# Bayesian Phylogenetics

Paul O. Lewis
Department of Ecology & Evolutionary Biology
University of Connecticut

25 January 2013
Workshop on Molecular Evolution
Český Krumlov





## An Introduction to Bayesian Phylogenetics

- Bayesian inference in general
- Markov chain Monte Carlo (MCMC)
- Bayesian phylogenetics
- Prior distributions
- Bayesian model selection

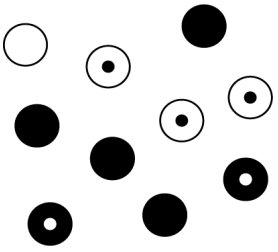


## I. Bayesian inference in general



## Joint probabilities

**B = Black S = Solid**
**W = White D = Dotted**

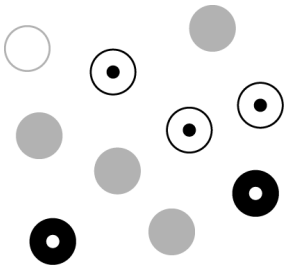


Pr(B) = 0.6 Pr(S) = 0.5
Pr(W) = 0.4 Pr(D) = 0.5

Pr( ) = Pr(B, D) = 0.2
Pr( ) = Pr(B, S) = 0.4
Pr( ) = Pr(W, D) = 0.3
Pr( ) = Pr(W, S) = 0.1



## Conditional probabilities



$$\Pr(B|D) = \frac{2}{5} = 0.4$$

Hide all solid marbles
(leaving 5 with dot)

Of those left, 2 are black

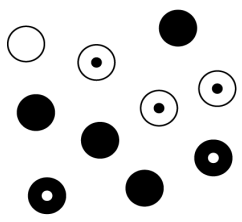


## Bayes' rule

$$\Pr(B, D)$$

$$\Pr(D)\Pr(B|D) = \Pr(B)\Pr(D|B)$$

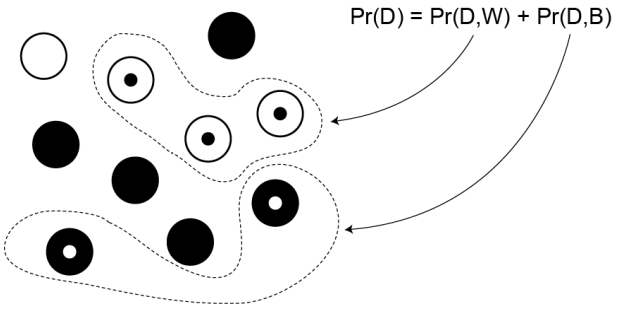
$$\frac{1}{2} \times \frac{2}{5} = \frac{3}{5} \times \frac{1}{3}$$



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

$$= \frac{\frac{3}{5} \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{5}$$

## Probability of "Dotted"






## Bayes' rule (cont.)

$$\Pr(B|D) = \frac{\Pr(B)\Pr(D|B)}{\Pr(D)}$$
$$= \frac{\Pr(D,B)}{\Pr(D,B) + \Pr(D,W)}$$

Pr(*D*) is the **marginal probability** of being dotted
To compute it, we **marginalize over colors**

It is easy to see that Pr(*D*) serves as a *normalization constant*, ensuring that Pr(*B*|*D*) + Pr(*W*|*D*) = 1.0

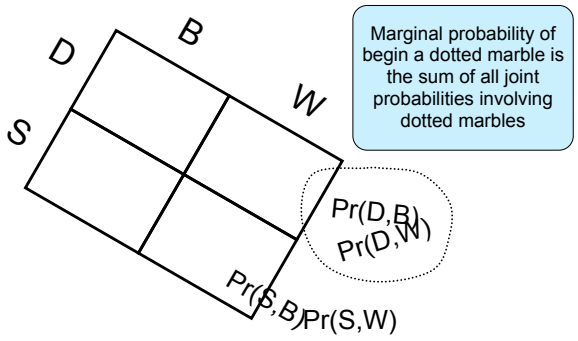


## Joint probabilities

| | B | W |
|---|---|---|
| D | Pr(D,B) | Pr(D,W) |
| S | Pr(S,B) | Pr(S,W) |



## Marginalizing over colors






## Marginal probabilities

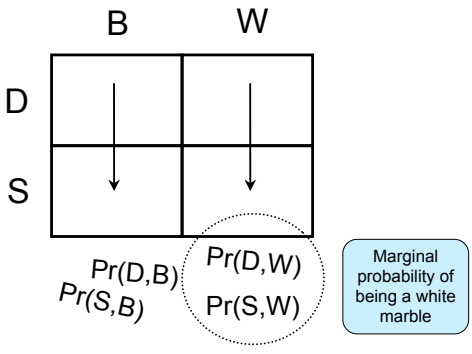
| | B | W | |
|---|---|---|---|
| D | | | Pr(D,B) + Pr(D,W) |
| S | | | Pr(S,B) + Pr(S,W) |

Marginal probability of being dotted

Marginal probability of being solid



## Marginalizing over "dottedness"



## Bayes' rule (cont.)

$$\begin{aligned}\Pr(B|D) &= \frac{\Pr(B)\Pr(D|B)}{\Pr(D,B)+\Pr(D,W)} \\ &= \frac{\Pr(B)\Pr(D|B)}{\Pr(B)\Pr(D|B)+\Pr(W)\Pr(D|W)} \\ &= \frac{\Pr(B)\Pr(D|B)}{\sum_{\theta\in\{B,W\}}\Pr(\theta)\Pr(D|\theta)}\end{aligned}$$



## Bayes' rule in Statistics

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

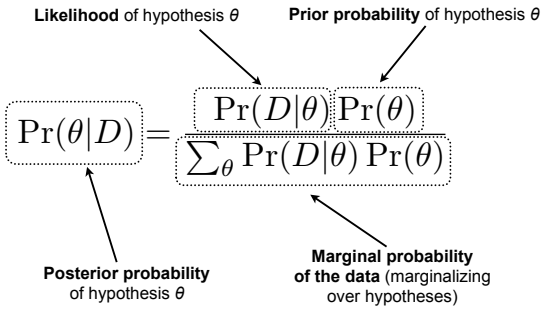
$D$ refers to the "observables" (i.e. the **Data**)

$\theta$ 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






## Simple (albeit silly) paternity example

$\theta_1$ and $\theta_2$ 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 | $\theta_1$ | $\theta_2$ | 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 |



## The prior can be your friend

Suppose the test for a **rare** disease is 99% accurate.

$$\begin{aligned}\Pr(+|\text{disease}) &= 0.99 \\ \Pr(+|\text{healthy}) &= 0.01\end{aligned}$$

datum hypothesis

Suppose further I **test positive** for the disease. How worried should I be?

(Note that we do not need to consider the case of a negative test result.)

It is very tempting to (mis)interpret the likelihood as a posterior probability and conclude "There is a 99% chance that I have the disease."



## The prior can be your friend

The posterior probability is 0.99 only if the **prior probability** of having the disease is 0.5:

$$\begin{aligned}\Pr(\text{disease}|+) &= \frac{\Pr(+|\text{disease})\left(\frac{1}{2}\right)}{\Pr(+|\text{disease})\left(\frac{1}{2}\right)+\Pr(+|\text{healthy})\left(\frac{1}{2}\right)} \\ &= \frac{(0.99)\left(\frac{1}{2}\right)}{(0.99)\left(\frac{1}{2}\right)+(0.01)\left(\frac{1}{2}\right)} = 0.99\end{aligned}$$

If, however, the prior odds against having the disease are a million to 1, then the posterior probability is much more reassuring:

$$\begin{aligned}\Pr(\text{disease}|+) &= \frac{(0.99)\left(\frac{1}{1000000}\right)}{(0.99)\left(\frac{1}{1000000}\right)+(0.01)\left(\frac{999999}{1000000}\right)} \\ &\approx 0.0001\end{aligned}$$

## An important caveat

This (rare disease) example involves a **tiny amount of data** (one observation) and an extremely **informative prior**, and gives the impression that maximum likelihood (ML) inference is not very reliable.

However, in phylogenetics, we often have **lots of data** and use much **less informative priors**, so in phylogenetics ML inference is generally **very reliable**.

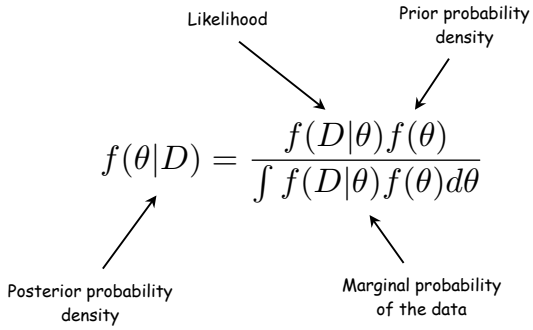


## Discrete vs. Continuous

- So far, we've been dealing with discrete hypotheses (e.g. either this father or that father, have disease or don't have disease)
- In phylogenetics, substitution models represent an infinite number of hypotheses (each combination of parameter values is in some sense a separate hypothesis)
- How do we use Bayes' rule when our hypotheses form a continuum?

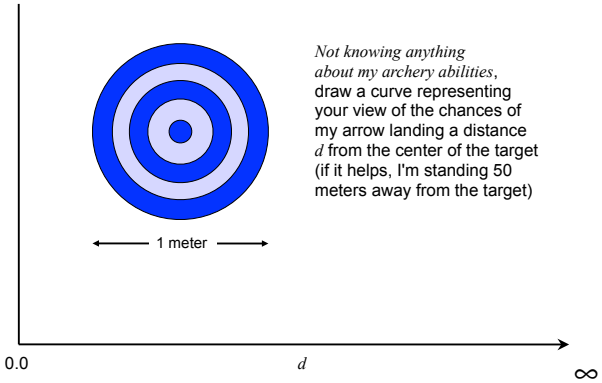


## Bayes' rule: continuous case



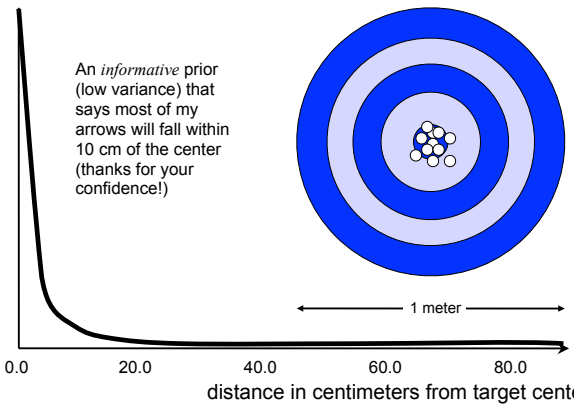



## If you had to guess...



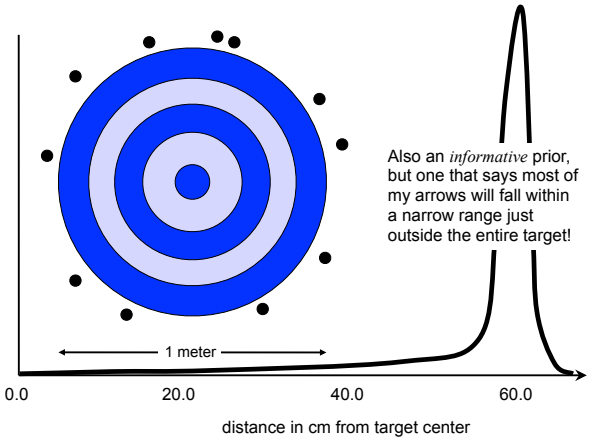



## Case 1: assume I have talent

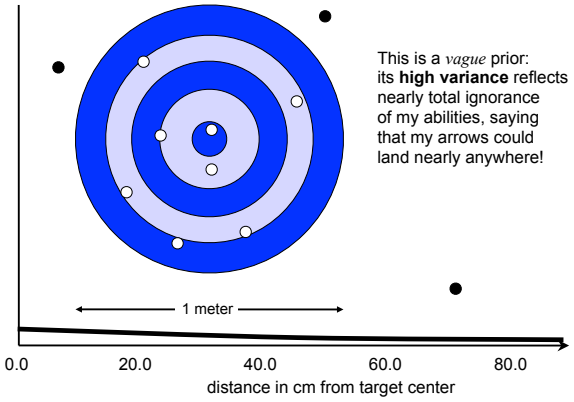





## Case 2: assume I have a talent for missing the target!



## 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!



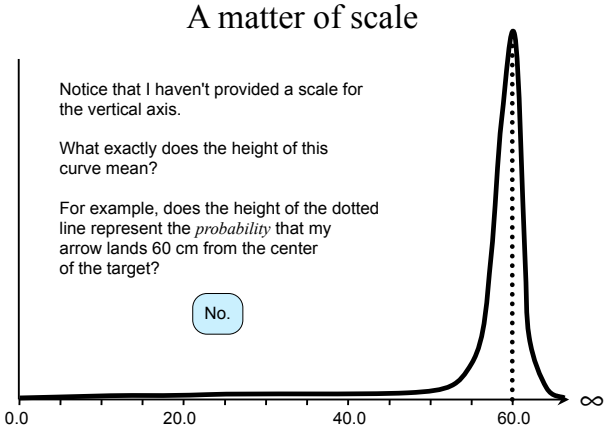
## A matter of scale

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.



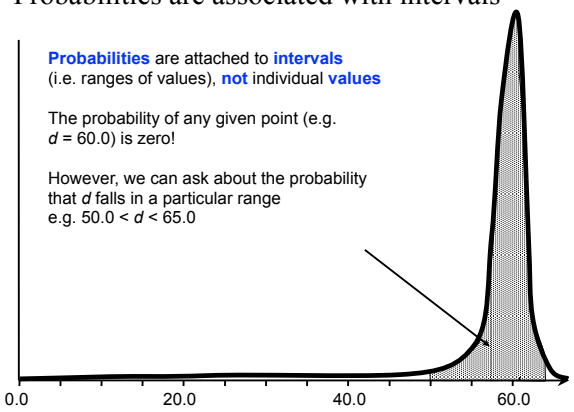



## Probabilities are associated with intervals

**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!

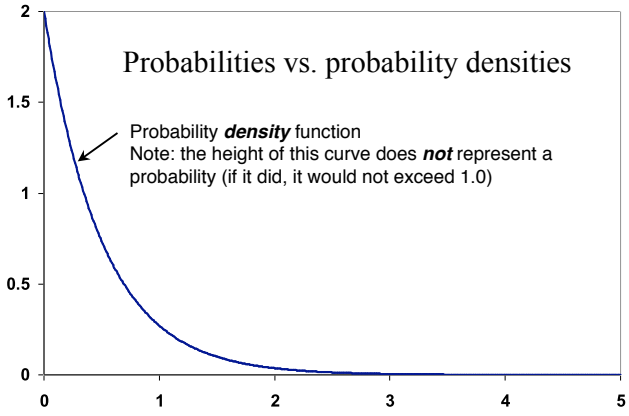
However, we can ask about the probability that $d$ falls in a particular range
e.g. $50.0 < d < 65.0$






## Probabilities vs. probability densities

Probability ***density*** function
Note: the height of this curve does ***not*** represent a probability (if it did, it would not exceed 1.0)






## Densities of various substances

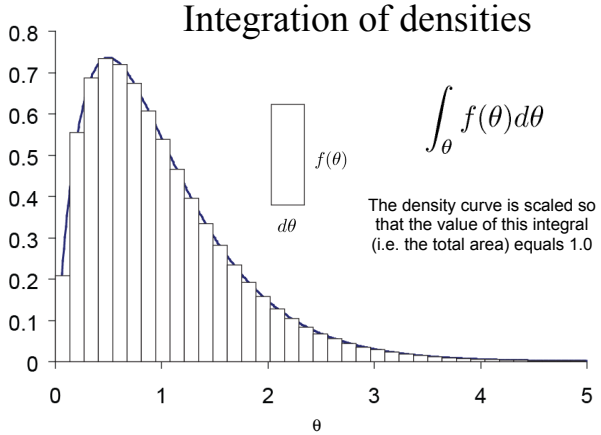
| Substance | Density (g/cm$^3$) |
|---|---|
| Cork | 0.24 |
| Aluminum | 2.70 |
| Gold | 19.30 |

*Density does not equal mass*
mass = density × volume

Note: *volume* is appropriate for objects of dimension 3 or higher
For 2-dimensions, *area* takes the place of volume
For 1-dimension, *linear distance* replaces volume.

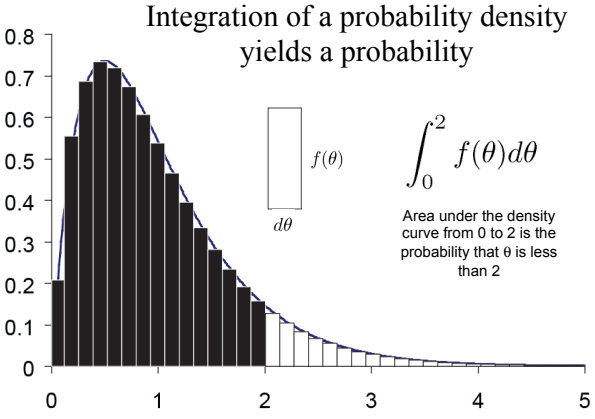


## Integration of densities



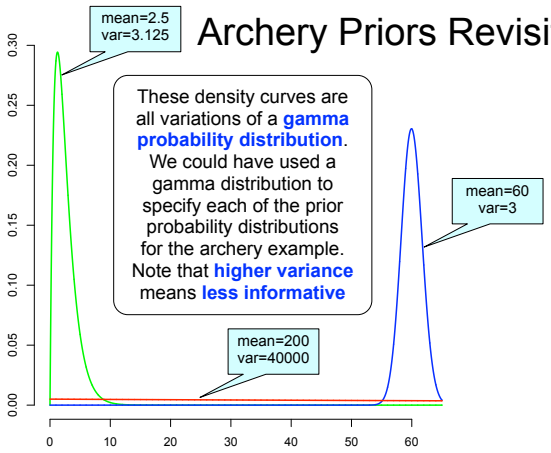

The density curve is scaled so that the value of this integral (i.e. the total area) equals 1.0

## Integration of a probability density yields a probability






## Archery Priors Revisited






## Coin-flipping

$y$ = observed number of heads
$n$ = number of flips (sample size)
$p$ = (unobserved) proportion of heads

$$\Pr(y|p) = \binom{n}{y} p^y (1-p)^{n-y} = L(p|y)$$

Note that the same formula serves as both the:
- probability of $y$ (if $p$ is fixed)
- likelihood of $p$ (if $y$ is fixed)



## Likelihood: why a new term?

| Outcome | Fair coin model | Two-heads model |
|---|---|---|
| H | 0.5 | 1.0 |
| T | 0.5 | 0.0 |
| | 1.0 | 1.0 |

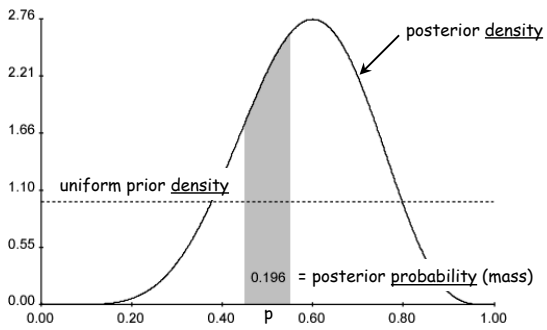
**Likelihoods** are functions of models (data fixed) Do not ordinarily sum to 1.0

**Probabilities** are functions of the data (the model is fixed) Sum to 1.0

Example usage:
- likelihood of the two-heads model
- probability of tails

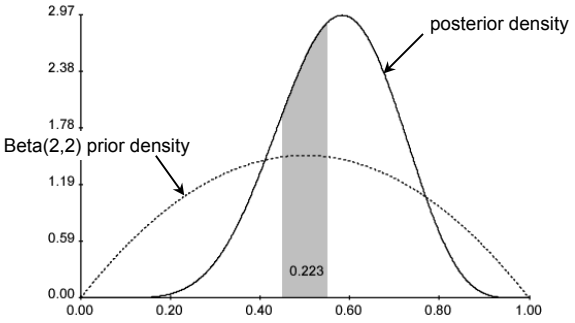


## The posterior is (almost always) more informative than the prior






## Beta(2,2) prior is vague but not flat




Posterior probability of $p$ between 0.45 and 0.55 is **0.223**

## Usually there are many parameters...

A 2-parameter example

Likelihood | Prior probability density

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

Posterior probability density

Marginal probability of data

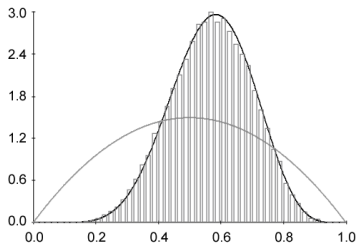
An analysis of **100 sequences** under the simplest model (JC69) requires 197 branch length parameters. The denominator is a **197-fold integral** in this case! Now consider summing over **all possible tree topologies**! It would thus be nice to avoid having to calculate the marginal probability of the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



## II. Markov chain Monte Carlo (MCMC)



## Markov chain Monte Carlo (MCMC)



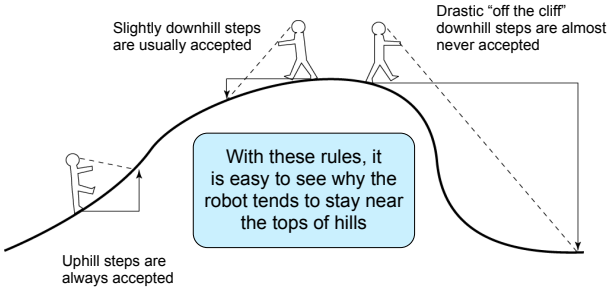

For more complex problems, we might settle for a

good approximation

to the posterior distribution

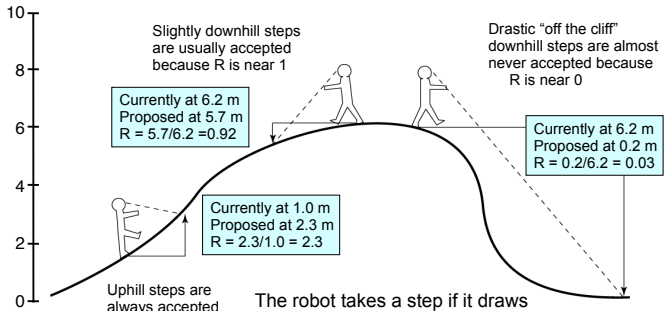


## MCMC robot's rules






## (Actual) MCMC robot rules




The robot takes a step if it draws a Uniform(0,1) random deviate that is less than or equal to R



## Cancellation of marginal likelihood

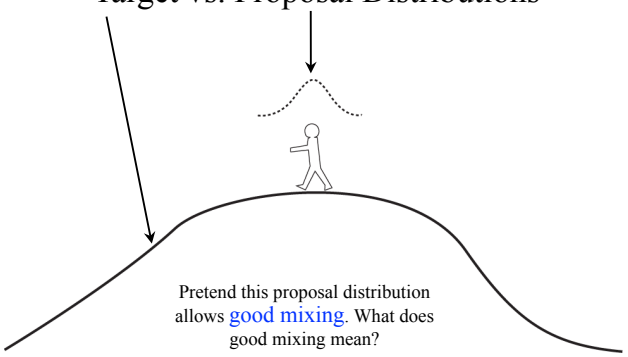
When calculating the ratio *R* of posterior densities, the marginal probability of the data cancels.

$$\frac{f(\theta^*|D)}{f(\theta|D)} = \frac{\frac{f(D|\theta^*)f(\theta^*)}{\cancel{f(D)}}}{\frac{f(D|\theta)f(\theta)}{\cancel{f(D)}}} = \frac{f(D|\theta^*)f(\theta^*)}{f(D|\theta)f(\theta)}$$

Posterior odds

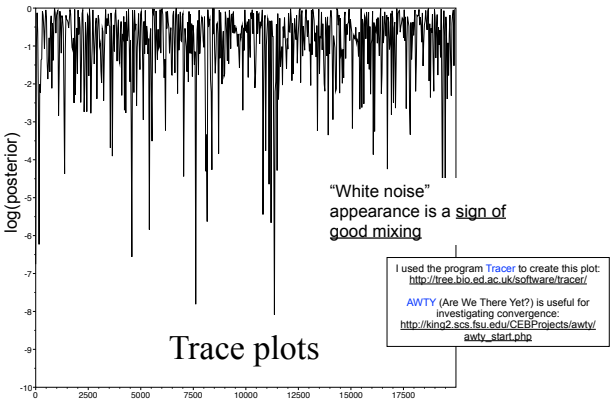
Likelihood ratio | Prior odds

# Target vs. Proposal Distributions



Pretend this proposal distribution allows good mixing. What does good mixing mean?






"White noise" appearance is a sign of good mixing

I used the program Tracer to create this 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

Trace plots



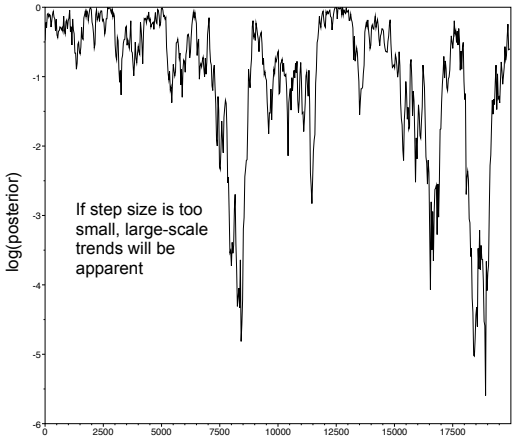
# Target vs. Proposal Distributions

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






If step size is too small, large-scale trends will be apparent



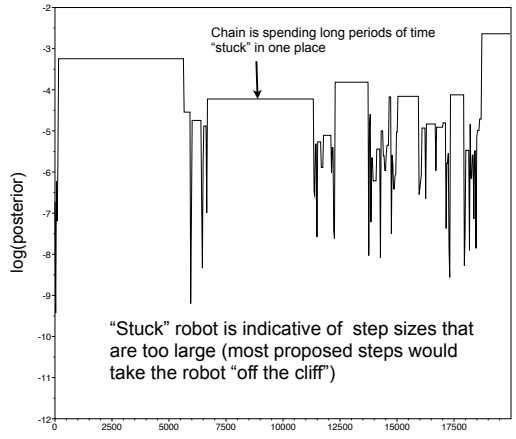
# Target vs. Proposal Distributions

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






Chain is spending long periods of time "stuck" in one place

"Stuck" robot is indicative of step sizes that are too large (most proposed steps would take the robot "off the cliff")

## MCRobot (or "MCMC Robot")

Free apps for **Windows** or **iPhone/iPad** available from http://mcmcrobot.org/

**Mac version:** some day
(but see John Huelsenbeck's iMCMC app for MacOS: http://cteg.berkeley.edu/software.html)

**Android:** some day



## Tradeoff

- Taking **big steps** helps in jumping from one "island" in the posterior density to another
- Taking **small steps** often results in better mixing
- How can we overcome this tradeoff? **MCMCMC**



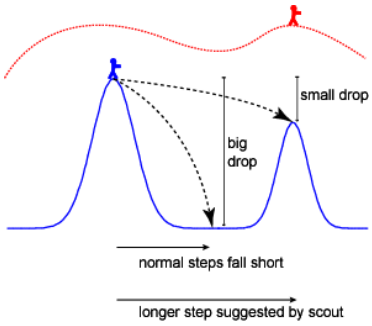
## Metropolis-coupled Markov chain Monte Carlo (MCMCMC)

- 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.).

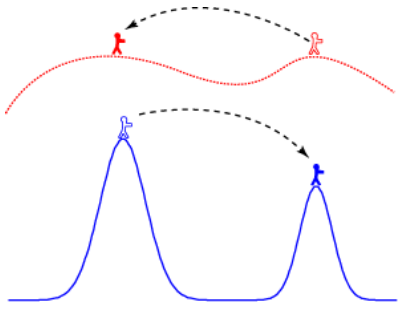


## Heated chains act as scouts for the cold chain






## Cold and hot chains swapped

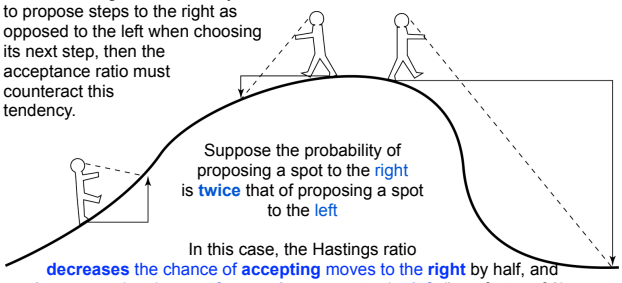




## Back to MCRobot...

# 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.



# 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

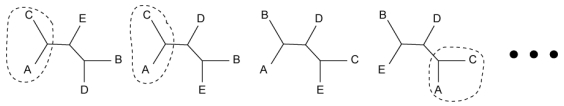
Note that if $q(\theta|\theta^*) = q(\theta^*|\theta)$, the Hastings ratio is 1



# III. Bayesian phylogenetics



# So, what's all this got to do with phylogenetics?

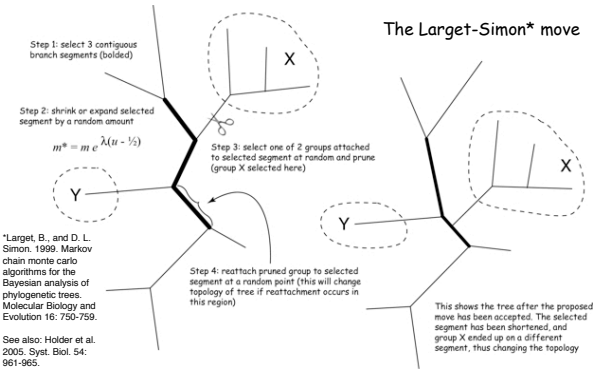


1 1 0 1

Imagine pulling out trees at random from a barrel. In the barrel, some trees are represented numerous times, while other possible trees are not present. Count 1 each time you see the split separating just A and C from the other taxa, and count 0 otherwise. Dividing by the total trees sampled approximates the true proportion of that split in the barrel.



# Moving through treespace

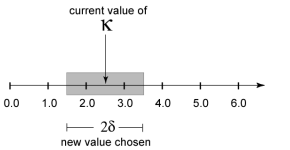


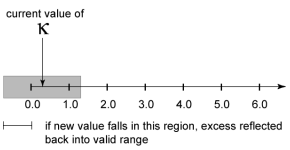

*Larget, B., and D. L. Simon. 1999. Markov chain monte carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution 16: 750-759.

See also: Holder et al. 2005. Syst. Biol. 54: 961-965.



# Moving through 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-d, κ+d)

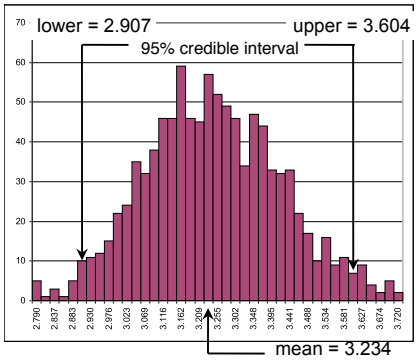
The "step size" of the MCMC robot is defined by d: a larger d means that the robot will attempt to make larger jumps on average.

## 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

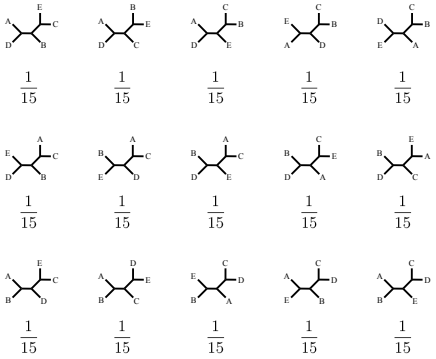


## Common Priors

- **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

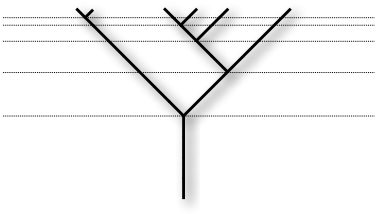


**Discrete Uniform** distribution for **topologies**



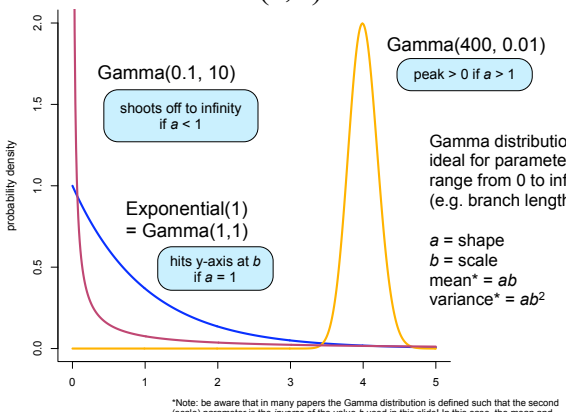



**Yule model** provides joint prior for both **topology** and **divergence times**



The rate of speciation under the Yule model ($\lambda$) 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* $\mu$ and $\sigma$...

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

X

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

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

log(X)

σ

μ

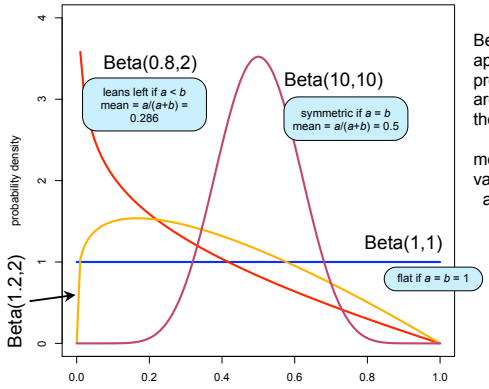
*Important*: $\mu$ *and* $\sigma$ *do* ***not*** *represent the mean and standard deviation of X: they are the mean and standard deviation of log(X)!*

To choose $\mu$ and $\sigma$ 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].

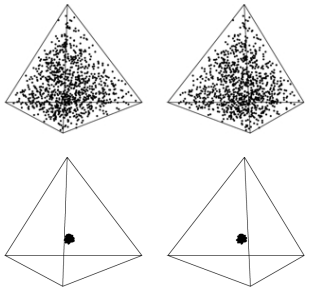
mean = a/(a+b)
variance = ab/[(a+b)²(a+b+1)]



# Dirichlet($a$,$b$,$c$,$d$) distribution

Used for nucleotide relative frequencies:

$a \rightarrow \pi_A$, $b \rightarrow \pi_C$, $c \rightarrow \pi_G$, $d \rightarrow \pi_T$



(stereo pairs)

Flat prior:
a = b = c = d = 1

(no scenario discouraged)

Informative prior:
a = b = c = d = 300

(equal frequencies strongly encouraged)

Dirichlet(a,b,c,d,e,f) used for GTR exchangeability parameters.

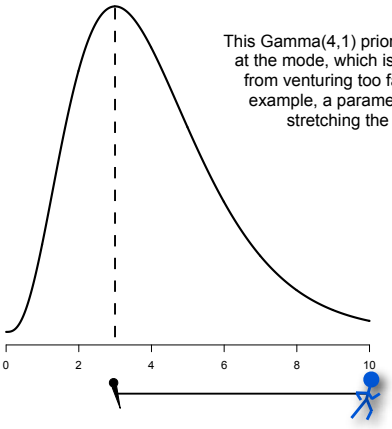
(Thanks to Mark Holder for suggesting the use of a tetrahedron)



# Prior Miscellany

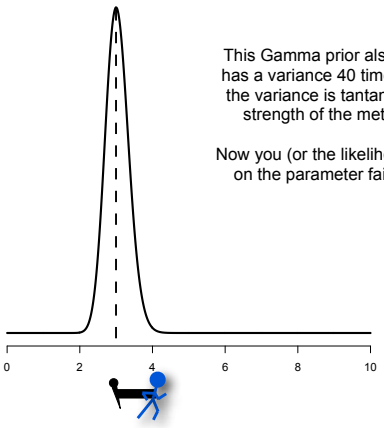
- priors as rubber bands
- running on empty
- hierarchical models
- empirical bayes





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)



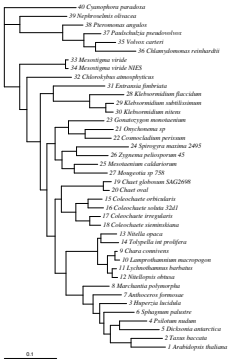

This Gamma prior also has a mode at 3, but has a variance 40 times smaller. Decreasing the variance is tantamount to increasing the strength of the metaphorical rubber band.

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



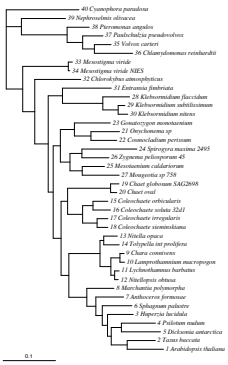
Separate priors applied to internal and external branches

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

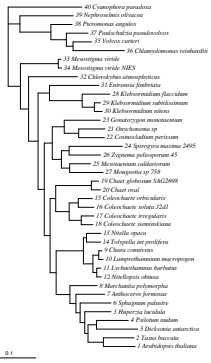




Internal branch length prior mean 0.01

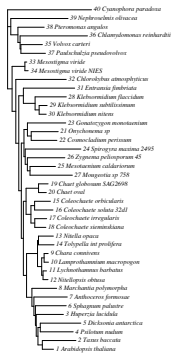
(external branch length prior mean always 0.1)





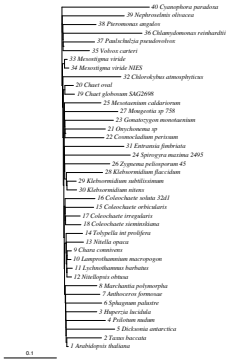
Internal branch length prior mean 0.001



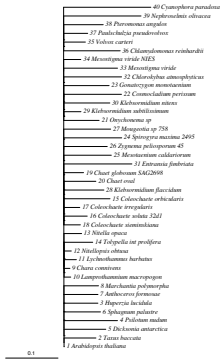


Internal branch length prior mean 0.0001





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.

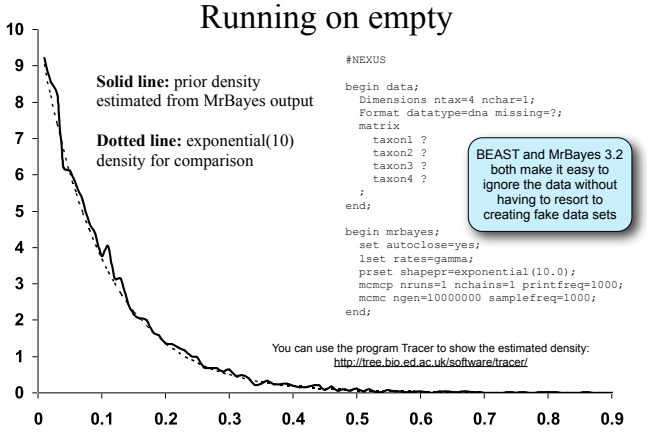


## Prior Miscellany

- 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

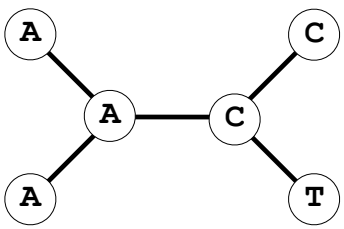
- priors as rubber bands
- running on empty
- hierarchical models
- empirical bayes



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

Prior: Exponential, mean=0.1

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






**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.

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. This uses the data twice, which is not acceptable to Bayesian purists

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

Prior: Exponential, mean=MLE

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



## V. Bayesian model selection



## AIC is not Bayesian. Why?

$$AIC = 2k - 2\log(\max L)$$

number of free (estimated) parameters — maximized log likelihood

AIC is not Bayesian because the **prior is not considered** (and the prior is an important component of a Bayesian model)

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

The **marginal likelihood** (denominator in Bayes' Rule) is commonly used for Bayesian model selection

Represents the (weighted) **average fit of the model** to the observed data (weights provided by the prior)



## An evolutionary distance example

X ——————— Y

– Let's compare models JC69 vs. K80
– Parameters:
  - ν is edge length (expected no. substitutions/site)
    – free in both JC69 and K80 models
  - κ is transition/transversion rate ratio
    – free in K80, set to 1.0 in JC69

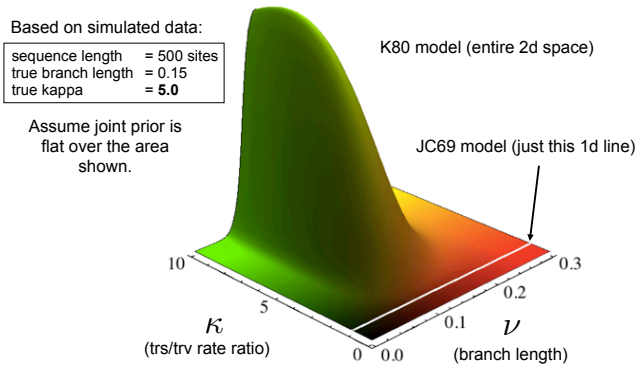


## Likelihood Surface when K80 true

Based on simulated data:

| | |
|---|---|
| sequence length | = 500 sites |
| true branch length | = 0.15 |
| true kappa | = **5.0** |

Assume joint prior is flat over the area shown.



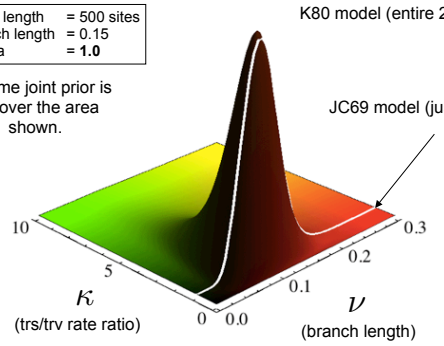

K80 wins

# Likelihood Surface when JC true

Based on simulated data:

| | |
|---|---|
| sequence length | = 500 sites |
| true branch length | = 0.15 |
| true kappa | = **1.0** |

Assume joint prior is flat over the area shown.




JC69 wins



# Harmonic mean method

$$\widehat{f(D|M)} = \frac{n}{\frac{1}{L^{(1)}} + \frac{1}{L^{(2)}} + \cdots + \frac{1}{L^{(n)}}}$$

$L^{(i)}$ = Likelihood (*not* log-likelihood) calculated for the *i*th sample from the MCMC analysis

$$\log \mathrm{BF}_{12} = \log\left(\frac{f(D|M_1)}{f(D|M_2)}\right)$$
$$= \log f(D|M_1) - \log f(D|M_2)$$

Most Bayesian programs provide the log of the harmonic mean of the sampled likelihoods for each model you run, so all you need to do is subtract.

Example: MrBayes output

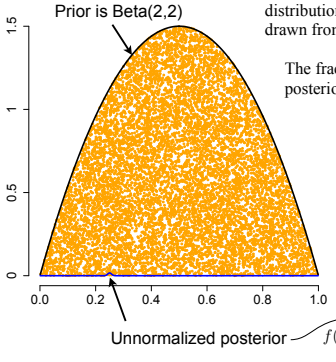
| Run | Arithmetic mean | Harmonic mean |
|---|---|---|
| 1 | -22913.52 | -22923.02 |
| 2 | -22913.52 | -22922.68 |
| TOTAL | -22913.52 | -22922.86 |

*Warning:* the harmonic mean method is **strongly biased** and **should not be used** if more accurate methods are available

Newton, M. A. and A. E. Raftery. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). J. Roy. Stat. Soc. B 56:3–48.



# Another approach




MCMC provides a way to sample from any distribution. The orange points are values of $p$ drawn from the Beta(2,2) prior.

The fraction of dots inside the unnormalized posterior is an estimate of this ratio:

$\frac{c_1}{c_0}$ ← Marginal likelihood (area under the unnormalized posterior)

1.0 (area under prior density)

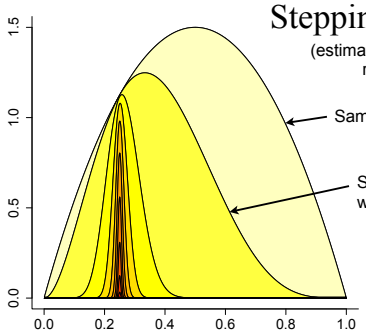
Would work better if unnorm. posterior represented a larger fraction of the area under the prior...

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



# Stepping-stone method

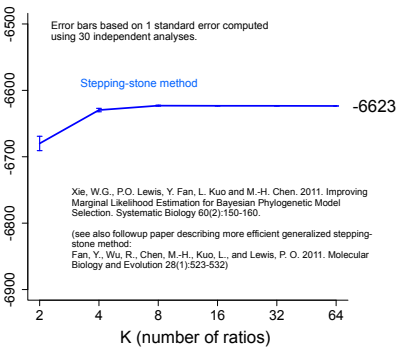
(estimates a series of ratios that each represent smaller jump)




$$\frac{c_{1.0}}{c_{0.0}} = \left(\frac{c_{1.0}}{c_{0.9}}\right)\left(\frac{c_{0.9}}{c_{0.8}}\right)\left(\frac{c_{0.8}}{c_{0.7}}\right)\left(\frac{c_{0.7}}{c_{0.6}}\right)\left(\frac{c_{0.6}}{c_{0.5}}\right)\left(\frac{c_{0.5}}{c_{0.4}}\right)\left(\frac{c_{0.4}}{c_{0.3}}\right)\left(\frac{c_{0.3}}{c_{0.2}}\right)\left(\frac{c_{0.2}}{c_{0.1}}\right)\left(\frac{c_{0.1}}{c_{0.0}}\right)$$



# How many "stepping stones" (i.e. ratios) are needed?



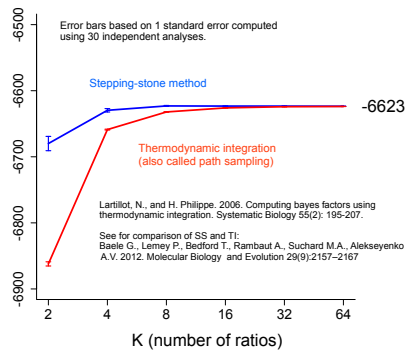

Xie, W.G., P.O. Lewis, Y. Fan, L. Kuo and M.-H. Chen. 2011. Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection. Systematic Biology 60(2):150-160.

(see also followup paper describing more efficient generalized stepping-stone method:
Fan, Y., Wu, R., Chen, M.-H., Kuo, L., and Lewis, P. O. 2011. Molecular Biology and Evolution 28(1):523-532)


- *rbc*L data
- 10 green plants
- GTR+G model
- 1000 samples/ steppingstone



# Is steppingstone sampling accurate?

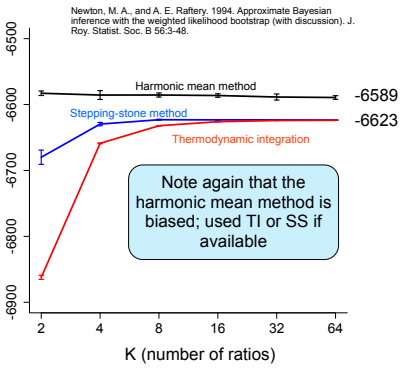



Lartillot, N., and H. Philippe. 2006. Computing bayes factors using thermodynamic integration. Systematic Biology 55(2): 195-207.

See for comparison of SS and TI:
Baele G., Lemey P., Bedford T., Rambaut A., Suchard M.A., Alekseyenko A.V. 2012. Molecular Biology and Evolution 29(9):2157–2167


- *rbc*L data
- 10 green plants
- GTR+G model
- 1000 samples/ steppingstone

# How about the harmonic mean method?





- *rbc*L data
- 10 green plants
- GTR+G model
- 1000 samples/ steppingstone



Konec