PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2011, 38(8): 768~776
www.pibb.ac.cn

一种新的 DNA 序列进化距离及其应用 *

梁丽萍[1)] 解小莉[1)] 程 健[1, 2)] 王振凤[1)] 郭满才[1)] 袁志发[1)**]
([1)] 西北农林科技大学理学院，杨凌 712100；[2)] 西北农林科技大学生物信息学研究中心，杨凌 712100)

摘要 为了研究核苷酸变异，通过 DNA 序列的同源率，建立了 DNA 序列进化的动力学方程，进而得到了一种新的物种间进化距离 d_y(选择进化距离). 由于核苷酸替代模型有很多，选用其中的 4 种模型，计算出其相应的选择进化距离 d_y，该进化距离包含了 4 种模型下的 p 距离、替代率为常数的距离 d 和替代率服从 Γ 分布的 Γ 距离 d_G. 进一步根据动力学方程的特点，将模型转化为一元线性回归问题，用最小二乘法求得选择模型中的动力学参数 b 和各核苷酸位点每年的平均替代速率 r. 以 16 个物种的线粒体基因序列为例，说明这种新的进化距离并通过构建不同进化距离下的基因进化树来对各进化距离进行比较. 结果表明：选择进化距离 d_y 是一种有效的构建进化距离的方法.

关键词 选择模型，DNA 序列，进化距离，进化树，核苷酸替代模型
学科分类号 Q332，Q349+.55 **DOI**: 10.3724/SP.J.1206.2010.00605

在分子进化中，DNA 序列中绝大部分核苷酸并不用于蛋白质编码基因，而且并不像氨基酸那样具有编码的兼并性等特点. 另外，DNA 中核酸的突变方式多，有多种多样的 DNA 区域，如编码区、外显子、内含子、重复 DNA 序列等，因而 DNA 序列比蛋白质序列包含着更多的进化信息. DNA 分子中的核苷酸变异也叫核苷酸替代，核苷酸序列中核苷酸替代数的估计是测定基因间进化距离的数学统计方法，是研究基因进化的基础. 我们可以通过对多种生物 DNA 不同区域的核苷酸替代研究，来推测这些生物的共同祖先及它们之间的亲缘关系，进行分子谱系树的构建.

在 DNA 进化中，为了描述同一祖先传衍的两个后裔序列 i 和 j (i、j 分化的时间分别为 t_i 和 t_j，t_i>t_j)的差异，人们通过比对，得到相同、转换和颠换核苷酸对的频率分别为 O_{ij}、P_{ij}和 Q_{ij}，并把 $p_{ij}=1-O_{ij}=P_{ij}+Q_{ij}=\frac{n_d}{n}$ 称为 p 距离[1-2]，其中 n_d 和 n 分别为所检测的两条序列间不同核苷酸数和配对数. 然而 DNA 进化是复杂的，有 GC 含量的偏倚、转换与颠换的偏倚和位点替代率的差异等，即突变并非是完全随机的. 这就是说 p 距离并不能真实地反映两个后裔序列的亲缘关系远近，能真实反映这种关系的是每个位点每年替代的核苷酸数(进化距离 d). 为了得到进化距离 d，人们建立了各种各样情况下的核苷酸替代数学模型(表 1). 进一步人们又假定每个位点的替代率服从 Γ 分布[3-6]，得到了不同核苷酸替代模型下的 Γ 距离. 这样，两后裔序列 i 和 j 的差异描述有 3 种：p 距离、替代率为常数的距离 d 和替代率服从 Γ 分布的 Γ 距离 d_G. 表 1 列出了各种核苷酸替代的数学模型以及在每个模型下的 p 距离、d 距离和 Γ 距离 d_G.

以上这些研究均在校正核苷酸替代率等参数的基础上进行的. 从分子进化角度讲，对一特定的 DNA 序列而言，各物种保留的 DNA 序列是中性选择的结果. 在时间顺序上的各物种，前者应是后者

* 08 回国人员科研启动基金资助项目(Z111020834).
** 通讯联系人.
Tel: 13709253016, E-mail: zhifayuan@nwsuaf.edu.cn
收稿日期：2011-01-10，接受日期：2011-03-31

Table 1 Some mathematical models of nucleotide substitution and their *p*-distance, *d*-distance and Γ distance d_G

		A	T	C	G	Evolution distance
Jukes-Cantor model	A	–	α	α	α	$p=\frac{n_d}{n}$ [1-2]
	T	α	–	α	α	$d_{J\text{-}C}=-\frac{3}{4}\ln(1-\frac{4}{3}p)$[7]
	C	α	α	–	α	
	G	α	α	α	–	$d_{J\text{-}C(G)}=\frac{3}{4}\alpha[(1-\frac{4}{3}p)^{-\frac{1}{\alpha}}-1]$[1]
Kimura model	A	–	β	β	α	$p=\frac{n_d}{n}$
	T	β	–	α	β	$d_K=-\frac{1}{2}\ln(1-2P-Q)-\frac{1}{4}\ln(1-2Q)$[8]
	C	β	α	–	β	
	G	α	β	β	–	$d_{K(G)}=\frac{\alpha}{2}[(1-2P-Q)^{-\frac{1}{\alpha}}+\frac{1}{2}(1-2Q)^{-\frac{1}{\alpha}}-\frac{3}{2}]$[9]
Tamura model	A	–	$\beta\theta_2$	$\beta\theta_1$	$\alpha\theta_1$	$p=\frac{n_d}{n}$
	T	$\beta\theta_2$	–	$\alpha\theta_1$	$\beta\theta_1$	$d_T=-h\ln(1-\frac{P}{h}-Q)-\frac{1}{2}(1-h)\ln(1-2Q)$[10]
	C	$\beta\theta_2$	$\alpha\theta_2$	–	$\beta\theta_1$	
	G	$\alpha\theta_2$	$\beta\theta_2$	$\beta\theta_1$	–	$d_{T(G)}=h\alpha[(1-\frac{P}{h}-Q)^{-\frac{1}{\alpha}}-1]+\frac{\alpha}{2}(1-h)[(1-2Q)^{-\frac{1}{\alpha}}-1]$
Tamura-Nei model	A	–	βg_T	βg_C	$\alpha_1 g_G$	$p=\frac{n_d}{n}$
	T	βg_A	–	$\alpha_2 g_G$	βg_G	$d_{T\text{-}N}=-\frac{2g_Ag_G}{g_R}\ln[1-\frac{g_R}{2g_Ag_G}P_1-\frac{1}{2g_R}Q]$ $-\frac{2g_Tg_C}{g_Y}\ln[1-\frac{g_Y}{2g_Tg_C}P_2-\frac{1}{2g_Y}Q]$ $-2(g_Rg_Y-\frac{g_Ag_Gg_Y}{g_R}-\frac{g_Tg_Cg_R}{g_Y})\ln(1-\frac{1}{2g_Rg_Y}Q)$[3]
	C	βg_A	$\alpha_2 g_G$	–	βg_G	$d_{T\text{-}N(G)}=2\alpha[\frac{g_Ag_G}{g_R}(1-\frac{g_R}{2g_Ag_G}P_1-\frac{1}{2g_R}Q)^{-\frac{1}{\alpha}}$ $+\frac{g_Tg_C}{g_Y}(1-\frac{g_Y}{2g_Tg_C}P_2-\frac{1}{2g_Y}Q)^{-\frac{1}{\alpha}}-g_Ag_G-g_Tg_C-g_Rg_Y$ $+(g_Rg_Y-\frac{g_Ag_Gg_Y}{g_R}-\frac{g_Tg_Cg_R}{g_Y})(1-\frac{1}{2g_Rg_Y}Q)^{-\frac{1}{\alpha}}]$[3]
	G	$\alpha_1 g_A$	βg_T	βg_C	–	

* P, Q are transition ratio and transversion ratio respectively, g_A, g_T, g_C and g_G are nucleotide frequencies, $\theta_1=g_C+g_G$, $\theta_2=g_A+g_T$. $d_{T(G)}$ is obtained by Nei, Gojobori[11] and Jin , Nei[9].

的祖先，后者都是前者选择的结果．本文拟从 DNA 序列变化的动力学过程建立确定进化距离的选择模型，把 p 距离、替代率为常数的距离 d 和替代率服从 Γ 分布的 Γ 距离 d_G 统一起来.

1 DNA 序列进化的选择模型

1.1 选择模型的主要思想

假设从共同祖先序列分化出两个序列 i 和 j，经过 t 世代进化，它们对应核苷酸位点的平均替代率为 $2r$，两序列的同源率(即相同率)记为 $x_{ij}(t)$，则不同率为 $1-x_{ij}(t)$．因为不同率 $1-x_{ij}(t)$随着进化时间 t 而非线性增加，故假定 $1-x_{ij}(t)$的适应度为 1；同源率 $x_{ij}(t)$随着进化时间 t 而非线性下降，因此我们认为 $x_{ij}(t)$的适应度为 $1-2rx^b_{ij}(t)$，其中 r 为各核苷酸位点每年的平均替代率．则在 t+1 世代的替代结果如表 2.

Table 2　The substitution results of $(t+1)^{th}$

	Homology (the sameness ratio)	The difference ratio
t^{th}	$x_{ij}(t)$	$1-x_{ij}(t)$
Fitness	$1-2rx^{b}_{ij}(t)$	1
$(t+1)^{th}$	$[1-2rx^{b}_{ij}(t)]\cdot x_{ij}(t)$	$1-x_{ij}(t)$
The average fitness	$1-2rx^{b+1}_{ij}(t)$	
The ratio of $(t+1)^{th}$	$\frac{x_{ij}(t)\cdot[1-2rx^{b}_{ij}(t)]}{1-2rx^{b+1}_{ij}(t)}$	$\frac{1-x_{ij}(t)}{1-2rx^{b+1}_{ij}(t)}$

由于 r 约为 10^{-9}/ 年，故可设 $1-2rx^{b+1}_{ij}(t)=1$．这样两世代间同源率 $x_{ij}(t)$的改变量为：

$$\Delta x_{ij}(t)=x_{ij}(t)(1-2rx^{b}_{ij}(t))-x_{ij}(t)=-2rx^{b+1}_{ij}(t)$$

连续化得核苷酸序列进化的动力学方程：

$$\begin{cases}\dfrac{\mathrm{d}x_{ij}}{\mathrm{d}t}=-2rx_{ij}^{b+1}\\ x_{ij}|_{t=0}=1\end{cases}\tag{1}$$

其解为：

$$x_{ij}=(1+2brt)^{-\frac{1}{b}}\tag{2}$$

当 $b=0$ 时，方程(1)变为：

$$\begin{cases}\dfrac{\mathrm{d}x_{ij}}{\mathrm{d}t}=-2rx_{ij}\\ x_{ij}|_{t=0}=1\end{cases}\tag{3}$$

其解为：

$$x_{ij}=\mathrm{e}^{-2rt}$$

由式(2)可得：

$$2rt=\frac{1}{b}(x_{ij}^{-b}-1)\tag{4}$$

经过 t 代进化后，物种 i 和 j 共同替代的核苷酸总数为 $2rt$，则选择进化距离为：

$$d_y=2rt=\frac{1}{b}(x_{ij}^{-b}-1)$$

从上式我们可以估计两个序列分化时间 $t=\frac{dy}{2r}$ 或核苷酸的平均替代率 $r=\frac{dy}{2t}$，它们的方差分别为 $V(t)=\frac{V(dy)}{(2r)^2}$ 和 $V(r)=\frac{V(dy)}{(2t)^2}$.

1.2　选择模型在不同核苷酸替代模型下的应用

运用 **1.1** 中提出的动力学思想，对不同替代模型计算相对应的选择进化距离 d_y，并研究在不同模型下与 p 距离、替代率为常数的距离 d 和 Γ 距离 d_G 的关系.

1.2.1　Jukes-Cantor 模型的选择进化距离 $d_{J\text{-}C(y)}$.

在单参数模型下，i 序列和 j 序列间的相同对的总频率 p_{ij} 为[1]：$p_{ij}=\frac{3}{4}-\frac{3}{4}\mathrm{e}^{-8\alpha t}$，则 $1-\frac{4}{3}p_{ij}=\mathrm{e}^{-8\alpha t}$. 令 $x_{ij}=1-\frac{4}{3}p_{ij}$，当 $b=0$ 时，$x_{ij}=\mathrm{e}^{-8\alpha t}$ 满足方程(3)，带入后解得 $2r=8\alpha$．则据式(4)有：

$$8\alpha t=\frac{1}{b}(x_{ij}^{-b}-1)$$

由于在单参数模型时，核苷酸替代总数为 $6\alpha t$，故：

$$d_{J\text{-}C(y)}=6\alpha t=\frac{3}{4b}[(1-\frac{4}{3}p)^{-b}-1]\tag{5}$$

显然，d_y 把 p、$d_{J\text{-}C}$ 和 $d_{J\text{-}C(G)}$统一起来：当 $b=-1$ 时，$d_{J\text{-}C(y)}=p$；当 $b\to 0$ 时，$d_{J\text{-}C(y)}\to-\frac{3}{4}\ln(1-\frac{4}{3}p)=d_{J\text{-}C}$; 当 $b=\frac{1}{a}>0$时，$d_{J\text{-}C(y)}=\frac{3}{4}a[(1-\frac{4}{3}p)^{-\frac{1}{a}}-1]=d_{J\text{-}C(G)}$．我们可以用图来说明 dy 与 p、$d_{J\text{-}C}$ 和 Γ 距离 $d_{J\text{-}C(G)}$之间的关系(图 1)，其中 $d_{J\text{-}C(G)}$中的参数 a 取 2.5，即 $b=0.4$.

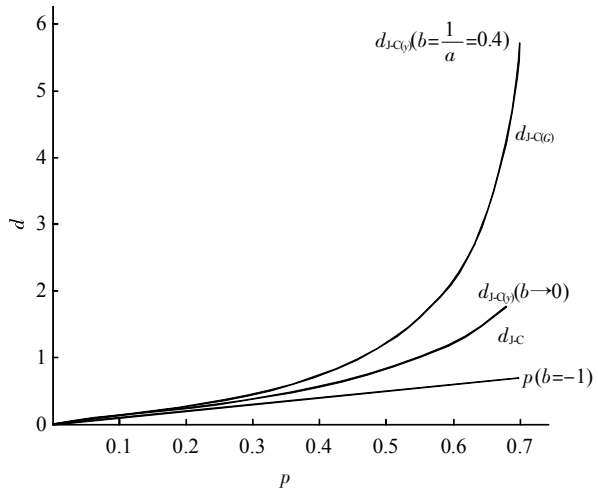

Fig. 1　The relationship among $d_{J\text{-}C(y)}$, p, $d_{J\text{-}C}$ and Γ distance $d_{J\text{-}C(G)}$

1.2.2 Kimura 双参数模型的选择进化距离 $d_{K(y)}$.

在双参数模型下，两序列间颠换、转换对的总频率 Q_{ij}、P_{ij} 分别为：$Q_{ij}=\frac{1}{2}(1-e^{-8\beta t})$, $P_{ij}=\frac{1}{4}(1-2e^{-4(\alpha+\beta)t}+e^{-8\beta t})$，则 $1-2Q_{ij}=e^{-8\beta t}$，$1-2P_{ij}-Q_{ij}=e^{-4(\alpha+\beta)t}$. 令 $x_{ij}=1-2Q_{ij}$，当 $b=0$ 时，$x_{ij}=e^{-8\beta t}$，它满足方程(3)，带入后解得 $2r=8\beta$. 则据式(4)有：

$$8\beta t=\frac{1}{b}(x_{ij}^{-b}-1)=\frac{1}{b}[(1-2Q_{ij})^{-b}-1]$$

又令 $x_{ij}=1-2P_{ij}-Q_{ij}$，当 $b=0$ 时，$x_{ij}=e^{-4(\alpha+\beta)t}$，它亦满足方程(3)，入后解得 $2r=4(\alpha+\beta)$. 则据式(4)有：

$$4(\alpha+\beta)t=\frac{1}{b}[(1-2P_{ij}-Q_{ij})^{-b}-1]$$

由于是同一对序列的进化，故 b 是不变的. 在双参数模型之下，经过进化时间 t，替代的核苷酸数为 $(2\alpha+4\beta)t$[9]，故

$$d_{K(y)}=(2\alpha+4\beta)t=\frac{1}{2b}[(1-2P-Q)^{-b}-1]+\frac{1}{4b}[(1-2Q)^{-b}-1]$$

$$=\frac{1}{2b}[(1-2P-Q)^{-b}+\frac{1}{2}(1-2Q)^{-b}-\frac{3}{2}] \quad (6)$$

显然，$d_{K(y)}$把 p 距离、d_K 及其 Γ 距离 $d_{K(G)}$统一起来了：

当 $b=-1$ 时，$d_{K(y)}=-\frac{1}{2}(1-2P-Q+\frac{1}{2}-Q-\frac{3}{2})=P+Q=p$；

当 $b\to 0$ 时，$d_{K(y)}\to-\frac{1}{2}\ln(1-2P-Q)-\frac{1}{4}\ln(1-2Q)=d_K$；

当 $b=\frac{1}{a}>0$ 时，$d_{K(y)}=\frac{a}{2}[(1-2P-Q)^{-\frac{1}{a}}+\frac{1}{2}(1-2Q)^{-\frac{1}{a}}-\frac{3}{2}]=d_{K(G)}$.

1.2.3 Tamura 模型的选择进化距离 $d_{T(y)}$.

在 Tamura 模型下，两序列间颠换和转换对的总频率 Q_{ij} 和 P_{ij} 分别为[10]：$Q_{ij}=\frac{1}{2}(1-e^{-4\beta t})$，$P_{ij}=\frac{h}{2}(1+e^{-4\beta t}-2e^{-2(\alpha+\beta)t})$，则 $1-2Q_{ij}=e^{-4\beta t}$，$1-\frac{P_{ij}}{h}-Q_{ij}=e^{-2(\alpha+\beta)t}$. 令 $x_{ij}=1-2Q_{ij}$，当 $b=0$ 时，$x_{ij}=e^{-4\beta t}$，它亦满足方程(3)，入后解得 $2r=4\beta$. 则据式(4)有：

$$4\beta t=\frac{1}{b}(x_{ij}^{-b}-1)=\frac{1}{b}[(1-2Q_{ij})^{-b}-1]$$

又令 $x_{ij}=1-\frac{P_{ij}}{h}-Q_{ij}$，当 $b=0$ 时，$x_{ij}=e^{-2(\alpha+\beta)t}$，它亦满足方程(3)，入后解得 $2r=2(\alpha+\beta)$. 则据式(4)有：

$$2(\alpha+\beta)t=\frac{1}{b}[(1-\frac{P_{ij}}{h}-Q_{ij})^{-b}-1]$$

其中 $h=2\theta(1-\theta)$，θ 为 GC 含量. 由于是同一对序列的进化，故 b 是不变的. 在 Tamura 模型之下，经过进化时间 t，替代的核苷酸数为 $[4\alpha\theta(1-\theta)+2\beta]t=(2\alpha h+2\beta)t$[10]，故

$$d_{T(y)}=(2\alpha h+2\beta)t=2(\alpha+\beta)ht+2\beta t-2\beta ht$$

$$=\frac{h}{b}[(1-\frac{P}{h}-Q)^{-b}-1]+\frac{1}{2b}(1-h)[(1-2Q)^{-b}-1]$$

同样，$d_{T(y)}$把 p 距离、d_T 及其 Γ 距离 $d_{T(G)}$统一起来了：

当 $b=-1$ 时，$d_{T(y)}=-h(1-\frac{P}{h}-Q-1)-\frac{1}{2}(1-h)(1-2Q-1)=P+Q=p$；

当 $b\to 0$ 时，$d_{T(y)}\to-h\ln(1-\frac{P}{h}-Q)-\frac{1}{2}(1-h)\ln(1-2Q)=d_T$；

当 $b=\frac{1}{a}>0$ 时，$d_{T(y)}=ha[(1-\frac{P}{h}-Q)^{-\frac{1}{a}}-1]+\frac{a}{2}(1-h)[(1-2Q)^{-\frac{1}{a}}-1)=d_{T(G)}$.

1.2.4 Tamura-Nei 模型的选择进化距离 $d_{T\text{-}N(y)}$.

在 Tamura-Nei 模型下，两序列间 AG 转换对、TC 转换对和颠换对的总频率 P_{1ij}、P_{2j} 和 Q_{ij} 分别为[3]: $P_{1ij}=2g_Ag_G(1+\frac{g_Y}{g_R}e^{-2\beta t}-\frac{1}{g_R}e^{-2(\alpha_1g_R+\beta g_Y)t})$，$P_{2ij}=2g_Tg_C(1+\frac{g_R}{g_Y}e^{-2\beta t}-\frac{1}{g_Y}e^{-2(\alpha_2g_Y+\beta g_R)t})$, $Q_{ij}=2g_Rg_Y(1-e^{-2\beta t})$. 则 $1-\frac{g_RP_{1ij}}{2g_Ag_G}-\frac{Q_{ij}}{2g_R}=e^{-2(\alpha_1g_R+\beta g_Y)t}$，$1-\frac{g_YP_{2ij}}{2g_Tg_C}-\frac{Q_{ij}}{2g_Y}=e^{-2(\alpha_2g_Y+\beta g_R)t}$，$1-\frac{Q_{ij}}{2g_Rg_Y}=e^{-2\beta t}$，其中 $g_Y=g_T+g_C$、$g_R=g_A+g_G$ 分别为嘧啶(T 和 C)、嘌呤(A 和 G)的相对频率. 令 $x_{ij}=1-\frac{Q_{ij}}{2g_Rg_Y}$，当 $b=0$ 时，$x_{ij}=e^{-2\beta t}$，它亦满足方程(3)，入后解得 $2r=2\beta$. 则据式(4)有：

$$2\beta t=\frac{1}{b}(x_{ij}^{-b}-1)=\frac{1}{b}[(1-\frac{Q_{ij}}{2g_Rg_Y})^{-b}-1]$$

令 $x_{ij}=1-\frac{g_RP_{1ij}}{2g_Ag_G}-\frac{Q_{ij}}{2g_R}$，当 $b=0$ 时，$x_{ij}=e^{-2(\alpha_1g_R+\beta g_Y)t}$，它亦满足方程(3)，入后解得 $2r=2(\alpha_1g_R+\beta g_Y)$. 则据式(4)有：

$$2(\alpha_1g_R+\beta g_Y)t=\frac{1}{b}[(1-\frac{g_RP_{1ij}}{2g_Ag_G}-\frac{Q_{ij}}{2g_R})^{-b}-1]$$

又令 $x_{ij}=1-\frac{g_YP_{2ij}}{2g_Tg_C}-\frac{Q_{ij}}{2g_Y}$，当 $b=0$ 时，$x_{ij}=e^{-2(\alpha_2g_Y+\beta g_R)t}$，它亦满足方程(3)，入后解得 $2r=2(\alpha_2g_Y+\beta g_R)$. 则据式(4)有：

$$2(\alpha_2 g_Y+\beta g_R)t=\frac{1}{b}[(1-\frac{g_Y P_{2ij}}{2g_T g_C}-\frac{Q_{ij}}{2g_Y})^{-b}-1]$$

由于是同一对序列的进化，故 b 是不变的．在 Timura-Nei 模型之下，经过进化时间 t，替代的核苷酸数为 $4\alpha_1 g_A g_G t+4\alpha_2 g_T g_C t+4\beta g_R g_Y t$[3]，则：

$$\begin{aligned}d_{T\text{-}N(y)}&=4\alpha_1 g_A g_G t+4\alpha_2 g_T g_C t+4\beta g_R g_Y t\\&=\frac{2g_A g_G}{g_R}\frac{1}{b}[(1-\frac{g_R P_1}{2g_A g_G}-\frac{Q}{2g_R})^{-b}-1]+\frac{2g_T g_C}{g_Y}\frac{1}{b}[(1-\\&\frac{g_Y P_2}{2g_T g_C}-\frac{Q}{2g_Y})^{-b}-1]+2g_R g_Y\frac{1}{b}[(1-\frac{Q}{2g_R g_Y})^{-b}-1]\\&-\frac{2g_A g_G g_Y}{g_R}\frac{1}{b}[(1-\frac{Q}{2g_R g_Y})^{-b}-1]\\&-\frac{2g_T g_C g_R}{g_Y}\frac{1}{b}[(1-\frac{Q}{2g_R g_Y})^{-b}-1]\end{aligned}$$

同样，$d_{T\text{-}N(y)}$把 p 距离、$d_{T\text{-}N}$ 及其 Γ 距离 $d_{T\text{-}N(G)}$统一起来了：

当 $b=-1$ 时，$d_{T\text{-}N(y)}=P_1+P_2+Q=p$

当 $b\to 0$ 时，

$$\begin{aligned}d_{T\text{-}N(y)}\to&-\frac{2g_A g_G}{g_R}\ln[1-\frac{g_R}{2g_A g_G}P_1-\frac{1}{2g_R}Q]-\frac{2g_T g_C}{g_Y}\ln[1-\\&\frac{g_Y}{2g_T g_C}P_2-\frac{1}{2g_Y}Q];\\&-2(g_R g_Y-\frac{g_A g_G g_Y}{g_R}-\frac{g_T g_C g_R}{g_Y})\ln(1-\frac{1}{2g_R g_Y}Q)=d_{T\text{-}N}\end{aligned}$$

当 $b=\frac{1}{a}>0$ 时，

$$\begin{aligned}d_{T\text{-}N(y)}=2a[&\frac{g_A g_G}{g_R}(1-\frac{g_R}{2g_A g_G}P_1-\frac{1}{2g_R}Q)^{-\frac{1}{a}}+\frac{g_T g_C}{g_Y}(1-\\&\frac{g_Y}{2g_T g_C}P_2-\frac{1}{2g_Y}Q)^{-\frac{1}{a}}\\&+(g_R g_Y-\frac{g_A g_G g_Y}{g_R}-\frac{g_T g_C g_R}{g_Y})(1-\frac{1}{2g_R g_Y}Q)^{-\frac{1}{a}}-g_A g_G-\\&g_T g_C-g_R g_Y]=d_{T\text{-}N(G)}.\end{aligned}$$

1.3　动力学参数 b 与 $\frac{dx}{dt}$ 之间的关系

由式(1)我们可以得到 $\frac{dx}{dt}=-2rx^{b+1}$，令 $r=10^{-9}$，则 $\frac{dx}{dt}$ 与 b 之间的关系如图 2．从图 2 可以看出，$\frac{dx}{dt}$ 随动力学参数 b 的增大而增大，当 $b>4$ 时，$\frac{dx}{dt}$ 几乎没有变化，即核苷酸的同源率几乎没有变化．当 b 很大时，选择进化距离 d_y 和 p 距离趋于一致．

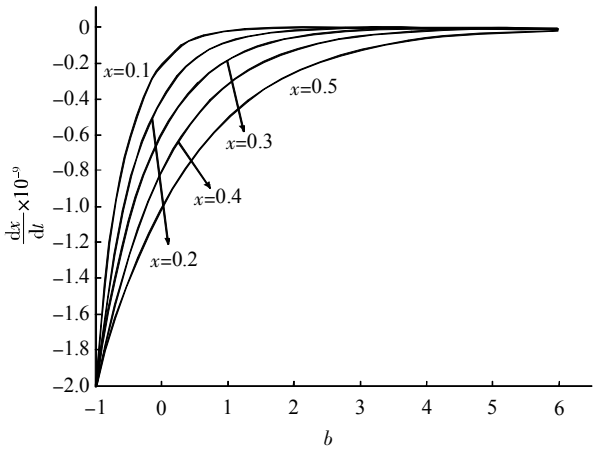

Fig. 2　The relationship between $\frac{dx}{dt}$ and b

2　b 的估计与应用举例

2.1　b 的估计

设物种 1，2，…，N，前者是后者的祖先，其出现的时间分别为 t_1，t_2，…，t_N．假设在某 DNA 序列中，同源率为 $x_{ij}(i<j)$，不同率为 $1-x_{ij}$．据模型 $\frac{dx_{ij}}{dt}=-2rx_{ij}^{b+1}$，由于 x_{ij} 是物种 i 与 j 物种联配的结果，故 $\frac{x_{ij}}{t_i-t_j}$ 是 $-\frac{dx_{ij}}{dt}$ 的估计，则有 $\frac{x_{ij}}{t_i-t_j}=2rx_{ij}^{b+1}$．这时令 $y_{ij}=\ln\frac{x_{ij}}{t_i-t_j}$，$A=\ln 2r$，$B=b+1$，$x'_{ij}=\ln x_{ij}$，则有线性回归方程 $y_{ij}=A+Bx'_{ij}$．由 $\frac{N(N-1)}{2}$ 个点(x'_{ij}, y_{ij})进行直线回归分析和相应的统计检验，可估计出 A，B，进而估计出 $b=B-1$ 和 $r=\frac{1}{2}e^A$．由于时间 t 是以百万年为单位，因此 r 最后的结果应该乘以 10^{-6}.

2.2　应用举例

文中以 16 个物种的线粒体基因序列为数据，根据以上方法计算了 Jukes-Cantor 模型的选择进化距离 $d_{J\text{-}C\ (y)}$．表 3 给出了人等 16 个物种的名称和它们在 GenBank 中的线粒体基因序列的检索号以及它们出现的地质年代，表 4 给出了差异率 p_{ij}(上对角矩阵)和相同率 q_{ij}(下对角矩阵).

Table 3 Names and GenBank accession numbers and geological time of 16 species

Name	Accession number	Geological time (*Ma*)
Human	V00662	8
Mouse	V00711	11.5
Cattle	V00654	23
Gibbon	X99256	29
Pig	AJ002189	51
Cat	U20753	53
Guinea pig	AJ222767	55.5
Squirrel	AJ238588	61
Opossum	Z29573	66
Shrew	AF348081	78
Platypus	X83427	177
Turtle	AB012104	225
Chicken	AP003580	300
Newt	AY458597	360
Carp	X61010	410
Shark	AJ310141	460

Table 4 Same ratio and difference ratio of nucleotide sequence in mitochondrial

	Shark	Carp	Newt	Chicken	Turtle	Platypus	Shrew	Opossum	Squirrel	Guinea pig	Cat	Pig	Gibbon	Cattle	Mouse	Human
Shark		0.2511	0.2851	0.3067	0.2823	0.3167	0.3121	0.3168	0.3185	0.3230	0.3134	0.3041	0.3296	0.3055	0.3116	0.3267
Carp	0.7489		0.2919	0.3042	0.2788	0.3286	0.3129	0.3220	0.3223	0.3229	0.3150	0.3036	0.3267	0.3101	0.3129	0.3271
Newt	0.7149	0.7081		0.3211	0.2981	0.3307	0.3239	0.3278	0.3336	0.3337	0.3248	0.3190	0.3462	0.3246	0.3303	0.3409
Chicken	0.6933	0.6958	0.6789		0.2669	0.3363	0.3277	0.3334	0.3339	0.3330	0.3225	0.3127	0.3243	0.3187	0.3259	0.3231
Turtle	0.7177	0.7212	0.7019	0.7331		0.3136	0.3037	0.3043	0.3107	0.3105	0.3014	0.2919	0.3205	0.2943	0.3008	0.3113
Platypus	0.6833	0.6714	0.6694	0.6638	0.6864		0.2726	0.2756	0.2786	0.2933	0.2822	0.2754	0.3092	0.2735	0.2842	0.3068
Shrew	0.6879	0.6871	0.6761	0.6723	0.6963	0.7274		0.2717	0.2353	0.2540	0.2125	0.2091	0.2608	0.2119	0.2453	0.2632
Opossum	0.6832	0.6780	0.6722	0.6666	0.6957	0.7244	0.7283		0.2735	0.2856	0.2768	0.2723	0.3080	0.2719	0.2737	0.3034
Squirrel	0.6815	0.6777	0.6664	0.6661	0.6893	0.7214	0.7647	0.7265		0.2507	0.2330	0.2302	0.2692	0.2369	0.2494	0.2650
Guinea pig	0.6770	0.6771	0.6663	0.6670	0.6895	0.7067	0.7460	0.7144	0.7493		0.2537	0.2471	0.2780	0.2533	0.2651	0.2784
Cat	0.6866	0.6850	0.6752	0.6775	0.6986	0.7178	0.7875	0.7232	0.7670	0.7463		0.1939	0.2572	0.2003	0.2503	0.2549
Pig	0.6960	0.6964	0.6810	0.6873	0.7081	0.7246	0.7909	0.7277	0.7698	0.7529	0.8061		0.2500	0.1885	0.2437	0.2493
Gibbon	0.6704	0.6733	0.6538	0.6757	0.6795	0.6908	0.7392	0.6920	0.7308	0.7220	0.7428	0.7501		0.2546	0.2775	0.1533
Cattle	0.6945	0.6899	0.6754	0.6813	0.7058	0.7265	0.7881	0.7281	0.7631	0.7467	0.7998	0.8115	0.7454		0.2465	0.2497
Mouse	0.6884	0.6871	0.6697	0.6741	0.6992	0.7158	0.7547	0.7263	0.7506	0.7349	0.7497	0.7564	0.7225	0.7536		0.2754
Human	0.6733	0.6729	0.6591	0.6769	0.6887	0.6932	0.7368	0.6967	0.7350	0.7216	0.7452	0.7508	0.8467	0.7503	0.7246	

在 Jukes-Cantor 模型下，令 $x'_{ij}=\ln x_{ij}=\ln(1-\frac{4}{3}p_{ij})$, $y_{ij}=\ln[\frac{1}{t_i-t_j}(1-\frac{4}{3}p_{ij})]$，则得到回归方程：$y=1.86+14.04\ x'$，其中 $A=1.86$，$B=14.04$，则 $b=B-1=13.04$，相关系数 $r=0.79>r_{0.05}(120)$，表明回归是极显著的，即 b 的估计是显著的．核苷酸的平均替代率为 $r=\frac{1}{2}e^{1.86}=3.2\times10^{-6}$/ 年，表 5 为 $b=13.04$ 时 Jukes - Cantor 模型下的选择进化距离 $d_{J\text{-}C\ (y)}=\frac{3}{52.16}[(1-\frac{4}{3}p)^{-13.04}-1]$.

Table 5　Evolution distances of b=13.04 in Jukes-Cantor′s selection model

	Carp	Newt	Chicken	Turtle	Platypus	Shrew	Opossum	Squirrel	Guinea pig	Cat	Pig	Gibbon	Cattle	Mouse	Human
Shark	11.661	29.316	54.610	27.104	73.574	64.064	73.729	77.728	89.068	66.519	50.508	109.070	52.732	63.136	99.809
Carp		35.599	50.715	24.588	105.810	65.552	86.259	87.185	88.688	69.804	49.789	99.593	60.311	65.689	100.890
Newt			83.902	42.483	112.690	91.582	103.100	123.520	124.060	93.971	78.722	184.410	93.568	111.470	155.820
Chicken				17.746	134.290	102.870	122.710	124.610	121.370	87.745	65.279	92.569	78.058	97.260	89.259
Turtle					67.078	49.993	50.819	61.450	61.068	46.735	35.528	82.485	38.028	45.887	62.482
Platypus						20.709	22.538	24.493	37.050	27.051	22.408	58.707	21.273	28.577	54.723
Shrew							20.239	7.742	12.576	4.373	4.020	15.064	4.313	10.000	16.056
Opossum								21.232	29.723	23.245	20.551	56.677	20.316	21.355	49.484
Squirrel									11.533	7.300	6.799	18.895	8.067	11.138	16.863
Guinea pig										12.460	10.485	24.069	12.346	16.927	24.304
Cat											2.784	13.696	3.245	11.406	12.858
Pig												11.302	2.448	9.591	11.097
Gibbon													12.787	23.745	1.076
Cattle														10.315	11.220
Mouse															22.364

2.3　各种进化距离的比较

用不同的进化距离可以构建不同的基因进化树，因此可通过构建的基因进化树来比较各进化距离．根井正利等[11]认为，在不同的模型比较中“一个参数很多的模型比简单模型会更好地解释数据，但是建立在很多参数模型的统计预测(或拓扑结构估计)易带来更多误差．因此，只要模型能很好地表达替代模式，使用一个简单的模型是明智的.”因此，我们只针对 Jukes-Cantor 模型下的选择进化距离 $d_{J\text{-}C(y)}$与 p 距离以及 $d_{J\text{-}C}$ 进行比较．本文用 Phylip 软件包中的程序构建基于 $d_{J\text{-}C(y)}$、p 以及 $d_{J\text{-}C}$ 三种进化距离的基因进化树并通过 Bootstrap 抽样进行自展检验(图 3～5).

其中树内部分支上所标注的数字为 Bootstrap 支持率，即自展值．从以下几个图中可以看出：a．图 3 和 4 的基因进化树的内部分支结构是完全一致的，图 5 与其他图形的进化树分支结构不一致的地方为负鼠和鸭嘴兽的位置有所区别；b．我们还可以看出图 3 和图 4 基因进化树的每个内部分支的自展值全为 100%，而图 5 内部分支的自展值分别出现了 85%，93%，没有达到 95%的置信度．综上，我们可以看出基于 $d_{J\text{-}C(y)}$和 p 距离的基因进化树不论拓扑结构还是各分支上的自展值都完全一致，即它印证了本文 **1.3** 中提到的“当 b 很大时，选择进化距离 d_y 和 p 距离趋于一致”这一选择模型特有的性质．这些比较说明，构建基因进化树时，用

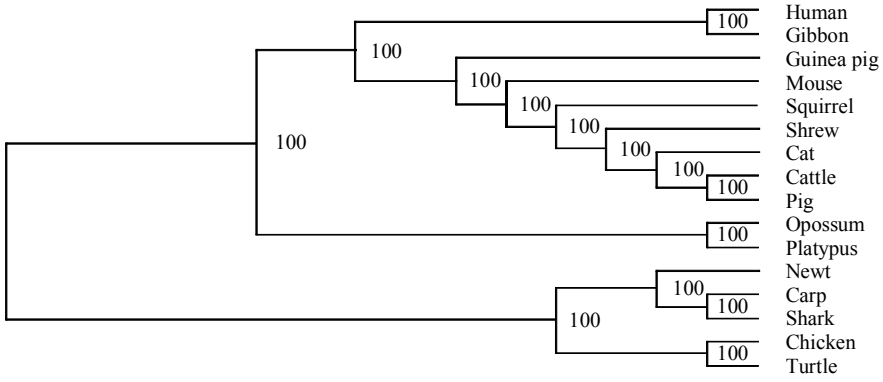

Fig. 3　Gene tree of 16 species based on $d_{J\text{-}C(y)}$

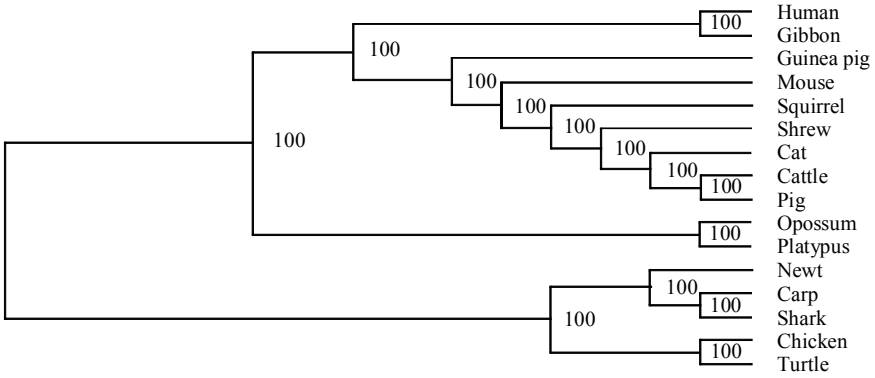

Fig. 4 Gene tree of 16 species based on *p* distance

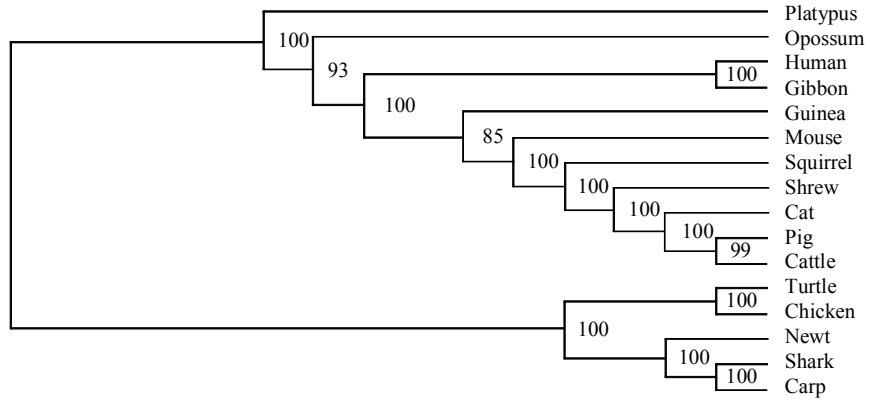

Fig. 5 Gene tree of 16 species based on $d_{J\text{-}C}$

该文提出的选择进化距离比用其他的进化距离能带来更正确的信息. 另外，我们也计算出了 Kimura 双参数模型下的选择进化距离 $d_{K(y)}$并构建其相应的进化树，但其进化树内部分支的自展值没有达到 100%，没有 Jukes-Cantor 模型下的结果好，因此我们没有显示其结果，这也进一步验证了根井正利等[1]关于模型选择的原则.

3 结论与讨论

在构建基因进化树时，需获得各物种在同一 DNA 序列间的进化距离，而进化距离是用 DNA 序列间的核苷酸替代数来估计. 但核苷酸替代模型有很多，并且在同一核苷酸替代模型下的进化距离有 p 距离、d 距离和 Γ 距离 d_G 三种. 在构建基因进化树时，如何对这三种距离进行选择，目前还没有一种普适的统计方法.

在 DNA 序列的进化演变中，两序列间核苷酸对的不同率随着进化时间 t 而非线性增加，同源率随着进化时间而非线性下降，这种现象是在随机突变和自然选择的联合作用下形成的. 本文在选择的意义下提出了 DNA 进化的动力学模型，由此获得了相应的进化距离 d_y，把各种模型下的 p 距离、d 距离和 Γ 距离 d_G 统一起来，并给出了选择模型中的参数 b 和核苷酸替代率 r 的估计方法，可以使研究者在已知各物种出现地质年代情况下，根据自己的联配数据，利用线性回归思想确定进化距离 d_y 中的 b，进而确定出进化距离.这就避免了对 p 距离、d 距离和 Γ 距离 d_G 的选择问题.

本文用 16 个物种的线粒体基因数据，建立在 Jukes-Cantor 单参模型下的进化距离 $d_{J\text{-}C\ (y)}$，进一步通过 Bootstrap 抽样建立 $d_{J\text{-}C(y)}$、p 和 $d_{J\text{-}C}$ 下的基因进化树来对选择进化距离进行评估. 结果表明，选择进化距离 d_y 是一种有效的构建进化距离的方法. 需要特别指出的是，动力学模型需要物种的分化时

间和核苷酸的平均替代速率，而不是核苷酸每个位点的替代速率．另外，要构建比较准确的进化树，建议采用更多不同的 DNA 序列进行分析．

参 考 文 献

[1] 根井正利, 苏德海尔·库马. 分子进化与系统发育. 吕宝忠, 钟扬, 高莉萍, 等. 北京: 高等教育出版社，2002: 29-39
Nei M, Kumar S. Molecular Evolution and Phylogenetics. Lü B Z, Zhong Y, Gao L P, *et al*. Beijing: Higher Education Press, 2002: 29-39

[2] 杨子恒. 计算分子进化. 钟 扬, 张文娟, 梅 旖, 等. 上海: 复旦大学出版社，2008: 3-21
Yang Z H. Computational Molecular Evolution. Zhong Y, Zhang W J, Mei Y, *et al*. Shanghai: Fudan University Press, 2008: 3-21

[3] Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution, 1993, **10**(3): 512-526

[4] Kocher T D, Wilson A C. Sequence evolution of mitochondrial DNA in human and chimpanzees: Control region and a protein-coding region//Osawa S, Honjo T. In Evolution of Life. New York: Springer-Verlag, 1991: 391-413

[5] Wakeley J. Substitution rate variation among sites in hypervariable I of human mitochondrial DNA. J Molecular Evolution, 1993, **37**(6): 613-623

[6] Wakeley J. Substitution rate variation among sites and the estimation of transition bias. Molecular Biology and Evolution, 1994, **11**(3): 436-442

[7] Jukes T H, Cantor C R. Evolution of protein molecules // Munro H N. In Mammalian protein metabolism. New York: Academic, 1969: 21-132

[8] Kimura M. A simple method for estimating evolution rates of base substitution through comparative studies of nucleotide sequences. J Molecular Evolution, 1980, **16**(2): 111-120

[9] Jin L, Nei M. Limitation of the evolutionary parsimony method of phylogenetic analysis. Molecular Biology and Evolution, 1990, **7**(1): 82-102

[10] Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G +C-content biases. Molecular Biology and Evolution, 1992, **9**(4): 678-687

[11] Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution, 1986, **3**(6): 418-426

A New Evolution Distance of DNA Sequence and Its Application*

LIANG Li-Ping[1)], XIE Xiao-Li[1)], CHENG Jian[1,2)], WANG Zhen-Feng[1)],
GUO Man-Cai[1)], YUAN Zhi-Fa[1)**]

([1)] *College of Science, Northwest A & F University, Yangling* 712100, China;
[2)] *Bioinformatics Center, Northwest A & F University, Yangling* 712100, China)

Abstract In order to study the variation, dynamic equation of DNA sequences was established by the homology in nucleotide sequences, and further evolution distance d_y (selected evolutionary distance) between species was obtained. Selected evolutionary distance d_y was calculated in 4 nucleotide substitution models, and indicated the relationship among p-distance, d-distance and Γ distance d_G. According to the characteristics of dynamic equations, selection model can be transformed into a linear regression problem. Then both of the parameter b and the average substitution ratio of nucleotide each year r were obtained by the Least Square Method. Take the mitochondrial DNA sequences of 16 species for example to illustrate the new evolution distance, and then evolution trees are constructed in order to compare different evolution distance. The results indicate that the new distance d_y is an efficient evolution distance to analyze DNA sequence.

Key words selection model, DNA sequences,evolution distance, evolution tree, nucleotide substitution model
DOI: 10.3724/SP.J.1206.2010.00605

*This work was supported by a grant from 08 special talent fund of Northwest A & F University(Z111020834).

**Corresponding author.
Tel: 86-13709253016, E-mail: zhifayuan@nwsuaf.edu.cn

Received: January 10, 2011 Accepted: March 31, 2011