2013 Applied Phylogenetics Workshop, Bodega Bay, CA USA

Divergence time estimation using BEAST v1.7.5

Central among the questions explored in biology are those that seek to understand the timing and rates of evolutionary processes. Accurate estimates of species divergence times are vital to understanding historical biogeography, estimating diversification rates, and identifying the causes of variation in rates of molecular evolution.

This tutorial will provide a general overview of divergence time estimation and fossil calibration in a Bayesian framework. The exercise will guide you through the steps necessary for estimating phylogenetic relationships and dating species divergences using the program BEAST v1.7.5.

Background: Divergence time estimation

Estimating branch lengths in proportion to time is confounded by the fact that the rate of evolution and time are intrinsically linked when inferring genetic differences between species. A model of lineage-specific substitution rate variation must be applied to tease apart rate and time. When applied in methods for divergence time estimation, the resulting trees have branch lengths that are proportional to time. External node age estimates from the fossil record or other sources are necessary for inferring the real-time (or absolute) ages of lineage divergences (Figure 1).

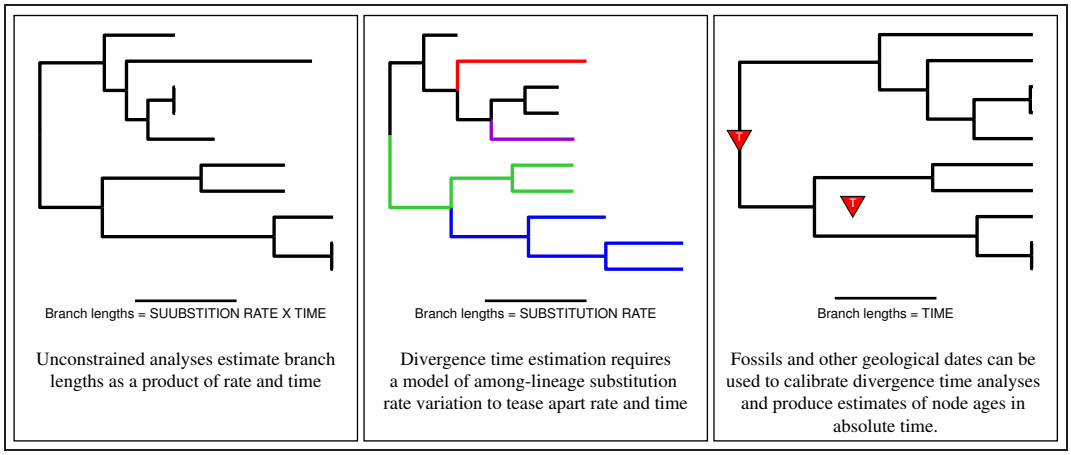


Figure 1: Estimating branch lengths in units of time requires a model of lineage-specific rate variation and external information to calibrate the tree.

Modeling lineage-specific substitution rates

Many factors can influence the rate of substitution in a population such as mutation rate, population size, generation time, and selection. As a result, many models have been proposed that describe how substitution rate may vary across the Tree of Life.

The simplest model, the molecular clock, assumes that the rate of substitution remains constant over time (Zuckerkandl and Pauling, 1962). However, many studies have shown that molecular data (in general) violate the assumption of a molecular clock and that there exists considerable variation in the rates of substitution among lineages.

Several models have been developed and implemented for inferring divergence times without assuming a strict molecular clock and are commonly applied to empirical data sets. Many of these models have been applied as priors using Bayesian inference methods. The implementation of dating methods in a Bayesian framework provides a flexible way to model rate variation and obtain reliable estimates of speciation times, provided the assumptions of the models are adequate. When coupled with numerical methods such as Markov chain Monte Carlo (MCMC), for approximating the posterior probability distribution of parameters, Bayesian inference methods can be extremely powerful for estimating the parameters of a statistical model and are widely used in phylogenetics.

Some models of lineage-specific rate variation:

- Global molecular clock: a constant rate of substitution over time (Zuckerkandl and Pauling, 1962)
- Local molecular clocks (Kishino, Miyata and Hasegawa, 1990; Rambaut and Bromham, 1998; Yang and Yoder, 2003; Drummond and Suchard, 2010)
  - Closely related lineages share the same rate and rates are clustered by sub-clades
- Compound Poisson process (Huelsenbeck, Larget and Swofford, 2000)
  - Rate changes occur along lineages according to a point process and at rate-change events, the new rate is a product of the old rate and a Γ-distributed multiplier.
- Autocorrelated rates: substitution rates evolve gradually over the tree
  - Log-normally distributed rates: the rate at a node is drawn from a log-normal distribution with a mean equal to the parent rate (Thorne, Kishino and Painter, 1998; Kishino, Thorne and Bruno, 2001; Thorne and Kishino, 2002)
  - Cox-Ingersoll-Ross Process: the rate of the daughter branch is determined a non-central χ^2 distribution. This process includes a parameter that determines the intensity of the force that drives the process to its stationary distribution (Lepage et al., 2006).
- Uncorrelated rates
  - The rate associated with each branch is drawn from a single underlying parametric distribution such as an exponential or log-normal (Drummond et al., 2006; Lepage et al., 2007; Heath, Holder and Huelsenbeck, 2012).

The variety of models for relaxing the molecular clock assumption presents a challenge for investigators interested in estimating divergence times. Some models assume that rates are heritable and autocorrelated over the tree, others model rate change as a step-wise process, and others assume that the rates on each branch are independently drawn from a single distribution. Furthermore, studies comparing the accuracy (using simulation) or precision of different models have produced conflicting results, some favoring uncorrelated models (Drummond et al., 2006) and others preferring autocorrelated models (Lepage et al., 2007). Because of this, it is important for researchers performing these analyses to consider and test different relaxed clock models (Lepage et al., 2007; Ronquist et al., 2012; Li and Drummond, 2012; Baele et al., 2013). It is also critical to take into account the scale of the question when estimating divergence times. For example, it might not be reasonable to assume that rates are autocorrelated if the data set includes very distantly related taxa and low taxon sampling. In such cases, it is unlikely that any signal of autocorrelation is detectible.

Fossil calibration

Without external information to calibrate the tree, divergence time estimation methods can only reliably provide estimates of relative divergence times (which are useful for some analyses: quantitative trait evolution, estimating relative diversification rates) and not absolute node ages. Calibration information

can come from a variety of sources including "known" substitution rates (often secondary calibrations estimated from a previous study), dated tip sequences from serially sampled data (typically time-stamped virus data), or geological date estimates (fossils or biogeographical data).

Age estimates from fossil organisms are the most common form of divergence time calibration information. These data are used as age constraints on their putative ancestral nodes. There are numerous difficulties with incorporating node age estimates from fossil data including disparity in fossilization and sampling, uncertainty in dating, and correct phylogenetic placement of the fossil. Thus, it is critical that careful attention is paid to the paleontological data included in phylogenetic divergence time analyses.

With an accurately dated and identified fossil in hand, further consideration is required to determine how to apply the node-age constraint. If the fossil is truly a descendant of the node it calibrates, then it provides a reliable minimum age bound on the ancestral node time. However, maximum bounds are far more difficult to come by. Bayesian methods provide a way to account for uncertainty in fossil calibrations. Prior distributions reflecting our knowledge (or lack thereof) of the amount of elapsed time from the ancestral node to is calibrating fossil are easily incorporated into these methods.

A nice review paper by Ho and Phillips (2009) outlines a number of different parametric distributions appropriate for use as priors on calibrated nodes. In this exercise we will use the uniform, normal, log-normal, and exponential distributions (Figure 2).

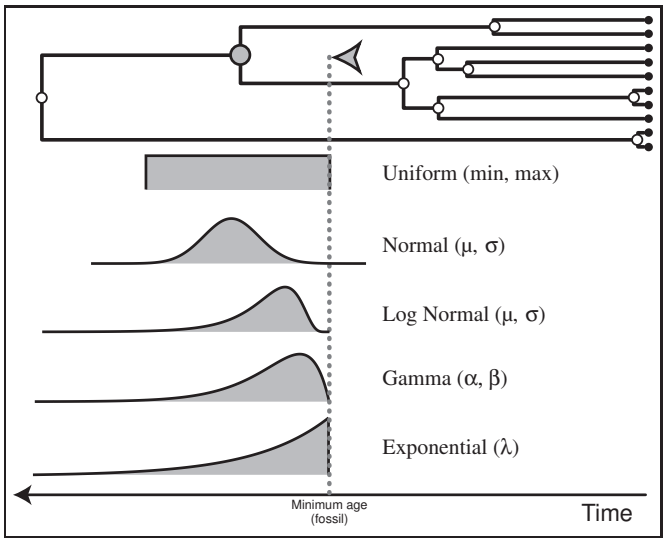


Figure 2: Five different parametric distributions that can be applied as priors on the age of a calibrated node.

Uniform distribution – Typically, you must have both maximum and minimum age bounds when applying a uniform calibration prior (though some methods are available for applying uniform constraints with soft bounds). The minimum bound is provided by the fossil member of the clade. The maximum bound may come from a bracketing method or other external source. This distribution places equal probability across all ages spanning the interval between the lower and upper bounds.

Normal distribution – The normal distribution is not always appropriate for calibrating a node using fossil information (though some methods allow for assigning a truncated normal prior density). When applying a biogeographical date (e.g. the Isthmus of Panama) or a secondary calibration (a node age estimate from a previous study), the normal distribution can be a useful calibration prior. This distribution is always symmetrical and places the greatest prior weight on the mean (μ). Its scale is determined by the standard

deviation parameter (σ).

Probability distributions restricted to the interval $[0, \infty)$, such as the log-normal, exponential, and gamma are appropriate for use as zero-offset calibration priors. When applying these priors on node age, the fossil age is the origin of the prior distribution. Thus, it is useful to consider the fact that the prior is modeling the amount of time that has elapsed since the divergence event (ancestral node) until the time of the descendant fossil (Figure 3).

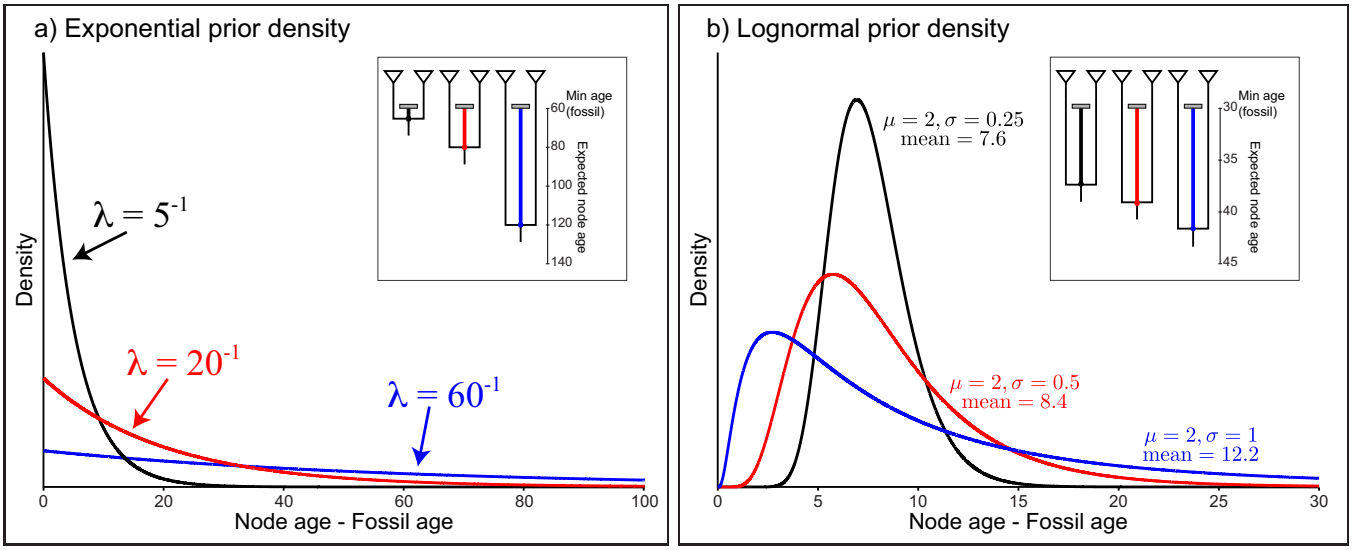


Figure 3: Two common prior densities for calibrating node ages. a) The exponential distribution with three different values for the rate parameter, λ. As the value of the λ rate parameter is decreased, the prior becomes less informative (the blue line is the least informative prior, $\lambda = 60^{-1}$). The inset shows an example of the three different priors and their expected values placed on the same node with a minimum age bound of 60. b) The lognormal distribution with 3 different values for the shape parameter, σ. For this distribution, even though μ is equal to 2.0 for all three, expected value (mean) is dependent on the value of σ. The inset shows an example of the three different priors and their expected values placed on the same node with a minimum age bound of 30.

Gamma distribution – The gamma distribution is commonly used as a prior on scalar variables in Bayesian inference. It relies on 2 parameters: the scale parameter (α) and a rate parameter (λ). More specifically, the gamma distribution is the sum of α independently and identical exponentially distributed random variables with rate λ. As α becomes very large ($\alpha > 10$), this distribution approaches the normal distribution.

Exponential distribution – The exponential distribution is a special case of the gamma distribution and is characterized by a single rate parameter (λ) and is useful for calibration if the fossil age is very close to the age of its ancestral node. The expected (mean) age difference under this distribution is equal to λ^{-1} and the median is equal to $\lambda^{-1} * ln(2)$. Under the exponential distribution, the greatest prior weight is placed on node ages very close to the age of the fossil with diminishing probability to ∞. As λ is increased, this prior density becomes strongly informative, whereas very low values of λ result in a fairly non-informative prior (Figure 3a).

Log-normal distribution – An offset, log-normal prior on the calibrated node age places the highest probability on ages somewhat older than the fossil, with non-zero probability to ∞. If a variable is log-normally distributed with parameters μ and σ, then the natural log of that variable is normally distributed with a mean of μ and standard deviation of σ. The median of the lognormal distribution is equal to e^{μ} and the mean is equal to $e^{\mu+\frac{\sigma^2}{2}}$ (Figure 3b).

Priors on node times

There are many component parts that make up a Bayesian analysis of divergence time. One that is often overlooked is the prior on node times, often called a *tree prior*.

In BEAST, the available tree priors for divergence time estimation using inter-species sequences are the ***Yule*** prior and the ***Birth-Death*** prior (Kendall, 1948; Nee, May and Harvey, 1994; Gernhard, 2008; Stadler, 2010). These priors are based on the birth-death process which assumes that lineages speciate and go extinct according to a stochastic process with a single rate for speciation (birth $= \lambda$) and a single rate for extinction (death $= \mu$). The Yule model is a special case of the birth-death process where $\mu = 0$. Extensions of these models are also available in BEAST, including the ***Calibrated Yule***, ***Birth Death Incomplete-Sampling***, and ***Birth-Death Serially Sampled***. Other programs also offer these speciation priors as well as some alternative priors such as a uniform prior (*PhyloBayes*, *MrBayes v3.2*, *DPPDiv*), a Dirichlet prior (*multidivtime*), and a birth-death prior with species sampling (*MCMCTree*). BEAST also offers tree priors based on the coalescent which are intended for population-level analyses or time-stamped virus data. The effect of different node-time priors on estimates of divergence times is not well understood and appears to be dataset-dependent (Lepage et al., 2007). Accordingly, it is important to account for the characteristics of your data when choosing a tree prior.

Programs used in this exercise

BEAST – Bayesian Evolutionary Analysis Sampling Trees

BEAST is a free software package for Bayesian evolutionary analysis of molecular sequences using MCMC and strictly oriented toward inference using rooted, time-measured phylogenetic trees. The development and maintenance of BEAST is a large, collaborative effort and the program includes a wide array of different types of analyses:

- Phylogenetic tree inference under different models for substitution rate variation
  - Constant rate molecular clock (Zuckerkandl and Pauling, 1962)
  - Uncorrelated relaxed clocks (Drummond et al., 2006)
  - Random local molecular clocks (Drummond and Suchard, 2010)
- Estimates of species divergence dates and fossil calibration
- Analysis of non-contemporaneous sequences
- Heterogenous substitution models across data partitions
- Population genetic analyses
  - Estimation of demographic parameters (population sizes, growth/decline, bottlenecks, migration)
  - Bayesian skyline plots
  - Phylogeography (Lemey et al., 2009)
- Gene-tree/species-tree inference (∗BEAST; Heled and Drummond, 2010)
- and more...

BEAST is written in java and its appearance and functionality are consistent across platforms. Inference using MCMC is done using the BEAST program, however, there are several utility applications that assist in the preparation of input files and summarize output (BEAUti, LogCombiner, and TreeAnnotator are all part of the BEAST software bundle; http://beast.bio.ed.ac.uk).

BEAUti – Bayesian Evolutionary Analysis Utility

BEAUti is a utility program with a graphical user interface for creating BEAST and *BEAST input files which must be written in the eXtensible Markup Language (XML). This application provides a clear way to specify priors, partition data, calibrate internal nodes, etc.

LogCombiner – When multiple (identical) analyses are run using BEAST (or MrBayes), LogCombiner can be used to combine the parameter log files or tree files into a single file that can then be summarized using Tracer (log files) or TreeAnnotator (tree files). However, it is important to ensure that all analyses reached convergence and sampled the same stationary distribution before combining the parameter files.

TreeAnnotator – TreeAnnotator is used to summarize the posterior sample of trees to produce a maximum clade credibility tree and summarize the posterior estimates of other parameters that can be easily visualized on the tree (e.g. node height). This program is also useful for comparing a specific tree topology and branching times to the set of trees sampled in the MCMC analysis.

Tracer – Tracer is used for assessing and summarizing the posterior estimates of the various parameters sampled by the Markov Chain. This program can be used for visual inspection and assessment of convergence and it also calculates 95% highest posterior density (95% HPD) intervals and effective sample sizes (ESS) of parameters (http://tree.bio.ed.ac.uk/software/tracer).

FigTree – FigTree is an excellent program for viewing trees and producing publication-quality figures. It can interpret the node-annotations created on the summary trees by TreeAnnotator, allowing the user to display node-based statistics (e.g. posterior probabilities) in a visually appealing way (http://tree.bio.ed.ac.uk/software/figtree).

The eXtensible Markup Language

The eXtensible Markup Language (XML) is a general-purpose markup language, which allows for the combination of text and additional information. In BEAST, the use of the XML makes analysis specification very flexible and readable by both the program and people. The XML file specifies sequences, node calibrations, models, priors, output file names, etc. BEAUti is a useful tool for creating an XML file for many BEAST analyses. However, typically, dataset-specific issues can arise and some understanding of the BEAST-specific XML format is essential for troubleshooting. Additionally, there are a number of interesting models and analyses available in BEAST that cannot be specified using the BEAUti utility. Refer to the BEAST web page (http://beast.bio.ed.ac.uk/XML_format) for detailed information about the BEAST XML format. Box 1 shows an example of BEAST XML syntax for specifying a birth-death prior on node times.

```
<!-- A prior on the distribution node heights defined given                    -->
<!-- a Birth-Death speciation process (Gernhard 2008).                         -->
<birthDeathModel id="birthDeath" units="substitutions">
    <birthMinusDeathRate>
        <parameter id="birthDeath.meanGrowthRate" value="1.0" lower="0.0" upper="Infinity"/>
    </birthMinusDeathRate>
    <relativeDeathRate>
        <parameter id="birthDeath.relativeDeathRate" value="0.5" lower="0.0" upper="Infinity"/>
    </relativeDeathRate>
</birthDeathModel>
```

Box 1: BEAST XML specification of the speciation model.

Practical: Divergence time estimation

For this exercise, we will estimate phylogenetic relationships and date the species divergences of the ten simulated sequences in the file called `bodega.nex`. This simple alignment contains two genes, each 500 nucleotides in length.

- Download all of the compressed directories from:
  https://molevol.mbl.edu/wiki/index.php/Divergence_Times
  and place them in a directory you've created named: `bodega_beast`. After uncompressing, you should have the files listed in Box 2.

```
• bodega_beast/data/
    – bodega.nex
    – bodega_start_tree.tre
• bodega_beast/output1/
    – bodega.log
    – bodega.prior.log
    – bodega.trees
    – bodega.prior.trees
    – bodega.ops
    – bodega.prior.ops
• bodega_beast/output2/
    – bodega_100m.1.log
    – bodega_100m.2.log
    – bodega_100m.prior.log
• bodega_beast/output3/
    – bodega_100m.1.trees
    – bodega_100m.2.trees
    – bodega.comb.trees
```

Box 2: The data files required for this exercise.

- Open the NEXUS file containing the sequences in your text editor. The `ASSUMPTIONS` block contains the commands for partitioning the alignment into two separate genes (Box 3). Tests for model selection indicated that `gene1` and `gene2` evolved under separate GTR+Γ models.

```
#NEXUS
BEGIN DATA;
    DIMENSIONS NTAX=10 NCHAR=1000;
    FORMAT DATATYPE = DNA GAP = - MISSING = ?;
    MATRIX
    T1   CTACGGGAGGGCAACGGGGCTAGATGGTAAACGCGCCATCGATCGCAAG...
    T2   CTACGGGAGGGCGACGGGGCTAGATGGTAAACGCGCCCTCGATCGCAAG...
    T3   CAGCGTGAGGGCCACGGGGCTGGCAGGTACTCCGGCCCACGAGTGGAAG...
    T4   CAGCGTGGGGGCCACGGGGCTAGAAGTTACTCCGGCCCACGAGTGGAAG...
    T5   CAGCGTGGGGGCCACGGGGCTAGAAGTTACTCCGGCCCACGAGTGGAAG...
    T6   CAGCGAGAAGCCGACGGGGATGGAAGGGACTCAGACGCACGAGTCCATG...
    T7   CATCGCGAGGGGGACGGGGCTCGTAGATTATCGTTCATGCAAGCTGAAG...
    T8   CATCGCGAGGGGGACGGGGCTCGTAGTTTATCGTTCAGGCAAGCTGAAG...
    T9   CAGCGTGACCACGACGGGGCTGGGGGTGATTCCCGCTGACAAGATGAAG...
    T10  CTGCGTGACAACGACGGGGCTGGGAGTTGTTCCCGCTCACAAGAGGAAG...
;
END;

BEGIN ASSUMPTIONS;
    charset gene1 = 1-500;
    charset gene2 = 501-1000;
END;
```

Box 3: A fragment of the NEXUS file containing the sequences for this exercise. The data partitions are defined in the `ASSUMPTIONS` block..

These 10 sequences consist of 6 ingroup taxa: `T1, T2, T3, T4, T5, T6` and 4 outgroup taxa: `T7, T8, T9, T10`. After performing an unconstrained analysis using maximum likelihood, we get the topology in Figure 4. The branch lengths estimated under maximum likelihood are indicative of variation in substitution rates. Divergence time estimation and fossil calibration require that you have some prior knowledge of the tree topology so that calibration dates can be properly assigned.

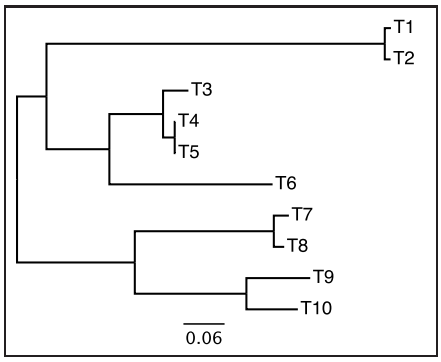


Figure 4: A maximum likelihood estimate of the phylogenetic relationships of `bodega.nex`. (Constructed using PAUP* v4.0a125; Swofford, 1998)

Fossil node calibrations are often difficult to obtain for every node and for many groups they are simply unavailable. In such cases, constraints can be applied to outgroup nodes. There are four external calibration points for this data set. These are illustrated in Figure 5. The oldest fossil belonging to the ingroup can calibrate the age of that clade. This fossil was identified as a member of the clade, falling within the crown group. Two fossils calibrate nodes within the outgroup clade and a well supported estimate of the root age from a previous study allows us to place a prior distribution on that node (Figure 5).

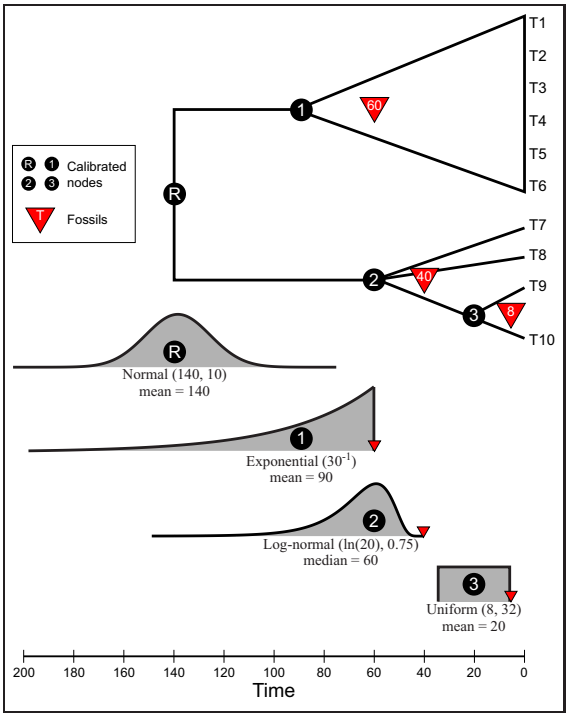


Figure 5: Four nodes with fossil calibration for the `bodega.nex` data set. There are calibrations on the root and 3 other internal nodes. Each calibration point is assumed to have a different prior density.

Getting started with BEAUti

Creating a properly-formatted BEAST XML file from scratch is not a simple task. However, BEAUti provides a simple way to navigate the various elements specific to the BEAST XML format.

- Begin by executing the BEAUti program. For Mac OSX and Windows, you can do this by double clicking on the application. For Unix systems (including Mac OSX), it is convenient to add the entire `BEASTv1.7.5/bin` directory to your path.

- Import the sequences from `bodega.nex` using the pull-down menu: ***File→Import Data***.

This example data set contains 2 different partitions labeled `gene1` and `gene2`. When the NEXUS file is imported into BEAUti, the ***Partitions*** tab lists each partition and their currently assumed substitution model, clock model, and tree.

- Double click on the file name (`bodega.nex`) next to one of the data partitions. This will bring up a window allowing you to visually inspect your alignment.

We would like to analyze each gene under separate substitution models, while assuming the clock and tree are linked.

- Select both `gene1` and `gene2`. While both partitions are highlighted click the ***Unlink Subst. Models*** button. You will notice that the site model listed for `gene2` has changed. [Figure 6]

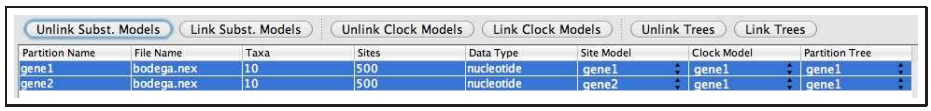


Figure 6: Unlink the substitution models for the two data partitions.

- For the sake of clarity, let's rename the ***Clock Model*** and ***Partition Tree***. While both partitions are highlighted, click the ***Link Clock Models*** button. This will bring up a window allowing you to rename the clock model. Check the box next to ***Rename clock model partition to:*** and provide a new name. The example below calls it `bodegaClock`. [Figure 7]

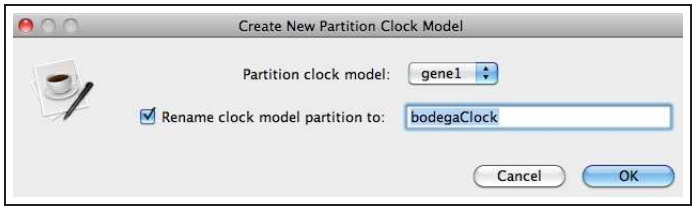


Figure 7: Rename the clock model.

- Similarly, rename the ***Partition Tree*** by clicking the ***Link Trees*** button. Perhaps call it `bodegaTree`.

Now your ***Partitions*** tab should show that the two genes are assumed to have separate substitution models, a single clock model called `bodegaClock`, and a single tree called `bodegaTree`. [Figure 8]

Now that you have set up your data partitions, you can move on to specifying ancestral nodes in the tree for calibration.

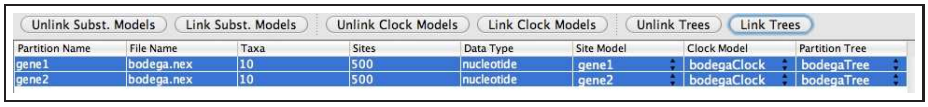


Figure 8: The data partitions with unlinked substitution models and linked clock model and tree.

- Go to the ***Taxa*** tab.

The options in the ***Taxa*** window allow the user to identify internal nodes of interest. Simply creating a taxon set does not necessarily force the clade to be monophyletic nor does it require you to specify a time calibration for that node. Once a taxon set is created, statistics associated with the most recent common ancestor (MRCA) of those taxa (e.g. node height) will be reported in the BEAST parameter log file.

For this data set, we will apply external time calibrations to 4 internal nodes (Figure 5), including the root. Because the root node is implicit, we only have to specify 3 internal nodes in the ***Taxa*** window.

- Create a new taxon set for calibration node 1 by clicking the [+] button in the lower, left corner of the window. Double click on the default name (`untitled1`) for the taxon set and rename it `mrca1`. These taxa form our "ingroup" clade, so check the box under the ***Monophyletic?*** column. There is also a check-box in the column labeled ***IncludeStem?*** this allows the user to specify a calibration for the stem of a clade. Leave this unchecked since we are assuming that the fossil calibration for `mrca1` is within the crown group. You will also see a text-entry box where you can provide a starting age for the node. This is a new feature of BEAUti 1.7.5 that allows you to use a randomly-generated starting tree that is consistent with your calibrations. We will specify our own starting tree, so leave this box empty. [Figure 9]

- In the center panel of the ***Taxa*** window, select the taxa descended from calibration node 1 (`T1, T2, T3, T4, T5, T6`). When you have highlighted each of these taxon names, move them from the ***Excluded Taxa*** column to the ***Included Taxa*** column by clicking the button with the green arrow. [Figure 9]

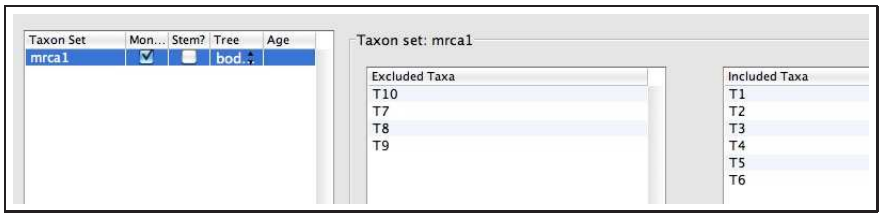


Figure 9: Defining the taxon set for the most recent common ancestor of T1, T2, T3, T4, T5, and T6.

- Create a new taxon set for calibration node 2 (`T7, T8, T9, T10`; Figure 5) and rename it `mrca2`. These taxa make up the "outgroup" and should also be monophyletic (though this is already done by constraining ingroup monophyly).

- Make a new taxon set for calibration node 3 called `mrca3` (`T9, T10`; Figure 5) and do not constrain the clade to be monophyletic.

The ***Tips*** menu contains the options necessary for analyses of data sets containing non-contemporaneous tips. This is primarily for serial sampled virus data and not applicable to this exercise.

The ***Traits*** tab is for gene-tree/species-tree analysis in *BEAST, which is not covered in this tutorial.

- Go to the ***Sites*** menu to specify a substitution model for each of our data partitions. Change the substitution model for `gene1` to ***GTR***, with ***Estimated*** base frequencies, and set the among-site heterogeneity model to ***Gamma***. [Figure 10]

- Specify a GTR+Γ model for `gene2` as well. [Figure 10]

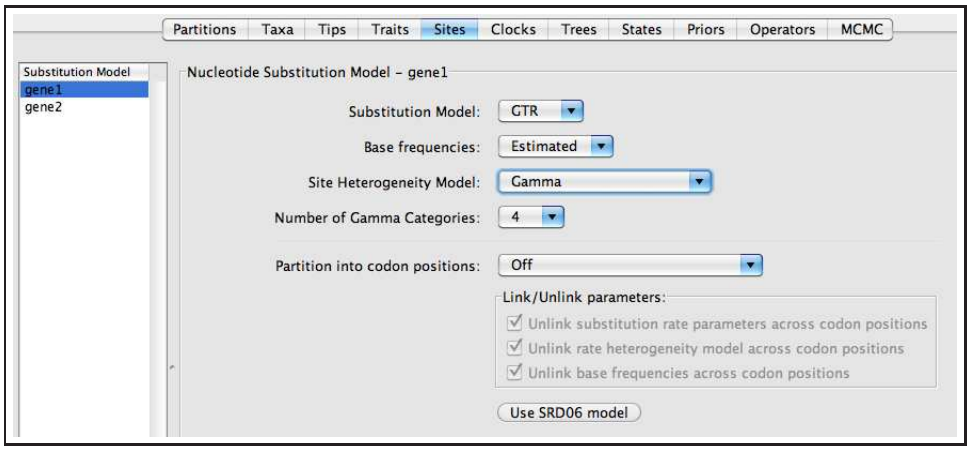


Figure 10: Defining a GTR+Γ model for `gene1`.

You will notice other options in the ***Sites*** menu which are specifically for protein-coding data, allowing you to partition your data set by codon position. The button labeled ***Use SRD06 Model*** will set the model parameters to those used in a paper by Shapiro, Rambaut and Drummond (2006) which partitions the data so that 3rd codon positions are analyzed separately from 1st and 2nd positions and assumes a HKY+Γ model. The sequences in this analysis are not protein-coding, so these options are not applicable to this exercise.

- Move on to the ***Clocks*** menu to set up the relaxed clock analysis.

Here, we can specify the model of lineage-specific substitution rate variation. The default model in BEAUti is the ***Strict Clock*** with a fixed substitution rate equal to `1`. Three models for relaxing the assumption of a constant substitution rate can be specified in BEAUti as well. The ***Lognormal relaxed clock (Uncorrelated)*** option assumes that the substitution rates associated with each branch are independently drawn from a single, discretized lognormal distribution (Drummond et al., 2006). Under the ***Exponential relaxed clock (Uncorrelated)*** model, the rates associated with each branch are exponentially distributed (Drummond et al., 2006). The ***Random local clock*** uses Bayesian stochastic search variable selection to average over random local molecular clocks (Drummond and Suchard, 2010).

- Set the ***Clock Model*** for `bodegaClock` to ***Lognormal relaxed clock (Uncorrelated)*** and check the box in the ***Estimate*** column. [Figure 11]

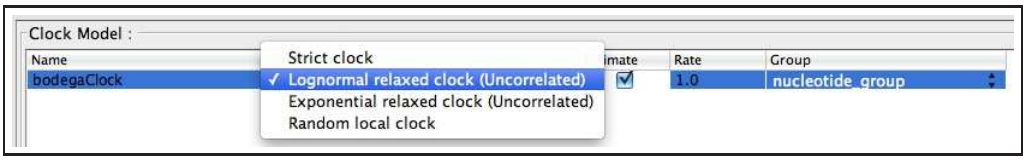


Figure 11: Options for modeling among-lineage substitution rate variation in the ***Clock Model*** menu.

Below the settings for the ***Clock Model*** there is a box labeled ***Clock Model Group***. The clock group table is used to specify shared molecular clocks, and to set apart clocks applied to microsatellite data.

- Move on to the ***Trees*** window.

In the ***Trees*** menu, you can specify a starting tree and the ***Tree Prior*** which is the prior on the distribution of node times.

- Go to the ***Tree Prior*** pull-down menu and choose ***Speciation: Birth-Death Process***. [Figure 12]

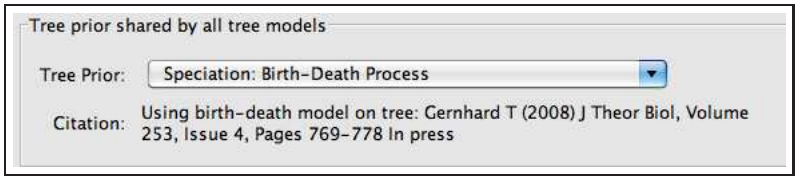


Figure 12: Setting the birth-death prior distribution on branching times.

This model and the Yule model (***Speciation: Yule Process***) are appropriate for analyses of inter-species relationships. Both of these models are stochastic branching processes with a constant rate of speciation (λ) and a constant rate of extinction (μ). In the case of the Yule model, the extinction rate is equal to 0. Both models are implemented in BEAST following Gernhard (2008) with prior distributions on the parameters of the model. Thus, for the birth-death model used in this analysis, our runs will sample the average net diversification rate (`meanGrowthRate` $= \lambda - \mu$) and the relative rate of extinction (`relativeDeathRate` $= \frac{\mu}{\lambda}$).

You will notice several other options for the ***Tree prior*** available in BEAUti. These priors are differentiated by the labels ***Coalescent***, ***Speciation***, or ***Epidemiology***. Coalescent tree priors are appropriate for population-level analyses. Conversely, when you are estimating relationships and divergence times of interspecies data, it is best to employ a speciation prior. Choosing a coalescent prior for estimating deep divergences, or a speciation prior for intra-species data, can often lead to problematic results due to interactions between the prior on the node ages and the prior on the branch rates. Thus, it is critical that these priors are chosen judiciously. Furthermore, it is important to note that our understanding of the statistical properties of speciation prior densities combined with calibration densities is somewhat incomplete, particularly when the tree topology is considered a random variable (Heled and Drummond, 2012; Warnock, Yang and Donoghue, 2012). The option ***Speciation: Calibrated Yule*** is a more statistically sound tree prior when applying a single calibration density. Refer to Heled and Drummond (2012) for more details about this issue.

BEAST also has a few options for initialization of the tree topology and branch lengths. Starting trees can be generated either randomly (under a coalescent model) or with UPGMA. Alternatively, the user can specify the tree by including a `trees` block in the NEXUS data file, importing it directly into BEAUti, or by pasting the Newick string and XML elements in the XML file.

We would like to use our own starting tree. The starting tree topology can come from any type of analysis. For this data set, we performed an initial analysis using maximum likelihood resulting in the tree in Figure 4. Using the maximum likelihood tree topology, branch lengths were generated by drawing them from a uniform distribution conditional on the node age constraints from the fossil record. For this exercise, the file called `bodega_start_tree.tre` is a NEXUS formatted file containing the starting tree. [Figure 13]

- Load the starting tree using the pull-down menu: ***File→Import Data***. Import the NEXUS file containing your starting tree: `bodega_start_tree.tre` (you may have to change the ***File Format***

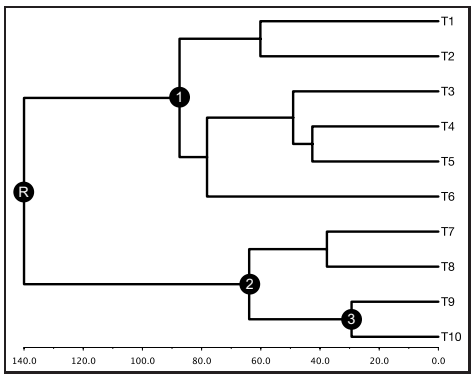


Figure 13: Starting tree. The topology was inferred using maximum likelihood and the node heights were drawn from a uniform distribution with constraints on the root and 3 internal nodes.

to ***All Files***). This will jump you back to the ***Partitions*** window.

- Return to the ***Trees*** window and select ***User-specified starting tree*** in the ***Tree Model*** section. Next to the option ***Select-user specified tree:*** choose `ml_tre`. [Figure 14]

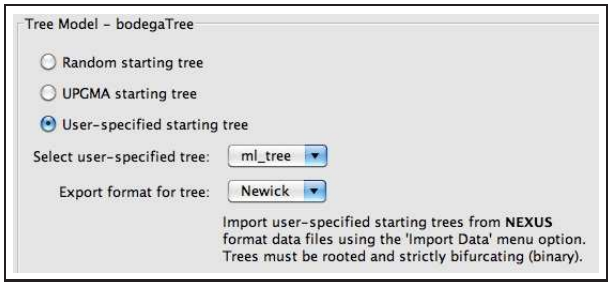


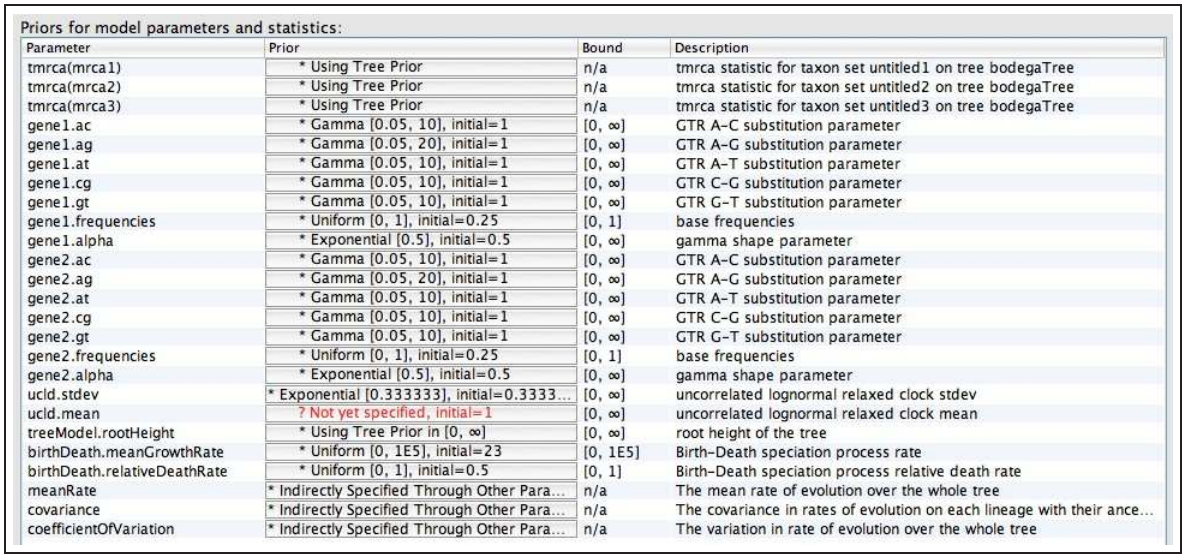
Figure 14: Specifying a starting tree.

The ***States*** panel allows one to specify analyses for ancestral state reconstruction and provides sequence error model options for any data partition. Leave these options unmodified.

- Navigate to the ***Priors*** menu.

In the ***Priors*** window, all of the model parameters and statistics specified in earlier menus are listed. Here you can set prior distributions on substitution model parameters, calibration nodes, and parameters associated with the clock model and tree model. [Figure 15]

The prior on the `ucld.mean` parameter is indicated in red because it is not set by default in this version of BEAUti. When you click on the box for this prior you will see a window allowing you to specify a prior distribution on the mean rate of substitution. In Bayesian terminology this parameter is a *hyperparameter* because it is a parameter describing a prior distribution and not a direct parameter of the data model (like base frequencies or branch lengths). In Bayesian inference a prior distribution can be placed on a hyperparameter, and this is called a *hyperprior*. By allowing the value of this hyperparameter to vary, we are freed from the responsibility of fixing the mean of the log-normal prior distribution on branch-specific substitution rates. Additionally, the Markov chain will sample this hyperparameter along with the other parameters directly associated with the models on our data, providing us with an estimate of the posterior distribution.



Priors for model parameters and statistics:

Parameter	Prior	Bound	Description
tmrca(mrca1)	* Using Tree Prior	n/a	tmrca statistic for taxon set untitled1 on tree bodegaTree
tmrca(mrca2)	* Using Tree Prior	n/a	tmrca statistic for taxon set untitled2 on tree bodegaTree
tmrca(mrca3)	* Using Tree Prior	n/a	tmrca statistic for taxon set untitled3 on tree bodegaTree
gene1.ac	* Gamma [0.05, 10], initial=1	[0, ∞]	GTR A-C substitution parameter
gene1.ag	* Gamma [0.05, 20], initial=1	[0, ∞]	GTR A-G substitution parameter
gene1.at	* Gamma [0.05, 10], initial=1	[0, ∞]	GTR A-T substitution parameter
gene1.cg	* Gamma [0.05, 10], initial=1	[0, ∞]	GTR C-G substitution parameter
gene1.gt	* Gamma [0.05, 10], initial=1	[0, ∞]	GTR G-T substitution parameter
gene1.frequencies	* Uniform [0, 1], initial=0.25	[0, 1]	base frequencies
gene1.alpha	* Exponential [0.5], initial=0.5	[0, ∞]	gamma shape parameter
gene2.ac	* Gamma [0.05, 10], initial=1	[0, ∞]	GTR A-C substitution parameter
gene2.ag	* Gamma [0.05, 20], initial=1	[0, ∞]	GTR A-G substitution parameter
gene2.at	* Gamma [0.05, 10], initial=1	[0, ∞]	GTR A-T substitution parameter
gene2.cg	* Gamma [0.05, 10], initial=1	[0, ∞]	GTR C-G substitution parameter
gene2.gt	* Gamma [0.05, 10], initial=1	[0, ∞]	GTR G-T substitution parameter
gene2.frequencies	* Uniform [0, 1], initial=0.25	[0, 1]	base frequencies
gene2.alpha	* Exponential [0.5], initial=0.5	[0, ∞]	gamma shape parameter
ucld.stdev	* Exponential [0.333333], initial=0.3333...	[0, ∞]	uncorrelated lognormal relaxed clock stdev
ucld.mean	? Not yet specified, initial=1	[0, ∞]	uncorrelated lognormal relaxed clock mean
treeModel.rootHeight	* Using Tree Prior in [0, ∞]	[0, ∞]	root height of the tree
birthDeath.meanGrowthRate	* Uniform [0, 1E5], initial=23	[0, 1E5]	Birth-Death speciation process rate
birthDeath.relativeDeathRate	* Uniform [0, 1], initial=0.5	[0, 1]	Birth-Death speciation process relative death rate
meanRate	* Indirectly Specified Through Other Para...	n/a	The mean rate of evolution over the whole tree
covariance	* Indirectly Specified Through Other Para...	n/a	The covariance in rates of evolution on each lineage with their ance...
coefficientOfVariation	* Indirectly Specified Through Other Para...	n/a	The variation in rate of evolution over the whole tree

Figure 15: The statistics, parameters, and hyperparameters specific to this analysis and default priors.

- Click on the prior for `ucld.mean` and specify an ***Exponential*** distribution on this hyperparameter with a mean equal to `10.0`. [Figure 16]

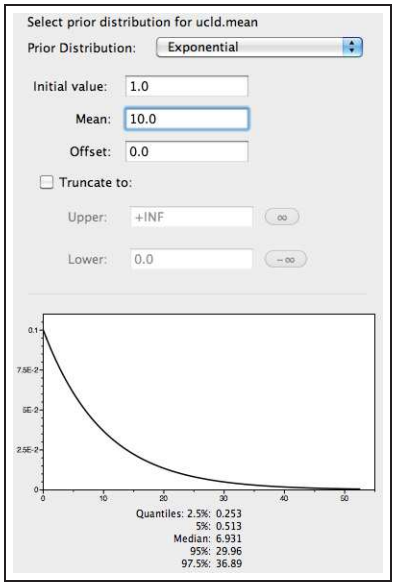


Figure 16: Exponential prior distribution on the `ucld.mean` hyperparameter.

Review the hyperparameters of the birth-death model. For this model the average net diversification rate is `meanGrowthRate` $= \lambda - \mu$ and the relative rate of extinction is `relativeDeathRate` $= \frac{\mu}{\lambda}$. Under the birth-death model implemented in BEAST, the net diversification rate (`meanGrowthRate`) must be greater than zero ($\mu < \lambda$). Therefore, the relative death rate can only take on values between 0 and 1. The default priors for these parameters are both uniform distributions.

Next, we can specify prior distributions on calibrated node ages. Each of the 4 calibrated nodes (Figure 5) requires a different type of prior distribution on their respective ages. Notice that the current prior for each of our calibrated nodes is set to: ***Using Tree Prior***. If we did not have constraints on the ages of the

clades defined in the ***Taxa*** menu, the times for these divergences will be sampled, just like all the others, with the prior being the birth-death model. By creating these `tmrca` parameters in the ***Taxa*** window, we have created a statistic that will be logged to our output file.

- Set a normal distribution on the age of the root. Click on the prior box for the `treeModel.rootHeight` parameter and choose ***Normal***. Parameterize the normal distribution so that the ***Mean*** and ***Initial Value*** are equal to `140` and the ***Stdev*** (standard deviation) is equal to `10`. For this prior, the value is truncated so that the age of the root cannot be less than 0. Leave the box indicating a truncated normal distribution checked. The normal distribution is not always appropriate for representing fossil age constraints, but is useful for imposing a prior requiring soft minimum and maximum bounds. Typically, this type of prior on the calibrated node age is based on biogeographical data or when a secondary calibration date is used. [Figure 17]

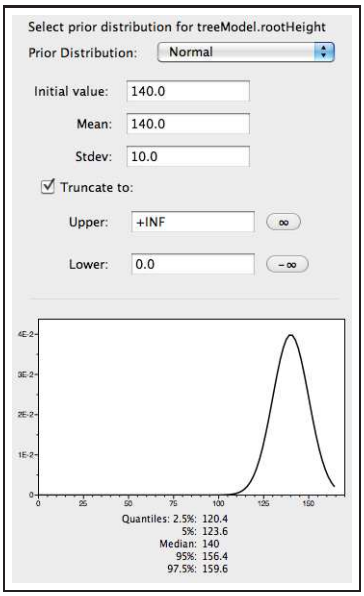


Figure 17: A normal prior distribution on the age of the root.

A fossil age constraint is available for the most recent common ancestor of our ingroup, `mrca1` (`T1, T2, T3, T4, T5, T6`). This minimum age estimate will serve as a hard lower bound on the age that node. Thus, the prior distribution on this node will be offset by the age of the fossil (`60`). When applying offset priors on nodes, it is perhaps easiest to consider the fact that the distribution is modeling the time difference between the age of the calibrated node and the age of the fossil (see Figure 3). We are applying an exponential prior on the age of `mrca1`. The exponential distribution is characterized by a single rate parameter (λ). The expected (mean) age difference under this distribution is equal to λ^{-1} and the median is equal to $\lambda^{-1} * ln(2)$. Specifying the exponential prior on a node age in BEAST requires that you set the *expected* age difference between the node and fossil (***Mean***) and the hard lower bound (***Offset***).

- Click on the prior box next to `tmrca(mrca1)` and set the prior distribution on the age of `mrca1` to an ***Exponential*** with a ***Mean*** equal to `30` and ***Offset*** equal to `60`. The quantiles for this prior are provided below the settings. And you can see that with an offset of 60 My, 97.5% of the prior probability will be below 170.67 My. [Figure 18]

The fossil calibration for the MRCA of the outgroup (`T7, T8, T9, T10`) provides a minimum age bound for the age of `mrca2`. The log-normal distribution is often used to describe the age of an ancestral node

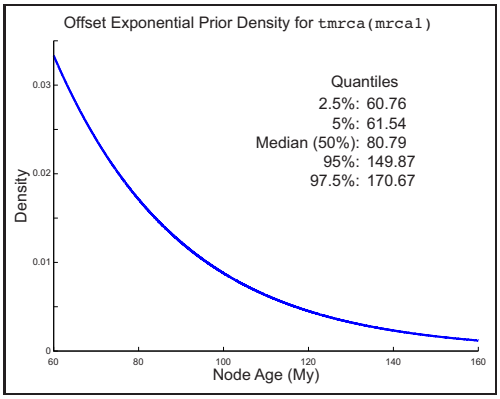


Figure 18: An offset exponential prior on the calibration node `mrca1`.

in relation to a fossil descendant. If a random variable, χ, is log-normally distributed with parameters μ and σ: $\chi \sim \text{LN}(\mu, \sigma)$, then the natural log of that variable is normally distributed with a mean of μ and standard deviation of σ: $\log(\chi) \sim \text{Norm}(\mu, \sigma)$. The median of the lognormal distribution is equal to e^{μ} and the mean (or expectation) is:

$$\mathbb{E}(\chi) = e^{\mu + \frac{\sigma^2}{2}}.$$

When applying the lognormal offset prior distribution on node age in BEAST, first consider the expected age difference between the MRCA and the fossil. Generally, it is difficult to know with any certainty the time lag between the speciation event and the appearance of the fossil and, typically, it is preferable to specify prior densities that are not overly informative (Heath, 2012).

For the prior density on `mrca2`, we will specify a lognormal prior density with an expected value equal to 20. Thus, if we want the expectation of the lognormal distribution to equal 20, we must determine the value for μ using the equation above solve for μ:

$$\begin{aligned} \mu &= \ln(20) - \frac{0.75^2}{2} \\ &= 2.714482, \end{aligned}$$

where `0.75` is the standard deviation parameter of the lognormal distribution for this particular fossil calibration.

- Select the prior box next to `tmrca(mrca2)` and specify a ***Lognormal*** prior distribution for `mrca2` and set ***Log(Mean)*** to $\mu = 2.714482$. The age of the fossil is `40` time units; use this date to set the ***Offset*** for the lognormal prior distribution. Finally, set the ***Log(Stdev)*** to `0.75`, so that 97.5% of the prior probability is below `105.7`. [Figure 19A]

Notice that the window for the ***Lognormal*** prior distribution allows you to specify ***Mean in Real Space***. If you choose this option, then the mean value you enter is the expected value of the lognormal distribution. You will specify the exact same distribution as above if you set the ***Mean*** to `20` while this option is selected. [Figure 19B]

It is important that you are very careful when specifying these parameters. If, for example, ***Mean in Real Space*** was checked and you provided a value of `2.714482` for the ***Mean***, then your calibration prior density would be very informative. In the case of the prior on `tmrca(mrca2)`, this would place 95.7% of the prior density below 47.03566.

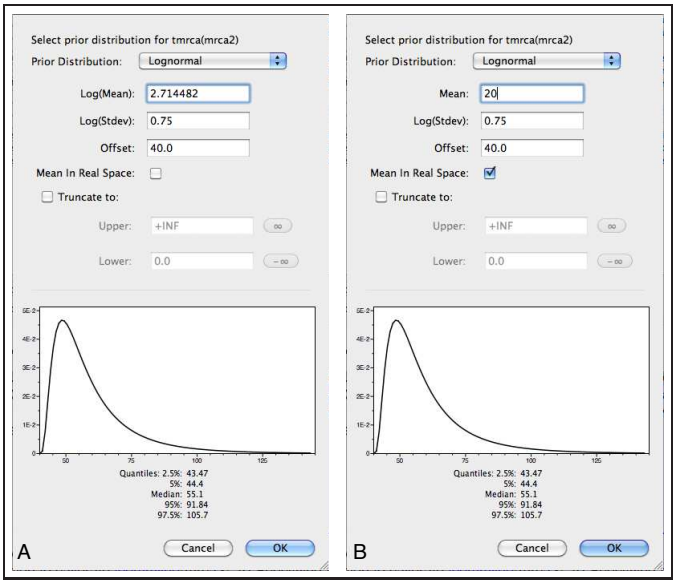


Figure 19: A log-normal prior distribution to calibrate `mrca2`. A) Specifying the ***Log(Mean)***. B) Selecting ***Mean in Real Space***. A bug in BEAUti v1.7.0 causes the plot of the prior density to disappear when you specify an ***Offset*** value.

The node leading to taxa `T9` and `T10` will be calibrated using a simple uniform distribution. For this calibration we will be placing a hard minimum and a hard maximum bound on the age of `mrca3`. Fossil data typically provide us with reliable minimum constraints on the age of a clade. And although absolute maximum bounds are very difficult to obtain, uniform priors with hard lower and upper age constraints are often used for divergence time estimation. Some methods using phylogenetic bracketing and stratigraphic bounding have been developed for determining possible maximum bounds (Benton and Donoghue, 2007), though it is best to apply soft age maxima when using these methods.

- Select the prior box next to `tmrca(mrca3)` and set the ***Uniform*** prior distribution on the time of `mrca3` with a ***Lower*** limit of `8` and an ***Upper*** limit of `32`. [Figure 20]

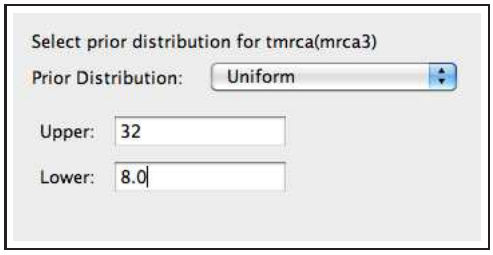


Figure 20: A uniform prior distribution on the age of `mrca3` with hard lower and upper bounds.

The remaining unmodified priors in the ***Priors*** window can be left at their default values for this exercise.

The ***Operators*** menu contains a list of the parameters and hyperparameters that will be sampled over the course of the MCMC run. In this window, it is possible to turn off any of the elements listed to fix a given parameter to its starting value. For example, if you would like to estimate divergence times on a fixed tree topology (using a starting tree that you provided), then disable proposals operating on the `Tree`. For this exercise, leave this window unmodified.

Now that you have specified all of your data elements, models, priors, and operators, go to the ***MCMC*** tab to set the length of the Markov chain, sample frequency, and file names. By default, BEAST sets the number of generations to 10,000,000.

- Since we have a limited amount of time for this exercise, change the ***Length of chain*** to 1,000,000. (Runtimes may vary depending on your computer, if you have reason to believe that this may take a very long time, then change the run length to something smaller.)

The frequency parameters are sampled and logged to file can be altered in the box labeled ***Log parameters every***. In general, this value should be set relative to the length of the chain. If a low value is specified, the output files containing the parameter values and trees will be very large, possibly without gaining much additional information. Conversely, if you specify an exceedingly large sample interval, then you will not get enough information about the posterior distributions of your parameters.

- Change ***Log parameters every*** to `100`.
- The frequency states are echoed to the screen is simply for observing the progress of your run, so set this to a satisfactory value (such as `1000`), keeping in mind that writing to the screen too frequently can cause the program to run slower. Specify the output file name prefix `bodega` in ***File name stem***.

Now you are ready to generate your BEAST XML file!

- Click on the button in the lower right corner: ☼ ***Generate BEAST File...***

You will see a window that will allow you to see the priors that you have not reviewed. This window also warns you about improper priors.

- It is safe to proceed, so click ***Continue*** and save your XML file with the name `bodega.xml`.

For the last step in BEAUti, generate an XML file that will run the analysis on an "empty" dataset and sample only from the prior. This allows us to evaluate the priors we have applied to the various parameters and hyperparameters in our analysis. This will produce output files that we can use to visualize the priors and compare them to our posterior estimates.

- Check the box next to ***Sample from prior only - create empty alignment***. Be sure to change the ***File name stem*** to `bodega.prior` and the XML file name to `bodega.prior.xml` after you click ☼ ***Generate BEAST File...***, so that you do not over-write your analysis files. Then close BEAUti.

Making changes in the XML file

BEAUti is a great tool for generating a properly-formatted XML file for many types of BEAST analyses. However, you may encounter errors that require modifying elements in your input file and if you wish to make small to moderate changes to your analysis, altering the input file is far less tedious than generating a new one using BEAUti. Furthermore, BEAST is a rich program and all of the types of analyses, models, and parameters available in the core cannot be specified using BEAUti. Thus, some understanding of the BEAST XML format is essential.

If you attempted to execute `bodega.xml` in BEAST right now, the analysis would terminate due to an error resulting from issues with the specification of the truncated normal prior on the root height in BEAUti. For other datasets, you may run into problems if you use a randomly-generated starting tree that is not compatible with calibrations on nodes.

- Open the `bodega.xml` file generated by BEAUti in your text editor and glance over the contents. BEAUti provides many comments describing each of the elements in the file.

As you look over the contents of this file, you will notice that the components are specified in an order similar to the steps you took in BEAUti. The XML syntax is very verbose. This feature makes it fairly easy to understand the different elements of the BEAST input file. If you wished to alter your analysis or realized that you misspecified a prior parameter, changing the XML file is far simpler than going through all of the steps in BEAUti again. For example, if you wanted to change bounds of the uniform prior on `mrca3` from (8, 32) to (12, 30), this can be done easily by altering these values in the XML file (Box 4), though leave these at 8 and 32 for this exercise.

```
<uniformPrior lower="8.0" upper="32.0">
    <statistic idref="tmrca(mrca3)">
</uniformPrior>
```

Box 4: The XML syntax for specifying a uniform prior distribution on `tmrca(mrca3)`. Changing the parameters (highlighted) of this prior is simply done by altering the XML file.

For our analysis, there is one setting causes an error when this XML file is executed in BEAST (Box 5). This is a problem with the truncated prior on the age of the root node.

```
Error parsing '<uniformPrior>' element with id, 'null':
Uniform prior uniformPrior cannot take a bound at infinity, because it returns 1/(high-low) = 1/inf
```

Box 5: BEAST error from truncated normal prior.

- We can easily correct this problem if we edit the XML file. Find the priors specified for the parameter called `treeModel.rootHeight` in your XML file. These priors are specified in the section delineated by the comment: `Define MCMC` (Box 6).

```
<uniformPrior lower="0.0" upper="Infinity">
    <parameter idref="treeModel.rootHeight">
</uniformPrior>
<normalPrior mean="140.0" stdev="10.0">
    <parameter idref="treeModel.rootHeight">
</normalPrior>
```

Box 6: The truncated normal prior distribution on the `treeModel.rootHeight`. The error results from the upper bound on the `uniformPrior`.

- Because the improper uniform prior doesn't integrate to 1.0, BEAST returns an error. This can be fixed by either giving the `uniformPrior` an upper bound equal to some arbitrarily chosen, high value (change `Infinity` to `100000.0`) or by deleting the syntax describing the `uniformPrior` on `treeModel.rootHeight` altogether (delete everything highlighted in yellow in Box 6). Since the height of the root node is always bounded by the ages of its descendant nodes, the age is already truncated (Box 6).

- Make these changes for both of the XML files created in BEAUti (`bodega.xml` and `bodega.prior.xml`). Look over the elements specified in each of the XML files and verify that everything is satisfactory. Save and close the input files.

Although running multiple, independent analyses is an important part of any Bayesian analysis, BEAST does not do this by default. However, setting up multiple runs is trivial once you have a complete XML file in hand and only requires that you make a copy of the input file and alter the names of the output files in the XML (it's also best to change the initial states for all of your parameters, including the starting tree).

Running BEAST

Now you are ready to start your BEAST analysis. BEAST allows you to use the BEAGLE library if you already have it installed. BEAGLE is an application programming interface and library that effectively takes advantage of your computer hardware (CPUs and GPUs) for doing the heavy lifting (likelihood calculation) needed for statistical phylogenetic inference (Ayers et al., 2012). Particularly, when using BEAGLE's GPU (NVIDIA) implementation, runtimes are significantly shorter.

- Execute `bodega.prior.xml` and `bodega.xml` in BEAST. You should see the screen output every 1,000 generations, reporting the likelihood and several other statistics.

- Once you have verified that your XML file was properly configured and you see the likelihood update, feel free to kill the run. I have provided the output files for this analysis and you can find them in `bodega_beast/output*`.

Summarizing the output

Once the run reaches the end of the chain, you will find three new files in your analysis directory. The MCMC samples of various scalar parameters and statistics are written to the file called `bodega.log`. The tree-state at every sampled iteration is saved to `bodega.trees`. The tree strings in this file are all annotated in extended Newick format with the substitution rate from the uncorrelated lognormal model at each node. The files called `bodega.ops` and `mcmc.operators` (these files may be identical) summarize the performance of each of the proposal mechanisms (operators) used in your analysis. Reviewing this file can help identify operators that might need adjustment if their acceptance probabilities are too high.

The main output files are the `.log` file and `.trees` file. It is not feasible to review the data contained in these files by simply opening them in a spreadsheet program or a tree viewing program. Fortunately, the developers of BEAST have also written general utility programs for summarizing and visualizing posterior samples from Bayesian inference using MCMC. Tracer is a cross-platform, java program for summarizing posterior samples of scaler parameters. This program is necessary for assessing convergence, mixing, and determining an adequate burn-in. Tree topologies, branch rates, and node heights are summarized using the program TreeAnnotator and visualized in FigTree.

Tracer

This section will briefly cover using Tracer and visual inspection of the analysis output for MCMC convergence diagnostics.

- Open Tracer and import the `bodega.log` file in the ***File→Import New Trace File***.

You will notice, in the ***Estimates*** tab, that many items in the ***ESS*** column are red. The MCMC runs you have performed today are all far too short to produce adequate posterior estimates of divergence times

and substitution model parameters and this is reflected in the ESS values. The ESS is the *effective sample size* of a parameter. The value indicates the number of effectively independent draws from the posterior in the sample. This statistic can help to identify autocorrelation in your samples that might result from poor mixing. It is important that you run your chains long enough and sufficiently sample the stationary distribution so that the ESS values of your parameters are all high (over 200 or so).

- Click on a parameter with a low ESS and explore the various windows in Tracer. It is clear that we must run the MCMC chain longer to get good estimates.

Provided with the files for this exercise are the output files from analyses run for 100,000,000 iterations. These files can be found in the `bodega_beast/output2/` directory and are all labeled with the file stem: `bodega_100m*`.

- Close `bodega.log` in Tracer using the [-] button and open `bodega_100m.1.log`, `bodega_100m.2.log` and `bodega_100m.prior.log`.

These log files are from much longer runs and since we ran two independent, identical analyses, we can compare the log files in Tracer and determine if they have converged on the same stationary distribution. Additionally, analyzing an empty alignment allows you to compare your posterior estimates to the prior distributions used for each parameter.

- Select and highlight all three files (`bodega_100m.1.log`, `bodega_100m.2.log` and `bodega_100m.prior.log`) in the ***Trace Files*** pane (do not include ***Combined***). This allows you to compare all three runs simultaneously. Click on the various parameters and view how they differ in their estimates and 95% HPDs for those parameters.

- Find the parameter `ucld.stdev` and compare the estimates of the standard deviation of the uncorrelated log-normal distribution.

The `ucld.stdev` indicates the amount of variation in the substitution rates across branches. Our prior on this parameter is an exponential distribution with $\lambda = 3.0$ ($mean = 0.33333$). Thus, there is a considerable amount of prior weight on `ucld.stdev` $= 0$. A standard deviation of 0 indicates support for no variation in substitution rates and the presence of a molecular clock.

- With `ucld.stdev` highlighted for all three runs, go to the ***Marginal Density*** window, which allows you to compare the marginal posterior densities for each parameter.

- Color (or "colour") the densities by ***Trace File*** next to ***Colour by*** at the bottom of the window (if you do not see this option, increase the size of your Tracer window). You can also add a ***Legend*** to reveal which density belongs to which run file. [Figure 21]

The first thing you will notice from this plot is that the marginal densities from each of our analysis runs (`bodega_100m.1.log` and `bodega_100m.2.log`) are nearly identical. If you click through the other sampled parameters, these densities are the same for each one. This indicates that both of our runs have converged on the same stationary distribution. Since some of the other parameters might not have mixed well, we may want to run them longer, but we can have good confidence that our runs have sampled the same distribution.

Second, notice how the marginal densities for the `ucld.stdev` parameter from each of the analysis runs are quite different from the marginal density of that parameter sampled from the prior. The signal in the

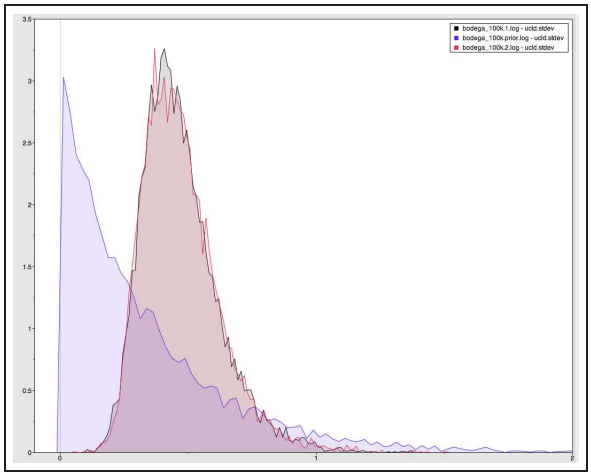


Figure 21: Comparing the marginal densities of the `ucld.stdev` parameter from 2 independent runs (red and gray) and the prior (blue) in Tracer.

data is not overwhelmed by the prior on this parameter. Moreover, our analysis runs do not have any significant density at zero, indicating no support for a constant rate of substitution (e.g. strict molecular clock). This is also evident if you view the 95% credible intervals (95% HPD) for each of the runs. When the analyses are run with data, a `ucld.stdev` of 0 does not fall within the credible interval.

When calibrating divergence time estimates using off-set parametric prior densities, it is *very* important to evaluate (and report) the marginal densities of both the prior and posterior samples of calibrated node heights.

- Highlight only the trace file containing MCMC samples under the prior (`bodega_100m.prior.log`). Then inspect the ***Marginal Density*** for each one. The prior densities are quite close to the calibration priors we specified in BEAUti and described in Figure 5.

- Specifically, look at the marginal prior density of `tmrca(mrca1)`. The calibration prior assigned to this node was an exponential density with a rate (or λ) equal to $\frac{1}{30}$. We expect the marginal density of the age of that node sampled under the prior to match the expected prior density. In Figure 22, you can see that the marginal prior density of the age of `mrca1` (purple line) closely fits an exponential distribution with a rate equal to $\frac{1}{30}$ (black line).

Because the two analysis runs (`bodega_100m.1.log` and `bodega_100m.2.log`) sampled from the same posterior distribution, particularly for the parameters we are interested in, they can be combined. Tracer does this when you import files and you will see a file called ***Combined*** in the ***Trace Files*** window. To use this option, however, you must first remove the prior trace file.

- Highlight only the prior file (`bodega_100m.prior.log`) in the ***Trace Files*** pane. Click the `-` button below the window to remove the file.

- Now highlight the ***Combined*** trace file. Navigate through the sampled parameters and notice how the ESSs have improved.

Continue examining the options in Tracer. This program is very useful for exploring many aspects of your analysis.

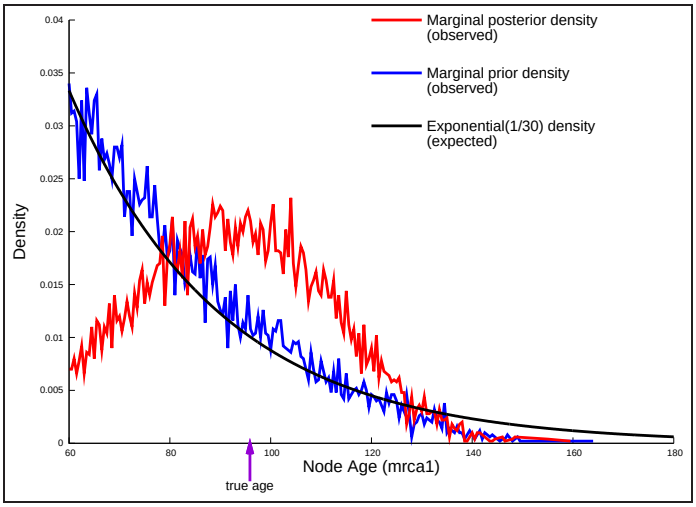


Figure 22: Comparing the marginal density of `tmrca(mrca1)` (purple line) to the exponential prior density specified for that calibration node (black line). (This figure was generated using `gnuplot` and can also be made in `R`.)

Summarizing the trees

After reviewing the trace files from the two independent runs in Tracer and verifying that both runs converged on the posterior distributions and reached stationarity, we can combine the sampled trees into a single tree file and summarize the results.

- Open the program LogCombiner and set the ***File type*** to ***Tree Files***. Next, import the two `trees` files in the `bodega_beast/output3/` directory (`bodega_100m.1.trees` and `bodega_100m.2.trees`) using the [+] button. [Figure 23]

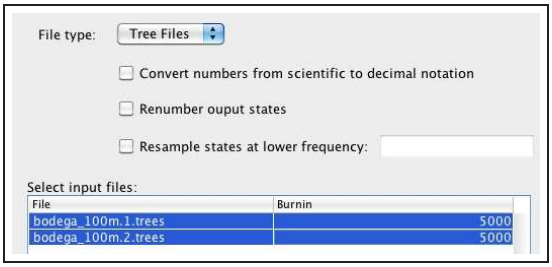


Figure 23: Use LogCombiner to combine the trees from two independent, identical runs.

These analyses were each run for 100,000,000 iterations with a sample frequency of 10,000. Therefore, each of the trees files contains 100,000 trees.

- Set a burn-in value of 2,500, thus discarding the first 25% of the samples in each tree file. Then click on the ***Choose file ...*** button to create an output file (call it `bodega.comb.trees`) and run the program. [Figure 23]

- Alternatively, use LogCombiner in unix with the command:

```
> logcombiner -trees -burnin 5000 bodega_100m.1.trees bodega_100m.2.trees bodega.comb.trees
```

Once LogCombiner has terminated, you will have a file containing 10,000 trees called `bodega.comb.trees` which can be summarized using TreeAnnotator. TreeAnnotator takes a collection of trees and summarizes

them by identifying the topology with the best support, calculating clade posterior probabilities, and calculating 95% HPD intervals for node-specific parameters. All of the node statistics are annotated on the tree topology for each node in the Newick string.

- Open the program TreeAnnotator. Since we already discarded a set of burn-in trees when combining the tree files, we can leave ***Burnin*** set to `0`. [Figure 24]

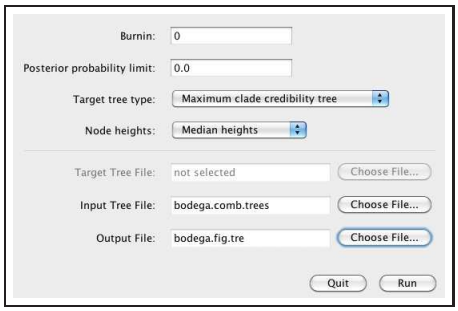


Figure 24: Set up tree summary in TreeAnnotator.

- For the ***Target tree type***, choose ***Maximum clade credibility tree***.

The ***Maximum clade credibility tree*** is the topology with the highest product of clade posterior probabilities across all nodes. Alternatively, you can select the ***Maximum sum of clade credibilities*** which sums all of the clade posteriors. Or you can provide a target tree from file.

The ***Posterior probability limit*** option applies to summaries on a target tree topology and only calculates posteriors for nodes that are above the specified limit.

- Choose ***Median heights*** or ***Mean heights*** for ***Node heights*** which will set the node heights of the output tree to equal the median or mean height for each node in the sample of trees.

- Choose `bodega.comb.trees` as your ***Input Tree File***. Then name the ***Output File:*** `bodega.fig.tre` and click ***Run***. [Figure 24]

Once the program has finished running, you will find the file `bodega.fig.tre` in your directory.

- Open `bodega.fig.tre` in your text editor. The tree is written in NEXUS format. Look at the tree string and notice the annotation. Each node in the tree is labeled with comments using the `[&parameter_name=<value>]` format.

An alternative program to LogCombiner and TreeAnnotator is **SumTrees**, a program in the **DendroPy** package (Sukumaran and Holder, 2010). SumTrees is very flexible and allows more options than TreeAnnotator and it provides a way to summarize sets of trees from a number of different programs and analyses.

The tree summaries produced by TreeAnnotator or SumTrees can be opened with FigTree. FigTree reads the comments for each each node and can display them in a variety of ways.

- Open FigTree and open the file `bodega.fig.tre`.

FigTree is a great tree-viewing program and it also allows you to produce publication-quality tree figures. This summary tree is shown in Figure 25. The posterior probabilities are labeled on each branch (they are all equal to 1). The branches of the tree are colored by the average substitution rate estimated for each lineage. On each node, bars are displayed representing the node age 95% HPD interval.

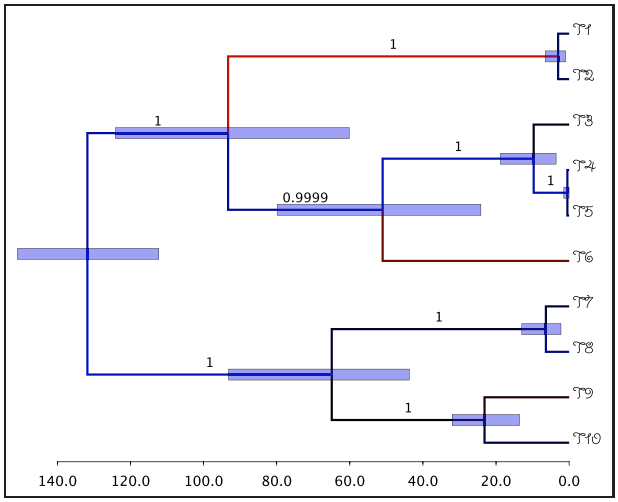


Figure 25: The final tree with node bars representing the age 95% HPD intervals and posterior probabilities of each clade labeled on the branches. Each branch is colored according to the average substitution rate sampled by the MCMC chain.

- Explore the various options for creating a tree figure and recreate the tree in Figure 25. The figure you create can be exported as a PDF or EPS file.

Divergence time estimation for this simulated data set is very straight-forward. To generate the sequences for this exercise I first simulated a tree topology and divergence times under a constant-rate birth-death process with 10 extant taxa (`T1, T2, T3, T4, T5, T6, T7, T8, T9, T10`). The simulated time-tree is shown in Figure 26A.

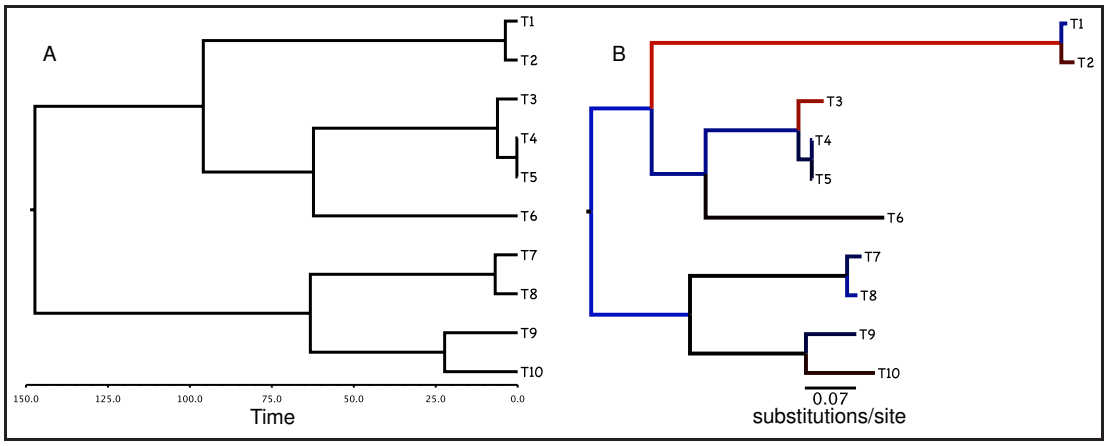


Figure 26: A) The true tree and branching times. The age of the root is equal to 147.28. B) The true tree and branch lengths in units of $rate * time$. The branches are colored according to their substitution rate.

With a tree topology and branch lengths in units of time, I simulated lineage-specific substitution rate variation under an uncorrelated model such that the rate associated with each branch was drawn from a log-normal distribution (Figure 26B). The tree with branch lengths in units of $rate * time$ was then used to simulate two separate genes each under a different GTR+Γ model. For the most part, the models assumed in this analysis matched the models used to generate the data, therefore our estimates of divergence time and tree topology are quite accurate. We inferred the correct tree topology and the true divergence times all fall within the node age 95% credible intervals.

Useful Links

- BEAST website and documentation: http://beast.bio.ed.ac.uk
- BEAST open source project: http://code.google.com/p/beast-mcmc
- Join the BEAST user discussion: http://groups.google.com/group/beast-users
- BEAST2 (under development, a complete rewrite of BEAST1): http://beast2.cs.auckland.ac.nz
- MrBayes: http://mrbayes.sourceforge.net/
- DPPDiv: http://phylo.bio.ku.edu/content/tracy-heath-dppdiv
- PhyloBayes: www.phylobayes.org/
- multidivtime: http://statgen.ncsu.edu/thorne/multidivtime.html
- MCMCtree (PAML): http://abacus.gene.ucl.ac.uk/software/paml.html
- BEAGLE: http://code.google.com/p/beagle-lib/
- A list of programs: http://evolution.genetics.washington.edu/phylip/software.html#Stratigraphy
- The Paleobiology Database: http://www.paleodb.org
- The Fossil Record & Date A Clade: http://www.fossilrecord.net

Questions about this tutorial can be directed to Tracy Heath (email: tracyh@berkeley.edu).

Relevant References

Ayers DL, Darling A, Zwickl DJ, et al. (12 co-authors). 2012. BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. Systematic Biology. 61:170–173.

Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P. 2013. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. Molecular Biology and Evolution. 30:239–243.

Benton MJ, Ayala FJ. 2003. Dating the tree of life. Science. 300:1698–1700.

Benton MJ, Donoghue PCJ. 2007. Paleontological evidence to date the tree of life. Molecular Biology and Evolution. 24:26–53.

Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology. 4:e88.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology. 7:214.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biology. 8:114.

Gaut BA, Weir BS. 1994. Detecting substitution-rate heterogeneity among regions of a nucleotide sequence. Molecular Biology and Evolution. 11:620–629.

Gernhard T. 2008. The conditioned reconstructed process. Journal of Theoretical Biology. 253:769–778.

Graur D, Martin W. 2004. Reading the entrails of chickens: Molecular timescales of evolution and the illusion of precision. Trends in Genetics. 20:80–86.

Hasegawa M, Kishino H, Yano T. 1989. Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea. Journal of Human Evolution. 18:461–476.

Heath TA. 2012. A hierarchical Bayesian model for calibrating estimates of species divergence times. Systematic Biology. 61:793–809.

Heath TA, Holder MT, Huelsenbeck JP. 2012. A Dirichlet process prior for estimating lineage-specific substitution rates. Molecular Biology and Evolution. 29:939–255.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution. 27:570–580.

Heled J, Drummond AJ. 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. Systematic Biology. 61:138–149.

Ho SYW. 2007. Calibrating molecular estimates of substitution rates and divergence times in birds. Journal of Avian Biology. 38:409–414.

Ho SYW, Phillips MJ. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Systematic Biology. 58:367–380.

Huelsenbeck JP, Larget B, Swofford DL. 2000. A compound Poisson process for relaxing the molecular clock. Genetics. 154:1879–1892.

Hug LA, Roger AJ. 2007. The impact of fossils and taxon sampling on ancient molecular dating analyses. Molecular Biology and Evolution. 24:1889–1897.

Hugall AF, Foster R, Lee MSY. 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. Systematic Biology. 56:543–63.

Inoue J, Donoghue PC, Yang Z. 2010. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. Systematic Biology. 59:74–89.

Kendall DG. 1948. On the generalized "birth-and-death" process. Annals of Mathematical Statistics. 19:1–15.

Kishino H, Hasegawa M. 1990. Converting distance to time: Application to human evolution. Methods in Enzymology. 183:550–570.

Kishino H, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and th origin of chloroplasts. Journal of Molecular Evolution. 31:151–160.

Kishino H, Thorne JL, Bruno W. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Molecular Biology and Evolution. 18:352–361.

Lee MSY, Oliver PM, Hutchinson MN. 2009. Phylogenetic uncertainty and molecular clock calibrations: A case study of legless lizards (Pygopodidae, Gekkota). Molecular Phylogenetics and Evolution. 50:661–666.

Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots. PLoS Computational Biology. 5:e1000520.

Lepage T, Bryant D, Philippe H, Lartillot N. 2007. A general comparison of relaxed molecular clock models. Molecular Biology and Evolution. 24:2669–2680.

Lepage T, Lawi S, Tupper P, Bryant D. 2006. Continuous and tractable models for the variation of evolutionary rates. Mathematical Biosciences. 199:216–233.

Li WLS, Drummond AJ. 2012. Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. Molecular Biology and Evolution. 29:751–761.

Magallón S. 2009. Using fossils to break long branches in molecular dating: A comparison of relaxed clocks applied to the origin of angiosperms. Systematic Biology. 59:384–399.

Marshall CR. 1990. Confidence intervals on stratigraphic ranges. Paleobiology. 16:1–10.

Marshall CR. 2008. A simple method for bracketing absolute divergence times on molecular phylogenies using multiple fossil calibration points. The American Naturalist. 171:726–742.

Muse SV, Weir BS. 1992. Testing for equality of evolutionary rates. Genetics. 132:269–276.

Nee S, May RM, Harvey PH. 1994. The reconstructed evolutionary process. Philosophical Transactions of the Royal Society B. 344:305–311.

Pyron RA. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Systematic Biology. 60:466–481.

Rambaut A, Bromham L. 1998. Estimating divergence dates from molecular sequences. Molecular Biology and Evolution. 15:442–448.

Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. Journal of Molecular Evolution. 43:304–311.

Rannala B, Yang Z. 2007. Inferring speciation times under an episodic molecular clock. Systematic Biology. 56:453–466.

Ronquist F, Klopfstein S, Vilhelmsen L, Schulmeister S, Murray DL, Rasnitsyn AP. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Systematic Biology. 61:973–999.

Rutschmann F, Eriksson T, Salim KA, Conti E. 2007. Assessing calibration uncertainty in molecular dating: the assignment of fossils to alternative calibration points. Systematic Biology. 56:591–608.

Sanderson MJ. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Molecular Biology and Evolution. 14:1218–1231.

Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Molecular Biology and Evolution. 19:101–109.

Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Molecular Biology and Evolution. 23:7–9.

Stadler T. 2009. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. Journal of Theoretical Biology. 261:58–66.

Stadler T. 2010. Sampling-through-time in birth-death trees. Journal of Theoretical Biology. 267:396–404.

Stadler T. 2011. Simulating trees on a fixed number of extant species. Systematic Biology. 60:668–675.

Sukumaran J, Holder MT. 2010. DendroPy: A Python library for phylogenetic computing. Bioinformatics. 26:1569–1571.

Swofford DL. 1998. PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods. Sunderland, Massachusetts: Sinauer Associates, Inc.

Thorne J, Kishino H. 2002. Divergence time and evolutionary rate estimation with multilocus data. Systematic Biology. 51:689–702.

Thorne J, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Molecular Biology and Evolution. 15:1647–1657.

Thorne JL, Kishino H. 2005. Estimation of divergence times from molecular sequence data. In: Nielsen R, editor, Statistical Methods in Molecular Evolution. New York: Springer, pp. 235–256.

Warnock RCM, Yang Z, Donoghue PCJ. 2012. Exploring the uncertainty in the calibration of the molecular clock. Biology Letters. 8:156–159.

Yang Z, Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Molecular Biology and Evolution. 14:717–724.

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Molecular Biology and Evolution. 23:212–226.

Yang Z, Yoder AD. 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Systematic Biology. 52:705–716.

Yoder AD, Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. Molecular Biology and Evolution. 17:1081–1090.

Zuckerkandl E, Pauling L. 1962. Molecular disease, evolution, and genetic heterogeneity. In: Kasha M, Pullman B, editors, Horizons in Biochemistry. Academic Press, New York, pp. 189–225.