

3.36pt

Bayesian Inference and MCMC

Emily Jane McTavish

Life and Environmental Sciences
University of California, Merced

`ejmctavish@ucmerced.edu`, `twitter:snacktavish`

(With thanks to Mark Holder and Paul Lewis for slides!)

Bayes' rule in Statistics

$$\Pr(\theta|D) = \frac{\Pr(D|\theta) \Pr(\theta)}{\sum_{\theta} \Pr(D|\theta) \Pr(\theta)}$$

D refers to the "observables" (i.e. the **Data**)

θ refers to one or more "unobservables"

(i.e. **parameters** of a model, or the **model itself**):

- *tree model* (i.e. tree topology)
- *substitution model* (e.g. JC, F84, GTR, etc.)
- *parameter* of a substitution model (e.g. a branch length, a base frequency, transition/transversion rate ratio, etc.)
- *hypothesis* (i.e. a special case of a model)
- a *latent variable* (e.g. ancestral state)

Bayes' rule in statistics

The diagram illustrates Bayes' rule with the following components and arrows:

- Likelihood of hypothesis θ** : An arrow points to the blue box containing $\Pr(D|\theta)$.
- Prior probability of hypothesis θ** : An arrow points to the orange box containing $\Pr(\theta)$.
- Posterior probability of hypothesis θ** : An arrow points to the purple box containing $\Pr(\theta|D)$.
- Marginal probability of the data (marginalizing over hypotheses)**: An arrow points to the green box containing the denominator $\sum_{\theta} \Pr(D|\theta) \Pr(\theta)$.

$$\Pr(\theta|D) = \frac{\Pr(D|\theta) \Pr(\theta)}{\sum_{\theta} \Pr(D|\theta) \Pr(\theta)}$$

Simple (albeit silly) paternity example

θ_1 and θ_2 are assumed to be the only possible fathers, **child** has genotype **Aa**, **mother** has genotype **aa**, so child must have received allele **A** from the true father. Note: the **data** in this case is the child's genotype (**Aa**)

Possibilities	θ_1	θ_2	Row sum
Genotypes	AA	Aa	---
Prior	1/2	1/2	1
Likelihood	1	1/2	---
Prior X Likelihood	1/2	1/4	3/4
Posterior	2/3	1/3	1

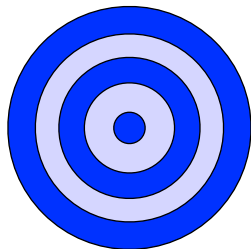
Bayes' rule: continuous case

The diagram illustrates Bayes' rule for the continuous case. The equation is presented with color-coded components and arrows indicating their roles:

- Likelihood:** An arrow points from the text "Likelihood" to the term $f(D|\theta)$ in the numerator, which is highlighted in a light blue box.
- Prior probability density:** An arrow points from the text "Prior probability density" to the term $f(\theta)$ in the numerator, which is highlighted in a light orange box.
- Posterior probability density:** An arrow points from the text "Posterior probability density" to the term $f(\theta|D)$ in the denominator, which is highlighted in a light purple box.
- Marginal probability of the data:** An arrow points from the text "Marginal probability of the data" to the integral term $\int f(D|\theta)f(\theta)d\theta$ in the denominator, which is highlighted in a light green box.

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{\int f(D|\theta)f(\theta)d\theta}$$

If you had to guess...



← 1 meter →

0.0

d

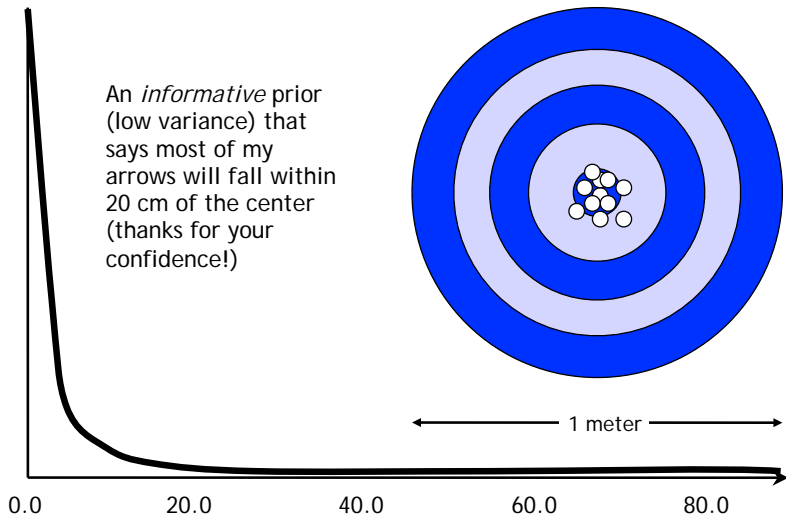
∞

Not knowing anything about my archery abilities, draw a curve representing your view of the chances of my arrow landing a distance d centimeters from the center of the target (if it helps, I'm standing 50 meters away from the target)

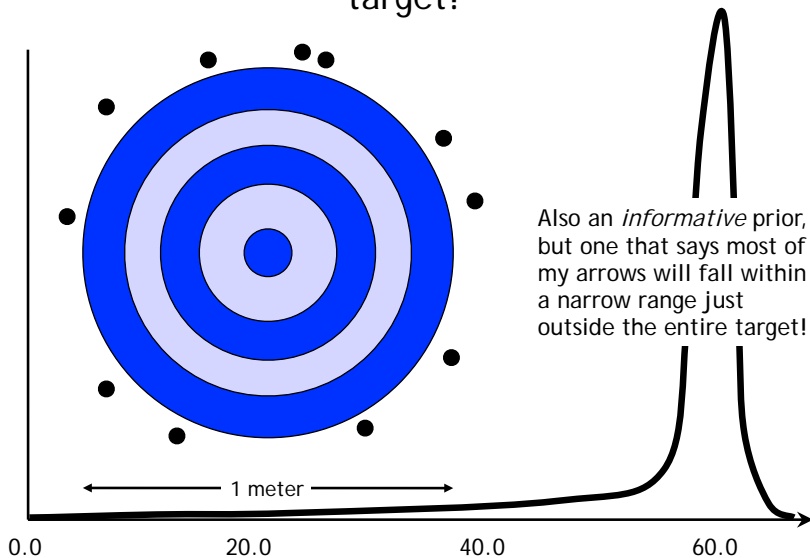


Photo by Tracy Heath

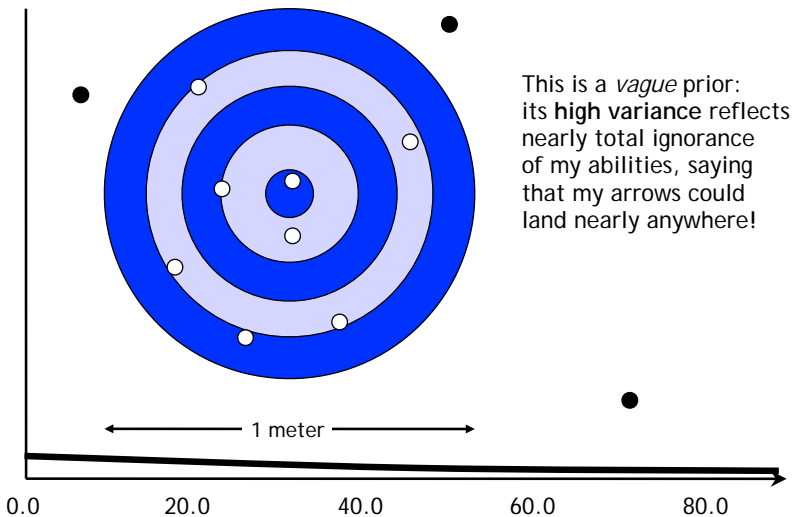
Case 1: assume I have talent



Case 2: assume I have a talent for missing the target!



Case 3: assume I have no talent



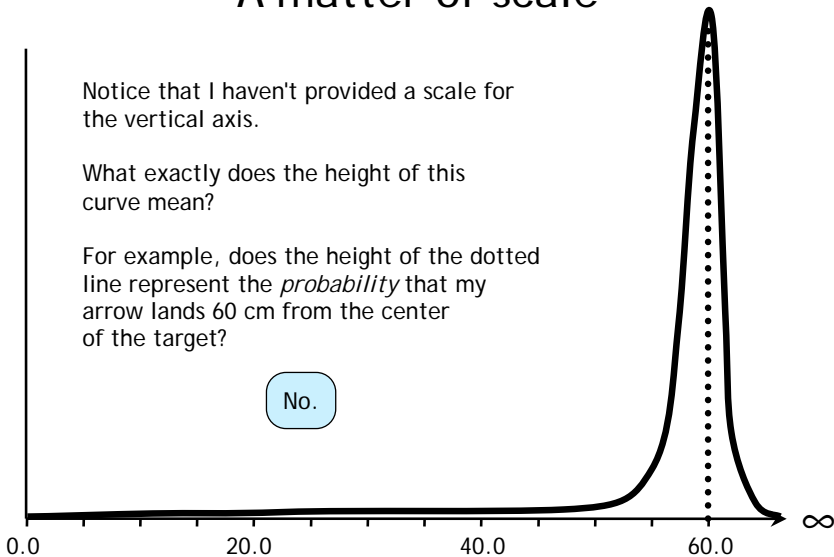
A matter of scale

Notice that I haven't provided a scale for the vertical axis.

What exactly does the height of this curve mean?

For example, does the height of the dotted line represent the *probability* that my arrow lands 60 cm from the center of the target?

No.

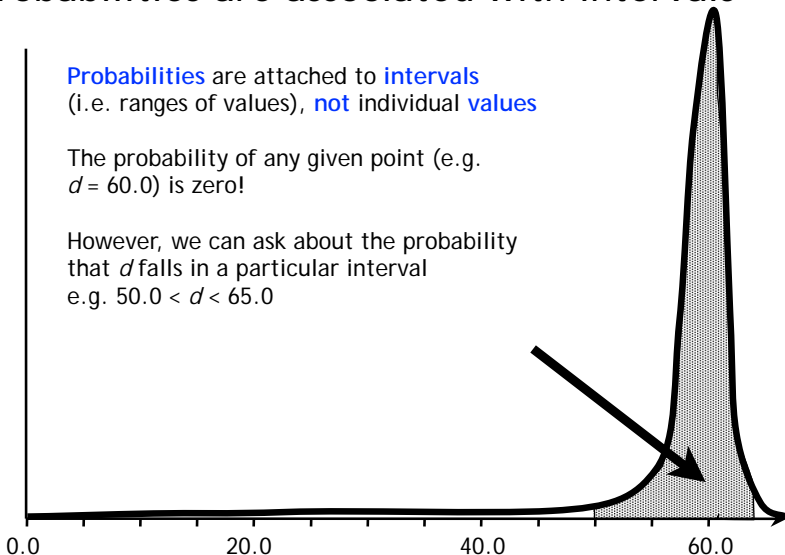


Probabilities are associated with intervals

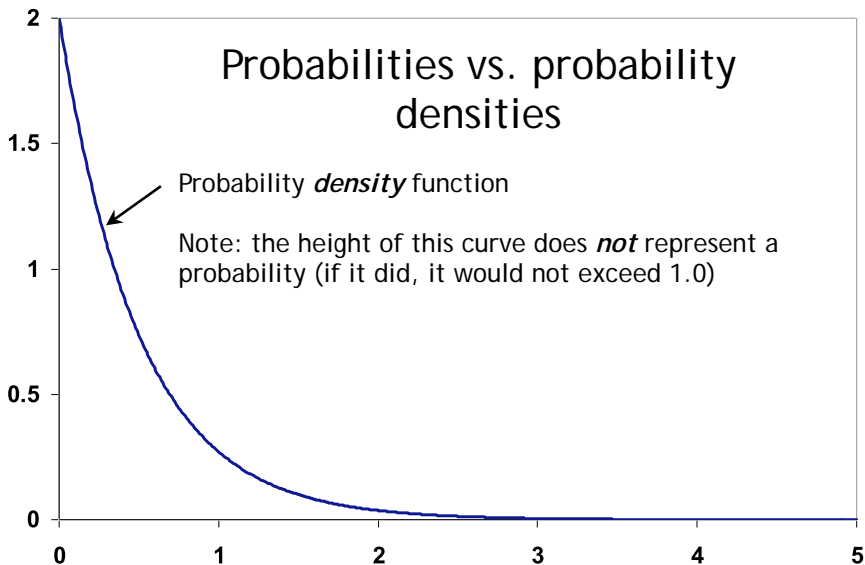
Probabilities are attached to **intervals**
(i.e. ranges of values), **not** individual **values**

The probability of any given point (e.g.
 $d = 60.0$) is zero!

However, we can ask about the probability
that d falls in a particular interval
e.g. $50.0 < d < 65.0$



Probabilities vs. probability densities



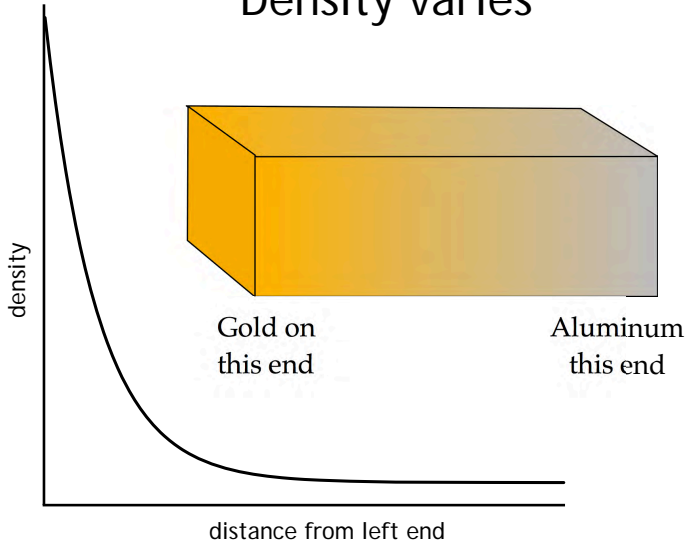
Densities of various substances

Substance	Density (g/cm ³)
Cork	0.24
Aluminum	2.7
Gold	19.3

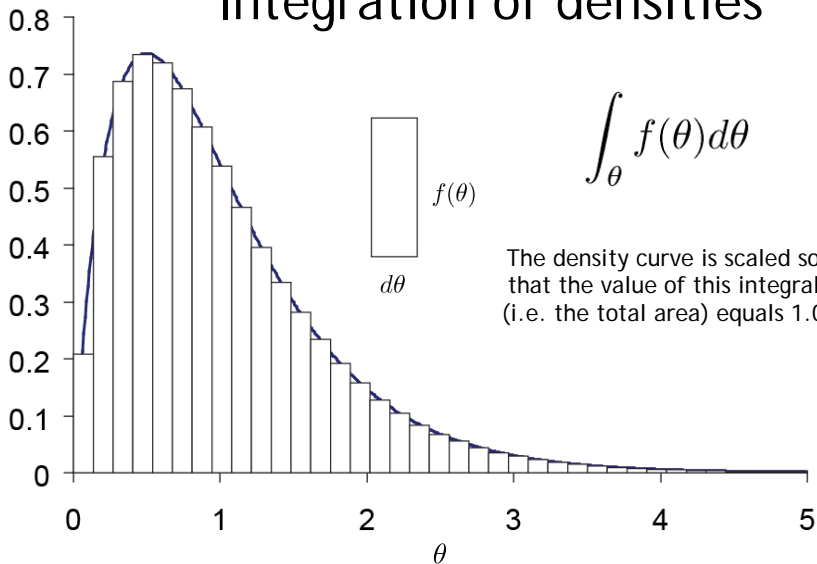
Density does not equal mass

$$\text{mass} = \text{density} \times \text{volume}$$

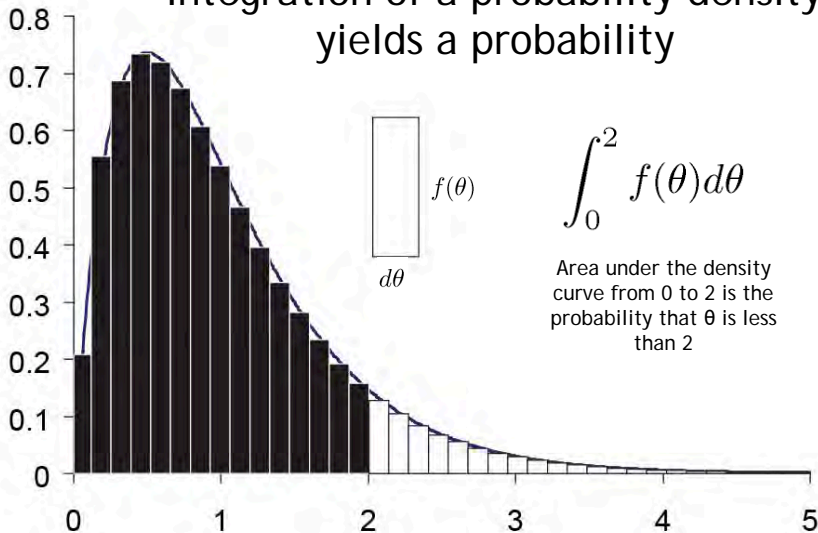
Density varies



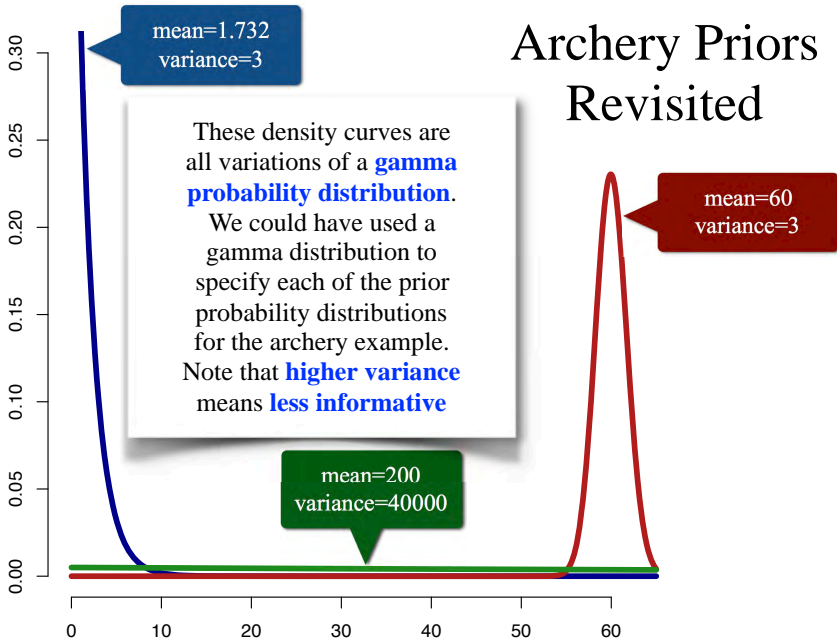
Integration of densities



Integration of a probability density yields a probability



Archery Priors Revisited



Usually there are many parameters...

A 2-parameter example

$$f(\theta, \phi | D) = \frac{f(D|\theta, \phi) f(\theta) f(\phi)}{\int_{\theta} \int_{\phi} f(D|\theta, \phi) f(\theta) f(\phi) d\theta d\phi}$$

Posterior probability density

Likelihood

Prior probability density

Marginal probability of data

An analysis of **100 sequences** under the simplest model (JC69) requires 197 branch length parameters. The denominator is a **197-fold integral** in this case! Now consider summing over **all possible tree topologies**! It would thus be nice to avoid having to calculate the marginal probability of the data...

Markov Chain Monte Carlo (MCMC)

- ▶ **Monte Carlo** - stochastic
- ▶ **Markovian** - present state depends only on the previous state and no other state

Markov chain Monte Carlo

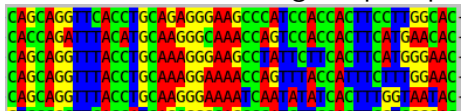
- Simulates a walk through parameter/tree space.
- Lets us estimate posterior probabilities for any aspect of the model
- Relies on the *ratio* of posterior densities between two points

$$R = \frac{Pr(\text{Point}_2|\text{Data})}{Pr(\text{Point}_1|\text{Data})}$$

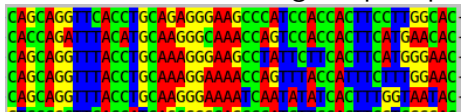
$$R = \frac{\frac{Pr(\text{Point}_2)L(\text{Point}_2)}{Pr(\text{Data})}}{\frac{Pr(\text{Point}_1)L(\text{Point}_1)}{Pr(\text{Data})}}$$

$$R = \frac{Pr(\text{Point}_2)L(\text{Point}_2)}{Pr(\text{Point}_1)L(\text{Point}_1)}$$

We can avoid calculating the prior probability of the data!

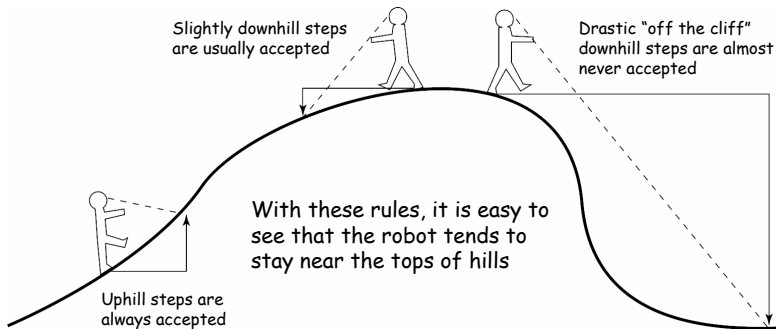
A 5x20 grid of DNA sequence data. Each row contains 20 nucleotides (A, C, G, T) separated by vertical bars. The background of each cell is color-coded: A is green, C is red, G is blue, and T is yellow. The sequence is as follows:
Row 1: CAGCAGGTTACCTGCAAGGGGAAGCCCATCCACCATTCTTGGCAC
Row 2: CACCAGATTTACATGCAAGGGCAAACAGTCCACCATTCTTGAACAC
Row 3: CAGCAGGTTTACCTGCAAAGGGGAAGCCTATTCTTCACTTTCATGGGAAC
Row 4: CAGCAGGTTTACCTGCAAAGGGAAAACAGTTTACCATTTCTTTGGAAC
Row 5: CAGCAGGTTTACCTGCAAGGGAAAAATCAATATATCACTTTGGTAATAC

We can avoid calculating the prior probability of the data!

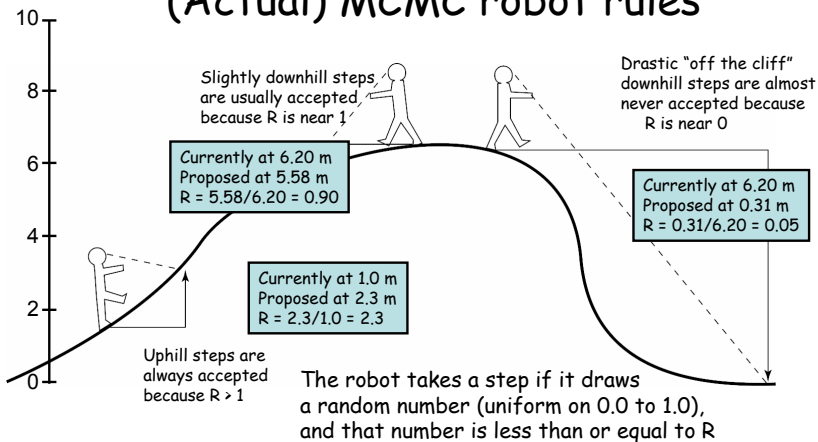


Imagine using a robot to survey a landscape...

MCMC robot's rules



(Actual) MCMC robot rules

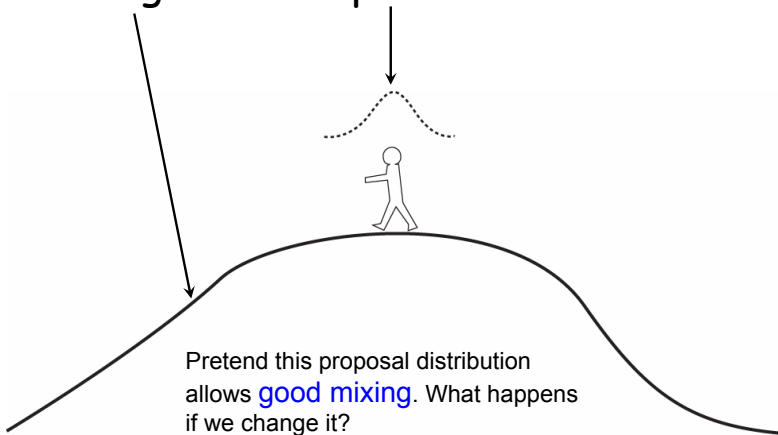


Target vs. proposal distributions

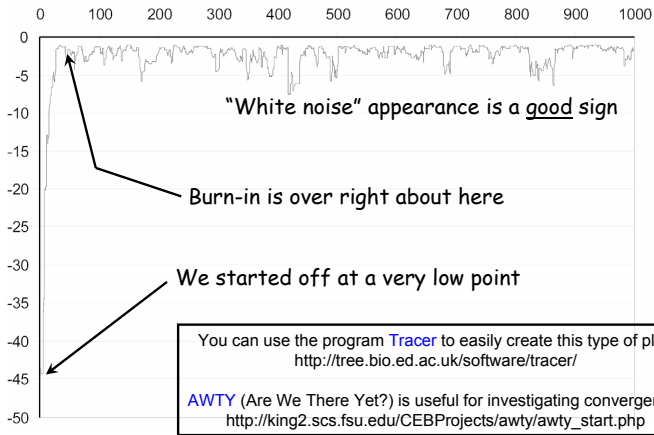
- The target distribution is the posterior distribution of interest
- The proposal distribution is used to decide which point to try next
 - you have much flexibility here, and the choice affects only the **efficiency** of the MCMC algorithm
 - MCMC using a **symmetric** proposal distribution is the Metropolis algorithm (Metropolis et al. 1953)
 - Use of an **asymmetric** proposal distribution requires a modification proposed by Hastings (1970), and is known as the Metropolis-Hastings algorithm

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087-1092.

Target vs. Proposal Distributions



Trace plots



Target vs. Proposal Distributions

Proposal distributions
with **smaller variance**...



Disadvantage: robot takes smaller steps, more time required to explore the same area



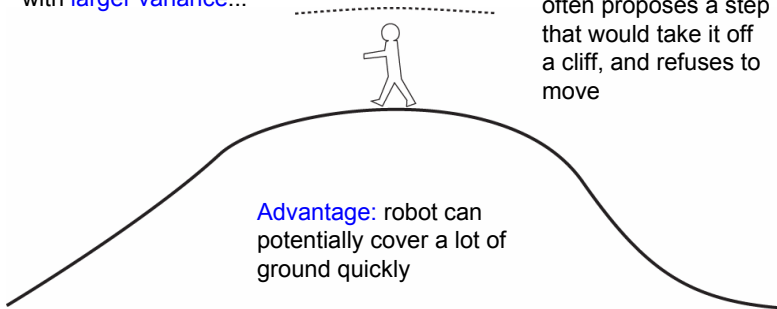
Advantage: robot seldom refuses to take proposed steps

Target vs. Proposal Distributions

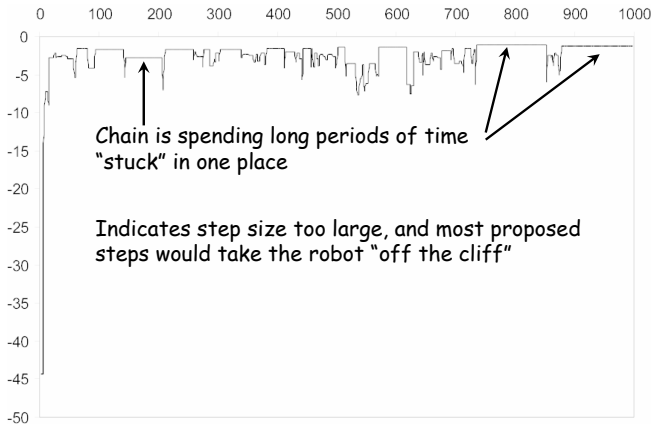
Proposal distributions
with **larger variance**...

Disadvantage: robot
often proposes a step
that would take it off
a cliff, and refuses to
move

Advantage: robot can
potentially cover a lot of
ground quickly

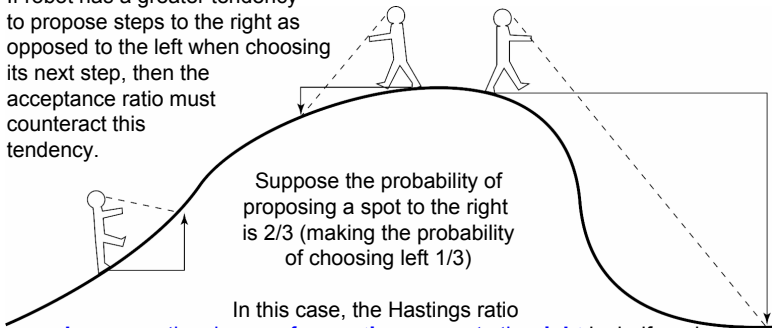


Poor mixing



The Hastings ratio

If robot has a greater tendency to propose steps to the right as opposed to the left when choosing its next step, then the acceptance ratio must counteract this tendency.



In this case, the Hastings ratio **decreases the chance of accepting moves to the right** by half, and **increases the chance of accepting moves to the left** (by a factor of 2), thus **exactly compensating** for the asymmetry in the proposal distribution.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97-109.

Paul Lewis' MCMC robot demo

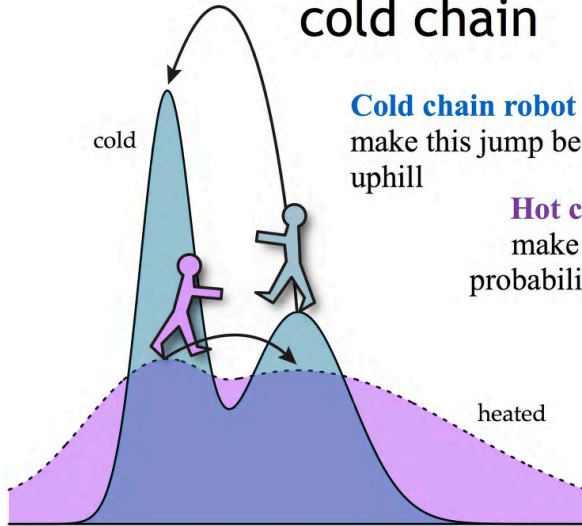
<http://phylogeny.uconn.edu/mcmc-robot/>

Metropolis-coupled Markov chain Monte Carlo (MCMCMC)

- MCMCMC involves running **several chains simultaneously**
- The **cold chain** is the one that counts, the rest are **heated chains**
- Chain is heated by raising densities to a power less than 1.0 (values closer to 0.0 are warmer)

Geyer, C. J. 1991. Markov chain Monte Carlo maximum likelihood for dependent data. Pages 156-163 *in* Computing Science and Statistics (E. Keramidas, ed.).

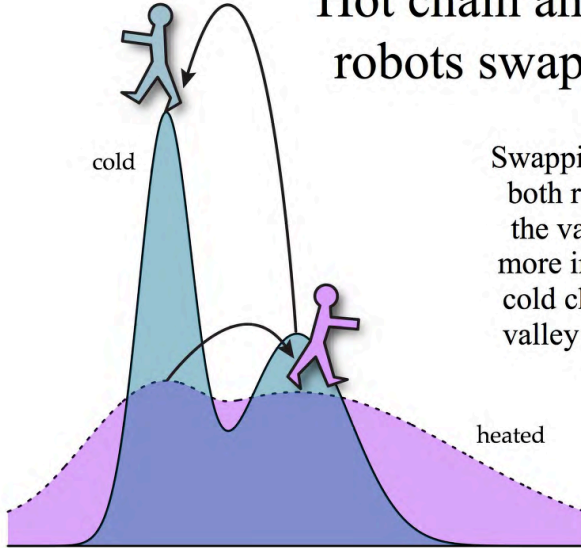
Heated chains act as scouts for the cold chain



Cold chain robot can easily make this jump because it is uphill

Hot chain robot can also make this jump with high probability because it is only slightly downhill

Hot chain and cold chain robots swapping places



Swapping places means both robots can cross the valley, but this is more important for the cold chain because its valley is much deeper

Back to MCRobot...

Paul Lewis' MCMC robot demo, with poor mixing
<http://phylogeny.uconn.edu/mcmc-robot/>

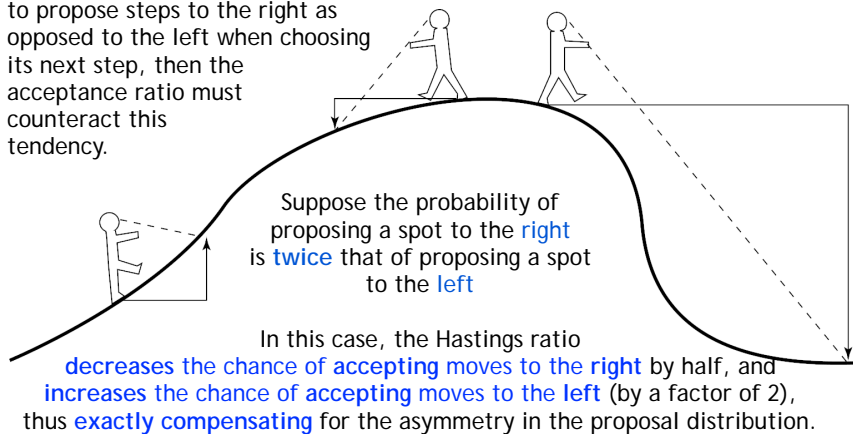
“Metropolis algorithm will produce a precise and accurate approximation of the posterior distribution if run long enough”. - Paul Lewis

“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

The Hastings ratio

If robot has a greater tendency to propose steps to the right as opposed to the left when choosing its next step, then the acceptance ratio must counteract this tendency.

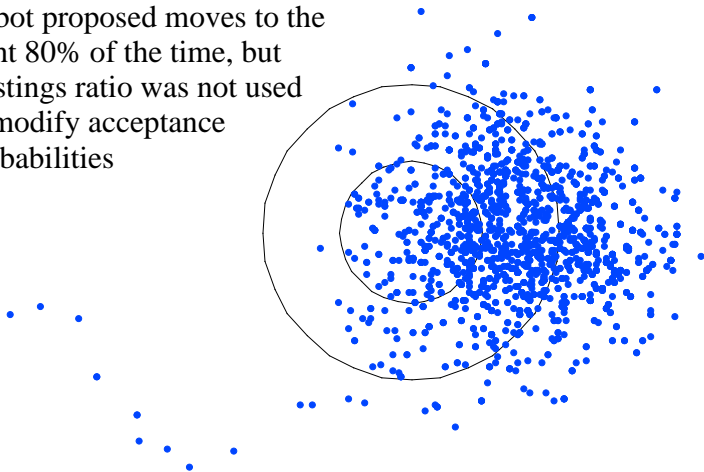


Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97-109.

The Hastings ratio

Example where MCMC

Robot proposed moves to the right 80% of the time, but Hastings ratio was not used to modify acceptance probabilities



Hastings Ratio

$$R = \left[\frac{f(D|\theta^*) f(\theta^*)}{f(D|\theta) f(\theta)} \right] \left[\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} \right]$$

Acceptance
ratio

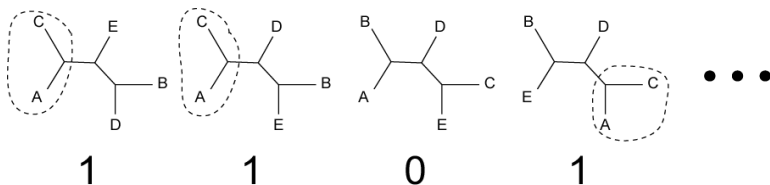
Posterior ratio

Hastings ratio

Note that if $q(\theta|\theta^*) = q(\theta^*|\theta)$, the Hastings ratio is 1

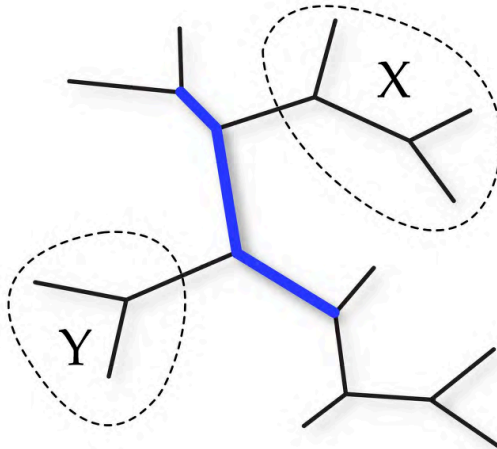
III. Bayesian phylogenetics

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.

Moving through treespace



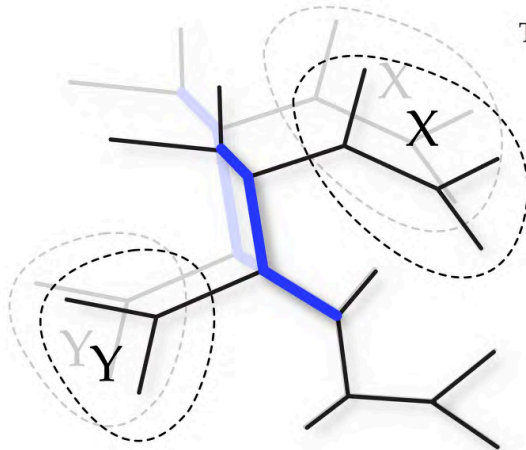
The Target-Simon move

Step 1:

Pick 3 contiguous edges randomly, defining two subtrees, X and Y

*Larget, B., and D. L. Simon. 1999. Markov chain monte carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* 16: 750-759. See also: Holder et al. 2005. *Syst. Biol.* 54: 961-965.

Moving through treespace



The Target-Simon move

Step 1:

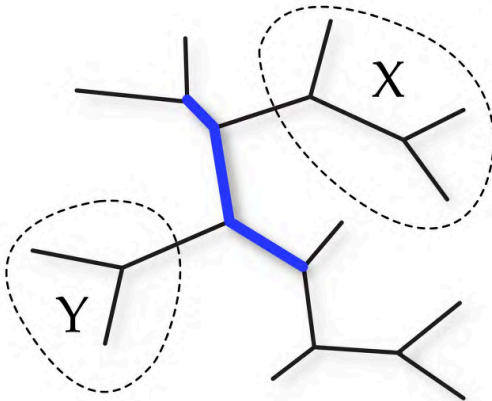
Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Step 2:

Shrink or grow selected 3-edge segment by a random amount

Moving through treespace

The Target-Simon move



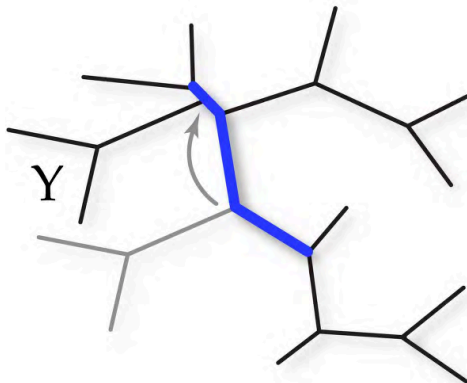
Step 1:

Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Step 2:

Shrink or grow selected 3-edge segment by a random amount

Moving through treespace



The Target-Simon move

Step 1:

Pick 3 contiguous edges randomly, defining two subtrees, X and Y

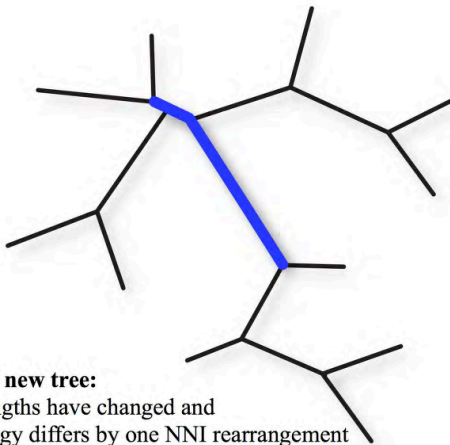
Step 2:

Shrink or grow selected 3-edge segment by a random amount

Step 3:

Choose X or Y randomly, then reposition randomly

Moving through treespace



Proposed new tree:
3 edge lengths have changed and
the topology differs by one NNI rearrangement

The Target-Simon move

Step 1:

Pick 3 contiguous edges
randomly, defining two
subtrees, X and Y

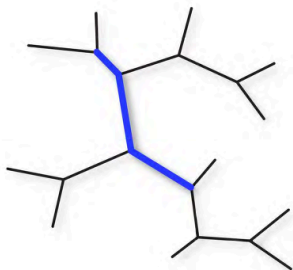
Step 2:

Shrink or grow selected
3-edge segment by a
random amount

Step 3:

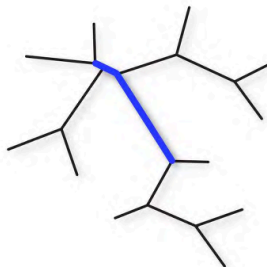
Choose X or Y randomly,
then reposition
randomly

Moving through treespace



Current tree

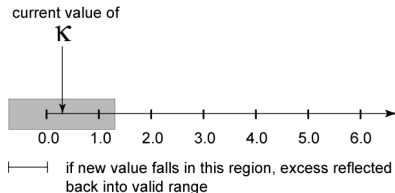
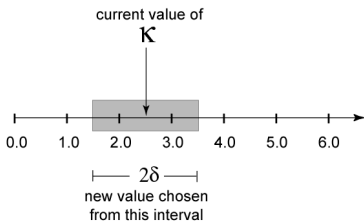
log-posterior = -34256



Proposed tree

log-posterior = -32519
(better, so accept)

Moving through parameter space



Using κ (ratio of the transition rate to the transversion rate) as an example of a model parameter.

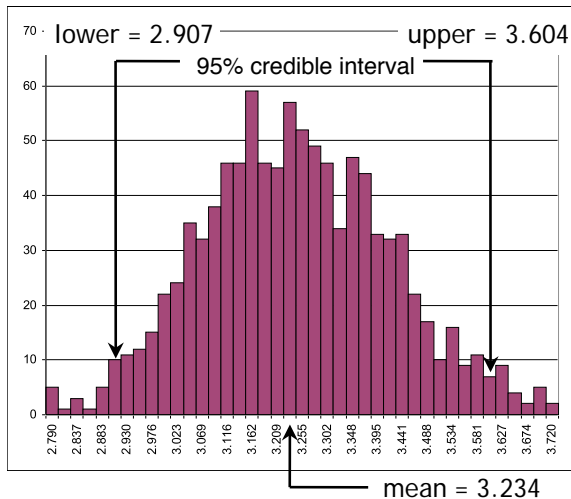
Proposal distribution is the uniform distribution on the interval $(\kappa - \delta, \kappa + \delta)$

The “step size” of the MCMC robot is defined by δ : a larger δ means that the robot will attempt to make larger jumps on average.

Putting it all together

- **Start with** random tree and arbitrary initial values for branch lengths and model parameters
- **Each generation** consists of one of these (chosen at random):
 - Propose a **new tree** (e.g. Target-Simon move) and either accept or reject the move
 - Propose (and either accept or reject) a **new model parameter value**
- Every k generations, save tree topology, branch lengths and all model parameters (i.e. **sample the chain**)
- After n generations, **summarize sample** using histograms, means, credible intervals, etc.

Marginal Posterior Distribution of κ



Histogram created from a sample of 1000 kappa values.

IV. Prior distributions

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

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



$$\frac{1}{15}$$



$$\frac{1}{15}$$



$$\frac{1}{15}$$



$$\frac{1}{15}$$



$$\frac{1}{15}$$



$$\frac{1}{15}$$



$$\frac{1}{15}$$



$$\frac{1}{15}$$



$$\frac{1}{15}$$



$$\frac{1}{15}$$



$$\frac{1}{15}$$



$$\frac{1}{15}$$



$$\frac{1}{15}$$

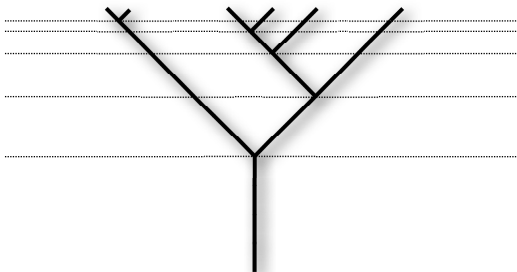


$$\frac{1}{15}$$



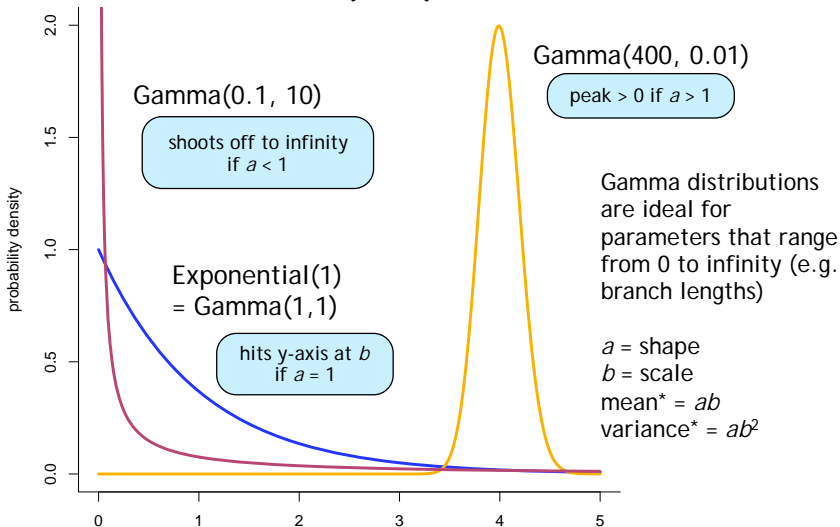
$$\frac{1}{15}$$

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.

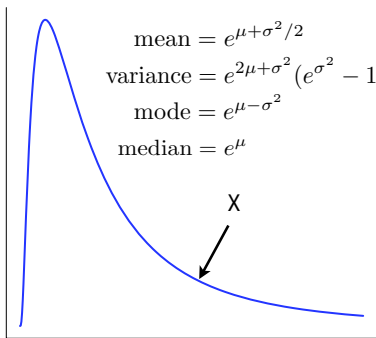
Gamma(a, b) distributions



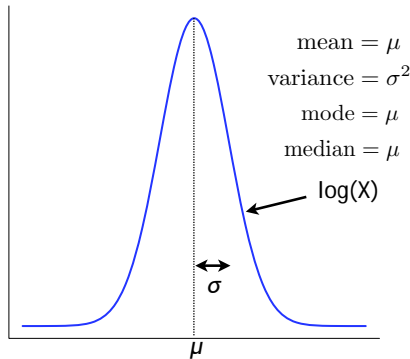
*Note: be aware that in many papers the Gamma distribution is defined such that the second (scale) parameter is the *inverse* of the value b used in this slide! In this case, the mean and variance would be a/b and a/b^2 , respectively.

Log-normal distribution

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



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



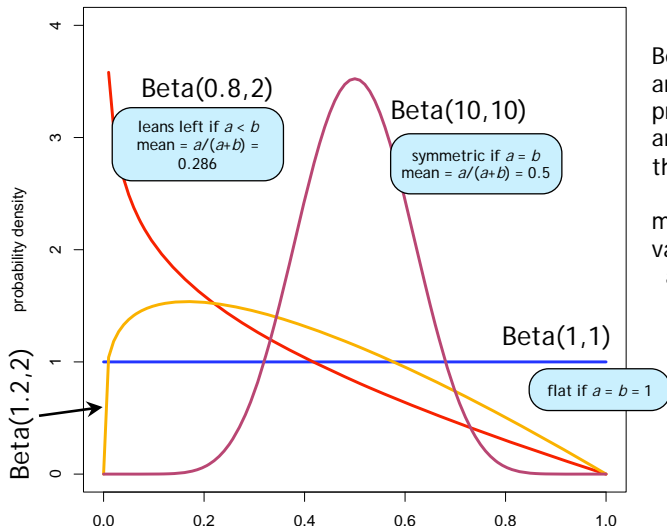
Important: μ and σ do not represent the mean and standard deviation of X : they are the mean and standard deviation of $\log(X)$!

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

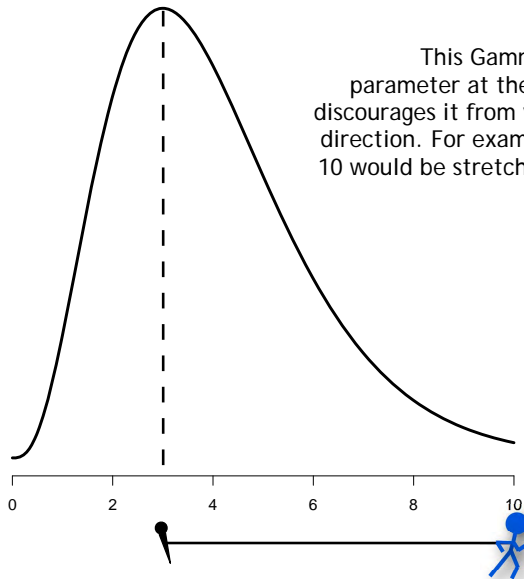


Beta distributions are appropriate for proportions, which are constrained to the interval $[0, 1]$.

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

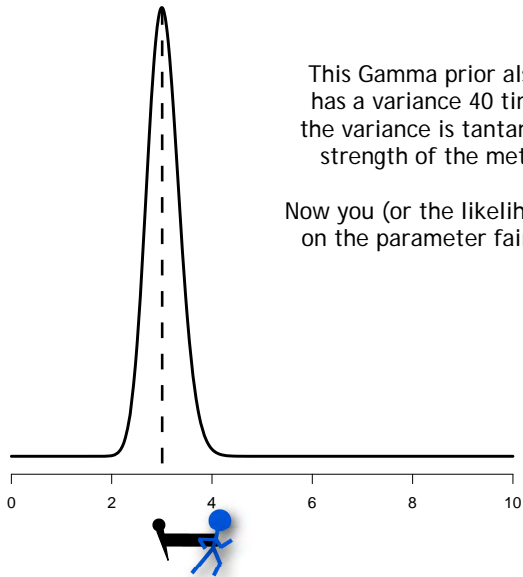
Prior Miscellany

- 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 $a > 1$)

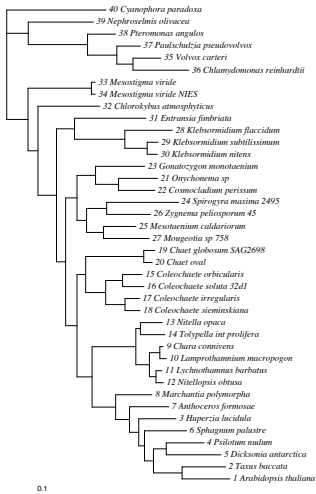


This Gamma prior also has a mode at 3, but has a variance 40 times smaller. Decreasing the variance is tantamount to increasing the strength of the metaphorical rubber band.

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

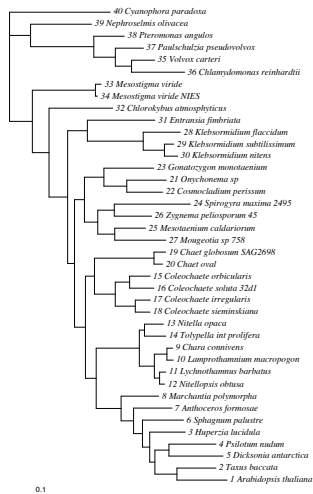


Separate priors applied to internal and external branches

External branch length prior is exponential with mean 0.1

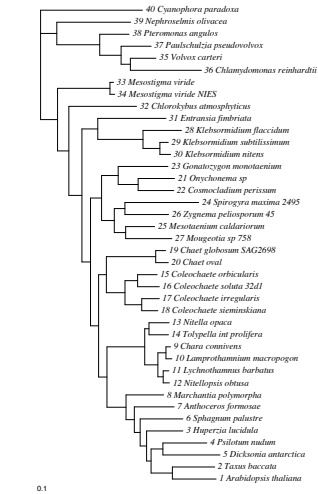
Internal branch length prior is exponential with mean 0.1

This is a reasonably vague internal branch length prior

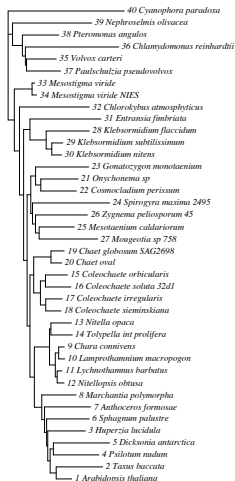


Internal branch length prior mean 0.01

(external branch length prior mean always 0.1)

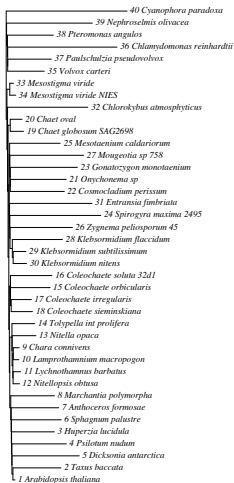


Internal branch length prior mean
0.001

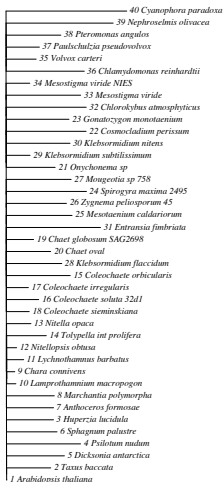


Internal branch length prior mean
0.0001

0.1



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.

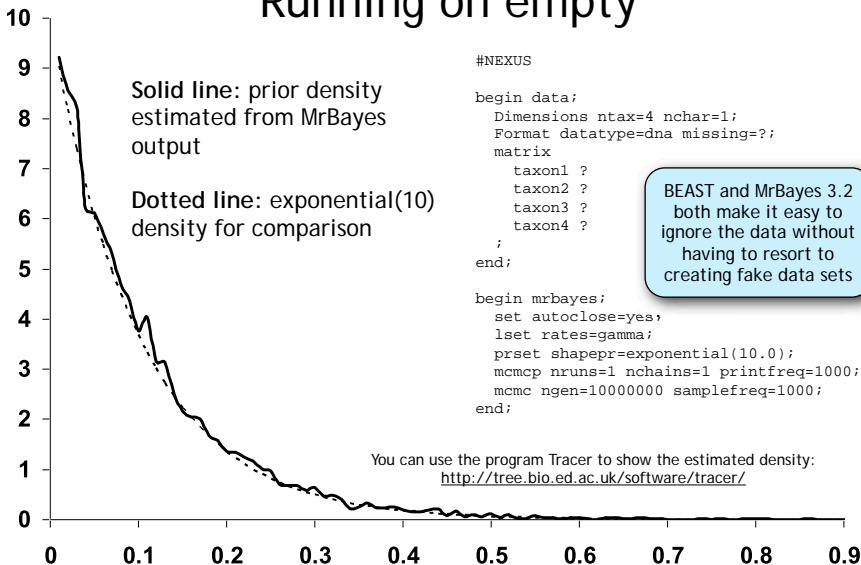
0.1

Prior Miscellany

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



Running on empty



Prior Miscellany

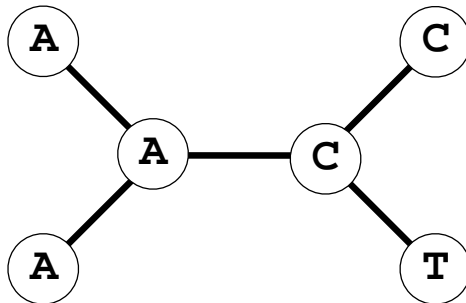
- priors as rubber bands
- running on empty
- hierarchical models
- empirical bayes



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

Prior: Exponential, mean=0.1

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



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.

Prior Miscellany

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



Empirical Bayes

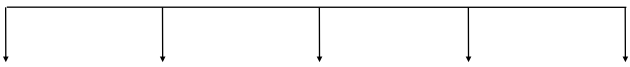
Empirical Bayes uses the data to determine some aspects of the prior, such as the prior mean.

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


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