

## 贝叶斯方法建树 MrBayes

网址：<http://mrbayes.sourceforge.net/>

MrBayes 是一款利用贝叶斯方法进行进化树构建的软件。贝叶斯方法建树与最大似然法建树有密切联系，最大似然法是寻找合适的参数（树型、枝长和进化模型），使得数据（多序列比对结果）的似然率最大，最大化  $P(\text{Data} \mid \text{Tree})$ ；而贝叶斯方法则是利用给定数据，寻找概率最大的树型，最大化  $P(\text{Tree} \mid \text{Data})$ 。不仅如此，贝叶斯方法还提供给定数据条件下，各种树型出现的概率，也称后验概率。但实际上各种树型的后验概率很难直接计算，一般是采用 MCMC 方法来近似。图 3.1 是利用 MCMC(Markov Chain Monte Carlo)方法构建贝叶斯树的流程图。

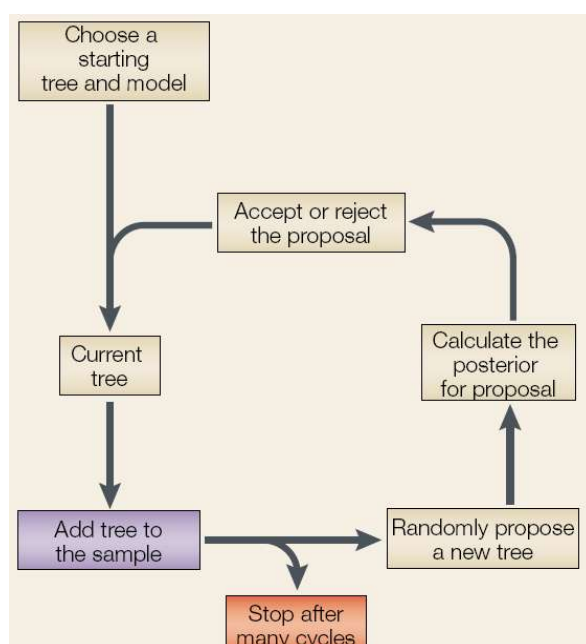


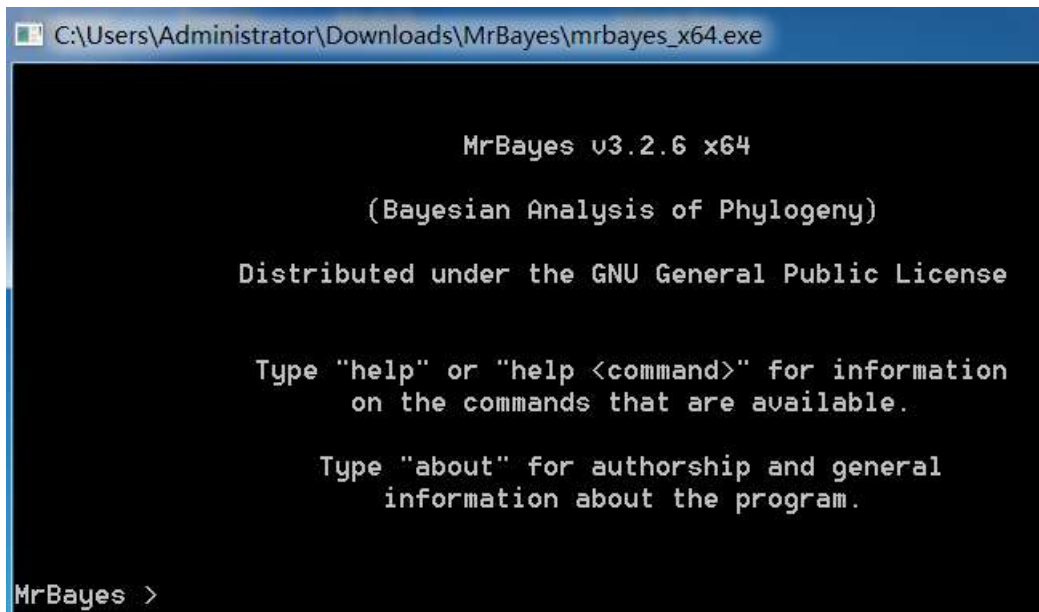
图 3.1 MCMC 方法构建贝叶斯树流程图

MrBayes 是 DOS 界面的，不需安装可直接使用。它默认读入 nexus 格式的多序列比对结果文件进行分析，我们首先用程序提供的演示序列学习使用方法。

### 3.1 运行 MrBayes，读入 nex 文件，确定参数。

解压程序文件夹，可以看到里面有两个可执行文件，mrbayes\_x64 适用于 64 位操作系统的

电脑，mrbayes\_X86 适合 32 位电脑。双击应用程序，打开界面（图 3.2）。



```
C:\Users\Administrator\Downloads\MrBayes\mrbayes_x64.exe

MrBayes v3.2.6 x64

(Bayesian Analysis of Phylogeny)

Distributed under the GNU General Public License

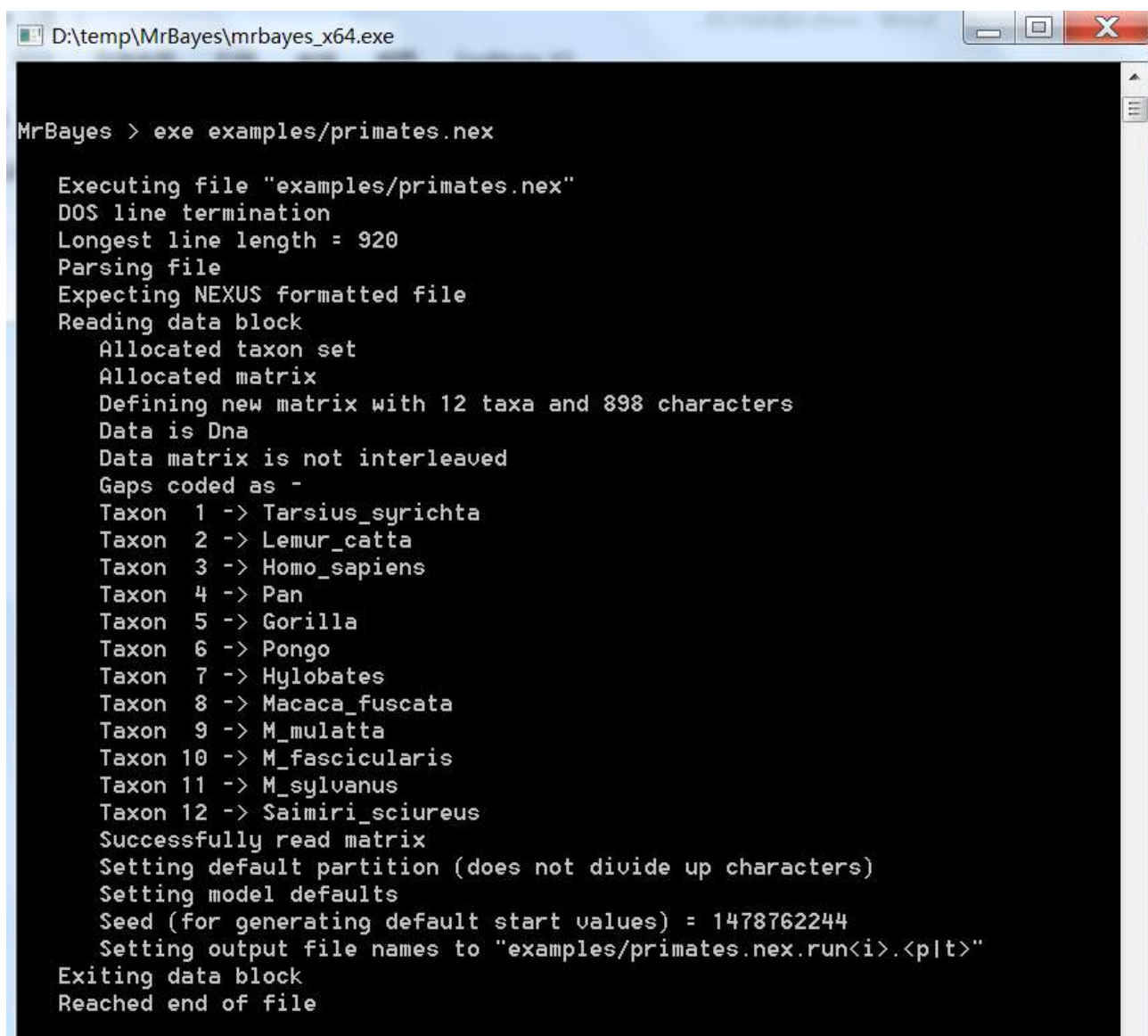
Type "help" or "help <command>" for information
on the commands that are available.

Type "about" for authorship and general
information about the program.

MrBayes >
```

图 3.2 MrBayes 打开界面

在“ MrBayes >” 控制符后输入 [execute examples/primates.nex](#) 点击回车，程序读入演示用的多序列比对文件 primates.nex，并输出相关数据供核对。（图 3.3）



```
MrBayes > exe examples/primates.nex

Executing file "examples/primates.nex"
DOS line termination
Longest line length = 920
Parsing file
Expecting NEXUS formatted file
Reading data block
  Allocated taxon set
  Allocated matrix
  Defining new matrix with 12 taxa and 898 characters
  Data is Dna
  Data matrix is not interleaved
  Gaps coded as -
  Taxon 1 -> Tarsius_syrichta
  Taxon 2 -> Lemur_catta
  Taxon 3 -> Homo_sapiens
  Taxon 4 -> Pan
  Taxon 5 -> Gorilla
  Taxon 6 -> Pongo
  Taxon 7 -> Hylobates
  Taxon 8 -> Macaca_fuscata
  Taxon 9 -> M_mulatta
  Taxon 10 -> M_fascicularis
  Taxon 11 -> M_sylvanus
  Taxon 12 -> Saimiri_sciureus
Successfully read matrix
Setting default partition (does not divide up characters)
Setting model defaults
Seed (for generating default start values) = 1478762244
Setting output file names to "examples/primates.nex.run<i>. <plt>"
Exiting data block
Reached end of file
```

图 3.3 MrBayes 读入序列

如果要分析自己的序列，需先用序列比对软件（如 ClustalX）进行多序列比对，输出比对结果为 MrBayes 可识别的为 **nexus** 文件，或者用比对结果转换软件将其它格式的比对结果转换成 nexus 格式。

**Tips:** 在用 ClustalX 做多序列比对前，先将要比对的序列 fasta 格式第一行（“>”后）只保留物种名\_序列名，这样方便后续分析，用 ClustalX 读入序列后，在 output format option 选中 nexus 格式，对序列进行比对（Do complete alignment）。生成的 nxs 文件可以用写字板打开浏览，里面内容是多条序列比对结果。将它与 mrbayes 可执行文件放在同一目录下即可读入分析。

可以在控制符后输入 **help lset** , 查看默认的参数(图 3.4)。

Model settings for partition 1:		
Parameter	Options	Current Setting
Nucmodel	4by4/Doublet/Codon/Protein	4by4
Nst	1/2/6/Mixed	1
Code	Universal/Uvertmt/Invermt/Yeast/Mycoplasma/ Ciliate/Echinoderm/Euplotid/Metmt	Universal
Ploidy	Haploid/Diploid/Zlinked	Diploid
Rates	Equal/Gamma/LNorm/Propinv/ Invgamma/Adgamma	Equal
Ngammacat	<number>	4
Nbetacat	<number>	5
Omegavar	Equal/Ny98/M3	Equal
Covarion	No/Yes	No
Coding	All/Variable/Informative/Nosingletons Noabsencesites/Nopresencesites/ Nosingletonabsence/Nosingletonpresence	All
Parsmodel	No/Yes	No

图 3.4 默认参数 ( 部分 )

**Nucmodel** 设定核酸模型，doublet 是二联子核苷酸，有 16 种状态，用于分析核糖体 DNA 茎环结构，codon 是将核酸作为三联体密码子分析的，默认的 4by4 就是将核酸按其四种核苷酸 ( ACGT ) 进行分析。

**Nst** 设定替换模型，1 是 JC Model，所有核苷酸替换速率相同，2 是 kimura Model，将转换和颠换区分开来，6 是 General Time Reversible (GTR) model，任意两个核苷酸之间替换速率均不相等，有 6 个速率参数。

**Code** 是选择密码子编码方式的，只有在核酸模型( Nucmodel )选择为 codon 时需要设定。

**Ploidy** 有两个选项，单倍体 haploid 和多倍体 diploid，只有在设定了 coalescence prior 时需要设定。

**Rates**是设定替换速率的，默认的是equal(所有位点替换速率相同)，还可选择gamma ( 位点间替换速率变化服从gamma分布 )，propinv ( 有一部分位点的替换速率是不变的 )，Invgamma (有一部分位点的替换速率是不变的，其余位点的替换速率服从gamma分布)

adgamma ( 相邻位点的替换速率是相关的, 其边缘分布为gamma分布 ) .

**Ngammacat** : 由于 gamma 分布是连续的, 无法计算某个点的概率, 所以这里采用了近似的方法, 将连续的 gamma 分布分成几个单元 ( categories ) ,每个单元的平均速率作为此单元的替换速率。这里设定的就是独立的单元数目(the number of discrete categories), 默认为 4, 这个数字越高, 近似效果越好, 但需要的时间也越长。

**Nbetacat** : 当使用表型数据 ( morphological data ) 建树时, 替换速率常采用 beta 分布近似, 与 gamma 分布相同, 需要将连续的分布近似为几个离散的分类, 这里设定的就是独立的分类数目。

**Omegavar** 设定各位点间的非同义替换与同义替换速率的比值。Equal 假定各位点间的 omega (nonsynonymous/synonymous rate)相同 ,Ny98 和 M3 各位点间 omega 值不同, 两个模型的取值范围有所不同。

**Covarion** 设定在进化历史上各位点替换速率是否恒定。选择 yes ,则各位点突变速率随时间会发生变化。

**Coding** 设定数据抽样方式, all 适用于所有类型的数据, 所有的字符都会被抽中; variable 只适合表型数据和限制性位点 ( restriction site ) 数据, 只有发生了变化的数据会被抽中; noabsencesite 适合限制性位点数据, 所有物种中不存在的位点不抽取。

**Parsmodel** 是否利用简约模型 ( parsimony model ) 。

在控制符后输入 **lset nst=6 rates=invgamma**, 点击回车(图 3.5)。将进化模型设定为 GTR 模型, 4 种核苷酸替换速率各不相同, 序列中有部分位点不发生替换, 其它位点的替换服从 Gamma 分布。

```
MrBayes > lset nst=6 rates=invgamma
Setting Nst to 6
Setting Rates to Invgamma
Successfully set likelihood model parameters
MrBayes >
```

图 3.5 设定参数

这时在控制符后输入 **Help lset** , 查看刚才设定的参数是改变 ( 图 3.6 )。

```
Model settings for partition 1:
```

Parameter	Options	Current Setting
Nucmodel	4by4/Doublet/Codon/Protein	4by4
Nst	1/2/6/Mixed	6
Code	Universal/Uertmt/Invermt/Yeast/Mycoplasma/ Ciliate/Echinoderm/Euplotid/Metmt	Universal
Ploidy	Haploid/Diploid/Zlinked	Diploid
Rates	Equal/Gamma/LNorm/Propinv/ Invgamma/Adgamma	Invgamma
Ngammacat	<number>	4
Nbetacat	<number>	5
Omegavar	Equal/Ny98/M3	Equal
Covarion	No/Yes	No
Coding	All/Variable/Informative/Nosingletons Noabsencesites/Nopresencesites/ Nosingletonabsence/Nosingletonpresence	All
Parsmodel	No/Yes	No

图 3.6 参数修改后页面

如果你分析的是蛋白质序列，需要提前设定氨基酸替换模型，可选的氨基酸替换模型包括 **poisson**(20 种氨基酸频率相同，替换速率相同) **jones**( Jones ,1992 )，**dayhoff**(Dayhoff, Schwartz and Orcutt, 1978), **mtrev**(Adachi and Hasegawa, 1996), **mtmam**(Cao et al., 1998; Yang, Nielsen, and Hasegawa, 1998), **wag**(Wheland and Goldman, 2001), **rtrev**(Dimmic et al., 2002), **cprev**(Adachi et al., 2000) , **vt**(Muller and Vingron, 2000) , **blosum**(Henikoff and Henikoff, 1992) , **equalin** , **gtr**。由于蛋白质替换符合哪个模型事先未知，可以设定 **prset aamodelpr=mixed** , 则 MCMC 算法对前 10 个模型分别尝试 ( jumping between fixed-rate amino acid models ) , 每个模型对结果的贡献与其后验概率成正比。

**prset** 命令是用来设定各个参数的先验概率分布，可以根据你對自己数据的了解来设定，也可以设定一个无信息的先验概率分布（所有模型的概率都均等）。MrBayes 默认的是无信息的先验分布，所以如果对自己的数据没有先验经验，可以不做设定。想了解系统对各个参

数的先验设定，在控制符后输入 **help prset** 即可查看。

### 3.2 MCMC

在控制符后输入 **mcmc ngen=100000 samplefreq=100 printfreq=100 diagnfreq=1000**，点击回车，程序开始运行（图3.7），首先输出你设定的参数和待分析数据情况，然后输出抽样结果。

Ngen设定MCMC算法进行的循环代数，也就是对树型、枝长或进化模型参数改变的次数，这里设定为10万次；默认值为100万。因为每次对树型、枝长或进化模型参数改变的并不大，如果每个循环（每次改变）都抽样，输出文件太大，样本相似度太高；所以这里samplefreq抽样频率设定为100个循环抽一次，100000个循环将会从后验概率分布中抽得1000个样本；默认抽样频率为500代抽一次。printfreq是分析结果输出屏幕的频率，默认是500代输出一次。MCMC默认同时运行两个独立的分析（run），也就是从八个不同的起始树（initial tree，每个分析4棵起始树）开始运行。同时运行两个独立的分析，可以在运行过程中诊断结果是否收敛，刚开始两个分析抽样出来的树差异比较大，但随着代数的增加，它们会收敛，越来越相似，说明我们已经得到了一个好的后验概率分布样本，诊断频率diagnfreq设定每过多少代，检查一下两个分析结果的差异，默认为5000代一次。如果你序列数目比较少，很快就达到收敛状态，ngen可以设置的小一点儿，抽样和诊断频率高一点儿（samplefreq小一点儿）；反之如果你的样本比较大，可能需要更多代数才能达到收敛，这时抽样和诊断频率可以低一点儿（samplefreq大一点儿）。

```
991000 -- [-5733.089] (-5726.151) (-5725.184) (-5743.025) * (-5732.126) [-5720.168] (-5725.802) (-5726.181) -- 0:00:06
992000 -- [-5723.272] (-5729.899) (-5742.193) (-5730.360) * [-5727.533] (-5724.606) (-5725.310) (-5727.388) -- 0:00:05
993000 -- (-5722.341) (-5722.122) (-5731.578) [-5729.470] * (-5724.375) (-5735.846) (-5726.495) [-5727.530] -- 0:00:04
994000 -- (-5729.299) [-5726.288] (-5734.573) (-5735.619) * [-5716.533] (-5722.941) (-5726.585) (-5728.799) -- 0:00:04
995000 -- (-5719.794) (-5722.460) (-5724.074) [-5724.331] * [-5723.534] (-5729.552) (-5726.432) (-5725.525) -- 0:00:03

Average standard deviation of split frequencies: 0.000684

996000 -- [-5721.965] (-5725.975) (-5732.159) (-5733.311) * (-5728.729) [-5723.971] (-5729.363) (-5726.425) -- 0:00:02
997000 -- (-5723.149) (-5725.954) (-5726.135) [-5722.036] * (-5725.540) (-5726.572) (-5727.264) [-5720.958] -- 0:00:02
998000 -- (-5732.913) (-5727.345) [-5721.449] (-5723.231) * (-5727.286) [-5727.528] (-5722.218) (-5724.112) -- 0:00:01
999000 -- (-5722.435) [-5722.312] (-5726.212) (-5727.267) * (-5728.184) (-5725.502) [-5723.982] (-5737.912) -- 0:00:00
1000000 -- (-5739.162) [-5727.845] (-5731.388) (-5727.674) * (-5721.851) (-5724.368) [-5716.052] (-5732.175) -- 0:00:00

Average standard deviation of split frequencies: 0.000680
Continue with analysis? (yes/no): n
```

图 3.7 MCMC 运行界面（部分）



图 3.7 第一列是代数（也即循环数），后面四列数字分别为第一轮运行的四条链各自的对数概率值（log likelihood），[ ]内的对应的是冷链(cold chain)，( )内的是热链(heated chains)。注意冷热线间应该能变换状态（图中上下两行间[]的位置发生变化），如果不能，则算法效率不高，可能需要延长代数或降低冷热线间温度差距（Temp）。\*号后的是第二轮运行的四条链的情况。最后一列是估计的完成设定代数还需要的时间。

运行结束后如果 standard deviation of split frequencies 小于 0.01，在程序询问“Continue the analysis?” 时回答 no,否则答 yes,增加世代数继续运行至小于 0.01。

运行结束可以看到在输出界面有一个链交换信息（chain swap information），矩阵左上部分是交换频率，这些值如果在 0.1~0.8 之间，说明结果合理，否则要重新设置增加参数，如增加代数 ngen，降低 Temp 等。

这时在 MrBayes 可执行程序文件夹内可看到生成五个文件，扩展名为 **mcmc** 的文件一个，扩展名为 **p** 和 **t** 的文件各两个，每个分析（run）一个，**mcmc** 文件记录的是抽样的信息，**p** 文件记录了每个抽样的模型参数，可以用写字板打开查看；**t** 文件是树型和枝长数据，可以用 TreeView 打开，每个 run 有一个树文件，里面包含了 1000 棵树。

```
Chain swap information for run 1:
      1      2      3      4
-----
1 |      0.77    0.58    0.42
2 | 166449      0.79    0.60
3 | 166818 166613      0.80
4 | 166974 166753 166393

Chain swap information for run 2:
      1      2      3      4
-----
1 |      0.77    0.58    0.42
2 | 166537      0.79    0.60
3 | 166930 166666      0.80
4 | 166949 166554 166364

Upper diagonal: Proportion of successful state exchanges between chains
Lower diagonal: Number of attempted state exchanges between chains
```

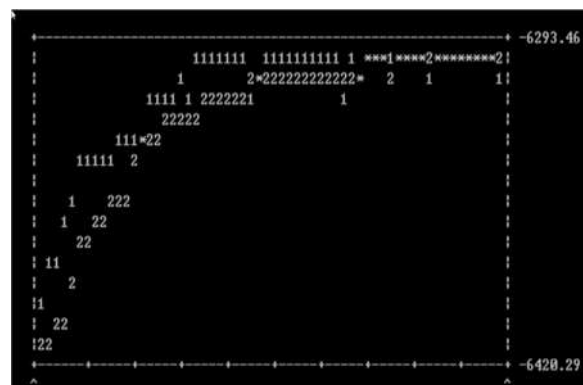
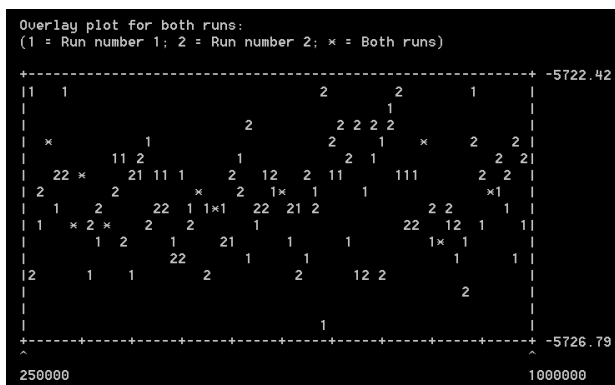
图 3.8 MCMC 运行结束输出界面（部分）

### 3.3 归纳参数和树型（sump & sumt）



在控制符后输入**sump burnin=250**，点击回车运行。

Sump对MCMC抽样的替换模型参数 (p文件) 进行总结，由于起始阶段的树型概率并不高，不是最优树，所以要将这部分样本抛弃。burnin后面的250就是这里要抛弃的样本数量。程序默认是将样本的前25%抛弃( ( burninfrac = 0.25 ) )后再总结。你可以改变burnin的数量或比例，如果就用默认的25%，则可以只输入**sump**即可。输出结果包括三部分内容：首先是数据的log-probability随世代数变化的图，图上的点应该没有明显趋势，说明抽样到达了平稳状态( stationarity )( 左图 )如果你的分析数据显示随世代数增加，log-probability有上升或下降的趋势（右图），那么需要增加代数，继续分析；第二部分是两轮MCMC分析的边缘概率估计；第三部分是用到的各个参数取值情况。



```

Estimated marginal likelihoods for runs sampled in files
"examples/primates.nex.run1.p" and "examples/primates.nex.run2.p":
(Use the harmonic mean for Bayes factor comparisons of models)

(Values are saved to the file examples/primates.nex.lstat)

```

Run	Arithmetic mean	Harmonic mean
1	-5719.47	-5735.92
2	-5719.31	-5732.49
TOTAL	-5719.39	-5735.25

Parameter	Mean	Variance	95% HPD Interval		Median	min ESS*	avg ESS	PSRF+
			Lower	Upper				
TL	3.197476	0.105942	2.617705	3.863267	3.168768	536.92	631.70	1.000
r(A<->C)	0.041935	0.000068	0.025966	0.057986	0.041625	606.06	636.19	1.000
r(A<->G)	0.491749	0.002216	0.392388	0.578128	0.490881	526.20	527.98	1.001
r(A<->T)	0.035252	0.000063	0.020835	0.051790	0.034887	818.20	839.70	1.000
r(C<->G)	0.029289	0.000170	0.005504	0.054835	0.028177	631.66	712.88	1.000
r(C<->T)	0.383365	0.001790	0.305871	0.472657	0.384067	561.66	649.37	1.001
r(G<->T)	0.018410	0.000148	0.000025	0.041289	0.016416	671.11	804.56	1.000
pi(A)	0.355009	0.000167	0.330407	0.380851	0.354649	813.37	943.68	1.000
pi(C)	0.321981	0.000130	0.298581	0.342770	0.321918	719.21	888.92	1.000
pi(G)	0.080009	0.000045	0.067008	0.093365	0.079702	568.58	643.34	1.001
pi(T)	0.243001	0.000111	0.221855	0.261997	0.242966	685.27	817.25	1.000
alpha	0.597019	0.032791	0.339827	0.951364	0.559231	295.48	403.03	1.000
pinvar	0.146899	0.006910	0.000466	0.281331	0.148732	219.40	364.10	1.000

\* 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.

图 3.9 Summarize the parameter 结果

图 3.9 第三张图片共输出了 13 个参数的情况，第一行 TL 是进化树的总枝长，接下来 6 个 r 是两两核苷酸替换速率，再下来 4 个 pi 是四种核苷酸频率，alpha 是 gamma 分布的形状参数，pinvar 是不变位点所占比例。每个参数分别输出其均值 mean，方差 variance，95% 置信区间的 lower 和 upper 值，中位数 median，如果 MCMC 运行结果收敛，最后一栏 PSRF(potential scale reduction factor)值应该在 1 左右。

然后可以通过 **sumt burnin=250** 对所有抽样树型进行总结（同样前 25% 的样本被抛弃），程序对树型进行总结，输出所有枝长参数，同时输出 cladogram，上面标注了每个分枝(clade)的后验概率，还输出 phylogram，各枝的长度代表对应的进化距离（图 3.10）。

Parameter	Mean	Variance	95% HPD Interval		Median	PSRF+	Nruns
			Lower	Upper			
length[1]	0.538764	0.009404	0.358520	0.725100	0.527726	1.000	2
length[2]	0.376980	0.006097	0.231891	0.529469	0.369955	1.000	2
length[3]	0.051161	0.000142	0.029824	0.076353	0.050331	1.001	2
length[4]	0.064179	0.000159	0.040483	0.090092	0.063425	1.000	2
length[5]	0.061580	0.000237	0.031666	0.091046	0.060528	1.001	2
length[6]	0.154605	0.000808	0.098458	0.205834	0.152749	1.000	2
length[7]	0.181388	0.001166	0.116921	0.251855	0.179496	1.001	2
length[8]	0.017657	0.000036	0.006761	0.029288	0.017166	1.000	2
length[9]	0.023755	0.000047	0.010876	0.036815	0.023276	1.001	2
length[10]	0.058620	0.000154	0.035715	0.082716	0.057836	1.002	2
length[11]	0.074368	0.000466	0.032970	0.115791	0.073640	1.001	2
length[12]	0.477032	0.007222	0.321181	0.646441	0.470236	1.000	2
length[13]	0.278032	0.002813	0.179277	0.382383	0.274033	1.000	2
length[14]	0.036896	0.000123	0.016918	0.058582	0.035963	1.000	2
length[15]	0.088766	0.000512	0.047601	0.133482	0.086762	1.000	2
length[16]	0.136878	0.001487	0.062898	0.210760	0.133331	1.000	2
length[17]	0.303625	0.005531	0.170759	0.457443	0.298724	1.000	2
length[18]	0.030680	0.000148	0.010252	0.056799	0.029353	1.000	2
length[19]	0.059975	0.000505	0.018600	0.103012	0.057780	1.000	2
length[20]	0.050194	0.000374	0.012785	0.087570	0.048561	1.001	2
length[21]	0.132916	0.002410	0.046517	0.235450	0.129603	1.000	2

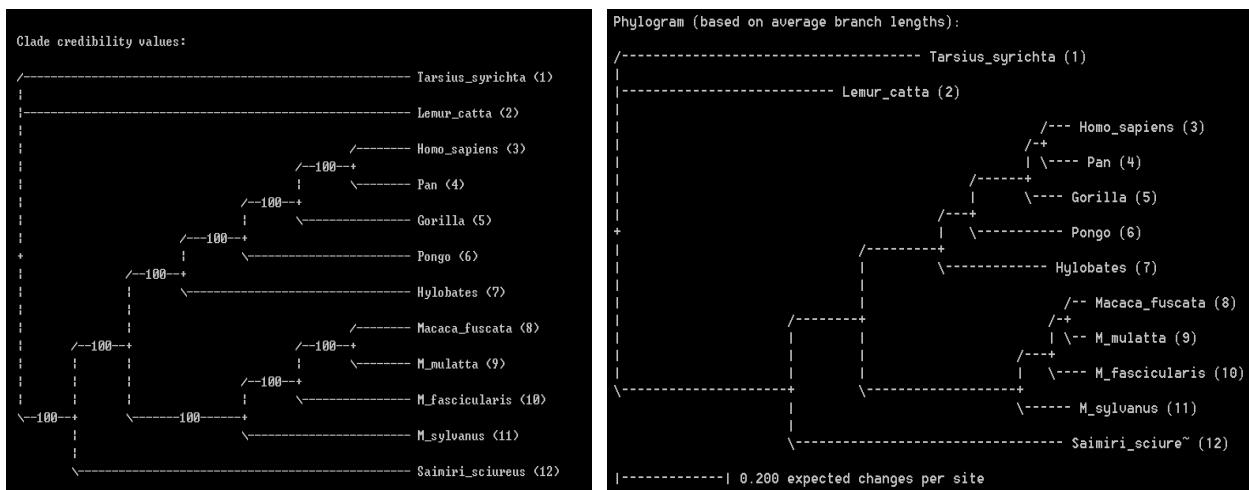


图 3.10 Summarize Trees 结果

观察程序所在文件夹，可以看到生成了六个文件，扩展名为 **parts** 的文件里记录的是构建出来的进化树的分枝类型 ( branch pattern, partition )。扩展名为 **tstat**、**vstat** 和 **lstat** 的文件分别总结了各种树型 ( topology )、枝长(branch length)和似然值 ( marginal likelihood ) 的统计数据，这三个文件可用写字板打开查看；扩展名为 **con.tre** 的文件就是总结出来的一致树，可用 Treeview 软件打开查看，可看到各分枝的后验概率及枝长 ( standard deviation from posterior probability ) 信息；**trprobs** 文件包含了 MCMC 搜索过程中得到的树，且按照后验概率大小排序。