遗传距离的错误观点和病毒进化踪迹的探索

杜念兴[1**]，潘芹芹[2]，张素芳[1]

(1 南京农业大学动物医学院 南京 210095 ;2 江苏省人民医院 HLA 实验室 南京 210029)

The Wrong Idea of Genetic Distance and the Exploration of the Trace of Virus Evolution

DU Nian-xing[1**], PAN Qin-qin[2], ZHANG Su-fang[1]

(1. *The Animal Medicine College of Nanjing Agricultral University, Nanjing* 210095, *China*; 2. *The HLA lab of Jiangsu Province Hospital, Najing* 210029, *China*)

Abstract: In the course of calculation of genetic distance of virus gene, p and q should be comprehended of the transition rate and transversion rate. The result K should response the time concept of evolution speed if K would be a percent. So genetic distance were proposed to change into evolution per cent, then exploring expediently the trace of virus evolution from the disciplinary change of evolution rate.

Key words: Genetic distance; Evolution per cent; Evolution speed; Transition; Transversion

摘要:通过对病毒基因序列的遗传距离(K)的演算,发现公式中的 p 和 q 应理解为转换率和颠换率;计算所得的 K 值是一个百分率,更适合于显示生物进化的时间概念。因此建议将遗传距离改名为进化率。并从进化率计算中的规律性变化,探索了病毒进化的信息。

关键词:遗传距离;进化率;进化速率;转换;颠换

中图分类号:Q7 **文献标识码**:A **文章编号**:1003-5125(2005)03-0335-05

作者从事兔出血症(RHD)研究,断断续续近 20 年,为弄清该病毒的起源和流行规迹,先后对 10 株来自国内外不同时间,不同地区从 RHD 分离的免嵌杯病毒(RCV)VP60 基因序列,进行遗传距离(K)的计算[1]。在计算中发现,对该公式中 K、p、q 的理解和对遗传距离的认识都存在一些疑点。对此,进行了分析和研讨,并进一步探索了病毒进化的信息。

1 问题的发现

计算遗传距离和绘制进化树,通常是将所测病毒的基因序列输入相应的电脑软件,即可完成:序列比较、同源性、遗传距离和进化树等图表,既快速又正确,很少用人工运用的。作者为了进一步认识遗传距离的性质,参照文献[2]所示的方法进行运算。

遗传距离(K)的计算公式如下:

$$K = \frac{1}{2}\ln[(1-2p-2q)(1-2q)^{\frac{1}{2}}] \quad \cdots\cdots\cdots\cdots (1)$$

上式 p 为转换,即胞嘧啶(C)与胸腺嘧啶(T),或腺嘌呤(A)与鸟嘌呤(G)之间的变换;q 为颠换,即嘧啶和膘呤间的变换。

当我们按上述公式来计算各毒株间遗传距离(K)时,发现如将公式中的 p 和 q 理解为转换和颠换的核苷酸数时,则该公式无法计算,因 1-2q 是个负数,不能开方。后将 p 和 q 理解为百分率,并将转换和颠换的核苷酸数用大写的 P 和 Q 表示,所测序列核苷酸总数以 T 表示。

将(1)式改为(2)式:

$$K\text{率}(\%) = \frac{1}{2}\ln[(1-2P/T-2Q/T)(1-2Q/T^{\frac{1}{2}})] \times 100 \quad \cdots\cdots\cdots\cdots (2)$$

这样计算结果和电脑计算完全一样,但电脑计算没有加百分符号,是错误的。

遗传距离(Genetic disdance)是反映空间关系的一个实数,不能用来显示生物进化的时间效应,因

收稿日期:2004-11-03,修回日期:2004-12-15

** 通讯作者:杜念兴(1922-),男,浙江青田籍,教授,主要从事兽医病毒学及免疫学研究。
Corresponding author. Tel: 025-84395844, E-mail: dunianxing@hotmail.com.

而提出应将其改称为进化率(evolution percent),即K率。

2 遗传距离改名为进化率的理由

查阅最早提出该公式的 Kimura M. 论文[3],他将公式中的 K 命名为"进化距离百分率(evolutionary distances per cent)",与本文作者提出的进化率仅一字之差。而 Kimura M. 最初加入"距离"一字的目的是与他同时提出的另一名词"进化速率"(evolution speed)一词区别。后者是用于2个已知分离时间的毒株之间,用以计算其进化速率。

Kimura M. 没有想到他的后人把他定的名词中去掉"进化"一字,配之以"遗传"改成"遗传距离"(genetic disdance)。这样一来与 K 的原意就背道而驰了。"遗传"是个静态词,是指生物进化的某个时期的特性,包括生物的各种性状、结构、基因的序列等,都是固定的,可以遗传的。而"进化"一词则是动态的。生物进化论的奠基人达尔文指出生物进化是遗传、变异和选择三种因素综合作用的过程,明确指出进化是一个过程。遗传和进化是不同属性的词,说明用"遗传"替代进化的不合理性。

其次是"距离"一词的不合理性。距离是反映空间关系的用词,以公里、米、厘米、毫米、微米、纳米为单位。而进化则与时间相关,以年、月、日、时、分、秒等为单位,不能用"距离"表示。

最后说说不用百分符号的不合理性。K 在运算时用的是百分率,而运算结果不用百分符号,这是不合理性之一。变异率、同源性等词均用百分率表示,而同样与测序总数相关的"遗传距离"不用百分符号,这是不合理性之二。

综上所述"遗传距离"一词的不合理性已经不言而喻。这就是本文作者提出改用"进化率"一词,并将 K 改为 K 率的充分理由。

3 进化率运算中的规律性

3.1 变异相同时 P、Q 不同对 K 率的影响

为了弄清进化率(K)计算中的规律性变化。我们将所测基因的核苷酸总数假设为 1000 和 500。对变异率自 1.0%～10.0%的转换数(P)和颠换数(Q)的各种可能性进行了测算。在大量运算中发现,总数 1000 和 500,计算结果出人意外的一致。这就是说,无论所测基因核苷酸数多少,其变化规律完全一样。呈现一种非常有趣的趣味数学现象。

鉴于同一型或亚型的病毒,变异率一般在 10% 以内,而转换数(P)远大于颠调数(Q),Q 占总变异数的百分率,称为 Q 率,一般在 30% 以内。根据这些基本规律,将核苷酸总数为 1000 时的 K 率计算见表 1。

从表 1 可以看出,在同一变变率时,随着 P 和 Q 的变化,$1-2p-q$(A)和$(1-2q)^{\frac{1}{2}}$(B)亦表现规律性地递增或递减。首先是 A 的变化规律:在变异率为 1.0%,Q 为 0 时,A 为 0.98;在变异率为 2.0% 时 A 为 0.96;变异率每增加 1 个百分点,A 则递减 0.02。在 Q 为 1 时,q = 0.001,A = 0.98 + 0.001 = 0.981。依次类推,Q 每增加 1,A 递增 0.001。B 的变化正好相反。Q 为 0 时,B = 1.0,Q 为 1 时,B = $1.0^{4/}$ 0.001 = 0.999,依次类推,Q 每增加 1,B 递减 0.001。

K 率的变化也非常有规律。在 Q 为 0 时,K 率最大,以后随 Q 的增加而递减。变异率 < 2.0% 时,变异率 = K 率;变异率从 3.0%～7.0% 时,K 率为 3.1～7.4%,变异率每递增 1.0%,K 率递增 0.1%,此时 Q 率均在 30% 范围内变化,不影响 K 率。变异率 8.0%～10.0%,Q 为 0 时,K 率为 8.7～11.1%,变异率每递增 1.0%,K 率递增 0.2%。变异率 8.0% 时,Q 率每增加 15%,K 率递减 0.1%。变异率9.0% 时,Q 率每增加 10%,K 率递减 0.1%。从表 2 可以看出 Q 对 K 率的影响随着变异率的增加而加大。

3.2 Q 为零时变异率与 K 率的变化

表 1 只计算了变异率在 10.0% 以下整数的 K 率,下面我们计算所有的小数点 1 位的变异率,在 Q = O 时的 K 率,结果见表 2。

从表 2 可以看出,变异率 < 3.0% 时,变异率 = K 率;自 2.3～3.7%,K 率递增 0.1%。以后 K 率每递增 0.1% 的变异率分别为:3.8%～4.8%、4.9%～5.6%、5.7%～6.6%、6.7%～7.2%、7.7%～8.1%、8.2%～8.7%、8.8%～9.1%、9.2%～9.5%、9.6%～9.9%,K 率分别自 0.2%、0.3%…1.1%,递增的间隔越来越短。

4 用查表法计算进化率

从病毒基因库中,检索到 10 株国内外由 RHD 分离到的 RCV VP60 基因序列。以最早自本病的发生地无锡(WX)株为标准株,用查表法计算了 WX 与其它 9 个毒株,以及其中 2 对 Q 率最高的毒株间的进化率。查表法是先按变异率从表 2 中查出 Q 为零时的进化率,然后根据 Q 数计算 Q 率,从表 1 查出 Q 率增加时,需递减的百分率即成。结果见表 3。

表 1 变异率(V)相同转换数(P)和颠换数(Q)不同的 K 率的变化规律

Table 1 The change disciplinarian of K with the same variance per cent and different P and Q number

V %	P	p	Q	q	1-2p-q	$(1-2q)^{\frac{1}{2}}$	K %	K-V %
1.0	10	0.01	0	0	0.98	1.0	1.0	0
	9	0.009	1	0.001	0.981	0.999	1.0	0
	8	0.008	2	0.002	0.982	0.998	1.0	0
	:	:	:	:	:	:	:	:
2.0	20	0.02	0	0	0.96	1.0	2.0	0
	19	0.019	1	0.001	0.961	0.999	2.0	0
	18	0.018	2	0.002	0.962	0.998	2.0	0
	:	:	:	:	:	:	:	:
3.0	30	0.03	0	0	0.94	1.0	3.1	0.1
	29	0.029	1	0.001	0.941	0.999	3.1	0.1
	:	:	:	:	:	:	:	:
	20	0.020	10	0.01	0.95	0.99	3.1	0.1
4.0	40	0.04	0	0	0.92	1.0	4.2	0.2
	:	:	:	:	:	:	:	:
	27	0.037	13	0.013	0.933	0.987	4.1	0.1
5.0	50	0.05	0	0	0.9	1.0	5.3	0.3
	:	:	:	:	:	:	:	:
	35	0.035	15	0.015	0.915	0.985	5.2	0.2
6.0	60	0.06	0	0	0.88	1.0	6.4	0.4
	:	:	:	:	:	:	:	:
	40	0.04	20	0.02	0.90	0.98	6.3	0.3
7.0	70	0.07	0	0	0.86	1.0	7.5	0.5
	:	:	:	:	:	:	:	:
	49	0.049	21	0.021	0.881	0.979	7.4	0.4
8.0	80	0.08	0	0	0.84	1.0	8.7	0.7
	:	:	:	:	:	:	:	:
	68	0.068	12	0.12	0.852	0.988	8.6	0.6
	:	:	:	:	:	:	:	:
	56	0.056	24	0.24	0.864	0.976	8.5	0.5
9.0	90	0.09	0	0	0.82	1.0	9.9	0.9
	:	:	:	:	:	:	:	:
	80	0.08	10	0.16	0.83	0.99	9.8	0.8
	70	0.07	20	0.2	0.84	0.98	9.7	0.7
	60	0.06	30	0.3	0.85	0.97	9.6	0.6
10.0	100	0.1	0	0	0.80	1.0	1.1	1.1

表 2 Q 为零时变异率与 K 率的变化(%)

Table 2 Variance and change of K if Q amount zero (%)

V	K	I	V	K	I	V	K	I	V	K	I	V	K	I	V	K	I	V	K	I	V	K	I
1.0	1.0	0	2.0	2.0	0	3.0	3.1	0.1	4.0	4.2	0.2	5.0	6.3	0.3	6.0	6.4	0.4	7.0	7.5	0.5	8.0	8.7	0.7	9.0	9.9	0.9
1.1	1.1	0	2.1	2.1	0	3.1	3.2	0.1	4.1	4.3	0.2	5.1	5.4	0.3	6.1	6.5	0.4	7.1	7.6	0.5	8.1	8.8	0.7	9.1	10.0	0.9
1.2	1.2	0	2.2	2.2	0	3.2	3.3	0.1	4.2	4.4	0.2	5.2	5.5	0.3	6.2	6.6	0.4	7.2	7.7	0.5	8.2	9.0	0.8	9.2	10.2	1.0
1.3	1.3	0	2.3	2.4	0	3.3	3.4	0.1	4.3	4.5	0.2	5.3	5.6	0.3	6.3	6.7	0.4	7.3	7.8	0.5	8.3	9.1	0.8	9.3	10.3	1.0
1.4	1.4	0	2.4	2.5	0	3.4	3.5	0.1	4.4	4.6	0.2	5.4	5.7	0.3	6.4	6.8	0.4	7.4	8.0	0.6	8.4	9.2	0.8	9.4	10.4	1.0
1.5	1.5	0	2.5	2.6	0	3.5	3.6	0.1	4.5	4.7	0.2	5.5	5.8	0.3	6.5	6.9	0.4	7.5	8.1	0.6	8.5	9.3	0.8	9.5	10.5	1.0
1.6	1.6	0	2.6	2.7	0	3.6	3.7	0.1	4.6	4.8	0.2	5.6	5.9	0.3	6.6	7.0	0.4	7.6	8.2	0.6	8.6	9.4	0.8	9.6	10.7	1.1
1.7	1.7	0	2.7	2.8	0	3.7	3.8	0.1	4.7	4.9	0.2	5.7	6.1	0.4	6.7	7.2	0.5	7.7	8.4	0.7	8.7	9.6	0.9	9.7	10.8	1.1
1.8	1.8	0	2.8	2.9	0	3.8	4.0	0.2	4.8	5.0	0.2	5.8	6.2	0.4	6.8	7.3	0.5	7.8	8.5	0.7	8.8	9.7	0.9	9.8	10.9	1.1
1.9	1.9	0	2.9	3.0	0	3.9	4.1	0.2	4.9	5.2	0.3	5.8	6.3	0.4	6.9	7.4	0.5	7.9	8.6	0.7	8.9	9.8	0.9	9.9	11.0	1.1

* V 变异率(variance per cent), K 进化率(evolution per cent), I 递增率(increase by degrees rate)

表3 用查表法计算10个RCV VP60基因序列的进化率(%)

Table 3 evolution rate of 10 strains RCV VP60 gene sequence by checking table

	毒株 Virus strains	序列数 Sequence number	变异数 Variance number	变异率 Variance rate(%)	P数 P number	Q数 Q numbe	Q率 rate(%)	进化率 Evolution rate(%)	递减 Degression
WX	无锡 Wuxi/1984	1740	0	0	0	0	0	0	0
U4	法国 France/1988	1740	29	1.7	27	2	6.9	1.7	0
M6	德国 Germany/1993	1740	32	1.7	27	5	15.6	1.7	0
U5	奥地利 Austria/1996	1740	38	2.2	33	5	13.2	2.2	0
X8	意大利 Italy/1995	1740	46	2.6	37	9	19.6	2.7	0
Z4	西班牙 Spain/1989	1740	59	3.4	53	6	10.2	3.5	0
Z2	法国 France/1995	1740	66	3.8	63	3	4.5	4.0	0
AF	北美 North America/2000	1740	111	6.4	96	15	13.5	6.8	0
HB	哈尔滨 haerbin/2002	1740	117	6.7	96	21	17.9	7.2	0
AY	南京 nanjing/2003	1740	120	6.9	103	17	14.2	7.4	0
X8	and HB	1740	149	8.6	117	27	18.1	9.3	0.1
HB	and AY	1740	149	8.3	110	34	23.6	9.0	0.1

由表3可见,以WX为参照株时,10个毒株间的变异率均在7.0%以下,其Q率最低4.5%,最高19.6%,也都没有超过30%,因此毋须递减,由表2查出的数据,即为进化率。表中最后2行为Q率最高的2对毒株间的比较,其变异率均在8.0%～9.0%范围内,其Q率则在15.0～30.0%范围内。按表1规定,需递减0.1%。X8与HB 2毒株间的变异率为8.6%,Q=0时其进化率为9.4%,减去0.1%,则其进化率为9.3%。按同法计算,HB与AY之间的变异率为8.3%,Q=0时,其进化率为9.1%减0.1%,即9.0%。

所有以上结果与电脑计算完全一致,表明进化率(K)的计算公式虽很复杂,但只要掌握其规律性,可以简化到用查表法来计算进化率。当然这仅仅是兴趣数学在病毒进化研究中的一个例证,并不是想用它来替代电脑计算。

5 对病毒进化研究的启示

5.1 变异式的分析

在分析核苷酸变异和氨基酸变异的关系时,作者将密码子3个位点的变异情况,列成变异式:1:P、1:Q、2:P、2:Q、3:P、3:Q,":"前面为位点,后面为变异性质。RCV 10个毒株间的变异式,见表4。

P加Q的总变异数为644,3:P占总变异数的66.5%,3:Q占12.9%,3:P远高于3:Q。P、Q相加,第3位变异占79.4%,第1、2位的变异只占20.6%。

而氨基酸变异都集中在第1、2位变异,其中为1:P的78例中,有59例氨基酸变异,另19例不变者,均为L→L,占1:P的24.4%,比例相当高。

所有第2位变异的,包括2:P 30例和2:Q 15例均引起氨基酸变异,而所有第3位变异的,包括3:P 428例和3:Q 83例均不引起氨基酸变异。在全部核苷酸变异的644例中,氨基酸变异114例,占17.7%,说明核苷酸变异率远远高于氨基酸变异率。

表4 10个RCV VP60基因序列的变异式

Table 4 variance model of VP60 of 10 strains RCV

变异式 variance model	毒株 virus strains										合计 total
	WX	U4	M6	U5	X8	Z4	Z2	AF	HB	AY	
1:P	0	1	4	5	2	6	7	18	18	17	78
2:P	0	3	0	0	1	4	3	5	6	8	30
3:P	0	24	23	28	34	43	53	73	72	78	428
total	0	27	27	33	37	53	63	96	96	103	536
1:Q	0	0	1	2	0	2	1	2	0	2	10
2:Q	0	0	2	1	0	1	1	3	4	3	15
3:Q	0	2	2	2	9	3	1	10	17	12	83
total	0	2	5	5	9	6	3	15	21	17	108

5.2 稳定区和高变区

RCV衣壳蛋白(VP60)氨基酸序列可以分成A～F 6个区,C区27个氨基酸为铰链区,使VP60形成内层和外层。A、B 2区为内层,D、E、F为外层。在总数580个氨基酸中,有30个位点引起氨基酸变异,变异率为5.2%。稳定氨基酸占94.8%,每个变异位点的毒株数自1～6不等,其中4个以上的13个,为高变氨基酸,占43.3%。变异氨基酸主要集中在B区的末端第293到E区的371位的78位氨基酸中,有14位氨基酸变异,占总变异率的46.7%,且其中70个为高变氨基酸,说明这一区域为氨基酸高变区。不变异氨基酸分散在各个区,其中B区

278 个氨基酸中,有 51～124 共 75 位、第 136～205 共 70 位、209～293 共 85 位均无氨基酸变异为氨基酸稳定区。E 区 145 个氨基酸中只有 1 个位点,1 个毒株变异,也为氨基酸稳定区。

核苷酸变异没有明显的高变区和稳定区。在 1740 个核苷酸中共有 247 个核苷酸发生变异,占 14.2 %,它们分散在各个区。大多呈连续的或间隔 1～2 个氨基酸连成一片。连号的(即间隔 2 个核苷酸有 1 个变异的)177 个占 71.7 %,还有一部分间隔 3～4 氨基酸有 1 个核苷酸变异的为数不多,间隔较大的,包括间隔 5 个氨基酸以上的有 16 处,其中有 2 处比较突出:一是在 B 区中部 104 位与 123 位氨基酸(57 个核苷酸)中间只有 2 个核苷酸变异;再有在 F 区第 562 位氨基酸至最后 580 位氨基酸间 54 个核苷酸中只有 1 个核苷酸变异。因此,可以认为核苷酸变异并没有大片明显的稳定区。

每个变异位点的毒株数不等,只有 1 个的 94 株,2 个的 61 株,3 个的 56 株,4 个的 6 株,5 个的 10 株,6 个的 4 株,8 个的 3 株。此外还有一些特殊的例子,如同一位核苷酸有 2 种变异的有 4 例;同一氨基酸第 1 位和第 3 位核苷酸均变异的有 5 例;第 1 位和第 2 位、第 3 位变异的各 1 例,更奇特的有 1 例 3 个密码均发生了变异。

再深入分析变异的毒株类型,发现上述特殊变异大多为 AF、HB 和 AY 3 个毒株,而这 3 个毒株与标准株(WX)的变异核苷酸有 83 位是相同的,其共变率分别为:AF (83/111) 74.8 %、HB (83/117) 70.9 %和 AY(83/120) 69.2 %,充分说明它们间亲缘关系的密切程度。以上数字虽然很枯燥,但对病毒的进化、变异及其在分子病毒和分子流行病学上意义都是深远的。

5.3 病毒是研究进化的最佳模型

生物的遗传基础是基因,研究生物进化首先应从基因变异看,而基因变异与物种的世代和基因的复制周期有关。动物进化周期以性成熟年龄为一个周期;大多数植物每年开花结籽,通常以年为世代;而病毒复制周期短,通常以小时计,其基因组成又最简单。因此研究生物进化和基因变异的规律,通常以病毒为研究模型,这说明为什么进化率(即遗传距离)和进化速率的计算主要集中在分子病毒学研究中。但通常这些数据均用电脑计算,一切过程全由电脑代替,研究者只需对所得的结果进行分析,探索毒株间的亲缘关系和进化脉络,从而忽视了基因变异的基本规律。本文通过实际运算,发现了很多值得进一步深入研究的迹象。

通过对病毒变异式、高变区和稳定区以及共变性的分析,有以下几点值得深思:

(1) 核苷酸变异中转换(P)数远高于颠换(Q)数,这是可以理解的,因为嘧啶(C、T)、嘌呤(A、G)内部互换,因其化学结构相似消耗能量少,并且也较少引起氨基酸致变异,有利于遗传稳定性的保持。

(2) 密码子第 3 位的变异占极大优势,尤其是 3 :P 的变异几乎占总变异数的 3/4,这是因为 3 :P 不引起氨基酸变异。病毒基因组在复制中,核苷酸变异是随机的,有很多因第 1、2 位变异而引起氨基酸变异的个体,在生存环境的选择下被淘汰。虽然其淘汰率很高,但因病毒复制周期短,少数适应者能迅速增殖到所需水平。在 1 :P 中氨基酸不变的均为 L →L,其比例很高,也是因为只有这一组密码子第 1 位 C、T 互变时,不影响氨基酸变异。

(3) RCV 核苷酸变异分散而比较均匀,没有明显的大片不变异的稳定区,而氨基酸变异则明显地表现有高变区和稳定区。这是因为病毒的表面结构及其致病因子均与宿主细胞的环境密切相关,它们必须随宿主环境的改变而改变,这是在适者生存的前提下,遗传稳定性和变异性的矛盾统一。

(4) 共变和共变率的分析对照该毒株间的亲缘关系十分有用,在病毒的分子流行病上有广泛应用的前景。

参考文献

[1] 杜念兴, 徐为燕, 等,兔出血症研究[J],中国农业科学,1991, 22(1).

[2] 金 奇, 等. 医学分子病毒学[M]. 北京:科学出版社,2001.

[3] Thiel H J, Kimura M. Caliciviruses an overview [J]. Vet Microliol,1999,69:55-62.