Dating Microbial Evolution with MCMCtree

Mario dos Reis
School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK
Email: m.dosreisbarros@qmul.ac.uk

Molecular sequences accumulate substitutions at an approximately constant rate in time ***(1)***. Thus, if we estimate an evolutionary tree using molecular data, it may be possible to calibrate the divergence events in the tree to geological time if we have temporal information about the nodes in the tree, for example, by using the fossil record, mutation rate estimates, or non-contemporaneous taxa with known sampling times ***(2–5)***. However, these sources of temporal information will have uncertainties associated with them, and ideally, these uncertainties should be taken into account in the analysis. The Bayesian method has gained popularity for molecular-clock dating of evolutionary trees because it allows incorporation of uncertainties from various sources as prior information in the analysis ***(6)***. The Bayesian method has now been used to date phylogenies from viruses ***(7)*** and prokaryotes ***(8, 9)*** to plants ***(10)*** and animals ***(11, 12)***.

In this chapter I guide the user on how to use the computer program MCMCtree ***(13)*** to date microbial divergences using the Bayesian method. MCMCtree's strength lies in the implementation of an approximation to the likelihood that can speed up the Bayesian computation over exact likelihood calculation ***(14)***. Because Bayesian molecular-clock dating relies on expensive MCMC sampling, the time savings obtained by using the approximation can be quite dramatic, up to 1,000x depending on the dataset ***(15)***. Thus, MCMCtree can be used to date evolutionary trees built using large molecular alignments or containing hundreds of taxa ***(8, 16–18)***. Other features of MCMCtree include the implementation of relaxed-clock models that allow the molecular clock to vary among the branches of a phylogeny ***(19)***; the use of flexible fossil calibration densities to specify the prior on node ages ***(20–22)***; models for dating phylogenies with taxa serially-sampled in time ***(23)***; and models for accommodating rate variation among alignment partitions ***(24)***. MCMCtree requires the topology of the phylogeny to be fixed. Uncertainty in topological placements of taxa can be addressed by dating separate phylogenies encompassing the relevant topological rearrangements ***(16, 25)***. Other Bayesian molecular-clock dating software that can co-estimate divergence times and topology include MrBayes ***(26)***, PhyloBayes ***(27)*** and Beast ***(28)***. However, the likelihood approximation can not be used when co-estimating times and topology, and thus, time and topology co-estimation is computationally very expensive.

Here I illustrate how to use MCMCtree to date two phylogenies. In the first exercise, I show how to date a deep phylogeny of prokaryotes and eukaryotes by using fossil calibrations ***(8)***. In the second exercise, I show how to to date a large virus phylogeny in which the sampling times of the viruses is known ***(23)***. Both exercises use the likelihood approximation to speed up the analyses. Important issues in MCMC sampling such as convergence to the posterior distribution, diagnosing the MCMC output, and verification of the prior are discussed, as well as strategies for dating microbial phylogenies when neither fossil calibrations nor sampling times of taxa are available. In this chapter I assume the reader is familiar with phylogenetic methods to build evolutionary trees from molecular sequence alignments. Excellent introductions to phylogenetic tree reconstruction methods can be found in Felsenstein ***(29)*** and Yang ***(30)*** (see also ***(31)***). Several excellent reviews on the theoretical aspects of Bayesian phylogenetics and Bayesian clock-dating are available elsewhere ***(32, 33)***.

Software and data

To carry out the exercises in this chapter, you will need to install the MCMCTree, CODEML and BASEML programs from PAML's software package for phylogenetic analysis ***(13)***. PAML is freely available from bit.ly/ziheng-paml. Please follow the instructions in PAML's website to install the programs. In particular, you should add PAML's executables to your system's path. See PAML's website for instructions on how to do this. The data files necessary for this tutorial are available from github.com/dosreislab/microdiv. Please download the data files and save them to a directory called `microdiv` in your computer. You are assumed to be familiar with the command line (also known as terminal, shell or prompt) from your operating system, as MCMCtree is a command line program. It is also useful (although not strictly necessary) if you install FigTree (useful for timetree visualization, tree.bio.ed.ac.uk/software/figtree/), and the R statistical environment (useful to diagnose the MCMC output, www.r-project.org). All exercises were tested with PAML version 4.9j.

Bayesian molecular-clock dating with approximate likelihood

The Bayesian method uses probability distributions to characterise the uncertainties in all unknowns in an inference problem ***(30)***. The Bayes theorem is then used to express the uncertainty in the estimated parameter values, known as the posterior distribution, as a function of our prior knowledge about the parameters and the likelihood of the data given the parameters. In molecular-clock dating, the posterior distribution of divergence times and rates, t and r, (our parameters), given a molecular sequence alignment D (our data) is

$$f(t, r|D) = Cf(t)f(r|t)f(D|r, t), \quad (1)$$

where $f(t)$ is the prior on the divergence times, $f(r|t)$ is the prior on the molecular evolutionary rates (i.e. the molecular clock model), and $f(D|r,t)$ is the likelihood of the molecular sequence alignment as a function of the times and rates. The likelihood function is constructed from a model describing how nucleotides or amino acids change within sequences along evolutionary time (i.e. the substitution model, ***(30)***).

Constant $C = 1/\int f(t)f(r|t)f(D|r,t)drdt$, known as the marginal likelihood of the data, is necessary to guarantee the posterior is a proper statistical distribution that integrates to one. Unfortunately, in most molecular-clock dating problems, C cannot be calculated analytically. Thus, the MCMC algorithm, which does not require calculation of C, is used to obtain a random sample of times and rates from the posterior distribution ***(30, 33)***. This random sample can then be summarised, for example, by computing histograms and sample means, giving approximations to the posterior distribution and the posterior means.

In a typical molecular-clock dating MCMC, the software generates a random set of initial values for the times and rates, (t,r), known as the initial state. The software then generates a random perturbation of one the values, say for t, to generate the proposed state, $(t*,r)$. The proposal ratio

$$\alpha = \frac{f(t*,r|D)}{f(t,r|D)} = \frac{f(t*)f(r|t*)f(D|t*,r)}{f(t)f(r|t)f(D|t,r)}, \qquad (2)$$

is then calculated (note constant C cancels out and thus is not needed). If the proposal, $(t*,r)$, increases the posterior (i.e. $\alpha > 1$), it is accepted. Otherwise it is accepted with probability α (or rejected with probability $1-\alpha$). If it is accepted, we set $(t,r) = (t*,r)$, otherwise we set $(t,r) = (t,r)$. The process is repeated to accept or reject perturbations for all the times and rates within a cycle. Then, N proposal cycles are carried out until a sufficiently large sample has been obtained.

MCMC is computationally expensive because each proposal requires calculation of the prior times the likelihood (Eq. 2). Furthermore, the time it takes to compute the likelihood is proportional to the number of site patterns in the alignment (a site pattern is a unique configuration of nucleotides or amino acids in an alignment column). Thus, long alignments lead to computationally expensive MCMCs: Typical molecular-clock dating MCMCs can take from several days to months to compute. Fortunately, the likelihood can be approximated by its Taylor expansion ***(14, 34)***, and the approximation is virtually indistinguishable from exact calculation when using long alignments ***(14)***. In molecular-clock dating, the likelihood depends on the product of the times and rates, the branch length, $b = tr$, and not on the times and rates separately, and thus we can write $f(D|t,r) = f(D|b)$. The approximation is

found by Taylor expansion of $log\ f(D|b)$ around the maximum-likelihood estimates (MLEs) of the branch lengths

$$log\ f(D|b) \approx log\ f(D|\hat{b}) + (b-\hat{b})'g + \frac{1}{2}(b-\hat{b})'H(b-\hat{b}), \quad (3)$$

where $\hat{b}$ are the MLEs, and g and H are the gradient vector and Hessian matrix of first and second derivatives of the log-likelihood evaluated at the MLEs. MCMCtree improves the accuracy of the Taylor approximation by using transformations of the branch lengths ***(14)***.

Thus, to use the approximation, one first calculates the MLEs of the branch lengths and the gradient and Hessian on a fixed tree topology for a given molecular alignment. The substitution model is chosen at this step (because different substitution models generate different values of $\hat{b}$, g and H). Then, MCMC sampling proceeds using Eq. (3) to approximate the likelihood during calculation of the proposal ratio (Eq. 2). The size of g and H depend on the number of taxa and not on the alignment length. Thus, approximate likelihood computation takes the same amount of time whatever the length of the original alignment. A detailed mathematical description of MCMC sampling in phylogenetics can be found in ***(30)*** and the full details of the approximation in ***(14)***.

Dating an microbial phylogeny using fossil calibrations

The data are an alignment of a DNA replication factor from 55 prokaryotic and eukaryotic taxa, with 300 amino acid sites. This is a subset of the data analysed by Betts et al. ***(8)*** to date the origin of life and of the eukaryotic cell. File `betts/data/dm_17.phy` contains the alignment in phylip format and file `betts/data/dm_17.tree` contains the phylogeny with fossil calibrations. You can use a text editor (e.g. Notepad or TextEdit) to look at the contents of both files.

Table 1 shows the fossil calibrations used. MCMCtree implements a broad set of calibration densities to represent uncertainty in node ages. The two most commonly used calibrations are the "B" and "L" calibrations (see MCMCtree's documentation for the full list of calibration densities and how to specify them). B calibrations use two bounds, t_{min} and t_{max}, to specify a uniform distribution between the minimum and maximum age of the node. The bounds are soft, meaning the minimum and maximum bounds are allowed to be violated with probabilities p_{min} and p_{max} respectively. L calibrations are specified by using a soft minimum bound, t_{min}, and a truncated Cauchy distribution with a heavy-tail decaying back in time ***(21)***. B calibrations are suitable when information is available to specify both the minimum and maximum age of a node. Unfortunately, the fossil record is rather incomplete and absence of evidence is not evidence of absence, and thus specifying maximum bounds may be very difficult ***(4)***. In such cases, the L calibrations, which only need the minimum bound, may be more appropriate.

For example, for the root of the phylogeny (the last universal common ancestor, node 56, Table 1) Betts et al ***(8)*** suggest a minimum bound of $t_{min} = 3.347$ billion years ago (Ba) based on stromatolite fossils of the Strelley Pool formation of Australia. For the maximum bound they suggest $t_{max} = 4.52$ Ba, based on the Moon-forming impact (Table 1). The Strelley formation has definitive evidence of cellular life and thus we assign a zero probability, $p_{min} = 0$, of the minimum bound being violated (in MCMCtree, $p_{min} = 0$, is specified as $p_{min} = 10^{-300}$ which is a tiny number), whereas the maximum is soft with $p_{max} = 0.025$. Five further nodes have L calibrations (Table 1). For the ancestor of Rhodophyta (node 76, Table 1), a minimum of 1.033 Ba is used based on fossil *Bangiomorpha*. This minimum is slightly older than the strata containing the fossil and thus a soft minimum with $p_{min} = 0.025$ is used. The other four L calibrations use hard minima as they are based on uncontroversial fossil members of the corresponding groups. You can use the mcmc3r package in R (available from https://github.com/dosreislab/mcmc3r) to plot the L distributions. If the L calibration appears unreasonable, parameters c and p can be adjusted to control the mode and the length of the tail of the calibration ***(21)***. You can visualise the fossil calibrations by opening the tree file in a text editor or by plotting the tree with FigTree.

Table 1. Fossil calibration densities used in the analysis of Betts et al. ***(8)*** phylogeny. The time unit is 100 Ma.

Node	**Calibration**[a]
56	B (33.4700, 45.2000, 1e-300, 0.025)
76	L (10.3300, 0.1, 1, 0.025)
80	L (4.2040, 0.1, 1, 1e-300)
81	L (1.2500, 0.1, 1, 1e-300)
94	L (5.5025, 0.1, 1, 1e-300)
97	L (3.9210, 0.1, 1, 1e-300)

a. B and L calibrations are specified in the tree file. B calibrations use the format B($t_{min}, t_{max}, p_{min}, p_{max}$). L calibrations use the format L(t_{min}, p, c, p_{min}), where p and c control the position of the mode and the length of the tail of a truncated Cauchy distribution ***(21)***.

The steps we will use to estimate the divergence times on the Betts et al. phylogeny using the approximate likelihood method are as follows:

1. Estimate the MLEs of branch lengths, gradient vector and Hessian matrix using CODEML.
2. Conduct MCMC sampling of the posterior of times and rates using MCMCtree and obtain summaries of the posterior distribution.
3. Conduct diagnostic tests of MCMC convergence to the posterior distribution using R.
4. Conduct MCMC sampling of the prior of times and rates using MCMCtree.

Estimation of the branch lengths MLEs, gradient and Hessian

Go to the betts/gH directory and open file mcmctree-outBV.ctl using a suitable text editor. MCMCtree will use the information in this file to prepare input files for CODEML. We will then use CODEML to estimate the MLEs of branch lengths, gradient and Hessian using the LG+G amino acid substitution model ***(35, 36)***. The contents of the mcmctree-outBV.ctl file are shown in Figure 1. The first two lines give the name of the alignment and tree files respectively. Then ndata = 1 indicates the alignment is contained in one partition. This is important because if your alignment is divided into two one more partitions, then you need to calculate one set of branch length MLEs, gradient and Hessian for each partition. The next line indicates the type of alignment (amino acids) and then usedate = 3 indicates we want to estimate the necessary MLEs for approximate likelihood calculation. The MLEs will be stored in a file called out.BV. The remaining lines indicate we will use an empirical amino acid substitution matrix (the LG model stored in the lg.dat file) with a discrete gamma model of rate variation among sites (with 5 categories).

```
      seqfile = ../data/dm_17.phy
     treefile = ../data/dm_17.tree

        ndata = 1
      seqtype = 2    * 0: nucleotides; 1:codons; 2:AAs
      usedata = 3    * 0: no data (prior); 1:exact likelihood;
                     * 2: approximate likelihood; 3:out.BV (in.BV)
        model = 2    * 0: poisson; 1: proportional; 2:Empirical;
                     * 3: Empirical+F; 6: FromCodon; 8: REVaa_0; 9: REVaa(nr=189)
        alpha = 0.5  * alpha for gamma rates at sites
        ncatG = 5    * No. categories in discrete gamma
    aaRatefile = lg.dat * Amino acid substitution model rate matrix

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

Figure 1. The mcmctree-outBV.ctl file for calculation of branch length MLEs, gradient and Hessian.

Open a terminal and, within the betts/gH directory, type

```
$ mcmctree mcmctree-outBV.ctl
```

Do not type in the $ sign as this represents the command prompt. If the command above does not work, it means you have not yet installed MCMCtree or you have not added PAML's programs to your system's path variable (see above). If the command is successful, MCMCtree will generate a set of special files (named with the tmp suffix), and then will call the CODEML program who will then calculate $\hat{b}$, g and H by numerical optimization of the likelihood function. CODEML will write out the values of $\hat{b}$, g and $H$to a file called rst2, and MCMCtre will rename this file as out.BV. CODEML will use the "simultaneous" algorithm of branch-length optimization, which is rather slow: In this case, it will take CODEML about 11 min (on an 2.8 GHz Intel Core i7) to do the calculations. If you want, you can kill CODEML (in Mac and Unix systems you may do this by pressing the CTL + C keys), and then open the tmp0001.ctl file (this is the control file MCMCtree created for CODEML)

and change the line `method = 0` to `method = 1`. This changes to the "one-by-one" branch-length optimization algorithm which is much faster ***(37)***. Then in the terminal type

```
$ codeml tmp0001.ctl
```

Which calls CODEML using the modified `tmp0001.ctl` file. It should take about 40 sec for CODEML to obtain the estimates (on the 2.8 GHz Intel Core i7). You then need to rename file `rst2` to `out.BV`. When analysing phylogenies with hundreds of species or very large alignments, it may be useful to use MCMCtree to prepare the `tmp` files, and then use these files to run CODEML (or BASEML in the case of nucleotide alignments, see below) in a high-performance computer cluster.

Figure 2 shows the part of the contents of the out.BV file (you can use a text editor to open the file). The first line indicates the number of species, 55. Then we see the unrooted phylogeny, the vector of branch lengths, the gradient vector (the line of mostly zeroes), and the Hessian matrix. There are 2 × 55 – 3 = 107 branches in the unrooted tree. Thus, the gradient vector has 107 elements, and the Hessian matrix is of size 107 x 107.

```
55

(Synechocystis_sp._PCC6803: 0.777504, (Clostridium_acetobutylicum: 0.738700, ((Treponema_pallidum:

  0.777504  0.210891  0.738700  0.097101  0.115624  0.965066  0.136945  0.679848  0.875403  0.058785

  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000


Hessian

     -75.09     -51.11     -9.421     -26.96     -24.77     -16.27     -12.48     -8.463      -14.3
     -51.11     -147.1     -12.11     -43.23      7.892     -12.16      3.393     -3.355     -9.761
     -9.421     -12.11     -110.3     -54.15     -17.86     -9.572     -28.76     -9.087     -13.73
     -26.96     -43.23     -54.15       -294     -27.06     -21.97     -11.18     -16.37     -22.33
     -24.77      7.892     -17.86     -27.06     -454.9     -30.15     -87.68     -38.58      -30.3
     -16.27     -12.16     -9.572     -21.97     -30.15     -77.94     -19.56     -5.906     -19.81
```

Figure. 2. File `out.BV` containing the estimates of branch lengths, gradient and Hessian for the Betts et al. phylogeny obtained using the LG + G substitution model.

MCMC sampling from the posterior distribution

Go into directory `betts/mcmc`. Copy the `out.BV` file into this directory and rename it to `in.BV`. In this directory you will see file `mcmctree.ctl`, which contains the instructions for the MCMC sampling. Figure 3 shows the contents of the file. Note the line with `usedata = 2`, this tells MCMCtree that we will conduct MCMC sampling of the posterior distribution using approximate likelihood. Line `clock = 2` indicates the independent rates model will be used (other options are the strict clock and the autocorrelated rates model ***(19)***). MCMCtree uses a birth-death process to specify the prior on node ages without fossil calibrations ***(20)***. Here we use `BDparas = 1 1 0`, which specifies a uniform prior. Lines `rgene_gamma` and `sigma2_gamma` specify the gamma prior on the mean evolutionary rate and on the rate dispersion parameter of the relaxed-clock models ***(19)***. It is important to specify a sensible rate prior. If the mean on the prior rate is too high or too

low, it will be in conflict with the fossil calibrations and may affect the posterior distribution of times. Here we use `rgene_gamma = 2 40 1`, which specifies a gamma distribution with mean = 2/40 = 0.05 amino acid substitutions per 100 million years, which is within the order of magnitude expected in ancient phylogenies ***(25)***. The last three lines control MCMC sampling. The initial state of the MCMC is usually far from the region of high probability mass in the posterior distribution. Thus, the MCMC will initially move from an area of very low probability to the high probability areas as sampling progresses. This initial movement of the MCMC is known as the burn-in, as samples obtained in this phase are usually discarded. Here we use `burnin = 1000`, which means we will discard the first 1,000 samples collected. Lines `sampfreq = 2` and `nsample = 20000` indicate we will obtain a sample of size 20,000 sampled every 2 iterations. Thus, the total number of cycles in our MCMC will be $N = 1000 + 2 \times 20000 = 41000$.

```
         seed = -1

      seqfile = ../data/dm_17.phy
     treefile = ../data/dm_17.tree
     mcmcfile = mcmc.txt
      outfile = out.txt

        ndata = 1
      seqtype = 2    * 0: nucleotides; 1:codons; 2:AAs
      usedata = 2    * 0: no data (prior); 1:exact likelihood;
                     * 2: approximate likelihood; 3:out.BV (in.BV)
        clock = 2    * 1: global clock; 2: independent rates; 3: correlated rates

       BDparas = 1 1 0   * birth, death, sampling
   kappa_gamma = 6 2     * gamma prior for kappa
   alpha_gamma = 1 1     * gamma prior for alpha

   rgene_gamma = 2 40 1   * gammaDir prior for rate for genes
  sigma2_gamma = 1 10 1   * gammaDir prior for sigma^2     (for clock=2 or 3)

         print = 1   * 0: no mcmc sample; 1: everything except branch rates 2: all
        burnin = 1000
      sampfreq = 2
       nsample = 20000
```

Figure 3. The `mcmctree.ctl` file for MCMC sampling of the posterior distribution of times and rates.

In the terminal type

```
$ mcmctree
```

This will initiate the MCMC sampling of the posterior. MCMCtree will print a lot of information to the screen as the MCMC progresses. It should take 1-2 min for MCMCtree to complete the sampling. Once MCMCtree is finished, you will see several new files have been created. File `out.txt` contains a summary of the analysis. You can open this file in a text editor and examine it. At the beginning of the file, you will see the alignment compressed into site patterns and other statistics on the sequences. You will then see three Newick trees. The first one simply has the species names and all internal nodes are labelled with numbers. These numbers correspond to the node ages in the output. The next two trees are calibrated to geological time and thus the branch lengths are given in

time units, but the last tree also has the 95% credibility intervals (CI) of node ages. At the bottom of the file you will see a table with all the node ages, named with a `t_` suffix followed by the node number. For each node you will see the posterior mean age, the 95% equal-tail CI, the 95% highest-posterior density (HPD) CI, and the HPD CI width. You will then see summary statistics for the posterior mean of the rate (`mu`), and the relaxed-clock parameter (`sigma2`), as well as for the log-likelihood. Table 2 shows example values of these parameters. File `FigTre.tre` contains the dated phylogeny, suitable for plotting with FigTree (Fig. 4). Finally, file `mcmc.txt` contains the raw MCMC sample: each column corresponds to a parameter and each row to a sampling cycle. Note by default, MCMCtree does not collect the samples of the rates for branches, only the mean rate is collected. If you are interested in the branch rates, set `print = 2` in the `mcmctree.ctl` file.

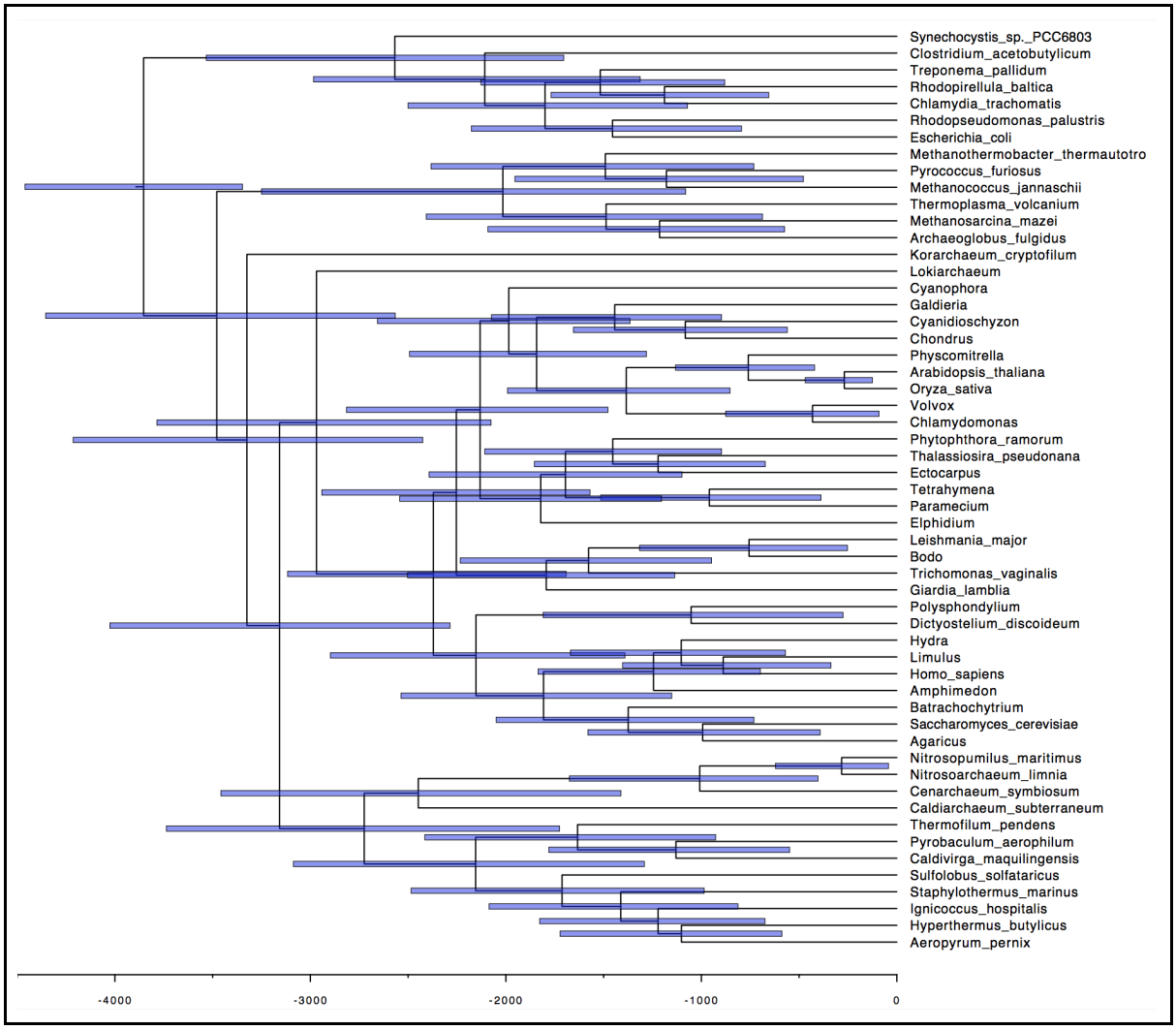

Figure 4. The dated phylogeny of Betts et al. as plotted by FigTree. The horizontal bars are the 95% HPD CI of node ages. Nodes are plotted at their posterior means. Time unit is 1 million years.

Table 2. Posterior summary of some parameters in Betts et al. phylogeny. Because the MCMC algorithm is stochastic, your values will be similar but different to those shown here. Time unit is 100 My.

Parameter	Posterior mean	95% Equal-tail CI	95% HPD CI	HPD width
t_56 (root age)	38.5	(33.7, 45.0)	(33.5, 44.6)	11.1
mu	0.0372	(0.0268, 0.0508)	(0.0263, 0.0499)	0.0236
sigma2	0.437	(0.300, 0.610)	(0.286, 0.594)	0.308
Log L	–49.14	(–63.39, –36.69)	(–63.03, –36.42)	26.61

MCMC diagnostics

An MCMC sample is guaranteed to converge to the posterior distribution as the size of sample approaches infinity. In practice, however, there is no guarantee that an MCMC chain has run long enough, and thus an actual MCMC sample may be too short and thus a poor approximation to the posterior. Because in practical problems we do not know what the posterior is supposed to look like, the way to guard against poor MCMC samples is to run the analysis several times and compare the results. If results from many runs are very similar, then we can have some confidence that the MCMCs have converged to the posterior distribution.

Rename files `mcmc.txt`, `out.txt` and `FigTree.tre` to `mcmc1.txt`, `out1.txt` and `FigTree1.tre`. Then in the terminal type

```
$ mcmctree
```

This will run the analysis again. By default, MCMCtree uses the current system's time as the random seed, and thus the random sampling will be different. Once the run has finished, rename the resulting files as `mcmc2.txt`, `out2.txt` and `FigTree2.txt`.

Many software packages exist to conduct MCMC diagnostics. In phylogenetics, the Tracer program (www.beast2.org/tracer-2/) is very popular and easy to use. MCMCtree's `mcmc.txt` files are compatible with Tracer. Here we will use R as it provides fine grain control over the diagnostics. Go into the `R/` directory and start R. If you use Rstudio, you can open file `R.Rproj` instead, and this will start R within Rstudio. The R commands to carry out the diagnostics are in file `betts.R`. You can open this file in a text editor or in a Rstudio. In R type

```
> mc1 <- read.table("../betts/mcmc/mcmc1.txt", head=TRUE)
> mc2 <- read.table("../betts/mcmc/mcmc2.txt", head=TRUE)
```

Do not type in the > character as this represents the R prompt. The MCMC samples are now loaded into R. We will now calculate the posterior means of times for each MCMC and plot them against each other:

```
> # posterior means:
> ti <- grep("t_", names(mc1))
> t.mean1 <- colMeans(mc1[,ti])
```

```
> t.mean2 <- colMeans(mc2[,ti])
> plot(t.mean1, t.mean2, main="(a) Posterior means")
> abline(0, 1)
```

As you can see (Fig. 5a), the posterior means are similar between the two MCMC runs, but the dots do not quite fall on the $x = y$ line. This suggests running the MCMCs for longer would be appropriate to improve convergence. We will now plot the histograms, traces and autocorrelation for the root age sample:

```
> # density histograms:
> ri <- which(names(mc1) == "t_n56")
> plot(density(mc1[,ri]), main = "(b) Density histograms of root age")
> lines(density(mc2[,ri]), lty=2)

> # trace plot for the root age:
> plot(mc1[,ri], ty='l', main = "(c) Trace plot of root age")

> # autocorrelation plot for root age:
> acf(mc1[,ri], main="(d) Autocorrelation of root age")
```

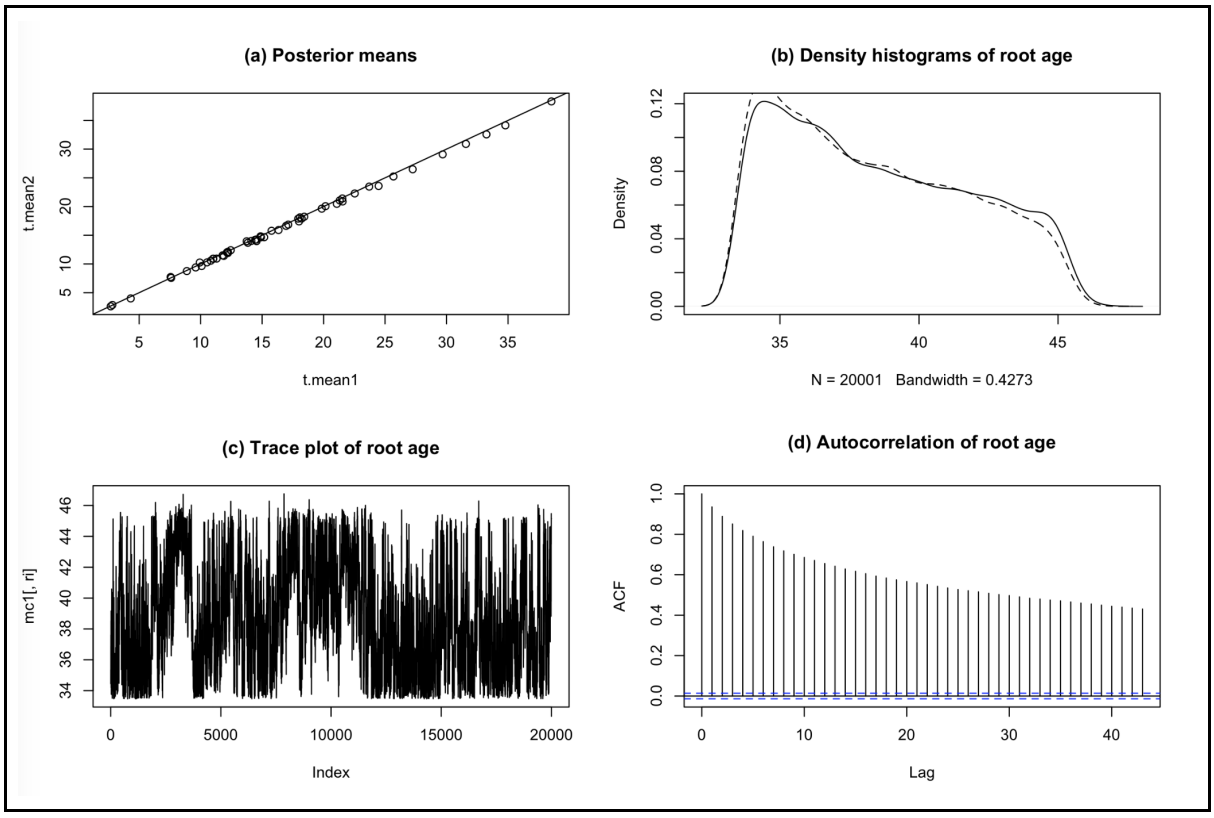

Figure 5. MCMC diagnostics for the Betts et al dataset.

The density histograms show noticeable differences (Fig. 5b) and the trace plot is not dense (Fig. 5c). This confirms convergence has not been fully achieved. MCMC samples are autocorrelated because an MCMC state is either equal to the previous state, or a

modification of the previous state. Inefficient chains have high autocorrelation and thus require longer runs to achieve a good approximation to the posterior. The high autocorrelation in this analysis can be appreciated in Fig. 5d. You can use the coda package in R to calculate the effective sample size of the MCMC, which is the equivalent sample size of an independent sample with the same sampling variance. In R

```
# calculate effective sample sizes
# note you need to have the coda package installed
coda::effectiveSize(mc1[,ri])
```

The effective size for the root age is 200. This is only 1% of the total sample size. To improve convergence and increase the effective size, we need to run the MCMC chains for much longer. Change line `nsample = 2` to `nsample = 100` in the mcmctree.ctl file. This will increase the total number of iterations in the MCMC to over two million. Then run MCMCtree twice to generate two independent MCMC samples. Note the new chains will require over an hour of computation time each. Once the two MCMC chains have completed, re-calculate the diagnostics and compare them to the previous results.

If you're keen on learning the practical details of MCMC sampling, you can do the MCMC tutorial in R prepared by Nascimento et al. ***(33)***.

Sampling from the prior

In a phylogeny, nodes are constrained to be older than their daughter nodes. These constraints must be applied during MCMC sampling when node ages are proposed, and this results in truncation of the fossil calibration densities used ***(21, 38)***. In other words, the calibration density applied to a node may be very different to the marginal prior used during MCMC sampling after truncation ***(39)***. Thus, it is advisable to run an MCMC with "no data" to sample from the prior and thus verify that the actual prior on node ages is sensible ***(39)***. Go into the `betts/prior` directory. File mcmctree-prior.ctl has one line changed, `usedata = 0`, which tells MCMCtree to carry out sampling from the prior. In the terminal

```
> mcmctree mcmctree-prior.ctl
```

Once MCMCtree has finished you will see files `mcmc.txt`, `out.txt` and `FigTree.tre` which in this case contain the prior sample and prior summaries. We will use R to plot the calibration densities and the marginal prior. Using the R session you started in the R/ directory, type

```
# load prior sample:
pmc <- read.table("../betts/prior/mcmc.txt", head=TRUE)

# plot prior root age:
plot(density(pmc[,ri]), lty=2, xlim=c(30, 50))
# add the calibration density on the root:
curve(mcmc3r::dB(x, 33.47, 45.2, 1e-300, .025), n=5e2, add=TRUE)
# add prior and calibrations for other nodes:
```

```
lines(density(pmc[,ni]), lty=2)
curve(mcmc3r::dL(x, 10.33, 0.1, 1, 1e-300), n=5e2, add=TRUE)
```

This will generate the plot seen in Figure 6. You can see the marginal prior on the root and Rhodophyta's age (dashed lines) are similar to the calibration densities (solid lines). However, the small differences observed between prior and calibration are due to slight truncation effects due to the long tail of the L calibrations. As an exercise, you may try plotting the calibration and prior densities for the other four calibrated nodes of Table 1.

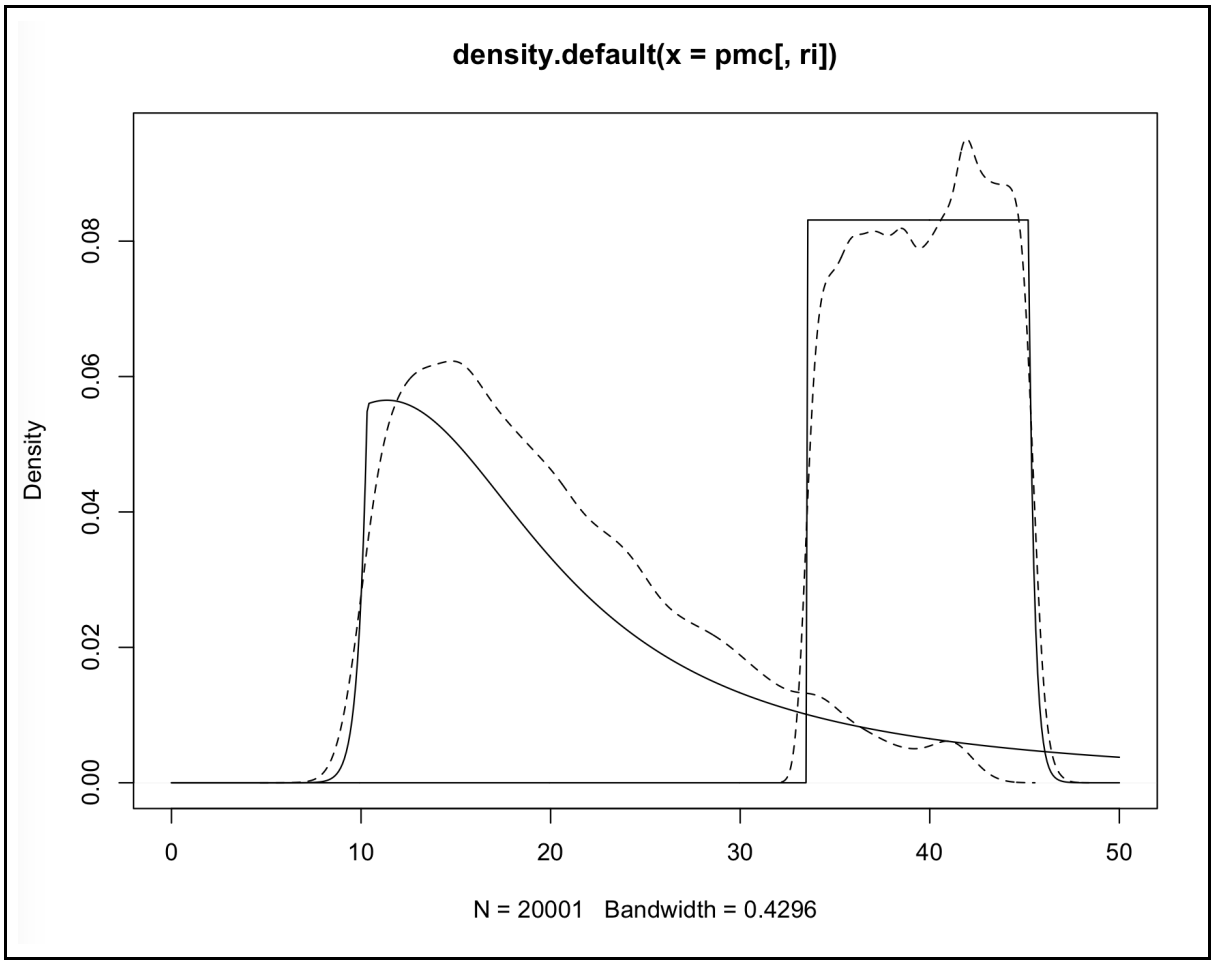

Figure 6. Marginal prior and fossil calibration density on the ages of the root and Rhodophyta in Betts et al. phylogeny.

Dating a phylogeny with serially-sampled taxa

Some microbes have their sampling dates recorded. For example, ancient DNA sequencing technologies allow the study of bacteria from ancient human remains ***(40, 41)***, while dated samples from environmental microbiology can be used to date and track disease spread ***(42)***. Influenza viruses from the 1918's to the present have had their genomes sequenced ***(43)***, while HIV genome sequences from mid-20th century isolates have been used to date the origin of the HIV pandemic ***(44)***. These serially sampled taxa with known sampling times have much information for molecular-clock dating ***(3, 45)***. When selecting serially-sampled taxa, it is important to have relatively old samples in the analysis. If the phylogeny to be

dated has “shallow” tips (i.e. the taxa have been recently sampled with respect to the age of the root), there may be relatively little information to date the phylogeny and time estimates may have large uncertainties ***(3)***.

Here we will date an influenza phylogeny ***(46)*** using the birth-death sequential sampling (BDSS) prior implemented in MCMCtree ***(23)***. The phylogeny and alignment of the H1 protein-coding gene for 289 influenza isolates are available in files `flu/data/flu.tree` and `flu/data/H1.tree`. Open the alignment file in a text editor. You will notice the isolates have the sampling year at the end of their name. For example, the oldest isolate is 1_Human_H1N1_USA_1918. MCMCtree will use the sampling year to calibrate the tree in the analysis.

Go into the flu/gH directory. In the terminal type

```
$ mcmctree mcmctree-gH.ctl
```

This will initiate calculation of the gradient and Hessian. MCMCtree will prepare the necessary `tmp` files and, because this is a DNA alignment, will call the BASEML program to carry out the calculations. The substitution model is HKY + G ***(36, 47)***.

Once the gradient and Hessian calculations are done, change to the `flu/mcmc` directory and copy `flu/gH/out.BV` into the directory and rename it to `in.BV`. Open the `mcmctree.ctl` file in a text editor. The file has a similar format to the one in the previous analysis but note the following lines:

```
TipDate = 1 100  * TipDate (1) & time unit
```

This tells MCMCtree that we will be activating tip dating with a time unit of 100 years.

```
RootAge = B(1, 5, .001, .001)  * used if no fossil for root
```

This is the calibration on the root age, to be from 100 to 500 years before the youngest isolate, which dates from 2009. This means the root age is calibrated between 1909 and 1509. MCMCtree always requires a calibration on the root, because both the BDS and BDSS processes are conditioned on the age of the root.

```
BDparas = 2 1 0 1.8  * lambda, mu, rho, psi for BDSS model
```

This specifies the parameters for the BDSS prior on node ages, which were chosen to produce a sensible prior density ***(23)***. You should make sure this prior is reasonable in your own analysis. The best way to verify this, is to run MCMCtree with no data and check that the prior node ages are sensible.

In the terminal type

```
$ mcmctree
```

MCMC sampling of the posterior distribution using approximate likelihood will take place. This analysis will take about 40 min depending on the computer system. As before, MCMCtree will generate files `mcmc.txt`, `out.txt` and `FigTre.tree`. Figure 7 shows the dated phylogeny plotted with FigTree.

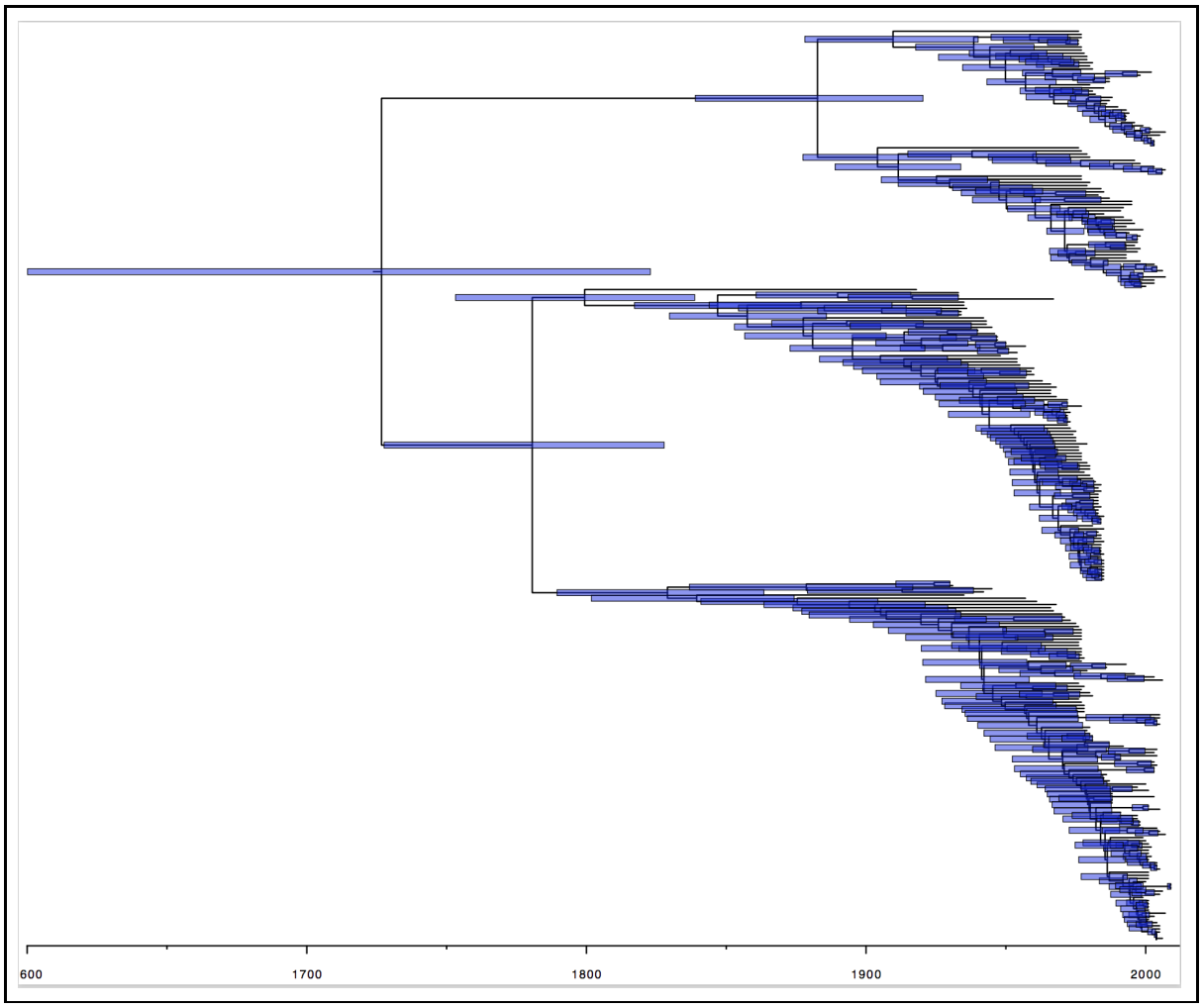

Figure 7. The dated influenza phylogeny as plotted by FigTree. Time unit is 1 year.

As an exercise, repeat the MCMC posterior sampling, then use the commands in file `R/flu.R` to carry out the MCMC diagnostics. You should also go into the `flu/prior` directory and conduct MCMC sampling from the prior as described above.

Dating microbial phylogenies without fossil calibrations nor sampling times

Some microbial groups have scant or no fossil records and may not have taxa with known sampling dates. Dating such microbial phylogenies is a difficult task. At least two approaches appear possible. If a microbial group, with a fossil record, has exchanged genetic material by horizontal transfer with another group, that has no fossil record, then it may be possible to conduct molecular-clock dating using the horizontally-transferred loci and the appropriate

fossil calibration for the outgroup ***(48)***. This approach has been used to, for example, date the origin of methanobacteria ***(9)*** and cyanobacteria, fungi and archaea ***(49)***. Chapter XX in this volume discusses how to date microbial phylogenies with this method. A second plausible approach is to use experimental estimates of bacteria mutation rates ***(50)***. For example, suppose the spontaneous mutation rate estimate for a bacterial group is 0.05 substitutions per 100 My (or 5×10^{-10} substitutions per site per year) with a ± 10% error on the estimate. We can construct a gamma prior on the rate in MCMCtree using 0.05 as the mean, and a standard deviation of 0.05 * 0.1 / 2 = 2.5×10^{-3}. The gamma distribution with parameters α and β has mean α/β and variance α/β^2. Thus, we can use the gamma prior with $\alpha = 400$and $\beta = 8000$, which results in the desired mean and standard deviation. This prior can then be specified in the control file as `rgene_gamma = 400 8000`. A diffuse prior on the root age would still be required. However, using such spontaneous mutation rates to date phylogenies is challenging. Different loci in the genome may evolve at vastly different rates and thus the spontaneous rate estimate may not be a good proxy for the rate of a particular locus. Also, the spontaneous rate estimated for a bacterial group may be very different to other related groups, and thus such a prior may not be appropriate for dating phylogenies with distantly related groups.

References

1. Zuckerkandl E and Pauling L (2014) Evolutionary Divergence and Convergence in Proteins, In: Evolving Genes and Proteins, pp. 97–166
2. Bromham L and Penny D (2003) The modern molecular clock. Nat Rev Genet 4:216–224
3. Drummond AJ, Pybus OG, Rambaut A, et al (2003) Measurably evolving populations. Trends Ecol Evol 18:481–488
4. Benton MJ and Donoghue PCJ (2007) Paleontological evidence to date the tree of life. Mol Biol Evol 24:26–53
5. Tiley GP, Poelstra JW, Reis M dos, et al (2020) Molecular clocks without rocks: New solutions for old problems.
6. Reis M dos, Donoghue PCJ, and Yang Z (2016) Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet 17:71–80
7. Rambaut A, Pybus OG, Nelson MI, et al (2008) The genomic and epidemiological dynamics of human influenza A virus. Nature 453:615–619
8. Betts HC, Puttick MN, Clark JW, et al (2018) Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. 2:1556–1562
9. Wolfe JM and Fournier GP (2018) Horizontal gene transfer constrains the timing of methanogen evolution. Nat Ecol Evol 2:897–903
10. Morris JL, Puttick MN, Clark JW, et al (2018) The timescale of early land plant evolution. Proc Natl Acad Sci U S A 115:E2274–E2283
11. Meredith RW, Janečka JE, Gatesy J, et al (2011) Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334:521–524
12. Misof B, Liu S, Meusemann K, et al (2014) Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763–767
13. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591
14. Reis M dos and Yang Z (2011) Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol Biol Evol 28:2161–2172
15. Battistuzzi FU, Billing-Ross P, Paliwal A, et al (2011) Fast and slow implementations of relaxed-clock methods show similar patterns of accuracy in estimating divergence times. Mol Biol Evol 28:2439–2442

16. Reis M dos, Inoue J, Hasegawa M, et al (2012) Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc Biol Sci 279:3491–3500
17. Jarvis ED, Mirarab S, Aberer AJ, et al (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331
18. Reis M dos, Gunnell GF, Barba-Montoya J, et al (2018) Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: Primates as a test case. Syst Biol 67:594–615
19. Rannala B and Yang Z (2007) Inferring speciation times under an episodic molecular clock. Syst Biol 56:453–466
20. Yang Z and Rannala B (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol 23:212–226
21. Inoue J, Donoghue PCJ, and Yang Z (2010) The impact of the representation of fossil calibrations on bayesian estimation of species divergence times. Syst Biol 59:74–89
22. Wilkinson RD, Steiper ME, Soligo C, et al (2011) Dating primate divergences through an integrated analysis of palaeontological and molecular data. Syst Biol 60:16–31
23. Stadler T and Yang Z (2013) Dating phylogenies with sequentially sampled tips. Syst Biol 62:674–688
24. Reis M dos, Zhu T, and Yang Z (2014) The impact of the rate prior on Bayesian estimation of divergence times with multiple Loci. Syst Biol 63:555–565
25. Reis M dos, Thawornwattana Y, Angelis K, et al (2015) Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. Curr Biol 25:2939–2950
26. Ronquist F, Teslenko M, Van Der Mark P, et al (2012) Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542
27. Lartillot N, Lepage T, and Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288
28. Bouckaert R, Heled J, Kühnert D, et al (2014) BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput Biol 10
29. Felsenstein J (2003) Inferring phylogenies, Sinauer Associates
30. Yang Z (2014) Molecular evolution: A statistical approach, Oxford University Press
31. Holder M and Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet 4:275–284
32. Heath T a. and Moore BR (2014) Bayesian inference of species divergence times, In: Chen, M.-H., Kuo, L., and Lewis, P.O. (eds.) Bayesian Phylogenetics: Methods, Algorithms, and Applications, pp. 277–318 CRC Press
33. Nascimento FF, Reis M dos, and Yang Z (2017) A biologist's guide to Bayesian phylogenetic analysis. Nat Ecol Evol 1:1446–1454
34. Thorne JL, Kishino H, and Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15:1647–1657
35. Le SQ and Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25:1307–1320
36. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314
37. Yang Z (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus {A}. J Mol Evol 51:423–432
38. Rannala B (2016) Conceptual issues in Bayesian divergence time estimation. Philos Trans R Soc Lond B Biol Sci 371
39. Warnock RCM, Parham JF, Joyce WG, et al (2015) Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. Proc Biol Sci 282:20141013
40. Bos KI, Harkins KM, Herbig A, et al (2014) Pre-Columbian mycobacterial genomes

reveal seals as a source of New World human tuberculosis. Nature 514:494–497
41. Arning N and Wilson DJ (2020) The past, present and future of ancient bacterial DNA. Microb Genom 6
42. Martinez-Urtaza J, Trinanes J, Gonzalez-Escalona N, et al (2016) Is El Niño a long-distance corridor for waterborne disease? Nat Microbiol 1:16018
43. Taubenberger JK, Reid AH, Lourens RM, et al (2005) Characterization of the 1918 influenza virus polymerase genes. Nature 437:889–893
44. Worobey M, Gemmel M, Teuwen DE, et al (2008) Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455:661–664
45. Rambaut A (2000) Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 16:395–399
46. Reis M dos, Tamuri AU, Hay AJ, et al (2011) Charting the host adaptation of influenza viruses. Mol Biol Evol 28:1755–1767
47. Yang Z (1994) Estimating the pattern of nucleotide substitution. J Mol Evol 39:105–111
48. Dos Reis M (2018), Fossil-free dating
49. Davín AA, Tannier E, Williams TA, et al (2018) Gene transfers can date the tree of life. Nat Ecol Evol 2:904–909
50. Rosche WA and Foster PL (2000) Determining mutation rates in bacterial populations. Methods 20:4–17