Applied Bayesian Phylogenetics

Emily Jane McTavish

Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu, twitter:snacktavish

BL exercise review:

Split into groups. Go over your HW assignment together.

Did you get the same answers? Check each one and see if you can come to an agreement.

How would you extend the branch length MCMC exercise to a full phylogenetic analysis?

"Metropolis algorithm will produce a precise and accurate approximation of the posterior distribution if run long enough". - Paul Lewis

"Metropolis algorithm will produce a precise and accurate approximation of the posterior distribution if run long enough". - Paul Lewis

"People always forget how long of a time infinity really is" - paraphrased from Dave Swofford

So, what's all this got to do with phylogenetics?
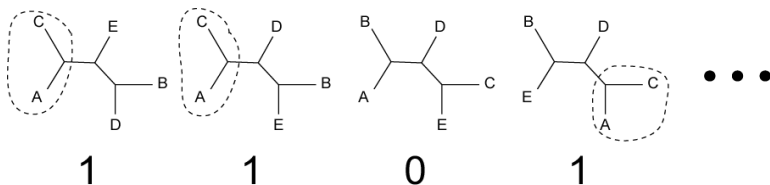
The posterior probability of the split AC|BDE may be approximated by the fraction of trees sampled from the posterior that contain that split.

How are these splits values this similar to those you get from bootstrapping?

How are they different?

While we often motivate Bayesian analysis by integrating prior information, setting up accurate, informative priors for phylogenetic inference is hard to do.

Common Priors

- **Discrete uniform** for topologies
  - exceptions becoming more common
- **Beta** for proportions
- **Gamma** or **Log-normal** for branch lengths and other parameters with support $[0,\infty)$
  - Exponential is common special case of the gamma distribution
- **Dirichlet** for state frequencies and GTR relative rates

Discrete Uniform distribution for **topologies**
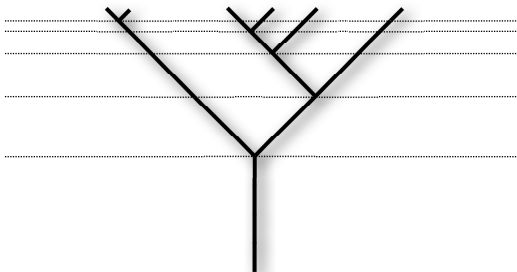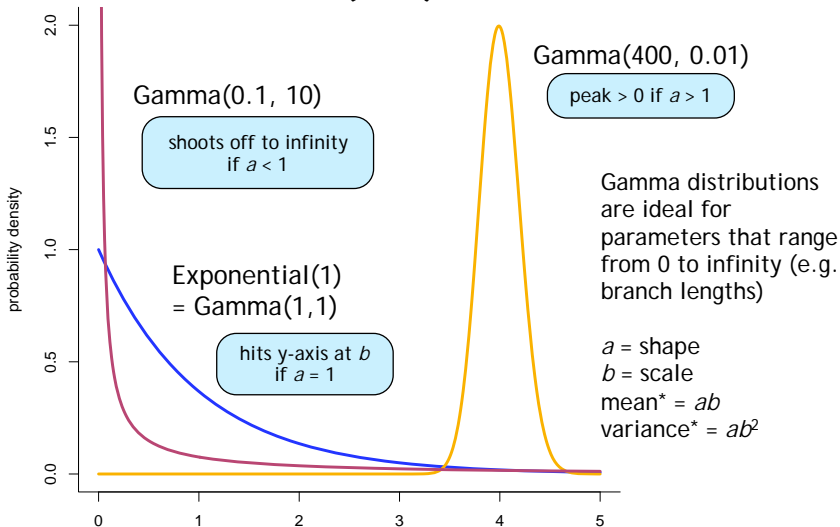
Yule model provides joint prior for both topology and divergence times

The rate of speciation under the Yule model (λ) is constant and applies equally and independently to each lineage. Thus, speciation events get closer together in time as the tree grows because more lineages are available to speciate.
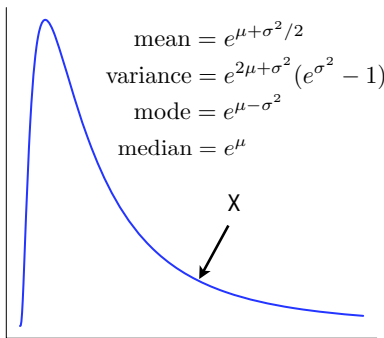
Gamma(a, b) distributions

Gamma(0.1, 10)

shoots off to infinity if $a < 1$

Gamma(400, 0.01)

peak > 0 if $a > 1$

Exponential(1) = Gamma(1,1)

hits y-axis at b if $a = 1$

Gamma distributions are ideal for parameters that range from 0 to infinity (e.g. branch lengths)

a = shape
b = scale
mean* = ab
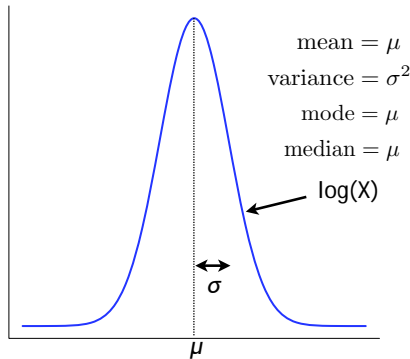variance* = ab^2

probability density

*Note: be aware that in many papers the Gamma distribution is defined such that the second (scale) parameter is the *inverse* of the value b used in this slide! In this case, the mean and variance would be a/b and a/b^2, respectively.

Log-normal distribution

If **X** is **log-normal** with *parameters* μ and σ…

$$\text{mean} = e^{\mu + \sigma^2/2}$$
$$\text{variance} = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)$$
$$\text{mode} = e^{\mu - \sigma^2}$$
$$\text{median} = e^{\mu}$$

X

…then **log(X)** is **normal** with *mean μ* and *standard deviation σ.*

$$\text{mean} = \mu$$
$$\text{variance} = \sigma^2$$
$$\text{mode} = \mu$$
$$\text{median} = \mu$$

log(X)

σ

μ

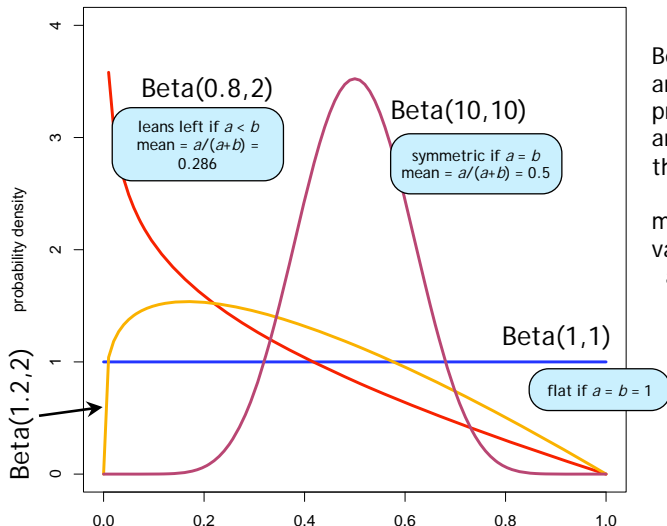*Important*: *μ and σ do **not** represent the mean and standard deviation of X: they are the mean and standard deviation of log(X)!*

To choose μ and σ to yield a particular mean (*m*) and variance (*v*) for X, use these formulas:
$$\mu = \log(m^2) - \log(m) - \frac{\log(v + m^2) - \log(m^2)}{2}$$
$$\sigma^2 = \log(v + m^2) - \log(m^2)$$

Beta(*a*, *b*) gallery

Beta(0.8,2)

leans left if *a* < *b*
mean = *a/(a+b)* =
0.286

Beta(10,10)

symmetric if *a* = *b*
mean = *a/(a+b)* = 0.5

Beta(1,1)

flat if *a* = *b* = 1

Beta(1.2,2)

probability density

Beta distributions are appropriate for proportions, which are constrained to the interval [0,1].

mean = a/(a+b)
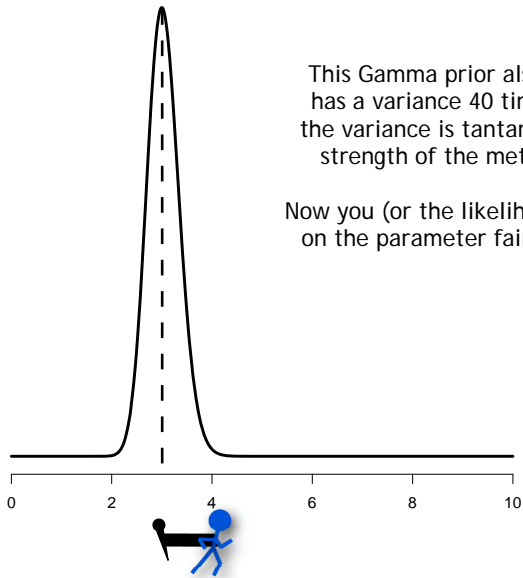variance =
  $ab/[(a+b)^2(a+b+1)]$

Prior Miscellany

- priors as rubber bands ◄
- running on empty
- hierarchical models
- empirical bayes
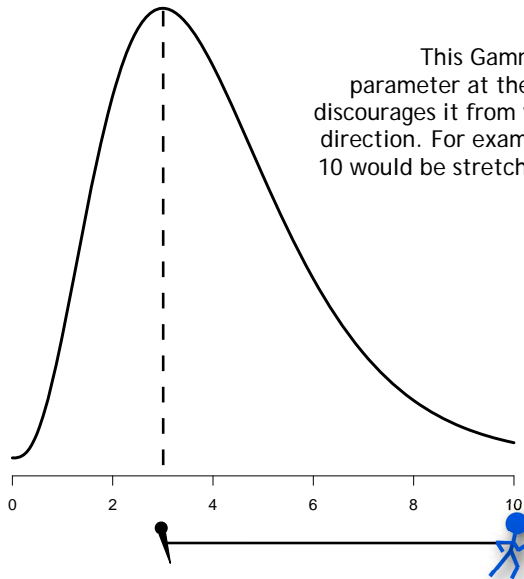
This Gamma(4,1) prior ties down its parameter at the mode, which is at 3, and discourages it from venturing too far in either direction. For example, a parameter value of 10 would be stretching the rubber band fairly tightly

The mode of a Gamma(a, b) distribution is (a-1)b (assuming $a > 1$)

0   2   4   6   8   10

This Gamma prior also has a mode at 3, but has a variance 40 times smaller. Decreasing the variance is tantamount to increasing the strength of the metaphorical rubber band.

Now you (or the likelihood) would have to tug on the parameter fairly hard for it to have a value as large as 4.

This gamma distribution has shape 91.989 and scale 0.032971
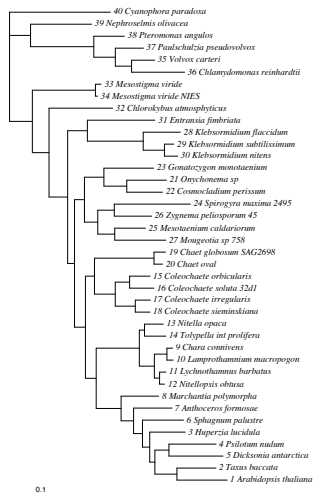
Example: Internal Branch Length Priors
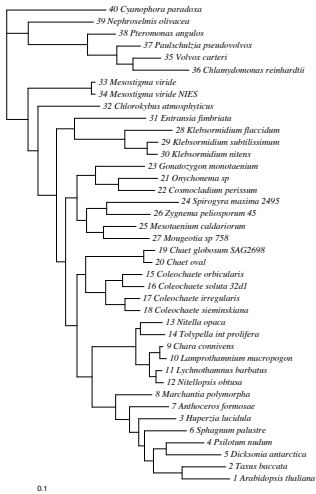
Separate priors applied to internal and external branches

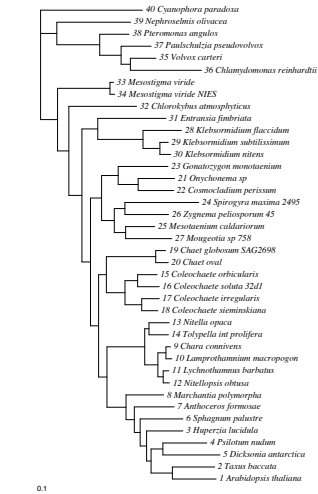External branch length prior is exponential with mean 0.1

Internal branch length prior is exponential with mean 0.1

This is a reasonably vague internal branch length prior

Internal branch length prior mean 0.01

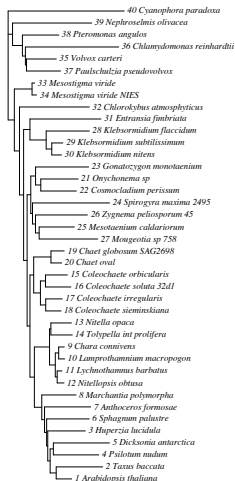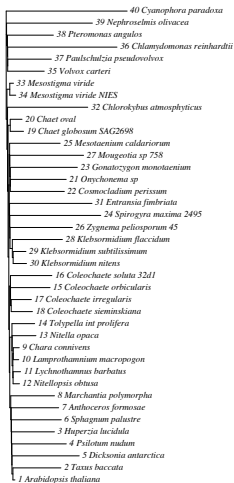(external branch length prior mean always 0.1)
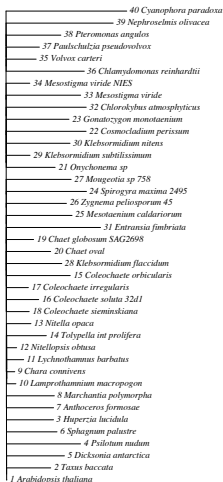
Internal branch length prior mean 0.001

Internal branch length prior mean
0.0001

Internal branch length prior mean 0.00001

Internal branch length prior mean 0.000001

The internal branch length prior is calling the shots now, and the likelihood must obey.

Prior Miscellany

- priors as rubber bands
- running on empty
- hierarchical models
- empirical bayes

Running on empty

Solid line: prior density estimated from MrBayes output

Dotted line: exponential(10) density for comparison

```
#NEXUS

begin data;
  Dimensions ntax=4 nchar=1;
  Format datatype=dna missing=?;
  matrix
    taxon1 ?
    taxon2 ?
    taxon3 ?
    taxon4 ?
  ;
end;

begin mrbayes;
  set autoclose=yes;
  lset rates=gamma;
  prset shapepr=exponential(10.0);
  mcmcp nruns=1 nchains=1 printfreq=1000;
  mcmc ngen=10000000 samplefreq=1000;
end;
```

> BEAST and MrBayes 3.2 both make it easy to ignore the data without having to resort to creating fake data sets

You can use the program Tracer to show the estimated density:
http://tree.bio.ed.ac.uk/software/tracer/
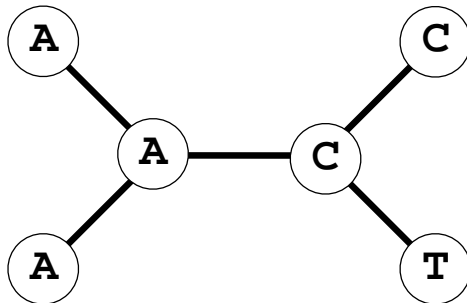
Prior Miscellany

- priors as rubber bands
- running on empty
- hierarchical models ◀
- empirical bayes

In a **non-hierarchical** model, all parameters are present in the likelihood function

Prior: Exponential, mean=0.1

$$L_k = \tfrac{1}{4}\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4v_1/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4v_2/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4v_3/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4v_4/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4v_5/3}\right]$$

Hierarchical models add *hyperparameters* not present in the likelihood function

μ is a *hyperparameter* governing the mean of the edge length prior

hyperprior

↓

Prior: Exponential, mean *μ*

$$L_k = \frac{1}{4}\left[\frac{1}{4}+\frac{3}{4}e^{-4v_1/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4v_2/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4v_3/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4v_4/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4v_5/3}\right]$$

During an MCMC analysis, *μ* will hover around a reasonable value, sparing you from having to decide what value is appropriate. You still have to specify a hyperprior, however.

For example, see Suchard, Weiss and Sinsheimer. 2001. MBE 18(6): 1001-1013.

Prior Miscellany

- priors as rubber bands
- running on empty
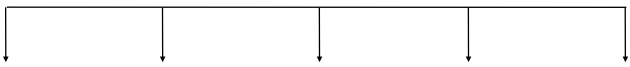- hierarchical models
- empirical bayes

Empirical Bayes

Empirical Bayes uses the data to determine some aspects of the prior, such as the prior mean.

Pure Bayesian approaches choose priors without reference to the data.

An empirical Bayesian would use the maximum likelihood estimate (MLE) of the length of an average branch here

Prior: Exponential, mean=MLE

$$L_k = \tfrac{1}{4} \left[\tfrac{1}{4} + \tfrac{3}{4} e^{-4v_1/3} \right] \left[\tfrac{1}{4} + \tfrac{3}{4} e^{-4v_2/3} \right] \left[\tfrac{1}{4} - \tfrac{1}{4} e^{-4v_3/3} \right] \left[\tfrac{1}{4} - \tfrac{1}{4} e^{-4v_4/3} \right] \left[\tfrac{1}{4} + \tfrac{3}{4} e^{-4v_5/3} \right]$$

There are several software packages to perform Bayesian phylogenetics.

MrBayes

MrBayes is a program for Bayesian inference and model choice across a wide range of phylogenetic and evolutionary models. MrBayes uses Markov chain Monte Carlo (MCMC) methods to estimate the posterior distribution of model parameters. Has mostly been replaced by RevBayes, but is still useful for some analyses.

`http://nbisweden.github.io/MrBayes/`

RevBayes

A fuller featured successor to MrBayes. RevBayes provides an interactive environment for statistical computation in phylogenetics. It is primarily intended for modeling, simulation, and Bayesian inference in evolutionary biology, particularly phylogenetics.

`https://revbayes.github.io/`

BEAST2
BEAST 2 is a cross-platform program for Bayesian phylogenetic analysis of molecular sequences. It estimates rooted, time-measured phylogenies using strict or relaxed molecular clock models. It can be used as a method of reconstructing phylogenies but is also a framework for testing evolutionary hypotheses without conditioning on a single tree topology. BEAST 2 uses Markov chain Monte Carlo (MCMC) to average over tree space, so that each tree is weighted proportional to its posterior probability. BEAST 2 includes a graphical user-interface for setting up standard analyses and a suit of programs for analysing the results.
`http://www.beast2.org/` Well developed set of tutorials at `https://taming-the-beast.org/`

MrBayes tutorial together on the cluster.
We will do a simplified version of Paul Lewis's tutorial at
`https://plewis.github.io/mrbayes/`