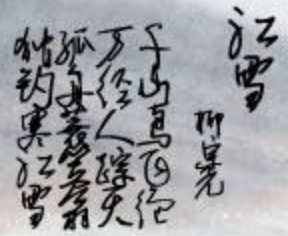


# 分子进化树

江山萬物皆有情  
孤舟蓑笠釣江雪  
柳宗元



# 进化生物学



- 进化生物学是研究生命的起源及进化的过程、原因、机制、速率和方向的科学。
- 进化生物学的基础理论就是进化论。
- 研究生物进化的三个途径：化石纪录，形态比较，大分子比较。

# 分子进化

江雪  
千山鳥飛絕  
萬徑人踪滅  
孤舟蓑笠翁  
獨釣寒江雪  
柳宗元

分子进化的特征

分子进化树(phylogenetic tree) 的构建

几个常见建树方法的介绍

可靠性检验

常用的软件包

# 生物大分子的进化特征

江山萬里  
孤舟蓑笠翁  
独釣寒江雪  
柳堤

- 如果以核酸或蛋白质的一级结构的改变(即分子序列中核苷酸或氨基酸的替换数)作为进化改变量的测度,那么生物大分子随时间的改变(即分子进化速率)是相当稳定的,其原因可能是“替换”是一个没有特殊驱动和控制的随机过程。

# 中性学说或中性漂移学说

江雪  
千山鳥飛絕  
萬徑人踪滅  
孤舟蓑笠翁  
獨釣寒江雪

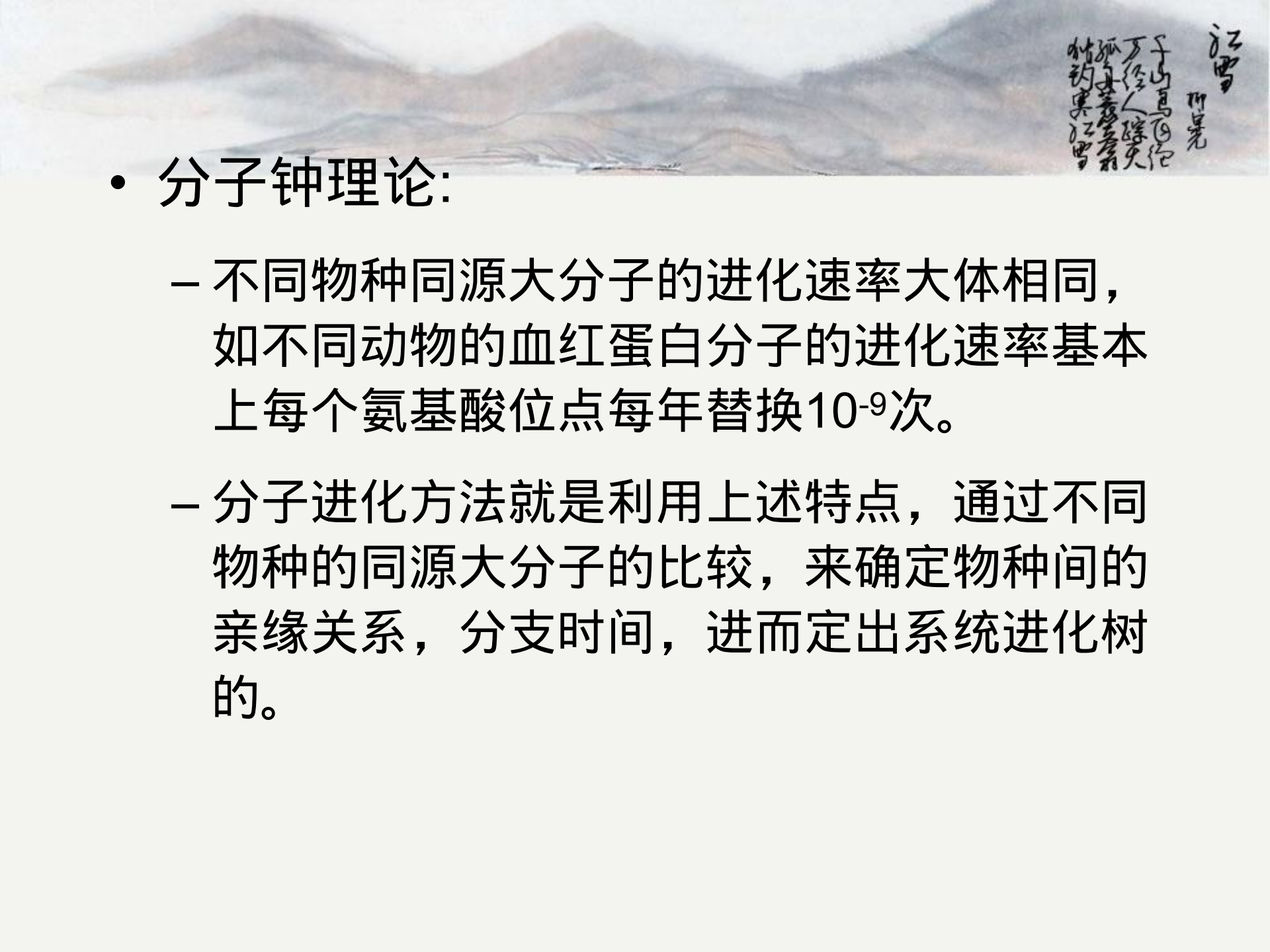


MOTOO KIMURA

日本遗传学家木村资生  
(Motoo Kimura, 1924-  
1994)

分子进化速率的恒定性: 进化过程中置换绝大部分是中性或近似中性的突变随机固定的结果。功能上对生命生存制约性低的分子或一个分子中不那么重要的部分, 较之对生命生存制约性高的分子或分子中重要的部分, 其突变率置换率高。

多态性的来源: 许多蛋白质多态性必须在选择上为中性或近中性, 并在群体中由突变引入与随机灭绝间两者的平衡维持。



江山萬里  
孤舟一葉  
釣寒江雪  
即景  
丁巳

- 分子钟理论:

- 不同物种同源大分子的进化速率大体相同，如不同动物的血红蛋白分子的进化速率基本上每个氨基酸位点每年替换 $10^{-9}$ 次。
- 分子进化方法就是利用上述特点，通过不同物种的同源大分子的比较，来确定物种间的亲缘关系，分支时间，进而定出系统进化树的。

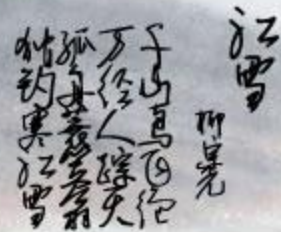


# 分子进化树

江雪  
千山鸟飞绝  
万径人踪灭  
孤舟蓑笠翁  
独钓寒江雪

- ✚ 拓扑图形
- ✚ 基于分子钟理论
- ✚ 相关的核酸或蛋白
- ✚ 反映一系列相关分子或生物如何演化或进化

# 分子进化树的构建



## 基本思想:

物种体内同功能生物分子（如蛋白质或核酸分子）的相似程度越高，则物种的亲缘关系越近。

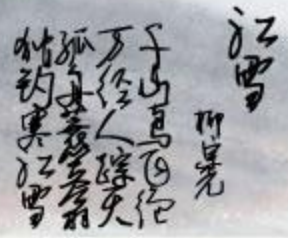
## 具体步骤:

- 选择“特征分子”，原则是：a. 各个物种都有的同源分子，b. 进化速率适当；
- 对这些同源分子的序列进行多序列比对(multi-sequences alignment), 截取比对的最好的区域作为物种的代表序列；





3. 按某种方法，算出代表序列两两之间的差异度(如: 比对打分,  $p=n/N$ ,  $D=-0.75\ln(1-4p/3)$  ...)
4. 基于这些差异度，绘制系统发生树
5. 对系统发生树进行可信度检验(bootstrap)



- 构建系统进化树的方法主要有：
  - 距离法（UPGMA、Neighbor-Joining）、
  - 最大简约法（Maximum parsimony, MP）
  - 最大似然法（Maximum likelihood, ML）
- 近缘序列: 最大简约法
- 远缘序列: 最大似然法或基于距离矩阵的方法
- 若序列有较高的相似性，各种方法都会得到不错的结果，模型间的差别也不大

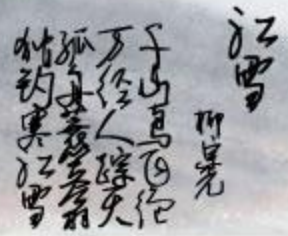
# 距离法 -Fitch and Margoliash Method

参考文献: (Science 155: 279-284, 1987)

江雪  
孤舟蓑笠翁  
独钓寒江雪  
柳宗元

步骤如下:

1. 对要研究的一组物种(或序列), 采用某种规则, 算出两两之间的距离, 构造出**距离矩阵**;
2. 选出距离最近的两个物种(比如A和C), 把其余的物种看成一个“**复合物种**”, A到“**复合物种**”的距离是A和构成“**复合物种**”的所有物种的距离的平均, C亦然;



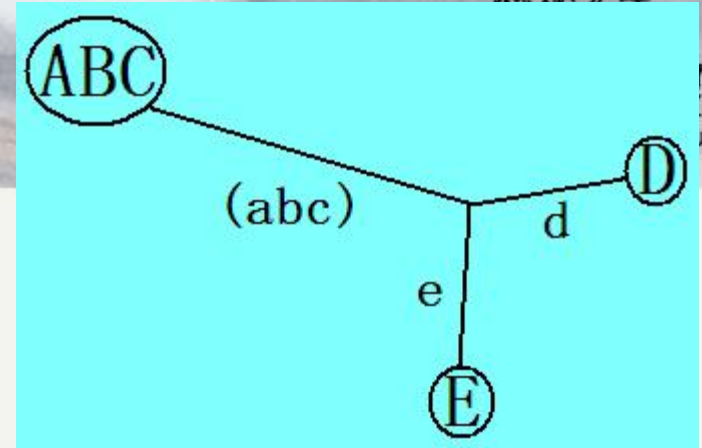
3. 求出A、C、“复合物种”三者构成的树的分支长度 $a, c, (\text{复合物种})$ ，这时，A、C的“紧邻分支长度已确定”；
4. 将“紧邻分支长度已确定”的A、C看成一个“复合物种”（AC），计算其它各物种到（AC）的距离，构造出新的（少了一维的）距离矩阵；
5. 按照同样的办法，一遍遍重复2到4的步骤，就可逐步确定出树的整体结构及全部分支长度。

江雪  
千山鸟飞绝  
万径人踪灭  
孤舟蓑笠翁  
独钓寒江雪  
柳宗元

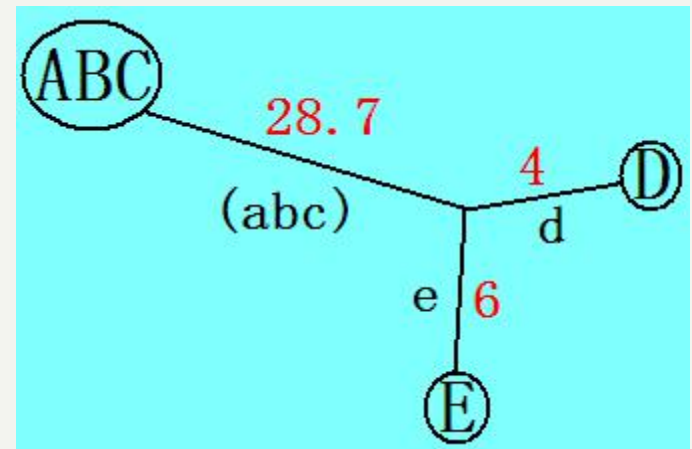
例子：

	A	B	C	D	E
A	—	22	39	39	41
B	—	—	41	41	43
C	—	—	—	18	20
D	—	—	—	—	10
E	—	—	—	—	—

	D	E	(ABC)
D	—	10	32.7
E	—	—	34.7
(ABC)	—	—	—



$$\begin{cases} d + e = 10 \\ d + (abc) = 32.7 \\ e + (abc) = 34.7 \end{cases} \rightarrow$$

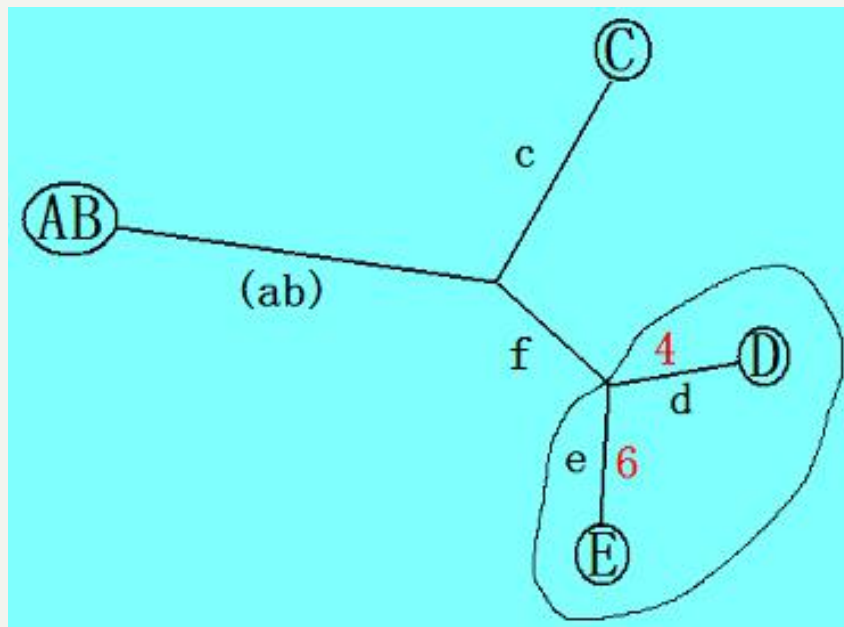




	A	B	C	D	E
A	—	22	39	39	41
B	—	—	41	41	43
C	—	—	—	18	20
D	—	—	—	—	10
E	—	—	—	—	—

	A	B	C	(DE)
A	—	22	39	40
B	—	—	41	42
C	—	—	—	19
(DE)	—	—	—	—

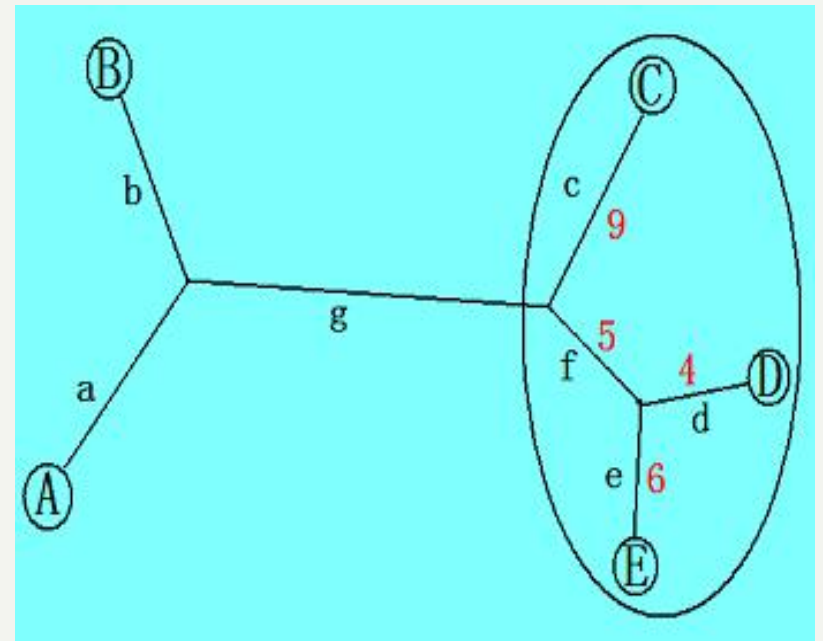
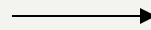
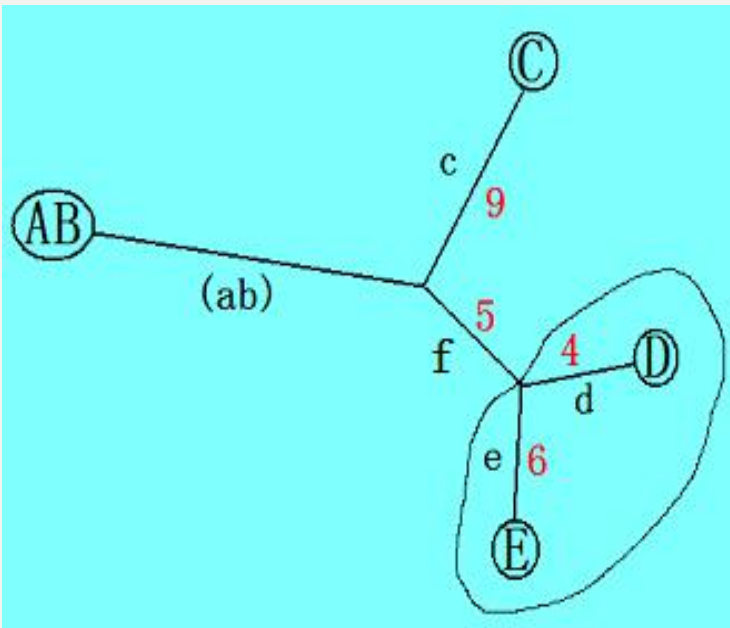
	(AB)	C	(DE)
(AB)	—	40	41
C	—	—	19
(DE)	—	—	—



$$\begin{cases} (ab) + c = 40 \\ (ab) + (de) = 41 \\ (de) + c = 19 \end{cases} \longrightarrow \begin{cases} c = 9 \\ (de) = 10 \\ (ab) = 31 \end{cases}$$

孤舟蓑笠翁  
独钓寒江雪  
江雪  
柳宗元

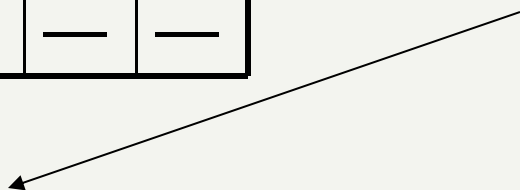
$$(de) = [(f + 6) + (f + 4)] \div 2 = 10 \quad \rightarrow \quad f = 5$$



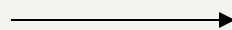
	A	B	C	D	E
A	—	22	39	39	41
B	—	—	41	41	43
C	—	—	—	18	20
D	—	—	—	—	10
E	—	—	—	—	—



	A	B	(CDE)
A	—	22	119/3
B	—	—	125/3
(CDE)	—	—	—

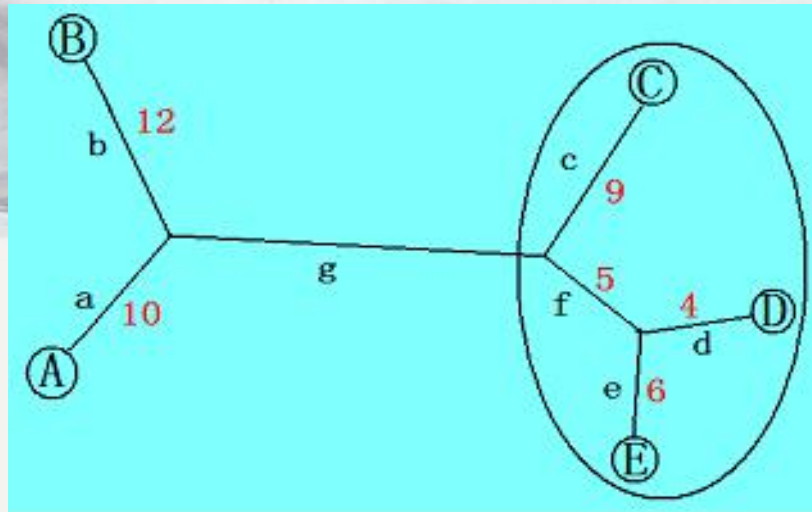


$$\begin{cases} a + b = 22 \\ a + (cde) = \frac{119}{3} \\ b + (cde) = \frac{125}{3} \end{cases}$$



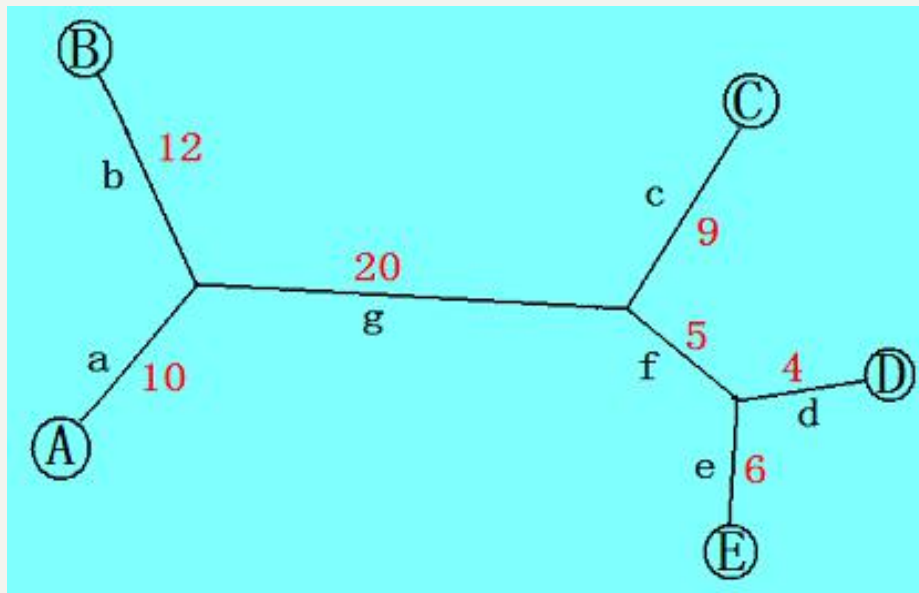
$$\begin{cases} a = 10 \\ b = 12 \\ (cde) = \frac{89}{3} \end{cases}$$

江雪  
千山鸟飞绝  
万径人踪灭  
孤舟蓑笠翁  
独钓寒江雪  
柳宗元



江雪  
柳岸花  
孤舟蓑  
笠翁  
独钓寒  
江雪

$$[(g + 9) + (g + 5 + 4) + (g + 5 + 6)] \div 3 = \frac{89}{3} \rightarrow g = 20$$





- 该方法的优点是：
  - 不要求物种(序列)的相似程度必须很高，并且计算速度快，能处理物种(序列)数目较大的情况；
- 缺点是：
  - 不能保证找到的是最优的树。



- 构建系统进化树的方法：
  - 距离法（UPGMA、Neighbor-Joining）、
  - 最大简约法（Maximum parsimony, MP）
  - 最大似然法（Maximum likelihood, ML）



# 进化树的可信度检验

江雪  
柳光  
千山鳥飛絕  
萬徑人踪滅  
孤舟蓑笠翁  
獨釣寒江雪

如何判断构建的进化树是否正确

# 可信度检验：Bootstrap

方法: 对数据进行干扰，判断结果的稳定性。

江雪  
千山鸟飞绝  
万径人踪灭  
孤舟蓑笠翁  
独钓寒江雪

S1: AACCAAC  
S2: AACCCC  
S3: ACCAAC  
S4: CCACCA  
S5: CCAAAC

从整个序列的碱基（氨基酸）中任意选取一半，剩下的一半序列随机补齐组成一个新的序列。这样，一个序列就可以变成了许多序列，一个多序列组也就可以变成许多个多序列组。每个多序列组都可以生成一个进化树。将生成的许多进化树进行比较，按照多数规则（majority-rule）我们就会得到一个最“逼真”的进化树。

S1: ACCCAC  
S2: ACCCCC  
S3: CCCCAC  
S4: CAAACA  
S5: CAAAAC

S1: AAAACC  
S2: AACCCC  
S3: ACAACC  
S4: CCCCAA  
S5: CCAACC

S1: AAAAAC  
S2: AACCCC  
S3: CCAAAC  
S4: CCCCCA  
S5: CCAAAC


.....



# The bootstrap

江雪  
千山鸟飞绝  
万径人踪灭  
孤舟蓑笠翁  
独钓寒江雪  
柳宗元

- BOOTSTRAP值的意义：根据你所选的统计计算模型，设定初始值1000次，即让模型计算并绘制1000株系统发育树，这是命令阶段产生的，而系统发育树中每个节点上的数字，则代表在命令阶段要求的1000次进化树分析中 有多少次是具有相同的进化分类的关系



江雪  
千山鸟飞绝  
万径人踪灭  
孤舟蓑笠翁  
独钓寒江雪

在任何一组新的序列中：

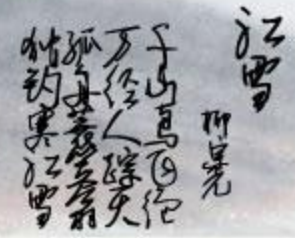
- 序列的长度和原始的长度一样；
- 某些“列”可能被使用多次，而某些“列”则可能没用到。
- 当Bootstrap的值 $>70$ ，一般都认为构建的进化树较为可靠。
- 对于进化树的构建，如果对理论的了解并不深入，则推荐使用缺省的参数，并启用Bootstrap检验。一般情况下，使用两种不同的方法构建进化树，如果得到的进化树基本一致，结果较为可靠。

# 分子进化的局限性

江雪  
千山鳥飛絕  
萬徑人踪滅  
孤舟蓑笠翁  
獨釣寒江雪

- 单个分子代表整个物种的片面性问题；
- LGT(基因横向迁移)问题；
- 难以区分直系和 旁系的问题；
- 无法考虑多次突变或回复突变；

# 常用软件



- PHYLIP
  - <http://evolution.genetics.washington.edu/phylip.html>
- Mega
  - <https://www.megasoftware.net>