第五章 分子进化：系统树的构建[1]

自 20 世纪中叶，随着分子生物学的不断发展，进化研究也进入了分子进化(molecular evolution)研究水平，并建立了一套依赖于核酸、蛋白质序列信息的理论和方法。随着基因组测序计划的实施，基因组的巨量信息对若干生物领域重大问题的研究提供了有力的帮助，分子进化研究再次成为生命科学中最引人注目的领域之一。这些重大问题包括：遗传密码的起源、基因组结构的形成与演化、进化的动力、生物进化等等。分子进化研究目前更多地是集中在分子序列上，但随着越来越多生物基因组的测序完成，从基因组水平上探索进化奥秘，将开创进化研究的新天地。人与老鼠的基因组大小相似，都含有约 30 亿碱基对，基因的数量也相近，可人与老鼠为何差异如此之大？从进化的角度如此解释？是否可以在浩如烟海的基因组密码中获得答案？

第一节 系统树及其它

一．系统树

分类学涉及的问题是将生物合理地分成一定的类群，使类群内的个体成员相同或非常相似。分类学可以进行物种的分类。对于进化研究，分类涉及到系统发育的重构(reconstruction of phylogenies)，构建系统发育过程有助于通过物种间隐含的种系关系揭示进化动力的实质。Nei(1987)、Li 和 Graur(1991)等人已对构建系统发育过程进行了全面的总结，本章只提示性地介绍相关方法。

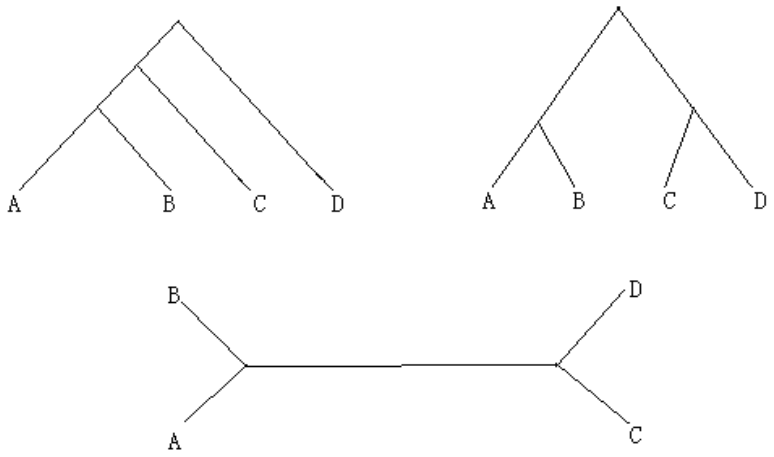
表型的(phenetic)和遗传的(cladistic)数据有着明显差异。Sneath 和 Sokal(1973)将表型性关系定义为根据物体一组表型性状所获得的相似性，而遗传性关系含有祖先的信息，因而可用于研究进化的途径。这两种关系可用于系统树(phylogenetic tree)或树状图(dendrogram)来表示。表型分枝图(phenogram)和进化分枝图(cladogram)两个术语已用于表示分别根据表型性的和遗传性的关系所建立的关系树。进化分枝图可以显示事件或类群间的进化时间，而表型分枝图则不需要时间概念。在本章，我们将不会十分注重这一区别，正如 Nei(1987)指出的，如果表型相似性的尺度意味着进化上的相似性的程度，则有关表型的方法就可以提供遗传上的关系树。文献中，更多地是使用“系统树”一词来表示进化的途径，另外还有系统发育树、物种树(species tree)、基因树等等一些相同或含义略有差异的名称。

系统树分有根(rooted)和无根(unrooted)树。图 5.1 中显示了 4 个物种部分有根树和无根树形式。有根树反映了树上物种或基因的时间顺序，而无根树只反映分类单元之间的距离而不涉及谁是谁的祖先问题。

用于构建系统树的数据有二种类型：一种是特征数据(character data)，它提供了基因、个体、群体或物种的信息；二是距离数据(distance data)或相似性数据(similarity data)，它涉及的则是成对基因、个体、群体或物种的信息。距离数据可由特征数据计算获得，但反过来则不行。这些数据可以矩阵的形式表

[1]本部分内容主要取自Weir B．S．(徐云碧等译)．遗传学数据分析—群体遗传学离散型数据分析方法，北京：中国农业出版社，1996

达。距离矩阵(distance matrix)是在计算得到的距离数据基础上获得的，距离



的计算总体上是要依据一定的遗传模型，并能够表示出两个分类单位间的变化量。系统树的构建质量依赖于距离估算的准确性。

图 5.1 4 个物种(A、B、C 和 D)的 2 种有根树和 1 种无根树形式

系统树的构建主要有三种方法。距离矩阵法(distance matrix method)是根据每对物种之间的距离，其计算一般很直接，所生成的树的质量取决于距离尺度的质量。距离通常取决于遗传模型。最大简约(maximum parsimony)法较少涉及遗传假设，它通过寻求物种间最小的变更数来完成的。对于模型的巨大依赖性是最大似然(maximum likelihood)法的特征，该方法在计算上繁杂，但为统计推断提供了基础。

二．遗传模型和序列距离

遗传模型在系统树构建中非常重要，因为距离计算过程必须在一定的遗传假设下才可能进行。以下以两个在 DNA 序列距离计算中最为常用的遗传模型为例，说明距离数据的计算由来。

在分子进化研究中，我们往往认定这样的一个假设，即序列是同源的，它们具有单一祖先序列；这一祖先序列在进化过程中发生了一系列的核苷酸突变。图 5.2 表示了各种核苷酸变化情况。

在以上的假设基础上，Judes 和 Cautor 进一步假设每一碱基具有同等机率突变为另外 3 种碱基中的任何一种，其频率常数为 μ/3，μ为碱基替换频率。Kimura(1980)考虑到转换(transition，两种嘧啶或两种嘌呤碱基之间的突变)和颠换(transversion，一个嘧啶和一个嘌呤碱基之间的突变)具有不同的频率，α和β。表 5.1 简要说明了以上两种遗传模型。

表 5.1 Judes-Cantor 单参数模型(上三角部分)和 Kimura 两参数模型(下三角部分)。α、β分别为两种碱基间 2 个不同的置换频率。

	A	T	G	C

A		a	a	a
T	β		a	a
G	a	β		a
C	β	a	β	

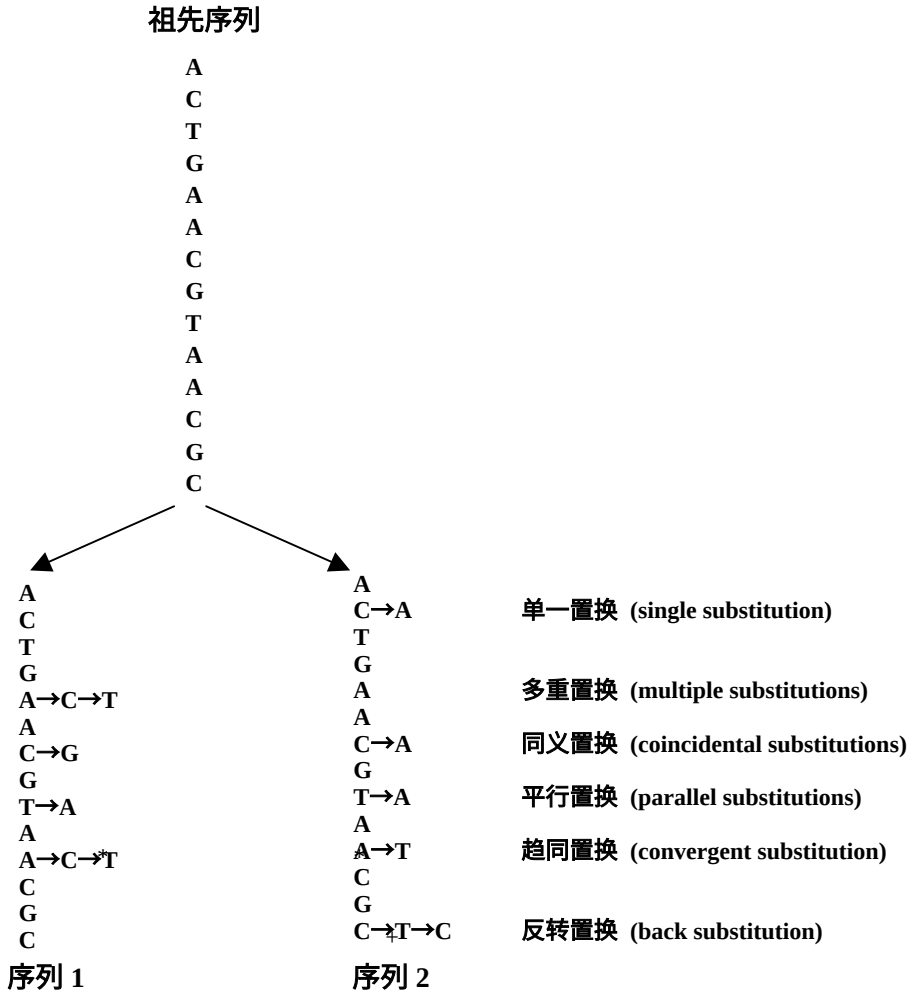


图 5.2 同源序列间的核苷酸置换(Li & Graur,1991)

根据以上遗传模型，Judes 和 Cantor(1969)提出了 DNA 序列距离 *K*(最早为氨基酸序列引入)计算公式：

$$K = \frac{3}{4}\ln(\frac{4}{4q-1}) \approx 2\mu t \qquad (5.1)$$

其中 q 为同源 DNA 序列中具有相同碱基的概率，经过 t 世代，由于祖先序列的趋异变化，其值为：

$$q_t = \frac{1}{4} + \frac{3}{4}(1-\frac{8\mu}{3})^t \qquad (5.2)$$

μ为碱基替换频率。

距离 K 适用于显示两条序列从一个祖先序列趋异进化以来的时间，并能用于序列间系统树的构建。在计算时，均需要将序列作初步的列线分析。Kimura 在其两参数模型下证实，由于趋异变化，由转换造成差异(Ⅰ型变化)或由颠换造成差异(Ⅱ型变化)的碱基，随时间而变化：

$$P_{It} = \frac{1}{4}(1-2e^{-4(\alpha+\beta)t} + e^{-8\beta t})$$

$$P_{It} = \frac{1}{4}(1 - e^{-8\beta t}) \qquad (5.3)$$

如果 k＝ a ＋2β是单位时间碱基替换的总频率，则适合作为系统树的距离尺度为：

$$K = -\frac{1}{2}\ln[(1 - 2p_I - p_{II})\sqrt{1 - 2P_{II}}] \approx 2kt \qquad (5.4)$$

该类距离可用于有关系统树距离矩阵中，用样本比值代入（5.4）式就可估计这些距离。

Kimura 以兔和鸡的β-球蛋白序列为例(见图 5.3)，计算了上述距离。序列长 438bp，有 58 个Ⅰ型变化、63 个Ⅱ型变化。因此，$\tilde{p}_I$=0.1324，$\tilde{P}_{II}$=0.1438，Kimura 距离为 0.3513。这与只根据相同碱基比例$\tilde{q}$=0.7237 所得 Jukes-Cantor 距离 0.3446 没有本质上的差异。

图 5.3 兔和鸡的β-球蛋白序列。每两条序列上下两行星号表示由转换（Ⅰ型变化)或颠换(Ⅱ型变化)造成的碱基差异。

DNA序列距离K又可称为DNA序列间的分歧度(sequence divergence)，即序列间相异性的一个指标。蛋白质序列的分歧度分为两序列同义变化的分歧度(K_S)和非同义变化的分歧度(K_A)，根据Jukes-Cautor单参数模型和Kimura两参数模型等遗传模型，可以分别计算得到两序列的分歧度(或称为蛋白质序列间的距离)。

三．分子进化与系统发育分析软件

软件名称	网址	说 明
PHYLIP	http://evolution.genetics.washinton.edu/phylip/software.html	目前发布最广，用户最多的通用系统树构建软件，由美国华盛顿大学 Felsenstein 开发，可免费下载，适用绝大多数操作系统
PAUP	scavotto@sinauer.com 或 ftp://onyx.si.edu/paup	国际上最通用的系统树构建软件之一，美国 simthsonion institute 开发，仅适用 Apple-Macintosh 和 UNIX 操作系统
Tree of Life	http://phylogeny.arizona.edu/tree/program/program.html	美国 University of Arizona 建立的系统发育方面网站
MEGA	http://bioinfo.weizmann.ac.il/databases/info/mega.sof	美国宾西法尼亚州立大学 MasatoshiNei 开发的分子进化遗传学软件
MOLPHY	ftp://ftp sunmh.ism.ac.jp/pub/molphy	日本国立统计数理研究所开发，最大似然法构树
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html	英国 University college London 开发，最大似然法构树和分子进化模型
PUZZLE	ftp://fx.zi.biologie.uni-muenchen.de/pub/puzzle	应用 quarter puzzling 方法(一种最大简约法)构建系统树
TreeView	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html	英国 University of Glasgow 开发
phylogeny	http://www.ebi.ac.uk/biocat/phylogeny.html	欧洲生物信息研究所(EBI)的系统发育分析软件

第二节 距离矩阵法

系统树可建立在(遗传)距离矩阵的基础上。这里的遗传距离为所有成对实用分类单位(operational taxonomic units，OTU)之间的距离。对于 t 个 OUT，每一对之间的距离矩阵列于表 5.2。

表 5.2 t 个实用分类单位（OUT）间的距离矩阵

		OUT 数				
		1	2	3	…	t
OUT 数	1	-	d_{12}	d_{13}	…	d_{1t}
	2	d_{21}	-	d_{23}	…	d_{2t}
	3	d_{31}	d_{32}	-	…	d_{3t}
	…	…	…	…	…	…
	t	d_{t1}	d_{t2}	d_{t3}	…	-

用这些距离对 OUT 进行表型意义的分类可借助于聚类分析(clustering)，聚类过程可以看作是鉴别具有相近 OUT 类群的过程。

一．平均连接聚类法(UPGMA 法)

可以采用几种聚类方法，这些方法包括序贯法(sequential)．聚合法(agglomerative)．分层法(hierarchical)和非重叠法(nonoverlapping)等。应用最广泛的是平均连接聚类法(average linkage clustering)或称为 UPGMA 法

(应用算术平均数的非加权成组配对法，unweighted pair-group method using an arithmetic average)。该法将类间距离定义为两个类的成员所有成对距离的平均值。

作为实例，我们考虑图 5.4 所列的线粒体 DNA 序列的资料。每对序列间的 Jukes-Cantor 距离取决于每对序列间差异核苷酸的观察数。如果在两条序列中相同碱基的比例为 q，则距离 K 可估计为

$$\tilde{K} = \frac{3}{4}\ln(\frac{3}{4q-1})$$

序列的差异和距离列于表 5.3

1．人类 GTAAATATAG TTTAACCAAA ACATCAGATT GTGAATCTGA CAACAGAGGC TTACGACCCC TTATTTACC
2．黑猩猩 GTAAATATAG TTTAACCAAA ACATCAGATT GTGAATCTGA CAACAGAGGC TCACGACCCC TTATTTACC
3．大猩猩 GTAAATATAG TTTAACCAAA ACATCAGATT GTGAATCTGA TAACAGAGGC TCACAACCCC TTATTTACC
4．猩猩 GTAAATATAG TTTAACCAAA ACATTAGATT GTGAATCTAA TAATAGGGCC CCACAACCCC TTATTTACC
5 长臂猿 GTAAACATAG TTTAATCAAA ACATTAGATT GTGAATCTAA CAATAGAGGC TCGAAACCTC TTGCTTACC

图 5.4 五种生物线粒体 DNA 序列

最近的距离是人类和黑猩猩之间的，将它们合并为一个类。其它序列与这个新类之间的距离就是该序列到新类各成员间的平均距离：

$$d_{(hu-ch),go} = \frac{1}{2}(d_{hu,go} + d_{ch,go}) = 0.037$$

$$d_{(hu-ch),or} = \frac{1}{2}(d_{hu,or} + d_{ch,or}) = 0.135$$

$$d_{(hu-ch),gi} = \frac{1}{2}(d_{hu,gi} + d_{ch,gi}) = 0.189$$

表 5.3 图 5.4 中 5 个线粒体序列的差异核苷酸数(对角线下)和 Jukes-Cantor 距离(对角线上)

	人类(hu)	黑猩猩(ch)	大猩猩(go)	猩猩(or)	长臂猿(gi)
人类(hu)	-	0.015	0.045	0.143	0.198
黑猩猩(ch)	1	-	0.030	0.126	0.179
大猩猩(go)	3	2	-	0.092	0.179
猩猩(or)	9	8	6	-	0.179
长臂猿(gi)	12	11	11	11	-

图 5.4
距离矩阵可简缩为：

	(hu-ch)	go	or	gi
hu-ch		0.037	0.135	0.189
go				0.179
or				0.179
gi				

其中人类－黑猩猩（hu-ch）与大猩猩（go）之间的距离最小。将它们合并为一类。新距离为：

$$d_{(hu-ch-go),or} = \frac{1}{3}(d_{hu,or} + d_{ch,or} + d_{gp.pr}) = 0.121$$

$$d_{(hu-ch-go),gi} = \frac{1}{3}(d_{hu,gi} + d_{ch,gi} + d_{go,gi}) = 0.185$$

下一个简缩后的距离矩阵为：

	(hu-ch-go)	or	gi
(hu-ch-go)		0.121	0.185
or			0.179
gi			

现在人类－黑猩猩－大猩猩（hu-ch-go）和猩猩（or）之间的距离最小，将其并为一类，从该四合体到猩猩序列的距离为：

$$d_{(du-ch-go-or),gi} = \frac{1}{4}(d_{hu,gi+dch,gi} + d_{go,gi} + d_{or,gi}) = 0.183$$

上述聚类结果可表示为图 5.5 所示的树状图。在构建树状图时，分枝点安置在两个序列或类的中点。图中成对序列间的距离为分枝长度之和。

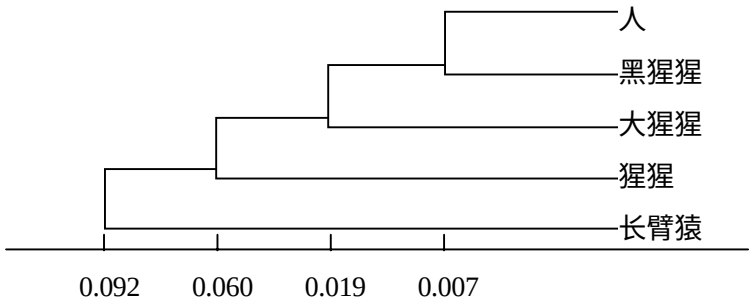


图 5.5 平均连接聚类法系统树

UPGMA 方法广泛用于距离矩阵。Nei 等（1983）模拟了构建树的不同方法，发现当沿树上所有分枝的突变率相同时， UPGMA 法一般能够得到较好的结果。但必须强调有关突变率相等(或几乎相等)的假设对于 UPGMA 的应用是重要的。另一些模型研究(如 Kim 和 Burgman，1988)已证实当各分枝的突变率不相等时，这一方法的结果不尽人意。当各分枝突变率相等时，认为分子钟(molecular clock)在起作用。

二．Fitch-Margoliash 算法

UPGMA 法包含这样的假定：沿着树的所有分枝突变率为常数。Fitch 和 Margoliash(1967)所发展的方法去除了这一假定。该法的应用过程包括插入“丧失的”OUT 作为后面 OUT 的共同祖先，并每次使分枝长度拟合于 3 个 OTU 组。现在用图 5.4 的线粒体资料来说明 Fitch-Margoliash 法则。

将 OUT 分为三组：距离最近的一对为 A=人类(hu)和 B=黑猩猩(ch)，剩下 X=(大猩猩 go，猩猩 or，长臂猿 gi)。引入树节 C 作为 A 和 B 的直接祖先。设从 C 到 A、B 的长度为 a、b，从 C 到 X 的为 x(图 8.4)。A、B、C 之间的 3 个成对距离提供了可解 3 个未知数的 3 个方程：

$$\begin{cases} a + x = d_{AX} = d_{AB} = \dfrac{1}{3}(0.045 + 0.143 + 0.198) = 0.129 \\ b + x = d_{BX} = d_{B\bar{A}} = \dfrac{1}{3}(0.030 + 0.126 + 0.179) = 0.112 \\ a + b = d_{AB} = 0.015 \end{cases}$$

设定如下符号约定：设d_{UV}为节点U到节点V的距离，$d_{U\bar{V}}$为节点U到V外所有节点的平均距离，d_{U^*V}为U以下所有末端节到V的平均距离。U^*表示从同一字母的节点U下的一组末端树节。

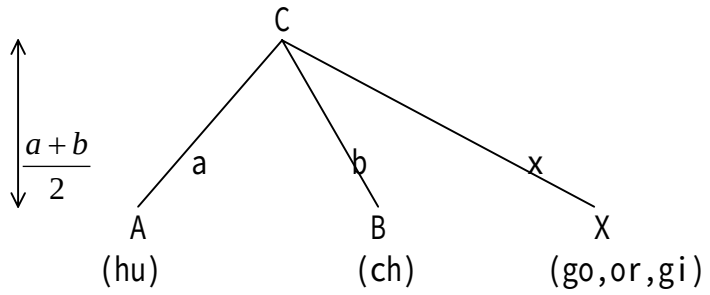


图 5.6 将 Fitch-Margoliash 算法应用于图 5.4 线粒体资料的初始步骤

第一个方程采用了从 A 到 X 的每一成员的平均距离。解以上三个方程得：

a=0.016，b=-0.001

为了方便以见，负的值定为 0，因此 b=0。a、b 的平均值为树节 C 的高度，该值为 0.008。

用C代替A、B，按UPGMA所采用的方式再计算距离值，得到下一个最近的一对为C和D(=go)。引入树节E作为C和D的直接祖先。如图 5.7 所示，节点C^*和E、D和E，E和X的分枝长度分别为c、d和x。现在X只包含猩猩(or)和长臂猿(gi)。要解的 3 个方程为：

$$\begin{cases} c+d=d_{C^*D}=\frac{1}{2}(0.045+0.030)=0.037 \\ c+x=d_{c^*x}=d_{(AB)^*}=\frac{1}{4}(0.143+0.198+0.126+0.179)=0.162 \\ d+x=d_{DX}=\frac{1}{2}(0.092+0.179)=0.136 \end{cases}$$

因此

c=0.032， b=0.006

节点E的高度为(c+d)/2=0.019。由于c度量了C到E距离以及从A和B到C的平均距离，所以c减去树节C的高度就得到C到E之间的分枝长度c'。换言之

c'=0.032-0.008=0.024

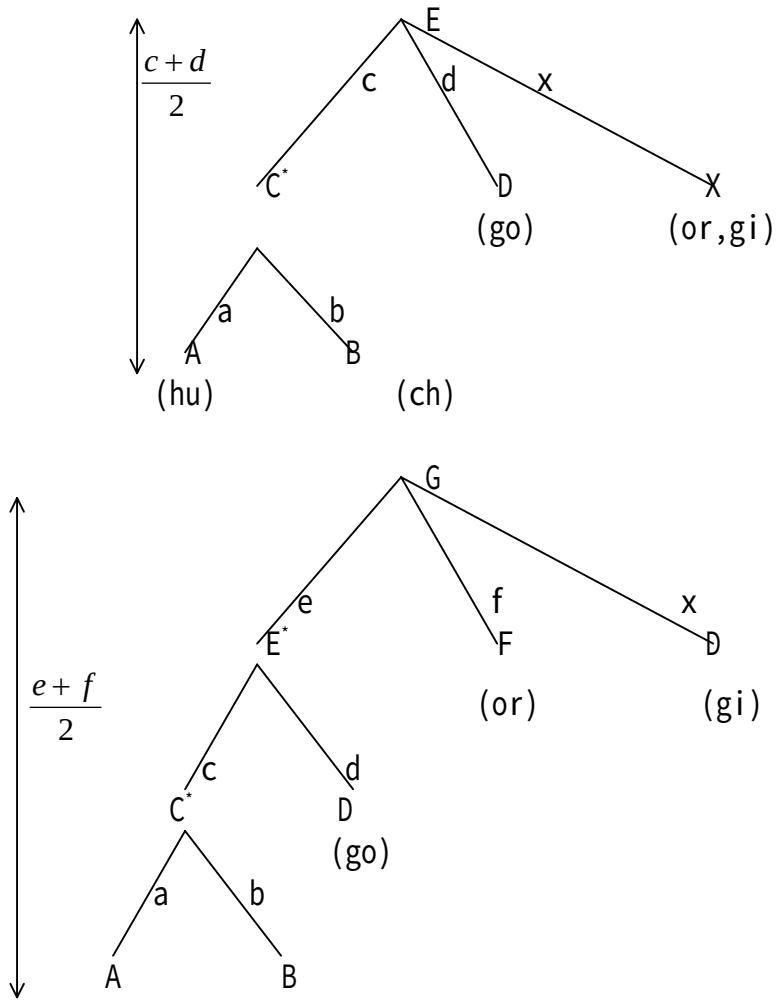


图 5.7 将 Fitch-Margoliash 算法应用于图 5.4 线粒体序列资料时的中间步骤

随着 OUT 简缩到 E、猩猩(or)和长臂猿(gi)。距离最近的一对就是 E 和 F(=or)了。引入 G 作为直接祖先，余下的 X=gi。要得到分枝长度所要解的方程为

$$\begin{cases} e+f = d_{E^*F} = \frac{1}{3}(0.143+0.126+0.092) = 0.121 \\ e+x = d_{E^*X} = \frac{1}{3}(0.198+0.179+0.179) = 0.185 \\ f+x = d_{FX} = 0.179 \end{cases}$$

故

e=0.063, f=0.057

节点G的高度为(e+f)/2=0.060，从E到G的分枝长度e' 为e与E的高度之差，即0.063-0.019=0.044。

Fitch-Margoliash 算法计算过程可以到此为止，图 5.8 给出了其无根系统树。

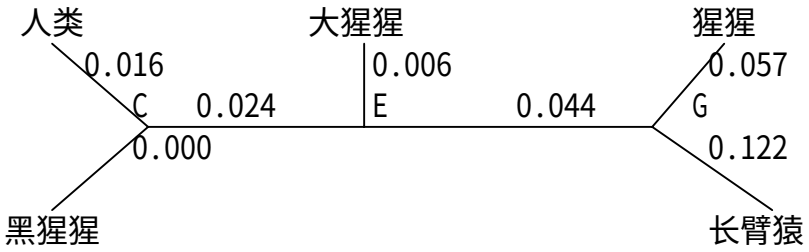


图 5.8 图 5.4 所列线粒体序列资料的 Fitch-Margoliash 无根系统树

如果不假定沿所有分枝具有相同的变更率,则由 Fitch-Margoliash 算法只能得到无根系统树。如果设置树根 I，并假定从 I 到现在所有序列的两个分枝具有相等的变更率，因而从 G 到 I 的距离 g 与从 H 到 I 的距离 h 是相等的，则有根树就可以采用与 UPGMA 提供的相同拓扑方法来获得。由于

$$g + h = d_{G^*H}$$

$$= \frac{1}{4}(0.198 + 0.179 + 0.179 + 0.179) = 0.184$$

所以g=h=0.092,且从G到I的距离g'为g减去G的高度，即 0.032。将所有这些分枝长度一起考虑便得到图 5.9 所示有根系统树。

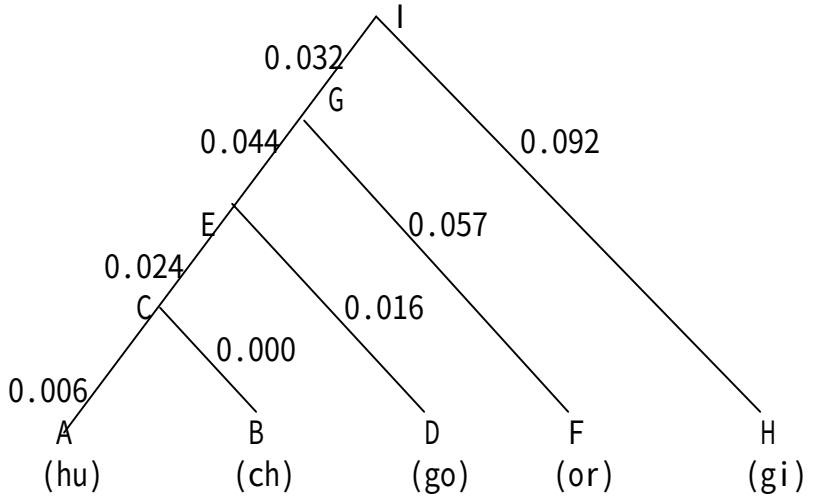


图 5.9 图 5.4 所列线粒体序列资料的 Fitch-Margoliash 有根树状图

Fitch和Margoliash承认他们的法则所得到的拓扑结构可能是不正确的，并建议考查其它的拓扑结构。可以采用Fitch和Margoliash(1967)称之为“百分标准差”的一种拟合优度来比较不同的系统树，最佳系统树应具有最小的百分标准差。如果d_{ij}为n个OUT中i和j的观测距离(即Judes-Cantor距离)，e_{ij}为i和j之间分枝长度之和，则

$$s = \left\{ \frac{\sum [(d_{ij} - e_{ij}) / d_{ij}]^2}{n(n-1)} \right\}^{\frac{1}{2}} \times 100 \qquad (5.5)$$

为百分标准差。考虑到可加性的假定，因而有任意两个节点之间的距离就是它们之间分枝长度之和。对于图 5.7 的系统树，观测距离和分枝长度列于表 5.4，其百分标准差为 1.94。通过调整适合系统树的分枝长度来降低 s 是可能的。

根据百分标准差选择系统树，其最佳系统树可能与由 Fitch-Margoliash 法则所得的不相同。当存在分子钟时，可以预期这一标准差的应用将给出类似于 UPGMA 方法的结果。如果不存在分子钟，因而在不同的世系(分枝)中的变更率是不同的，则 Fitch-Margoliash 标准就会比 UPGMA 好得多。

表 5.4 图 5.4 中 5 种线粒体序列的观测距离(对角线上)和采用 Fitch-Margoliash 算法计算所得距离(对角线下)

	人类	黑猩猩	大猩猩	猩猩	长臂猿
人类	—	0.015	0.045	0.143	0.198
黑猩猩	0.016	—	0.030	0.126	0.179
大猩猩	0.046	0.030	—	0.092	0.179
猩猩	0.141	0.125	0.107	—	0.179
长臂猿	0.208	0.192	0.174	0.181	—

通过选择不同的 OUT 作为初始配对单位，就可以选择其它的系统树进行考查。具有最低百分标准差的系统树即被认为是最佳的，并且这个标准是建立在应用 Fitch-Margoliash 算法的基础上的。例如，首先将人类和大猩猩分为一类，然后依次将黑猩猩、猩猩和长臂猿增加进去。但是，在这种情况下，第二个内部节点 E 的高度低于第一个内部节点 C 的高度，观测距离和计算距离之间的适合度就不如第一种情形那么好。

三．邻接法

邻接法(Neighbor-joining Method)由 Saitou 和 Nei(1987)提出。该方法通过确定距离最近(或相邻)的成对分类单位来使系统树的总距离达到最小。相邻是指两个分类单位在某一无根分叉树中仅通过一个节点(node)相连。图 5.2 中，人与黑猩猩是相邻的，人与大猩猩则不是；如果人与黑猩猩组成一个新类，则该新类与大猩猩又成为相邻。总之，通过循序地将相邻点合并成新的点，就可以建立一个相应的拓扑树。

邻接法的一般步骤：

①计算第i终端节点(即分类单位i)的净分歧度r_i

$$r_i = \sum_{k=1}^{N} d_{ik} \qquad (5.6)$$

其中N为终端节点数，d_{ik}为节点i和节点k之间的距离，有$d_{ik}=d_{ki}$

②计算并确定最小速率校正距离(rate-corrected distance)M_{ij}：

$$M_{ij} = d_{ij} - \frac{r_i + r_j}{N-2} \qquad (5.7)$$

③定义一个新节点 u，u 节点由节点 i 和 j 组合而成。节点 u 与节点 i 和 j 的距离为：

$$S_{iu} = \frac{d_{ij}}{2} + \frac{r_i + r_j}{2(N-2)}$$

$$S_{ju} = d_{ij} - S_{iu} \qquad (5.8)$$

节点 u 与系统树其它节点 k 的距离为：

$$d_{ku} = \frac{d_{ik} + d_{jk} - d_{ij}}{2} \qquad (5.9)$$

④从距离矩阵中删除列节点 i 和 j 的距离，N 值(总节点数)减去 1

⑤如果尚余 2 个以上终端节点，返回到步骤①继续计算，直至系统树完全建成。

以上每一步可以产生一个中间节点，并最终画出系统树。图中各分枝的角度是随意的。

现还以表 5.3 线粒体序列为例说明以上计算过程。表 5.5 列出了各步计算的结果，其中最小M_{ij}值用星号注明。第一步，星号(or)和长臂猿(gi)之间的M_{ij}值最小，则它们用节点 1 取代，进入第 2 步，则新节点(节点 1)到这二个节点的距离为：

$$d_{or,\text{节点1}} = \frac{1}{2} d_{or,gi} + \frac{r_{or} - r_{gi}}{6} = 0.057$$

$$d_{gi,\text{节点1}} = d_{or,gi} - d_{or,\text{节点1}} = 0.122$$

节点 1 到其它各节点的距离见表 5.5 第二步矩阵。在该矩阵中，人(hu)和黑猩猩(ch)的M_{ij}值最小，则它们又形成一个新节点(节点 2)……依次类推，便可最终完成矩阵的计算和邻接法无根系统树。

表 5.5 邻接法计算线粒体序列(图 5.4)的距离d_{ij}(上对角线部分)和M_{ij}(下对角线部分)

		hu j=1	ch j=2	go j=3	or j=4	gi j=5	净分歧度 r_i
hu	i=1	0.000	0.015	0.045	0.143	0.198	0.401
ch	i=2	-0.235	0.000	0.030	0.126	0.179	0.350
go	i=3	-0.204	-0.202	0.000	0.092	0.179	0.346
or	i=4	-0.171	-0.171	-0.203	0.000	0.179	0.540
gi	i=5	-0.181	-0.183	-0.181	-0.246	0.000	0.735

		hu j=1	ch j=2	go j=3	节点 1 j=4	r_i
hu	i=1	0.000	0.015	0.045	0.081	0.141
ch	i=2	-0.110	0.000	0.030	0.063	0.108
go	i=3	-0.086	-0.084	0.000	0.046	0.121
节点 1	i=4	-0.085	-0.086	-0.110	0.000	0.190

		go j=1	节点 1 j=2	节点 2 j=3	r_i
go	i=1	0.000	0.046	0.030	0.076
节点 1	i=2	-0.141	0.000	0.065	0.111
节点 2	i=3	-0.141	-0.141	0.000	0.095

		go j=1	节点 3 j=2
go	i=1	0.000	0.005
节点 3	i=2		0.000

*hu、ch、go、or 和 gi 分别代表人、黑猩猩、大猩猩、猩猩和长臂猿

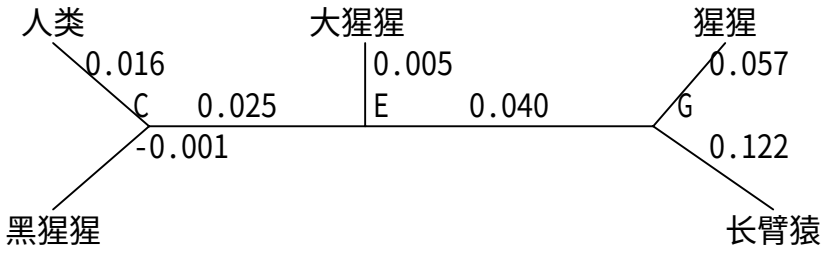


图 5.10 根据线粒体序列(图 5.4)构建邻接法无根系统树

第三节 简约法

简约法(Parsimony)明显注重每一物种观测的特征值，而不是概括特征值之间差异的序列间距离。该法由 Edwards 和 Cavalli-Sforza(1963)以“最小进化原理”的名称应用于基因频率资料。如果有一组物种的序列可供利用，那么连接它们的最为简约的拓扑结构就可能得到。但一般无法获得分枝长度。

对于每种可能的拓扑结构，每一节点的序列就是产生两个直接后裔序列所需变更最小的序列。然后可以找到整个系统树所需的变更总数，具有最小总数的系统树就是最简约的。为说明这一方法，我们讨论 Fitch(1971)所给的例子。有 6 个物种 A～F 的序列可以利用，并且在某一特定位置，它们分别具有碱基 C、T、G、T、A、A。存在许多可能的拓扑结构，其中之一如图 5.11 所示。从离现

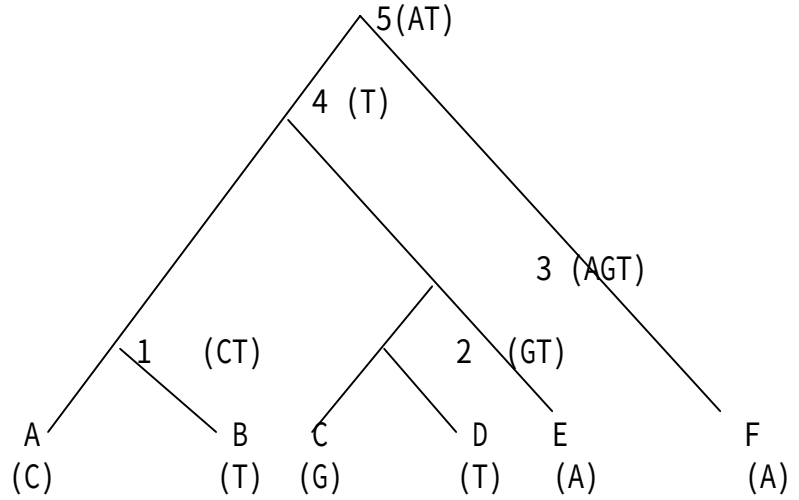


图 5.11 在 6 条序列的一个位点上寻找最简约树的过程

存序列最近的节点开始，依次考虑节点 1～5 中的每一个。在每一节点，写出两后裔序列的“简约式”。这一计算(这里记为◇)是一个集运算，如果交集不是空的，则定义此运算为两个集的交；如果交集是空的，则定义为两个集的并。对于不同的集(序列)X、Y、Z，并和交的集合运算可以与简约运算对比如下：

交 [X、Y]∩[X、Z] = [X] [X]∩[Y]=φ

并 [X、Y]∪[X、Z] = [X] [X]∪[Y]=[X，Y]

约减 [X、Y]◇[X，Z]＝[X] [X]◇[Y]=[X，Y]

如果两个序列在某位置具有相同碱基，则当它们的共同祖先也具有该碱基时就产生最小的变更数。如果它们具有不同的碱基，最小变更数则要求它们的祖先具有这两个碱基的其中之一。在图 5.11 中，节点 1 和 2 分别为(CT)和(GT)，意味着所列两个碱基之一将给出最小的变更数。对于节点 3 有 3 种可能性，但对于节点 4 只有 1 种可能性，节点 5 有 2 种可能性。如果节点 1～5 都具有碱基 T，则这一拓扑方法所得最小变更数为 4。但正如 Nei(1987)指出的，如果每节都有碱基 A，则产生相同的最小变更数。同时存在另外 9 种产生最小变更数的可能性，即 5 个节点具有碱基 TTTTA、TAAAA、CAAAA、AGAAA、ATAAA、CGAAA、CTAAA、TTAAA 或 TGAAA 之一者。

重复进行上述过程得到其它的拓扑结构，需要最小变更数的拓扑结构可看成为最后的系统树。对于最大化的简约，只需考虑那些信息位点(Informative

site)。对于 DNA 序列，信息位点是指那些至少存在 2 个不同的碱基且每个不同碱基至少出现两次的位点。只有一个碱基且只在一个序列中出现的位点不属于信息位点，因为那种独特的碱基位点是由于在直接通向它所在序列的分枝上发生单个碱基变更所引起的。这种碱基变更可与任何拓扑结构相容。以表 5.6 为例，只有位点 5、7、9 为信息位点。

表 5.6 信息位点列举(以 4 条序列共 9 个位点为例)

序列	位点								
	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	A
3	A	G	A	T	A	T	C	C	G
4	A	G	A	G	A	T	C	C	G

对于图 5.4 中的线粒体序列，存在 5 个信息位点：25、39、44、47、54。图 5.12 显示了根据这 5 个位点所得到的简约系统树。象构建其它可能系统树那样，它有 6 个碱基变更。尽管获得了与距离矩阵法找到的系统树相同的拓扑结构，但非常有限的资料已产生了某些惊人的效果。图 5.8 中节点 E 的 G 之间的分枝短于节点 G 的 F 之间的，而在信息位点间，前一分枝上有 3 个碱基变更，而在后一分枝上未发生碱基变更。

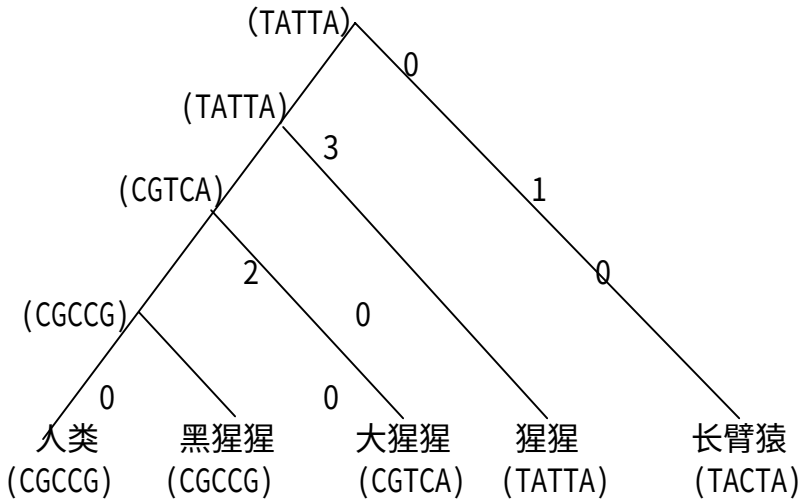


图 5.12 图 5.4 线粒体序列资料的最简约系统树
（数字为节点间的碱基变更数）

Felsenstein(1983)已批评了约减法，因为该法不是以统计原理为基础。Felsenstein 指出，在试图使进化事件的次数最小时，简约法隐含地假定这类事件是不可能的。如果在进化时间范围内碱基变更的量较小，则简约法是很合理的，但对于存在大量变更的情形，随着所用资料的增加，简约法可能给出实际上更为错误的系统树(Felsenstein,1978)。

第四节 似然法

一．DNA 序列的似然模型

构建系统树的似然法试图避免其它方法的局限性，尽管它需要的计算量大得惊人。与距离矩阵法不同，似然法试图充分有效地利用所有资料而不是将资料简缩为距离的集合。它们与简约法的不同之处在于其进化概率模型采用了标准的统计方法(Felsenstein，1981)。

当考虑实施最大似然法时，该方法先假定系统树的形式，然后选择分枝长度以使产生特定系统树的资料的似然值最大化。通过比较不同系统树的似然函数值，将具有最大似然值的系统树看作最佳估计。一个直接的问题是随着OUT的增加，系统树的数目迅速增加。当树端具有n个OUT时，无根分歧树(在每一内部树节上连接着两个分枝的树)的数目为$(2n-5)!/[(n-3)!2^{n-3}]$。当n=3、4、6、8 和 10 时，该数分别为 1、3、105、10395、2027025。具有n个树端的有根树数目与具有n+1 个树端的无根树数目相同(Felsenstein，1978)。实际应用时，只研究所有系统树的一个亚集。

对于DNA序列资料，似然法依据的模型规定了在特定时间内由于突变使一个序列变更为另一序列的概率。尽管DNA序列中的毗邻碱基不是独立的，但是模型的确假定了不同位点上进化的独立性，从而某系统树上一组序列的概率就是序列上每一位点概率的乘积。在任何单一位点，在经过时间T后，碱基i将变更为碱基j的概率为$P_{ij}(T)$。设定对于碱基A、C、G、T，下标i、j的值为 1、2、3、4。

最为简单的碱基替换突变模型假定突变率为常数。当碱基突变时，它以常数π_i的突变率变更为i型碱基。这包括了一个碱基突变为与之相同的类型，尽管这种类型的替代是观察不到的。当单位时间(世代)的碱基替换率为u时，则经过T世代后某一位点不发生突变的概率为$(1-u)^T$，因此突变概率p为：

$$P = 1-(1-u)^T \approx 1-e^{uT} \qquad (5.10)$$

经过时间 T 后由碱基 i 变更为碱基 j 的概率可写为(Felsenstein，1981)：

$$P_{ii}(T) = (1-p) + p\pi_i$$

$$P_{ij}(T) = p\pi_j \quad , \qquad (\ j \neq i\ ) \qquad (5.11)$$

当设定所有π_i均为 1/4 时，这就是Jukes-Cantor突变模型，但有关突变率的解释略有不同。本模型中突变率u是对所有碱基替换而言，且u等于 4/3 乘以Jukes-Cantor模型中的可检测替换率μ。

注意到概率只涉及突变率和时间的乘积，采用这里讨论的方法无法对二者作分别估计。因此，我们只讨论乘积$v = uT$，即沿系统树分枝碱基替换的期望数。如果树的所有分枝以相同的速率发生碱基替换，则分枝长度将显示出树上每对树节间的相对时间。

似然法假定了系统树的结构。现存的序列形成系统树的树端，而其它树节的序列均不知道。有关系统树资料的似然值必须考虑这些未知序列的所有可能性。

在这里所描述的一个参数突变模型下，预期 4 种碱基变具有相等频率，结果对于i=1、2、3、4，π_i设定为 0.25。另一可能的方式是利用从构建系统树的序列得到的碱基平均突变率。

二．两条序列系统树

具有两个序列的一个有根系统树如图 5.13 所示。对于这个序列的第j个核苷酸位置，观测到的碱基为S_1、S_2。设在未知祖先序列中该位点碱基为k。将所有可能为k碱基的概率相加，则该位点似然值$L(j)$为：

$$L(j)=\sum_{k=1}^{4}\pi_k P_{ks_1}(v_1)P_{ks_2}(v_2) \qquad (5.12)$$

对于所有 m 个位点，似然值为：

$$L=\prod_{j=1}^{m}L(j) \qquad (5.13)$$

该似然值是两个未知分枝长度v_1、v_2的函数。

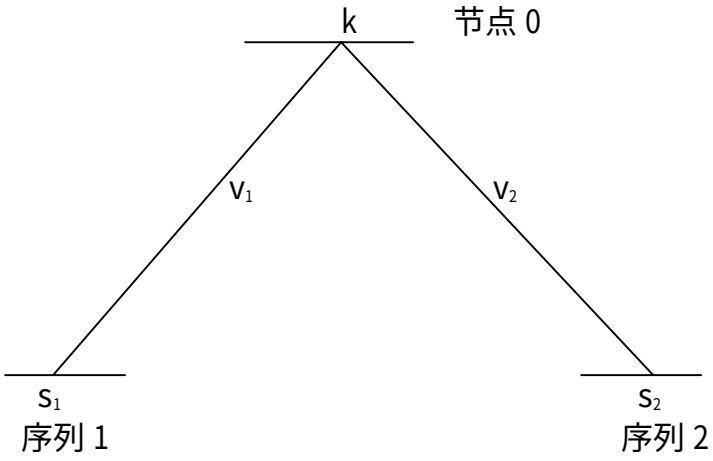


图 5.13 两个序列的有根树状图
（在j位点，两个序列具有碱基s_1和s_2和相应节点具有碱基k）

由于只存在一组从序列 1 到序列 2 的可观测的转换，因而内部节点 0 不能唯一定位。可以从 Felsenstein(1981)的“滑轮原理”来证实这一点。例如，在 j 位点序列 1 具有碱基 A，序列 2 具有碱基 C，考虑用似然函数显示该位点内部节点的 4 种碱基之和：

$$\begin{aligned}L(j)&=\pi_A P_{AA}(v_1)P_{AC}(v_2)+\pi_c P_{cA}(v_1)P_{CC}(v_2)\\&\quad+\pi_G P_{GA}(v_1)P_{GC}(v_2)+\pi_T P_{TA}(v_1)P_{TC}(v_2)\\&=\pi_A[(1-p_1)+p_1\pi_A]p_2\pi_C+\pi_C p_1\pi_A[(1-p_2)+p_2\pi_C]\\&\quad+\pi_G p_1\pi_A p_2\pi_C+\pi_T p_1\pi_A p_2\pi_C\\&=\pi_A(p_1+p_2-p_1p_2)\pi_C\\&=\pi_A p_{12}\pi_C\end{aligned} \qquad (5.14)$$

换言之，涉及突变概率为p_1和p_2的两条通径(由k到A和由k到C)的似然值，与涉及概率为p_{12}的一条通径(A到C)的似然值相同。注意到

$$p_{12} = p_1 + p_2 - p_1 p_2 = 1 - e^{-(v_1+v_2)} \qquad (5.15)$$

因而图 5.13 系统树的似然值只取决于两个物种 1 和 2 间总的分枝长度(v_1+v_2)，而与节点 0 的位置无关。不可能分别估计v_1和v_2，因而系统树简缩成两个序列间的单个分枝。换言之，可估计得到的系统树是无根的。

当 4 种碱基的概率相等时，即$\pi_i=1/4$(i=1、2、3、4)，则该一分枝系统树的似然值简缩为：

$$L = \left(\frac{4-3p}{64}\right)^s \left(\frac{p}{64}\right)^{m-s} \qquad (5.16)$$

其中 p 是该分枝的突变概率，且两个序列的 m 个位点中有 s 个具有相同的碱基。将似然值最大化，得到

$$\hat{p} = \frac{4(m-s)}{3m} \qquad (5.17)$$

分枝长度的最大似然估计值为

$$\hat{v} = \ln\left(\frac{3}{4\tilde{q}-1}\right) \qquad (5.18)$$

其中

$$\tilde{q} = \frac{s}{m}$$

回顾一下，u 与 Jukes-Cantor 模型中的 $4\mu/3$ 相对应，且两序列间的时间 T 在那个模型中写作 $2t$(从每一序列到祖先序列的时间的两倍)。这些关系表明，分枝长度也可以从两个序列间的 Jukes-Cantor 距离 K 得到：

$$v = uT = \ln\left(\frac{3}{4q-1}\right)$$

$$K = 2\mu t = \frac{3}{4}\ln\left(\frac{3}{4q-1}\right) \qquad (5.19)$$

长度 v 是所有碱基替换的期望数，而长度 K 是指可检测到的替换，且 $v=4k/3$。

三．三条及多条序列系统树

对于三个序列则存在三种有根系统树形式，其中之一如图 5.14 所示。除了三个可观测的序列外，在节点 0 与 4 还有未定的序列，且有 4 个分枝长度有待确定。可依次考虑三种树状图，给出最大似然值的就是估计得到的系统树。但事实上，没有必要这样做，因为三种树状图具有相同的似然函数。

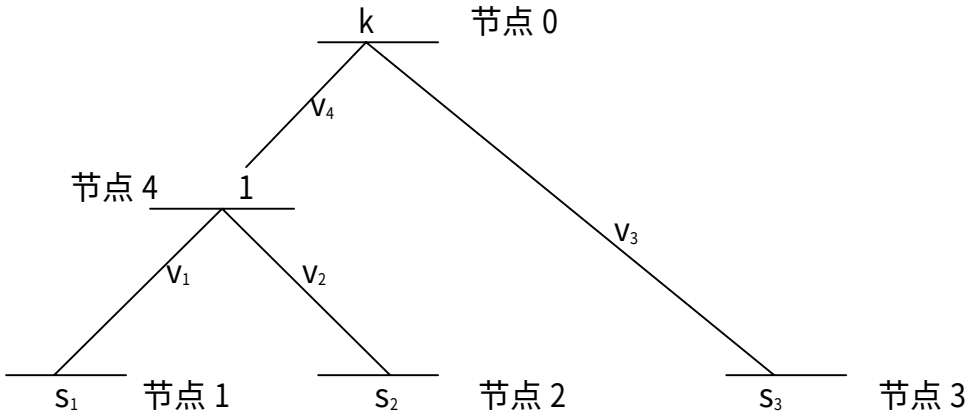


图 5.14 三个序列的一种有根系统树形式
（在位点j，三个序列具有碱基s_1、s_2、s_3，节点 0 和 4 具有碱基k和l）

对于图 5.14 所示的排列方式，位点 j 的似然值可以用节点 4 的碱基 l、节点 0 的碱基 k 表示如下：

$$L(j)=\sum_k\sum_l \pi_k P_{kl}(v_4)P_{ks_3}(v_3)P_{ls_1}(v_1)P_{ls_2}(v_2) \qquad (5.20)$$

如果节点 0 移动到节点 3 和 4 之间的任何位置，则Felsenstein滑轮原理的应用不会改变该似然值。似然值只取决于总距离v_3+v_4。如果使节点 0 和 4 叠合，则似然值可写作：

$$L(j)=\sum_k \pi_k P_{ks_1}(v_1)P_{ks_2}(v_2)P_{ks_3}(v_3) \qquad (5.21)$$

无法唯一地确定接点 0 的位置，且对于三个序列只有图 5.15 中星状系统树需要考虑。

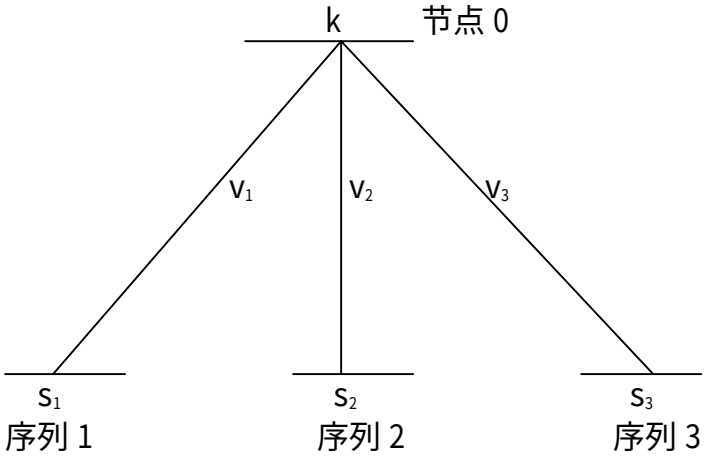


图 5.15 三个序列的星状系统树
(三个序列 1、2、3 来自于同一祖先序列 0)

在相等碱基频率的假定下，由于存在三个未知的分枝长度且有三个成对的 Jukes-Cantor 距离可供利用，所以利用 Bailey 法可从下列等式得到最大似然估

计：

$$\hat{v}_1 + \hat{v}_2 = K_{12}$$

$$\hat{v}_1 + \hat{v}_3 = K_{13}$$

$$\hat{v}_2 + \hat{v}_3 = K_{23}$$

估值为

$$\hat{v}_1 = \frac{1}{2}(K_{12} + K_{13} - K_{23})$$

$$\hat{v}_2 = \frac{1}{2}(K_{12} + K_{23} - K_{13})$$

$$\hat{v}_3 = \frac{1}{2}(K_{13} + K_{23} - K_{12})$$

实际序列并非具有相等的碱基频率，因而Jukes-Cantor距离不会使似然值最大，但它们的确为迭代法提供了很好的初始值。Newton-Raphson迭代法为找到最大似然值的数值解提供了直接的方法，且从寻求$p_i=1-e^{-v_i}$的估值来看，这一方法在描述上是最为简单的。

表 5.7 给出了图 5.4 中人类(1)、大猩猩(2)、长臂猿(3)线粒体序列收敛过程的例子。三个序列间的平均碱基频率用作模型中的概率项π_i。

表 5.7 图 5.4 中人类、大猩猩和长臂猿线粒体序列非约束型最大似然树分枝长度的连续迭代

迭代	v_1	v_2	v_3
初始值	0.0423	0.0174	0.2215
1	0.0420	0.0196	0.2230
2	0.0420	0.0199	0.2299
3	0.0420	0.0199	0.2299
标准差	0.0297	0.0218	0.0600

用几个序列作为树端来构建系统树时，可采用以上所述的一般方法。先指定一种系统树，然后对来自该系统树似然函数的方程进行 Newton-Raphson 迭代来估计分枝长度。在理论上，应研究所有可能的系统树来寻找具有最大似然值的系统树。Fukami 和 Tateno(1989)证实至多存在一组对于 L 给出平稳值的分枝长度，且这组分枝长度提供了所需的最大似然估计。将这一方法应用于图 5.4 所列的 5 种线粒体序列，获得了图 5.16 所示的无根树状图。

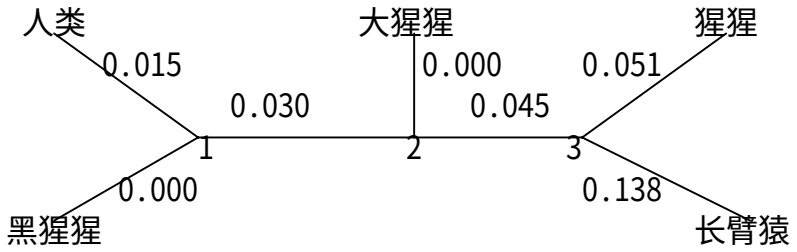


图 5.16 利用 Felsenstein 的 PHYLIP 软件构建的图 5.4 线粒体序列资料的最大似然树

四．对系统树 Bootstrap 抽样

在任一特定的树状拓扑结构内，已知最大似然值提供了分枝长度的一致估计值，这意味着随着资料量的增加，估计值逐渐接近真值。但是，与所有拓扑结构相比，具有最大似然值的系统树特性是怎样的？在何种意义上它可以认为估计了真实的系统树？尽管这是一个难以解决的理论问题，但在实际上可采用数值重复抽样来获得经验性的证据。

Felsenstein(1985)建议在所研究序列的各位点进行 Bootstrap 抽样。当序列长度为 m 时，Bootstrap 样本就包括从原始 m 个位点进行有返回抽样所得每一序列在 m 个位点的那些碱基。每一 Bootstrap 样本象原始资料一样进行相同的似然估计。对所有 Bootstrap 样本范围内应注意单源(monophyletic)物种的集合。如果发现一组物种它与 95%的 Bootstrap 系统树一起出现，则可以认为这组物种在 5%显著水平上是单源的。还有一个有用的概念，即由“多数规则”(majority rule)建立一致树(consensus tree)(Margush 和 McMorris，1981)，它由在 Bootstrap 样本所得的大多数系统树中出现的那些物种所组成。在系统发育分析中获得不同的系统树时，往往需要将这些系统树组合成一致树。