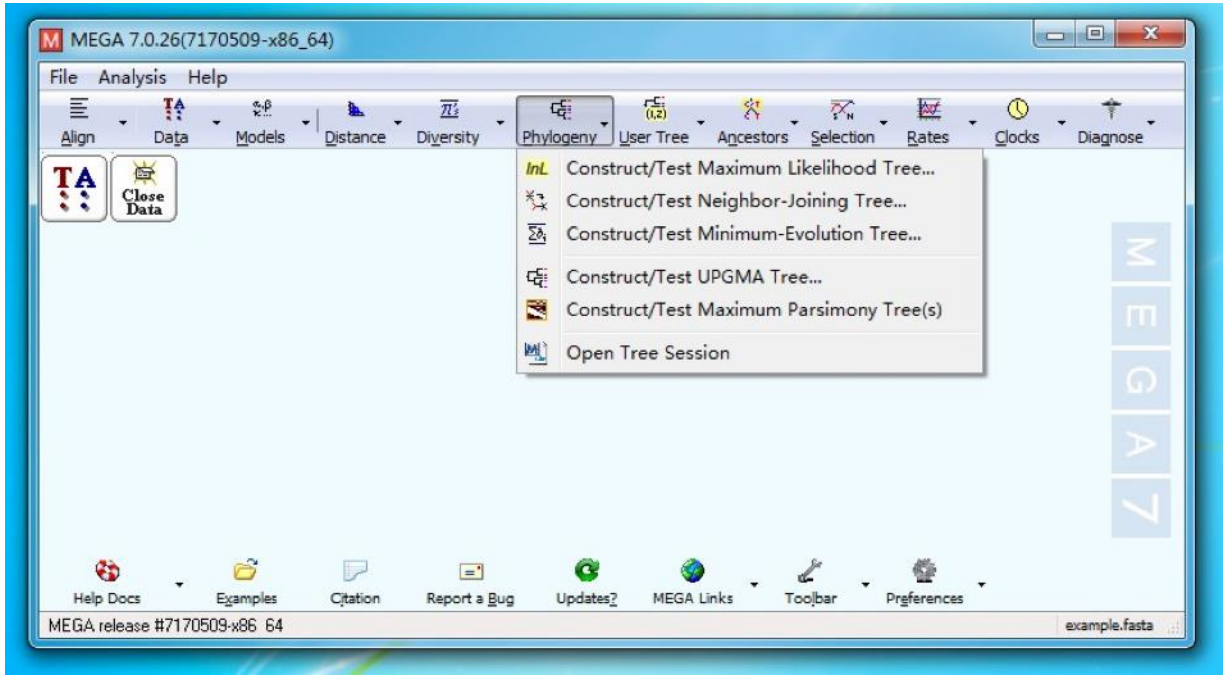
基于生成模型的系统发生推断

Phylogenetic inference within generative framework

G16: **曹智杰** 房巧 孙元强 温正扬
G06: **唐小鹿** 徐可言 黄钰 郭歌

课堂回顾



- 基于概率模型的方法具有更明确的生物学解释，在很多情况下更为可靠，可扩展性也较强，我们将对其原理进行介绍

提要

1. 描述系统发生的概率模型
2. 系统发生树的最大似然估计
3. 系统发生树的贝叶斯估计
4. 应用与比较

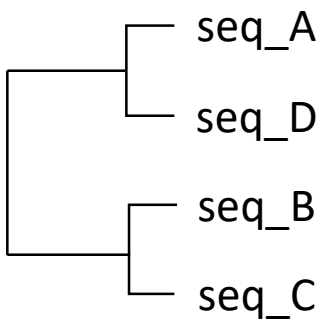
提要

1. **描述系统发生的概率模型**
2. 系统发生树的最大似然估计
3. 系统发生树的贝叶斯估计
4. 应用与比较

起点：多序列比对 (Multiple Sequence Alignment)

seq_A	A	G	C	A	A	T	T	G
seq_B	A	G	C	G	A	G	T	G
seq_C	A	C	C	G	A	C	T	G
seq_D	A	G	C	A	A	C	T	G

终点：系统发生树 (Phylogenetic Tree)



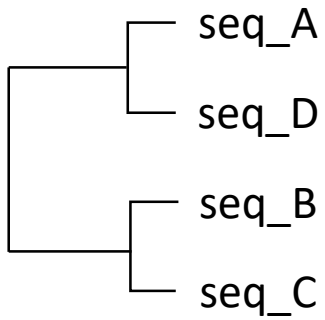
系统发生的生成模型

- 生成模型描述的是：在特定的演化历史（系统发生树）下，产生当前观测序列的概率

- 数学表示：$p(MSA|tree)$

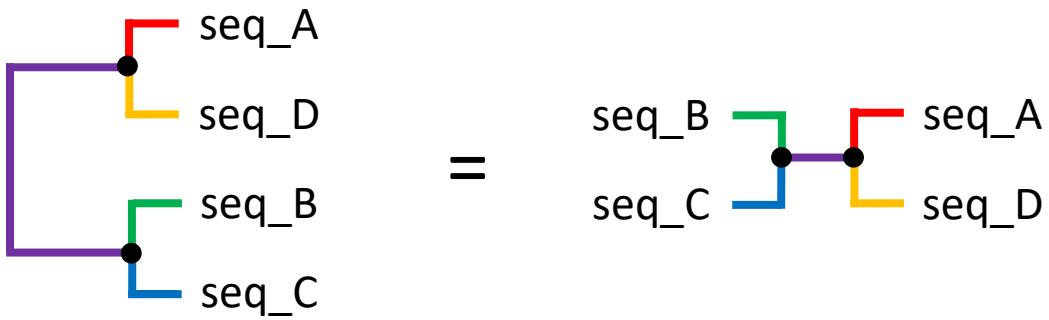
- **基于概率模型的建树方法都依赖于$p(MSA|tree)$的计算**

seq_A	A	G	C	A	A	T	T	G
seq_B	A	G	C	G	A	G	T	G
seq_C	A	C	C	G	A	C	T	G
seq_D	A	G	C	A	A	C	T	G



生成模型中的系统发生树

- 我们这里主要讨论无根树，有根树的计算思想也是一样的



- 它包括两个方面：
  - 树的拓扑结构
  - 树中每个树枝的长度（演化距离）
- 叶子结点是现在观测到的序列，祖先结点是某祖先物种的序列
- 每个树枝代表了一段演化过程，是概率计算的基本单位

演化枝的概率描述

- 简单情况：假定树枝两端的序列都是已知的

seq_A	A	G	C	A	A	T	T	G
seq_D	A	G	C	A	A	C	T	G

seq_A —— l —— seq_D

- 这一枝上的概率可以写为 $p(seq_A|seq_D;l) = p(seq_D|seq_A;l)$

  表示在演化距离l下，从其中一条序列转变为另一条序列的概率

- 位点独立假定：

碱基替换模型

$$p(seq_A|seq_D;l) = \prod_i p\left(seq_{A_i}|seq_{D_i};l\right)$$

碱基替换模型

- 替换概率矩阵 P ：

  每一行的和为1

	A	G	C	T
A	0.8	0.05	0.05	0.1
G	0.05	0.8	0.1	0.05
C	0.05	0.1	0.8	0.05
T	0.1	0.05	0.05	0.8

- 演化过程可以看作突变逐步累积的过程，随着演化距离的增加，序列可能发生的变化也逐步增加，比如上面的替换概率矩阵，在更长演化时间下，可能变成：

	A	G	C	T
A	0.6	0.1	0.1	0.2
G	0.1	0.6	0.2	0.1
C	0.1	0.2	0.6	0.1
T	0.2	0.1	0.1	0.6

碱基替换模型

- 替换概率矩阵可以写为**演化距离l**的函数 $P(l)$
- 通常采用连续马尔科夫过程来描述：
  - $P(0) = \boldsymbol{I}$
  - $\frac{dP(l)}{dl} = Q \cdot P(l)$，即$P(l) = e^{Ql}$
- 其中，Q是碱基替换概率随时间的变化率矩阵，

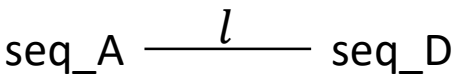
  比如：

	A	G	C	T
A	-0.1	0.02	0.02	0.06
G	0.02	-0.1	0.06	0.02
C	0.02	0.06	-0.1	0.02
T	0.06	0.02	0.02	-0.1

小结

- 到此为止，给定枝两端的序列、枝的长度，我们已经能计算这一条枝的概率
- 实际计算一下枝长为5或10时，这条枝的概率分别是多少：

seq_A	A	G	C	A	A	T	T	G
seq_D	A	G	C	A	A	C	T	G



假定替换速率矩阵Q如下：

	A	G	C	T
A	-0.1	0.02	0.02	0.06
G	0.02	-0.1	0.06	0.02
C	0.02	0.06	-0.1	0.02
T	0.06	0.02	0.02	-0.1

小结

- 首先，根据$P(l) = e^{Ql}$计算$P(5)$和$P(10)$，结果如下所示：

$P(l = 5)$

	A	G	C	T
A	0.64	0.08	0.08	0.20
G	0.08	0.64	0.20	0.08
C	0.08	0.20	0.64	0.08
T	0.20	0.08	0.08	0.64

$P(l = 10)$

	A	G	C	T
A	0.46	0.14	0.14	0.26
G	0.14	0.46	0.26	0.14
C	0.14	0.26	0.46	0.14
T	0.26	0.14	0.14	0.46

- 然后按照位点独立假设，计算给定替换矩阵下，观测到这两条序列的概率：

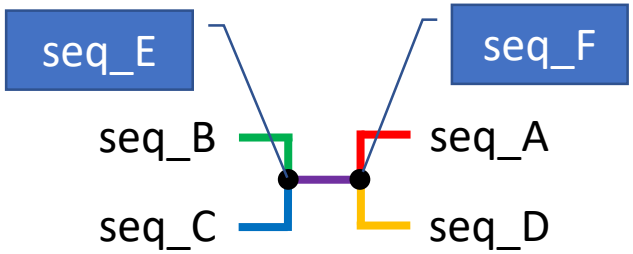
$$p(seq_A | seq_D; l = 5) = 0.64^7 \times 0.08 = 0.0035$$

$$p(seq_A | seq_D; l = 10) = 0.46^7 \times 0.14 = 0.0006$$

seq_A	A	G	C	A	A	**T**	T	G
seq_D	A	G	C	A	A	**C**	T	G

系统发生树的似然计算

给定树结构和枝长，如何计算似然函数 $p(MSA|tree)$？



下面用粗体的**A-F**表示6条序列中的某个位点

$$p(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D}|tree) = \sum_{\boldsymbol{E},\boldsymbol{F}\in\{A,G,C,T\}} p(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D},\boldsymbol{E},\boldsymbol{F}|tree)$$

$$= \sum_{\boldsymbol{E},\boldsymbol{F}\in\{A,G,C,T\}} p(\boldsymbol{F})p(\boldsymbol{A}|\boldsymbol{F};l_{\boldsymbol{AF}})p(\boldsymbol{D}|\boldsymbol{F};l_{\boldsymbol{DF}})p(\boldsymbol{E}|\boldsymbol{F};l_{\boldsymbol{EF}})p(\boldsymbol{B}|\boldsymbol{E};l_{\boldsymbol{BE}})p(\boldsymbol{C}|\boldsymbol{E};l_{\boldsymbol{CE}})$$

位点独立假定

系统发生树的似然计算

seq_A	A	G	C	A	A	**T**	T	G
seq_B	A	G	C	G	A	**G**	T	G
seq_C	A	C	C	G	A	**C**	T	G
seq_D	A	G	C	A	A	**C**	T	G

$$p(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D}|tree) = \sum_{\boldsymbol{E},\boldsymbol{F}\in\{A,G,C,T\}} p(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D},\boldsymbol{E},\boldsymbol{F}|tree)$$

$$= \sum_{\boldsymbol{E},\boldsymbol{F}\in\{A,G,C,T\}} p(\boldsymbol{F})p(\boldsymbol{A}|\boldsymbol{F};l_{\boldsymbol{AF}})p(\boldsymbol{D}|\boldsymbol{F};l_{\boldsymbol{DF}})p(\boldsymbol{E}|\boldsymbol{F};l_{\boldsymbol{EF}})p(\boldsymbol{B}|\boldsymbol{E};l_{\boldsymbol{BE}})p(\boldsymbol{C}|\boldsymbol{E};l_{\boldsymbol{CE}})$$

提要

1. 描述系统发生的概率模型
2. **系统发生树的最大似然估计**
3. 系统发生树的贝叶斯估计
4. 应用与比较

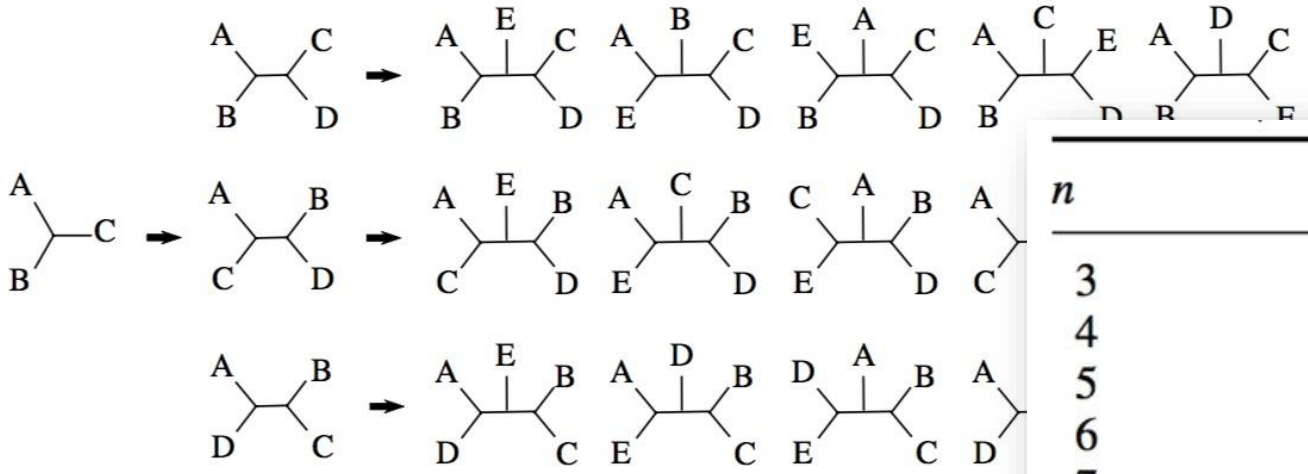
最大似然估计

$$tree^* = argmax_{tree}\, p(MSA|tree)$$

找出一棵最优的树（包括拓扑和枝长），使得似然函数$p(MSA|tree)$最大

naive的想法：遍历所有可能的树？

树结构的搜索空间



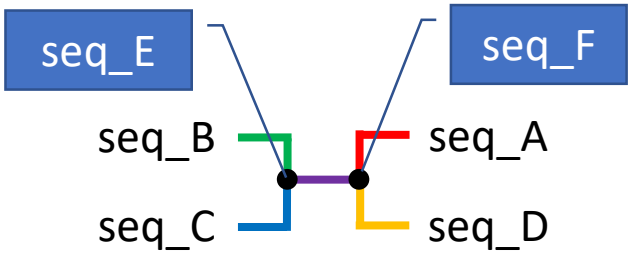
n	T_n
3	1
4	3
5	15
6	105
7	945
8	10 395
9	135 135
10	2 027 025
20	$\sim 2.22 \times 10^{20}$
50	$\sim 2.84 \times 10^{74}$

- n个物种可能的树结构有$1 \times 3 \times 5 \times \cdots \times (2n -$
  个

给定树结构，枝长的搜索空间也非常大（与枝数指数相关）

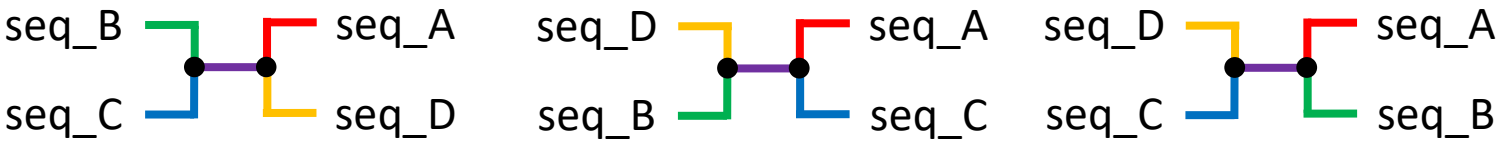
最大似然的启发式搜索

- 基于随机初始化的树，迭代改进系统生成树，两种改进树的方法：
  - 给定树结构，优化特定枝的长度（通用数值优化算法）



$$\mathcal{L}(\boldsymbol{l_{EF}}) = \sum_{E,F} p(B,C|E)p(E)p(F|E;\boldsymbol{l_{EF}})p(A,D|F)$$

  - 优化树结构



对于一条特定的边，与其相接的四个子树一共有三种拓扑，选择似然最高的一种

Guindon, Stéphane, and Olivier Gascuel. *Systematic biology*, 2003

提要

1. 描述系统发生的概率模型
2. 系统发生树的最大似然估计
3. **系统发生树的贝叶斯估计**
4. 应用与比较

贝叶斯估计

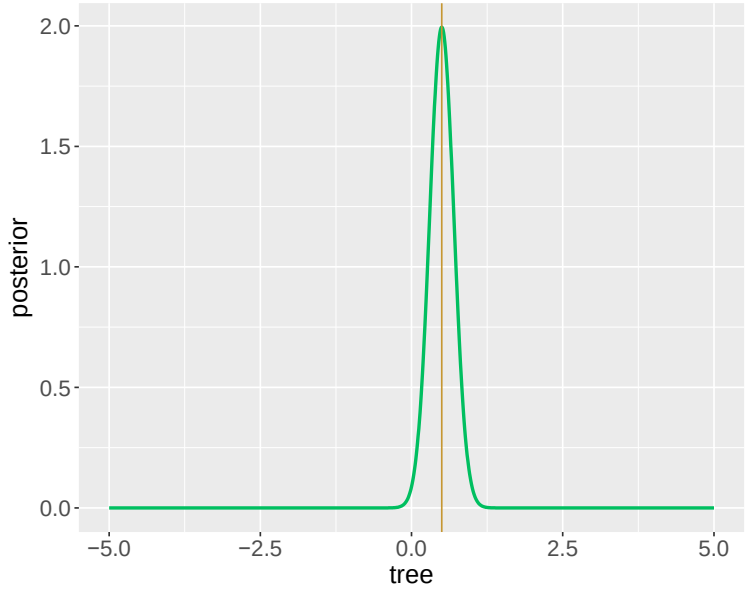
似然函数

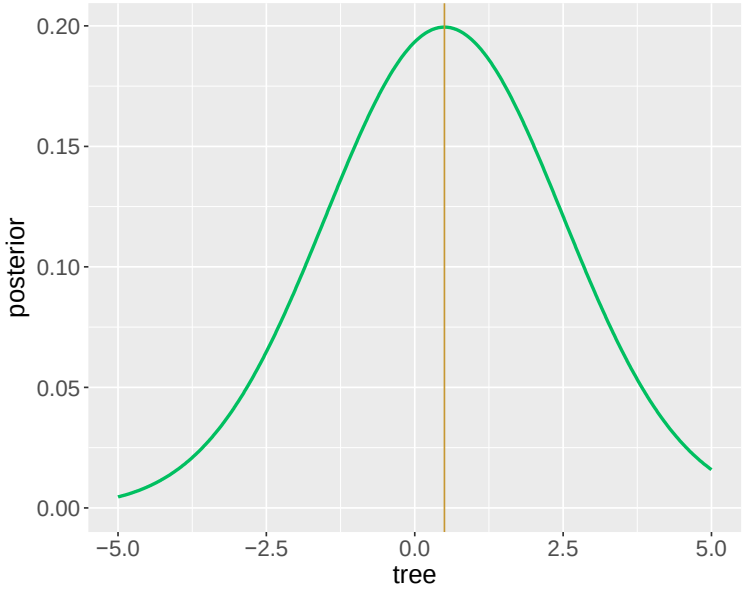
先验分布

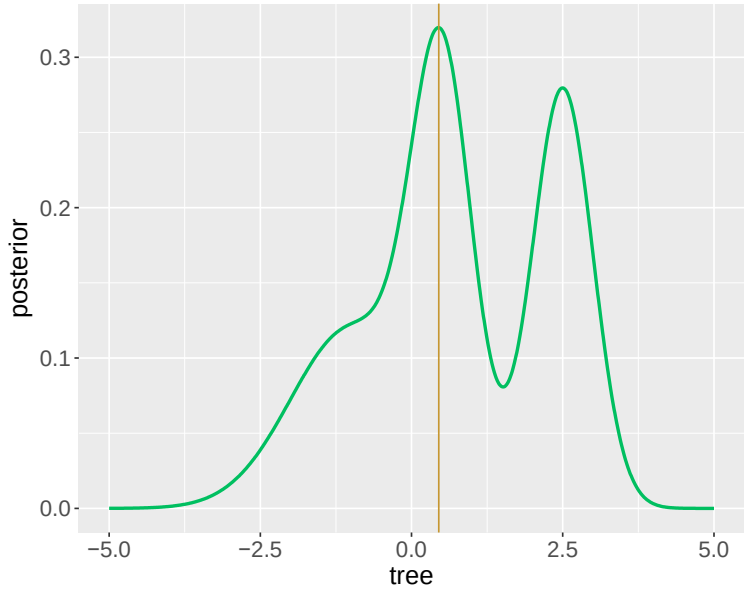
$$p(tree|MSA) = \frac{p(MSA|tree)\, p(tree)}{\sum_{tree} p(MSA|tree)\, p(tree)}$$

后验分布

贝叶斯估计求的是后验分布，而不是一个最优值







- 最大似然估计只能告诉我们哪个点是最优的
- 贝叶斯估计得到的后验概率分布能进一步给出树的可靠性

贝叶斯估计

似然函数

先验分布

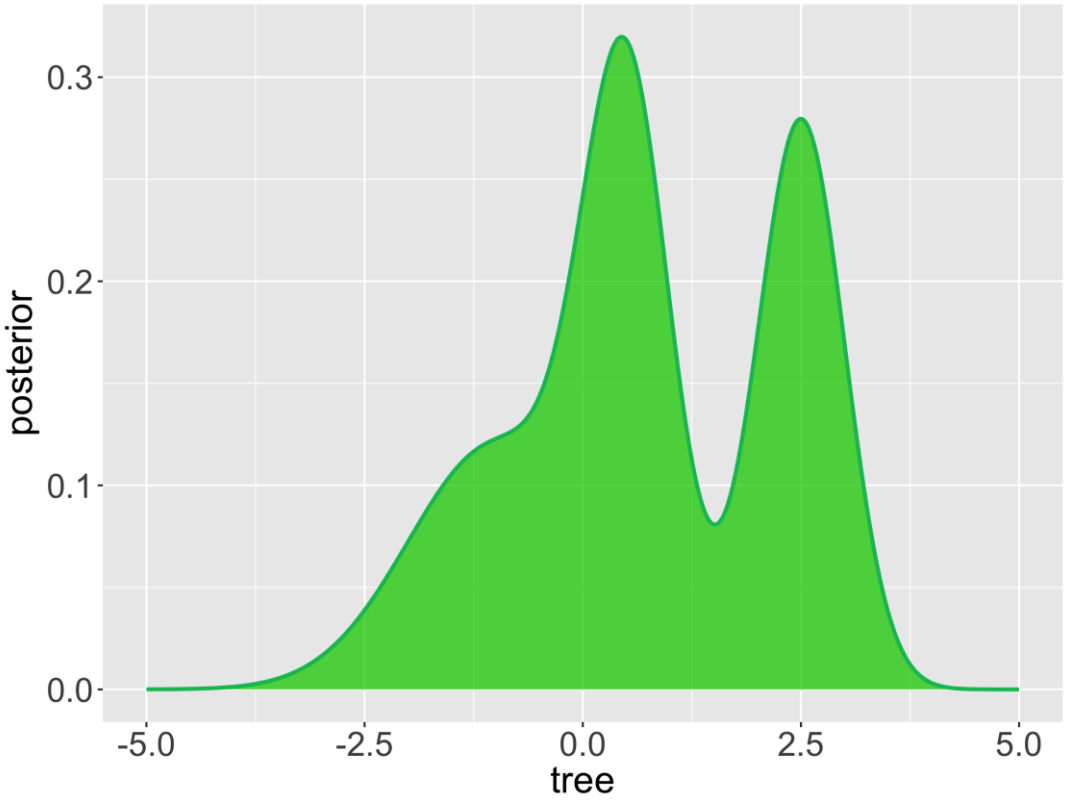
$$p(tree|MSA) = \frac{p(MSA|tree)\, p(tree)}{\sum_{tree} p(MSA|tree)\, p(tree)}$$

后验分布

问题在于，对所有可能的树求和是不可能的

n	T_n
3	1
4	3
5	15
6	105
7	945
8	10 395
9	135 135
10	2 027 025
20	$\sim 2.22 \times 10^{20}$
50	$\sim 2.84 \times 10^{74}$

马尔科夫随机采样法 (MCMC)



从初始的树出发：

- 按照$p(tree^*|tree)$，基于当前树propose一棵新树
- 计算p_{accept} =

$$\min\left(\frac{p(MSA|tree^*)p(tree^*)}{p(MSA|tree)p(tree)}, 1\right)$$

- 按照p_{accept}决定是否接受proposal，如果接受，把这棵树放到采样序列里
- 基于更新的树重复上述操作

数学上可以证明，序列中所有的树的集合，服从后验分布$\boldsymbol{p(tree|MSA)}$，即可以把它当做后验概率的采样

提要

1. 描述系统发生的概率模型
2. 系统发生树的最大似然估计
3. 系统发生树的贝叶斯估计
4. **应用与比较**

应用与比较

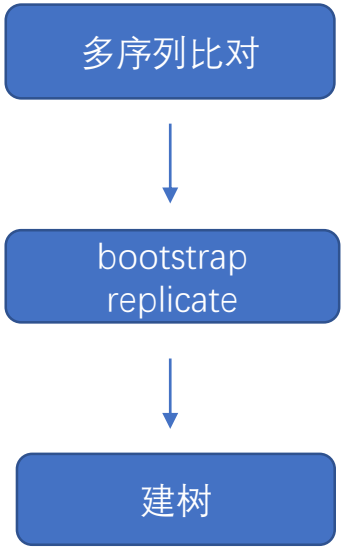
使用软件及简述方法：

- PHLIP：最大似然法
- MrBayes：贝叶斯算法
- MEGA：最大似然法

数据来源：

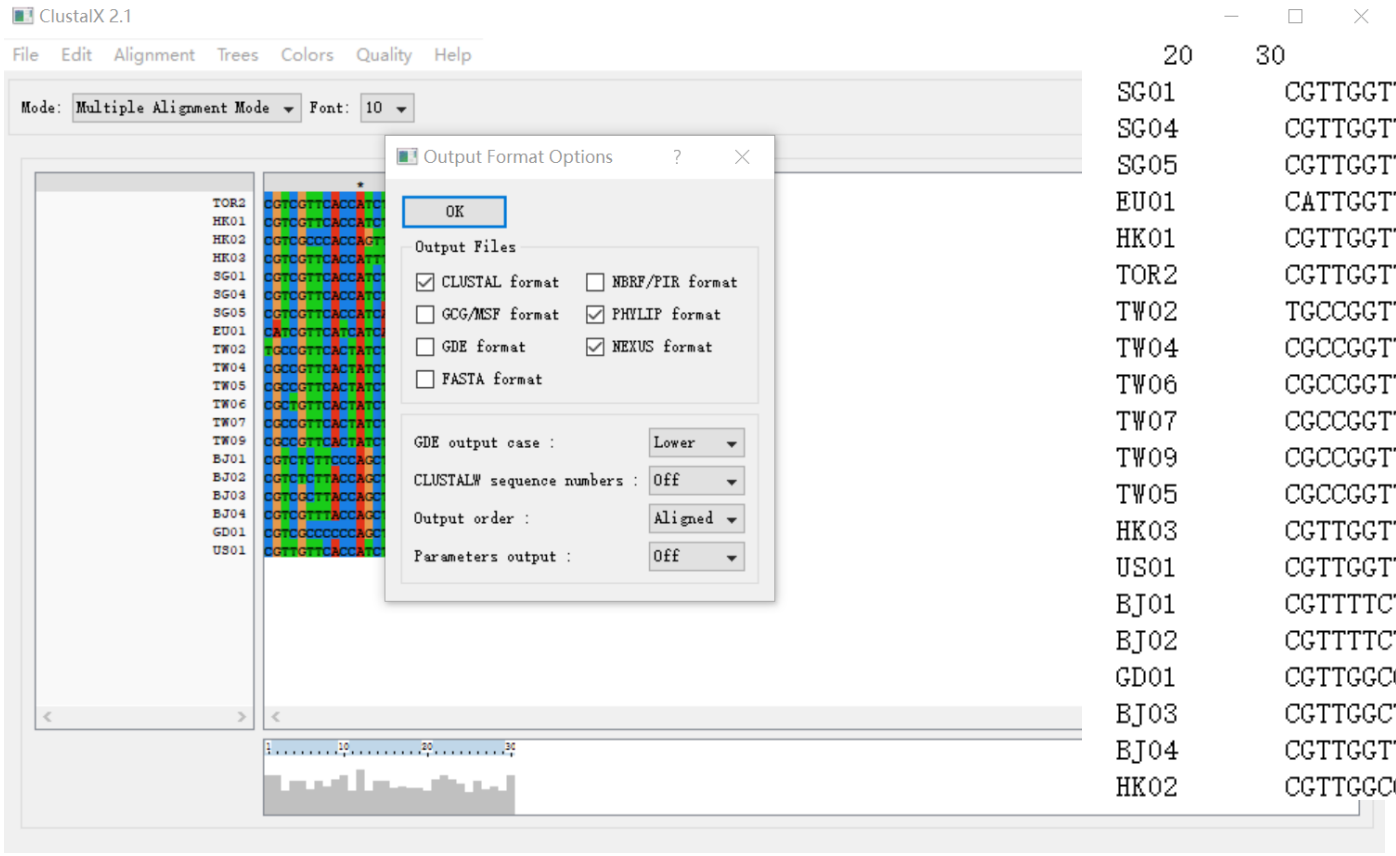
ABC网站上关于SARS病毒的20条核酸序列

PHYLIP（最大似然法）构建进化树



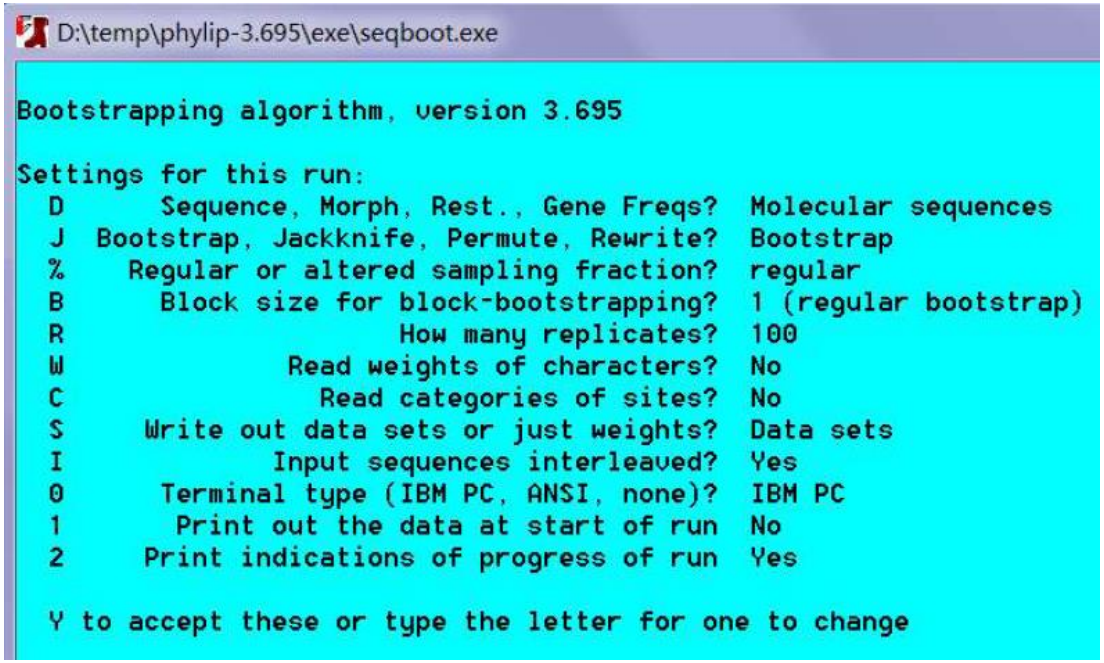
- 进行多序列比对，生成 phylip 格式的比对结果；
- 用 PHYLIP 产生 bootstrap replicate（伪样本）；
- 最大似然法建树(Maximum Likelihood)构建进化树。

ClustalX进行多序列比对




```
                20        30
SG01       CGTTGGTTAA CCCCTTCCCT TAAAAATCCC
SG04       CGTTGGTTAA CCCCTTCCCT TAAAAATCCC
SG05       CGTTGGTTAA CCCCTTCCCA TAAAAATCCC
EU01       CATTGGTTAA TTCCTTCCCA TAAAAATTCT
HK01       CGTTGGTTAA CCCCTTCCCT CAAAAATTCC
TOR2       CGTTGGTTAA CCCCTTCCCT CAAAAATCCC
TW02       TGCCGGTTAA CCTTTTCCCT CAAAAAGCCC
TW04       CGCCGGTTAA CCTTTTCCCT CAAAAAGCCC
TW06       CGCCGGTTAA CCTTTTCCCT CAAAAAGCCC
TW07       CGCCGGTTAA CCTTTTCCCT CAAAAAGCCC
TW09       CGCCGGTTAA CCTTTTCCCT CAAAAAGCCC
TW05       CGCCGGTTAA CCTTTTCCCT CAAAAAGCCC
HK03       CGTTGGTTAA CCCCTTTTTT CAAAAAGCCC
US01       CGTTGGTTAA CCCCTTCCCT CAAAAATCCC
BJ01       CGTTTTCTCC CCCCGGCCCT CGGCCCTCTC
BJ02       CGTTTTCTAA CCCCGGCCCT CGGAAATCTC
GD01       CGTTGGCCCC CCCCGGCCCT CGGAAATCTC
BJ03       CGTTGGCTAA CCCCGGCCCT CGGCCCTCTC
BJ04       CGTTGGTTAA CCCCGGCCCT CGGAAATCTC
HK02       CGTTGGCCAA CCCCGGTTTT CAAAAATCCC
```

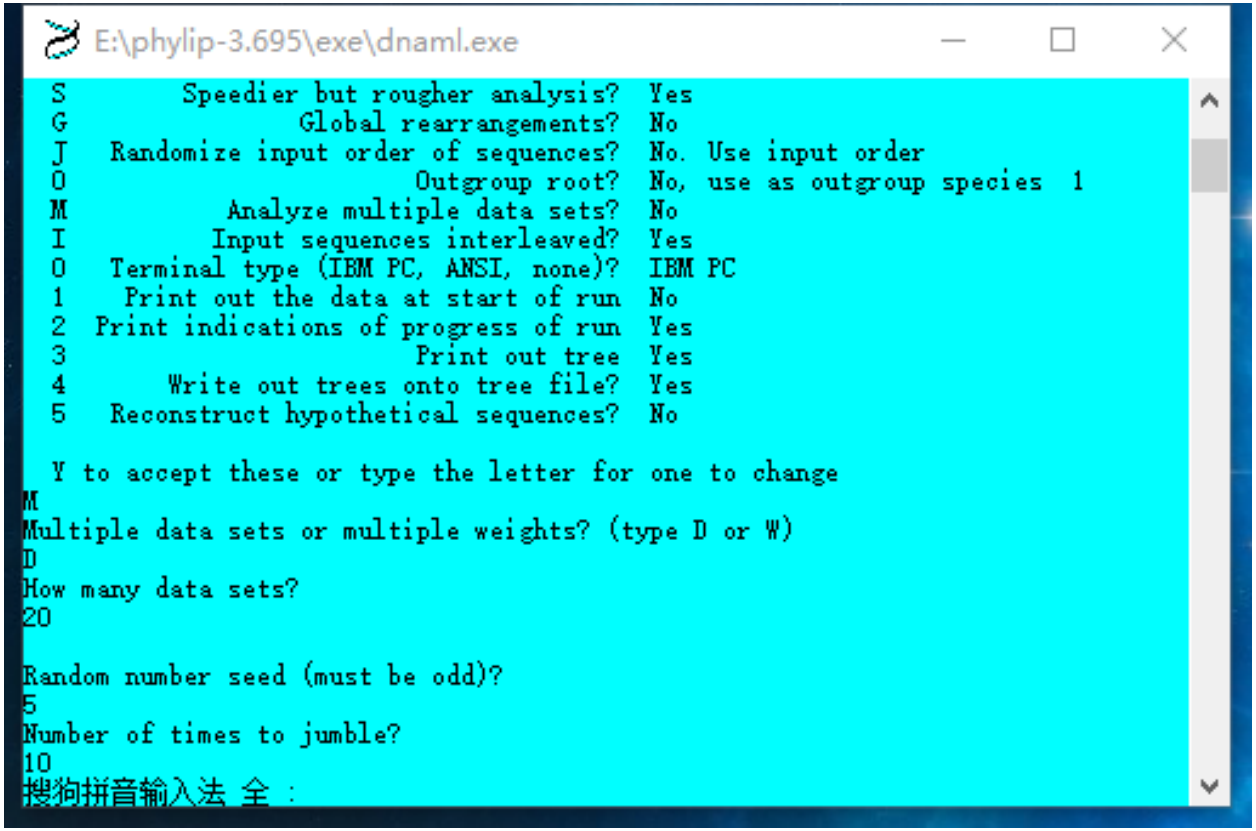
用 PHYLIP 产生 bootstrap replicate



产生bootstrap replicate(伪样本)，与原始序列一起建树

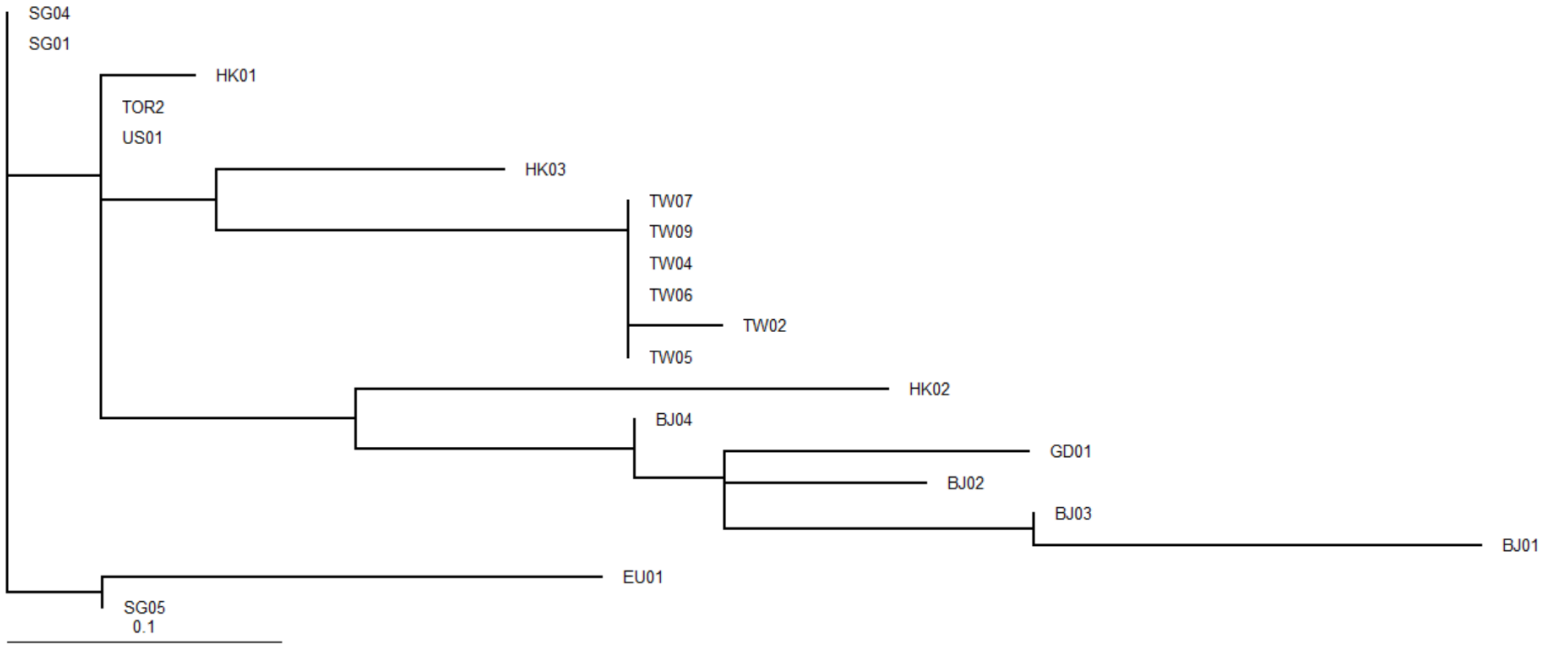
- D：数据类型；
- J：伪样本产生方法；
- B：自举法窗口大小选择，默认为 1，也可任意设定。
- **R：产生伪样本的数目，默认 100 个。**
- W：输入文件为字符还是权重；
- S：输出文件为字符数据还是权重，与输入要保持一致。
- I：phy 文件格式。

DNAML构建进化树

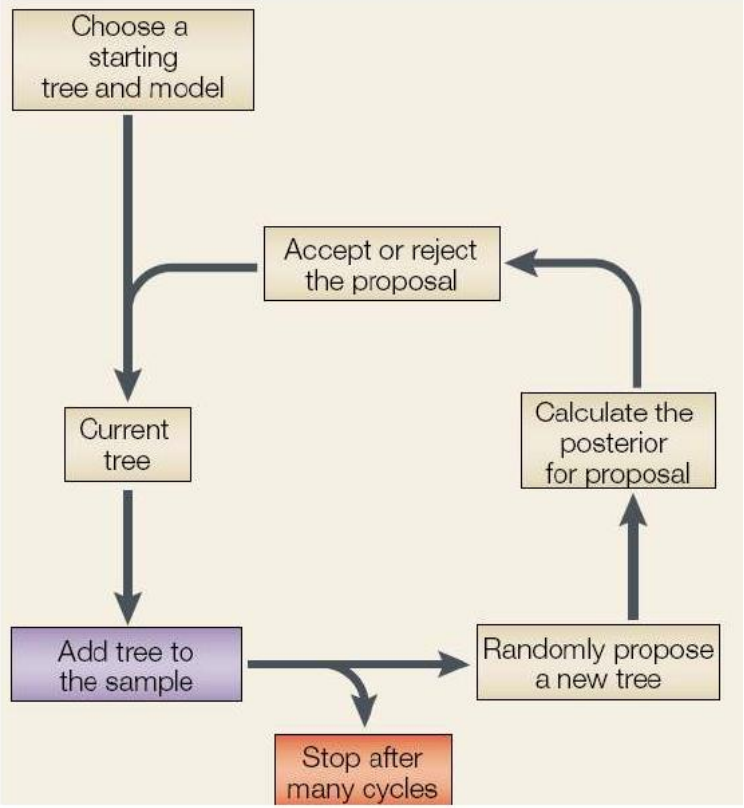


- 参数S：设置是否对所有树型和枝长依次迭代计算；
- 参数G：设置拓朴结构是否要迭代计算。

PHYLIP :


SG04
SG01
HK01
TOR2
US01
HK03
TW07
TW09
TW04
TW06
TW02
TW05
HK02
BJ04
GD01
BJ02
BJ03
BJ01
EU01
SG05
0.1

MrBayes



1. Read the Nexus data file
2. Set the evolutionary model
3. Run the analysis
4. Summarize the samples

(JP Huelsenbeck *et al*.Bioinformatics,2001)

1. Read the Nexus data file

```
[gaojd@mendel SARS]$ less SARS.nxs
#NEXUS
BEGIN DATA;
dimensions ntax=20 nchar=30;
format missing=?
interleave datatype=DNA gap= -;

matrix
SG01        CGTCGTTCACCATCTATAGTACACTCCCTC
SG04        CGTCGTTCACCATCTATAGTACACTCC-TC
SG05        CGTCGTTCACCATCAATAGTACACTCCCTC
EU01        CATCGTTCATCATCAATAGTATACTTCCTT
HK01        CGTCGTTCACCATCTACAGTACACTTCCTC
TOR2        CGTCGTTCACCATCTACAGTACACTCCCTC
TW02        TGCCGTTCACTATCTACAGTACACGCCCTC
TW04        CGCCGTTCACTATCTACAGTACACGCCCTC
TW06        CGCTGTTCACTATCTGCAGTACATGCCTTC
TW07        CGCCGTTCACTATCTGCAGTACATGCCTTC
TW09        CGCCGTTCACTATCTACAGTACATGCCTTC
TW05        CGCCGTTCACTATCTGCAGTACACGCCCTC
HK03        CGTCGTTCACCATTTACAGTACACGCCCTC
US01        CGTTGTTCACCATCTGCAGTACACTCCCTC
BJ01        CGTCTCTTCCCAGCTACGACACCCTCTCCC
BJ02        CGTCTCTTACCAGCTACGACGCACTCTCCC
GD01        CGTCGCCCCCCAGCTACGACGCACTCTCCC
BJ03        CGTCGCTTACCAGCTACGACACCCTCTCCC
BJ04        CGTCGTTTACCAGCTACGGCACACTCTCCC
HK02        CGTCGCCCACCAGTTGCAACACACTCCCCC
;
end;
```

```
MrBayes > Execute ../SARS/SARS.nxs

   Executing file "../SARS/SARS.nxs"
   DOS line termination
   Longest line length = 44
   Parsing file
   Expecting NEXUS formatted file
   Reading data block
      Allocated taxon set
      Allocated matrix
      Defining new matrix with 20 taxa and 30 characters
      Missing data coded as ?
      Data is Dna
      Gaps coded as -
      Taxon  1 -> SG01
      Taxon  2 -> SG04
      Taxon  3 -> SG05
      Taxon  4 -> EU01
      Taxon  5 -> HK01
      Taxon  6 -> TOR2
      Taxon  7 -> TW02
      Taxon  8 -> TW04
      Taxon  9 -> TW06
      Taxon 10 -> TW07
      Taxon 11 -> TW09
      Taxon 12 -> TW05
      Taxon 13 -> HK03
      Taxon 14 -> US01
      Taxon 15 -> BJ01
      Taxon 16 -> BJ02
      Taxon 17 -> GD01
      Taxon 18 -> BJ03
      Taxon 19 -> BJ04
      Taxon 20 -> HK02
      Successfully read matrix
      Setting default partition (does not divide up characters)
      Setting model defaults
      Seed (for generating default start values) = 1515725017
      Setting output file names to "../SARS/SARS.nxs.run<i>.<p|t>"
   Exiting data block
   Reached end of file
```

2. Set the evolutionary model

```
MrBayes > lset nst=6 rates=invgamma

   Setting Nst to 6
   Setting Rates to Invgamma
   Successfully set likelihood model parameters
```

- **Nst** 是设定替换模型，6 是General Time Reversible (GTR) model，任意两个核苷酸之间替换速率均不相等（C42 = 6 参数）
- gamma（位点间替换速率变化服从gamma分布

3. Run the analysis

```
MrBayes > mcmc ngen=20000 samplefreq=100 printfreq=100 diagnfreq=1000

   Setting number of generations to 20000
   Setting sample frequency to 100
   Setting print frequency to 100
   Setting diagnosing frequency to 1000
   Running Markov chain
   MCMC stamp = 7107536226
   Seed = 1537625422
   Swapseed = 965369756
```

- **Ngen**（number of generations）设置分析要跑的代数。
- **Samplefreq**定义对链取样的频率。
- **Diagnfreq**默认的每1000代计算各种运行（分析）的诊断，并把它们保存在<filename>.mcmc的文件中。
- **Seed**是随机数产生器随机输出的一个种子数值。
- 默认状态下，bayes会同时运行两个(Nruns = 2)完全独立的但由不同的随机树开始的分析。

3.1总结采样取代模型参数

[ID: 7107536226]

Gen	LnL	LnPr	TL	r(A<->C)	r(A<->G)	r(A<->T)	r(C<->G)	r(C<->T)	r(G<->T)	pi(A)	pi(C)	pi(G)	pi(T)	alpha	pinvar	
0	-2.204258e+02	2.824498e+01	1.859166e+00	6.993839e-02	4.926987e-01	1.568211e-02	5.274900e-03	3.288468e-01	8.755909e-02	2.058344e-01	3.708123e-01	1.32				
100	-2.213993e+02	3.019200e+01	1.716744e+00	6.993839e-02	4.926987e-01	1.568211e-02	5.274900e-03	3.288468e-01	8.755909e-02	1.728026e-01	3.708123e-01	1.65				
200	-2.246698e+02	3.731089e+01	1.450949e+00	6.242355e-02	5.201656e-01	1.920025e-02	4.413126e-03	3.011068e-01	9.269061e-02	1.434678e-01	3.735539e-01	1.98				
300	-2.273297e+02	3.012172e+01	1.760605e+00	6.242355e-02	5.201656e-01	1.920025e-02	4.413126e-03	3.011068e-01	9.269061e-02	1.434678e-01	4.264059e-01	1.98				
400	-2.276601e+02	2.979116e+01	1.777784e+00	6.242355e-02	5.201656e-01	1.920025e-02	4.413126e-03	3.011068e-01	9.269061e-02	1.604042e-01	3.943599e-01	2.19				
500	-2.294396e+02	3.519869e+01	1.444977e+00	7.436981e-02	5.033689e-01	3.180386e-02	9.870466e-03	2.842009e-01	9.638609e-02	1.817125e-01	3.598110e-01	1.98				
600	-2.348283e+02	3.852714e+01	1.355094e+00	5.816209e-02	4.318916e-01	8.831592e-02	2.008328e-03	2.650579e-01	1.545641e-01	1.606966e-01	3.534500e-01	1.15				
700	-2.354423e+02	2.313287e+01	2.068534e+00	5.816209e-02	4.318916e-01	8.831592e-02	2.008328e-03	2.650579e-01	1.545641e-01	1.606966e-01	3.511522e-01	1.18				
800	-2.308357e+02	2.938278e+01	1.615210e+00	6.805930e-02	5.781604e-01	4.823798e-02	2.927562e-02	2.004594e-01	7.580733e-02	1.491871e-01	3.461840e-01	2.36				

取样来源的随机生成的ID号；

第2行为标题，从左到右依次为：

1) 代数（the generation number，Gen）；

2) 冷链对数似然值（the log likelihood of the cold chain，LnL）；

3) 树长（the total tree length，TL）；

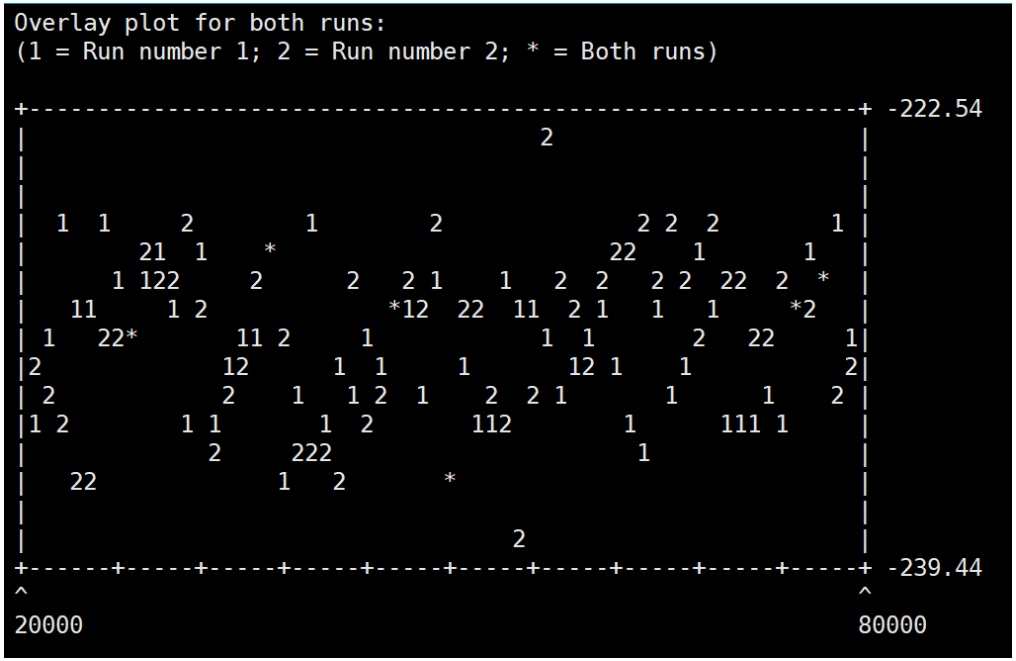
4) 6个GTR比率参数（the six GTR rate parameters，如r(A<->C), r(A<->G)等）；

5) 4个核苷酸发生频率（the four stationary nucleotide frequencies，如pi(A), pi(C)等）；

6) 取代速率变化Gamma分布的形状参数（the shape parameter of the gamma distribution of rate variation，alpha）；

7) 不变位点的比例（the proportion of invariable sites，pinvar）。

4. Summarize the samples



- sump命令会首先生成一个代数（ngen）和对数似然值（the log likelihood values）的关系图。如果分析代数已经足够充分的话，图看起来很平稳，没有上升或者下降的趋势。如果有任何上升或者下降的趋势，可能需要延长分析时间以获得充分的后验概率分布取样。

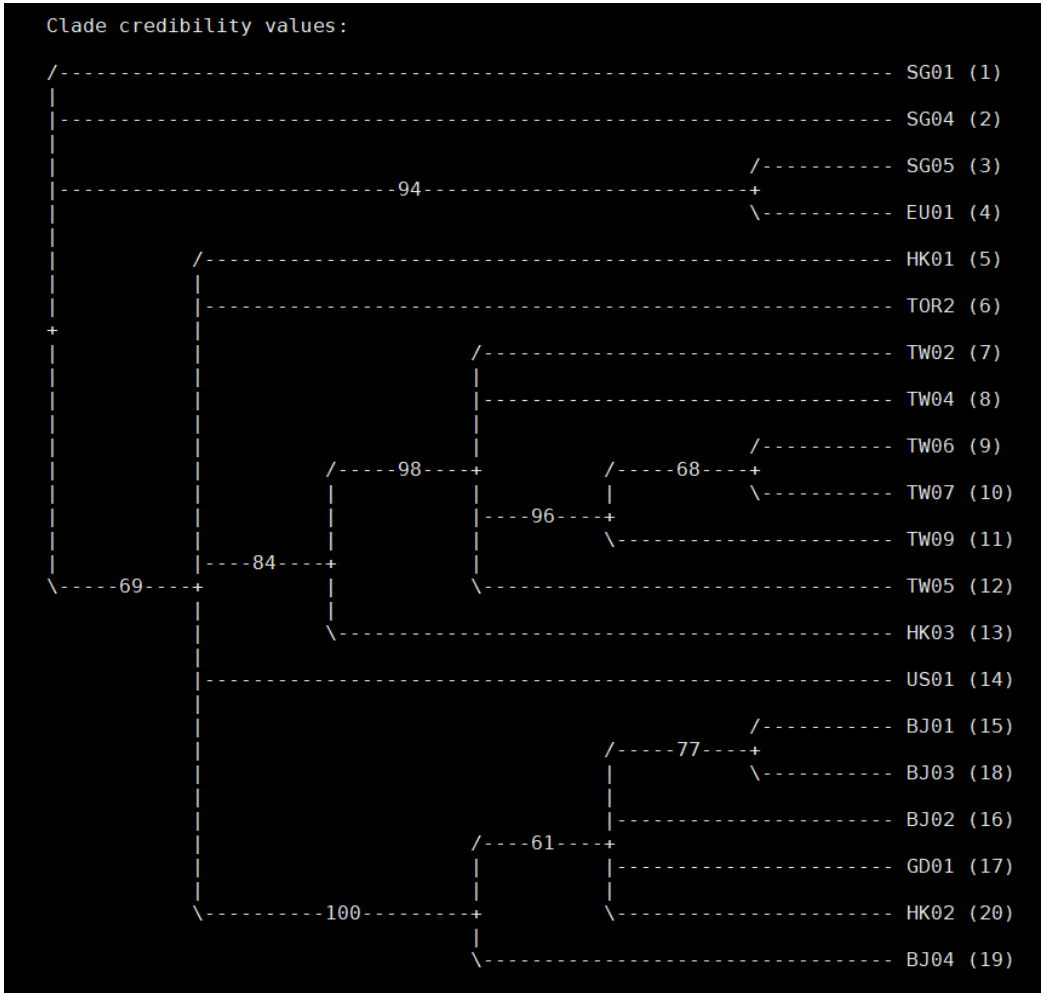
4. Summarize the samples

			95% HPD Interval					
Parameter	Mean	Variance	Lower	Upper	Median	min ESS*	avg ESS	PSRF+
TL	1.616051	0.109078	1.027827	2.302609	1.596137	263.06	275.07	1.001
r(A<->C)	0.061382	0.001414	0.009900	0.133501	0.052353	37.88	44.81	1.013
r(A<->G)	0.450887	0.010049	0.246358	0.640673	0.458672	31.99	35.55	1.000
r(A<->T)	0.043247	0.000990	0.000019	0.109993	0.036380	35.23	77.75	1.024
r(C<->G)	0.019068	0.000321	0.000095	0.053759	0.013027	43.04	73.32	1.023
r(C<->T)	0.279824	0.006599	0.144515	0.449598	0.273725	23.68	30.89	1.003
r(G<->T)	0.145591	0.003043	0.044046	0.247183	0.139208	34.42	131.41	1.006
pi(A)	0.164983	0.002528	0.075426	0.260346	0.161117	57.16	59.00	1.000
pi(C)	0.398525	0.004010	0.273514	0.514495	0.398717	86.23	130.90	0.999
pi(G)	0.144113	0.001902	0.060482	0.221567	0.141301	94.51	143.91	1.013
pi(T)	0.292379	0.003115	0.195922	0.401034	0.292907	35.77	114.12	1.006
alpha	3.706789	2.039074	1.453895	6.498836	3.501712	197.00	236.20	1.000
pinvar	0.035162	0.001127	0.000091	0.101543	0.024978	120.21	219.46	0.999

* Convergence diagnostic (ESS = Estimated Sample Size); min and avg values correspond to minimal and average ESS among runs.
ESS value below 100 may indicate that the parameter is undersampled.
\+ Convergence diagnostic (PSRF = Potential Scale Reduction Factor; Gelman and Rubin, 1992) should approach 1.0 as runs converge.

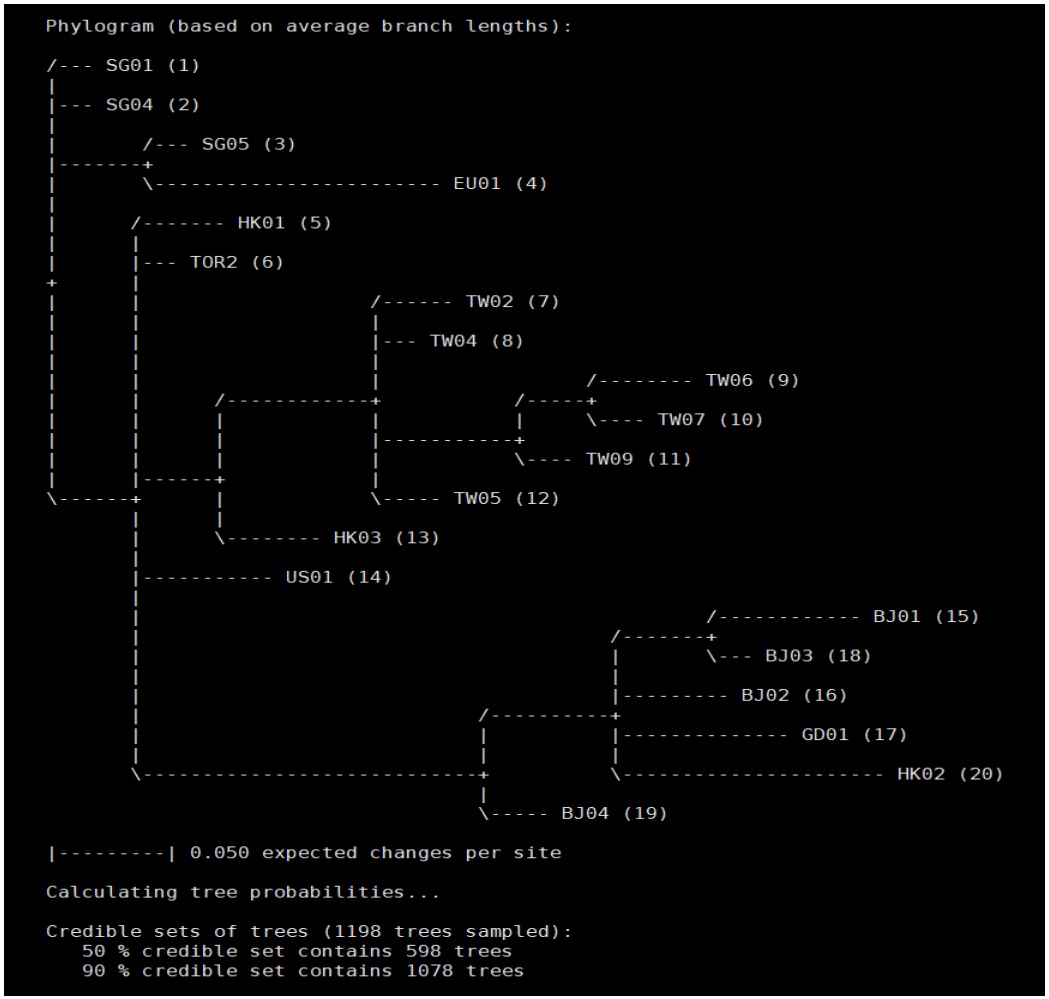
- 各参数的平均值（Mean）、方差（Variance）、95%可靠间区（Credible interval）的下界（Lower）和上界（Upper）、中位值（Median）和PSRF（Potential Scale Reduction Factor）。
- PSRF所有参数的潜在缩放因子（PSRF）也是一种收敛诊断方式，如果分析较彻底，该值应接近1.0；如果不是，则需要更长时间运行分析

4. Summarize the samples



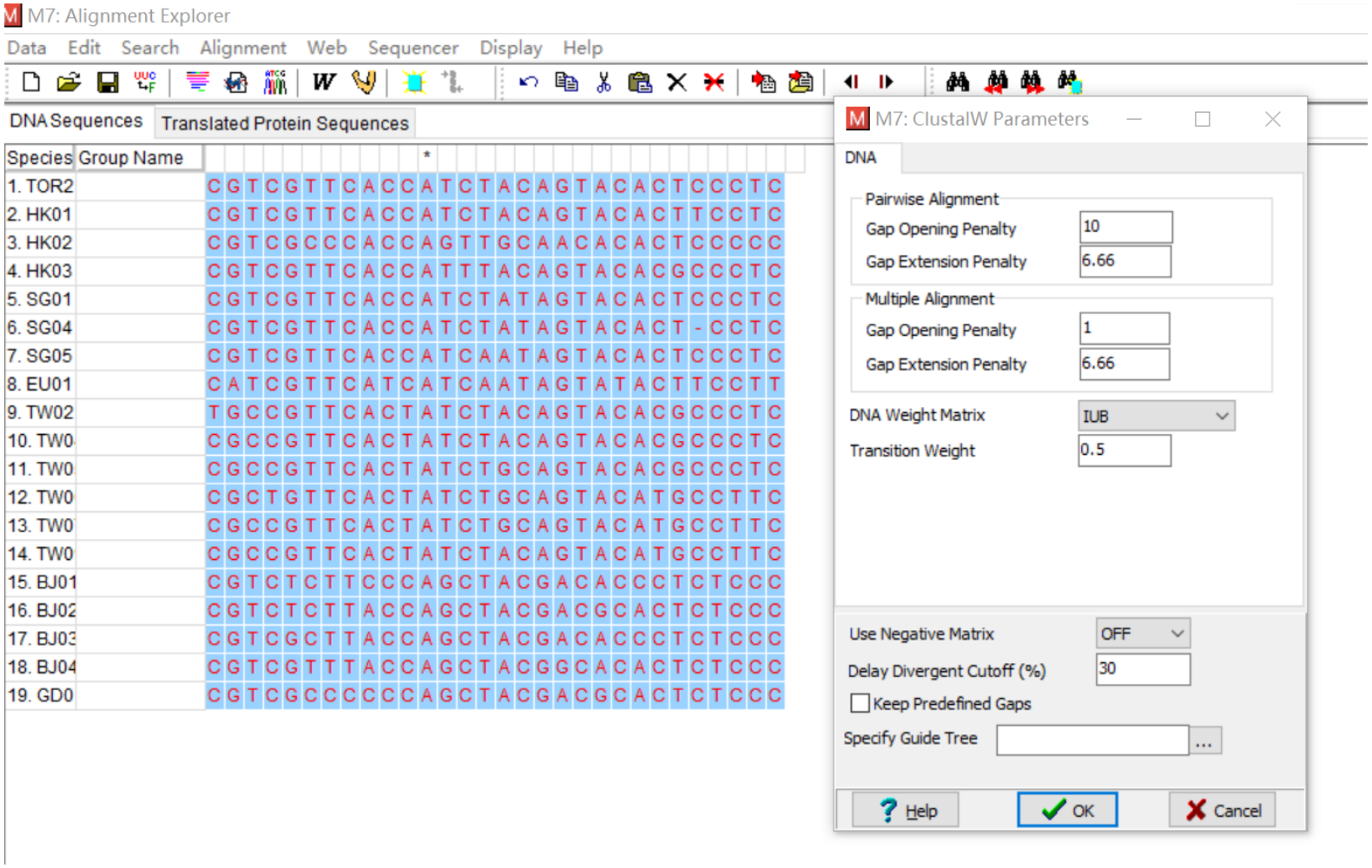
- 枝长可信度树clade credibility (posterior probability) values），上面标注了每个分枝(branch pattern)的后验概率

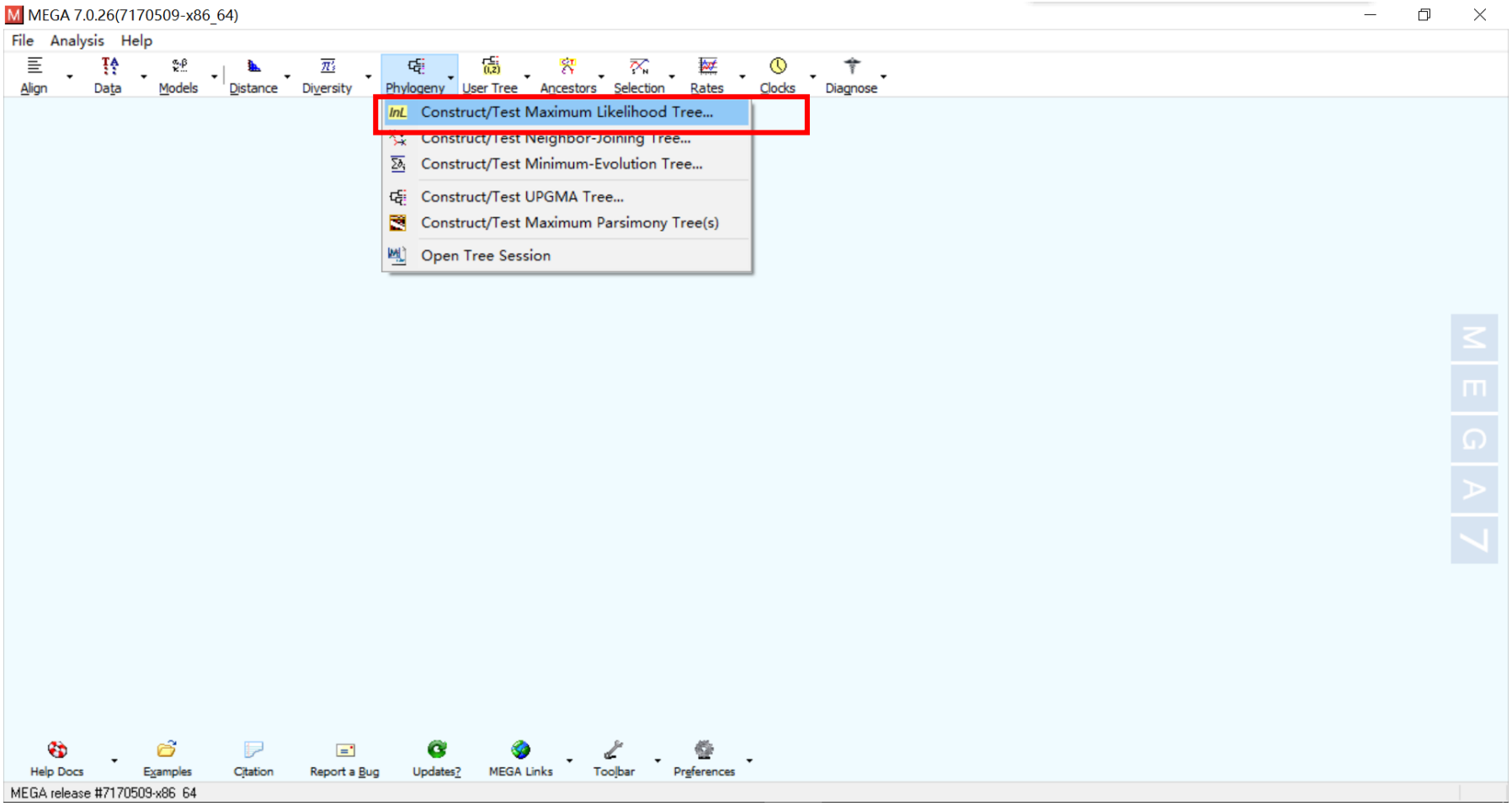
4. Summarize the samples

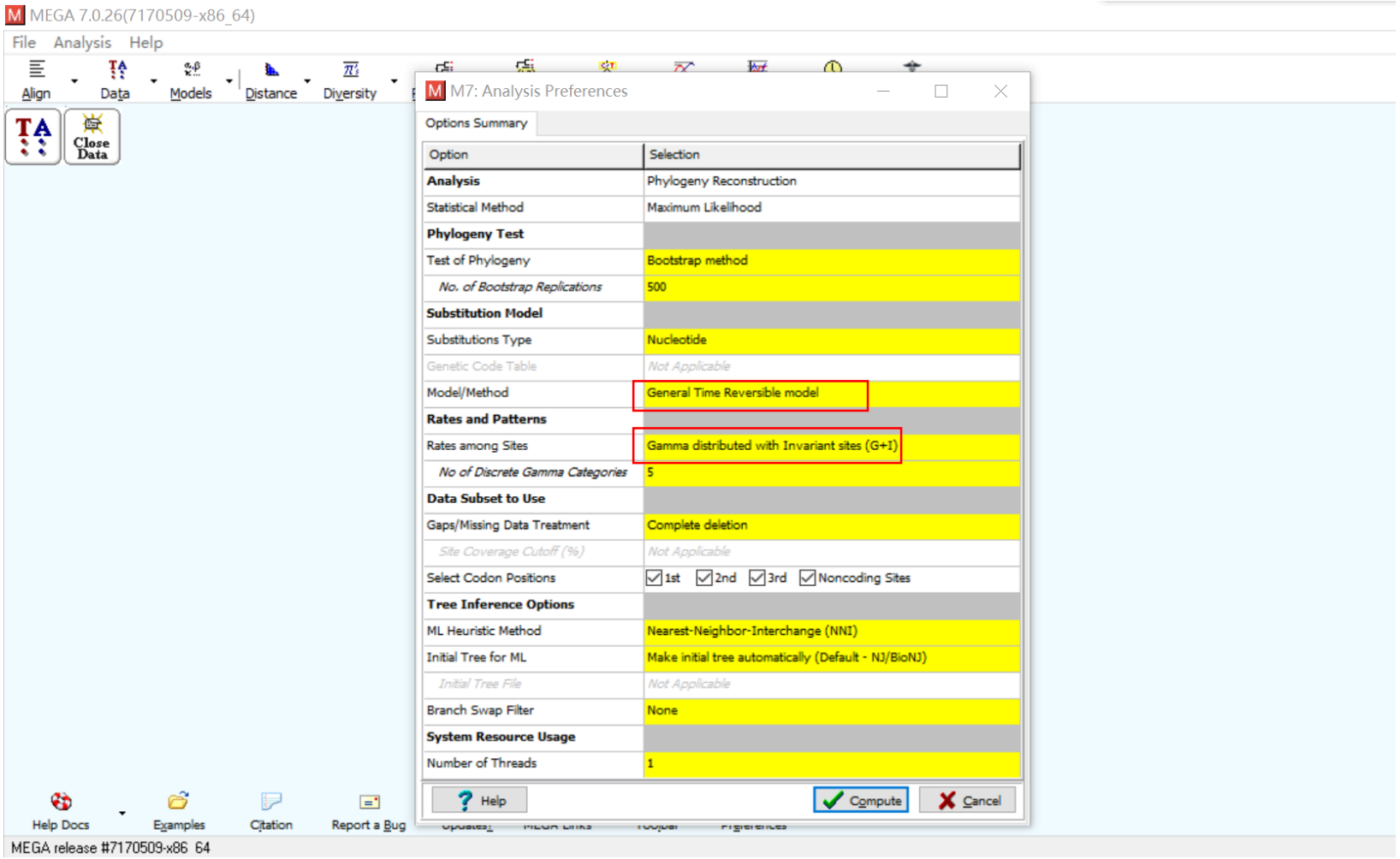


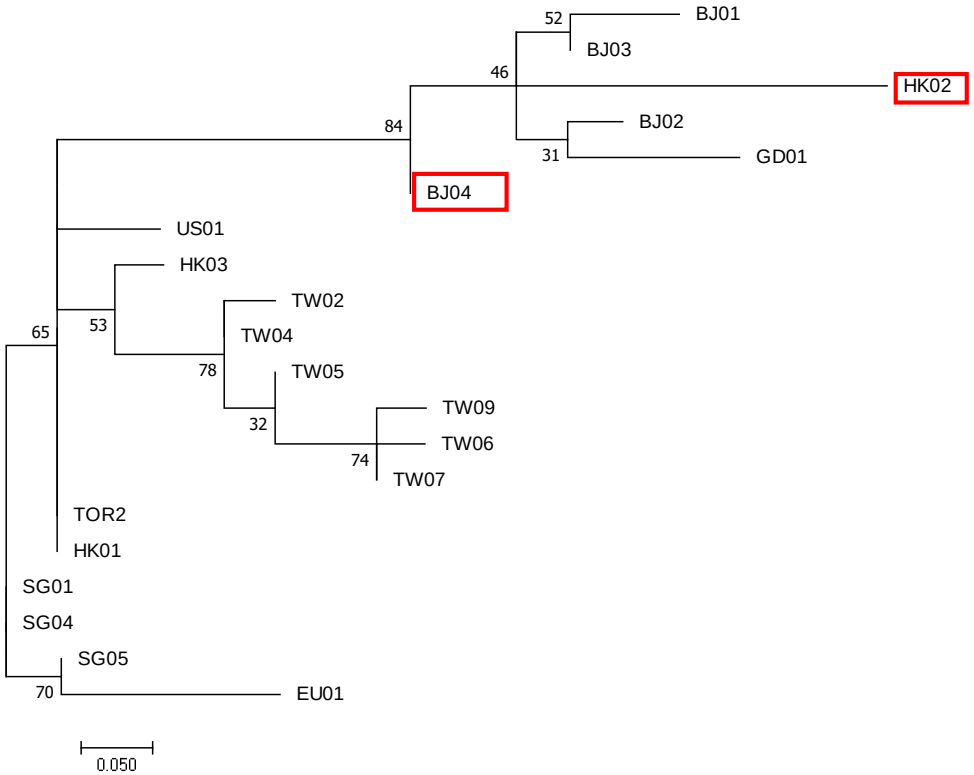
- 系统演化树给出枝长branch lengths measured in expected substitutions per site
- 各枝的长度与其遗传距离相关

MEGA（最大似然法）构建进化树



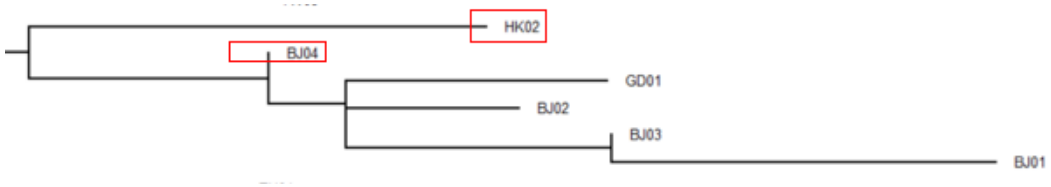

MEGA 7.0.26(7170509-x86_64)
File
Analysis
Help
Align
Data
Models
Distance
Diversity
Phylogeny
User Tree
Ancestors
Selection
Rates
Clocks
Diagnose
Construct/Test Maximum Likelihood Tree...
Construct/Test Neighbor-Joining Tree...
Construct/Test Minimum-Evolution Tree...
Construct/Test UPGMA Tree...
Construct/Test Maximum Parsimony Tree(s)
Open Tree Session
Help Docs
Examples
Citation
Report a Bug
Updates?
MEGA Links
Toolbar
Preferences
MEGA release #7170509-x86_64


MEGA 7.0.26(7170509-x86_64)
File Analysis Help
Align Data Models Distance Diversity
Close Data
M7: Analysis Preferences
Options Summary
Option
Selection
Analysis
Phylogeny Reconstruction
Statistical Method
Maximum Likelihood
Phylogeny Test
Test of Phylogeny
Bootstrap method
No. of Bootstrap Replications
500
Substitution Model
Substitutions Type
Nucleotide
Genetic Code Table
Not Applicable
Model/Method
General Time Reversible model
Rates and Patterns
Rates among Sites
Gamma distributed with Invariant sites (G+I)
No of Discrete Gamma Categories
5
Data Subset to Use
Gaps/Missing Data Treatment
Complete deletion
Site Coverage Cutoff (%)
Not Applicable
Select Codon Positions
1st 2nd 3rd Noncoding Sites
Tree Inference Options
ML Heuristic Method
Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML
Make initial tree automatically (Default - NJ/BioNJ)
Initial Tree File
Not Applicable
Branch Swap Filter
None
System Resource Usage
Number of Threads
1
Help
Compute
Cancel
Help Docs
Examples
Citation
Report a Bug
MEGA release #7170509-x86_64

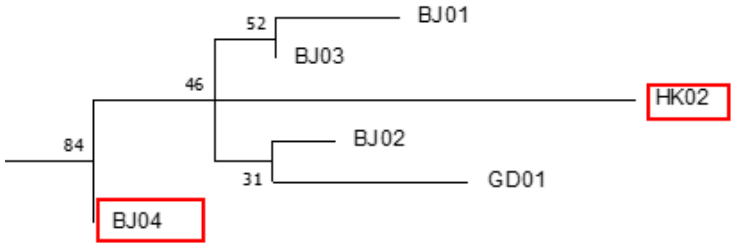

BJ01
52
BJ03
46
HK02
BJ02
84
31
GD01
BJ04
US01
HK03
TW02
65
53
TW04
78
TW05
32
TW09
TW06
74
TW07
TOR2
HK01
SG01
SG04
SG05
70
EU01
0.050

MEGA最大似然法

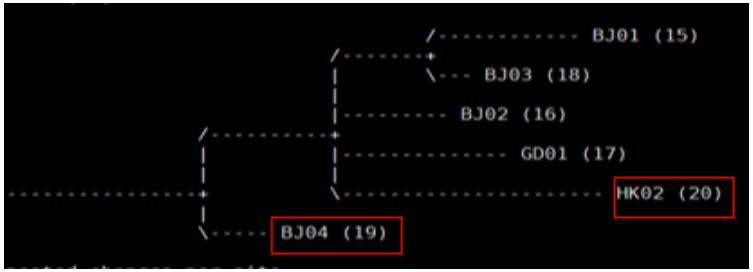
结果比较：



PHLIP最大似然法



MEGA最大似然法



贝叶斯法

总结

- 介绍了基本的分子演化概率模型
- 启发式最大似然搜索可以给出一棵局部最优的树
- 贝叶斯估计可以给出树的后验分布，从而准确地评估树的可靠性
- 实际例子中可以看到，最大似然的不同启发式算法可能给出不同的结果，而贝叶斯估计可以更全面、客观地评估给定观测序列背后的演化过程

Thanks !