Phylogenetic Tree Reconstruction: Theory and Practice

重建系统发生树：理论与实践

蒋陈焜*, 李玲, 韩翔 & 王志娟†

北京大学生命科学学院

2015年1月7日

Outline

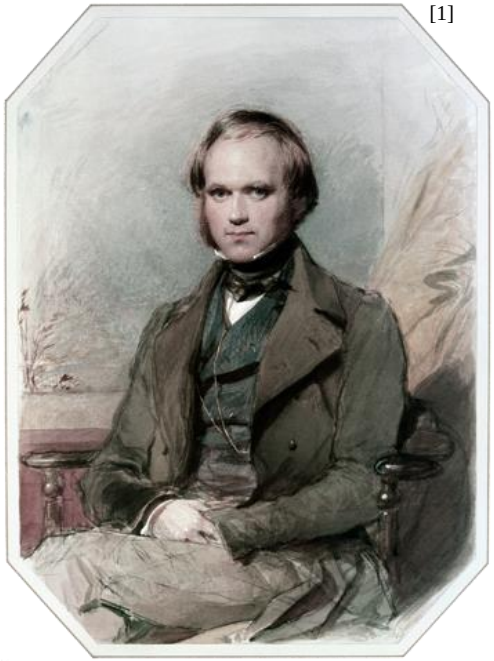
- What is phylogenetic tree: a historical perspective
- Why do we reconstruct trees?
- ABCs of molecular phylogenetics
- Methods in phylogenetic reconstruction
  - Data sets
  - Alignment
  - Phylogenetic Analysis

Outline

- What is phylogenetic tree: a historical perspective
- Why do we reconstruct trees?
- ABCs of molecular phylogenetics
- Methods in phylogenetic reconstruction
  - Data sets
  - Alignment
  - Phylogenetic analysis

什么是系统发生树？

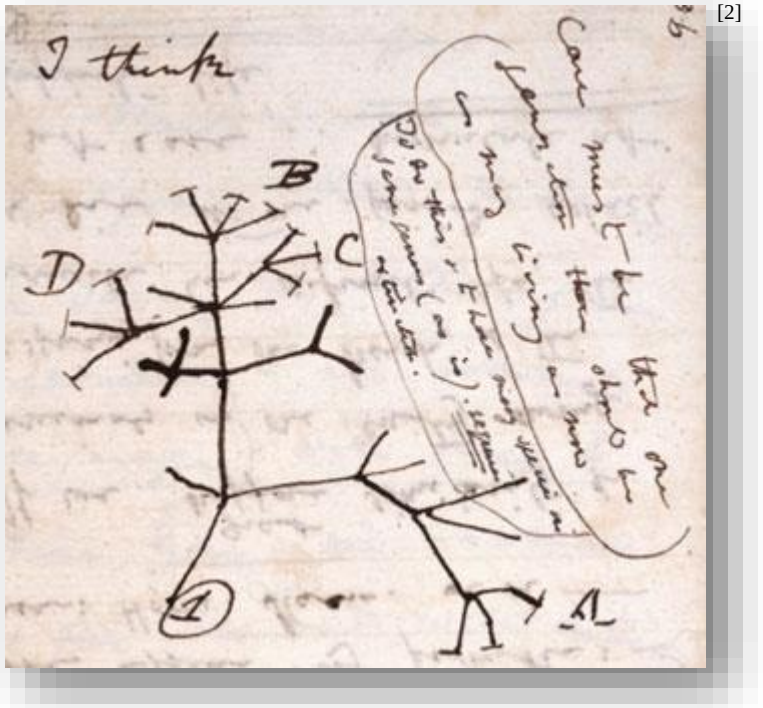
- One of the most famous tree: “I think” tree



[1]

Charles R. Darwin (1809-1882)

查尔斯 • 达尔文

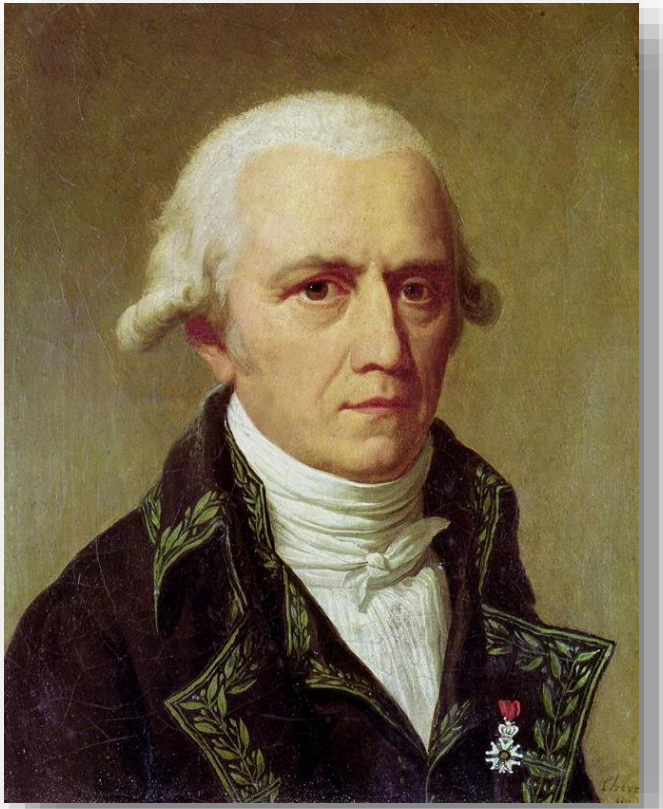


[2]

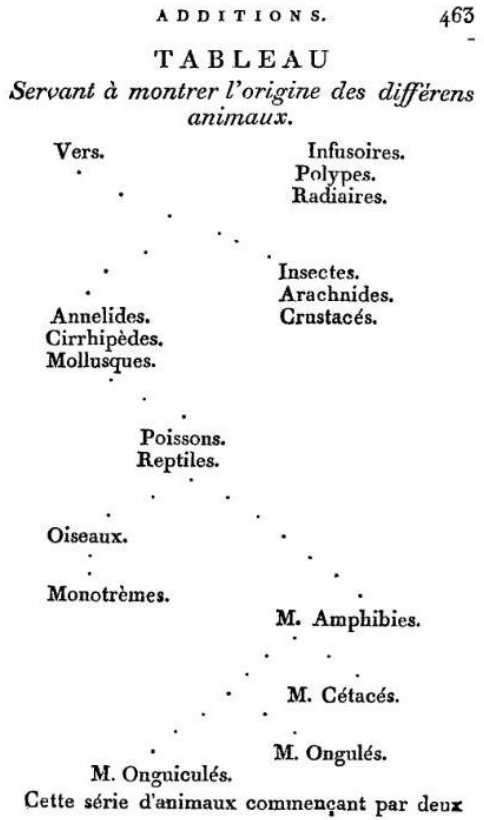
Found in his notebook from 1837

是最早的吗？

最早的进化树


[3]

Jean-Baptiste Lamarck (1744-1829)
拉马克


ADDITIONS. 463

TABLEAU

Servant à montrer l'origine des différens animaux.

Vers.
Infusoires.
Polypes.
Radiaires.
Insectes.
Arachnides.
Crustacés.
Annelides.
Cirrhipèdes.
Mollusques.
Poissons.
Reptiles.
Oiseaux.
Monotrèmes.
M. Amphibies.
M. Cétacés.
M. Ongulés.
M. Onguiculés.

Cette série d'animaux commençant par deux

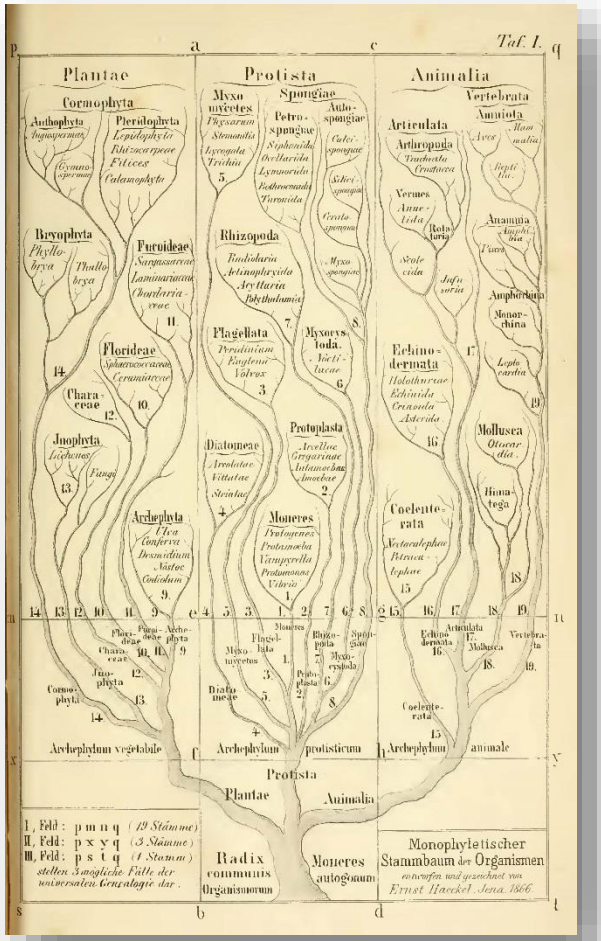
Published in *Philosophie zoologique* of 1809

“ Monophyletischer Stambaum der Organismen” (1866)

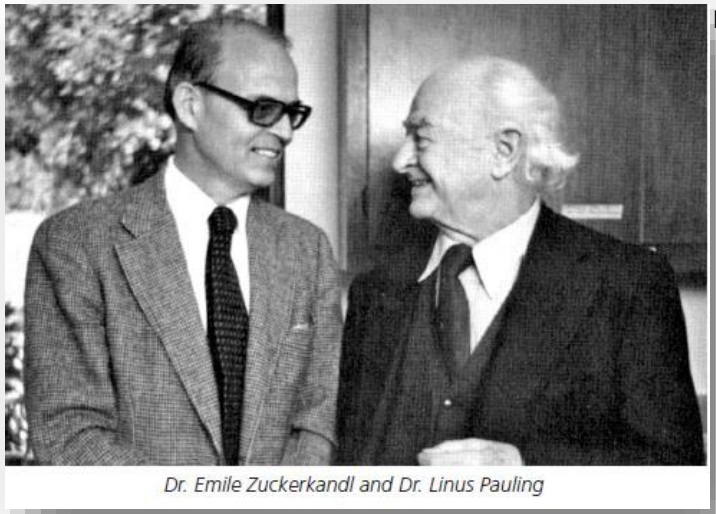
[4]

Ernst Haeckel (1834-1919)

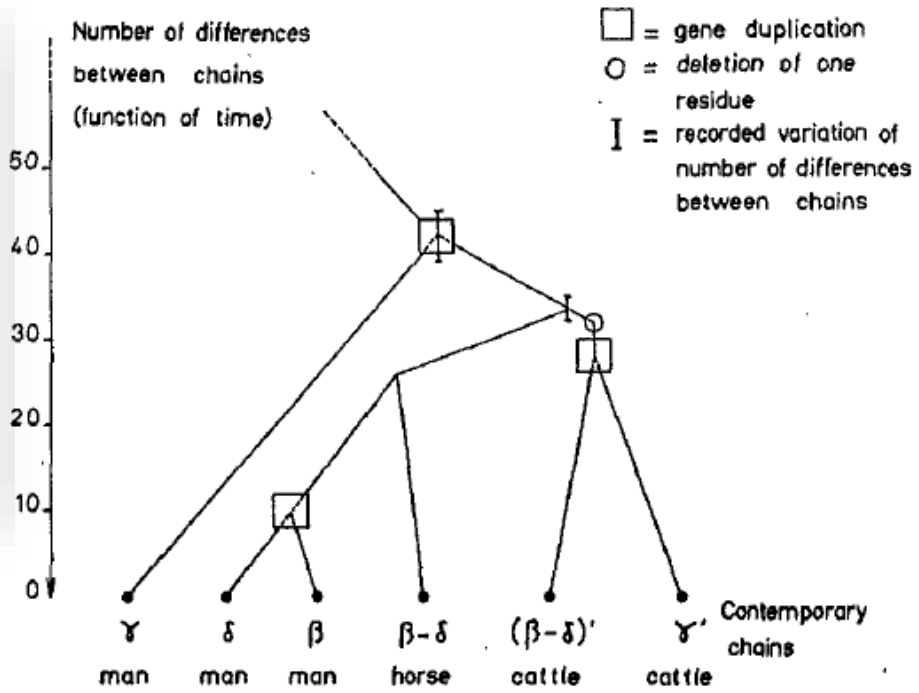
恩斯特・海克尔



系统发生树诞生前的“假定”树


[5]

Dr. Emile Zuckerkandl and Dr. Linus Pauling

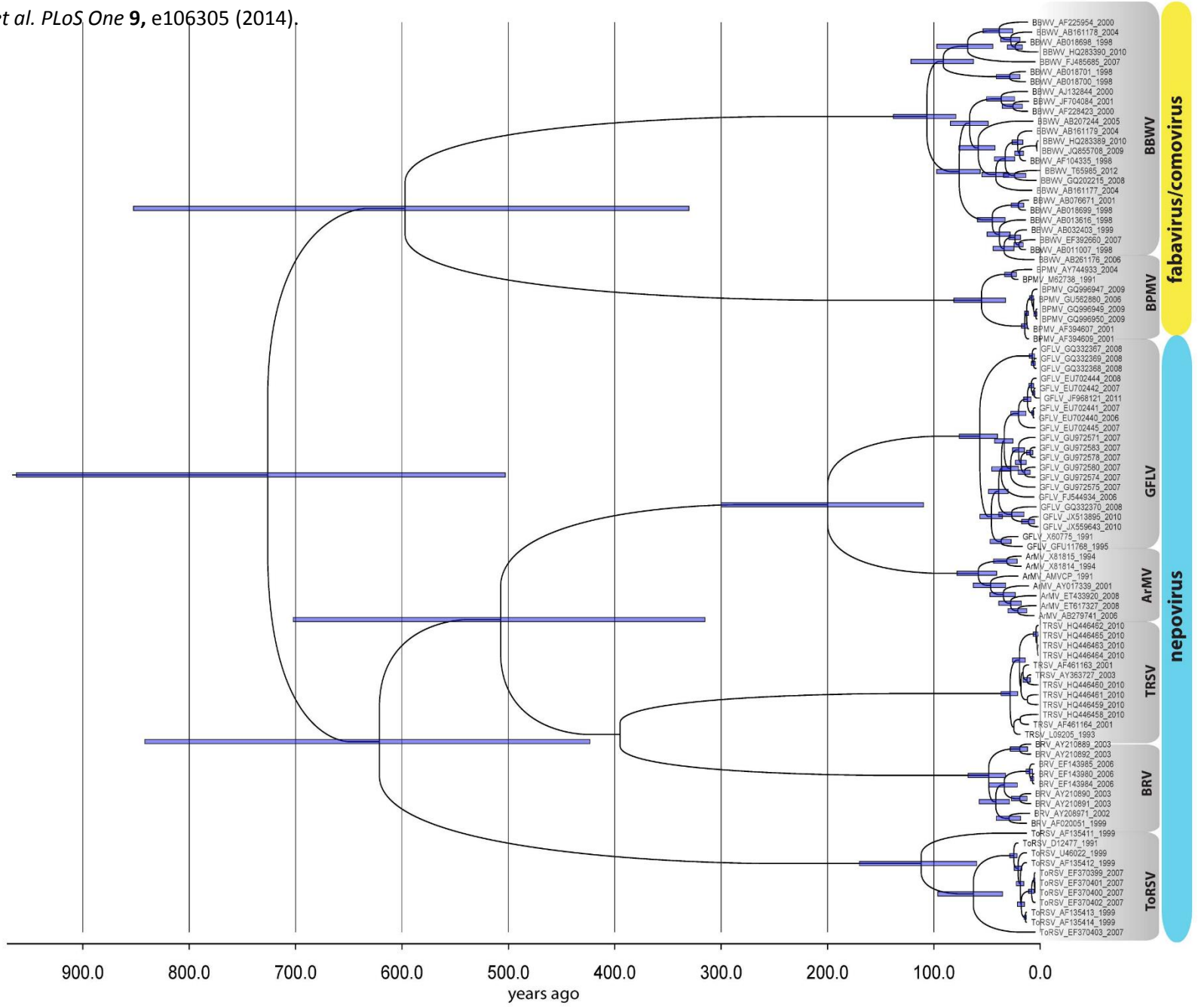


“Probable evolutionary relationship of some mammalian hemoglobin chains.”

Molecular evolutionary clock

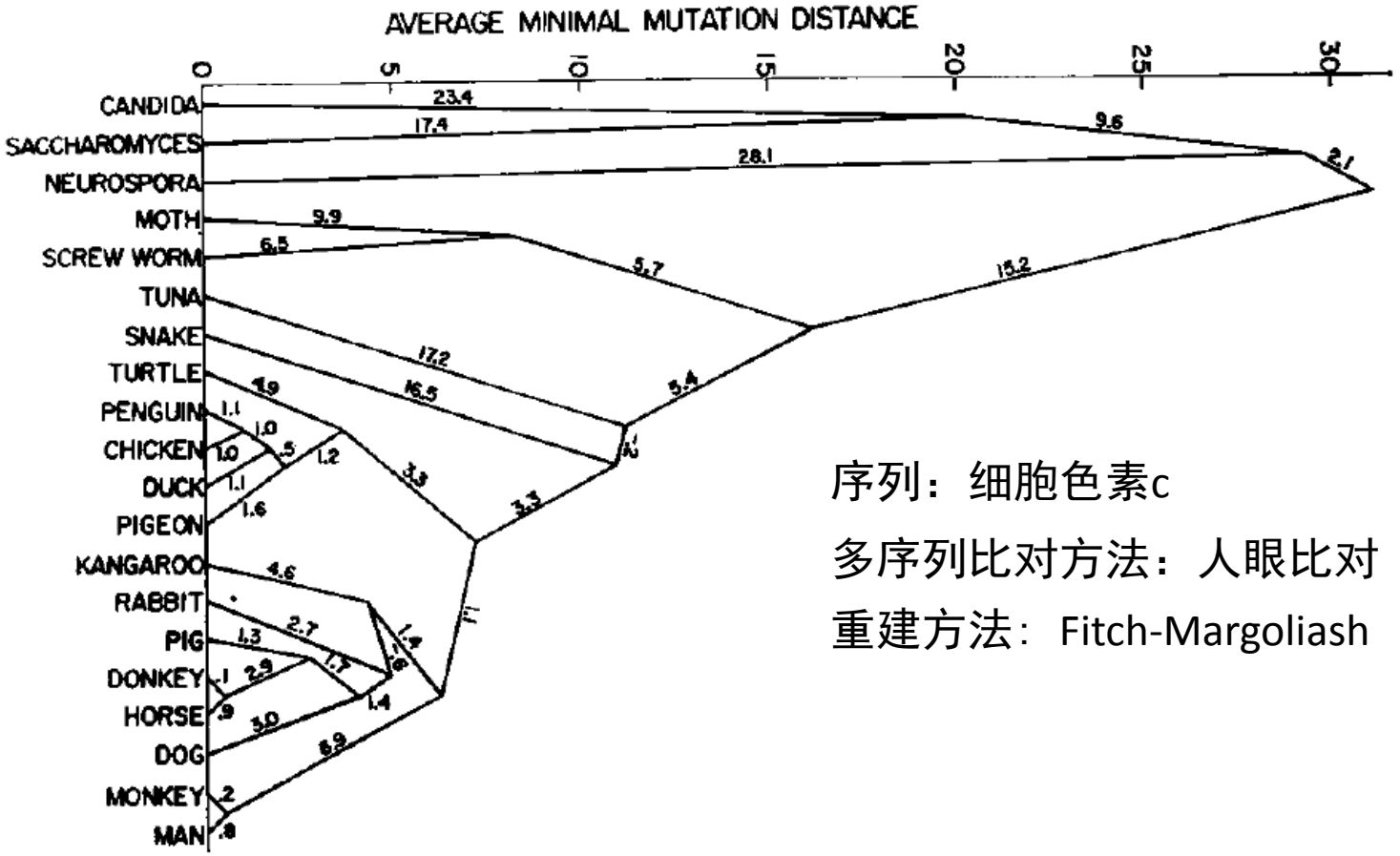
一个特定的大分子在所有的演化谱系中具有**恒定的演化速率**

Thompson *et al. PLoS One* **9,** e106305 (2014).



豇豆花叶病毒亚科(Comovirinae)壳蛋白序列构建的系统树

使用序列信息重建的第一棵系统发生树



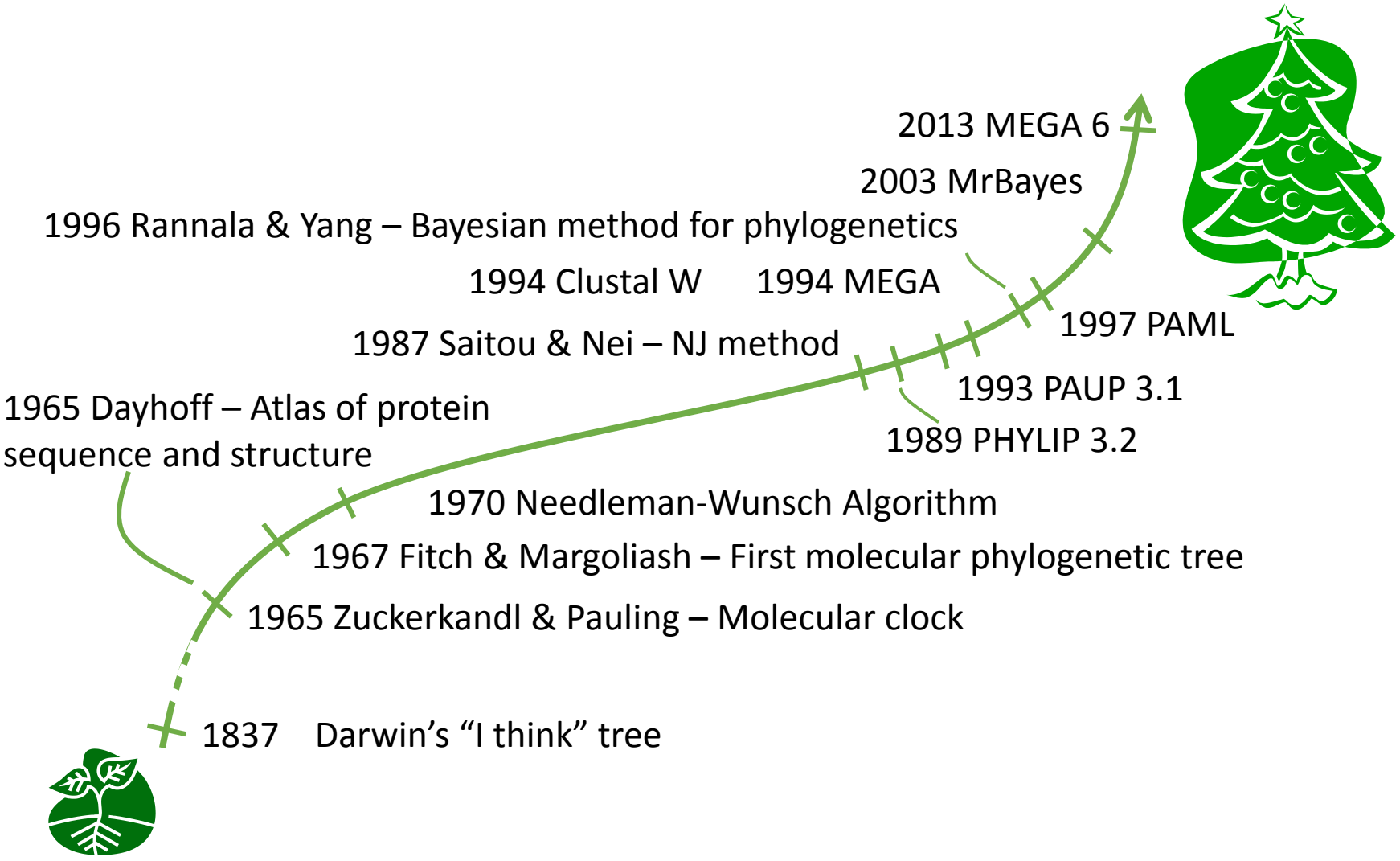
序列：细胞色素c

多序列比对方法：人眼比对

重建方法：Fitch-Margoliash

Fitch & Margoliash. *Science* **155,** 279-284 (1967).

系统发生树的时间轴

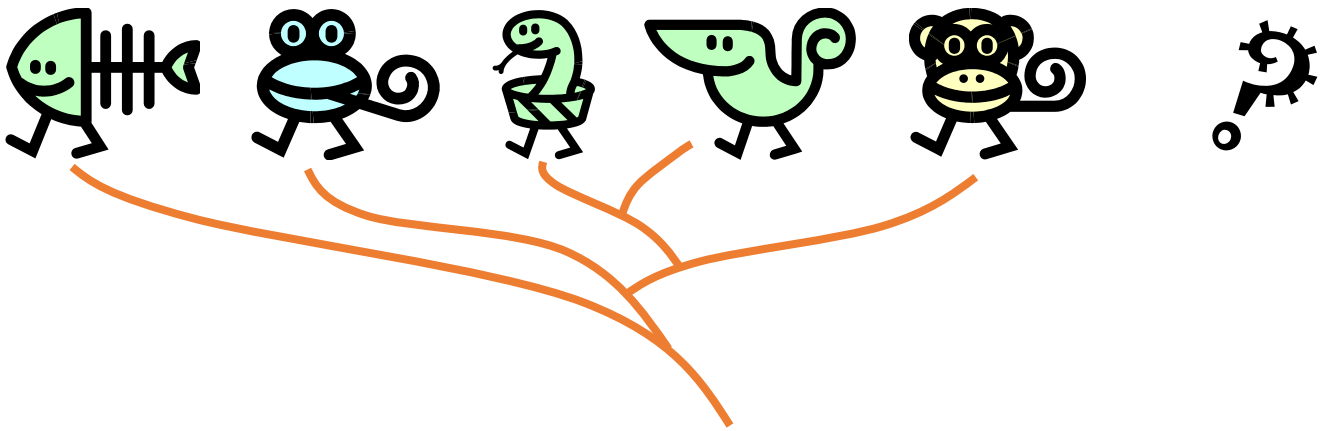


Outline

- What is phylogenetic tree: a historical perspective
- Why do we reconstruct trees?
- ABCs of molecular phylogenetics
- Methods in phylogenetic reconstruction
  - Data sets
  - Alignment
  - Phylogenetic analysis

为何要重建系统发生树？

- 了解和描绘物种间的相互关系和分子间的相互关系

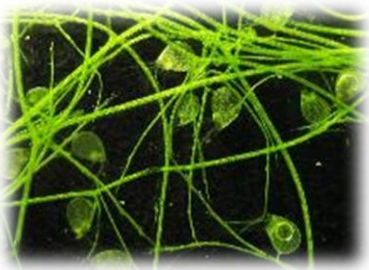


更实际一点呢？

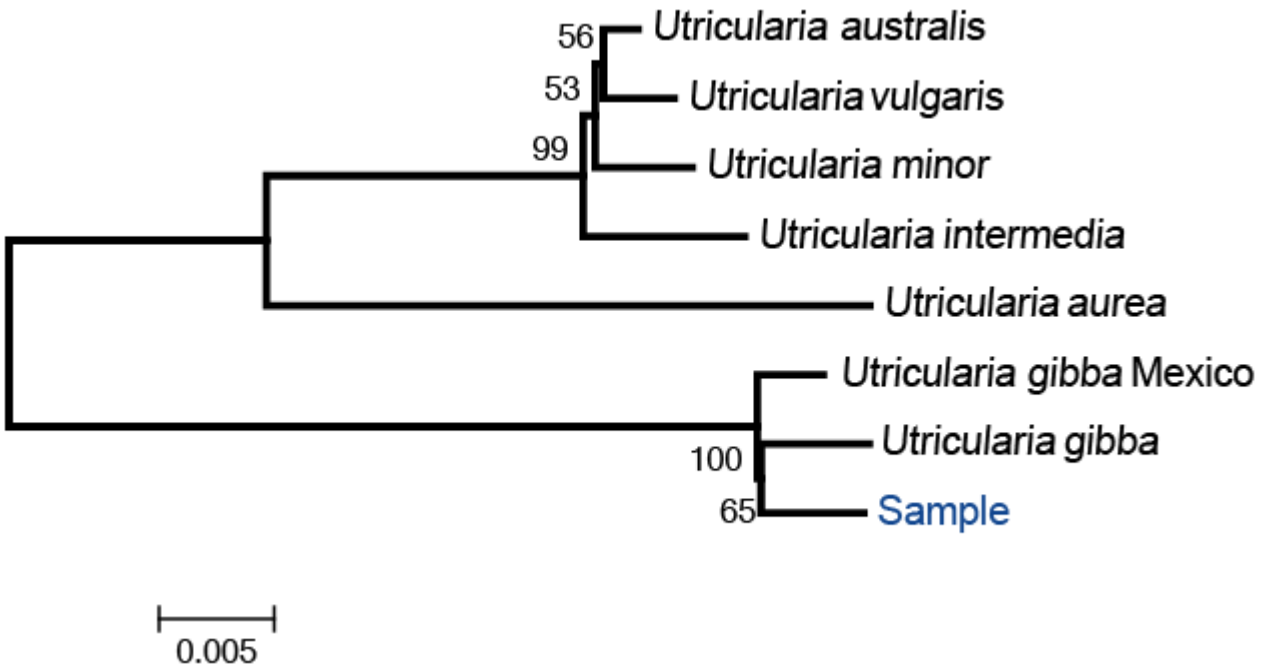
为何要重建系统发生树？

- Case 1

小雨同学希望在被子植物中，研究水生植物的RNA编辑与陆生植物的差异。她希望利用2013年完成基因组测序的少花狸藻(*Utricularia gibba*)作为研究材料。苦于匮乏分类学知识和缺少野外采集植物经验，她不得不在网上购买标有“少花狸藻”名字的植物。但得到材料后，小雨如何能知道手里的这种水草就是真正的少花狸藻呢？



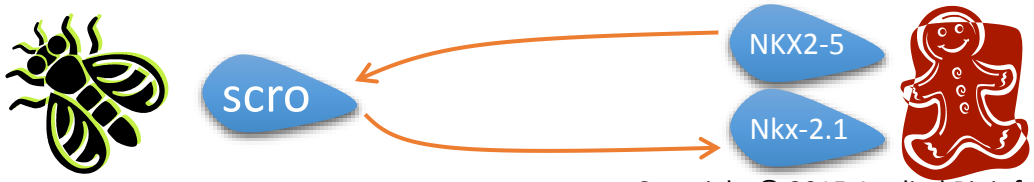
通过分子系统发生学鉴定物种



为何要重建系统发生树？

- Case 2

小黄在一个研究果蝇的实验室轮转，但他却对人类疾病很感兴趣。通过上课的机会他了解到人*NKX2-5*基因的突变可能导致先天性心脏病，他灵光一闪：*NKX2-5*是homeobox家族的基因，是否可能在果蝇中存在同源的基因呢？通过BLAST，他在果蝇中搜索到scro的蛋白质序列与NKX2-5的序列最相似；但当他把scro作为检测序列，在人中找到最相似的序列却是Nkx-2.1。小黄很困惑，到底果蝇的scro和人类的NKX2-5与Nkx-2.1有什么关系？



Outline

- What is phylogenetic tree: a historical perspective
- Why do we reconstruct trees?
- ABCs of molecular phylogenetics
- Methods in phylogenetic reconstruction
  - Data sets
  - Alignment
  - Phylogenetic analysis

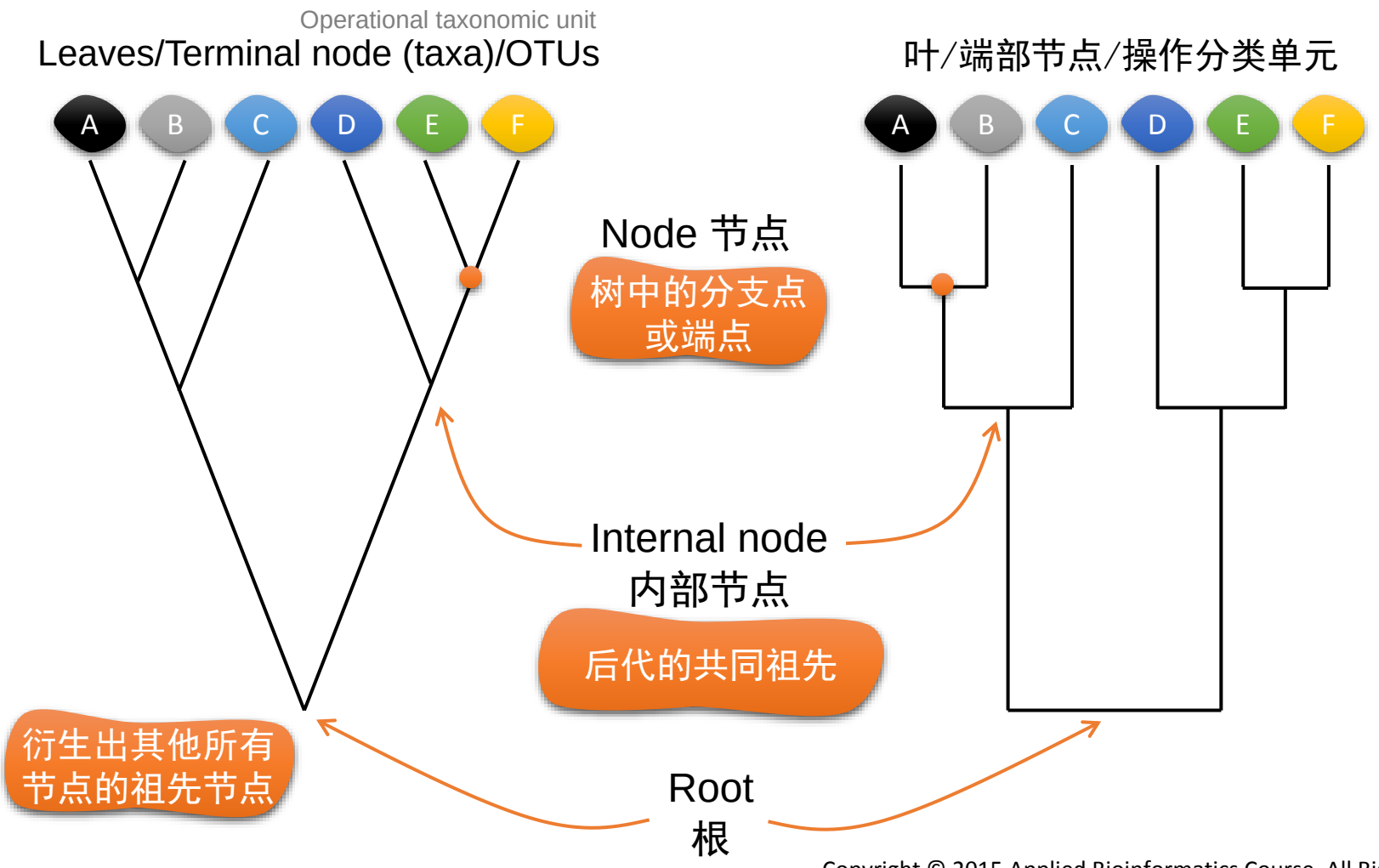
什么是分子系统发生学？

- 系统发育(Phylogeny)：生物体或分子间的演化关系
- 系统发生学(Phylogenetics)：利用各种性状，通过构建系统发生树研究上述关系的学科
- 分子系统发生学(Molecular phylogenetics)
- 系统发生树(Phylogenetic tree)：系统发生分析结果之“图像化隐喻”

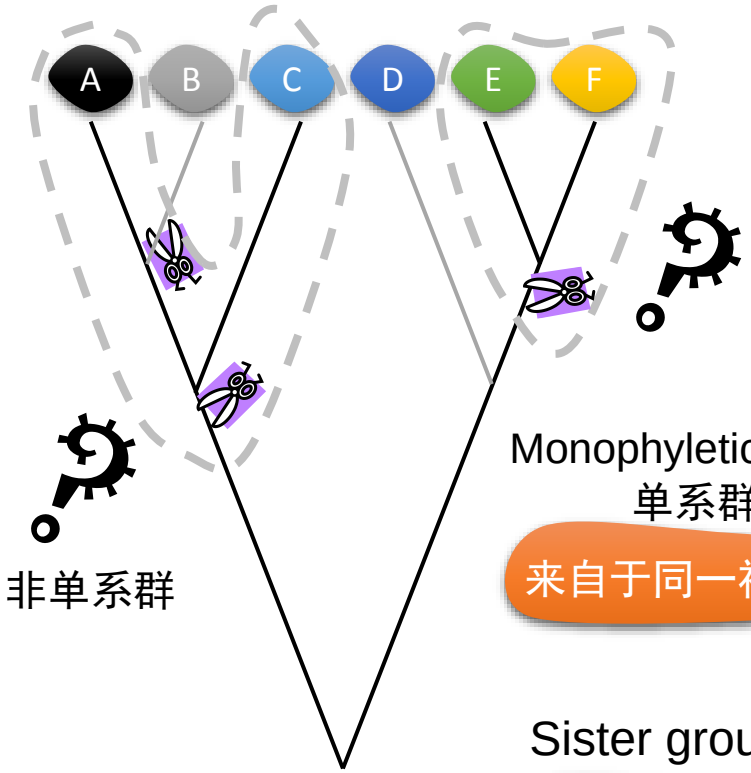
Introduction to Phylogenetic Tree

- Terms
- Rooted tree vs. Unrooted tree
- Cladogram vs. Phylogram
- Scaled tree vs. Unscaled tree
- Gene tree vs. Species tree

有根树的术语



有根树的术语



非单系群

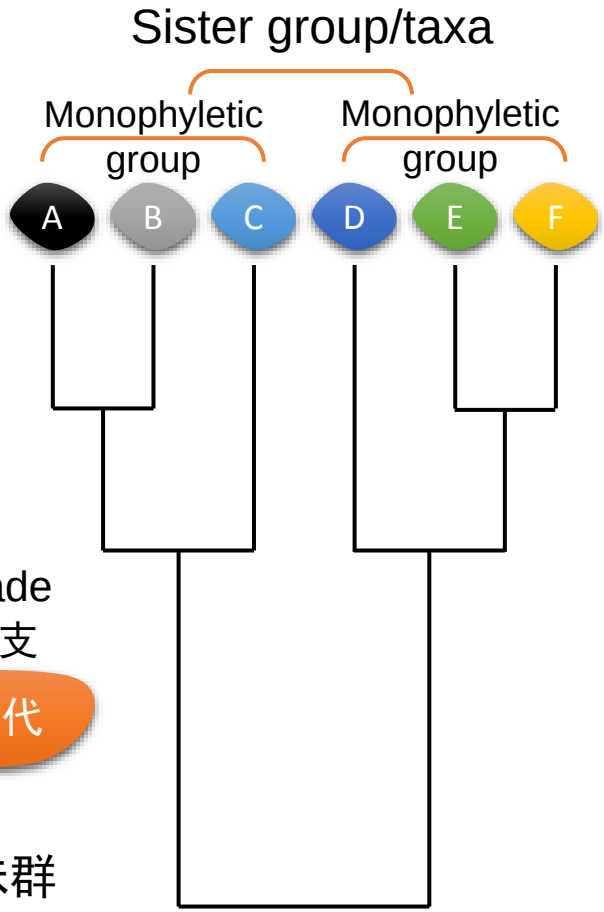
Monophyletic group
单系群

Clade
分支

来自于同一祖先的全部后代

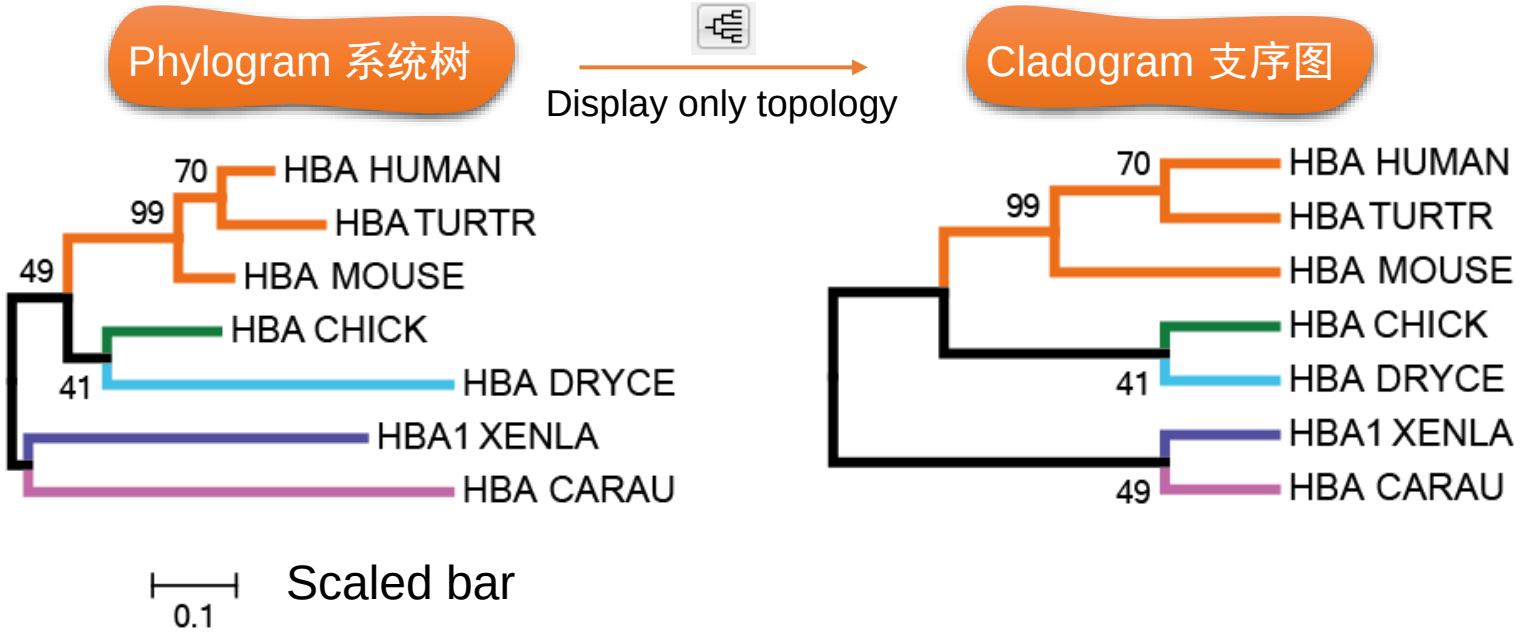
Sister group/taxa 姐妹群

任一节点上分歧的两个分支



系统树 vs. 支序图

一棵树的分支式样叫做一个**拓扑结构**。

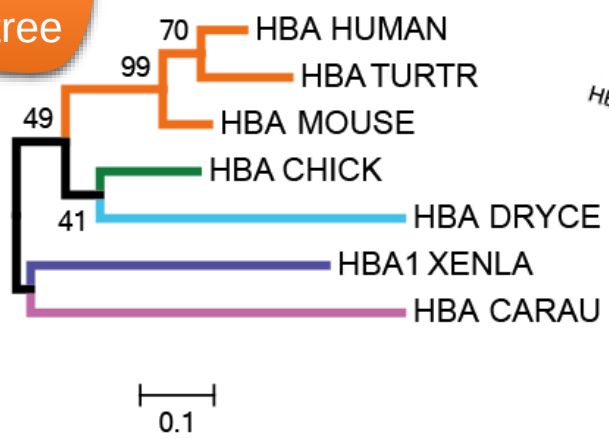


Scaled tree 标度树

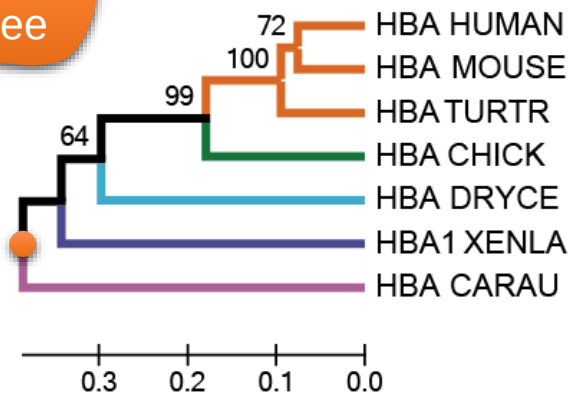
Unscaled tree 非标度树

无根树 vs. 有根树

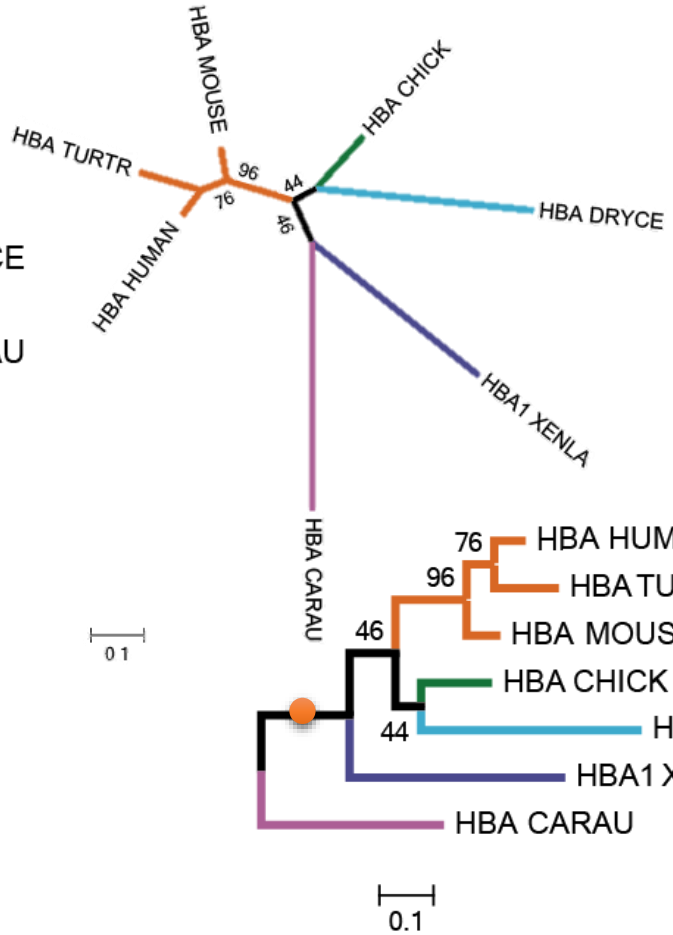
无根树
Unrooted tree



有根树
Rooted tree

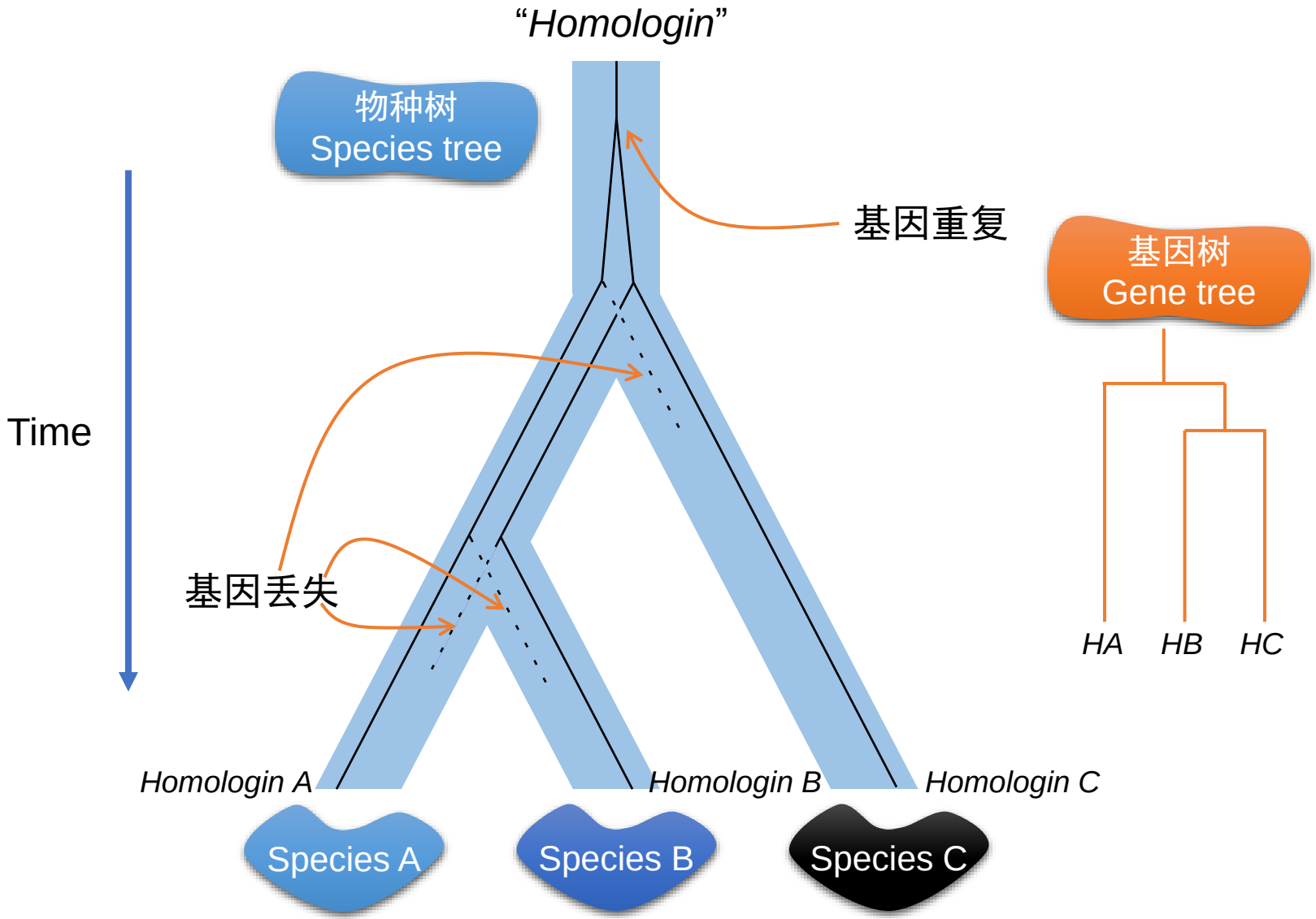


中点赋根法 Midpoint rooting



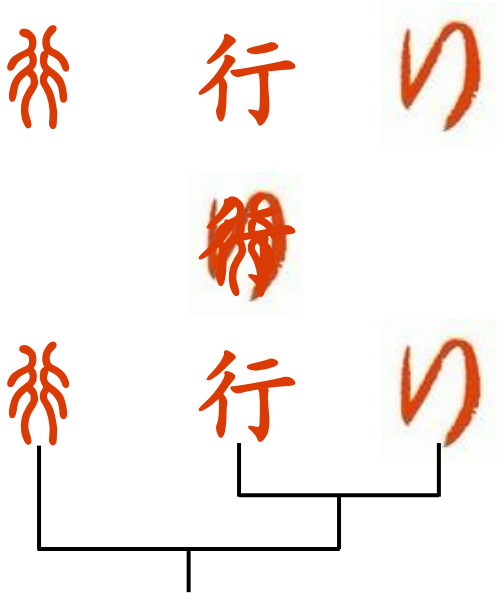
外类群赋根法 Outgroup rooting

基因树 vs. 物种树



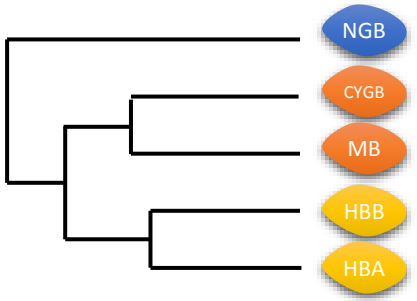
Outline

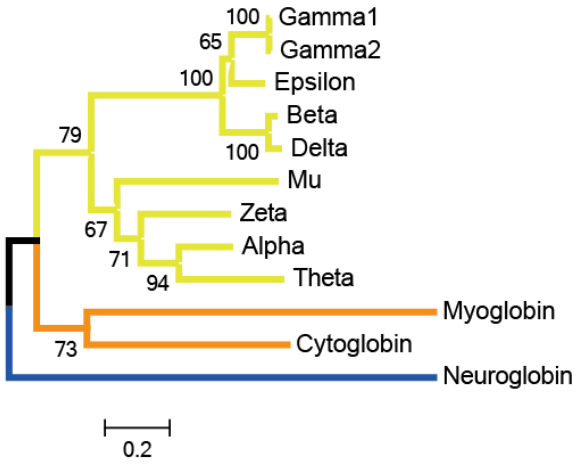
- What is phylogenetic tree: a historical perspective
- Why do we reconstruct trees?
- ABCs of molecular phylogenetics
- Methods in phylogenetic reconstruction
  - ◆数据集 Data sets
  - ◆序列比对 Alignment
  - ◆Phylogenetic analysis

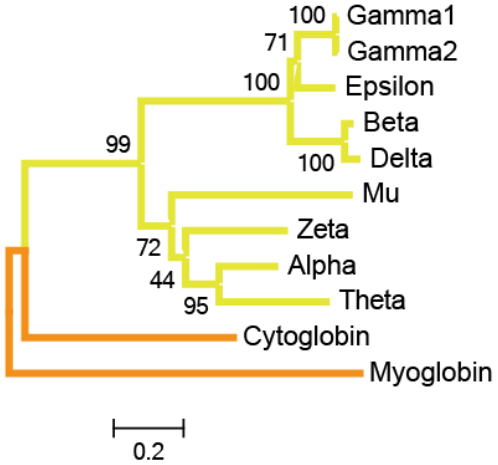


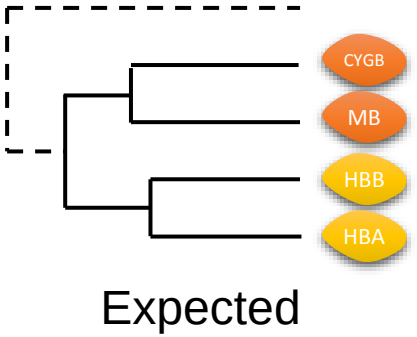
数据集的建立

- Well begun is half done.
- Sometimes things don't go well.
- If NGB was ignored…









Expected

合适的数据集才会告诉你真实的故事

氨基酸序列还是核苷酸序列?

Amino acid

- More possible character states (20)
- Alignment is easier
- **More conserved**
- No preferential codon usage

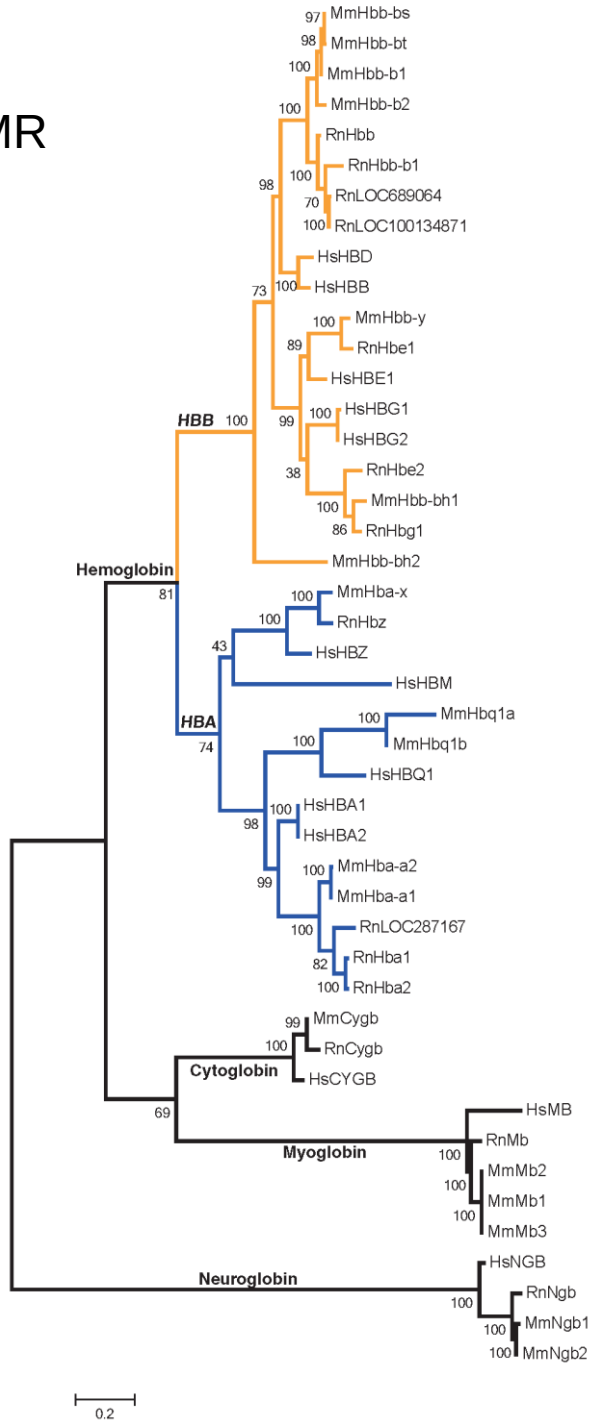
Nucleotide

- 4 character states
- More characters - better resolution
- **Evolutionary rate is rapid**
- Depicts synonymous and nonsynonymous substitutions

Globin_HMR

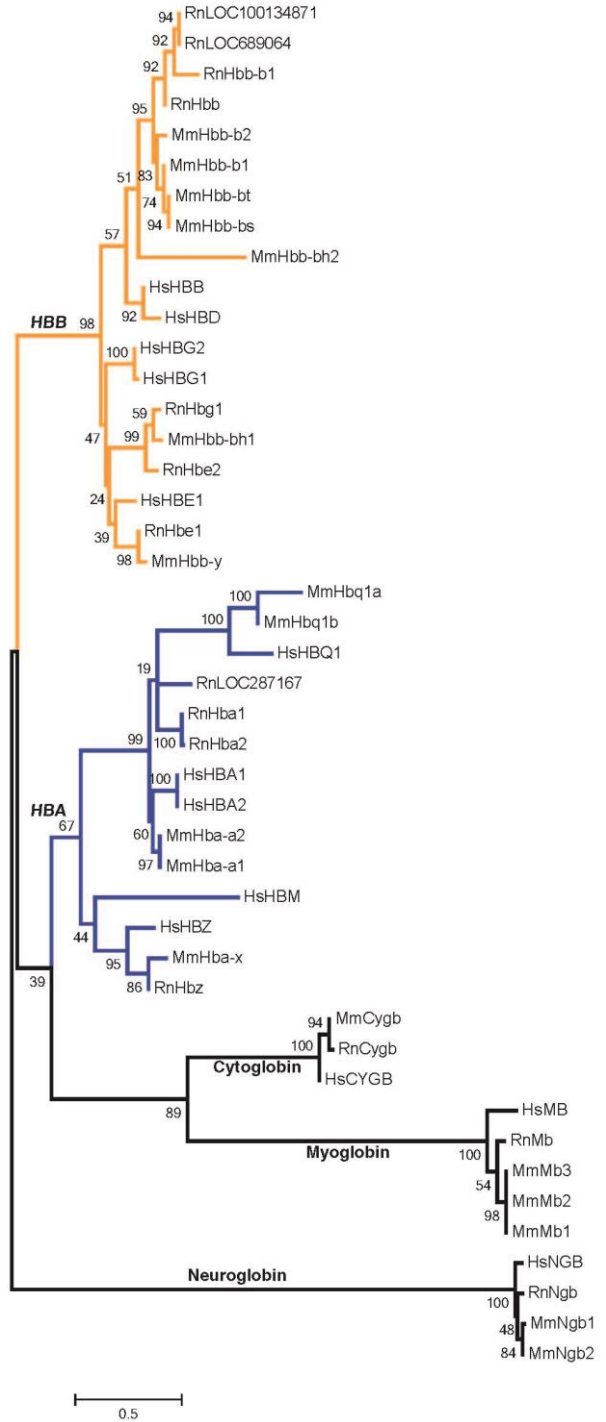
CDS

ML



PEP

ML



Alignment

Alignment may be the most critical step because it establishes positional correspondence in evolution.

- **What can you do?**

①Manual editing: correcting mismatching of key cofactor residues and residues of similar physicochemical properties

②Full alignment or parts of it (domain only)

③Remove ambiguously aligned regions(subjective process)

④Automatic approach: Rascal, NorMD and Gblocks

⑤Statistical models to correct homoplasy

⑥Using a γ correction factor to correct site-dependent rate variation

- **Choosing substitution models**

①Nucleotide: Juke-Cantor Model and Kimura Model

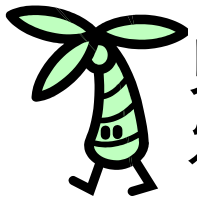
②Protein: PAM or JTT amino acid substitution matrix

系统发生分析方法

- Distance-matrix methods 距离矩阵方法
  - UPGMA
  - Neighbor-joining method
  - Minimum evolution method

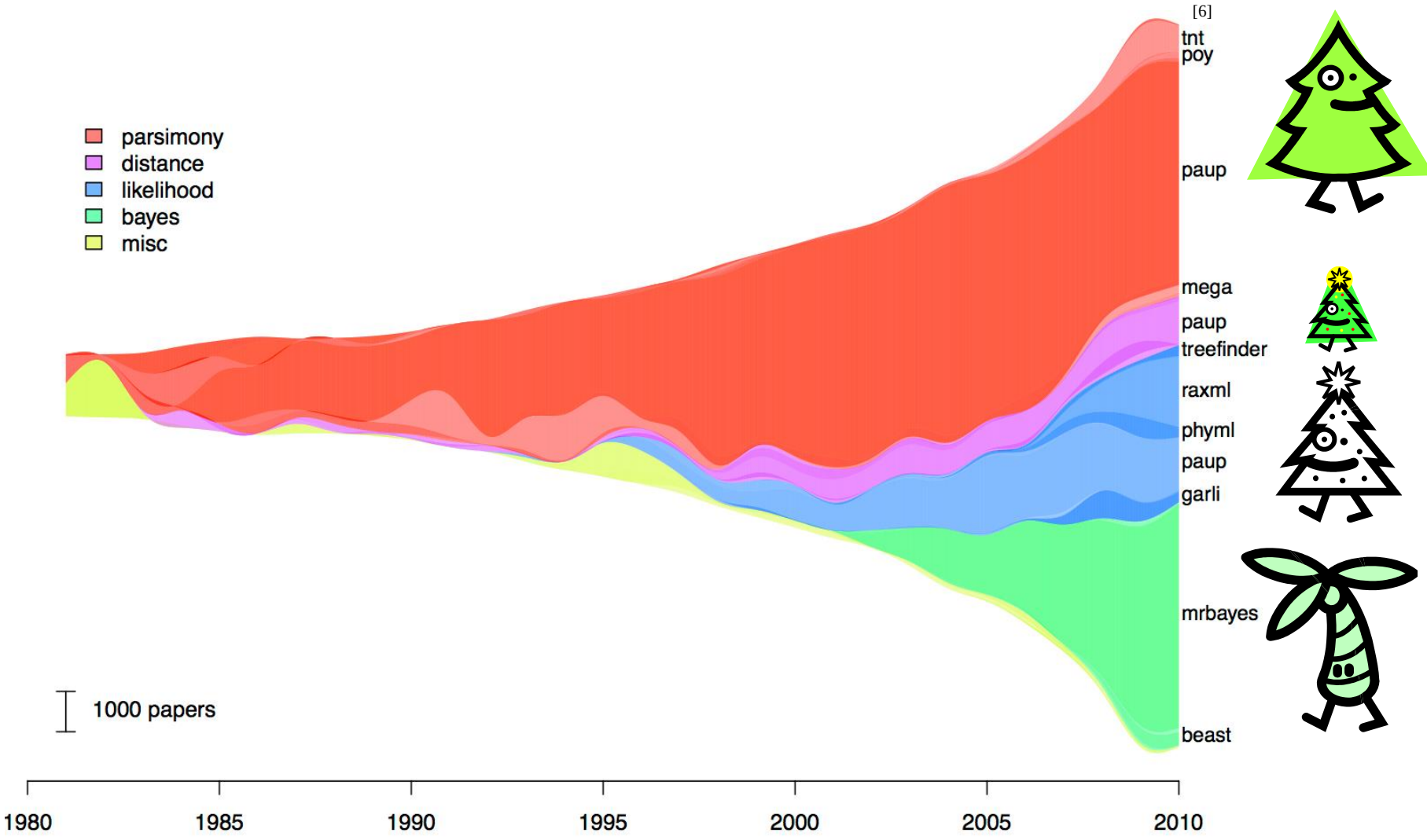
- Maximum parsimony method 最大简约法

- Maximum likelihood method 最大似然法



- Bayesian inference of phylogeny 贝叶斯系统发生推断法

系统发生分析方法的增长趋势



系统发生分析方法基本原理

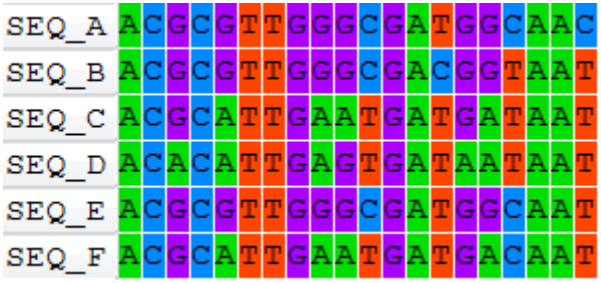
距离矩阵方法	• 将比对的序列数据转化为距离矩阵，根据序列间距离拟合成一棵树
最大简约法	• 在所有可能的系统树中，寻找解释数据集中性状状态改变需要最少额外步骤的树。
最大似然法	• 选择一个适合数据集的模型，对指定拓扑结构的树优化分支长度，使所计算的拓扑结构的似然值最大化。通过对不同拓扑结构树的似然值进行计算，将具有最大似然值的树作为指定模型下的最佳估计。
贝叶斯系统发生推断法	• 基于进化树的先验概率，在指定进化模型和树的分支形式的条件下，利用得到的似然函数计算树的后验概率。得到具有最高后验概率值的树即为系统发育关系的最佳估计。

系统发生分析方法

- Distance-matrix methods
  - UPGMA
  - Neighbor-joining method 邻接法
  - Minimum evolution method
- Maximum parsimony method
- Maximum likelihood method
- Bayesian inference of phylogeny

距离是什么？怎么计算？

演化距离是衡量一对序列间核苷酸（或氨基酸）替代数目的值



No. of difference

d_{ij}	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B	3				
SEQ_C	7	6			
SEQ_D	8	7	3		
SEQ_E	1	2	6	7	
SEQ_F	6	7	1	4	5

理想很骨感，现实很丰满。

SEQ_anc. ACGCATCGAGTGATGATAGT

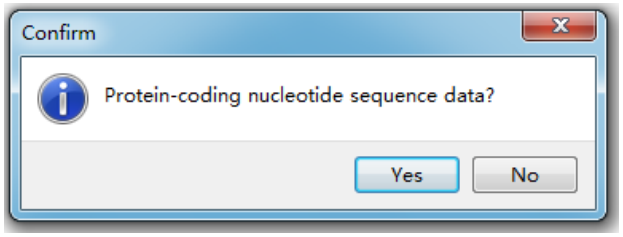
ACGCATCGAGTGATGATAGT
T A C A
G

SEQ_C ACGCATTGAATGATGATAAT

ACGCATCGAGTGATGATAGT
A G C A A
T

SEQ_D ACACATTGAGTGATAATAAT

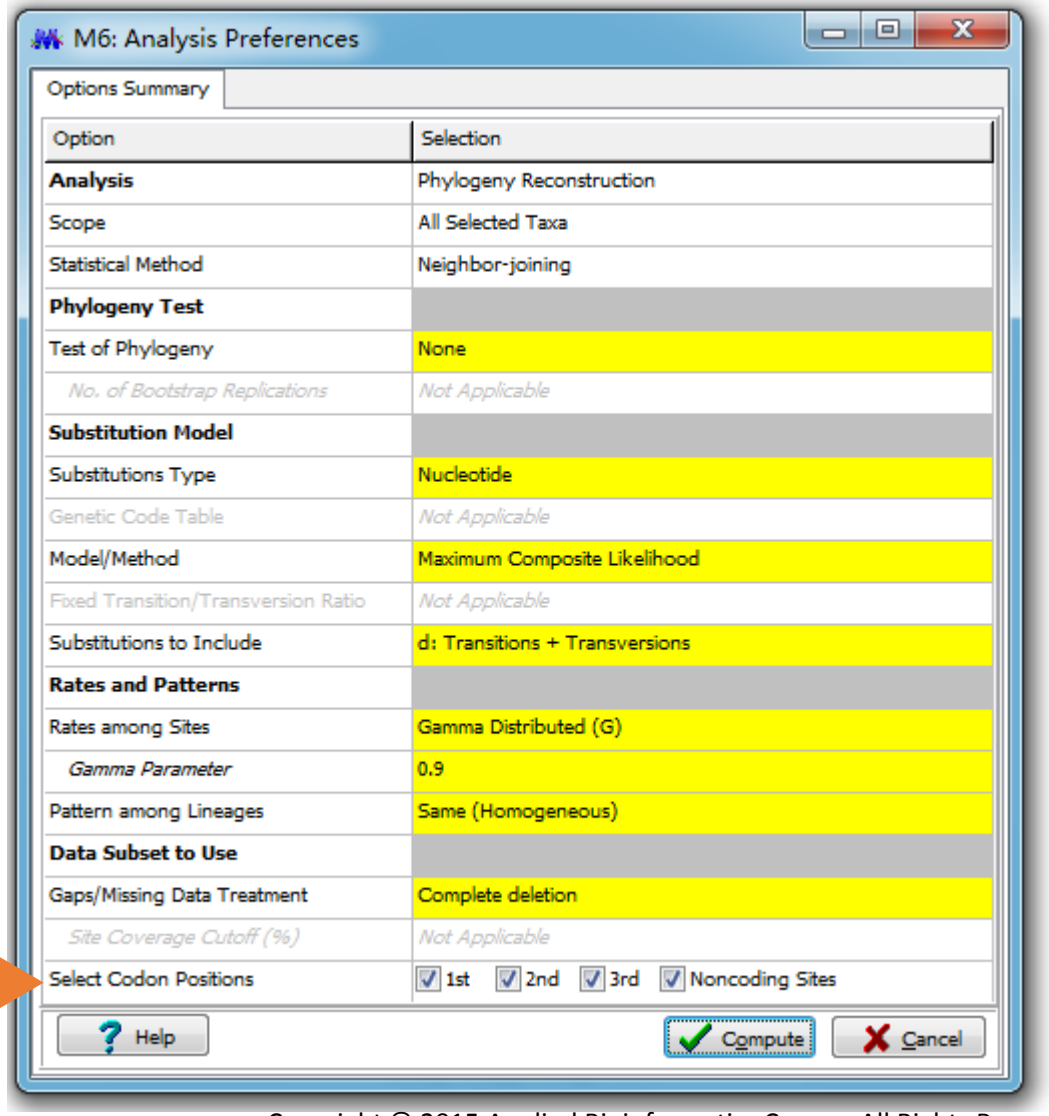
核苷酸替代模型



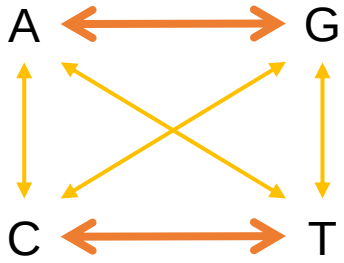
If yes…

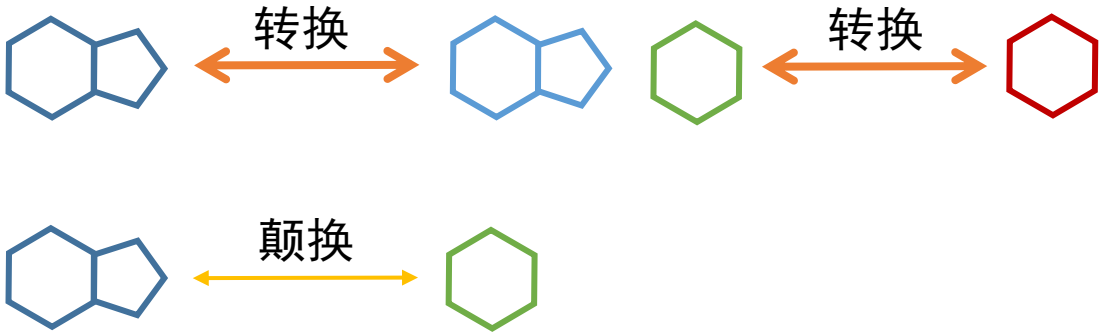
Model - Nucleotide

No. of differences
p-distance
Jukes-Cantor Model
Tajima-Nei Model
Kimura 2-Parameter Model
Tamura 3-Parameter Model
Tamura-Nei Model
Log-Det Method
Maximum Composite Likelihood Model

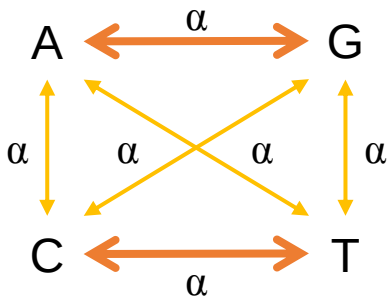


碱基替代的类型





Jukes-Cantor Model



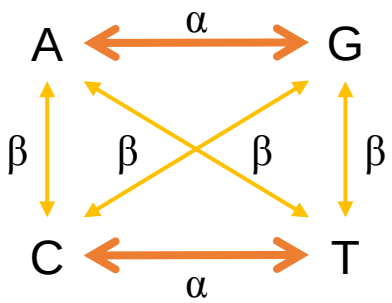
距离

每个位点上核苷酸的替代数

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

p 是差异核苷酸位点的比例

Kimura 2-Parameter Model



距离

$$d = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q)$$

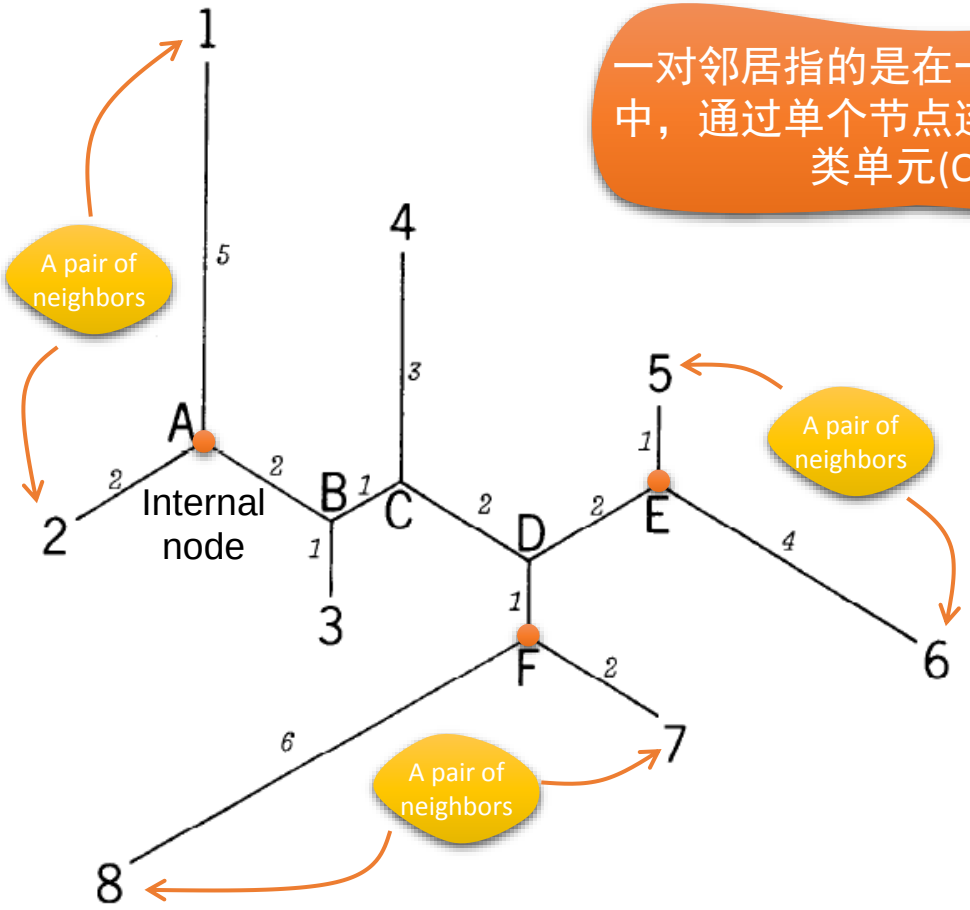
P 和 Q 分别是转换和颠换的频率

TABLE 3.1 Models of nucleotide substitution

O\S[a]	A	T	C	G
a. Two-parameter model (Kimura 1980)				
A	$1-\alpha-2\beta$	β	β	α
T	β	$1-\alpha-2\beta$	α	β
C	β	α	$1-\alpha-2\beta$	β
G	α	β	β	$1-\alpha-2\beta$
b. Four-parameter model (Blaisdell 1985)				
A	$1-\alpha-2\gamma$	γ	γ	α
T	δ	$1-\alpha-2\delta$	α	δ
C	δ	β	$1-\beta-2\delta$	δ
G	β	γ	γ	$1-\beta-2\gamma$
c. Six-parameter model (Kimura 1981a)				
A	$1-2\alpha-\gamma$	γ	α	α
T	δ	$1-2\alpha-\delta$	α	α
C	β	β	$1-2\beta-\varepsilon$	ε
G	β	β	ξ	$1-2\beta-\xi$
d. Nine-parameter model				
A	$1-g_T\beta_1-g_C\gamma_1-g_G\alpha_1$	$g_T\beta_1$	$g_C\gamma_1$	$g_G\alpha_1$
T	$g_A\beta_1$	$1-g_A\beta_1-g_C\alpha_2-g_G\gamma_2$	$g_C\alpha_2$	$g_G\gamma_2$
C	$g_A\gamma_1$	$g_T\alpha_2$	$1-g_A\gamma_1-g_T\alpha_2-g_G\beta_2$	$g_G\beta_2$
G	$g_A\alpha_1$	$g_T\gamma_2$	$g_C\beta_2$	$1-g_A\alpha_1-g_T\gamma_2-g_C\beta_2$
e. General model				
A	$1-\alpha_{12}-\alpha_{13}-\alpha_{14}$	α_{12}	α_{13}	α_{14}
T	α_{21}	$1-\alpha_{21}-\alpha_{23}-\alpha_{24}$	α_{23}	α_{24}
C	α_{31}	α_{32}	$1-\alpha_{31}-\alpha_{32}-\alpha_{34}$	α_{34}
G	α_{41}	α_{42}	α_{43}	$1-\alpha_{41}-\alpha_{42}-\alpha_{43}$

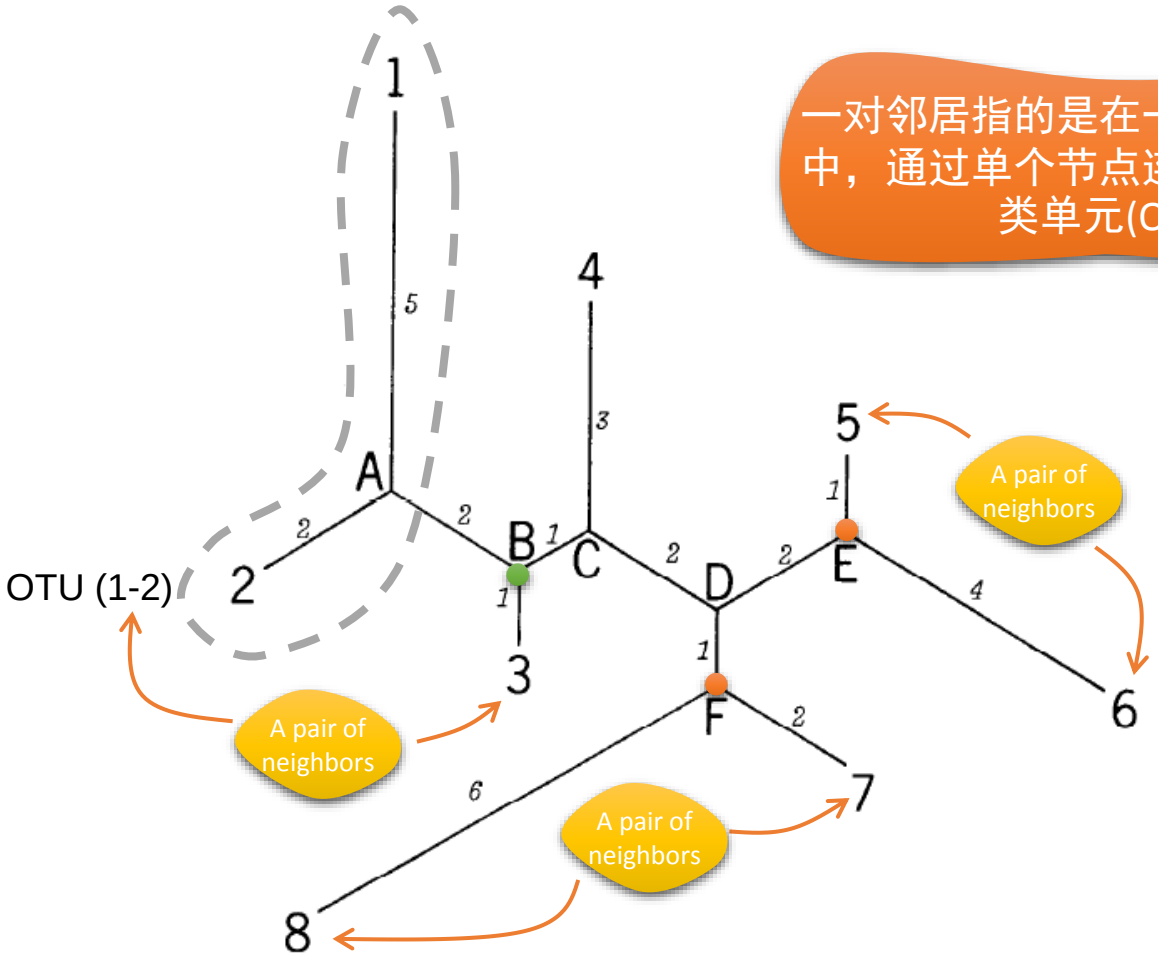
[a]O, Original nucleotide; S, substitute nucleotide.

何为邻居？

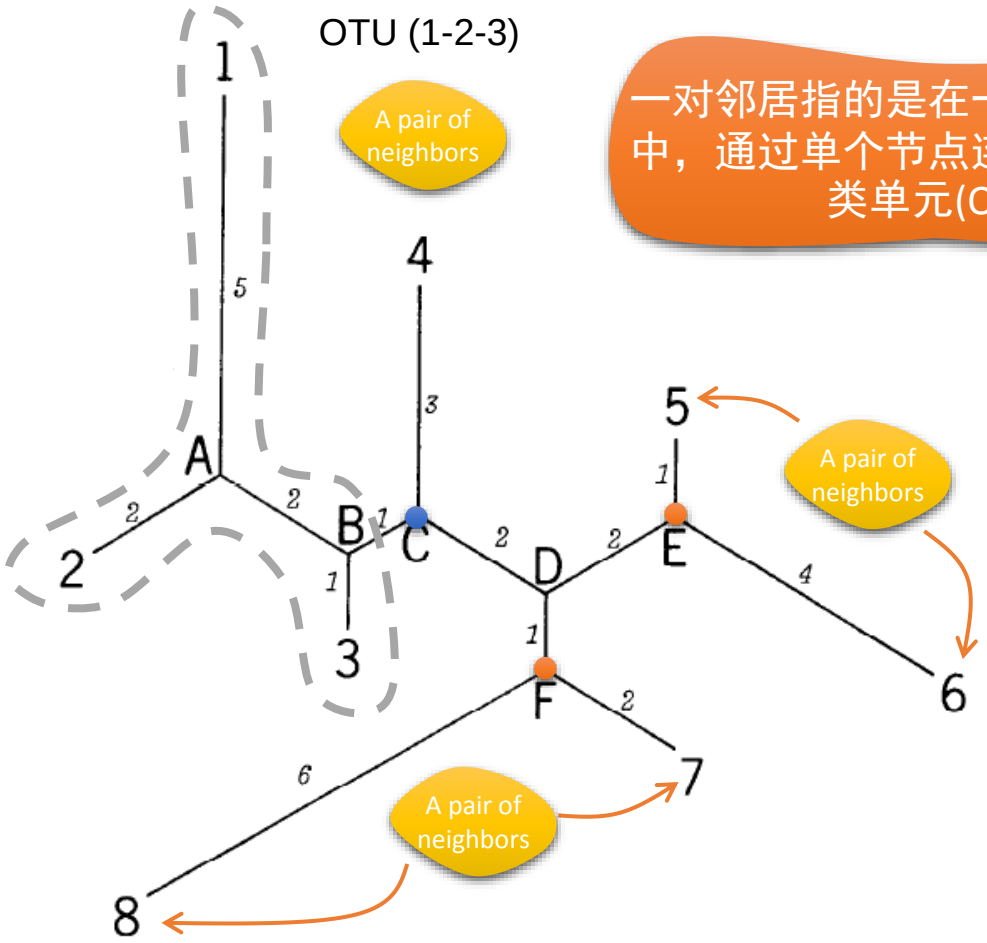


一对邻居指的是在一个无根的二叉树中，通过单个节点连接的一对操作分类单元(OTUs)。

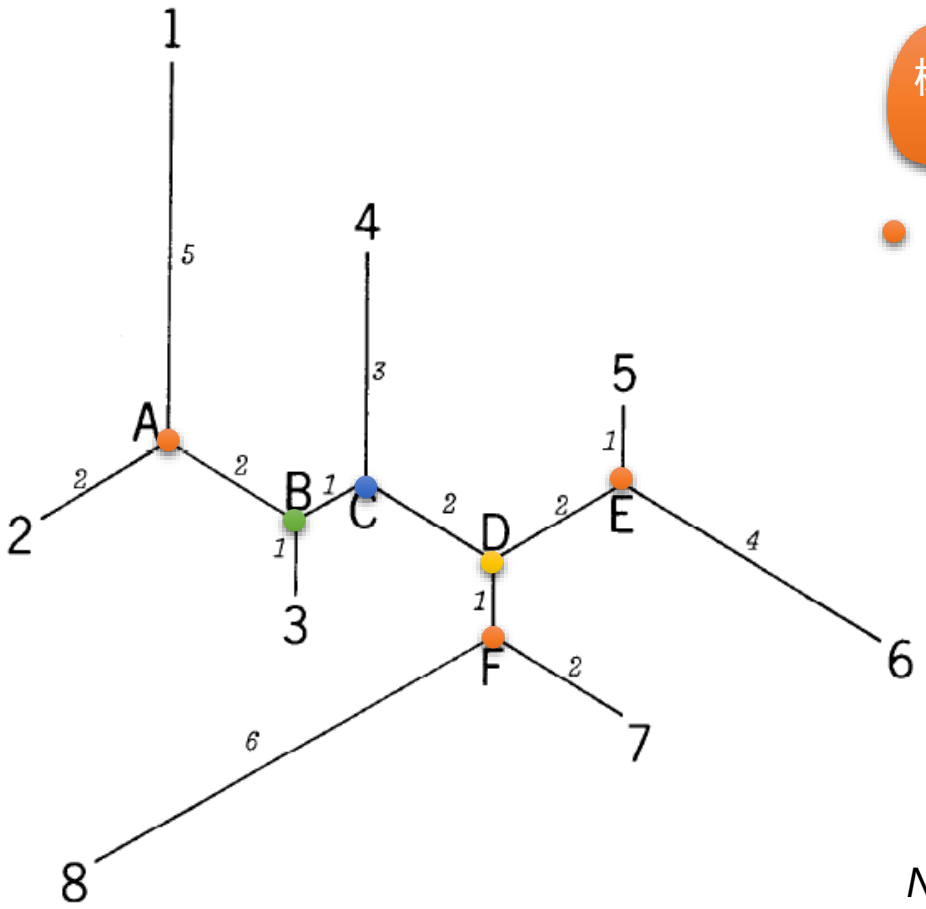
何为邻居？



何为邻居？



何为邻居？

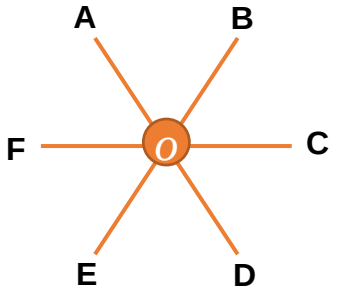


树的拓扑结构可以通过下述几对邻居进行描述：

- [1, 2]　[5, 6]　[7, 8]
- [1-2, 3]
- [1-2-3, 4]
- [5-6, 7-8]

$N - 2 = 6$ pairs of neighbors

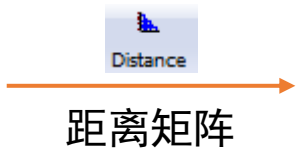
邻接法的算法



最开始的星状树——
不知道谁与谁是邻居

SEQ_A	ACGCGTTGGGCGATGGCAAC
SEQ_B	ACGCGTTGGGCGACGGTAAT
SEQ_C	ACGCATTGAATGATGATAAT
SEQ_D	ACACATTGAGTGATAATAAT
SEQ_E	ACGCGTTGGGCGATGGCAAT
SEQ_F	ACGCATTGAATGATGACAAT

操作分类单元数目 $N = 6$



距离矩阵

d_{ij}	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B	3				
SEQ_C	7	6			
SEQ_D	8	7	3		
SEQ_E	1	2	6	7	
SEQ_F	6	7	1	4	5

邻接法使用修正后的距离矩阵M，而不是直接使用距离矩阵决定哪两个OTUs聚在一起。

$$r_i = \sum_{k=1}^{N} d_{ik}$$

$$M_{ij} = (N - 2)d_{ij} - r_i - r_j$$

$r_{SEQ\_A} = 3 + 7 + 8 + 1 + 6 = 25$

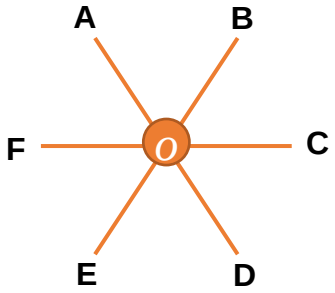
$r_{SEQ\_B} = 3 + 6 + 7 + 2 + 7 = 25$

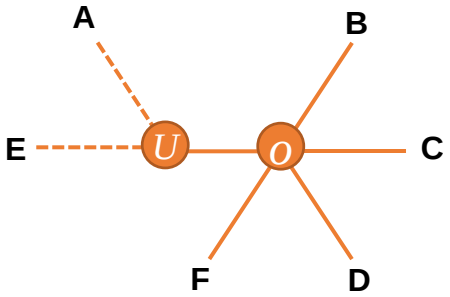
$r_{SEQ\_C} = 23$ $\quad r_{SEQ\_D} = 29$

$r_{SEQ\_E} = 21$ $\quad r_{SEQ\_F} = 23$

M_{ij}	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B	-38				
SEQ_C	-20	-24			
SEQ_D	-22	-26	-40		
SEQ_E	-42	-38	-20	-22	
SEQ_F	-24	-20	-42	-36	-24

邻接法的算法





M_{ij}	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B	-38				
SEQ_C	-20	-24			
SEQ_D	-22	-26	-40		
SEQ_E	**-42**	-38	-20	-22	
SEQ_F	-24	-20	-42	-36	-24

d_{ij}	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B	3				
SEQ_C	7	6			
SEQ_D	8	7	3		
SEQ_E	1	2	6	7	
SEQ_F	6	7	1	4	5

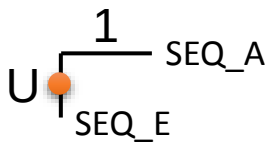
挑选M中最大的值，对应的两条序列为“邻居”。制造一个新节点u，它是最近“邻居”i和j的一个内部节点，然后计算u到i和u到j的分支长度。

$$S_{iu} = \frac{1}{2(N-2)}[(N-2)d_{ij} + r_i - r_j]$$

$$S_{ju} = d_{ij} - S_{iu}$$

$$\therefore \quad S_{\text{SEQ\_A-U}} = \frac{1}{2(6-2)}[(6-2)\times 1 + 25 - 21] = 1$$

$$S_{\text{SEQ\_E-U}} = 1 - 1 = 0$$



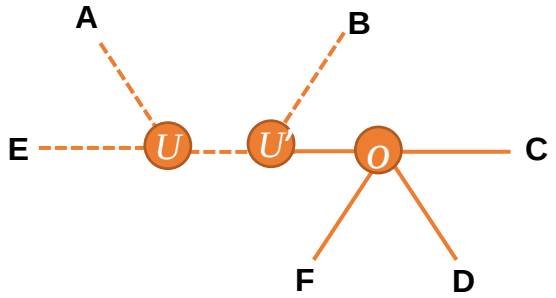
从节点u到节点k（除了i和j）的距离：

$$d_{ku} = \frac{1}{2}(d_{ik} + d_{jk} - d_{ij})$$

从矩阵中删去节点i和j。

d_{ij}	U	SEQ_B	SEQ_C	SEQ_D
SEQ_B	2			
SEQ_C	6	6		
SEQ_D	7	7	3	
SEQ_F	5	7	1	4

邻接法的算法



d_{ij}	U	SEQ_B	SEQ_C	SEQ_D
SEQ_B	2			
SEQ_C	6	6		
SEQ_D	7	7	3	
SEQ_F	5	7	1	4

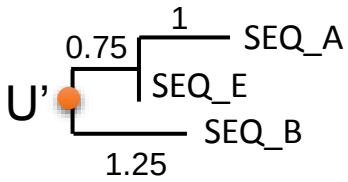
$$r_i = \sum_{k=1}^{N} d_{ik}$$

$$M_{ij} = (N - 2)d_{ij} - r_i - r_j$$

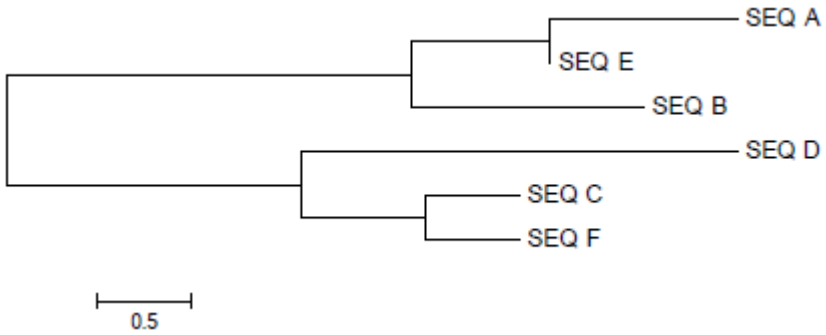
M_{ij}	U	SEQ_B	SEQ_C	SEQ_D
SEQ_B	**-36**			
SEQ_C	-18	-20		
SEQ_D	-20	-22	-28	
SEQ_F	-22	-18	-30	-26

$$S_{\text{U-U'}} = \frac{1}{2(5-2)}[(5 - 2)\times 2 + 20 - 22] = 0.67$$

$$S_{\text{SEQ\_B-U'}} = 2 - 0.67 = 1.33$$

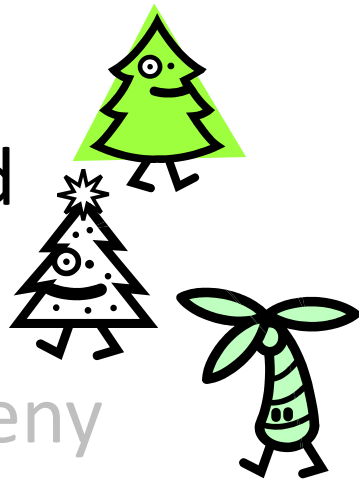


重复上述过程直到只有两个邻居为止。



Phylogenetic methods

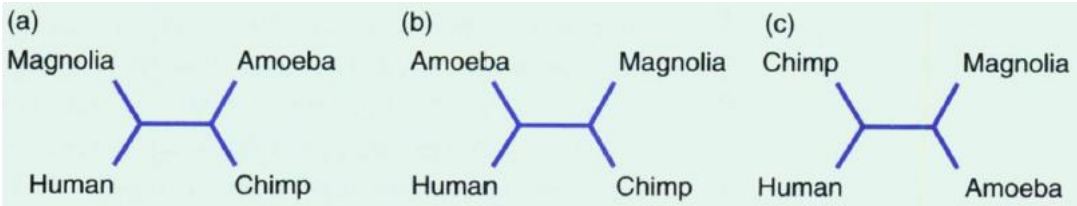
- Distance-matrix methods
  - UPGMA
  - Neighbor-joining method
  - Minimum evolution method
- Maximum parsimony method
- Maximum likelihood method
- Bayesian inference of phylogeny



最大简约法

简约性原则：生物演化“按照”最短的步骤进行（演化的步骤越少，其实际发生的可能性越大）
例：阿米巴虫，玉兰树，黑猩猩和人之间的相互关系如何？

4物种共有特征：1000
每个物种特有：10
人与黑猩猩共有：100



Number of evolutionary events

	(a)	(b)	(c)
Characters shared by all 4 species	1000	1000	1000
Characters unique to each of the 4 species	40	40	40
Characters shared by chimps and humans	200	200	100
	1240	1240	1140

按简约性原则，只有(c)符合要求

Test of Phylogeny

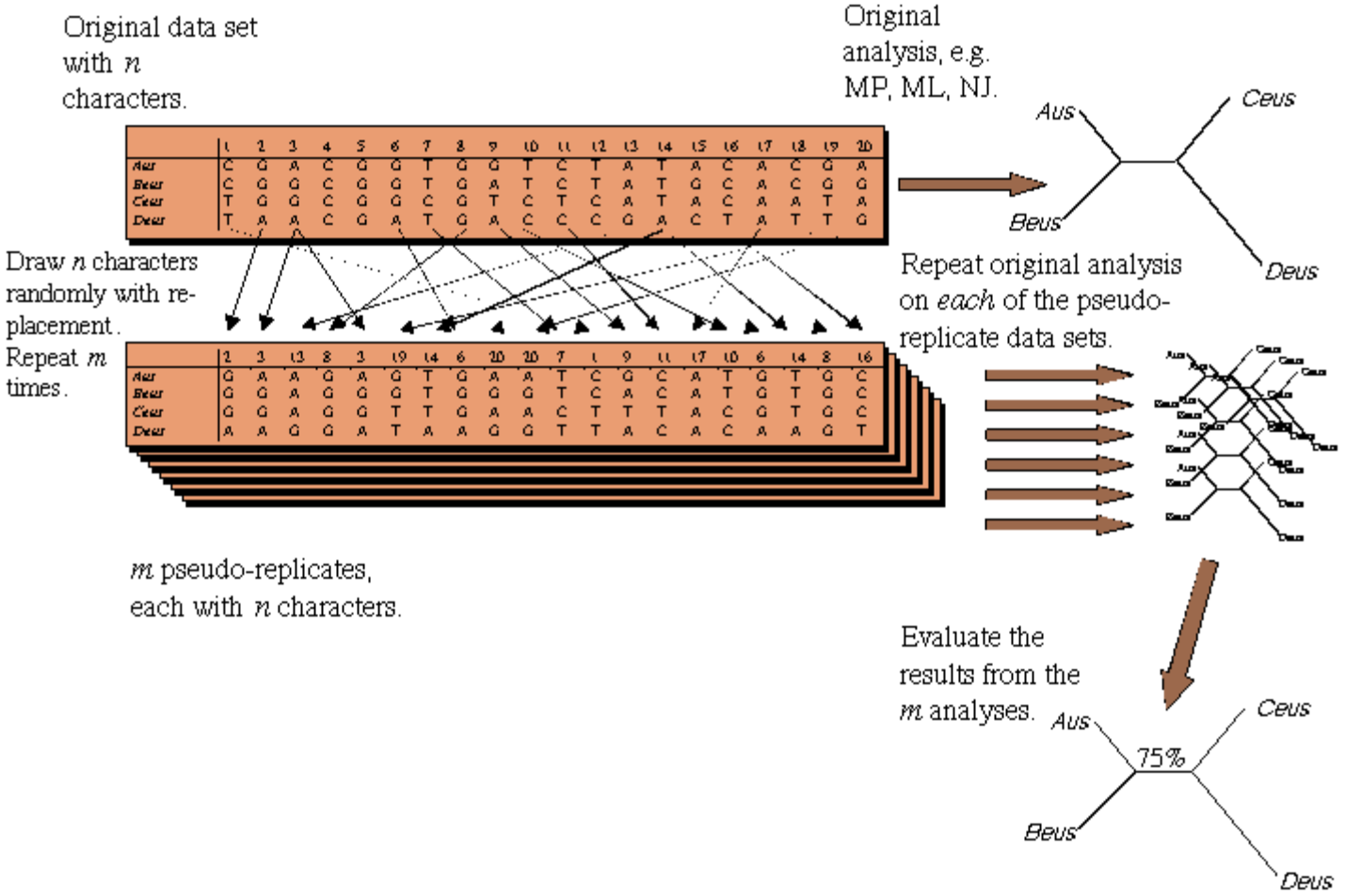
- Distance-matrix methods
  - UPGMA
  - Neighbor-joining method
  - Minimum evolution method
- Maximum parsimony method
- Maximum likelihood method
- Bayesian inference of phylogeny

Bootstrap

Posterior probability

Principle of Bootstrap

[7]

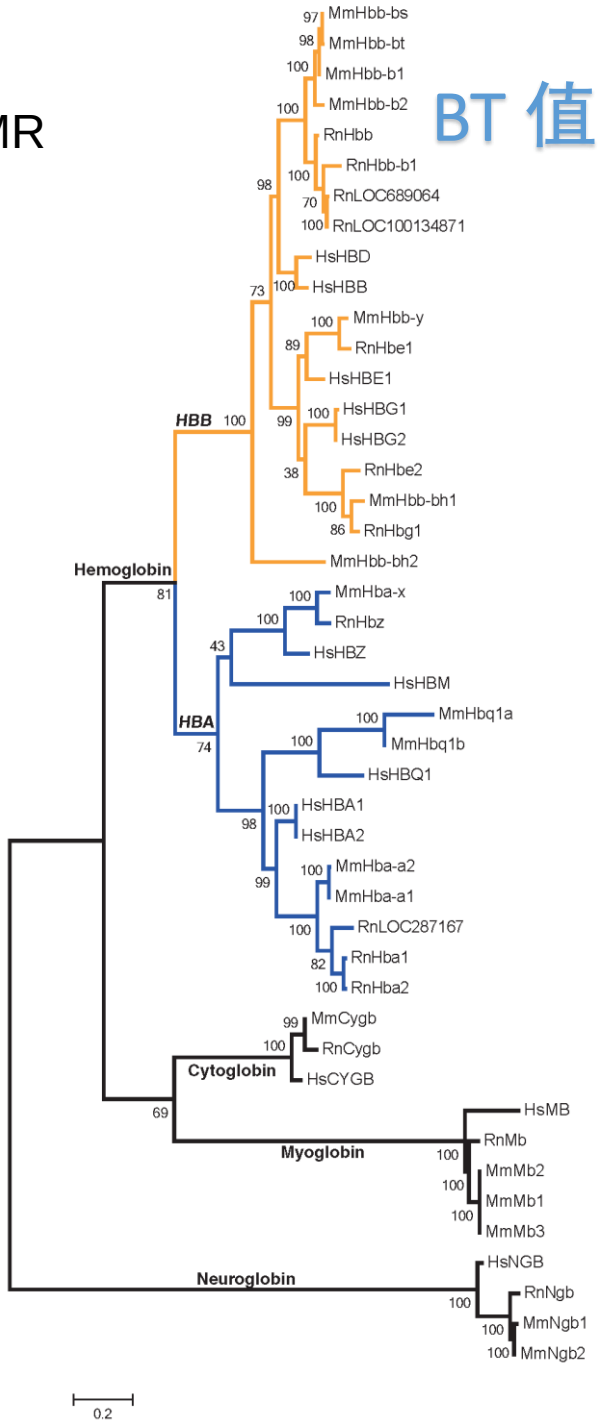


Globin_HMR

BT 值越高越可靠

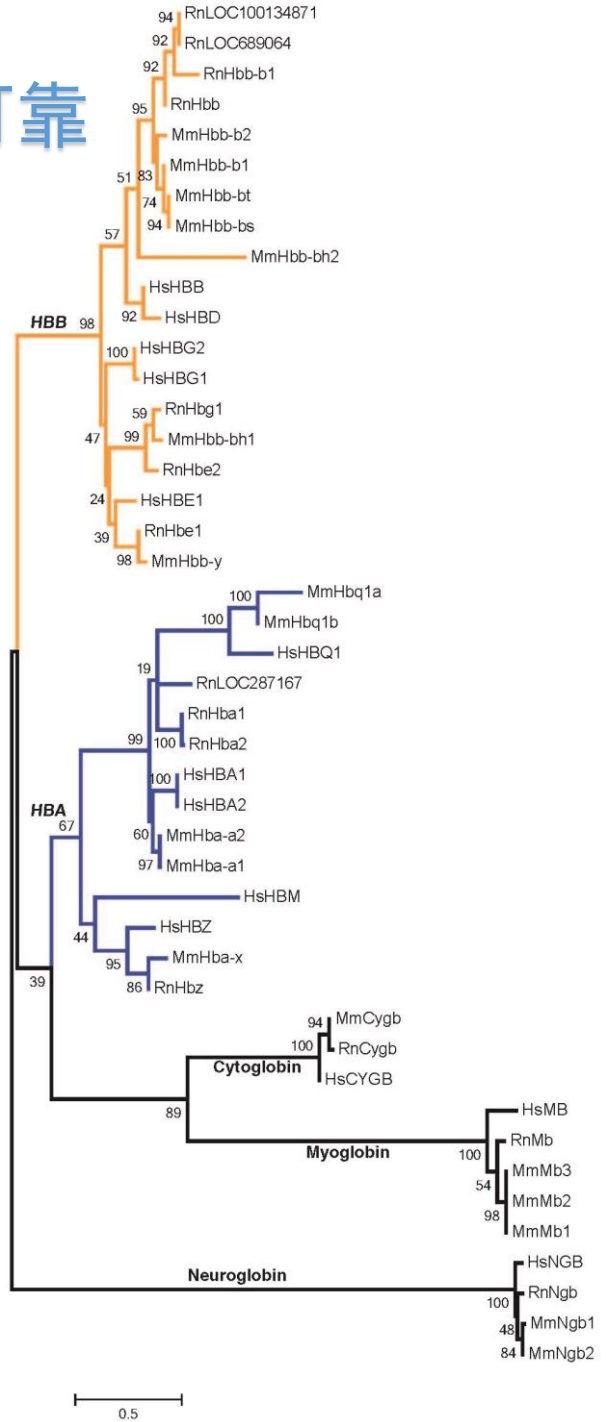
CDS

ML

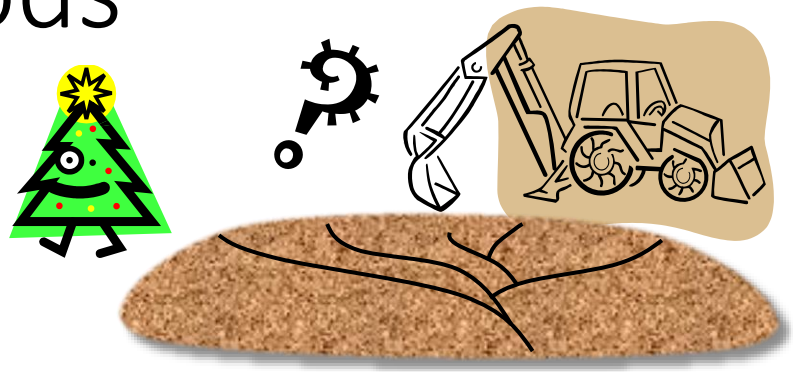


PEP

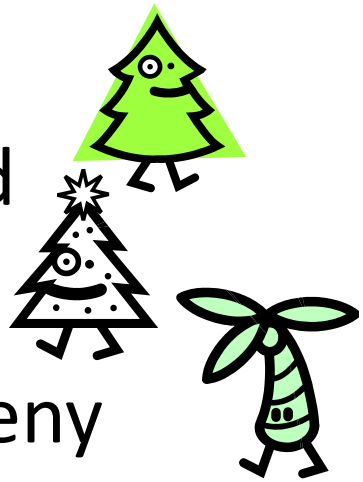
ML

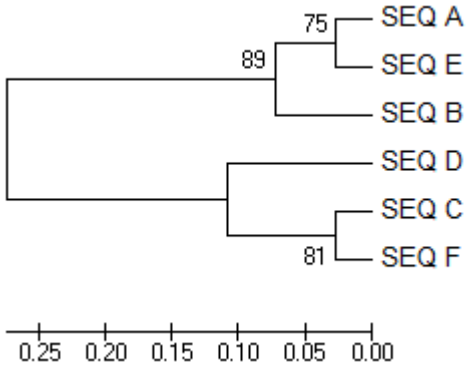


Phylogenetic methods

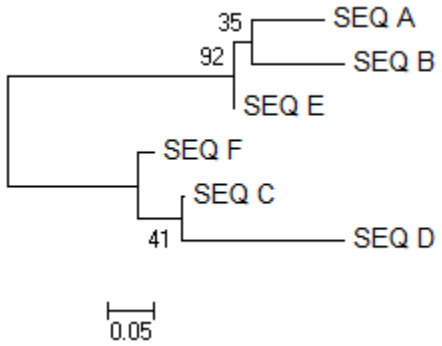


- Distance-matrix methods
  - UPGMA
  - Neighbor-joining method
  - Minimum evolution method
- Maximum parsimony method
- Maximum likelihood method
- Bayesian inference of phylogeny

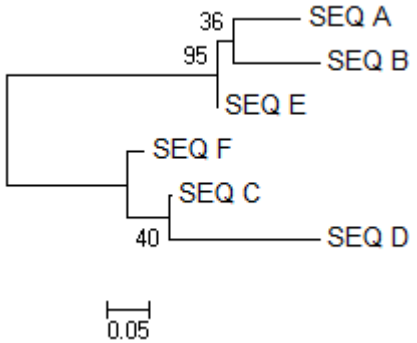




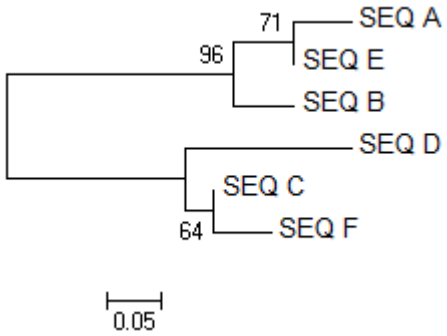
UPGMA (K2-model)



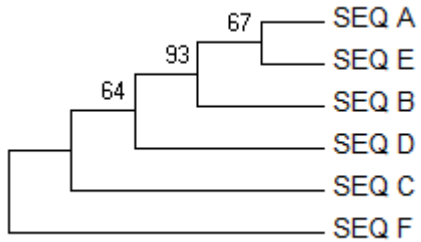
NJ (K2-model)



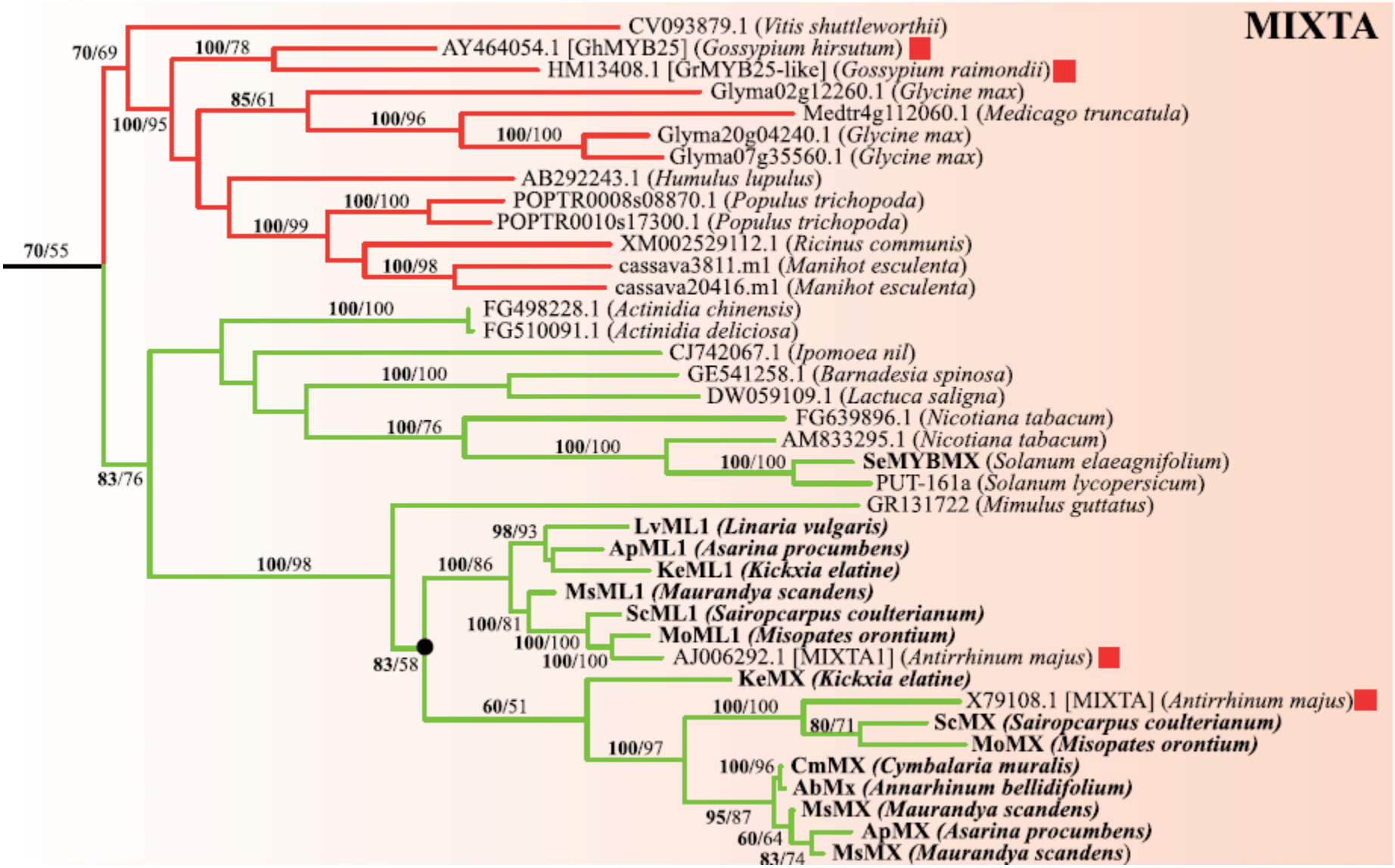
ME (K2-model)



ML (K2-model)

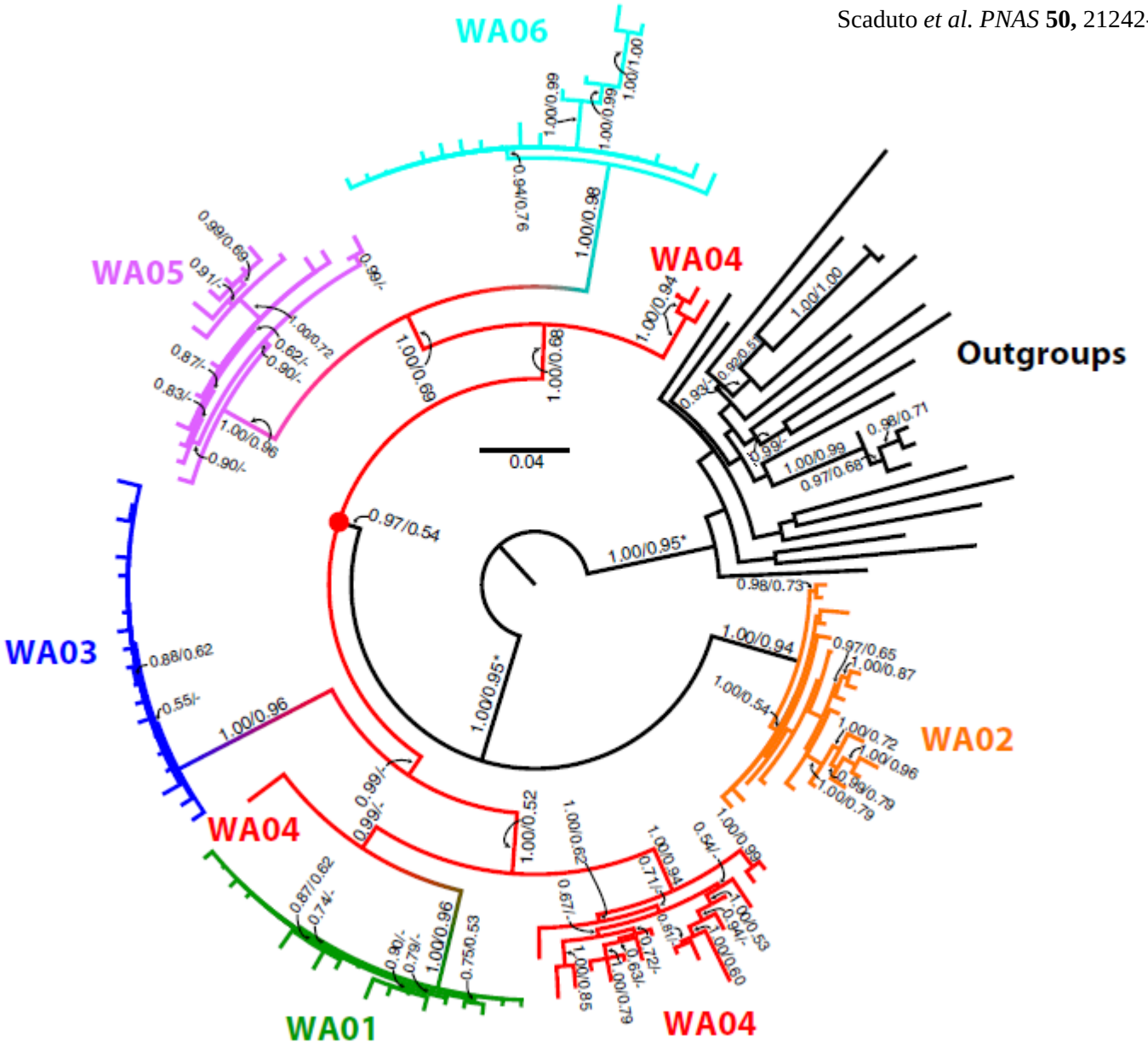


MP [Subtree-Pruning-Regrafting (SPR)]



Numbers next to nodes are Bayesian Posterior probabilities (in bold) and ML BS support values from 100 replicates.

Scaduto *et al. PNAS* **50,** 21242-21247 (2010).



6个HIV携带者的*env*系统发生树。支持率（贝叶斯后验概率/ML自举比例）标注在分支处。

THE END

Reference

- [1] http://en.wikipedia.org/wiki/Charles_Darwin#mediaviewer/File:Charles_Darwin_by_G._Richmond.png
- [2] http://www.nhm.ac.uk/nature-online/evolution/tree-of-life/darwin-tree/
- [3] http://en.wikipedia.org/wiki/Jean-Baptiste_Lamarck#mediaviewer/File:Jean-Baptiste_de_Lamarck.jpg
- [4] http://en.wikipedia.org/wiki/Ernst_Haeckel#mediaviewer/File:Ernst_Haeckel_1860.jpg
- [5] http://lpi.oregonstate.edu/ss14/zuckerkandl.html
- [6] http://www.brianomeara.info/figures
- [7] http://artedi.ebc.uu.se/course/X3-2004/Phylogeny/Phylogeny-Credibility/Phylogeny-Credibility.html