第 1-6 章 系统发生树构建

自 20 世纪中叶,随着分子生物学的不断发展,进化生物学也进入了分子水平,并建立了一套依赖于核酸、蛋白质序列变异的理论和方法,由此也开创了生物信息学新领域。如本书绪论所述, Pauling 等第一次(1962 年)将蛋白质序列变异用于分子进化研究,标志着生物信息学学科的起始。

第一节 系统发生树与遗传模型

一、系统发生树概述

分类学涉及的问题是将不同生物合理地分成不同的类群,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,分类涉及到系统发生(发育)或系统进化的重构(reconstruction of phylogenies),构建系统发生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 等人已对构建系统发育过程进行了全面的总结(见《分子进化与系统发育》(Molecular Evolution and Phylogenetics)一书),本章仅简要介绍相关方法。

不同生物表型(phenotype)和基因型(genotype)数据有着明显差异。Sneath 和 Sokal (1973)将表型性关系定义为根据生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遗传信息,这两种关系可用系统进化树或系统发生树(phylogenetic tree)来表示。Nei (1987)指出,如果表型相似性的尺度意味着进化上的相似性程度,则有关表型的方法就可以提供遗传上的判断。系统发生树分有根(rooted)和无根(unrooted)树(图 1-6.1 给出 4 个物种部分有根树和无根树)。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。通常用 Newick 格式来对系统树进行文本表示,有根树的文本表示是唯一的,而无根树可以通过指定不同树根位置进行文本表示,因此其文本表示也就不是唯一的。

用于构建系统树的数据有两种类型: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。这些数据可以以矩阵的形式表达。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

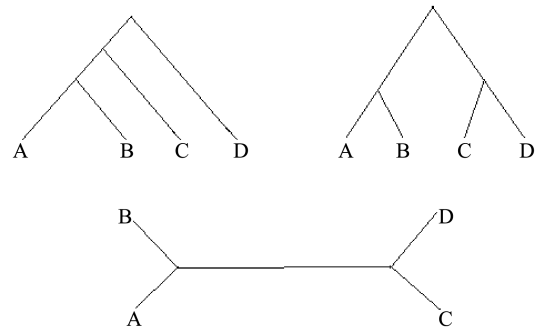

图 1-6.1 四个物种(A、B、C 和 D)的 2 种有根树和 1 种无根树形式

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。

1. 系统发生树构建方法

系统发生树的构建并非易事。随着用于构建发生树的个体或称为实用分类单位(OTU, operational taxonomic units)数量的增加,可能的树形数量将以 2^S 数量级增加(S = OTU 数量)。当 OTU 数量为 5 时,可能的有根树和无根树树形分别为 105 和 15 个,OTU 数量为 10 时,可能的树形分别为 34 459 425 和 2 027 025 个,而当 OTU 数量增加到 20 时,可能的树形分别约为 $8.20e^{21}$ 和 $2.22e^{20}$ 个。如果我们考虑树的枝长(进化距离)的话,问题将变得更加复杂。同时,我们实际研究中,OTU 数量往往不止 20 个,这就使可能的树形数量非常巨大。如何在如此众多的有根和无根树树形中,确定最佳树形或最优树形?

系统发生树的构建主要有三种方法:距离矩阵法、最大简约法和最大似然法。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。最大简约法(maximum parsimony, MP)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。由于该方法将基于数据测算所有树形的可能性,计算量很大,一般超过 12 个 OTU 就难以建树。对于模型的巨大依赖性是最大似然法(maximum likelihood, ML)的特征,该方法在计算上繁杂,但为统计推断提供了良好基础。该方法特别适用于那些序列间差异非常明显的进化分析,同时它可以利用不同进化模型构建最佳系统发生树。综上所述,三种方法在实际分析中的适用性总结如图1-6.2:

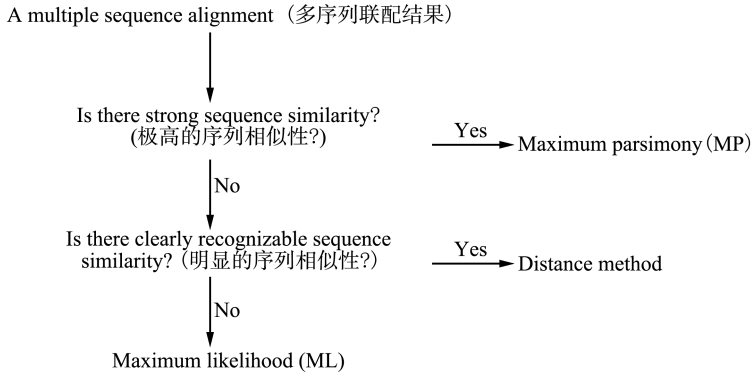

图 1-6.2　三种建树方法适用性

2. 树形统计测验

当我们构建了一个系统发生树,有多大把握认为它的结构反映了真实的进化关系呢?这就涉及到发生树的稳健性和可靠性问题。由于发生树具有复杂的结构,很难将传统构建置信区间及进化统计显著性假设检验等方法照搬过来,往往采取再抽样统计检验。

统计学教材介绍再抽样方法时,常常用这样一个例子:坛子里面装有大量颜色不同但质地手感相同的小球。要求用再抽样的方法估计坛子中各种颜色小球的比例。抽样检验的步骤如下:

(1) 摇晃罐子使其中小球分布均匀(随机化);

(2) 闭眼从坛子里抽取 100 个小球(随机采样);

（3）清点并记录刚才取出的各种颜色小球的数目；

（4）把刚才采样取出的小球全部放回坛子里；

（5）继续从第一步开始操作；上述操作总共循环 1 000 次以上。

对记录下来的数字求平均值和相对比例，最终得出坛中各色小球分布的估计。

由于第(4)步要求把取出的小球送回坛子里去，这一方法被称为“自举法”(bootstrap)。如果把第(4)步改成：“把采样取出的小球均扔掉”，那相应的方法称为“刀切法”(jackknife)。如果坛子中的小球总数不够多(采样空间有限)，人们就只能“自举”，不敢“刀切”。这是文献中“自举”比“刀切”法常见的缘由。

“自举法”是推断进化树可靠性的常用方法，简单来说当序列长度为 m 时，把序列的位点都重排，进行 m 次有放回的抽样，然后将这些抽样得到的新的 m 列数据，重新使用相同的方法构建得到 bootstrap 树，并重复一定次数(如 1 000 次)。对于各种树形中可靠的分枝，必定有大量重排数据支持这一分枝，如 95%甚至更高比例的支持率。“自举法”通常需要较大的计算量，特别是对于似然法建树。另外，也有其他的统计测验方法，例如 Kishino 与 Hasegawa 提出的一种基于似然度比较两个候选进化树的 KH 近似检验方法以及 Shimodaira 与 Hasegawa 提出的 SH 检验。

3. 主要建树软件

目前构建系统进化树的算法及其软件很多，如维基百科中列举的就多达 50 种以上(见 https://en.wikipedia.org/wiki/List_of_phylogenetics_software)。我们列出了几个目前常用的软件(表 1-6.1)，其中尤以 MEGA 最为流行。MEGA 软件目前已更新至第 7 版，包含 Windows、Unix 和 Mac 等多个操作系统下的软件包，当用户需要构建进化树时，只需要将原始序列输入至 MEGA 界面，选择特定的联配方法(如 CLUSTALW、MUSCLE)进行联配，同时用户也可直接输入已经联配好的序列进行建树。MEGA 软件提供了多种建树方法包括最大似然法、最大简约法、距离矩阵法，可以选择 Bootstrap 法测验，每个方法也提供了不同进化模型供选择使用。

表 1-6.1 分子进化与系统发育主要分析软件

软件名称	网址	说 明
MEGA	http://www.megasoftware.net/	美国宾西法尼亚州立大学 Masatoshi Nei 开发的分子进化遗传学软件
PHYLIP	http://evolution. genetics. washington. edu/phylip.html	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
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
PAUP	http://paup.csit.fsu.edu/	国际上最通用的系统树构建软件之一，美国 Simthsonion Institute 开发
RAxML	http://sco.h-its.org/exelixis/web/software/raxml/index.html	大量数据的最大似然法建树常用方法
MrBayes	http://mrbayes.sourceforge.net/	基于贝叶斯方法的建树工具

二、遗传模型

当我们说两条序列为同源序列(homologous sequence),意味着它们有共同的祖先。同源序列的来源主要通过物种分化、基因和基因组片段水平倍增等机制产生。同源基因包括两种类型——旁系(paralogous)和直系(orthologous)同源基因(图 1-6.3),其中旁系同源基因由同一个物种内的基因倍增而来,而直系同源基因是指物种分化后产生的同源基因。

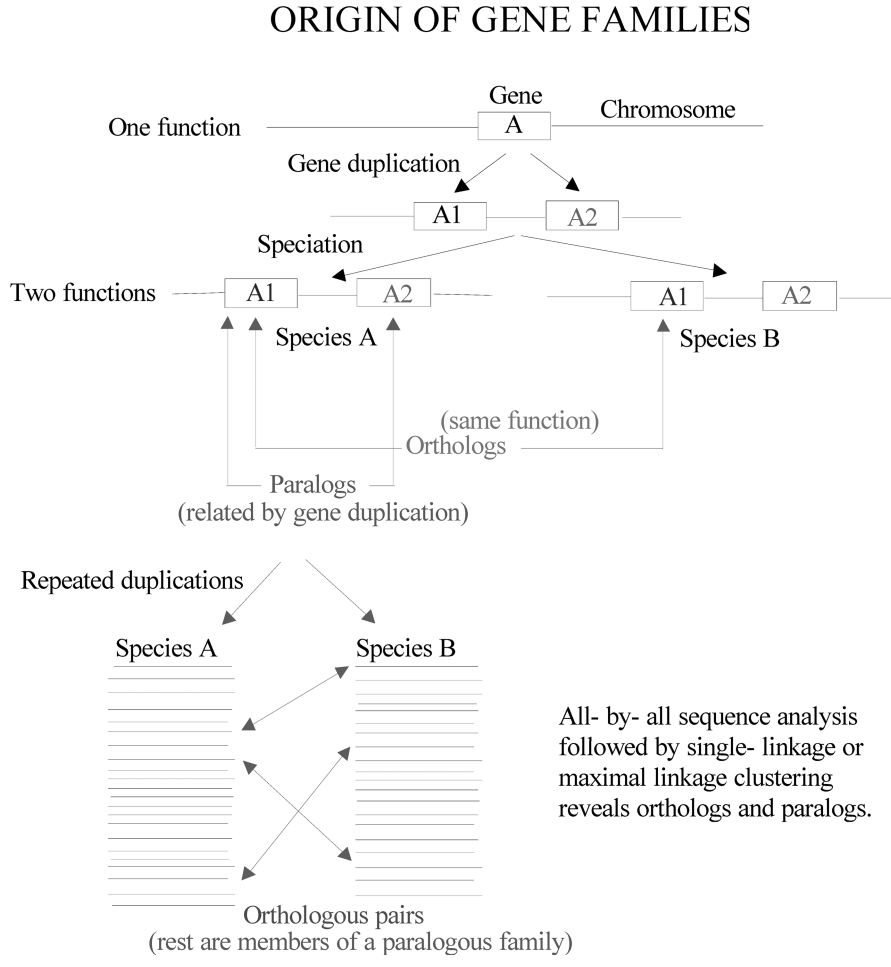

图 1-6.3　旁系同源基因(paralog)和直系同源基因(ortholog)的产生机制

在分子进化研究中,我们往往假定两条序列同源。这样它们就具有单一祖先序列,而这一祖先序列在进化过程中,会发生一系列的核苷酸突变(图 1-6.4)。

个体发生碱基变异后,新基因型可能在群体内扩散。对环境适应性具有优势的变异,在自然选择下,新基因型比例会明显上升,甚至在群体内固定下来。因此,在群体水平上,在旁系同源基因上可以发现大量单碱基变异,当该变异基因型的频率很低(<1%),我们一般认定其为新的突变(mutation),但其频率超过 1%后,我们称其为单碱基变异多态性(SNP)。当然除了碱基变异,序列还会在进化过程中发生碱基的插删即插入与删除(缩写为 Indel)。例如,我们测定一段来自栽培稻(*Oryza sative*)两个亚种(*indica/japonica*)及其祖先野生种

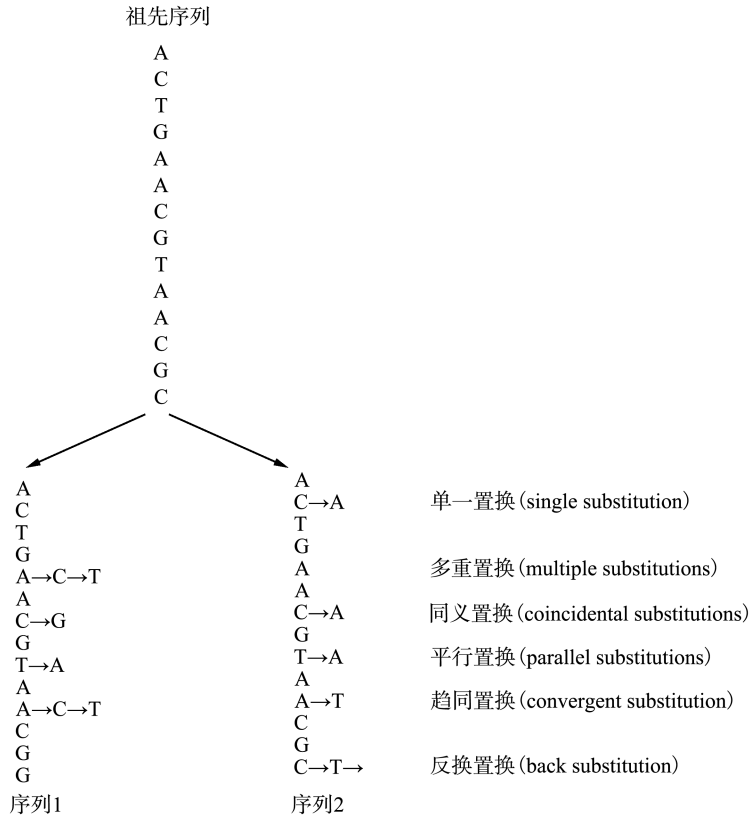

图 1-6.4 同源序列间的核苷酸变异机制

(*O.rufipogon*和 *O.nivara*)基因片段,我们会发现不少碱基变异情况,如 SNP 和 Indel 等(图 1-6.5)。分子进化中,由于碱基变异遗传模型等研究比较清楚,目前主要用 SNP 数据构建系统发生树。

*indica*_1	ATG CGG GAT CCA TTC CTT AAT GAG TTT CCT AAA ACG GTG CAG CAC GGT TTT
*indica*_2	ATG TGG GAT CCA TTC CTT AAT GAG TTT CCT GAA ACG GTG CAG CAC GGT TTT
*indica*_3	ATG TGG GAT CCA TTC CTT AAT GAG TTT CCT GAA ACG GTG CAG CAC GGT TTT
*japonica*_1	ATG TGG --- CCA TTC CTT AAT GAG TTT CCT GAA ACC GTG CAG CAC GGT TTT
*japonica*_2	ATG CGG --- CCA TTG CTT AAT GAG TTT CCT GAA ACC GTG CAG CAC GGT TTT
*japonica*_3	ATG TGG GAT CCA TTG CTT AAT GAG TTT CCT GAA ACC GTG CAG CAC GGT TTT
*O.rufipogon*_1	ATG TGG GAT CCA TTC CTT AAT GAG TTT CCT GAA ACG GTG CAG CAC GGT TTT
*O.rufipogon*_2	ATG TGG GAT CCA TTG CTT AAT GAG TTT CCT GAA ACC GTG CAG CAC GGT TTT
*O.rufipogon*_3	ATG TGG GAT CCA TTG CTT AAT GAG TTT CCT GAA ACC GTG CAG CAC GGT TTT
O.nivara	ATG TGG GAT CCA TTC CTT AAT GAG TTT CCT GAA ACG GTG CAG CAC GGT TTT

图 1-6.5 水稻及其祖先野生种碱基变异情况

其中 3 个位点我们定义为 SNP 位点,一个位点发生插删

遗传模型在系统树构建中非常重要,因为距离计算等建树过程必须在一定的遗传假设下才可能进行。目前主要遗传进化模型包括 Jukes-Cantor 模型、Kimura 模型、Felsenstein 模型和 Hasegawa-Kishino-Yano(HKY)模型。

以下主要介绍在 DNA 序列距离计算中最为常用的两个遗传模型。

在分子进化研究中,我们往往假设序列是同源的,它们具有单一祖先序列,且这一祖先

序列在进化过程中发生了一系列的核苷酸突变。在该假设基础上,Jukes 和 Cantor(1969)进一步假设每一种碱基具有同等机率突变为另外 3 种碱,其频率常数为 $\mu/3$,μ 为碱基替换频率。因此 Jukes-Cantor 模型通常也被称为单参数进化模型。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)具有不同的发生频率,a 和 β,提出一种新模型。该模型由于考虑转换率和颠换率的不同,又通常称为双参数进化模型。表 1-6.2 简要说明了以上两种遗传模型。

表 1-6.2　Jukes-Cantor 单参数模型(上三角部分)和 Kimura 两参数模型(下三角部分)

a、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	

根据以上遗传模型,Jukes 和 Cantor(1969)提出了 DNA 序列距离 K 计算公式:

$$K = \frac{3}{4}\ln\left(\frac{4}{4q-1}\right) \approx 2\mu t \qquad (1\text{-}6.1)$$

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:

$$q_t = \frac{1}{4} + \frac{3}{4}\left(1 - \frac{8\mu}{3}\right)^t \qquad (1\text{-}6.2)$$

μ 为碱基替换频率。

距离 K 适用于两条序列从一个祖先序列趋异进化以来的时间估计,并可用于序列间系统发生树的构建。在计算时,均首先需要进行序列联配分析。Kimura 在其两参数模型下证实,由于趋异变化,随时间由转换(Ⅰ型变化)或颠换(Ⅱ型变化)造成的碱基替换率为

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

$$P_{II} = \frac{1}{4}(1 - e^{-8\beta t}) \qquad (1\text{-}6.3)$$

如果 $k = a + 2\beta$ 是单位时间碱基替换的总频率,则适合作为系统发生树的距离尺度为

$$K = -\frac{1}{2}\ln[(1 - 2p_I - p_{II})\sqrt{1 - 2P_{II}}] \approx 2kt \qquad (1\text{-}6.4)$$

Kimura 以兔和鸡的 β-球蛋白序列为例(GenBank 记录号 J00860 和 J00659),计算了上述距离。序列长度 438bp,有 58 个Ⅰ型变化、63 个Ⅱ型变化。因此,$\tilde{p}_I = 0.132\,4$,$\tilde{P}_{\Pi} = 0.143\ 8$,Kimura 距离为 0.351 3。这与只根据相同碱基比例 $\tilde{q} = 0.723\,7$ 所得 Jukes-Cantor 距离 0.344 6 差异不大。

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。由于密码子的简并性,碱基变异不导致氨基酸的变异,这种情况称为同义突变(synonymous mutation),而导致氨基酸变异的突变则为非同义突变(non-synonymous mutation),因此蛋白质序列的分歧度可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-Cantor 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)。

Felsenstein(1981)模型是 Jukes-Cantor 模型的另一种推广模型。该模型满足稳态概率分布 $q_A+q_C+q_C+q_T=1$,当取 $q_A=q_C=q_C=q_T=1/4$ 时,该模型即简化为 Jukes-Cantor 模型。Hasegawa 等人 1985 年提出的 HKY 模型则是对 Felsenstein 进化模型的进一步推广,类似于 Kimura 模型对 Judes-Cantor 模型的推广,即对转换和颠换突变进行了区分。此外,核苷酸替代模型还有许多,包括 1986 年 Tavaré 提出的 GTR(Generalised time-reversible)模型、1992 年 Tamura 和 1993 年 Tamura 和 Nei 提出的模型等。

第二节 距离法

系统发生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(OTU)之间的距离。对于 t 个 OTU,每一对之间的距离矩阵列于表 1-6.3。

表 1-6.3 实用分类单位(OTU)间的距离矩阵

OTU	OTU				
	#1	#2	#3	…	#t
#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 OTU 进行表型意义的分类可借助于聚类分析,聚类过程可以看作是鉴别具有相近 OTU 类群的过程。

距离法主要包括 3 个主要方法:非加权平均连接聚类法(UPGMA 法)、Fitch-Margoliash 法和邻接法(Neighbor-joining 法)。

一、非加权平均连接聚类法(UPGMA)

非加权平均连接聚类法(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)是早期应用最广泛的一种聚类方法。该法将类间距离定义为两个类内成员所有成对距离的平均值。

作为实例,我们考虑图 1-6.6 所列的 5 条来自人类等线粒体 DNA 序列数据。每对序列间的 Jukes-Cantor 距离取决于每对序列间核苷酸替换率。根据距离 K 估计,5 条线粒体 DNA 序列的差异和距离列于表 1-6.4。

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 1-6.6 五种生物线粒体 DNA 序列

表 1-6.4　来自人类等五条线粒体序列(图 1-6.6)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	-

五种生物线粒体 DNA 序列中,人类与黑猩猩之间的距离最近(0.015),首先将它们合并为一个 OTU 新类(hu-ch)。然后计算这个新类与其它序列之间的距离,即其他序列到新类中各成员间的平均距离:

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

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

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

因此,表 1-6.4 距离矩阵可更新为:

	(hu-ch)	go	or	gi
(hu-ch)		0.037	0.135	0.189
go			0.092	0.179
or				0.179
gi				

该表中人类-黑猩猩(hu-ch)与大猩猩(go)之间的距离最小。将它们再合并为一新类(hu-ch-go),新距离计算为:

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

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

距离矩阵进一步更新为:

	(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,再将其并为一新类。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 1-6.7 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距离的中值点,成对序列间的距离为分枝长度之和。

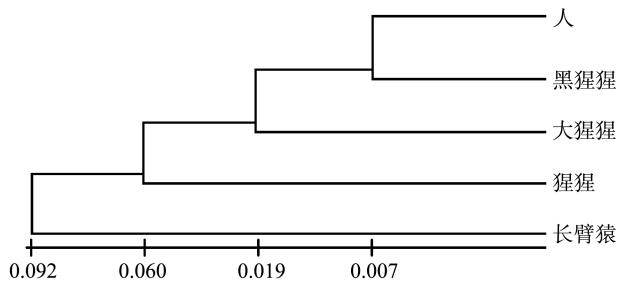

图 1-6.7 平均连接聚类法(UPGMA)系统树

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。当各分枝突变率相等时,认为分子钟(molecular clock)在起作用。因此,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,使用 UPGMA 时必须注意。另一些模拟研究已证实,当各分枝的突变率不相等时,这一方法的结果不尽人意。

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)发展的 Fitch-Margoliash 算法去除了这一假定。该法的应用过程包括插入“丧失的”OTU 作为后面 OTU 的共同祖先,并每次使分枝长度拟合于 3 个 OTU 组。同样用图 1-6.6 的线粒体资料来说明 Fitch-Margoliash 算法。

将 OTU 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 1-6.8)。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:

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

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。

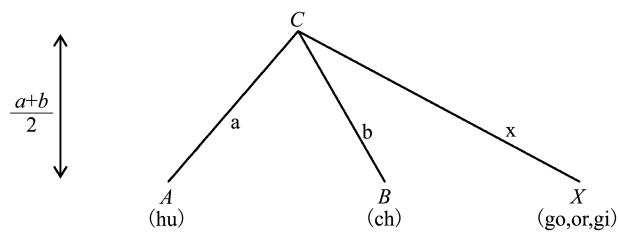

图 1-6.8 将 Fitch-Margoliash 算法应用于图 1-6.6 线粒体资料的初始步骤

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 3 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 作为新节点,按 UPGMA 所采用的方式再计算距离值,得到下一个最近的一对节点为 C 和 D(=go)。引入树节 E 作为 C 和 D 的直接祖先。如图 1-6.9 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:

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

因此

$$c=0.032, \qquad b=0.006$$

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

$$c'=0.032-0.008=0.024$$

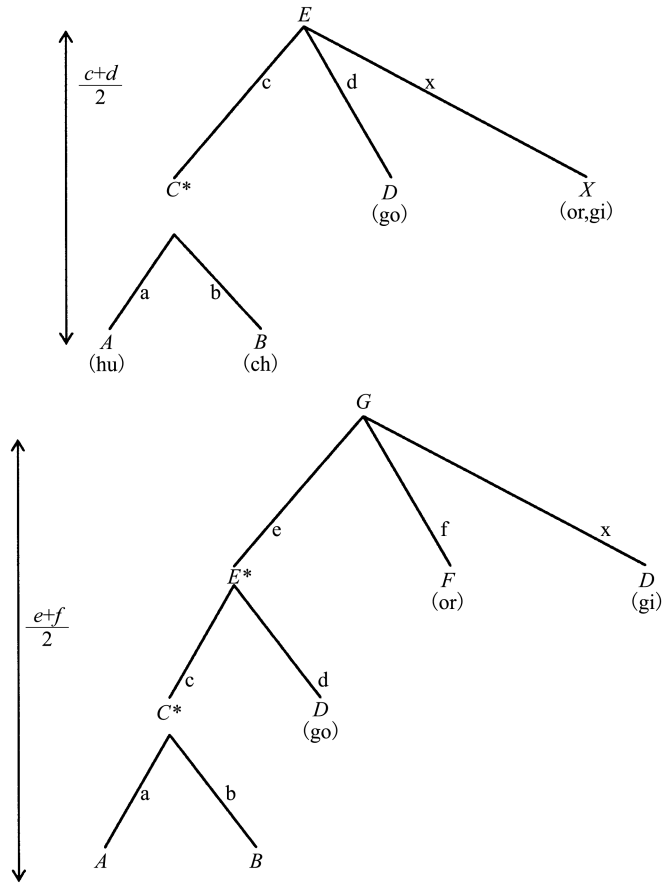

图 1-6.9　将 Fitch-Margoliash 算法应用于图 1-6.6 线粒体序列分析的中间步骤

随着 OTU 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

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

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 1-6.10 给出了其无根系统树。

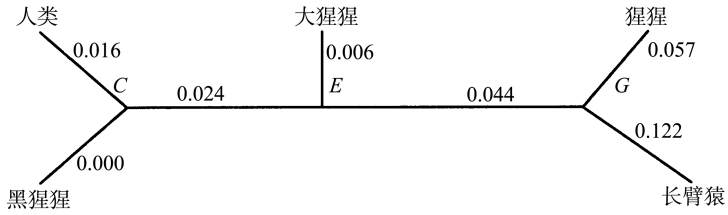

图 1-6.10 图 1-6.6 所列线粒体序列资料的 Fitch-Margoliash 无根系统树

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 1-6.11 所示的有根系统树。

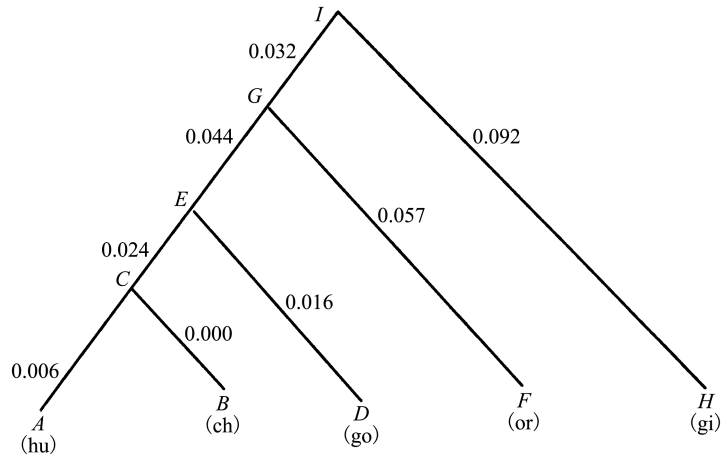

图 1-6.11 基于图 1-6.6 所列线粒体序列的 Fitch-Margoliash 有根树状图

Fitch 和 Margoliash 承认他们的法则所得到的拓扑结构有可能是不正确的，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 OTU 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

$$s=\left\{\frac{\sum\left[(d_{ij}-e_{ij})/d_{ij}\right]^2}{n(n-1)}\right\}^{\frac{1}{2}}\times 100 \qquad (1\text{-}6.5)$$

为百分标准差。考虑到可加性的假定，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 1-6.10 的系统树，观测距离和分枝长度列于表 1-6.5，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，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 1-6.5　五条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 OTU 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（NJ 法）

邻接法（Neighbor-joining Method，NJ）由 Saitou 和 Nei（1987）提出。与 UPGMA 方法类似，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。上述例举（图 1-6.6），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。与 UPGMA 方法不同之处，是其确定距离的方法不一样。

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

$$r_i=\sum_{k=1}^{N}d_{ik} \qquad (1\text{-}6.6)$$

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}:

$$M_{ij}=d_{ij}-\frac{r_i+r_j}{N-2} \tag{1-6.7}$$

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:

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

$$S_{ju}=d_{ij}-s_{iu} \tag{1-6.8}$$

节点 u 与系统树其它节点 k 的距离为:

$$d_{ku}=\frac{d_{ik}+d_{jk}-d_{ij}}{2} \tag{1-6.9}$$

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,并最终画出系统发生树。

现以图 1-6.6 线粒体序列为例,说明以上计算过程。表 1-6.6 列出了各步计算的结果。第一步,猩猩(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 1-6.6 第二步矩阵。在该矩阵中,人类(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无根系统发生树构建。

表 1-6.6 邻接法计算线粒体序列(图 1-6.6)的距离 d_{ij}(上对角线部分)和 M_{ij}(下对角线部分)

第一步		hu	ch	go	or	gi	净分歧度
		$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	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

第二步		hu	ch	go	节点 1	
		$j=1$	$j=2$	$j=3$	$j=4$	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

第三步		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

第四步		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 分别代表人类、黑猩猩、大猩猩、猩猩和长臂猿

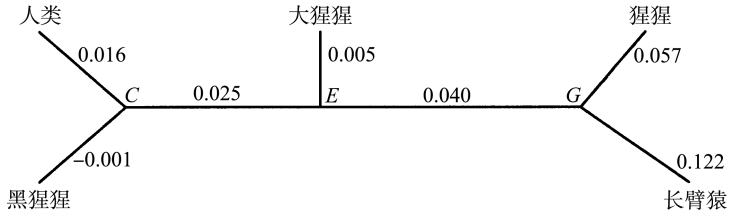

图 1-6.12　利用邻接法构建的五条线粒体序列无根系统发生树

第三节　简约法

简约法(Parsimony)由 Edwards 和 Cavalli-Sforza(1963)以“最小进化原理”应用于基因频率数据。如果有一组来自不同物种的序列可供利用,那么连接它们最为简约的拓扑结构就可能得到。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系统发生树。简约法的基本假设是,生物序列总是采用某种“最节约成本”或“最经济”的方法完成进化过程。

为说明这一方法,我们举如下一个例子:六个物种(#A～#F)的序列可以利用,并且在某一特定联配位置,它们分别具有碱基 C、T、G、T、A 和 A。如何构建它们的拓扑结构? 其存在许多可能的拓扑结构,其中之一如图 1-6.13 所示。

从离现存序列最近的节点开始,依次考虑节点 1～5 中的每一个节点。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　$[X,Y]\cap[X,Z]=[X]$　　$[X]\cap[Y]=\varphi$

并　$[X,Y]\cup[X,Z]=[X]$　　$[X]\cup[Y]=[X,Y]$

简约　$[X,Y]\diamond[X,Z]=[X]$　　$[X]\diamond[Y]=[X,Y]$

对于简约运算,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

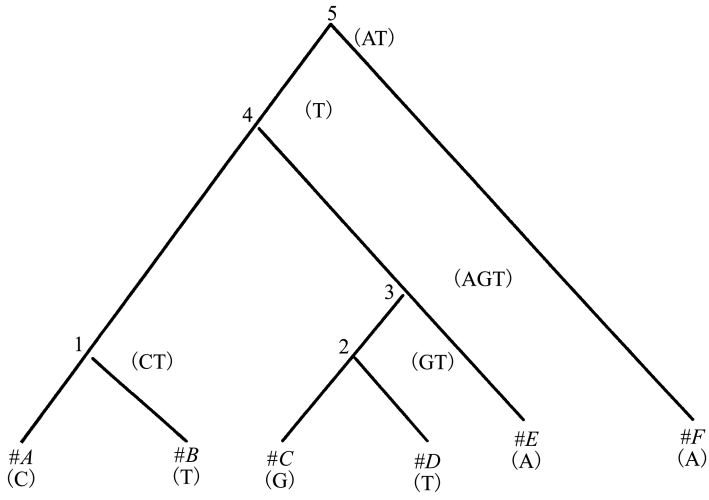

图 1-6.13 基于 6 条序列(物种#A-#F)一个位点碱基变异确定最简约树的过程

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 1-6.13 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。

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优的系统发生树。

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。以表 1-6.7 为例,只有位点 5,7,9 为信息位点。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。例如位点 4,其碱基变异无法为评判哪种树型提供任何依据,因为基于该位点,每种树型都需要 3 次碱基变更;相反,位点 5 可以给出碱基最小变更的树型,因此它提供了有用信息,为信息位点。

表 1-6.7 序列信息位点列举(以 4 条序列共 9 个位点为例)

仅有三个位点(位点 5、7、9)为信息位点;同时列出基于位点 4 和位点 5 构建系统发生树所需碱基变更次数

序列编号	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

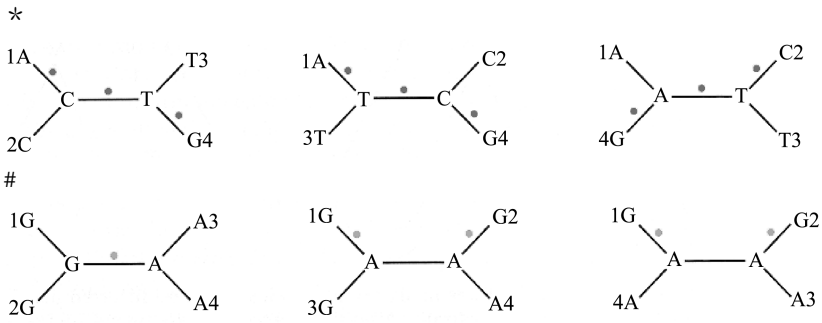

图 1-6.6 例举中的线粒体序列,基于上述标准存在 5 个信息位点(位点 25、39、44、47 和 54)。根据这 5 个位点可以构建它们的简约系统发生树(图 1-6.14)。与其它可能系统树一样存在碱基变更,但该树仅有 6 个碱基变更,为最简约系统树。尽管我们仅利用了非常有限的位点信息,但获得了与距离矩阵法找到的系统树相同的拓扑结构,可见信息位点提供的可靠系统发生信息。

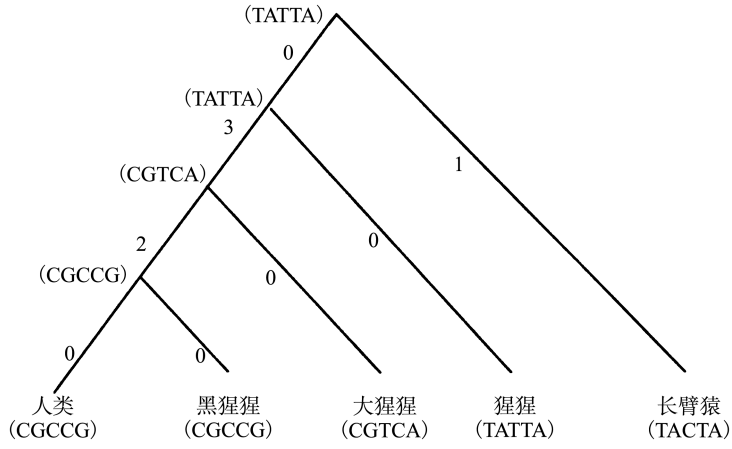

图 1-6.14　基于图 1-6.6 中来自五个物种线粒体序列的最简约系统树

图中数字为节点间的碱基变更数

部分科学家对简约法提出批评,因为该法不是以统计原理为基础。例如 Felsenstein (1983)指出,在试图使进化事件的次数最小时,简约法隐含地假定碱基多次突变事件是不可能的。如果在进化时间范围内碱基变更的数量较小,则简约法是很合理的,但对于存在大量变更的情形,随着所用数据的增加,简约法可能给出错误的系统树。

第四节　似然法

一、DNA 序列的似然模型

似然法试图避免其它方法构建系统发生树的局限性,尽管它需要的计算量大得惊人。与距离矩阵法不同,似然法试图充分和有效地利用所有数据,而不是将数据简缩为距离的集合。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。

当考虑实施最大似然法时,该方法先假定系统发生树的结构,然后选择分枝长度,以使

产生特定系统树的数据似然值最大化。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发生树看作最佳估计。一个直接的问题是随着 OTU 的增加,系统树的数目迅速增加。当树端具有 n 个 OTU 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,该数目分别为 1,3,105,10 395 和 2 027 025。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。实际应用时,只能测验所有系统树的一个亚集。

对于 DNA 序列数据,似然法依据的模型,规定了在特定时间内由于突变使一条序列变更为另一条序列的概率。尽管 DNA 序列中的毗邻碱基不是独立的,但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$(1-u)^T$,因此发生突变概率 P 为:

$$P=1-(1-u)^T\approx 1-e^{uT} \tag{1-6.10}$$

经过时间 T 后由碱基 i 变更为碱基 j 的概率可写为:

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

$$P_{ij}(T)=p\pi_j \qquad (j\neq i) \tag{1-6.11}$$

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。

上述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 $\nu=uT$,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。

在这里所描述的一个参数突变模型下,预期 4 种碱基变化具有相等频率,即对于 $i=1,2,3,4$,π_i 设定为 0.25。另一可能的方式,是估计用于构建系统树的序列碱基平均突变率,作为 π_i 值。

二、两条序列系统发生树

具有两条序列的一个有根系统发生树如图 1-6.15 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(节点 0)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:

$$L(j)=\sum_{k=1}^{4}\pi_k P_{ks_1}(v_1)P_{ks_2}(v_2) \tag{1-6.12}$$

对于所有 m 个位点,似然值为:

$$L=\prod_{j=1}^{m}L(j) \tag{1-6.13}$$

该似然值是两个未知分枝长度 v_1、v_2 的函数。

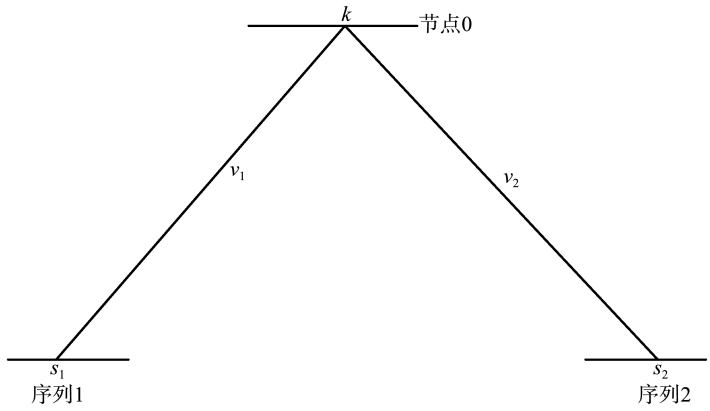

图 1-6.15　两个序列的有根树状图

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 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:

$$\begin{aligned} L(j) &= \pi_A P_{AA}(v_1) P_{AC}(v_2) + \pi_c P_{cA}(v_1) P_{CC}(v_2) + \pi_G P_{GA}(v_1) P_{GC}(v_2) + \pi_T P_{TA}(v_1) P_{TC}(v_2) \\ &= \pi_A[(1-p_1) + p_1\pi_A]p_2\pi_C + \pi_C p_1 \pi_A[(1-p_2) + p_2\pi_C] + \pi_G p_1\pi_A p_2\pi_C + \pi_T p_1\pi_A p_2\pi_C \\ &= \pi_A(p_1 + p_2 - p_1p_2)\pi_C \\ &= \pi_A p_{12}\pi_C \end{aligned} \tag{1-6.14}$$

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

$$p_{12} = p_1 + p_2 - p_1p_2 = 1 - e^{-(v_1+v_2)} \tag{1-6.15}$$

因此图 1-6.15 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条序列间的单个分枝。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:

$$L = \left(\frac{4-3p}{64}\right)^s \left(\frac{p}{64}\right)^{m-s} \tag{1-6.16}$$

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

$$\hat{p} = \frac{4(m-s)}{3m} \tag{1-6.17}$$

分枝长度的最大似然估计值为

$$\hat{v} = \ln\left(\frac{3}{4\tilde{q} - 1}\right) \tag{1-6.18}$$

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：

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

$$K = 2\mu t = \frac{3}{4}\ln\left(\frac{3}{4q - 1}\right) \tag{1-6.19}$$

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发生树

对于三条序列，则存在三种有根系统树形式，其中之一如图 1-6.16 所示。除了三条可观测的序列外，在节点 0 与节点 4 还有未定的序列，且有 4 个分枝长度有待估计。可依次考虑三种树状图，其中给出最大似然值的树型就是估计得到的系统发生树。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。

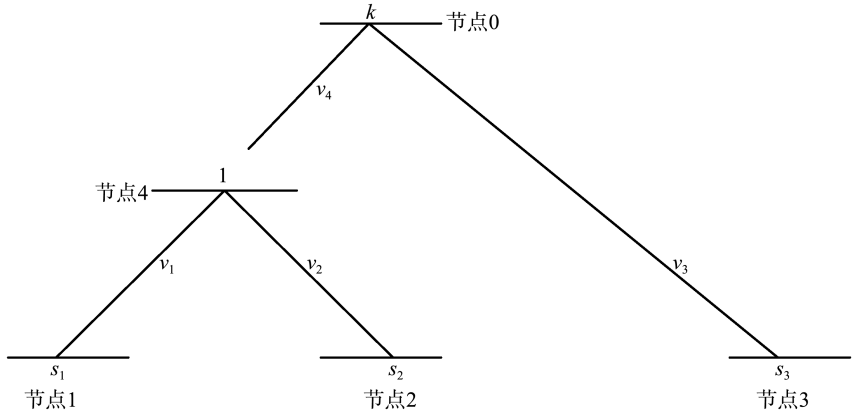

图 1-6.16 三条序列的一种有根系统发生树树型

在位点 j，三个序列具有碱基 s_1、s_2、s_3，节点 0 和节点 4 具有碱基 k 和 l

对于图 1-6.16 所示的树型，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：

$$L(j) = \sum_k \sum_l \pi_k P_{kl}(v_4) P_{ks_3}(v_3) P_{ls_1}(v_1) P_{ls_2}(v_2) \tag{1-6.20}$$

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：

$$L(j) = \sum_k \pi_k P_{ks_1}(v_1) P_{ks_2}(v_2) P_{ks_3}(v_3) \tag{1-6.21}$$

无法唯一地确定接点 0 的位置，且对于三条序列只有图 1-6.17 中星状系统树需要考虑。

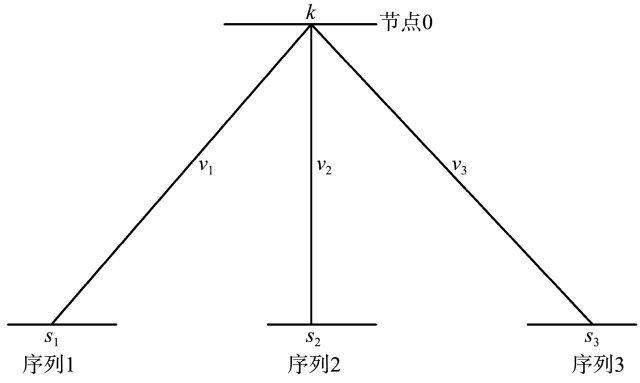

图 1-6.17　三个序列的星状系统发生树

三条序列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 $p_i=1-e^{-vi}$ 的估值来看,这一方法是最为简单的。

用多条序列作为树端来构建系统发生树时,可采用以上所述的一般过程。先指定一种系统发生树树型,然后对来自该系统树似然函数的方程进行 Newton-Raphson 迭代,估计其分枝长度。在理论上,应研究所有可能系统树来寻找具有最大似然值的系统树。研究证实,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。将这一方法应用于前述的 5 种线粒体序列,获得了图 1-6.18 所示的无根树状图。

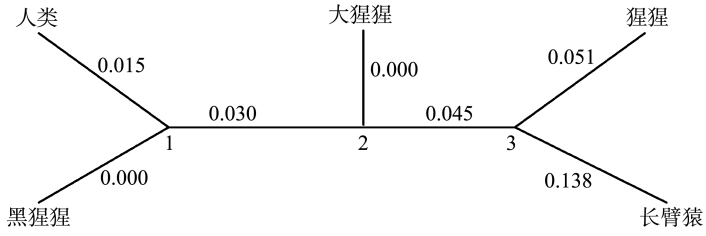

图 1-6.18　利用 PHYLIP 软件构建五条线粒体序列(图 1-6.6)的最大似然树

第五节 基因组组分矢量方法

随着测序技术的进步,大量生物物种基因组被测序完成。例如,目前细菌基因组测序项目超过 6 万个,其基因组数据已公开的超过 3 万个。利用基因组进行系统发生关系研究有其特有优势,可以避免单一基因由于横向转移、基因进化速率等因素的干扰,从而获得更加准确和高分辨率的亲缘关系。同时,在估计序列间变异时往往基于序列联配。序列联配的算法依赖于许多参数,其打分矩阵等参数会影响真实序列关系的确定。不使用序列联配方法,才可能成为无参数的算法。

为此,郝柏林院士课题组发展了一个基于基因组组分矢量方法(CVTree 算法)用于系统发生树构建。该方法把 20 个字母组成的氨基酸字母集合,扩大成由 20^K 个 K 肽字符串组成的集合。这样做,可以把 K 个字母以内的短程关联都自动包含进来,而且对于结构域交换、小段 DNA 或氨基酸段落的增删等等都变得不太敏感。该方法不依赖于序列联配,应用于细菌等微生物亲缘关系分析取得了很好效果,特别是在细菌的大尺度分类和亚种以下株系的精细分类(郝柏林,2015)。

一、组分矢量方法(CVTree 算法)

对于来自一个基因组所编码的全部 M 条蛋白质,设第 i 条蛋白质长度为 L_i。固定一个整数 K,用宽度为 K 的窗口从左向右滑动,可以得到(L_i-K+1)个 K 肽字符串。对所有的蛋白质序列进行这样的计数,得出各种 K 肽串的数目。这些 K 肽都包含在 20^K 这个总种类数之内。把所有可能的 K 肽串按氨基酸字母顺序排列起来,即从 AA...A 排列到 WW...W。构造一个长度为 20^K 的矢量,把刚才统计出来的各种 K 肽串的数目,对号入座地填写到矢量的相应位置。这样就得到一个初始的组分矢量。

组分矢量的每个分量,是一个特定 K 肽串的出现频度。为了继续发展算法,把字母串的出现频度转换成出现概率。为简单起见,考虑由 L_i 个氨基酸组成的第 i 条蛋白质。把 K 字母串 $\alpha_1\alpha_2\ldots\alpha_K$ 的出现频度标记为$f(\alpha_1\alpha_2\ldots\alpha_K)$,其中 α_i 是 20 种氨基酸之一的单字母标号。如果只有这一条蛋白质,出现频度除以这条蛋白质所包含的 K 肽串总数(L_i-K+1),

当 K 很大时,这个比值就逼近该字母串的出现概率(大数定律):

$$p(\alpha_1\alpha_2\cdots\alpha_k)=\frac{f(\alpha_1\alpha_2\cdots\alpha_k)}{(L_i-K+1)}$$

如果有 M 条蛋白质,上式要换成

$$p(\alpha_1\alpha_2\cdots\alpha_k)=\frac{f(\alpha_1\alpha_2\cdots\alpha_k)}{[L-M(K-1)]}$$

式中的 L 是所有蛋白质的总长度。用这些 K 串频度或概率做分量的组分矢量,反映了演化历史上突变和选择的结果。这样用组分矢量做每个物种的代表,可以寻求它们之间的关系。然而,这样得到的结果并不好。问题在于普遍存在于基因组中的中性突变,K 计数的结果里包含着中性突变的贡献,该突变扰乱了上面的概率计算,要设法把它们减除掉,以突出自然选择的结果。因此,需要对计数结果实行背景减除的处理(具体减除方法略)。

二、基因组关联“距离”与系统发生树构建

为了简化书写,把所有可能的 K 肽串用下标 i 编号,$i=1,2,\ldots,20^K$。两个基因组 A 和 B 的组分矢量及其分量记为

$$A=(a_1,a_2,\cdots,a20^k)$$

$$B=(b_1,b_2,\cdots,b20^k)$$

首先计算 A 和 B 两个矢量的关联 $C(A,B)$,办法是把一个矢量往另一个矢量上投影:

$$C(A,B)=\frac{\sum_{i=1}^{20k} a_i b_i}{\sqrt{\sum_{i=1}^{20k} a_i^2}\sqrt{\sum_{j=1}^{20k} b_j^2}}$$

$C(A,B)$是已经归一化的关联,它的变化范围是$[-1,+1]$。进一步定义物种 A 和 B 的关联距离 $D(A,B)$:

$$D(A,B)=\frac{1}{2}[1-C(A,B)]$$

关联距离也是归一的:它的变化范围是从 0 到 1。根据上式,计算所有基因组两两之间的距离,形成距离矩阵。然后就可以利用某种距离方法来构树,如邻接法。大量研究经验表明,邻接法是一种稳定的从距离出发的构树方法。

作为不依靠序列联配的方法,CVTree 方法采取统计再抽样的办法。大量分析结果证明,CVTree 方法很好地通过了“自举”和“刀切”检验,构建的细菌系统发生树与当前细菌体系高度一致。

肽段长度 K 的意义和选择。对多个基因组所编码的蛋白质集合,进行不依靠序列联配的比较,其基本办法是把单个氨基酸的技术($K=1$)扩展到对 K 肽片段或字符串的计数。定性地说,长 K 串具有较大的物种特异性。但如果 K 取得太大,即过分强调物种特异性,在极端情形下就会获得一棵星形树(star tree),即每个物种各成一支,这样就不能反映出物种之间的联系。物种之间的关系是靠较多物种所共有的较短的 K 串来体现的。

对于细菌,理论推断获得的 K 整数范围应该为 5 和 6,该结果同我们多年的实际计算经验一致。在 $K=5$ 或 6 时,物种距离之间的三角形不等式全部成立(具体详见郝柏林,2015)。在这两个 K 值下,CVTree 结果通过“自举”和“刀切”检验的效果最好;更为重要的,在 $K=5$ 或 6 时,古菌、真细菌和真核生物这三个生命“超界”在 CVTree 结果中明确分开。对于病毒和真菌,可以相应地选取的氨基酸总数量级分别为 10^5 和 10^7,基于上述估计,我们把得到它们最佳 K 串长度范围分别为 3 或 4(病毒)和 6 或 7(真菌)。

由此可见,对于细菌 $K=5$ 或 6 给出最好的构树和分类结果。其实,上述的肽段长度观察与生物学家们的经验是一致的。诺贝尔化学奖获得者米歇尔(Hartmut Michel)曾说过,只要知道蛋白质的一小部分(6 个氨基酸就足够了),就可以在数据库中确定整个序列;蛋白质组学研究也指出,6 肽或稍长的肽链在一个物种的蛋白质组中几乎是唯一的。他们谈及的还仅是单个 6 肽,而 CVTree 方法则使用了全部 5 肽或 6 肽的集合,因而分辨力和物种特异性更强。

习 题

1. 构建进化树有哪几种方法？分别适用于何种情况？

2. 试列举构建进化树时影响最终树形的因素。

3. 进化树可靠性一般用什么方法评估？

4. 熟悉并使用目前常见的 1-2 个建树软件流程，如 MEGA 或 PHYLIP 等。

5. 请在 GenBank 数据库中搜索 10-20 条植物已知 NBS 类抗性基因，并构建其系统发生树。