5 分子系统发育

分子系统发育是用分子生物学技术对生物间的进化关系进行的研究。它是分子进化中的一个领域，近十年来已引起了广泛关注，这主要是因为在许多情况下系统发育关系用任何别的方法都难以估价的缘故。本章的目的是，解释如何用分子数据来构建系统树，以及给出一些例子，以说明对一些存在已久的系统发育问题，分子方法能够提供比传统方法要清晰得多的解答。

5.1 分子数据对系统发育研究的影响

分子系统发育的研究发端于本世纪开始前后，甚至早于孟德尔定律被重新发现的１９００年。免疫化学研究表明，血清学交叉反应在亲缘关系较近的生物间比在关系较远的生物间强。这些发现的进化含义被纳托尔（Nutall,1904）用来推论各不同动物类群间的系统发育关系。例如，他断定人类最近缘的亲属是猿，其后，亲缘关系由近及远依次为：古世界猴，新世界猴和原猴。

自十九世纪六十年代后期以来，分子生物学中的各种技术都得到发展，从而开始了将分子数据广泛应用于系统发育的研究。特别是在十九世纪七十年代和十九世纪八十年代，用蛋白质顺序资料研究分子系统发育进展极快。花费不大但便利得多的方法，象蛋白质电泳，ＤＮＡ－ＤＮＡ杂交和免疫学方法，虽然不如蛋白质顺序测定来得精确，却也被广泛用于群体间或亲缘关系较近的物种间的系统发育关系的研究中。这些方法的应用也刺激了遗传距离的测定和系统树构建法的发展（见 Fitch 和 Margoliash,1967;Nei,1975;Felsenstein,1988)。

自十九世纪八十年代后期以来的ＤＮＡ顺序资料的积累，已经对分子系统发育产生了巨大的影响。ＤＮＡ顺序资料不仅更为丰富，而且比蛋白质顺序资料更容易分析。因此，它们一方面已被用于推断象人与猿那样亲缘关系很近的物种间的系统发育关系（见第 81 页），另一方面又被用来研究一些非常古老的进化事件，象线粒体和叶绿体的起源（见第 86 页），门和界的分化（Woese,1987）等。将来，ＤＮＡ定序看来将用于解决系统发育研究中许多长期得不到解决的问题，象细菌和单细胞真核生物间的进化关系（Sogin 等,1986,1989）,这用任何传统的进化研究方法都不可能得到解决。事实上,分子数据在进化史的研究方面被证明是极为有用的,藉此也许我们最终能完美地构建出生物界主要类群的系统发育树。　　当然，我们也不放弃进化研究的传统手段，象形态学、解剖学、生理学和古生物学等。因为，不同方法可提供互相补充的数据。我们注意到，分类学主要以形态学和解剖学资料为依据，而古生物学信息则是能提供用于进化研究的时间框架的唯一资料。

5.2 系统树

地球上的一切生命形式，不管是现存的还是已经灭绝了的，都有一个共同的起源，它们的祖先可以追溯到大约在４０亿年以前生存的一种或几种生物。因此，所有动物、植物、细菌通过祖藉而相互关联。亲缘关系近的生物是由一个较近代的共同祖先传下来的，亲缘关系远的生物则由较远古的共同祖先传下来。系统发育研究的目的是：（１）建立各生物间正确的系谱学联系，和（２）估计各生物自它们从最后一个共同祖先那里分歧以来的分歧的时间。

在系统发育研究中，一组生物类群间的进化关系常用系统树（phylogenetic tree）来图示说明。系统树是一种由节点和分枝组成的图象，其中任何两个邻近的节点都只由一个分枝来联结（图 5-1)。节点（nodes）代表分类学单位，而分枝（branches）则用祖藉和祖先来定义这些单位间的关系。一个树的分

枝模式被称为拓扑图（topology)。枝长（branch length） 通常代表在该分枝中曾发生过的变化数。由节点表示的分类学单位可以是物种、群体、个体或基因。

在处理系统树时，我们要分清外节点（external nodes）和内节点(internal nodes)。例如，图５－１ａ中，节点Ａ、Ｂ、Ｃ、Ｄ和Ｅ是外节点，其余的都是内节点。外节点代表处于比较中的现存分类学单位，并被命名为操作中的分类学单位（operational taxonomic units）即ＯＴＵ。内节点则代表祖先单位。

图５－１为处理一个系统树的两种普通方法的示意图。在图５－１ａ中，其分枝是无尺度的（unscaled）,它们的长度与已注明在分枝上的变化数不成比例。这种表示法使我们

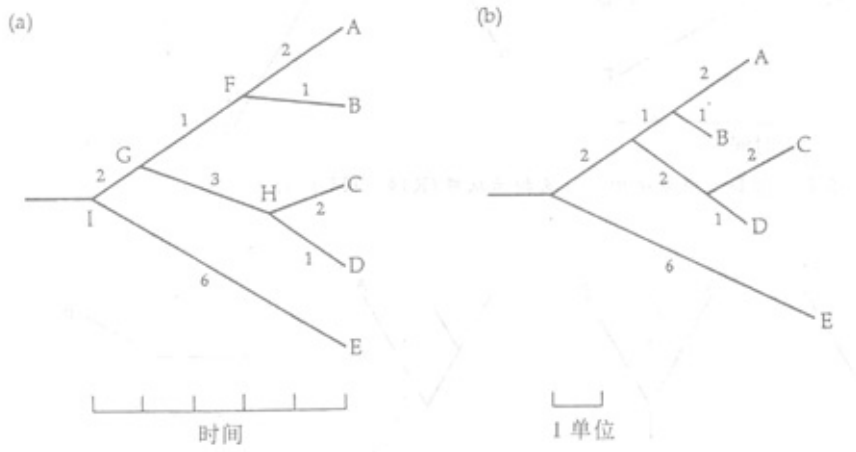


图５－１ 一个关于５个ＯＴＵ的系统树的两种可供采用的表示法。（ａ）无尺度的分枝：现存的ＯＴＵ排成一条线，节点位于与分歧时间成比例的地方。（ｂ）有尺度的分枝：分枝的长度与分子变化数成比例。

能将现存的ＯＴＵ排成一条直线，而且在分歧时间已知或已估出时，还可把代表分岐事件的节点按时间尺度来排列。在图５－１ｂ中，分岐是有尺度的（scaled）,且其长度正比于变化数。

如果任何两个ＯＴＵ间的距离等于将它们联起来的所有分枝的长度之和，则这样的树即被称为是加性的（additive)。例如，若可加性成立，则图５－１ａ中ＯＴＵＡ和Ｃ间的距离应等于 2+1+3+2=8。两个0TU 间的距离可直接从分子数据（例如ＤＮＡ顺序）算出，而枝长则可按某些规则从ＯＴＵ间距离估出（见第 77 页)。如果在任一核苷酸位点上曾发生过多重替换，则可加性一般不成立（图３－５)。

若一个节点只有两个直接的后代谱系，则它是两分叉的（bifurcating)，若它有两个以上的直接后代谱系则是多分叉的（multifurcating)。为简便起见,我们将只考虑象图 5-1 中那样的两分叉树。

有根树和无根树

系统树可以是有根的（rooted）也可以是无根的(unrooted)(图 5-2)。在有根树中，存在一个被称为根（图５－２ａ中的Ｒ）的特殊节点，由此导向任何别的节点都只有唯一图５－２ （ａ）有根系统树和无根系统树。箭头指示从根（Ｒ）到ＯＴＵＤ的唯一途径。途径。每一途径中的方向与进化时间相对应，而根则是所有正被研究的ＯＴＵ的共同祖先。无根树是一种只将各ＯＴＵ间的关系具体化而未定义进化途径的树（图５－２ｂ)。

对３个物种来说，存在着３种可能的有根树，但只有一种无根树（图５－３)。图５－３ 从３种ＯＴＵ可能会构建出３种不同的有根树（ａ)，但却只有一种无根树（ｂ)。对ｎ个ＯＴＵ而言，两分叉有根树的数目（N_R）由：

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!} \qquad (5.1)$$

给出，其中 n≥2。对 n≥3，两分叉无根树的数目(N_U)为：

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!} \qquad (5.2)$$

注意， ｎ个ＯＴＵ的可能无根树数等于（ｎ－１）个ＯＴＵ的可能有根树数。ＯＴＵ从２到１０的可能

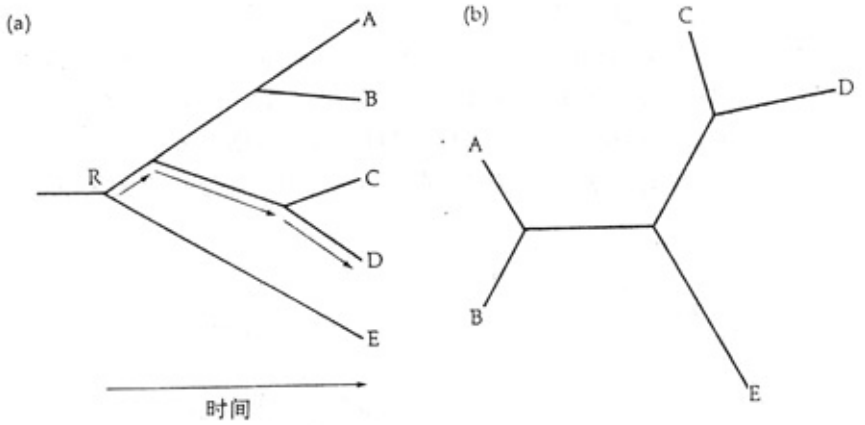
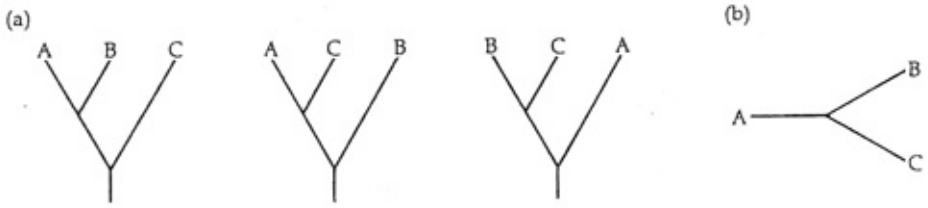


图5－2 (a)有根系统树和无根系统树 箭头指示从根(R)到OTUI)的唯一途径



有根树和无根树的数目列于表５－１。我们看到，随着ｎ的增加N_U和N_R都极快地增大，而到１０ＯＵＴ时已经有２００多万两分叉无根树和将近３５００万有根树了。由于这些树中只有一种能正确地表示这些ＯＴＵ间真实的进化关系，所以，当ｎ较大时，推论出真实的系统树通常是非常困难的。

表５－１ 对1－１０ ＯＴＵ可能的有根树和无根树的数目

OTU数	有根树数	无根树数
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10395	954
8	135135	10395
9	2027025	135135
10	34459425	2027025

自 Felsenstein(1987)

真实树和推测树

导致任何类群的ＯＴＵ形成的物种形成事件的秩序，在历史上是唯一的。所以，在用某一给定数的ＯＴＵ建立的所有可能的树中，只有一种能代表真实的进化历史。这样一种系统树称为真实树（true tree)。用某一组数据和某种构树法得到的树称推测树（inferred tree)。推测树可能与真实树等同，也可能与真实树不同。

基因树和物种树

表示一群物种的进化途径的系统树称物种树（species tree)。若系统树是根据来自各物种的一个基

因构成的，则该推测树即为基因树（gene tree）(Nei,1987)。它与物种树有两个方面不同。第一，从两不同物种取样的两基因的分岐可能在时间上早于两物种的分岐（图 5-4)。这将造成高估了枝长的结果，但如果我们处理的是长期进化，在这个过程中由物种内遗传多态性造成的分岐成分可以忽略的话，则还不至于出现严重问题。

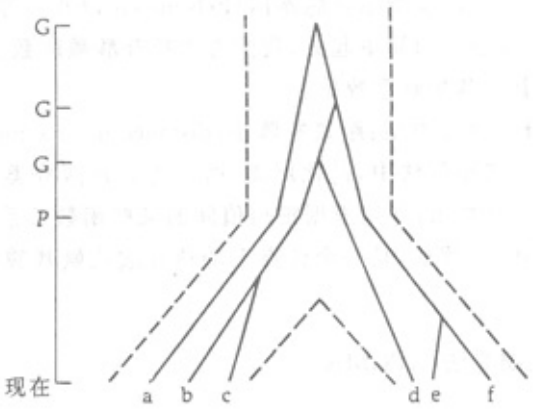


图 5 － 4 本图表示：若群体在时刻 P 是遗传上多态的，则基因分化（G）通常比群体分化(P)出现得早。结果形成用 a －f 表示的 6 个等位基因的基因进化史用实线表示，群体分岐则以虚线表示。自Ｎｅｉ（１９８７)。

基因树的第 2 个问题是，基因树的分枝模式（即它的拓扑图）可能与物种树的不同。图 5 — 5 展示了这两种树间的 3 种不同的可能关系。基因树（ａ）与（ｂ）的拓扑图和与其对应的物种树的图等同（例

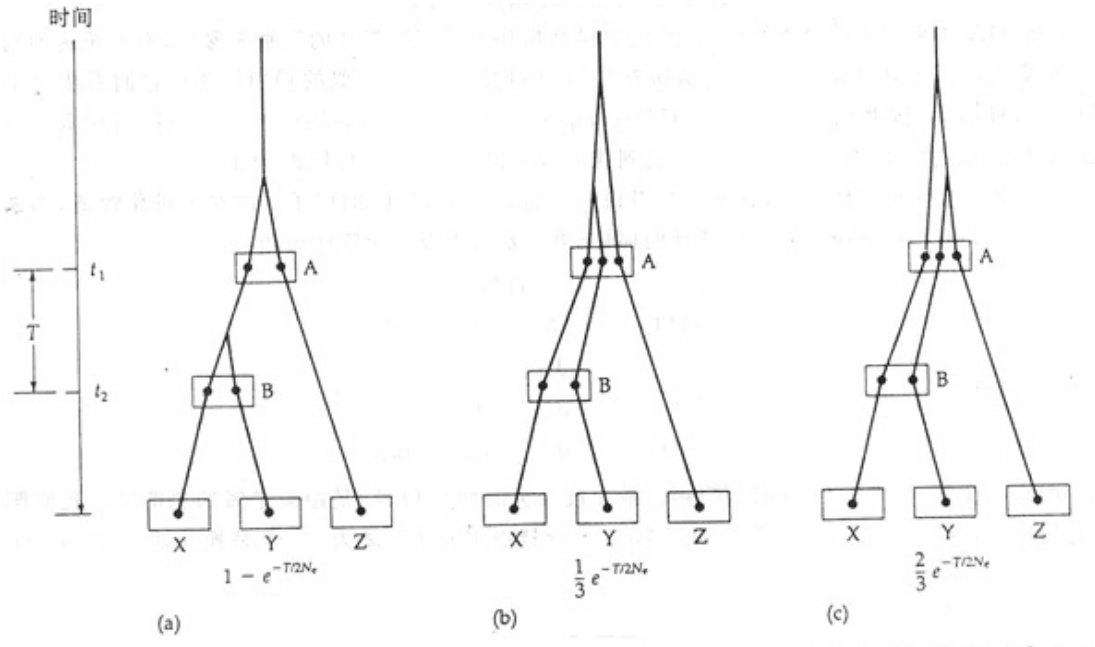


图 5 － 5 物种树（矩形）和基因树（圆点）间的 3 种可能关系。在（ａ）和（ｂ）中，物种树的拓扑图与基因树的等同。注意，（ａ）中基因的分岐时间大致等于群体的分岐时间。而（ｂ）中却不然，基因Ｘ和Ｙ的分岐时间大大早于各自所属的群体的分岐时间。在（ｃ）中，基因树的拓扑图与物种树的不同。对于中性等位基因，每种树出现的概率已在该树的下面列出。t_1 是第一次物种形成事件发生的时间，t_2 是第二次物种形成事件发生的时间。图中 $T = t_1 - t_2$，Ne 则是有效群体大小。自Ｎｅｉ(１９８７)。

如，Ｘ和Ｙ形成一簇)。然而基因树（ｃ）却与真实的物种树不同，因为此时Ｙ和Ｚ是姐妹类群。当第一次物种分化和第二次物种分化间的时间间隔（t_1-t_2）较短时，得到错误的树（ｃ）的概率将是相当高的，这种情况在确定人、黑猩猩和大猩猩间的系统发育关系时可能确实存在。为了避免得到这种错误类型，在构建系统树时就必须用许多基因。为了避免随机误差也需要大量数据。因为核苷酸替换是随机地发生的，所以有可能会产生随机误差。例如图 5 — 5 ａ中，谱系Ｚ累积的替换可能会因机遇而比谱系Ｘ和Ｙ的少，尽管它发生分枝的时间要更早一些。

5.3 系统树的构建方法

文献中已有许多构树方法被提出来。有关的详细处理读者可参考史尼斯和索卡尔（Sneath 和 Sokal,1970）、根井（Nei,1987） 以及费尔森施泰因（Felsenstein,1988）等的论述。这里我们讲 4 种在系统发育研究中常被用到的方法。为简单起见，我们考虑核苷酸顺序数据，但所讲到的方法同样也可用于其他类型的分子数据，象氨基酸顺序数据。

以下所讲的方法可分成两种类型：距离矩阵法（distance matrix methods） 和最节省法(maxmum parsimony methods)。在距离矩阵法中，进化距离（通常为分开两分类学单位的核苷酸或氨基酸替换数）按所有可能的分类单位对算出，并用根据距离值间的某些函数关系建立的算法来构建系统树。在最节省法中，应用了特性状态（例如，某一个位点上的核苷酸或氨基酸），而导致这些特性状态的最短途径即被选取为系统树。

不加权算术平均组对法（UPGMA）

不加权算术平均组对法（unweighted pair group method with arithmetic mean）(UPGMA)是一种简单的构树法。它最初本用来构建分类学表型关系图,即能反映出各 OTU 间的表型相似性的树(Sokal 和 Michener,1958;并见第 77 页)，但如果进化速率在不同谱系间近似地恒定，以至进化距离与分岐时间之间存在近似的线性关系，则它也可用来构建系统树（Nei,1975)。如果这种关系成立，则象核苷酸（或氨基酸）替换数这样的线性距离尺度就应该被用到了。

ＵＰＧＭＡ法采用连续聚类算法，在此法中局部拓扑关系按相似性的级别来鉴别，而系统树的构建是分步实现的。也就是说，我们首先从所有ＯＴＵ中找出两个最相似的ＯＴＵ，并把它们看成一个新ＯＴＵ。这样的ＯＴＵ我们称之为复合ＯＴＵ（composite OTU)。接着，从这种新的ＯＴＵ群中我们再找出有最大相似性的对子，如此继续下去，直到我们最后仅留下 2 个ＯＴＵ时为止。 为了对此法作出具体说明，让我们来考虑一个包含 4 个ＯＴＵ的例子。成对的进化距离，如朱克斯和坎托（Jukes 和 Cantor,1969)所估计的那样（第三章），由以下矩阵所给出：

	OTU		
OTU	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

在此矩阵中，d_{ij}代表ＯＴＵ i 和 j 间的距离。最先被聚类的两个ＯＴＵ是距离最短的那两个。我们假定 dAB 有最小值，则ＯＴＵＡ和Ｂ是最先被聚类的，而分枝点则放在距离为 $d_{AB}/2$ 替换的地方（图 5 — 6 a)。

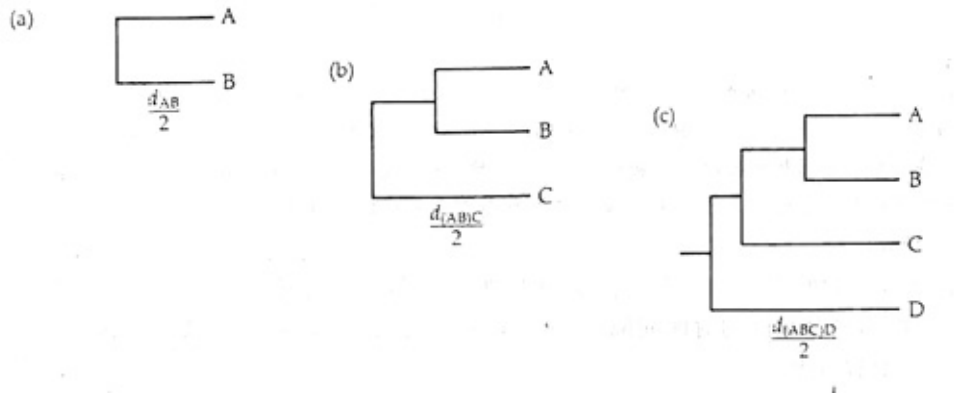


图 5 - 6 对 4 个ＯＴＵ按ＵＰＧＭＡ法分步构建系统树的示意图（详见正文)。

第一次聚类后，Ａ和Ｂ被看成一个复合ＯＴＵ，据此而算出一个新距离矩阵：

	OTU	
OTU	(AB)	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

在此矩阵中，$d_{(AB)C}=(d_{AC}+d_{BC})/2$，$d_{(AB)D}=(d_{AD}+d_{BD})/2$。换言之，一个单ＯＴＵ和一个复合ＯＴＵ间的距离是，该单ＯＴＵ与该复合ＯＴＵ的各组成单ＯＴＵ间距离的平均值。如果 $d_{(AB)C}$ 是新矩阵中的最小距离，那么，OTU C 将加入到复合 OTU(AB)中，且在 $d_{(AB)C}/2$ 处有一个分枝节点(图 5-6b)。

最后一步是将最后剩下的ＯＴＵＤ与复合ＯＴＵ（ＡＢＣ）聚类。整个树的根位于 $d_{(ABC)D}/2=〔(d_{AD}+d_{BD}+d_{CD})/3〕/2$ 处。用ＵＰＧＭＡ法推测出的最后树如图５－６ｃ所示。

在ＵＰＧＭＡ法中，两个复合ＯＴＵ间的距离用每一复合ＯＴＵ中的所有组成ＯＴＵ间距离的算术平均算出。例如，复合ＯＴＵ（ｉｊ）和复合ＯＴＵ（ｍｎ）间的距离为：

$$d_{(ij)(mn)}=(d_{im}+d_{in}+d_{jm}+d_{jn})/4 \qquad (5.3)$$

在复合 OTU(ijk)和(mn)的例子中，距离为：

$$d_{(ijk)(mn)}=(d_{im}+d_{in}+d_{jm}+d_{jn}+d_{km}+d_{kn})/6 \qquad (5.4)$$

变形距离法

如果假定速率恒定在各谱系间不成立，则ＵＰＧＭＡ法可能会给出错误的拓扑图。例如，假定图５－７ａ中的系统树是真实树，而各ＯＴＵ对的进化距离则由以下矩阵给出：

	OTU		
OTU	A	B	C
B	8		
C	7	9	
D	12	14	11

用ＵＰＧＭＡ法，我们得到的推测树在分枝模式上与真实树不同（图５－７ｂ）。如，推测树中ＯＴＵＡ和Ｃ被组合在一起了，而在真实树中，Ａ和Ｂ才是姐妹ＯＴＵ。（注意，此例中可加性不存在。如，Ａ和Ｂ间的真实距离为８，而估出的联结Ａ和Ｂ的分枝，其枝长之和为 3.50+0.75+4.25=8.50）。

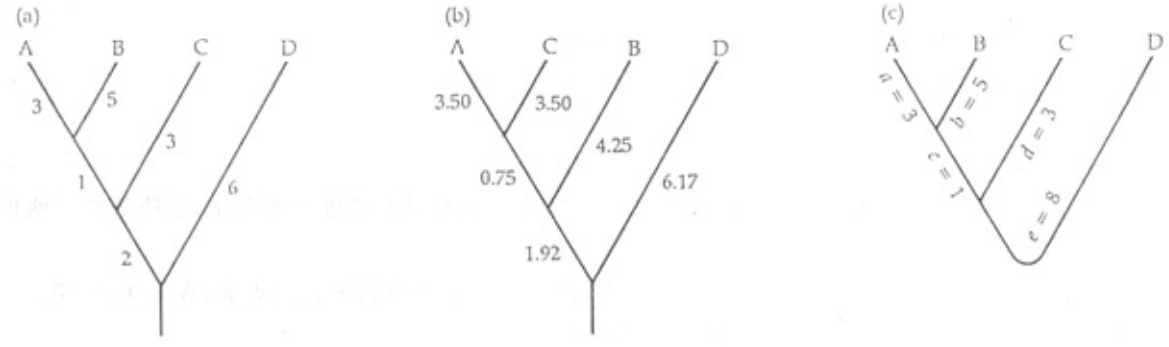


图 5－7 （ａ）有４个ＯＴＵ和（ｂ）有５个ＯＴＵ的两分叉无根树

不过，该拓扑图错误也许能用被称为变形距离法（transformed distance method ）的校正方法来订正（Farris,1977;Klotz 等,1979)。简单地说，此法是用一个组外单位（outgroup） 作参照物，来对被研究谱系间进化速率不等的情况作一些校正，然后对新得到的距离矩阵应用ＵＰＧＭＡ法，从而推测出该树的拓扑图。组外单位是一个我们对其有外在认识的ＯＴＵ。这些外在认识，如分类学的或古生物学的知识，能清楚地表明，该ＯＴＵ已先于所有其他被研究的ＯＴＵ而从其共同祖先分岐出来。

在现在的例子中，我们假定分类单位D对所有别的分类单位来说是一个组外单位。那么D即可当作参照物，并用下式来变换距离：

$$d'_{ij} = d_{ij} - d_{iD} - d_{jD}/2 + d_D \qquad (5.5)$$

这里 d'_{ij} 是变形距离，i=A,B 或 C， $d_D=(d_{AD}+d_{BD}+d_{CD})/3$。$d_D$ 项的引入是为了保证所有 d'_{ij} 的值都是正的。做此手续是因为，实际上距离不可能是负的。对于有 n 个ＯＴＵ（不包括组外单位）的一般情形，$d_D=\sum d_{iD}/n$。

在我们的例子里， $d_D=37/3$ 而新距离矩阵中关于分类单位Ａ、Ｂ和Ｃ的值为

	OTU	
OTU	A	B
B	10/3	
C	13/3	13/3

由于 d'_{AB} 有最小值,所以 A 和 B 最先被聚类在一起,然后,C 再加入到该树中。根据定义，组外单位ＯＴＵＤ决定该树的根，因而最后加入该树。这样就给出了正确的拓扑图（图５－７ｃ）。在上例中，我们仅考虑了 3 个分类单位和 1 个组外单位，但此法可很容易地推广到更多分类单位和/或更多组外单位的场合。

在许多具体情况下，不可能预先决定被研究的分类单位中哪一个是组外单位。为了克服这一困难，两步进行法已被提了出来（Li,1981）。第一步，用ＵＰＧＭＡ法先推出该树的根。然后，处于根一侧的分类单位被当作参照物（组外单位），对处于根的另一侧的谱系间不相等的进化速率做校正，以后再做相反方向的校正。在我们的例子中，此法也能找出正确的树。

近邻关系法

在一个无根两分叉树中，如果两个ＯＴＵ通过一个内部节点联结，则它们就被说成是近邻（neighbors)。例如，图５－８ａ中，Ａ和Ｂ是近邻，Ｃ和Ｄ也是如此。与之相比，图５－８ｂ中，Ａ和Ｃ不是近邻，Ｂ和Ｃ也不是。然而，如果我们把ＯＴＵＡ和Ｂ结合成一个复合ＯＴＵ，则复合ＯＴＵ（ＡＢ）和单ＯＴＵＣ就变成了一对新近邻。

现在我们假定，图５－８ａ中展示的树是真实树。那么，若可加性存在则我们应有：

$$d_{AC}+d_{BD}=d_{AD}+d_{BC}=a+b+c+d+2x=d_{AB}+d_{CD}+2x \qquad (5.6)$$

这里 x 是内部分枝的长度。因此，以下两个条件成立：

$$d_{AB}+d_{CD}<d_{AC}+d_{BD} \qquad (5.7a)$$

和

$$d_{AB}+d_{CD}<d_{AD}+d_{BC} \qquad (5.7b)$$

这两个条件总称为四点条件（four point condition)。即使可加性只不过近似地存在，这些条件可能也是成立的。

反之，对系统发育关系不知的４个ＯＴＵ而言，以上两个条件可用来找出近邻（Ａ和Ｂ；Ｃ和Ｄ)。一旦这两对近邻被确定，系统树的拓扑图也就被确定了。

萨塔斯和特韦尔斯基（Sattath 和 Tversky,1977）提出以下方法来处理多于４个ＯＴＵ的情况。首先，象在ＵＰＧＭＡ中的情况一样算出距离矩阵。对每种可能的四单位组，比如ＯＴＵｉ，ｊ，ｍ和ｎ，算出 $d_{ij}+d_{mn}$，$d_{im}+d_{jn}$ 和 $d_{in}+d_{jm}$。假定第一个和的值最小，那么，我们把ｉ和ｊ对以及ｍ和ｎ对都记１分，而ｉ和ｍ对、ｉ和ｎ对、ｊ和ｍ对、ｊ和ｎ对则都记０分。否则，如果 $d_{im}+d_{jn}$ 有最小值，则我们就给ｉ和ｍ对以及ｊ和ｎ对记１分，而给其他４种可能的对子记０分。在对所有可能的四单位组都评过分以后，得总分最高的对子即被选为第一个近邻对，并把它当成一个ＯＴＵ来处理，下一步，如在ＵＰＧＭＡ中的情况一样算出新距离矩阵，然后重复以上过程选出第二个近邻对。此过程继续到所有ＯＴＵ被聚类时止。在我们考虑猿和人的系统发育时将给出此法的详细说明（见第 72 页)。

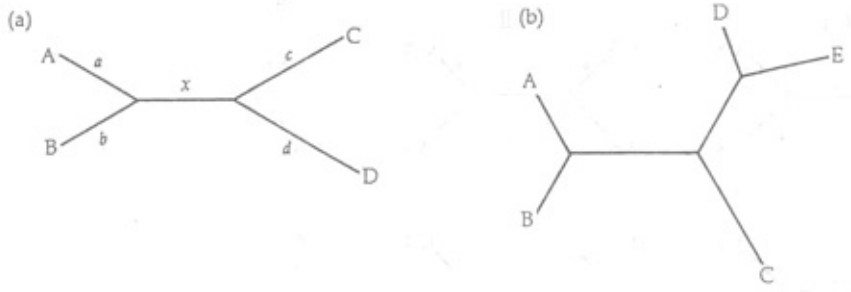


图５－８ （ａ）有４个ＯＴＵ和（ｂ）有５个ＯＴＵ的两分叉无根树.

以近邻关系概念为根据的另一种方法曾由菲奇(Fitch,1981)提出。塞图和根井(Saitou 和 Nei,1987)曾提出一个称为近邻结合法（neighbor joining method）的方法。这是连续地找出使树的总长度最小的近邻对的方法。

最节省法

最节省或最少进化的原则涉及：找出一个要求最小进化变化数的树，以解释被研究ＯＴＵ间观察到的差异，这样一种树称为最节省树（maximum parsimony tree)。常常可以发现不止一种树有同样的最小变化数，所以，推测出的树可能不是唯一的。

下面讨论的方法最初是为了处理氨基酸顺序数据而提出的（Eck 和 Dayhoff,1966）,后来经修改后被用于核苷酸顺序（Fitch,1977)。

我们先对信息位点（informative sites）进行定义。一个核苷酸位点是具有系统发育方面的信息的，仅当它在众多的树中偏向其中一些树时。为了说明信息位点和非信息位点间的区别，让我们考虑以下４个假想的顺序：

	位点								
顺序	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	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G
					*		*		*

对４个ＯＴＵ有３种可能的无根树（图５－９）。位点１不是有信息的，因为所有顺序在该位点上都有Ａ，以至在３种可能的树中都不需要任何变化。在位点２上，顺序１有Ａ，而所有其他顺序都有Ｇ，所以只需作一简单假定，即谱系中核苷酸从Ｇ变成Ａ导至了顺序１。于是，这一位点也不是有信息的，因为对这３种可能的树来说每种都只需要１次变化。如图５－９ａ所示，对位点３来说，这３种可能的树中的每一种都需要２次变化，所以也不是有信息的。注意，如果假定图５－９ａ的树Ⅰ中联结ＯＴＵ　１和２的节点上的核苷酸是Ｃ（或Ａ）而不是Ｇ，则该树所需要的变化数仍为２。图５－９ｂ表明对于位点４这３种树的每一种都要求３次变化，所以位点４也是非信息的。对于位点５，树Ⅰ只要求１次变化，而树Ⅱ和Ⅲ则各要求２次变化（图５－９ｃ），所以该位点是有信息的。

从这些例子中我们看到，就我们所谈到的分子数据而言，仅当一个位点上至少有两种不同类型的核苷酸、每种类型至少在两个被研究顺序中出现时，该位点才是有信息的。在上例中，信息位点用星号（*）指出。

为了推出一个最节省树，我们首先要找出所有信息位点。接着，对每种可能的树我们算出每一信息位点上的最小替换数。在上例中，对于位点５、７和９，树Ⅰ分别需要１、１和２次变化，树Ⅱ分别需要２、２和１次变化，树Ⅲ分别需要２、２和２次变化。最后一个步聚是，对每种树将所有信息位点上的变化数加和，并选出总替换数最小的那种树。在我们的例子中，因为树Ⅰ在信息位点上需要的变化数

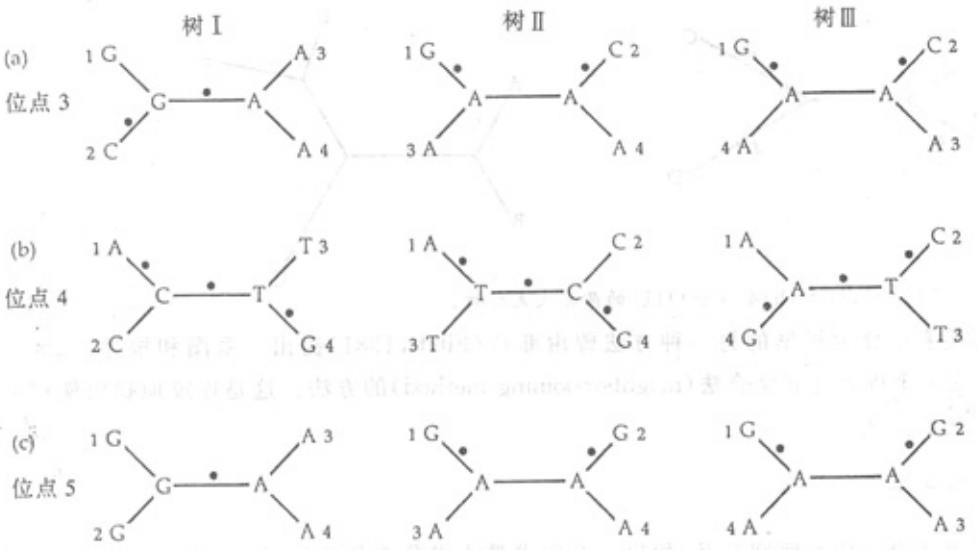


图5－9 关于4个DNA顺序（1、2，3和4）的3种可能的无根树（Ⅰ，Ⅱ和Ⅲ），这4个顺序曾被用来选取最节省树（见正文）。终端节点处指出了现存物种中该同源位置上的核苷酸种类。分枝上的每一个圆点表示一次替换，它被放在发生替换的分枝上。注意，每种树的两个内部节点上的核苷酸只表示了几种可供选择的构建中的一种。例如，树Ⅲ（c）的两个内部节点上可能都是G而不是A（右下角）。在这种情况下，两次替换将位于导致物种3和4的分枝上。而所要求的替换最小数却保持相同。

最小（4），所以选中树Ⅰ。

在有4个OTU的情况下，一个信息位点只偏向3种可供选取的树中的一种。例如，位点5偏向树Ⅰ甚过树Ⅱ和树Ⅲ，因而被说成是支持树Ⅰ的。容易看出，受信息位点支持数最大的那种树就是最节省树。例如，上例中，树Ⅰ受2个位点支持，树Ⅱ受1个位点支持，而树Ⅲ无位点支持，结论不言自明。

5.4 表型学与进化枝学

分类学中长期存在的争论就是那些一直在“进化枝学家”和“表型学家”间发生的言辞刻薄的争论。进化枝学（cladistics）这个词可以定义为研究进化途径的科学。换句话说，进化枝学家对这样一些问题感兴趣：在一群生物中存在多少分枝？ 哪些分枝与别的哪些分枝相联？ 分枝次序如何（Sneath和Sokal，1973）？一种表示这种祖先—后代关系的树状网络称进化分枝图（cladogram）。换句话说，进化分枝图即指一种有根系统树的拓扑图。

另一方面，表型学（phenetics）是根据一群生物间的相似程度来研究它们间的关系的科学，引以为据的类似性可以是分子上的、表型上的或解剖学上的。表示表型学关系的树状网络称表型关系图（phenogram）。虽然表型关系图可以当作进化枝学关系的指示物来用，但它可不必与进化分枝图等同。如果在分岐时间和遗传学上的（或形态学上的）分岐程度间存在线性关系，则这两种树可能会变得相互等同。

在上面所讨论的方法中，最节省法是进化枝学方法的一个典型代表，而UPGMA法则是一种典型的表型学方法。然而，其他方法却不能按以上标准简单地进行分类。例如，变形距离法和近邻关系法常常被说成是表型学方法，但这是一种不确切的说法。虽然这些方法用了类似性（或不似性，即距离）尺度，但它们并没有假定类似性和进化关系间直接关联，也并不打算推测表型学关系。

在分子系统发育中， 用更合理一些的分类方法应把它们分成距离法和特性状态法（distance and character state approaches）。属于前者的方法根据距离尺度，象核苷酸或氨基酸的替换数等，而后者则依靠该特性的状态，如某一特定位点上的核苷酸或氨基酸，或某一DNA位置上缺失或插入的出现或不出现等。根据这种分类，则UPGMA法、变形距离法和近邻关系法都是距离法，而最节省法则是一种特性状态法。

一直有这样一种说法，即，特性状态法比距离法更有效用，因为该原始资料是一串特性状态（例如核苷酸顺序），而将特性状态数据转变成距离矩阵以后，有些信息就丢失了。然而我们注意到，虽然最节省法事实上应用的是原始资料，但通常它仅用了可用资料的一小部分。例如，第 75 页上的例子中，只有 3 个位点被用于分析中，倒有 6 个位点被排除在外了。为此，此法常比有些距离矩阵法的效率低（例如见 Saitou 和 Nei,1986）。当然，如果信息位点数多，则最节省法一般是非常有效的。

最后，应该注意到，不管用何种方法，一个推测树常难免有拓扑图错误。为了得到正确的树，通常需要大量数据资料。

5.5 枝长的估计

除了在ＵＰＧＭＡ法中外，我们都不曾讨论过怎样去估计枝长的问题。现在，我们在该树的拓扑图已被推测出来的假定下，来处理这一问题。我们只考虑用距离矩阵法推测出来的树对于最节省法来说，这个问题要复杂得多（见 Fitch,1971）。下面讨论的方法是菲奇和马戈利阿什（Fitch 和 Margoliash,1967）的方法。

首先，我们考虑一种最简单的情形，即一个有３个ＯＴＵ（Ａ，Ｂ和Ｃ）和１个节点的无根树（图５－１０ａ）。设ｘ、ｙ和ｚ分别是导向Ａ、Ｂ和Ｃ的分枝的长度。很容易看出以下等式成立：

$$d_{AB}=x+y \qquad (5\ \ 8a)$$

$$d_{AC}=x+z \qquad (5\ \ 8b)$$

$$d_{BC}=y+z \qquad (5\ \ 8c)$$

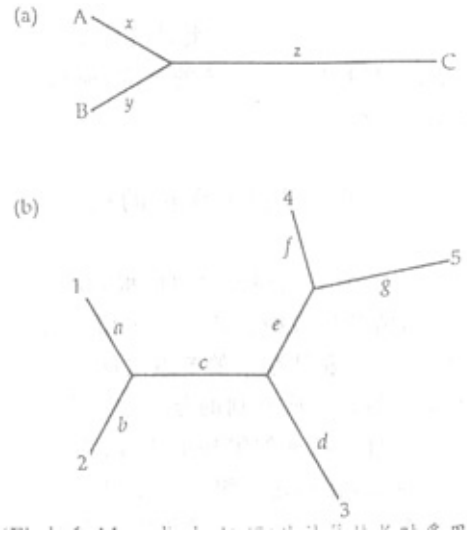


图５－１０ 用菲奇和马戈利阿什（Fitch 和 Margoliash,1967）法计算枝长时采用的无根系统树。（ａ）有３个ＯＴＵ的树，（ｂ）有５个ＯＴＵ的树。

从这些等式，我们得到以下解：

$$x=d_{AB}+d_{AC}-d_{BC}/2 \qquad (5.9a)$$

$$y=d_{AB}+d_{BC}-d_{AC}/2 \qquad (5.9b)$$

$$z=d_{AC}+d_{BC}-d_{AB}/2 \qquad (5.9c)$$

我们现在来处理有３个以上ＯＴＵ的情形。为简单计，我们假定有５个ＯＴＵ（１，２，３，４和５），而其拓扑图和各枝长如图５－１０ｂ所示。假定ＯＴＵ１和２是构树过程中最先被聚类的ＯＴＵ，则我们分别用Ａ和Ｂ来表示ＯＴＵ１和２，而把所有别的ＯＴＵ（３、４和５）放在一个复合ＯＴＵ中，并表示成Ｃ。通过这种安排，我们可以用等式 5.9a-5.9c 来估计导向Ａ、Ｂ和Ｃ的分枝的长度。

不过，现在 $d_{AC}=d_{1(345)}=(d_{13}+d_{14}+d_{15})/3, d_{BC}=d_{2(345)}=(d_{23}+d_{24}+d_{25})/3$ 于是，我们有 a=x 和 b=y。接着ＯＴＵ１和２被看成一个复合ＯＴＵ。下面再假定，复合ＯＴＵ（１２）与单ＯＴＵ３是下一个要结合在一起的对子。于是，我们分别把复合ＯＴＵ（１２）和单ＯＴＵ３用Ａ和Ｂ表示，把其余的ＯＴＵ（即４和５）放在新的复合ＯＴＵＣ中。以上述同样的方式，我们将得到ｘ，ｙ和ｚ。注意，这时ｄ＝ｙ，c+(a+b)/2=x。从前一次我们得到的关于ａ和ｂ的值，可以算出ｃ。继续进行这样的过程，直到所有枝长都得到时为止。

注意，有时某一被估出的枝长可能会是负的。由于实际枝长绝不可能是负的，所以，最好用０来代替这样的估值。

作为应用上述方法的一个例子，让我们来算一下图５—7c 中那个树的各枝长。为便于说明，我们把在推测该树的拓扑图时用过的距离矩阵再次列出。为了避免与等式 5.9 中的符号混淆，我们把ＯＴＵＡ，Ｂ，Ｃ和Ｄ分别地重新命名为ＯＴＵ１，２，３和４。

	OTU		
OTU	1	2	3
2	8		
3	7	9	
4	12	14	11

因为ＯＴＵ１和２最先被聚类，所以，我们把ＯＴＵ３和４放在一个复合ＯＴＵＣ中，从而先算出导向ＯＴＵ１和２的两个分枝的长度（ａ和ｂ）。而我们有 $d_{AB}=d_{12}=8, d_{AC}=(d_{13}+d_{14})/2=(7+12)/2=9.5$ 和 $d_{BC}=(d_{23}+d_{24})/2=11.5$。由等式 5.9a-5.9c, 我们有：a=x=(8+9.5-11.5)/2=3，b=y=(8+11.5-9)/2=5。下一步我们把ＯＴＵ１和２处理成ＯＴＵ（１２）并用Ａ表示。于是，我们有：$d_{AB}=d_{(12)3}=(d_{13}+d_{23})/2=(7+9)/2=8, d_{AC}=d_{(12)4}=(d_{14}+d_{24})/2=(12+14)/2=13, d_{BC}=d_{34}=11$。因此，根据等式 5.9a-5.9c, 我们有：x=(8+13-11)/2=5, d=y=(8+11-13)/2=3, e=z=(13+11-8)/2=8。由图５—７ｃ可以看出(a+b)/2+c=x, 所以, c=1。至此计算完毕。然而请注意，由于我们不知道根的准确位置，所以我们不能估出联系根和ＯＴＵＤ的分枝的长度，而只能估出从ＯＴＵＡ，Ｂ和Ｃ的共同祖先节点通过根到ＯＴＵＤ的长度，即ｅ＝８。

5.6 寻找无根树的根

大多数构树法得到的都是无根树。为了找到无根树的根，我们通常都需要一个组外单位（即一个ＯＴＵ，其外在信息，如古生物学证据，清楚地表明它比那些被研究的分类单位更早地发生分枝）。而根则被放在该组外单位和将它和别的ＯＴＵ联结起来的节点之间。

虽然我们必须能肯定该组外单位的确比所有别的分类单位更早地发生分岐，但奉劝各位，不要选一个与被研究的各分类单位亲缘关系过于遥远的组外单位，因为，在这样的情况下很难得到该组外单位与其他分类单位间的可靠估值。例如，在构建一组有胎盘哺乳类间的系统发育关系时，我们可以用一种有袋类作为组外单位。仅当所用的ＤＮＡ序列在进化中是高度保守的时，鸟类才可能被当成可靠的组外单位。在此例中植物与真菌显然也够资格作为组外单位，不过，仅仅因为它们与哺乳类的关系太过遥远用它们作组外单位，结果就可能会出现严重的拓扑图错误。用一个以上的组外单位(假若它们与别的分类单位都不是关系遥远)，则一般能改善该树拓扑图的估计结果。该组外单位还必须在系统发育上与别的ＯＴＵ不至于太接近，因为，在这种情况下我们不能肯定它是一个真正的组外单位，即不能肯定它与其他ＯＴＵ的分岐早于这些ＯＴＵ相互间的分岐。

在缺少组外单位的情况下，我们可以通过假定进化速率在所有分枝上近似一致来放置根。在这种假定下，我们把根放在两个ＯＴＵ间的最长路线的中点处。

5.7 物种分岐时间的估计

因为古生物学记录远不够完全，所以我们常常不去管物种分岐的年代。ＤＮＡ顺序数据在这方面可能

会给出很大帮助。从以前的研究中获知我们假定某一ＤＮＡ序列的进化速率，为每年每位点 r 替换。为了得到物种Ａ和Ｂ间的分岐时间（Ｔ），我们对来自这两个物种的该顺序进行比较，并算出每位点替换数（Ｋ）。如第四章所示，我们有以下等式：

$$r=K/2T \qquad (5.10)$$

因此，Ｔ估计为

$$T=K/2r \qquad (5.11)$$

如在第四章所说明过的那样，从一组生物得到的核苷酸替换速率也许不能用于另一组生物。为了避免出现这样的问题，我们可通过加进第３个物种Ｃ（它与物种对Ａ和Ｂ的分岐时间已知），来估计替换速率（图５－１１）。设Ｋｉｊ是物种ｉ和ｊ间每位点核苷酸替换数，那么，核苷酸替换速率可由：

$$r=K_{AC}+K_{BC}/2(2T_1) \quad (5.12)$$

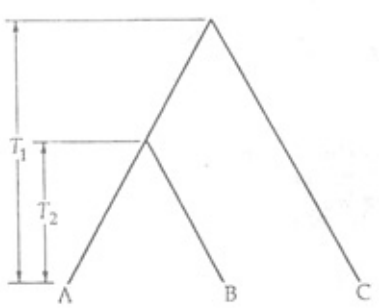


图５－１１ 估计分岐时间用的模型树（见正文）

估出。物种Ａ和Ｂ间的未知分岐时间（Ｔ2）用：

$$T_2= K_{AB}/2r=2K_{AB}\ T_1/K_{AC}+K_{BC} \qquad (5.13)$$

估出。反之，在 T2 已知而 T1 未知的情况下，T1 由下式给出：

$$T_1= (K_{AC}+K_{BC})T_2/2K_{AB} \qquad (5.14)$$

以上公式均假定速率恒定。但如前一章中所讨论的那样，此假定常常是不成立的，所以估出的分岐时间应该小心处理。李和谷村（Li 和 Tanimura,1987）曾提出一种方法，此法能降低替换速率不等对分岐时间估计的影响。

早先我们已经注意到，两顺序间的分岐时间可能比携带这两个顺序的两物种间的分岐时间早。不过，如果我们关心的是长期分岐事件，比如说几百万年或更长的时间等级，则这种误差通常并不很严重。还应该注意，分岐时间的估值通常承受着较大的随机误差。为了减少这类误差，在估计中应该用许多顺序。

5.8 进化枝

系统发育研究的目的是建立不同物种间的进化关系。特别地，我们对找出自然进化枝（clades)感兴趣。一个进化枝被定义成一组有共同祖先的物种，而该祖先却不为这一进化枝外的其他物种所共有。

图５－１２显示脊椎动物的３个纲：鸟纲、爬行纲和哺乳纲间的进化关系。我们看到，经典分类学把爬行类指定为一个独立的纲，这与进化枝的定义不符，因为这３类爬行动物与另一个类群——鸟类有共同祖先，而后者则并不被包括在爬行纲的定义范围内。另一方面，鸟类与鳄类倒的的确确地构成了一个自然进化枝，初龙亚纲，因为它们所共有的一个共同祖先不为除鸟类和鳄类以外的任何现存生物所共有。类似地，所有鸟类和所有爬行类合在一起构成一个自然进化枝。

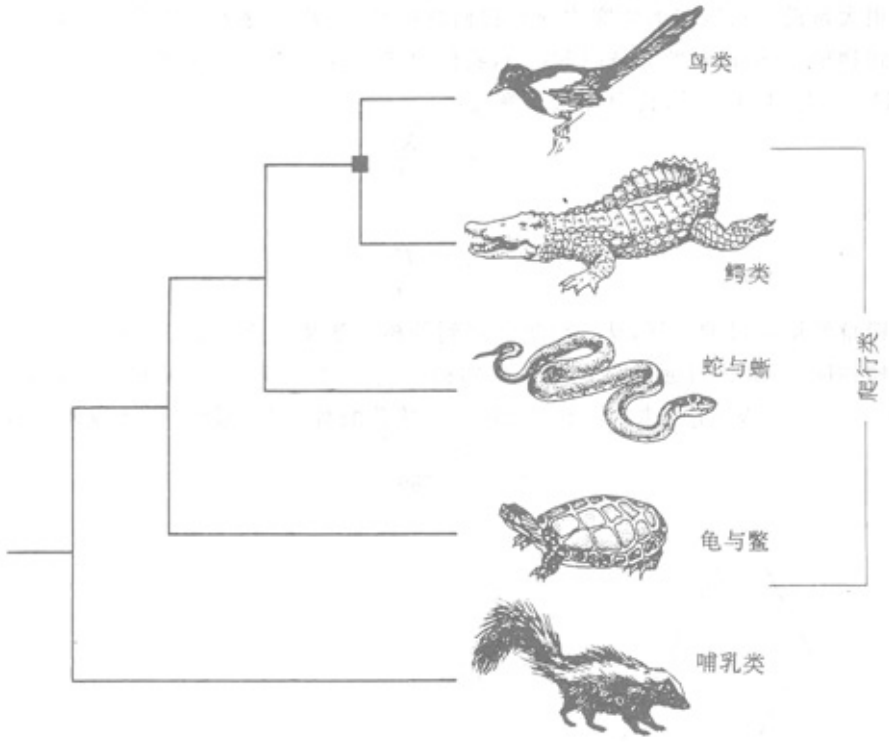


图 5 - 1 2 鸟类，爬行类和哺乳类的分枝进化图。爬行类不构成一个自然进化枝，因为它们与鸟类有共同祖先，而后者并不被包括在爬行纲中。另一方面，鸟类与鳄类却构成一个自然进化枝（初龙亚纲），因为它们所共有的一个共同祖先（黑色方块）不为任何别的生物所共有。

5.9 人和猿的系统发育

与人类在进化上最接近的亲缘关系问题一直是令生物学家们感兴趣的问题。例如，达尔文（Darwin,1871）曾认为，非洲猿，黑猩猩（Pan）和大猩猩(Gorilla)是人类的近亲，所以他的结论是，人类的进化起源将被发现一定是在非洲。。由于各种原因达尔文的观点没有引起人们的兴趣，且在一个相当长的时期里，分类学家们认为人属（Homo） 是唯一与猿的亲缘关系较远的一类，所以它自成一科，即人科。黑猩猩、大猩猩和马来猩猩（Pongo）通常被放在独立的一个科，即猩猩科（图 5 — 1 3 a ）中。长臂猿（Hylobates）或者独立成科或者被分在猩猩科（图 5 — 1 3 b ； 见 Simpson,1961）。古德曼（Goodman,1963）曾正确地认识到，这种系统安排是以人类为中心的，因为人类代表着“系统发育所达到的一个新等级，一个比猿和所有别的以前存在的等级更高的等级”。然而，把各种猿放在一个科里而把人放在另一个科里，这就意味着这些猿相互间共有一个比人更近的共同祖先。当把人与一种现存的猿放在同一进化枝中时，它通常是与亚洲猿，即马来猩猩放在一起（图 5 — 1 3 c ； Schultz,1963）。

由于血清沉淀法的应用，古德曼（Goodman,1962）已能证明人、 黑猩猩和大猩猩构成一个自然进化枝（图 5 — 1 3 d ），而马来猩猩和长臂猿与别的猿类发生分岐的年代则要早得多。根据微量互补结合资料，萨里奇和威尔逊（Sarich 和 Wilson,1967）估计人与黑猩猩或大猩猩间的分岐时间可能近至 5 0 0 万年前，而不是当时普遍为古生物学家所接受的至少 1 5 0 0 万年前。

然而，血清学方法、电泳法以及氨基酸顺序等都不能解决人与非洲猿间的进化关系问题，而所谓人—大猩猩—黑猩猩的三分法仍未得到解决，并继续是一个极有争议的问题（图 5 — 1 4 ）。下面，我们将用 M.古德曼及其同事（见 Miyamoto 等,1987）和其它学者（Maeda 等,1988）得到的ＤＮＡ顺序资料，以表明分子证据支持人—黑猩猩进化枝，同时对前几节所讨论的构树法加以解释说明。

表 5 — 2 展示了以下几个ＯＴＵ每一对子间每 1 0 0 位点的核苷酸替换数：人（Ｈ），黑猩猩（Ｃ），大猩猩（Ｇ），马来猩猩（Ｏ）和罗猴（Ｒ）。我们先对这些距离用ＵＰＧＭＡ法。人与黑猩猩间的距离最短（d_{HC}=1.45），因此，我们首先把这两个ＯＴＵ结合，并把节点放在 1.45/2=0.73 处（图 5 — 1 5

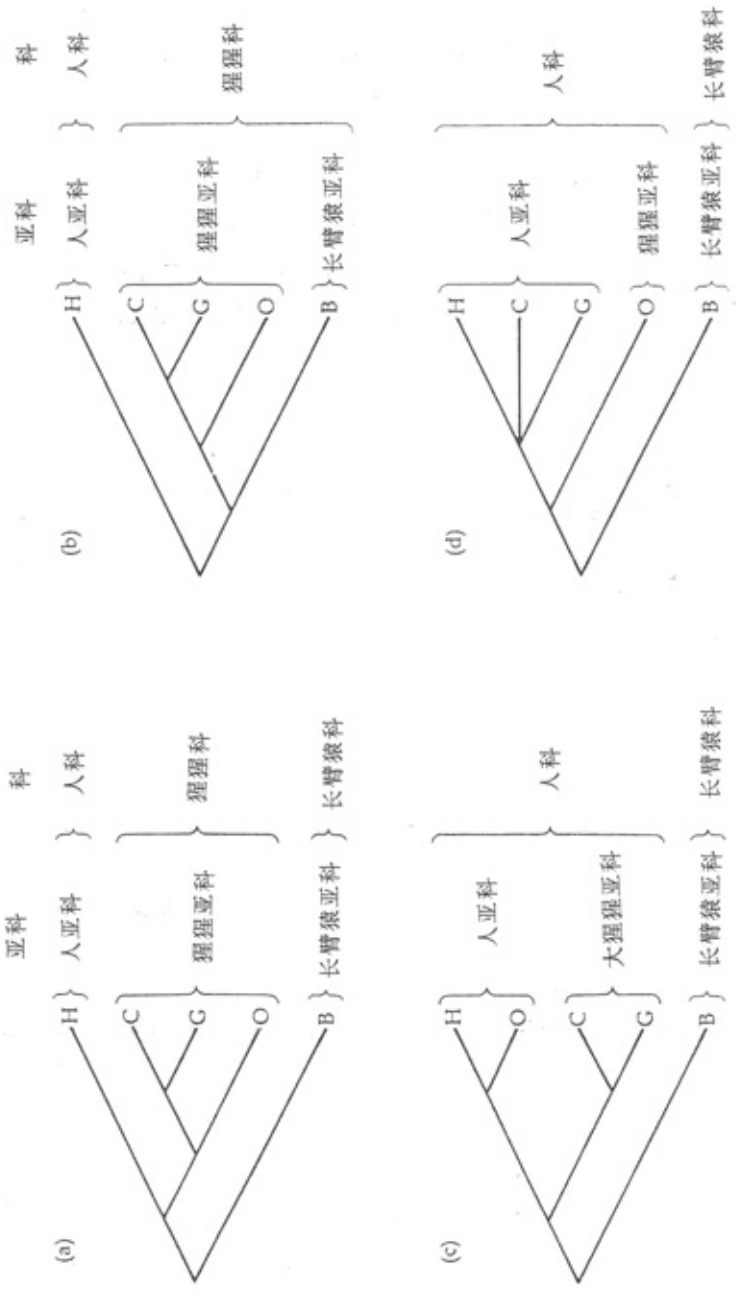


图 5 - 1 3 现代猿和人（人总科）的 4 种供选择的系统发育和分类方式。传统分类法把人类独立出来，如（ a ）和（ b ）中所展示的那样。人与马来猩猩的聚类如（ c ）中所示。把分子的和形态学的证据结合起来则偏向（ d ）中所示的那种分类方式。

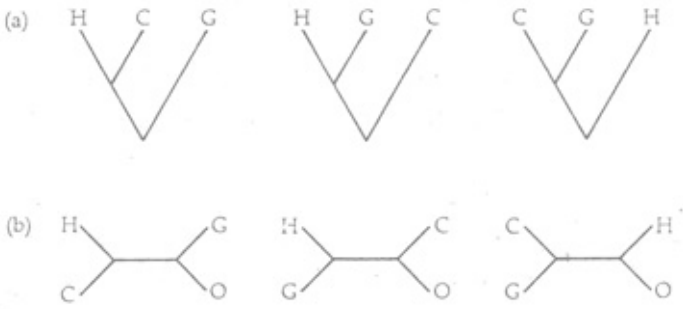


图 5 - 1 4 （ a ）关于大猩猩、黑猩猩和人的 3 种可能的有根树。（ b ）以马来猩猩作为组外单位的相应无根树。物种缩写：C，黑猩猩（ Pan troglodytes ）；H，人（ Homo sapiens ）；G，大猩猩（ Gorilla gorilla ）；O，马来猩猩（ Pongo pygmaeus ）

a）。然后，我们算出该复合ＯＴＵ（ＨＣ）与其余每一物种间的距离，得出一个新的距离矩阵：

	OTU		
OTU	(HC)	G	O
G	1.54		
O	2.96	3.04	
R	7.53	7.39	7.10

因为（ＨＣ）和Ｇ为最短的距离所隔开，所以它们是下一个要结合起来的对象，且该联结的结节位于1.54/2=0.77处。继续此类过程，我们即得到图５－１５ａ中那样的树。我们注意到，估出的关于Ｈ和Ｃ的分枝结点非常接近于（ＨＣ）和Ｇ的分枝节点。事实上，这两个节点间的距离，比Ｈ、Ｃ和Ｇ间所

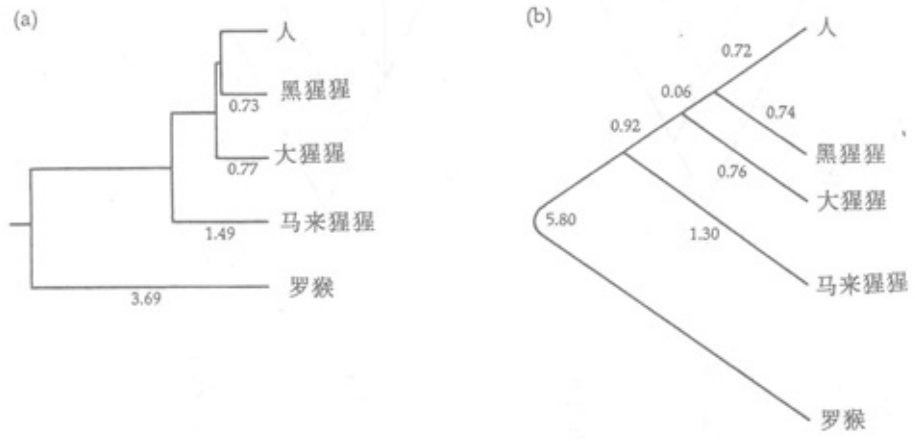


图５－１５ 由ＵＰＧＭＡ法（ａ）以及萨塔特和斯韦尔斯基的近邻关系法（ｂ）推测出的、关于人、黑猩猩、大猩猩和罗猴的进化树。

有关于成对距离估值的标准误差都小（表５－２）。所以，虽然该资料表示人的最近现存亲属是黑猩猩，但该资料并没对分枝次序问题给出结论性的解决。另一方面，马来猩猩对于人—黑猩猩—大猩猩进化枝而言是一个组外单位，这一点则是毫不含糊的。

表５－２ 各ＯＴＵ间每１００位点核苷酸替换数的平均值（对角线下）和标准误差（对角线上） a

	OTU				
OTU	人	黑猩猩	大猩猩	马来猩猩	罗猴
人		0.17	0.18	0.25	0.41
黑猩猩	1.45		0.18	0.25	0.42
大猩猩	1.51	1.57		0.26	0.41
马来猩猩	2.98	2.94	3.04		0.40
罗猴	7.51	7.55	7.39	7.10	

自 Li 等(1987b)

a 所用顺序资料为长 5.3Kb 的非编码 DNA,它由两个分开的区域构成: (1)由 Koop 等(1986b)报导的η 珠蛋白基因座位(2 2kb)和由 Maeda 等(1983,1988)定序的长 3.1kb 的η－δ珠蛋白基因间区域。

下面，我们用萨塔特和斯韦尔斯基的近邻关系法。我们一次考虑４个ＯＴＵ，由于现有５个ＯＴＵ，所以有 5!/〔4!(5-4)!〕=5 种可能的四单位组。我们从ＯＴＵＨ、Ｃ、Ｇ和Ｏ开始，并算出以下距离之和（数据自表５－２）：$d_{HC}+d_{GO}=1.45+3.04=4.49$,$d_{HG}+d_{CO}=4.45$ 和 $d_{HO}+d_{CG}=4.55$。因为第２个和最小，所以我们选 Ｈ和Ｇ为一个近邻对，Ｃ和Ｏ为另一对（表５－３）。类似地，我们再考虑另外４种可能的四单位组，结果如表５－３所示。从表５－３的最后一行我们看到，其在所有近邻对中（ＯＲ）有最高的近邻关系得分，所以我们选取（ＯＲ）为第一个近邻对。把这个对子当作一个ＯＴＵ，则我们得到以下新的距离矩阵：

	OTU		
OTU	H	C	G
C	1.45		
G	1.51	1.57	
(OR)	5.25	5.25	5.22

表 5 － 3 用表 5 － 2 中的距离矩阵得到的近邻关系得分

用于比较的 OTU[a]	成对距离之和	选取的近邻对
H,C,G,O	$d_{HC}+d_{GO}=4.49$ $d_{HG}+d_{CO}=4.45$ $d_{HO}+d_{CG}=4.55$	(HG),(CO)
H,C,G,R	$d_{HC}+d_{GR}=8.84$ $d_{HG}+d_{CR}=9.06$ $d_{HR}+d_{CG}=9.08$	(HC),(GR)
H,C,O,R	$d_{HC}+d_{OR}=8.55$ $d_{HO}+d_{CR}=10.53$ $d_{HR}+d_{CO}=10.45$	(HC),(OR)
H,G,C,R	$d_{HG}+d_{OR}=8.61$ $d_{HO}+d_{GR}=10.37$ $d_{HR}+d_{GO}=10.55$	(HG),(OR)
C,G,O,R	$d_{CG}+d_{OR}=8.67$ $d_{CO}+d_{GR}=10.33$ $d_{CR}+d_{GO}=11.59$	(CG),(OR)
总分	(HC)=2,(HG)=2,(HO)=0,(HR)=0,(CG)=1 (CO)=1,(CR)=0,(GO)=0,(GR)=1,(OR)=3	

a.H,人; C，黑猩猩; G，大猩猩; O，马来猩猩; R，罗猴。

由于仅留下 4 个 O T U，所以很容易看出，$d_{HC}+d_{G(OR)}=6.76<d_{HG}+d_{C(OR)}=6.76<d_{H(OR)}+d_{CG}=6.82$。因此，我们选 H 和 C 作为一个近邻对，G 和（O R）为另一近邻对。用此法最后得到的树如图 5 － 1 5 b 所示。该树的拓扑图与图 5 － 1 5 a 中的图等同。然而请注意，在此法中最先相互结合成对的是 O 和 R，而不是 H 和 C。这是因为，在无根树中 O 和 R 事实上就是近邻。图 5 － 1 5 b 中的各枝长由更早些时给出的方法估出。

最后，我们看看最节省法。为简单起见，我们只考虑人、黑猩猩、大猩猩和马来猩猩。表 5 － 4 列出了包括 η－珠蛋白假基因及其周围区域的、长 10.2kb 的片段中，有关的信息位点(Koop 等,1986a;Miyamoto 等,1987; Maeda 等,1988)。一位点所支持的假定在最后一列中给出。如果我们只考虑碱基改变，则 1 5 个信息位点中，有 8 个支持人—黑猩猩进化枝（假定Ⅰ），4 个支持黑猩猩—大猩猩进化枝（假定Ⅱ），3 个支持人—大猩猩进化枝（假定Ⅲ)。而且，4 个涉及裂缝的信息位点全部支持人—黑猩猩进化枝。因此，人—黑猩猩进化枝被选作最可能是真实系统发育方式的代表。在一次更详细的分析中，威廉姆斯和古德曼（Williams 和 Goodman,1989）曾证明，有关支持人—黑猩猩进化枝的结果，有按 1 %水平衡量的统计意义。

表 5 － 4　人、黑猩猩、大猩猩和马来猩猩的顺序中的信息位点

位点[a]	顺序				受支持的假定[b]
	人	黑猩猩	大猩猩	马来猩猩	
来自 Miyamoto 等(1987)的数据					
34	A	G	A	G	Ⅲ
560	C	C	A	A	Ⅰ
1287	*[c]	*	T	T	Ⅰ
1338	G	G	A	A	Ⅰ
3057－3060	* * * *	* * * *	TAAT	TAAT	Ⅰ
3272	T	T	*	*	Ⅰ
4473	C	C	T	T	Ⅰ
5153	A	C	C	A	Ⅲ
5156	A	G	G	A	Ⅲ
5480	G	G	T	T	Ⅰ
6368	C	T	C	T	Ⅲ
6808	C	T	T	C	Ⅲ
6971	G	G	T	T	Ⅰ
来自 Maeda 等(1988)的数据					
127－132	* * * * * *	* * * * * *	AATATA	AATATA	Ⅰ
1472	G	G	A	A	Ⅰ
2131	A	A	G	G	Ⅰ
2224	A	G	A	G	Ⅲ
2341	G	C	G	C	Ⅲ
2635	G	G	A	A	Ⅰ

自 Williams 和 Goodman(1989)修改而成。　a、位点序数对应于原始来源中给出的序数。所用序列的总长度为 10.2 kb, 约为表 5-2 中所用序列的 2 倍。　b、假定：Ⅰ，人和黑猩猩在一个进化枝中；Ⅱ，黑猩猩和大猩猩在一个进化枝中；Ⅲ人和大猩猩在一个进化枝中。　c、一个星号表示在该位点上缺失一个核苷酸。

另外还有些类型的分子数据也支持人和黑猩猩聚类的结论。例如，ＤＮＡ－ＤＮＡ杂交数据（Sibley 和 Ahlguist,1984;Caccone 和 Powell,1989)明确地表示，在图 5 － 1 4 里 3 种供选取的系统发育关系中，人和黑猩猩聚被类成一个进化枝。于是，与人关系最近的现存亲属是两种黑猩猩，其后为大猩猩、马来猩猩和 9 种长臂猿。

5.10 线粒体和叶绿体的内共生起源

基本上有两种理论解释真核生物中与细胞核分开的线粒体和叶绿体基因组的存在。　第一种理论（例如，　Cavalier-Smith,1975）认为，细胞器的基因是自体起源的，是通过直接的父子关系从细胞核基因传下来的。细胞核基因组的某些部分被并入到一个由膜包被的细胞器中，接着假定它的准独立存在，从而完成了这一起源过程。与其不同的是，内共生学说（例如 Margulis,1981）认为核外ＤＮＡ的起源是外源性的。按照这种看法真核生物的祖先吞噬了原核生物，接着，由于互利或有共生关系的缘故，后者被保留了下来。随着时间的流逝，这种内共生通过某些基因的丢失而变得更为密切，以至最后成了强制性共生者（即，不能在其宿主体外独 立地存在的生物)。

现在，分子证据绝对地支持内共生学说。使叶绿体和原核生物的基因组与真核生物的基因组区分开来的生化特性，列于表 5 － 5 中。然而，最根本的支持来自 rRNA 顺序数据。由于 rRNA 顺序有较低的替换速率，所以，它们已被证明对处理涉及非常古老的进化分岐事件的问题是很有用的。

施瓦茨和科塞尔（Schwarz 和 Kossel,1980）证明，来自玉米(Zea mays)叶绿体中的 16SrRNA，其核苷酸顺序与来自大肠杆菌（E. coli）的 16SrRNA 的极为相似。在细胞核的和叶绿体的 rRNA 顺序间，这种类似程度就要低得多。来自光合蓝细菌的１６ＳrRNA 顺序的详细分析（Giovannovi 等，1988）　支持这样的看法：即，绿色的叶绿体来自一群称为蓝细菌的光合细菌（Bonen 等,1979）（见图 5 － 1 6 a)。

表５－５ 将叶绿体和原核生物的基因组与真核生物的核基因组区别开的分子特征

1. 无组蛋白 DNA
2. 大小为 120,000－150,000 碱基对
3. 环状基因组
4. 转录的利福平敏感性
5. 核糖体受链霉素、氯霉素、spectromycin 和巴龙霉素抑制
6. 翻译对放线菌酮的不敏感性
7. 翻译以甲酰甲硫氨酸为起始
8. mRNA 的多聚腺苷缺失或很短
9. 原核促进子结构

rRNA 顺序的系统发育分析表明，线粒体来自紫色细菌的α分枝（图５－１６ａ；Cedergren 等,1988)。然而，所用的细胞核 rRNA 顺序指出，高等植物在大约与动物和真菌相同的时刻发生分枝（图５－１６ｂ），这与传统观点一致；而线粒体 rRNA 顺序却指出，高等植物在非常接近紫色细菌的根处聚类，与真菌、绿色藻类和动物则是分离的（图５－１６ａ）。这是与传统观点，即把高等植物和绿色藻类组合在一个进化枝中的观点相矛盾的。为此缘故，格雷等（Gray 等,1989)曾提出，高等植物的线粒体中的 rRNA 基因，有比其他生物线粒体中的 rRNA 基因更近的进化起源。此假说是否会被进一步的证据所支持，我们将拭目以待。

5.11 分子古生物学

从微克量保存下来的组织中得到的不纯样本的ＤＮＡ，现在已有可能对它的片段加以定序了。所采用的方法是聚合酶链式反应法（ＰＣＲ)。ＰＣＲ是通过应用两种已知引物来使某序列混合物中的一种独特序列扩增的方法（图５－１７；Saiki 等,1985,1988；Scharf 等,1986;Engelke 等,1988)。每一引物和一个ＤＮＡ的互补小段附着，从而引发ＤＮＡ多聚酶的结合，然后拷贝该片段。由于每一新产生的拷贝都可充作进一步复制时的模板，所以靶片段的拷贝数将呈指数增长。用此法有可能将混合物中的某一被选定的ＤＮＡ片段合成产生许多拷贝，此时混合物中其他ＤＮＡ序列可能会远为超量地存在（Kocher 等,1989)。因此，采用ＰＣＲ我们可以从博物馆标本，象保存下来的有机物质（主要是皮肤和肌肉)、严重损坏的考古学遗物甚至骨骼中找回某些特别的ＤＮＡ序列（Hagelberg 等,1989)。
应用此法现在已可能建立某些已灭绝物种，象南非斑驴和澳大利亚袋狼间的系统发育关系，或决定斯堪的那维亚铁器时代的沼泽地人和埃及木乃伊，这些已灭绝人类群体间的祖先 后裔关系。对后两个人类群体的形态学比较，得到的结果比较含糊。

应用ＰＣＲ法，托马斯等（Thomas 等,1989)已能对来自袋狼 Thylacinus cynocephalus 和其他澳大利亚与南美有袋类的线粒体的２１９个核苷酸加以定序和比较，并且与来自有胎盘类的同源顺序比较。应用此法，他们已能在以下两种说法间作出取舍：（ａ）袋狼与南美的有袋动物类群有亲缘关系，和（ｂ）袋狼与其他澳大利亚有袋类的亲缘关系很近。从这些顺序比较中得出的结论是，袋狼与另外两种澳大利亚有袋类，几近灭绝的塔斯马尼亚魔鬼（一种袋獾 Sarcophilus harrisii)和澳大利亚虎猫（一种袋鼬， Dasyurus maculatus) 有很近的亲缘关系,但与一种南美有袋类,负鼠(一种毛鼹， Philander opossum andersoni) 只有较疏远的亲缘关系（图５－１８)。于是，袋狼和南美有袋动物间形态学上的相似，被认为是一个形态学水平上趋同进化的例子，它在线粒体ＤＮＡ中没有平行关系。

5.12 深色海滩雀：物种保护生物学中的一次教训

最后一只深色海滩雀死于１９８７年６月１６日，地点在佛罗里达州俄伦多附近沃特. 迪斯尼世界的动物园里。深色海滩雀在１８７２年被发现。因其黑色的外貌而被明确地归类成一个亚种（ Ammodramus maritimus nigrescens)。A.m.nigrescens 的地理 分布被限制在佛罗里达州布列伐得县中的盐碱沼泽

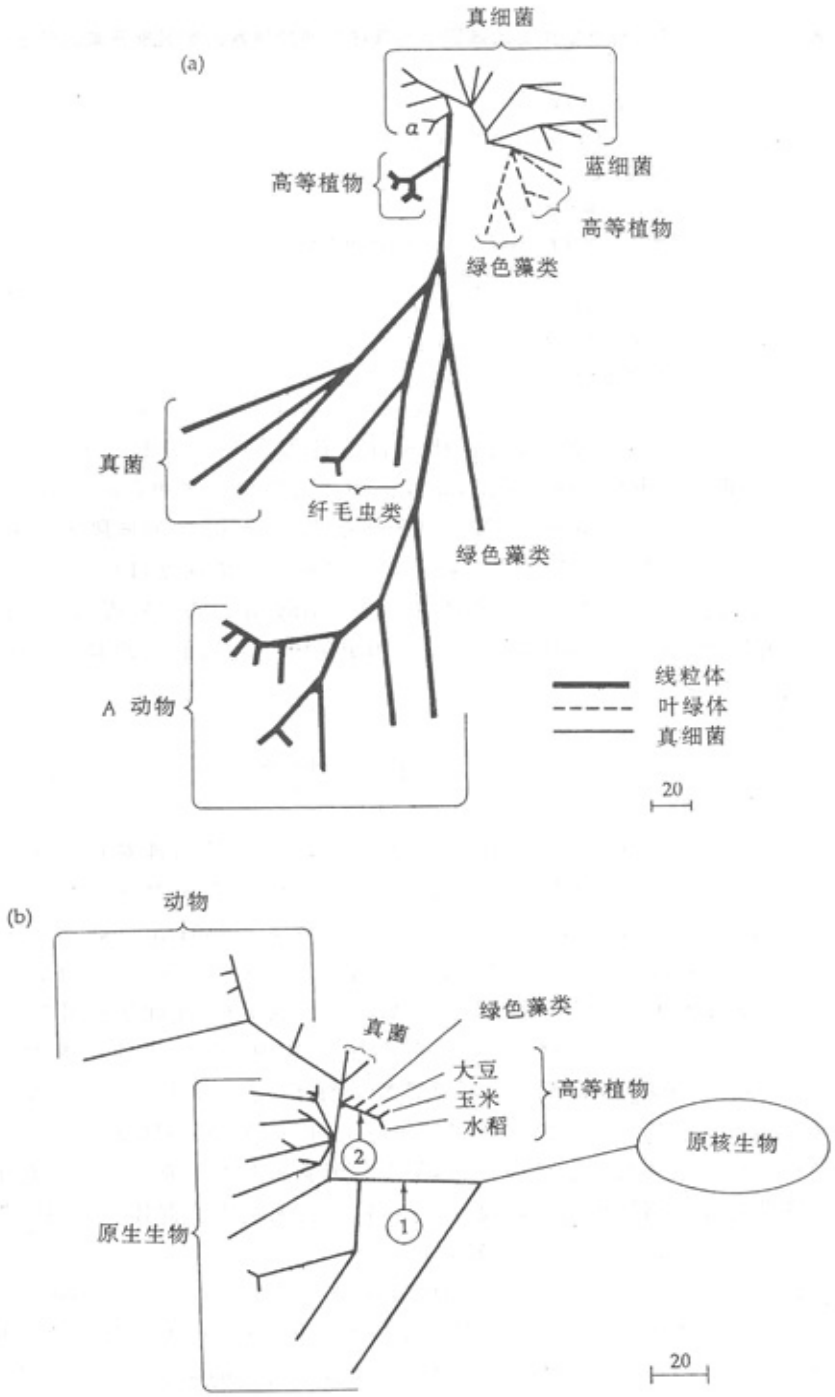


图 5－16 （a）从小亚基 rRNA的顺序推测出的无根树中真细菌—叶绿体—线粒体部分。注意，来自绿色藻类和植物的叶绿体顺序都是单源发生的，而线粒体顺序则是多源发生的。（b）从细胞核小亚基 rRNA顺序推测出的无根树。步聚 1（用圆圈标出）是早期内共生、假定已产生了大多数真核生物的线粒体基因组，步聚 2 是后期共生，Gray 等(1989)假定它贡献了高等植物线粒体的 rRNA基因。每一分枝的长度与其两个内部点间的替换数成比例，比例尺度如图中所示。自 Gray 等(1989)修改而成。

地带（图5－19），到了1980年，自然界中只能找到6个个体，全是雄性。显然该群体的灭绝已命中注定，于是，一个人工杂交计划被当作为保护该亚种的基因的最后一搏而展开了。

在这种情况下，保护方案就是将这些濒临灭绝的亚种的雄性个体，与来自我们所能找到的与其亲缘关系最近的亚种的雌体交配。然后，子一代的杂种雌体再与该雄体回交，它们产生的后代又再次与原雄体回交，此过程循环往复直至该雄体死去为止。该实验的关键是要决定应从哪种群体中选取雌体，即决定哪一个亚种在系统发育上最接近该濒临灭绝的亚种。

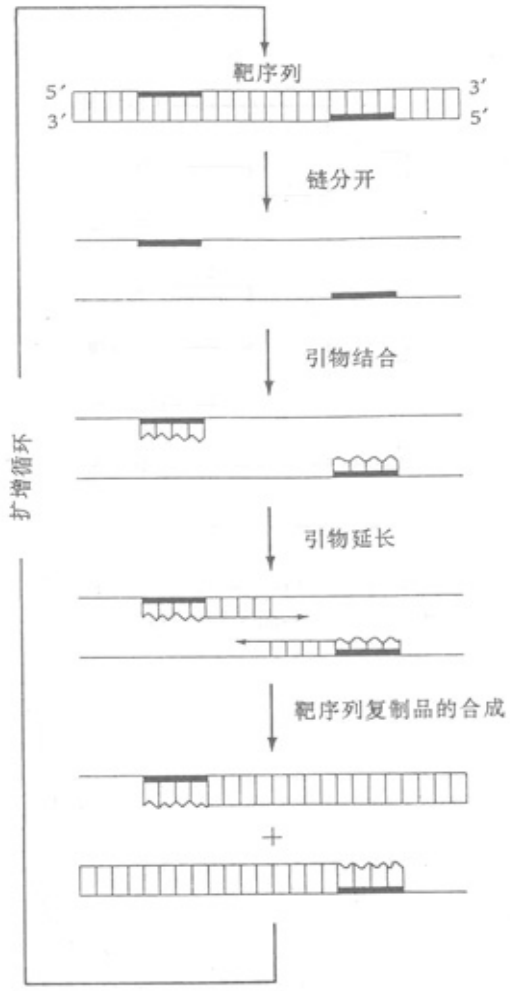


图 5 - 1 7 用聚合酶链式反应法（ＰＣＲ）使混合物中的特定ＤＮＡ序列扩增。ＤＮＡ经加热解链成单链分子。两个合成的ＤＮＡ小片段（锯齿状波纹线）各与靶序列（黑色粗线）一端上的特别顺序互补，充作引物。每一引物各与不同链上的互补顺序结合。然后，ＤＮＡ多聚酶通过在引物上添加核苷酸来使其延长。很快地，靶序列的精确复制品即产生了。在以后的循环中，原始靶序列及其复制品都可充作模板。相反，不含与引物互补的顺序的ＤＮＡ分子则不被扩增。因此，不纯的ＤＮＡ混合物也可用来只对其中的一种序列进行扩增，即使在其他序列远为过量地存在的情况下，这一点也能办到。更详细说明见 Mullis(1990)

在海滩雀的例子中，有 8 个可区别的亚种供选取。这些物种的地理分布范围如图 5 — 1 9 所示。根据形态学和行为特征所作出的决定是，最接近深色海滩雀的亚种为斯科特的海滩雀（ A.m.peninsulae ），它们栖息在佛罗里达湾的沿岸。在此决定下， nigrescens 亚种的几个雄体与来自 peninsulae 亚种的雌体交配。已取得了两次成功的回交，而由此产生的群体则此后维持近交，以期有一天能将这“重建”的亚种放回到它原来的栖息地中去。

为了看一看该雌体的选择是否正确，阿维斯和纳尔逊（Avise 和 Nelson,1989）曾进行过有关线粒体ＤＮＡ的限制酶模式的比较，材料的一方来自已故的纯种 nigrescens 的标本，另一方来自海滩雀现存 8 个亚种中的 5 个，共３９个个体。他们选择线粒体ＤＮＡ有几个原因。首先，脊椎动物中线粒体ＤＮＡ已知进化非常迅速（第四章），因此它可以为区别亲缘关系较近的生物提供一个分辨率较高的解决途径。其次，线粒体是母系遗传的，所以不会发生因等位基因分离而造成的复杂局面。最后，由于母系传递模式来自已故雄性深色海滩雀中的线粒体ＤＮＡ，并不曾转移到复原育种方案中的杂种个体里。所以，虽然细胞核中的某些基因已在杂种中生存了下来，但深色海滩雀的线粒体基因却不一样，它们是真正的灭绝了。

根据限制酶模式，阿维斯和纳尔逊（Avise 和 Nelson,1989）用ＵＰＧＭＡ法和最节省法，构建出了海滩雀几个亚种间的进化关系（图 5 — 2 0 ）。从图中可以看到，大西洋沿岸群体，包括深色海滩雀

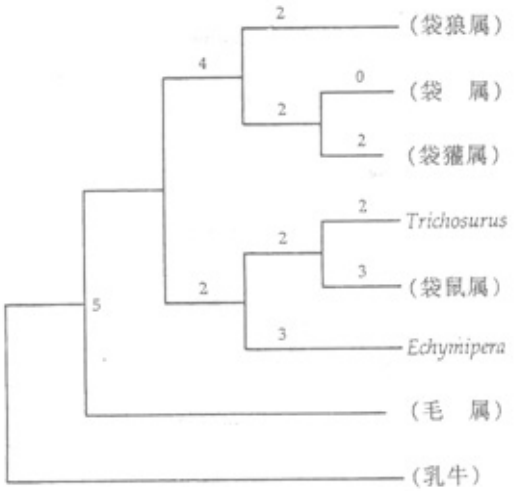


图 5 - 1 8 袋狼（　Thylacinus　）和 6 种其他有袋类根据线粒体 1 2　S rRNA 顺序作出的最节省树。该树用乳牛（　B o s　）作组外单位找到了根. 分枝上的数字代表替换数。自 Thomas 等(1989)。

图 5 - 1 9 海滩雀 (Ammodramus maritimus) 的 9 个在分类学上有区别的亚种的地理分布。自 Avise 和 Nelson(1989)。

在内，相互间几乎是不可区别的。在此研究中 3 个海湾沿岸的亚种也有同样的关系。相比之下，大西洋沿岸亚种与海湾沿岸亚种间则有相当明显的区别。这两大类群间的每位点核苷酸替换数估计约为 1 %。如果海滩雀的线粒体 D N A 以与哺乳类和其他鸟类的线粒体 D N A 大致相同的速率进化(即每百万年 2 — 4 % 的顺序分歧)，那么，这两大类群的群体是在大约 2 5 万— 5 0 万年前发生分离的。虽然这些对分歧的绝对年代的估值，由于缺乏确定的统一标准而停留在定性水平，但它们却与已得到的海平面开始下降的年代一致得相当好。由于海平面下降，佛罗里达半岛露出水面，从而对半岛两侧的群体起了生殖屏障的作用。

更重要的是，阿维斯和纳尔逊 (Avise 和 Nelson, 1989)的分子研究表明，深色海滩雀亚种与另两个大西洋亚种 (即　A.m.maritima 和 A.m.macgillivraii)在分子上不可区别，而它与海湾亚种，象 A.m.peninsulae　却是相当不同的，但繁殖计划中却从后者里选出雌体。结果，深色海滩雀的救援方案也许是建立在错误的系统发育前提上的。该方案可能不是在重建一个灭绝了的亚种，而是创造了一个新亚种。所以，系统发育关系方面的知识在生物多样性的保护中，对作出合理决定是至关重要的。分类学上的失误也许会使那些愿望良好的努力变得劳而无功。

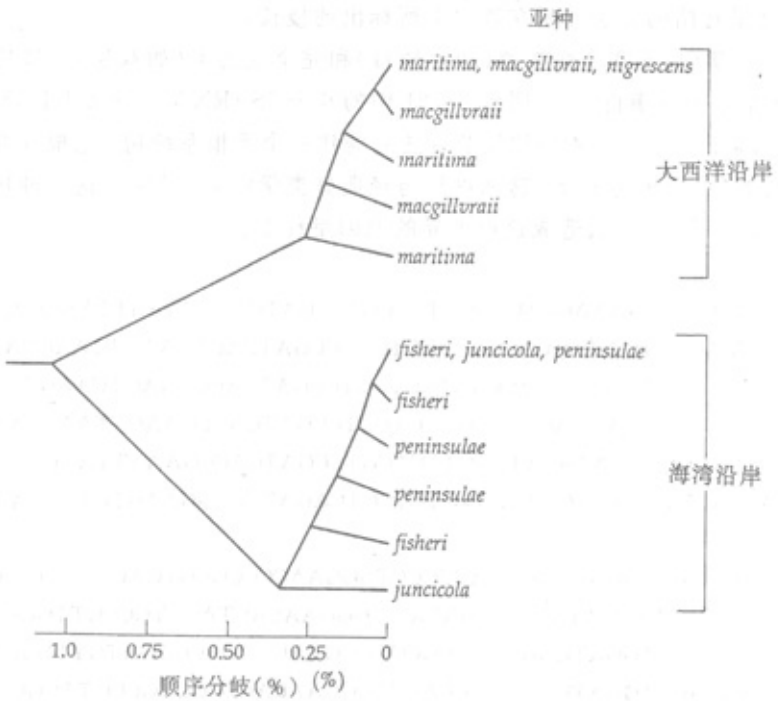


图 5-20 UPGMA 系统树图，表示海滩雀的大西洋沿岸群体和海湾沿岸群体间的线粒体ＤＮＡ基因型区别。应用最节省法可得到许多同样节省的树，其中包括与所示图等同的树。所有供选取的树都涉及大西洋进化枝内部或海湾进化枝内部少量的分枝重排，但这两个类群间的界线却维持不变。系统树图中同样的亚种名多次出现，这表示属于同一亚种的不同个体展现出不同的限制酶模式。反之，几个亚种名在一个分枝的末端处出现则表示，根据形态学和动物地理学而被归于不同亚种的某些个体，对所使用的那些酶表现出相同的限制酶模式。自 Avise 和 Nelson(1989)。

习题

1　对４个ＯＴＵ：Ａ，Ｂ，Ｃ和Ｄ，画出所有可能的有根和无根树。

2　找出以下５个假想顺序中的所有信息位点：

12345678

(a) ＡＴＧＡＣＴＡＡ

(b) ＧＴＧＡＴＴＧＡ

(c) ＡＣＧＧＡＴＡＡ

(d) ＡＴＧＣＡＴＴＡ

(e) ＡＣＧＣＡＴＣＡ

3　（ａ）用ＵＰＧＭＡ法，（ｂ）用变形距离法，和（ｃ）用萨塔特和特韦尔斯基的近邻关系法，对以下距离矩阵构建系统树。在变形距离法的场合，假定ＯＴＵ　Ｅ对所有其他ＯＴＵ而言是一个已知的组外单位。

	OTU			
OTU	A	B	C	D
B	3			
C	8	7		
D	7	6	3	
E	11	10	13	12

4　图５－１５ｂ中的树是用萨塔特和特韦尔斯基的近邻关系法得到的（见第 66—67 页和 74—75

页)。请用第65—66页里介绍的方法证明在该树上所标出的枝长。

5　在经典昆虫分类学中，不完全变态类（如直翅目）和完全变态类（如双翅目、鳞翅目，同翅目）间的分歧曾被假定是非常远古的事件。（ａ）用图５－２１中的６个５Ｓ　rRNA顺序和ＵＰＧＭＡ法来构建系统树。（ｂ）用其中5个顺序（ａ、ｂ、ｃ、ｅ和ｆ）和最节省法来构建一个无根系统树。把根放在联结组外单位　Artemia salina　(f)和其他OTU的分枝上。这两种树与经典分类学结果一致吗？这两种构建法结果能得到相同的拓扑图吗？如果不同，那么造成这种差异的原因是什么？

(a) GCCAACGTCCATACCACGTTGAAAGCACCGGTTCTCGTCCGATCACCGAAGTTAAGCAGC
(b) GGCAACGACCATACCACGTTGAATACACCAGTTCTCGTCCGATCACTGAAGTTAAGCAAC
(C) GCCAACGTCCATACCACGTTGAAAACACCGGTTCTCGTCCGATCACCGAAGTCAAGCAAC
(d) GCCAACGTCCATACCACGTTGAAAACACCGGTTCTCGTCCGATCACCGAAGTTAAGCAAC
(e) GCCAACGACCATACCACGCTGAATACATCGGTTCTCGTCCGATCACCGAAATTAAGCAGC
(f) ACCAACGGCCATACCACGTTGAAAGTACCCAGTCTCGTCAGATCCTGGAAGTCACACAAC

(a) GTCGGGCGCGGTTAGTACTTGGATGGGTGACCGCCTGGGAACCCCGCGTGACGTTGGCA
(b) GTCGGGCGTAGTTAGTACTTGGATGGGTGACCGCTTGGGAACACTACGTGCCGTTGGCA
(c) GTCGGGCGTAGTCAGTACTTGGATGGGTGACCGCCTGGGAACACTACGTGATGTTGGCT
(d) GTCGGGCGCGGTCAGTACTTGGATGGGTGACCACCTGGGAACACCGCGTGCCGTTGGCT
(e) ETCGGGCGCGGTTAGTACTTAGATGGGGGACCGCTTGGGAACACCGCGTGTTGTTGGCC
(g) GTCGGGCCCGGTCAGTACTTGGATGGGTGACCGCCTGGGAACACCGGGTGCTGTTGGCA

图５－２１　来自6种节肢动物的５ＳrRNA基因的ＤＮＡ顺序。（ａ）蝗虫（Acheta　domesticus，直翅目）（ｂ）蚜虫（Acyrthosiphon magnoliae　，同翅目）（ｃ）家蚕（Bombyx mori　，鳞翅目）(d) 蛾(Philosamia cynthia　，鳞翅目)(e)果蝇（Drosophila melanogaster　，双翅目）（ｆ）甲壳动物（　Artemia salina　，甲壳纲）。资料来自Kawata和Ishikawa（1982），Morton和Sprague(1982)，Gu等(1982)，Bagshaw等(1987)，Cave等(1987)，以及Samson和Wegnez(1988)。

后继阅读文献

Felsenstein, J. 1988. Phylogenies from molecular sequences: Inference and reliability. Annu. Rev. Genet. 22:521-565

Goodman, M. (ed.) . 1982. Macromolecular Sequences in Systematic and Evolutionary Biology. Plenum, New York.

Hillis, D. M. and C. Moritz (eds.). 1990. Molecular Systematics. Sinauer Associates , Sunderland, MA.

Margulis, L. 1981. Symbiosis in Cell Evolution. Freeman, San Francisco.

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

Sneath, P. H. A. and R. R. Sokal. 1973. Numerical Taxonomy. Freeman, San Francisco.

Woese, C. R. 1987, Bacterial evolution. Microbiol. Rev. 51:221-227.