遗传学报，21 (1): 17—23，1994
Acta Genetica Sinica

DNA 进化马尔可夫过程模型的评价与推广①

杨子恒　　Nick Goldman
（北京农业大学畜牧系　北京　100094）　（英国国家医学研究所　伦敦）

摘要　本文对 DNA 序列进化过程中核苷酸替代的随机模型进行了评价，对替代速率在时间和空间上不恒定的情形进行了考察与推广。Lanave 等(1984)曾提出一个模型，宣称对替代的模式未做任何假定，但事实上我们证明它假定替代过程是可逆的。运用 2-p、4-p 和 6-p 模型进行的计算表明替代速度在位点间的差异会造成估计的替代数严重偏低，并且替代数越大，偏差也越大。替代模式在位点间的差异也会造成估计值偏低，但偏差不严重。运用非齐次马氏过程模型研究了替代速度在时间上的变异，结果表明估计的替代数反映了速率在时间上的平均值。文中还讨论了本文结果与进化树推断的关系。

关键词　核苷酸替代，随机过程模型，分子进化

DNA 序列资料的飞速积累使得我们可以在 DNA 水平上直接研究进化过程。通过估计同源序列间的进化距离我们可以推断物种间的分化顺序及分化年代。为此，通常假定核苷酸的替代过程为时齐马尔可夫过程，这样序列间的距离，即每位点核苷酸替代数就跟时间呈线性关系。然而，这些模型包含了一些假定[2,15]：(1)替代速率在进化路径间及进化路径内都是恒定的。(2)不同位点上的替代速率相等且替代过程独立。(3)没有(正)选择的作用，即只有消除不利突变、影响替代速率的纯化选择。近来有不少的证据表明这些假定有时是不现实的[2,3,4,7]，可是对于违反这些假定会有什么效果尚缺乏研究。

在上述假定下，齐次马氏过程模型的替代速率矩阵是一个 4 × 4 矩阵，有 12 个自由参数。为了估计两序列间的进化距离，通常需要对速率矩阵的结构做进一步的约束以减少未知参数的个数，例如 1-p（参数）[8]、2-p[9]、4-p[14] 及 6-p[6] 模型即是如此。Lanave 等[10]提出的模型则宣称对速率矩阵未做任何约束，从而为"最通用的模型"，但事实并非如此。

本文中我们首先对 Lanave 等的模型做一考察，然后对速率在位点上及时间上的变异的效果进行研究。

1　速率矩阵 Q 的结构及 Lanave 等[10]模型的评价

假定序列中任一位点上核苷酸的替代过程为一时齐马尔可夫过程。以 $P(t)=\{p_{ij}(t)\}$ 表示时间 t 内的转移概率矩阵，其中 $p_{ij}(t)$ 为一位点在时间 0 时为 i 的条件下，时间 t 时为 j 的概率，i,j 取值 1、2、3、4，对应 T、C、A、G。这时

$$Q = P'(0) = \left.\frac{d(P)(t)}{dt}\right|_{t=0} = \lim_{t\to 0}(P(t) - I)/t \tag{1}$$

称之为速率矩阵，因为 $q_{ij}\Delta t(i \neq j)$ 表示核苷酸 i 在时间间隔 Δt 内突变为 j 的概率。

① 本文于 1992 年 4 月 28 日收到

$$q_{ii} = -\sum_{j \neq i} q_{ij} = -q_i$$

Q 与 $P(t)$ 互相唯一决定[1]。且 $P(t) = e^{Qt}$，对 Q 进行谱分解(即对角化)有

$$Q = \sum_{\alpha=1}^{4} \lambda_\alpha u_\alpha v'_\alpha$$

其中 $\lambda_1 = 0, \lambda_2, \lambda_3, \lambda_4 < 0$ 为 Q 的特征根，而 $u_\alpha = (u_{1\alpha} u_{2\alpha} u_{3\alpha} u_{4\alpha})'$ 和 $v_\alpha = (v_{\alpha1} v_{\alpha2} v_{\alpha3} v_{\alpha4})'$ 为 λ_α 所对应的右和左特征向量，并且 $v_\alpha u_\beta = \delta_{\alpha\beta}$，这样有

$$P(t) = \sum_{\alpha=1}^{4} e^{t\lambda_\alpha} u_\alpha v'_\alpha \tag{2}$$

或

$$p_{ij}(t) = \sum_{\alpha=1}^{4} e^{t\lambda_\alpha} u_{i\alpha} v_{\alpha j}$$

我们可以观察到的是 x_{ij}，即一序列中为 i，另一序列中为 j 的位点的比例：

$$x_{ij}(t) = \sum_{k=1}^{4} \pi_k p_{ki}(t) p_{kj}(t) \tag{3}$$

其中 π_k 为祖先序列中的核苷酸频率，通常假定处于平衡并以现存序列中估计。两序列分化后每位点核苷酸替代数为

$$K = 2t \sum_{i=1}^{4} \pi_i q_i \tag{4}$$

上面诸式建立了观察频率 x_{ij} 与待估计距离 K 之间的关系。为了估计 K，通常还需对 Q 的结构做某些假定。Lanave 等[10]曾尝试不用这种假定。按本文记号，他们对特征向量重新进行了标记。

$$z_{i\alpha} = \pi_i^{\frac{1}{2}} u_{i\alpha}, \quad w_{\alpha i} = \pi_i^{-\frac{1}{2}} v_{\alpha i}$$

w_α 与 $z_\alpha(\alpha = 1、2、3、4)$ 也还可做为一套特征向量，因为 $w'_\alpha z_\beta = \sum_i w_{\alpha i} z_{i\beta} = \sum_i v_{\alpha i} u_{i\beta} = \delta_{\alpha\beta}$。但是 Lanave 等理所当然地认为

$$z_\alpha = w_\alpha$$

或

$$u_{i\alpha}\pi_i = v_{\alpha i} \tag{5}$$

这时对(3)式进行简化可得替代数 K 与 x_{ij} 间较简单的关系式。问题是对于什么样的 Q 阵(5)式才能得到满足？容易看出对于对称的 Q 阵(如 1-p、2-p 模型)(5)式是成立的。对于一般的 Q 阵我们将证明(5)式成立的充分必要条件是马氏过程可逆，即 Q 阵满足

$$\pi_i q_{ij} = \pi_j q_{ji} \tag{6}$$

根据特征向量的定义，有 $\sum_j q_{ij} u_{j\alpha} = \lambda_\alpha v_{i\alpha}$，$\sum_j v_{\alpha j} q_{ji} = \lambda_\alpha v_{\alpha i}$ 对所有 i 成立。首先，利用式(6)，$\sum_j \pi_i q_{ij} u_{j\alpha} = \lambda_\alpha \pi_i u_{i\alpha}$，即 $\sum_j \pi_j q_{ji} u_{j\alpha} = \lambda_\alpha \pi_i u_{i\alpha}$，也就是说 $v_{\alpha i} = \pi_j u_{j\alpha}$ 构成的向量 v_α 为 Q 的左特征向量，即当 Q 阵满足(6)式时(5)式成立。

另一方面，如(5)式成立，则 $v_{\alpha i} = \pi_i u_{i\alpha}$ 构成左特征向量。$\sum_j \pi_j u_{j\alpha} q_{ji} = \lambda_\alpha \pi_i u_{i\alpha} = \sum_j \pi_i q_{ij} u_{j\alpha}$，从而有 $\sum_j (\pi_j q_{ji} - \pi_i q_{ij}) u_{j\alpha} = 0$ 对任何 i 成立。

由于矩阵 $U = (u_1 u_2 u_3 u_4)$ 是可逆阵，所以有 $\pi_i q_{ij} = \pi_j q_{ji}$，即 Q 阵为可逆马氏

过程的速率阵。

可逆过程在 DNA 序列进化的研究中具有重要意义，此时无论两个序列中何者为祖先、或者两个都为某共同祖先的后代，其概率行为相同。Tavare[15] 曾对可逆过程模型下进化距离的估计做过研究。此处值得指出的是 Lanave 等[10]的方法由于忽略了其内在的假定，计算上存在着自相矛盾的地方，如[10]中表 6 所估计的速率阵不满足行和为零的数学要求。

2　替代速率的空间变异

通常假定序列中各位点上核苷酸替代的速率是相同的。不过这种假设看来是不现实的[7]。不同的位点可能会由于它们编码蛋白质的不同区域而受到不同的选择约束，而且不同位点上的突变率也可能不同[3]。本节中我们考察存在位点间速率变异时传统距离估计的误差。

2.1　空间替代速度的变异

我们对替代的模式和速度加以区别：模式是指速率阵 $\boldsymbol{Q}$ 中元素的相对大小，而速度则指其绝对大小。例如两个速率阵 $\boldsymbol{Q}$ 与 $c\boldsymbol{Q}$ 具有不同的速度和相同的模式。对于 2-p 替代方案，模式则指转换/颠换替代率之比 α/β。此处，我们考虑位点间替代模式相同 $(\alpha/\beta = c)$、而速度 $r = \alpha + 2\beta$ 不同的情形。

用一连续分布 $f(r)(r > 0)$ 来近似 r 在位点间的变异。由于替代速率为 r 的位点在时间 t 内期望发生的替代数为 $r \cdot 2t$，则易见就位点平均后的真实替代数还是时间 t 的线性函数

$$\bar{K}\int_0^\infty r \cdot 2t \cdot f(r)dr = 2t \cdot \int_0^\infty (\alpha + 2\beta)f(r)dr \tag{7}$$

而我们观察到的转换型差异和颠换型差异比例分别为

$$P = \frac{1}{4} + \frac{1}{4}\int_0^\infty f(r)e^{-4rt/(c+2)}dr - \frac{1}{2}\int_0^\infty f(r)e^{-2rt/(c+1)\cdot(c+2)}dr$$

$$Q = \frac{1}{2} - \frac{1}{2}\int_0^\infty f(r)e^{-4rt/(c+2)}\alpha r$$

从而估计的平均每位点替代数为(据[9])

$$\hat{K} = -\frac{1}{4}\log(1 - 2Q) - \frac{1}{2}\log(1 - 2P - Q) \tag{8}$$

当 r 服从均匀分布或指数分布时由上面诸式计算的 $\bar{K}$ 和 $\hat{K}$ 列入表 1。显然 $\hat{K}$ 严重低估 $\bar{K}$，并且偏差随真值增大而很快变得不可接受。这一结论跟 Palumbi[13] 相一致，他考察了只有速率为零的和速率恒定的两种类型位点的情况。

对于一般的替代方案，设 $r_i\boldsymbol{Q}$ 为第 i 位点上的速率矩阵。注意到 $r_i\boldsymbol{Q}$ 的特征根为 $r_i\lambda$，而特征向量与 $\boldsymbol{Q}$ 的相同，观察到的矩阵 $\boldsymbol{X}$ 现在是对各位点的平均值。我们用 Monte Carlo 方法对 4-p 和 6-p 替代方案进行了计算。所用 $\boldsymbol{Q}$ 阵的参数取自文献[6]：对 4-p 模型为 $\alpha = 0.00375$，$\beta = 0.015$，$\gamma = 0.001$，$\boldsymbol{Q} = 0.5$；对 6-p 模型为 $\alpha = 0.00125$，$\beta = 0.005$，$\alpha_1 = 0.008$，$\alpha_2 = 0.0118$，$\beta_1 = 0.004$，$\beta_2 = 0.0059$。我们考虑了 500 个位点。对每个位点，从一均值为 $\bar{K}$ 的分布中抽取 r_i 来构造该位点的速率阵 $r_i\boldsymbol{Q}$，然后据式(2)、(3)得到该位点的 $\boldsymbol{X}$ 阵，平均后即为观察到的 $\boldsymbol{X}$ 阵，运用各种估计方

法可得到估计值 $\hat{K}$。结果跟表 1 相一致所以不再列出。我们的一般性结论为：存在速率空间变异时，常规的估计方法都给出过低的估值，并且实际的替代模式越复杂，这种差异也越大。指数分布下比均匀分布下偏差也更大，据推测前者更符合生物学实际。

表 1 位点间替代模式相同而替代速度服从均匀或指数分布时估计的每位点核苷酸替代数($\hat{K}$)
使用 2-p 替代方案，$\alpha/\beta = c$

Table 1 Estimated numbers of substitutions per site ($\hat{K}$) when the substitution speed over sites obey a uniform or exponential distribution with mean $\bar{K}$ and the substitution pattern over sites is the same ($\alpha/\beta = c$)

C	$\hat{K}$						
	0.01	0.1	0.5	0.75	1.0	1.5	2.0
	均匀分布 Rate uniform						
0.5	0.010	0.098	0.444	0.627	0.786	1.049	1.253
1.0	0.010	0.098	0.445	0.629	0.790	1.054	1.259
2.0	0.010	0.098	0.444	0.626	0.784	1.045	1.248
5.0	0.010	0.097	0.437	0.612	0.762	1.006	1.197
10.0	0.010	0.097	0.431	0.599	0.741	0.969	1.145
	指数分布 Rate ~~uniform~~ exponential						
0.5	0.011	0.103	0.411	0.555	0.675	0.869	1.024
1.0	0.010	0.094	0.383	0.520	0.636	0.824	0.975
2.0	0.009	0.083	0.343	0.468	0.576	0.753	0.896
5.0	0.007	0.067	0.280	0.385	0.476	0.628	0.753
10.0	0.006	0.058	0.240	0.329	0.404	0.533	0.641

2.2 空间替代模式的变异

现在我们假定各位点上替代的速度相同，而转换/颠换速率比不同。我们考虑两种类

表 2 位点间替代速度恒定而替代模式不同时估计的每位点核苷酸替代数 $\hat{K}$，假定两类位点各占一半，第一类转换/颠换比 $\alpha/\beta = c_1 = 1$，另一类上 $\alpha/\beta = c_2$

Table 2 Estimated numbers of substitutions per site ($\hat{K}$) when the substitution speed is constant but the transition-transversion *ratio* varies over sites. Two classes of sites with equal frequency are assumed, the one with $\alpha/\beta = c_1 = 1$ and the other $\alpha/\beta = c_2$

$\bar{K}$	c_2			
	0.5	0.2 2	4.0	8.0
0.1000	0.1000	0.0999	0.0998	0.0996
0.2000	0.1999	0.1998	0.1992	0.1984
0.3000	0.2997	0.2995	0.2981	0.2963
0.4000	0.3995	0.3992	0.3967	0.3935
0.5000	0.4992	0.4987	0.4948	0.4899
0.6000	0.5988	0.5981	0.5925	0.5854
0.7000	0.6984	0.6975	0.6899	0.6803
0.8000	0.7979	0.7967	0.7868	0.7743
0.9000	0.8973	0.8958	0.8833	0.8676
1.0000	0.9967	0.9948	0.9794	0.9602

型的位点(各占一半)：第一类 $\alpha/\beta = c_1 = 1.0$，第二类中 $\alpha/\beta = c_2$。仿照与前述相似的方法用两类位点上差异比例 P、Q 的平均值来计算替代数 $\hat{K}$，并与真值相比较。当然如 $c_2 = 1 = c_1$，则 $\hat{K} = \bar{K}$，结果列于表 2。

显然 $\hat{K}$ 低估 $\bar{K}$。鉴于这些结果只跟估计公式所用函数的凹凸性质有关，所以将此结论推广到一般情形应该没有什么危险：位点间替代模式有差异时常规估计公式给出偏低的值，但偏差不大。实际序列中替代模式在位点间的可能差异难以分析，所以有了此处的结论我们可以满意了。

3　替代速率的时间变异

有证据表明不同物种中替代的速率可能是不同的[4,11]。例如 Li 和 Tanimura[11] 估计啮齿类中进化的速率要高于灵长类中 4—10 倍，并认为是由于世代间隔的差异。而更重要的速率的差异也可能会因为同源基因在不同物种中位于不同的基因区域、从而受不同的选择约束和突变压力而造成[2,3]。

3.1　物种间的速率变异

我们仅限于分析速度的变异。假定两序列在分化后以不同但恒定的速率进化，速率矩阵分别为 $\boldsymbol{Q}$ 和 $\boldsymbol{Q}^* = c\boldsymbol{Q}$。它们的特征根 λ 和 λ^* 间有关系 $\lambda = \lambda^*/c$，而其特征向量是相同的。因此代替式(3)我们有

$$x_{ij} = \sum_{k=1}^{4} \pi_k \cdot p_{ki}(t) \cdot p_{kj}^*(t) = \sum_{k=1}^{4} \pi_k \sum_{\alpha=1}^{4} \sum_{\beta=1}^{4} e^{t(\lambda_\alpha + c\lambda_\beta)} u_{k\alpha} v_{\alpha i} u_{k\beta} v_{\beta j}$$

在 2-p 方案下，可以取 $\boldsymbol{u}_\alpha = \boldsymbol{v}_\alpha$，所以

$$x_{ij} = \sum_{\alpha=1}^{4} e^{\lambda_\alpha(1+c)t} u_{i\alpha} v_{\alpha j}$$

从而可以看出用 2-p 方法所估计的距离为

$$K = (\alpha + 2\beta)(1 + c)t \tag{9}$$

如果有多个物种的序列就有可能估计速率常数 c。

3.2　物种内速率变异

设速率矩阵为 $\boldsymbol{Q}(t) = f(t)\boldsymbol{Q}$，即替代过程为非时齐马氏过程。以 $\boldsymbol{P}(s,t) = \{p_{ij}(s,t)\}$ 表示从时间 s 到时间 t 的转概阵，则后退 Kolmogorov 方程为[1]

$$\frac{\partial \boldsymbol{P}(s,t)}{\partial s} = -\boldsymbol{Q}(s)\boldsymbol{P}(s,t),\ \boldsymbol{P}(t,t) = \boldsymbol{I}$$

因为 $\boldsymbol{Q}(t)$ 满足下列可交换条件，所以该微分方程可以解出：

$$\boldsymbol{Q}(t)\left[\int_{t0}^{t} \boldsymbol{Q}(\tau)d\tau\right] = \left[\int_{t0}^{t} \boldsymbol{Q}(\tau)d\tau\right]\boldsymbol{Q}(t)$$

其解为

$$\boldsymbol{P}(s,t) = \exp\left\{\int_s^t [-\boldsymbol{Q}f(\tau)]d\tau\right\} = \exp\left\{\boldsymbol{Q}\int_s^t f(\tau)d\tau\right\}$$

或

$$\boldsymbol{P}(t) = \boldsymbol{P}(0,t) = \sum_{\alpha=1}^{4} \exp\left\{\lambda_\alpha \int_0^t f(\tau)d\tau\right\} \boldsymbol{u}_\alpha \boldsymbol{v}_\alpha'$$

其中 λ_α、$\boldsymbol{u}_\alpha$、$\boldsymbol{v}_\alpha$ 跟以前一样是 $\boldsymbol{Q}$ 的特征对。

以 $f_1(t)$ 和 $f_2(t)$ 分别表示序列 1 和 2 的"速率因子函数"。在 2-p 模型下，容易证明式(8)之 $\hat{K}$ 所估计的距离是

$$K = (\alpha + 2\beta)\int_0^t [f_1(\tau) + f_2(\tau)]d\tau = (\alpha + 2\beta)t[\bar{f}_1(t) + \bar{f}_2(t)] \tag{10}$$

其中 $\bar{f}_1(t)$ 和 $\bar{f}_2(t)$ 是两序列的替代速率在时期 $(0,t)$ 内的平均。当 $f_1(t) = f_2(t) = 1$ 时上式右侧简化为 $2t(\alpha + 2\beta)$；而当 $f_1(t) = 1$，$f_2(t) = c$ 时简化为 $(\alpha + 2\beta)(1 + c)$，正如所期望的那样。

值得注意的是 K 反映的是替代速率在时间上的平均。看来要区分恒定的替代速率和平均的可变的速率是困难的。只有两个序列时显然设法做到这一点。其次，如果速率函数不是常数的话，估计的距离就不再是时间的线性函数。

4 讨论

本文对替代速率(速度与模式)在位点间有变异时序列间进化距离估计的误差做了研究。看来位点间模式的变异不是大问题，而位点间速度的差异则会造成估计值很大的偏差。Jin 和 Nei[12] 最近提出了一个假定替代速率在位点间服从 Γ 分布时的距离估计，此估计值依赖于 Γ 分布的参数 α。我们对实际序列资料进行了分析，对参数 α 进行了估计，结果将另文发表。

本文中未讨论替代模式在物种间的差异，初步研究表明存在这种差异时传统估计方法高估真值，但偏差不大。总之，最严重的问题是位点间替代速率(度)的变异，当然位点间替代过程的非独立性、核苷酸频率的变化乃至选择的作用等都需要研究。

在以序列为资料推断进化树的方法中，Felsenstein[5] 的最大似然法由于有着坚实的统计学基础、且能给出树的可靠性评价而成为目前最好的方法。不过 Felsenstein[5] 假定替代速率在位点间是恒定的。那么在位点间速率有变异的情况下，似然法的稳定性如何，以及如何在似然法中考虑这种速率的变异，显然具有重要意义，对此我们正在进行研究，并已经取得了一些成果。

参 考 文 献

1 胡迪鹤，1983。可数状态马尔可夫过程论，武汉：武汉大学出版社
2 杨子恒，1990。遗传学报，17(4)：354—359
3 Bernardi *et al*, 1988. J. Mol. Evol., 28: 7—18
4 Britten R J, 1986. Science, 231: 1393—1398
5 Felsenstein J, 1981. J. Mol. Evol., 17: 368—376
6 Gojobori T, Ishii K, and M Nei, 1982. J. Mol. Evol., 18: 414—423
7 Holmquist R, Goodman M, Conry T, and J Czelusniak, 1983. J. Mol. Evol., 19: 137—448
8 Jukes T H, and C R Cantor, 1969. In: Mammalian protein metabolism (ed. H. N. Munro), New York: Academic press, 21—123
9 Kimura M, 1980. J. Mol. Evol., 16: 111—120
10 Lanave C, preparata G, Saccone C, and G Serio, 1984. J. Mol. Evol., 20: 86—93
11 LiW—H, and M Tanimura, 1987. Nature, 326: 93—96
12 Nei M, and L Jin, 1990. Mol. Biol. Evol., 7: 82—102
13 Palumbi S R, 1989. J. Mol. Evol., 29: 180—187
14 Takahata N, and M Kimura, 1981. Genetics, 98: 641—657
15 Tavare S, 1986. In: Lectures in Mathematics in the life Sciences, Vol. 17, 57—86

Evaluation and Extension of Markov Process Models for the Evolution of DNA①

Yang Ziheng
(*Department of Animal Sciences, Beijing Agricultural University, Beijing* 100094)

Nick Goldman
(*National Institute for Medical Research London NW7 1AA, UK*)

Abstract

Markov process models of nucleotide substitution are evaluated. A model proposed by Lanave *et al* (1984), alleged to need no priori assumption about the substitution pattern, is found to have the assumption of reversibility. Calculations based on the 2-p, 4-p, and 6-p substitution schemes show that site variation of substitution speed leads to serious under-estimation of sequence divergence by various methods. Spatial pattern variation also leads to under-estimation, but the discrepancy is slight. A nonhomogeneous Markov process model is used to study the temporal variations of rates and it is shown that the estimated number of substitutions reflects a rate averaged over time. The implications of those results to evolutionary phylogenetics are discussed.

Key words Nucletide substitution, Markov process models, Molecular evolution

① Received April 28, 1992

Evaluation and Extension of Markov Process Models for the Evolution of DNA

Zi-Heng Yang

(Department of Animal Sciences, Beijing Agricultural University)

ABSTRACT

Stochastic models for studying the nucleotide substitution process during the evolution of DNA are evaluated. A model proposed by Lanave *et al.* (1984), alleged to need no *a priori* assumption about the substitution pattern, is found to have very stringent restrictions on the pattern. Under the 2-p, 4-p, and 6-p substitution schemes it is shown that when the substitution rates differ over sites the estimated number of nucleotide substitutions seriously underestimates the real rates, more seriously when the number of substitutions is greater. A nonhomogeneous Markov process model is proposed to study the temporal variations of rates and it is shown that the estimated number of substitutions reflects a rate averaged over time. The implications of those results to evolutionary phylogeny are discussed.

1. INTRODUCTION

The availability of great amount of nucleic acids data has made it possible to study the evolutionary process at the DNA level directly. Estimated distances among homologous sequences have been used to date major evolutionary events and infer phylogeny of species. In doing this the nucleotide substitution process has been regarded as a homogeneous Markov process. Then the distance, *i.e.*, the number of substitutions per site after divergence, will be linearly related to time. However, certain assumptions underlie those models[15,9,29,33]: *i)*. The substitution rates are constant over within lineages. *ii)*. The nucleotides at different sites, or codon sites for protein-coding genes, are changing independently and with the same rates (spatial uniformity). *iii)*. No selection exists except for that which only affects substitution rates. Recently there have been accumulated evidences that these assumptions are not realistic[3,30,32,33]. On the other hand, what effects the violation of these assumptions will have remains largely untackled.

?? ok – means actual rate than visible.

Under the aforementioned Markov process models the substitution rate matrix is a 4x4 matrix with 12 free parameters. More assumptions are needed about the structure of the matrix to reduce the unknowns, as in the 1-p (one-parameter)[14], 2-p[16], 4-p[28], and 6-p models[10,17]. Lanave *et al.*[19] proposed a new method which was claimed to have needed no *a priori* assumption about the structure of the rate matrix, but in fact very stringent conditions underlie the model.

1-p = Jukes/Cantor
2-p = Kimura
4-p = Takahata & Kimura
6-p = GTN; Kimura

In this paper Lanave *et al.*'s model will first be evaluated. Then the effects of rate variations over sites and over time will be studied.

2. STRUCTURE OF RATE MATRIX Q: EVALUATION OF LANAVE *ET AL.*'S MODEL

Suppose the nucleotide substitution process at any site of the sequence to be a homogeneous Markov process. Let the transition matrix be $\mathbf{P}(t)=\{p_{ij}(t)\}$, where $p_{ij}(t)$ is the probability that the nucleotide at a site is base j given that it was base i at time zero and i,j=1,2,3,4 corresponding to bases T,C,A,G, respectively. Then

$$\mathbf{Q}=\mathbf{P}'(0)=\left.\frac{d\mathbf{P}(t)}{dt}\right|_{t=0} = \lim_{t\to 0}\frac{1}{t}(\mathbf{P}(t)-\mathbf{I}) \qquad (1)$$

is known as the rate matrix, since $q_{ij}\delta t$ $(i=j)$ represents the probability that base i 'mutates' to base j in time δt, and $q_{ii}=-\Sigma_{j\neq i}q_{ij}=-q_i$. $\mathbf{Q}$ and $\mathbf{P}(t)$ are mutually and uniquely determined by Eq.(1) and the following relationship

$$\mathbf{P}(t)=e^{t\mathbf{Q}}.$$

The spectral decomposition of $\mathbf{Q}$ leads to

$$\mathbf{Q}=\Sigma_\alpha \lambda_\alpha \mathbf{u}_\alpha \mathbf{v}'_\alpha,$$

where $\lambda_1=0$, λ_2, λ_3, $\lambda_4<0$ are eigenvalues of $\mathbf{Q}$, and $\mathbf{u}_\alpha=(u_{1\alpha}\ u_{2\alpha}\ u_{3\alpha}\ u_{4\alpha})'$ and $\mathbf{v}_\alpha=(v_{\alpha1}\ v_{\alpha2}\ v_{\alpha3}\ v_{\alpha4})'$ the right and left eigenvectors related to λ_α respectively, with $\mathbf{v}'_\alpha\mathbf{u}_\alpha=\delta_{\alpha\beta}$. Furthermore,

$$\mathbf{P}(t)=\Sigma_\alpha e^{t\lambda_\alpha}\mathbf{u}_\alpha\mathbf{v}'_\alpha \qquad (2)$$

or $$p_{ij}(t)=\Sigma_\alpha e^{t\lambda_\alpha}u_{i\alpha}v_{\alpha j}.$$

The observable quantity x_{ij}, which is the probability of finding base i in one sequence and base j in another at a given site, is given by

$$x_{ij}(t)=\Sigma_k \pi_k p_{ki}(t)p_{kj}(t) \qquad (3)$$

where π_k is the nucleotide frequency in the ancestor sequence, most often assumed to be in equilibrium and estimated from the extant sequences. The number of substitutions per site since divergence is

$$K=2t\Sigma_i \pi_i q_i. \qquad (4)$$

It seems necessary to make certain assumptions as to the structure of matrix $\mathbf{Q}$ to simplify Eq.(3) so as to estimate $\mathbf{Q}$ and/or the parameter K. Lanave *et al.*[19], however, attempted to proceed without them. In the present notation, they rescaled the eigenvectors as

$$z_{i\alpha}=\pi_i^{\frac{1}{2}}u_{i\alpha}$$

$$w_{\alpha i}=\pi_i^{-\frac{1}{2}}v_{\alpha i}.$$

$\mathbf{w}_\alpha$ and $\mathbf{z}_\alpha$ $(\alpha=1,2,3,4)$ could still be a set of eigenvectors since $\mathbf{w}'_\alpha\mathbf{z}_\beta=\Sigma_i w_{\alpha i}z_{i\beta}=\Sigma_i v_{\alpha i}u_{i\beta}=\delta_{\alpha\beta}$. What is striking is that Lanave *et al.* took for granted that

YES! V. TRUE!

$$\mathbf{z}_\alpha=\mathbf{w}_\alpha \quad \text{or} \quad u_{i\alpha}\pi_i=v_{\alpha i} \qquad (5)$$

Then simplification of Eq.(3) leads to the following simple equation (Eq.(18) in [19]):

$$\{x_{ij}(t)/(\pi_i\pi_j)\}=\Sigma_\alpha e^{2t\lambda_\alpha}\mathbf{w}_\alpha\mathbf{w}_\alpha'$$

The question is whether such $\mathbf{u}_\alpha$ and $\mathbf{v}_\alpha$ that satisfy Eq.(5) can be found. Consider the case where the rate matrix $\mathbf{Q}$ has distinct eigenvalues, which is probably true for a general $\mathbf{Q}$. Then any orthogonal eigenvectors related to λ_α can be expressed as $c_\alpha\mathbf{u}_\alpha$ and $c_\alpha^{-1}\mathbf{v}_\alpha$, where c_α is a nonzero constant. To find the scale constant c_α the following relationship must be satisfied:

$$v_{\alpha i}/(\pi_i u_{i\alpha})=c_\alpha^2=\text{Constant irrespective of i.} \qquad (6)$$

Eq.(6) turns out to be a very stringent condition, as it does not hold for the 6-p model with the eigenvalues and eigenvectors from [10]. It does hold, however, when $\mathbf{Q}$ is symmetrical because in such a case $\pi_i=1/4$ for all i and eigenvectors which satisfy $\mathbf{u}_\alpha=\mathbf{v}_\alpha$ can be found. What property a nonsymmetrical $\mathbf{Q}$ should possess to make Eq.(6) hold is unclear.

could it possibly be reversibility (only a wild guess)?

The inconsistency of Lanave *et al*'s model can also be seen from the fact that the rate matrix obtained in their paper does not satisfy $\mathbf{Q1=0}$ (Table 6 in [19]).

3. SPATIAL VARIATIONS OF MUTATION RATES

To estimate x_{ij} (Eq.(3)) by counting nucleotide differences in extant homologous sequences, it is commmonly asssumed that each site in the sequence has an equal mutation rate. However, this assumption is seldom realistic[5,12]. Different sites might be under different selective restrictions because of the requirements of protein folding, polarity and because the amino acids coded by them lie in different domains of the protein. Before discussing the effect of spatial variation of rates, we first discuss the measurement of spatial variation.

Measure of The Spatial Variation of Rates

Unfortunately no appropriate measure of the variability of rates over sites is available, because such a measure would require pre-estimation of the numbers of substitutions at every site among many homologous sequences. Thus knowledge of the phylogeny is necessary. For instance, a homologous site with base frequencies A_4G_3 and $A_4G_1C_1U_1$ in seven sequences could be considered the same or not in the extent of variation according to different phylogenies. For a 'star' phylogeny, where all the extant sequences diverged from an ancestor sequence at about the same time, the proportion of sequences that differ from the most common nucleotide (consensus base, assumed to be the base in the ancestor sequence) would be a reasonable approximation. If multiple 'hits' and different rates of transitions and transversions are considered, the following estimator can be used after the 2-p model[16]:

$$r_i = -\tfrac{1}{4}\ln(1-2Q_i)-\tfrac{1}{2}\ln(1-2P_i-Q_i) \qquad (7)$$

where P_i and Q_i are proportions of sequences that have a transition- and transversion-type difference at site i from the consensus base respectively. Thus P=3/7, Q=0 for A_4G_3 while P=1/7 and Q=2/7 for $A_4G_1C_1U_1$. Then the variance of r_i can be used to measure the rate variation. Eq.(7) is at least consistent in theory in that r_i approaches the true value as the number of sequences increases. An entropy-like measure, called RNV, proposed by Manske & Chapman[23], is not linearly related to the number of mutations at that site, as seen from their Eq.(2), and is thus not consistent.

The Effect of Spatial Variation of Rates

In this paper the pattern (mode) of substitutions is assumed to be the same, only variations in the speed (tempo) of evolution is considered. By 'pattern' we mean the relative size of the elements of the matrix **Q**. For instance, two substitution processes with rate matrices **Q** and c**Q** have different speeds but the same pattern. For the 2-p mutation scheme this restriction means a fixed transition/ transversion ratio over sites, *i.e.* $\alpha/\beta = c$. Let f_i be the proportion of sites with transition and transversion mutation rates α_i, β_i respectively. Then the mean number of substitutions averaged over sites is still a linear function of time:

variations between sites, but not within lineages for a given site

$$\bar{K}=2t\Sigma_i f_i(\alpha_i+2\beta_i).$$

However this quantity can not be estimated directly. The expected proportions of transitional and transversional differences are now

$$P=\Sigma f_i P_i = \tfrac{1}{4}+\tfrac{1}{4}\Sigma f_i e^{-4\beta t}-\tfrac{1}{2}\Sigma f_i e^{-2(\alpha+\beta)t}$$

$$Q=\Sigma f_i Q_i = \tfrac{1}{2}-\tfrac{1}{2}\Sigma f_i e^{-4\beta t}$$

from which we get the following estimated number of substitutions per site since divergence:

$$\hat{K}=-\tfrac{1}{4}\ln(1-2Q)-\tfrac{1}{2}\ln(1-2P-Q) \qquad (8)$$

When a continuous distribution f(r) of r, where $r=(\alpha+2\beta)t$, of rates is used as an approximation,

$$P=\tfrac{1}{4}+\tfrac{1}{4}\int f(r)e^{-4r/(c+2)}dr-\tfrac{1}{2}\int f(r)e^{-2r(c+1)/(c+2)}dr$$

$$Q=\tfrac{1}{2}-\tfrac{1}{2}\int f(r)e^{-4r/(c+2)}dr$$

The results for uniform and exponential distributions of f(r) are shown in Table 1. It is apparent that K seriously underestimates K, and the discrepancy increases with K or time, quickly becoming unacceptable. The result is consistent with that of Palumbi(1989), who considered only 2 classes of sites with zero and constant rates respectively, whish is apparently a special case of the present analysis.

For a general substitution scheme, let $r_i\mathbf{Q}$ be the rate matrix at the *i*th site. Note that the eigenvalues of $r_i\mathbf{Q}$ are r_i , and the eigenvectors of $r_i\mathbf{Q}$ are the same as those of **Q**. The observed matrix **X** in Eq. (3) is now an average over sites. Under the 4-p and 6-p schemes, Monte Carlo method has been used to calculate the estimated number of substitutions. The parameters used in matrix **Q** are from Gojobori *et al.*'s simulations[10]: α=.00375, β=.015, γ=.001, θ=.5 for the 4-p scheme, and α=.00125, β=.005, α_1=.008, α_2=.0118, β_1=.004, β_2=.0059 for the 6-p scheme. **Q** has been rescaled to make $\Sigma\pi_i q_i=1$ so that **Q** only reflects the 'pattern' of mutation. r_i's are sampled from a certain distribution with mean K to get the rate matrix $r_i\mathbf{Q}$ for every site. Fig.1 shows the results when r_i's are exponentially distributed, with other sets of parameters basically producing the same results. It is also found that the discrepancy is greater when the rate obeys an exponential distribution than when it obeys a uniform distribution (See also Table 1). The exponential distribution of rates implies that there are a few 'hot' sites with great mutation rates while most sites are very conservative with very low mutation rates, biologically more realistic than the uniform distribution.

That's what I do, too!

4. TEMPORAL VARIATION OF MUTATION RATES

Recent evidences show that substitution rates might be different in different lineages[3,22,25,32]. For instance, Li & Tanimura[21] estimated that the rate of evolution in rodents was 4-10 times faster than in primates, 2-4 times faster than in artiodactyls. The variation was attributed to differences in generation time[21]. More importantly the rates might vary within a lineage due to changes in restrictions on the DNA sequence. Such changes could be caused by mutations at other sites or in other related genes; by changes of the function of the protein coded by the gene; by changes of the location of the sequence on the chromosome which would have brought it into a different isochore, and thus under different selective or mutational pressures [1,2,26,31,33].

Rate Variations Along Different Lineages

Suppose two sequences evolve at different but constant rates **Q** and **Q*** upon divergence, and $\mathbf{Q}=c\mathbf{Q}^*$. The eigenvalues of **Q** and **Q*** are related by $\lambda=\lambda^*/c$ and their eigenvectors are the same. Thus instead of Eq.(3) we get

$$x_{ij}=\Sigma_k\pi_k p_{ki}(t)p_{kj}^*(t)$$
$$=\Sigma_k\pi_k\Sigma_\alpha\Sigma_\beta e^{t(\lambda_\alpha+c\lambda_\beta)}u_{k\alpha}v_{\alpha i}u_{k\beta}v_{\beta j}$$

Under the 2-p scheme, we can assume $\mathbf{u}_\alpha=\mathbf{v}_\alpha$, so that

$$x_{ij}(t)=\Sigma_\alpha e^{\lambda_\alpha(1+c)t}u_{i\alpha}v_{\alpha j},$$

and the quantity estimated by Eq.(8), *i.e.*, K in Eq.(4), is

$$K=(\alpha+2\beta)(1+c)t \qquad (9)$$

It would be possible to estimate c or t when sequences from several species are available.

Rate Variation Within a Lineage

Consider the case when the rate matrix is time dependent and the nucleotide substitution process is a nonhomogeneous Markov process. Suppose $\mathbf{Q}(t)=f(t)\mathbf{Q}$, where $f(t)$ is a nonnegative continuous function of time t. Let $\mathbf{P}(s,t)=\{p_{ij}(s,t)\}$ be the transition matrix from time s to time t. Then the backward Kolmogorov equation is [13]

$$\frac{\partial \mathbf{P}(s,t)}{\partial s} = -\mathbf{Q}(s)\mathbf{P}(s,t), \quad \text{with } \mathbf{P}(t,t)=\mathbf{I}.$$

This differential equation can be solved when, as it does here, $\mathbf{Q}(t)$ satisfies the following exchangeability condition[34]:

$$\mathbf{Q}(t)\left[\int_{t_0}^{t} \mathbf{Q}(\tau)d\tau\right] = \left[\int_{t_0}^{t} \mathbf{Q}(\tau)d\tau\right]\mathbf{Q}(t).$$

The solution is

$$\mathbf{P}(s,t) = e^{\left(\int_t^s [-\mathbf{Q}f(\tau)]d\tau\right)} = e^{\left(\mathbf{Q}\int_s^t f(\tau)d\tau\right)}$$

or

$$\mathbf{P}(t) = \mathbf{P}(0,t) = e^{\left(\mathbf{Q}\int_0^t f(\tau)d\tau\right)} = \Sigma_\alpha e^{\left(\lambda_\alpha\int_0^t f(\tau)d\tau\right)}\mathbf{u}_\alpha\mathbf{v}_\alpha',$$

where λ_α, $\mathbf{u}_\alpha$, $\mathbf{v}_\alpha$ (α=1,2,3,4) are eigenvalues and eigenvectors of $\mathbf{Q}$.

Let the 'rate functions' $f(t)$ for sequences 1 and 2 be $f_1(t)$ and $f_2(t)$ respectively. Then under the 2-p scheme, it can be shown that the quantity K in Eq.(8) estimates:

$$K = (\alpha+2\beta)\int_0^t [f_1(\tau)+f_2(\tau)]d\tau = (\alpha+2\beta)t(\bar{f}_1(t)+\bar{f}_2(t)) \quad (10)$$

where $\bar{f}_i(t) = \int_0^t f_i(\tau)d\tau/t$ (i=1,2) is the substitution rate per site in sequence i averaged over time period (0,t). The right hand side of Eq.(10) simplifies to $2t(\alpha+2\beta)$ when $f_1(t)=f_2(t)=1$ or to $(\alpha+2\beta)(1+c)t$ when $f_1(t)=1$, $f_2(t)=c$, as expected. It is interesting to note that K is an average of the substitution rates over time. So it seems difficult to distinguish a constant rate from an averaged variable rate. Formerly Lewontin[20] and later Gillespie[6,7,8] proposed that the observed relatively constant rate of molecular evolution might be an average of variable rates over extended time. They considered selection to be the main cause of rate variation, a view strongly opposed by Kimura[18] and Takahata[27]. Whatever the cause(s) might be (see the review at the beginning of this section), it should be noted that K will not be a linear function of time t when $f_1(t)$ or $f_2(t)$ is not constant.

Li & Tanimura[21] proposed a method of estimating the divergence times between sequences under the assumption of different but constant rates within lineages. It is apparent that the rates could not be constant within lineages when they were different over lineages. So it can be expected that their method still under-compensates the rate differences, since the differences were probably formed gradually after divergence.

5. DISCUSSION

Phylogeny inferring by nucleic acids data depends heavily on the molecular clock hypothesis, *i.e.*, on the hypothesis of a linear relationship between numbers of nucleotide substitutions and time. In this regard, spatial variation of rates seems to be a more serious problem than temporal one because it damages the linearity more severely (Table 1 and Fig.1).

There are many other factors not discussed in this paper that affect the relationship between the estimated distance and time and thus affect the reliability of the inferred phylogeny, as pointed out in the Introduction. Among them is natural selection, whose effect is controversial. However, though the effect of Darwinian selection caused by environmental changes is a moot issue, it is generally agreed that restrictive selection which eliminates deleterious mutations has been at work. This kind of selection not only affects the substitution rates, but might also cause non-independency of substitutions over sites, among other factors such as multiple substitution[11]. For instance, restrictions at the amino acid level would cause associated substitutions at the three codon sites. What effects this association would have remains unknown.

For distantly related sequences, systematic pressures might have been at work in different lineages, as revealed by base compositions at the third codon position and codon usage patterns in homologous sequences[30,33]. Such pressures, whether selective or mutational, would cause variation in both speed and mode of substitutions, and they must have been related to evolutionary changes at the chromosome level[4]. A more complete understanding would need studies at both the molecular and chromosome level.

REFERENCES

[1] Bernardi, G. *et al.*(1985). The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953-958.

[2] Bernardi, *et al.* (1988). Compositional patterns in veterbrate genomes, conservation and change in evolution. *J. Mol. Evol.*, **28**, 7-18.

[3] Britten, R.J. (1986). Rates of DNA sequence evolution differ between taxonomic groups. *Science*, **231**, 1393-1398.

[4] Dover, G.A. (1987). DNA turnover and the molecular clock. *J. Mol. Evol.*, **26**, 47-58.

[5] Fitch, W.M. & Markowitz, E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.*, **4**, 579-593.

[6] Gillespie, J.H. (1984a). The molecular clock may be a episodic clock. *Proc. Natl. Acad. Sci. USA*, **81**, 8009-8013.

[7] Gillespie, J.H. (1984b). Molecular evolution over the molecular landscape. *Evolution*, **38**, 1116-1129.

[8] Gillespie, J.H. (1986a). Variability of evolutionary rates of DNA. *Genetics*, **113**, 1077-1091.

[9] Gillespie, J.H. (1986b). Rates of molecular evolution. *Ann. Rew. Ecol. Syst.*, **17**, 637-65.

[10] Gojobori, T., Ishii, K. & Nei, M. (1982). Estimation of average number of nucleotide substitutions when the rate of susbstitution varies with nucleotides. *J. Mol. Evol.*, **18**, 414-423.

[11] Golding, G.B. (1987). Multiple substitutions create biased estimates of divergence times and small increases in the variance to mean ratio. *Heredity*, **58**, 331-339.

[12] Holmquist, R., Goodman, M., Conry, T. & Czelusniak, J. (1983). The spatial distribution of fixed mutations within genes coding for proteins. *J. Mol. Evol.*, **19**, 137-448.

[13] Hu, D.H. (1983). *Denumerable Markov Processes*. Wuhan University Press, Wuhan. pp. 250-256.

[14] Jukes, T.H. & Cantor, C.R. (1969). Evolution of protein molecules. In: *Mammalian Protein Metabolism* (ed. H.N. Munro), Academic Press, New York. pp. 21-123.

[15] Kaplan, N. (1983). Statistical analysis of restriction enzyme map data and nucleotide sequence data. In: *Statistical Analysis of DNA Sequence Data* (ed. B.S. Weir), Marcel Dekker Inc., New York, pp. 75-106.

[16] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111-120.

[17] Kimura, M. (1981). Estimation of evolutionary distances between

homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA*, **78**,454-458.

[18] Kimura, M. (1983). *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

[19] Lanave, C., Preparata, G., Saccone, C. & Serio, G. (1984). A new method for calculating evolutionary substitution rates. *J. Mol. Evol.*, **20**, 86-93.

[20] Lewontin, R. (1974). *The genetic basis of evolutionary change.* Columbia University Press, New York.

[21] Li, W.-H. & Tanimura, M. (1987). The molecular clock runs more slowly in man than in apes and monkeys. *Nature*, **326**, 93-96.

[22] Li, W.-H., Tanimura, M. & Sharp, P.M. (1987). An evaluation of the molecular clock hypothesis using mammallian DNA sequences. *J. Mol. Evol.*, **25**, 330-342.

[23] Manske, C.L. & Chapman, D.J. (1987). Nonuniformity of nucleotide substitution rates in molecular evolution, Computer simulation and analysis of 5S ribosomal RNA sequences. *J. Mol. Evol.*, **26**, 226-251.

[24] Palumbi, S.R. (1989). Rates of molecular evolution and the function of nucleotide positions free to vary. *J. Mol. Evol.*, **29**,180-187.

[25] Sharp, P.M. & Li, W.-H. (1989). On the rate of DNA sequence evolution in Drosophila. *J. Mol. Evol.*, **28**, 398-402.

[26] Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA*, 85, 2653-2657.

[27] Takahata, N. (1987). On the overdispersed molecular clock. *Genetics*, **98**, 641-657.

[28] Takahata, N. & Kimura, M. (1981). A model of evolutionary base substitutions and its application with special referrence to rapid change of pseudogenes. *Genetics*, **98**, 641-657.

[29] Tavare,S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences. In: *Lectures in Mathematics in the life Sciences*, Vol.17, pp. 57-86.

[30] Wells, D. Bains, W. & Kedes, L. (1986). Codon usage in histone gene families of high eukaryotes reflects functional rather than phylogenetic relationships. *J. Mol. Evol.*, **23**, 224-241.

[31] Wolfe, K.H. Sharp, P.M. & Li, W.-H. (1989). Mutation rates differ among regions of the mammalian genome. *Nature*, **337**, 283-285.

[32] Wu, C.-I. & Li. W.-H. (1985). Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci. USA*, **82**, 1741-1745.

[33] Yang, Z.-H. (1989). *Genetica Sinica*, In press.

[34] Yu, E.X. (1988). *Matrix Theory in Science and Technology*. Central China Technical University Press, Wuhan, pp. 144-148.

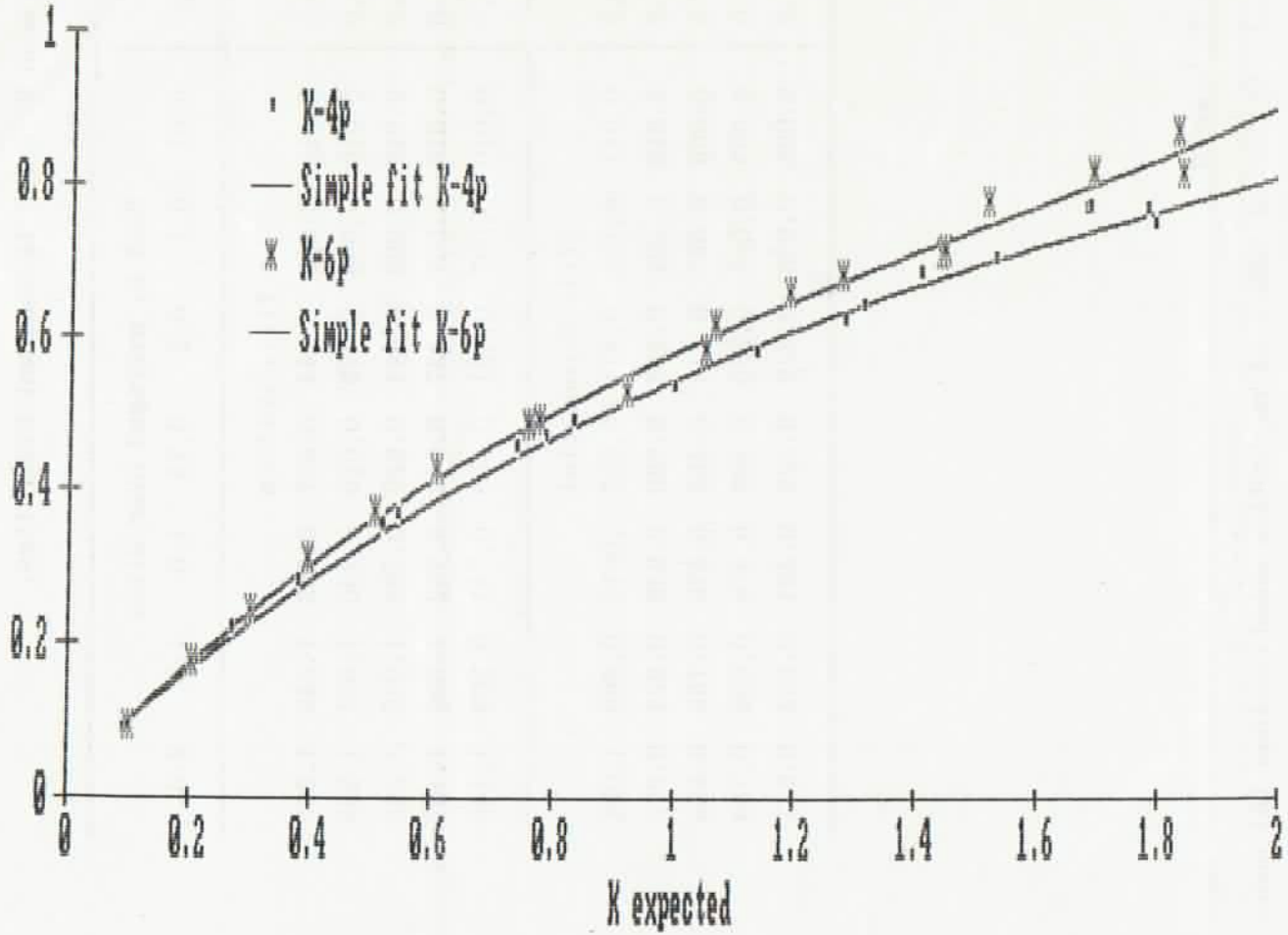
Table 1 Estimated numbers of substitutions per site (K) under the 2-p scheme when the rates over sites obey a uniform or exponential distribution with mean K. $c=\alpha/\beta$ is constant over sites.

c	Real K, averaged over sites						
	0.01	0.1	0.5	0.75	1.0	1.5	2.0
	f(r), Uniform						
0.5	0.010	0.098	0.444	0.627	0.786	1.049	1.253
1.0	0.010	0.098	0.445	0.629	0.790	1.054	1.259
2.0	0.010	0.098	0.444	0.626	0.784	1.045	1.248
5.0	0.010	0.097	0.437	0.612	0.762	1.006	1.197
10.0	0.010	0.097	0.431	0.599	0.741	0.969	1.145
	f(r), Exponential						
0.5	0.011	0.103	0.411	0.555	.0675	0.869	1.024
1.0	0.010	0.094	0.383	0.520	0.636	0.824	0.975
2.0	0.009	0.083	0.343	0.468	0.576	0.753	0.896
5.0	0.007	0.067	0.280	0.385	0.476	0.628	0.753
10.0	0.006	0.058	0.240	0.329	0.404	0.533	0.641

Figure Legend

Fig. 1 Estimated number of nucleotide substitutions (K) under the 4-p and 6-p mutation schemes when mutation rates are exponentially distributed over sites. The expected K is an average of 500 such sampled rates.

Running Headline: On Markov Process Models of DNA Evolution


K-4p
Simple fit K-4p
K-6p
Simple fit K-6p
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
K expected