3 核苷酸顺序中的进化变化

ＤＮＡ序列进化中的基本形式是核苷酸随时间的改变。这一过程值得详细考虑，因为核苷酸顺序中的变化在分子进化研究中，既用来估计进化的速率又用于重建生物进化的历史。然而，核苷酸替换的过程通常是极其缓慢的，以至它不可能在研究者所生存的时间里被观察到。因此，为了检出ＤＮＡ序列中的进化变化，我们依靠比较法，即让某一给定顺序与另一个与它在进化上过去有共同祖先的顺序比较。这种比较要用到统计学方法，其中几种将在本章中讨论。

3.1 ＤＮＡ序列中的核苷酸替换

前一章中，我们把进化过程描绘成一系列的基因替换，在这一过程中新等位基因以单个突变的形式产生，继而增加其频率，最终则在群体中固定。现在我们从不同的观点来看这一过程。我们注意到，要被固定的等位基因其顺序不同于它们所替代的等位基因。如果我们使用的时间尺度中一个时间单位比固定时间长，那么，任何给定基因座位上的ＤＮＡ顺序都将表现出连续变化。为此，研究一下一个ＤＮＡ序列中的核苷酸如何随时间改变将是有趣的。象以后我们要解释的那样，这一研究结果可被用来建立估计两序列间替换数的方法。

为了研究核苷酸替换的动力学，我们必须作出几项关于一个核苷酸被另一个替换的概率的假定。文献中已提出了许多这样的数学方案。我们将把讨论仅限制在那些最简单且最常用的方法上：朱克斯和坎托（Jukes 和 Cantor,1969）的一参数模型（one-parameter model）和木村(Kimura ,1980)的两参数模型(two-parameter model)。关于更一般的模型的评论，读者可参考李等(Li 等,1985a)的论述。

朱克斯和坎托的一参数模型

朱克斯和坎托模型的替换方案如图３－１所示。该模型假定替换在４种核苷酸类型中随机地发生。换言之，在变化方向上没有任何倾斜。例如，如果所考虑的核苷酸是Ａ，则它将以相同的概率改变

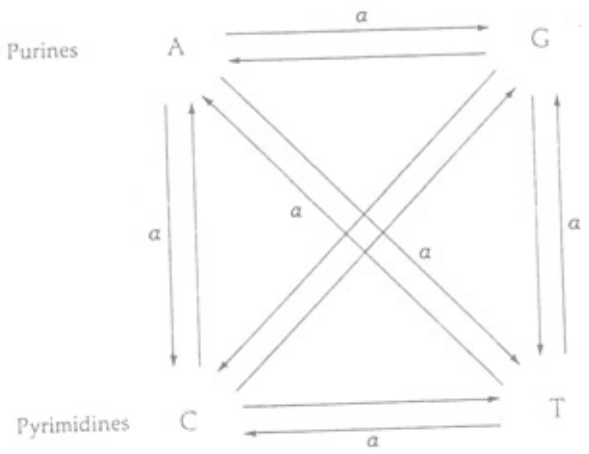


图３－１ 核苷酸替换的一参数模型。在此模型中，每一方向的替换速率都是α

成Ｔ、Ｃ或Ｇ。在此模型中，对每种核苷酸来说，替换速率为每单位时间３α，且３种可能的变化方向中每种的替换速率都是α。因为该模型只涉及一个参数α，所以它又叫一参数模型。

我们假定一个DNA序列中某一位点上座落的核苷酸在时刻 0 时为A。首先，我们要问：“该位点在时刻 t 时被A占据的概率是多少？”该概率用$P_{A(t)}$表示。

因为我们从A开始，所以该位点在 0 时刻被A占据的概率是$P_{A(0)}=1$。在时刻 1，该位点上仍为A的概率由

$$P_{A(1)}=1-3\alpha \tag{3.1}$$

给出，它反映出核苷酸保持不变的概率，即 1−3α。

在时刻 2 仍有 A 的概率为

$$P_{A(2)}=(1-3\alpha)P_{A(1)}+\alpha[1-P_{A(1)}] \tag{3.2}$$

为了得出此式，我们考虑两种可能的局面：（1）核苷酸保持不变，和（2）核苷酸已变成T、C或G但随后又回复到A（图 3－2）。在时刻 1 核苷酸为A的概率是$P_{A(1)}$，而在时刻 2 保持为A的概率是 1－3α。这两个独立变量的乘积给出了第一种局面的概率，它构成等式 3.2 的第一项。在时刻 1 核苷酸不是A的概率为 1－$P_{A(1)}$，而在时刻 2 变成A的概率为α。这两个概率的乘积给出了第二种局面的概率，它即等式 3.2 中的第 2 项。

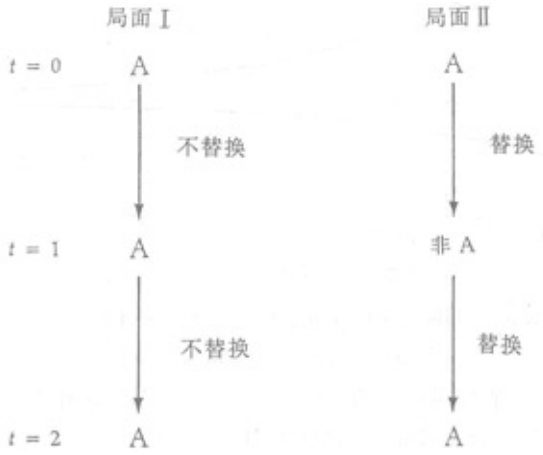


图 3 - 2 假定在时刻 0 某一位点上为A，而在时刻 2 该位点上仍有A的两种可能的局面[TS]]

用以上公式，我们可以证明，以下递推式可用于任何的 t：

$$P_{A(t+1)}=(1-3\alpha)P_{A(t)}+\alpha[1-P_{A(t)}] \tag{3.3}$$

我们可以按每单位时间$P_{A(t)}$的改变量重写等式 3.3，为：

$$P_{A(t+1)}-P_{A(t)}=-3\alpha P_{A(t)}+\alpha[1-P_{A(t)}] \tag{3.4a}$$

或

$$\Delta P_{A(t)}=-3\alpha P_{A(t)}+\alpha[1-P_{A(t)}]=-4\alpha P_{A(t)}+\alpha \tag{3.4b}$$

至此我们考虑的是一个离散的时间过程。不过，我们可以用连续时间模型来近似这一过程，把$\Delta P_{A(t)}$看成是时刻 t 时的变化率。以这一近似，等式 3.4b 被重写成

$$\frac{dP_{A(t)}}{dt}=-4\alpha P_{A(t)}+\alpha \tag{3.5}$$

这是一个一阶线性微分方程，其解由

$$P_{A(t)}=\frac{1}{4}+(P_{A(0)}-\frac{1}{4})e^{-4\alpha t} \tag{3.6}$$

给出。因为我们从 A 开始，所以$P_{A(0)}=1$。故而

$$P_{A(t)}=\frac{1}{4}+(\frac{3}{4})e^{-4\alpha t} \tag{3.7}$$

事实上，等式 3.6 不管在什么起始条件下都成立。例如，若该起始核苷酸不是 A，则$P_{A(0)}=0$，而在时刻 t 该位置上有 A 的概率为

$$P_{A(t)}=\frac{1}{4}-(\frac{1}{4})e^{-4\alpha t} \tag{3.8}$$

等式 3.7 和 3.8 对描绘替换过程来说是充分的。从等式 3.7,我们可以看到,如果起始核苷酸是 A,那么 $P_{A(t)}$ 将呈指数地从 1 降到 1/4(图 3-3)。另一方面,从等式 3.8 我们看到,如果起始核苷酸不是 A,那么 $P_{A(t)}$ 将从 0 单调地上升到 1/4。所以,不管起始条件如何,$P_{A(t)}$ 最终都将达到 1/4(图 3-3)。这对 T、C 和 G 而言也是正确的。因此，在朱克斯坎托模型下 4 种核苷酸中每种的平衡频率都是 1/4。达到平衡后，在概率上将没有进一步的变化，即，对所有 t 都有 $P_{A(t)}=P_{T(t)}=P_{C(T)}=P_{G(t)}=1/4$。然而，核苷酸的频率仅在无限长的 DNA 序列中保持不变。实际上，DNA 序列的长度是有限的，所以核苷酸频率上的波动看来是会发生的。

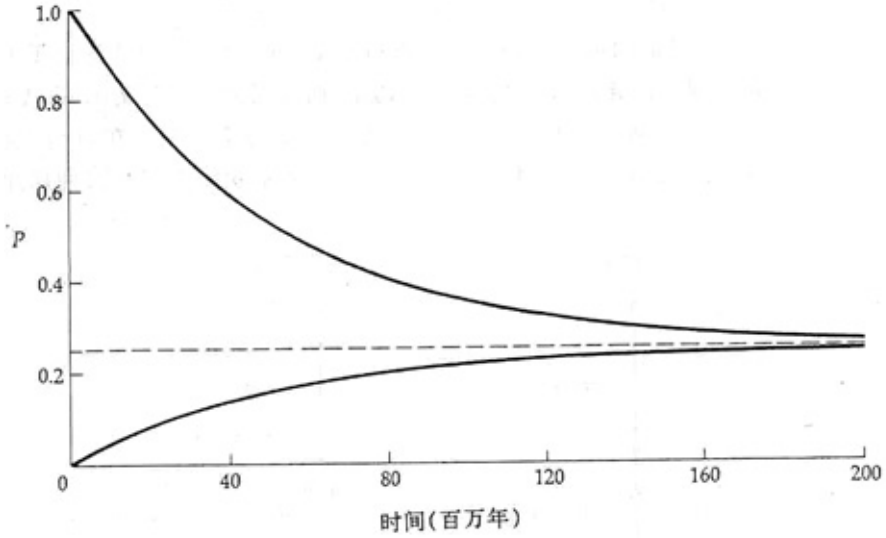


图 3 - 3 一个位置上有某一核苷酸的概率随时间的变化：由同样的核苷酸开始（上线）或由不同核苷酸开始（下线）。虚线表示平衡频率（0.25）。α = 5 × 10-9 核苷酸/位点/年

上面，我们的注意力集中在一个特定的核苷酸位点上，而把 $P_{A(t)}$ 处理为一种概率。然而，$P_{A(t)}$ 也可被解释成某一 DNA 序列中 A 的频率。例如，如果我们从一个仅由腺嘌呤构成的序列开始，那么 $P_{A(0)}=1$，而 $P_{A(t)}$ 则是在时刻 t 该序列中 A 的期望频率。

把起始核苷酸是 A 且时刻 t 的核苷酸还是 A 这一事实考虑进去，我们可以把等式 3.7 以更明确的形式重写成：

$$P_{AA(t)}=\frac{1}{4}+(\frac{3}{4})e^{-4\alpha t} \tag{3.9}$$

如果起始核苷酸是 G 而不是 A，那么由等式 3.8 我们得

$$P_{GA(t)}=\frac{1}{4}-(\frac{1}{4})e^{-4\alpha t} \tag{3.10}$$

因为在朱克斯-坎托模型下所有核苷酸都是等价的，所以 $P_{GA(t)}=P_{CA(t)}=P_{TA(t)}$。事实上，我们可以考虑一个一般性的概率，$P_{ij(t)}$，这是某一核苷酸在给定起始核苷酸为 i 的条件下在时刻 t 变为 j 的概率。应用这个一般化了的概念和等式 3.9，我们得

$$P_{ii(t)}=\frac{1}{4}+(\frac{3}{4})e^{-4\alpha t} \tag{3.11}$$

且由等式 3.10：

$$P_{ij(t)}=\frac{1}{4}-(\frac{1}{4})e^{-4\alpha t} \tag{3.12}$$

这里 $i\neq j$。

木村的两参数模型

朱克斯和坎托模型那样，假定所有核苷酸替换随机发生，这是不现实的。例如，转换（即 A 和 G 之间或 C 和 T 之间的变化）一般比颠换（即所有其他类型的变化）更频繁一些（第 4 章）。考虑到这一事实，木村（Kimura,1980）曾提出一个两参数模型，如图 3 — 4 所示。在此方案中，每一核苷酸位点上转换型替换的速率为每单位时间 α，而每种颠换型替换类型的速率则为每单位时间 β。

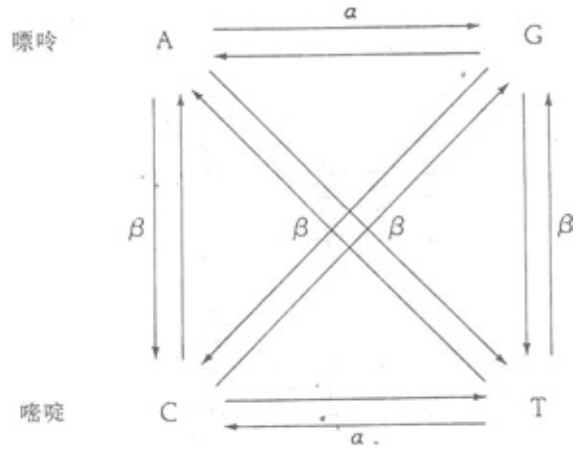


图 3 - 4 核苷酸替换的两参数模型。在此模型中，转换的速率（α）可能不等于颠换的速率（β）。

该模型比朱克斯　坎托模型复杂，而我们将只给出最后结果。从等式 3.11 我们看到，在朱克斯坎托模型中，某一位点上在时刻 t 时的核苷酸与时刻 0 时的相同的概率，对 4 种核苷酸来说是相同的。即，$P_{AA(t)}=P_{GG(t)}=P_{cc(t)}=P_{TT(t)}$。由于替换方案的对称，这种等同性对木村的两参数模型也是成立的。我们将用 X（t）表示该概率。可以证明：

$$X(t)=\frac{1}{4}+(\frac{1}{4})e^{-4\beta t}+(\frac{1}{2})e^{-2(\alpha+\beta)t} \qquad (3.13)$$

在朱克斯-坎托模型下，等式 3.12 不管从核苷酸 i 到核苷酸 j 的变化是转换还是颠换都成立。与之不同，在木村的两参数模型下，我们必须对转换和颠换两种变化加以区别。我们用 $Y_{(t)}$ 表示起始核苷酸和时刻 t 时的核苷酸经转换而互不相同的概率。我们看到，由于替换方案的对称，所以 $Y_{(t)}=P_{AG(t)}=P_{GA(t)}=P_{TC(t)}=P_{CT(t)}$ 可以证明

$$Y(t)=\frac{1}{4}+(\frac{1}{4})e^{-4\beta t}-(\frac{1}{2})e^{-2(\alpha+\beta)t} \qquad (3.14)$$

时刻 t 时的核苷酸与起始核苷酸经某一特定类型的颠换而互不相同的概率，$Z_{(t)}$，由下式给出：

$$Z(t)=\frac{1}{4}-(\frac{1}{4})e^{-4\beta t} \qquad (3.15)$$

注意，每种核苷酸只有一种转换类型，但却经受着两种类型的颠换。例如，若起始核苷酸是A，那么这两种可能的颠换变化即为A→C和A→T。因此，起始核苷酸与时刻 t 时的核苷酸经两种颠换类型之一的变化而互不相同的概率，将是由等式 3.15 给出的概率的两倍。还要注意，$X_{(t)}+Y_{(t)}+2Z_{(t)}=1$。

3.2 两 DNA 序列间的核苷酸替换数

一个群体中的等位基因替换一般要花成千甚至上百万年来完成（第二章）。为此，我们不能靠直接观察来处理核苷酸替换的过程，核苷酸替换常常是从那些有共同起源的ＤＮＡ分子的成对比较中推断出来的。

两个核苷酸序列相互分岐以后，每一个都将积累核苷酸替换。所以，自两序列发生分岐以来所出现的核苷酸替换数，就是分子进化中最通常用到的变量。

当两核苷酸序列间的分岐程度较小时，在任一位点上发生一次以上替换的机会可以忽略，则两序列间被观察到的差异数将接近实际替换数。另一方面如果分岐程度突出，那么，由于在同一位点上的多重替换（multiple substitution）或多次“击中”（multiple “hits”），观察差异数看来将小于实际替换数。例如，如果某一位点上的核苷酸，在一个序列中从A变到C再变到T，在另一个序列中则从A变到T，那么，尽管已发生了 3 次替换，但两序列在该位点上却是相同的（图３－５）。文献中已有几种修正这种偏差的方法被提了出来。

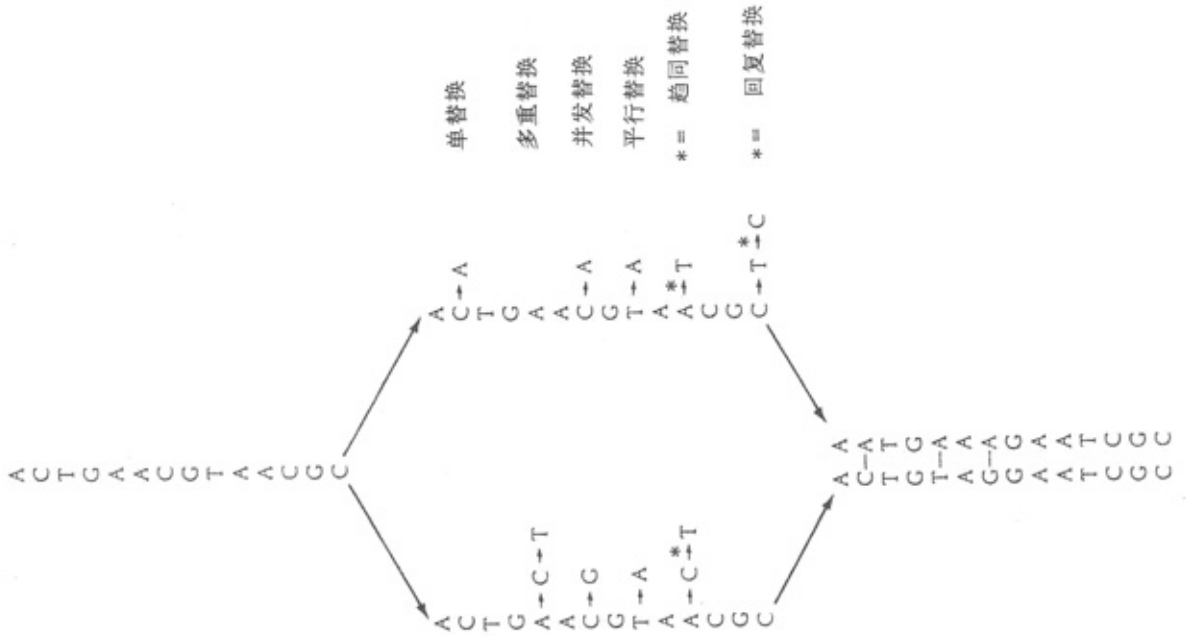


图 3 - 5 来自一个祖先序列且从它们开始分歧时起累积突变的两个同源ＤＮＡ序列。注意，虽然已经累积了１２次突变，但可被检出的差异却只有 3 个核苷酸位点。再注意，“并发替换”，“平行替换”，“趋同替换”和“回复替换”，都涉及同一位点上的多重替换，虽然这些替换可能发生在不同的品系中

替换数通常是以每核苷酸位点的替换数的形式表示，而不是以两序列间的总替换数表示。这有利于长度不相同的序列对间的分岐程度的比较。

为蛋白质编码的序列和非编码序列应分别处理，因为它们通常以不同的速率进化。在前一种情况下，建议将同义替换和非同义替换加以区别，因为已知它们是以显然不同的速率进化着的（第四章）。另一方面，在非编码区，则可假定所有位点以同样的速率进化。

两非编码序列间的替换数

我们在本章前部分对单个ＤＮＡ序列得到的结果，可用于研究两个有共同起源的序列间的核苷酸分岐。我们先从一参数模型开始。在该模型中，只考虑Ｉ$_{(t)}$就够了。Ｉ$_{(t)}$是在时刻ｔ某一给定位点上的核苷酸在两个序列中相同的概率。假定某一给定位点上的核苷酸在时刻 0 是Ａ。在时刻ｔ，一个后代序列在该位点上有Ａ的概率为Ｐ$_{ＡＡ(t)}$，因此两个后代序列在该位点上都有Ａ的概率则为Ｐ$_{ＡＡ(t)}^{2}$。类似地，两个序列在该位点上都有Ｔ、Ｃ或Ｇ的概率分别应为Ｐ$_{ＡＴ(t)}^{2}$，Ｐ$_{ＡＣ(t)}^{2}$和Ｐ$_{ＡＧ(t)}^{2}$因此，

$$I_{(t)} = P_{AA(t)}^2 + P_{AT(t)}^2 + P_{AC(t)}^2 + P_{AG(t)}^2 \qquad (3.16)$$

由等式 3.11 和 3.12，我们得

$$I_{(t)} = \frac{1}{4} + (\frac{3}{4})e^{-8\alpha t} \qquad (3.17)$$

等式 3.17 对Ｔ、Ｃ或Ｇ也成立。因此，不管某一位点上的起始核苷酸如何，Ｉ$_{(t)}$表示从ｔ时间单位以前开始分岐的两个序列间相同核苷酸的比例。注意，在时刻ｔ两序列在某一位点上不同的概率为Ｐ＝1－Ｉ$_{(t)}$。所以，

$$P = \frac{3}{4}(1 - e^{-8\alpha t}) \qquad (3.18a)$$

或

$$8\alpha t = -\ln(1 - \frac{4}{3}P) \qquad (3.18b)$$

两序列发生分岐的时间通常是未知的，这样我们就不能估出α。所以我们不去求它而去算Ｋ，Ｋ是两序列间自分岐以来的每位点替换数。在一参数模型的情况下，Ｋ＝２（３αｔ），这里３αｔ是两个品系的每一个中每位点的替换数。应用等式 3.18b，我们可算出Ｋ为

$$K = -\frac{3}{4}ln(1-\frac{4}{3}P) \tag{3.19}$$

其中 P 是两序列间不同核苷酸的比例(Jukes 和 Cantor,1969)。对于长为 L 的序列,取样方差由

$$V(k) = \frac{P(1-P)}{L\left(1-\frac{4}{3}P\right)^2} \tag{3.20}$$

近似给出(Kimura 和 Ohta,1972)。

在两参数模型的情况下,两序列间的差异被分类成转换型和颠换型。设 P 和 Q 分别为两序列间转换型和颠换型差异的比例。那么,两序列间核苷酸替换数 K,由

$$K = \frac{1}{2}ln(a) + \frac{1}{4}ln(b) \tag{3.21}$$

来估出,这里 $a=\frac{1}{1-2p-Q}$,$b=\frac{1}{1-2Q}$。取样本方差则由

$$V(K) = \frac{a^2P + c^2Q - (aP+cQ)^2}{L} \tag{3.22}$$

近似给出,其中 c=(a+b)/2,而 L 则为这些序列的长度(Kimura,1980)

让我们来考虑一个假想的数字例,设长为 2 0 0 的两个序列有 2 0 个转换和 4 个颠换差异。则 L = 2 0 0 , P =20/200=0.1,Q=4/200=0.02 。此例据两参数模型我们有:a=1/(1-0.2-0.02)=1.28,b=1/(1-0.04)=1.04,和 K=(1/2)ln(1.28)+(1/4)ln(1.04)≈0.13。替换总数可由每位点替换数 K,乘以位点数 L 求得。在此例中,从两序列间的 2 4 个差异,我们得到一个约为 2 6 次替换的估值。按照一参数模型,p=24/200=0.12 和 K≈0.13。于是,用一参数模型,我们达到了与两参数模型情况一样的结果。

上例中两种模型基本上给出了同样的估值,这是因为岐化程序低,以至修正后的岐化度(K=0.13)只略大于未经修正的值(p=24/200=0.12)的缘故。在这样的情况下,我们可用较简短的朱克斯和坎托的模型。

当两序列间的岐化度较大时,由两种模型得出的估值就可能差异显著。例如,两个具 L=200、相互有 50 个转换和 16 个颠换差异的序列,有 p=50/200=0.25,Q=16/200=0.08。按两参数模型,我们有 a=2.38,b=1.19 及 k≈0.48。而根据一参数模型,p=66/200=0.33, 和 K≈0.43。可见,按一参数模型 K 的估值小于用两参数模型得到的估值。当两序列间的岐化度较大、且特别地在有预先存在的原因相信转换的速率与颠换速率很不一样的情况下,两参数模型看来比一参数模型更精确。

两个为蛋白质编码序列间的替换数

在研究为蛋白质编码序列中,我们通常将起始和终止密码子排除在外,因为这两个密码子几乎不随时间而变。

为了分别处理同义替换和非同义替换,我们首先要把余下的核苷酸位点按以下方式分类:考虑一个密码子中的某一特别位置。设 i 为该位点上可能的同义变化数. 那么该位点被算作 i/3 同义的和 3-i/3 非同义的。例如,在密码子 TTT(Phe)中,最初两个位置被算作非同义的,因为在这两个位置上不发生同义变化;而第 3 位被 算作三分之一同义的和三分之二非同义的,因为该位置上三种可能的变化中有一种是同义的。另一个例子,密码子ACT(Thr)有两个非同义位点(前两个位置)和一个同义位点(第 3 位),因为前两位上所有可能的变化都是非同义的,而第 3 位上所有可能的变化都是同义的。在比较两个序列时,我们先要算出每一序列中同义位点的数目和非同义位点的数目,然后计算这两个序列间的平均值。我们用 Ns 表示同义位点的平均数,用 N_A 表示非同义位点的平均数。

其次,我们把核苷酸差异分成同义的差异和非同义的差异两类。对于只有一个核苷酸差异的两个密码子,这种差异很容易判断。例如,GTC(Val)和 GTT(Val)这两个密码子间的差异是同义的,而 GTC(Val) 和 GCC(Ala)这两个密码子间的差异则是非同义的。对于有不止一个核苷酸差异的两个密码子,我们必须考虑导致该观察到的变化的所有可能的进化途径。例如对AAT(Asn)和ACG(Thr)

这两个密码子，即有两种可能的途径：

途径Ⅰ：AAT(Asn)　ACT(Thr)　ACG(Thr)

途径Ⅱ：AAT(Asn)　AAG(Lys)　ACG(Thr)

途径Ⅰ需要一次同义变化和一次非同义变化，而途径Ⅱ则需要二次非同义变化。已知同义替换远比非同义替换发生得频繁（第四章），所以我们可以假定途径Ⅰ比途径Ⅱ可能性更大一些。例如，如果我们假定途径Ⅰ的权重为 0.7 而途径Ⅱ的权重为 0.3，那么两密码子间同义的差异数估计为 0.7×1+0.3×0=0.7，而非同义的差异数为 0.7×1+0.3×2=1.3。这里所用的权重是假设的。对所有可能的密码子对的权重作经验性的估计，宫田和安永（Miyata 和 Yasunaga,1980） 曾根据蛋白质顺序数据作出，李等（Li 等,1985b）则根据ＤＮＡ顺序数据而得到。如果我们假定两种途径可能性相同，那么，上例的非同义差异数为（1+2）/2=1.5, 而同义差异数则为(1+0)/2=0.5。可见，加权法和非加权法可能会给出有点不同的结果。实际上，两种方法的估值间的差异一般较小（Nei 和 Gojobori,1986），但对于那些高度保守的蛋白质，如组蛋白和肌动蛋白，对编码的基因而言它们可能非常重要（Li 等,1985b）。用任何一种方法，我们都能估出两编码序列间的同义差异数（M_S）和非同义差异数（M_A）。

从以上结果我们可以用 P_S=Ms/Ns 算出每同义位点的同义差异数，并用 pA=MA/NA 算出每非同义位点的非同义差异数。这些公式显然没有把同一位点上多次击中的效应考虑进去。我们可用朱克斯和坎托的公式：

$$K_s = -\frac{3}{4}ln\left(1-\frac{4Ms}{3Ns}\right) \qquad (3.23)$$

和

$$K_A = -\frac{3}{4}ln\left(1-\frac{4M_A}{3N_A}\right) \qquad (3.24)$$

来做这样的修正。

一种可采用的处理编码区的方法是，把核苷酸位点分成非简并的（nondegenerate）,两重简并的(twofold degenerate)和四重简并的(fourfold degenerate)位点(Li 等, 1985b)。如果一个位点上所有可能的变化都是非简并的，则该位点是非简并的；如果三种可能的变化中一种是同义的，则该位点是两重简并的；如果所有可能的变化都是同义的，则该位点即为四重简并的。例如，密码子ＴＴＴ（Ｐｈｅ）的前两位是非简并的，而第 3 位则是两重简并的（见第一章中的表１－１）。相比之下，密码子ＧＴＴ（Ｖａｌ）的第 3 位是四重简并的。 3 个异亮氨酸密(Ile)码子中的第 3 位被简化处理成两重简并位点，尽管事实上该位置上的简并是三重的。在哺乳动物的线粒体基因中，异亮氨酸只有两个密码子，所以其第 3 位事实上就是两重简并位点（见第一章中的表１－３）。

将核苷酸位点经上述分类分成各种简并类型（degeneracy classes）之后,我们即可对这 3 类位点分别计算两编码序列间的替换数。注意,根据定义所有非简并位点上的替换都是非同义的。类似地，所有四重简并位点上的替换都是同义的。在两重简并位点上，转换型变化（Ｃ-Ｔ和Ａ-Ｇ）是同义的，而所有其他变化，即颠换型变化，都是非同义的。在哺乳动物线粒体的遗传密码里，此规则一无例外。另一方面，在通用的细胞核遗传密码中，却有两个例外：精氨酸密码子（CGA 和 AGA，CGG 和 AGG）的第 1 位, 其上的一种颠换型变化是同义的,而其他类型的颠换和所有转换都是同义的;以及 3 个异亮氨酸密子(AUU、AUC 和 AUA)中的最后一位也是如此。

根据两种方法计算替换速率的计算机程序可由作者提供，若需要，请寄一个格式化了的ＩＢＭ　ＰＣ－兼容软磁盘来拷贝。

3.3 核苷酸序列和氨基酸序列的线性排比

两个同源序列的比较涉及对缺失和插入位置的鉴别问题，因为两个品系从其共同祖先分岐演化以来，任何一个中都可能发生这类变化。这一过程称为顺序线性排比（sequence alignment)。两个 DNA

序列的比较通常不能告诉我们,是其中一个序列中发生了丢失呢还是另一个序列中出现了插入。因此,这两类事件的后果统称为裂缝。

虽然我们是用ＤＮＡ序列来说明线性排比的过程，但同样的原则和程序也可用于氨基酸序列的排比。事实上，用氨基酸顺序与用ＤＮＡ顺序比起来，前者通常能得到更可靠的线性排比。

线性排比由一系列成对的碱基组成，其中每一个碱基各来自一个序列。有３种线列的对：（１）匹配的碱基对，（２）匹配错误的碱基对，和（３）由来自一个序列的碱基与另一序列的空缺碱基（null base）组成的对子。空缺碱基用—表示。一个匹配的对子表示一个自两序列分岐以来没有发生变化的位点，一个匹配错误的对子表示一次替换，而一个空缺对子则表示，在这两个序列之一的该位置上曾经发生过一次缺失或者插入。

考虑两ＤＮＡ序列Ａ和Ｂ，其长度分别为m和ｎ的例子。如果我们用ｘ表示匹配的对子数，用ｙ表示匹配错误的对子数，而用ｚ表示含有一个空缺碱基的对子数，则我们有：

$$n + m = 2(x + y) + z \qquad (3.25)$$

点阵法

当只有少数裂缝且两序列在其他任何方面差异都不太大时，一种合理的线性排比可以由视觉观察得到，或者也可用被称为点阵法（dot matrix method）的方法得到。在此法中，被线排的两个序列作为一个矩阵的首列和首行而写出（图３－６）。在两序列中核苷酸相同的矩阵位置处记上圆点。如果两序列等同，那么该矩阵对角线上的所有元素都将是圆点（图３－６ａ）。如果两序列有差异但可被无裂缝地线排，则对角线元素的大多数是圆点（图３－６ｂ）。如果两序列之一中出现一个裂缝，则线性排比的对角线将垂直或水平地移动（图３－６ｃ）。如果两序列间的差异既有裂缝又有替换（图３－６ｄ），则找出裂缝的位置并从几种可能的线性排比中挑出一种可能是很困难的。在这样的情况下，视觉观察和点阵法就不可靠了，而为了得到客观的线性排比已有几种计算方法被提出来了。

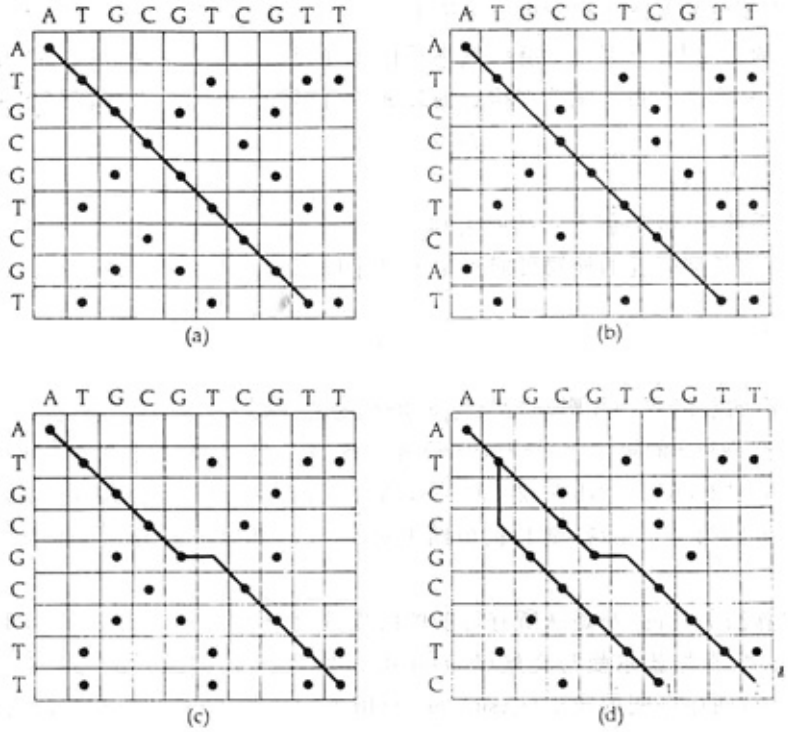


图３－６ 用于线排核苷酸　顺序的点阵。（ａ）两序列等同；（ｂ）两序列有差异但不含裂缝；（ｃ）两序列含有一个裂缝，但此外则相互等同；（ｄ）两序列既有替换又有裂缝。在（ｄ）中，途径１由６个对角线步骤，其中无空格，和２个垂直步骤组成。途径２含有８个对角线步骤，其中２个是空格，和１个水平步骤。途径１和途径２间的选取靠裂缝处罚来决定，即根据哪种进化序列更有可能：一次两核苷酸缺失（途径１）或一次一核苷酸缺失和两次替换（途径２）来决定

顺序—距离法

两序列间最为可能的线性排比是，根据某种标准使线列中匹配错误和裂缝的数目最小的那种。不幸地是，降低匹配错误数结果常会导致裂缝数增加，反之亦然。

例如，考虑以下两个序列：

A：TCAGACGATTG (m=11)

B：TCGGAGCTG (n=9)

我们可按如下排比将匹配错误数降到零：

（Ⅰ）ＴＣＡＧ－ＡＣＧ－ＡＴＴＧ

ＴＣ－ＧＧＡ

在这种情况下裂缝数为 6。反之，裂缝数可降低到由｜m-n｜个核苷酸组成的一个裂缝，结果匹配错误数却增加了：

（Ⅱ） TCAGACGATTG

TCG*GAG*KC*TG*--

在这种情况下，我们只有一个位于末端，因而也是不可避免的裂缝，但匹配错误（用星号标出）的数目却为 5 。

或者，我们可以选一个裂缝数和替换数都不是最小的线性排比。例如，

(Ⅲ)TCAG-ACGATTG

TC-GGA-GC*TG*-

在这种情况下匹配错误数是 2 ，裂缝数为 4 。

那么，这 3 种线性排比中哪一种最可取？显然，将替换与裂缝比较就好象将苹果和桔子比较一样。所以，我们必须找到一个共同标准，藉此来比较裂缝和替换。此共同标准被称为裂缝处罚（gap penalty)。

有几种指定裂缝处罚的系统。所有系统都是在相对于点状的替换、缺失和插入出现的频繁程度如何，这类问题的预先理解的基础上建立的。在第 1 个系统中，裂缝的总长度（z）用恒定的裂缝处罚（w）来乘。该系统背后的假定是，有某一裂缝的概率反比于裂缝的大小。举例说来，有一个由两核苷酸组成的裂缝的概率，与有两个各由一核苷酸构成的裂缝的概率相同。这样，对任何线性排比，我们都能用

$$D=y+wz \qquad (3.26)$$

来计算两序列间的距离尺度（ D)。

在第 2 种处罚系统中，我们假定长的缺失和插入在进化中与短的比较，出现的可能性是不同的。在这种情况下，对不同长度裂缝的处罚可能正比于裂缝长度也可能并非如此。根据这一系统，与某一特定线性排比有关的距离尺度是

$$D=y+\sum w_k z_k \qquad (3.27)$$

其中，z k 是长度为 K 的裂缝数，w k 则是对长为 K 的裂缝的处罚。

现在让我们用有 w＝ 2 的第 1 个系统来比较线性排比Ⅰ，Ⅱ和III。得到的距离（ D)，对线性排比Ⅰ，Ⅱ和III分别为：0+(2×6)=12,5+(2×2)=9,和 2+(2×4)=10。我们将选取线排Ⅱ。如果我们用有 w_1=2 和 w_2=6 的第 2 个系统,则 D 的值结果对Ⅰ、Ⅱ和III分别为 1 2 ， 1 1 和 1 0 。在这种情况下，我们选取线排III。

任何线性排比算法的目的，都是从所有可能的线性排比中，选取具有最小 D 值的那种线性排比。在最常应用的方法中，有尼德尔曼与文施（Neeldleman 和 Wunsch,1970）法，和塞勒斯（Seller,1974）法. 在前一种方法中,两序列间的类似性(similarity)用类似指数(similarity index) 来测度，而具最大类似性的那种线性排比将被从所有候选者中选取出来。在塞勒斯法中，两序列间的不相似性（dissimilarity）用距离指数（distance index）来测度，具最小距离的那种线性排比将被选出。这两种方法曾被证明在某些条件下是等价的（Smith 等,1981)。

在必须从许多线性排比中选出一种时，寻找最佳排比的任务若无计算机的帮助常常难以完成。在

尼德尔曼与文施（Needleman 和 Wunsch,1970）算法或其修订法的基础上，已有许多关于线排顺序的常用

计算机程序建立起来。

要记住的最重要的一点是，作为最后结果的线性排比常有赖于裂缝处罚的选取，而后者又有赖于，相对于点替换的频率裂缝事件在ＤＮＡ和蛋白质进化中的频率究竟是多少的这样一些关键的假定上。

3.4 核苷酸替换数的间接估计

在估计两序列间核苷酸替换数方面，最完全的解决可通过比较它们的核苷酸顺序而得到。不过，替换数也可从其他类型的分子数据，象限制酶图谱或者ＤＮＡ－ＤＮＡ杂交得到的数据间接地推断出来。

限制性核酸内切酶片段模式和位点图谱

限制性核酸内切酶（restriction endonucleases)或限制酶(restriction enzymes)能识别被称为识别顺序（recognition sequences)的特殊双链ＤＮＡ序列，并在识别顺序上或其近旁切开该ＤＮＡ。识别顺序通常长为4或6碱基对，它们中许多都是回文（即它们是旋转对称的)。识别顺序可能是唯一的（例如EcoRI),也可能不是唯一的(例如HindⅡ)(见表3-1)。切点称为拼接位点（splicing site)或限制位点（restriction site)。许多限制性内切核酸酶以一种错开的方式切开双链ＤＮＡ，所以将产生“粘性末端”(sticky ends)，以后它们可在连接酶（ligase)的作用下相互连接(ligated)。这就是为什么限制酶能在遗传工程中成为一种极有用的工具的原因。表３－１列出了几种限制酶的识别顺序和拼接位点。

表３－１ 几种限制性核酸内切酶的识别顺序和切点

酶（生物来源）	识别位点	识别顺序(RS) 大小	不确定性	回文	邻接	切割 在RS中	错开式
EcoR Ⅰ (*Escherichia coli*)	5′－G↓A－A－T－T－C－3′ 3′－C－T－T－A－A－G↑－5′	6	－	＋	＋	＋	＋
Hind Ⅱ (*Haemophilus influenzae*)	5′－G－T－Py↓Pu－A－C－3′ 3′－C－A－Pu↑Py－T－G－5′	6	＋	＋	＋	＋	－
Hae Ⅲ (*Haemophilus aegyptus*)	5′－G－G↓C－C－3′ 3′－C－C↑G－G－5′	4	－	＋	＋	＋	－
Bbv Ⅰ (*Bacillus brevis*)	5′－G－C－A－G－C－(N_8)↓3′ 3′－C－G－T－C－G－(N_{12})↑－5′	5	－	－	＋	－	＋
Nci Ⅰ (*Neisseria cinerea*)	5′－C－C↓C/G－G－G－3′ 3′－G－G－G/C↑C－C－3′	5	＋	＋	＋	＋	＋
Not Ⅰ (*Nocardia otitidis-caviarum*)	5′－G－C↓G－G－C－C－G－C－3′ 3′－C－G－C－C－G－G↑C－G－5′	8	－	＋	＋	＋	＋
Hinf Ⅰ (*Haemophilus influenzae*)	5′－G↓A－N－T－C－3′ 3′－C－T－N－A↑G－5′	4	－	＋	－	＋	＋

a、识别顺序用黑体字母表示。切点用箭头指出。不确定的地方，象Pu：嘌呤;Py:嘧啶；C/G:C或G；N：任何核苷酸。Ｎｎ表示由ｎ个任意的核苷酸组成的序列。

当一个双链的ＤＮＡ片段受到水解时，即有各种不同长度的片段产生出来。它们可因其各自的长度而在电泳凝胶上分开，因为在凝胶上较短的片段比较长的要跑得更快也移动得更远。通过用已知长度的ＤＮＡ片段作基准，限制性片段的长度即可被估计出来。不同的ＤＮＡ序列根据其识别位点的数目和位置的差异而受到限制酶的不同切割。由一个ＤＮＡ序列水解产生的片段的数目和大小被称为限制片段模式（restriction fragment pattern)。连续而交互地应用几种能将ＤＮＡ水解成重叠片段的限制酶，常常可推断出ＤＮＡ上限制位点的大概位置（图３－７）。表示某一ＤＮＡ序列上限制位点的位置的方案图称限制图谱（restriction map)。

应用限制酶来推断两序列间的替换数，其背后的理由是，两DNA序列的类似性越大则其限制片段模式就越相似。通过对DNA序列内限制位点的分布作出某些假定，例如，象4种核苷酸有相同的频率以及它们在序列中的空间分布是随机的，这样的假定，则从限制位点数据就可以对DNA序列间限制模式方面的进化变化进行研究，从而估计出每位点的核苷酸替换数（K）。

首先，我们考虑从限制片段模式来估计K。从共有片段数来估计K，要求我们对由限制性核酸内切酶水解的DNA的电泳模式进行直接比较。这里提供的方法是由根井和李（Nei 和 Li,1979）创导的。文献中曾报导过另外两种方法(Upholt,1977;Engels,1981a)，而卡普兰(Kaplan,1983)曾证明了这3种方法给出类似的结果。

DNA的两序列间共有DNA片段的期望比例（F），可由

DNA 的两序列间共有 DNA 片段的期望比例(F),可由

$$\hat{F} = \frac{2m_{XY}}{m_X + m_Y} \tag{3.28}$$

来估计,其中 m_X 和 m_Y 分别是序列 X 和 Y 水解后产生的限制片段的数目,而 m_{XY}则是两序列共有的片段数。

根井和李(Nei 和 Li,1979)曾证明,共有片段的期望比例(F)可用在 t 时间内某一限制位点保持不变的概率(G)来表示,两者之间的近似关系为

$$F \approx \frac{G^4}{3 - 2G} \tag{3.29}$$

这里,$G=e^{-r\lambda t}$,G 中的 r 为识别位点中核苷酸的数,λ 是核苷酸的替换速率,t 是两序列间发生分歧的时间。这些序列间每位点的替换数为 $K=2\lambda t$。为了估出 G,我们重新安排等式 3.29,得

$$G = [F(3 - 2G)]^{\frac{1}{4}} \tag{3.30}$$

该方程可通过一个反复尝试过程解出。根井(Nei,1987) 建议用 $G=F^{1/4}$作为最初尝试值。一般只需要很少几次反复循环。G 的估值可使我们得到 K 的估值,关系如下:

$$K = -\frac{2}{r} ln(G) \tag{3.31}$$

让我们来考虑下面这样一个例子：取自两种野生小麦（Aegilops sharonensis 和 Ae. bicornis ）的相应线粒体DNA片段，用3种限制酶， Bam Ⅰ, HindⅢ和 EcoRI 来水解,它们的识别顺序都为 6 碱基对长(数据自 Graur 等,1989a)。Ae. sharonensis 水解产生 4 个片段,而 Ae. bicornis 水解则产生 5 个片段。两个片段为两种小麦所共有。用等式 3.28，我们估出 F 为 2/9=0.222。现在我们可以开始由等式 3.30 给出的反复尝试过程。我们采用的G的初值为 $0.222^{1/4}$ =0.687。第一次循环后我们得G＝0.775，而下一次循环G＝0.753。随着尝试的进行摆幅将越来越小，而在弟5次和第6次循环后我们都得到G＝0.758。因此，我们终止反复尝试过程。为了得到两序列间替换数的估值，我们用等式 3.31。最后结果是，两线粒体序列的相互差异用每核苷酸位点替换数K＝0.092 来估计。

现在我们考虑由限制位点图谱估计两序列间的核苷酸替换数。在前面的例子中，限制位点的位置是未知的。如果限制位点已在DNA序列上定位，那么，我们可以直接从图谱上找出共有和非共有的位点，并估出替换数。设 m_X 和 m_Y 分别为 DNA 序列 X 和 Y 中限制位点的数目,而 m_{XY}为两序列间共有的限制识别位点数。X 和 Y 在某一给定位点上共有同样的识别顺序的概率用 S 表示,此值可用

$$\hat{S} = \frac{2m_{XY}}{m_X + m_Y} \tag{3.32}$$

估出(Nei 和 Li,1979)。核苷酸差异的比例,p,可用

$$\hat{p} = 1 - \hat{S}^{\frac{1}{r}} \tag{3.33}$$

估计，其中 r 是识别顺序中核苷酸的数目。两序列间的每位点替换数可用等式 3.19 从已知的 $\hat{p}$ 估出。

限制位点图谱法比限制片段模式法要乏味一些,但却可靠得多。前者在K值高达 0.25 时仍可应用, 而后者对于 K>0.05 的情况就可能是不精确的。

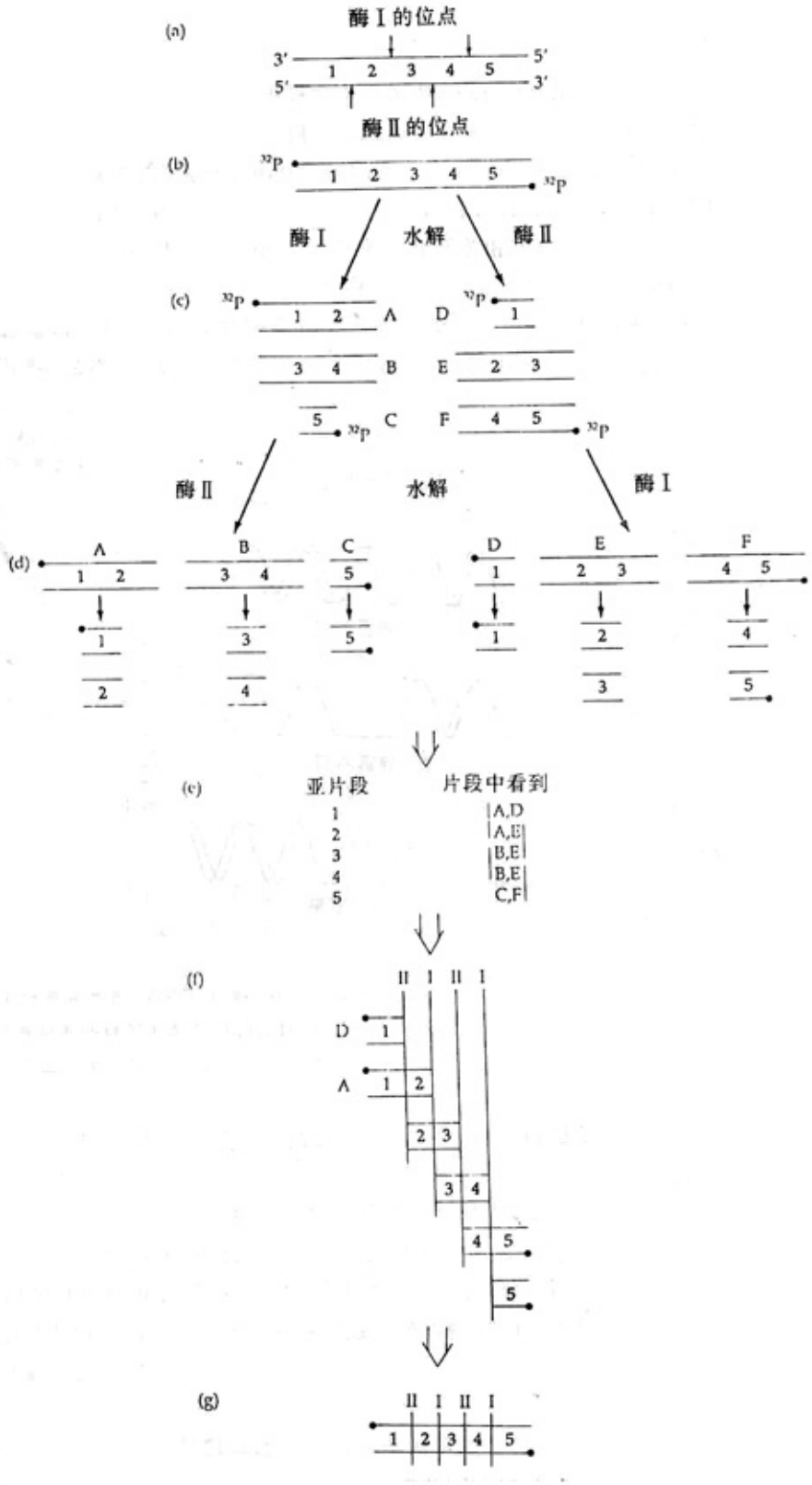


图３－７ ＤＮＡ序列上限制位点的定位图。（ａ）一个假想的ＤＮＡ，具有由两种不同的限制酶识别的识别位点。该序列的识别图谱未知。（ｂ）3′ 端用放射性标记. (C) DNA 由酶Ⅰ水解产生片段 A, B 和 C, 而由酶Ⅱ水解则产生片段 D、E 和Ｆ。（ｄ）用一种限制酶得到的每一片段再用另一种限制酶水解，产生出亚片段１－５。（ｅ）这些亚片段被用于鉴别重叠片段。（ｆ）这些亚片段按它们间的重叠模式指示的秩序排列。(g) 推理出的 DNA 序列的限制图谱。片段和亚片段可根据它们的长度一个个地鉴别出来，而长度则是从它们在电泳凝胶上的位置推出的。末端片段和亚片段则由其放射性标记来鉴别。自 Suzuki 等 (1989) 修改而成。

ＤＮＡ－ＤＮＡ杂交

ＤＮＡ－ＤＮＡ杂交（DNA-DNA hybridization）技术是以这样的事实为根据的：双链ＤＮＡ分子的热稳定性有赖于两条链间核苷酸匹配的比例。随着匹配比例的降低，双链的热稳定性也降低。在两条链来自同一序列的双链ＤＮＡ（即同源双链（homoduplex）分子）中,匹配的比例根据定义应为100%。另一方面，在两条链来源不同的双链ＤＮＡ（即异源双链（heteroduplex）分子）中，匹配的比例则小于１。其大小有赖于这两个序列自它们从某一共同祖先分化以来究竟累积了多少核苷酸差异。所以，异源双链ＤＮＡ将会在比同源双链ＤＮＡ低的温度下变性或熔解成单链。

ＤＮＡ杂交试验的基本实验程序如图３－８所示。大致上，在重复序列被除去以后，该程序包括将来自两不同物种的变性ＤＮＡ的混合物缓慢冷却，以制造出人工杂种ＤＮＡ分子。然后，将该混合物逐

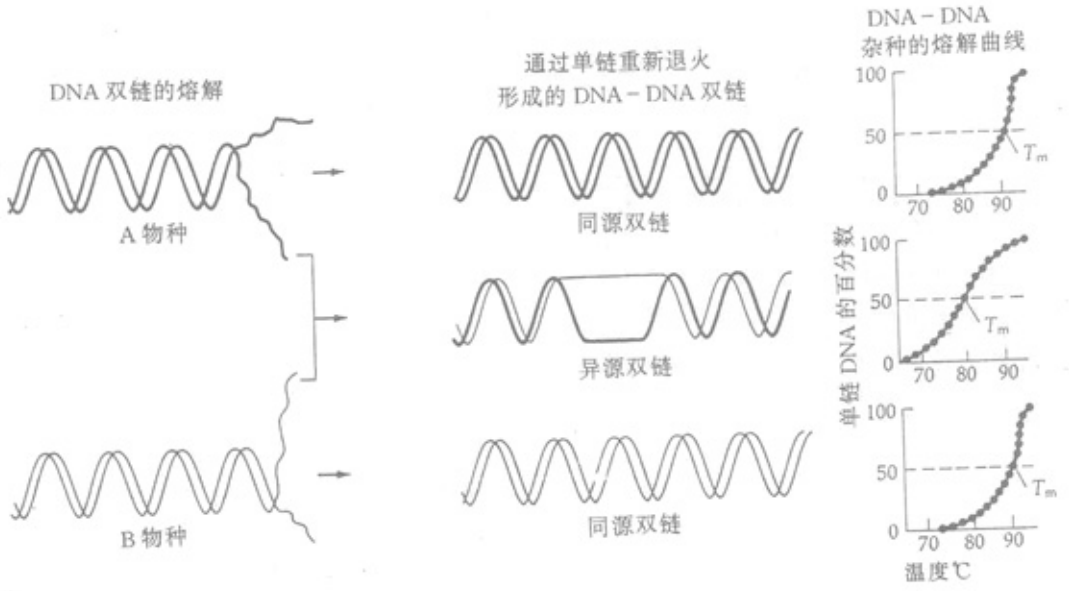


图３－８ 从ＤＮＡ－ＤＮＡ杂交研究推论出的顺序类似性。双链分子被熔解成单链的ＤＮＡ。同源双链和异源双链通过单链的重新退火而形成。５０％的ＤＮＡ熔解成单链的温度用 Tm 表示，测定两种同源双链和异源双链的 Tm。两种同源双链间的Ｔｍ值可能是不同的，同样两互不相同的异源双链类型间的Ｔｍ值也可能是不同的。自 Avers(1989)修改而成。

渐加热，并在每一温度下测定溶液中单链ＤＮＡ的百分比。关于此法（TEACL 法）的详细介绍，可见例如享特等(Hunt 等,1981)的论述。

杂种ＤＮＡ的热稳定性，用５０％的杂种ＤＮＡ解离成单链时的温度来度量。然后将此半熔解温度与５０％的同源双链ＤＮＡ变成单链时的温度比较。注意，在每一次种间比较中，我们有两种同源双链，每物种各有一种，所以，习惯上我们用它们的半熔解温度的平均值。同源双链和异源双链的半熔解温度间差异，ΔTm,由经验证明,与碱基对误配的比例近似地线性相关(Britten 等,1974)。我们将这种关系表示成

$$p=C\Delta T_m \quad (3.34)$$

这里ｐ是误配的比例，Ｃ是一个常数。Ｃ的值通过对碱基对误配度已知的异源双链进行ＤＮＡ－ＤＮＡ杂交试验，而从经验上得到。Ｃ值被发现随实验条件的不同而在Ｃ＝0.01 和 C=0.015 之间变化。已知ΔTm 的实验误差是非常大的，因此，对同样的物种对应该做许多次重复观察。

现在让我们来考虑下面的一个数字例（数据来自 Caccone 和 Powell,1989)。来自人类和矮黑猩猩（Pan paniscus　）雄性的同源双链ＤＮＡ的平均Ｔｍ值，分别为 59.50℃和 59.12℃。于是，同源双链分子的 Tm 平均值为 59.31℃。两交互的异源双链 DNA 的 Tm 平均值则为 57.59℃。因此，ΔTm 为 1.72℃。由等式 3.34，我们得到一个每核苷酸位点大约 0.017-0.026 次替换的差异。

习题

1　证明等式 3.3 对 t=0 成立,即,若 t=0 则它将简化成等式 3.1。

2　导出等式 3.10，并证明在朱克斯　坎托模型下 $P_{GA}(t)$ ＝ $P_{CA}(t)$ ＝ $P_{TA}(t)$ 。

3　当 α ＝ β 时，木村的两参数模型将变得与朱克斯和坎托的一参数模型等同。为了证实这一点。证明：当此条件满足时，等式 3.13 将变得与等式 3.11 相同,且等式 3.14 和 3.15 都变得与等式 3.12 相同。

4　用等式 3.13,3.14 和 3.15 证明，在木村的两参数模型下一个序列中 4 种核苷酸的平衡频率都是相同的（即 1/4）,这与一参数模型的结果一样。

5　从等式 3.16 导出等式 3.17。

6　对以下两个序列：

Ser	Thr	Glu	Met	Cys	Leu	Met	Gly	Gly
TCA	ACT	GAG	ATG	TGT	TTA	ATG	GGG	GGA
TCG	ACA	GGG	ATA	TAT	CTA	ATG	GGT	ATA
Ser	Thr	Gly	Ile	Tyr	Leu	Met	Gly	Ile

计算（a）每同义位点的同义替换数和（b）每非同义位点的非同义替换数。

7　根据木村的两参数模型，两序列间的差异数为 P ＋ Q，证明当转换和颠换合起来考虑时等式 3.21 将被简化成等式 3.19。

8　用点阵法线性排比以下两个顺序：

AATGCTTGCATGGGGCTAGTT

ATTGCTGCATGAGGCGCGCTAGT

选出两种可能的线性排比，并决定用每核苷酸为 2 的恒定裂缝处罚时哪一种要更好一些。若用更大的裂缝处罚，比如说 1 0，该选择会受到影响吗？

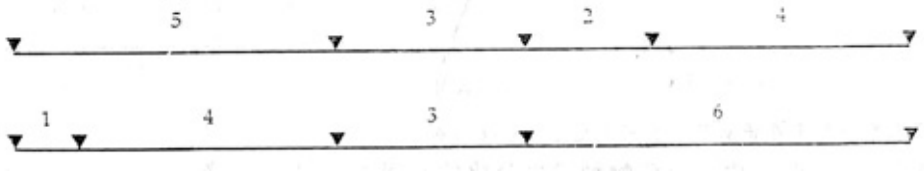


图 3 - 9　两限制性内切核酸酶图谱的假想例。序列上的数字代表这些片段的长度（以 kb 为单位. ）

9　从图 3 — 9 中的两序列的限制位点图谱，估计两序列间核苷酸的替换数，用（a）共有片段的比例，和（b）共有限制位点的比例。限制性内切核酸酶的识别顺序为 4 个核苷酸。用有 6 核苷酸识别位点的限制性内切核酸酶，将会有什么样的结果？该差异的原因是什么？

后继阅读文献

Doolittle,R.F.1990. Molecular Evolution:Computer Analysis of Protein and Nucleic Acid Sequences, Academic Press,San Diego,CA

Li, W. H.,C. C.Luo and C. I.Wu.1985.Evolution of DNA sequences . pp. 1-94. In R. J. MacIntyre(ed.),Molecular Evolutionary Genetics, Plenum, New York.

Nei, M. 1987 Molecular Evolutionary Genetics, Columbia University Press, New York.

4 核苷酸替换的速率和模式

前一章中导出的数学理论可用于核苷酸替换速率的研究之中，而该速率则是分子进化研究里的一个基本量。事实上，为了阐明某一ＤＮＡ序列进化的特性，我们需要知道，它进化得究竟有多快，以及其组成部分的核苷酸替换速率是多少。比较一下基因和不同ＤＮＡ区域间的替换速率也是很有趣的，因为这可以帮助我们理解进化中核苷酸替换的机制。知道了核苷酸替换的速率，还使我们能对物种间的分岐演化这样的进化事件，给出一个时间年代来。不过，要想做到这一点，我们必须知道从一组物种估出的速率是否能适用于另一组生物种群。这就提出了这样一个问题，即，速率在不同的进化谱系间
是怎样变化的。

4.1 核苷酸替换的速率

核苷酸替换的速率（rate of nucleotide substitution）被定义成每年每位点的替换数，并可用两同源序列间的替换数Ｋ，除以２Ｔ来算出，这里Ｔ是两序列间发生分岐的时间（图４－１）。即，

$$r=K/2T \quad (4.1)$$

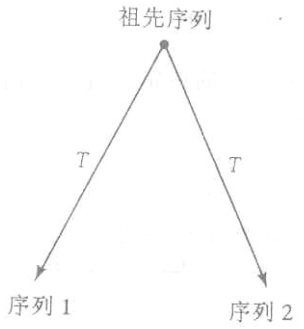


图４－１ 两同源序列在Ｔ年前从某一共同祖先序列分岐而来两序列发生分岐的时间Ｔ，假定与两物种发生分化的时间相同，且通常都用古生物学数据资料来推算。

本节我们将处理不同基因间，和某一基因的不同区域间的速率变异问题。为此目的，建议对所有被考虑的基因使用同样的物种对。这有两重原因。首先，关于分岐时间的古生物学估计通常都有很大的不确定性。用同一对物种，我们就可以无须知道分岐时间而去比较各基因间的进化速率。其次，替换速率在各谱系间可能变化很大（见第 53 页），在这种情况下，两基因间速率上的差异可能是由谱系间的差异所造成，而不是由两基因本身的差异所造成的。

目前研究核苷酸替换速率的最合适的数据来自哺乳动物，这是因为，有关哺乳动物的ＤＮＡ序列的数据最为丰富，有关哺乳类的化石记录相对而言特征比较明确、全面，再加上可以得到哺乳类之间相当可靠的分岐时间的缘故。

编码区

我们在表４－１中列出了３６种为蛋白质编码的基因的同义替换速率和非同义替换速率。这些速率是从人类与啮齿类同源基因间的比较中算出的。根据与真兽亚纲哺乳类的辐射演化有关的古生物学证据，人类　啮齿类的分岐时间已被设定为８０００万年前。

表４－１ 哺乳类各为蛋白质编码的基因的同义替换速率和非同义替换速率 a

基因	L[b]	非同义替换速率($\times 10^{9}$)	同义替换速率($\times 10^{9}$)
组蛋白			
组蛋白 3	135	0.00±0.00	6.38±1.19
组蛋白 4	101	0.00±0.00	6.12±1.32
收缩系统蛋白			
肌动蛋白 α	376	0.01±0.01	3.68±0.43
肌动蛋白 β	349	0.03±0.02	3.13±0.39
激素、神经肽和其他活性肽			
生长激素释放抑制因子－28	28	0.00±0.00	3.97±2.66
胰岛素	51	0.13±0.13	4.02±2.29
促甲状腺素	118	0.33±0.08	4.66±1.12
胰岛素样生长因子Ⅱ	179	0.52±0.09	2.32±0.40
促红细胞生成素	191	0.72±0.11	4.34±0.65
胰岛素 C 肽	35	0.91±0.30	6.77±3.49
甲状旁腺素	90	0.94±0.18	4.18±0.98
促黄体生成激素	141	1.02±0.16	3.29±0.60
生长激素	189	1.23±0.15	4.95±0.77
尿激酶—血纤蛋白溶酶原活化因子	435	1.28±0.10	3.92±0.44
白细胞中介素 I	265	1.42±0.14	4.60±0.65
松弛肽	54	2.51±0.37	7.49±6.10
血红蛋白和肌红蛋白			
α-珠蛋白	141	0.55±0.11	5.14±0.90
肌红蛋白	153	0.56±0.10	4.44±0.82
β-珠蛋白	144	0.80±0.13	3.05±0.56
载脂蛋白			
E	283	0.98±0.10	4.04±0.53
A－Ⅰ	243	1.57±0.16	4.47±0.66
A－Ⅳ	371	1.58±0.12	4.15±0.47
免疫球蛋白			
IgV_H	100	1.07±0.19	5.66±1.36
$Ig\gamma_1$	321	1.46±0.13	5.11±0.64
Igk	106	1.87±0.26	5.90±1.27
干扰素			
α1	166	1.41±0.13	3.53±0.61
β1	159	2.21±0.24	5.88±1.08
γ	136	2.79±0.31	8.59±2.56
其他蛋白质			
醛缩酶 A	363	0.07±0.03	3.59±0.52
羟黄嘌呤磷酸核糖基	217	0.13±0.04	2.13±0.35
转移酶			
肌酸激酶 M	380	0.15±0.03	3.08±0.37
甘油醛-3-磷酸脱氢酶	331	0.20±0.05	2.84±0.37
乳酸脱氢酶 A	331	0.20±0.04	5.03±0.61
乙酰胆碱受体 γ 亚基	540	0.29±0.04	3.23±0.31
血纤蛋白原 γ	411	0.55±0.06	5.82±0.67
白蛋白	590	0.91±0.07	6.63±0.61
平均[c]		0.85(0.73)	4.61(1.44)

a　所有速率都以人和啮齿类基因间的比较为根据，且分歧时间设定为８０００万年以前。速率以每１０^{9}年每位点替换数为单位。　b．L＝受比较的密码子数。　c　平均指算术平均值，括号内的值为标准偏差，都是用所有基因的值算出的。

我们注意到，非同义替换的速率在基因间变化极大。变化幅度从组蛋白 3 和组蛋白 4 的有效数字为零，到干扰素 γ 的每年每非同义位点 2.79×10^{-9} 替换。某些激素（例如生长激素释放抑制因子-28 和胰岛素）是极其保守的，而另一些激素则或者以中间速率（例如促红细胞生成素）进化，或者以高速率（例如白细胞中介素Ⅰ和松弛肽）进化。血红蛋白和肌红蛋白以中速进化，而载脂蛋白和免疫球蛋白则进化得非常迅速。

同义替换的速率变化也相当大，不过比起非同义替换来速率变化要小得多。可以证明，基因间同义替换速率方面的变异明显地大于仅由统计学波动造成的期望变异。

对表４－１中绝大多数基因来说，同义替换的速率大大超过非同义替换的速率。如在一个最极端的例子，组蛋白３中，虽然从其氨基酸顺序看它是进化上最为保守的蛋白质中的一种，但其同义替换的速率却非常高。对表４－１中的基因来说，非同义替换的平均速率为每年每非同义位点 0.85×10^{-9} 替换。同义替换的平均速率为每年每同义位点 4.6×10^{-9} 替换，即为非同义替换平均速率的５倍。

非编码区

来自非编码区的数据远不如来自编码区的数据丰富，所以目前只做过有限的比较分析工作。（注意，要估出某一序列中的替换速率，我们必须至少有来自两个物种的数据。）因为大多数已发表的序列为ｍＲＮＡ，它们不含内含子和侧区域，所以，其５′和３′不翻译区是唯一能进行仔细研究的非编码区.表 4-2 列出了根据人与啮齿类比较得到的 16 种基因中这两个区域的替换速率。在这两个区域中不同基因间的速率变化都非常大,但这种变异可能很大程度上代表了抽样所造成的影响，因为这两区域通常都非常短。在几乎所有基因中，５′和３′不翻译区中的速率都低于四重简并位点上的替换速率(即,其上所有可能的核苷酸替换都是同义替换的位点)。５′和３′不翻译区的平均速率分别为每年 1.96×10^{-9} 和 2.10×10^{-9} 替换,它们都约为四重简并位点上的平均速率,每年 3.55×10^{-9} 替换的 60%。

表 4-2 根据人与小鼠或大鼠的基因间比较,得到的为蛋白质编码的基因的 5′及 3′ 不翻译区和四重简并位点上的核苷酸替换速率[a]

基　因	5′不翻译区		3′不翻译区		四重简并位点	
	L[b]	速率	L	速率	L	速率
ACTH	99	1.87±0.41	97	2.32±0.49	275	2.78±0.34
醛缩酶 A	124	1.08±0.26	154	1.73±0.32	195	3.16±0.48
载脂蛋白 A－Ⅳ	83	3.06±0.68	134	1.73±0.33	160	3.38±0.50
载脂蛋白 E	23	1.27±0.69	84	1.70±0.42	153	4.00±0.60
Na,K-ATP 酶β	118	2.45±0.45	1117	0.57±0.06	118	2.87±0.54
肌酸激酶 M	70	1.71±0.46	168	1.79±0.30	178	2.81±0.41
α-胎蛋白	47	3.64±1.13	144	2.79±0.49	225	4.14±0.54
α-珠蛋白	34	1.56±0.65	90	2.21±0.50	81	4.47±0.98
β-珠蛋白	50	1.30±0.46	126	2.85±0.49	78	2.42±0.56
甘油醛-3-磷酸脱氢酶	70	1.34±0.38	121	1.74±0.36	170	2.43±0.39
生长激素	21	1.79±0.85	91	1.83±0.41	83	3.82±0.78
胰岛素	56	2.92±0.80	53	3.09±0.81	62	4.19±1.00
白细胞中介素Ⅰ	59	1.09±0.38	1046	2.02±0.14	105	2.97±0.60
乳酸脱氢酶 A	95	2.79±0.55	470	2.48±0.23	152	3.64±0.60
金属巯基组氨酸三甲基内盐Ⅰ	61	1.88±0.52	111	2.57±0.48	23	2.37±1.00
甲状旁腺素	84	1.79±0.43	228	2.21±0.30	38	3.85±1.21
平均[c]		1.96(0.78)		2.10(0.61)		3.33(0.69)

a　速率以每１０^{9}年每位点替换数为单位。　b．Ｌ＝位点数。　c、平均指算术平均，括号内的值为标准偏差，都是用所有基因的值算出的。

假基因（pseudogenes）是一些由功能基因派生，但由于发生了阻止其正常表达的突变而退化成无功能的ＤＮＡ序列（第六章和第七章）。由于它们不受功能限制，所以，它们可以期望以较高的速率进化。表４－３列出了乳牛和山羊的ψβ^{X}和ψβ^{Z}假基因中替换速率间的比较，以及β－和γ－珠蛋白基因中非编码区和四重简并位点上的速率间的比较。这些假基因中的速率事实上略高于其他区域中的速率。看来这一点对假基因来说是普遍成立的，尽管目前有关假基因的资料尚属有限。

表 4 － 3　乳牛和山羊的 β －和 γ －珠蛋白基因间的分歧，以及 β －珠蛋白假基因间的分歧

统计量	β 和 γ-珠蛋白基因[a]						
	5′FL	5′UT	四重简并	内含子	3′UT	3′FL	假基因
百分比分歧	5.3	4.0	8.6	8.1	8.8	8.0	9.1
标准误差	1.2	2.0	2.5	0.7	2.2	1.5	0.9

a．F L = 侧区域；U T = 不翻译区域；四重简并 = 四重简并位点。

图 4 － 2 中，我们对基因的不同区域中，以及假基因中的替换速率进行了比较。关于 5′ 和 3′ 不翻译

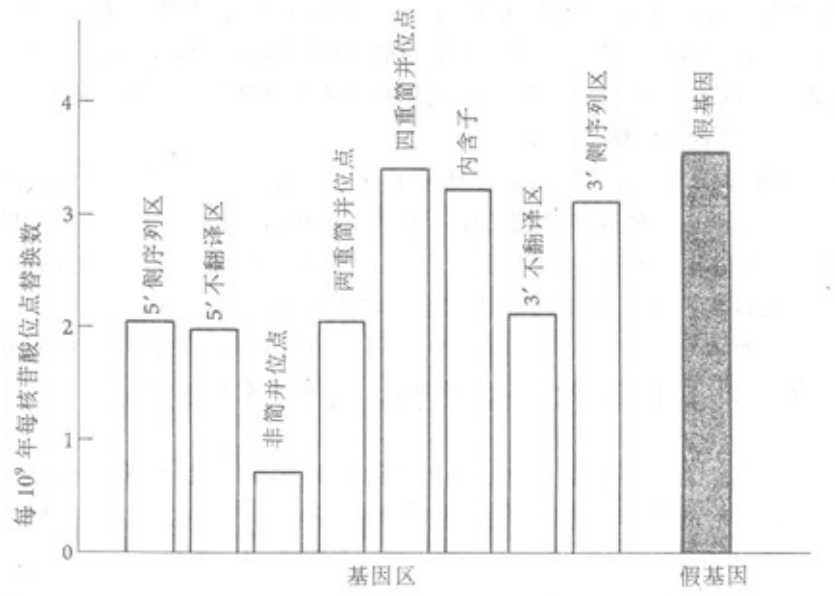


图 4 － 2　基因的不同部位中以及假基因中的替换平均速率。

区域，非简并位点，两重简并位点和四重简并位点的速率，都是将表 4 － 2 中所列基因加以平均后的平均速率。5′ 侧区域的速率，通过假定该速率与四重简并位点上的速率之比为 5.3/8.6（即由表 4 － 3 得出的值），和四重简并位点上的平均速率为每年 3.33×10^{-9} 替换(表 4-2)而算出。内含子的速率，3′ 侧区域的速率以及假基因的速率也按同样方式算出。由于以有限的资料为基础而作出的估计已经够多了，又由于一个区域中的速率会因基因的不同而有差异，所以，图 4 － 2 中展示的速率可能对任何一个具体的基因都是不适用的，但它却提供了不同 D N A 区域中的替换速率间一个大致的、一般性的比较。有了这种思想准备，我们将看到，一个基因中的替换速率以四重简并位点上的为最高，内含子中和 3′ 侧区域中要略低一些，3′ 不翻译区域，5′ 侧区域，5′ 不翻译区域和两重简并位点有中等大小，而非简并位点上的为最低。平均下来假基因有最高的替换速率，虽然它只比一个功能基因的四重简并位点上的速率稍高一些。

4.2 替换速率变异的原因

为了推理出 D N A 区域间替换速率出现变异的原因，我们应注意到，替换速率是由两个因子所决定的：（1）突变率和（2）一个突变的固定概率（第二章）。后者又与该突变是有利的、中性的还是有害的有关。由于突变率看来在一个基因内变化不大而在不同基因间则可能变化较大，所以，我们将对一个基因的不同区域间的速率变异和不同基因间的速率变异分别讨论。

不同基因区域间的变异

我们首先考虑一个基因中同义位点和非同义位点间的大差异。由于一个基因内同义位点与非同义位点上的突变率应该相同，或者至少是非常相似，所以，替换速率上的差异就可归因于两种不同类型位点间纯洁化选择的强度上的差异。这可用分子进化的中性学说来理解（第二章）。结果会导致氨基酸替换的突变比同义的改变对该蛋白质的功能造成有害影响的机会要高。所以，绝大多数非同义突变都将受纯洁化选择

而从群体中清除。其结果将使非同义位点上的替换速率降低。相比之下，同义的改变有较高的机会是中性的，而它们中在群体中固定的也要多些。

当然，非同义替换可能有幸使蛋白质的功能得到改善。然而，如果有利选择在该蛋白质的进化中起主要作用的话，则非同义替换的速率应该超过同义替换的速率。事实上，在某些免疫球蛋白基因里，决定互补性的区域（CDRs，又以高可变区著称）中非同义的速率高于同义的速率。这种较高的速率已经归因于对抗体多样性的超显性选择（Tanaka 和 Nei,1989）。不过，当考虑的是整个免疫球蛋白基因时，非同义的速率仍然大大低于同义的速率（表 4 － 1 ）。这个结果指出，即使在免疫球蛋白中，大多数非同义突变也是不利的，并且将从群体中清除。休斯和根井（Hughes 和 Nei,1989）曾报导在主组织相容性复合体基因的某些区域中有类似情形，即非同义替换的速率超过同义替换的速率。他们把非同义替换有更高的速率归因于超显性选择。

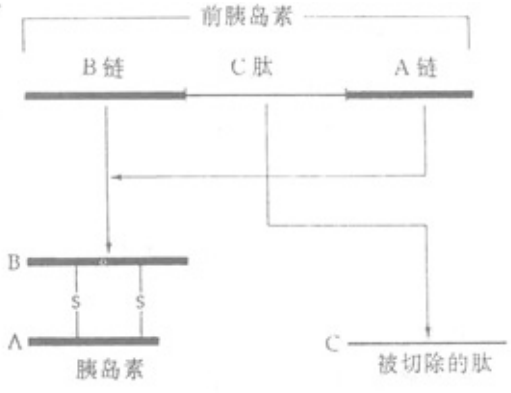
一个基因中同义的和非同义的速率间的对比证明了分子进化中一个众所周知的原则，即，对某一大分子的功能限制越强，则其进化的速率就越缓慢。木村（Kimura,1983）曾用一个简单模型将此原则表达成一个公式。假定某一分子中所有突变的某一部分 f_0，是选择中性或近中性的，而其余的则是有害的（有利突变被假定仅极偶然地出现，以至其相对频率的有效数字为零，而且它们对整个分子进化的速率没有多大贡献）。如果我们用 v_T 表示每单位时间每位点的总突变率，那么，中性突变的突变率即为 $v_0 = v_T f_0$。根据分子进化的中性学说，替换速率为 $K = v_0$（第二章）。因此，

$$K = v_T f_0 \qquad (4\ \ 2)$$

在任一给定基因内，该 vT 值可假定对同义位点和非同义位点都是相同的。然而，f_0 值则是同义位点的比非同义位点的高，所以前者要比后者进化得快。虽然该模型是过于简单了，但它对解释不同ＤＮＡ区域间速率上的差异却很有帮助。

依上述模型看，最高的速率预期应出现在一个没有任何功能的序列中，由于没有功能，所以它里面的所有突变都是中性的（即 $f_0 = 1$ ）。事实上，假基因看起来的确有最高的核苷酸替换速率（表 4 － 3 和图 4 － 2 ）。5′ 和 3′ 不翻译区有比编码区中的同义替换更低的替换速率，这一观察事实进一步支持解释问题的中性路线，因为这些区域含有关于转录起始和终止的信号。

在一个蛋白质内，有不同结构和功能的区域看来受着有差别的功能限制，并以不同速率进化着。胰岛素原为此提供了一个极好的例子。它由Ａ，Ｂ和Ｃ三个片段组成（图 4 － 3 ），片段Ｃ位于分子的中间，并在活性激素（胰岛素）形成期间被除去。即胰岛素是由余下的Ａ和Ｂ两个片段所构成的。片段Ｃ



进化速率 0.13*10 －9/位点/年　　进化速率 0.97*10 -9/位点/年

图 4 － 3 为有功能的胰岛素（Ａ和Ｂ链）和Ｃ肽编码的ＤＮＡ区域中核苷酸替换速率间的比较。成熟的胰岛素分子由一条Ａ链和一条Ｂ链，通过二硫键（ｓ）联结而成。自 Kimura(1983)修改而成。

对胰岛素的激素活性不起任何作用，而被认为只对产生该激素的正确三级结构有促进效果。结果，为Ｃ片段编码的区域的非同义替换速率，为Ａ链和Ｂ链编码的区域的平均非同义替换速率的 7 倍多（图 4 － 3 ）。然而，Ｃ片段上一定仍受着相当程度的限制，因为该区域中的非同义替换速率还是比较低的，它与β－珠蛋白中的相应速率大致相当（表 4 － 1 ）。

基因间的变异

为了对基因间非同义替换速率方面的大变异作出解释，我们必须再次考虑这样两个可能的肇事者：突变率和选择强度。不同的基因有相同的突变率，这样的假定在这种情况下可能不能成立，因为基因组的不同区域可能有着不同的突变倾向。沃尔夫等（Wolfe 等,1989a）曾提出，哺乳动物细胞核基因组的不同区域，在突变率方面的差异相互间以一个数值为 2 的因子来区别。然而，不同基因组区域间突变率上出现 2 倍的差异，甚至不能算作造成非同义替换速率方面将近 1 0 0 0 倍的出入的部分原因。所以，决定非同义替换速率的最重要因素看来还是选择强度，它又转而由功能限制所决定。

为了说明功能限制的效应，让我们考虑一下载脂蛋白和组蛋白 3，它们有着差异显著的非同义替换速率。载脂蛋白是脊椎动物血液中各种脂类的主要载体，而它们的脂类结合区大部分由疏水性残基所组成。对来自哺乳纲各目的载脂蛋白的顺序比较分析表明，该区域内由一个疏水性氨基酸（比如缬氨酸、亮氨酸）去替换另一个疏水性氨基酸，这在许多位点上都是可以接受的（Luo 等,1989）。这种不太严格的结构要求可用来解释为什么这些基因中的非同义的速率会相当高（表 4 － 1 ）。

处于另一个极端的是组蛋白 3 。因为组蛋白 3 中的大多数氨基酸在核小体（图 4 － 4 ）形成时，将直接与 D N A 或其他核心组蛋白相互作用，所以，可以合理地假定，只有很少几种可能的替换能在不妨

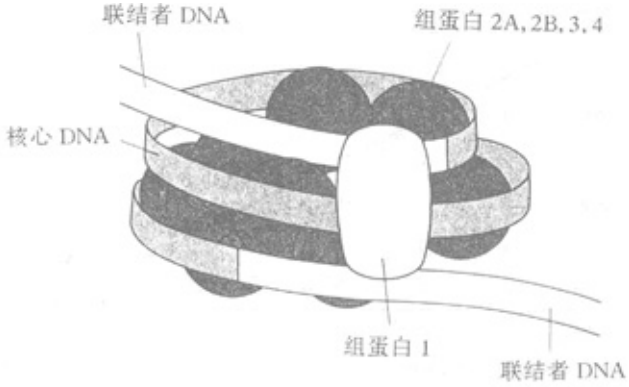


图 4 － 4 一个核小体的示意图。D N A 双螺旋（黑色带）围着核心组蛋白（组蛋白 2 A， 2 B， 3 和 4 各二个）缠绕。组蛋白 1 （淡灰色）与该核心粒子的外部和联结者 D N A （白色带）相结合。自 Stryer (1988)修改而成,

碍该蛋白质功能的条件下发生。此外，组蛋白 3 还必须维持其严格的致密性和高度碱性这对与酸性的 DNA 分子相互作用是必要的。结果,组蛋白 3 对大多数分子变化都绝不容忍。事实上，这种蛋白质是已知的进化最为缓慢的蛋白质之一，比载脂蛋白要慢 1 0 0 0 多倍。

同义替换的速率为什么也会因基因而异还不太清楚。出现这种变异可能有两个原因。首先，基因组的不同区域间突变率可能是不同的，因而同义替换速率上的变异可能简单地反映出基因所处的染色体位置（Wolfe 等,1989a）。这种可能性因这样的事实而得到进一步的支持，即，真核生物的基因组是由被称为同质段的明显地以 G C 为内容的片段所构成的，这些片段可能是独立复制的因而可能表现出不同的突变率（第八章）。第二个原因可能是，在某些基因中，并不是所有密码子都是在适合度上等价的。结果，有些同义替换可能会受到选择的排斥。这种纯洁化选择就会在基因间产生同义替换速率方面的变异。然而，虽然纯洁化选择已被证明能影响同义替换的速率，能影响细菌、酵母和果蝇的基因组中同义密码子的使用模式，但现在还不清楚这类选择是否在哺乳类中发生作用（见第 6 1 页）。

还有一种现象也曾受到注意，即一个基因中的同义替换速率和非同义替换速率间存在一种正相关（Graur,1985; Li 等,1985 b ）。若假定突变率随基因而变（因此有些基因的同义和非同义的替换速率将都很高），或者假定同义位置上的选择大小受邻近的非同义位置上核苷酸组成的影响，则该现象即可得到解释（Ticher 和 Graur,1989）。

4.3 一个正选择例子：乳牛和叶猴的溶菌酶

如前面的章节所讨论的那样，基因组的绝大多数基因和非基因区域中核苷酸替换的速率和模式，都可以通过①突变输入，②中性或近中性等位基因的随机遗传漂变，和③排斥有害等位基因的纯洁化选择，这三方面的结合来加以解释。然而，在溶菌酶的例子中，对有利突变的正选择曾被证明在某些哺乳类谱系

中起作用。

前肠发酵消化曾在有胎盘哺乳类的进化中独立地两次出现，一次在反刍动物（例如乳牛）中，另一次在疣猴类（例如叶猴）中。在这两种情况中，对别的哺乳动物来说通常不在胃中分泌产生的溶菌酶，它能补充进来，以消化在前肠执行发酵任务的细菌的细胞壁。斯图尔特和威尔逊（Stewart 和 Wilson,1987）曾对来自乳牛、叶猴、狒狒、人、大鼠、马和鸡的溶菌酶进行过氨基酸顺序比较（表 4 — 4 ）。他们注意到，乳牛和叶猴间有 4 个独特地共有的氨基酸。对这一观察结果有两种可能的解释。第一种：有可能乳牛在进化上与叶猴的亲缘关系比与马的更近，于是这些独特地共有的氨基酸，只代表出现在它们的共同祖先中未发生改变的氨基酸顺序。乳牛和叶猴间系统发生关系更近的假定已知是错误的。另一种，这些在该两物种中独特地共有的氨基酸，可能是独立地发生在两个谱系中的一系列平行替换的结果。事实上，将氨基酸替换的顺序重建后，斯图尔特和威尔逊（Stewart and Wilson,1987）发现在乳牛和叶猴谱系中有 7 个平行或趋同替换（图 4 — 5 ）。

表 4 — 4　不同物种的溶菌酶间顺序的成对比较 a

物种	物种					
	叶猴	狒狒	人	大鼠	乳牛	马
叶猴		14	18	38	32	65
狒狒	0		14	33	39	65
人	0	1		37	41	64
大鼠	0	1	0		55	64
乳牛	4	0	0	0		71
马	0	0	0	0	1	

自 Stewart 和 Wilson(1987)。

a　对角线上的数为物种间氨基酸的差异数，而对角线下的数为物种间独特地共有的残基数。

而且，已经确定，这些替换中有些对溶菌酶在低 pH 值下更好地行使功能有贡献，象在反刍动物消化系

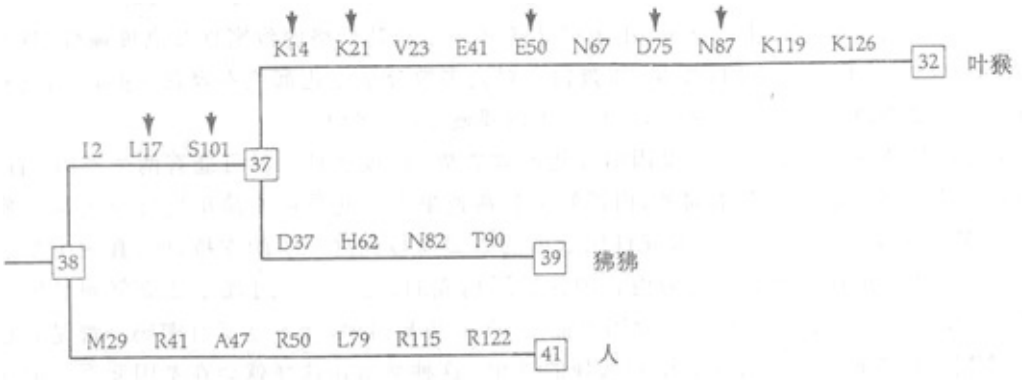


图 4 — 5　乳牛和叶猴溶菌酶中的平行或趋同氨基酸替代。谱系的长度与沿该谱系发生的氨基酸替代数成正比。每次替代用替代后氨基酸的一字母缩写（表 1 — 1 ）表示，其后续数字表示替代发生的位置。箭头指出了叶猴中发生的 7 次替代，它们以与乳牛谱系平行或趋同的形式发生。与乳牛胃溶菌酶（没有画出）的氨基酸差异数写在方块之中。自 Stewart 和 Wilson,(1987)。

统中发现的那些酶。相反，叶猴和乳牛溶菌酶在高 pH 值下都不如人溶菌酶有效。最后，看来可以不出什么差错地推论，我们这里处理的是一个有利替换在不同进化路线中平行地发生，表现出对类似的选择因子平行地适应的例子。

4.4 分子钟

在对来自不同物种的血红蛋白和细胞色素 c 的蛋白质顺序比较研究中，朱克坎德尔和波林（Zuckerkandl 和 Pauling,1962,1965）以及马戈利阿什（Margoliash,1963）首次注意到，这些蛋白质中的氨基酸替换速率在不同的哺乳类谱系中近似相同。因此朱克坎德尔和波林（Zuckerkandl 和

Pauling,1965）提出,对任何给定的蛋白质而言,分子进化的速率在所有谱系中都对时间近似地恒定,换言之,就是存在着一种分子钟(molecular clock)。这一提议马上激起了人们对将大分子用于进化研究的极大兴趣。事实上，如果蛋白质是以恒定的速率进化着的，那么，它们将可用于决定物种分歧的年代，并用来重建生物间的系统发育关系。这将与通过测定放射性元素的衰变来决定地质年代类似。

分子钟假说也激起了大量的争论。例如，经典进化论学者们就反对这种说法，因为进化速率恒定的说法与在表型和生理学水平上进化速率的变化无常对不上号。当速率恒定假说用于估计人与非洲猿间的分歧时间，得到一个５００万年的估值（Sarich 和 Wilson,1967）时，该假说受到了特别强烈的反对。因为，在古生物学家中占统治地位的观点是，人和猿的分歧至少应在１５００万年前，两者差得太远。许多分子进化科学家们也对分子钟假说的正确性提出了异议。特别是古德曼（Goodman,1981） 以及他的同事们（Czelusniak 等,1982)。他们认为，进化速率常常在基因重复之后出现加速，而蛋白质顺序在适应性辐射的年代里进化要快得多。例如，他们主张，在基因重复使α和β血红蛋白分开之后出现了极高的氨基酸替换速率，而这种高替换速率则是由于改善血红蛋白功能的有利突变所造成的。

虽然速率恒定假说一直是有争议的，但它已广泛用于分岐时间的估计和系统发育树的重建中（Nei,1975; Wilson 等,1977)。所以，分子钟假说的正确性在分子进化中是一个生死攸关的问题。近几年中ＤＮＡ顺序数据的迅速积累为检验该假说提供了一个全新的机会。用这类数据与用蛋白质顺序数据相比可使我们更近地检验该假说，而与ＤＮＡ－ＤＮＡ杂交数据和免疫距离数据相比，则可得到更直接的解释。

相对速率测验

关于分子钟假说的争论常常引起有关物种分岐年代的异议。为了避免这一问题，萨里奇和威尔逊（Sarich 和 Wilson,1973）提出了一种不需要知道分岐年代的检验法。该检验法称相对速率测验（relative-rate test)，如图４－６所示。假定我们要比较谱系Ａ和Ｂ中的速率。于是，我们用第３个物种Ｃ作为参照物。我们应该确定，该参照物种的分岐过程发生得比物种Ａ和Ｂ间的分岐更早。例如，为了比较人和马来猩猩谱系中的速率，我们用一种猴作为参照物。

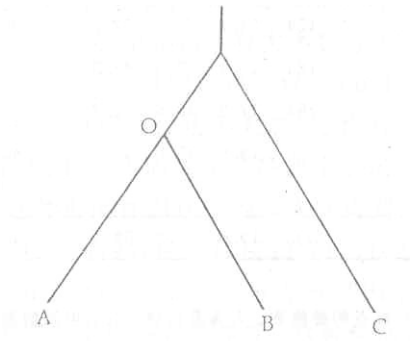


图４－６ 用于相对速率测验的系统树。O表示物种Ａ和Ｂ的共同祖先。

从图４－６很容易看出，物种Ａ和Ｃ间的替换数KAC 等于从点 0 到点 A 发生的替换数(K_{OA})和从点 0 到点 C 发生的替换数(K_{OC})之和。即，

$$K_{AC} = K_{OA} + K_{OC} \qquad (4.3a)$$

类似地，有：

$$K_{BC} = K_{OB} + K_{OC} \qquad (4.3b)$$

和

$$K_{AB} = K_{OA} + K_{OB} \qquad (4.3c)$$

因为K_{AC}, K_{KBC}和K_{AB}能从核苷酸顺序直接估出（第三章），所以，我们可以很容易地解出这３个方程，找到K_{OA}, K_{OB}和K_{OC}的值:

$$K_{OA} = \frac{K_{AC} + K_{AB} - K_{BC}}{2} \qquad (4.4a)$$

$$K_{OB} = \frac{K_{AB} + K_{BC} - K_{AC}}{2} \qquad (4.4b)$$

$$K_{OC} = \frac{K_{AC} + K_{BC} - K_{AB}}{2} \qquad (4.4c)$$

现在我们可以通过 K_{OA} 和 K_{OB} 的值,来决定替换速率在谱系 A 和谱系 B 中是否相等。自物种A和B最后地共有一个共同祖先以来所经过的时间，定义为对两个谱系是相等的。所以，按分子钟假说，K_{OA} 和 K_{OB} 应该相等，即，$K_{OA}-K_{OB}=0$。从等式 4.3a 和 4.3b 可以看出,$K_{OA}-K_{OB}=K_{AC}-K_{CB}$。故而，我们可以从 K_{AC} 和 K_{BC} 来直接比较 A 和 B 中的替换速率。

小鼠和大鼠中接近相等的速率

表 4-5 展示了用相对速率测验法进行的小鼠和大鼠中同义替换速率的比较。表中的物种 A 皆指小鼠,而物种 B 则全表示大鼠。因此,若 $K_{AC}-K_{BC}$ 的值为一正号,则表示小鼠中的速率高于大鼠中的，而若为负号则指示实际情况正好相反。

表 4-5 小鼠(物种 A)和大鼠(物种 B)间每 100 位点的同义替换数差异(KAC-KBC) a

基因	L	K_{AB}	K_{AC}	K_{BC}	$K_{AC}-K_{BC}$
载脂蛋白 E	201	7.4	61.3	59.5	1.8±5.3
肌动蛋白 α	249	17.9	58.2	59.1	−0.9±4.8
肌动蛋白 β	233	19.7	50.1	45.1	5.0±4.6
Thy-1 抗原	116	19.3	51.8	57.3	−5.5±6.9
乳酸脱氢酶 A	219	30.9	80.4	80.3	0.1±8.2
糖蛋白激素,α 亚基	58	30.8	97.7	84.3	13.4±18.5
胰岛素样生长因子 Ⅰ	130	4.8	37.0	40.9	−3.9±2.8
心房钠尿因子	107	20.4	69.7	57.4	12.3±8.3
生长激素	124	14.1	80.9	79.2	1.7±7.7
甲状腺球蛋白 β	90	25.7	77.4	92.7	−15.3±12.9
鸦片黑素皮质激素原	154	21.4	61.5	52.7	8.8±6.5
醛缩酶 A	184	15.4	57.5	63.3	−5.8±5.3
肌酸激酶 M	251	17.2	48.6	52.2	−3.6±4.3
金属巯基组氨酸三甲基内盐 Ⅰ	35	19.0	45.5	36.7	8.8±10.2
总计	2187	19.0	59.8	59.4	0.4±1.5

自 Li 等(1987a)

a. L = 受比较位点数。K_{ij} = 物种 i 和 j 间每 100 位点的替换数。人为参照物种（c），但肌酸激酶M为一例外，该行数据是用兔的顺序作为参照物得出的。

由于受资料限制，我们用作参照物的是人或兔的顺序，而不是从与小鼠和大鼠的亲缘关系更近的物种，象仓鼠或豚鼠中得到的顺序。结果，$K_{CA}-K_{BC}$ 的估值表现出较大的统计误差(表 4 — 5)。不过，小鼠和大鼠中的替换速率接近相等，这一点是相当明显的。换句话说，当将所有顺序一起考虑时，该速率差接近于 0 。关于这两个物种中非同义替换的速率，可得出同样的结论（Li 等,1987a)。

人中的速率低于猴的速率

根据免疫距离和蛋白质顺序数据，古德曼（Goodman,1961）及其同事们(Goodman 等,1971)提出，人科动物（人和猿）自它们从远古时代与猴分离后，出现了速率减缓。然而，威尔逊等（Wilson 等,1977）反驳说，减缓是一种人为现象，是用了对人—猿分岐时间的错误估值的结果。他们用免疫距离数据和蛋白

质顺序数据做了两次相对速率测验，结论是，没有任何表明人科动物速率减缓的证据。

ＤＮＡ顺序数据可对上述争论给出一个更好的解决。在表４－６中，相对速率测验法被用于比较人谱系和古世界猴谱系的核苷酸替换速率。在所有检验中，谱系Ｂ是人的谱系，而谱系Ａ为猴的谱系。谱系Ｃ是参照物种（见表下的注释）。因此，速率差（$K_{AC}-K_{BC}$）为正号意味着人谱系曾较慢地进化着，而为负号则意思相反。

表４－６ 古世界猴谱系（Ａ）和人谱系（Ｂ）间每１００位点的核苷酸差异数（KAC-KBC）a

顺序	位点数	K_{AB}	$K_{AC}-K_{BC}$
η-珠蛋白假基因	2000	7.4	2.1±0.7**
同义位点			
β-珠蛋白	71	8.9	2.8±5.6
载脂蛋白 A-1	158	7.9	−5.3±4.8
促红细胞生成素	145	11.2	5.1±5.9
α_1-抗胰蛋白酶	140	10.9	6.7±6.8
胰岛素	84	18.6	−7.5±7.2
内含子			
δ-珠蛋白	601	4.7	3.4±1.4
不翻译区和侧区域			
β-珠蛋白	179	4.6	1.2±1.7
δ-珠蛋白	172	8.8	6.1±3.2
总计	3550	6.7	2.3±0.6**

a 所用参照物种为枭猴（η－珠蛋白假基因），狐猴（β－和δ珠蛋白），小鼠或大鼠（促红细胞生成素，载脂蛋白Ａ－Ⅰ，和α１－抗胰蛋白酶）和狗（胰岛素）。

** 在１％水平与０差异显著。

我们注意到，９个所用的顺序中只有２个为负号。速率上的差异即使在只用η假基因时也是很显著的。若将所有顺序一起考虑，则 $K_{AC}-K_{BC}=2.3\%$，而 $K_{AB}=6.7\%$。因此，古世界猴谱系中的Ｋ值（K_{OA}）为(6.7%+2.3%)/2=4.5%，人谱系中的Ｋ值(K_{OB})只有 6.7%-4.5%=2.2%，表明猴谱系以 4.5/2.2≈2 倍于人谱系的速率更快地进化着。

啮齿类中的速率高于灵长类中的速率

吴和李（Wu 和 Li，1985）曾用相对速率测验法来比较啮齿类谱系和人类谱系中的替换速率，参照物则用偶蹄类或食肉类谱系。他们的结论是，啮齿类谱系同义替换的速率约为人类谱系的２倍。不过要注意，速率上的估计差异指长期内（即从啮齿类—灵长类分岐到现在的时期）的平均值。由于在发生分岐的那一时期和其后的短时期里两谱系中的替换速率应是相似的，所以，速率差异一定是随时间而增加的。因此，以上估计的速率差异值可能是一个低估值。要想知道更近时期的速率差异，我们需要分别估出啮齿动物内和灵长动物内的替换速率。

表４－７展示了灵长类和啮齿类中的替换速率的比较。如果我们假定人—黑猩猩的分岐发生在七百万年前，而人—古世界猴的分岐发生在二千五百万年前，则人和黑猩猩的顺序间平均速率为每年每位点 1.3×10^{-9} 替换，而人与古世界猴的顺序间则为每年每位点 2.2×10^{-9} 替换。这一结果与古世界猴谱系以２倍于人谱系的速率进化的结论是一致的。如果我们假定小鼠—大鼠的分岐发生在一千五百万年前，则小鼠和大鼠的顺序间平均速率为每年每位点 7.9×10^{-9} 替换。因此，啮齿类中的速率可能是较高等的灵长类中的速率的４到６倍。虽然关于所用的分岐时间还不是很可靠，但啮齿类顺序的进化比灵长类顺序要快得多，这一点是很明显的。

表４－７ 灵长类和啮齿类中每年每位点同义替换的平均速率 a

物种对	位点数	百分比分歧	替换速率($\times10^{-9}$)
人对黑猩猩	921	1.9	1.3(0.9—1.9)[b]
人对古世界猴	998	11.0	2.2(1.8—2.8)
小鼠对大鼠	3886	23.7	7.9(3.9—11.8)

自 Li 等(1987a)

a 所用分歧时间，人—黑猩猩为７（５－１０）百万年前，人—古世界猴为２５（２０－３０）百万年前，小鼠—大鼠为１５（１０－３０）百万年前。 b 括号内的值为从分歧时间的上限估值和下限估值得到的速率估值构成的范围。

不同进化谱系间替换速率上出现变异的原因

猴的替换速率高于人的以及啮齿类的替换速率高于灵长类的，这也许能用所谓世代时间效应（generation time effect） 来加以解释（Kohne,1970）。啮齿类的世代时间比人的要短得多，所以，如果在这些生物间每世代的种系复制没有很大差别的话，则每年的种系ＤＮＡ复制的次数在啮齿类中就可能比在人类中要高许多倍。因为突变大多在ＤＮＡ复制的过程期间积累，所以，复制的周期数越多，则突变错误也将发生得越多。这一因素也许能在很大程度上解释啮齿类的替换速率高于人类的替换速率的现象。类似地，猴有比人短的世代时间，所以，应该预期它将有较高的替换速率。

替换速率上的差异也可部分地归因于ＤＮＡ修复系统的效率方面的差异（Britten,1986）。已有的有限资料表明，啮齿类有比人类效率低的ＤＮＡ修复系统，因而，在每一复制周期中将积累更多突变。 以上结果不应拿来作为不存在分子钟的证据。我们注意到，替换速率上的差异是在具有很不相同的世代时间的生物间被观察到的。当具有相近的世代时间的生物，象小鼠和大鼠进行比较时，速率恒定规律表现得相当明显。所以，虽说没有一个关于所有哺乳类的全球性时钟，但关于许多亲缘关系较近的物种类群的地方性时钟也许是存在的。

4.5 细胞器ＤＮＡ中的替换速率

与细胞核基因组相比，细胞器基因组要小得多，也更容易进行实验研究。而且，在哺乳动物线粒体基因组中替换速率特别高（Brown 等,1979）。这一发现激起了人们对细胞器ＤＮＡ的进化问题的更大兴趣。

哺乳动物线粒体基因组由一个环状、双链ＤＮＡ组成。长为 15,000-17,000 碱基对（ｂｐ），近似地相当于最小的动物细胞核基因组的 1/10,000。它只含有单一的（即非重复的）序列：１３个为蛋白质编码的基因，２个ｒＲＮＡ基因，２２个ｔＲＮＡ基因和一个调控区，后者含有复制和转录的起始位点。该基因组在结构上是非常稳定的，这一点，从不同种的哺乳动物间其基因组大小变异不大即可看出。

与其成鲜明对照的是，植物的线粒体基因组却展现出较大的结构变异性。它们经历了频繁的重排、重复和缺失（Palmer,1985）。为此，基因组大小在 40,000bp 到 2,500,000bp 的范围内变化。植物中的线粒体基因组可能是线状的，也可能是环状的，而在许多情况中，遗传信息被分割成相互独立的ＤＮＡ分子，后者被称为亚基因组环。植物线粒体的编码内容还没有全部确定；不过，我们确已知道，有３个确定ｒＲＮＡ的基因、数目还不清楚的ｔＲＮＡ基因和大约１５个—３０个为蛋白质编码的基因，其中有些已被鉴别出来了。(在植物线粒体基因组中结构基因可能以多重拷贝出现。)目前，尽管植物线粒体基因组在大小上有很大变异性，但在编码内容上却没有表现出性质变异的迹象。

维管植物的叶绿体基因组是环状的，大小在 120,000 到 220,000bp 的范围内变化,平均大小为 150,000bp(Palmer,1985)。尽管在大小上有如此大的变异，但该基因组已知在结构上是稳定的。烟草(Nicotiana tabacum) 的叶绿体基因组已经被完全定序了（Shinozaki 等,1986）。它是一个环状分子，长 155,844bp 含有 37 个 tRNA 基因（其中 8 个含有单内含子），８个ｒＲＮＡ基因，和４５个为蛋白质编码的基因（其中 5 个含有单内含子，其中 2 个含有两内含子)。两条链都用于编码。 Nicotiana tabacum

的叶绿体基因组还含有 5 9 个功能不知的外加开读框架，其中 2 个则插进了内含子。

哺乳动物线粒体基因中同义替换的速率已估出，为每年每同义位点 5.7×10^{-8} 替换(Brown 等,1982)。这大约是细胞核中为蛋白质编码的基因的同义替换值的 1 0 倍。非同义替换的速率在 1 3 个为蛋白质编码的基因中变化很大，但通常都比细胞核基因的平均非同义替换速率大得多。哺乳动物线粒体中这些高替换速率的原因，看来是相对于细胞核而言它有较高的突变率。高突变率则是由于（ a ）线粒体中的 D N A 复制过程保真度低，（ b ）缺乏修复机制或修复机制效率极差，和（ c ）诱变剂浓度高（例如超氧化物基团 O_2^-），后者是线粒体执行代谢功能的结果。另一方面，作用在非同义突变上的纯洁化选择的强度，看来与作用在细胞核基因上的属同一数量级。

根据几个基因顺序或限制酶图谱资料进行的早期研究指出，叶绿体基因有比哺乳动物细胞核基因低的核苷酸替换速率(Curtis 和 Clegg,1984;Palmer,1985)，用核苷酸替换表示则植物线粒体 D N A 进化缓慢，虽然它频繁地经历着顺序重排（Palmer 和 Hebron,1987)。这些结果近来被更广泛的 D N A 顺序分析所证实（Wolfe 等,1987,1989)。

表 4 — 8 展示了高等植物的这 3 种基因组中替换速率的比较。每非同义位点的平均替换数（K_A）在叶绿体和线粒体基因组中是相似的，但每同义位点的平均替换数（K s ）却很不相同，在单子叶植物与双子叶植物间比较，叶绿体基因组中的K s 几乎是线粒体基因组中的 3 倍；而在玉米与小麦或大麦间比较则前者是后者的 6 倍。以下，我们将采用前一个比值，因为它是根据更大的数据组而得到的。植物细胞核基因中的平均同义替换速率约为叶绿体基因的 4 倍。于是植物线粒体、叶绿体和细胞核基因中的同义替换速率，近似地呈 1∶3∶12 这样的比例。

表 4 — 8 植物叶绿体、线粒体和细胞核基因中核苷酸替换速率的比较[a]

基因组	K_S	L_S	K_A	L_A
单子叶与双子叶植物间的比较				
叶绿体基因	0.58±0.02	4177	0.05±0.00	14421
线粒体基因	0.21±0.01	1219	0.04±0.00	4380
玉米与小麦或大麦间的比较				
细胞核基因	0.71±0.04	1475	0.06±0.00	5098
叶绿体基因	0.17±0.01	2068	0.01±0.00	7001
线粒体基因	0.03±0.01	413	0.01±0.00	1526

自 Wolfe 等, (1987,1989b)

a . K s : 每同义位点的替换数; KA: 每非同义位点的替换数; L s : 同义位点数; LA: 非同义位点数。

如果我们把玉米与小麦间的分岐时间取为 5 0 — 7 0 百万年（Stebbins , 1981;Chao 等,1984)，那么，表 4 — 8 中关于细胞核的数据则表现出一个每年每位点 5.1-7.1×10^{-9} 替换的平均同义速率。这与在哺乳类细胞核基因中看到的同义替换速率（表 4 — 1 ）相似。

有趣的是，细胞器的基因组中核苷酸替换的速率与结构变化的速率无关。在哺乳类中，以核苷酸替换表示的线粒体 D N A 的进化非常迅速，但其基因的空间排列和基因组的大小却在各物种间保持稳定。相反，植物的线粒体基因组经历了频繁的结构变化，但其核苷酸替换速率却极低。在叶绿体 D N A 中，核苷酸替换速率和结构进化都很缓慢。替换速率和结构进化速率间无相关性，这表示两个过程是独立地进行的。

4.6 假基因中的核苷酸替换模式

因为点突变是 D N A 序列进化中最重要的因素之一，所以分子进化学家们长期以来一直对决定自发突变的模式（pattern of spontaneous mutation） 怀有兴趣（例如，Beale 和 Lehmann,1965;Zuckerkandl 等,1971)。该模式可以被当作一个标准，用于推论：任一给定 D N A 序列中核苷酸间相互变换的观察频率与在无选择下，即在选择中性（selective neutrality）下，预期的值究竟偏离多远。

研究点突变模式的途径之一，是检验不受选择限制的 D N A 区域中的替换模式。假基因在这里特别

有用。由于它们无功能，所以，所有发生在假基因中的突变都是选择中性的，且以相同的概率在群体中固定。于是，假基因中的核苷酸替换模式预期将反映出自发点突变的模式。

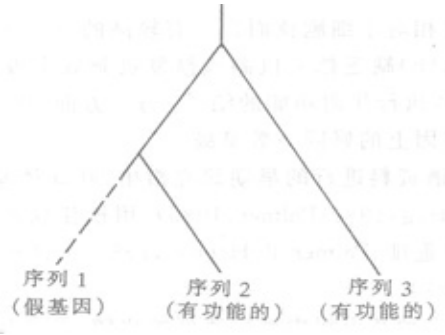


图 4 - 7 表示推论假基因序列中的核苷酸替换的一种简单方法（Gojobori 等,1982;Li 等,1984）。

序列 1 是一个假基因，序列 2 是其有功能的对应物，来自同一物种，而图 4 － 7 推论假基因序列中核苷酸替换模式的系统树。虚线的含义是无功能。序列 3 则是假基因出现前即发生分岐的功能序列。假定在某一核苷酸位点上，序列 1 和 2 分别各为Ａ和Ｇ。那么，我们可以假定，若序列 3 中该位点上为Ｇ则假基因序列中该核苷酸从Ｇ变为Ａ，但若序列 3 中该位点上为Ａ则序列 2 中该核苷酸从Ａ变成Ｇ。不过，如果序列 3 中的是Ｔ或Ｃ，则我们不能决定变化的方向，而若出现这种情况则该位点将从比较中排除。由于假基因中的替换速率通常大大高于其同源的功能基因中的速率，所以，一个基因与一个假基因间核苷酸顺序上的差异，在绝大多数情况下都被认为是假基因中的改变所造成的。

表 4 － 9 中的矩阵代表从 1 3 种哺乳类假基因序列推论出的替换的联合模式。矩阵中的每一个项 f_{ij} ，都代表一个随机序列（即四种碱基以相同的频率出现其中的序列）中每 1 0 0 次替换里碱基从 i 变成 j 的期望次数。例如， f_{AT}＝4.7，表示所有替换的 4.7%为从Ａ变成Ｔ。

表 4 － 9　假基因中的替换模式　a

从	到 A	到 T	到 C	到 G	行总计
A	—	4.7±1.3 (5.3±1.4)	5.0±0.7 (5.6±0.8)	9.4±1.3 (10.3±1.4)	19.1 (21.2)
T	4.4±1.1 (4.8±1.1)	—	8.2±1.3 (9.2±1.3)	3.3±1.2 (3.6±1.3)	15.9 (17.6)
C	6.5±1.1 (7.1±1.3)	21.0±2.1 (18.2±2.3)	—	4.2±0.5 (4.2±0.6)	31.7 (29.5)
G	20.7±2.2 (18.6±1.9)	7.2±1.1 (7.7±1.3)	5.3±1.0 (5.5±1.3)	—	33.2 (31.8)
列总计	31.6 (30.5)	32.9 (31.2)	18.5 (20.3)	16.9 (18.1)	

自 Gojobori 等,(1982)和 Li 等,(1984)

a　表中项为以 1 3 种哺乳类假基因序列为根据推论出的，碱基从 i 变为 j 的百分数（ f_{ij} ）。括号中的值是把所有ＣＧ二核苷酸从比较中排除后得到的。

我们注意到，突变的方向是非随机的。例如，Ａ变成Ｇ比变成Ｔ或Ｃ更常发生。从右上角到左下角的对角线上的 4 个元是转换的 f_{ij} 值，其余的 8 个元代表颠换。所有转换，特别是Ｃ→Ｔ和Ｇ→Ａ，都比颠换更常发生。转换的相对频率之和为 59.2% （若ＣＧ二核苷酸被排除则为 54.4%，见下）。我们注意到在随机突变下转换的期望比例仅３３％，因为只有 4 种转换却有 8 种颠换。该观察比例几乎是在随机突变下预期值的两倍。

我们也注意到，有些核苷酸比另一些要更容易突变。在表 4 － 9 的最后一列，我们列出了从Ａ、Ｔ、Ｃ和Ｇ突变成别的核苷酸的相对频率。如果 4 种核苷酸都有相同的突变性，则我们应期望该列中的每一个元都有一个２５％的值。实际上，我们看到，Ｇ以 33.2%的相对频率突变（即，Ｇ是一种高可突变的核苷

酸），而T则以15.9%的相对频率突变（即它达不到那种可变程度）。在表4－9的最后一行，我们列出了所有经突变而变成A、T、C和G的相对频率。我们注意到，所有突变的64.5%是变成A或T的，而随机过程期望的值应为50%。由于C和G有一种频繁地变成A或T的倾向，又由于A和T不如C和G那样可突变，所以，假基因预期应变得富含A和T。这对其他一些不受功能限制的非编码区应该也是成立的。事实上，非编码区普遍发现是富AT的。

表4－9中的结果是根据有意义的链，即未被转录的链而得出的。所以，从G到A的变化实际上意味着一个G：C对被一个A：T对所取代。这种情况的出现，可以是有意义的链中G突变成A的结果，也可能是与前者互补的链中C突变成T的结果。类似地，从C到T的变化，也可能是在一条链中C突变成T或在另一条链中G突变成A的结果。如果两条链间突变的模式没有差别，那么我们应有 $f_{GA}=f_{CT}$ 。类似地，我们应能得到 $f_{AG}=f_{TC}$,$f_{AT}=f_{TA}$,$f_{AC}=f_{TG}$,$f_{CA}=f_{GT}$ 和 $f_{CG}=f_{GC}$。这些等式仅近似地成立，且事实上这两条链间的突变模式可能存在较小的不对称性。这种不对称性可能是由于DNA复制期间，先导链和滞后链间在复制机制方面有差异所造成（Wu和Maeda, 1987）。

已知从C到T的转换，除了碱基误配以外，还可能是从甲基化了的C残基经脱氨而变成T残基，这样一种转变过程而实现的（Coulonder等,1978;Razin和Riggs,1980）。该作用将提高C：G→T：A和G：C→A：T，即 f_{CT} 和 f_{GA} 的频率。由于脊椎动物DNA中约90%的甲基化了的C残基发生在5′－CG-3′二核苷酸中（Razin和Riggs,1980），所以，该效应将主要以CG二核苷酸变成TG或CA的形式表现出来。一个基因变成假基因后，这类变化将不再受任何功能限制，因而，如果在基因的沉默化（silencing）（即失去功能）发生前CG的频率相对而言较高，则它能为C→T和G→A转换作出显著贡献。对看来曾在这假基因的祖先序列中出现过CG二核苷酸的那些位点予以排除，由此而得到的替换模式在表4－9的括号中给出。此模式也许更适合于：预测一个长期不受功能限制的序列（例如一个内含子的某些部分）中的突变模式，因为在这样的序列中将只有少量CG二核苷酸存在。排除CG二核苷酸后得到的模式有点不同于不经排除而得到的模式。特别地，4种转换间的相对频率差异显著性略有降低，而各颠换的相对频率，除G→C和C→G外，则略有升高。

4.7 同义密码子的非随机应用

由于遗传密码的简并，20种氨基酸中大多数都是由一个以上的密码子编码的（第一章）。因为同义突变不造成氨基酸顺序中的任何变化，且因为自然选择被认为主要在蛋白质水平上起作用，所以，同义突变曾被当作选择上呈中性的突变的候选者（Kimura,1968;King和Jukes, 1969）。然而，若所有同义突变事实上都是选择中性的，那么，为同一个氨基酸编码的同义密码子就应该以多少有点相同的频率应用。不幸的是，随着DNA顺序资料的积累，逐渐表明同义密码子的应用，在原核生物和真核生物的基因中都显然是非随机的（Grantham 等,1980）。事实上，在许多酵母的基因和大肠杆菌的基因中，应用上的偏斜是极显著的。例如，在大肠杆菌（*Escherichia coli*）外膜蛋白Ⅱ（*omp* A）中的23个亮氨酸残基里，有21个由密码子CUG编码，尽管为亮氨酸编码的还有5种密码子。这种偏斜不能用非随机突变来解释。如何解释这种广泛存在的密码子非随机应用现象，成了一个有争议的问题，好在对此问题看来已出现了一些明确的答案。

有助于理解非随机应用现象的一个观察事实是，一个生物中或有亲缘关系的物种中的基因，一般表现出对同义密码子的选取有同样的模式（Grantham 等,1980）。于是，哺乳动物、大肠杆菌和酵母的基因被归为不同的密码子应用类型。格兰瑟姆等（Grantham 等,1980） 因此而提出了基因组假说（genome hypothesis）。按其假说，任何给定基因组中的基因在同义密码子的选取方面都采用同样的编码策略，即在密码子应用上的偏斜是物种特异的。基因组假说被证明一般说来是正确的，虽然在一个基因组的不同基因间密码子应用有着相当大的异质性（见下面）。

大肠杆菌和酵母中密码子应用的研究，极大地增加了我们对影响同义密码子选取的因素的认识。波斯特等（Post 等,1979）发现,大肠杆菌核糖体蛋白基因,优先应用于被含量最多的 tRNA 种类识别的同义密码子。他们认为，这种偏向是自然选择的结果，因为应用由含量最多的tRNA种类翻译的密码子，将会增加翻译的效率和精确性。他们的发现曾激励池村（Ikemura,1981,1982）去收集有关大肠杆菌和酵母

Saccharomyces cerevisiae （酿酒酵母）中各ｔＲＮＡ种类的相对丰度的资料。他证明，在这两个物种中，一个基因中同义密码子的相对频率与识别它们的ｔＲＮＡ种类的相对丰度间存在着正相关。对于高度地表达的基因而言，这种相关非常强。例如，在大肠杆菌中４种亮氨酸ｔＲＮＡ里含量最多的是$tRNA_1^{Leu}$，它识别ＣＵＧ密码子，而 ompA 基因也主要用这种密码子为亮氨酸编码（见上面）。 图４－８表示６个亮氨酸密码子的频率和识别它们的ｔＲＮＡ的相对丰度间的对应关系。在大肠杆

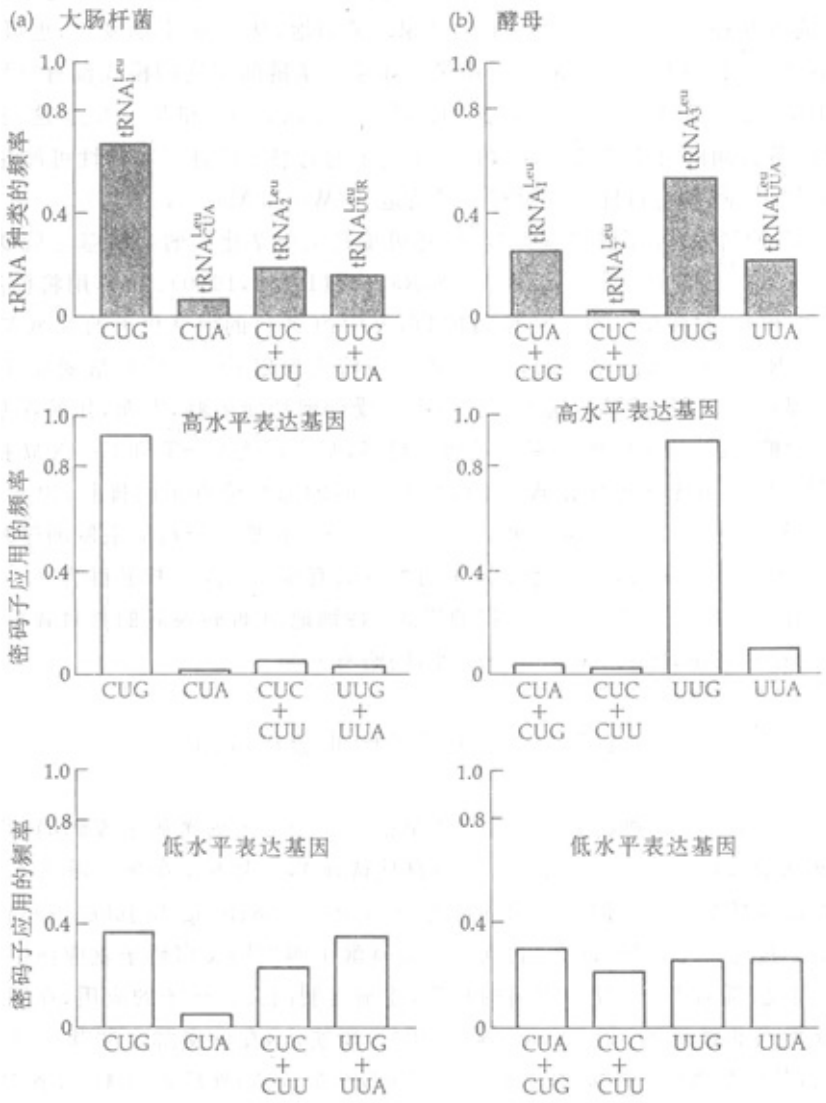


图４－８ 关于亮氨酸的密码子应用的相对频率（白柱体），和相应的识别ｔＲＮＡ种类的相对丰度（黑柱体）间关系的图解说明，（ａ）在大肠杆菌中，和（ｂ）在 Sacharomyces cerevisiae 中。加号（例如 E.coli 中密码 CUC 和 CUU 间的符号）表示这类密码子对中的密码子都是由同一种 tRNA 来识别的(例如，E. coli 中 CUC 和 CUU 都由 $tRNA_2^{Leu}$ 来识别)。

菌中，$tRNA_1^{Leu}$是含量最多的ｔＲＮＡ种类，而事实上，在高水平表达的基因中，ＣＵＧ（由这种ｔＲＮＡ识别的密码子）的应用比另外５种密码子要频繁得多。另一方面，在酵母中，含量最丰富的亮氨酸ｔＲＮＡ种类是$tRNA_3^{Leu}$，而被这种ｔＲＮＡ识别的密码子（ＵＵＧ）也是数量上占优势的密码子。对比之下，在以较低水平表达的基因中，ｔＲＮＡ丰度和各密码子间的对应在这两个物种里都要弱得多（图４－８）。

在决定高水平表达的基因中密码子的应用模式方面，翻译效率的重要性已得到以下观察的进一步支持(Ikemura, 1981)。已知密码子－反密码子配对在第３位上出现摇摆（wobbling）。例如，反密码子的第１位上的Ｕ既可与Ａ也可与Ｇ配对。类似地，Ｇ既可与Ｃ也可与Ｕ配对。但是，反密码子第１位上的Ｃ则只能与密码子第３位上的Ｇ配对，以及Ａ只能与Ｕ配对。摇摆还可能通过这样的事件来实现：有些ｔＲＮＡ在第１反密码位置上含有经修饰过的碱基，而这类ｔＲＮＡ能识别一种以上的密码子。例如，次黄嘌呤（一

种经过修饰的腺嘌呤）可与U、C、A三种碱基中的任何一种配对。有趣的是，大多数能识别一种以上密码子的tRNA，都表现出对其中的某一种有不同的偏爱。例如，反密码子的摇摆位置上的4－硫尿嘧啶（S^4U），可以识别密码子摇摆位置上的A和G;然而，与以G结尾的密码子相比，它对以A结尾的密码子表现出明显的偏爱。这种偏爱在高度地表达的基因中应会反映出来。大肠杆菌中两个为赖氨酸编码的密码子是由一种tRNA识别的，该tRNA分子在反密码子的摇摆位置上有S4U，而事实上，在大肠杆菌的ompA基因中，19个赖氨酸密码子里15个是AAA，只有4个是AAG 。

表4－10列出了由夏普等（Sharp等，1988）广泛收集的密码子应用资料中的一部分。对每一组同义密码子来说，如果应用机会均等，则每种密码子的相对频率应该是1。然而，大多数情况下显然并非如此。而且，在大肠杆菌和酵母这两个物种中，密码子应用偏斜都是在高水平表达的基因中比在低水平表达的基因中更严重。对此差异的一个简单解释是，在高水平表达的基因中对翻译效率和精度的选择要强一些，所以密码子应用偏斜也就显著一些。另一方面，在低水平表达的基因中选择相对而言较弱，所以，该应用模式主要受选择压力和随机遗传漂变的影响，因而偏斜程度也低一些（Sharp和Li，1986）。 **表4－10 4个物种中的密码子应用** a

氨基酸	密码子	*Escherichia Coli*		*Saccharomyces Cerevisiae*		*Drosophila Melanogaster*		人	
		高	低	高	低	高	低	G+C	A+T
Leu	UUA	0.06	1.24	0.49	1.49	0.03	0.62	0.05	0.99
	UUG	0.07	0.87	5.34	1.48	0.69	1.05	0.31	1.01
	CUU	0.13	0.72	0.02	073	0.25	0.80	0.20	1.26
	CUC	0.17	0.65	0.00	0.51	0.72	0.90	1.42	0.80
	CUA	0.04	0.31	0.15	0.95	0.06	0.60	0.15	0.67
	CUG	5.54	2.20	0.02	0.84	4.25	2.04	3.88	1.38
Val	GUU	2.41	1.09	2.07	1.13	0.56	0.74	0.09	1.32
	GUC	0.08	0.99	1.91	0.76	1.59	0.93	1.03	0.69
	GUA	1.12	0.63	0.00	1.18	0.06	0.53	0.11	0.80
	GUG	0.40	1.29	0.02	0.93	1.79	1.80	2.78	1.19
Ile	AUU	0.48	1.38	1.26	1.29	0.74	1.27	0.45	1.60
	AUC	2.51	1.12	1.74	0.66	2.26	0.95	2.43	0.76
	AUA	0.01	0.50	.00	1.05	0.00	0.78	0.12	0.64
Phe	UUU	0.34	1.33	0.19	1.38	0.12	0.86	0.27	1.20
	UUC	1.66	0.67	1.81	0.62	1.88	1.14	1.73	0.80
Met	AUG	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

自Sharp等，(1988)

a 对每一同义密码子组来说，相对频率之和等于该组中的密码子数。例如，亮氨酸有6个密码子，所以关于这6个密码子的相对频率之和即应为6。在均等地应用下，一组中每一密码子的相对频率应为1，所以各值与1的偏差指示应用上偏斜的程度。“高”和“低”表示以高水平表达和以低水平表达的基因。对人来说，“G＋C”意味着高GC区，而“A＋T”意味着高AT区。

总之，在大肠杆菌和酵母中，同义密码子的选取受着 RNA可用性和其他与翻译效率有关的因素的限制。这些限制结果将以纯洁化选择表现出来，从而减缓了同义替换的速率（Ikemura，1981；Kimura，1983）。事实上，已经得到证明的是，肠道细菌的基因中同义替换的速率与密码子应用的偏斜程度呈负相关（Sharp和 Li，1986）。因此，同义密码子的非随机应用现象不能被当作反对分子进化的中性学说的证据，因为它可用选择限制越强结果进化速率就越低这一原理来加以解释（见第50页和Kimura，1983）。

表4－10还展示出，果蝇中的密码子偏斜在高水平表达的基因中比在低水平表达的基因中要严重得多，这指出对翻译效率的选择在决定这种生物里同义密码子的选取方面也起重要作用。

在许多人的基因中，密码子倾向于以G或C结尾（即在第3位置上有较高的GC含量），而在另一些基因中却有较低的第3位置GC含量。不过，有几种原因可用来说明为什么该偏斜可能与基因表达的水平无关。首先，α－和β－珠蛋白基因在密码子的第3位上有不同的GC含量（分别为高含量和低含量），

但它们在同样的组织（红细胞）中以近似相等的量表达，因而它们应该有相同的表达水平。其次，在鸡的基因中，密码子应用的频率与ｔＲＮＡ的可用性无关（Ouenzar 等,1988），虽然该观察也许不能直接地用于人的基因。最后，第3密码子位置上的ＧＣ含量与侧区域中和内含子中的ＧＣ水平有很强的相关性（第八章，Bernardi 和 Bernardi,1985;Aota 和 Ikemura,1986）。例如，α－珠蛋白基因ＧＣ含量高且它位于高ＧＣ区域，β－珠蛋白基因ＧＣ含量低而它位于低ＧＣ区域（第八章，Bernardi 等,1985）。于是，看来人基因中的密码子应用偏斜，极大地因含有该基因的区域中ＧＣ的含量而决定。正如将要在第八章讨论的那样，一个区域中的ＧＣ含量是由自然选择还是由突变偏斜所决定，这仍是一个有争议的问题。不过，由于一个基因中第3密码子位置上的ＧＣ含量倾向于高于其周围区域中的含量（第八章；Aota 和 Ikemura,1986），所以，有可能人的基因中密码子应用模式受到某种程度的自然选择的影响。要得到对影响人中密码子应用的因素的更多知识，还需要做进一步的研究。

习题

1　等式4.1的分母为什么是2ｐ不是ｐ？

2　图４－９表示来自橄榄狒狒（Papio anubis)和马来猩猩（Pongo pygmaeus)的Θ１－珠蛋白基因的第1和第2外显子的ＤＮＡ顺序。用一参数模型分别算出3个密码子位置中每一个的每位点替换数。哪一个位置进化得最快？为什么？

b: ATG GCG CTG TCC GCG GAG GAC CGG GCGGCT GTG CGC GCC CTG
o: ATG CGC CTG TCC GCG GAG GAC CGG GCGCTG GTG CGT GCC CTG
b: TGG AAG AAA CTG GGA AGC AAT GTT GGCGTC TAT GCT ACT GAG
o: TGG AAG AAG CTG GGC AGC AAC GTC GGCGTC TAC ACG ACA GAG
b: GCC CTG GAG AGG ACC TTC CTG GCT TTCCCC GCC ACG AAG ACC
o: GCC CTG GAG AGG ACC TTC CTG GCC TTCCCC GCA ACG AAG ACC
b: TAC TTC TCC CAC CTA GAC CTG AGC CCCGGC TCC GCC CAG GTT
o: TAC TTC TCC CAC CTG GAC CTG AGC CCCGGC TCC TCA CAG GTC
b: AGA GCA CAC GGC CAG AAG GTG GCG GACGCG CTG AGC CTC GCC
o: AGA GCC CAC GGC CAG AAG GTG GCG GACGCG CTG AGC CTC GCC
b: GTG GAG CGC CTA GAC GAC CTA CCC CGCGCG CTG TCC GCT CTG
o: GTG GAG CGC CTG GAC GAC CTA CCC CACGCG CTG TCC GCG CTG
b: AGC CAT CTG CAC GCT TGC CAG CTG CGAGTG GAC CCA GCT AAC
o: AGC CAC CTG CAC GCG TGC CAG CTG CGAGTG GAC CCG GCC AGC
b: TTC CCG
o: TTC CAG

图４－９ 来自橄榄狒狒（ｂ）和马来猩猩（ｏ）的Θ１－珠蛋白基因中，外显子1和2的ＤＮＡ顺序。资料取自 Shaw 等,(1987)和Marks 等,(1986)。

3　图４－１０表示来自橄榄狒狒和马来猩猩的Θ１－珠蛋白基因中第1个内含子的ＤＮＡ顺序。用（ａ）一参数模型，和（ｂ）两参数模型，算出替换数。这两个估计有差别吗？将此结果与你在习题2中得到的结果相比较，那么，该内含子的进化比外显子中3个密码子位置上的进化是快还是慢？

b: TGCGGCGAGGCTGGGCGCCCCCGCCCTCCGGGGCCCTGCCTCCCCAAGCC
o: TGCGGCGAGGCTGGGCGCCCCCGCCCCC-AGGGCCCTCCCTCCCCAAGCC
b: CCCCGGACGCGCCTCACCGCCGTTCCTCTCGCAG
o: CCCCGGACTCGCCTCACCCACGTTCCTCTCGCAG

图４－１０ 来自橄榄狒狒（b）和马来猩猩(o)的θ１－珠蛋白基因中第1个内含子的ＤＮＡ顺序。一个裂缝用　标出。资料取自 Shaw 等,(1987)和Marks 等,(1986)

4　图４－１１表示来自仓鼠、大鼠和小鼠的核仁素基因中的第1个内含子的部分ＤＮＡ顺序。以仓

鼠的顺序作为参照物，用相对速率检验法来决定，大鼠谱系和小鼠谱系间替换速率上是否存在差异。

m: GTAAGAGGCCTGGCGCGCCGACGCGGACGACTAGGCCTGCTTTCGGAGGG
r: GTAATAGGCCTGACGCGCGAACACGGACGACTAGGCCTGCTTTCTGAGAG
h: GTGAGAGGCCTCGCGCGCGCCGACGGACGGACGGGCCTGCTTTCTGAGGG
m: GCGCGCGCGCCGTCGCGGAGGGGAGGAGGGCTTGCGCGCAATCCCGGGCG
r: GCGCGCGCGCCGTCGCGGAGGGGAGGAGGGCCTGCGCACAGTCCCGGGCG
h: GCGCGCGCGCGGTCGCTCAGGGGAGGAGGGCCTGCGCGCAATCCCGGGCG
m: CGTTCGAGGGCGCCAGCTGGGGAACTCTCGCGCGACTAGCGGGAGGTCTC
r: CGTTCGAGGGCGCATGCTGGGGAAGTCTCGCGCGACTAGCGGAGGGTCTC
h: CGTTCGAGGGCGCATGCTGGGGAAGTCTCGCGCGACTAGCGGAGGGTCTC

图４－１１ 来自小鼠（ｍ）、大鼠（ｒ）和仓鼠（ｈ）的核仁素基因中第１个内含子的部分ＤＮＡ顺序。裂缝（缺失和插入）已被略去。资料取自 Bourbon 等,(1988)

５　艾滋（ＡＩＤＳ）病毒的两个品系用ＷＭＪ １和ＷＭＪ ２表示，在１９８４年１０月3日和１９８５年1月１５日从一个两岁的孩子身上分离出来(Hahn 等,1986)。这个孩子假定只受过一次感染（由她的母亲在围产期传给)。这两个分离物间外壳蛋白（env)基因中每同义位点的同义替换数为 0.0164(Li 等,1988)。（ａ）假定ＷＭＪ ２直接从ＷＭＪ １进化而来并假定这两个顺序在 1984 年 10 月 3 日分开,求同义替换速率的最大估值。（ｂ）假定这两个品系在被感染时即开始分岐，并假定它们已独立地进化了两年，求该基因同义替换速率的最小估值。这些替换速率比对哺乳动物的基因加以平均的同义替换速率（表４－１）要快多少？

６　从大肠杆菌、酵母和人中各找出一个完整的ｃＤＮＡ或基因顺序。对每一种基因编一个密码子应用表（即列出每种密码子在基因中被使用的次数)。那么，该密码子应用是否偏斜？在哪方面偏斜？密码子应用模式在这 3 种基因中是否相似？如果不是，则差异如何？用X^2检验法去判断，在每一种基因中缬氨酸密码子的应用与该族密码子机会均等地应用的偏差是否具有统计学意义。

后继阅读文献

Ikemura,T.1985.Codon usage and tRNA content in unicellular and multicellusa r organisms.Mol.Biol.Evol.2:13-34

Kimura,M.1983.The Neutral Theory of Molecular Evolution.Cambridge University Pre ss,Cambridge.

MacIntyre,R.J.(ed.).1985.Molecular Evolutionary Genetics.Plenum,New York.

Nei,M.1987.Molecular Evolutionary Genetics.Columbia University Press,New York.

Sharp,P.M.and W.　H.Li.1986 An evolutionary perspective on synonymous codon usag e in unicellular organisms.J.Mol.Evol.24:28-38

Steinhauer,D.A.and J.J.Holland.1987.Rapiel evolution of RNA viruses.Annu.Rev.Mic robiol.41:409-433.