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!)

Many many slides here, borrowed from both Mark Holder and Paul Lewis, because sometimes differnet ways of explaining concepts work better for different people. I will work through examples on the board, but am posting these for additional refence.

Bayes' Rule

$$Pr(A|B) = \frac{Pr(A)Pr(B|A)}{Pr(B)}$$

$$Pr(\text{Hypothesis}|\text{Data}) = \frac{Pr(\text{Hypothesis})Pr(\text{Data}|\text{Hypothesis})}{Pr(\text{Data})}$$

$$Pr(\text{Tree}|\text{Data}) \propto \mathbf{Pr(Tree)}Pr(\text{Data}|\text{Tree})$$

$Pr(\text{Tree})$ is the *prior* probability of the tree.

$$\mathbf{Pr}(\mathbf{Tree}|\mathbf{Data}) \propto Pr(\text{Tree})L(\text{Tree})$$

$Pr(\text{Tree})$ is the *prior* probability of the tree.

$L(\text{Tree})$ is the likelihood of the tree.

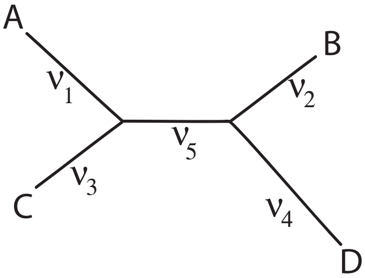
$\mathbf{Pr}(\mathbf{Tree}|\mathbf{Data})$ is the *posterior* probability of the tree.

The posterior probability is a great way to evaluate trees:

- Ranks trees
- Intuitive measure of confidence
- Is the ideal “weight” for a tree in secondary analyses
- Closely tied to the likelihood

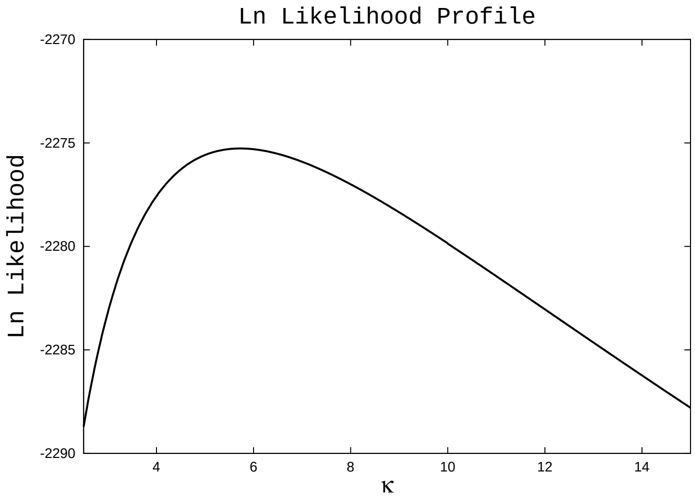
Our models don't give us $L(\text{Tree})$

They give us things like $L(\text{Tree}, \kappa, \alpha, \nu_1, \nu_2, \nu_3, \nu_4, \nu_5)$

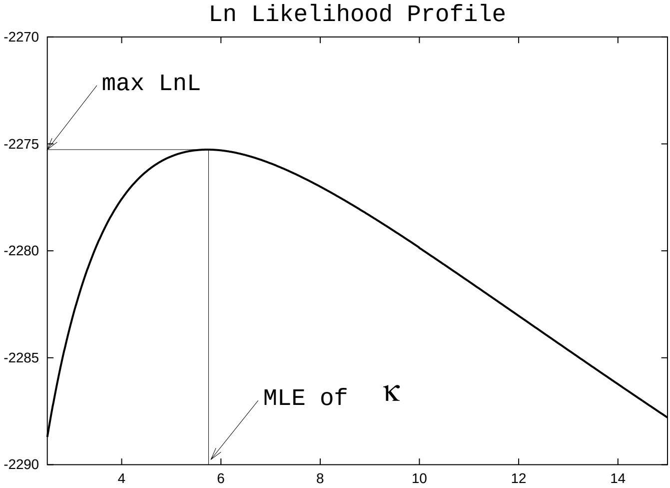


“Nuisance Parameters”

Aspects of the evolutionary model that we don’t care about, but are in the likelihood equation.



Ln Likelihood Profile
-2270
-2275
-2280
-2285
-2290
Ln Likelihood
4
6
8
10
12
14
κ


Ln Likelihood Profile
-2270
-2275
-2280
-2285
-2290
max LnL
MLE of κ
4
6
8
10
12
14

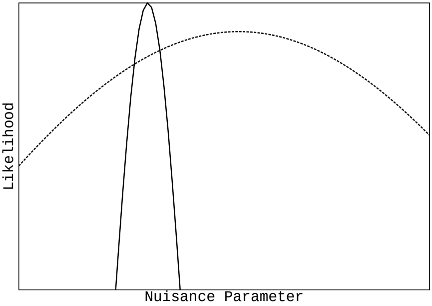
Marginalizing over (integrating out) nuisance parameters

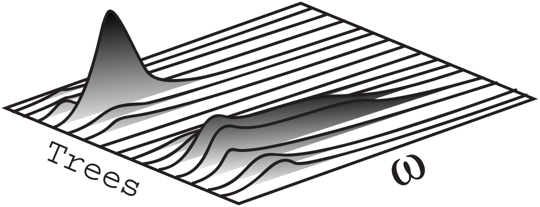
$$L(\text{Tree}) = \int L(\text{Tree}, \kappa) Pr(\kappa) d\kappa$$

- Removes the nuisance parameter
- Takes the entire likelihood function into account

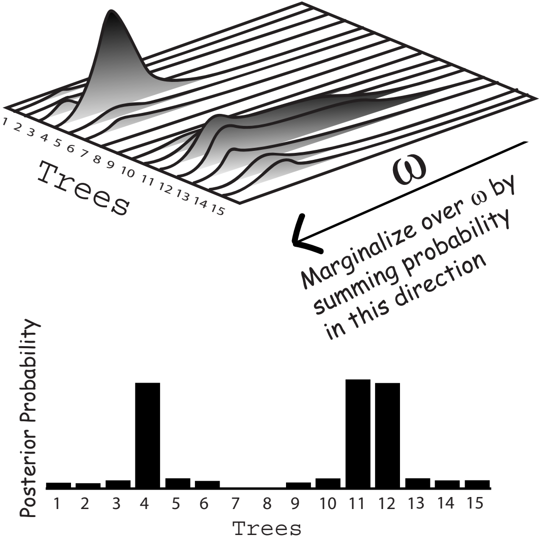
- Avoids estimation errors
- Requires a prior for the parameter

When there is substantial uncertainty in a parameter's value, marginalizing can give qualitatively different answers than using the MLE.

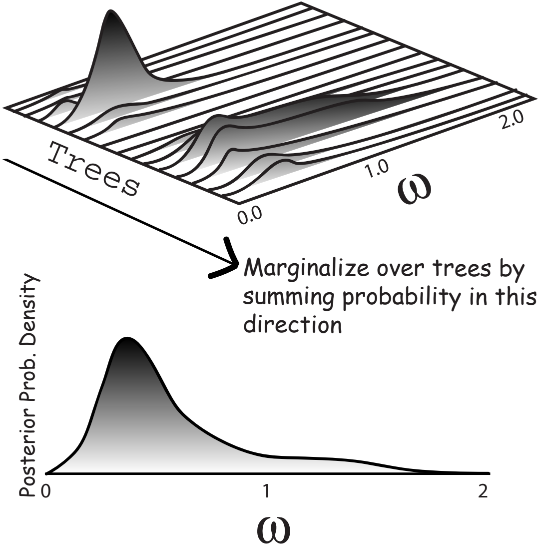




Joint posterior probability density for trees and ω



1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Trees
ω
Marginalize over ω by summing probability in this direction
Posterior Probability
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Trees


Trees
0.0
1.0
2.0
ω
Marginalize over trees by summing probability in this direction
Posterior Prob. Density
0
1
2
ω

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

Likelihood of hypothesis θ

Prior probability of hypothesis θ

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

Posterior probability of hypothesis θ

Marginal probability of the data (marginalizing over hypotheses)

Bayes' rule: continuous case

Likelihood

Prior probability ***density***

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

Posterior probability ***density***

Marginal probability of the data

If you had to guess…

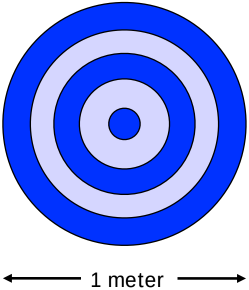


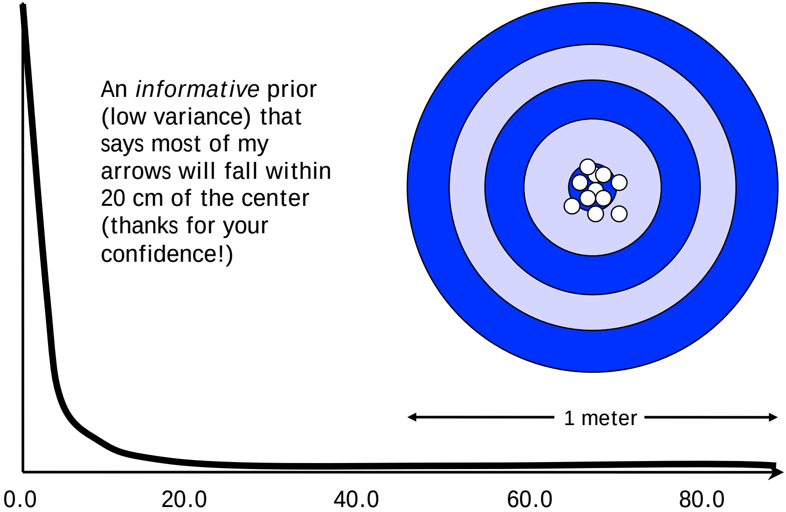
Photo by Tracy Heath

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)

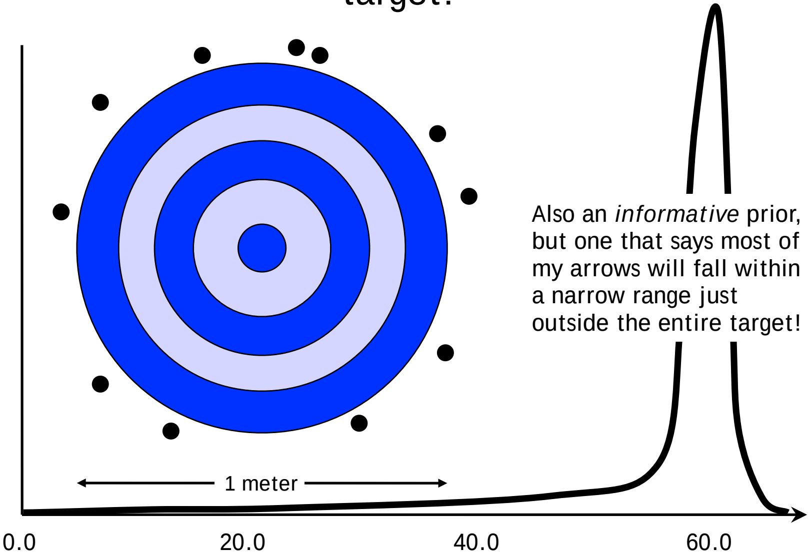
0.0 d ∞

Case 1: assume I have talent

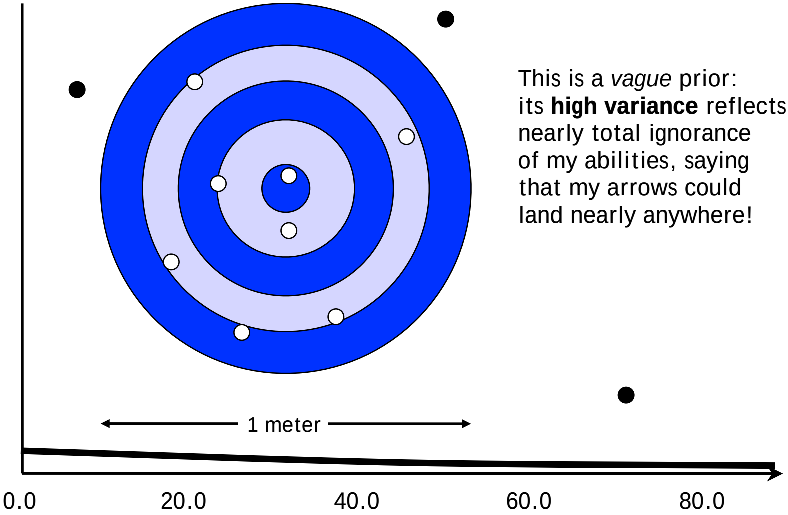
An *informative* prior (low variance) that says most of my arrows will fall within 20 cm of the center (thanks for your confidence!)



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



Case 3: assume I have no talent



This is a *vague* prior:
its **high variance** reflects
nearly total ignorance
of my abilities, saying
that my arrows could
land nearly anywhere!

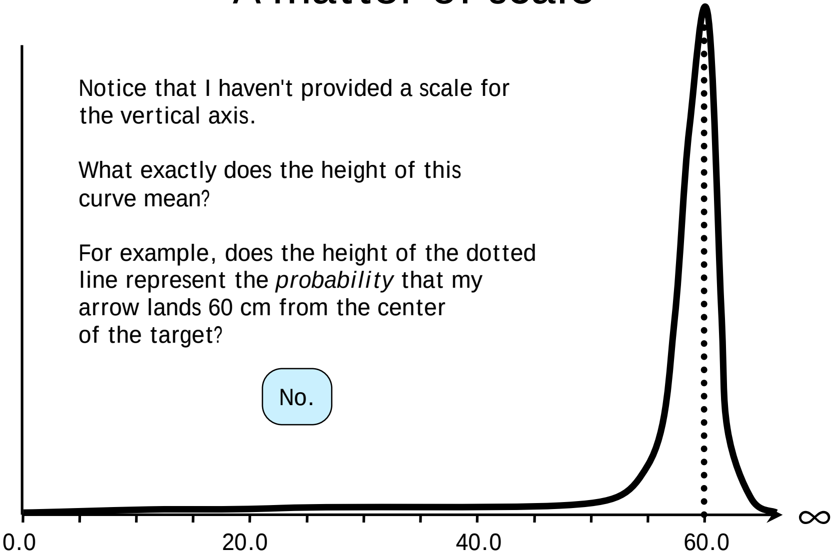
A matter of scale

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.

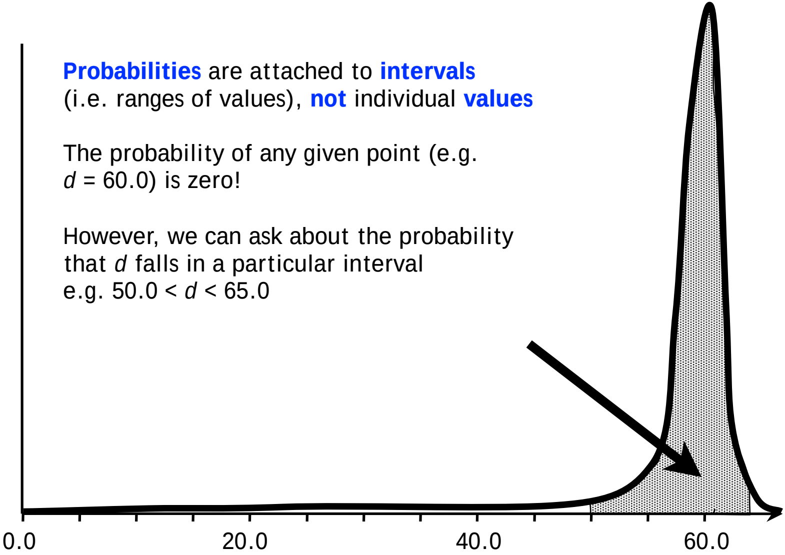


Probabilities are associated with intervals

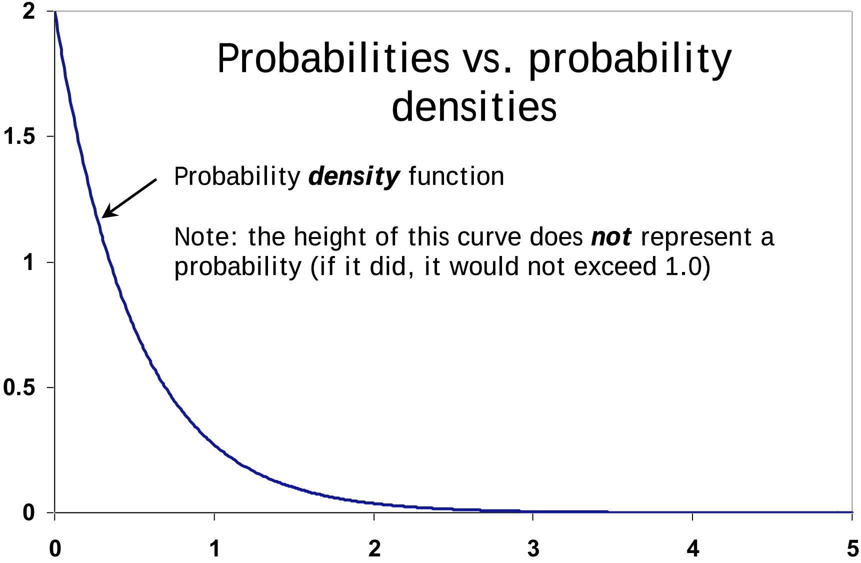
Probabilities are attached to **intervals** (i.e. ranges of values), **not** individual **values**

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



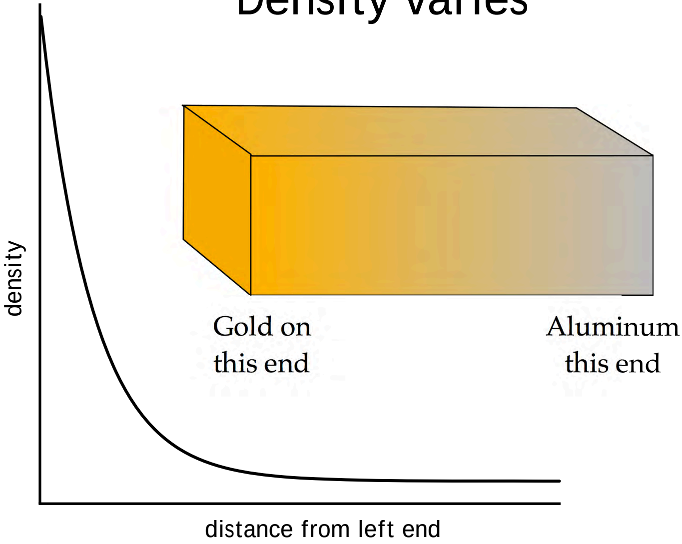
Densities of various substances

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

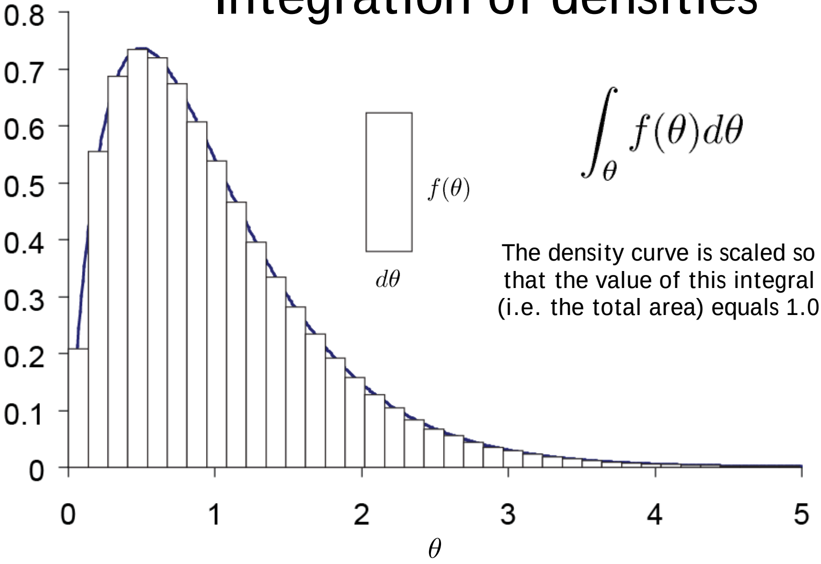
Density does not equal mass

mass = density × volume

Density varies



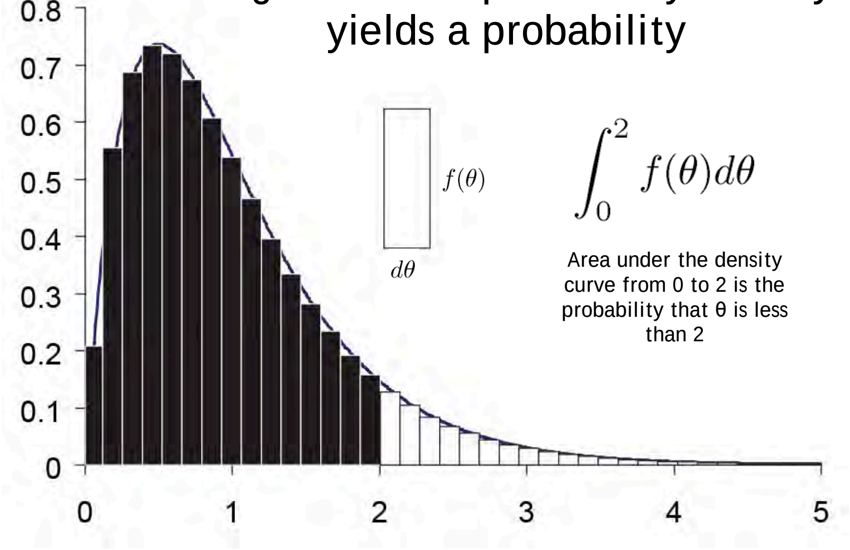
Integration of densities



$$\int_\theta f(\theta)d\theta$$

The density curve is scaled so that the value of this integral (i.e. the total area) equals 1.0

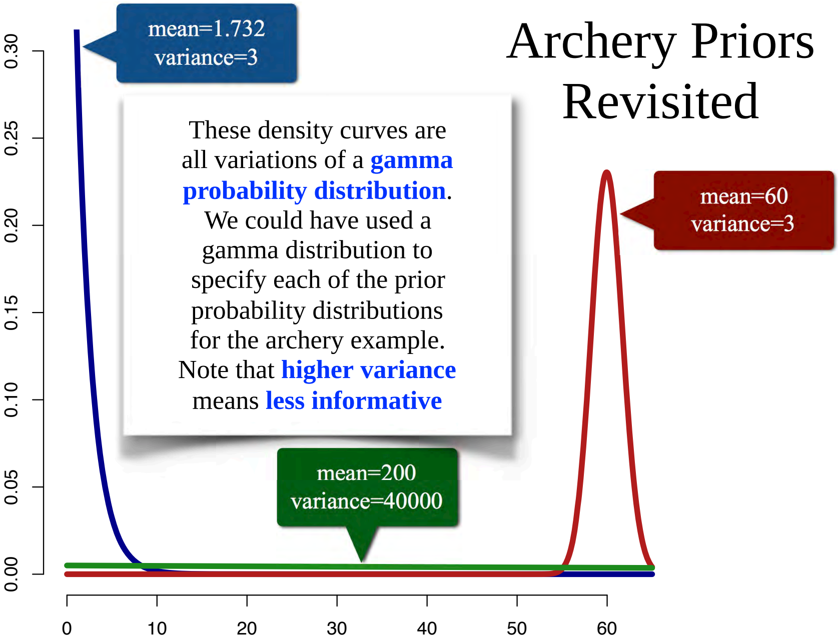
Integration of a probability density yields a probability



$$\int_0^2 f(\theta)d\theta$$

Area under the density curve from 0 to 2 is the probability that θ is less than 2

Archery Priors Revisited



These density curves are all variations of a **gamma probability distribution**. We could have used a gamma distribution to specify each of the prior probability distributions for the archery example. Note that **higher variance** means **less informative**

Usually there are many parameters...

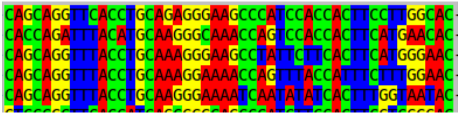
A 2-parameter example

$$f(\theta,\phi|D) = \frac{\overbrace{f(D|\theta,\phi)}^{\text{Likelihood}}\ \overbrace{f(\theta)f(\phi)}^{\text{Prior probability density}}}{\underbrace{\int_\theta \int_\phi f(D|\theta,\phi)\ f(\theta)f(\phi)\ d\theta d\phi}_{\text{Marginal probability of data}}}$$

Posterior probability density

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...

It is very hard to calculate the prior 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 **c**hain **M**onte **C**arlo

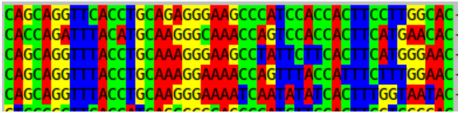
- Simulates a walk through parameter/tree space.
- Lets us estimate posterior probabilities for any aspect of the model
- Relies on the ***ratio*** of posterior densities between two points

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

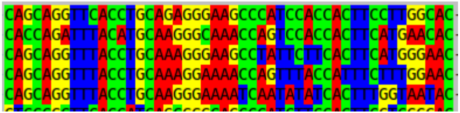
$$R = \frac{\frac{Pr(\mathsf{Point}_2)L(\mathsf{Point}_2)}{Pr(\mathsf{Data})}}{\frac{Pr(\mathsf{Point}_1)L(\mathsf{Point}_1)}{Pr(\mathsf{Data})}}$$

$$R = \frac{Pr\left(\mathsf{Point}_2\right)L\left(\mathsf{Point}_2\right)}{Pr\left(\mathsf{Point}_1\right)L\left(\mathsf{Point}_1\right)}$$

We can avoid calculating the prior probability of the data!

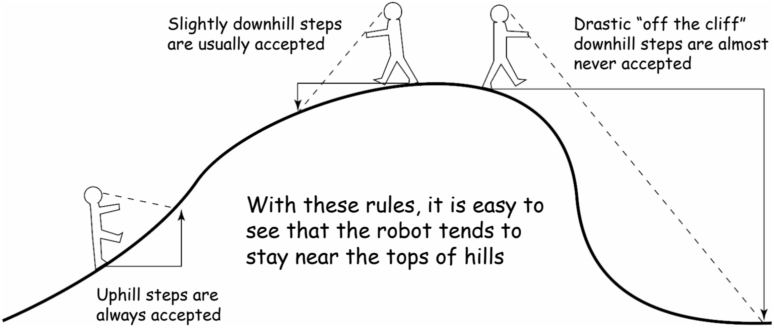


We can avoid calculating the prior probability of the data!

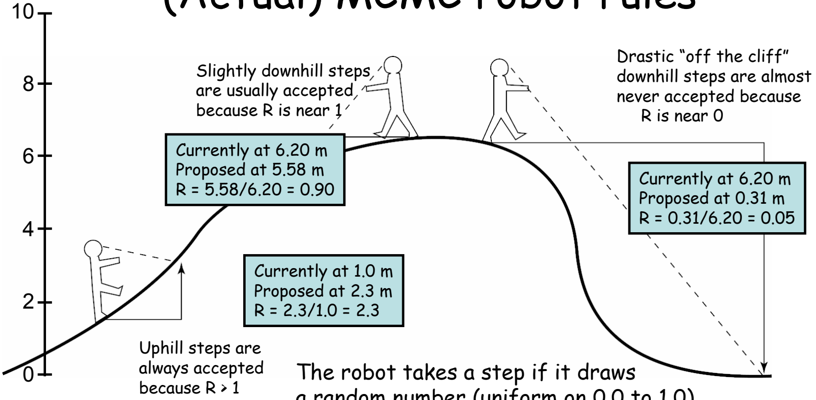


Imagine using a robot to survey a landscape...

MCMC robot's rules



(Actual) MCMC robot rules



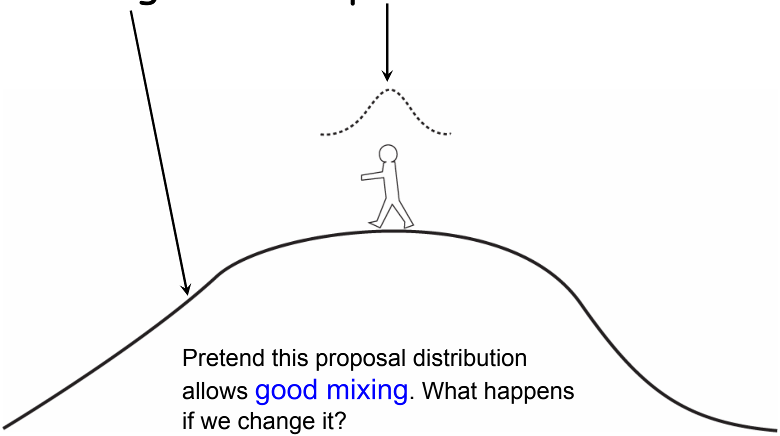
The robot takes a step if it draws a random number (uniform on 0.0 to 1.0), and that number is less than or equal to R

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

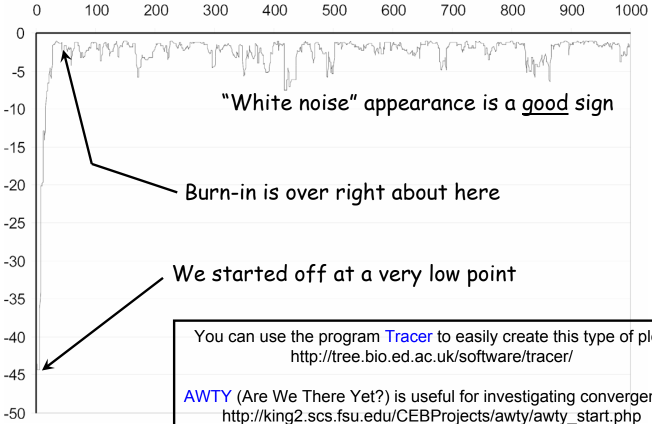
Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087-1092.

Target vs. Proposal Distributions



Pretend this proposal distribution allows good mixing. What happens if we change it?

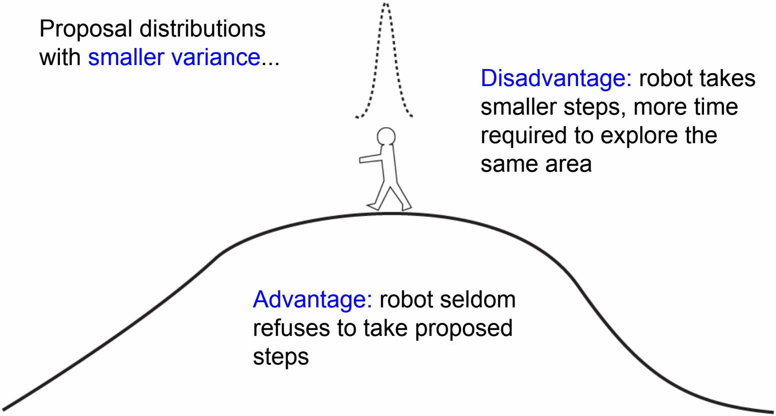
Trace plots



You can use the program Tracer to easily create this type of plot:
http://tree.bio.ed.ac.uk/software/tracer/

AWTY (Are We There Yet?) is useful for investigating convergence:
http://king2.scs.fsu.edu/CEBProjects/awty/awty_start.php

Target vs. Proposal Distributions



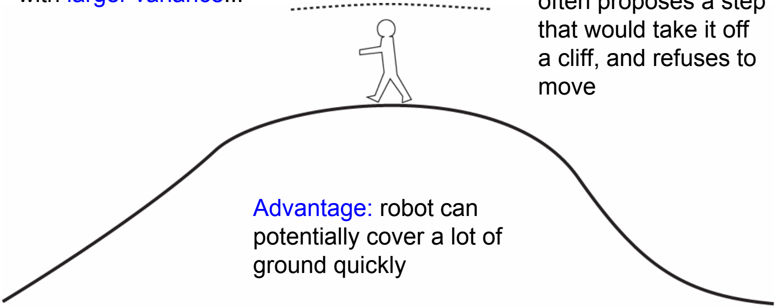
Proposal distributions with smaller variance...

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

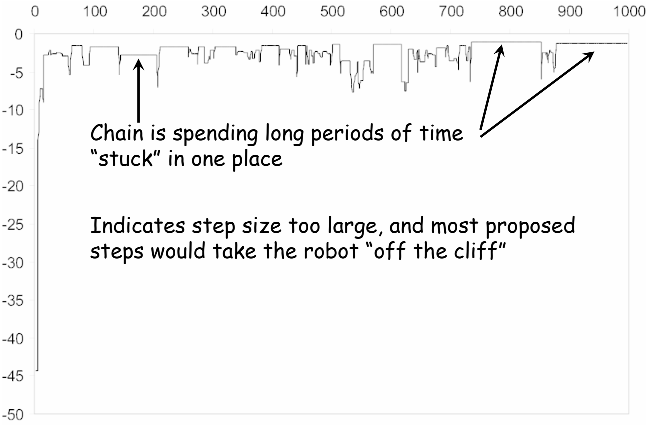
Proposal distributions with larger variance...



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



0
100
200
300
400
500
600
700
800
900
1000
0
-5
-10
-15
-20
-25
-30
-35
-40
-45
-50
Chain is spending long periods of time "stuck" in one place
Indicates step size too large, and most proposed steps would take the robot "off the cliff"

Working through example together!

https://hydrodictyon.eeb.uconn.edu/people/plewis/courses/phylogenetics/homeworks/2020/hw7.pdf

We will split into groups, and try different proposal distributions, and share our posterior sample.

Paul Lewis' MCMC robot demo

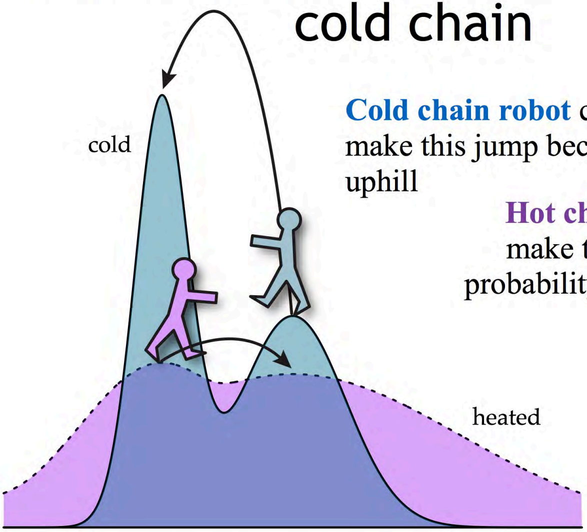
http://phylogeny.uconn.edu/mcmc-robot/

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.).

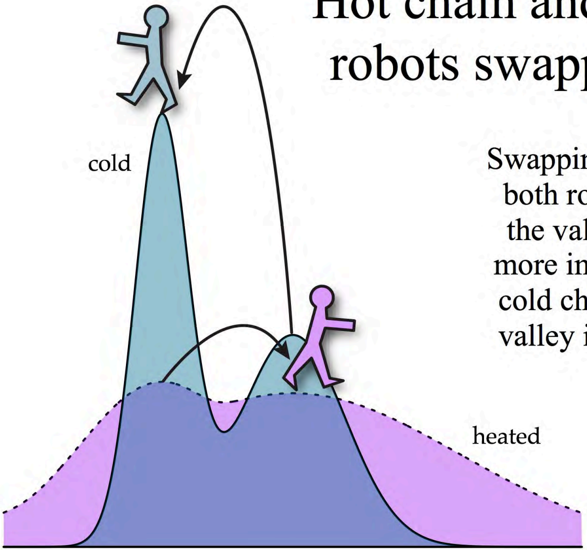
Heated chains act as scouts for the cold chain



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

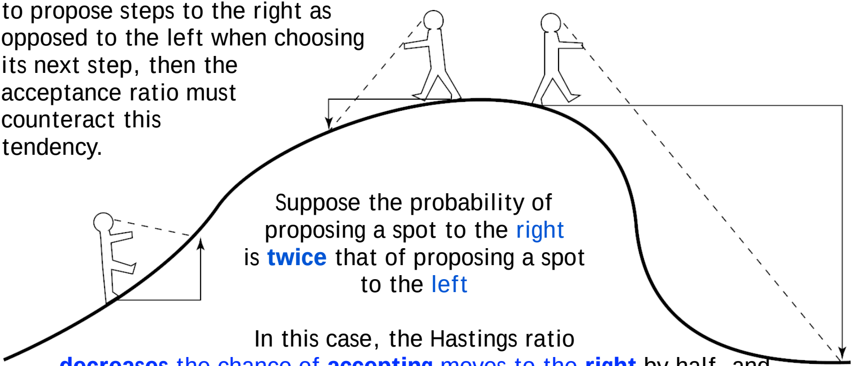
`http://phylogeny.uconn.edu/mcmc-robot/`

“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.



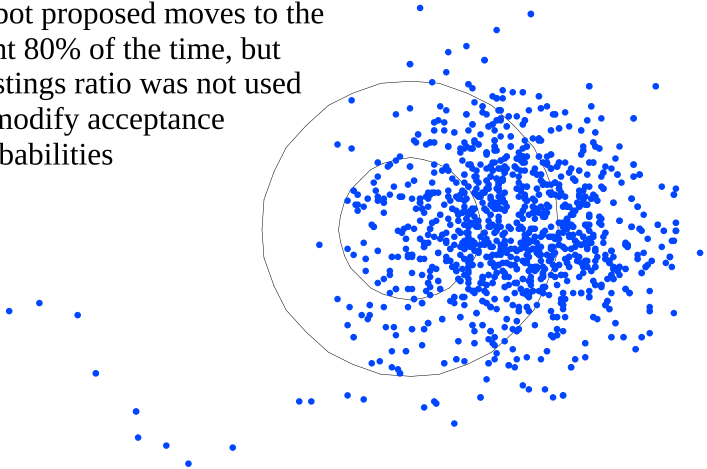
Suppose the probability of proposing a spot to the right is **twice** that of proposing a spot to the left

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.

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

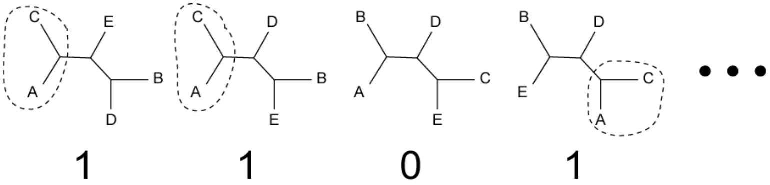
Posterior ratio

Hastings ratio

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

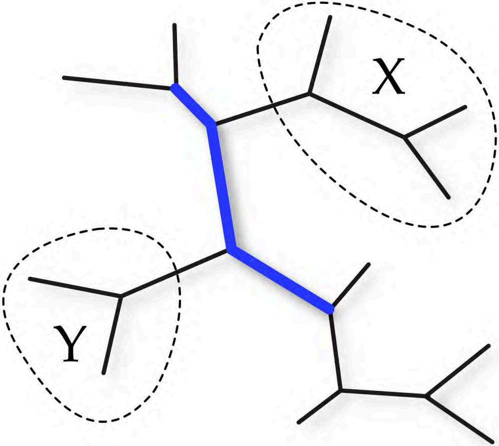
III. Bayesian phylogenetics

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

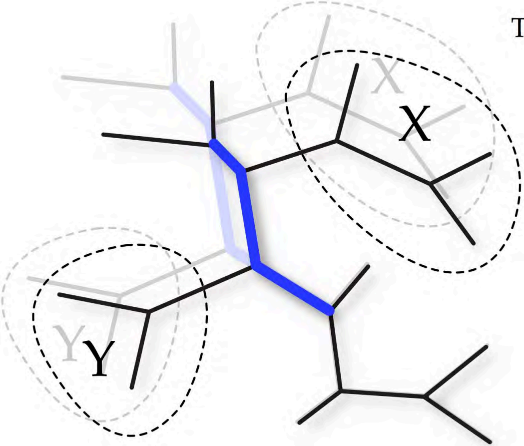


The Larget-Simon move

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

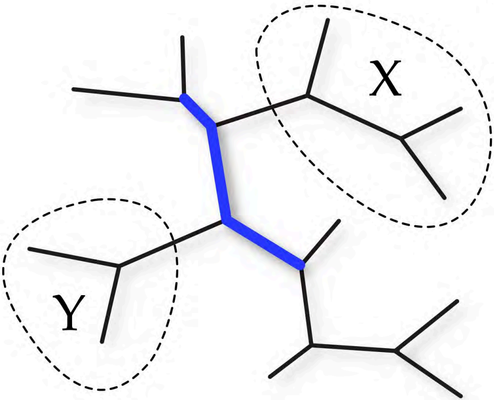


The Larget-Simon move

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

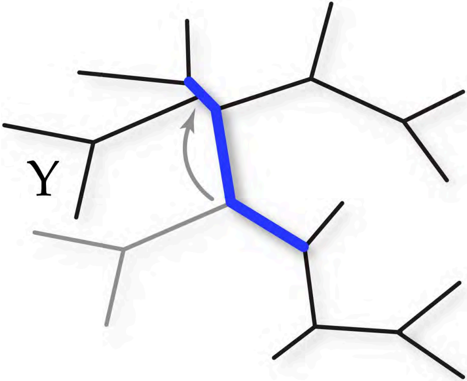


The Larget-Simon move

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



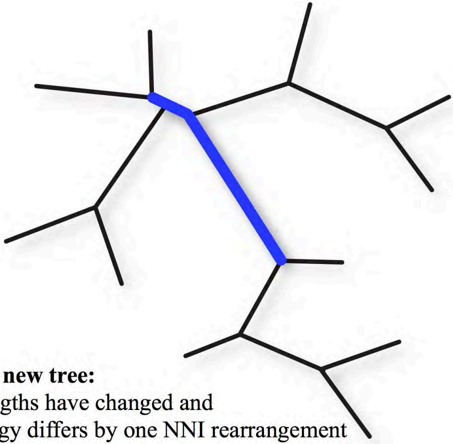
The Larget-Simon move

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

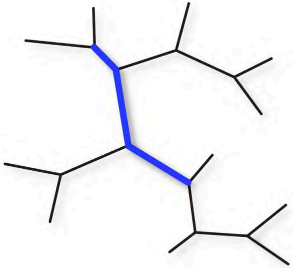
The Larget-Simon move

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

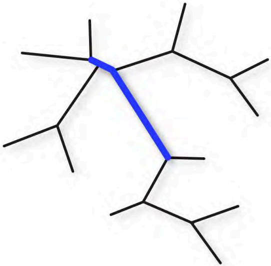
Step 3:
Choose X or Y randomly, then reposition randomly

Moving through treespace



Current tree

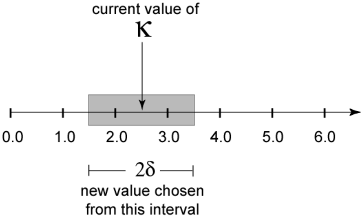
log-posterior = -34256

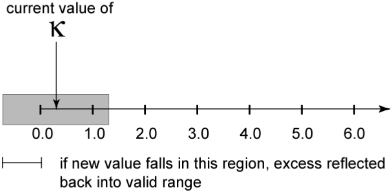


Proposed tree

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.

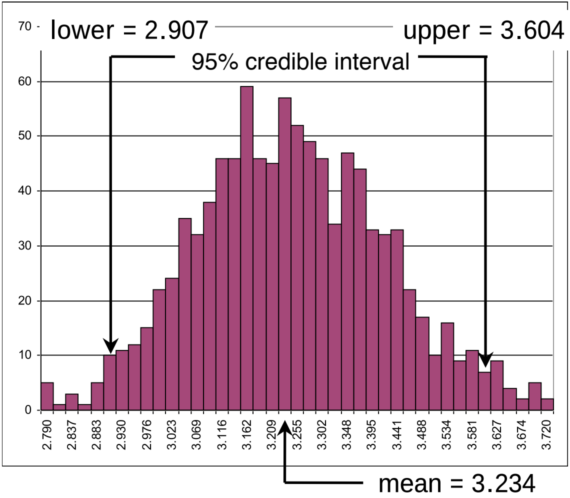
Proposal distribution is the uniform distribution on the interval (κ-δ, κ+δ)

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. Larget-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.

Data from Lewis, L., and Flechtner, V. 2002. Taxon 51: 443-451.

IV. Prior distributions

Common Priors

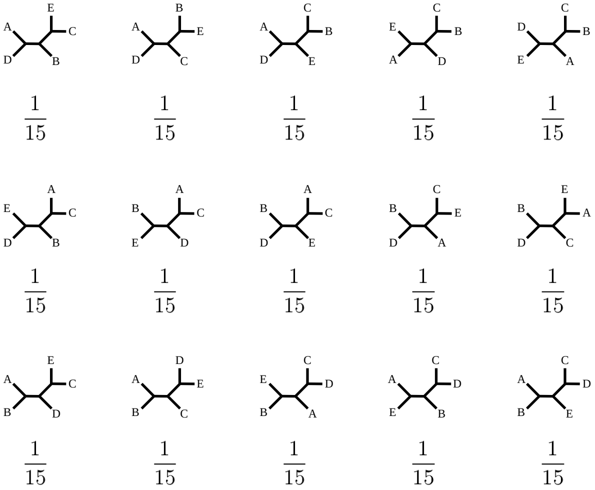
- **Discrete uniform** for topologies
  - exceptions becoming more common
- **Beta** for proportions
- **Gamma** or **Log-normal** for branch lengths and other parameters with support $[0,\infty)$
  - Exponential is common special case of the gamma distribution
- **Dirichlet** for state frequencies and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

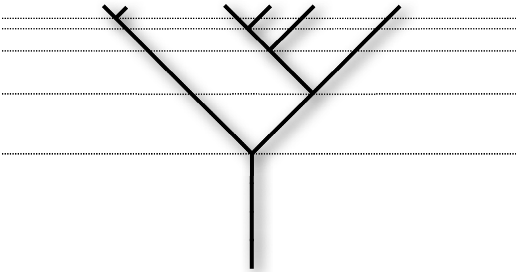
Common Priors

- **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**

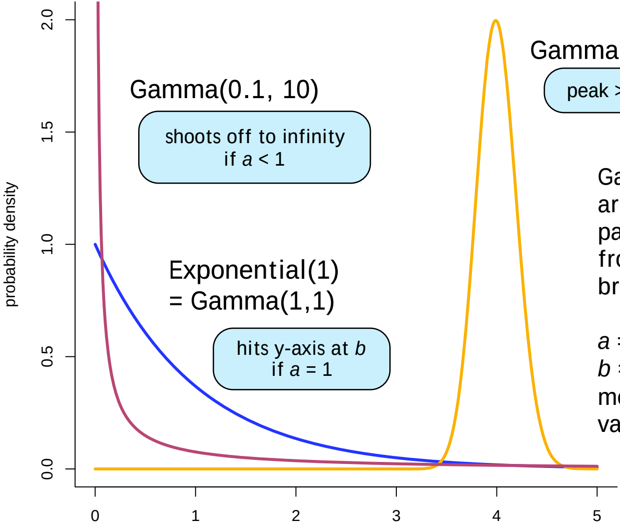


Yule model provides joint prior for both **topology** and **divergence times**



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



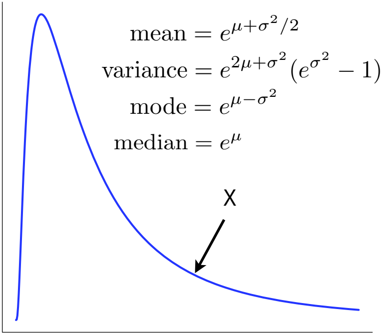
Gamma distributions are ideal for parameters that range from 0 to infinity (e.g. branch lengths)

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

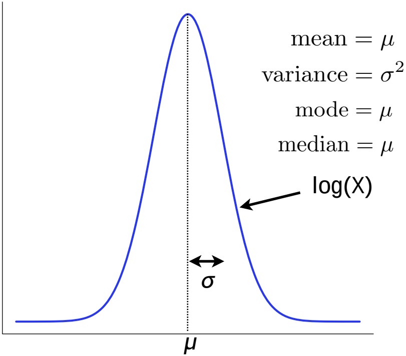
*Note: be aware that in many papers the Gamma distribution is defined such that the second (scale) parameter is the *inverse* of the value b used in this slide! In this case, the mean and variance would be a/b and a/b^2, respectively.

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.

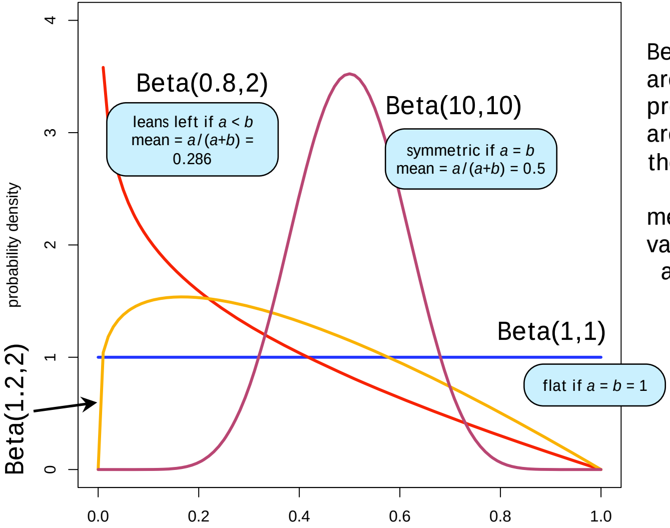


Important: *μ and σ do **not** represent the mean and standard deviation of X: they are the mean and standard deviation of log(X)!*

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

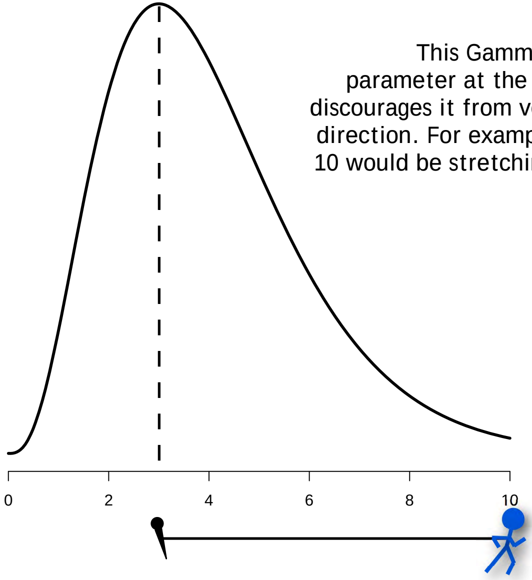


Beta distributions are appropriate for proportions, which are constrained to the interval [0,1].

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

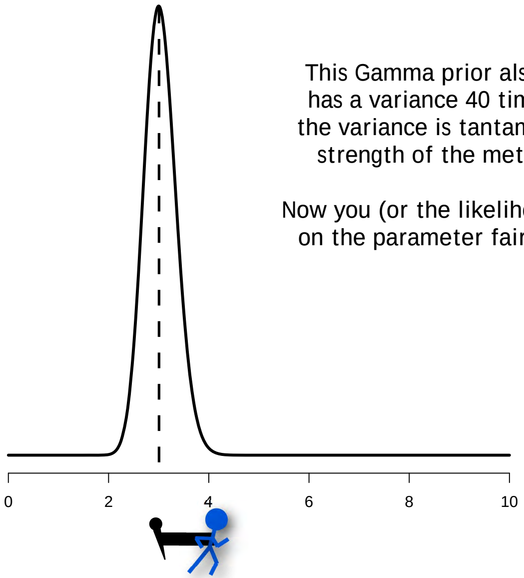
Prior Miscellany

- 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)

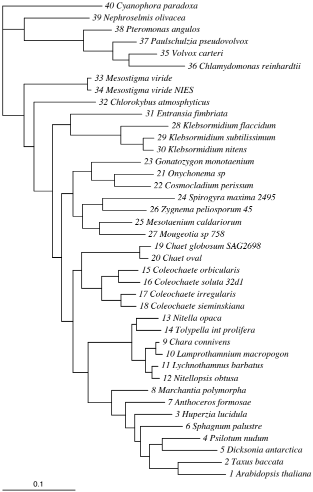


This Gamma prior also has a mode at 3, but has a variance 40 times smaller. Decreasing the variance is tantamount to increasing the strength of the metaphorical rubber band.

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

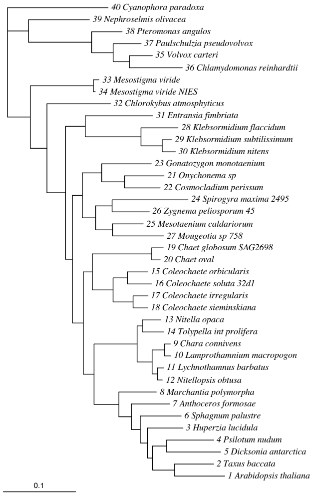


Separate priors applied to internal and external branches

External branch length prior is exponential with mean 0.1

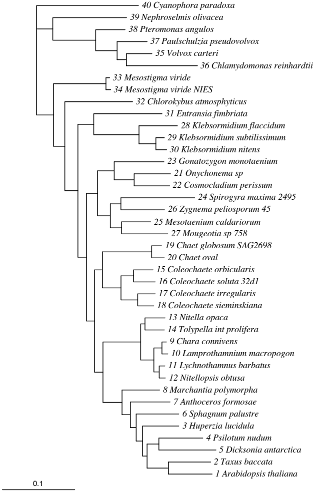
Internal branch length prior is exponential with mean 0.1

This is a reasonably vague internal branch length prior

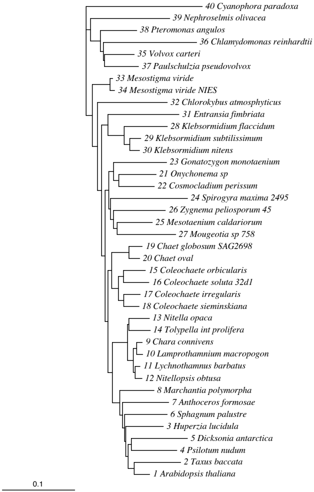


Internal branch length prior mean 0.01

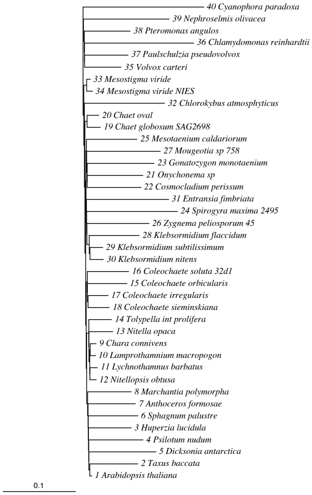
(external branch length prior mean always 0.1)



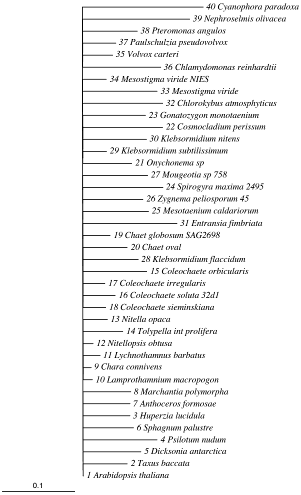
Internal branch length prior mean 0.001



Internal branch length prior mean 0.0001



Internal branch length prior mean
0.00001



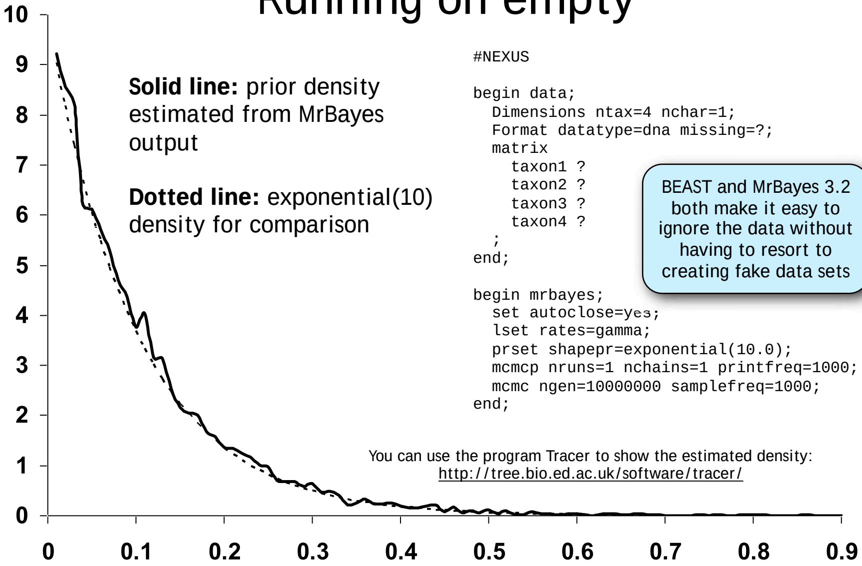
Internal branch length prior mean 0.000001

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/

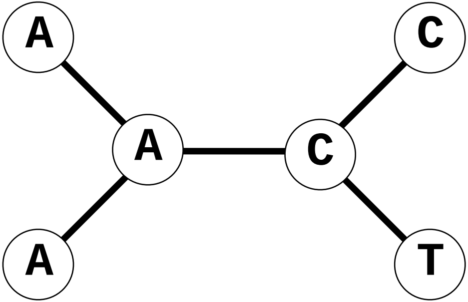
Prior Miscellany

- 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

$$L_k = \tfrac{1}{4}\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_1/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_2/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4\nu_3/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4\nu_4/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_5/3}\right]$$



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

$$L_k = \frac{1}{4}\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_1/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_2/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4\nu_3/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4\nu_4/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_5/3}\right]$$

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

$$L_k = \tfrac{1}{4}\left[\tfrac{1}{4} + \tfrac{3}{4}e^{-4\nu_1/3}\right]\left[\tfrac{1}{4} + \tfrac{3}{4}e^{-4\nu_2/3}\right]\left[\tfrac{1}{4} - \tfrac{1}{4}e^{-4\nu_3/3}\right]\left[\tfrac{1}{4} - \tfrac{1}{4}e^{-4\nu_4/3}\right]\left[\tfrac{1}{4} + \tfrac{3}{4}e^{-4\nu_5/3}\right]$$