

Bayesian Phylogenetics

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

25 January 2013
Workshop on Molecular Evolution
Český Krumlov

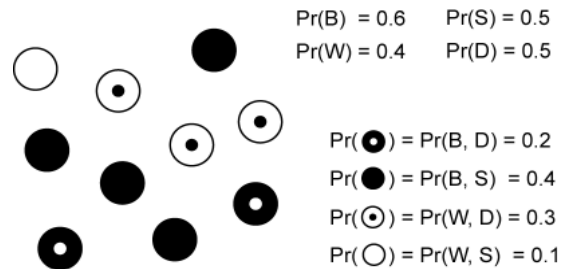
An Introduction to Bayesian Phylogenetics

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

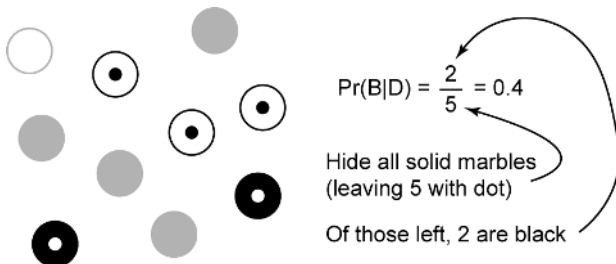
I. Bayesian inference in general

Joint probabilities

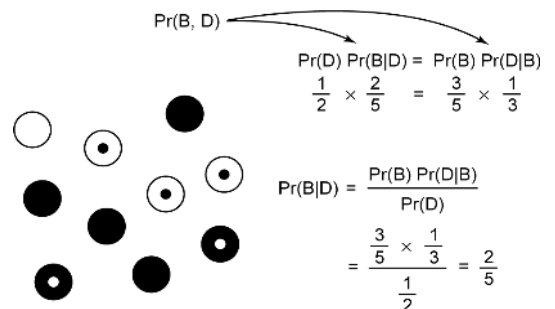
B = Black S = Solid
W = White D = Dotted



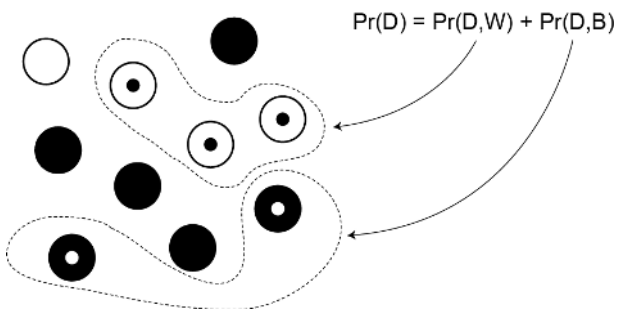
Conditional probabilities



Bayes' rule



Probability of "Dotted"



7

Bayes' rule (cont.)

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

$\Pr(D)$ 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 $\Pr(B|D) + \Pr(W|D) = 1.0$

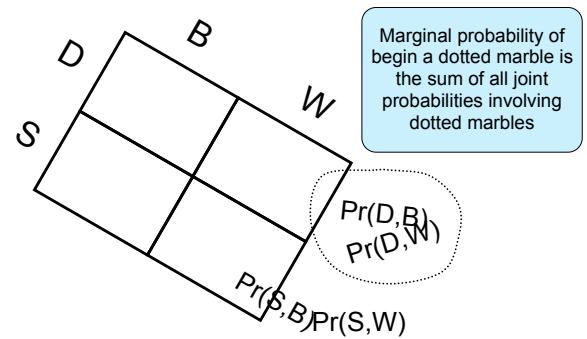
8

Joint probabilities

	B	W
D	$\Pr(D,B)$	$\Pr(D,W)$
S	$\Pr(S,B)$	$\Pr(S,W)$

9

Marginalizing over colors



10

Marginal probabilities

	B	W
D		
S		

Marginal probability of being dotted: $\Pr(D,B) + \Pr(D,W)$

Marginal probability of being solid: $\Pr(S,B) + \Pr(S,W)$

11

Marginalizing over "dottedness"

	B	W
D		
S		

$\Pr(D,B)$ $\Pr(D,W)$

$\Pr(S,B)$ $\Pr(S,W)$

Marginal probability of being a white marble

12

Bayes' rule (cont.)

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

13

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**)

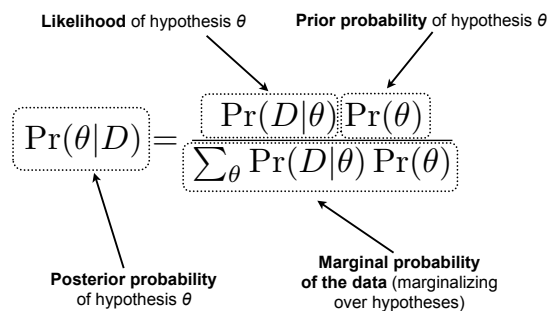
θ refers to one or more "unobservables"

(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)

14

Bayes' rule in statistics



15

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

16

The prior can be your friend

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

$$\Pr(+|\text{disease}) = 0.99$$

$$\Pr(+|\text{healthy}) = 0.01$$

datum hypothesis

Suppose further I **test positive** for the disease. (Note that we do not need to consider the case of a negative test result.)
How worried should I be?

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."

17

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:

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

18

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?

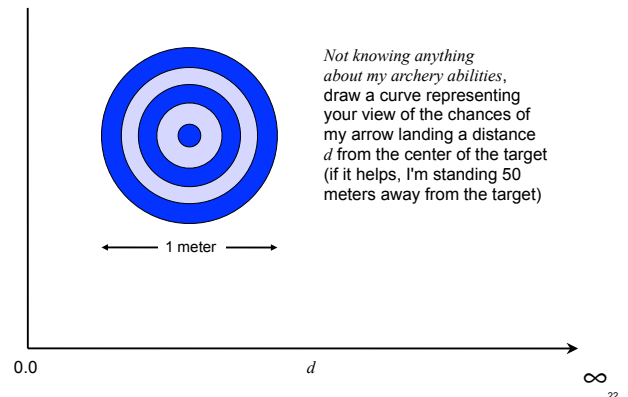
Bayes' rule: continuous case

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

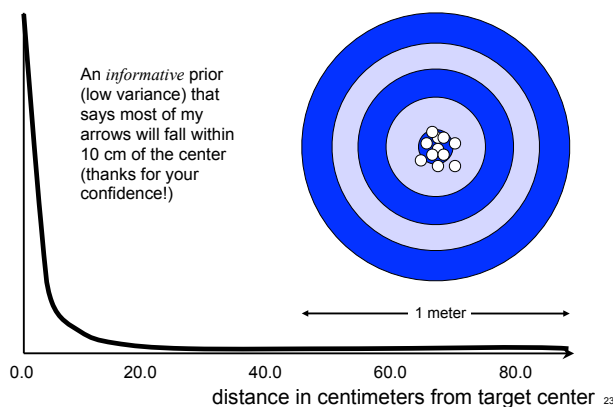
Labels in the diagram:

- Likelihood: $f(D|\theta)$
- Prior probability density: $f(\theta)$
- Posterior probability density: $f(\theta|D)$
- Marginal probability of the data: $\int f(D|\theta)f(\theta)d\theta$

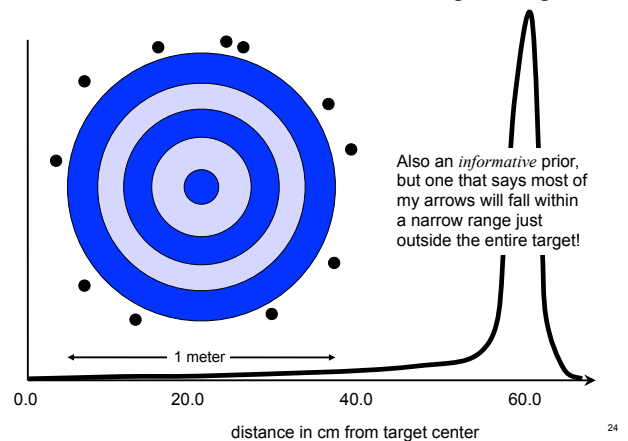
If you had to guess...



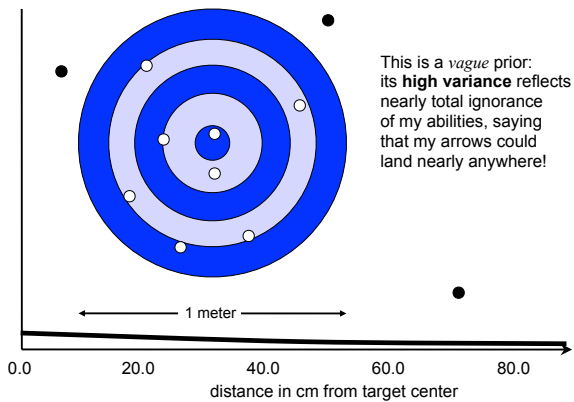
Case 1: assume I have talent



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

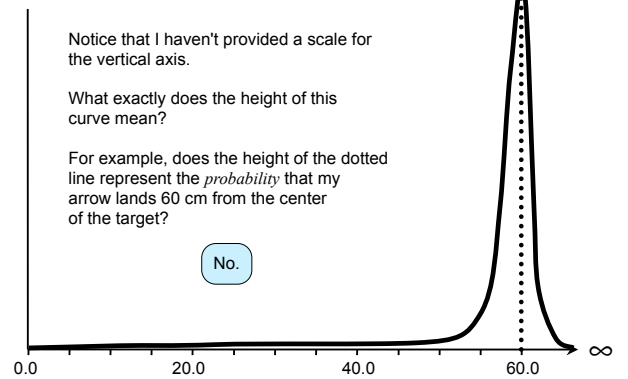


Case 3: assume I have no talent



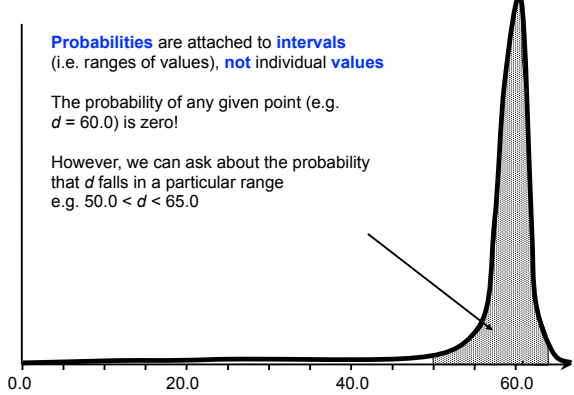
25

A matter of scale



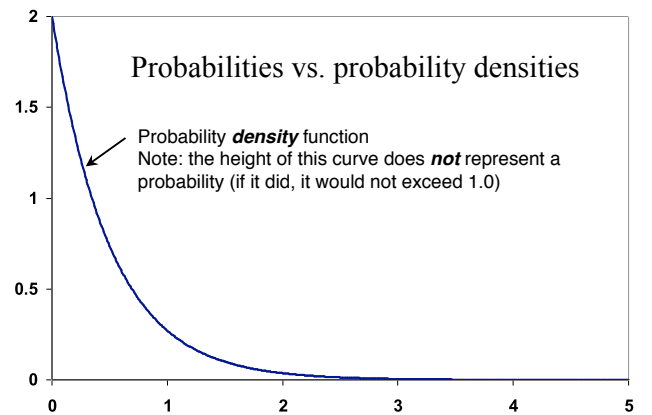
26

Probabilities are associated with intervals



27

Probabilities vs. probability densities



28

Densities of various substances

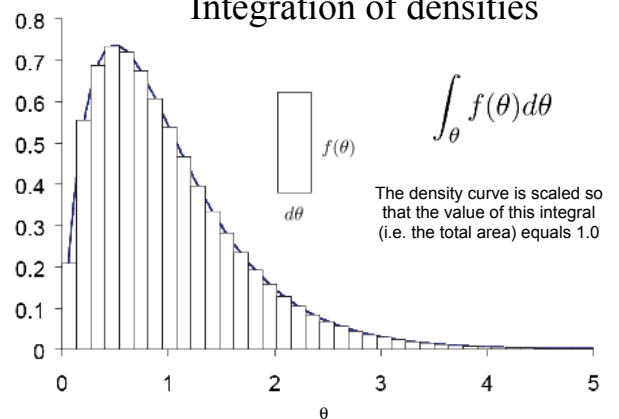
Substance	Density (g/cm ³)
Cork	0.24
Aluminum	2.70
Gold	19.30

Density does not equal mass
mass = density $\times$ volume

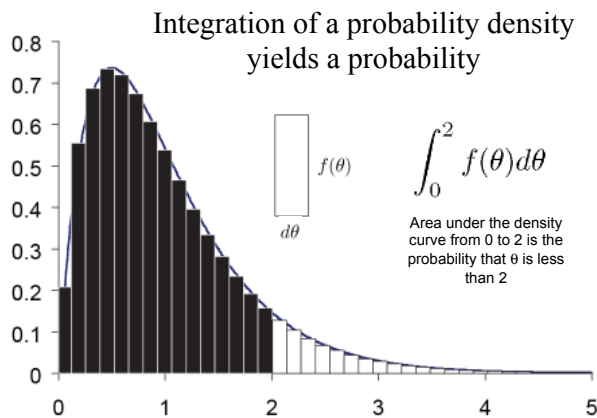
Note: *volume* is appropriate for objects of dimension 3 or higher
For 2-dimensions, *area* takes the place of volume
For 1-dimension, *linear distance* replaces volume.

29

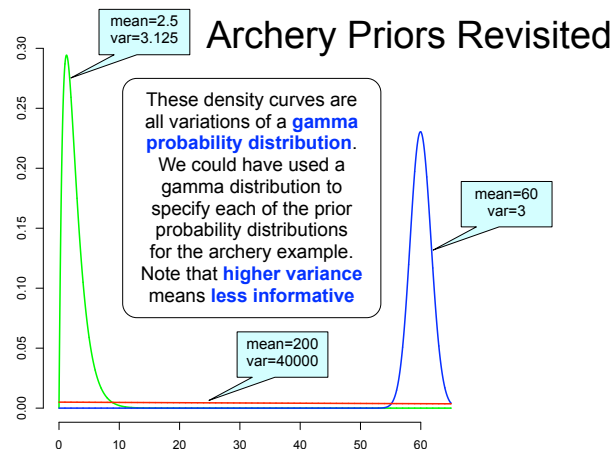
Integration of densities



30



31



32

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)

33

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

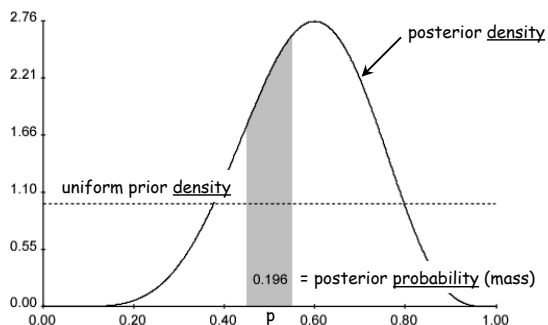
Probabilities are functions of the data (the model is fixed)
Sum to 1.0

Example usage:

- likelihood of the two-heads model
- probability of tails

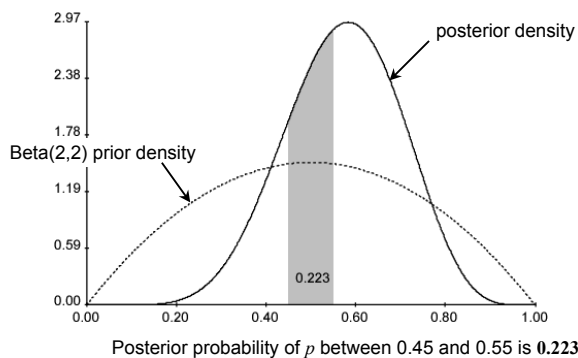
34

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



35

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



36

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

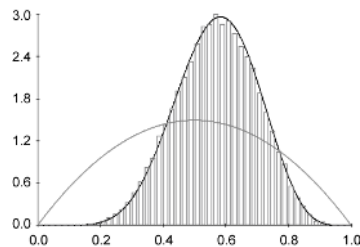
An analysis of **100 sequences** under the simplest model (JC69) requires 197 branch length parameters. The denominator is a **197-fold integral** in this case! Now consider summing over **all possible tree topologies**! It would thus be nice to avoid having to calculate the marginal probability of the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

37

II. Markov chain Monte Carlo (MCMC)

38

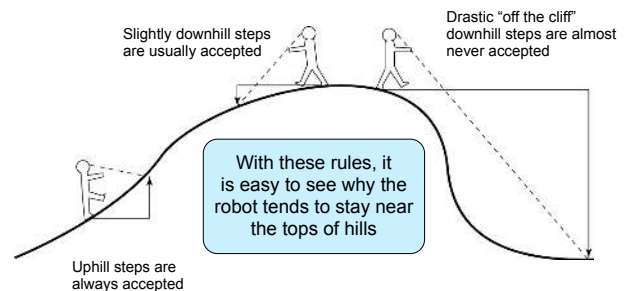
Markov chain Monte Carlo (MCMC)



For more complex problems, we might settle for a **good approximation** to the posterior distribution

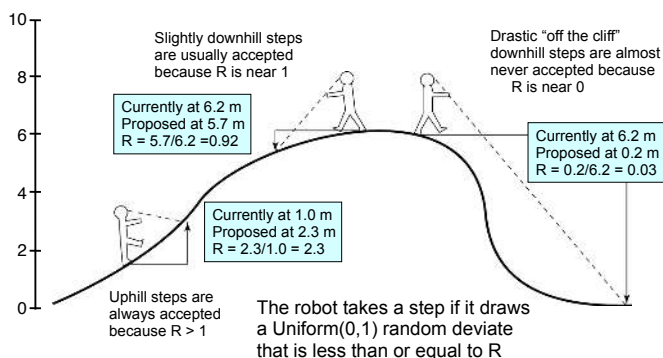
39

MCMC robot's rules



40

(Actual) MCMC robot rules



41

Cancellation of marginal likelihood

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

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

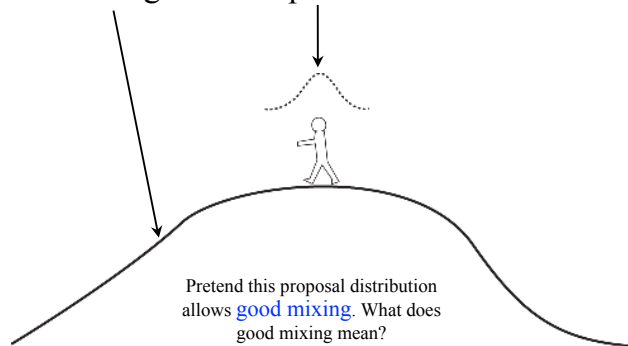
Posterior odds

Likelihood ratio

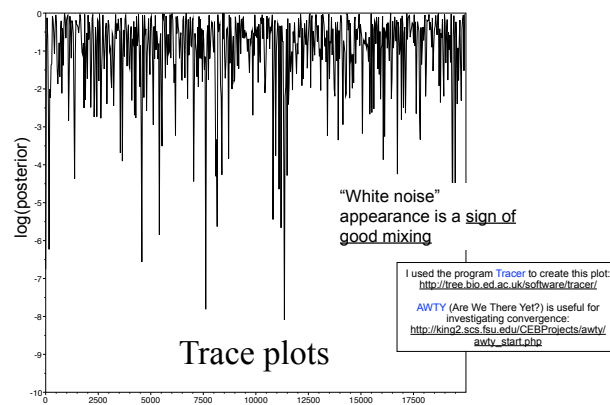
Prior odds

42

Target vs. Proposal Distributions

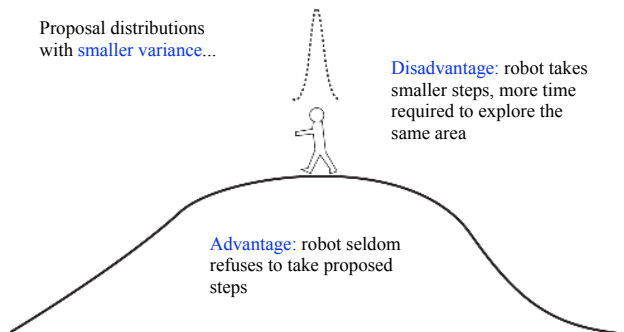


43

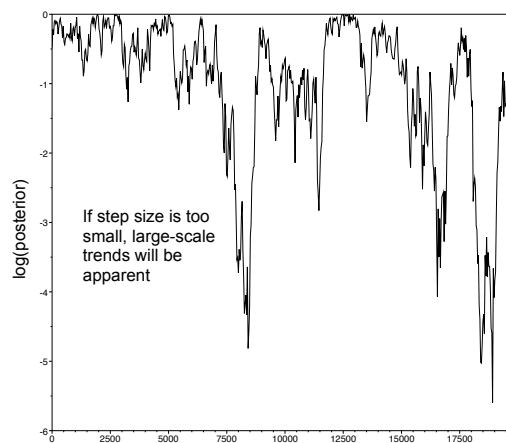


44

Target vs. Proposal Distributions

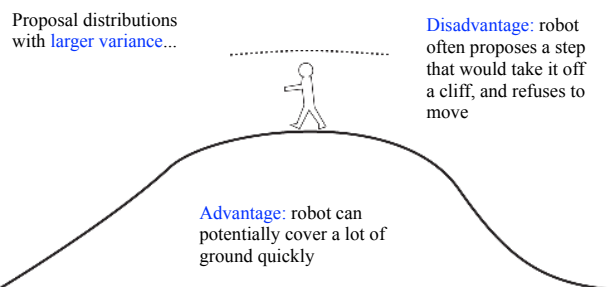


45

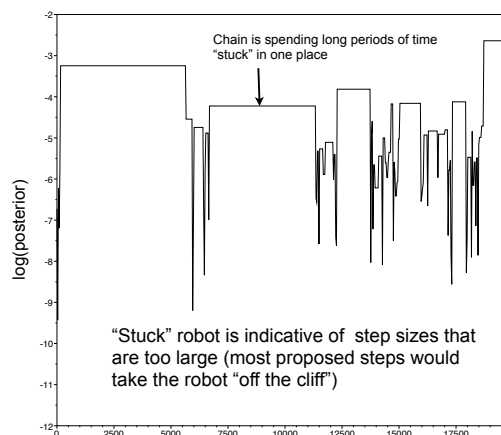


46

Target vs. Proposal Distributions



47



48

Tradeoff

MCRobot (or "MCMC Robot")

Free apps for **Windows** or **iPhone/iPad** available from <http://mcmicrobot.org/>

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

Android: some day

- 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**

49

50

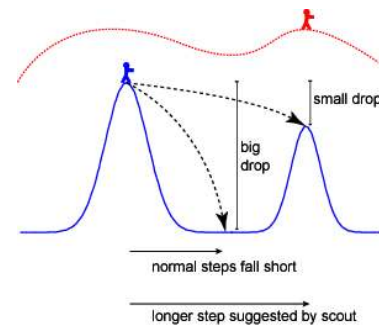
Metropolis-coupled Markov chain Monte Carlo (MCMCMC)

- 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.).

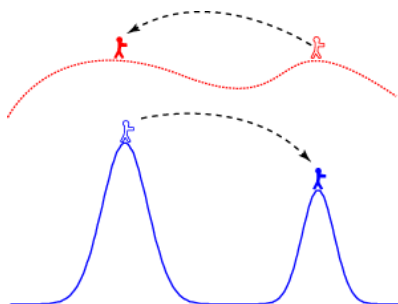
51

Heated chains act as scouts for the cold chain



52

Cold and hot chains swapped



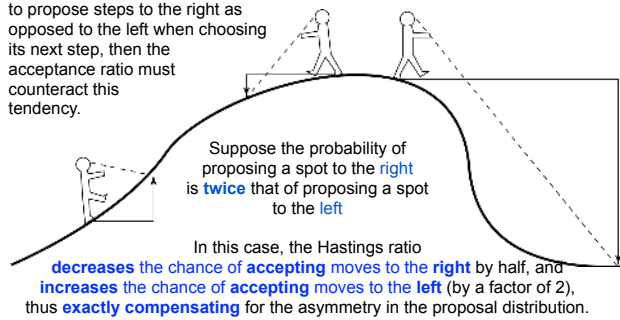
53

Back to MCRobot...

54

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.

55

Hastings Ratio

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

Acceptance ratio

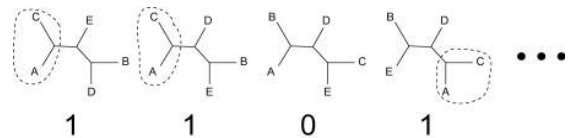
Posterior ratio

Hastings ratio

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

56

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



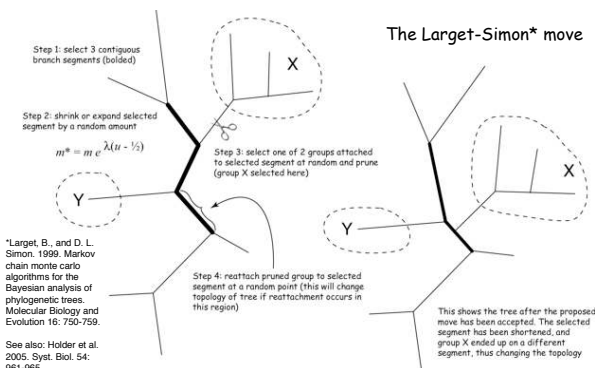
III. Bayesian phylogenetics

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**.

57

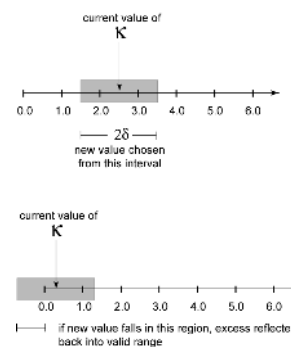
58

Moving through treespace



59

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 $(\kappa-d, \kappa+d)$

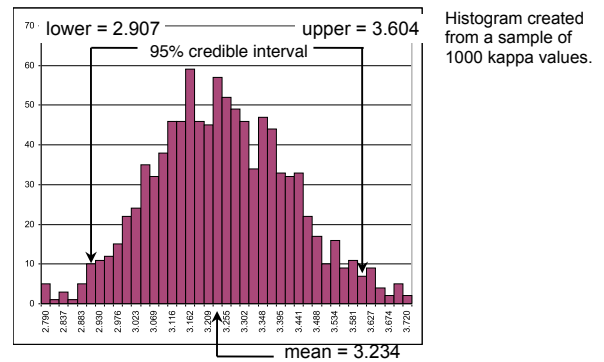
The "step size" of the MCMC robot is defined by d : a larger d means that the robot will attempt to make larger jumps on average.

60

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. Largert-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



61

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

62

Common Priors

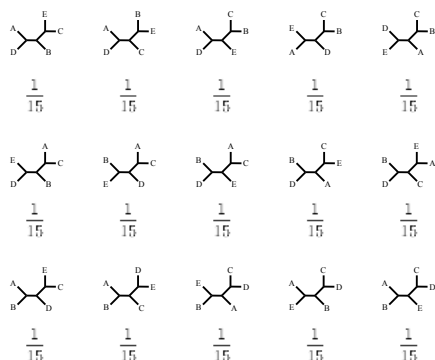
- **Discrete uniform** for topologies
 - exceptions becoming more common
- **Beta** for proportions
- **Gamma** or **Log-normal** for branch lengths and other parameters with support $[0, \infty)$
 - Exponential is common special case of the gamma distribution
- **Dirichlet** for state frequencies and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

IV. Prior distributions

63

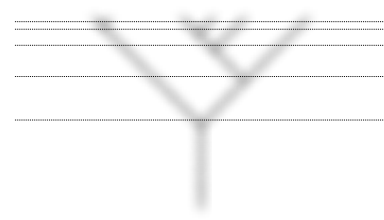
64

Discrete Uniform distribution for topologies



65

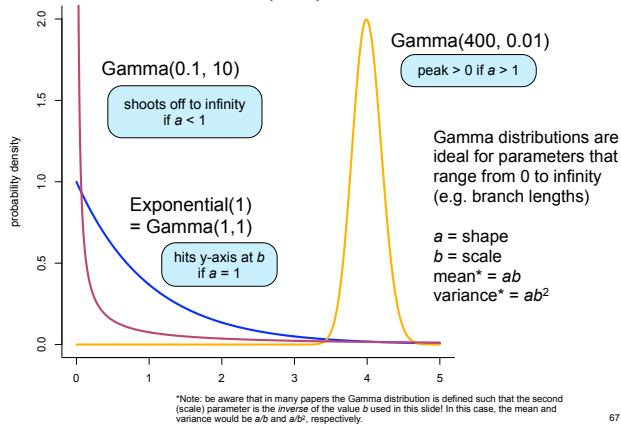
Yule model provides joint prior for both **topology** and **divergence times**



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.

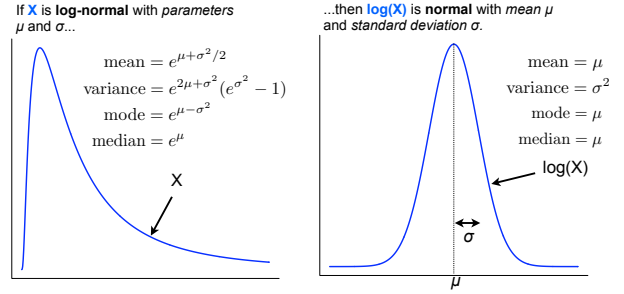
66

Gamma(a, b) distributions



67

Log-normal distribution



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

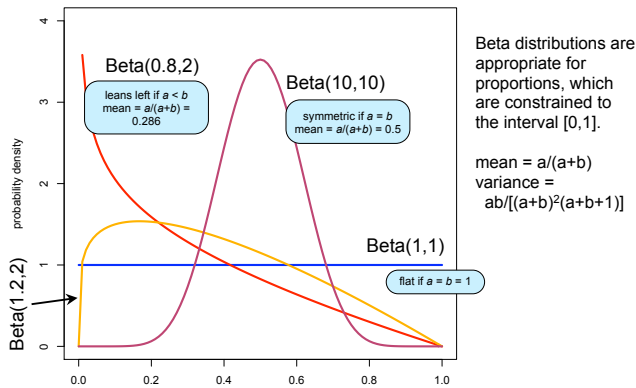
To choose μ and σ to yield a particular mean (m) and variance (v) for X , use these formulas:

$$\mu = \log(m^2) - \log(m) - \frac{\log(v + m^2) - \log(m^2)}{2}$$

$$\sigma^2 = \log(v + m^2) - \log(m^2)$$

68

Beta(a, b) gallery

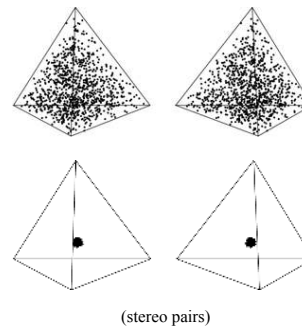


69

Dirichlet(a, b, c, d) distribution

Used for nucleotide relative frequencies:

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



Flat prior:

$$a = b = c = d = 1$$

(no scenario discouraged)

Informative prior:

$$a = b = c = d = 300$$

(equal frequencies strongly encouraged)

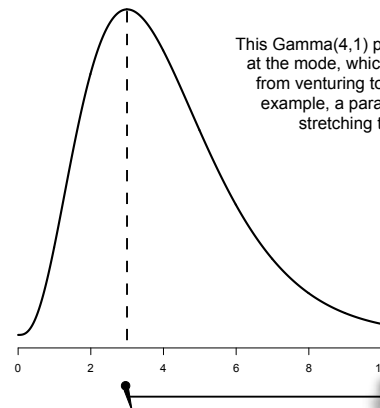
Dirichlet(a, b, c, d, e, f) used for GTR exchangeability parameters.

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

70

Prior Miscellany

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

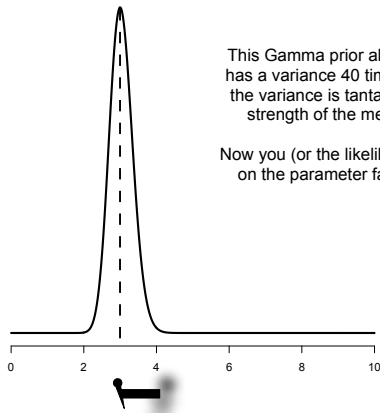


This Gamma(4, 1) prior ties down its parameter at the mode, which is at 3, and discourages it from venturing too far in either direction. For example, a parameter value of 10 would be stretching the rubber band fairly tightly

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

71

72



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

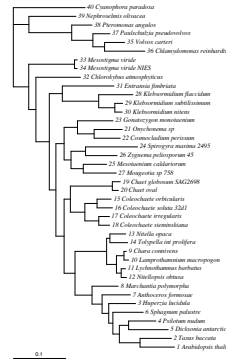
Example: Internal Branch Length Priors

Separate priors applied to internal and external branches

External branch length prior is exponential with mean 0.1

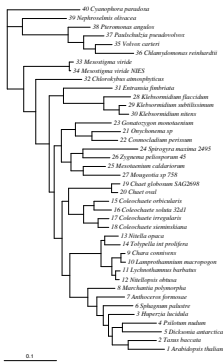
Internal branch length prior is exponential with mean 0.1

This is a reasonably vague internal branch length prior



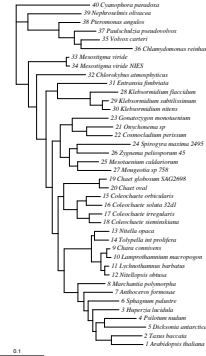
73

74



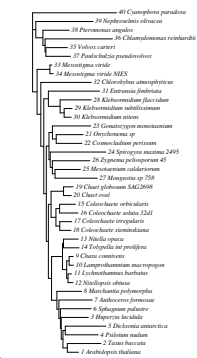
Internal branch length prior mean 0.01
(external branch length prior mean always 0.1)

75



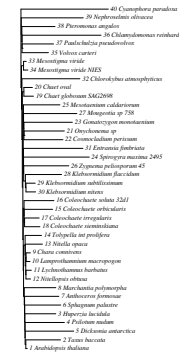
Internal branch length prior mean 0.001

76



Internal branch length prior mean 0.0001

77



Internal branch length prior mean 0.00001

78

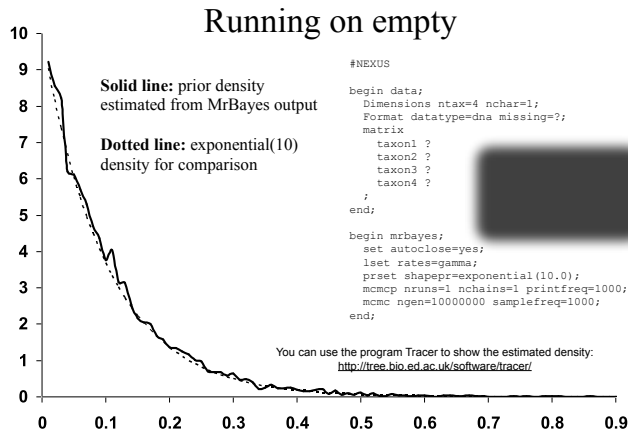
40) *Cymophora parvula*
 39) *Hypocistis albicoma*
 38) *Peromyscus atropis*
 37) *Psittacus parvirostris*
 36) *Vireo carolin*
 35) *Chrysomitris tristis*
 34) *Myiarchus cinerascens*
 33) *Myiarchus cinerascens*
 32) *Chrysomitris tristis*
 31) *Chrysomitris tristis*
 30) *Chrysomitris tristis*
 29) *Chrysomitris tristis*
 28) *Chrysomitris tristis*
 27) *Chrysomitris tristis*
 26) *Chrysomitris tristis*
 25) *Chrysomitris tristis*
 24) *Chrysomitris tristis*
 23) *Chrysomitris tristis*
 22) *Chrysomitris tristis*
 21) *Chrysomitris tristis*
 20) *Chrysomitris tristis*
 19) *Chrysomitris tristis*
 18) *Chrysomitris tristis*
 17) *Chrysomitris tristis*
 16) *Chrysomitris tristis*
 15) *Chrysomitris tristis*
 14) *Chrysomitris tristis*
 13) *Chrysomitris tristis*
 12) *Chrysomitris tristis*
 11) *Chrysomitris tristis*
 10) *Chrysomitris tristis*
 9) *Chrysomitris tristis*
 8) *Chrysomitris tristis*
 7) *Chrysomitris tristis*
 6) *Chrysomitris tristis*
 5) *Chrysomitris tristis*
 4) *Chrysomitris tristis*
 3) *Chrysomitris tristis*
 2) *Chrysomitris tristis*
 1) *Chrysomitris tristis*

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

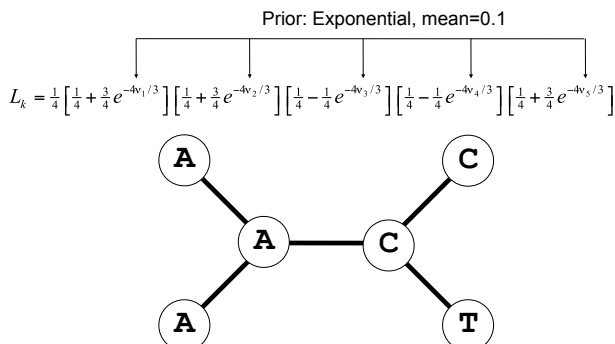


Prior Miscellany

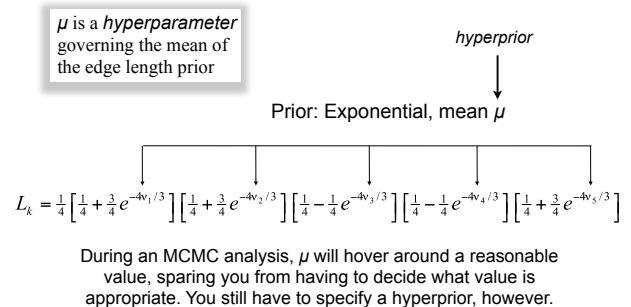
- priors as rubber bands
- running on empty
- hierarchical models
- empirical bayes



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



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



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

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

85

86

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)

87

88

V. Bayesian model selection

An evolutionary distance example

X ————— Y

– Let's compare models JC69 vs. K80

– Parameters:

- v 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

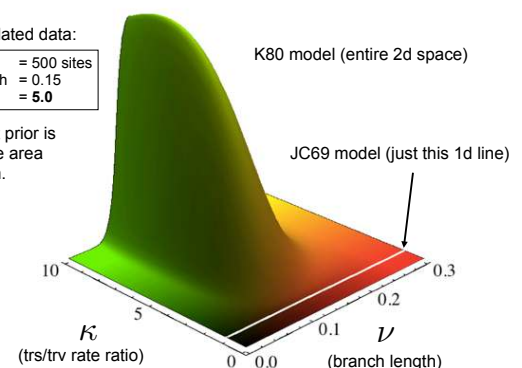
Based on simulated data:

sequence length = 500 sites
true branch length = 0.15
true kappa = 5.0

K80 model (entire 2d space)

Assume joint prior is flat over the area shown.

JC69 model (just this 1d line)



K80 wins

89

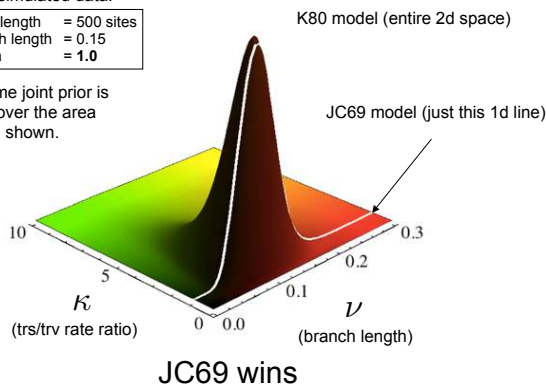
90

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.



Harmonic mean method

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

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

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

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.

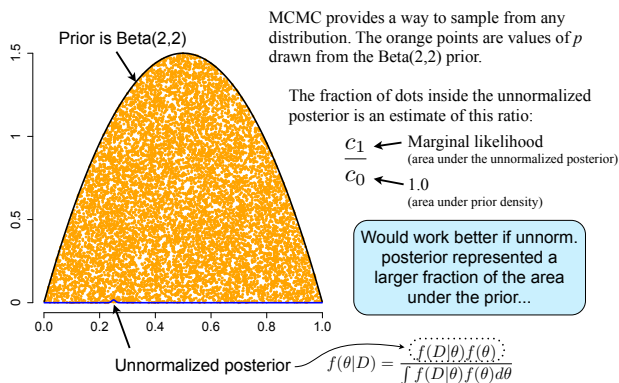
Run	Arithmetic mean	Harmonic mean
1	-22913.52	-22923.02
2	-22913.52	-22922.68
TOTAL	-22913.52	-22922.86

Example:
MrBayes
output

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

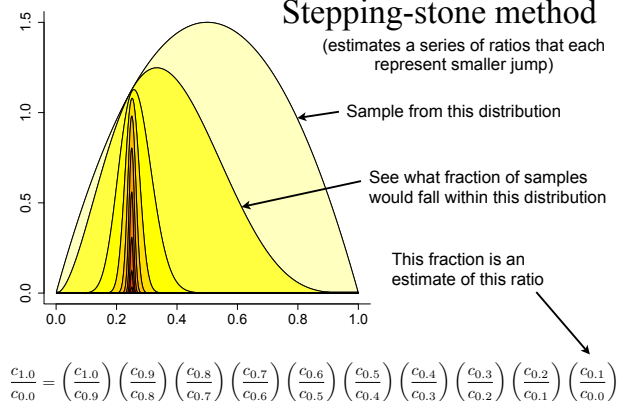
Newton, M. A. and A. E. Raftery. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *J. Roy. Stat. Soc. B* 56:3-48.

Another approach

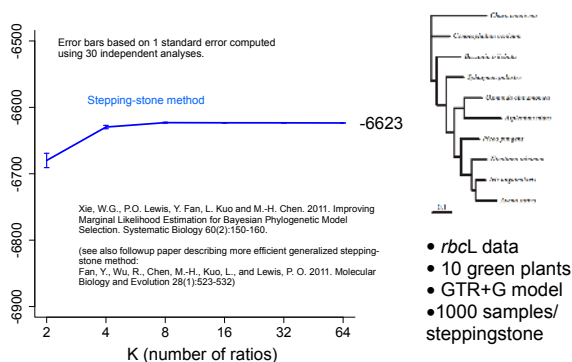


Stepping-stone method

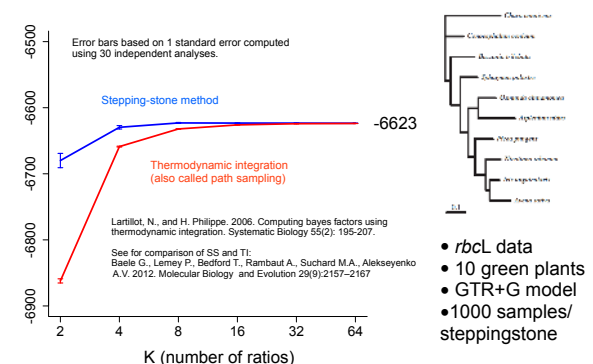
(estimates a series of ratios that each represent smaller jump)



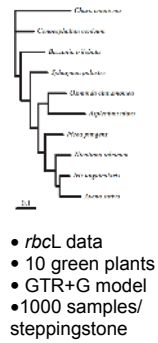
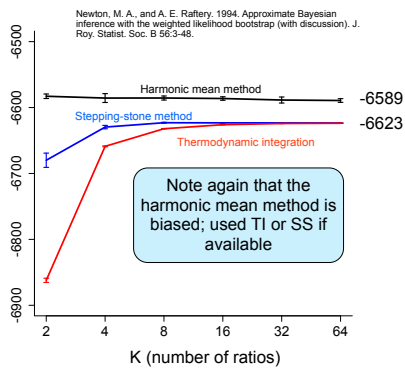
How many “stepping stones” (i.e. ratios) are needed?



Is steppingstone sampling accurate?



How about the harmonic mean method?



Konec