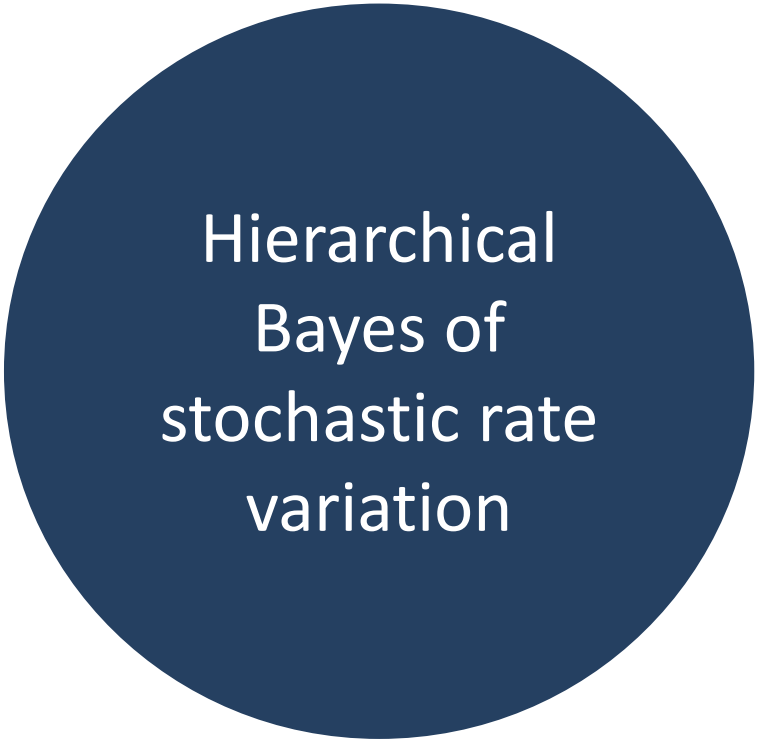
Rates of molecular evolution


Hierarchical
Bayes of
stochastic rate
variation

Mutation rate, functional constraints, and rate of molecular evolution

$$r = \mu \times 2N \times f$$

$$= \mu \times 2N \times p \times \frac{1}{2N}$$

$$= \mu p$$

p: proportion of neutral mutations

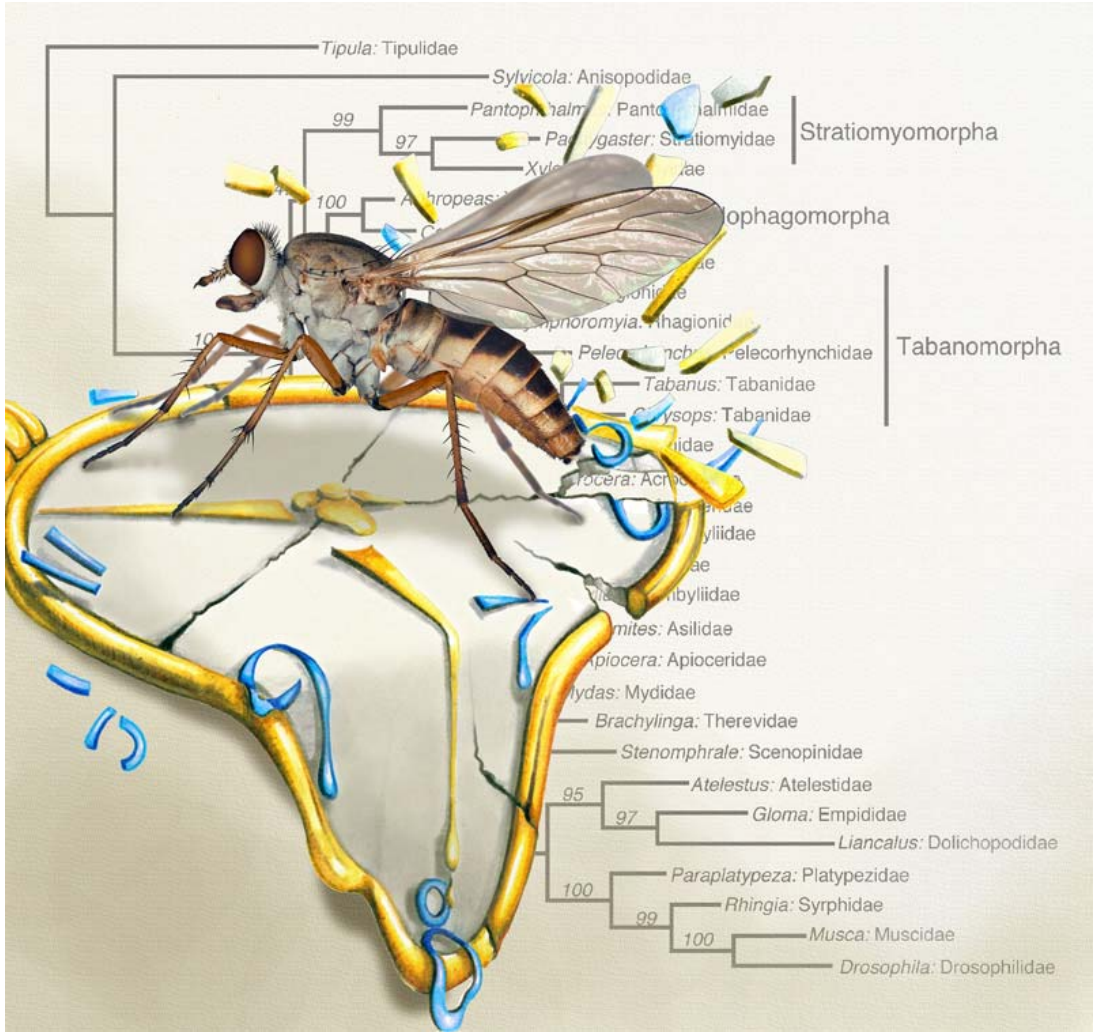
Mutation rate, functional constraints, and rate of molecular evolution

$$r = \mu \times 2N \times f$$

$$= \mu \times 2N \times p \times \frac{1}{2N}$$

$$= \mu p$$

p: proportion of neutral mutations



Time flies: tree of 28S rDNA

Wiegmann et al (2003).
Systematic Biology. **52**: 745-756

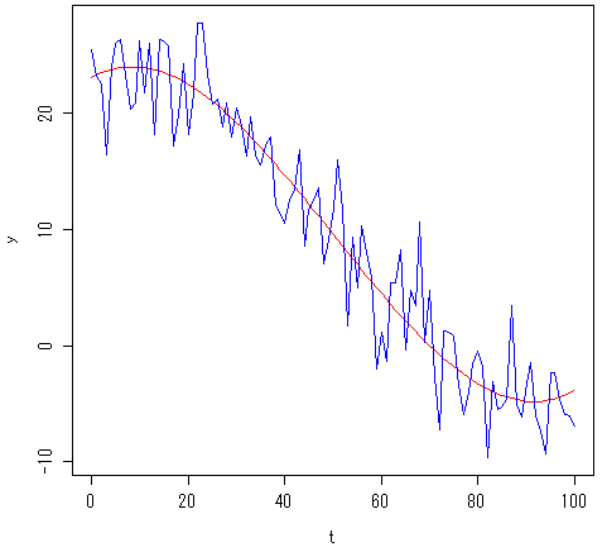
The case of trend estimation and hierarchical Bayes

$$y_t = T_t + \varepsilon_t$$

Strength of penalty

Minimize $S = \sum_t (y_t - T_t)^2 + \lambda \sum_t (T_t - T_{t-1})^2$

Fitting to the data Penalty on the local variation



Maximize $\exp\left(-\frac{S}{2\sigma^2}\right)$

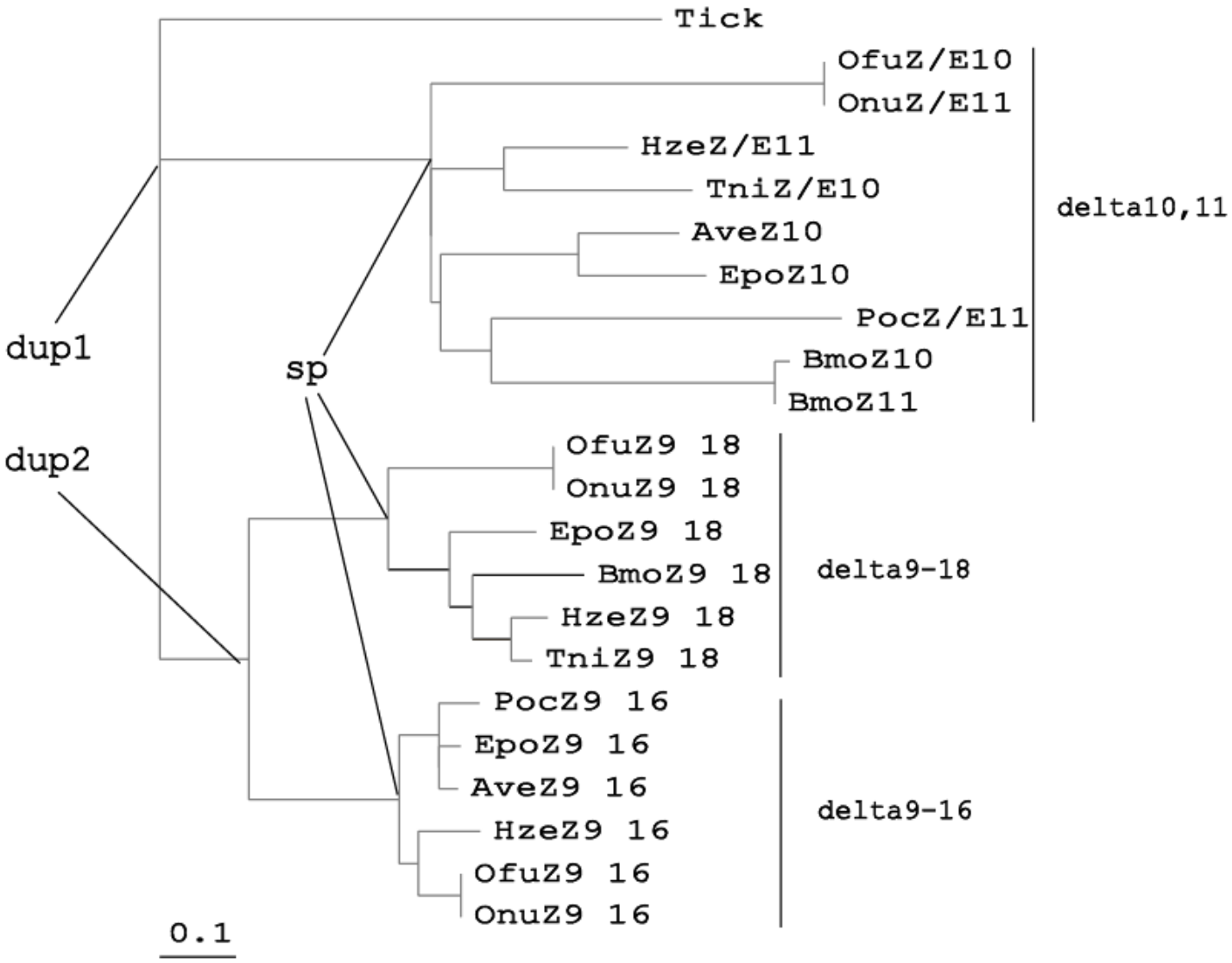
Hyper parameter

$$\propto \prod_t \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - T_t)^2}{2\sigma^2}\right) \times \prod_t \frac{1}{\sqrt{2\pi\sigma^2/\lambda}} \exp\left(-\frac{\lambda(T_t - T_{t-1})^2}{2\sigma^2}\right)$$

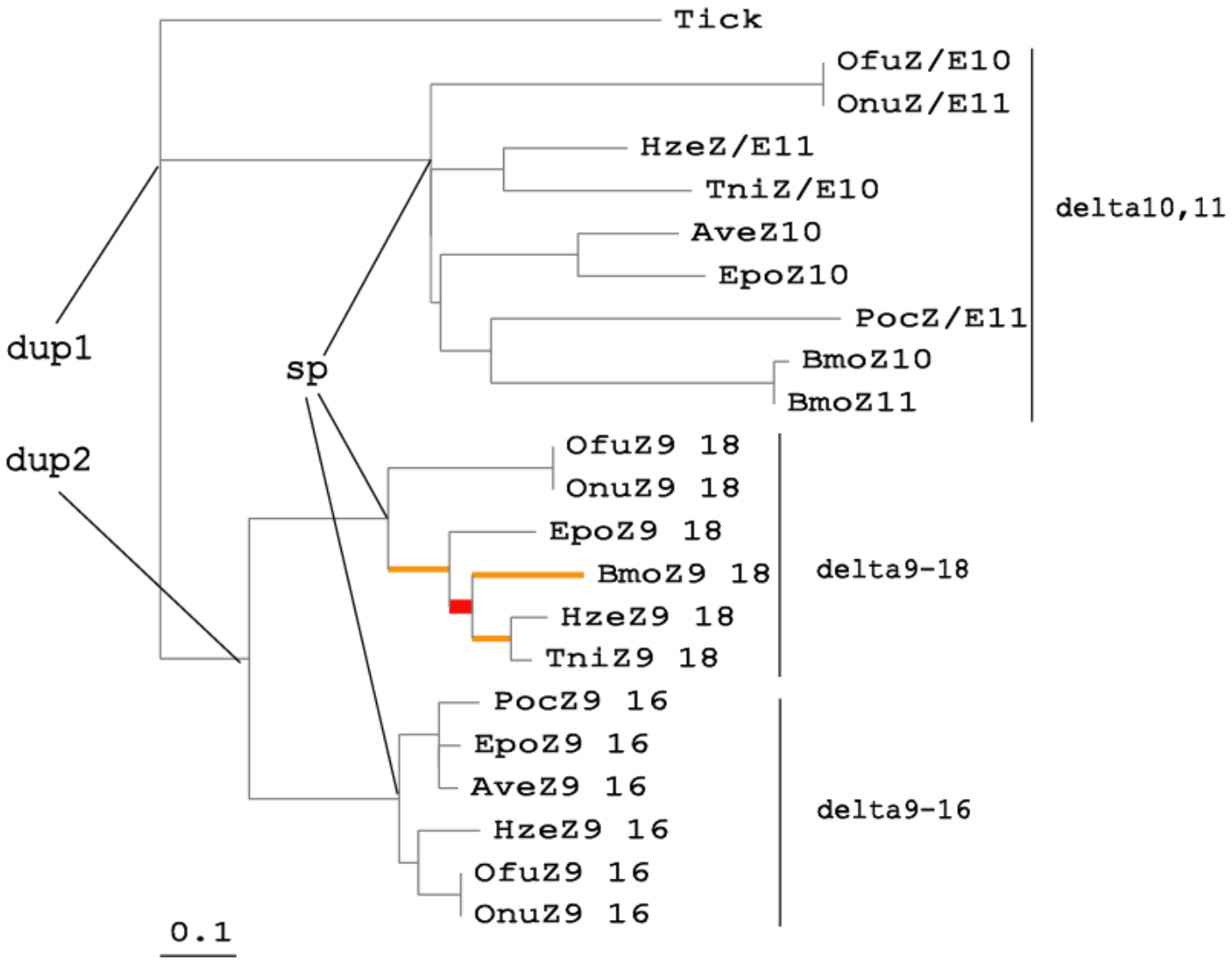
likelihood Prior distribution

Hierarchical Bayes model

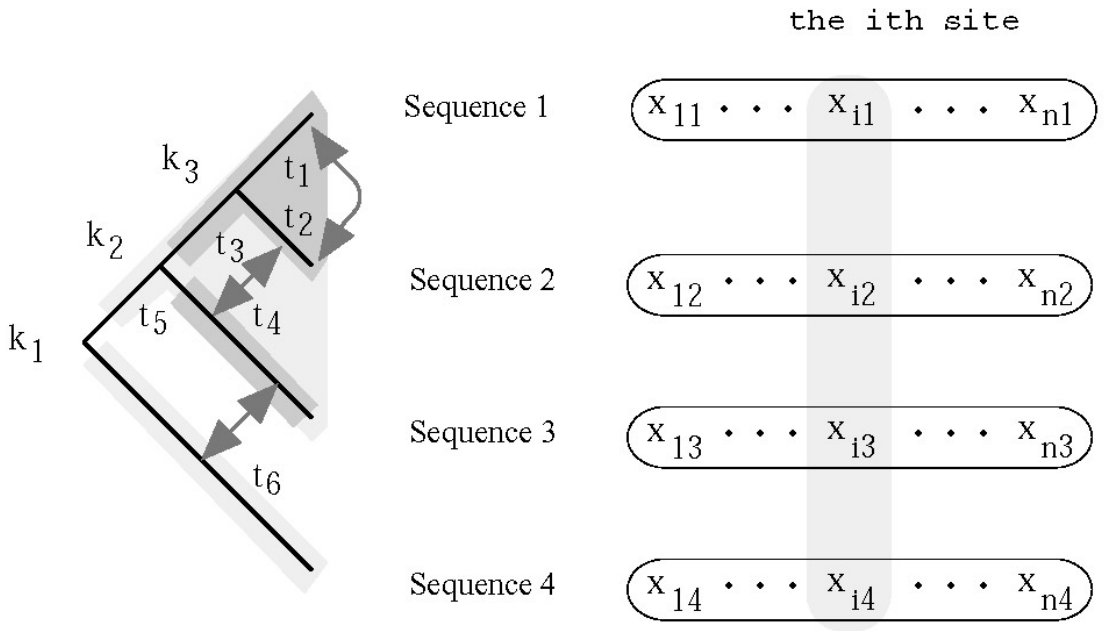
Desaturase and the origin of sex pheromone:



Desaturase and the origin of sex pheromone:



Likelihood of the phylogenetic tree



$$
\begin{aligned}
& f(x_{i1},\ldots,x_{i4}|\lambda,t_1,\ldots,t_6) \\
= & \sum_{k_1}\pi_{k_1}p_{k_1,x_{i4}}(t_6)\Big(\sum_{k_2}p_{k_1,k_2}(t_5)\Big(p_{k_2,x_{i3}}(t_4)\Big(\sum_{k_3}p_{k_2,k_3}(t_3)\big(p_{k_3,x_{i2}}(t_2)p_{k_3,x_{i1}}(t_1)\big)\Big)\Big)\Big)
\end{aligned}
$$

The likelihood of a site consists of the transition probabilities.

Hierarchical Bayes models of stochastic rate variation

$\mathbf{X} = (X_1, \mathrm{K}\ , X_n)'$ — data

$P(X \mid \mathbf{rt})$ — likelihood

$p(\mathbf{r} \mid \nu)$

$t = t_0 \boldsymbol{\pi}$

Prior

$\tilde{r}(t) \equiv \log r(t)$

$E[\tilde{r}(t) \mid \tilde{r}(s)] = \tilde{r}(s) - \frac{\nu}{2}(t-s)$

$Var[\tilde{r}(t) \mid \tilde{r}(s)] = \nu(t-s)$

$\approx CV[r(t) \mid r(s)]$

$p(t_0)$ Gamma

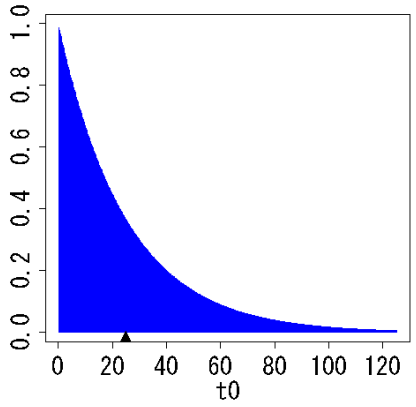
$$p(\boldsymbol{\pi}) = \frac{\Gamma\left(\sum_{m=1}^{M} \alpha_m\right)}{\prod_{m=1}^{M} \Gamma(\alpha_m)} \prod_{m=1}^{M} \pi_m^{\alpha_m - 1}$$

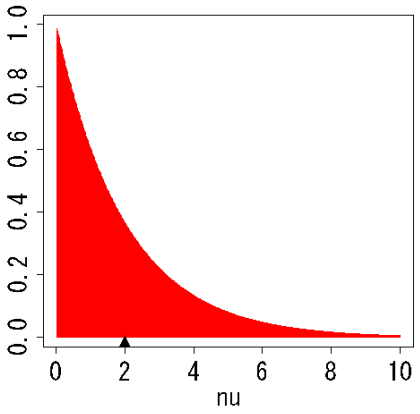
ν : the size of rate variation

$p(\nu)$

Hyper-prior

Thorne et al. (1998)
MBE 15: 1647-1657
Drummond et al. (2006).
PLoS Biol. 4(5): e88.
Rannala and Yang (2007)
Syst. Biol. 56: 453–466





Hierarchical Bayes models of stochastic rate variation

Penalty for the residuals of molecular clock hypothesis

$$\log L - \lambda \sum \left(\tilde{r}_i - \tilde{r}_{a(i)}\right)^2 \;\Leftrightarrow\; L \times \left(\frac{\lambda}{\pi}\right)^{\frac{n}{2}} \exp\left(-\lambda \sum \left(\tilde{r}_i - \tilde{r}_{a(i)}\right)^2\right)$$

auto-correlated log-normal

$$\log L - \lambda \sum \left(\tilde{r}_i - \mu\right)^2 \;\Leftrightarrow\; L \times \left(\frac{\lambda}{\pi}\right)^{\frac{n}{2}} \exp\left(-\lambda \sum \left(\tilde{r}_i - \mu\right)^2\right)$$

uncorrelated log-normal

$$\log L - \lambda \sum \left|\tilde{r}_i - \mu\right| \;\Leftrightarrow\; L \times \left(\frac{\lambda}{2}\right)^{n} \exp\left(-\lambda \sum \left|\tilde{r}_i - \mu\right|\right)$$

uncorrelated (double) exponential

- correlated vs uncorrelated
- lognormal vs exponential
- random local clock

Hierarchical Bayes models of stochastic rate variation

- correlated vs uncorrelated
- lognormal vs exponential
- random local clock

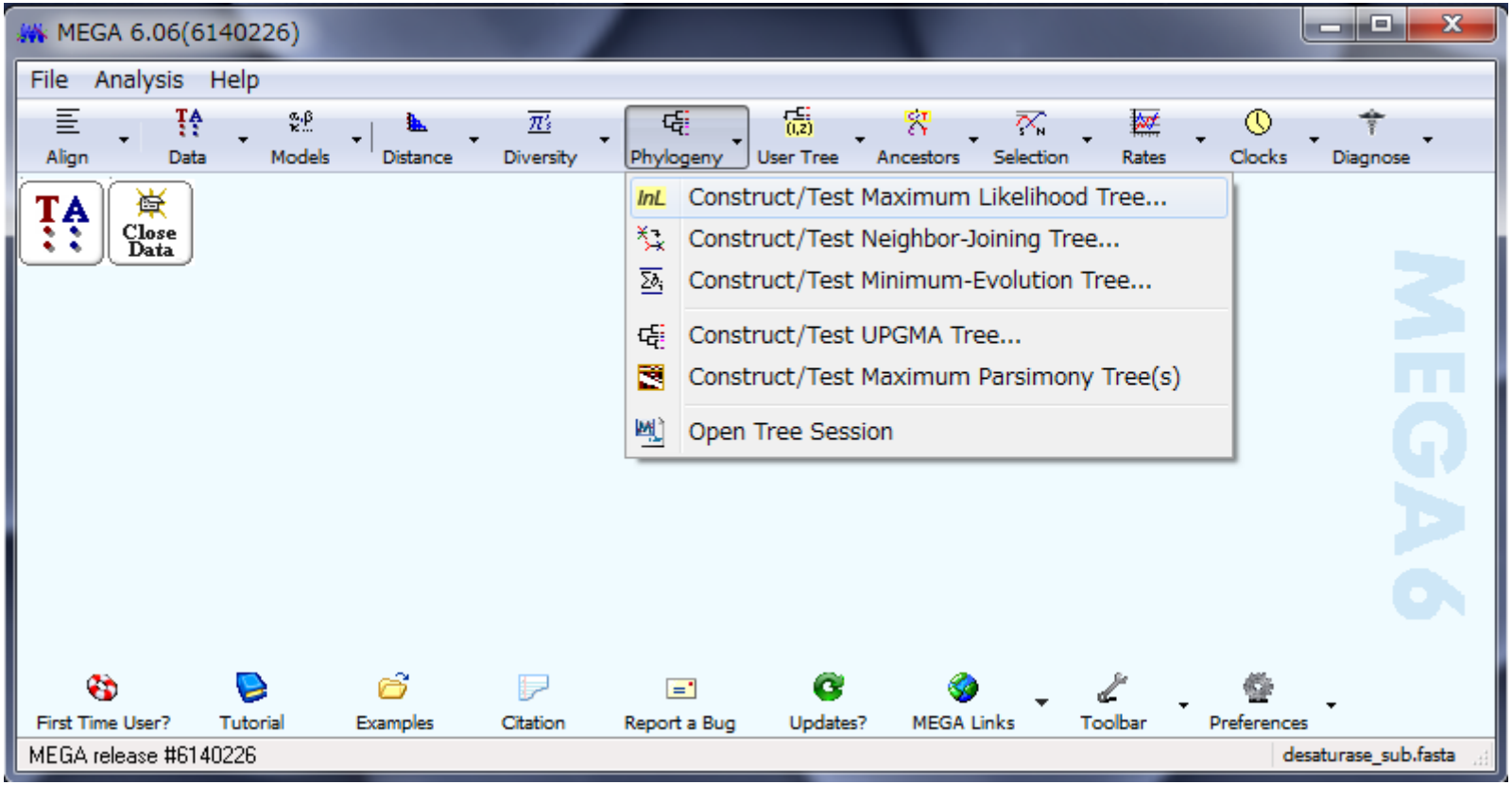
Phylogenetic tree of desaturase

desaturase.fasta

```
>HzeZ9_18
MPPQGQTGGSWVLYETDAVNEDTDAPVIVPPSAEKREWKIVWRNVILMGMLHIGGVYGAY
LFLTTAMWRTCIFAVVLYICSGLGITAGAHRLWAHKSYKARLPLRLMLTLFNTLAFQDAV
IDWARDHRMHHKYSETDADPHNATRGFFFAHVGWLLVRKHPQIKAKGHTIDLSDLKSDPI
LRFQKKYYLFLMPLVCFILPCYIPT-LWGESLWNAYFVCSIFRYVYVLNVTWLVNSAAHL
WGAKPYDKNINPVETRPVSLVVLGEGFHNYHHTFPWDYKTAELGDYSLNLTKLFIDTMAA
IGWAYDLKTVSTDVIQKRVKRTGDGSHPVWGWDDHEVHQADKKLAAIINPEKT
>TniZ9_18
MPPQGQTGGSWVLYETDAVNTDTDAPVIVPPSAEKREWKIVWRNVILMGMLHIGGVYGAY
LFLTKAMWLTDLFAFFLYLCSGLGITAGAHRLWAHKSYKARLPLRLLLTLFNTLAFQDAV
IDWARDHRMHHKYSETDADPHNATRGFFFSHVGWLLVRKHPQIKAKGHTIDLSDLKSDPI
LRFQKKYYLTLMPLICFILPSYIPT-LWGESAFNAFFVCSIFRYVYVLNVTWLVNSAAHL
WGSKPYDKNINPVETRPVSLVVLGEGFHNYHHTFPWDYKTAELGDYSLNFTKMFIDFMAS
IGWAYDLKTVSTDVIQKRVKRTGDGSHAVWGWDDHEVHQEDKKLAAIINPEKT
. . . . . .
>HzeZ9_18
MPPQGQTGGSWVLYETDAVNEDTDAPVIVPPSAEKREWKIVWRNVILMGMLHIGGVYGAY
LFLTTAMWRTCIFAVVLYICSGLGITAGAHRLWAHKSYKARLPLRLMLTLFNTLAFQDAV
IDWARDHRMHHKYSETDADPHNATRGFFFAHVGWLLVRKHPQIKAKGHTIDLSDLKSDPI
LRFQKKYYLFLMPLVCFILPCYIPT-LWGESLWNAYFVCSIFRYVYVLNVTWLVNSAAHL
WGAKPYDKNINPVETRPVSLVVLGEGFHNYHHTFPWDYKTAELGDYSLNLTKLFIDTMAA
IGWAYDLKTVSTDVIQKRVKRTGDGSHPVWGWDDHEVHQADKKLAAIINPEKT
>TniZ9_18
MPPQGQTGGSWVLYETDAVNTDTDAPVIVPPSAEKREWKIVWRNVILMGMLHIGGVYGAY
LFLTKAMWLTDLFAFFLYLCSGLGITAGAHRLWAHKSYKARLPLRLLLTLFNTLAFQDAV
IDWARDHRMHHKYSETDADPHNATRGFFFSHVGWLLVRKHPQIKAKGHTIDLSDLKSDPI
LRFQKKYYLTLMPLICFILPSYIPT-LWGESAFNAFFVCSIFRYVYVLNVTWLVNSAAHL
WGSKPYDKNINPVETRPVSLVVLGEGFHNYHHTFPWDYKTAELGDYSLNFTKMFIDFMAS
IGWAYDLKTVSTDVIQKRVKRTGDGSHAVWGWDDHEVHQEDKKLAAIINPEKT
```

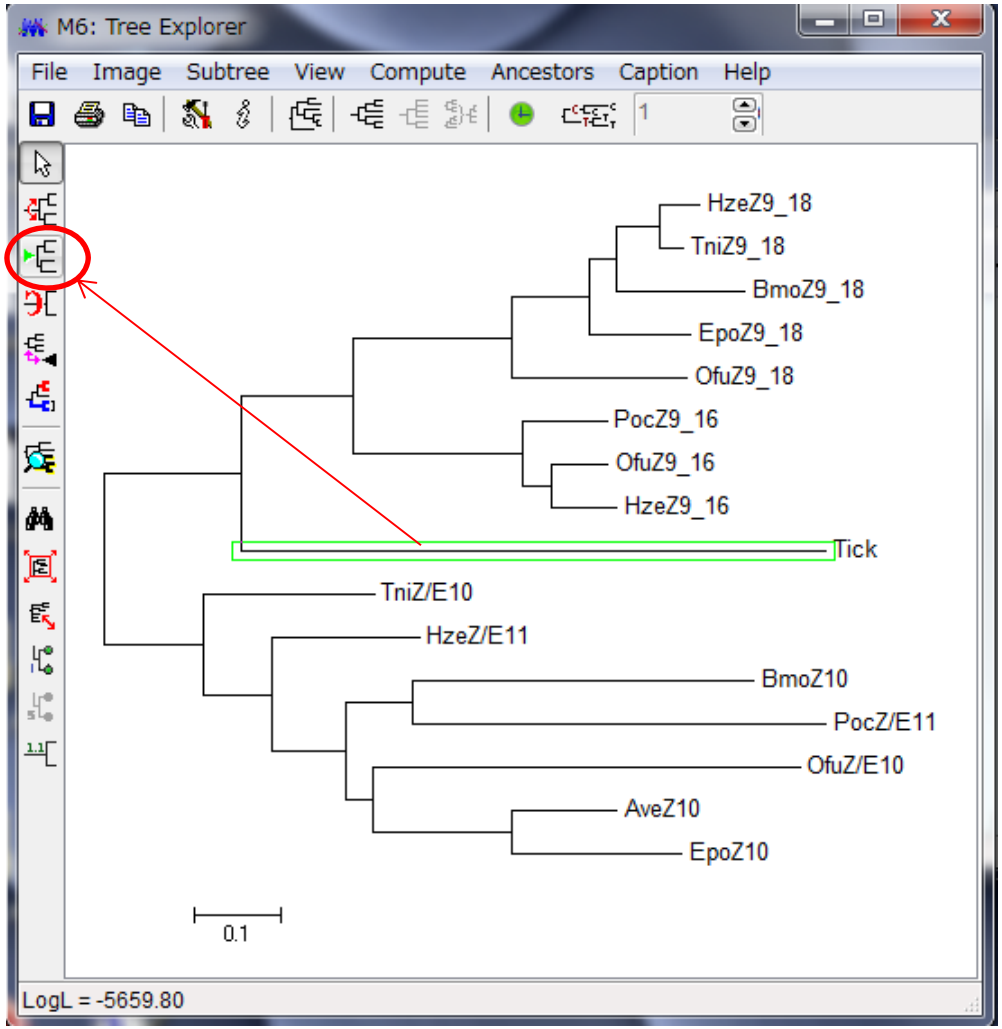
Analyze desaturase_sub.fasta by MEGA

Construct the phylogenetic tree by LG+G model



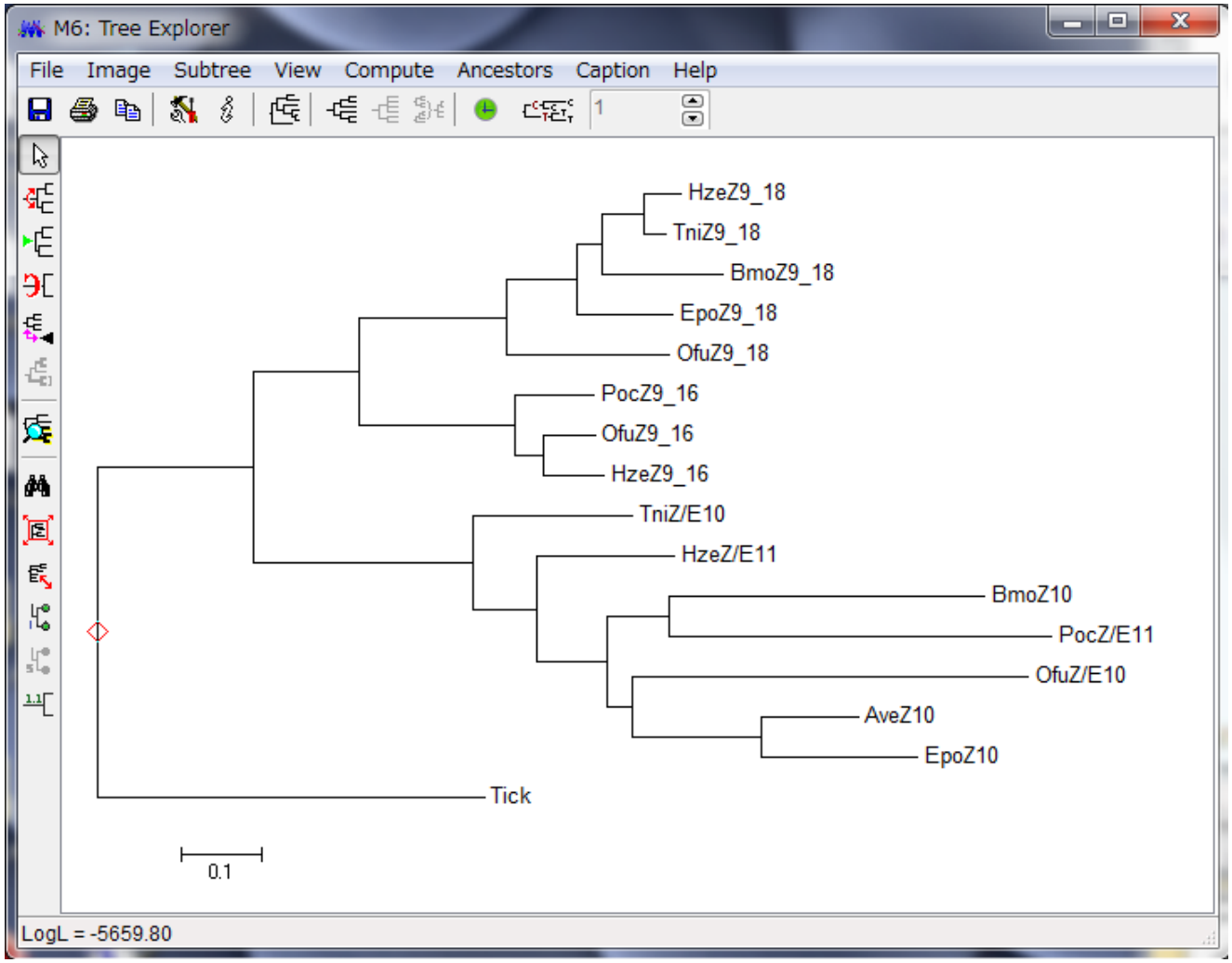
Analyze desaturase_sub.fasta by MEGA

Set Tick to be the outgroup

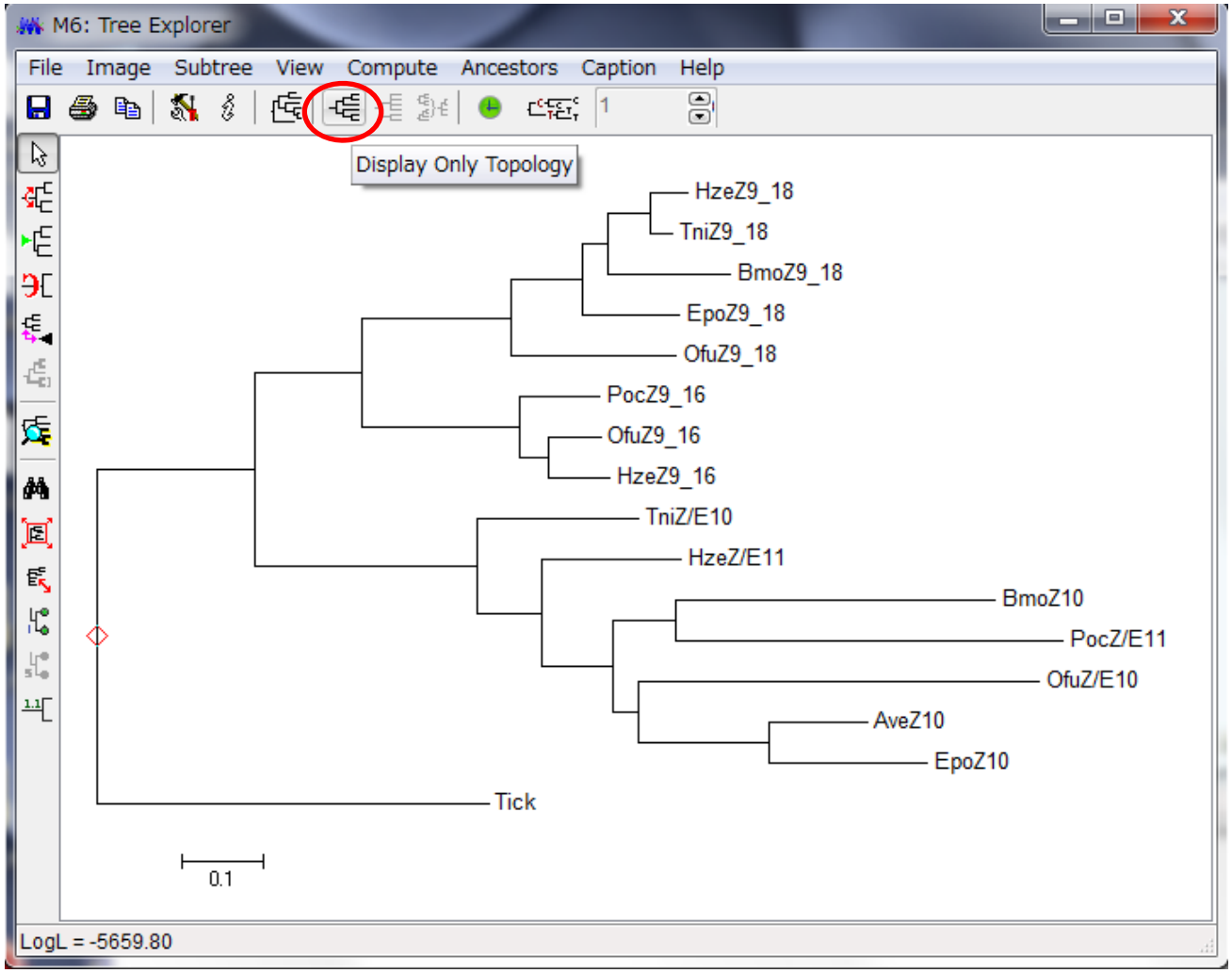


Analyze desaturase_sub.fasta by MEGA

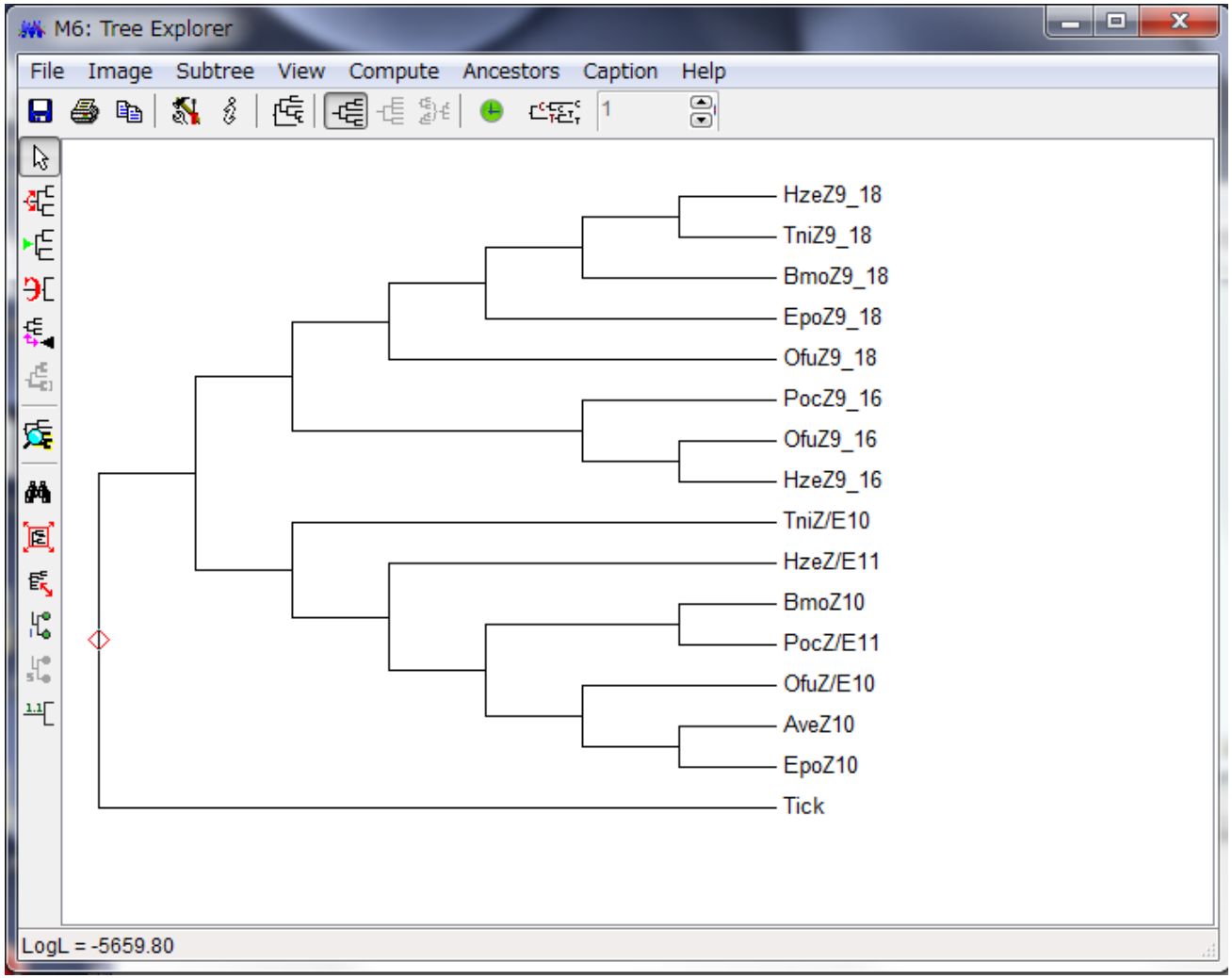
Set Tick to be the outgroup, and get the rooted tree



Extract the topology without the information on the branch lengths

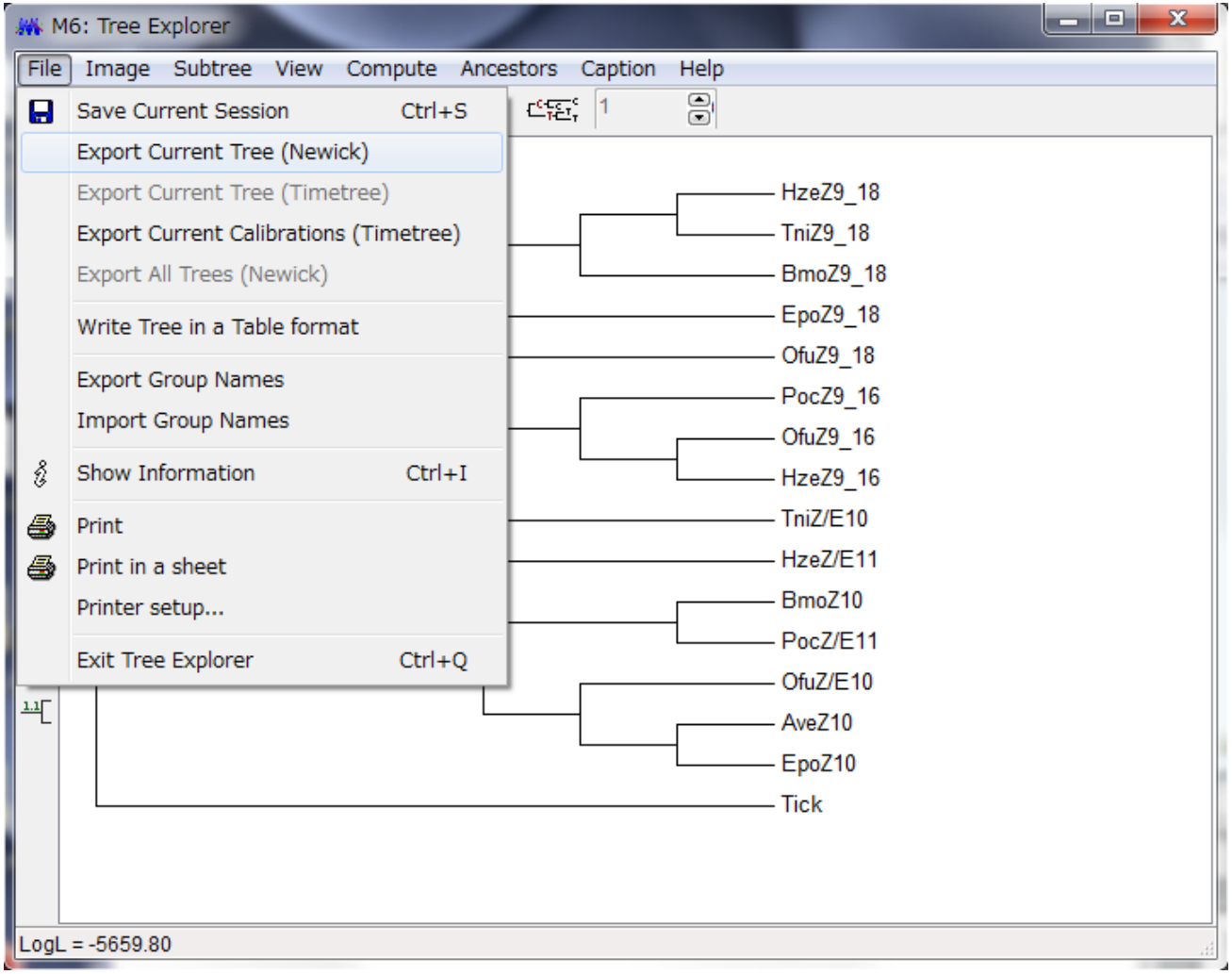


Extract the topology without the information on the branch lengths



Save the topology as desaturase_sub.nwk

[File][Export Current Tree (Newick)]



desaturase_sub.nwk

```
(((((((HzeZ9_18,TniZ9_18),BmoZ9_18),EpoZ9_18),OfuZ9_18),(PocZ9_16,(OfuZ9_16,HzeZ9_16))),(TniZ/E10,(HzeZ/E11,((BmoZ10,PocZ/E11),(OfuZ/E10,(AveZ10,EpoZ10))))),Tick);
```

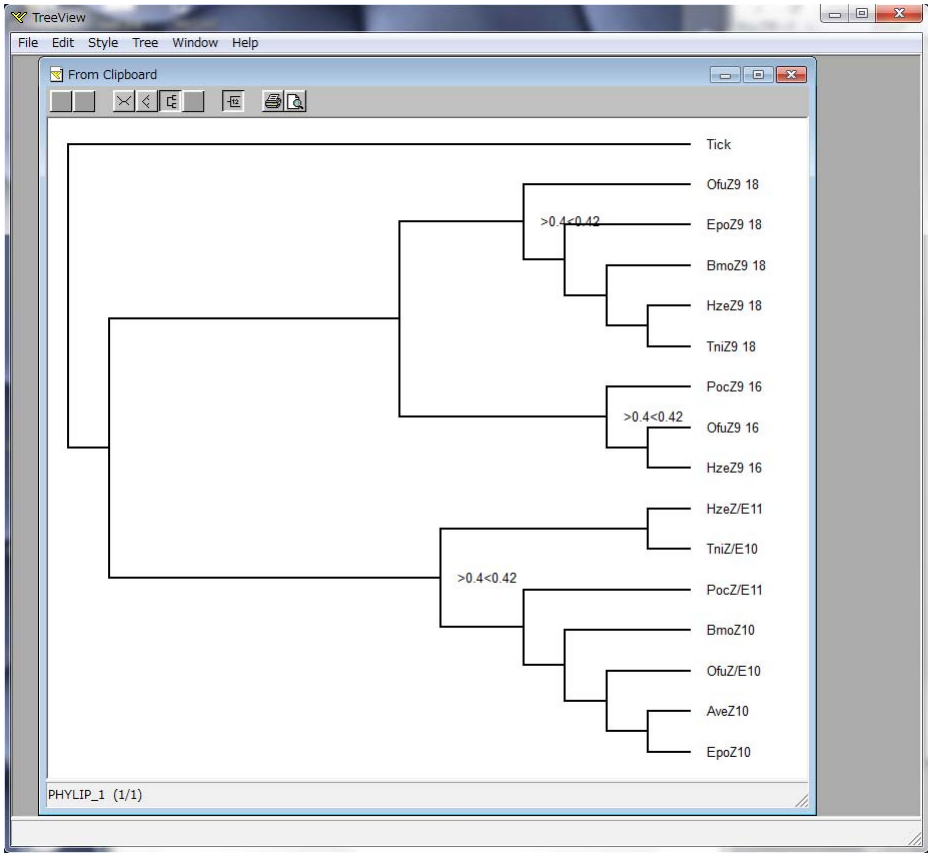
desaturase_sub.nwk

Specify the constraints on some reference node(s)

Set the constraints on the three nodes corresponding to the origin of Lepidoptera

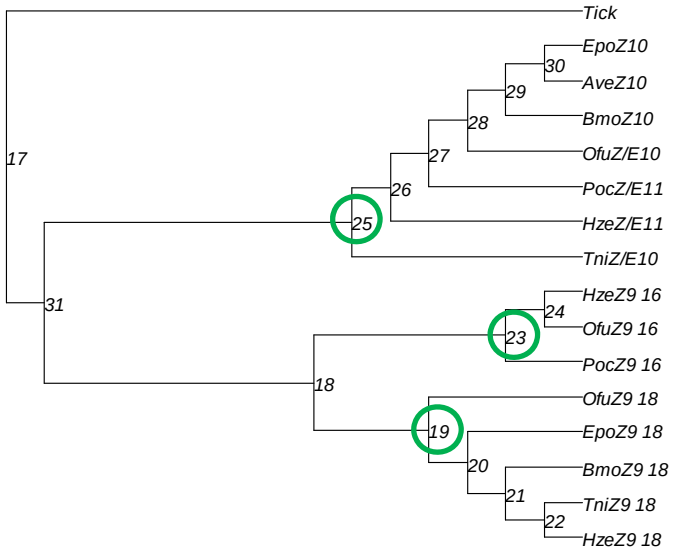
(((((((HzeZ9_18,TniZ9_18),BmoZ9_18),EpoZ9_18),OfuZ9_18)>0.4<0.42,(PocZ9_16,(OfuZ9_16,HzeZ9_16))>0.4<0.42),((HzeZ/E11,TniZ/E10),(PocZ/E11,(BmoZ10,(OfuZ/E10,(AveZ10,EpoZ10)))))>0.4<0.42),Tick);

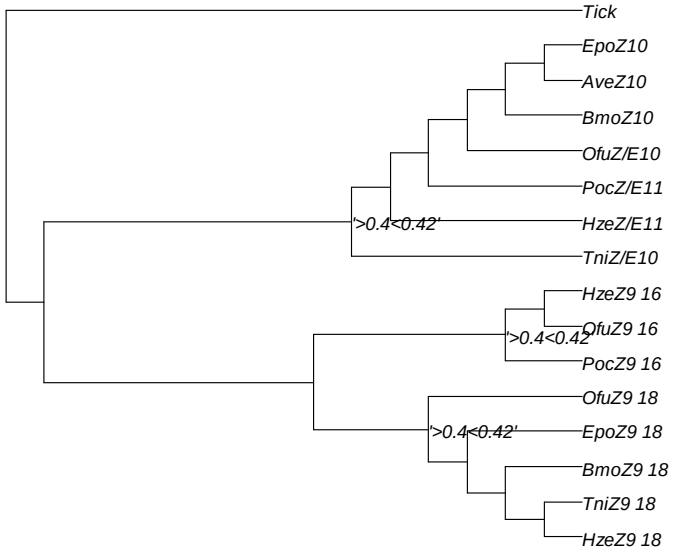
Check by TreeView that the constraints are specified on the correct nodes.



Alternatively, specify the constraints by setting the node labels in ape

```
library(ape)
read.tree("data//desaturase_sub.nwk")->tree
summary(tree)
is.rooted(tree)
tree$tip.label
tree <- root(tree,outgroup=16, resolve.root = T)
is.rooted(tree)
plot(tree)
summary(tree)
tree0$node.label <- 16+1:15
plot(tree0,show.node.label=T)


tree
tree$node.label <- rep("",15)
tree$node.label[tree0$node.label%in%c(19,23,25)] <- "'>0.4<0.42'"
plot(tree,show.node.label=T,main="tree")

write.tree(tree,"desaturase_sub_node_ref_for_paml.nwk")
```





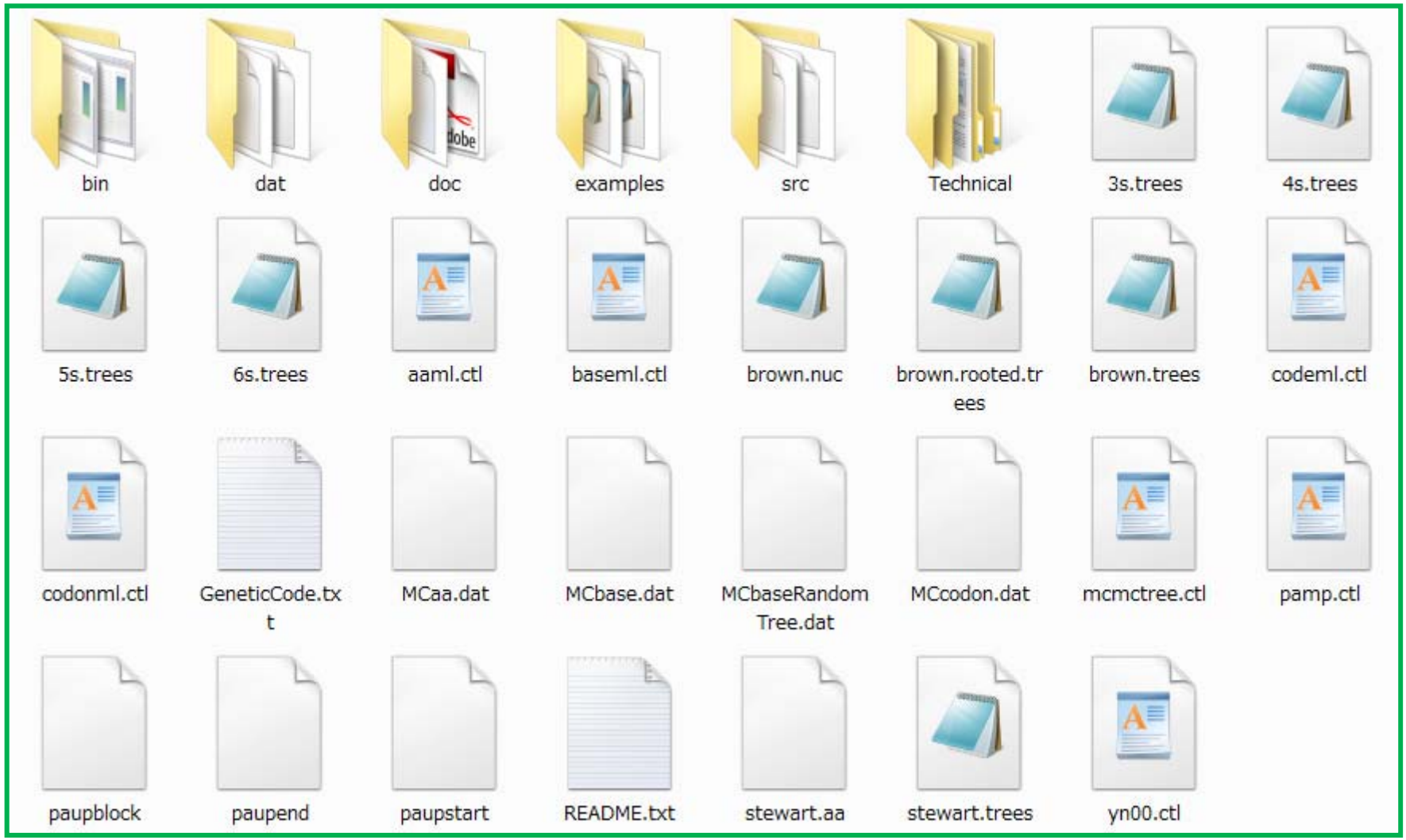
Write the number of sequences (16) and save as a nwk file

```
16
(((((((HzeZ9_18,TniZ9_18),BmoZ9_18),EpoZ9_18),OfuZ9_18)>0.4<0.42,(PocZ9_16,(OfuZ9_16,HzeZ9_16))>0.4<0.42),((HzeZ/E11,TniZ/E10),(PocZ/E11,(BmoZ10,(OfuZ/E10,(AveZ10,EpoZ10)))))>0.4<0.42),Tick);
```

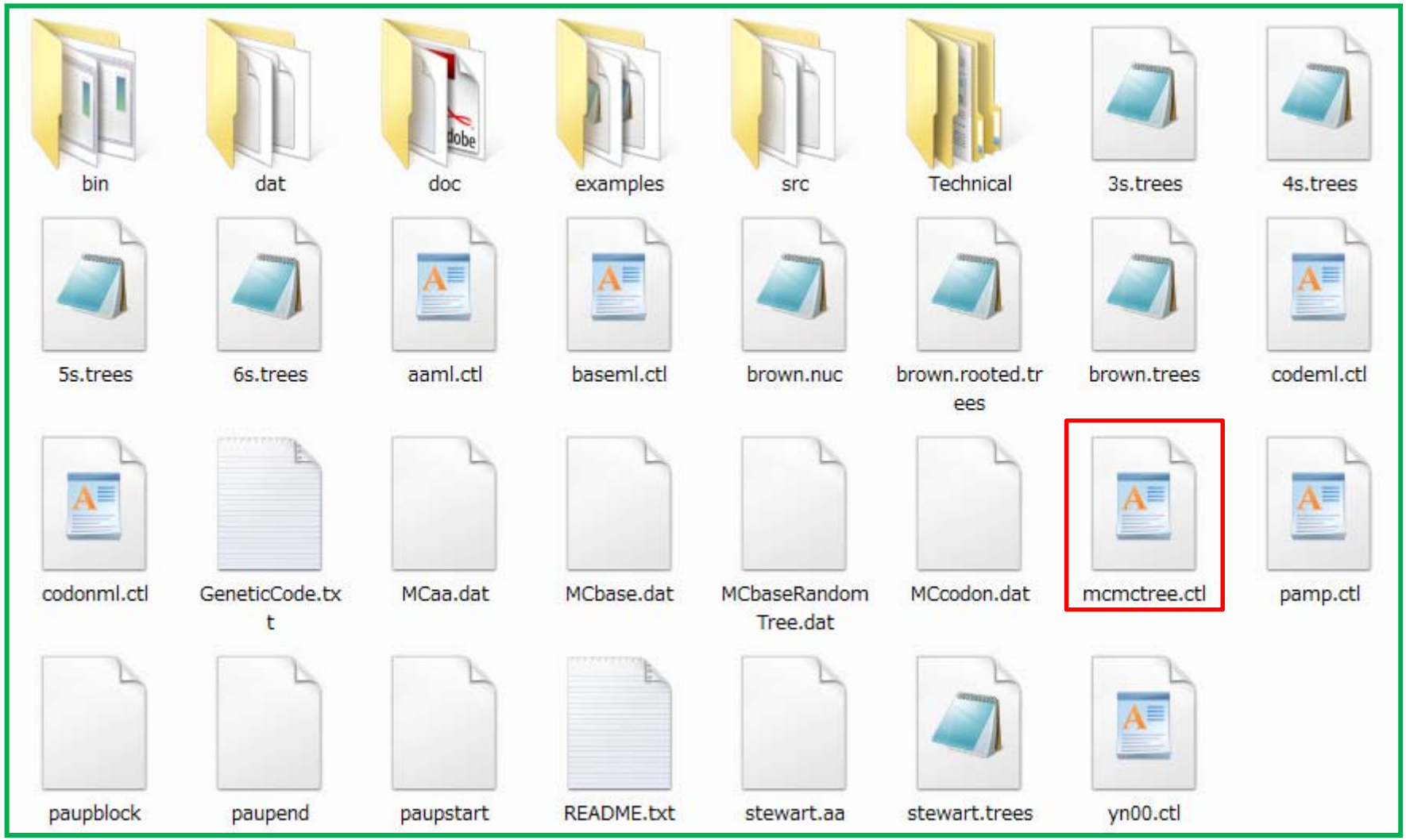
desaturase_sub_node_ref_for_paml.nwk


Run mcmctree in
paml
Prior dirstibutions
without data

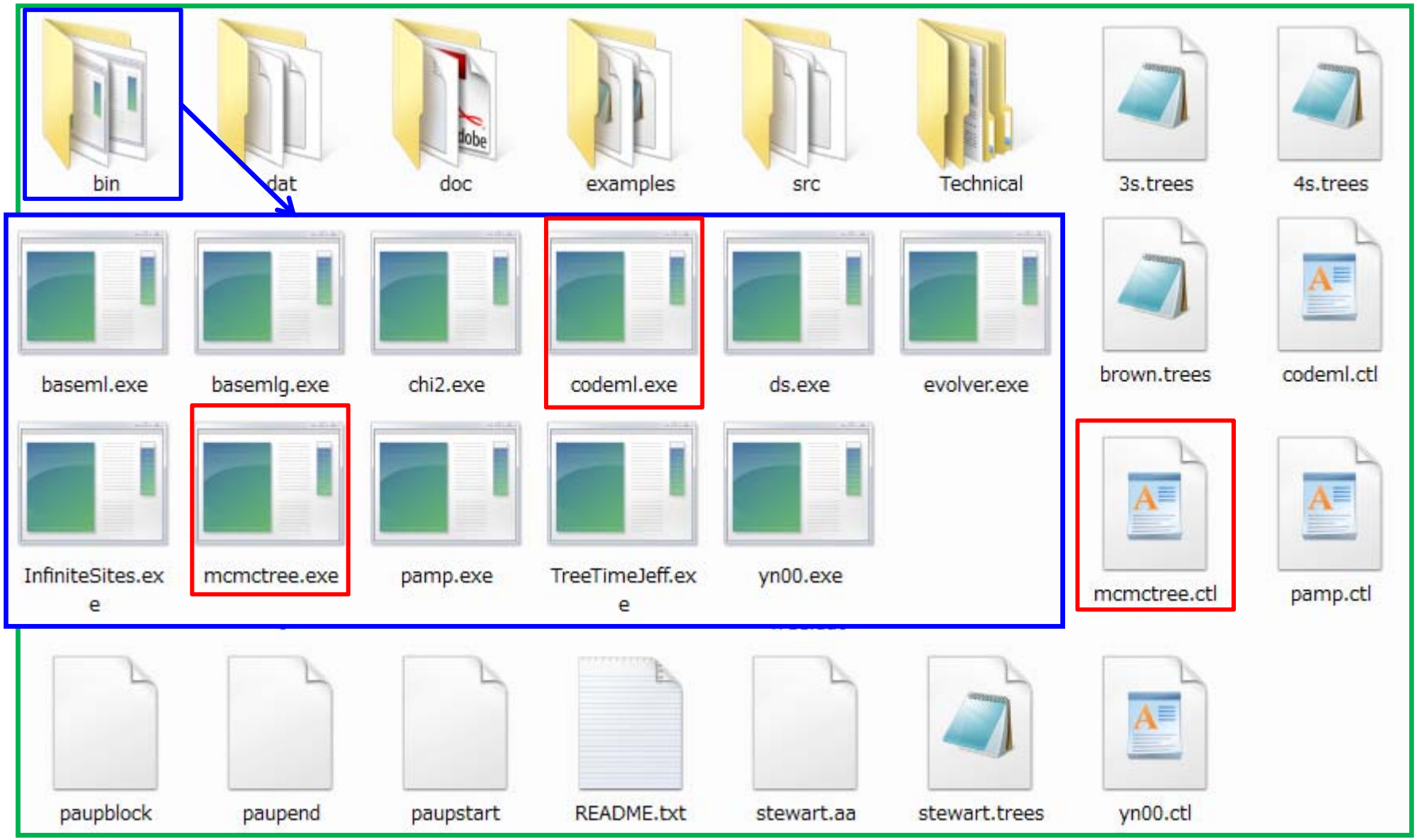
Prepare the input data, exec file, and cntl file



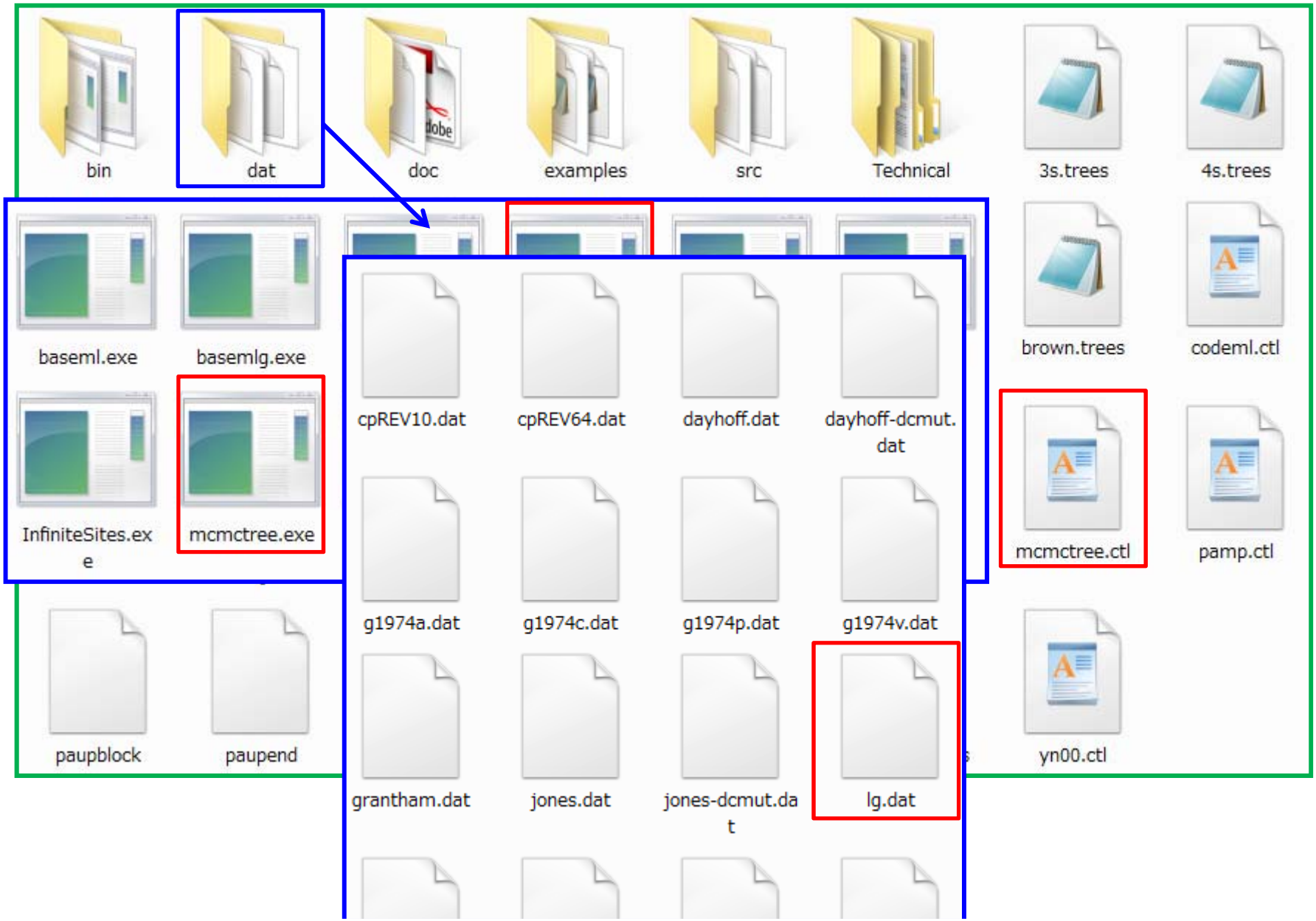
Cntl file mcmctree.ctl



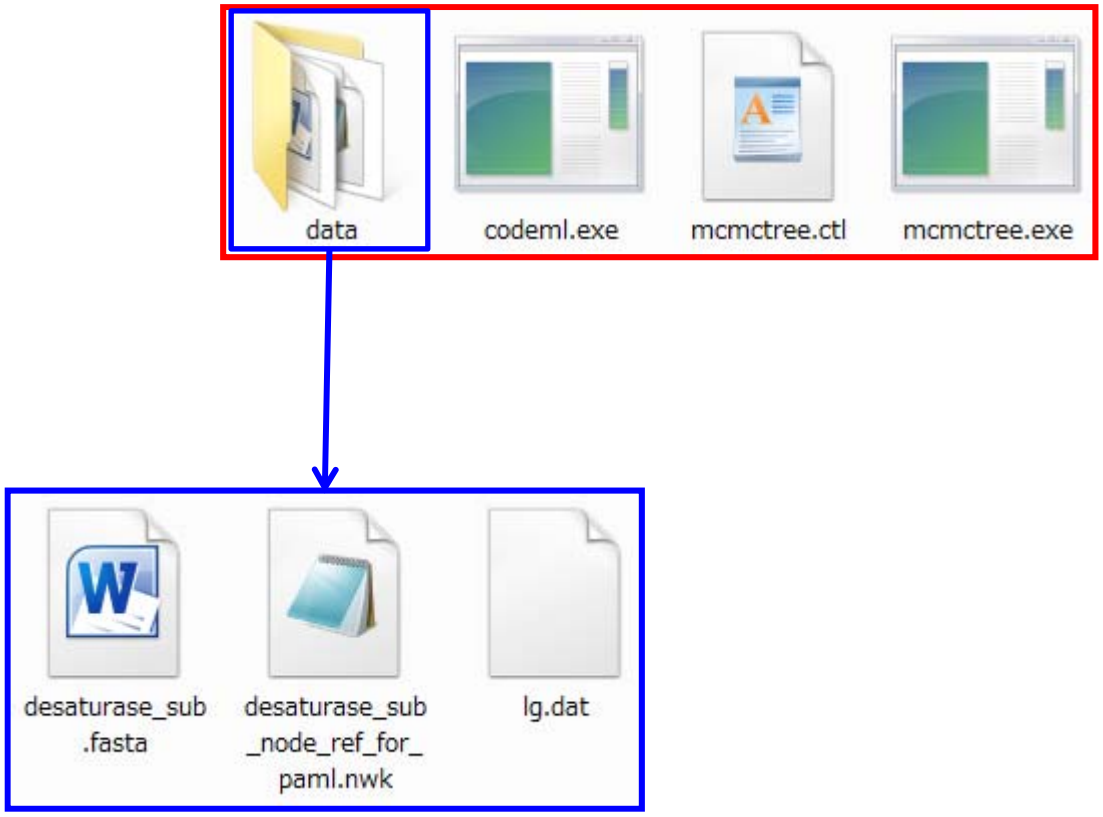
Exec file: codeml.exe mcmctree.exe



Amino acid replacement matrix: lg.dat



Input data, exec file, and cntl file, put in order



Write the required information in the cntl file

```
          seed = -1
       seqfile = data/desaturase_sub.fasta
      treefile = data/desaturase_sub_node_ref_for_paml.nwk
       outfile = desaturase_rate_time_out.txt

         ndata = 1
       seqtype = 0    * 0: nucleotides; 1:codons; 2:AAs
       usedata = 1    * 0: no data; 1:seq like; 2:normal approximation; 3:out.BV (in.BV)
         clock = 2    * 1: global clock; 2: independent rates; 3: correlated rates
       RootAge = <1.0  * safe constraint on root age, used if no fossil for root.

    aaRatefile = data/lg.dat  * only used for aa seqs with model=empirical(_F)
         model = 0    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
         alpha = 0    * alpha for gamma rates at sites
         ncatG = 5    * No. categories in discrete gamma

     cleandata = 0    * remove sites with ambiguity data (1:yes, 0:no)?

        BDparas = 1 1 0.1   * birth, death, sampling
   kappa_gamma = 6 2      * gamma prior for kappa
   alpha_gamma = 1 1      * gamma prior for alpha

   rgene_gamma = 2 2     * gamma prior for rate for genes
  sigma2_gamma = 1 10    * gamma prior for sigma^2     (for clock=2 or 3)

      finetune = 1: .1 .1 .1 .1 .1 .1
* auto (0 or 1) : auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

*      finetune = 1: 0.05  0.2  0.15  0.1 .5
* auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

         print = 2 (print = 2 to print rates for branches with clock=2 or 3)
        burnin = 2000
      sampfreq = 2
       nsample = 20000

*** Note: Make your window wider (100 columns) before running the program.
```

```
Seed of MCMC
Input file: sequences
Input file: tree
Output file
```

mcmctree.ctl

Write the required information in the cntl file

```
          seed = -1
       seqfile = data/desaturase_sub.fasta
      treefile = data/desaturase_sub_node_ref_for_paml.nwk
       outfile = desaturase_rate_time_out.txt

         ndata = 1
       seqtype = 2    * 0: nucleotides; 1:codons; 2:AAs
       usedata = 1    * 0: no data; 1:seq like; 2:normal approximation; 3:out.BV (in.BV)
         clock = 2    * 1: global clock; 2: independent rates; 3: correlated rates
       RootAge = <1.0  * safe constraint on root age, used if no fossil for root.

    aaRatefile = data/lg.dat  * only used for aa seqs with model=empirical(_F)
         model = 0    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
         alpha = 0    * alpha for gamma rates at sites
         ncatG = 5    * No. categories in discrete gamma

     cleandata = 0    * remove sites with ambiguity data (1:yes, 0:no)?

        BDparas = 1 1 0.1   * birth, death, sampling
   kappa_gamma = 6 2      * gamma prior for kappa
   alpha_gamma = 1 1      * gamma prior for alpha

   rgene_gamma = 2 2     * gamma prior for rate for genes
  sigma2_gamma = 1 10    * gamma prior for sigma^2     (for clock=2 or 3)

      finetune = 1: .1 .1 .1 .1 .1 .1
* auto (0 or 1) : auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

*      finetune = 1: 0.05  0.2  0.15  0.1 .5
* auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

         print = 2 (print = 2 to print rates for branches with clock=2 or 3)
        burnin = 2000
      sampfreq = 2
       nsample = 20000

*** Note: Make your window wider (100 columns) before running the program.
```

```
Number of loci: 1
Sequence type: amino
               acids
```

mcmctree.ctl

Write the required information in the cntl file

```
          seed = -1
       seqfile = data/desaturase_sub.fasta
      treefile = data/desaturase_sub_node_ref_for_paml.nwk
       outfile = desaturase_rate_time_out.txt

         ndata = 1
       seqtype = 2    * 0: nucleotides; 1:codons; 2:AAs
       usedata = 0    * 0: no data; 1:seq like; 2:normal approximation; 3:out.BV (in.BV)
         clock = 2    * 1: global clock; 2: independent rates; 3: correlated rates
       RootAge = <1.0  * safe constraint on root age, used if no fossil for root.

    aaRatefile = data/lg.dat  * only used for aa seqs with model=empirical(_F)
         model = 0    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
         alpha = 0    * alpha for gamma rates at sites
         ncatG = 5    * No. categories in discrete gamma

     cleandata = 0    * remove sites with ambiguity data (1:yes, 0:no)?

        BDparas = 1 1 0.1   * birth, death, sampling
   kappa_gamma = 6 2      * gamma prior for kappa
   alpha_gamma = 1 1      * gamma prior for alpha

   rgene_gamma = 2 2     * gamma prior for rate for genes
  sigma2_gamma = 1 10    * gamma prior for sigma^2     (for clock=2 or 3)

      finetune = 1: .1 .1 .1 .1 .1 .1
* auto (0 or 1) : auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

*      finetune = 1: 0.05  0.2  0.15  0.1 .5
* auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

         print = 2 (print = 2 to print rates for branches with clock=2 or 3)
        burnin = 2000
      sampfreq = 2
       nsample = 20000

*** Note: Make your window wider (100 columns) before running the program.
```

Data to be used (first, analyze “no data” to confirm the prior distributions

mcmctree.ctl

Write the required information in the cntl file

```
          seed = -1
       seqfile = data/desaturase_sub.fasta
      treefile = data/desaturase_sub_node_ref_for_paml.nwk
       outfile = desaturase_rate_time_out.txt

         ndata = 1
       seqtype = 2    * 0: nucleotides; 1:codons; 2:AAs
       usedata = 0    * 0: no data; 1:seq like; 2:normal approximation; 3:out.BV (in.BV)
         clock = 2    * 1: global clock; 2: independent rates; 3: correlated rates
       RootAge = <1.0  * safe constraint on root age, used if no fossil for root.

    aaRatefile = data/lg.dat  * only used for aa seqs with model=empirical(_F)
         model = 0    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
         alpha = 0    * alpha for gamma rates at sites
         ncatG = 5    * No. categories in discrete gamma

     cleandata = 0    * remove sites with ambiguity data (1:yes, 0:no)?

        BDparas = 1 1 0.1   * birth, death, sampling
   kappa_gamma = 6 2      * gamma prior for kappa
   alpha_gamma = 1 1      * gamma prior for alpha

   rgene_gamma = 2 2     * gamma prior for rate for genes
  sigma2_gamma = 1 10    * gamma prior for sigma^2     (for clock=2 or 3)

      finetune = 1: .1 .1 .1 .1 .1 .1
* auto (0 or 1) : auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

*      finetune = 1: 0.05  0.2  0.15  0.1 .5
* auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

         print = 2 (print = 2 to print rates for branches with clock=2 or 3)
        burnin = 2000
      sampfreq = 2
       nsample = 20000

*** Note: Make your window wider (100 columns) before running the program.
```

Pattern of clock: rates are independent among branches

mcmctree.ctl

Write the required information in the cntl file

```
         seed = -1
       seqfile = data/desaturase_sub.fasta
      treefile = data/desaturase_sub_node_ref_for_paml.nwk
       outfile = desaturase_rate_time_out.txt

         ndata = 1
       seqtype = 2    * 0: nucleotides; 1:codons; 2:AAs
       usedata = 0    * 0: no data; 1:seq like; 2:normal approximation; 3:out.BV (in.BV)
         clock = 2    * 1: global clock; 2: independent rates; 3: correlated rates
       RootAge = <1.0  * safe constraint on root age, used if no fossil for root.

    aaRatefile = data/lg.dat  * only used for aa seqs with model=empirical(_F)
         model = 0    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
         alpha = 0    * alpha for gamma rates at sites
         ncatG = 5    * No. categories in discrete gamma

     cleandata = 0    * remove sites with ambiguity data (1:yes, 0:no)?

        BDparas = 1 1 0.1   * birth, death, sampling
   kappa_gamma = 6 2      * gamma prior for kappa
   alpha_gamma = 1 1      * gamma prior for alpha

   rgene_gamma = 2 2     * gamma prior for rate for genes
  sigma2_gamma = 1 10    * gamma prior for sigma^2     (for clock=2 or 3)

      finetune = 1: .1 .1 .1 .1 .1 .1
* auto (0 or 1) : auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

*      finetune = 1: 0.05  0.2  0.15  0.1 .5
* auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

         print = 2 (print = 2 to print rates for branches with clock=2 or 3)
        burnin = 2000
      sampfreq = 2
       nsample = 20000

*** Note: Make your window wider (100 columns) before running the program.
```

Safe constraint on the root age

mcmctree.ctl

Write the required information in the cntl file

```
          seed = -1
       seqfile = data/desaturase_sub.fasta
      treefile = data/desaturase_sub_node_ref_for_paml.nwk
       outfile = desaturase_rate_time_out.txt

         ndata = 1
       seqtype = 2    * 0: nucleotides; 1:codons; 2:AAs
       usedata = 0    * 0: no data; 1:seq like; 2:normal approximation; 3:out.BV (in.BV)
         clock = 2    * 1: global clock; 2: independent rates; 3: correlated rates
       RootAge = <1.0  * safe constraint on root age, used if no fossil for root.

    aaRatefile = data/lg.dat  * only used for aa seqs with model=empirical(_F)
         model = 0    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
         alpha = 0    * alpha for gamma rates at sites
         ncatG = 5    * No. categories in discrete gamma

     cleandata = 0    * remove sites with ambiguity data (1:yes, 0:no)?

        BDparas = 1 1 0.1   * birth, death, sampling
   kappa_gamma = 6 2      * gamma prior for kappa
   alpha_gamma = 1 1      * gamma prior for alpha

   rgene_gamma = 2 2     * gamma prior for rate for genes
  sigma2_gamma = 1 10    * gamma prior for sigma^2     (for clock=2 or 3)

      finetune = 1: .1 .1 .1 .1 .1 .1
* auto (0 or 1) : auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

*      finetune = 1: 0.05  0.2  0.15  0.1 .5
* auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

         print = 2 (print = 2 to print rates for branches with clock=2 or 3)
        burnin = 2000
      sampfreq = 2
       nsample = 20000

*** Note: Make your window wider (100 columns) before running the program.
```

Use lg.dat for LG model

mcmctree.ctl

Write the required information in the cntl file

```
          seed = -1
       seqfile = data/desaturase_sub.fasta
      treefile = data/desaturase_sub_node_ref_for_paml.nwk
       outfile = desaturase_rate_time_out.txt

         ndata = 1
       seqtype = 2    * 0: nucleotides; 1:codons; 2:AAs
       usedata = 0    * 0: no data; 1:seq like; 2:normal approximation; 3:out.BV (in.BV)
         clock = 2    * 1: global clock; 2: independent rates; 3: correlated rates
       RootAge = <1.0  * safe constraint on root age, used if no fossil for root.

    aaRatefile = data/lg.dat  * only used for aa seqs with model=empirical(_F)
         model = 0    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
         alpha = 0    * alpha for gamma rates at sites
         ncatG = 5    * No. categories in discrete gamma

     cleandata = 0    * remove sites with ambiguity data (1:yes, 0:no)?

        BDparas = 1 1 0.1   * birth, death, sampling
   kappa_gamma = 6 2      * gamma prior for kappa
   alpha_gamma = 1 1      * gamma prior for alpha

   rgene_gamma = 2 2     * gamma prior for rate for genes
  sigma2_gamma = 1 10    * gamma prior for sigma^2     (for clock=2 or 3)

      finetune = 1: .1 .1 .1 .1 .1 .1
* auto (0 or 1) : auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

*      finetune = 1: 0.05  0.2  0.15  0.1 .5
* auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

         print = 2 (print = 2 to print rates for branches with clock=2 or 3)
        burnin = 2000
      sampfreq = 2
       nsample = 20000

*** Note: Make your window wider (100 columns) before running the program.
```

Model of sequence evol.
Rate heterogeneity among sites (gamma distribution)

Prior on div times
Prior on Ts/Tv ratio
Prior on site heterogeneity
Gamma prior on the rate
Gamma prior on the rate variation

gamma distribution
$\Gamma(x|\alpha,\beta)$
mean= $\frac{\alpha}{\beta}$
CV= $\frac{1}{\sqrt{\alpha}}$

Length of burin of MCMC
Sampling frequency
Size of MCMC sample

mcmctree.ctl

Write the required information in the cntl file

```
          seed = -1
       seqfile = data/desaturase_sub.fasta
      treefile = data/desaturase_sub_node_ref_for_paml.nwk
       outfile = desaturase_rate_time_out.txt

         ndata = 1
       seqtype = 2    * 0: nucleotides; 1:codons; 2:AAs
       usedata = 0    * 0: no data; 1:seq like; 2:normal approximation; 3:out.BV (in.BV)
         clock = 2    * 1: global clock; 2: independent rates; 3: correlated rates
       RootAge = <1.0  * safe constraint on root age, used if no fossil for root.

    aaRatefile = data/lg.dat  * only used for aa seqs with model=empirical(_F)
         model = 0    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
         alpha = 0    * alpha for gamma rates at sites
         ncatG = 5    * No. categories in discrete gamma

     cleandata = 0    * remove sites with ambiguity data (1:yes, 0:no)?

        BDparas = 1 1 0.1   * birth, death, sampling
   kappa_gamma = 6 2      * gamma prior for kappa
   alpha_gamma = 1 1      * gamma prior for alpha

   rgene_gamma = 2 2     * gamma prior for rate for genes
  sigma2_gamma = 1 10    * gamma prior for sigma^2     (for clock=2 or 3)

      finetune = 1: .1 .1 .1 .1 .1 .1
* auto (0 or 1) : auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

*      finetune = 1: 0.05  0.2  0.15  0.1 .5
* auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

         print = 2 (print = 2 to print rates for branches with clock=2 or 3)
        burnin = 2000
      sampfreq = 2
       nsample = 20000

*** Note: Make your window wider (100 columns) before running the program.
```

This is the case where these files are in the folder named “data” in the folder that includes codeml.exe, mcmctree.exe, and mcmctree.ctl.

Model of sequence evol.
Rate heterogeneity among sites (gamma distribution)

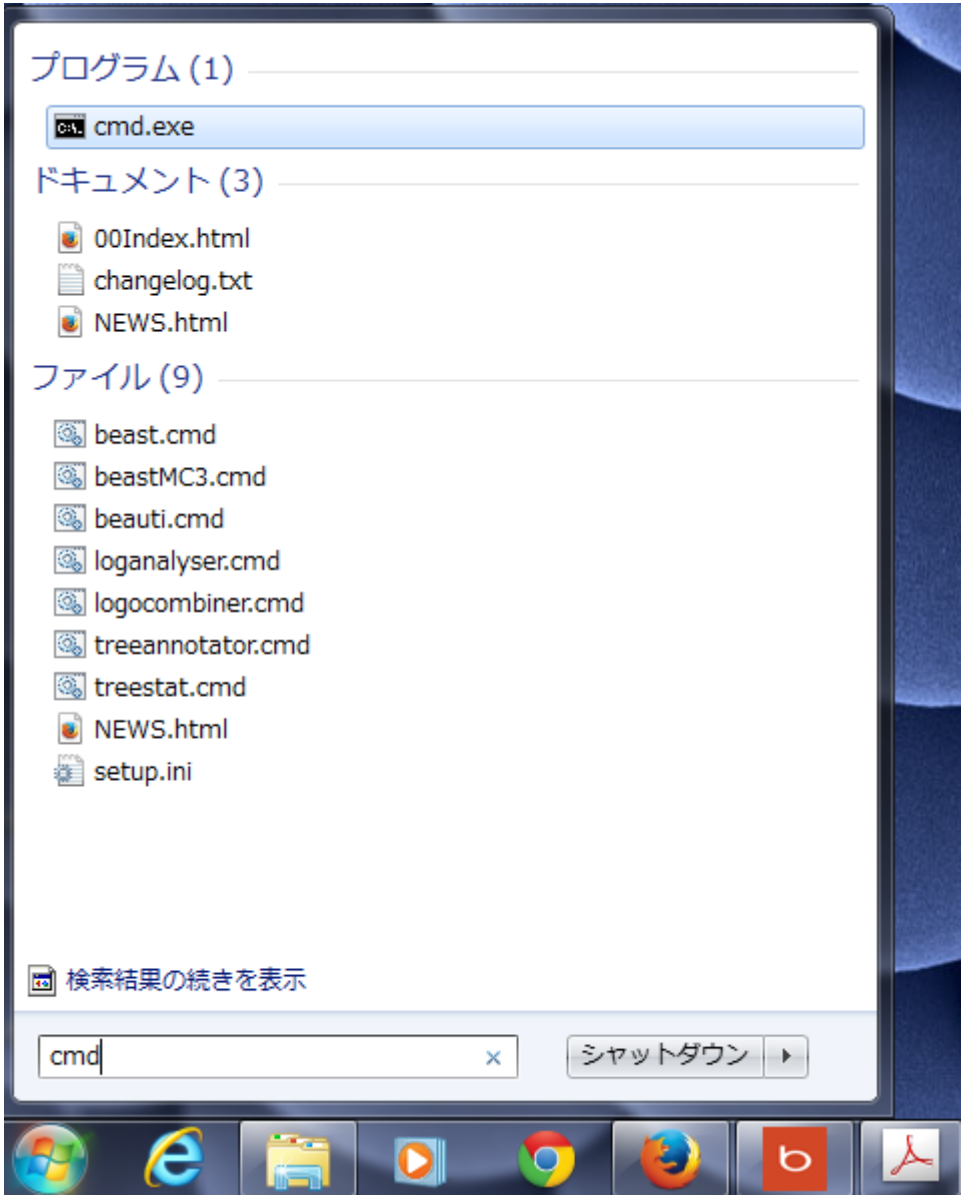
Prior on div times
Prior on Ts/Tv ratio
Prior on site heterogeneity
Gamma prior on the rate
Gamma prior on the rate variation

gamma distribution
$\Gamma(x|\alpha,\beta)$
mean= $\frac{\alpha}{\beta}$
CV= $\frac{1}{\sqrt{\alpha}}$

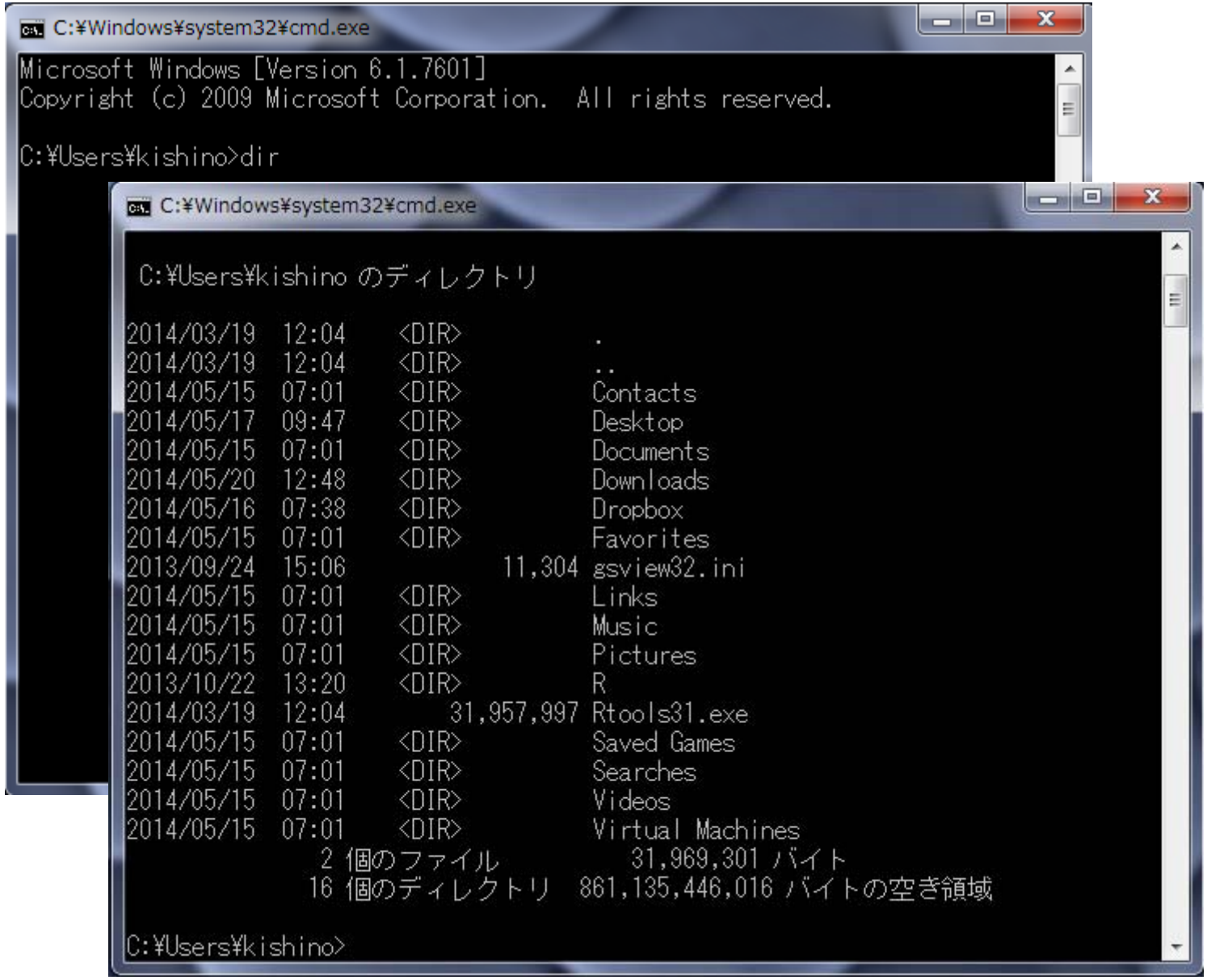
Length of burin of MCMC
Sampling frequency
Size of MCMC sample

mcmctree.ctl

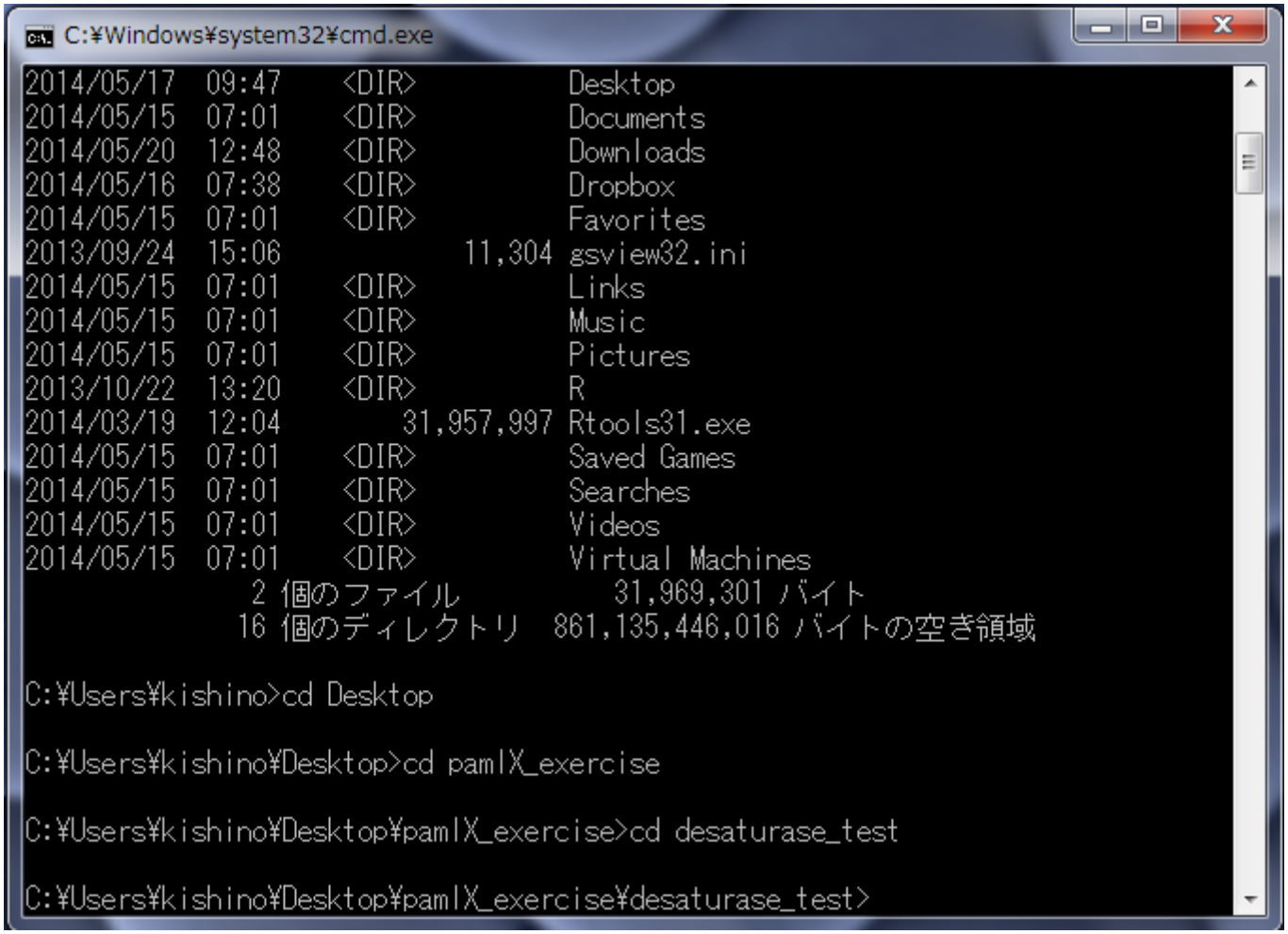
Command prompt



Check the content of the folder by dir

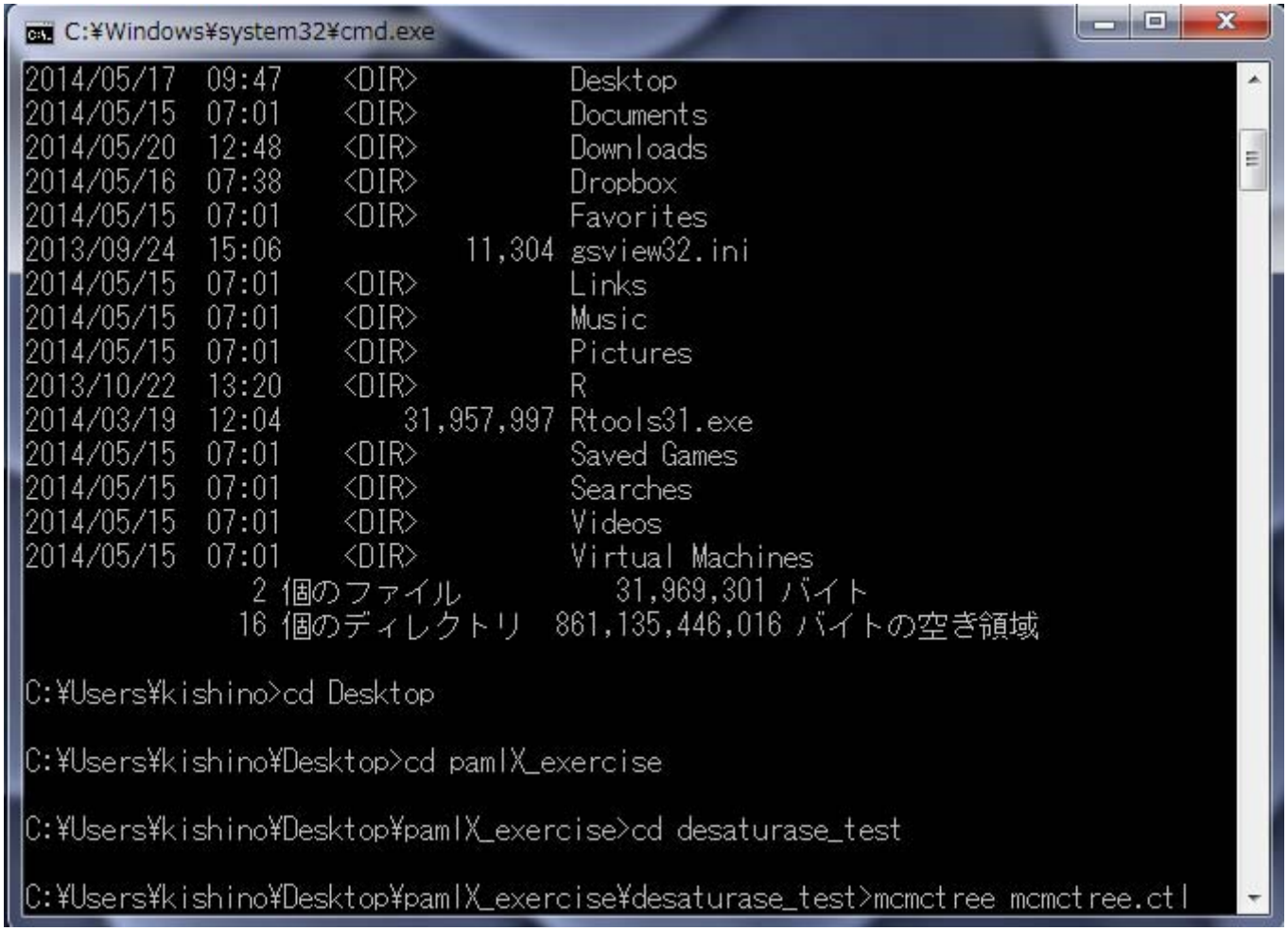


Move by “cd” to the folder that includes the exec files

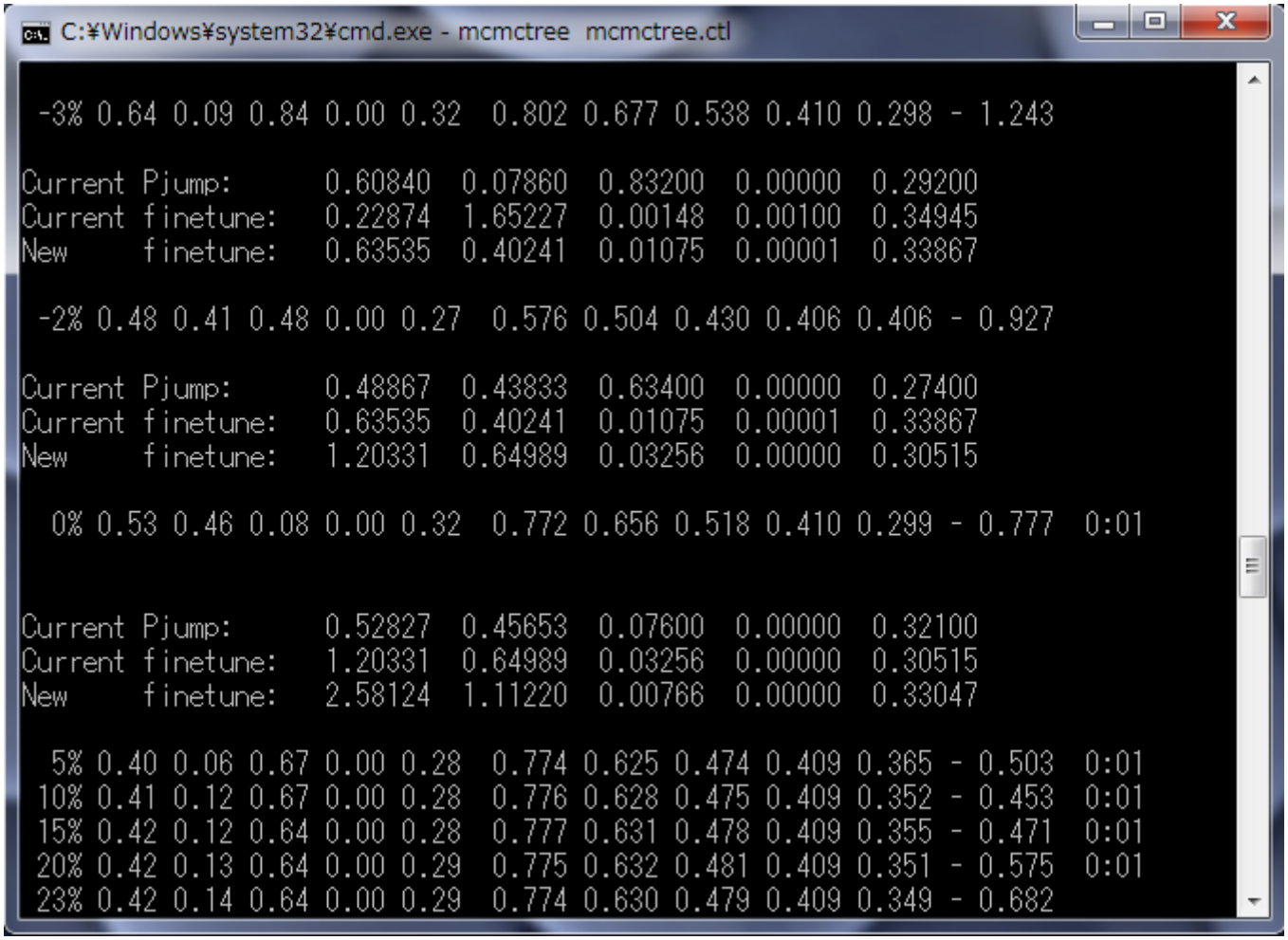


Explorer makes it easy

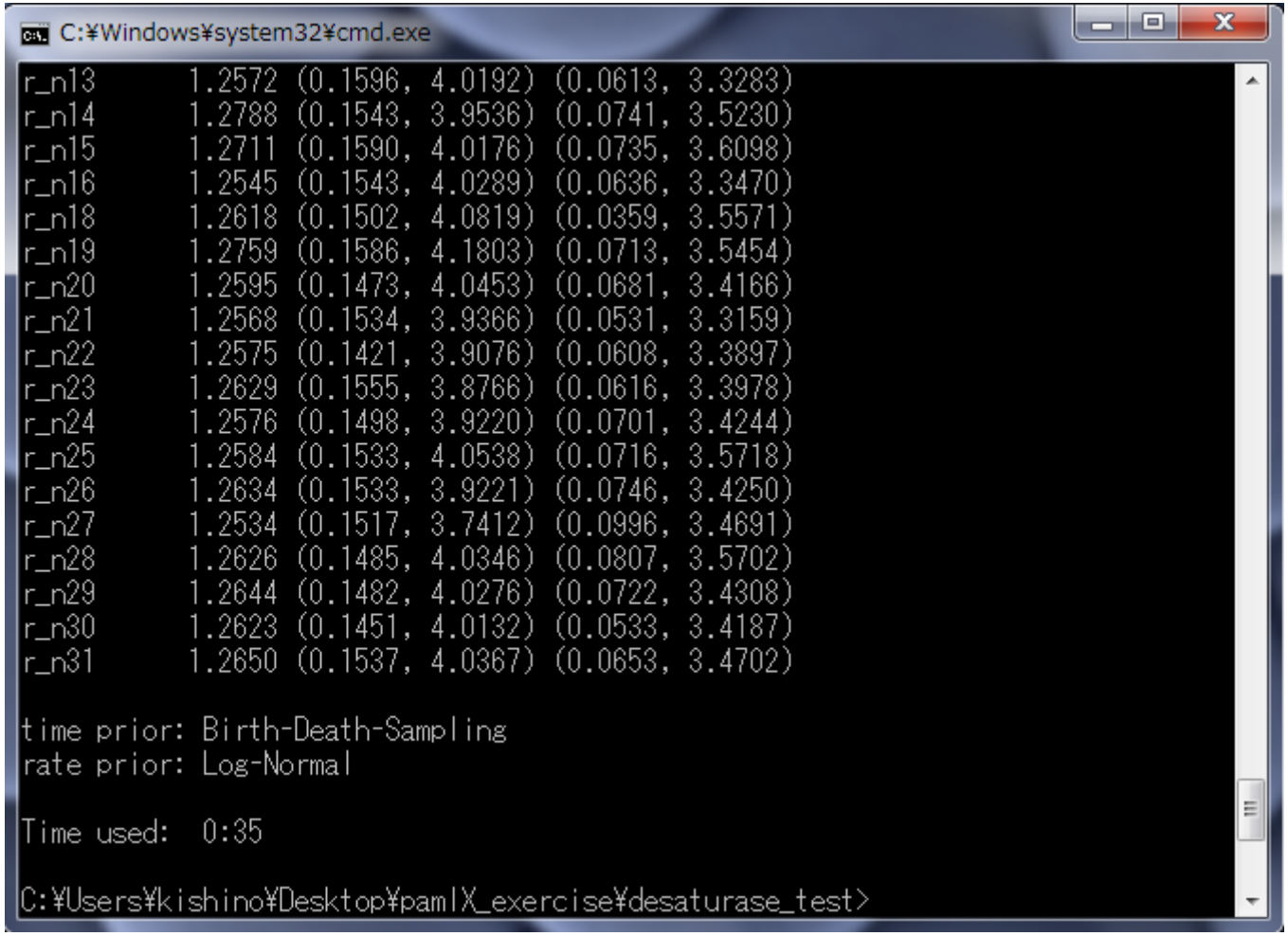
Run mcmctree.exe



MCMC starts


C:¥Windows¥system32¥cmd.exe - mcmctree mcmctree.ctl
-3% 0.64 0.09 0.84 0.00 0.32 0.802 0.677 0.538 0.410 0.298 - 1.243
Current Pjump: 0.60840 0.07860 0.83200 0.00000 0.29200
Current finetune: 0.22874 1.65227 0.00148 0.00100 0.34945
New finetune: 0.63535 0.40241 0.01075 0.00001 0.33867
-2% 0.48 0.41 0.48 0.00 0.27 0.576 0.504 0.430 0.406 0.406 - 0.927
Current Pjump: 0.48867 0.43833 0.63400 0.00000 0.27400
Current finetune: 0.63535 0.40241 0.01075 0.00001 0.33867
New finetune: 1.20331 0.64989 0.03256 0.00000 0.30515
0% 0.53 0.46 0.08 0.00 0.32 0.772 0.656 0.518 0.410 0.299 - 0.777 0:01
Current Pjump: 0.52827 0.45653 0.07600 0.00000 0.32100
Current finetune: 1.20331 0.64989 0.03256 0.00000 0.30515
New finetune: 2.58124 1.11220 0.00766 0.00000 0.33047
5% 0.40 0.06 0.67 0.00 0.28 0.774 0.625 0.474 0.409 0.365 - 0.503 0:01
10% 0.41 0.12 0.67 0.00 0.28 0.776 0.628 0.475 0.409 0.352 - 0.453 0:01
15% 0.42 0.12 0.64 0.00 0.28 0.777 0.631 0.478 0.409 0.355 - 0.471 0:01
20% 0.42 0.13 0.64 0.00 0.29 0.775 0.632 0.481 0.409 0.351 - 0.575 0:01
23% 0.42 0.14 0.64 0.00 0.29 0.774 0.630 0.479 0.409 0.349 - 0.682

Done


C:¥Windows¥system32¥cmd.exe
r_n13 1.2572 (0.1596, 4.0192) (0.0613, 3.3283)
r_n14 1.2788 (0.1543, 3.9536) (0.0741, 3.5230)
r_n15 1.2711 (0.1590, 4.0176) (0.0735, 3.6098)
r_n16 1.2545 (0.1543, 4.0289) (0.0636, 3.3470)
r_n18 1.2618 (0.1502, 4.0819) (0.0359, 3.5571)
r_n19 1.2759 (0.1586, 4.1803) (0.0713, 3.5454)
r_n20 1.2595 (0.1473, 4.0453) (0.0681, 3.4166)
r_n21 1.2568 (0.1534, 3.9366) (0.0531, 3.3159)
r_n22 1.2575 (0.1421, 3.9076) (0.0608, 3.3897)
r_n23 1.2629 (0.1555, 3.8766) (0.0616, 3.3978)
r_n24 1.2576 (0.1498, 3.9220) (0.0701, 3.4244)
r_n25 1.2584 (0.1533, 4.0538) (0.0716, 3.5718)
r_n26 1.2634 (0.1533, 3.9221) (0.0746, 3.4250)
r_n27 1.2534 (0.1517, 3.7412) (0.0996, 3.4691)
r_n28 1.2626 (0.1485, 4.0346) (0.0807, 3.5702)
r_n29 1.2644 (0.1482, 4.0276) (0.0722, 3.4308)
r_n30 1.2623 (0.1451, 4.0132) (0.0533, 3.4187)
r_n31 1.2650 (0.1537, 4.0367) (0.0653, 3.4702)
time prior: Birth-Death-Sampling
rate prior: Log-Normal
Time used: 0:35
C:¥Users¥kishino¥Desktop¥pamlX_exercise¥desaturase_test>

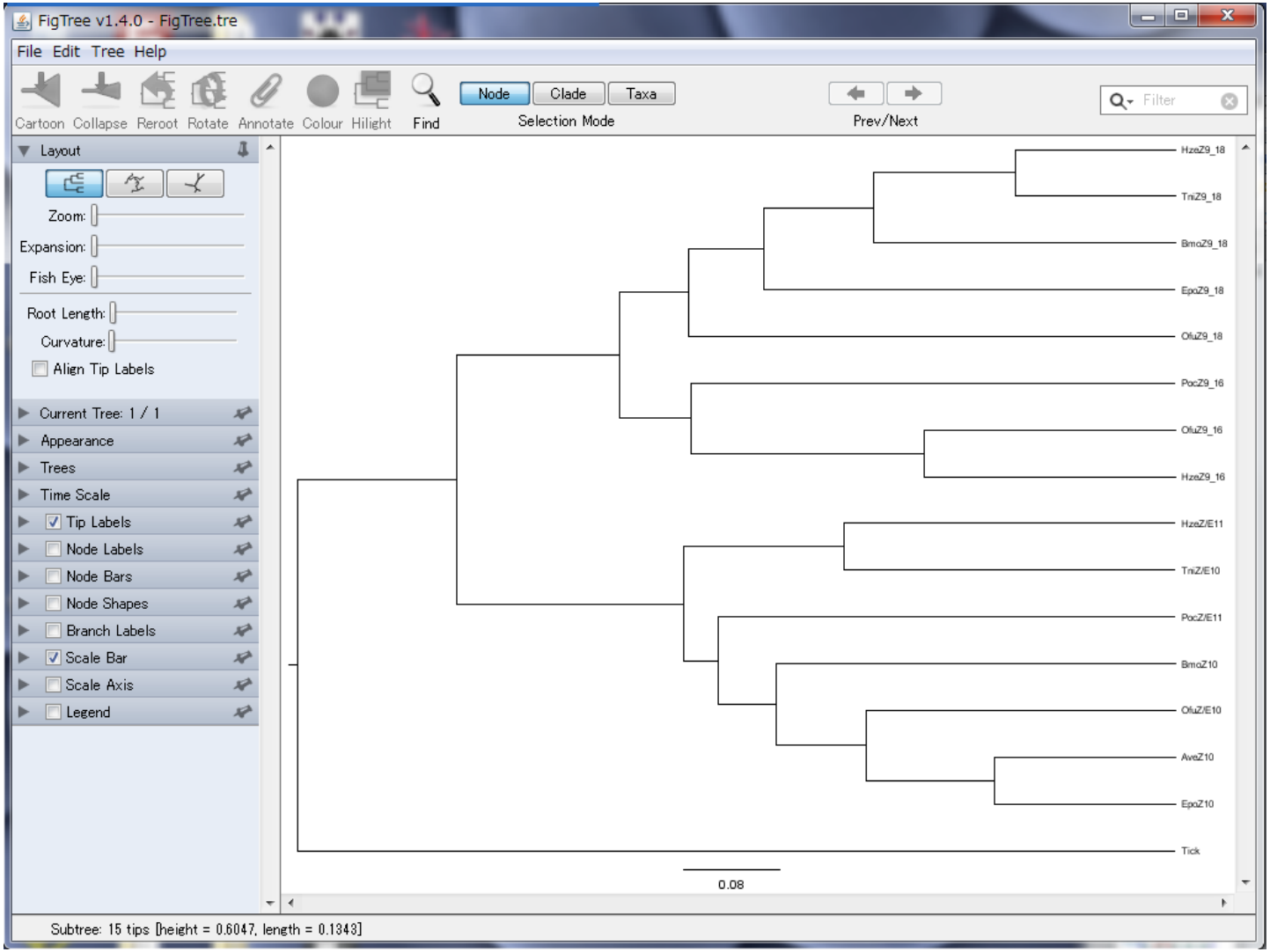
Times on the tree: FigTree.tre

95% credibility interval of the divergence time

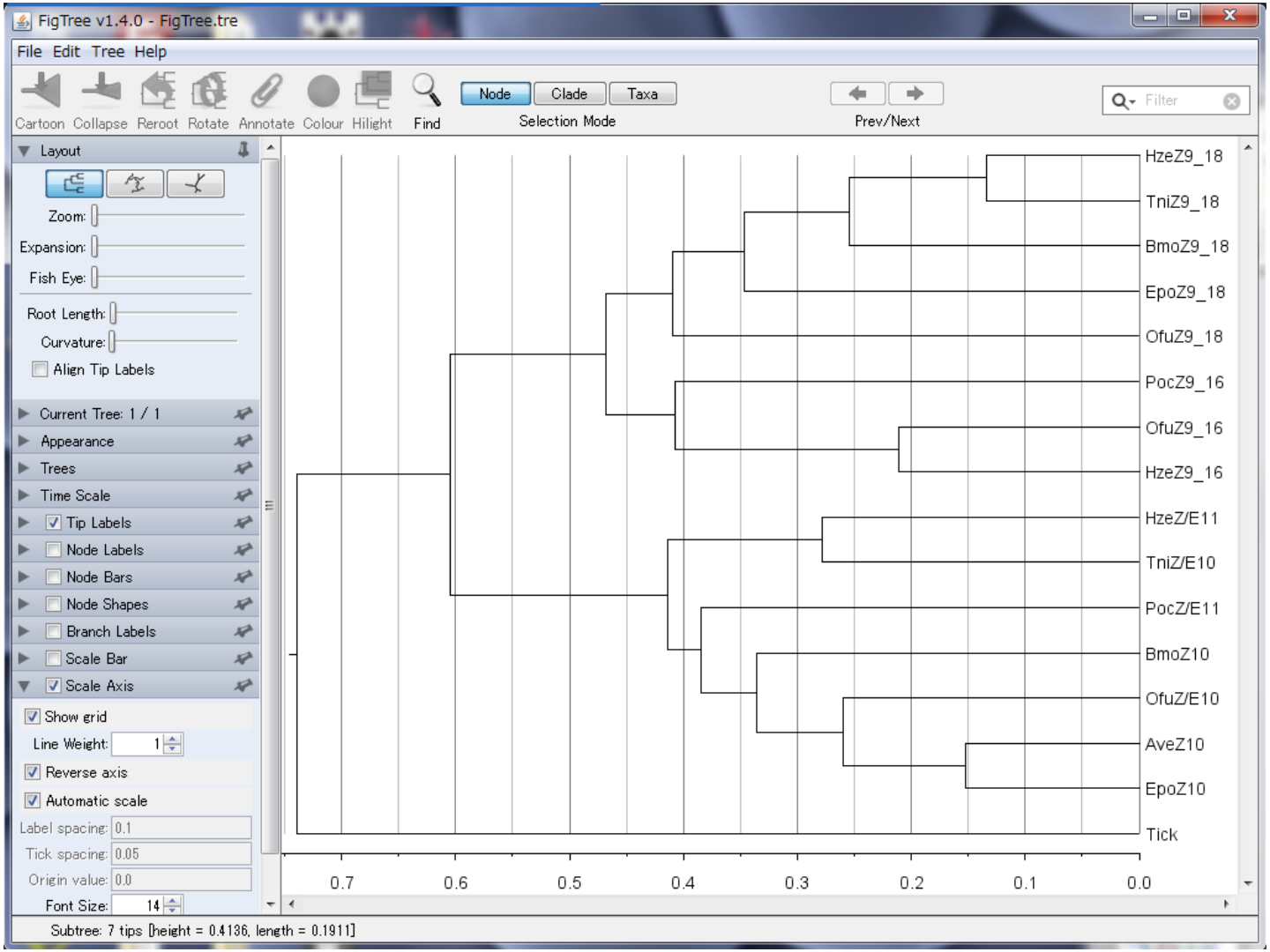
```
#NEXUS
BEGIN TREES;

        UTREE 1 = (((((((HzeZ9_18: 0.13418, TniZ9_18: 0.13418) [&95%={0.004, 0.405}]: 0.11971,
BmoZ9_18: 0.25389) [&95%={0.045, 0.412}]: 0.09272, EpoZ9_18: 0.34660) [&95%={0.142, 0.416}]:
0.06257, OfuZ9_18: 0.40917) [&95%={0.400, 0.420}]: 0.05879, (PocZ9_16: 0.40704, (OfuZ9_16:
0.21095, HzeZ9_16: 0.21095) [&95%={0.011, 0.408}]: 0.19610) [&95%={0.400, 0.419}]: 0.06092)
[&95%={0.405, 0.733}]: 0.13670, ((HzeZ/E11: 0.27878, TniZ/E10: 0.27878) [&95%={0.015, 0.418}]:
0.13482, (PocZ/E11: 0.38484, (BmoZ10: 0.33570, (OfuZ/E10: 0.26000, (AveZ10: 0.15221, EpoZ10:
0.15221) [&95%={0.004, 0.410}]: 0.10779) [&95%={0.040, 0.415}]: 0.07570) [&95%={0.109, 0.417}]:
0.04915) [&95%={0.213, 0.419}]: 0.02876) [&95%={0.401, 0.420}]: 0.19106) [&95%={0.423, 0.911}]:
0.13429, Tick: 0.73894) [&95%={0.463, 1.009}];

END;
```

FigTree.tree

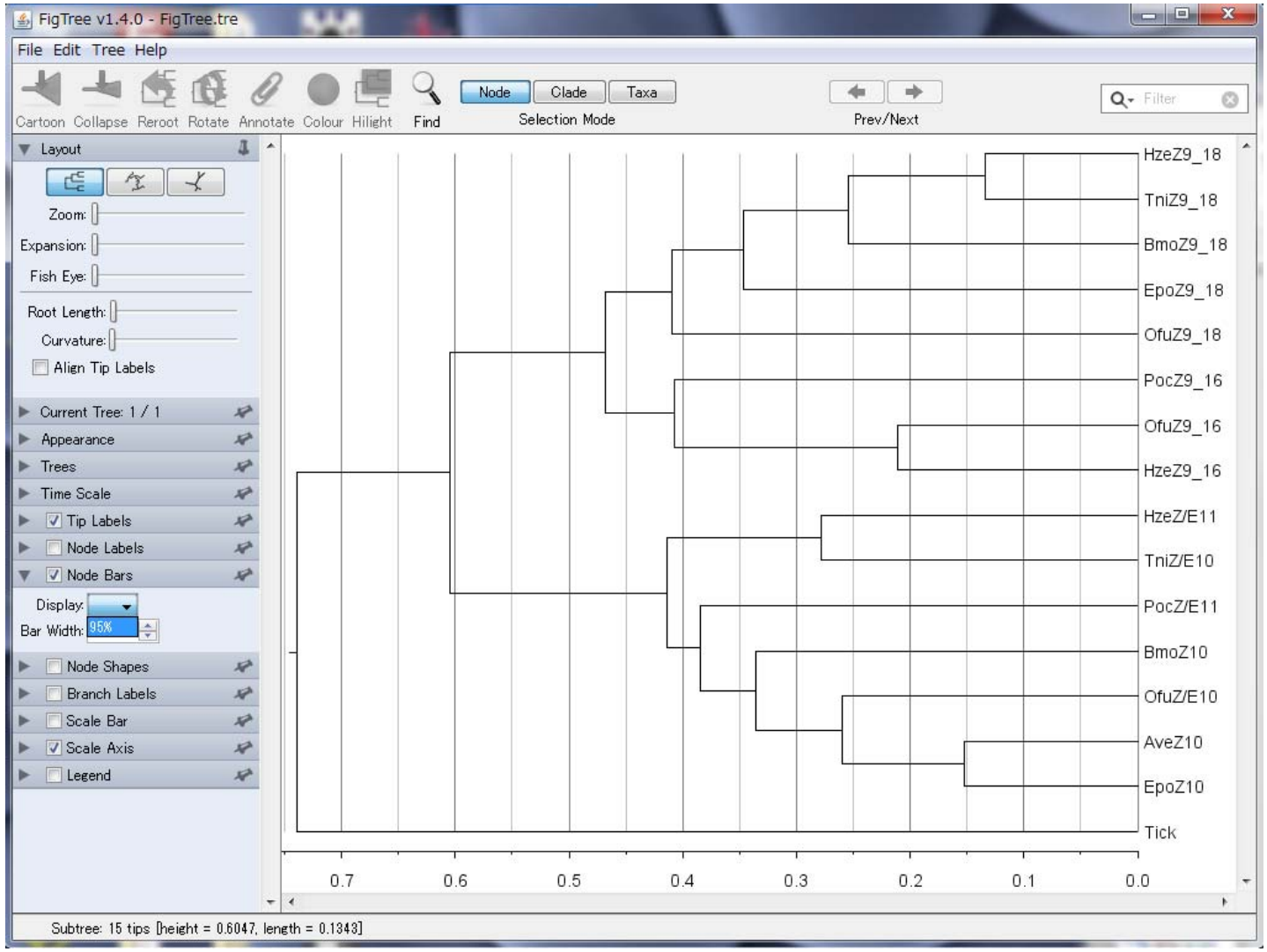
Open FigTree.tre by FigTree



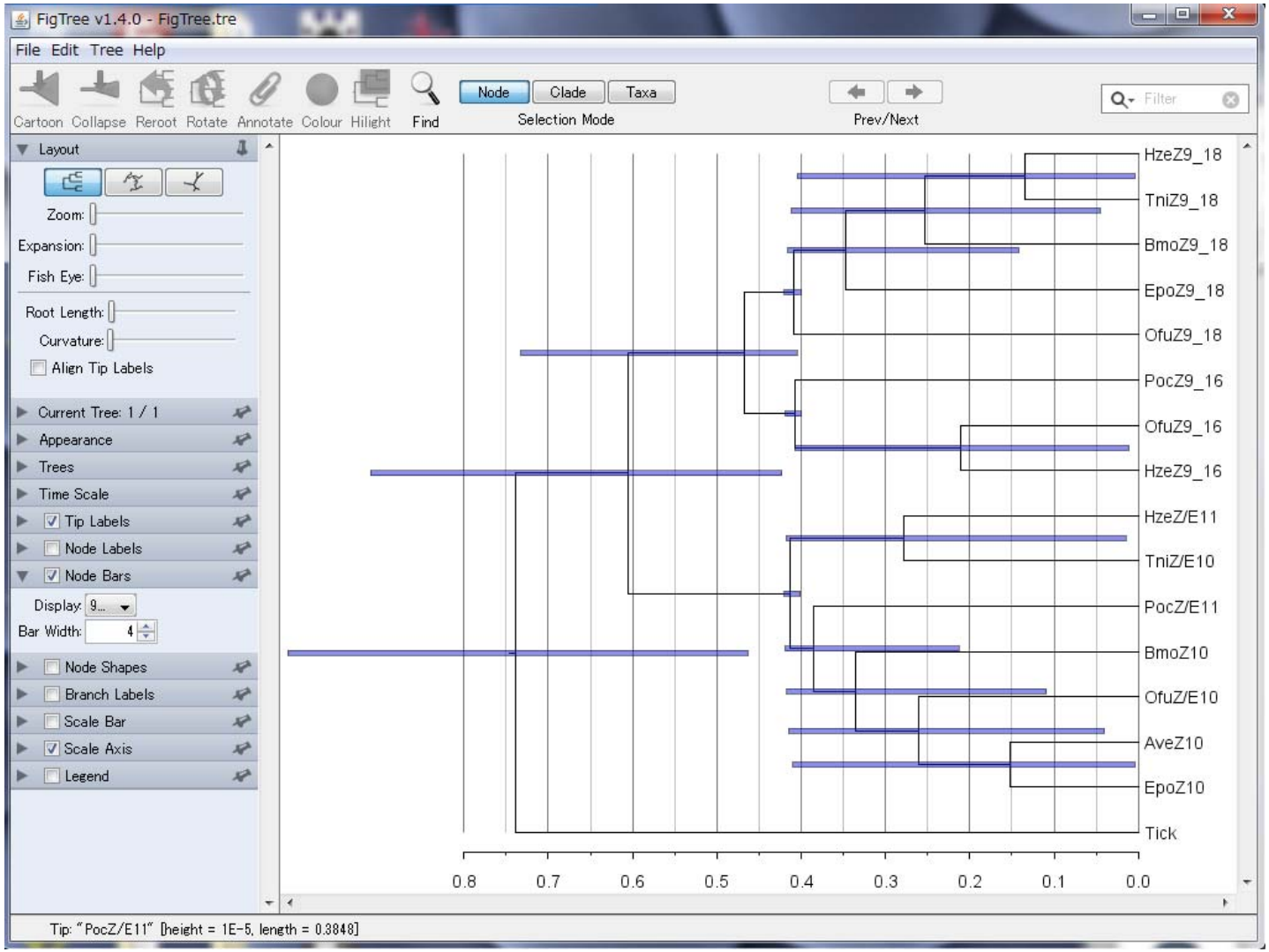
Specify labels and scale axis



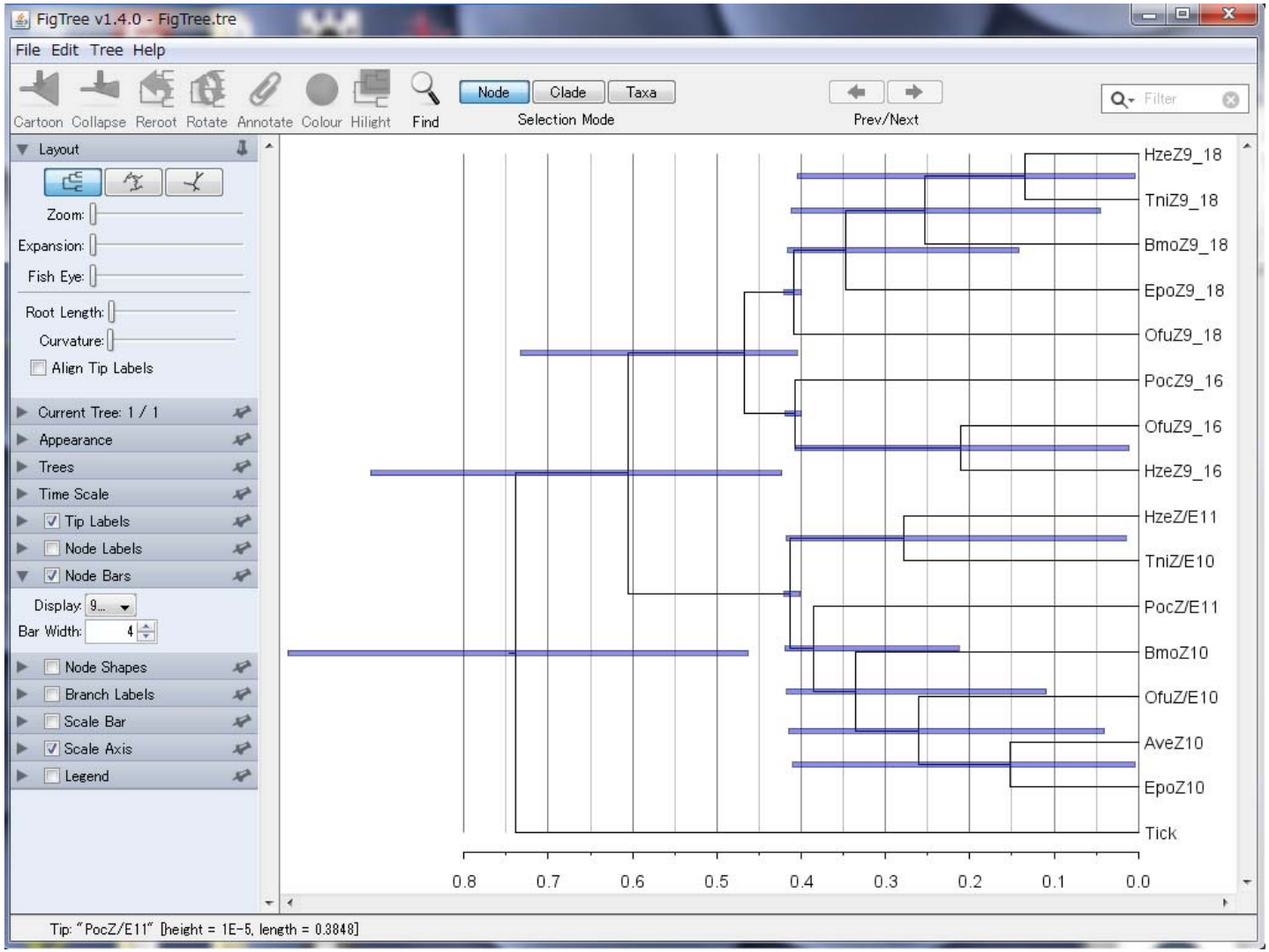
Node Bars: the credibility intervals of divergence times



Node Bars: the credibility intervals of divergence times



MCMC sample from no data represents the prior distribution



Run mcmctree in paml (continued) Posterior distributions

Two stage procedure of rate-time estimation

The two-stage procedure

- Estimate branch lengths and their variances
- Estimate the times and rates based on the normal approximation with the estimated branch lengths and their variances (Laplace method)

reduces the computational burden of analyzing amino acid sequences, with little loss of information.

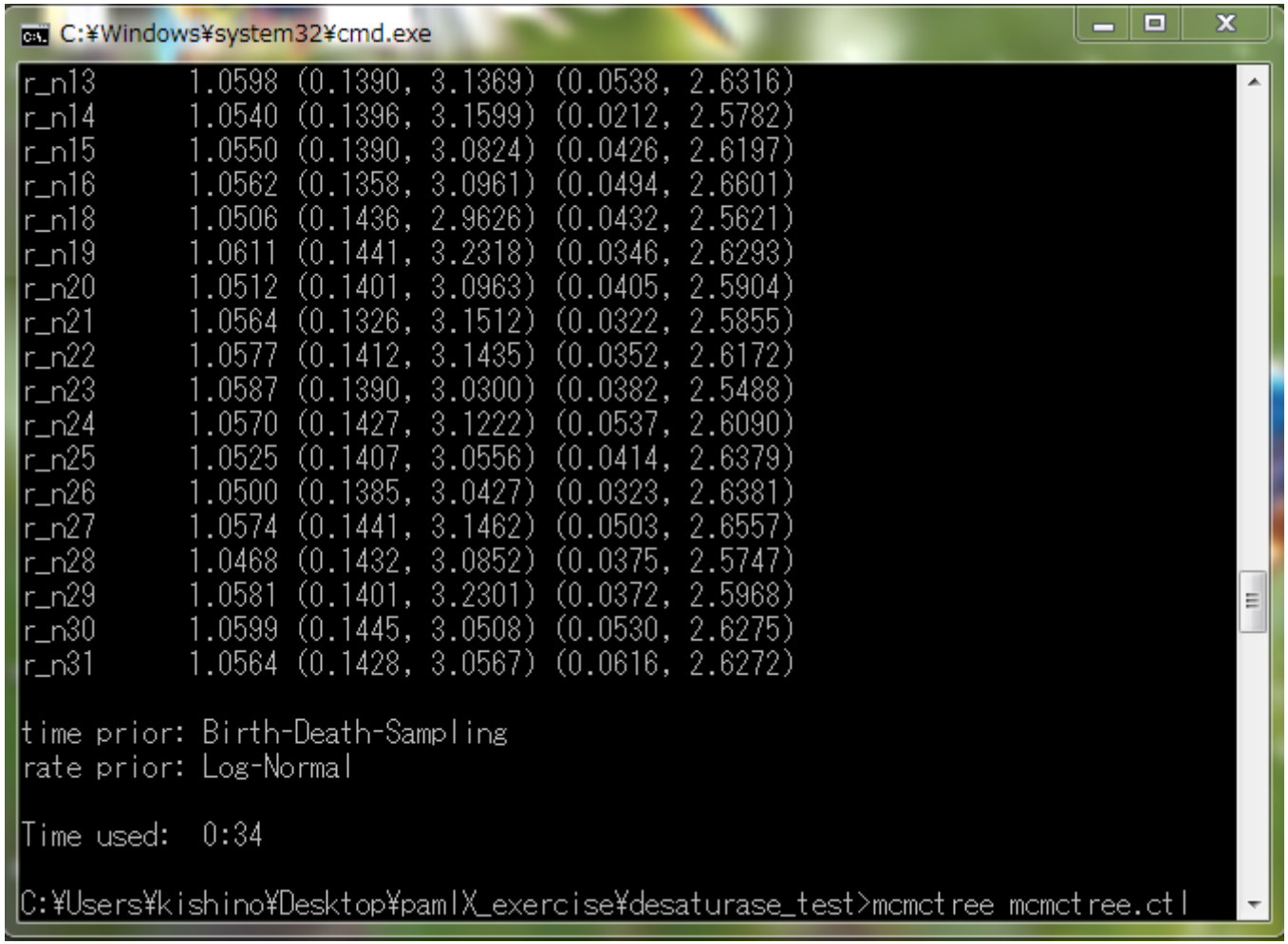
Set “usedata” to 3 in the cntl file (the first stage analysis)

```
          seed = -1
       seqfile = data/desaturase_sub.fasta
      treefile = data/desaturase_sub_node_ref_for_paml.nwk
       outfile = desaturase_rate_time_out.txt

         ndata = 1
       seqtype = 2    * 0: nucleotides; 1:codons; 2:AAs
       usedata = 3    * 0: no data; 1:seq like; 2:normal approximation; 3:out.BV (in.BV)
         clock = 2    * 1: global clock; 2: independent rates; 3: correlated rates
       RootAge = <1.0  * safe constraint on root age, used if no fossil for root.

    aaRatefile = data/lg.dat  * only used for aa seqs with model=empirical(_F)
         model = 0    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
         alpha = 0    * alpha for gamma rates at sites
         ncatG = 5    * No. categories in discrete gamma

     cleandata = 0    * remove sites with ambiguity data (1:yes, 0:no)?

        BDparas = 1 1 0.1   * birth, death, sampling
   kappa_gamma = 6 2      * gamma prior for kappa
   alpha_gamma = 1 1      * gamma prior for alpha

   rgene_gamma = 2 2     * gamma prior for rate for genes
  sigma2_gamma = 1 10    * gamma prior for sigma^2     (for clock=2 or 3)

      finetune = 1: .1 .1 .1 .1 .1 .1
* auto (0 or 1) : auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

*      finetune = 1: 0.05  0.2  0.15  0.1 .5
* auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

         print = 2 (print = 2 to print rates for branches with clock=2 or 3)
        burnin = 2000
      sampfreq = 2
       nsample = 20000

*** Note: Make your window wider (100 columns) before running the program.
```

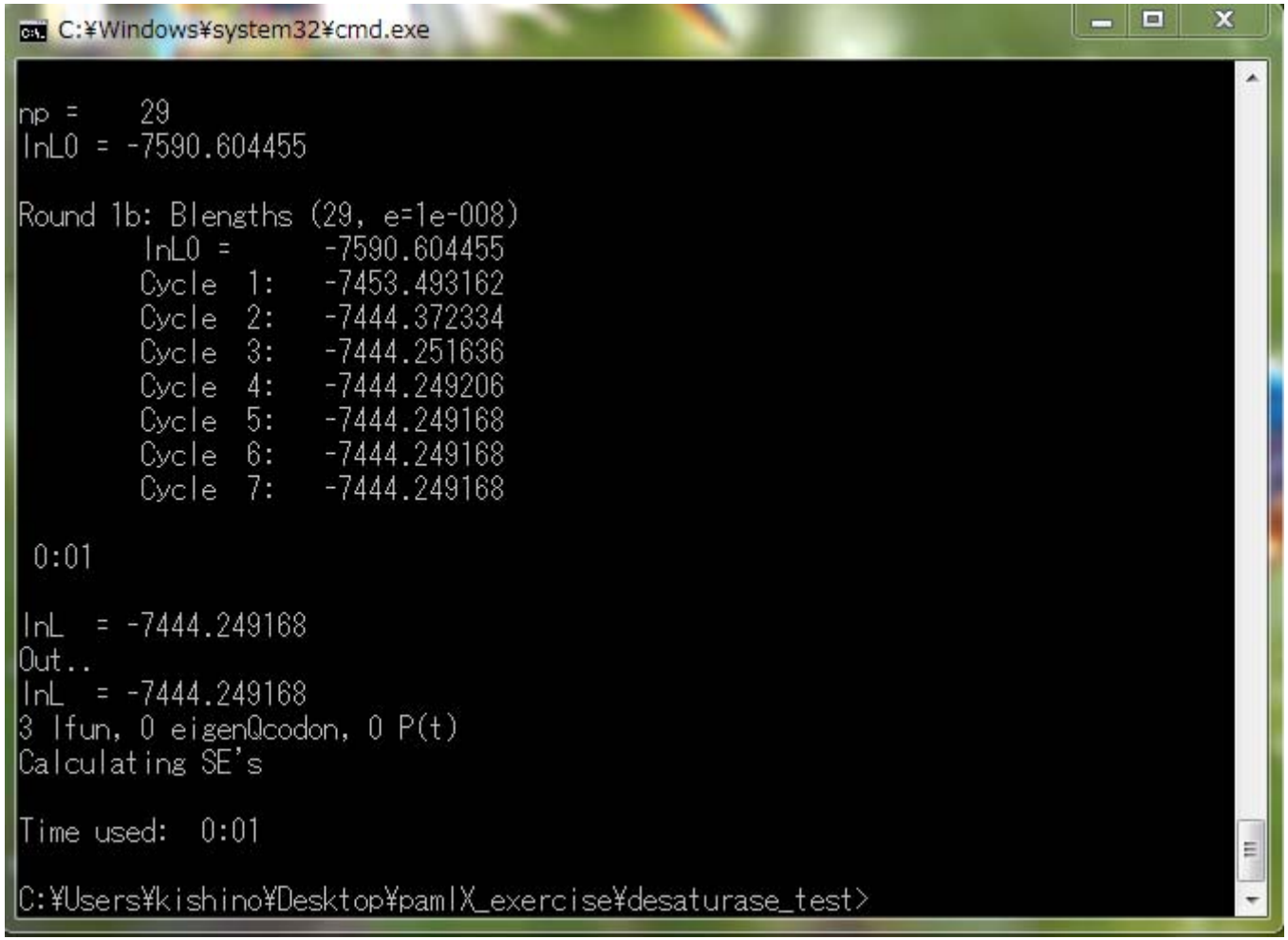
Estimated branch lengths together with their variances are saved as out.BV.

mcmctree.ctl

Run mcmctree.exe



Done

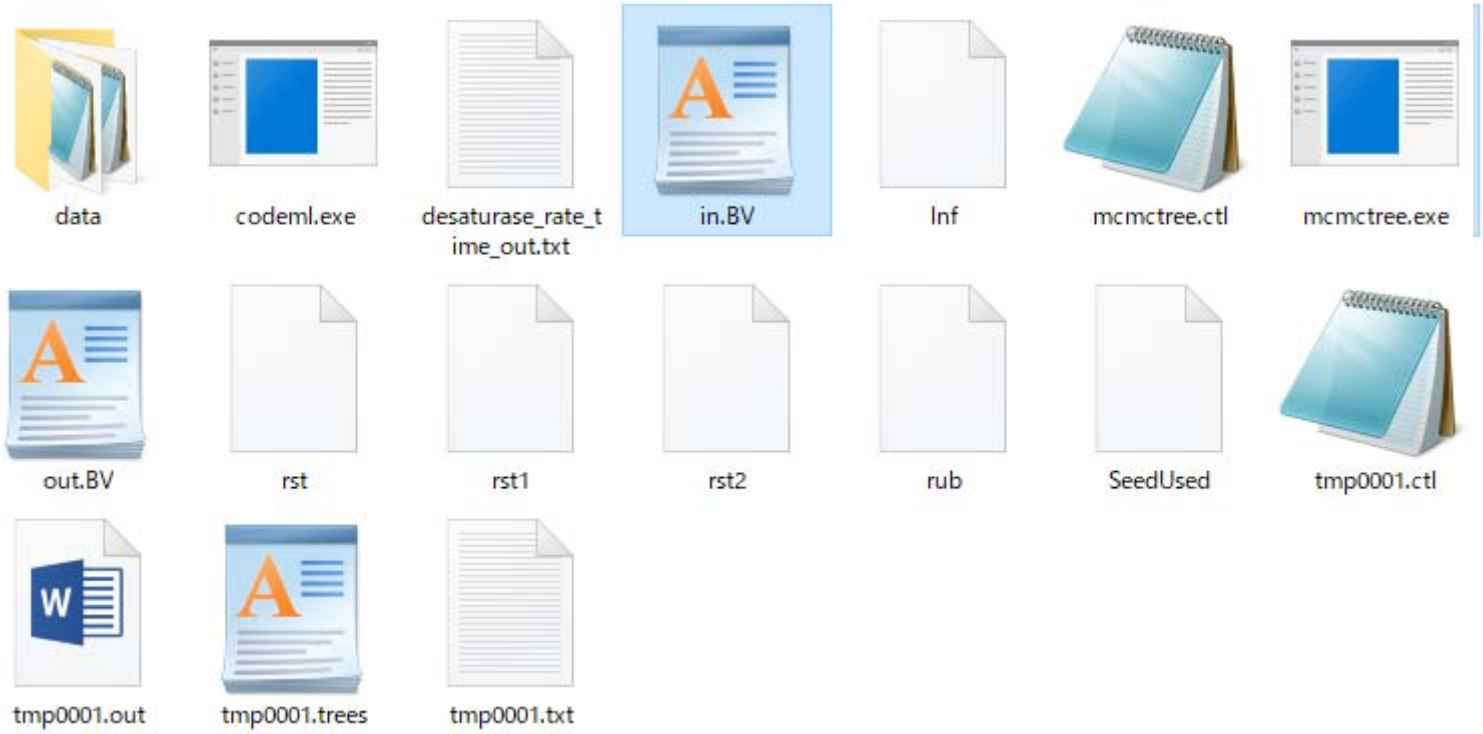

C:¥Windows¥system32¥cmd.exe
np =   29
lnL0 = -7590.604455
Round 1b: Blengths (29, e=1e-008)
lnL0 =   -7590.604455
Cycle 1:   -7453.493162
Cycle 2:   -7444.372334
Cycle 3:   -7444.251636
Cycle 4:   -7444.249206
Cycle 5:   -7444.249168
Cycle 6:   -7444.249168
Cycle 7:   -7444.249168
0:01
lnL = -7444.249168
Out..
lnL = -7444.249168
3 lfun, 0 eigenQcodon, 0 P(t)
Calculating SE's
Time used: 0:01
C:¥Users¥kishino¥Desktop¥paml¥_exercise¥desaturase_test>

Check the Hessian matrix (Fisher information matrix, the inverse of variance matrix)

```
 16

((((((HzeZ9_18: 0.04483, TniZ9_18: 0.02377): 0.04134, BmoZ9_18: 0.17606): 0.04922,
EpoZ9_18: 0.11299): 0.07354, OfuZ9_18: 0.17289): 0.11344, (PocZ9_16: 0.07747, (OfuZ9_16:
0.05978, HzeZ9_16: 0.08408): 0.03100): 0.16540): 0.09478, ((HzeZ/E11: 0.15490, TniZ/E10:
0.18498): 0.04105, (PocZ/E11: 0.32604, (BmoZ10: 0.28133, (OfuZ/E10: 0.29105, (AveZ10:
0.12347, EpoZ10: 0.16571): 0.12707): 0.04372): 0.03392): 0.06189): 0.20800, Tick:
0.44407);

  0.094784  0.113441  0.073539  0.049216  0.041343  0.044832  0.023770  0.176059
0.112987  0.172889  0.165402  0.077466  0.031004  0.059777  0.084081  0.208004  0.041046
0.154899  0.184983  0.061891  0.326038  0.033923  0.281326  0.043718  0.291054  0.127074
0.123466  0.165712  0.444066

  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
0.000000  0.000000  0.000000


Hessian

      -1610     -196.7     -153.7     -711.5     -258.2     -495.1     -340.9     -170.2
-370.3     -253.3     -193.7     -111.7     -345.5     -226.9     -559.8     -308.6
-127.9     -301.7     -92.83     -196.4     -66.51     -295.5     -52.87     -98.44
-69.44       -363      33.59     -111.8     -147.7

. . . . . .
```

out.BV

Copy out.BV and rename as in.BV as an input file of the second stage analysis



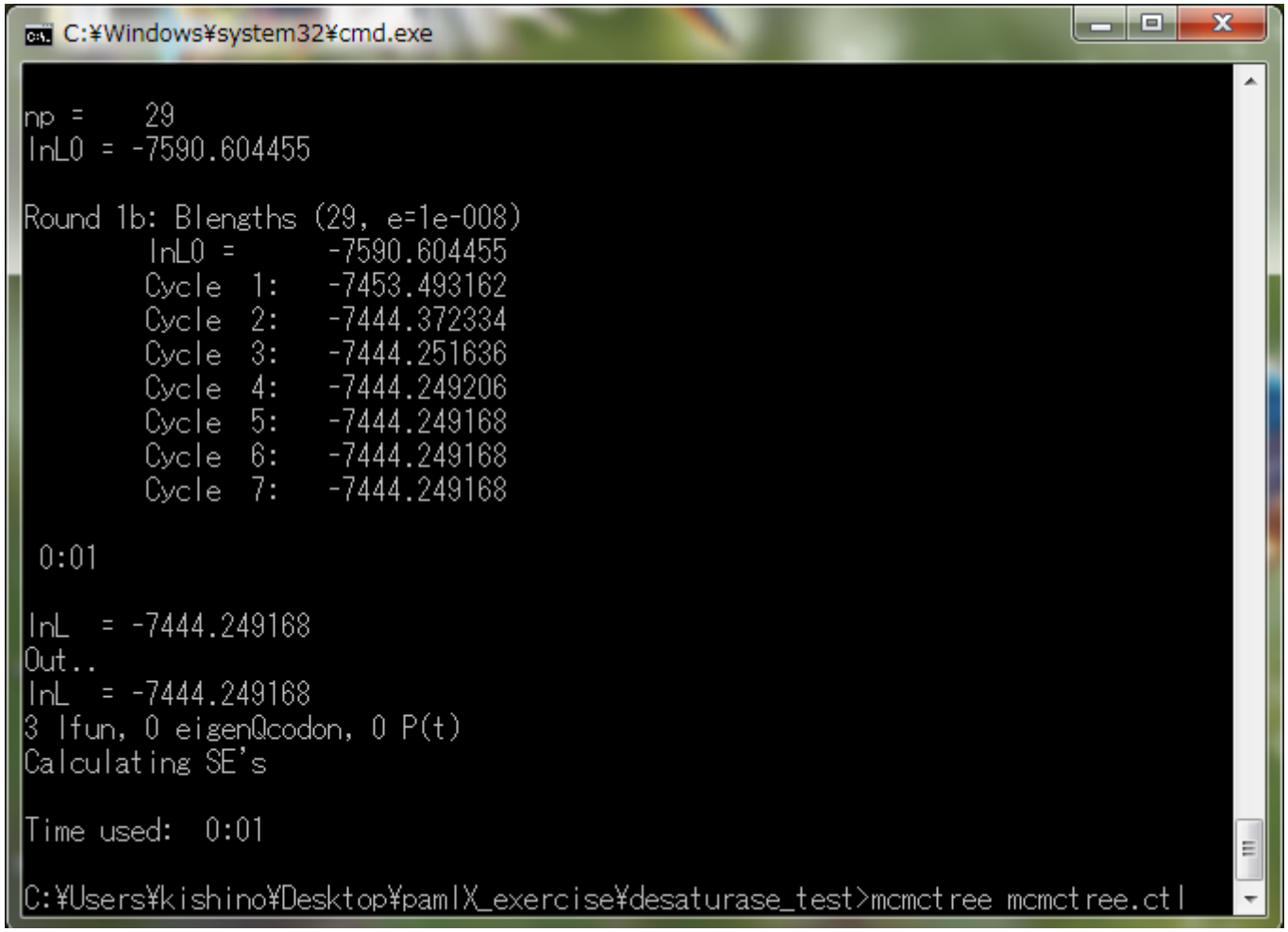
Set “usedata” to 2 in the cntl file (the second stage analysis)

```
          seed = -1
       seqfile = data/desaturase_sub.fasta
      treefile = data/desaturase_sub_node_ref_for_paml.nwk
       outfile = desaturase_rate_time_out.txt

         ndata = 1
       seqtype = 2    * 0: nucleotides; 1:codons; 2:AAs
       usedata = 2    * 0: no data; 1:seq like; 2:normal approximation; 3:out.BV (in.BV)
         clock = 2    * 1: global clock; 2: independent rates; 3: correlated rates
       RootAge = <1.0  * safe constraint on root age, used if no fossil for root.

    aaRatefile = data/lg.dat  * only used for aa seqs with model=empirical(_F)
         model = 0    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
         alpha = 0    * alpha for gamma rates at sites
         ncatG = 5    * No. categories in discrete gamma

     cleandata = 0    * remove sites with ambiguity data (1:yes, 0:no)?

        BDparas = 1 1 0.1   * birth, death, sampling
   kappa_gamma = 6 2      * gamma prior for kappa
   alpha_gamma = 1 1      * gamma prior for alpha

   rgene_gamma = 2 2     * gamma prior for rate for genes
  sigma2_gamma = 1 10    * gamma prior for sigma^2     (for clock=2 or 3)

      finetune = 1: .1 .1 .1 .1 .1 .1
* auto (0 or 1) : auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

*      finetune = 1: 0.05  0.2  0.15  0.1 .5
* auto (0 or 1) : times, rates, mixing, paras, RateParas, FossilErr

         print = 2 (print = 2 to print rates for branches with clock=2 or 3)
        burnin = 2000
      sampfreq = 2
       nsample = 20000

*** Note: Make your window wider (100 columns) before running the program.
```

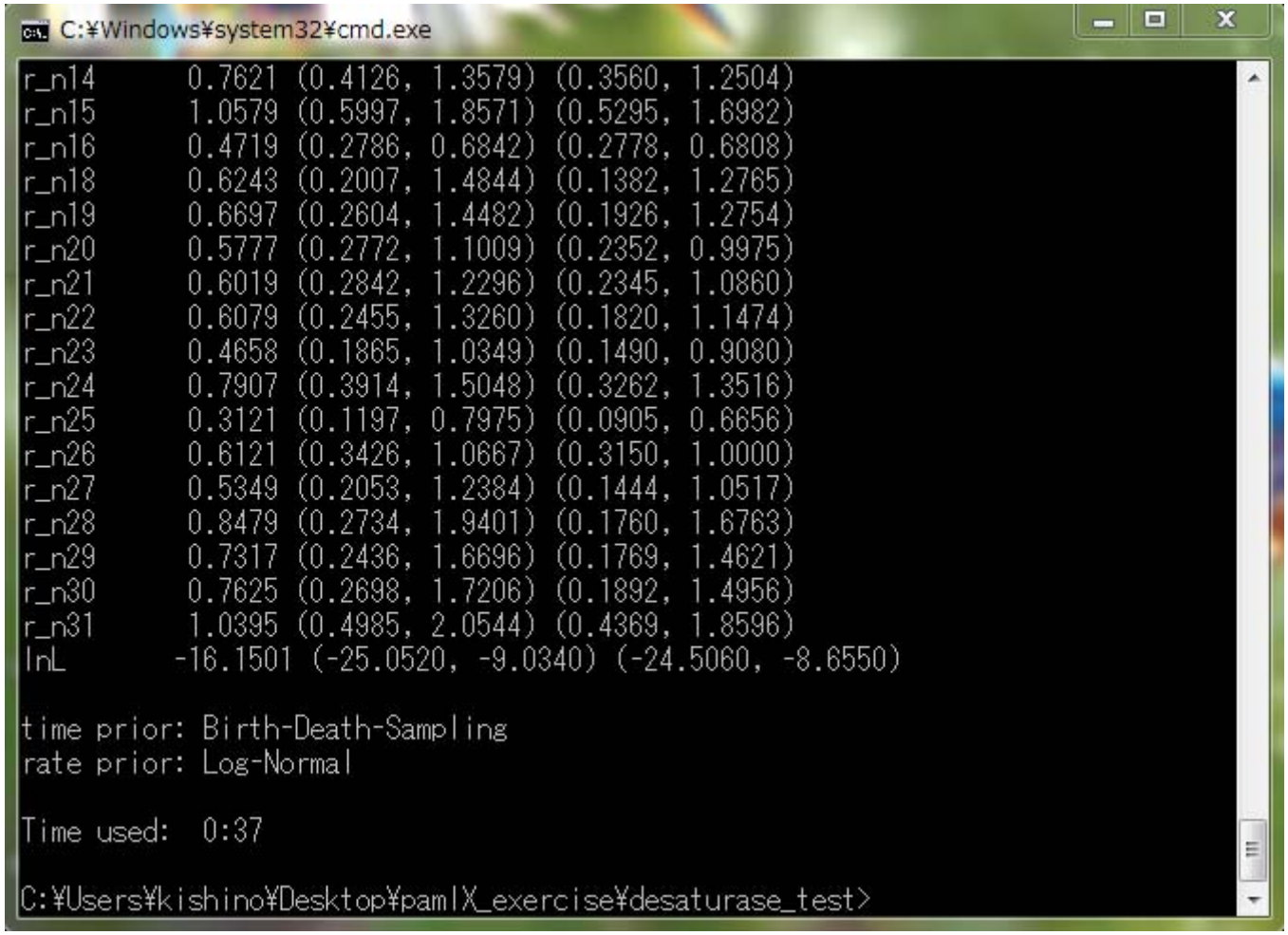
Estimate the times and rates using in.BV as an input data

mcmctree.ctl

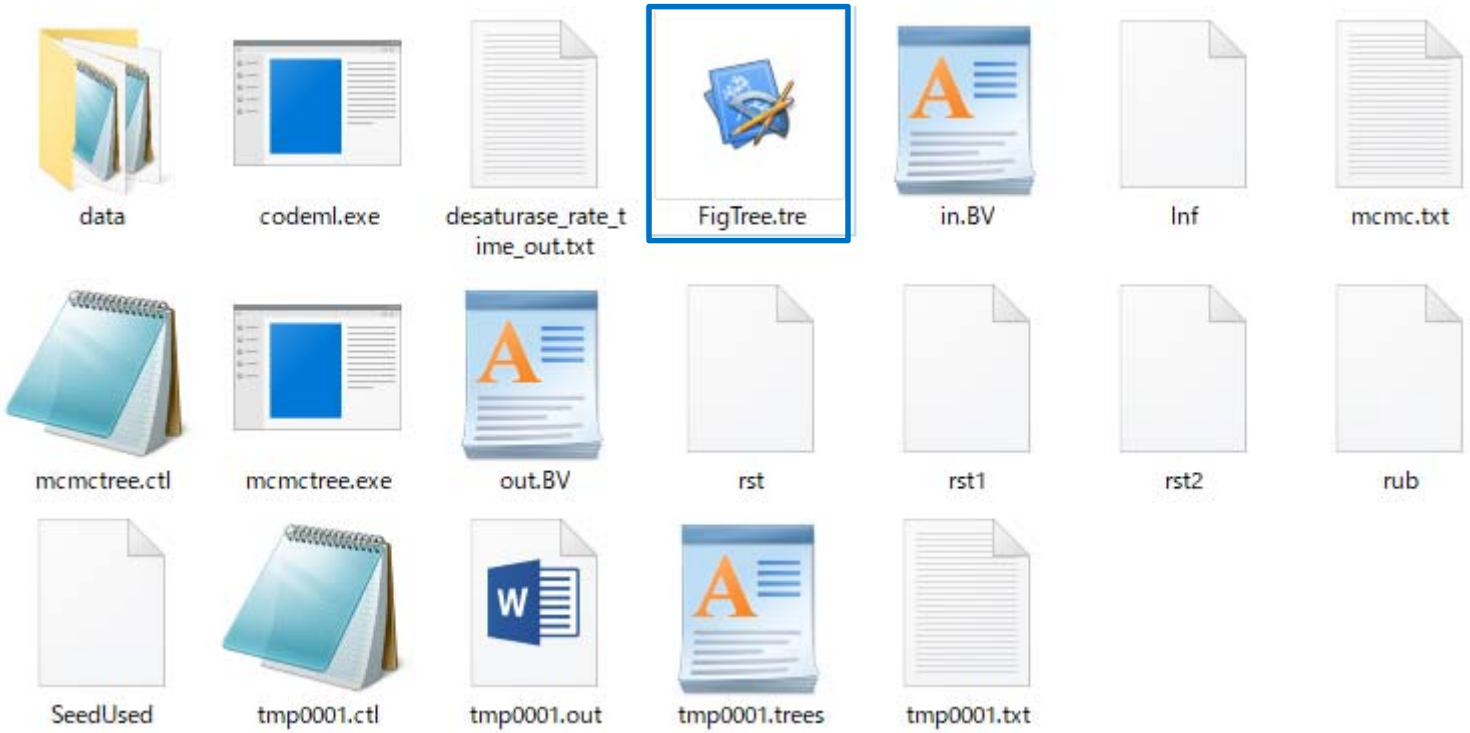
Run mcmctree.exe



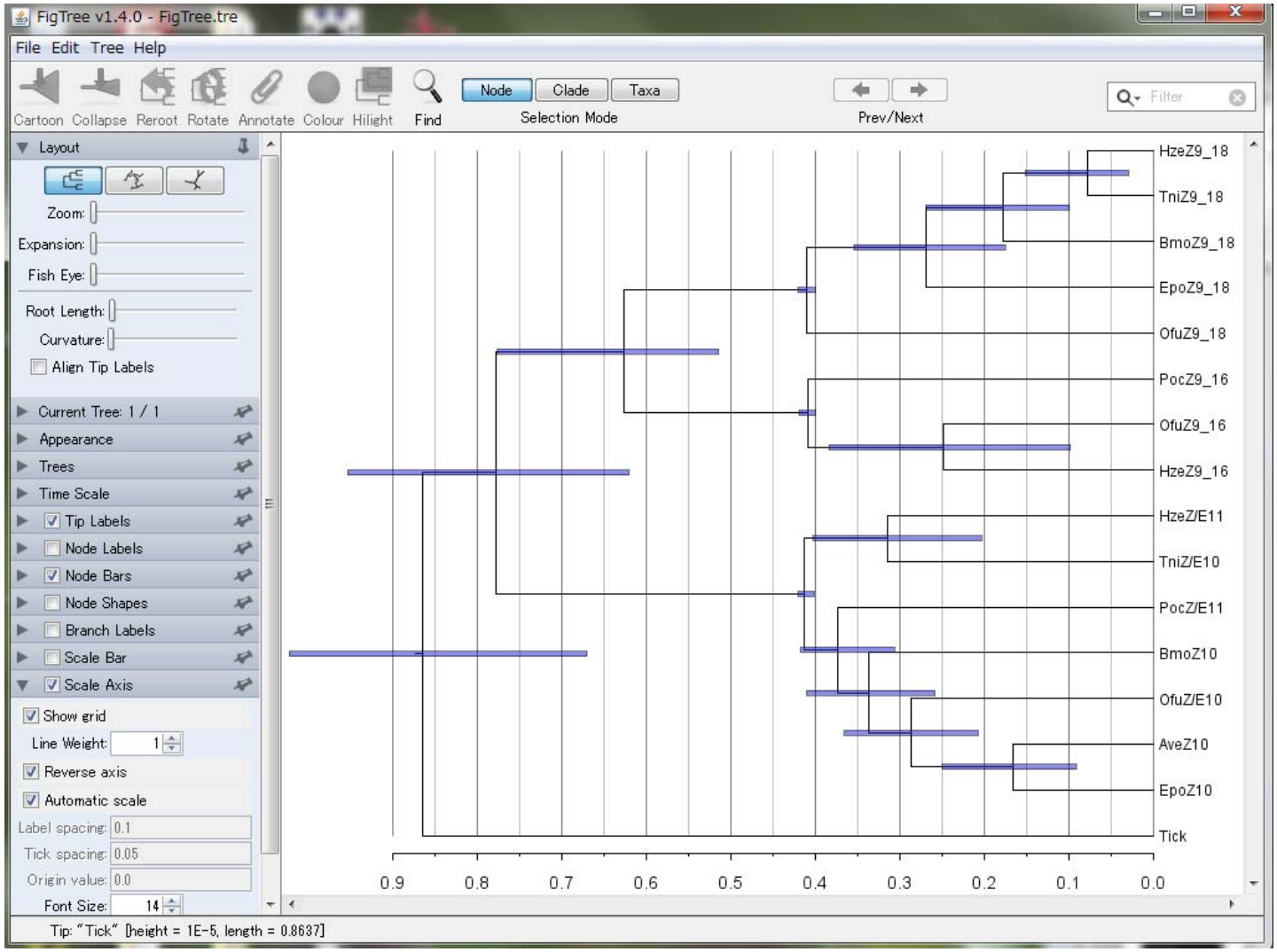
Done


C:¥Windows¥system32¥cmd.exe
r_n14 0.7621 (0.4126, 1.3579) (0.3560, 1.2504)
r_n15 1.0579 (0.5997, 1.8571) (0.5295, 1.6982)
r_n16 0.4719 (0.2786, 0.6842) (0.2778, 0.6808)
r_n18 0.6243 (0.2007, 1.4844) (0.1382, 1.2765)
r_n19 0.6697 (0.2604, 1.4482) (0.1926, 1.2754)
r_n20 0.5777 (0.2772, 1.1009) (0.2352, 0.9975)
r_n21 0.6019 (0.2842, 1.2296) (0.2345, 1.0860)
r_n22 0.6079 (0.2455, 1.3260) (0.1820, 1.1474)
r_n23 0.4658 (0.1865, 1.0349) (0.1490, 0.9080)
r_n24 0.7907 (0.3914, 1.5048) (0.3262, 1.3516)
r_n25 0.3121 (0.1197, 0.7975) (0.0905, 0.6656)
r_n26 0.6121 (0.3426, 1.0667) (0.3150, 1.0000)
r_n27 0.5349 (0.2053, 1.2384) (0.1444, 1.0517)
r_n28 0.8479 (0.2734, 1.9401) (0.1760, 1.6763)
r_n29 0.7317 (0.2436, 1.6696) (0.1769, 1.4621)
r_n30 0.7625 (0.2698, 1.7206) (0.1892, 1.4956)
r_n31 1.0395 (0.4985, 2.0544) (0.4369, 1.8596)
lnL -16.1501 (-25.0520, -9.0340) (-24.5060, -8.6550)
time prior: Birth-Death-Sampling
rate prior: Log-Normal
Time used: 0:37
C:¥Users¥kishino¥Desktop¥paml¥_exercise¥desaturase_test>

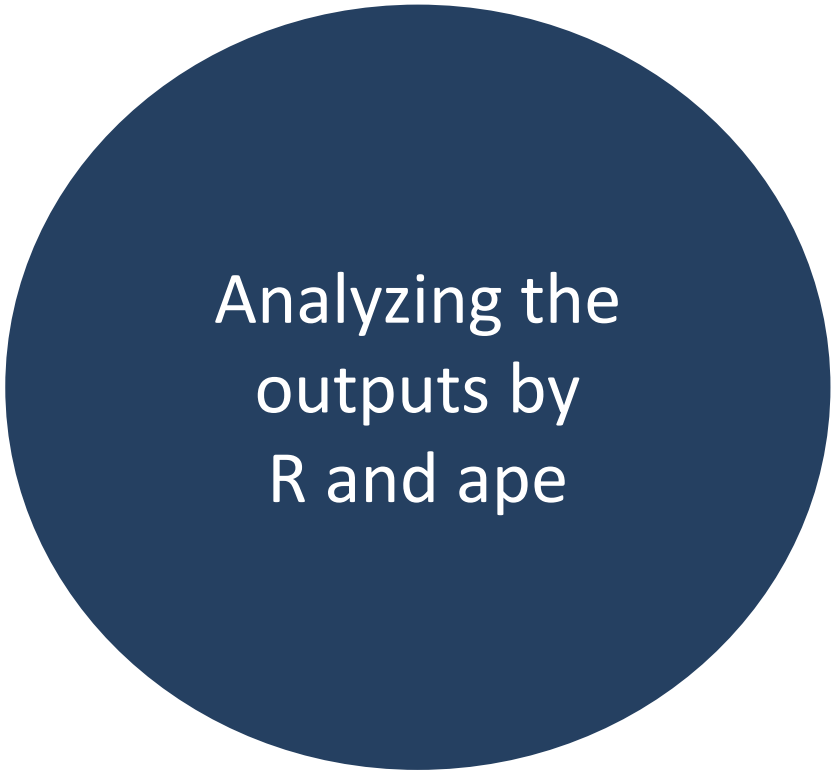
Times on the tree: FigTree.tre



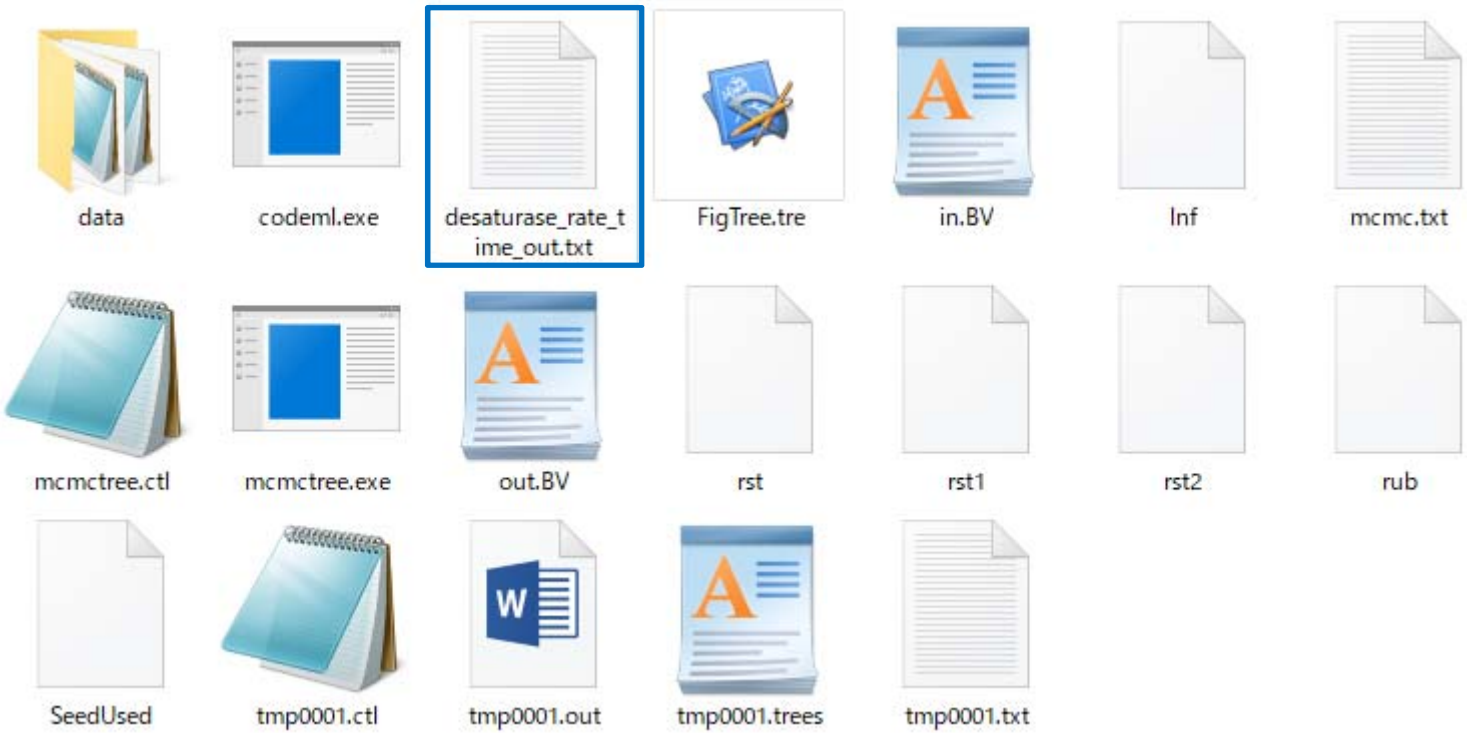
Look at the posterior distributions by FigTree



Confirm that the credibility intervals are much narrower than priors.


Analyzing the
outputs by
R and ape

Check paml output file for further analysis



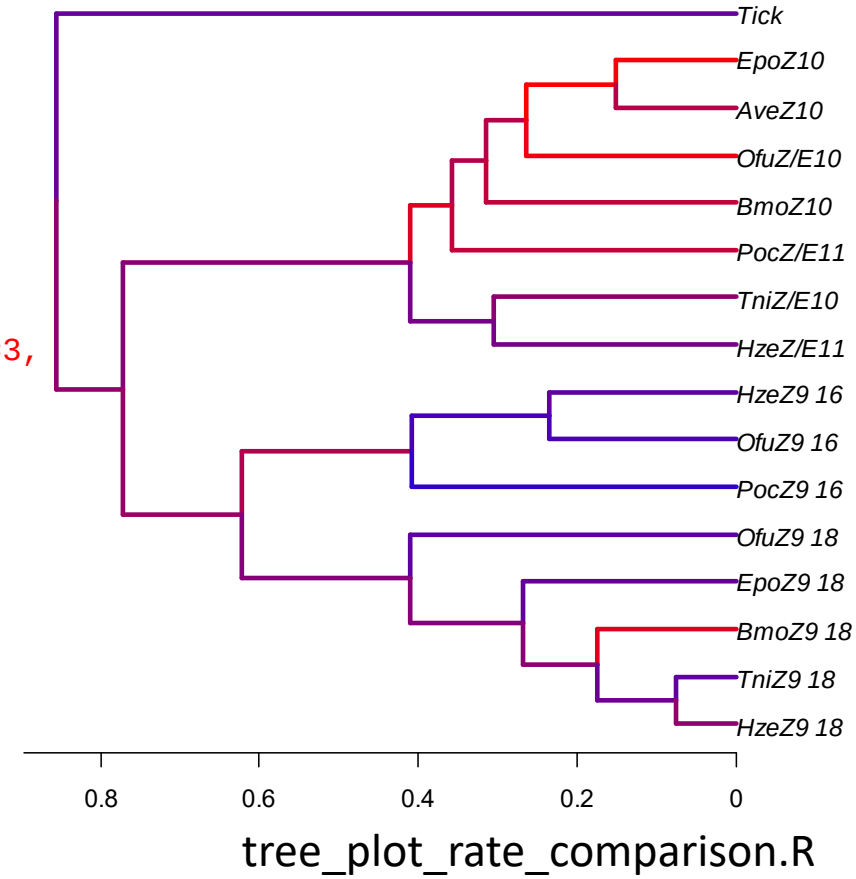
Check paml output file for further analysis

```
MCMCTREE (paml version 4.8a, July 2014) data/desaturase_sub.fasta

*** Locus 1 ***
     16    353

HzeZ9_18                       MPPQGQTGGS WVLYETDAVN EDTDAPVIVP PSAEKREWKI VWRNVILMGM LHIGGVYGAY LFLTTAMWRT CIFAVVLYIC SGLGITAGAH
RLWAHKSYKA RLPLRLMLTL FNTLAFQDAV IDWARDHRMH HKYSETDADP HNATRGFFFA HVGWLLVRKH PQIKAKGHTI DLSDLKSDPI LRFQKKYYLF LMPLVCFILP CYIPT-LWGE
   . . . . .

Species tree for FigTree.  Branch lengths = posterior mean times; 95% CIs = labels    Topology with node label -> tree_label.tre
(((((((1_HzeZ9_18, 2_TniZ9_18) 23 , 3_BmoZ9_18) 22 , 4_EpoZ9_18) 21 , 5_OfuZ9_18) 20 , (6_PocZ9_16, (7_OfuZ9_16, 8_HzeZ9_16) 25 ) 24 ) 19 ,
((9_HzeZ/E11, 10_TniZ/E10) 27 , (11_PocZ/E11, (12_BmoZ10, (13_OfuZ/E10, (14_AveZ10, 15_EpoZ10) 31 ) 30 ) 29 ) 28 ) 26 ) 18 , 16_Tick) 17 ;
                                                                                    Tree with branch length -> mean_tree.tre
(((((((HzeZ9_18: 0.075897, TniZ9_18: 0.075897): 0.099629, BmoZ9_18: 0.175527): 0.092779, EpoZ9_18: 0.268306): 0.141744, OfuZ9_18: 0.410049):
0.211851, (PocZ9_16: 0.408442, (OfuZ9_16: 0.235627, HzeZ9_16: 0.235627): 0.172815): 0.213458): 0.149084, ((HzeZ/E11: 0.304699, TniZ/E10:
0.304699): 0.106586, (PocZ/E11: 0.357231, (BmoZ10: 0.314663, (OfuZ/E10: 0.263869, (AveZ10: 0.153037, EpoZ10: 0.153037): 0.110832):
0.050794): 0.042567): 0.054054): 0.359699): 0.084940, Tick: 0.855924);
                                                                                    node times with 95% credibility intervals
(((((((HzeZ9_18: 0.075897, TniZ9_18: 0.075897) [&95%={0.0286713, 0.151527}]: 0.099629, BmoZ9_18: 0.175527) [&95%={0.0979479, 0.268361}]:
0.092779, EpoZ9_18: 0.268306) [&95%={0.17363, 0.355521}]: 0.141744, OfuZ9_18: 0.410049) [&95%={0.399884, 0.419941}]: 0.211851, (PocZ9_16:
0.408442, (OfuZ9_16: 0.235627, HzeZ9_16: 0.235627) [&95%={0.0832137, 0.3774}]: 0.172815) [&95%={0.399804, 0.419801}]: 0.213458)
[&95%={0.511311, 0.77471}]: 0.149084, ((HzeZ/E11: 0.304699, TniZ/E10: 0.304699) [&95%={0.182274, 0.385049}]: 0.106586, (PocZ/E11: 0.357231,
(BmoZ10: 0.314663, (OfuZ/E10: 0.263869, (AveZ10: 0.153037, EpoZ10: 0.153037) [&95%={0.0861759, 0.225652}]: 0.110832) [&95%={0.192473,
0.324582}]: 0.050794) [&95%={0.246292, 0.368669}]: 0.042567) [&95%={0.304718, 0.399086}]: 0.054054) [&95%={0.400375, 0.420166}]: 0.359699)
[&95%={0.616124, 0.949013}]: 0.084940, Tick: 0.855924) [&95%={0.666276, 1.01898}];

rategram locus 1:                                                                   Tree with edge length of rates -> mean_tree_rate.tre
(((((((HzeZ9_18: 0.671116, TniZ9_18: 0.438984): 0.499720, BmoZ9_18: 1.017230): 0.634089, EpoZ9_18: 0.452075): 0.614769, OfuZ9_18: 0.431126):
0.601047, (PocZ9_16: 0.219907, (OfuZ9_16: 0.334987, HzeZ9_16: 0.439319): 0.325804): 0.818299): 0.706660, ((HzeZ/E11: 0.544436, TniZ/E10:
0.641355): 0.542417, (PocZ/E11: 0.902217, (BmoZ10: 0.894612, (OfuZ/E10: 1.083523, (AveZ10: 0.826472, EpoZ10: 1.136717): 1.164701):
0.832367): 0.811453): 1.015589): 0.628988): 0.667502, Tick: 0.472998);


Posterior mean (95% Equal-tail CI) (95% HPD CI) HPD-CI-width

t_n17         0.8561 (0.6663, 1.0190) (0.6772, 1.0265) 0.3493 (Jnode 30)
t_n18         0.7710 (0.6161, 0.9490) (0.6145, 0.9464) 0.3319 (Jnode 29)
 . . . . .
```

desaturase_rate_time_out.txt

Draw the resultant tree with times and rates by R (ape)

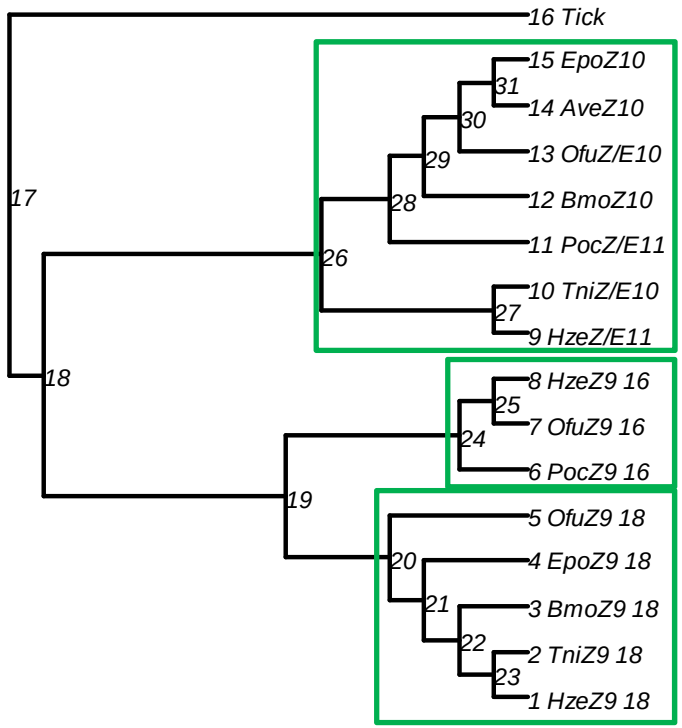
```
library(ape)
read.tree("mean_tree.tre")->tree
read.tree("mean_tree_rate.tre")->tree.rate
  ##### mean_tree.tre, mean_tree_rate.tre                  #####
  ##### were made by copy/paste from output file           #####
  ##### desaturase_rate_time_out.txt                       #####
  ##########################################################

names(tree)
tree$edge
tree$tip.label

rate0 <- tree.rate$edge.length
rel_rate <- rate0/max(rate0)
edge.colors <- rgb(rel_rate,0,1-rel_rate)

plot(tree,edge.color=edge.colors,edge.width=3,
     show.node.label=TRUE)
axisPhylo(side = 1)
```



tree_plot_rate_comparison.R

Checking the node numbers for further analysis

```
read.tree("tree_label.tre")->tree.label
  ##### tree_label.tre                                          #####
  ##### was made by copy/paste from output file                 #####
  ##### desaturase_rate_time_out.txt                            #####
  ###############################################################

plot(tree.label,edge.width=3,show.node.label=TRUE)
```



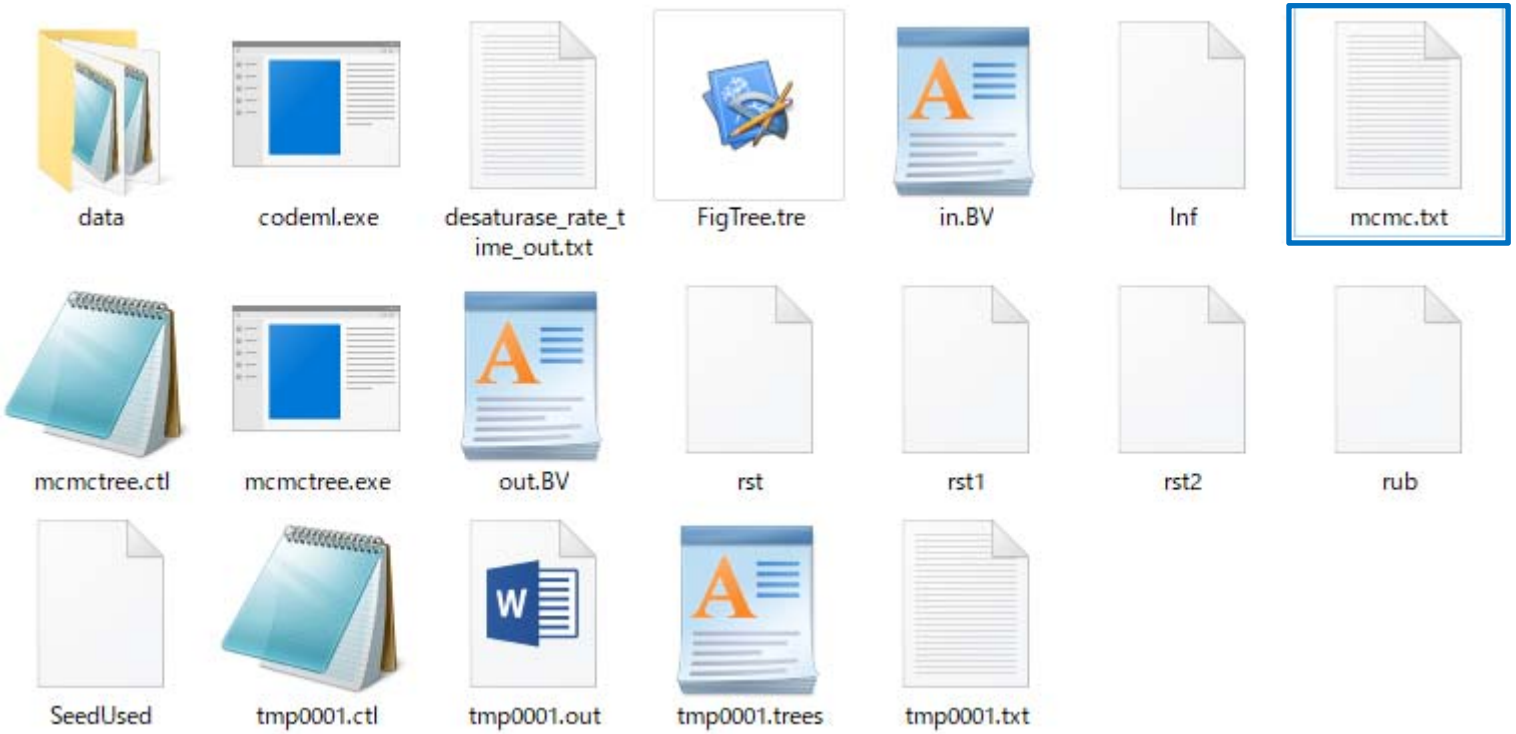

```
branches3 <- c(9:15, 27:31)
```

```
branches2<- c(6:8, 25)
```

```
branches1<-c(1:5, 21:23)
```

tree_plot_rate_comparison.R

MCMC sample of parameters: mcmc.txt



mcmc.txt by excel

Divergence times at nodes (node ages)

Evolutionary rates along branches

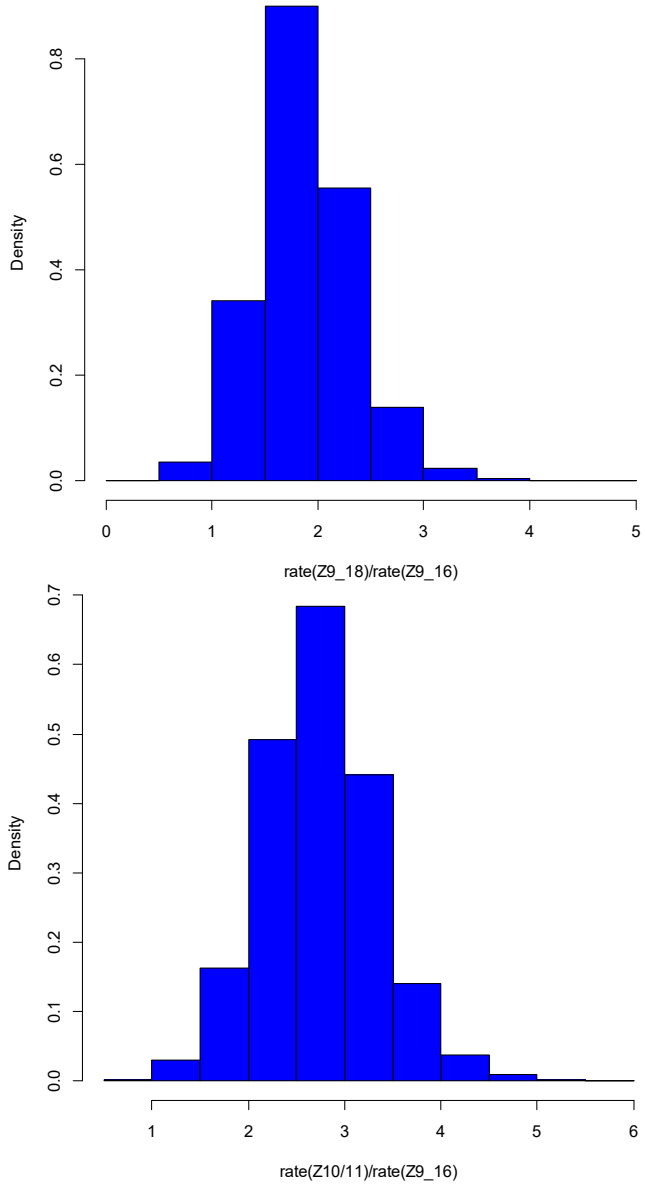
Log likelihood

Gen	t_n17	t_n18	t_n19	…	r_n1	r_n2	r_n3	……	lnL
1	0.693	0.656	0.594	…	1.309	0.618	2.286	…	-11.845
2	0.684	0.647	0.586	…	1.327	0.391	2.318	…	-10.640
4	0.649	0.647	0.586	…	1.327	0.289	2.318	…	-16.021
6	0.658	0.656	0.613	…	0.954	0.441	2.287	…	-13.464
8	0.648	0.641	0.587	…	1.819	0.491	2.294	…	-9.788
10	0.648	0.623	0.548	…	1.191	0.718	2.294	…	-11.819
12	0.602	0.586	0.562	…	1.053	0.682	2.254	…	-19.611
14	0.641	0.630	0.576	…	1.043	0.831	2.254	…	-12.292
16	0.675	0.637	0.583	…	1.032	1.165	1.517	…	-19.440
18	0.732	0.678	0.573	…	1.359	1.184	1.542	…	-17.279
20	0.700	0.677	0.572	…	1.279	0.710	1.695	…	-16.417
22	0.696	0.673	0.569	…	1.794	0.838	2.322	…	-10.390
24	0.686	0.664	0.561	…	1.315	1.163	1.511	…	-14.138
26	0.686	0.664	0.561	…	0.942	0.450	1.856	…	-16.249
28	0.696	0.673	0.569	…	0.564	0.650	1.829	…	-24.558
30	0.681	0.673	0.569	…	0.412	1.282	1.297	…	-20.841
32	0.681	0.673	0.569	…	0.464	1.029	1.297	…	-22.368
34	0.672	0.665	0.562	…	0.593	0.599	1.313	…	-21.875
36	0.667	0.659	0.557	…	0.923	0.604	1.324	…	-18.483
…	…	…	…	…	…	…	…	…	…

mcmc.txt

Comparing rate(Z9-16), rate(Z9-18), and rate(Z10/11)

```
read.table("mcmc.txt",header=T)->data0
names(data0)

branches1 <- c(1:5,21:23)
branches2 <- c(6:8,25)
branches3 <- c(9:15,27:31)

b1_label <- paste("r_n",branches1,sep="")
b2_label <- paste("r_n",branches2,sep="")
b3_label <- paste("r_n",branches3,sep="")

b1.mean <- apply(data0[,b1_label],1,mean)
b2.mean <- apply(data0[,b2_label],1,mean)
b3.mean <- apply(data0[,b3_label],1,mean)

hist(b1.mean/b2.mean,main="",probability=T,col="blue",
   xlab="rate(Z9_18)/rate(Z9_16)")
hist(b3.mean/b2.mean,main="",probability=T,col="blue",
   xlab="rate(Z10/11)/rate(Z9_16)")
```



tree_plot_rate_comparison.R

Comparing rate(Z9-16), rate(Z9-18), and rate(Z10/11)

- The posterior probability that average rate of Z9-18 clade is larger than that of Z9-16 clade

```
mean(b1.mean/b2.mean>1)
```

```
[1] 0.9826009
```

- The posterior probability that average rate of Z10/11 clade is larger than that of Z9-16 clade

```
mean(b3.mean/b2.mean>1)
```

```
[1] 0.99935
```

- The posterior probability that average rate of Z9-18 clade and average rate of Z10/11 clade are **both** larger than that of Z9-16 clade

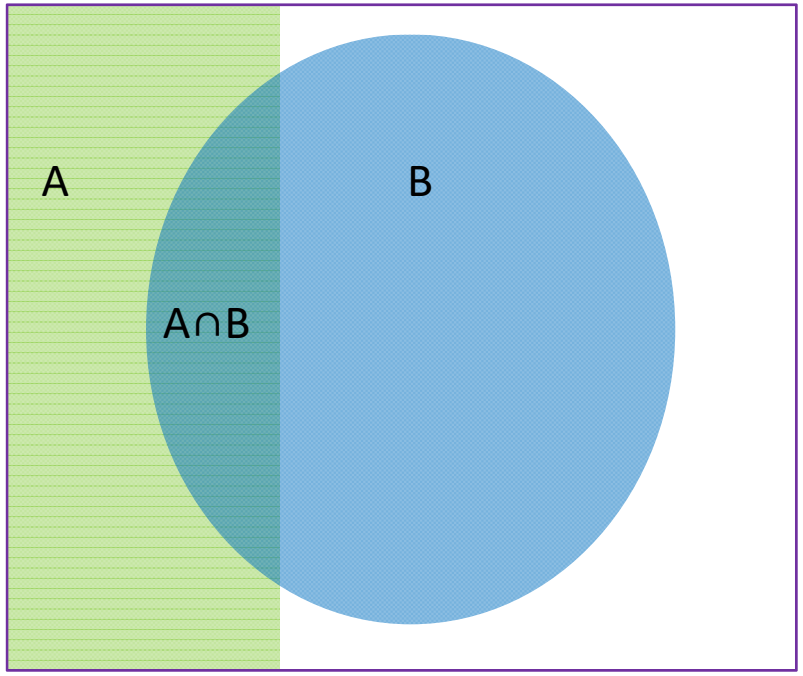
```
mean(b1.mean/b2.mean>1 & b3.mean/b2.mean>1)
```

```
[1] 0.9826009
```

tree_plot_rate_comparison.R

[coffee break]
Bayesian inference and MCMC

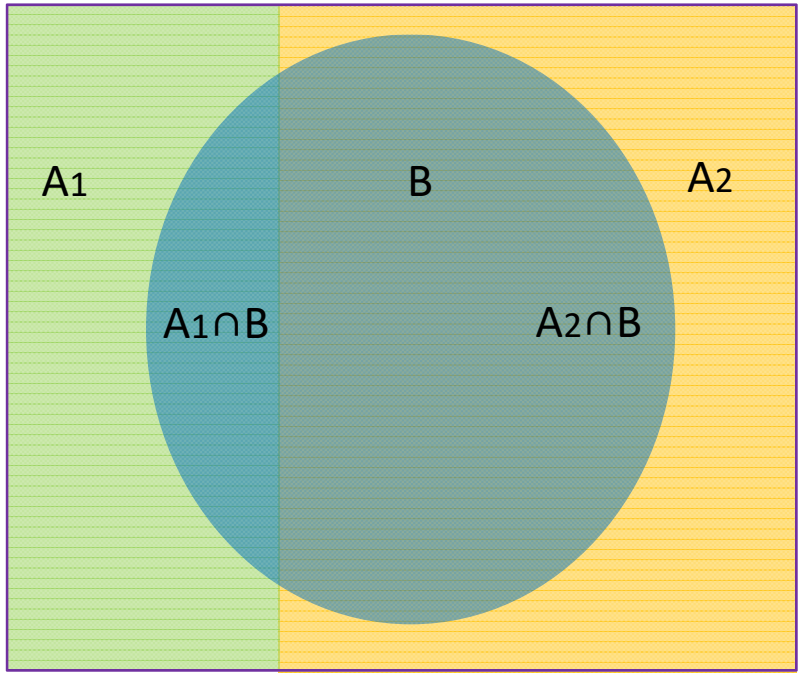
Conditional probability and Bayes formula



$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A} \mid \mathrm{B})P(\mathrm{B}) = P(\mathrm{B} \mid \mathrm{A})P(\mathrm{A})$$

$$P(\mathrm{A} \mid \mathrm{B}) = \frac{P(\mathrm{B} \mid \mathrm{A})P(\mathrm{A})}{P(\mathrm{B})}$$

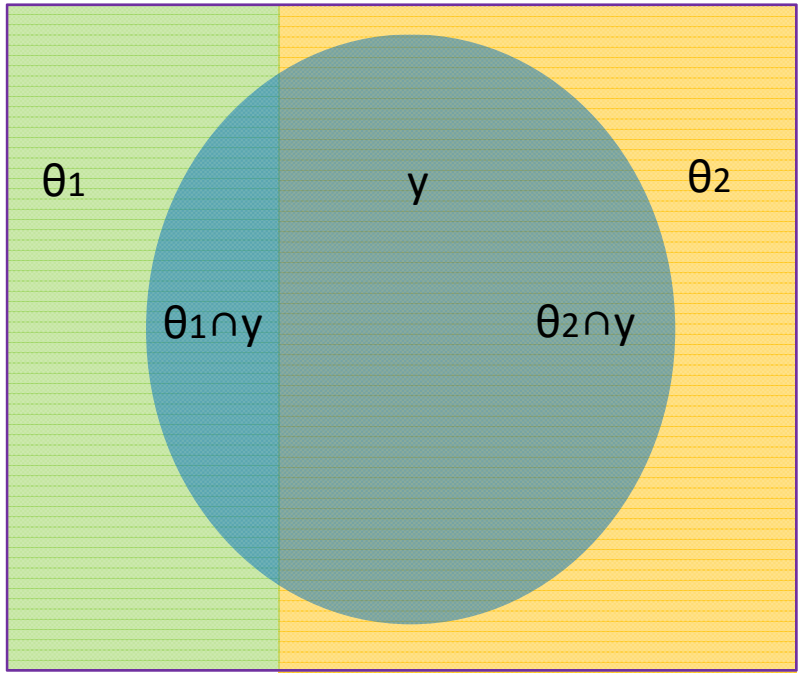
Conditional probability and Bayes formula



$$P(\mathrm{A}_1 \mid \mathrm{B}) = \frac{P(\mathrm{B} \mid \mathrm{A}_1)P(\mathrm{A}_1)}{P(\mathrm{B})}$$

$$P(\mathrm{A}_2 \mid \mathrm{B}) = \frac{P(\mathrm{B} \mid \mathrm{A}_2)P(\mathrm{A}_2)}{P(\mathrm{B})}$$

Conditional probability and Bayes formula

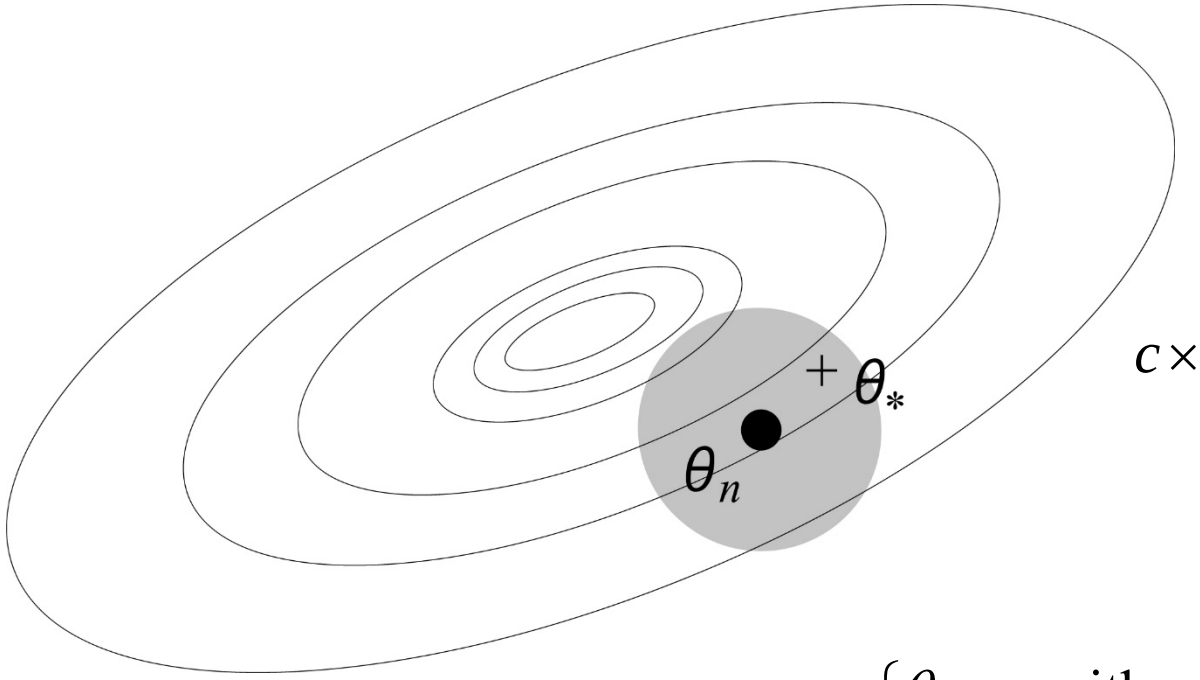


likelihood prior

$$P(\theta_1 \mid \mathrm{y}) = \frac{P(\mathrm{y} \mid \theta_1)P(\theta_1)}{P(\mathrm{y})}$$

$$P(\theta_2 \mid \mathrm{y}) = \frac{P(\mathrm{y} \mid \theta_2)P(\theta_2)}{P(\mathrm{y})}$$

Simulating posterior distribution by Markov chain Monte Carlo (MCMC)



$$c \times p(y \mid \theta) p(\theta)$$

$$\theta_{n+1} = \begin{cases} \theta_* & \text{with probability } r \\ \theta_n & \text{with probability } 1 - r \end{cases}$$

$$r = \min\left\{1, \frac{p(y \mid \theta_*) p(\theta_*)}{p(y \mid \theta_n) p(\theta_n)} \times \frac{h(\theta_n \mid \theta_*)}{h(\theta_* \mid \theta_n)}\right\}$$

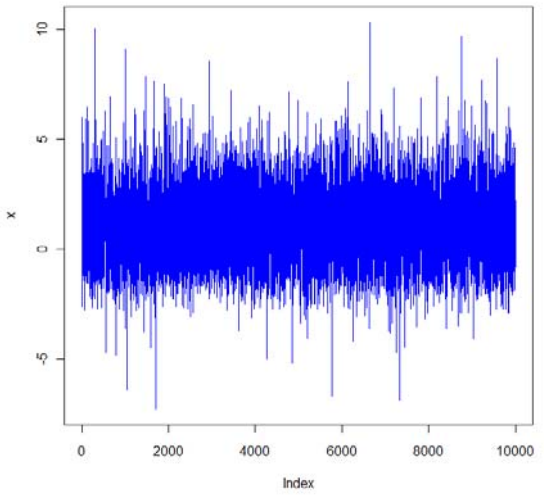
MCMC by R

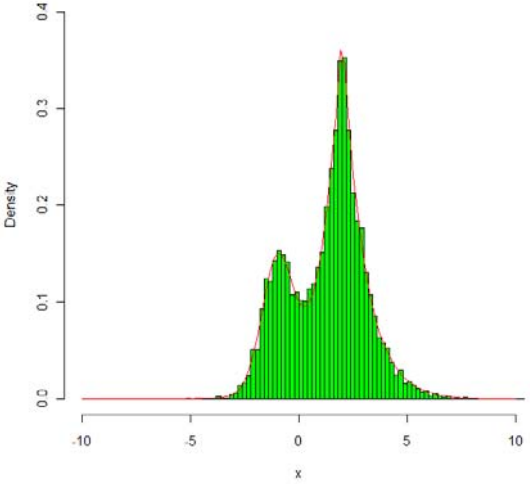
```
MCMC1 <- function(f,init,nsample,nthin,updatewidth){
  x <- NULL
  x0 <- init; y0 <- f(x0)
  for (i in 1:nsample){
    for (j in 1:nthin){
      x1 <- x0 + rnorm(1,sd=updatewidt)
      y1 <- f(x1)
      if(runif(1)<y1/y0) {x0 <- x1; y0 <- y1}
    }
    x <- c(x,x0)
  }
  x
 }
```

init	Initial value
nsample	Sample size to be simulated
nthin	Size of thinning to save the memory
updatewidth	Width for the proposal step
nburin	Burn-in period to be discarded from the sample

MCMC by R

```
func <- function(x){exp(-(x+1)^2)+3*exp(-abs(x-2))}

x <- MCMC1(func,0,10000,10,2)

plot(x,type="l",col="blue")

hist(x,probability=T,nclass=100,
     xlim=c(-10,10),ylim=c(0,0.4),col="green")

z <- seq(-10,10,length=100)
ftot <- 0
for (t in z) {ftot <- ftot+func(t)*20/100}
lines(z,func(z)/ftot,col="red")
```





Assignment 2

Please answer to either of the two:

1. We compared the average molecular evolutionary rates among the three paralogues, Z9_16, Z9_18, and Z/E10-11, after the origin of Lepidoptera. Please compare the molecular evolutionary rates of the two paralogues Z9_16 and Z9_18 before the origin of Lepidoptera.

2. In the exercise of desaturase, the time at the most recent common ancestor Lepidoptera (TMRCA) was assumed between 0.4 and 0.42. To see the effect of the assumption on the estimated divergence times and the pattern of rate variation, please analyze the data, assuming 0.5<TMRCA<0.6.

Please send the word file / pdf file named **agri2_name.doc** /**agri2_name.pdf** to:

Hirohisa Kishino (kishino@lbm.ab.a.u-tokyo.ac.jp).

Here, “name” should be replaced by your name.
Deadline: 5 June 2019 (Sunday)