PHYBAYES: A PROGRAM FOR PHYLOGENETIC ANALYSIS IN A BAYESIAN FRAMEWORK

Stéphane Aris-Brosou and Ziheng Yang
Department of Biology (Galton Laboratory), University College London, England

December 01

Suggested citation:

ARIS-BROSOU, S. AND Z. YANG. 2001a. Phybayes: a program for phylogenetic analyses in a Bayesian framework. Bioinformatics.

Disclaimer

This program is provided “as is”, without warranty of any kind. In no event shall the authors be held responsible for any damage resulting from the use of this software, including the frustration you may experience using it. This program is unsophisticated and unintentionally crude and user-rude, but those used to PAML should feel at home. This program is distributed free of charge for academic use only.

Using it assumes some basic knowledge Bayesian analysis and Markov chain Monte Carlo (MCMC) techniques for running the analyses and interpreting the results.

Reports of problems comments and questions concerning the program or this document are welcome. You can contact any of us via email (s.aris-brosou@ucl.ac.uk or z.yang@ucl.ac.uk) and check for updates regularly (http://abacus.gene.ucl.ac.uk/stephane/).

What can be done with PHYBAYES?

(1) estimating divergence times when the molecular clock is relaxed

The main feature of this program is to estimate divergence times under different models of rate change, either under the molecular clock, or under five models (lognormal, stationarised lognormal, gamma, exponential distributions or Ornstein-Uhlenbeck process) that relax the clock assumption.

For the theory and applications, see Aris-Brosou and Yang (2001b).

(2) testing phylogenetic trees as estimated from nucleotide sequences

As PHYBAYES computes the Bayes factor, it is possible to test phylogenetic hypotheses. Two types of tests can be performed: (i) test of a tree against, e.g., the maximum likelihood tree; (ii) construction of a confidence set of trees. The details of how to carry out these tests are given below.

For the theory and applications, see Aris-Brosou (2001).

How to format the data file?

Data file are formatted in the same way as for any program in PHYLIP (Felsenstein 1995) or PAML (Yang 1997), and several options are available. Following the PHYLIP format, the first line contains the number of taxa and the number of sites included in the analysis.

```
4  60
sequence 1
AAGCTTCACCGGCGCAGTCATTCTCATAAT
CGCCCACGGACTTACATCCTCATTACTATT
sequence 2
AAGCTTCACCGGCGCAATTATCCTCATAAT
CGCCCACGGACTTACATCCTCATTATTATT
sequence 3
AAGCTTCACCGGCGCAGTTGTTCTTATAAT
TGCCCACGGACTTACATCATCATTATTATT
sequence 4
AAGCTTCACCGGCGCAACCACCCTCATGAT
TGCCCATGGACTCACATCCTCCCTACTGTT
```

Species names should be less than 30 characters and should not contain symbols such as , : # (or) as these are reserved for special definitions. Species names can include spaces, but two consecutive space characters signify the end of the name to the program. In a sequence, three characters can be used, “.”, “–” or “?”, respectively for

“same character as in the first sequence”, “alignment gap” and undetermined site. Any site column with at least one of the last two characters (“–” or “?”) is excluded from the analysis. Characters `A`, `T`, `C`, `G`, `U`, `a`, `t`, `c`, `g` and `u` are recognized as nucleotides. Lines do not have to be equally long and the whole sequence can be put on one single line, or on several lines. Different options can be used.

Option G: this option is for combined analyses of heterogeneous data sets (multiple gene, the three codon positions, etc.). The sequences are concatenated and the option is used to specify which gene each site is from. The example data of Brown et al. (1982) are an 895-bp segment from the mitochondrial genome, which codes for parts of two proteins (ND4 and ND5) at the two ends and three tRNAs in the middle. Sites in the sequence fall naturally into 4 classes: the three codon positions and the tRNA coding region. The first line of the file contains the option character G. The second line begins with a G at the first column, followed by the number of site classes. The following lines contain the site marks, one for each site in the sequence. The site mark specifies which class each site is from. If there are *g* classes, the marks should be 1, 2, ..., *g*, and if *g* > 9, the marks need to be separated by spaces. The total number of marks must be equal to the total number of sites in each sequence.

```
5 895 G
G 4
3 123123123123123123123123123123123123123123123123123123123123
123123123123123123123123123123123123123123123123123123123123
123123123123123123123123123123123123123123123123123123123123
123123123123123123123123123123123123123123123123123123123123
123123123123123123123123123123123123123123123123123123123123
123123123123123123123123123123123123123123123123123123123123
123123123123123123123123123123123123123123123123123123123123
1231231231231231231231231231231231231
444444444444444444444444444444444444444444444444444444444444
444444444444444444444444444444444444444444444444444444444444
444444444444444444444444444444444444444444444444444444444444 444444444444444444
123123123123123123123123123123123123123123123123123123123123
123123123123123123123123123123123123123123123123123123123123
123123123123123123123123123123123123123123123123123123123123
12312312312312312312312312312312312312312312312312312312312
Human  AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGACTTACATCCTCATTACTATT
CTGCCTAGCAAACTCAAACTACGAACGCACTCACAGTCGCATCATAATC........
Chimpanzee  AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGACTTACATCCTCATTACTATT
CTGCCTAGCAAACTCAAACTACGAACGCACTCACAGTCGCATCATAATC........
........
```

If the data are concatenated sequences of multiple genes, a simpler format, shown below for an example data set, may be used. This sequence has 1000 nucleotides from 4 genes, obtained from concatenating four genes with 100, 200, 300, and 400 nucleotides from genes 1, 2, 3, and 4, respectively. The "lengths" for the genes must be on the line that starts with G, *i.e.*, on the second line of the sequence file. (This

requirement allows the program to determine which of the two formats is being used.) The sum of the lengths for the genes should be equal to the number of nucleotides.

```
5 1000 G
G 4 100 200 300 400
Sequence 1

TCGATAGATAGGTTTTAGGGGGGGGGGTAAAAAAAAA.......
```

How to format the tree file?

The file name is specified in the appropriate control file. Its default name is `mcmctree.ctl` (see below), but can be changed (the program is the run by giving the name of the user-defined control file as an argument). Two methods for representing a tree topology are used in PAML. The first is the familiar parenthesis representation, that is used in virtually any phylogenetic software. The species can be represented using either their names or their indexes corresponding to the order of their occurrences in the sequence data file. If species names are used, they have to match exactly those in the sequence data file (including spaces or strange characters). Branch lengths are allowed. The following is a possible tree structure file for a data set of four species (human, chimpanzee, gorilla, and orang-utan, occurring in this order in the data file). The first tree is a star tree, while the next four trees are the same.

```
4 1 // 4 species, 1 tree (the first one) taken into account by PHYBAYES
(((12)3)4)    // species 1 and 2 are clustered together
(((1,2),3),4) // Commas are needed with more than 9 species
(((human,chimpanzee),gorilla),orangutan);
(((human:.1,chimpanzee:.2):.05,gorilla:.3),orangutan:.5);
```

Input trees must be rooted. However, whether the tree will be considered rooted or unrooted depends on the type of analysis. To estimate divergence times, the trees are rooted. When testing hypotheses, the trees are considered unrooted, so that ((12)(34)) is the same as ((12)3)4)).

How to format control file?

The various options of PhyBayes are set by means of a control file (e.g. Figure 1). Defaults values are supplied for all the parameters of the model as well as for run settings. These must be optimised for each data set, in order to reach a satisfactory balance between acceptance rate and mixing of the Markov chain.

The default control file is `mcmctree.ctl`, for which examples are shown below (Figures 1 to 4), one for each type of analyses possible with PhyBayes. Note that spaces are required on both sides of the equal sign, and blank lines or lines beginning with "*" are treated as comments. The control file defines which type of analysis is performed. Examples are given below as indicative guidance.

(1) estimation of divergence times when the molecular clock is relaxed

This implements the models of speciation and of rate change as described in Aris-Brosou and Yang (2001b). They are respectively set with the options `EstTime=1` and `EstRates=1`, and by choosing the appropriate prior distributions for the divergence times (`PriorT`) and the rates of evolution (`PriorR`). The control file allows the user to set the starting values for hyperparameters of the generalised Birth-Death process for divergence times (`birth`, `death` and `sample`). The hyperparameters of the prior distribution for rates (`beta_oup` and `sigma2_oup`) are set to a value kept constant along the MCMC here.

Several possibilities exist to estimate divergence times under the molecular clock assumption. The first is to set the `clock` option to `1`; `EstTimes` and `EstRates` can be left to any value. The second set of options is to estimate divergence dates under a model of rate change (`clock=1`, `EstTimes=1` and `EstRates=1`, `PriorR` set to `0`, `1` or `4`), but not allowing rates to change along the lineages (e.g. `sigma2_oup=.0001`).

Note that for all these analyses, the tree file must contain a rooted tree.

```
** Files ********************************************************
   seqfile = HIV.nuc * sequence data file name
   outfile = HIV.out * main result file name
  treefile = HIV.tre * tree file for initial rooted tree topology
      seed = 1234567 * random number seed

** Model ********************************************************
     model = 4    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
     clock = 0    * 1:molecular clock; 0:no molecular clock
     kappa = 5    * kappa in K80, F84, or HKY85
     alpha = 0.5  * alpha for gamma rates at sites (0 if homogeneous rates)
       mut = .514 * initial mutation rate (for all the lineages)
     ncatG = 8    * No. categories in discrete gamma

** Parameters to integrate out by MCMC ****************************
 EstSimult = 3      * estimate all the parameters:
                    * 0: cyclically; 1: simultaneously;
                    * 2: random 1 @ a time; 3:==0+sample all parameters
   EstTopo = 0      * estimate topology: - NNI (EstTimes=EstRates=0)
  EstTimes = 1      * estimate divergence times
  EstRates = 1      * estimate rates of evolution
****when these last 2 are set to 0, only branch lengths are estimated**
  EstBases = 0      * estimate base composition
  EstAlpha = 0      * estimate alpha
  EstKappa = 0      * estimate kappa

** Running the MCMC & starting values for the tunings ***************
    burnin = 10000  * length of the burn-in period
 NbOfSteps = 10000  * number of sampled steps along the MCMC
  sam_intv = 100    * interval between two sampled steps

     tuneT = .01    * tuning parameter for time (ChangeTree)
     tuneK = .05    * tuning parameter for kappa in K80, F84, or HKY85
     tuneA = .025   * tuning parameter for alpha (gamma rates at sites)
     tuneR = .01    * tuning parameter for the rate of evolution (or branches)

     xMore = 0      * will U be likely 2 sample more steps from the MCMC?
     ndata = 1      * number of data sets (for simulation purpose)

** Specifications for the prior on rates ******************************
    PriorR = 3      * 0:lognormal; 1:"stationarized"-lognormal;
                    * 2:truncated normal; 3:Ornstein-Uhlenbeck process
                    * 4:Gamma; 5:Exponential.

  beta_oup = .001   * drift parameter (OUP: PriorR=3)
sigma2_oup = 10     * measure of variance for the rates (PriorR=0,1,2,4)
                    * diffusion parameter (OUP: PriorR=3)

** Specifications for the prior on times ******************************
    PriorT =  1     * prior 4 node times: 1:BDP; 2:Uniform; 3:Beta
     birth = 15.    * upper limit on birth rate (U prior)
     death =  5.    * upper limit on death rate (U prior)
    sample =   .01  * upper limit on sampling fraction (U prior)
```

Figure 1. Default control file to estimate divergence times and rates for a HIV data set.

```
** Files ********************************************************
   seqfile = HIV.nuc * sequence data file name
   outfile = HIV.out * main result file name
  treefile = HIV.tre * tree file for initial rooted tree topology
      seed = 1234567 * random number seed

** Model ********************************************************
     model = 4    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
     clock = 0    * 1:molecular clock; 0:no molecular clock
     kappa = 5    * kappa in K80, F84, or HKY85
     alpha = 0.5  * alpha for gamma rates at sites (0 if homogeneous rates)
       mut = .514 * initial mutation rate (for all the lineages)
     ncatG = 8    * No. categories in discrete gamma

** Parameters to integrate out by MCMC ****************************
 EstSimult = 3      * estimate all the parameters:
                    * 0: cyclically; 1: simultaneously;
                    * 2: random 1 @ a time; 3:==0+sample all parameters
   EstTopo = 0      * estimate topology: - NNI (EstTimes=EstRates=0)
  EstTimes = 0      * estimate divergence times
  EstRates = 0      * estimate rates of evolution
****when these last 2 are set to 0, only branch lengths are estimated**
  EstBases = 0      * estimate base composition
  EstAlpha = 0      * estimate alpha
  EstKappa = 0      * estimate kappa

** Running the MCMC & starting values for the tunings ***************
    burnin = 10000  * length of the burn-in period
 NbOfSteps = 10000  * number of sampled steps along the MCMC
  sam_intv = 100    * interval between two sampled steps

     tuneT = .01    * tuning parameter for time (ChangeTree)
     tuneK = .05    * tuning parameter for kappa in K80, F84, or HKY85
     tuneA = .025   * tuning parameter for alpha (gamma rates at sites)
     tuneR = .01    * tuning parameter for the rate of evolution (or branches)

     xMore = 0      * will U be likely 2 sample more steps from the MCMC?
     ndata = 1      * number of data sets (for simulation purpose)

** Specifications for the prior on rates ****************************
    PriorR = 3      * 0:lognormal; 1:"stationarized"-lognormal;
                    * 2:truncated normal; 3:Ornstein-Uhlenbeck process
                    * 4:Gamma; 5:Exponential.

  beta_oup = .001   * drift parameter (OUP: PriorR=3)
sigma2_oup = 10     * measure of variance for the rates (PriorR=0,1,2,4)
                    * diffusion parameter (OUP: PriorR=3)

** Specifications for the prior on times ****************************
    PriorT =  1     * prior 4 node times: 1:BDP; 2:Uniform; 3:Beta
     birth = 15.    * upper limit on birth rate (U prior)
     death =  5.    * upper limit on death rate (U prior)
    sample =   .01  * upper limit on sampling fraction (U prior)
```

Figure 2. Control file to estimate the probability of the data of the tree(s) defined in the file HIV.tre; all the parameters.

(2) test phylogenetic trees as estimated from nucleotide sequences

All the options concerning the prior for times and rates are ignored. Only branch lengths are integrated over along the Markov chain, and possibly the parameters of the nucleotide substitution model.

a – tests of phylogenetic hypotheses

Whichever test is aimed at (test of a tree against the maximum likelihood tree, or construction of a confidence set of trees), the basic step is to compute the probability of the data for different trees. These tests are not preformed directly by PHYBAYES, in the sense that carrying them out demands some manipulation of the quantities entering the Bayes factor, i.e. the probability of the data. The estimation of the probability of the data $p_{Ti}(X)$, for each tree topology T_i defined in the tree file, can be done using the control file as in Figure 2. The distribution of $p_{Ti}(X)$ is in the `logPost.out` file (second column), and the value of the log-probability of the data is approximated by taking the average of this distribution. The Bayes factor of T_i vs. T_j is then computed as the difference of the corresponding two averages. As the probability of the data may be a small number such as exp(–5123), a mathematical program may have to be used to perform simple vs. composite tests as required to construct confidence sets of trees.

b – posterior probability of trees

When the objective is to estimate the posterior probability of the trees on the tree space, $p(X \mid T_i)$, the control file of Figure 3 can be used. This setting can also be used to estimate the probability of the data $p(X)$ for each tree, although the number of the trees sampled is proportional to their posterior probability (by definition), which make inference for trees with a small posterior probability unreliable (as they are rarely sampled along the MCMC), whence the first procedure (Figure 2).

```
** Files ******************************************************
   seqfile = HIV.nuc * sequence data file name
   outfile = HIV.out * main result file name
  treefile = HIV.tre * tree file for initial rooted tree topology
      seed = 1234567 * random number seed

** Model ******************************************************
     model = 4    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
     clock = 0    * 1:molecular clock; 0:no molecular clock
     kappa = 5    * kappa in K80, F84, or HKY85
     alpha = 0.5  * alpha for gamma rates at sites (0 if homogeneous rates)
       mut = .514 * initial mutation rate (for all the lineages)
     ncatG = 8    * No. categories in discrete gamma

** Parameters to integrate out by MCMC ****************************
 EstSimult = 3      * estimate all the parameters:
                    * 0: cyclically; 1: simultaneously;
                    * 2: random 1 @ a time; 3:==0+sample all parameters
   EstTopo = 1      * estimate topology: - NNI (EstTimes=EstRates=0)
  EstTimes = 0      * estimate divergence times
  EstRates = 0      * estimate rates of evolution
****when these last 2 are set to 0, only branch lengths are estimated**
  EstBases = 0      * estimate base composition
  EstAlpha = 0      * estimate alpha
  EstKappa = 0      * estimate kappa

** Running the MCMC & starting values for the tunings ***************
    burnin = 10000  * length of the burn-in period
 NbOfSteps = 10000  * number of sampled steps along the MCMC
  sam_intv = 100    * interval between two sampled steps

     tuneT = .01    * tuning parameter for time (ChangeTree)
     tuneK = .05    * tuning parameter for kappa in K80, F84, or HKY85
     tuneA = .025   * tuning parameter for alpha (gamma rates at sites)
     tuneR = .01    * tuning parameter for the rate of evolution (or branches)

     xMore = 0      * will U be likely 2 sample more steps from the MCMC?
     ndata = 1      * number of data sets (for simulation purpose)

** Specifications for the prior on rates ***************************
    PriorR = 3      * 0:lognormal; 1:"stationarized"-lognormal;
                    * 2:truncated normal; 3:Ornstein-Uhlenbeck process
                    * 4:Gamma; 5:Exponential.

  beta_oup = .001   * drift parameter (OUP: PriorR=3)
sigma2_oup = 10     * measure of variance for the rates (PriorR=0,1,2,4)
                    * diffusion parameter (OUP: PriorR=3)

** Specifications for the prior on times ***************************
    PriorT =  1     * prior 4 node times: 1:BDP; 2:Uniform; 3:Beta
     birth = 15.    * upper limit on birth rate (U prior)
     death =  5.    * upper limit on death rate (U prior)
    sample =   .01  * upper limit on sampling fraction (U prior)
```

Figure 3. Control file to estimate the posterior probability of the tree; all the parameters but the tree are fixed.

(4) other options of the control file

`EstSimult`: defines how the states are sampled with respect to the updating scheme of the parameters.

`burnin`: defines the period of time during which the MCMC is run without keeping samples for inference.

`NbOfSteps`: this is how many states are to be sampled along the MCMC and kept for inference.

`sam_intv`: allows for thinning of the chain (the outputs): reduces autocorrelation of the sampled states, and thereby improves the estimates.

`tuneT/K/A/R`: tuning parameters of the parameters updated along the MCMC (variance of the proposal distributions, which are normal centred on the current state). The smaller, the higher the acceptance rate along the chain, but the slower the mixing.

`xmore`: set to `0` by default, which means that when the `NbOfSteps` are collected, the program stops and the summary statistics are computed. If set to `1`, the MCMC will be stopped and the user will be prompted how many extra steps to collect by running the chain further. Caution: running the program in the background (unix and linux users) with this option will create troubles.

`ndata`: used to analyse more than one data set. The extra data set(s) must be included in the data file.

How to run the program?

The default control file is `mcmctree.ctl`; if all the options set in this file are correct, the program can be run by typing `PhyBayes` at the command prompt (or double clicking on the icon). Otherwise, use `PhyBayes myControlFile.ctl` at the command prompt.

As the program uses some temporary files (`rst` and `rub`) and all the outfile names are independent of the control file, it is not a good idea to run the same executable (in the same directory) more than once at a time. Some operating systems prevent this (Win9x / Win2000), others do not (unix). Several copies of the program can be run simultaneously from a single location (typically a /bin directory where you have to set the path variable accordingly – please, refer to your system administrator), as long as each control file, tree file and data file, for each run, are in different working directories.

The outputs and how to use them

When all the options are checked, the output files are: `branch.out`, `d_alpha.out`, `d_kappa.out`, `d_pi.out`, `HIV.out`, `logPost.out`, `rates.out`, `ResTrees.out`, `ResTrees.trees`, `sampTree.tre`, `Time.trees`, `times.out`. Depending on the analysis performed, some of the files may not be created or appended if the file already exists in the folder. The files contain headers and are formatted with tabulations, which make their opening and reading with *Excel*™ or *S–plus*® / *R*® for instance easier. The first column generally contains the sampling stets, the likelihood value calculated at the state sampled, the tree ID, the tree (with no branch lengths) and the parameters.

`branch.out`:	branch lengths;
`d_alpha.out`:	alpha (parameter of the among-site rate variation) (†);
`d_kappa.out`:	kappa (transition / transversion rate ratio) (†);
`d_pi.out`:	bases frequencies (for `model` > 0) (†);
`HIV.out`:	general output file (contains basic statistics);
`logPost.out`:	contains the values used to compute $p(X)$;
`rates.out`:	rates of evolution for each branch (†);
`ResTrees.out`:	summary statistics of the MCMC;
`ResTrees.trees`:	summary trees sampled, with averaged branch lengths;
`sampTree.tre`:	actual trees sampled, with sampled branch lengths;
`Time.trees`:	summary trees sampled, branches scaled to clock;
`times.out`:	divergence times (†).

Note – †: created only when the parameter is integrated out along the MCMC.

Parameters must be checked for convergence. Generally, the likelihood converges quickly, as well as the parameters of the substitution model. However, when estimating divergence times, it is imperative to check convergence of the rates and the times. Failure to do so may result in spurious results. It is also desirable to run at least two to four chains starting form different points in the parameter space.

PHYBAYES in action: the HIV test data set

To demonstrate how to analyse the results, we ran PHYBAYES on a small HIV data set consisting of six sequences of 2,000 nucleotides long (Goldman et al. 2000; Aris-Brosou 2001). This sample data set can be found with the distribution of the program. Analyses are performed under the HKY85+Γ nucleotide substitution model (Hasegawa et al. 1985; Yang 1994), for which we are interested in estimating the posterior distribution of the parameters under three of the 105 possible six species trees (Figure 5a). The control file used is given in Figure 4.

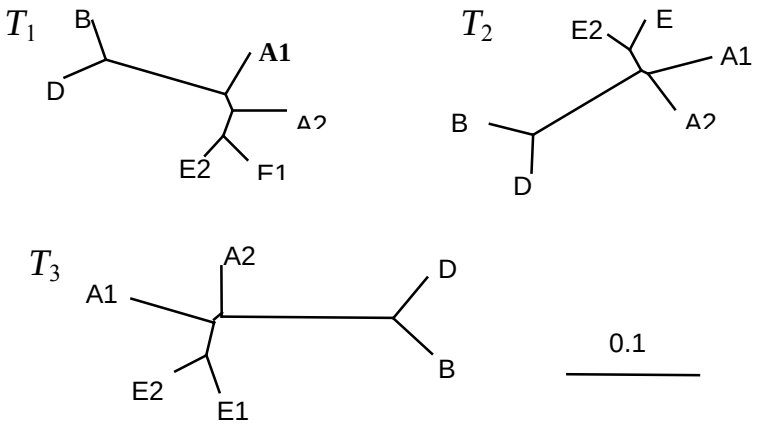
Multiple chains were run from different starting points. Linear regressions were performed on times series of each variable, testing the significance of the slope: the *P*-value should be large, indicating a slope not significantly different from zero, and the autocorrelation functions should not detect any structure in the samples.

With the options chosen (see Figure 4), each run can be examined by plotting the integrated log-likelihood (Figure 5b). The median *Med* (ℓ_i) is given under panel (5b) for tree T_i, as labelled in panel 5a. Similar distributions are obtained for the shape parameter of the gamma distribution modelling among-site rate variation (Figure 5c), the transition-transversion rate ratio (Figure 5d) or the base frequencies (not shown).

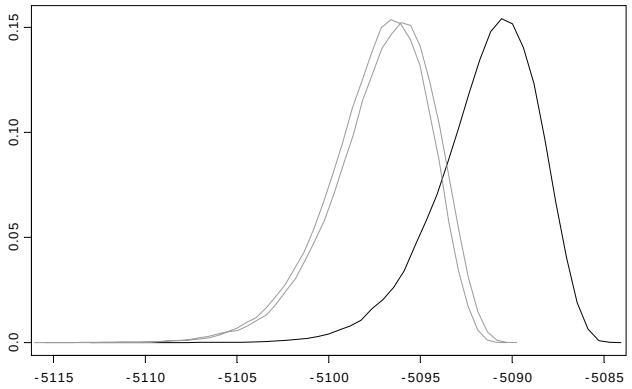
```
** Files ********************************************************
   seqfile = HIV.nuc * sequence data file name
   outfile = HIV.out * main result file name
  treefile = HIV.tre * tree file for initial rooted tree topology
      seed = 1234567 * random number seed

** Model ********************************************************
     model = 4    * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
     clock = 0    * 1:molecular clock; 0:no molecular clock
     kappa = 5    * kappa in K80, F84, or HKY85
     alpha = 0.5  * alpha for gamma rates at sites (0 if homogeneous rates)
       mut = .4   * initial mutation rate (for all the lineages)
     ncatG = 8    * No. categories in discrete gamma

** Parameters to integrate out by MCMC ****************************
 EstSimult = 3      * estimate all the parameters:
                    * 0: cyclically; 1: simultaneously;
                    * 2: random 1 @ a time; 3:==0+sample all parameters
   EstTopo = 0      * estimate topology: - NNI (EstTimes=EstRates=0)
  EstTimes = 0      * estimate divergence times
  EstRates = 0      * estimate rates of evolution
****when these last 2 are set to 0, only branch lengths are estimated**
  EstBases = 1      * estimate base composition
  EstAlpha = 1      * estimate alpha
  EstKappa = 1      * estimate kappa

** Running the MCMC & starting values for the tunings ***************
    burnin = 10000  * length of the burn-in period
 NbOfSteps = 100000 * number of sampled steps along the MCMC
  sam_intv = 100    * interval between two sampled steps

     tuneT = .01    * tuning parameter for time (ChangeTree)
     tuneK = .05    * tuning parameter for kappa in K80, F84, or HKY85
     tuneA = .025   * tuning parameter for alpha (gamma rates at sites)
     tuneR = .01    * tuning parameter for the rate of evolution (or branches)

     xMore = 0      * will U be likely 2 sample more steps from the MCMC?
     ndata = 1      * number of data sets (for simulation purpose)

** Specifications for the prior on rates ****************************
    PriorR = 3      * 0:lognormal; 1:"stationarized"-lognormal;
                    * 2:truncated normal; 3:Ornstein-Uhlenbeck process
                    * 4:Gamma; 5:Exponential.

  beta_oup = .001   * drift parameter (OUP: PriorR=3)
sigma2_oup = 10     * measure of variance for the rates (PriorR=0,1,2,4)
                    * diffusion parameter (OUP: PriorR=3)

** Specifications for the prior on times ****************************
    PriorT =  1     * prior 4 node times: 1:BDP; 2:Uniform; 3:Beta
     birth = 15.    * upper limit on birth rate (U prior)
     death =  5.    * upper limit on death rate (U prior)
    sample =   .01  * upper limit on sampling fraction (U prior)
```

Figure 4. Control file used to analyse the HIV data set and obtain the probability of the data for each tree defined in the tree file as well as the posterior distributions for the parameters of the substitution model (HKY85+Γ here).

(a)

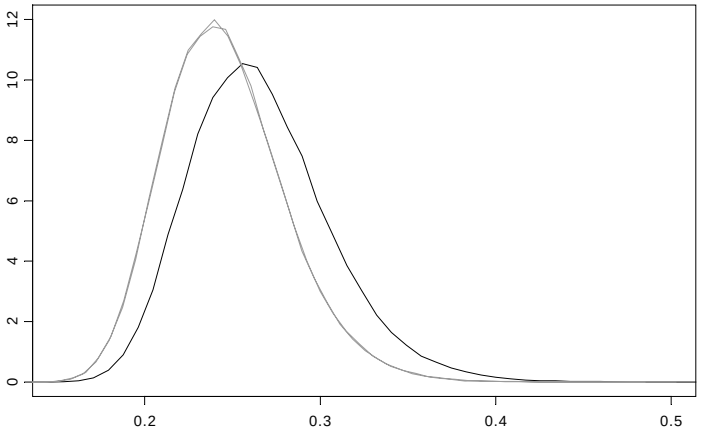


(b)



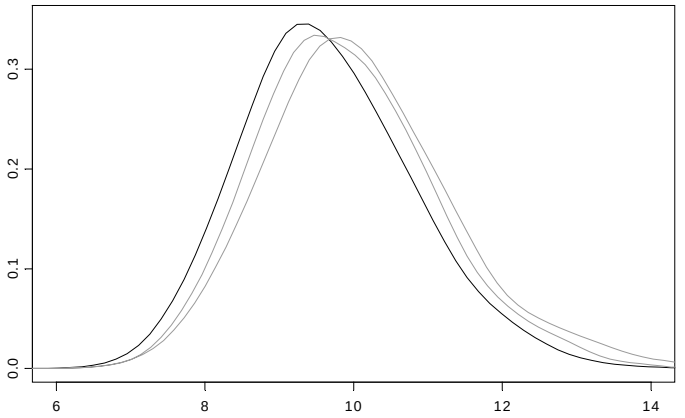
Med (ℓ_1) = –5091.06; *Med* (ℓ_2) = –5097.08; *Med* (ℓ_3) = –5096.7

(c)



Med (α_1) = 0.26; *Med* (α_2) = 0.24; *Med* (α_3) = 0.24

(d)



Med (κ_1) = 9.57; *Med* (κ_2) = 9.80; *Med* (κ_3) = 9.98

Figure 5. Outputs from PhyBayes for a HIV data set. See text for details. The *S–plus*® program was used to produce the plots.

References

ARIS-BROSOU, S. 2001. Importance of the null hypothesis in phylogenetics: test of hypotheses about the correct tree or significance tests of trees? Mol. Biol. Evol. (*submitted*)

ARIS-BROSOU, S. AND Z. YANG. 2001a. Phybayes: a program for phylogenetic analyses in a Bayesian framework. Bioinformatics. (*to be submitted*)

ARIS-BROSOU, S. AND Z. YANG. 2001b. The effects of models of rate evolution on estimation of divergence dates with a special reference to the metazoan 18S rRNA phylogeny. Syst. Biol. (*submitted*)

BROWN, W. M., E. M. PRAGER, A. WANG, AND A. C. WILSON. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18**:225–239.

FELSENSTEIN, J. 1995. PHYLIP (phylogeny inference package). University of Washington, Seattle. http://www.evolution.genetics.washington.edu/phylip.html.

GOLDMAN, N., J. P. ANDERSON, AND A. G. RODRIGO. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. **49**:652–670.

HASEGAWA, M., H. KISHINO, AND T. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555–556.