相对 Shannon信息量与基因变异的测量

袁志发 郭满才 宋世德 边宽江 周静芋
(西北农业大学基础科学系,陕西杨凌 712100)

摘 要 在 shannon信息量的基础上,建立了相对纯合度信息量 $S'_J(A^2)$、相对杂合度信息量 $S'_H(A^2)$和相对信息量 $S'(A^2)$的概念,并赋予它们以遗传学意义。与纯合度 J和杂合度 H进行了理论比较,结果表明,二者在数量规律性上有很好的一致性,而各相对信息量有更好的性质。$1-S'_H$主要反映基因的确定度,$S'_H(A^2)$主要反映基因的变异度;$S'(A^2)$与 K无关,它能反映群体的遗传变异程度,亦能比较不同位点间的遗传变异程度。

关键词 Shannon信息量,遗传变异,纯合度,杂合度
分类号 S813.1

自然群体中存在大量的质量性状变异,这些变异在 DNA水平上是基因变异引起的。设群体是由一个复等位基因位点控制的,群体遗传学家认为纯合度 J和杂合度 H是测定基因变异的理想尺度,适用于平衡群体和非平衡群体[1,2]。 Lewontin以及 Selander和 Johnson[1]采用 Shannon信息指数(或称信息量、信息熵)来测量基因变异。 Nei[1]认为"这个指数是用于信息工程的信息量的测定,它与任何遗传物质无关。如何用遗传物质的术语来表达这个量的绝对值,还不很清楚。"本文拟对 J、H和 Shannon信息量在理论上进行了比较研究,以阐述 Shannon信息量的遗传学意义和应用。

1 J和 H及其性质

设复等位基因位点 A为:

$$(A_1, A_2, \cdots, A_K) = (P_1, P_2, \cdots, P_K) \tag{1}$$

其中 P_i为 A_i的频率,$P_i \geqslant 0$,且$\sum_{i=1}^{K} P_i = 1$.群体平衡时,各基因型的频率为:

$$(P_1A_1 + P_2A_2 + \cdots + P_KA_K)^2 = \sum_{i=1}^{K} P_i^2 A_iA_i + 2\sum_{i<j} P_iP_jA_iA_j \tag{2}$$

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.纯合度 J和杂合度 H分别定义为:

$$J = \sum_{i=1}^{K} P_i^2, \qquad H = 2\sum_{i<j} P_iP_j \tag{3}$$

显然有

$$J + H = 1 \tag{4}$$

式中,J为群体中纯合基因型所占的比例,H为杂合基因型所占的比例。

杂合基因型携带着不同的等位基因,代表着变异的存在,因而杂合度就成为估计群体内遗传变异的重要尺度。式(4)的成立与 K无关。 J与 H的另一个性质为:

收稿日期 1998-03-26
作者简介 袁志发,男,1938年生,教授,博士生导师

$$\frac{1}{K} \leqslant J \leqslant 1, \qquad 0 \leqslant H \leqslant \frac{K-1}{K} \tag{5}$$

它们很容易化为在$\sum_{i=1}^{K} P_i = 1$条件下的极值问题. 容易推出当 P_i之一等于 1而其他均为 0时,有 $J_{max} = 1$, $H_{min} = 0$;当 P_i均等于$\frac{1}{K}$时,有 $J_{min} = \frac{1}{K}$, $H_{max} = \frac{K-1}{K}$. 显然由式(3)知二者的相关系数为 - 1.

2 相对 *Shannon* 信息量及其性质

据 *Shannon* 信息量的定义,复等位基因库 A的信息量为:

$$S(A) = -\sum_{i=1}^{K} P_i \ln P_i \tag{6}$$

$$0 \leqslant S(A) \leqslant \ln(K) \tag{7}$$

对于平衡群体(2)的信息量为:

$$S(A^2) = -\left(\sum_{i=1}^{K} P_i^2 \ln P_i^2 + 2\sum_{i<j} P_i P_j \ln P_i P_j\right) = -2\left(\sum_{i=1}^{K} P_i^2 \ln P_i + \sum_{i<j} P_i P_j \ln P_i P_j\right) \tag{8}$$

且有

$$S(A^2) = 2S(A) \tag{9}$$

事实上 $S(A^2) = -\sum_{i=1}^{K}\sum_{j=1}^{K} P_i P_j (\ln P_i + \ln P_j) = -2\sum_{i=1}^{K} P_i \ln P_i = 2S(A)$

在上述表述中,$\ln P_i$在信息论中往往用 $lb\ P_i$,这只能影响信息量的单位,并不影响信息量的性质. 另外在式(8)中,$2\sum_{i<j} P_i P_j \ln P_i P_j$不写成 $2\sum_{i<j} P_i P_j \ln 2P_i P_j$,是因为把正反交分开来的缘故. 由(9)与(7)有

$$0 \leqslant S(A^2) \leqslant 2\ln K \tag{10}$$

据式(7)和(10)可定义基因库(1)和群体(2)的相对信息量 $S'(A)$和 $S'(A^2)$:

$$S'(A) = \frac{1}{\ln K} S(A), \qquad S'(A^2) = \frac{1}{2\ln K} S(A^2) \tag{11}$$

显然有

$$S'(A^2) = \frac{1}{2\ln K} S(A^2) = \frac{1}{2\ln K} 2S(A) = S'(A) \tag{12}$$

对于相对信息量有:

$$0 \leqslant S'(A^2) = S'(A) \leqslant 1 \tag{13}$$

有了式(12), $S'(A^2)$的计算可由 $S'(A)$代替. 另外,也与式(3)相对应,即 $0 \leqslant S'(A) \leqslant 1$与 K无关. $S(A^2)$为群体(2)和位点(1)的不肯定性的度量,则 $S'(A^2)$的意义为群体(2)和位点(1)的不肯定性程度,即它占最大不肯定性的比例.

为了和 J, H相对应,可定义相对纯合度信息量 $S'_J(A^2)$和相对杂合度信息量 $S'_H(A^2)$:

$$\begin{cases} S'_J(A^2) = -\dfrac{1}{\ln K}\displaystyle\sum_{i=1}^{K} P_i^2 \ln P_i \\ S'_H(A^2) = -\dfrac{1}{\ln K}\displaystyle\sum_{i<j} P_i P_j \ln P_i P_j \end{cases} \tag{14}$$

$S'_J(A^2)$ 的意义是指群体中纯合体的不肯定性占群体最大不肯定性的比例，$S'_H(A^2)$可仿此解释。显然

$$S'(A^2) = S'_J(A^2) + S'_H(A^2) \tag{15}$$

$S'_J(A^2)$和 $S'_H(A^2)$有如下性质：

$$0 \leqslant S'_J(A^2) \leqslant \frac{1}{K}, \qquad 0 \leqslant S'_H(A^2) \leqslant \frac{K-1}{K} \tag{16}$$

$S'_J(A^2)$与 $S'_H(A^2)$的非负性是显然的，当 P_i中之一为 1其余为 0时，它们都等于 0. $S'_J(A^2)$和 $S'_H(A^2)$的极大值问题可化为在基因频率之和等于 1时的条件极值(证明略)。

3 模拟结果与分析

对上述理论结果，在各基因频率步长取 0. 1时，进行计算机模拟。当 K= 3时，模拟了 66组；当 K= 5时，模拟了 1 001组。模拟结果与理论结果完全相符。在对模拟数据进行相关分析时，得出了如下有意义的结果(表 1、表 2)：

从简单相关看，$S'_J(A^2)$、$S'_H(A^2)$和 $S'(A^2)$均与 J有强的负相关，与 H有强的正相关；$S'_J(A^2)$、$S'_H(A^2)$、$S'(A^2)$间有强的正相关。从偏相关看，$S'_J(A^2)$与 $S'_H(A^2)$为 - 1；$S'_J(A^2)$、$S'_H(A^2)$与 $S'(A^2)$均为 1；$S'_J(A^2)$、$S'_H(A^2)$与 J、H的偏相关为负，$S'(A^2)$与 J、H的偏相关为正，且随着 K的增大而加强。

表 1 S'_J, S'_H, S', J, H间的相关分析

K	相关系数	S'_J	S'_H	S'	J	H
3	S'_J	1. 000 000	0. 793 241	0. 904 315	- 0. 970 436	0. 970 436
	S'_H		1. 000 000	0. 977 261	- 0. 914 759	0. 914 759
	S'			1. 000 000	- 0. 979 209	0. 979 209
	J				1. 000 000	- 1. 000 000
5	S'_J	1. 000 000	0. 516 398	0. 682 630	- 0. 849 083	
	S'_H		1. 000 000	0. 987 968	- 0. 911 704	0. 911 704
	S'			1. 000 000	- 0. 963 763	0. 963 703
	J				1. 000 000	- 1. 000 000

表 2 S'_J, S'_H, S', J, H之间的偏相关分析

K	偏相关系数	S'_J	S'_H	S'	J	H
3	S'_J	1. 00 000	- 1. 000 000	1. 000 000	- 0. 000 584	- 0. 000 584
	S'_H		1. 000 000	1. 000 000	- 0. 000 584	- 0. 000 584
	S'			1. 000 000	0. 000 584	0. 000 584
	J				1. 000 000	- 1. 000 000
5	S^J	1. 000 000	- 1. 000 000	1. 000 000	- 0. 154 990	- 0. 154 990
	S'_H		1. 000 000	1. 000 000	- 0. 154 990	- 0. 154 990
	S'			1. 000 000	0. 154 990	0. 154 990
	J				1. 000 000	- 1. 000 000

4 结论与讨论

对一个位点的平衡群体来讲，基因型有两种或两种以上者称为多型性基因位点；否则

称为单型性基因位点.多型性位点是有基因变异的,基因变异又意味着遗传变异的存在.

1) $S'_H(A^2)$与 H具有类似的性质,其取值范围均为 [0, (K- 1) /K].当位点基因库中各等位基因频率相等时,它们同时取极大值 (K- 1) /K,当基因库中只有一个基因时,它们同时取最小值 0,因而它们的遗传学意义是相同的,都可描述位点内各等位基因或相应平衡群体中各杂合基因型的不肯定性程度,都可以作为位点基因变异的测量尺度.从相关性上看,二者具有强的正相关 (K= 3时为 0.914 6, K= 5时为 0.911 7).对于具有多个位点的平衡群体,它们各自的平均值均可以表示基因多样度.

2) 1- $S'_H(A^2)$与 J具有类似的性质,其取值范围均为 [1/K, 1].当位点基因库中各等位基因频率相等时,它们同时取最小值 1/K,当基因库中只有一个基因时,它们均取最大值 1,因而它们都描述了位点内各等位基因的确定性程度,都可以作为位点内各等位基因的确定性程度的测量尺度.从相关上看,二者具有强的正相关 (K= 3时为 0.914 6, K= 5时为 0.911 7).对于具有多个位点的平衡群体,它们各自的平均值都可作为基因一致度的测量尺度.

3) $S'_J(A^2)$反映的是位点的平衡群体中纯合基因型的不肯定性程度,其取值范围为 [0, 1/K],它与 J的意义相反.由于随着 K的增大, $S'_J(A^2)$的取值范围 [0, 1/K]愈来愈窄, J的取值范围 [1/K, 1]则愈来愈宽,因而其负相关程度很不稳定 (K= 3时为 - 0.970 4, K= 5时为 - 0.849 1),因此 $S'_J(A^2)$和 J的遗传学意义是不同的.

4)作为遗传变异的描述,最好的指标是遗传方差,它表示了遗传型值参差不齐的程度.对于可以量化的质量性状,如 (A, a)= (p, q)的平衡群体的遗传方差为 2pq(当 p= q= 1/2时最大),它是基因库的不肯定性度量,亦是群体各基因型间的不肯定性度量.因而, $S'(A^2)$既然是位点平衡群体的不肯定性度量,当然可以作为群体遗传变异程度的度量.这样 $S'_J(A^2)$、 $S'_H(A^2)$分别作为群体中纯合基因型和杂合基因型的遗传变异程度就是很合理的了.

5) J与 H的简单相关与偏相关系数均为 - 1. $S'_J(A^2)$与 $S'_H(A^2)$的简单相关系数为正,而偏相关为 - 1; J、 H与 $S'_J(A^2)$、 $S'_H(A^2)$间有微弱的偏相关,说明二者本质上有同样的作用但又是几乎独立的两个指标系统.

6)对于具有多个位点的平衡群体,可以建立相应的平均相对信息量.

5 实　例

以文献 [3]中辽宁绒山羊群体血液蛋白位点基因频率资料 (表 3)为例,其基因变异测度计算结果见表 4.

表 3　辽宁绒山羊两个血液蛋白位点基因频率估测结果

群　体	亚群	*Tf*		*Pep-B*		
		Tf^A	Tf^B	$Pep\text{-}B^A$	$Pep\text{-}B^B$	$Pep\text{-}B^C$
辽宁绒山羊	Ⅰ	0.792	0.208	0.958	0.000	0.042
	Ⅱ	0.958	0.042	0.958	0.042	0.000
	Ⅲ	0.944	0.056	0.991	0.009	0.000

表 4 辽宁绒山羊两个血液蛋白位点基因变异测度计算结果

群 体	亚群	*Tf*						*Pep-B*					
		J	H	$1-S'_H$	S'_J	S'_H	S'	J	H	$1-S'_H$	S'_J	S'_H	S'
辽 宁 绒山羊	Ⅰ	0.671	0.329	0.571	0.309	0.429	0.738	0.920	0.980	0.882	0.041	0.118	0.159
	Ⅱ	0.920	0.080	0.814	0.065	0.186	0.251	0.920	0.080	0.882	0.041	0.118	0.159
	Ⅲ	0.894	0.106	0.776	0.087	0.224	0.311	0.982	0.018	0.962	0.009	0.038	0.047

表 4表明,两个位点三个亚群的 J与 $1-S'_H$排序一致, H与 S'_H一致。这说明 J与 S'_H在描述基因确定度上是一致的, H与 S'_H在描述基因变异度上是一致的。在描述遗传变异方面,各相对信息量还有它的特别之处,如亚群Ⅰ 的 *Tf* 位点, $S'=0.738$,说明遗传变异达到最大变异的 73.8%,其中纯合体占 30.9%,杂合体占 42.9%;对于 *Pep-B* 位点,遗传变异达到最大变异的 15.9%,其中纯合体占 4.1%,杂合体占 11.8%.由这两个位点可见,来源于杂合体的遗传变异大于纯合体

参 考 文 献

1 根井正利著;王家玉译.分子群体遗传学.北京: 农业出版社, 1983. 122~ 123

2 [美]*Bruce S, Weir* 著;徐云碧,王志宁,俞志华译.遗传学数据分析—— 群体遗传学离散数据分析方法.北京: 中国农业出版社, 1996. 115~ 125

3 孙金梅,常 洪,秦国庆,等.山羊群体血液蛋白位点遗传分化研究.西北农业大学学报, 1998, 26(1): 21~ 25

Relative Shannon Information Capacity and Gene Variation Measurement

Yuan Zhifa Guo Mancai Song Shide Bian Kuanjiang Zhou Jingyu

(*Department of Basic Science, Northwestern Agricultural University, Yangling, Shaanxi* 712100)

Abstract Based on Shannon information theory, the relative information capacity of homozygosity $S'_J(A^2)$ and heterozygosity $S'_H(A^2)$, relative information capacity $S'(A^2)$ were set up, also meanings in genetics given to them, then theorical comparison with homozygosity J and heterozygosity H made. The results showed that they had very good uniformity in numerical regularity, each relative information capacity had better character. The $S'_J(A^2)$ mainly reflects gene identity and the $S'_H(A^2)$ mainly reflected gene diversity; $S'(A^2)$ had no relation to K, which could reflect the genetic variation degrees of population, and also could distinguish the genetic variation degrees among different loci.

Key words Shannon information capacity, genetic variation, homozygosity, heterozygosity