Molecular Phylogeny and Phylogenetic Tree Construction

分子系统发生学及系统发生树构建

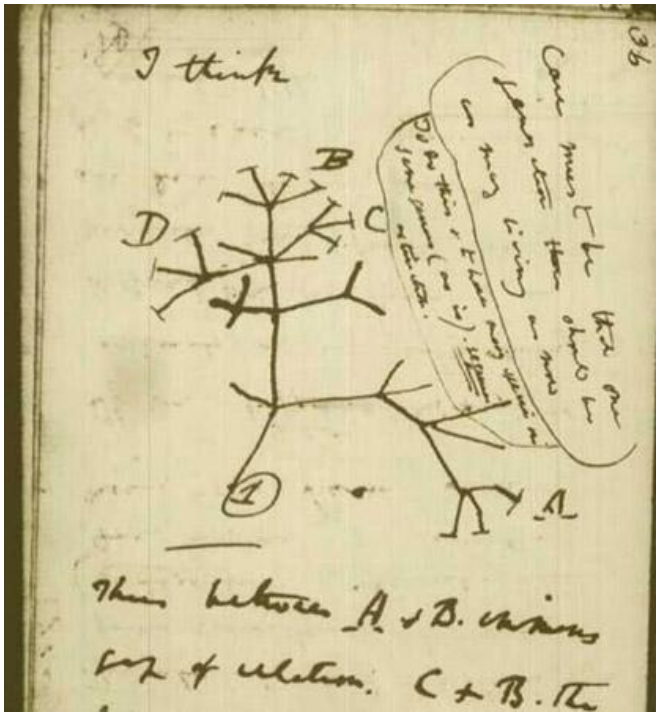
2019/4/27

杨继轩（顾红雅课题组）

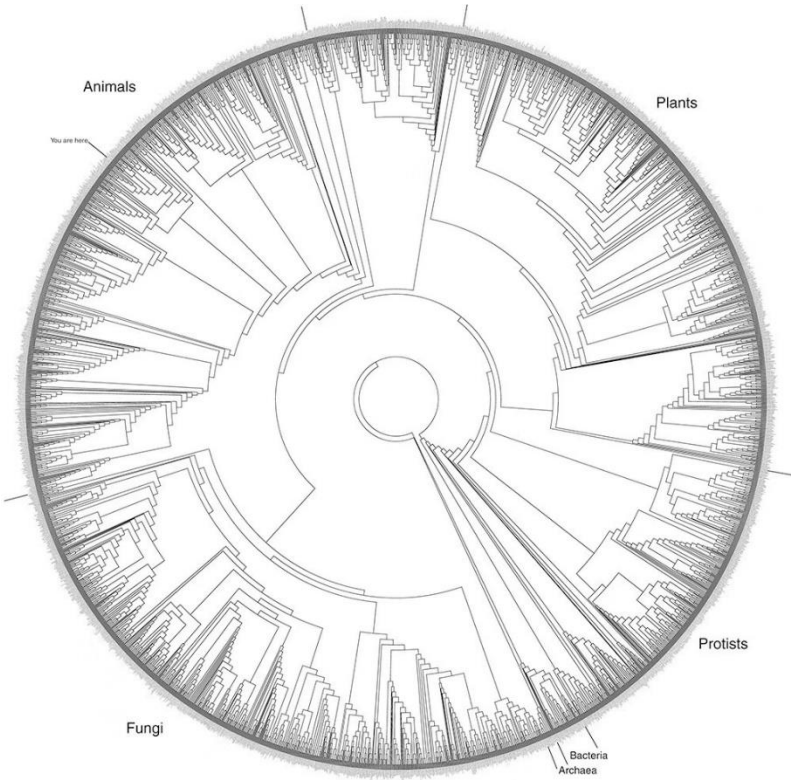
北京大学生命科学学院

Definitions

- Phylogeny: The evolutionary history of a biological group.
- Phylogenetic tree: Graphically represent evolutionary history and record the interrelationships between species.

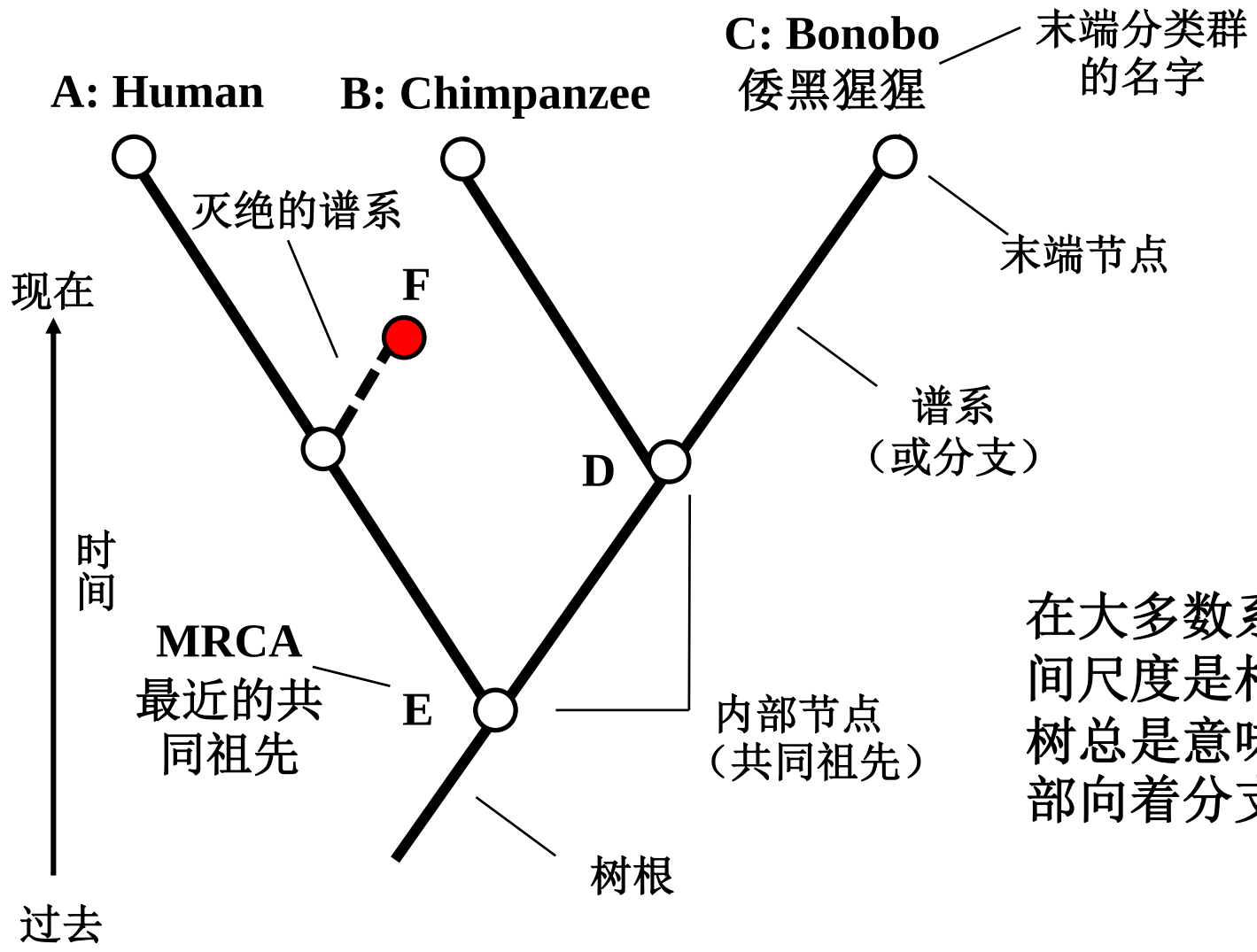


Found in Darwin's notebook



The tree of Life

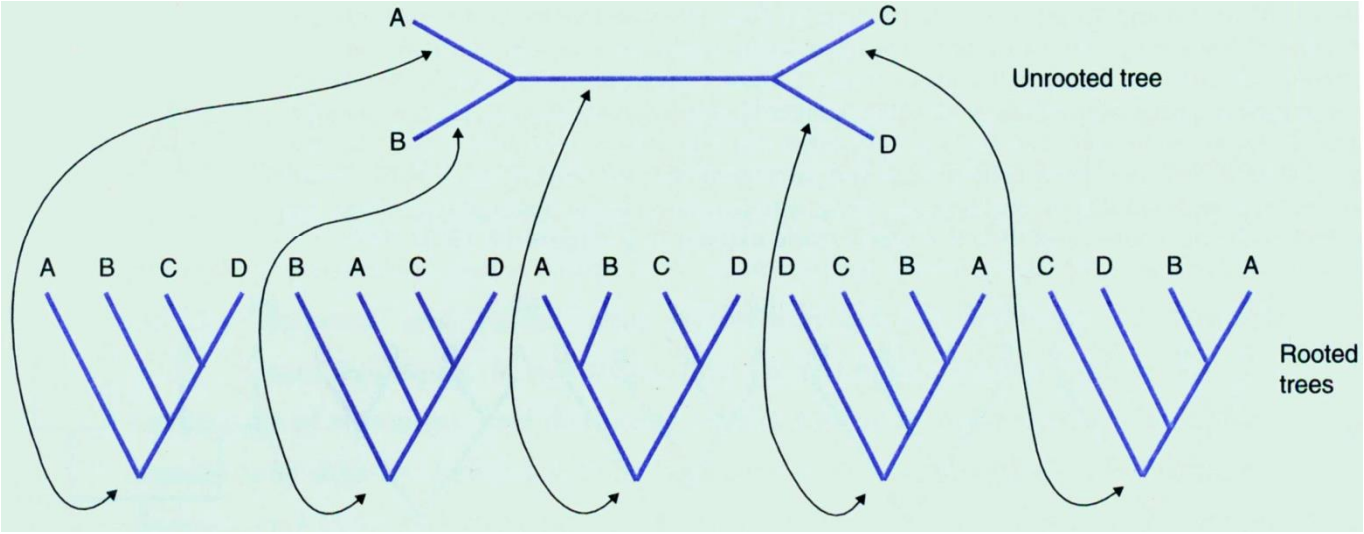
Definitions



在大多数系统发生树中，时间尺度是相对的，但是演化树总是意味着时间由树的根部向着分支末端推移。

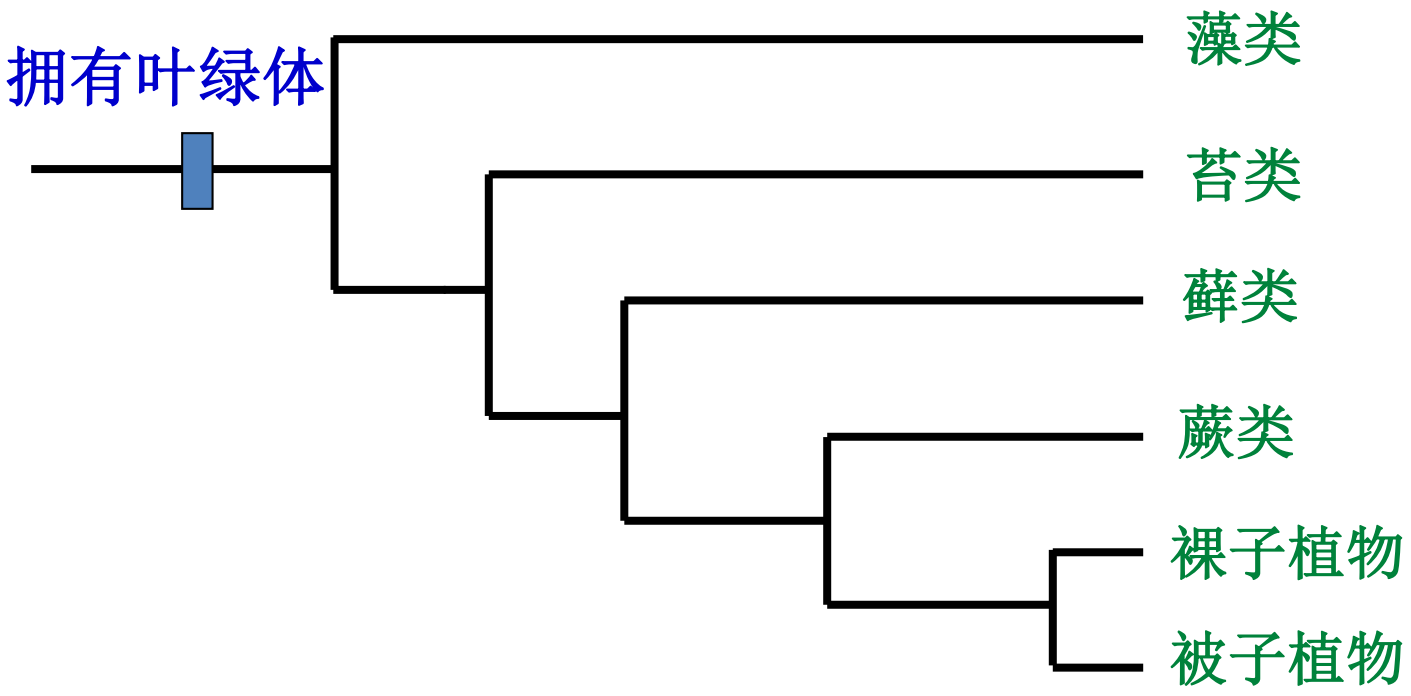
Definitions

- Rooted tree: The oldest species are roots, and the trees represent the order of species formation.
- Unrooted tree: There are no ancient species in the tree as a reference, only the interrelationship between species.



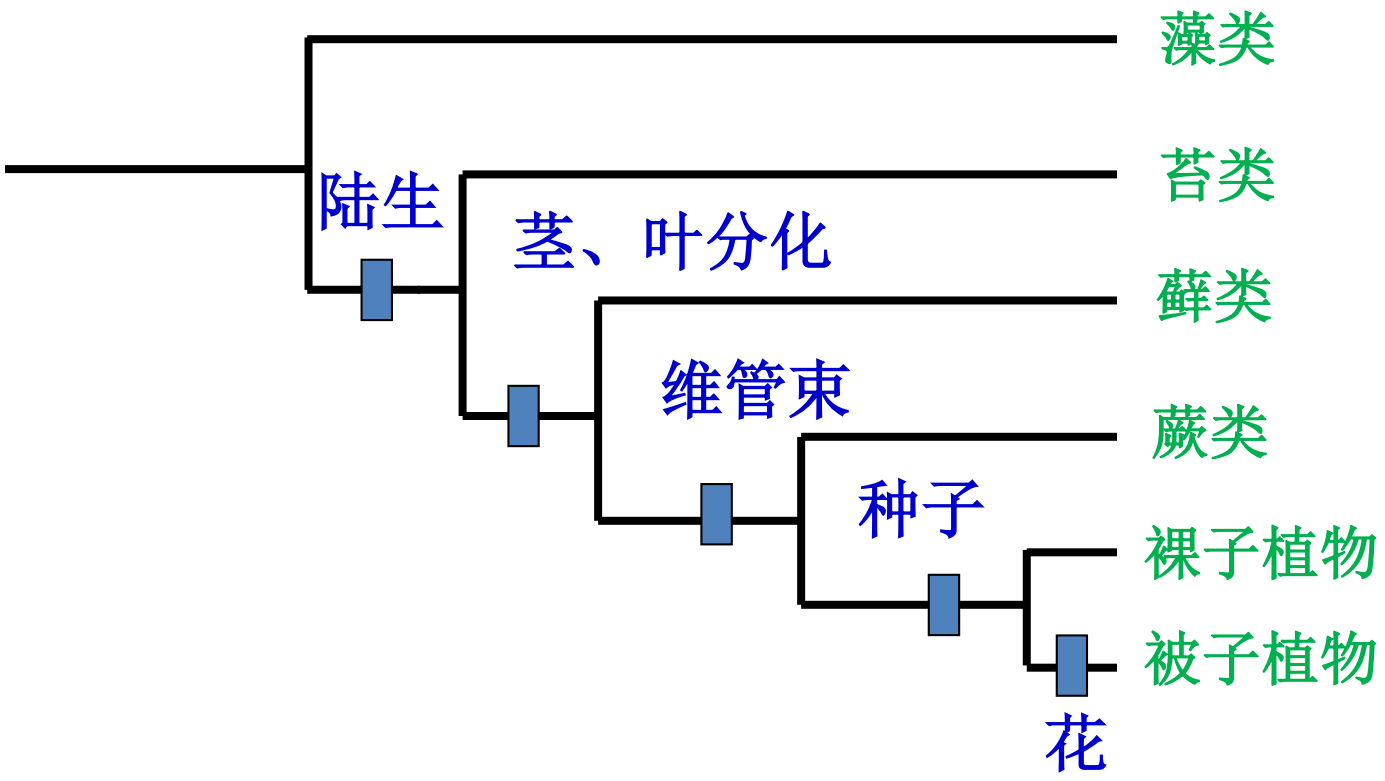
Plesiomorphic character

- Features similar to ancestral feature states - "ancient" features



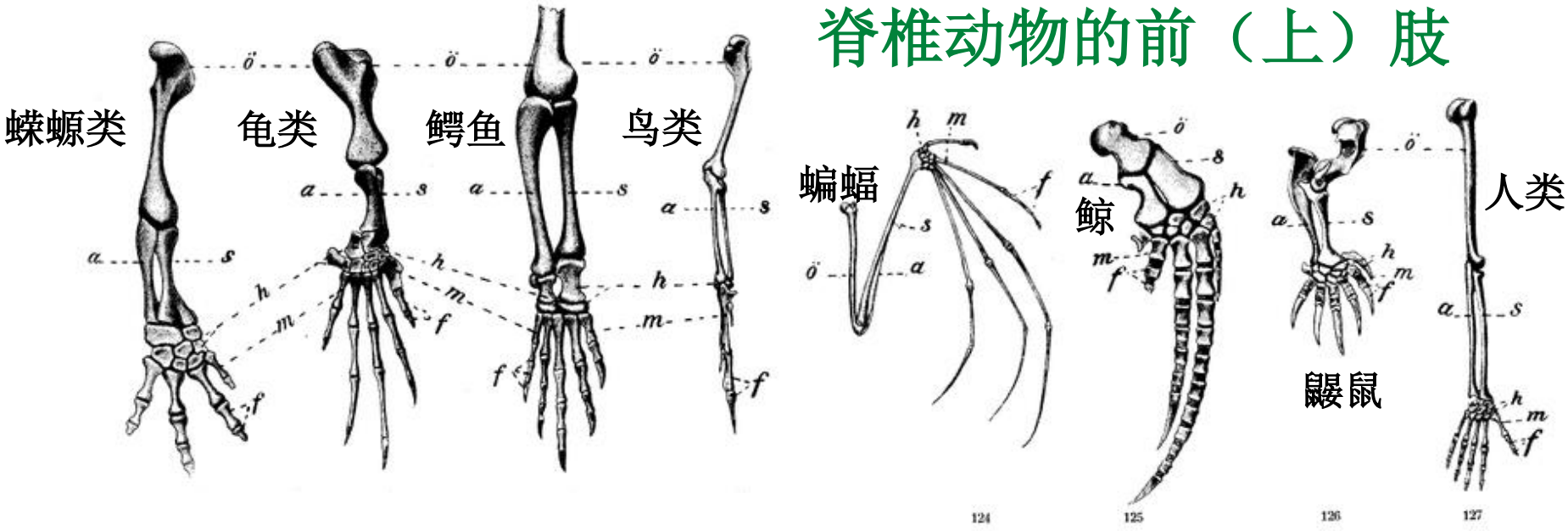
Apomorphic character

- A feature evolved from ancestral features - “later" features



Divergent evolution

- They originate from the same type of organisms, living in different environments and eventually forming new forms (groups) that adapt to different environments.



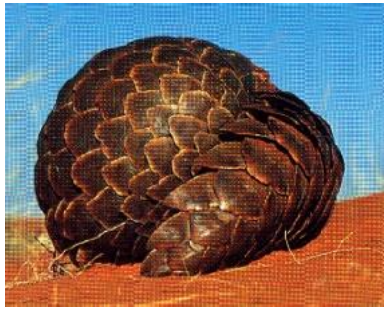
Convergent evolution

- These similar features come from distantly related species and are not from the most recent common ancestor.

不同纲动物的“翅膀”

Convergent evolution

穿山甲/鱗甲目



来自不同目的动物，有相似的外表

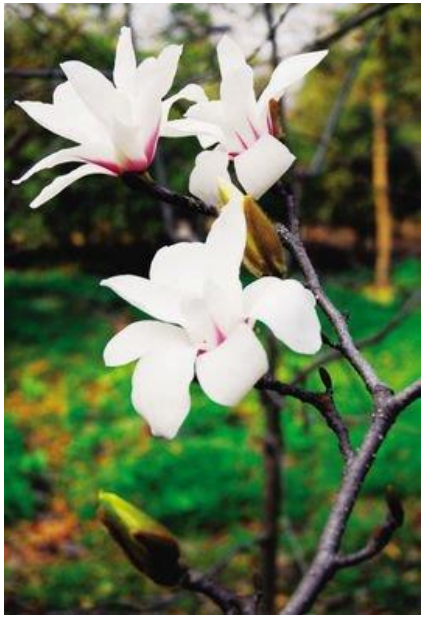
犰狳/有甲目

金毛鼹鼠(非洲猬目)

日本鼹鼠 (鼩形目)

Principle of parsimony

- 生物演化“沿着”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：四个物种：阿米巴虫(Amoeba)、玉兰树(Magnolia)、黑猩猩(Chimp)、人(Human)之间的相互关系如何？

Principle of parsimony

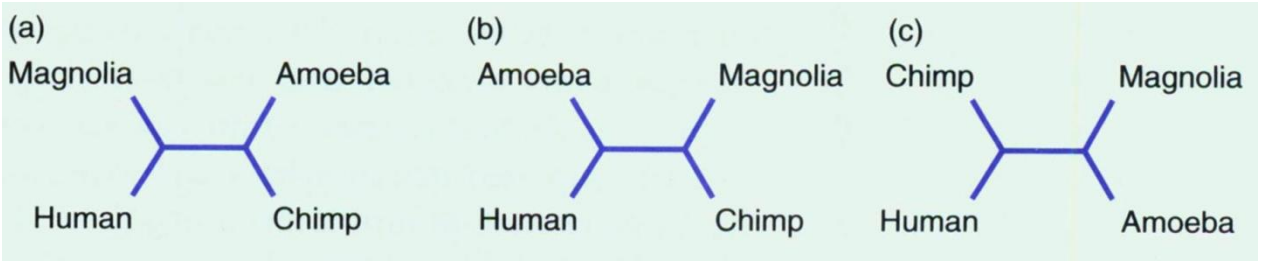
假设：

他们共有的性状：**1000**

他们特有的性状：**10**

人和黑猩猩共有性状：**100**

有三种可能的亲缘关系　哪种最客观？



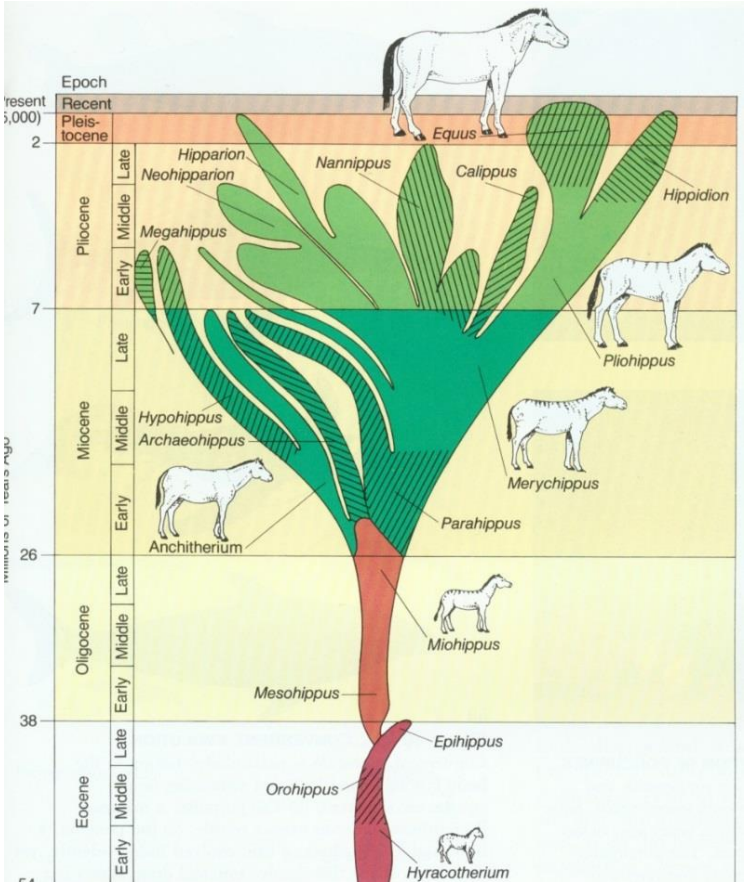
每个性状就是一次演化事件

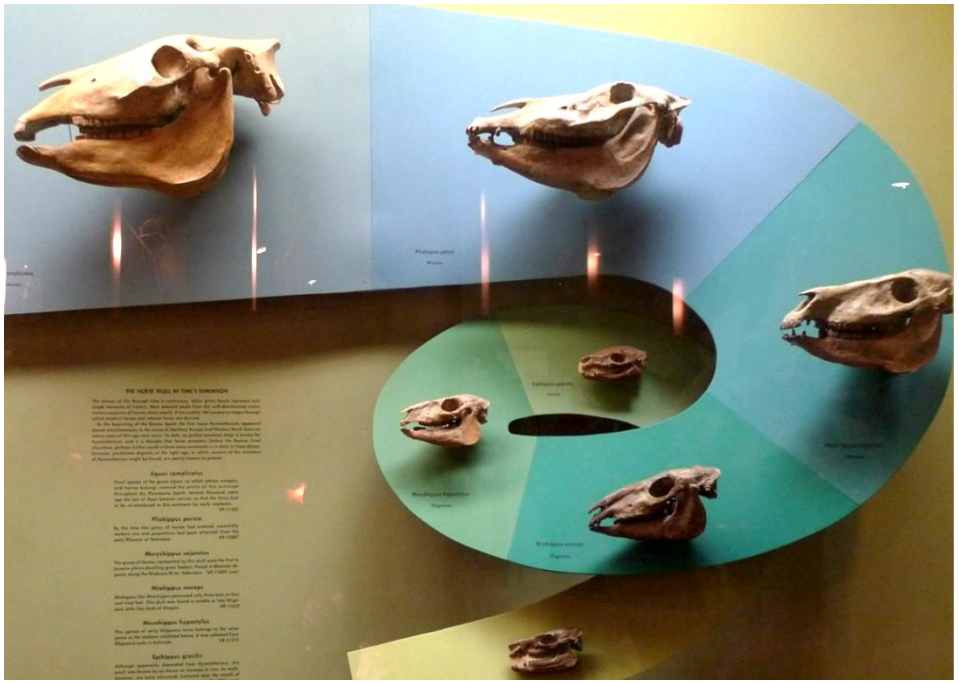
4个物种共同经历的演化事件：	1000	1000	1000
每个物种单独经历的演化事件之和：	40	40	40
人和黑猩猩经历的演化事件之和：	200	200	100
形成上述关系总共要经历的演化事件：	1240	1240	1140

按简约性原则，C 所经历的演化事件最少，C 符合要求

What method (feature) is chosen to build a phylogenetic tree?

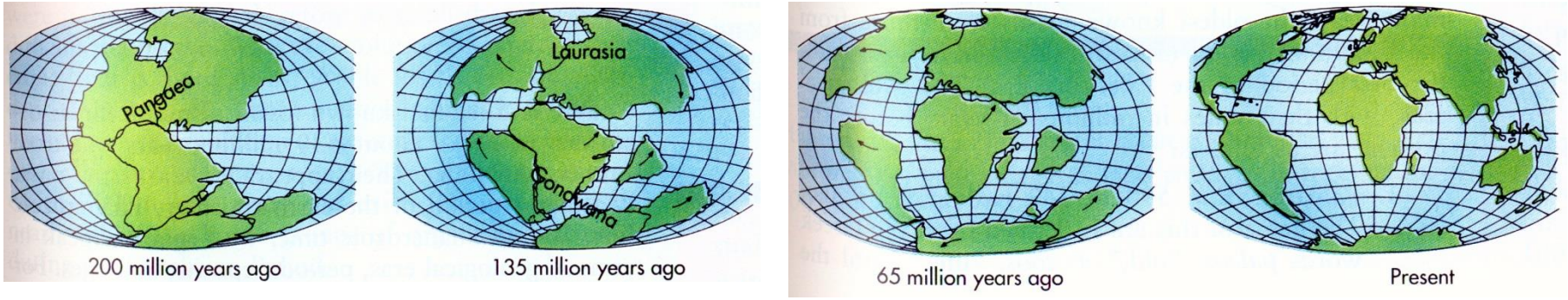
- 化石材料：客观、直观、真实地反映生物的过去；可遇不可求；如何确定质地年代





What method (feature) is chosen to build a phylogenetic tree?

- 形态：直观、易测量；相似似性状的性质（同源？趋同？）
- 行为
- 生态
- 地理位置（地理距离相近的物种亲缘关系也近？）



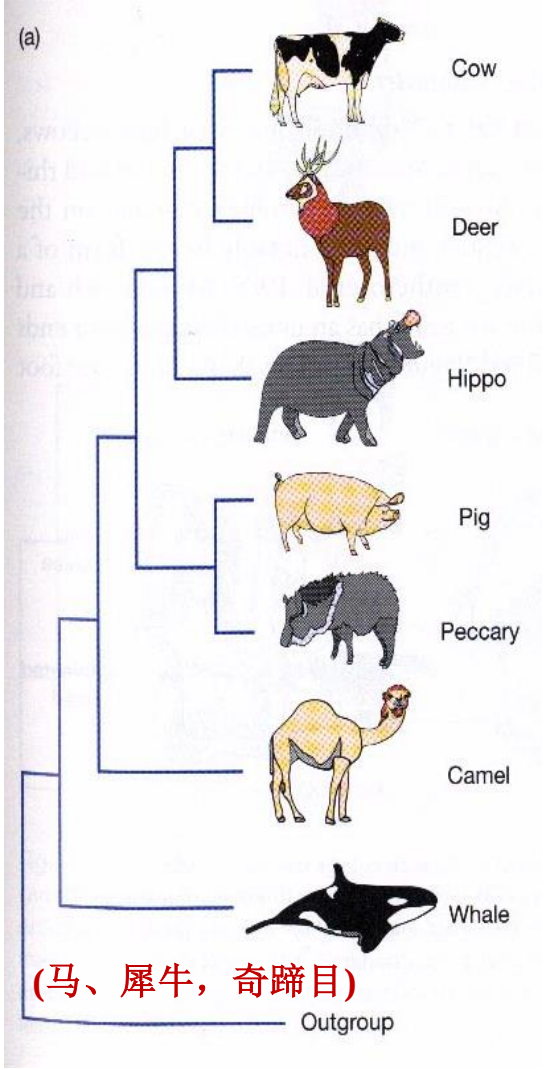
侏罗纪　　白垩纪　　古近纪（古新世）　　现代

How do we treat trees with different results?

- 在数据来源不同的情况下，即有化石数据、形态数据、分子数据，再加上有趋同演化和回复突变，将会有多个“树”，如何处理？

- 举例说明 — 鲸的亲缘关系研究

How do we treat trees with different results?



(马、犀牛，奇蹄目)

鲸“鱼”

- 一种特化的哺乳动物
- 最早的化石发现于5千5百万年前，有后肢

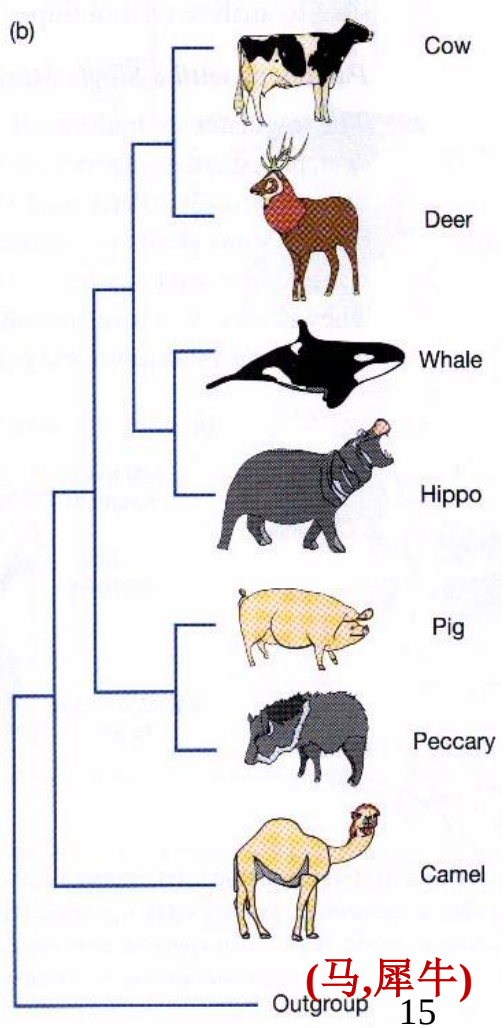
基于形态特征的演化树

Gingerich *et al.*, Science, 1990

基于线粒体DNA演化树

Graur & Higgins, Mol Biol Evol, 1994

Irwin & Arnason, J. Mammal. Evol, 1994



(马,犀牛)

15

How do we treat trees with different results?

引起了很大的争议，哪个树更接近真实情况？

1. *Hervé & Douzery, J. Mammal. Evol., 1994*

认为至少要30个物种的线粒体全基因组序列。

2. *Hasegawa & Adachi, Mol Biol Evol., 1996*

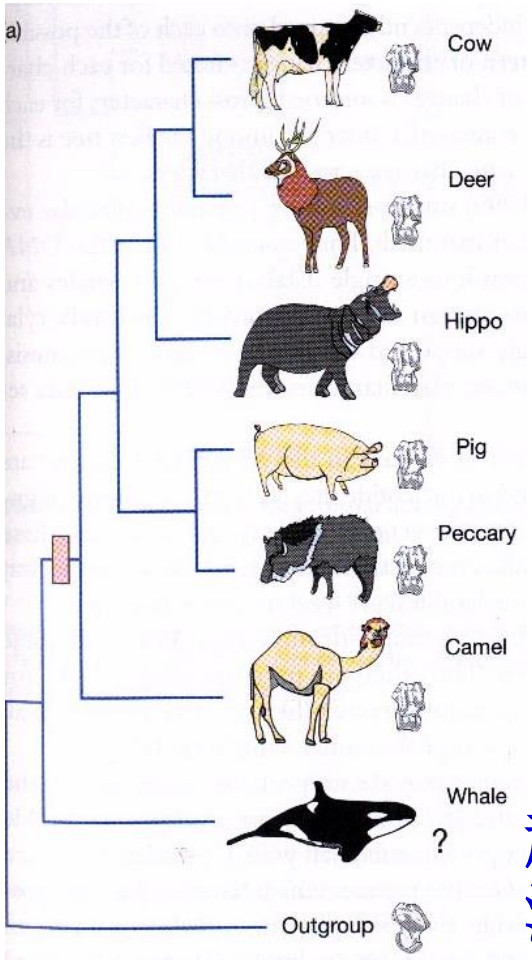
不同基因、不同外类群进行研究，结果多变。

3. *Gatesy et al., Mol Biol Evol., 1996*

Gatesy, Mol Biol Evol., 1997

两个乳蛋白编码基因和编码纤维蛋白原的基因的研究，更多的物种，分别支持鲸与河马最近。

How do we treat trees with different results?



来自转座子的证据支持鲸与河马的亲缘关系最近

Shimamura *et al.*, Nature, 1997

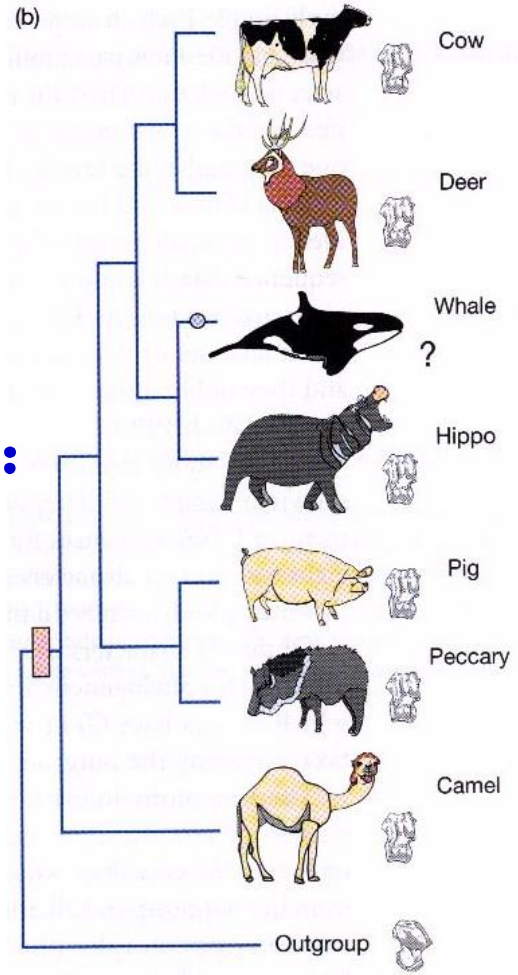
Nikaido *et al.*, Proc Natl Acad Sci, 1999

基于踝关节的形态来构树：

奇蹄目   偶蹄目

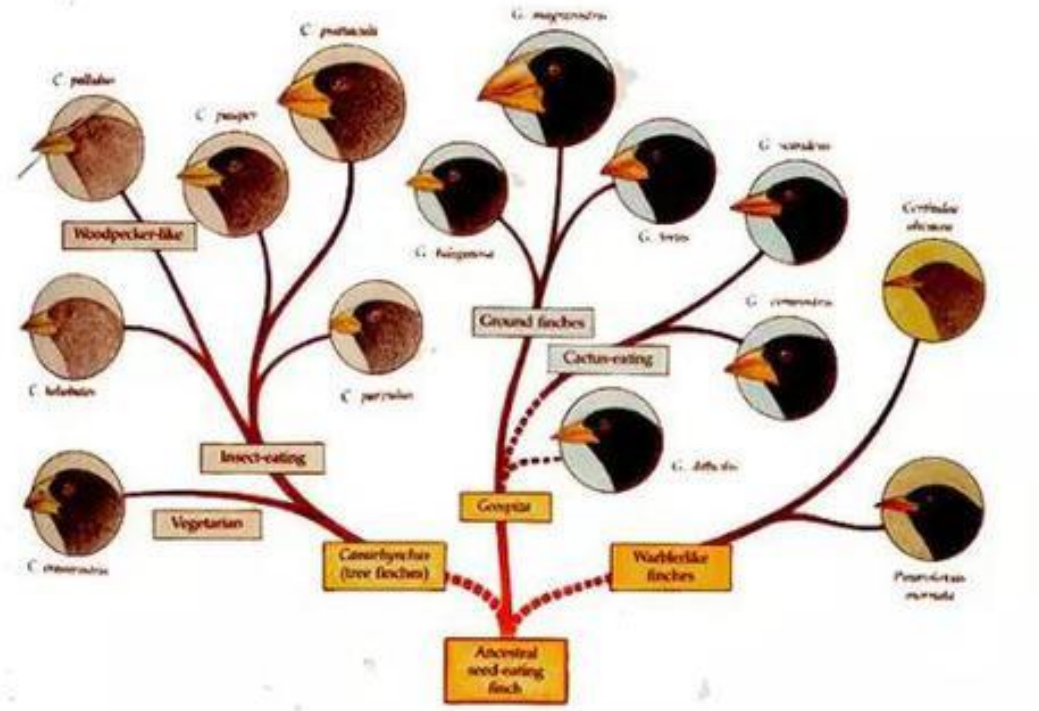
Thewissen & Madar, Syst Biol., 1999

越来越多的证据支持鲸与河马亲缘关系最近



How to construct phylogenetic tree?

- **邻接法（Neighbor-joining）**
- **最大简约法（Maximum parsimony）**
- **最大似然法（Maximum Likelihood）**



Phylogenetic tree of birds on the Galapagos

How to construct phylogenetic tree?

- 邻接法（**Neighbor-joining**）
- 首先把整体上（例如在**DNA**序列上）最相似的物种相连接，然后将这样的聚簇与下一个与它们最相似的物种相连接以形成更大的簇，重复这一过程直到所有物种都连接到了一个簇。
- 适用范围：远缘序列，进化距离不大，信息位点少的短序列。
- 优点：假设少，树的构建相对准确，计算速度快，只得到一棵树，可以分析较多序列，运行速度优于最大简约法。
- 缺点：序列上所有位点等同对待，且所分析的序列进化距离不能太大。

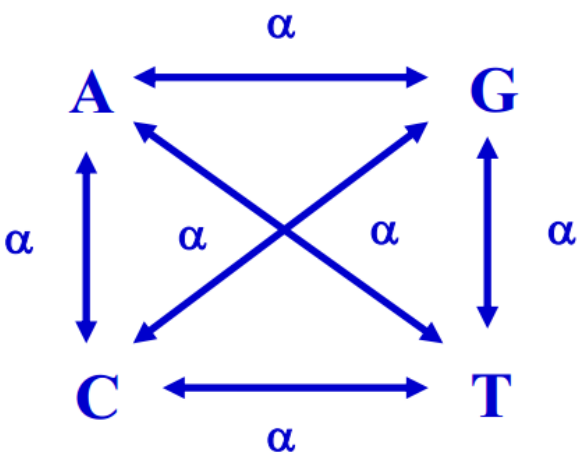
How to construct phylogenetic tree?

- **最大简约法（Maximum parsimony）**
- **根据序列的多重比对结果，对所有可能正确的拓扑结构进行计算并挑选出所需替代数最少的拓扑结构作为最优树，即能够利用最少的步骤去解释多重比对中的碱基差异。**
- **适用范围：近缘序列，物种序列数目≤ 12。**
- **优点：善于分析某些特殊的分子数据如插入、缺失等序列。**
- **缺点：只适用于数据中的序列数目≤12，存在较多回复突变或平行突变时，结果较差，对变异大的序列进行分析时易产生错误。**

How to construct phylogenetic tree?

- 最大似然法（**Maximum Likelihood**）
- 基于数据本身推断或估计**DNA**序列的进化模型，然后考察所有可能的树中的哪一棵树能使所观察到的特征状态进化的可能性（即似然性）达到最大。
- 最大似然法基于的**DNA**序列进化模型：

（1）



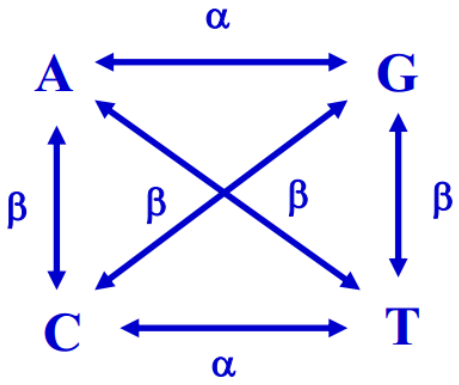
Jukes and Cantor's （JC） 模型或单参数模型（one parameter model）

假设每种碱基被另一种碱基替代的速率为 α
每种碱基被其他碱基的替代速率为 3α

How to construct phylogenetic tree?

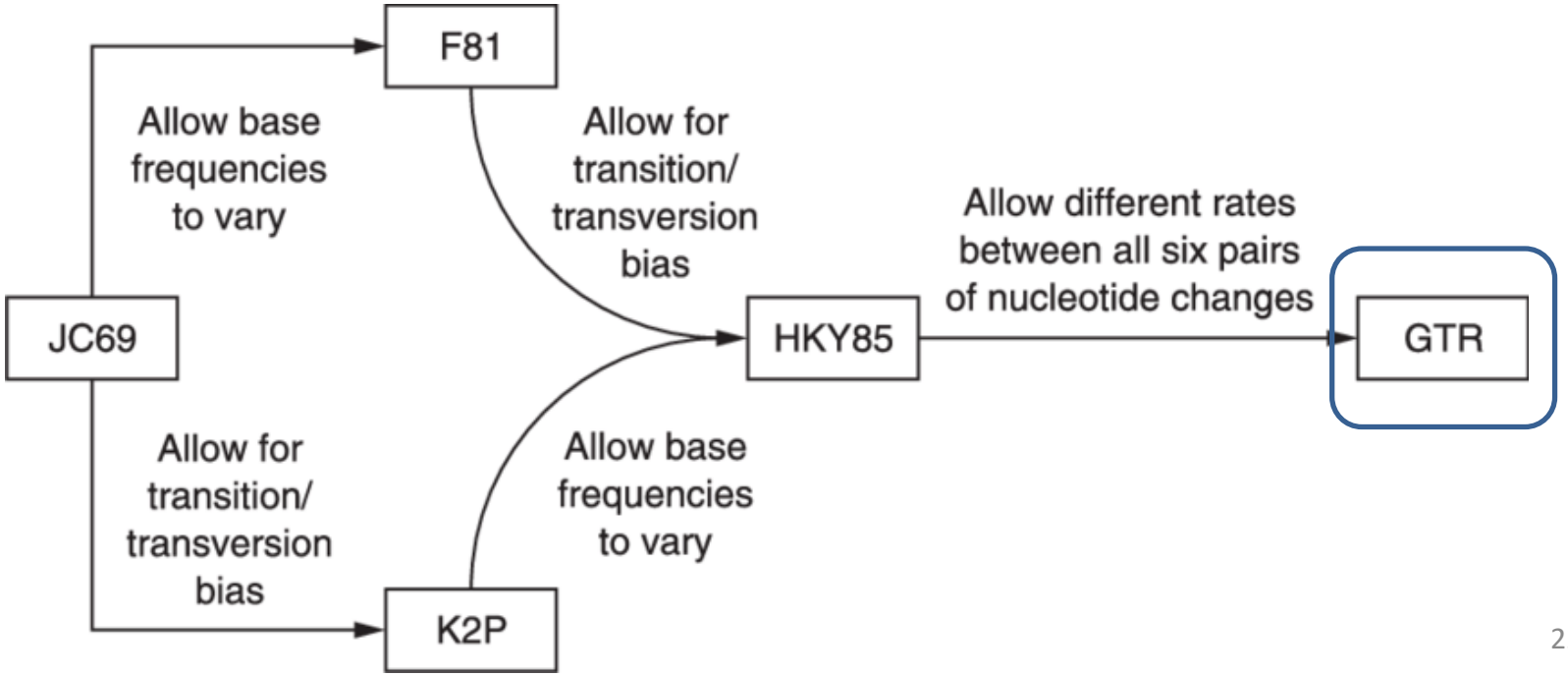
- 最大似然法（**Maximum Likelihood**）

（2）



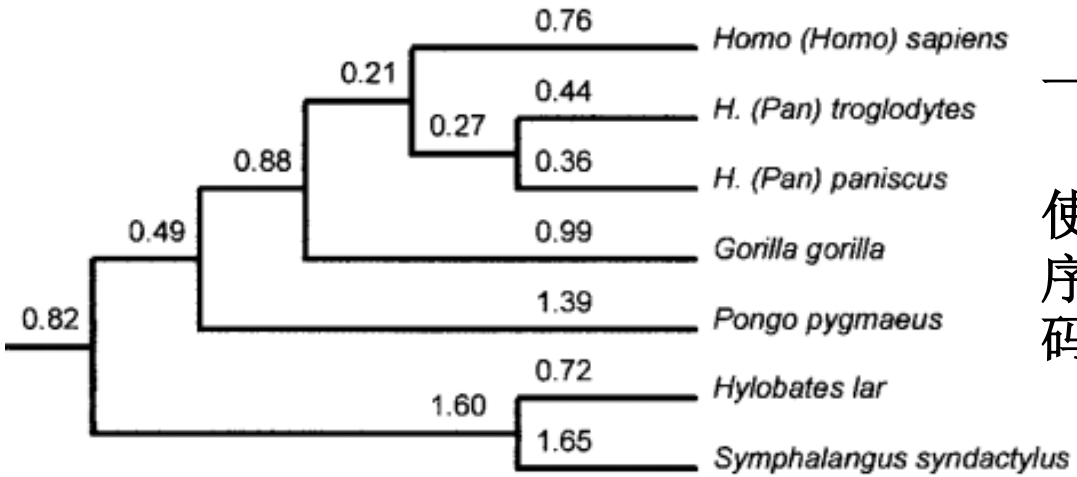
Kimura Two-Parameter 模型或 K-2 模型

碱基置换和颠换的速率不一样
α为一种碱基被同类另一种碱基替代（置换） 的速率
β为一种碱基被另一类碱基替代（颠换） 的速率



How to construct phylogenetic tree?

- 最大似然法（**Maximum Likelihood**）
- 适用范围：特定的替代模型，远缘序列。
- 优点：具有较好的统计学基础，大样本时似然法可以获得参数统计的最小方差，在进化模型确定的情况下，最大似然法是与演化事实吻合最好的建树方法。
- 缺点：计算量大，耗时较长，依赖于合适的替代模型。



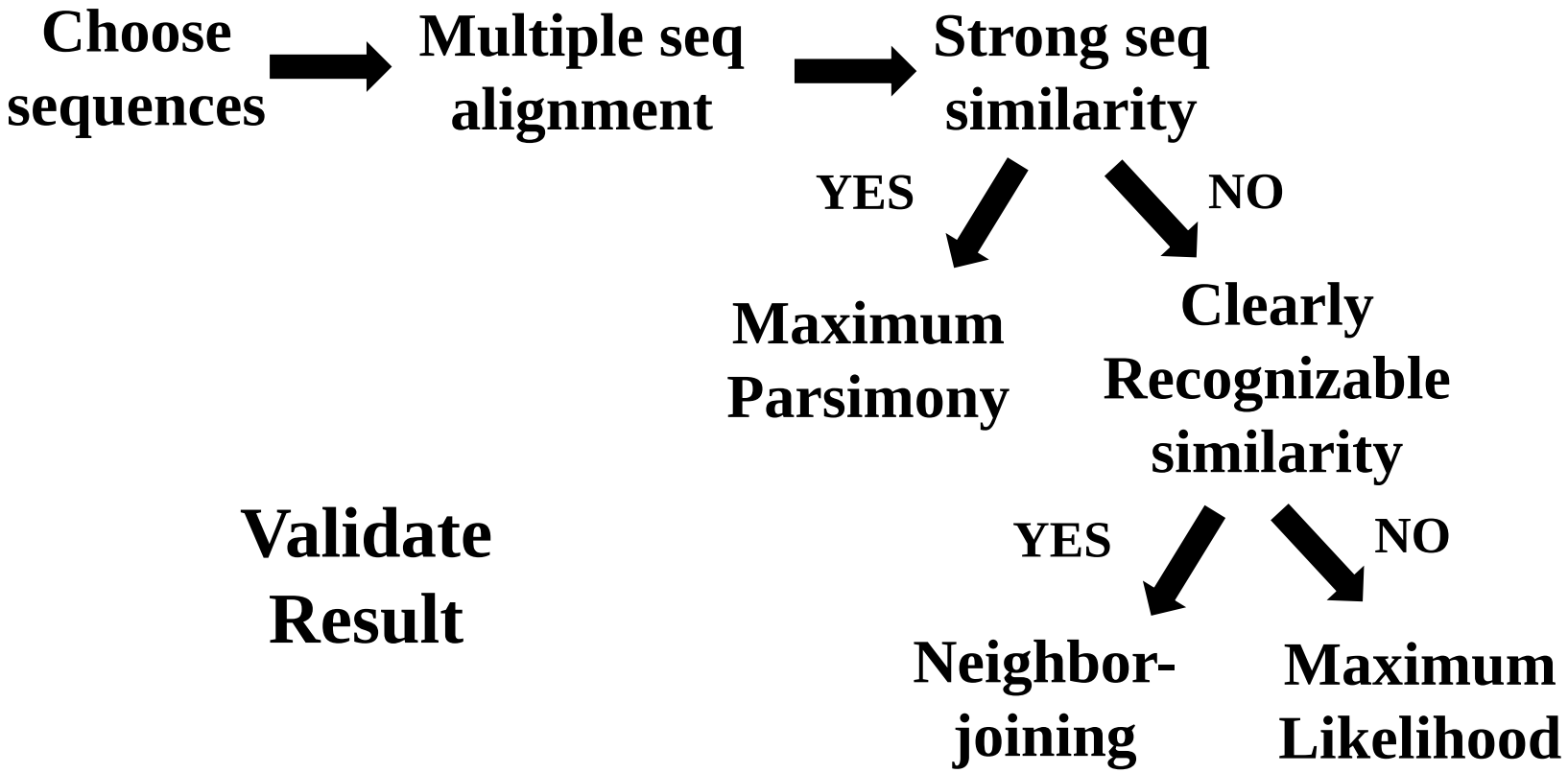
一些灵长类的最大似然系统发生树

使用一种类似于木村双参数模型的序列进化模型，通过分析两个非编码区细胞核**DNA**序列所获得

Page & Goodman, Mol Phylogenet Evol., 2001

How to construct phylogenetic tree?

如何选择一种最佳的构建系统发生树的方法?



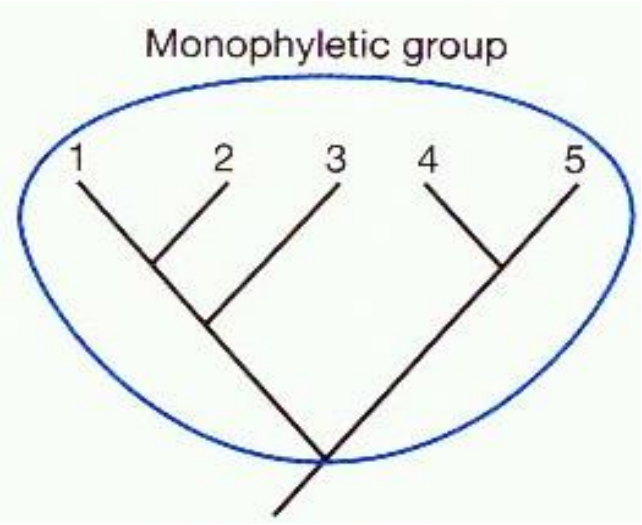
What does the rooted tree tell us?

- 所研究的生物类群是否来源于同一祖先或是同一祖先的所有的后代

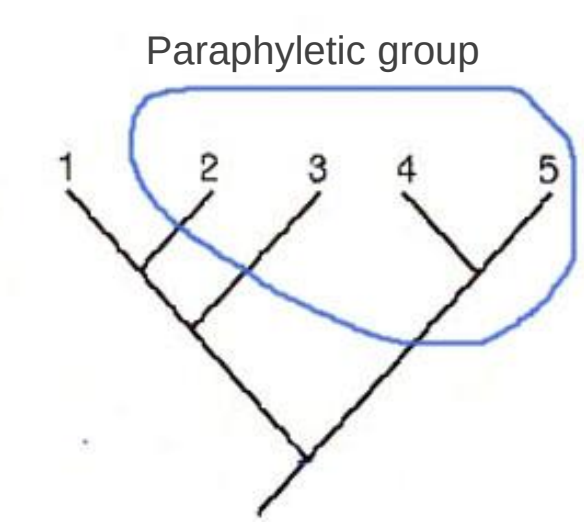
单系类群 (monophyletic)：
来自于同一最近祖先的全部后代

并系类群 (paraphyletic)：
来自于同一最近祖先的后代，但不是全部后代（一般来说含有一个完整的支）

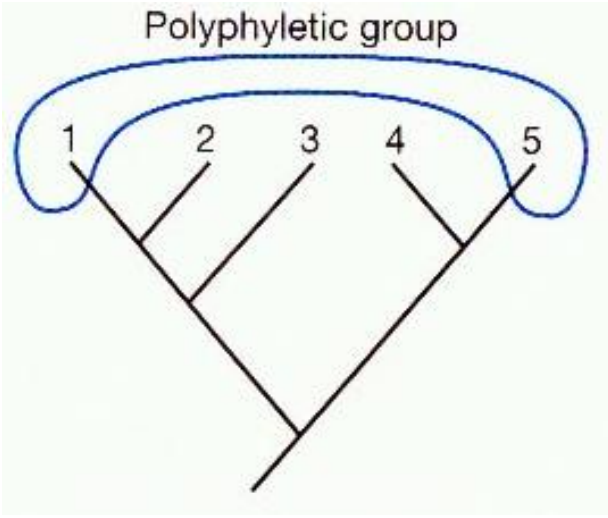
多系类群 (polyphyletic)：
来自于不同最近祖先的后代



单系类群



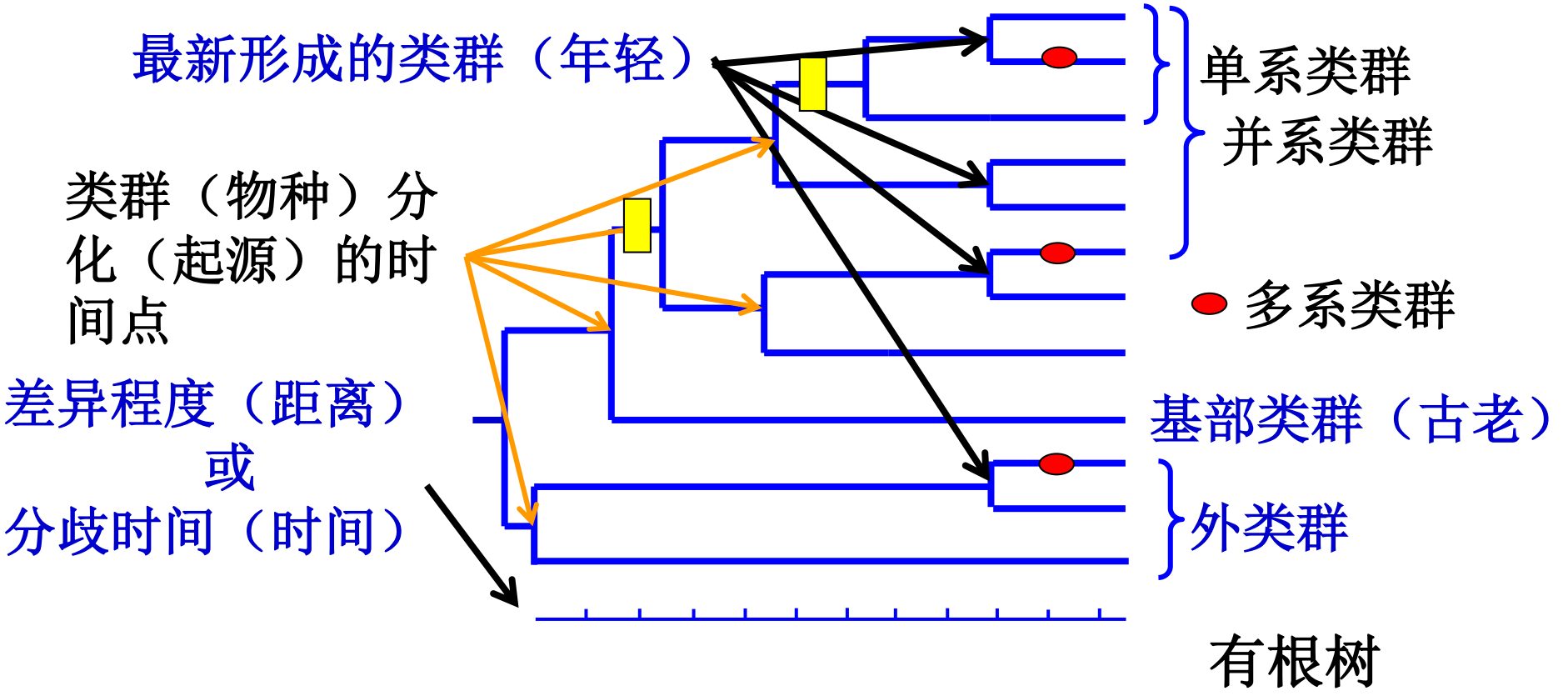
并系类群



多系类群

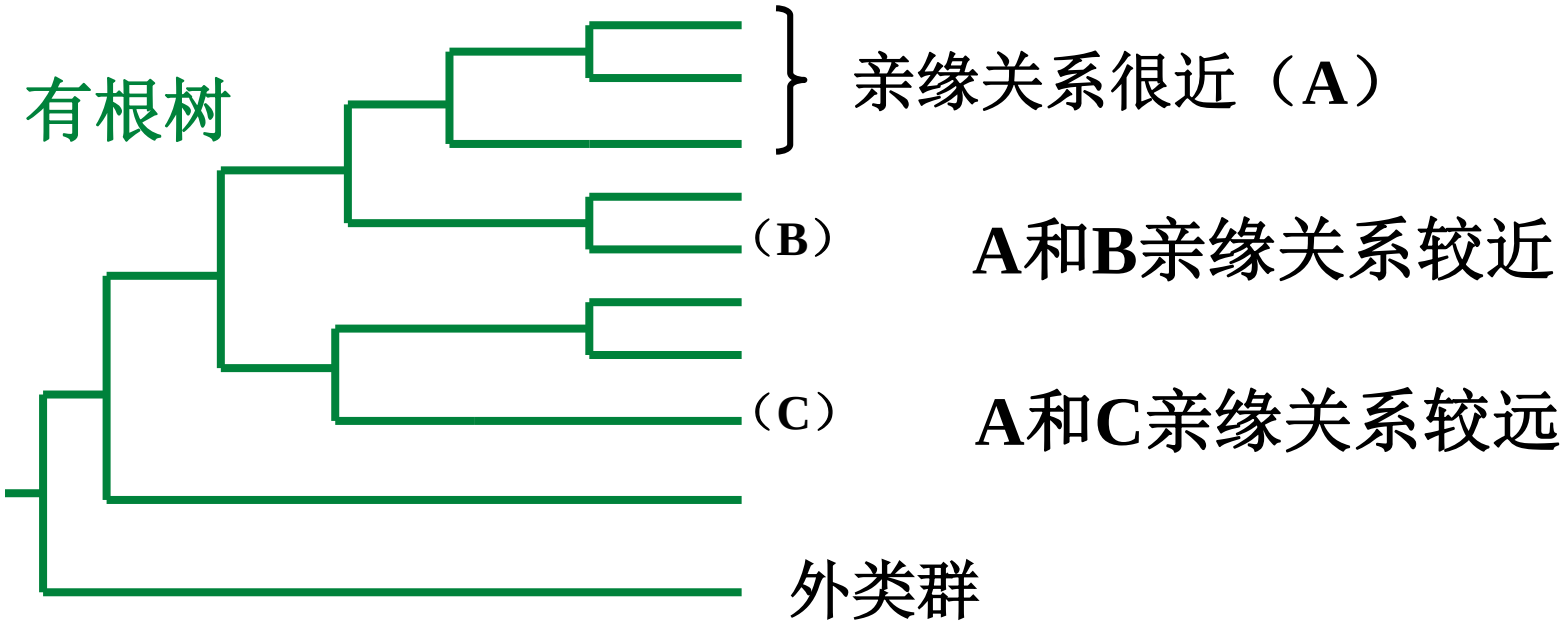
What does the rooted tree tell us?

- 物种形成的顺序（起源的时间）

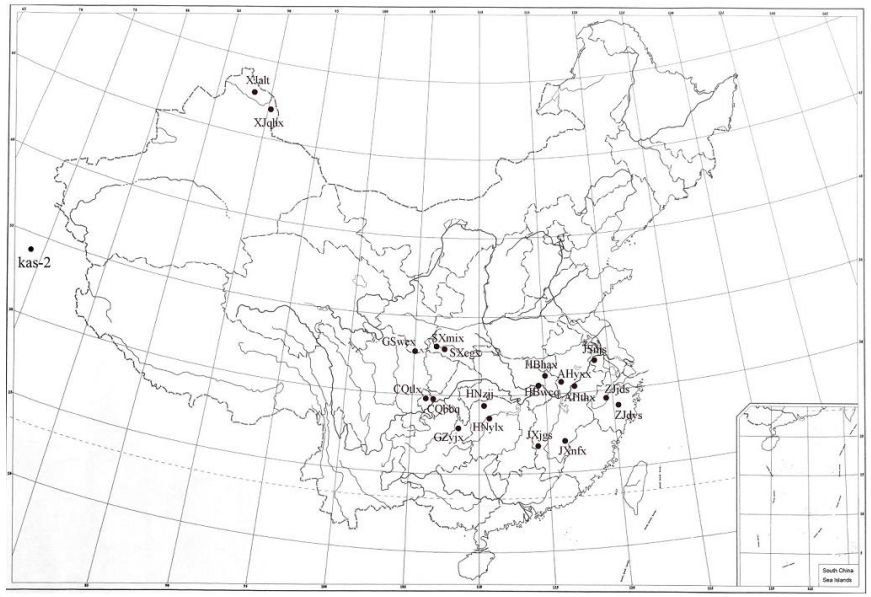


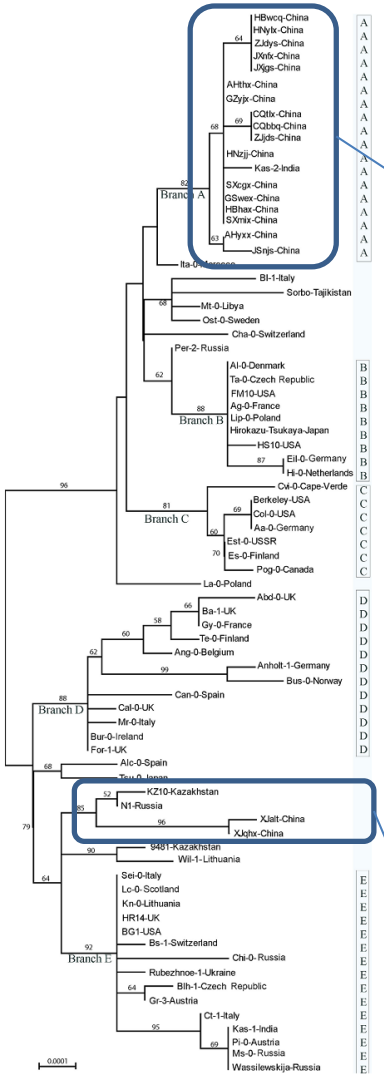
What does the rooted tree tell us?

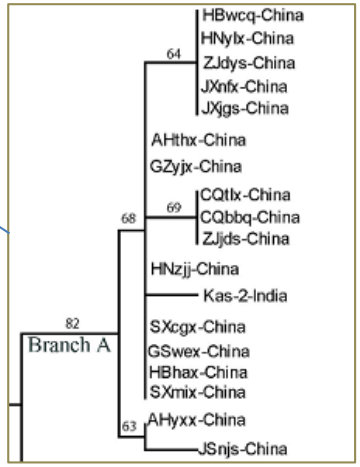
- 物种之间的相互关系



Evolution history of *Arabidopsis thaliana*



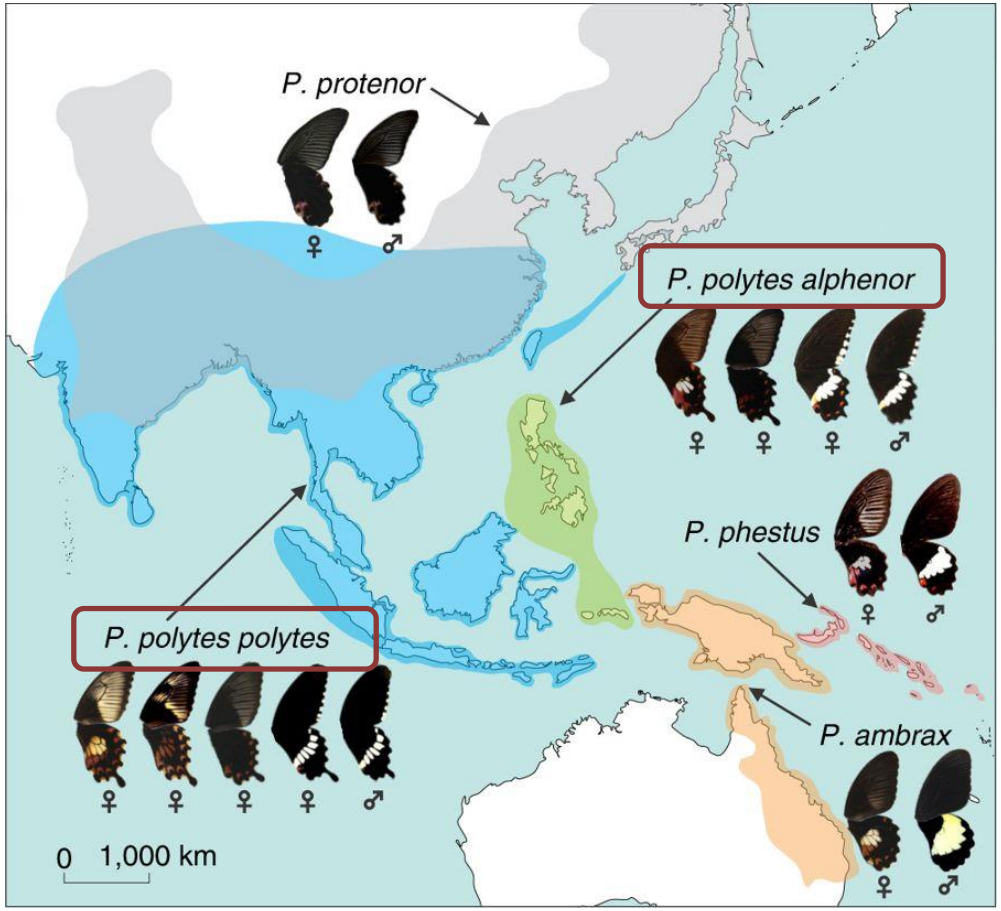




- 中国境内的拟南芥野生居群可能由两个较为独立的途径传入
- 长江流域居群（江西南丰、重庆铜梁、陕西城固、浙江东阳等）
- 新疆北部居群（阿勒泰、清河等）

He *et al.*, Heredity, 2007
Yin *et al.*,BMC Plant Biology, 2010

Identify the interrelationships between species

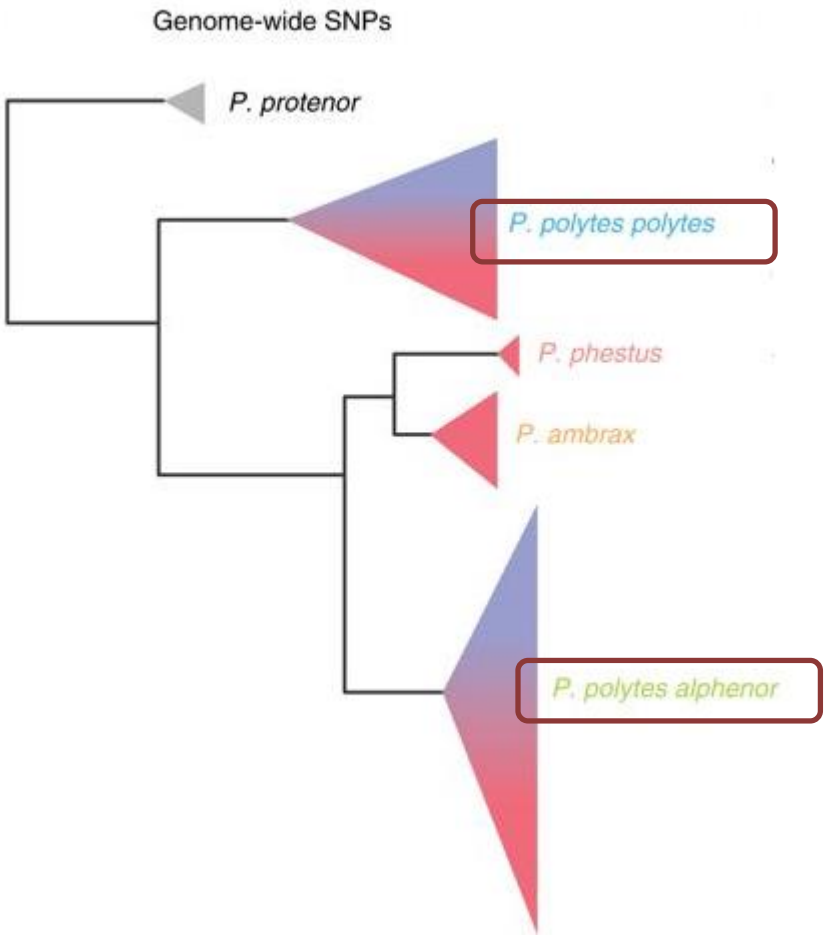


根据亚型的种类（表征种的分类方式）将***P. polytes alphenor***和***P. polytes polytes***归为一类。

表征种的定义：
形态上相似的一类生物
- 人为性很强， 没有任何遗传学、发育背景
- 没有 一个一致公认的标准
- 较为实用

Zhang *et al.*, Nature Communications, 2017

Identify the interrelationships between species



基于基因组SNP的系统发生树结果表明， ***P. polytes alphenor***和***P. polytes polytes***并不属于关系最近的两类物种。

P. phestus 和***P. ambrax***以前可能也有很多类型的亚种，但在演化历史中消失了。

系统发生树分析有助于确定物种关系，避免表征种分类造成的错误。

Zhang *et al.*, Nature Communications, 2017

The Great Wall of China: a physical barrier to gene flow?

居庸关长城始建于1386年，平均高约6米宽约5.8米。

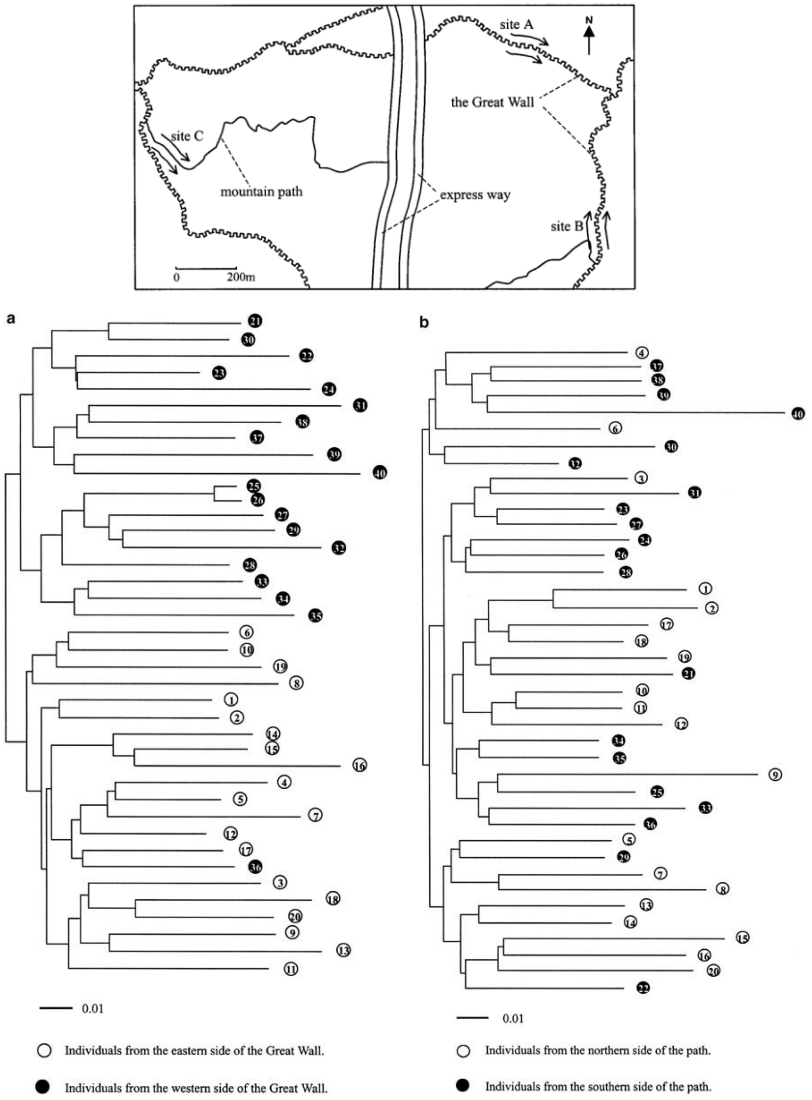
问题：长城对被其隔离的不同生活型、传粉方式植物亚居群之间的遗传分化是否有影响？

对照组，山间小路

Su *et al.*, Heredity, 2003

The Great Wall of China: a physical barrier to gene flow?

- 山杏（***Prunus armeniaca***）
- 酸枣（***Ziziphus jujuba***）
- 黄荆（***Vitex negundo***）
- 狗娃花（***Heteropappus hispidus***）
- 丛生隐子草（***Cleistogenes caespitosa***）
- 榆树（***Ulmus pumila***）
- 大果榆（***Ulmus macrocarpa***）
- 被居庸关长城隔离的植物亚居群间具有极显著的遗传分化
- 长城具有阻碍基因交流的作用



Su *et al.*, Heredity, 2003

Track the source of fungal infections

蛙壶菌, 感染无尾两栖类

- 上个世纪70年代人们注意到两栖类动物数量迅速减少　与气候和环境变化相关性不太大
- 1998年发现是一种真菌造成的
- 发现一个菌株毒性最强，广布于世界各地
- 问题：

起源地点：非洲，北美、南美、东亚？ 单基因序列研究

广布菌株起源时间：100年前，26000年前？

Track the source of fungal infections

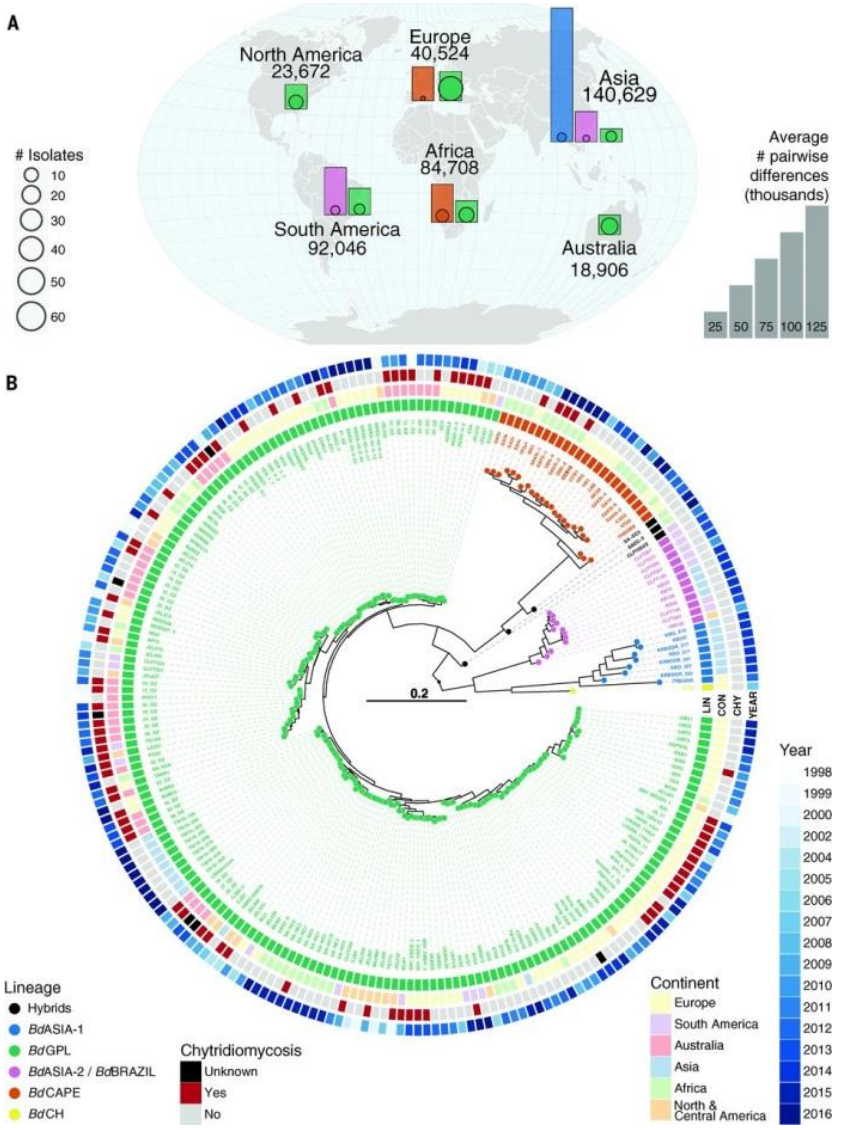
基因组测序分析：234 分离菌株

四个分支：

1. 广布菌株
2. 非洲菌株
3. 巴西/亚洲2菌株
4. 亚洲1/欧洲菌株 多样性最大

巴西/亚洲2菌株来自巴西牛蛙

亚洲1/欧洲菌株来自韩国

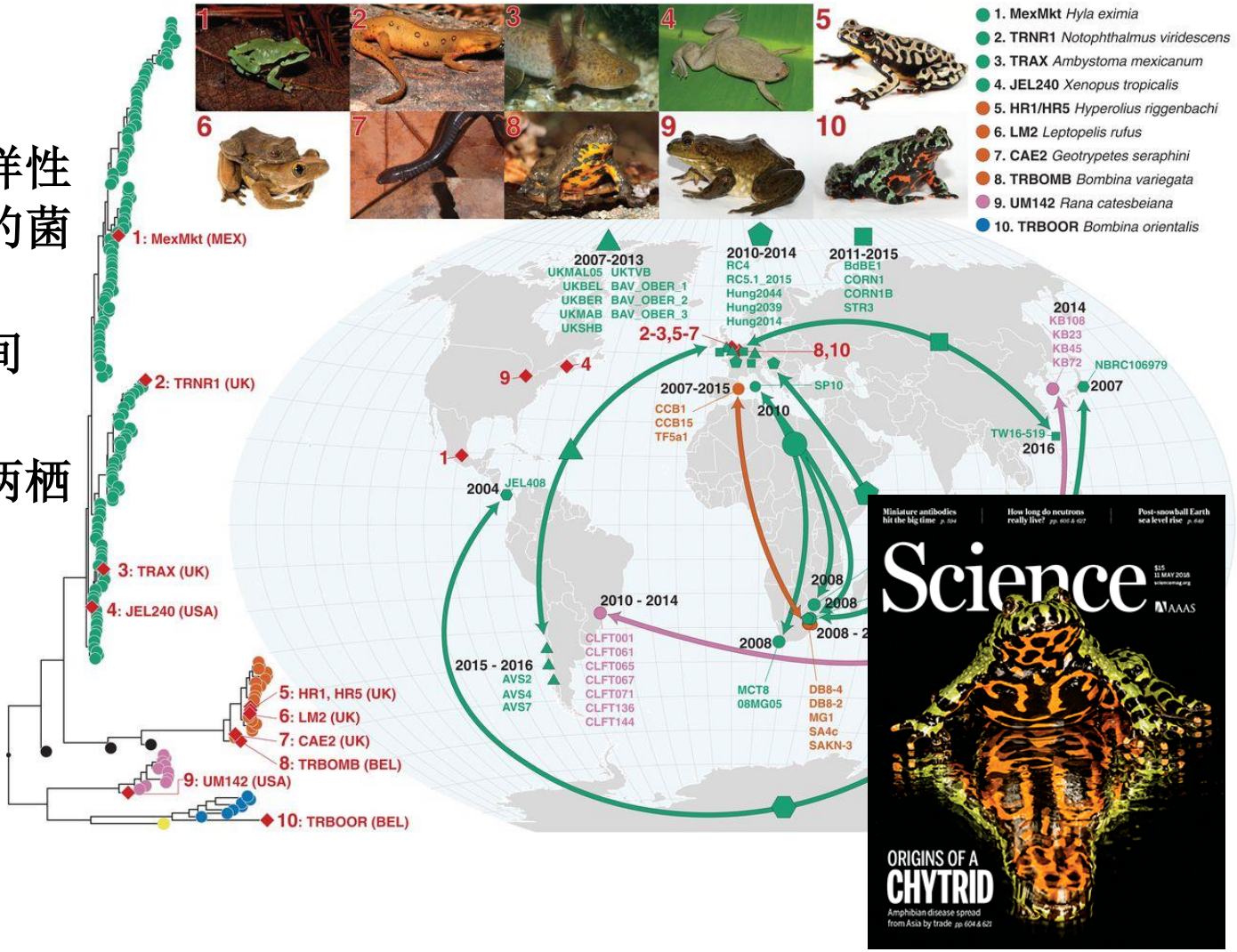


O'Hanlon *et al.*, Science, 2018

Track the source of fungal infections

结论：

1. 韩国是蛙壶菌多样性中心，保留了古老的菌株 ？
2. 广布菌株起源时间50-120年
3. 快速传播时间与两栖类贸易相同
4. 全球化贸易
5. 预防为主



O'Hanlon *et al.*, Science, 2018

A lawsuit in the United States in the 1990s

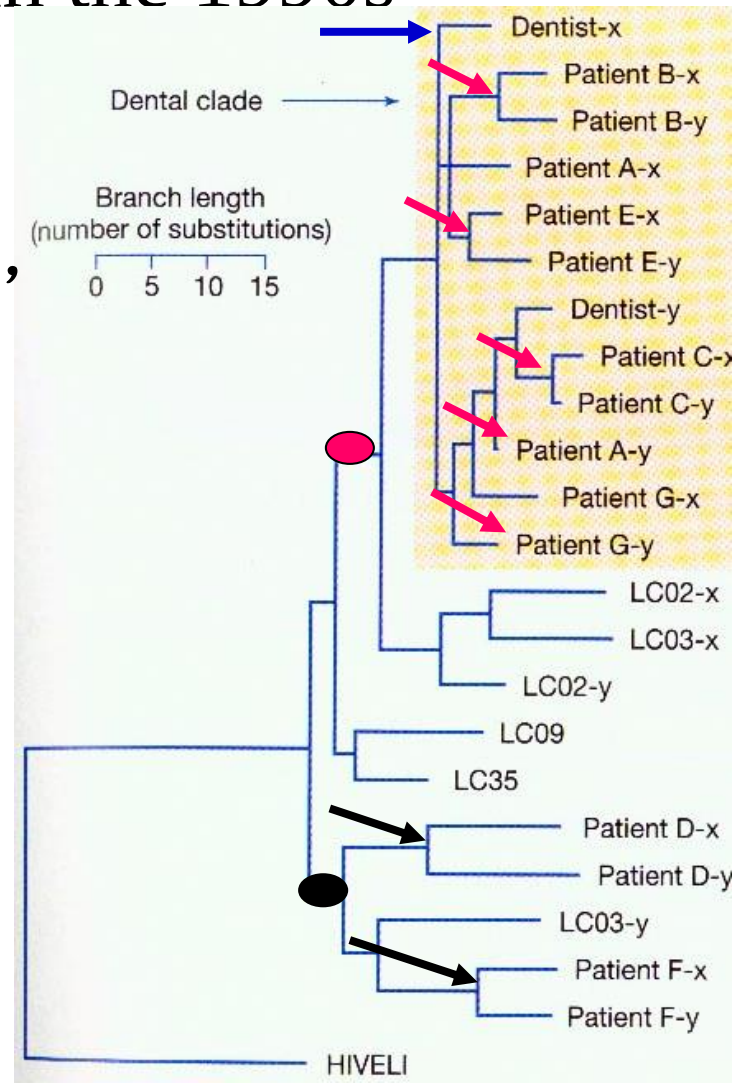
Florida的一位牙医得了艾滋病，他的一位病人被诊断为HIV阳性，其他病人均去检查，又发现了几个HIV阳性，便起诉牙医

美国CDC用牙医及其病人体内不同时间分离的HIV基因序列建树

结论：

- **有两位病人从其他来源感染了HIV**
- **有5位病人从牙医那感染了HIV**

该事件促使医院制定了相应的预防措施



Ou *et al.*, Science, 1992
DeBry *et al.*, Nature, 1993
Hillis *al.*, Science, 1994

A murder case in the United States in the 1990s

1994年美国路易斯安那州一位大夫给其诊所的护士打了一针

1995年护士发现HIV1阳性，控告大夫

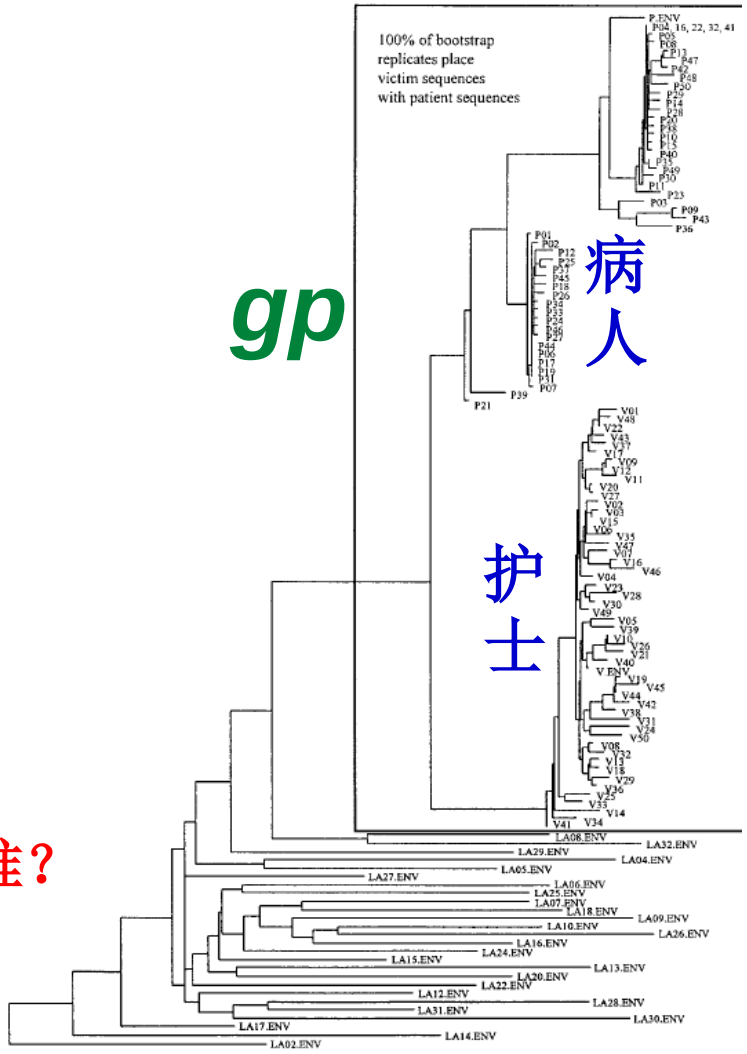
法院委托科学家调查

方法:构建系统发生树

材料:病人、护士、当地其他HIV病人血样

检测: 外壳蛋白 (*gp*)、逆转录酶基因(*RT*)

***gp*: 病人和护士样本在同一支上，但谁感染谁？**



Metzker *et al.*, Proc Natl Acad Sci U S A., 2002

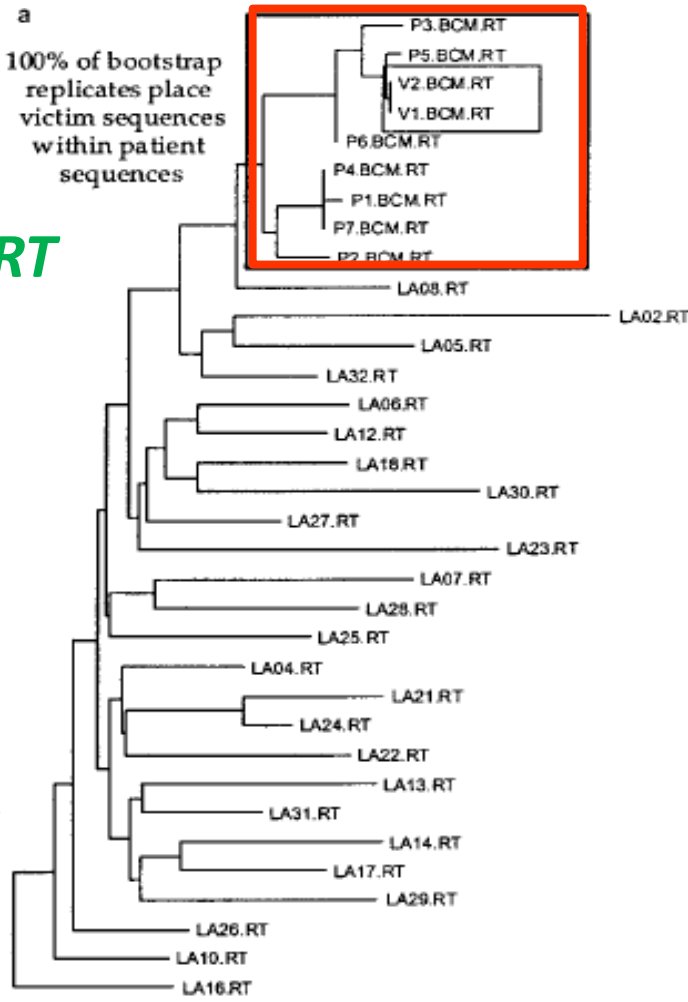
A murder case in the United States in the 1990s

***RT*: 护士的样本“埋”在病人的样本中**

结论：
护士体内的HIV病毒来自病人

州法院判决大夫有罪，大夫上诉

2002年最高法院判大夫二级谋杀罪名成立
从此，DNA证据被美国司法机构正式采用
两次抽样，两个实验室，双盲



Metzker *et al.*, Proc Natl Acad Sci U S A., 2002

Reference

- **张昀，1998，生物进化，北京大学出版社**
- **《生物进化》，2016，Futuyma著，葛颂、顾红雅等译**
- **《生物演化》，顾红雅、周忠和， 2016，高等教育出版社**
- **中国大学MOOC，生物演化，顾红雅，北京大学生命科学学院**

Acknowledgement

- **Prof. Luo Jingchu, Prof. Gu Hongya, Prof. Qu Li-jia**
- **Dr. Kang Juqing, Dr. Sun Tianshu**
- **Wang Xuefei, Zhu Chenqi**

Thanks