483 8	0.41	0.833 9	7.225 9	-47.682 6	-34.657 8
(0.42, 0.58)	85	0.487 2	0.42	0.831 4	6.579 6	-46.879 1	-33.635 9
(0.43, 0.57)	87	0.490 2	0.43	0.827 9	5.894 3	-45.993 8	-32.532 1
(0.44, 0.56)	89	0.492 8	0.44	0.823 2	5.169 7	-44.999 5	-31.330 2
(0.45, 0.55)	91	0.495 0	0.45	0.817 4	4.406 0	-43.910 0	-30.038 4
(0.46, 0.54)	93	0.496 8	0.46	0.810 3	3.608 1	-42.697 8	-28.640 3
(0.47, 0.53)	95	0.498 2	0.47	0.801 8	2.761 0	-41.349 5	-27.127 8
(0.48, 0.52)	97	0.499 2	0.48	0.791 8	1.879 9	-39.851 3	-25.492 6
(0.49, 0.51)	99	0.499 8	0.49	0.780 0	0.959 5	-38.162 5	-23.710 2
(0.50, 0.50)	101	0.500 0	0.50	0.766 0	0.000 0	-36.228 5	-21.748 0

表2 $\bar{S}_J(G)$, $\bar{S}_H(G)$, $\bar{S}(G)$, J 和 D 间的相关分析

相关系数	$\bar{S}_J(G)$	$\bar{S}_H(G)$	$\bar{S}(G)$	J	D
$\bar{S}_J(G)$	1.000 000	0.936 564	0.993 423	-0.992 351	0.992 351
$\bar{S}_H(G)$		1.000 000	0.970 536	-0.891 137	0.891 137
$\bar{S}(G)$			1.000 000	-0.973 324	0.973 324
J				1.000 000	-1.000 000
D					1.000 000

表3 $\bar{S}_J(G), \bar{S}_H(G), \bar{S}'(G), J$ 和 D 间的偏相关分析

偏相关系数	$\bar{S}_J(G)$	$\bar{S}_H(G)$	$\bar{S}'(G)$	J	D
$\bar{S}_J(G)$	1.000 000	-0.999 303	0.999 709	-0.036 334	0.036 334
$\bar{S}_H(G)$		1.000 000	0.999 905	-0.000 196	0.000 196
$\bar{S}'(G)$			1.000 000	0.012 416	-0.012 416
J				1.000 000	-1.000 000
D					1.000 000

4 结论与讨论

一个具有多态性的非平衡位点,存在着基因水平与基因型水平上的变异,如何客观地测量变异的程度,是多样性研究与应用的基础。

1) D 与 $\bar{H}$ 的差异表明,文献[1]、[2]、[4]、[5]等用 D 作为基因多样度的测量指标是合理的,用 $\bar{H}$ 不能测量非平衡群体的基因变异。

2) D 与 $\bar{S}'(G), \bar{S}_J'(G), \bar{S}_H'(G)$ 的强正相关性表明, $\bar{S}'(G), \bar{S}_J'(G), \bar{S}_H'(G)$ 也可作为度量群体遗传变异的指标。 D 的取值范围为 $[0, \frac{K-1}{K}]$, $\bar{S}'(G)$ 的取值范围为 $[0, 1]$, 它们的遗传学意义是相同的,都可描述该位点的基因型不肯定性程度。在实际应用时,当群体中某些基因型频率过小 (≤ 0.03) 时,由于 $\bar{S}'(G)$ 与 $\bar{S}'(G)$ 间存在显著差异,因此,该群体的 $\bar{S}'(G)$ 就成为表明其基因型变异的恰当尺度。

3) 由表1模拟结果, J 与 D 的简单相关与偏相关系数均为 -1 , $\bar{S}_J'(G)$ 与 $\bar{S}_H'(G)$ 的简单相关系数为 $0.936\ 564$, 而偏相关系数为 $-0.999\ 303$, 说明 J 与 $D, \bar{S}_J'(G)$ 与 $\bar{S}_H'(G)$ 在本质上是対立的, 但 $\bar{S}_J'(G)$ 与 $\bar{S}_H'(G)$ 的简单相关却为正, 其原因有待进一步研究。 J 和 D 与 $\bar{S}_J'(G), \bar{S}_H'(G)$ 和 $\bar{S}'(G)$ 间仅有微弱的偏相关, 因而是两个独立的指标体系, 且后者还有其信息论内涵。

4) D 虽然可作为度量非平衡群体遗传变异的指标, 但由于它只利用了基因资料, 而没有利用基因型的信息, 因而本质上是用平衡群体作为同一基因库的代表。这种方法本身存在对信息的损失, 反映在用 D 来度量具有同一基因库的所有群体的基因型变异时, 其数值与 $\bar{S}'(G)$ 比较显著偏小。

5) $\bar{S}_J'(G)$ 与 $\bar{S}_H'(G)$ 可实现对基因型水平上的遗传变异进行剖分。 $\bar{S}_J'(G)$ 反映的是纯合基因型遗传变异对总变异的贡献, $\bar{S}_H'(G)$ 反映的是杂合基因型的遗传变异对总变异的贡献。这种剖分可为相关遗传现象的解释, 遗传变异的定向控制, 生物多样性的保持机制及保持方法提供理论依据。

6) 对非平衡群体, 其遗传变异不可忽视 $\bar{S}'(G)$ 。也就是说, 当杂合体很少时, 遗传变异主要由纯合体引起, 用 $\bar{S}_J'(G)$ 或 $\bar{S}'(G)$ 度量群体遗传变异都是合适的。对基因型频率适中的群体 (尽管不平衡), $\bar{S}'(G)$ 是度量遗传变异的恰当尺度。

7) 无论是平衡群体还是非平衡群体, 基因水平上的变异与基因型水平上的变异都可

以用 $S'(G)$ 来表示。

5 实 例

文献[2]中用大麦复合杂交组合 V 第 4 代的酯酶基因座 A, B 和 C (即 Est1, Est2 和 Est4) 的一些数据^[6]证实, 对于自交物种, 基因多样性作为变异性的指标要优于杂合性。然而, 应用相对信息量指标体系, 会得出更为清晰的结果。

大麦 3 个酯酶基因座基因型频率估测结果及基因型变异测度计算结果见表 4, 5。

表 4 大麦 3 个酯酶基因座基因型频率估测结果

基因座	Est1			Est2			Est4		
基因型	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
基因型频率	0.440 0	0.064 0	0.496 0	0.886 5	0.038 0	0.075 4	0.656 4	0.091 6	0.252 0

表 5 大麦 3 个酯酶基因座基因型变异测度计算结果

基因座	Est1				Est2				Est4			
基因型	D	$S'(G)$	$S'_J(G)$	$S'_H(G)$	D	$S'(G)$	$S'_J(G)$	$S'_H(G)$	D	$S'(G)$	$S'_J(G)$	$S'_H(G)$
变异测度	0.498 4	0.905 6	0.645 4	0.160 2	0.171 0	0.387 8	0.274 5	0.113 3	0.418 2	0.767 0	0.567 7	0.199 3

表 5 表明, 3 个位点的 D 与 $S'(G)$ 排序一致, 说明 D 与 $S'(G)$ 在描述基因变异上是一致的, 显然 D 比相应的 $S'(G)$ 小, 因此, 用 $S'(G)$ 来描述基因变异度比用 D 更为清晰。在描述遗传变异方面, 相对信息量还有它的特别之处, 即可实现对基因型水平上变异的剖分。如 Est2 位点, 总遗传变异度为 38.78%, 其中纯合体变异为 27.45%, 杂合体变异为 11.33%。上述 3 个位点上纯合体频率分别为 0.936 0, 0.961 9 与 0.908 4, 可见纯合体频率远大于杂合体频率, 其遗传变异主要是由于纯合体的连续存在所致, 因而表现出 $S'_J(G)$ 也远大于 $S'_H(G)$, 对这类群体, 仅用 $S'_J(G)$ 来度量遗传变异也是合适的。

〔参考文献〕

- [1] 根井正利. 分子群体遗传学与进化论[M]. 王家玉译. 北京: 农业出版社, 1983. 122—123.
- [2] [美] Weir B S. 遗传学数据分析——群体遗传学离散型数据分析方法[M]. 徐云碧、王志宁、愈志华译. 北京: 中国农业出版社, 1996. 115—134.
- [3] 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.
- [4] 郭平仲. 群体遗传学导论[M]. 北京: 农业出版社, 1993. 314—326, 362—373.
- [5] 戴君锡, 王身立. 遗传分析导论[M]. 长沙: 湖南科学技术出版社, 1989. 212—217.
- [6] Weir B S, Allard R W, Kahler A L. Analysis of complex allozyme polymorphisms in a barley population[J]. Genetics, 1972, 72: 505—523.

Shannon entropy method in disequilibrium population gene variation measurement

GUO Man-cai, SONG Shi-de, ZHOU Jing-yu, YUAN Zhi-fa

(Department of Basic Sciences, Northwest Science and Technology University of
Agriculture and Forestry, Yangling, Shaanxi 712100, China)

Abstract: Based on Shannon information theory, the relative information entropy of population genotype $S'(G)$, homozygote $S'_I(G)$ and heterozygote $S'_H(G)$ were set up for the disequilibrium population, and gave it the genetic meaning. The theoretical comparison between gene identity J and gene diversity D showed that they had very good uniformity in numerical regularity, but it was an independent index system, with each relative information entropy having better character. $S'(G)$ could not only convey gene variation, but also indicated genetic variation of genotype level. The $S'_I(G)$ mainly reflected the genetic variation of homozygote, and $S'_H(G)$ the genetic variation of heterozygote. Every relative information entropy could reflect the genetic variation degree of population and distinguish the genetic variation degree among different loci.

Key words: disequilibrium population; Shannon information entropy; gene identity; gene diversity; population genotype entropy