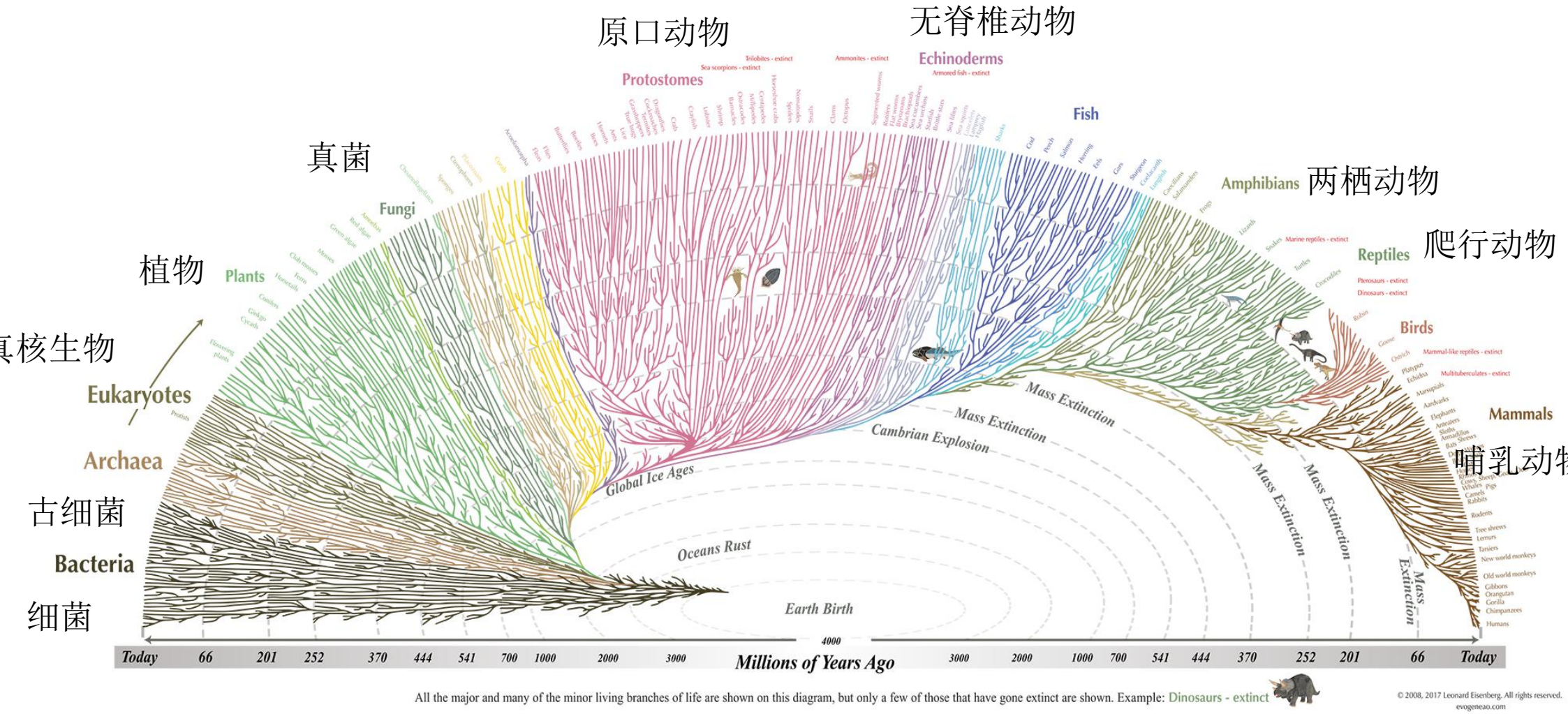
分子进化与系统发育分析（0）

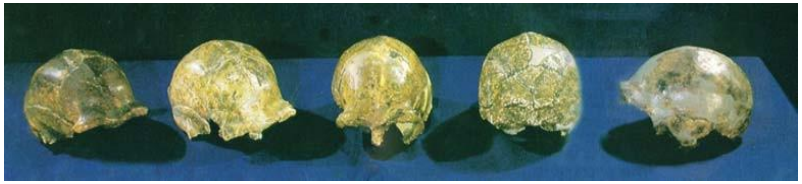
生物学家：We have a dream…

- **Tree of Life:** 重建所有生物的进化历史并以系统 树的形式加以描述

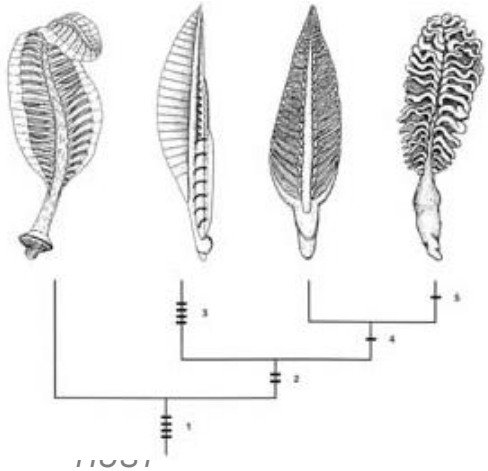

原口动物
无脊椎动物
真菌
植物
两栖动物
爬行动物
真核生物
哺乳动物
古细菌
细菌
Protostomes
Echinoderms
Fish
Fungi
Plants
Amphibians
Reptiles
Birds
Mammals
Eukaryotes
Archaea
Bacteria
Trilobites - extinct
Sea scorpions - extinct
Ammonites - extinct
Armored fish - extinct
Marine reptiles - extinct
Pterosaurs - extinct
Dinosaurs - extinct
Mammal-like reptiles - extinct
Multituberculates - extinct
Mass Extinction
Cambrian Explosion
Global Ice Ages
Oceans Rust
Earth Birth
Today
66
201
252
370
444
541
700
1000
2000
3000
4000
Millions of Years Ago
All the major and many of the minor living branches of life are shown on this diagram, but only a few of those that have gone extinct are shown. Example: Dinosaurs - extinct
© 2008, 2017 Leonard Eisenberg. All rights reserved.
evogeneao.com

梦想走进现实：How?

- 最理想的方法：化石！—— 零散、不完整



- 比较形态学和比较生理学：确定大致的进化框架 —— 细节存很多的争议



第三种方案：分子进化

- 1964年，Linus Pauling提出分子进化理论
- DNA & RNA: 4种碱基；蛋白质分子：20种 氨基酸
- 发生在分子层面的进化过程：DNA, RNA和 蛋白质分子
- 基本假设：核苷酸和氨基酸序列中含有生物 进化历史的全部信息

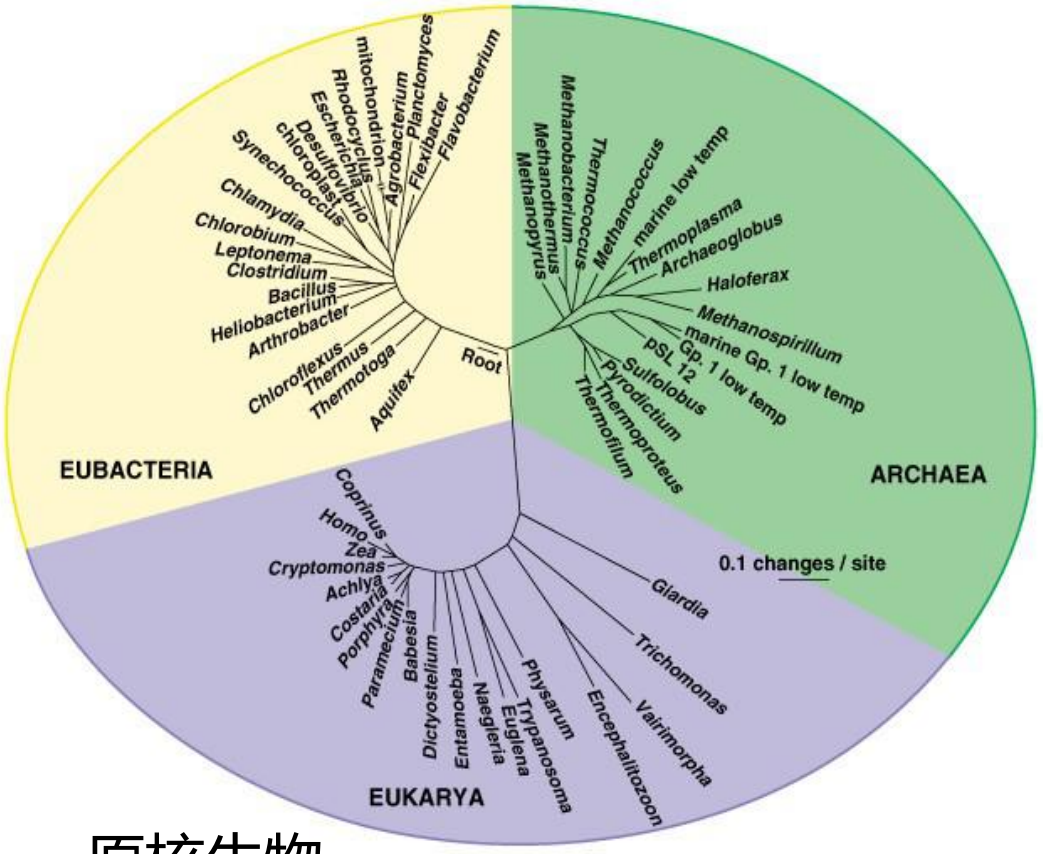
分子进化研究的目的

- 从物种的一些分子特性出发，构建系统发育树，进 而了解物种之间的生物系统发生的关系 —— tree of life; 物种分类
- 大分子功能与结构的分析：同一家族的大分子，具 有相似的三级结构及生化功能，通过序列同源性分 析，构建系统发育树，进行相关分析；功能预测
- 进化速率分析：例如，HIV的高突变性；哪些位点 易发生突变？

物种树是基于每个物种整体的进化关系，也就是基于整个基因组构建的，而分子树是基于不同物种里某一个基因或蛋白质序列之间的关系构建的。

Tree of Life:
16S rRNA

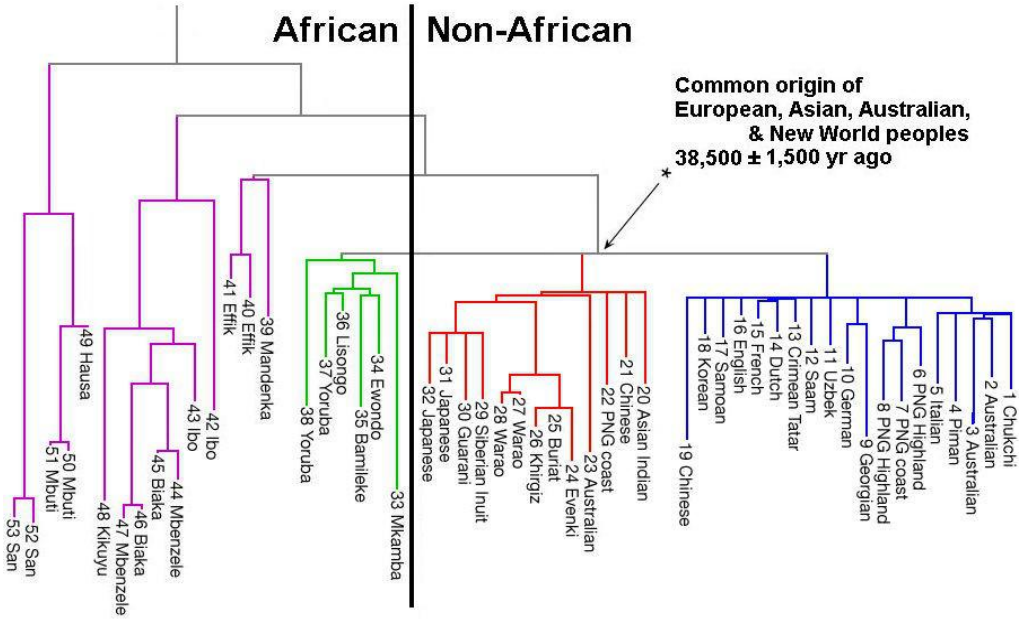
真核生物



古细菌

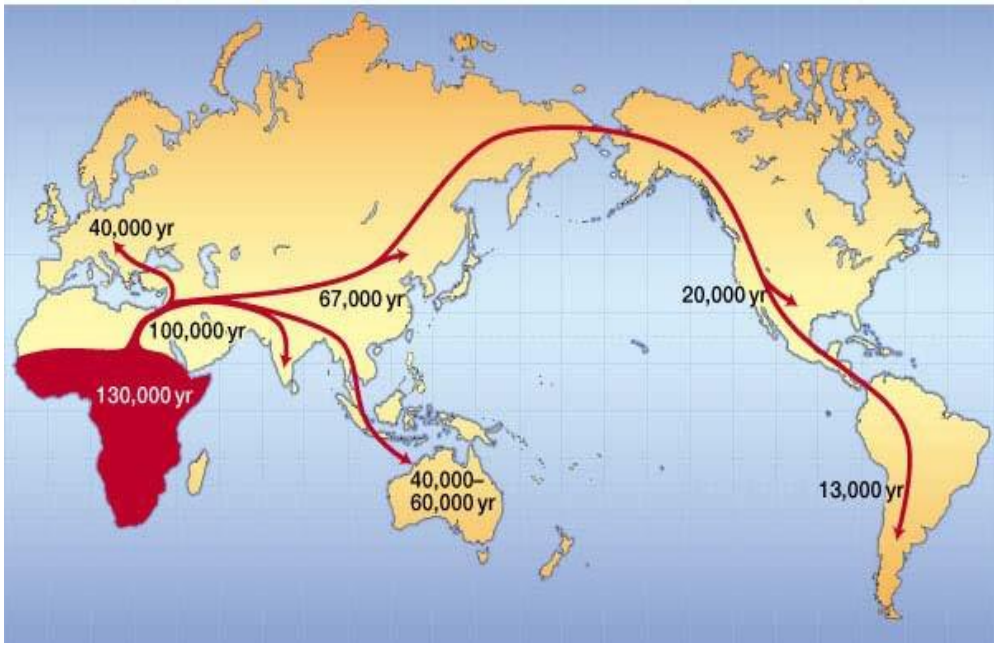
原核生物

Out of Africa



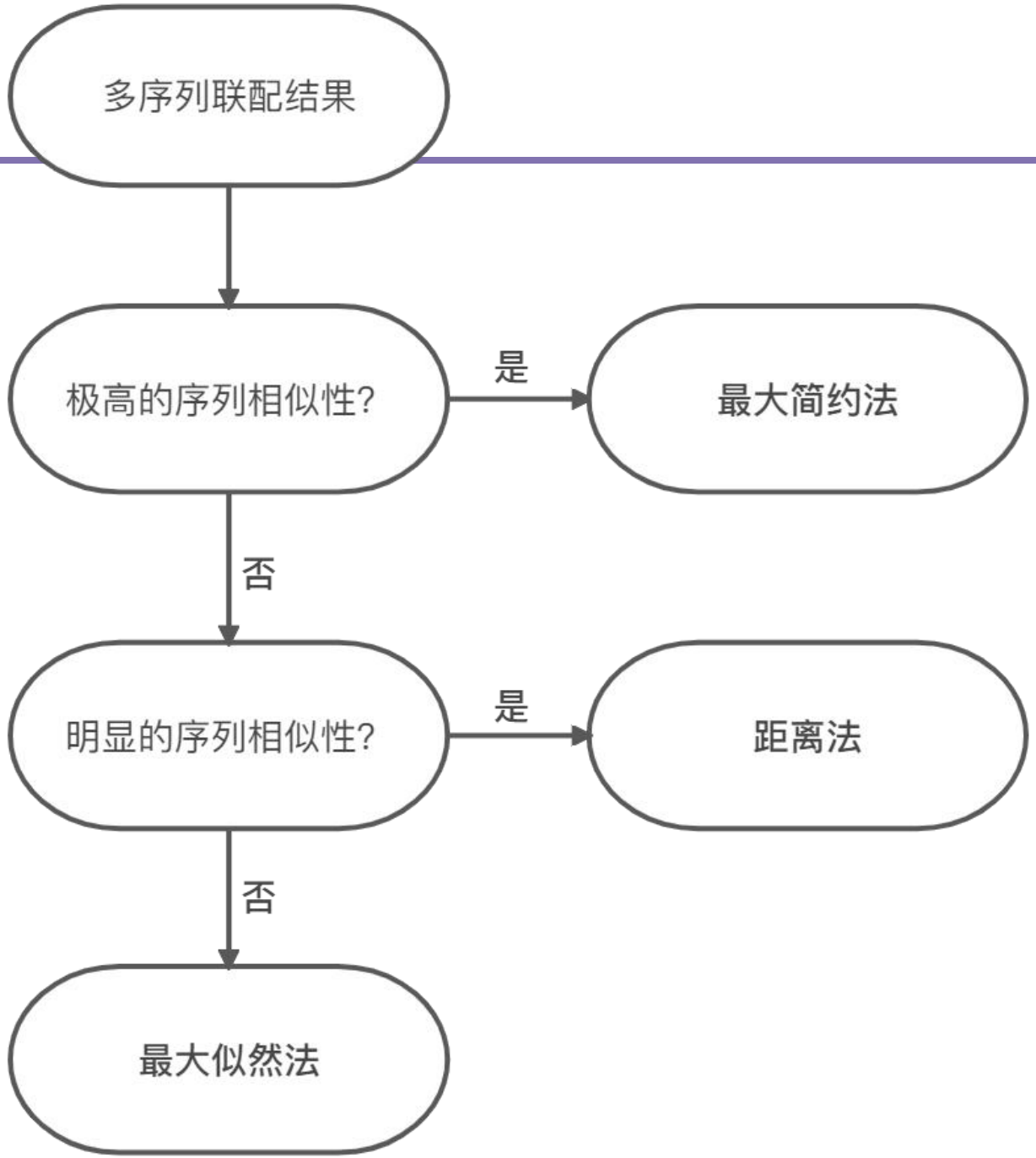
53个人的线粒体基因组(**16,587bp**)

人类迁移的路线



系统发育树的构建

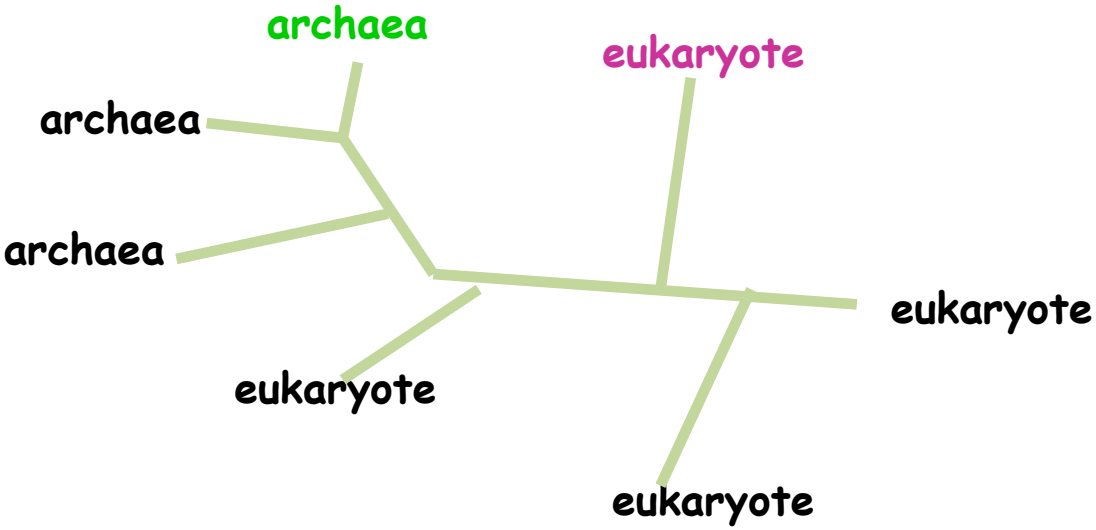
- 系统发育树：分子进化树/分子进化分析
- 通过进化树的构建，分析分子之间的起源关系，预测分子的功能
- 建树方法：
  - 距离法（Distance-based methods）
  - 最大简约法（Maximum Parsimony）
  - 极大似然性法（Maximum Likelihood）
  - 贝叶斯方法（Bayesian method）


多序列联配结果
极高的序列相似性?
是
最大简约法
否
明显的序列相似性?
是
距离法
否
最大似然法

术语

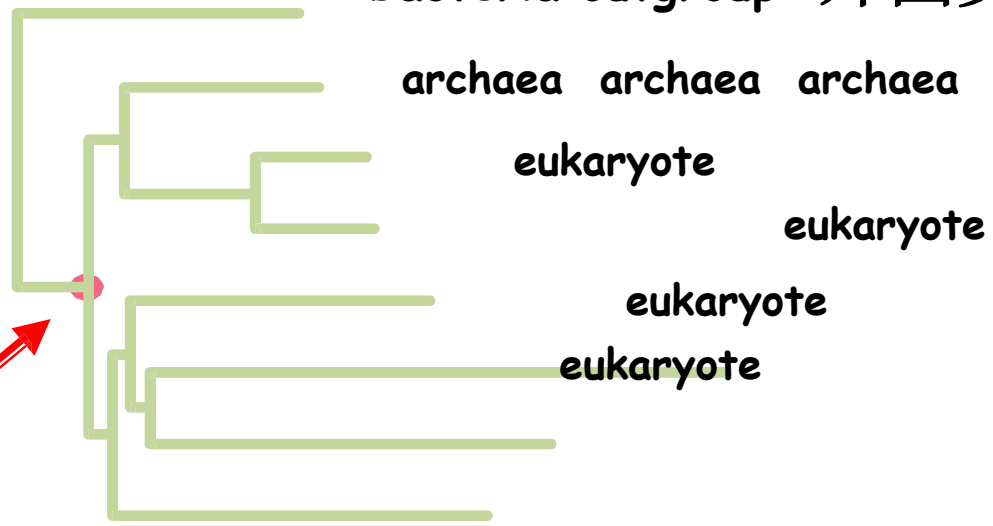
- **有根树指明了进化的方向以及物种的祖先**
- **无根树仅仅给出了物种之间的关系**
- **无根树可以通过下面两种方法构造有根树**
  - 外群或外围枝(outgroup): 事先知道某些物种之间的关系，通过这种关系来确定根的位置
  - 分子钟: 如果我们假设进化速率在物种之间以及各个不同的时期都是相等的，这时，进化树的根距离各个物种的距离都应该是相等的。通过这个关系，可以确定根的位置

无根树，有根树，外围支



无根树

通过外围支
来确定树根



外围支

有根树

根

无根树和有根树：潜在的数目

#节点数	无根树	有根树
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
…		
30	~$3.58X10^{36}$	~$2.04X10^{38}$

n个叶子节点，共有$3*5*7*\cdots*(2n-5)=(2n-5)!!$ 无根树；$(2n-5)!!*(2n-3)$个有根树。

Taxa增多，计算量急剧增加，因此，目前算法都为 优化算法，不能保证最优解

系统发育树重建分析步骤

多序列比对（自动比对，手工校正）

选择建树方法以及替代模型 建

立进化树

进化树评估

系统发育树重建的基本方法

- 距离法 (distance)
- 最大简约法 (maximum parsimony, MP)
- 极大似然法 (maximum likelihood, ML)
- 贝叶斯方法 (Bayesian method)

非如何从多序列比对得到距离矩阵

AGGCCATGAATTAAGAATAA 物种1
AGCCCATGGATAAAGAGTAA 物种2
AGGACATGAATTAAGAATAA 物种3
AAGCCAAGAATTACGAATAA 物种4

Distance Matrix

	1	2	3	4
1	-	0.2	0.05	0.15
2		-	0.25	0.4
3			-	0.2
4				-

- 距离矩阵取决于所使用的相似性矩阵，例如，蛋白质可以使用BLOSUM，PAM等
- 右侧，我们使用最为简单的Hamming距离除以总长度作为距离度量.
  - 例如
    distance(1,2)=4/20=0.2

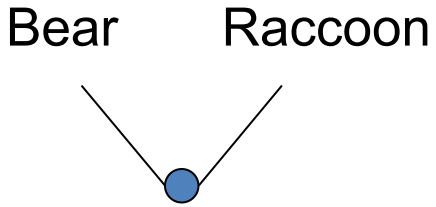
非加权分组平均法(unweighted pair group method using arithmetic averages, UPGMA)

该方法缩写让人望而生畏。其实此方法比较简单：对序列聚类，每一步融合两个类，同时创建一个新的节点。从而自下而上地构建一个树。

非加权分组平均法(UPGMA)

UPMGA (Michener & Sokal 1957)

D_{ij}	Bear	Raccoon	Weasel	Seal
Bear	-	**0.26**	0.34	0.29
Raccoon		-	0.42	0.44
Weasel			-	0.44
Seal				-

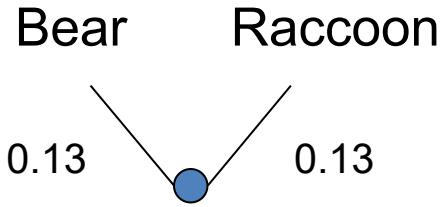


1. **在距离矩阵中找到距离最小的元素**

非加权分组平均法(UPGMA)

UPMGA (Michener & Sokal 1957)

D_{ij}	Bear	Raccoon	Weasel	Seal
Bear	-	**0.26**	0.34	0.29
Raccoon		-	0.42	0.44
Weasel			-	0.44
Seal				-



2. 把此元素所对应的物种合并一个节点。
此结点我们认为是这两个物种的祖先，并且此结点到这个物种的距离是元素值的一半。

非加权分组平均法(UPGMA)

3.计算新结点到其他各个物种之间的距离(算术平均)

$$D_{W(BR)} = \frac{D_{WB} + D_{WR}}{2} = \frac{0.34 + 0.42}{2} = 0.38$$

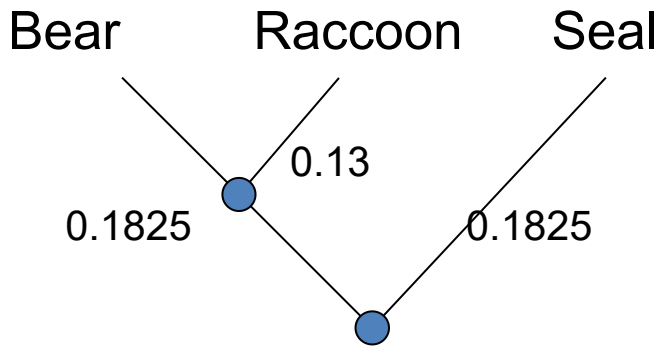
$$D_{S(BR)} = \frac{D_{SB} + D_{SR}}{2} = \frac{0.29 + 0.44}{2} = 0.365$$

D_{ij}	Bear	Raccoon	Weasel	Seal
Bear	-	**0.26**	0.34	0.29
Raccoon		-	0.42	0.44
Weasel			-	0.44
Seal				-

D_{ij}	BR	Weasel	Seal
BR	-	**0.38**	**0.365**
Weasel		-	0.44
Seal			-

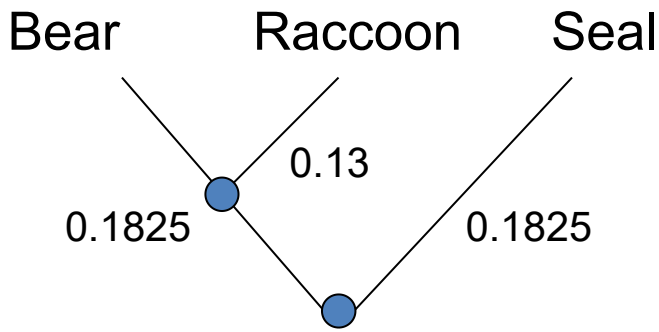
非加权分组平均法(UPGMA)

D_{ij}	BR	Weasel	Seal
BR	-	0.38	**0.365**
Weasel		-	0.44
Seal			-



1. 重新选择距离矩阵最小的元素
2. 同样的把此元素所对应的物种合并一个节点，此结点我们认为是这两个物种的祖先，并且此结点到这个物种的距离是元素值的一半

D_{ij}	BR	Weasel	Seal
BR	-	0.38	0.365
Weasel		-	0.44
Seal			-

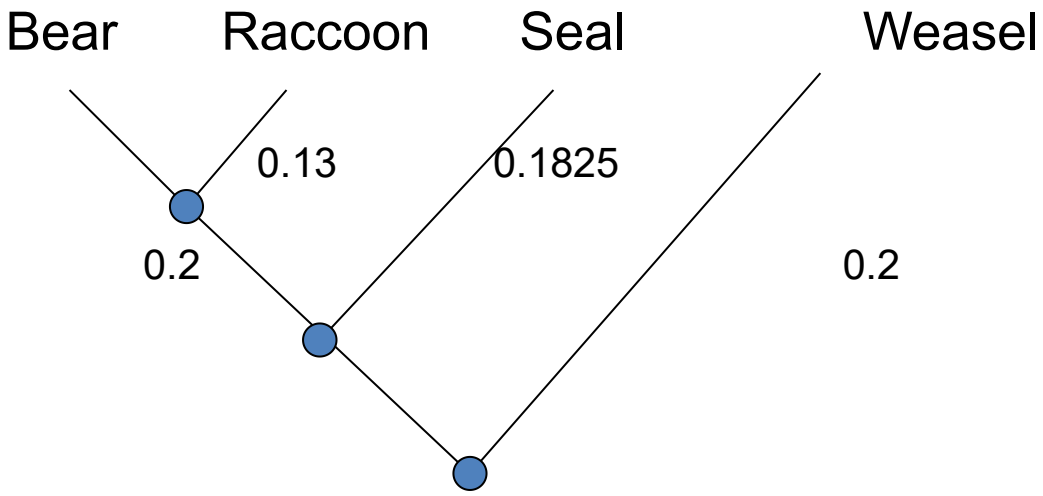


3. 计算新结点到其他各个物种之间的距离(算术平均)

$$D_{W(BRS)} = \frac{D_{WB} + D_{WR} + D_{WS}}{3} = \frac{0.34 + 0.42 + 0.44}{3} = 0.4$$

非加权分组平均法(UPGMA)

D_{ij}	BRS	Weasel
BRS	-	**0.4**
Weasel		-



1. 挑选最小值
2. 合并新节点
3. 完成

UPGMA算法（适用一般的情形）

初始化： 每条序列i自身作为一类C_i;为每条序列定义树T的一个叶子节点。高度为0.

迭代:

(1)找到距离最小的两个类C_i, C_j,距离 d_{ij}

(2)定义个新类$k, C_k = C_i \cup C_j$ （混合类）；

$$d_{kl} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{|C_i| + |C_j|},$$

其中 d_{il}, d_{jl} 分别表示两个集合之间的距离。以d_{il}为例

$$d_{il} = \frac{1}{|C_i||C_l|} \sum_{\substack{p\ in\ C_i; \\ q\ in\ C_l}} d_{pq}$$

(3)定义一个带子结点i, j的节点k，并设置高度为$\frac{d_{ij}}{2}$

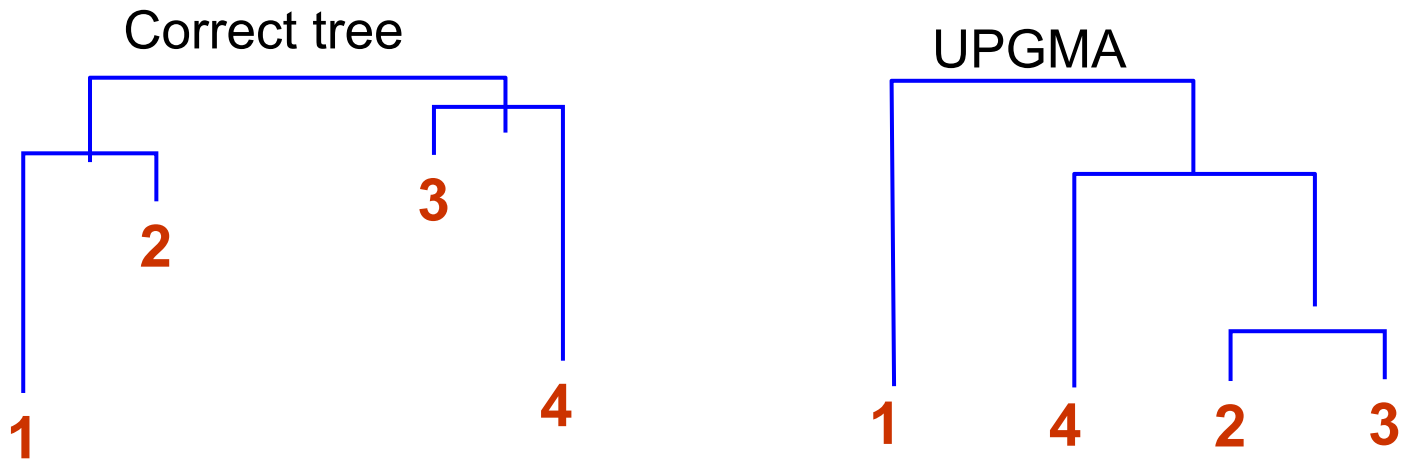
(4)将k添加到现有类集合中，删除i, j

非加权分组平均法(UPGMA)

UPGMA构建了一棵有根树

注意到使用UPGMA构造的进化树，从祖先结点到任意一个“叶子”的距离都是相等的

UPGMA假定分子钟的存在(树上所有点序列分化按照相同的固定速率发生）。



实际情况是，有可能一棵树（左）被UPGMA错误地重建了（右）

邻接法(Neighbor Joining)

在使用UPGMA算法的时候，我们默认任意一对叶子节点之间的距离是链接它们路径上边长之和。这种性质我们称为可加性。

然而存在满足可加性，但不满足分子钟情形。邻接法就是针对这种情形提出的。

邻接法(Neighbor Joining)

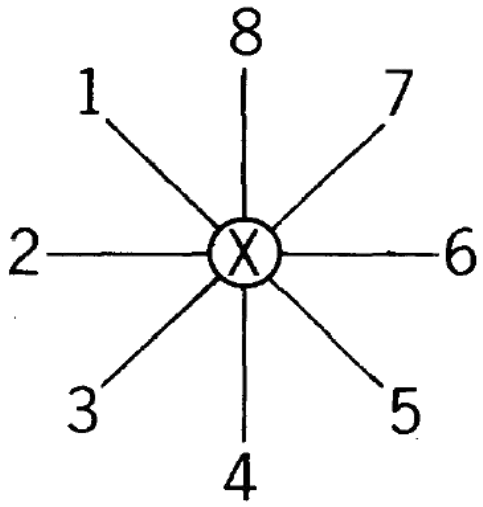
邻接法同样是基于距离的进化树构建方法是目前最流行的基于距离的进化树构建方法

邻接法的特点

- 同样是渐进的合并“邻居” (类同与UPGMA)
- 保证“邻居”之间的距离很近，同时与其他物种之间的“距离”很远
- 构造了无根树
- 不必假定分子钟

邻接法(Neighbor Joining)

0	d12	d13	d14	d15	d16	d17	d18
	0	d23	d24	d25	d26	d27	d28
		0	d34	d35	d36	d37	d38
			0	d45	d46	d47	d48
				0	d56	d57	d58
					0	d67	d68
						0	d78
							0



1. 计算N条序列两两距离，得到距离矩阵。
2. 每条序列当做一个节点，首先考虑星型结构其中D_{ij} 待分类i和j之间的距离。L_{ab} 是节点a和节点b之间的枝长。

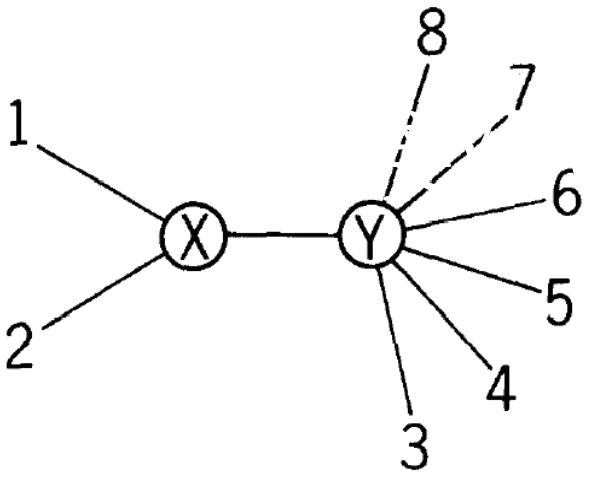
星型总枝长：

$$S_0 = \sum_{i=1}^{N} L_{ix} = \frac{1}{N-1}\sum_{i<j}^{N} D_{ij} = \frac{T}{N-1},$$

其中$T = \sum_{i<j}^{N} D_{ij}$

每一个枝被计算了$N-1$次，因为 $d_{12} = d_{1x} + d_{2x}$ 等等。

$$So = \sum_{i=1}^{8} L_{ix} = \frac{1}{7}\sum_{i<j}^{8} D_{ij} = \frac{T}{N-1}$$



3.任选两个节点，计算把这两个节点合并以后的总枝长。

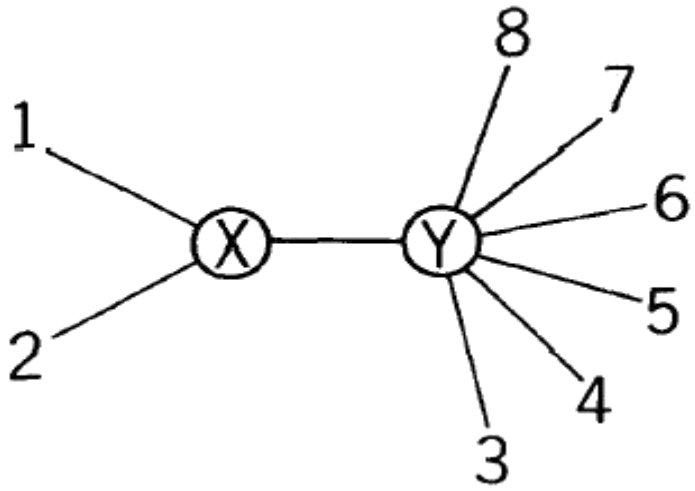
上图中，不妨将1,2两个节点合并。1，2合并为一类，其余节点为一类。内部节点为X，Y。节点X和节点Y之间边的长度L_{XY}:

$$L_{XY}=\frac{1}{2(N-2)}\left[\sum_{k=3}^{N}(D_{1k}+D_{2k})-(N-2)(L_{1X}+L_{2X})-2\sum_{i=3}^{N}L_{iY}\right]$$

第一项表示包含XY边的总支长度，Lxy计算了2(N-2)次。第 二项，第三项与Lxy不相关的部分。

接下来，利用距离矩阵Dij表示 L1x,L2X,LiY

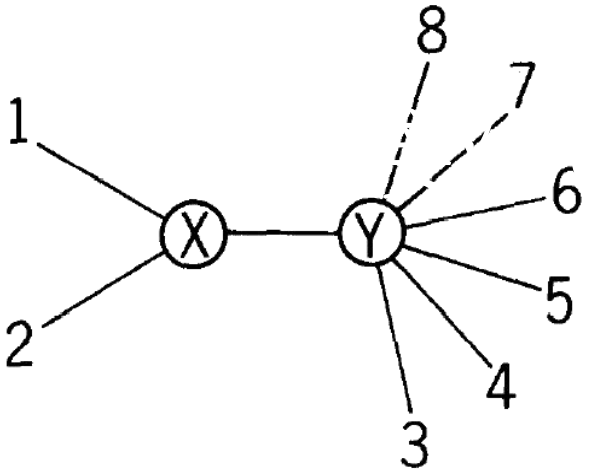
邻接法(Neighbor Joining)



$$L_{XY} = \frac{1}{2(N-2)}[\sum_{k=3}^{N}(D_{1k} + D_{2k}) - (N-2)(L_{1X} + L_{2X}) - 2\sum_{i=3}^{N} L_{iY}]$$

$$L_{1X} + L_{2X} = D_{12}$$

$$\sum_{i=3}^{N} L_{iY} = \frac{1}{N-3}\sum_{3 \le i < j} D_{ij}$$



$$L_{XY}=\frac{1}{2(N-2)}[\sum_{k=3}^{N}(D_{1k}+D_{2k})-(N-2)(L_{1X}+L_{2X})-2\sum_{i=3}^{N}L_{iY}]$$

$$L_{1X}+L_{2X}=D_{12}$$

$$\sum_{i=3}^{N}L_{iY}=\frac{1}{N-3}\sum_{3\le i<j}D_{ij}$$

$$L_{XY}=\frac{1}{2(N-2)}\left[\sum_{k=3}^{N}(D_{1k}+D_{2k})-(N-2)D_{12}-2*\frac{1}{N-3}\sum_{3\le i<j}D_{ij}\right]$$

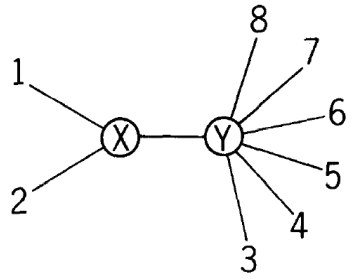
将节点12，合并之后总枝长

$$S_{12}=L_{XY}+(L_{1X}+L_{2X})+\sum_{i=3}^{N}L_{iY}$$

$$\boldsymbol{S_{12}=\frac{1}{2(N-2)}\sum_{k=3}^{N}(D_{1k}+D_{2k})+\frac{D_{12}}{2}+\frac{1}{N-2}\sum_{3\le i<j}D_{ij}}$$

邻接法(Neighbor Joining)

4.计算合并“邻居”后的各个枝长，总枝长最小的两个节点，称为邻居。把这两个节点合并。如下图。并计算这两个节点的相关枝长。



不妨设最小的邻居为1,2；计算L_{1X}，L_{2X}；
设Z是待分组节点中不包含1,2的那些节点集合。

$$L_{1X} = \frac{D_{12} + D_{1Z} - D_{2Z}}{2},$$

$$L_{2X} = \frac{D_{12} + D_{2Z} - D_{1Z}}{2}$$

其中，$D_{1Z} = \frac{\sum_{i=3}^{N} D_{1i}}{N-2}, D_{2Z} = \frac{\sum_{i=3}^{N} D_{2i}}{N-2}$

邻接法(Neighbor Joining)

5. 计算合并邻居之后，新节点到各节点之间的距离。

不妨设新节点由1,2两个节点组成。那么，新节点与j之间的距离：

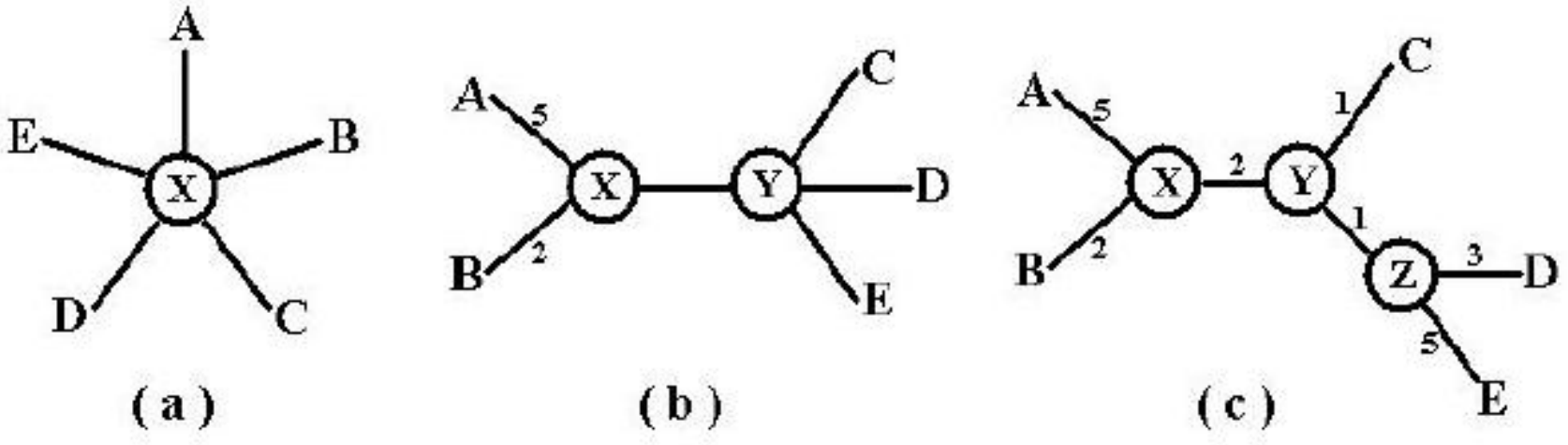
$$D_{(1,2)j} = \frac{D_{1j} + D_{2j}}{2}, (\text{其中} 3 \leq j \leq N)$$

这样总的外部结点数由N减少为N-1.内部节点有1个增加到2个。

邻接法(Neighbor Joining)

6. 重复上述步骤，直到外部待分类节点为3个时停止。此时便得到了一个无根树，即为所求。

例：	A	B	C	D	E
A	0				
B	7	0			
C	8	5	0		
D	11	8	5	0	
E	13	10	7	8	0



改进后的邻接法

1. 首先计算第i个终端节点的净分歧度$r_i = \sum_{k=1}^{N} d_{ik}$，
其中N为终端节点总数，$d_{ik} = d_{ki}$表示第i个节点和第k个节点的距离。

2. 计算最小速率校正距离$M_{ij} = d_{ij} - \frac{r_i + r_j}{N-2}$

3. 定义一个新节点u，节点u由节点i和节点j组合而成。节点u与节点i和j的距离分别为：

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

节点u与系统树上其他节点k的距离为$d_{uk} = \frac{d_{ki} + d_{kj} - d_{ij}}{2}$

4. 从距离矩阵中删除节点i，j，总结点数N减1

5. 如果剩下两个以上的节点，重复上述3个步骤，直到系统树构建完成。

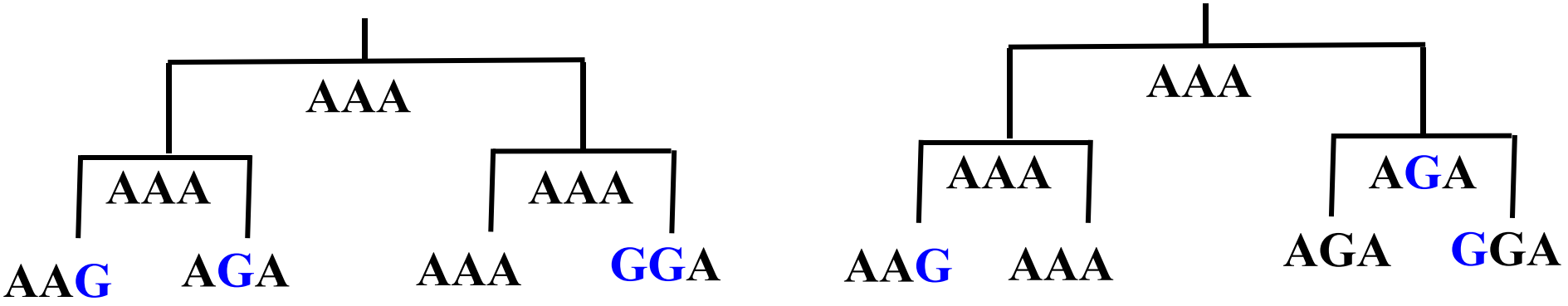
最大简约法(Maximum Parsimony)

最小进化原理：进化总是朝着最短的路径，或者说代价最小的路径进行。
最大简约法的目的是要构建一个树，使得从祖先序列按照这棵树进行进化时“代价”最小，也就是说突变最小。

- 优点：不需要在处理核苷酸或者氨基酸替代的时候 引入假设（替代模型）
- 缺点：分析序列上存在较多的回复突变或平行突变，而被检验的序列位点数又比较少的时候，可能会给 出一个不合理的或者错误的进化树推导结果

最大简约法(Maximum Parsimony)

AAG, AGA, AAA, GGA所构建的进化树



- MP方法会选择第二棵树，因为这棵树的突变次树为3，第一棵树的突变次数为4

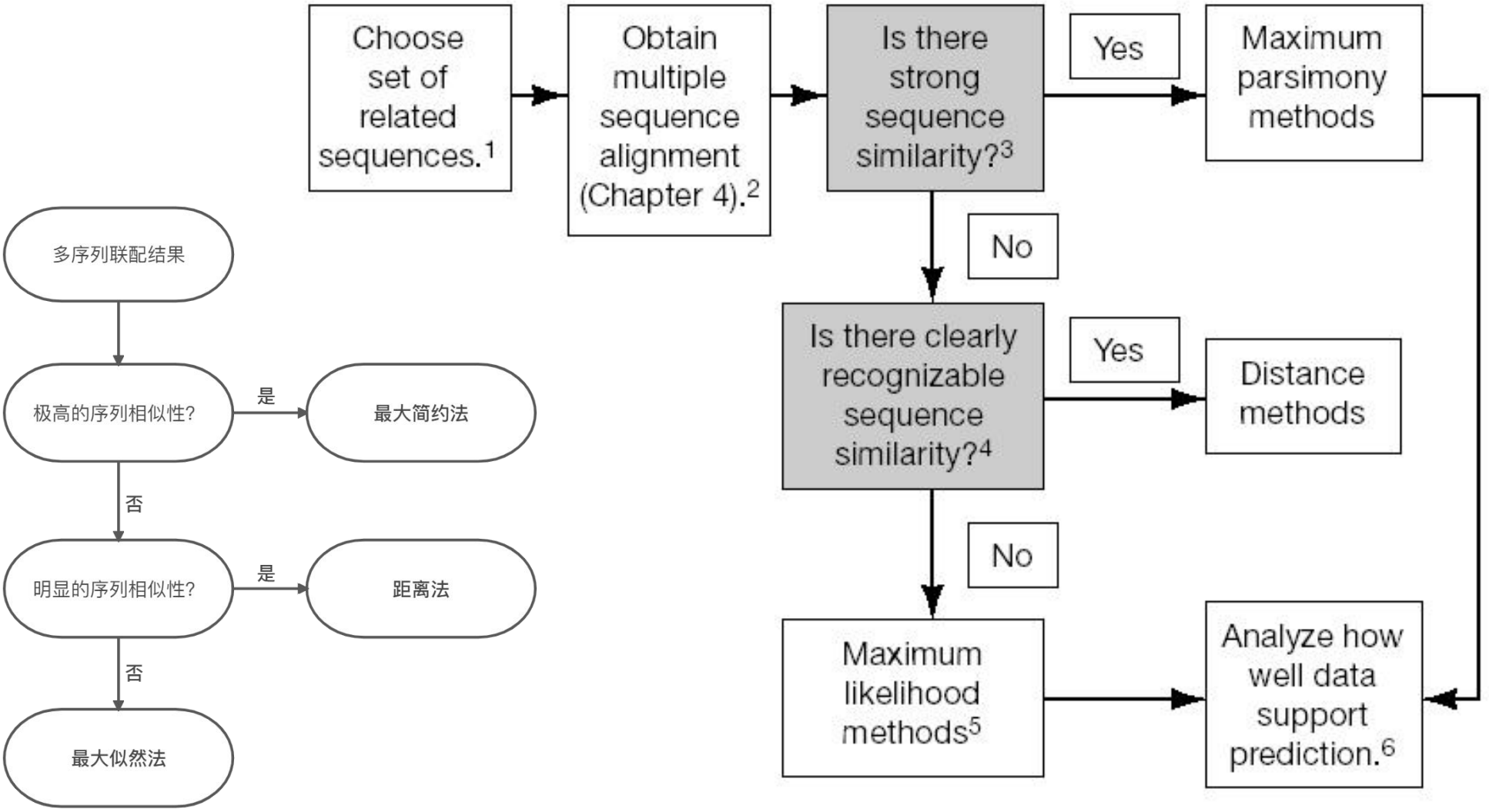
建树方法总结

快 慢

Neighbor-Joining UPGMA | Parsimony Method | Maximum Likelihood | Bayesian Method

不够准确 准确

构建进化树的一般原则


Choose set of related sequences.[1]
Obtain multiple sequence alignment (Chapter 4).[2]
Is there strong sequence similarity?[3]
Yes
Maximum parsimony methods
No
Is there clearly recognizable sequence similarity?[4]
Yes
Distance methods
No
Maximum likelihood methods[5]
Analyze how well data support prediction.[6]
多序列联配结果
极高的序列相似性?
是
最大简约法
否
明显的序列相似性?
是
距离法
否
最大似然法

构建进化树的一般原则 (2)

- 可靠的待分析数据
- 准确的多序列比对
- 选择合适的建树方法：
  - 序列相似程度高，**NJ, MP**首选
  - 序列相似程度较低，**ML**, 贝叶斯首选
  - 序列相似程度太低，无意义
- 一般采用两种及以上方法构建进化树，无显著区别可接受

选择外围支 (Outgroup)

- 选择一个或多个已知与分析序列关系较远的序列作为外围支
- 外围支可以辅助定位树根
- 外围支序列必须与剩余序列关系较近，但外围支序列与其他序列间的差异必须比其他序列之间的差异更显著

自展法

- 进化树的可靠性分析:自展法 (Bootstrap Method)
- 从排列的多序列中随机有放回的抽取某一列，构 成相同长度的新的排列序列
- 重复上面的过程，得到多组新的序列
- 对这些新的序列进行建树，再观察这些树与原始 树是否有差异，以此评价建树的可靠性

软件	网址	说明
ClustalX	http://bips.u-strasbg.fr/fr/Documentation/ClustalX/	图形化的多序列比对工具
ClustalW	http://www.cf.ac.uk/biosi/research/biosoft/Downloads/clustalw.html	命令行格式的多序列比对工具
GeneDoc	http://www.psc.edu/biomed/genedoc/	多序列比对结果的美化工具
BioEdit	http://www.mbio.ncsu.edu/BioEdit/bioedit.html	序列分析的综合工具
MEGA	http://www.megasoftware.net/	图形化、集成的进化分析工具，不包括ML
PAUP	http://paup.csit.fsu.edu/	商业软件，集成的进化分析工具
PHYLIP	http://evolution.genetics.washington.edu/phylip.html	免费的、集成的进化分析工具
PHYML	http://atgc.lirmm.fr/phyml/	最快的ML建树工具
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html	ML建树工具
Tree-puzzle	http://www.tree-puzzle.de/	较快的ML建树工具
MrBayes	http://mrbayes.csit.fsu.edu/	基于贝叶斯方法的建树工具
MAC5	http://www.agapow.net/software/mac5/	基于贝叶斯方法的建树工具
TreeView	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html	进化树显示工具