Model adequacy using BEAST v2.4.2

Assessing clock and substitution models

David A. Duchene

1 Background

This tutorial will guide you through methods to assess model adequacy in BEAST v2.4.2. It is common practice to select evolutionary models based on their statistical fit relative to each other. This might be sufficient in some cases, but there is a risk that all of the candidate models lead to inferences that poorly represent the true evolutionary process. Indeed, even the most complex or best fitting model from a set of candidates can produce highly erroneous estimates of parameters of interest. In this tutorial we will explore methods to investigate the absolute merits of the model. This kind of assessment of the absolute performance of models is also known as model checking or assessment of model adequacy or plausibility.

Before starting the tutorial, it is important that you understand the methods used in Bayesian inference for assessing model adequacy. A typical assessment of model adequacy is done by comparing a test statistic, calculated from the empirical data set, with the values of the statistic calculated from a large number of data sets simulated under the model. The test statistic should be informative about the assumptions of the model in question. The simulated data from the model are frequently referred to as posterior predictive (pps) simulations, and they can be considered as future or alternative data sets under the candidate model.

Several test statistics have been proposed to assess several of the components of phylogenetic analysis. In this tutorial we will investigate two test statistics, one for assessing the substitution model, and one for assessing the priors on rates and times used for molecular dating (Figure 1). The first statistic is called the multinomial likelihood, which is the likelihood of the data under a model with only a single general assumption: that substitution events are independent and identically distributed. This statistic can be used to assess overall substitution model fit. Note that sites with missing data or indels should not be included when estimating the multinomial likelihood. The second statistic is the A index, and assesses the power of the molecular dating model to estimate the number of substitutions across branches, assuming an adequate substitution model is used. We will go through this statistic in more detail during the tutorial.

Please also add a section after the exercise interpreting the results. End your tutorial with some useful links.

2 Programs used in this Exercise

- BEAST2.
- R programming environment.
  - R packages *ape* and *phangorn*.

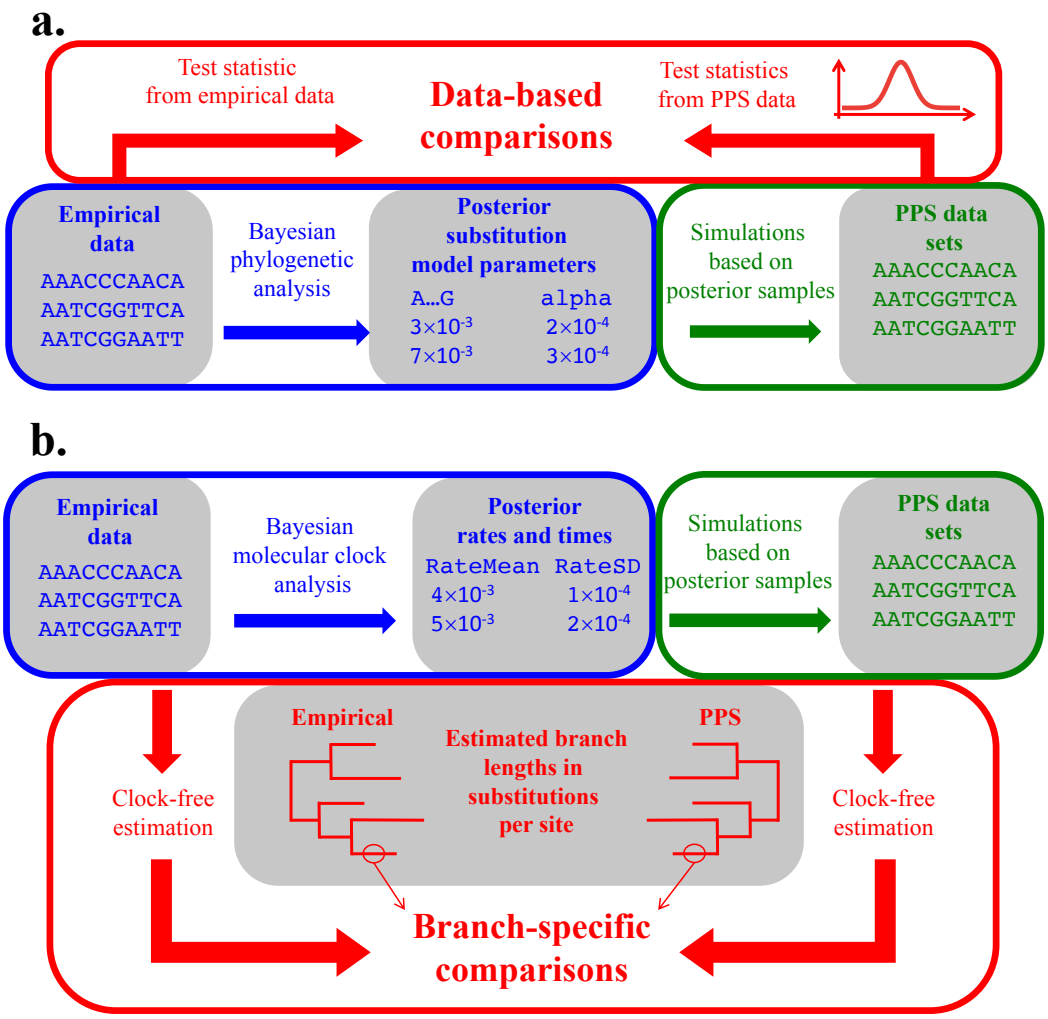

Figure 1: Two of the existing approaches to using posterior predictive simulations to assess model adequacy in Bayesian phylogenetics. (a) One group of methods use characteristics of the data for model assessment, like the multinomial likelihood or the GC content. (b) One method exists that can assess clock models using estimates from clock-free methods. The number of substitutions per site expected along each branch under the clock hierarchical model are then compared with those inferred in a clock-free analysis.

3 Practical: Steps for assessing model adequacy

3.1 Step 1: run the empirical data

In the data folder you will find a simulated sequence alignment with 2000 nucleotides in nexus format (al.nex) and a chronogram in newick format (chrono.tre) with 50 taxa. The alignment was simulated along the chronogram under a Jukes-Cantor substitution model and with rate auto-correlation among lineages. You will also find an XML file (sim.xml) in the xml folder to run BEAST 2 for this data set under a strict clock and a Jukes-Cantor model of substitution and a root calibration.

In this first step, we will run the XML file using BEAST 2. The output of this run can also be found in the folder preekooked runs.

3.2 Step 2: reading the runs and simulating data

To cement the next steps for assessing model adequacy, we will use some of the code in R that we have provided. Specifically, we will explore in detail the file adeq.R. We will first study the code in R (or the comments) in detail before running the analysis.

The following is the full adeq.R file.

```
adeq <- function(trees.file, log.file, empdat.file, Nsim = 100){
    empdat <- as.phyDat(as.DNAbin(read.nexus.data(empdat.file)))
    seqlen <- ncol(as.matrix(as.DNAbin(empdat)))
    tree.topo <- read.nexus(trees.file)[[1]]
    sims <- make.pps.als(trees.file, log.file, Nsim, seqlen)
    sims <- make.pps.tr(sims, empdat, tree.topo)
    bls <- compile.results(sims)
    return(bls)
}
```

One way to simulate data sets from the model is to use the R package *phangorn*. We first need to read the posterior trees and parameter estimates of your analysis in R. We do this using the function make.pps.als, which identifies the model being assessed and simulates data accordingly. The input for this section includes:

1. The paths of the posterior files for your analysis.
2. The number of simulations you want to perform.
3. The sequence length (number of sites in your alignment).

```
    sims <- make.pps.als(trees.file, log.file, Nsim, seqlen)
```

If the input file has a greater number of samples than the number of simulations requested, this will randomly select samples from the posterior. It is useful at this stage if you make sure you understand how an alignment can be simulated using data from the posterior.

3.2.1 Step 3: calculate test statistics

Once we have simulated data sets, we can calculate the test statistics. The function make.pps.trs takes the assumed tree, the substitution model, and the empirical and simulated data sets. The function estimates

phylogenetic branch lengths for the empirical data set and each of the simulated data sets. In this step we also estimate the multinomial likelihood test statistic for the empirical data and each of the simulated data sets.

```
sims <- make.pps.tr(sims, empdat, tree.topo)
```

The output of this function is what we need for model assessment: the test statistics for the empirical data, and the distribution of values for the simulated data sets.

3.2.2 Step 4: calculate *P*-values

We can now compare the test statistic for the empirical data and each of the simulated data sets. The most common way to do this is to calculate the tail area probability, which is the number of simulations with a test statistic greater than the value for empirical data.

```
bls <- compile.results(sims)
```

This function will provide the test statistics for simulations, as well as P-values for each of the test statistics. Following practice from frequentist statistics, we can consider the model to be inadequate if the P-value for a given test statistic is below 0.05. Importantly, the assessment of the clock model allows us to identify for which branches the molecular dating model can estimate the number of substitutions. We will explore this in the following section.

3.2.3 Performing model assessment

The results from the previous steps can be found in the precooked runs folder. The following is example code to run and save the results for clock model adequacy after the BEAST 2 run has completed.

After the BEAST 2 run using the empirical data has completed, the following example code can be run to perform model assessment.

Begin by opening R. The following will set the working directory to the scripts folder, and then source all the functions in the folder.

```
setwd("[INSERT THE PATH TO SCRIPTS FOLDER]")
for(i in dir()) source(i)
```

Next, we set the directory to precooked runs, and run the function adeq(). The arguments for this function are the posterior of trees in nexus format, the log file, the alignment in nexus format, and the number of posterior predictive simulations to be performed.

```
setwd("../example_run_and_results")
clock_adequacy_example <- adeq(trees.file = "sim.trees", sim.log ="sim.log", empdat.file = "↩
    al.nex", Nsim = 100)
names(clock_adequacy_example)
```

The elements in clock adequacy example are each of the components that can be used to assess clock and substitution model adequacy.

The clock adequacy example object should have the same contents as object "allres" in the file results.Rdata:

```
load("results.Rdata")
names(allres)
```

The following section uses the results from results.Rdata to produce a variety of useful graphics.

3.2.4 Visualising the results of clock model assessement

It is strongly recommended to use qualitative checks of models using graphical analyses. This section uses the results in precooked runs/results.Rdata to graph different components for assessing clock model adequacy using posterior predictive simulations.

The following script shows a simple example to explore the branch wise posterior predictive P-values. It requires to have the tree loaded. In this case we will use the original simulated tree, but usually the tree with the median posterior branching times would be appropriate. We will colour the branches with good accuracy in blue, and the branches that have poor accuracy in green (Figure 2).

```
tr <- read.tree("chrono.tre")
plot(tr, edge.col = rainbow(length(allres$branch_wise_pppvalues), start = 2/6, end = 4/6)[
    rank(allres$branch_wise_pppvalues)], edge.width = 6, cex = 1.5)
edgelabels(allres$branch_wise_pppvalues, bg = "white", cex = 1.5, frame = "none")
```

The following script shows a simple example to explore the branch wise length deviation, which is another metric for accuracy. We apply the same colouring system as the plot above, but note that in the case of branch length deviation larger numbers indicate greater deviation from the empirical branch length, and therefore lower accuracy (Figure 3).

```
plot(tr, edge.col = rainbow(length(allres$branch_length_deviation), start = 4/6, end = 2/6)[↩
    rank
        (allres$branch_length_deviation)], edge.width = 6, cex = 1.5)
edgelabels(round(allres$branch_length_deviation, 2), bg = "white", cex = 1.5, frame = "none")
```

Note that in this simple method to graph the results, the branches in the two plots above have been coloured by their rank, rather than their magnitude.

4 Useful references

- Goldman, N. (1993). Statistical tests of models of DNA substitution. Journal of molecular evolution, 36(2), 182-198.
- Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. Molecular Biology and Evolution, 19(7), 1171-1180.
- Duchene, D. A., Duchene, S., Holmes, E. C., Ho, S. Y. (2015). Evaluating the adequacy of molecular clock models using posterior predictive simulations. Molecular biology and evolution, 32(11), 2986-2995.

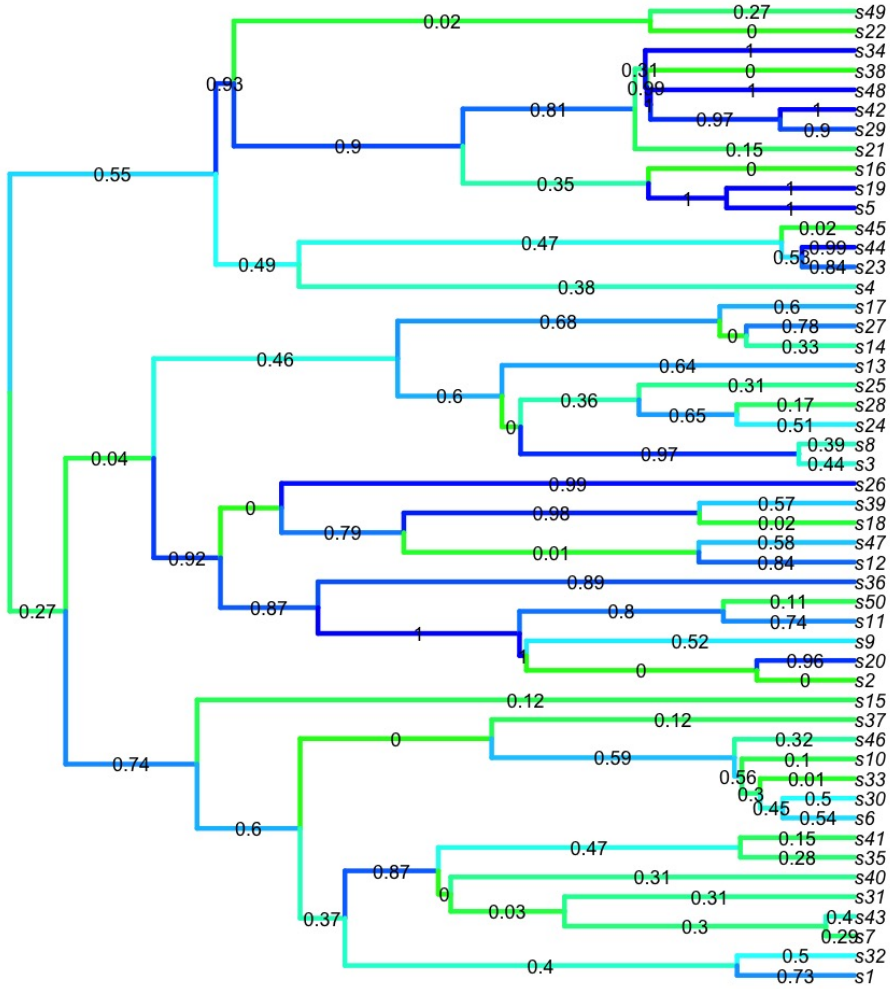

Figure 2: Estimated chronogram with branches coloured by their clock adequacy P-value.

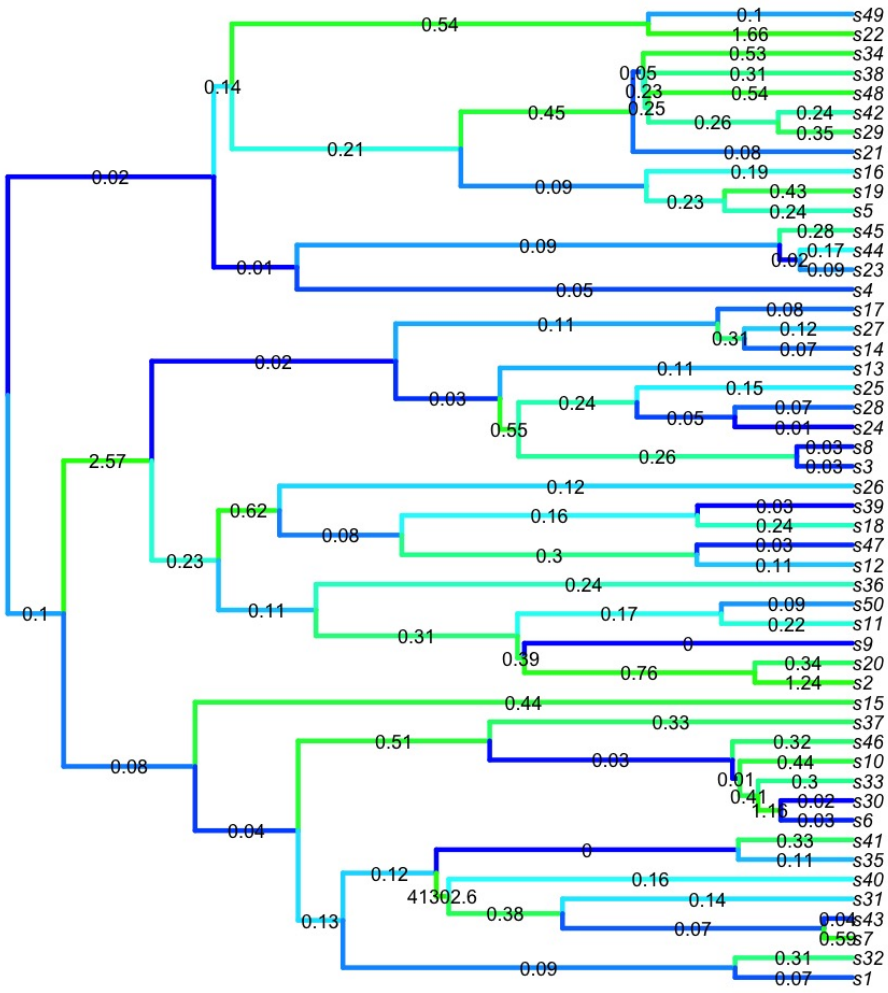

Figure 3: Estimated chronogram with branches coloured by their deviation between the empirical and simulated lengths.

Version dated: January 25, 2017