中国科学院植物研究所

生物多样性与生物安全研究组组会报告

从DNA序列到分化时间
——进化树与分子钟

博士研究生 张金龙

2010-1-29

报告内容

一 DNA序列

二 序列比对

三 碱基替换模型及其筛选

四 进化树的构建

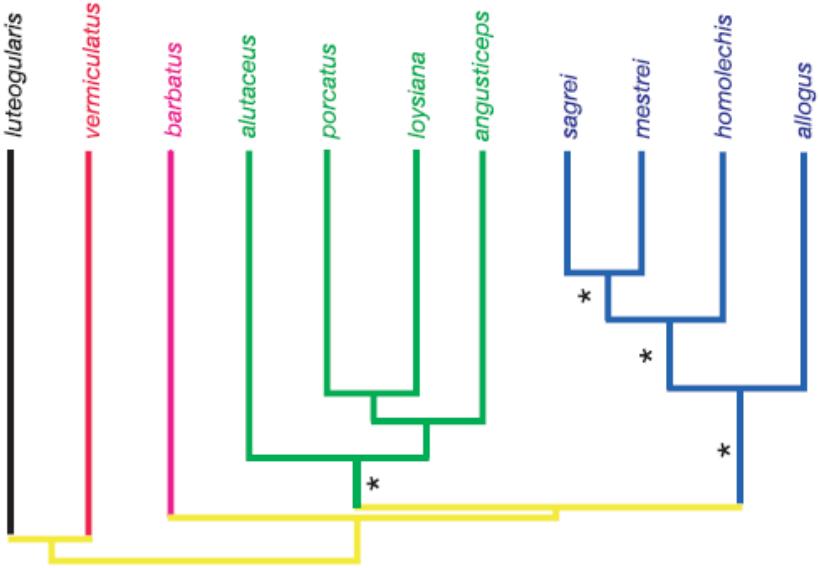
五 树的可信度 Bootstrap

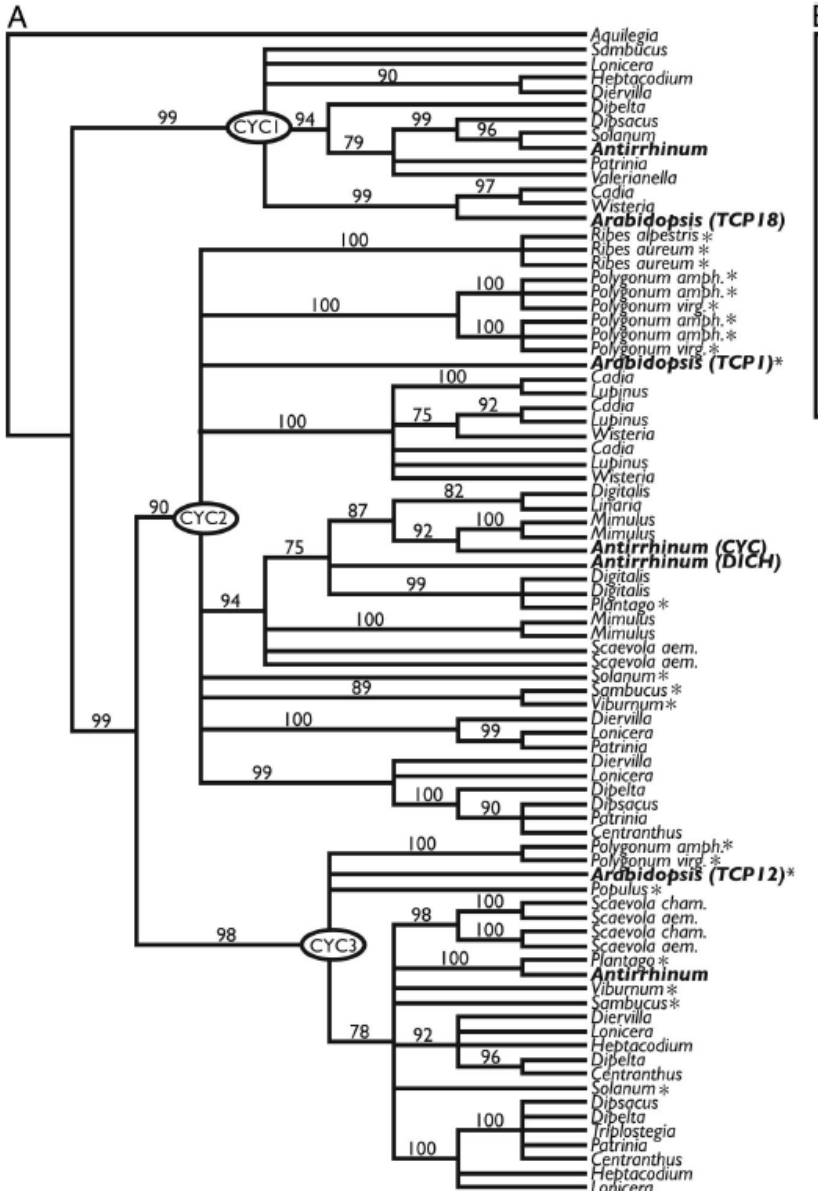
六 分子钟

进化树

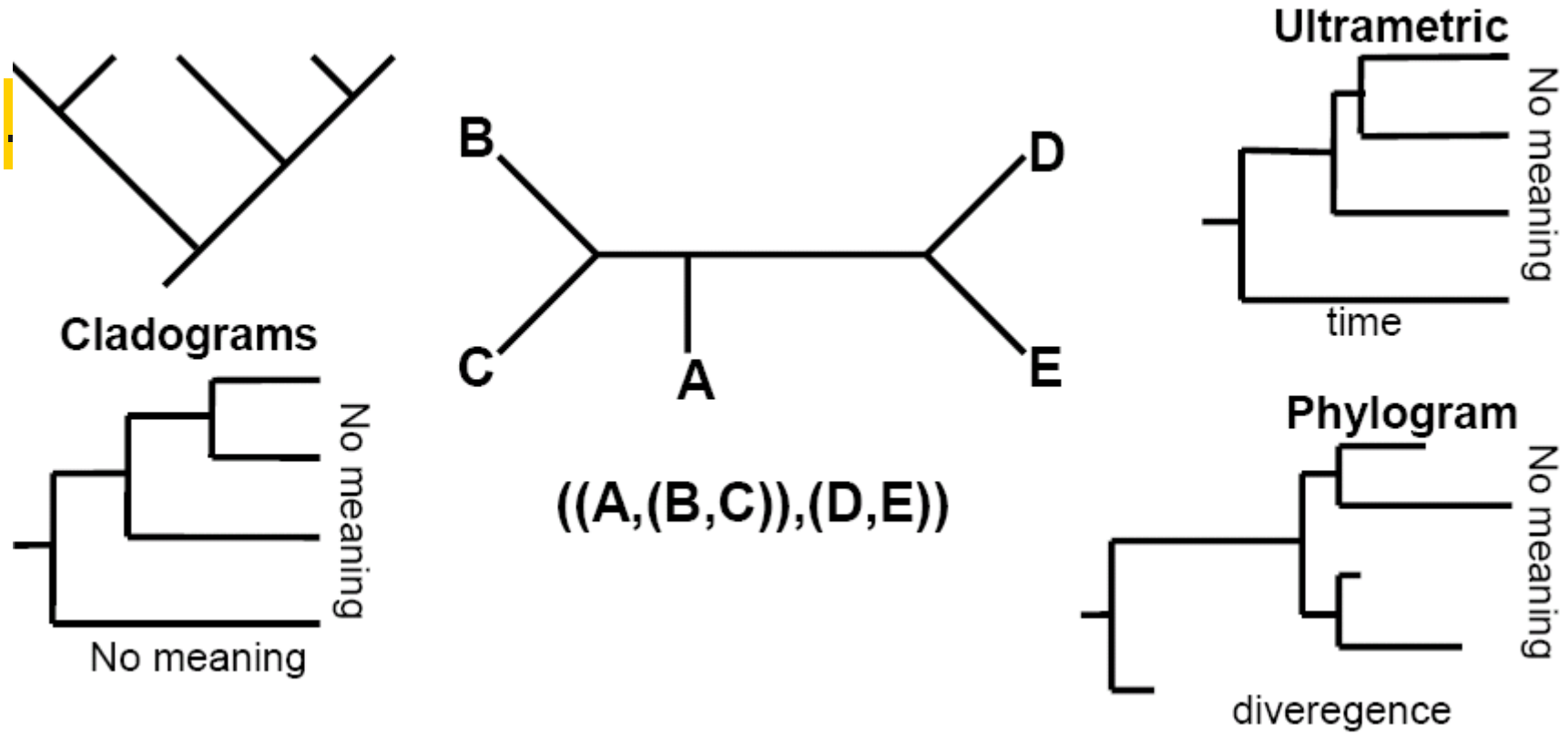
- 什么是进化树？
- 进化树是表示物种间系统发育关系的树状图，枝长的长短表示进化距离的差异。系统关系越近的物种，在进化树中的距离越近。
- 主要内容： 用DNA序列构建进化树的原理和方法

进化树





- 1 Losos et al. *nature*,2003 2 Howarth et al. *PNAS* ,2006

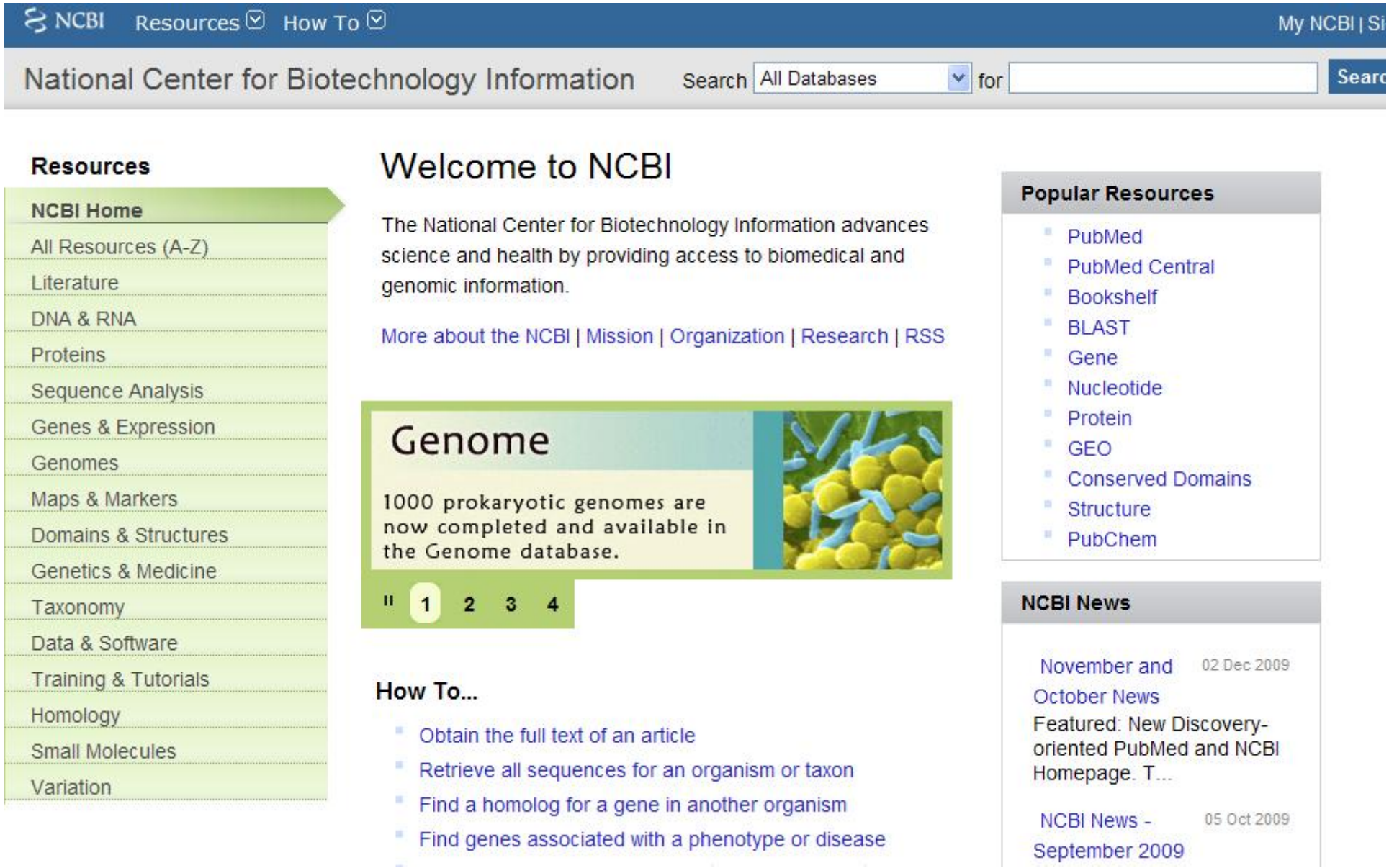


3. 各种进化树 (自N. Nikolaidis）

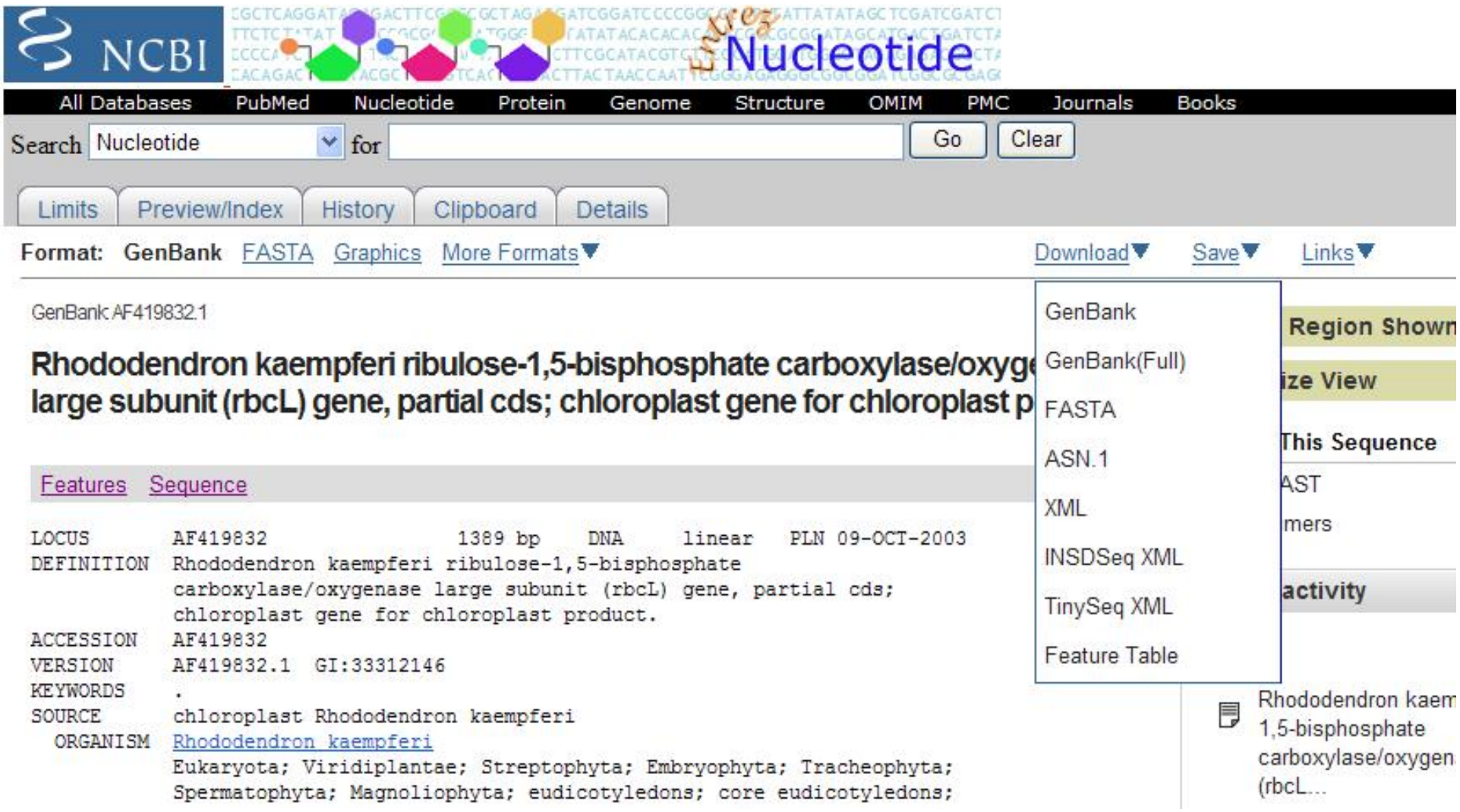
一 DNA序列

DNA序列

- 由ATCG四个碱基组成，一般从其3′端作为起始。
- 一个基因的长度在几百到几千个bp(碱基对)不等。
- 现有的序列可在NCBI 检索，下载。
- NCBI National Center for Biotechnology Information
- www.ncbi.nlm.nih.gov
- 可利用Paradis 的 ape程序包、BioPython等软件在线读取 。


NCBI
Resources
How To
My NCBI | Si
National Center for Biotechnology Information
Search
All Databases
for
Searc
Resources
NCBI Home
All Resources (A-Z)
Literature
DNA & RNA
Proteins
Sequence Analysis
Genes & Expression
Genomes
Maps & Markers
Domains & Structures
Genetics & Medicine
Taxonomy
Data & Software
Training & Tutorials
Homology
Small Molecules
Variation
Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
More about the NCBI | Mission | Organization | Research | RSS
Genome
1000 prokaryotic genomes are now completed and available in the Genome database.
1 2 3 4
How To...
Obtain the full text of an article
Retrieve all sequences for an organism or taxon
Find a homolog for a gene in another organism
Find genes associated with a phenotype or disease
Popular Resources
PubMed
PubMed Central
Bookshelf
BLAST
Gene
Nucleotide
Protein
GEO
Conserved Domains
Structure
PubChem
NCBI News
November and October News
02 Dec 2009
Featured: New Discovery-oriented PubMed and NCBI Homepage. T...
NCBI News - September 2009
05 Oct 2009

- NCBI检索



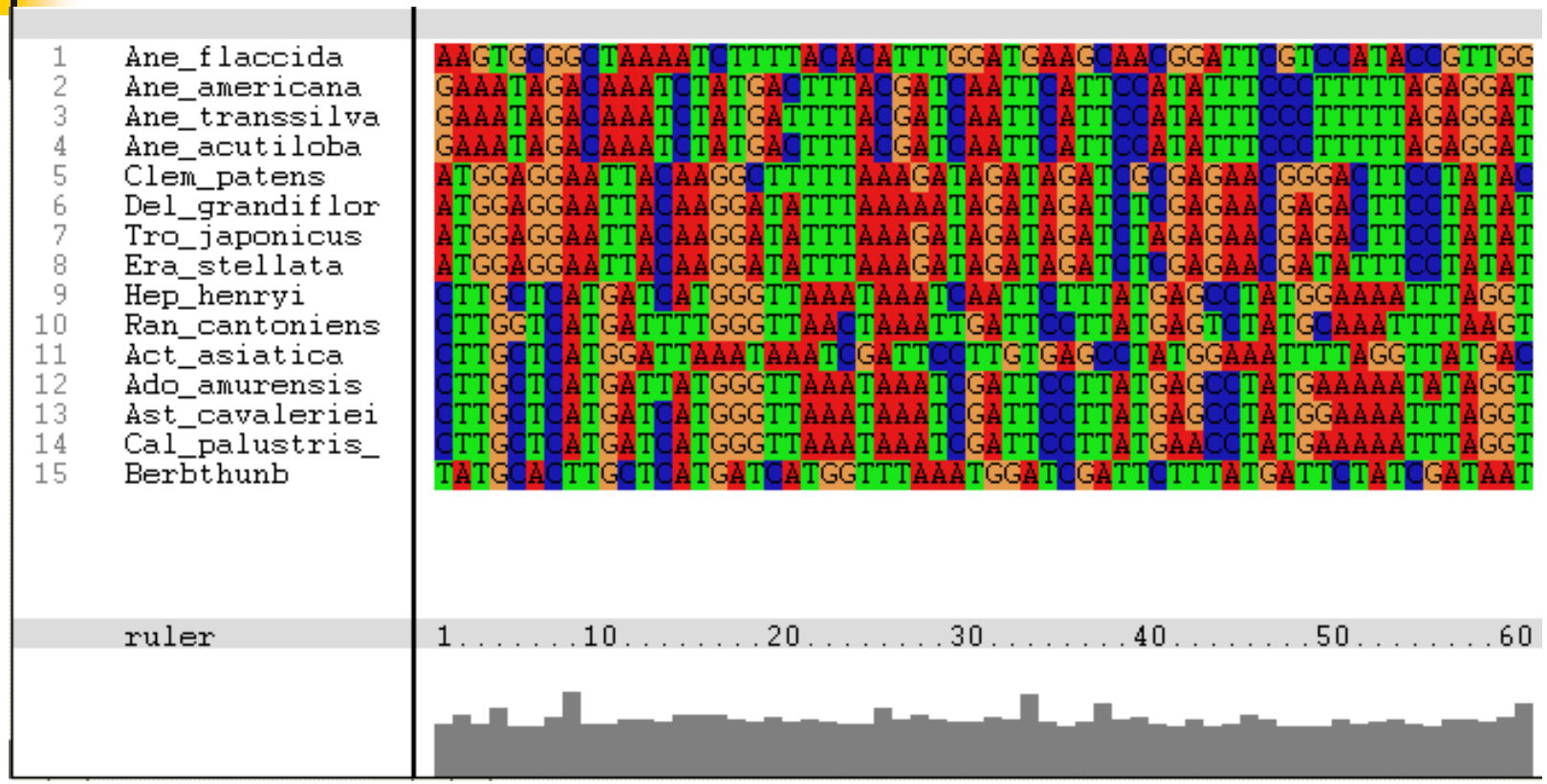
将序列信息另存为Fasta格式

二 序列比对

序列比对

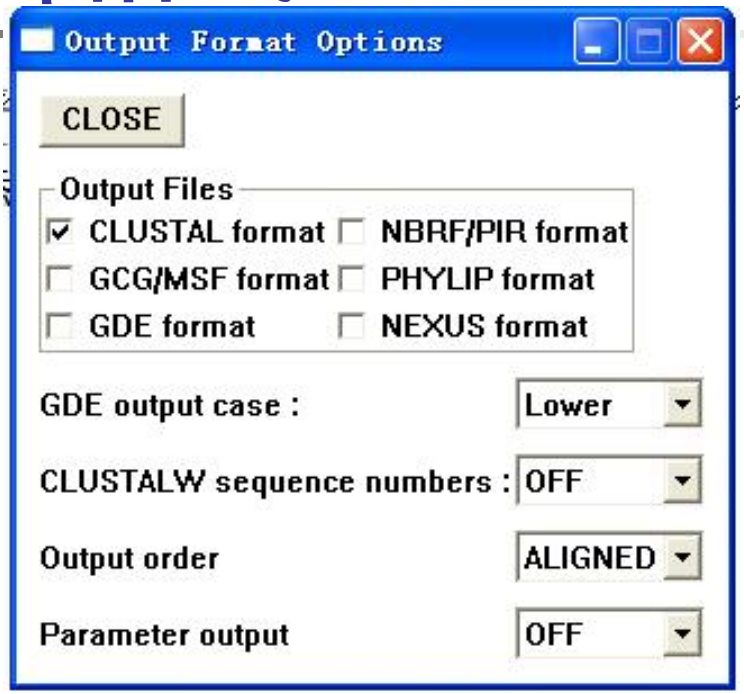
- 为什么要进行序列比对？
- 把相同位点调整到一起，从而发现某些物种在该位点的变异，寻找信息位点。
- 软件：ClustalX, MUSCLE　以ClustalX为例

序列比对ClustalX

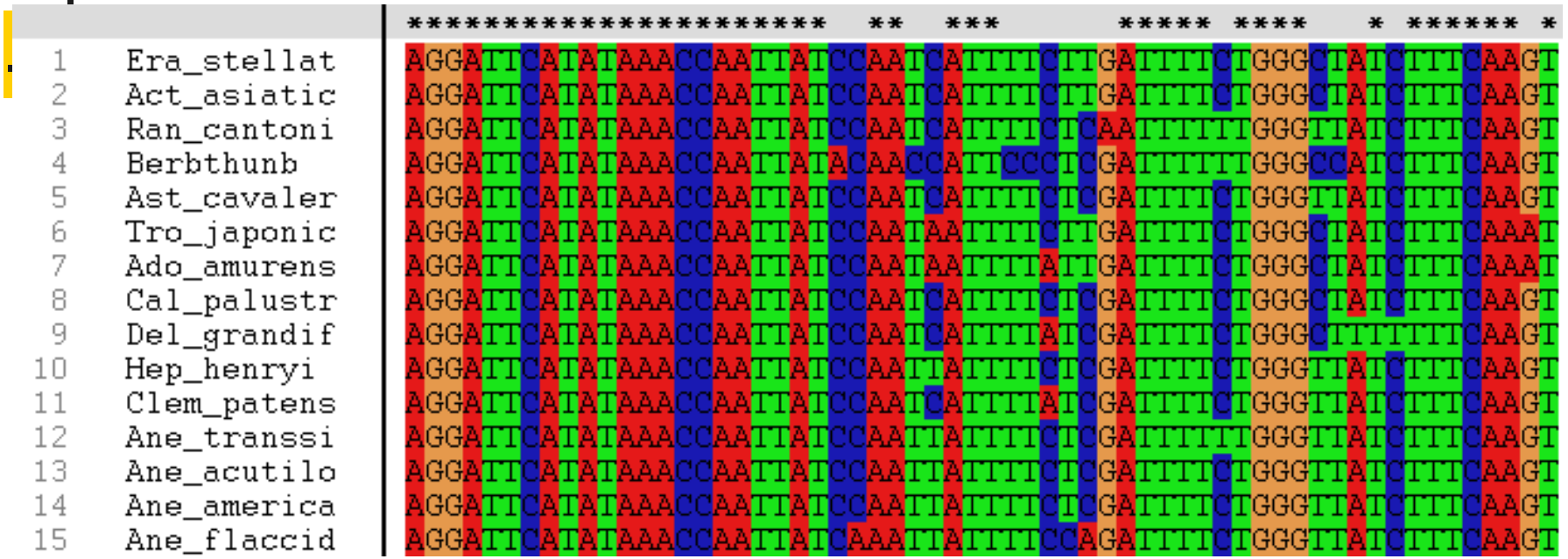


- 将下载的文件全部转移粘贴到一个fasta文件中，物种名前加“>”。

设定输出格式



- PHYLIP格式（PHYLIP， RAMXL,）
- NEXUS格式 （PAUP*, MrBayes等）
- Do Complete Alignment



- 序列经过ClustalX比对之后

比对后的校正

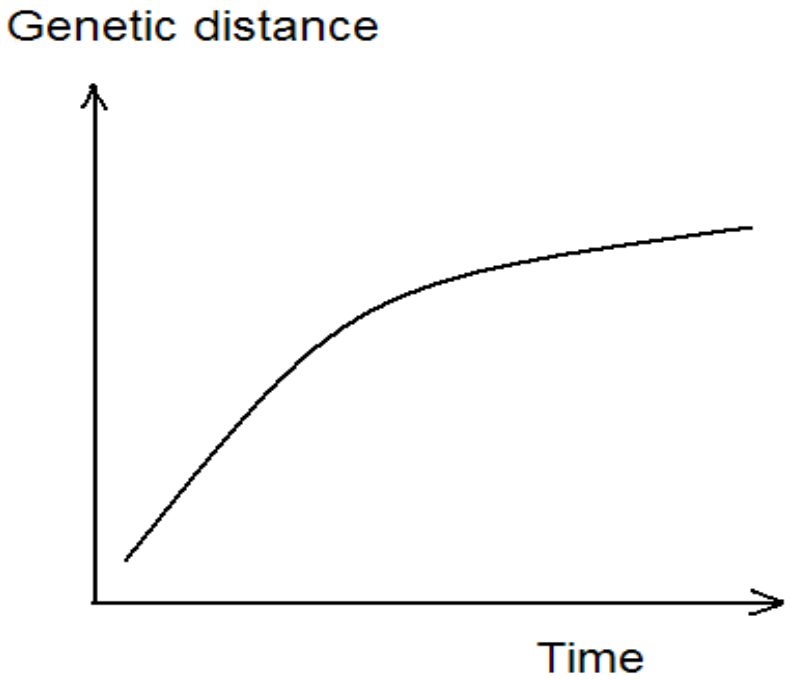
- 比对出错的地方进行手工校对。
- 对编码基因，三联体密码子，因此，碱基的增加或缺失应该是三的倍数，如果出现其他情况，很可能是比对错误或测序错误。
- 可用MEGA软件转换成BioEdit可以编辑的格式，再进行编辑。
- 此外，信息缺失的位点应该用“?”代替，经过比对间隔开的位点应该用“-”代替。

如何建树？

- 以NEXUS文件或PHYLIP文件为基础。
- 建树的方法包括：包括UPGMA法， 邻位法，最少进化法等还有极大似然法、贝叶斯法、最大简约法等。
- 有为数众多的软件可以实现，
- J. Felsenstein 给出了非常详尽的列表
- 参见 http://evolution.genetics.washington.edu/phylip/software.html
- 基于以上方法等都需要考虑ATCG相互替换的关系，即碱基替换模型。

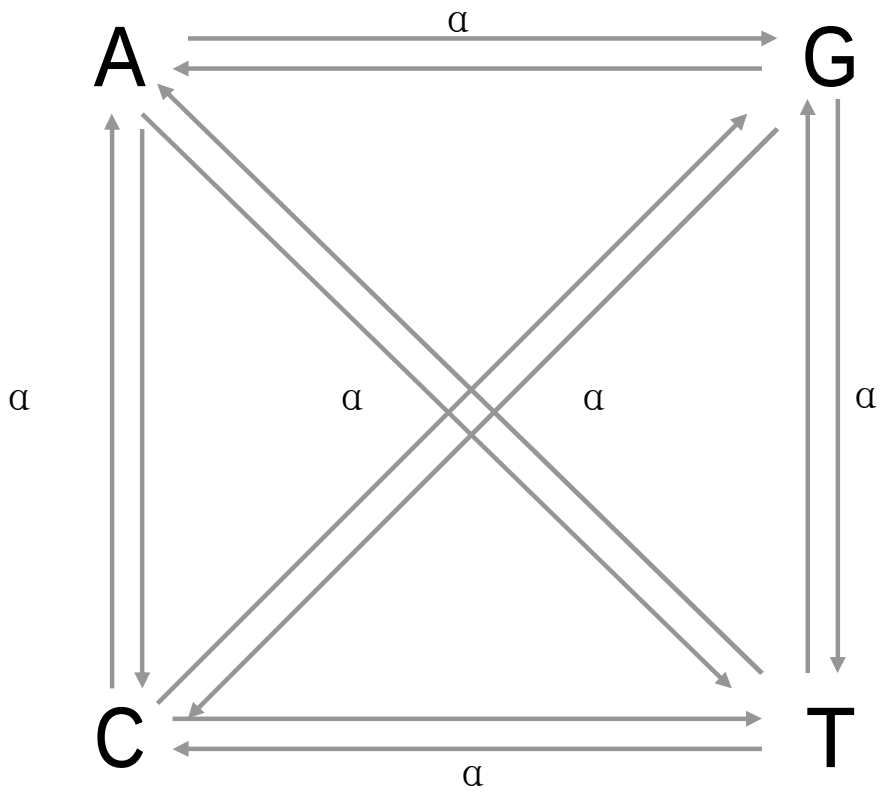
三 碱基替换模型及其筛选

序列之间的遗传距离



- 计算序列之间的遗传距离？ 如果只根据碱基组成的相似性，则经过一段时间，部分碱基又将突变回原有的碱基，因此，必须考虑碱基替换

最简单的碱基替换模型
Jukes-Cantor模型（JC69）



Jukes-Cantor模型（JC69）

$$M=\begin{array}{c} \begin{matrix} \quad A & \quad C & \quad G & \quad T \end{matrix} \\ \begin{bmatrix} 1-3a & a & a & a \\ a & 1-3a & a & a \\ a & a & 1-3a & a \\ a & a & a & 1-3a \end{bmatrix} \end{array}$$

P随时间的变化

- ACGT现有相同的频率（均为25%），假设某一位点，在起始时刻（t=0时）碱基为A
- t=1 时，即经历一个单位时间后(t=1)，该位点仍然为A的概率为

$$P_{A(1)} = (1-3a)$$

- t=2时 $P_{A(2)} = (1-3a)P_{A(1)} + a[1-P_{A(1)}]$

P随时间的变化

- 当时间为t+1时

$$P_{\mathrm{A}(t+1)} = (1-3a)P_{\mathrm{A}(t)} + a[1-P_{\mathrm{A}(t)}]$$

- 对于连续的时间过程，则有

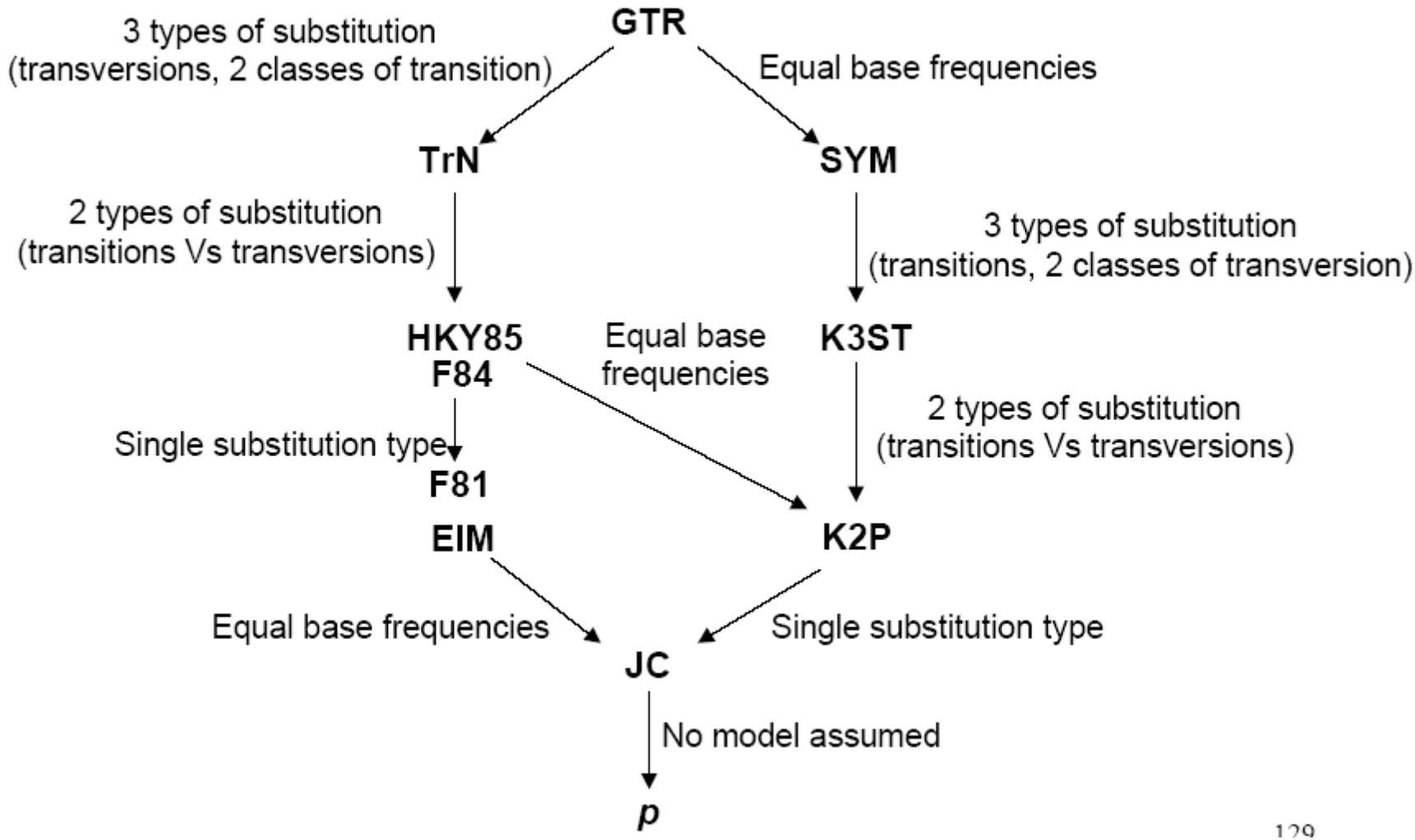
$$\frac{dP_{A(t)}}{dt} = -4a\,P_{A(t)} + a$$

P随时间的变化

- 若令$P_{A(0)}=1$,则

$$P_{A(t)} = \frac{1}{4} + \frac{3}{4}e^{-4at}$$

- 其他的进化模型如K2P、F81、 HKY85等都是从JC69模型发展起来的。主要考虑到不同碱基间置换的速率不同，添加了更多的参数。



129

- 碱基替换模型间的关系 自 N. **Nikolaidis**

碱基替换模型

- General Time Reversible model (GTR)
- 是所有碱基替换模型中考虑参数最多的，之前的模型都可以看做GTR模型的特例。
- 实际的碱基比例是不等的（两两之间），两两之间的替换率也是不等的，而所有这些参数的均已经以整合到GTR模型中。

不同位点进化速率的差异

- 密码子的最后一位比前两位受到的限制少，变化速率更快。为了在建树时表示各位点的进化速率的差异，人们引入了gamma分布

$$Pdf(r) = a^{a} r^{a-1} / \exp(a r)\Gamma(a)$$

- 用Gamma分布的形状表示碱基位点之间进化速率的异质性

模型的筛选

- 不同的模型预测的精度甚至结果会完全不同
- 模型的参数过多带来的后果
- 1 需要对每个参数进行估计，计算会更困难
- 2 参数过多，带来的误差可能会增加
- 如何在参数的数量和结果的精度之间权衡，即选择不同的进化模型？
- 常用如下方法： hLRTs和AIC

(1) hLRTs

- hierarchical Likelihood Ratio Tests
  - 似然函数：给定进化模型M，模型的K个参数 θ ，进化树拓扑结构 τ ，枝长 υ 之后，当前序列出现的可能性。

$$L = P(D \mid M, q, t, u)$$

  - 如何取这些参数，使得该序列出现的可能性最大？

$$\hat{q}, \hat{t}, \hat{u} = \max_{q,t,u} L(q, t, u)$$

LRT (Likelihood Ration Test)

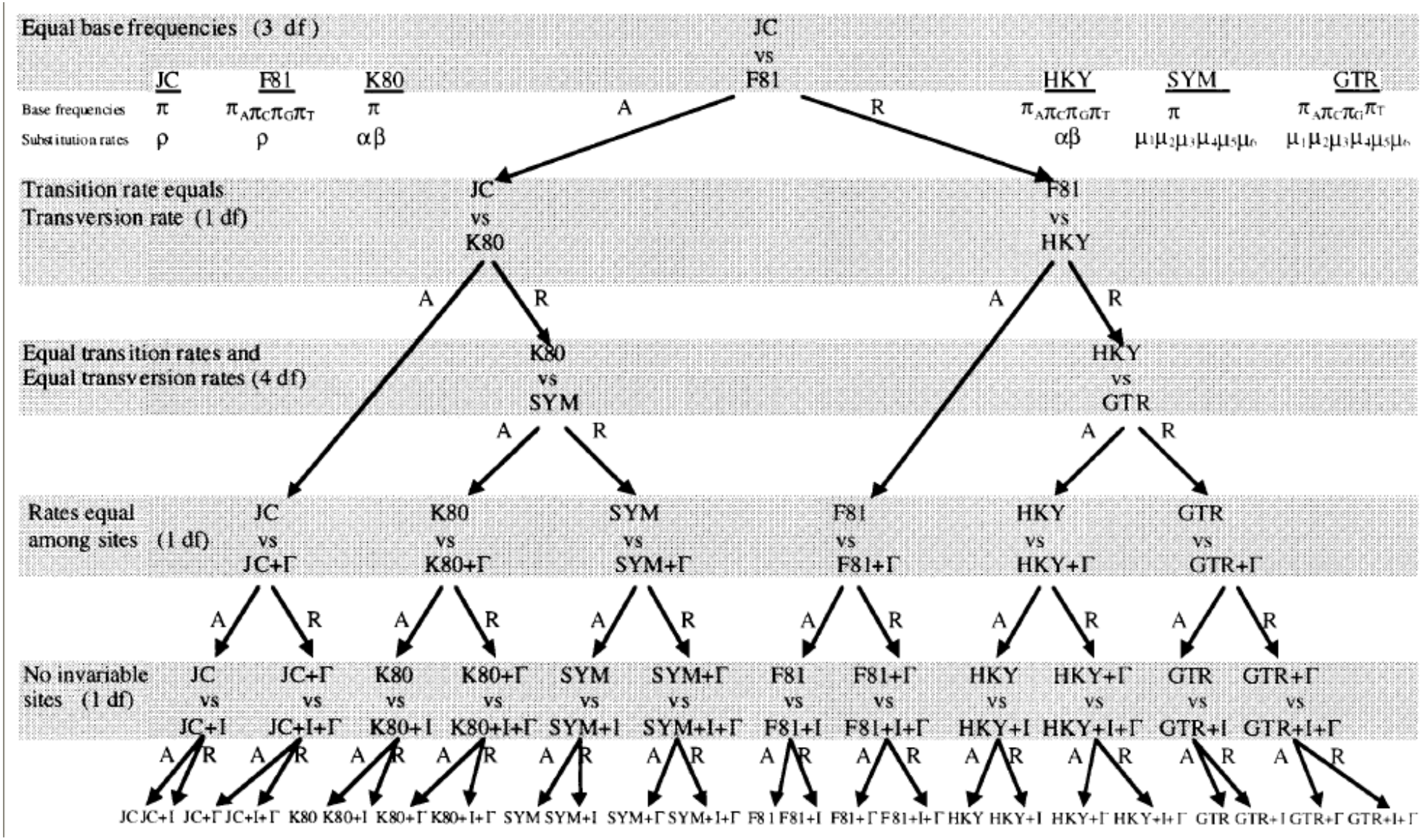
- 取对数，便于求极值

$$\mathbf{l} = \ln P(D \mid M, \hat{q}, \hat{t}, \hat{u})$$

- Likelihood Ratio Test (LRT)

$$LRT = 2(\mathbf{l}_1 - \mathbf{l}_2)$$

- 第一个模型参数多，第二个模型参数少，两者的极大似然函数对数的差值小（如果第二个模型是第一个模型的特例，可以用chisq test），表明参数的增加对预测的精度没有显著影响。



- Hierachical likelihood ratio tests in ModelTest

(2)AIC Akaike Information Criterion

- 赤池信息量

$$AIC = -2\mathbf{l} + 2K$$

- 样本量很少的时候，建议用改进的AIC
- Corrected AIC

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1}$$

模型筛选的软件

- ModelTest（用于PAUP*等软件）
- MrModelTest等（用于MrBayes）
- ModelTest 可以在DOS环境下运行
- 例如
- **cd c:\modeltest**
- **modeltest3.7 –n896 –t18 < mydata.scores > mydata.modeltest**

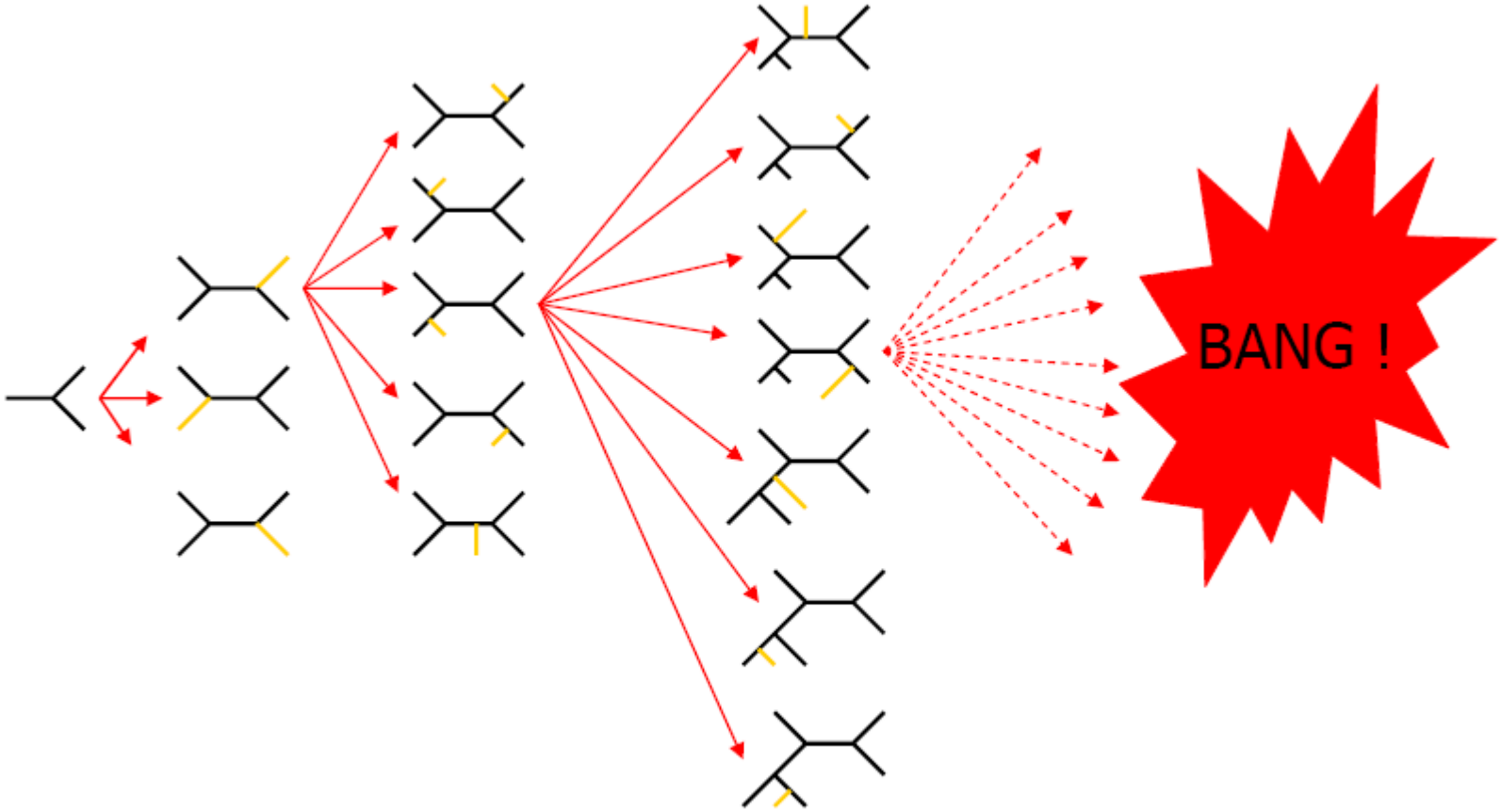
四 进化树构建

进化树的数量-天文数字

- 给出若干的物种，其进化树的结构有多少种可能？
- 进化树的数量随着要分析的物种数的增加迅速增加，服从以下公式：

$$t_n = \frac{(2n-5)!}{2^{n-3}(n-3)!} = \prod_{i=1}^{n}(2i-5)$$

n物种数, t进化树的数目

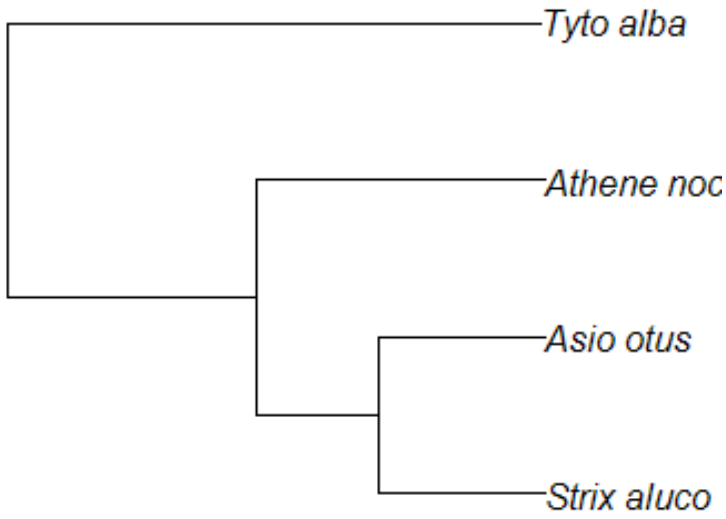


- 进化树数量随物种数增加的变化 自 A.Stamatakis 2007

建立进化树的软件

- PHYLIP 距离法、极大似然法等
- PAUP* 最大简约法、极大似然法、距离法等
- MrBayes 贝叶斯法
- PHYML 极大似然法（大样本量）
- RAMxL 极大似然法 （大样本量）
- MEGA 距离法、极大似然法等

进化树的表示 :Newick格式



- Newick格式，用于RAMxL, PHYLIP, PhyloCom等等

owls((Strix_aluco:4.2,Asio_otus:4.2):3.1,Athene_noctua:7.3)
:6.3,Tyto_alba:13.5);

NEXUS格式

- NEXUS 在newick格式的基础上增加了一些信息，NEXUS格式用于PAUP*, Mesquite, MacClade

```
Begin trees;
tree winteraceae =
owls(((Strix_aluco:4.2,Asio_otus
:4.2):3.1,Athene_noctua:7.3):6.3
,Tyto_alba:13.5);
End;
```

4.1 距离法

距离法

- 依据进化模型，首先计算各序列两两间的距离
- 采用UPGMA或WPGMA方法，进行聚类分类。但是这类方法产生错误率较高。随着新算法的不断提出，以聚类分析为基础的算法已经很少使用了。
- 当前流行的构建系统树的方法为：
- 极大似然法，最大简约法和贝叶斯法等。
- 不过在距离法中值得提及的有以下两种方法：

最少进化法与邻位法

- 最少进化法（Minimum Evolution ME）,计算每对序列间的遗传距离，并使进化树的总遗传距离最小。
- 而可能的进化树随着物种数的增加而爆炸式的增长，只能借助更好的搜索方法。
- 邻位法（Neighbour Joining）可以获得与最少进化法非常接近甚至一致的结果。

4.2 极大似然法

极大似然估计与极大似然法

- “在进化速率可变的假设下， 最大简约法略差于转换距离法和邻接法的结果，最大似然法的结果最优。”

——钟扬 《简明生物信息学》

什么是极大似然(Maxim likelihood)?

- 问题一，硬币的两面一致(正反面出现的概率均等），投硬币投掷100次，其中21次为正面，求该事件发生的概率.

$$\Pr[H=h]=\binom{n}{h}q^{h}(1-q)^{n-h}$$

$$\Pr[H=21]=\binom{21}{100}(\frac{1}{2})^{21}(1-\frac{1}{2})^{100-21}$$

问题二

- 问题二，硬币的两面不一致(正反面出现的概率不相等），投硬币投掷100次，其中21次为正面，这枚硬币正面出现的概率为多少时最有可能产生上述结果？

$$\Pr[H = h] = \binom{n}{h} q^h (1-q)^{n-h}$$

- 即已知n, h如何求 $\hat{q}$

似然函数定义

- 定义似然函数为

$$L(q) = \Pr[H = h] = \binom{n}{h} q^h (1-q)^{n-h}$$

- 两边取似然函数的对数，以便求其极值

$$\log[L(q)] = \log\binom{n}{h} + h\log q + (n-h)\log(1-q)$$

极大似然估计

- 取极大值时，似然函数对 θ 的导数为0

$$L'(q)=\frac{\partial \log[L(q)]}{\partial q}=\frac{h}{q}-\frac{n-h}{1-h}$$

- 当 $q=\frac{h}{n}$ 时，似然函数的值最大。
- 我们称 $\hat{q}=\frac{h}{n}$ 为 θ 的极大似然估计。

进化树的极大似然估计

- 假定序列是从一条碱基进化而来（拥有共同祖先）
- 给定一定的进化模型后，什么样的拓扑结构，什么样的枝长，什么样的进化模型参数最有可能产生出当前各序列的碱基差异？
- 极大似然估计

$$L(t,q) = \Pr(Data \mid t,q)$$

两条序列的极大似然估计

- 最简单的情况，碱基替换模型为JC69，两条序列
- 似然函数为

$$L(d)=\prod_{j=1}^{l} p_{s_1^j} P_{s_1^j s_2^j}\left(-\frac{4d}{3}\right)$$

- d为每个位点替换的次数；$P_{s_1^j s_2^j}$ 为原有位点为x碱基，而被替换为y碱基的概率
- $p_{s_1^j}$ 为第i个碱基在平衡状态的概率

多条序列的极大似然估计

- 似然率与进化模型、进化树的结构，每个枝长上的碱基替换个数密切相关。这些参数包括：每个位点的进化模型（GTR）参数, 每个节点的枝长。
- 简化处理：
- 1 假定每个位点都拥有一致的进化模型
- 2 每个位点的进化速率都为 μ
- 为了不过度简化，引入 ρ 作为进化速率因子

多条序列的极大似然估计

- 观察到某个位点当前替换格局的概率

$$\Pr[D_j \mid t, M, r_j], j = 1,...,l$$

- 所有位点当前替换格局的概率

$$L(t, M, r \mid D) \equiv \Pr[D \mid t, M, r] = \prod_{j=1}^{l} \Pr[D_j \mid t, M, r_j]$$

- 如果给定进化树（拓扑结构和枝长），给定进化模型，可以计算当前碱基替换格局出现的概率。
- 如果给定当前的碱基替换格局，可以估算进化树

计算机中的实现

- 计算机中寻找极大似然树按照如下思路：
- 给出所有的进化树，计算每一棵树的似然值。
- 但是由于可能的进化树太多，难以实现。
- 为此，人们发展出启发式搜索和进化树重排，以便快速发现满足需求的进化树

启发式搜索

- 由于进化树的数目太多，难以逐一进行计算，因此需要较为巧妙的算法寻找满足需求的进化树。
- 物种逐步添加（stepwise addition ） PHYLIP
- 星状分解法 （star decomposition） MOLPHY
- 邻位法 （neighbor joining ） 等

启发式搜索-逐步添加

- 在n个物种中随机挑选3种，建立极大似然树
- 再从剩余的n-3个种中随机挑选一个，随机加到上一步树的每个节点之间，挑选新树中似然值最高的。相应的分支称为插入分支（insertion branch）。
- 如此下去，每一步都计算2k-3个树，最终要计算的进化树的数量为

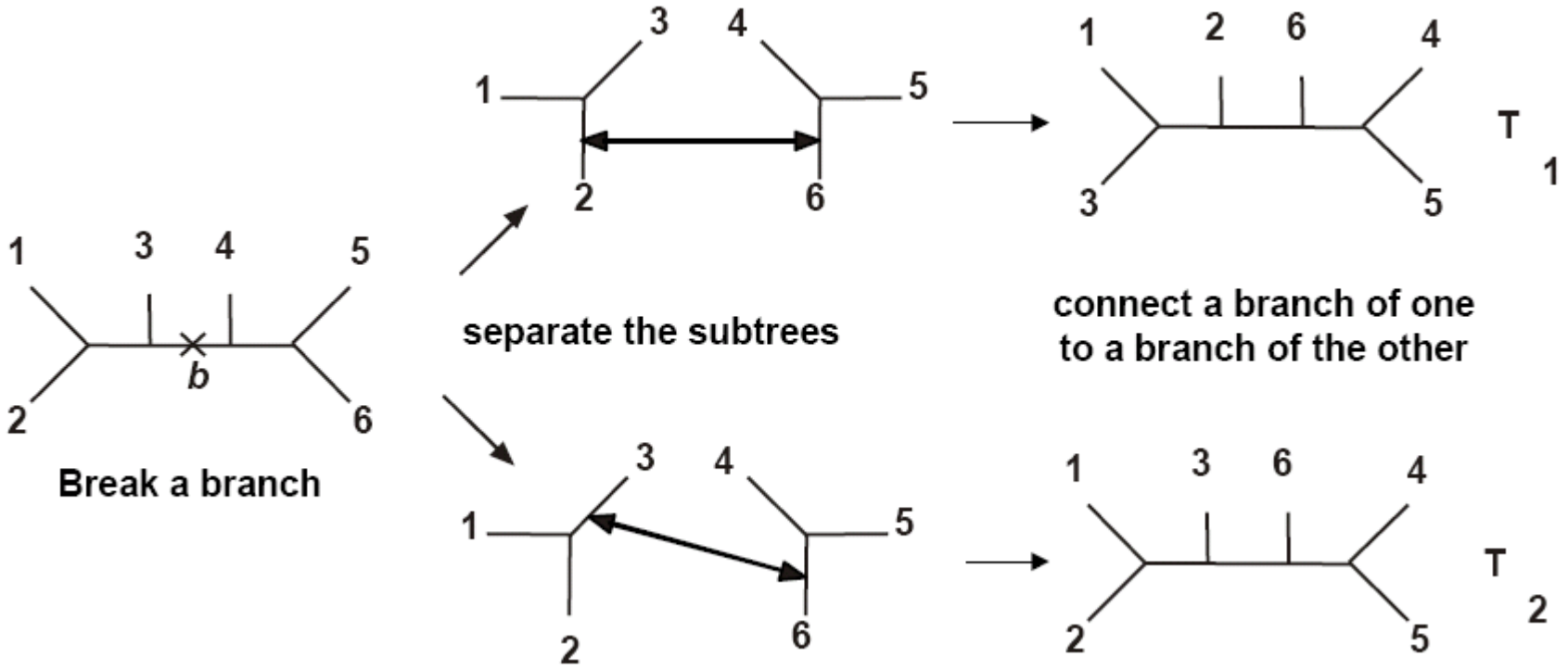
$$\sum_{i=3}^{n}(2i-5)=(n-2)^2$$

- 启发式搜索有时仍然难以找到进化树的极大似然估计，此时人们提出，对启发式搜索到的进化树进行重排，以提高搜索的精度。

进化树重排（Branch Swapping）

- 进化树按照一定的规则重排,重排后计算每棵树的似然值，并保留似然值最大的，继续重排，直到找不出更大的值为止。此时得到树称为局部最优树（locally optimal tree）
- 局部最优树与实际最优树相差多远取决于数据本身。
- NNI (Nearest neighbour interchange)
- SPR (sub-tree pruning and regrafting)
- TBR (Tree Bisection and Reconection)

TBR Tree-bisection reconnection



- 该方法包括了NNI和SPR 自 ***Nei and Kumar 2000***

寻找全局最优树

- 为了寻找全局最优树，往往从多个随机序列开始，进行多次启发式搜索，每次启发式搜索得到相应的局部最优树。
- 在局部最优树中挑选最为似然值最高的，从而保证尽最大可能发现全局最优树。

R软件+PHYML

ape程序包可以驱动PHYML软件，建立极大似然树

```
phymltest(seqfile, format = "interleaved",
  itree = NULL, exclude = NULL, execname,
  path2exec = NULL)
```

参见 E. Paradis 2006

Analysis of Phylogenetics and Evolution with R

4.3 贝叶斯法

后验概率

- 现有一袋子装了100个球分为黑白两种，如何得知其中白球或黑球的比例？
- 解决办法（1）全拿出来，查看一下 （2）随机抽取
- 假设白球的真实概率为p， 则在抽取n次之后，得到a个白球，b个黑球的概率为：

$$f(a,b \mid p) = p^a (1-p)^b \binom{a+b}{a}$$

- Probability mass function

解决方法

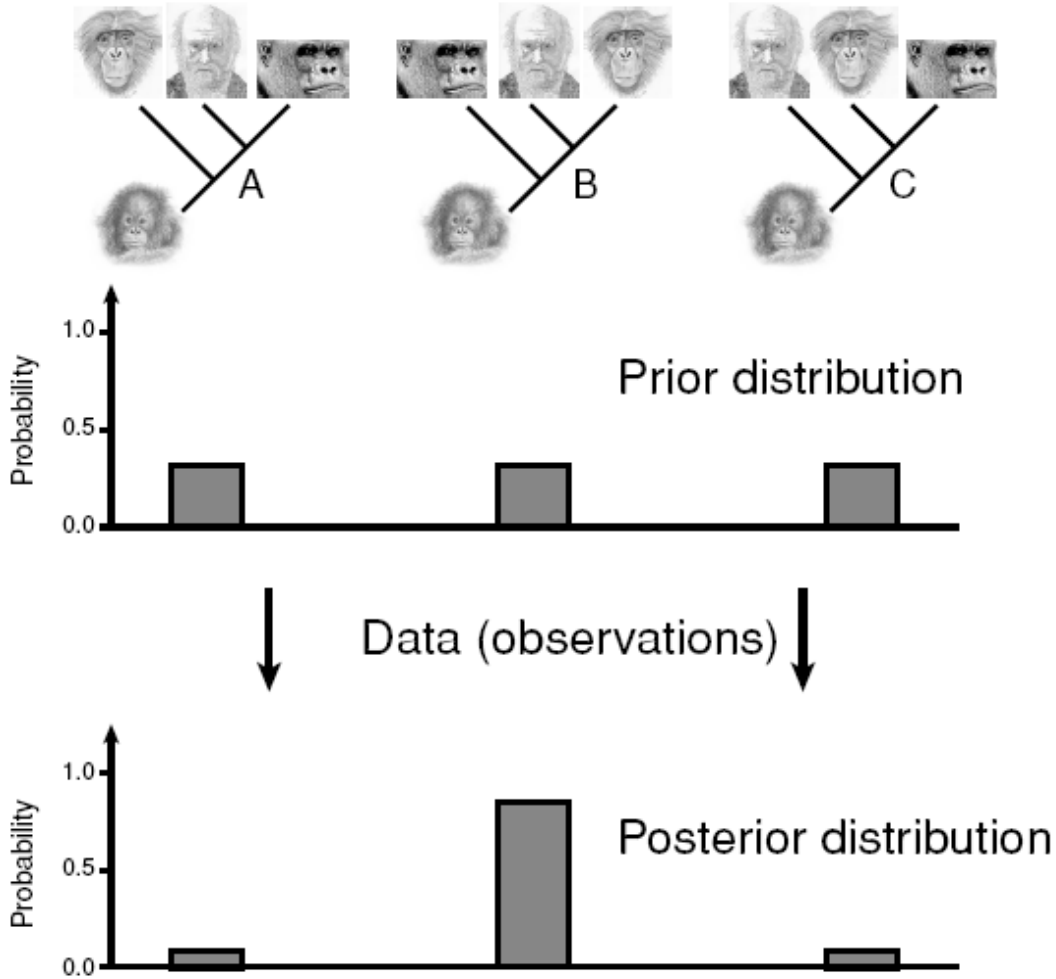
- 如果p已知，则 该函数可以计算以上事件出现的概率。
- 当前p未知，如何p进行估计？
- 解决办法： 抽样，获得p的分布，直到p趋于一个稳定值为止。
- 连续的数据 probability density function
- 离散的数据 probability mass function

Posterior probability distribution

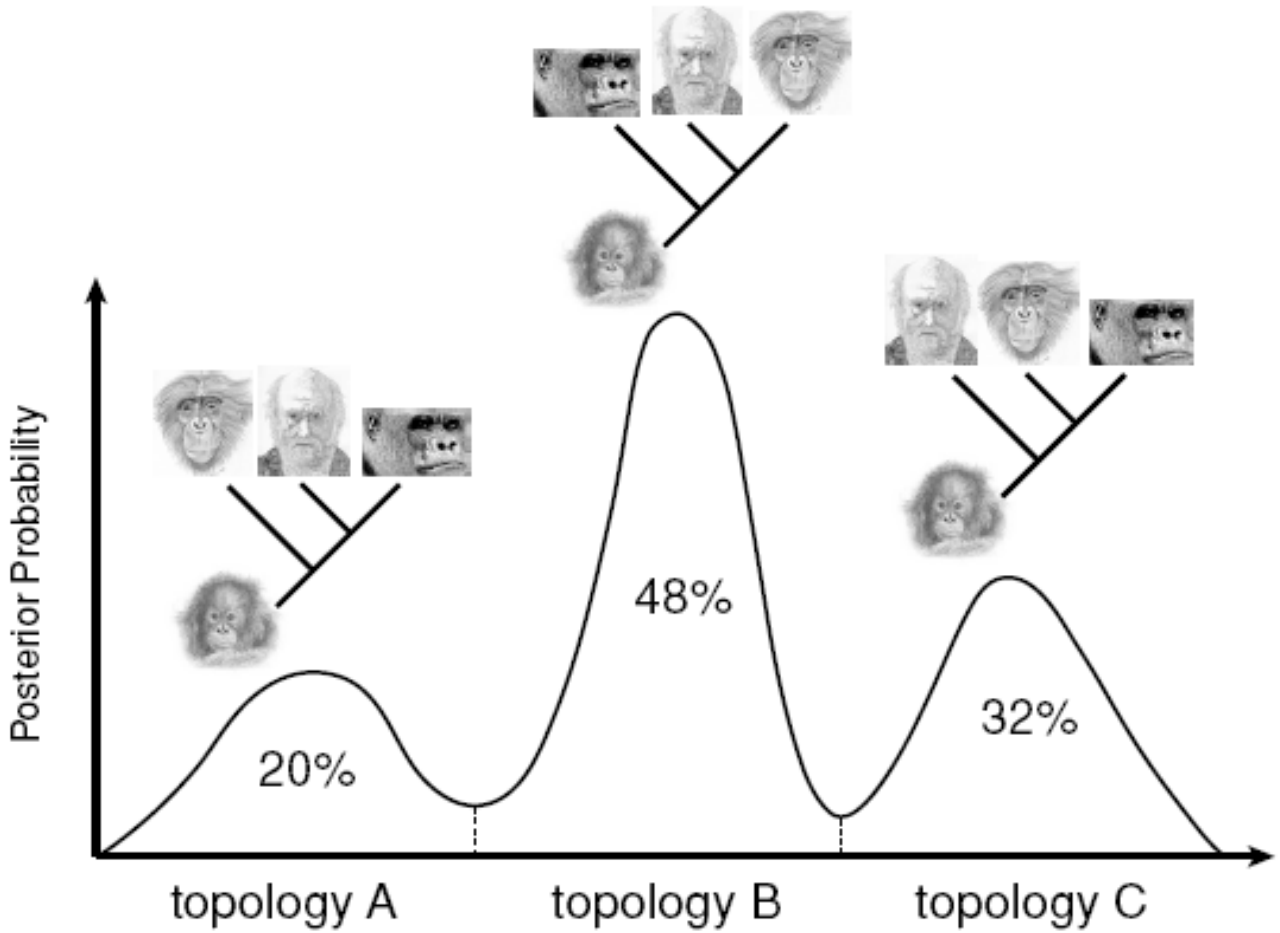
- 贝叶斯公式

$$f(p \mid a,b) = \frac{f(p)\, f(a,b \mid p)}{f(a,b)}$$

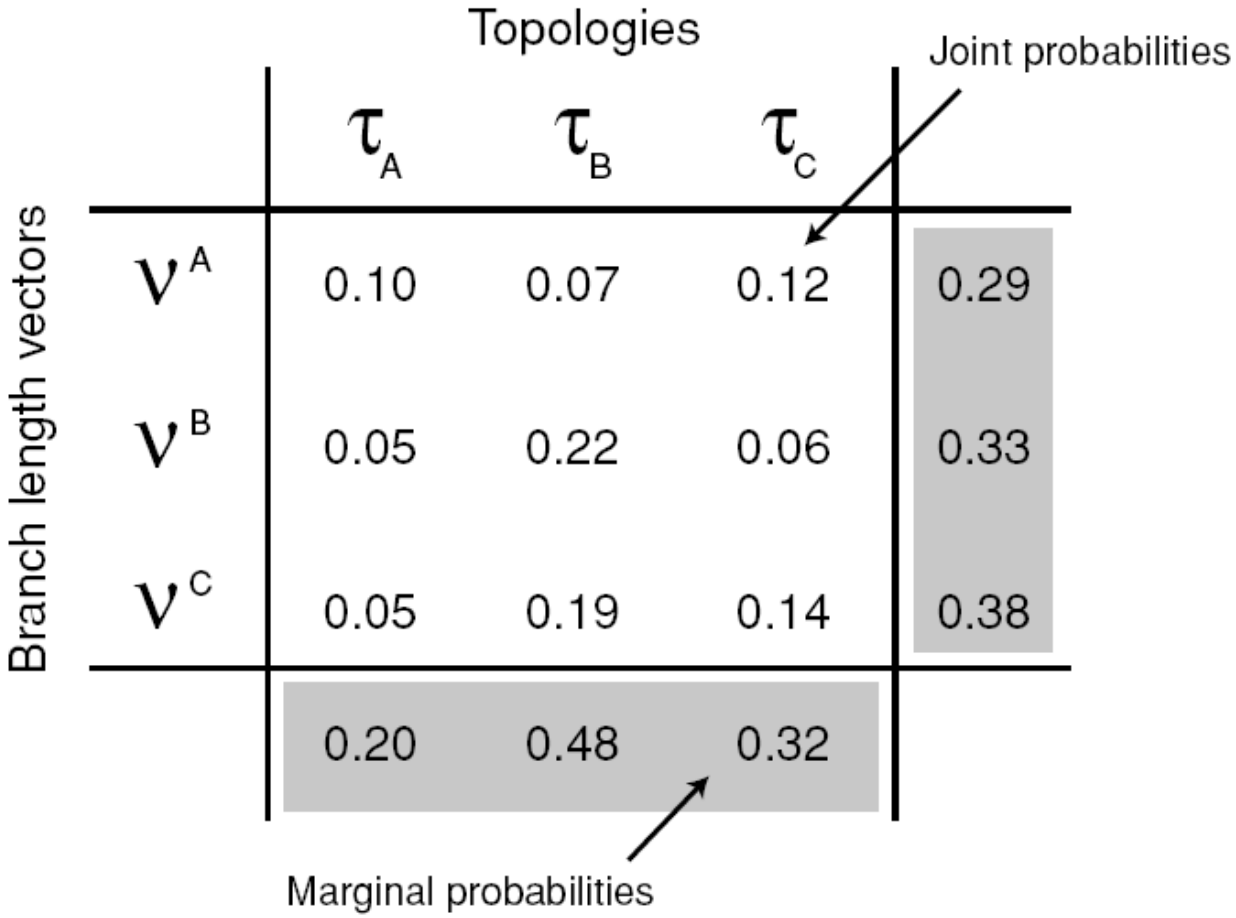
$$f(a,b) = \int_0^1 f(p)\, f(a,b \mid p)$$



- 在不知道树的概率时，首先假设每棵树的可能性都是相等的，将现有序列信息和进化模型参数代入贝叶斯公式计算每棵树的可能性，取概率最大者。（自 Ronquist F. et al. 2008)



- 三个系统树的拓扑结构分布在三个区间
- 每棵树的位置受到拓扑结构及枝长的影响。(自 Ronquist F. et al. 2008)



- 边缘概率(marginal probabilities) 联合概率(Joint probabilities)
- (自 Ronquist F. et al. 2008)

进化树及其参数的概率

1 后验分布

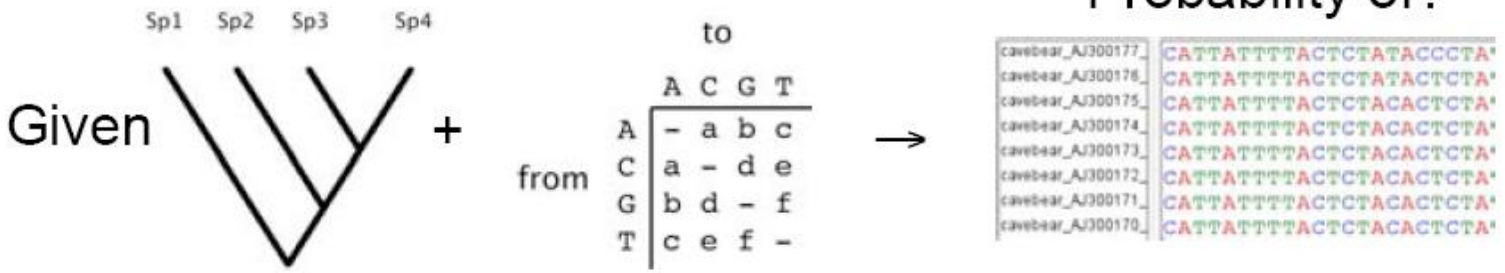
$$\Pr(\text{进化树},\text{参数}|\text{序列})$$

2 先验分布（可计算）　　3 likelihood（可计算）

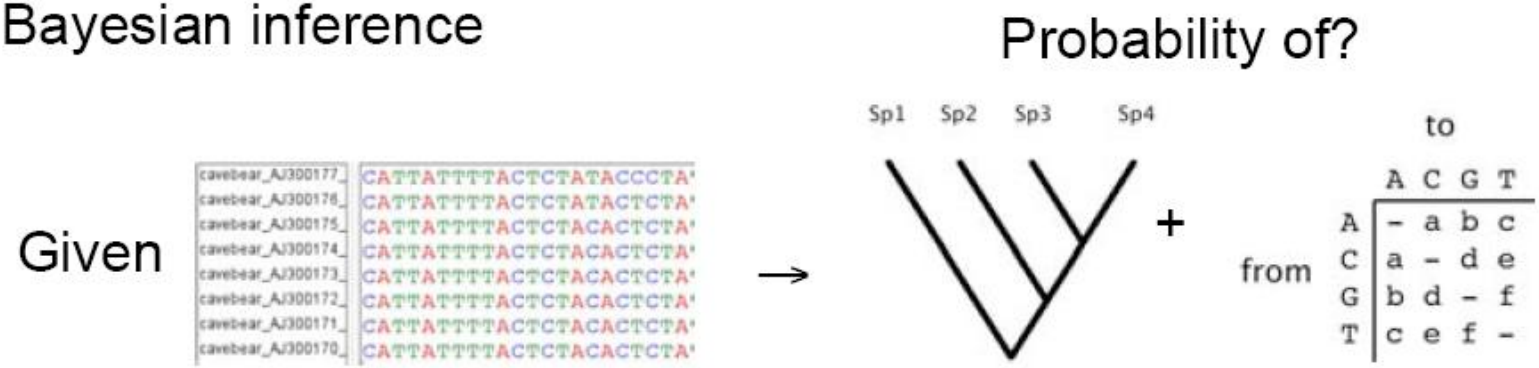
$$= \frac{\Pr(\text{进化树},\text{参数})\Pr(\text{序列}|\text{进化树},\text{参数})}{\Pr(\text{序列})}$$

4 对所有边缘概率求和
（如何计算？）

Maximum likelihood



Bayesian inference



- 极大似然法 指定树的结构和进化模型，计算序列组成的概率
- 贝叶斯法 给定序列组成， 计算进化树和进化模型的概率（自 Simon Ho）

进化树的后验分布计算

$$f(T,\theta|\xi,\mathbf{X}) = \frac{f(\mathbf{X}|T,\theta)f(\theta|\xi)f(T|\xi)}{\int_T \int_\theta f(\mathbf{X}|T,\theta)f(\theta|\xi)f(T|\xi)\,d\theta dT}.$$

自 Bruce Rannala

- 将进化树及模型参数整合到后验分布中，如何计算？

碱基替换：时间齐性马尔科夫过程

- Time-homogeneous stationary Markov process
- 1 任何一个位点上由碱基i变化为碱基j的变化率与碱基i之前的碱基无关。（马尔科夫性质）。
- 2 变化率不随时间而变化（齐次性）。
- 3 碱基AGCT相对频度处于平衡状态。

MCMC-Metropolis-Hastings 算法

- 对于系统发育推断的问题，难以得到各概率的解析解。
- 蒙特卡罗马尔科夫链使得我们无需计算分母，即可近似的得到后验分布,且经历的代数越多，结果越精确。
- 现有的解决办法，将进化树（拓扑结构与进化模型参数)转换为马尔科夫链，待马尔科夫链收敛于后验分布即可。
- Markov Chain Monte Carlo sampling

MCMC-Metropolis-Hastings 算法

- MCMC-当前Bayesian进化树构建中最主要的算法 Metropolis-Hastings 算法

1 以任意一棵树T_i开始

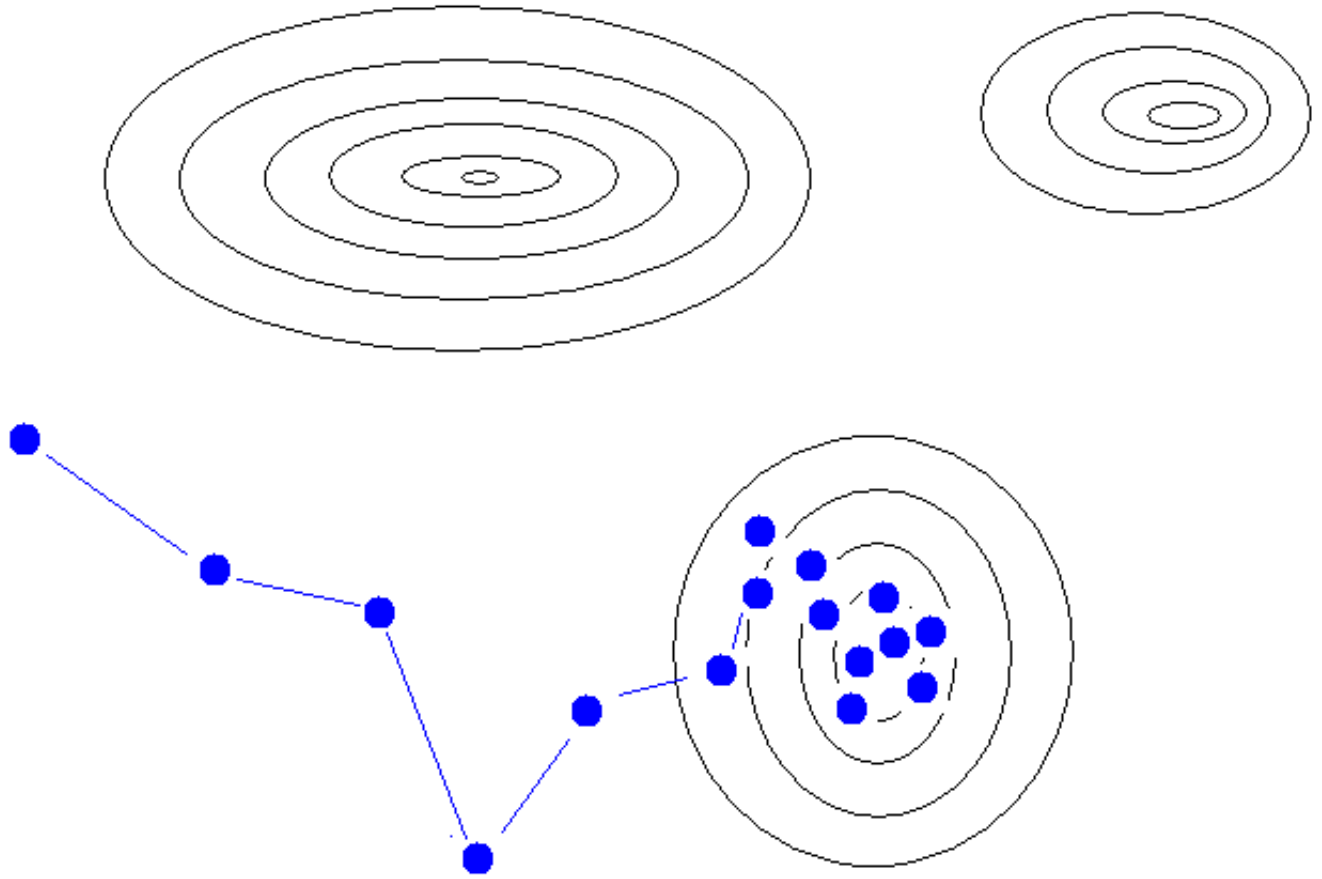
2 随机挑选一个与T_i相邻的树T_j.

3 计算两棵树（似然率*先验分布）的比值R $R = \frac{f(T_j)}{f(T_i)}$

3 如果$R \geqslant 1$，接受 T_j

如果$R<1$，在[0,1]之间取随机数k，若$k<R$，则接受T_j。否则拒绝接受 T_j,

5 回到第2步，继续…



- 马尔科夫链在局部的进化树参数空间停滞不前

MCMCMC （MC^3）

- 防止马尔科夫链在局部进化树参数空间停滞不前。
- 解决方法： MCMCMC （MCMC Metropolis Coupling）
- 设定热链，使该链空间变的较为平滑，令其取样点活动性更强，便于在不同的局部区域间跨越。
- 热链一旦找到后验分布更大的地点，冷链和热链的位置互换，继续搜寻，最终使冷链收敛于全局最优后验分布。

树的确定

- 链收敛后，每隔100代，保存一次进化树，经过1000000次后，保存了大量的进化树。
- 根据后验分布排序，保存最高的5%, 计算一致性树。

贝叶斯法建树：MrBayes软件

```
begin mrbayes;
set autoclose=yes nowarn=yes;
execute primates.nex;
lset nst=6 rates=gamma;
mcmc nruns=1 ngen=1000000 samplefreq=1000
   file=primates.nex1;
mcmc file=primates.nex2;
mcmc file=primates.nex3;
end;
```

4.4 最大简约法

简约分析-Parsimony

- 目的： 寻找一棵或多棵树，这些树满足在进化的过程中，碱基变化最少。
- 哲学原理：在解释某种现象时，如果某种推断所需的步骤最少，即认为这种推断是合理的。
- 思路：
- （1）计算给定的一棵树内，所有种与祖先不同的性状（碱基的变化）的总和。
- （2）给出所有可能的进化树，选出最小进化树。

碱基变化的计算

- 在进化速率缓慢时，简约法是最为有效的算法。
- 对于有n个种的二叉分枝树，包含n个末端节点，n-2个内部节点，2n-3个连接节点之间的分支。
- 设定 τ 为进化树的结构，则树的长度为

$$L(t) = \sum_{j=1}^{N} l_j$$

- N为位点数，l_j是位点的长度，即该位点性状变化数目

碱基总数变化的计算

- 一棵树内所有节点的变化总和

$$l_j = \sum_{k=1}^{2N-3} c_{a(k),b(k)}$$

$c_{a(k),b(k)}$ 是第k个位点从状态a到状态b的变化数

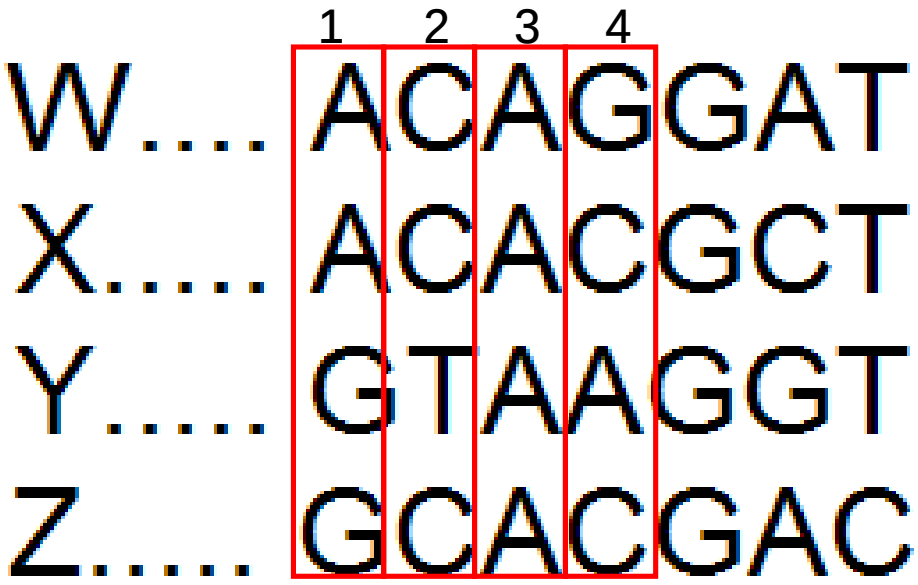
计算机如何实现？

- 思路
- 计算每个位点所有可能的进化树的总性状变化数。
- 对于DNA序列，需要计算的进化树的数目为

$$n = 4^{T-2}$$

- T为序列数目， 4即有四个碱基

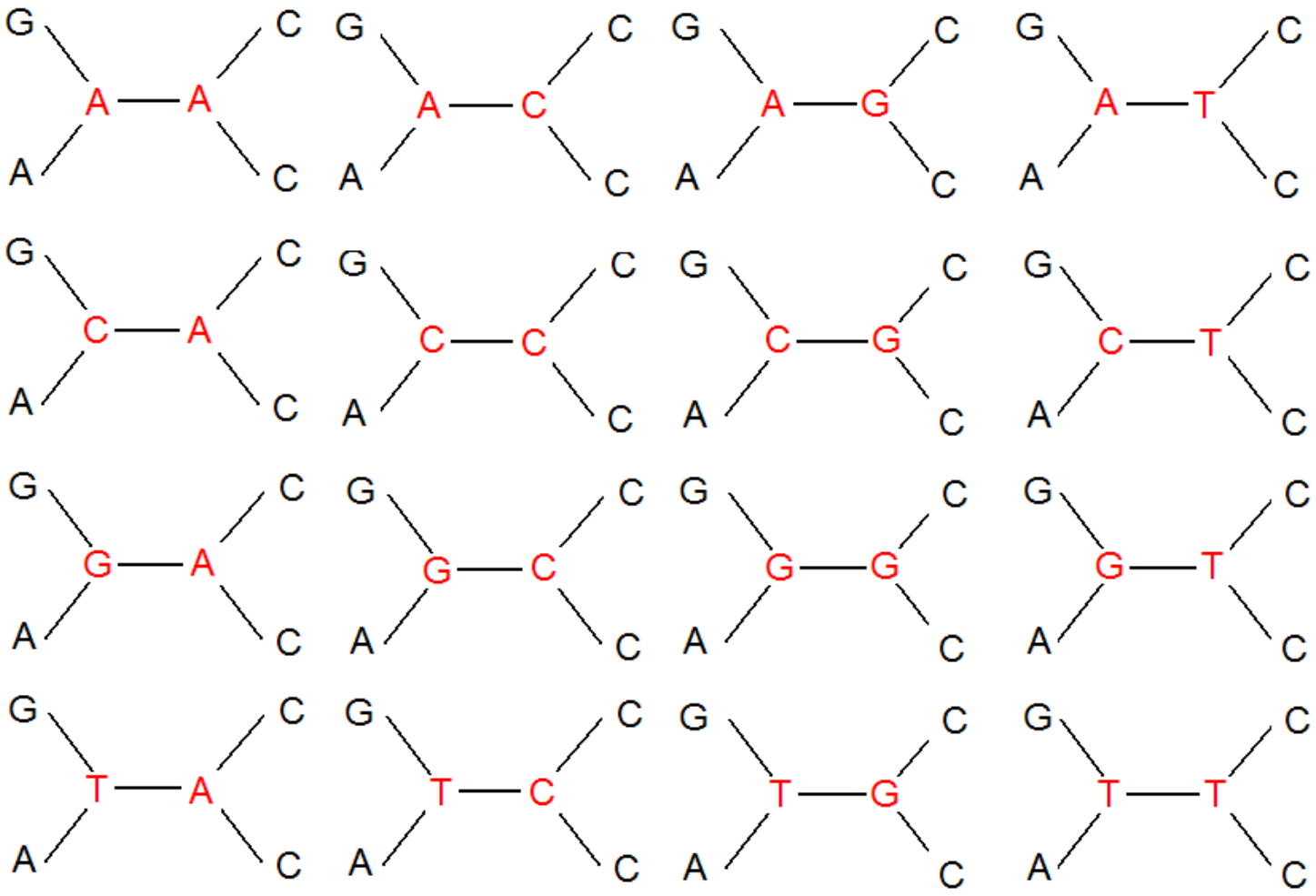
信息位点



- 第四个位点，在列出所有的可能替换中，变化数目有差异，则该位点为信息位点。

简约分析 举例

- 考虑第四个位点
- W（G）,X（C）,Y（A）,Z（C）
- 而这样的序列组成最可能是什么样的系统关系进化而来？
- 穷举法：计算所有可能的进化树，并给出当前种祖先的所有可能碱基组成，并计算其变化。
- 如何评估某棵树某个碱基位点变化的最小值？
- 如：计算 ((W,Y)(X,Z)) 第四位点碱基变化的最小值？



- 进化树((W,Y)(X,Z))内第四位点 四个物种祖先所有可能的碱基组成 W:G, X:C, Y:A, Z:C，挑选变化最少的

建树步骤

- 四个种的形成，首先要由共同祖先分化成两个种，而这两个种再分别演化成各自的后代。
- 最大简约法建树的步骤：
- （1）列出所有可能的进化树
- （2）计算基于每棵树每个位点的变化数，选择使该位点变化的总和最小的树
- （3）挑选出满足所有位点性状变化总和最小的树，即最大简约树

碱基变化的差异 Cost matrix

$$equal = \begin{array}{c} \begin{matrix} A & C & G & T \end{matrix} \\ \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \end{array} \quad unequal = \begin{array}{c} \begin{matrix} A & C & G & T \end{matrix} \\ \begin{bmatrix} 0 & 4 & 1 & 4 \\ 4 & 0 & 4 & 1 \\ 1 & 4 & 0 & 4 \\ 4 & 1 & 4 & 0 \end{bmatrix} \end{array}$$

- 考虑颠换和替换的速率不同，则矩阵从equal 变为 Unequal

实际的简约算法

- 实际的算法不会穷举如此多性状变化的可能，而是优化了一系列算法
- Sankoff 法：假设祖先性状是某一个分支的性状，计算总和，取最小值
- Fitch法：假设祖先性状是某个分支的性状，取并集，按照并集的数量寻求最大简约树。

启发式搜索

1 Exact Methods : branch and bound method

为简约步骤设定一个阈值，如果某个进化树超过该阈值，便不再该树的基础上继续搜索。

2 Approximate methods: Greedy algorithms 物种逐步添加

依据某种原则建立一棵树如（NJ树），并计算其性状变化的数目作为标准； 另建树，逐步添加物种，计算其变化的数目，舍弃不满足条件的树，如果该树的碱基变化数低于原来的树，则以当前的进化树作为新的标准。继续计算，直到找不出更优的树为止。

简约树重排

- 逐步添加物种仍难以找到最大简约树，为此需要对逐步添加物种找到的简约树进行重排：
- 需要对树进行重排 Branch Swapping 的方法
- Nearest- neighbor interchange (NNI)
- Subtree prunning and regraphting (SPR)
- Tree bisection and reconnection (TBR)

最大简约法实现：PAUP*软件

```
begin paup;
log start replace=yes file=model_log.txt;
set autoclose=yes criterion=parsimony
root=outgroup storebrlens=yes increase=auto;
outgroup SP4Ranunculus;
hsearch addseq=random nreps=1000 swap=tbr hold=1;
savetrees file=modeltrees.tre format=altnex
brlens=yes;
log stop;
END;
```


五 树的可信度 Bootstrap

什么是Bootstrap?

- Bootstrap
- 欲估算某一值的可信度（如标准误Standard Error），则在原有样本的基础上进行有放回的抽样，生成一系列伪样本，对伪样本的计算结果重新生成某一个值的分布，从而对原有样本计算的值的可信度进行估算。
- Bootstrap是在样本分布难以获得情况下，进行的一种抽样方法。

树的可信度 Bootstrap

- 对系统树的节点进行Bootstrap检验是J. Felsenstein于1985年引入的。
- 假设当前比对好的30个种的DNA序列，长度为550bp,已经用极大似然法建立进化树，如何对这30个种的系统关系的可信度进行Bootstrap检验？

进化树Bootstrap 1000次的过程

- 1 随机并有放回的抽取550个bp的任意一个位点，直到抽取到550个位点为止。对新形成的序列建立进化树。
- 2 重复1000次，有放回的随机抽取该550个位点，形成1000套的新的序列（注意：每套序列有550个位点），对以上序列分别建立系统树。即共建立1000棵进化树。
- 3 计算这1000棵树与初始树一致性，保留节点支持率>50%的进化树节点。

Bootstrap的不足

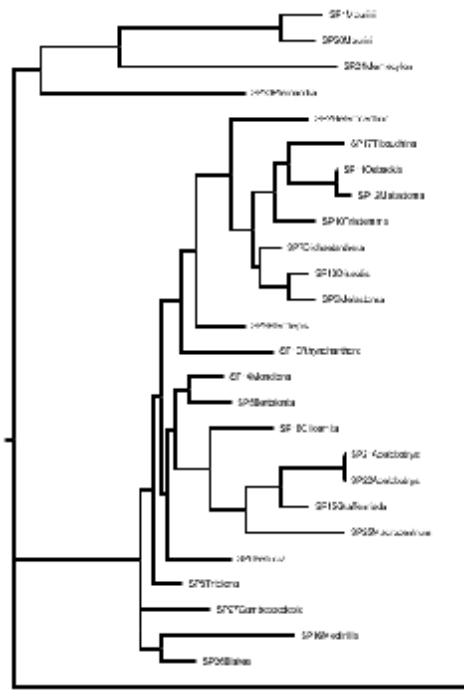
- Bootstrap 并不能给出所选模型的可信度
- 即如果进化模型选择的不合理，Bootstrap值再高也是没有意义的。

六 分子钟

6 分子钟

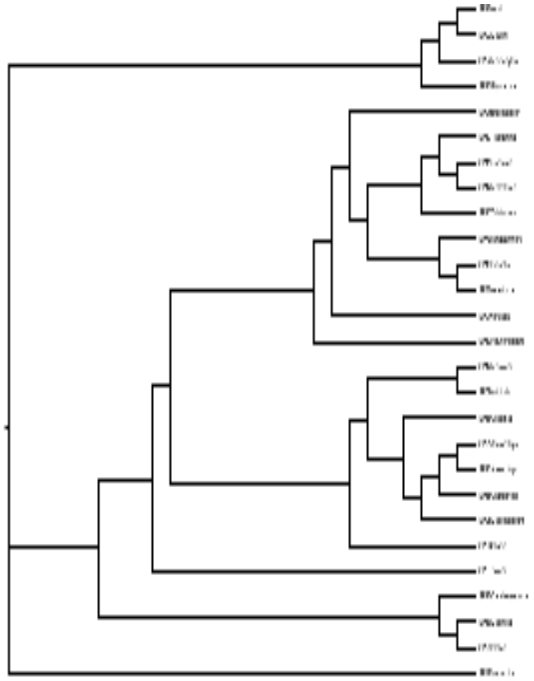
- 物种碱基组成的差异随着分化时间的增加而增加。
- 但是真实情况下，碱基替换与似然率大部分并不满足分子钟的假设。
- 进化树各末端的节点到根的长度不同，此时进化树就分成了Ultrametric树和Non Ultrametric树。

Ultrametric and non ultrametric



Non Ultrametric

（枝长为似然率或碱基替换数）



Ultrametric

时间校正

如何校正时间？

- 不满足分子钟假设时，需要对分化时间进行校订则需采用如下方法:
- 非参数速率平滑 NPRS (Non Parametric rate smoothing)
- 罚分似然法 Penalized likelihood
- 贝叶斯法
- 软件：r8s , ape, BEAST, Multidivtime等

r8s的主要命令

blformat 进化树的基本信息

mrca 为节点定名

fixage 设定节点的分化时间

constrain 限定节点的分化时间

divtime 分化时间估算

showage 显示分化时间和分化速率

describe 显示进化树及树的说明

set 设定参数

calibrate 时间校对

profile 从多个树中提取某个节点的信息

rrlike 检验进化速率

ape校对分子钟的函数

- NPRS的实现

```
chronogram(phy, scale = 1, expo = 2,
  minEdgeLength = 1e-06)
```

Penalized likelihood的实现

```
chronopl(phy, lambda, age.min = 1,
  age.max = NULL,node = "root", S = 1,
  tol = 1e-8,CV = FALSE, eval.max = 500,
  iter.max = 500, ...)
```


谢 谢！
敬请指正！