Bayesian Phylogenetics

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

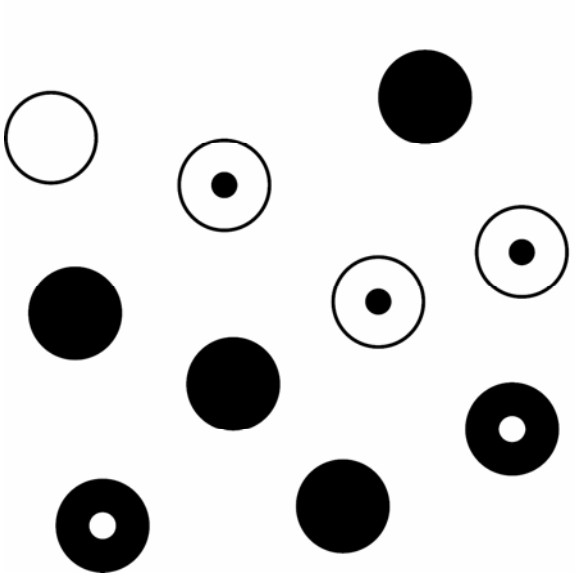
Woods Hole Molecular Evolution Workshop,
July 27, 2006

An Introduction to Bayesian Phylogenetics

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

I. Bayesian inference in general

Joint probabilities



Pr(B) = 0.6 Pr(S) = 0.5

Pr(W) = 0.4 Pr(D) = 0.5

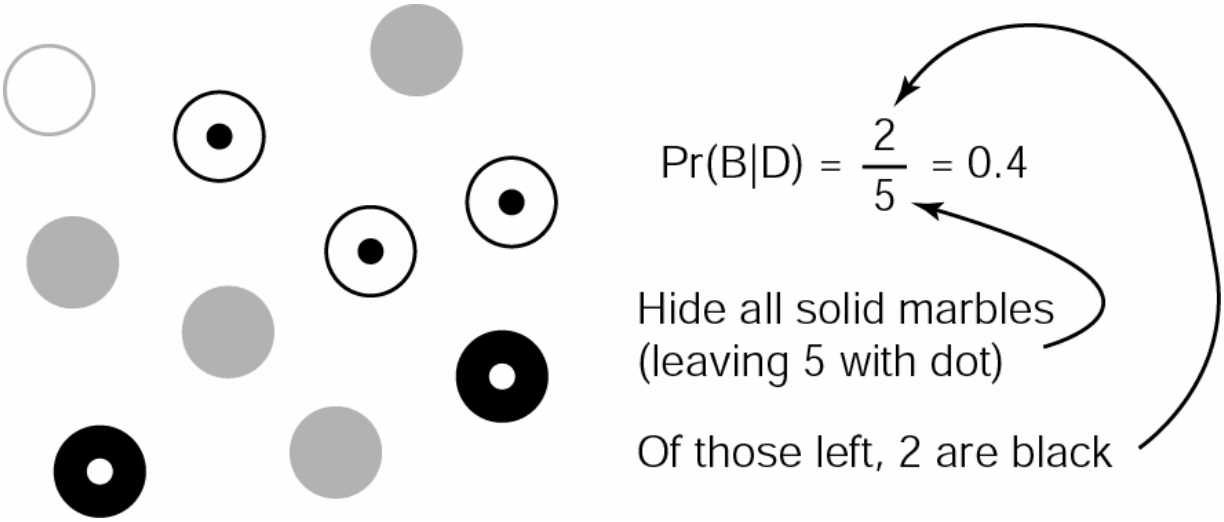
Pr( ) = Pr(B, D) = 0.2

Pr( ) = Pr(B, S) = 0.4

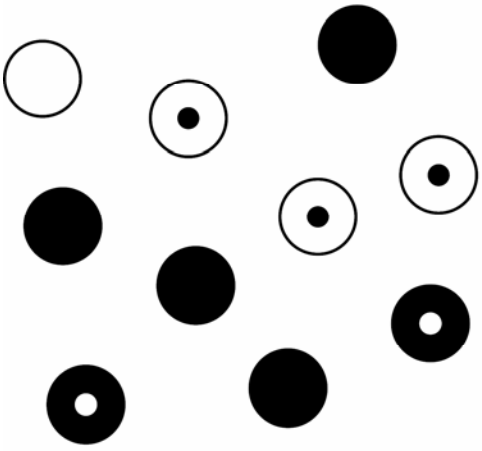
Pr( ) = Pr(W, D) = 0.3

Pr( ) = Pr(W, S) = 0.1

Conditional probabilities



Bayes' rule



$$\Pr(B, D)$$

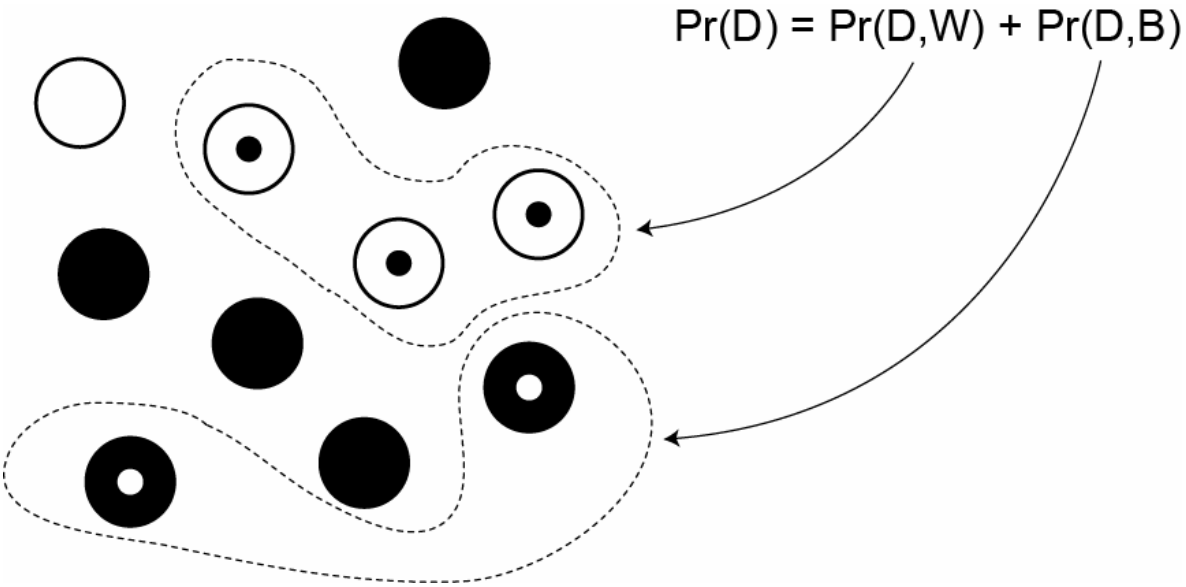
$$\Pr(D)\,\Pr(B|D) = \Pr(B)\,\Pr(D|B)$$

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

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

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

Probability of "Dotted"



Bayes' rule (cont.)

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

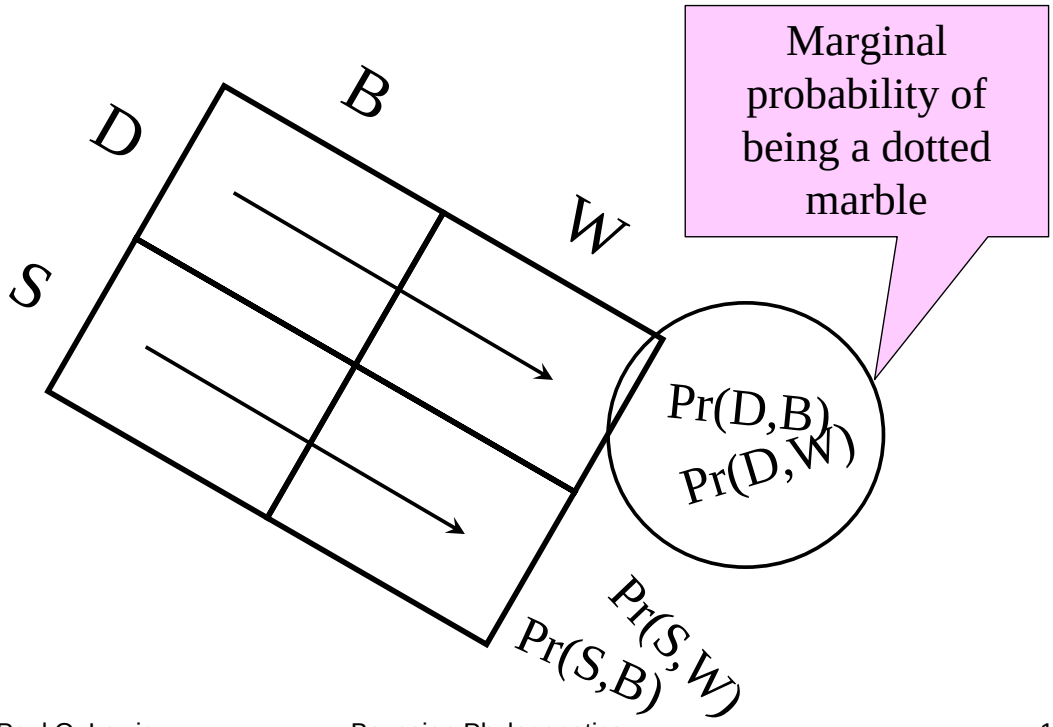
$$= \frac{\Pr(B)\Pr(D|B)}{\Pr(D,B)+\Pr(D,W)}$$

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

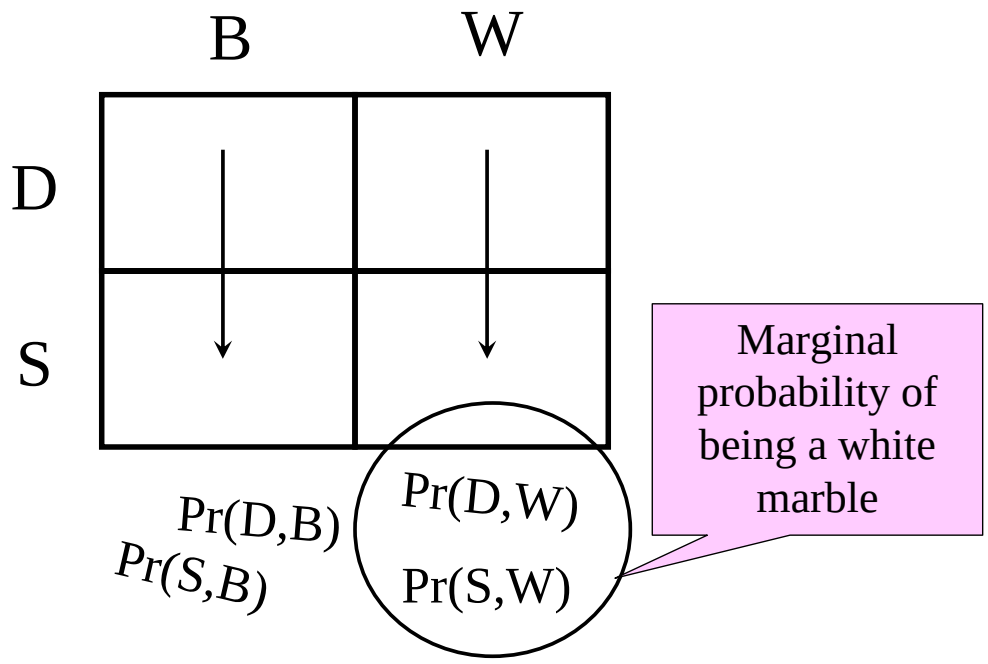
Joint probabilities

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

Marginalizing over colors



Marginalizing over "dottedness"



Bayes' rule (cont.)

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

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

$$= \frac{\Pr(B)\Pr(D|B)}{\sum_{\theta\in\{B,W\}}\Pr(\theta)\Pr(D|\theta)}$$

Bayes' rule in Statistics

Likelihood of hypothesis θ

Prior probability of hypothesis θ

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

Posterior probability of hypothesis θ

Marginal probability of the data

Simple (albeit silly) paternity example

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

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

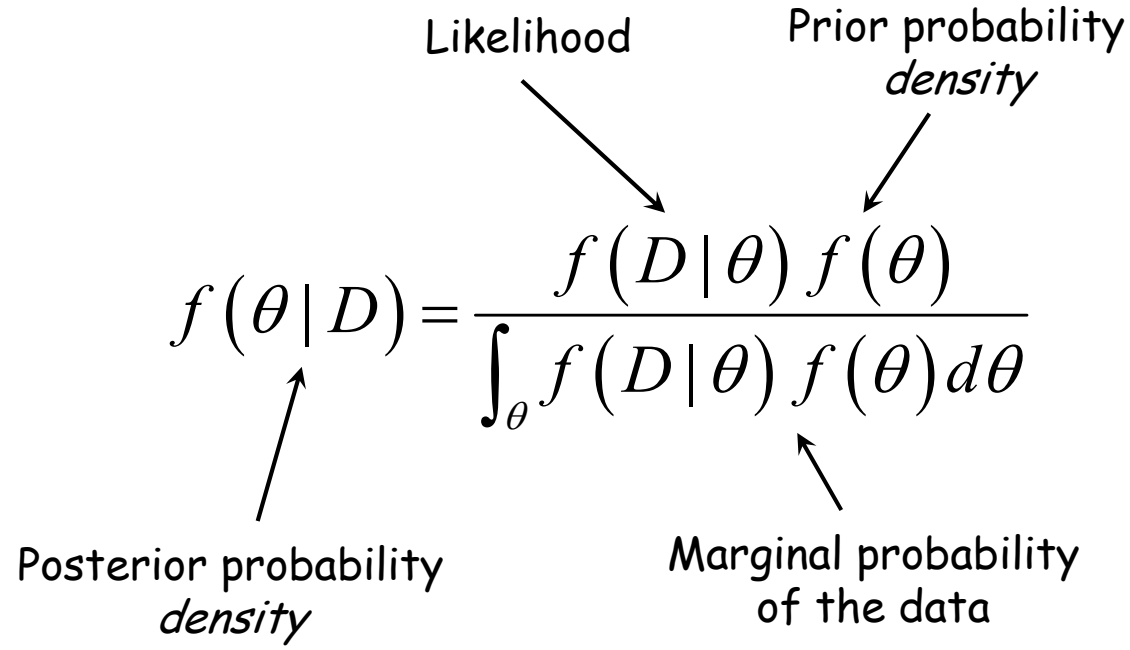
Bayes' rule in Statistics

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

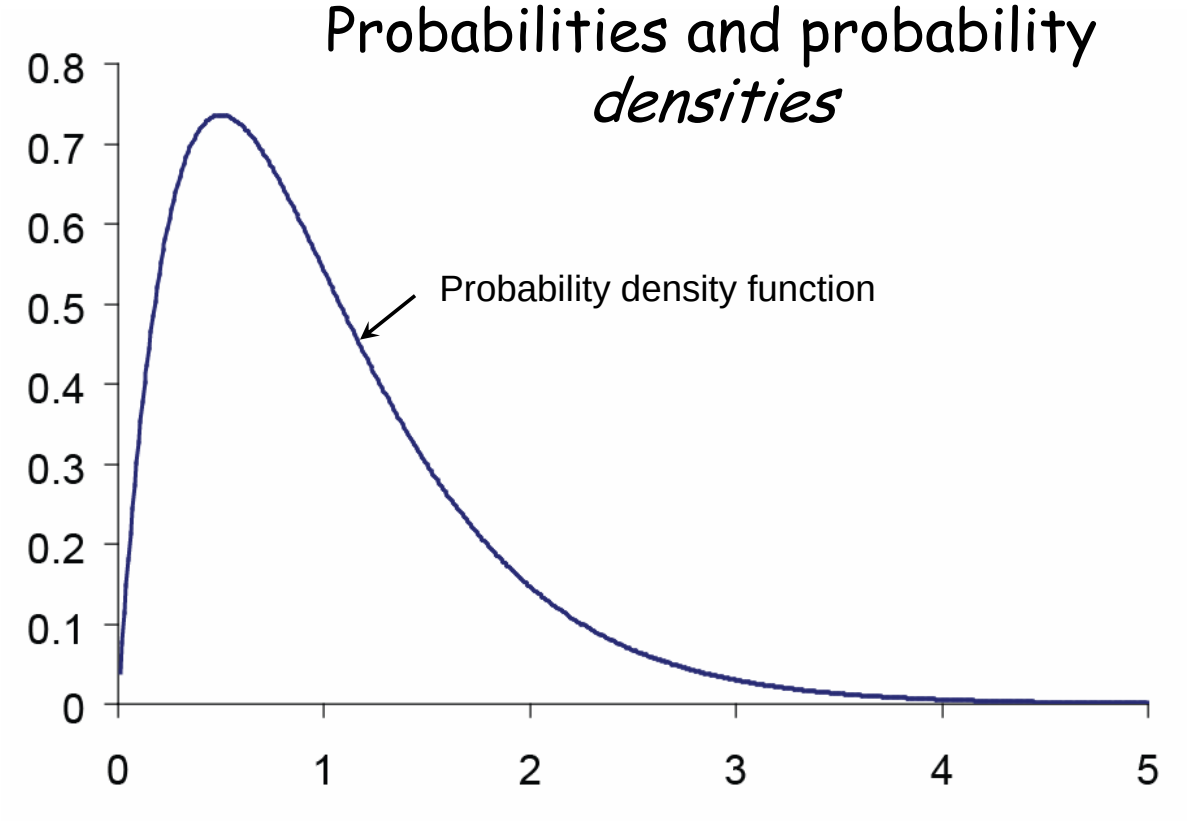
θ refers to one or more "unobservables" (i.e. parameters of the model):

- a tree model (i.e. tree topology)
- a substitution model (e.g. JC, F84, GTR, etc.)
- a parameter of a substitution model (e.g. a branch length, a base frequency, transition/transversion rate ratio, etc.)

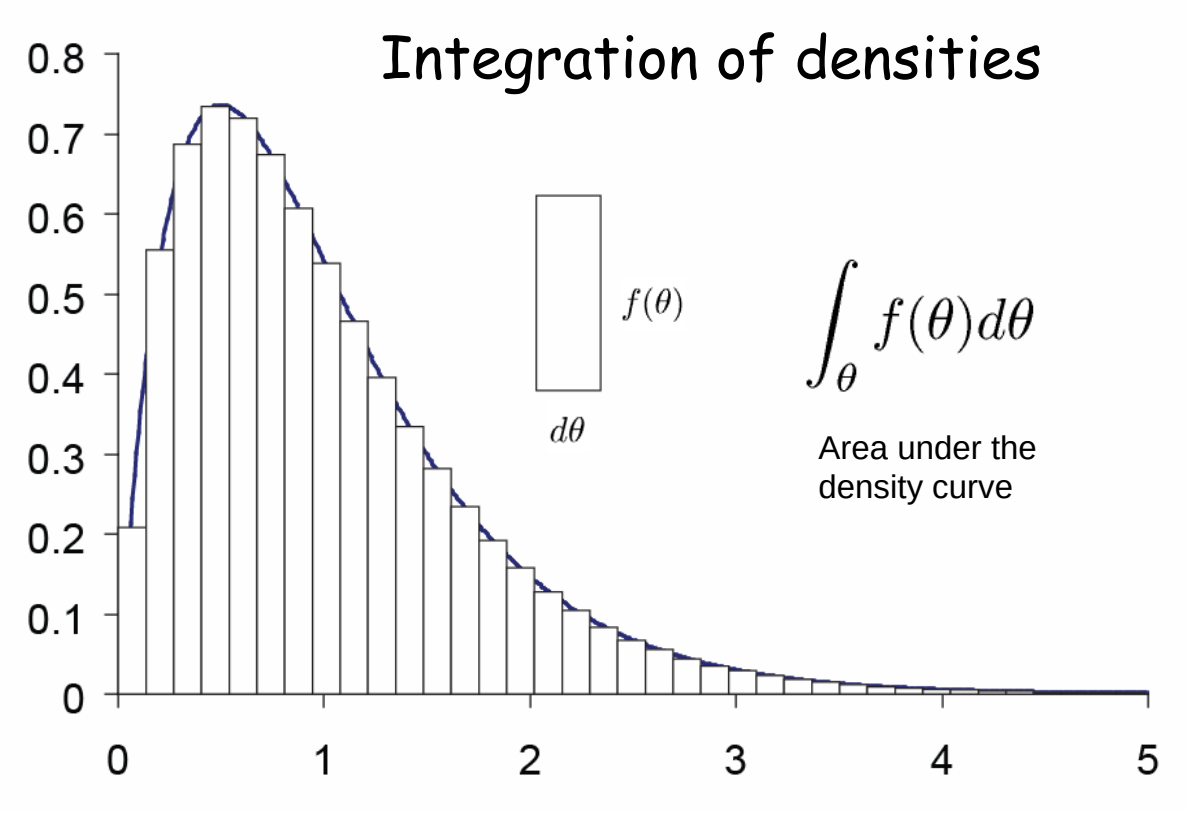
Bayes' rule: continuous case



Probabilities and probability *densities*



Integration of densities



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

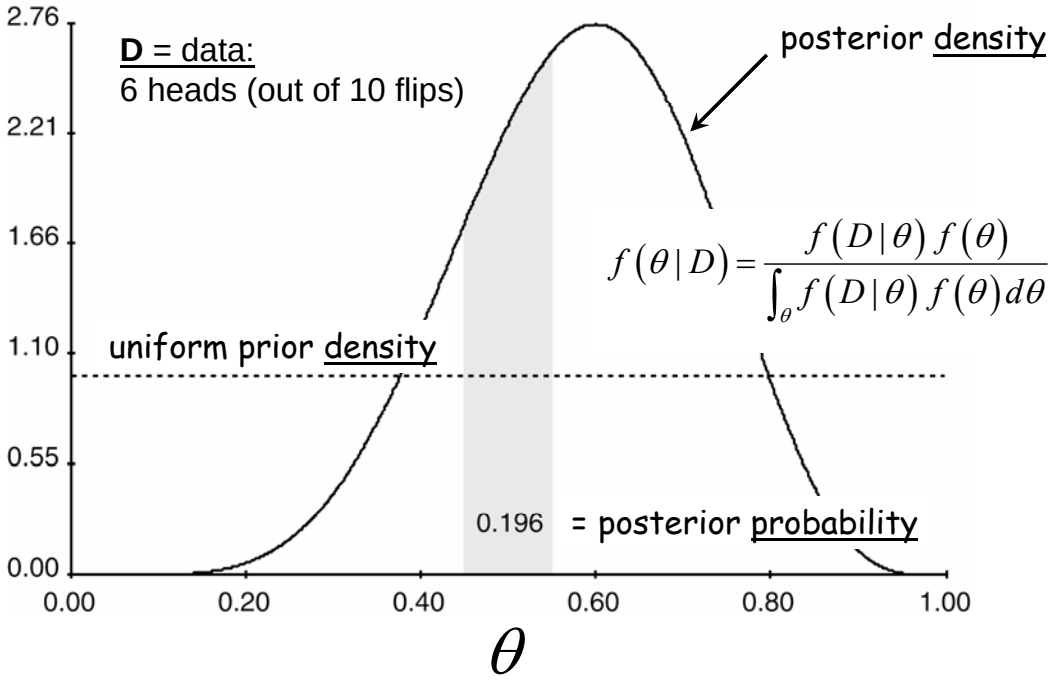
Area under the density curve

Coin-flipping example

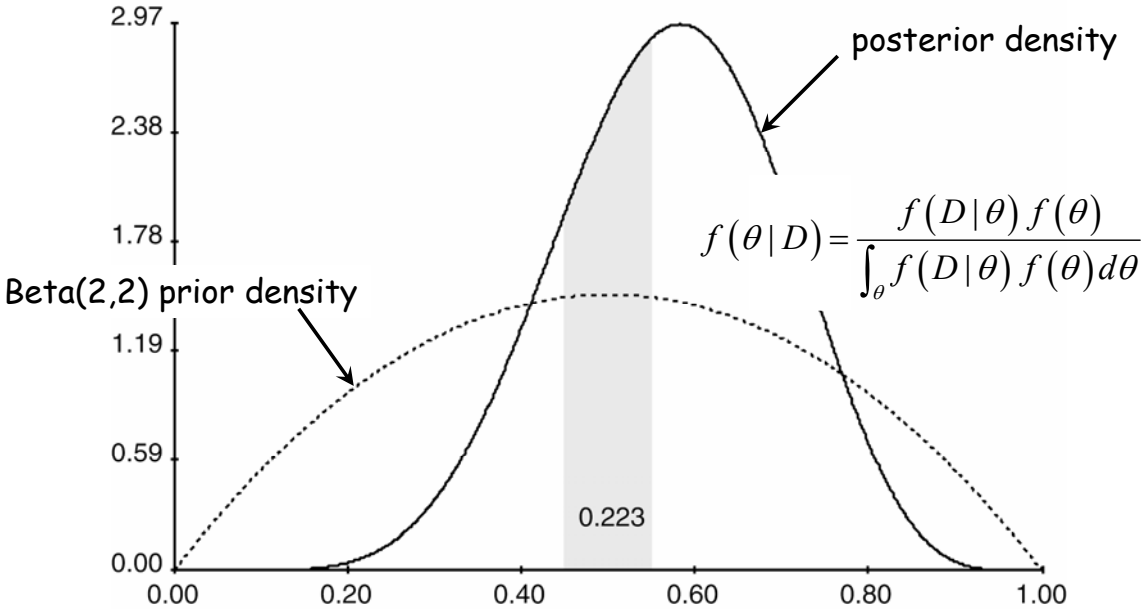
- **D**: 6 heads (out of 10 flips)
- θ = true underlying probability of heads on any given flip
  - if θ = 0.5, coin is perfectly fair
  - if θ = 1.0, coin always comes up heads (e.g. it is a trick coin with heads on both sides)
- Likelihood:

$$\Pr(D|\theta) = 210\,\theta^6\,(1-\theta)^4$$

Density vs. probability



Beta prior gives more flexibility



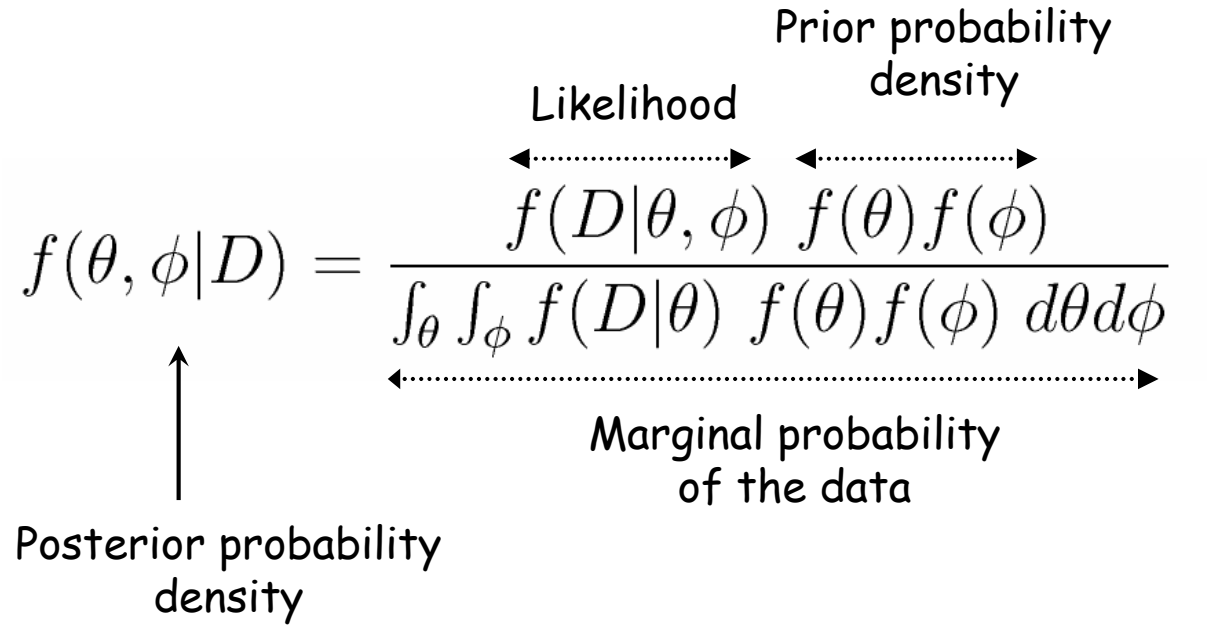
Posterior probability that θ lies between 0.45 and 0.55 is **0.223**

Bayesian Coin Flipper

Windows program download from:
http://hydrodictyon.eeb.uconn.edu/people/plewis/

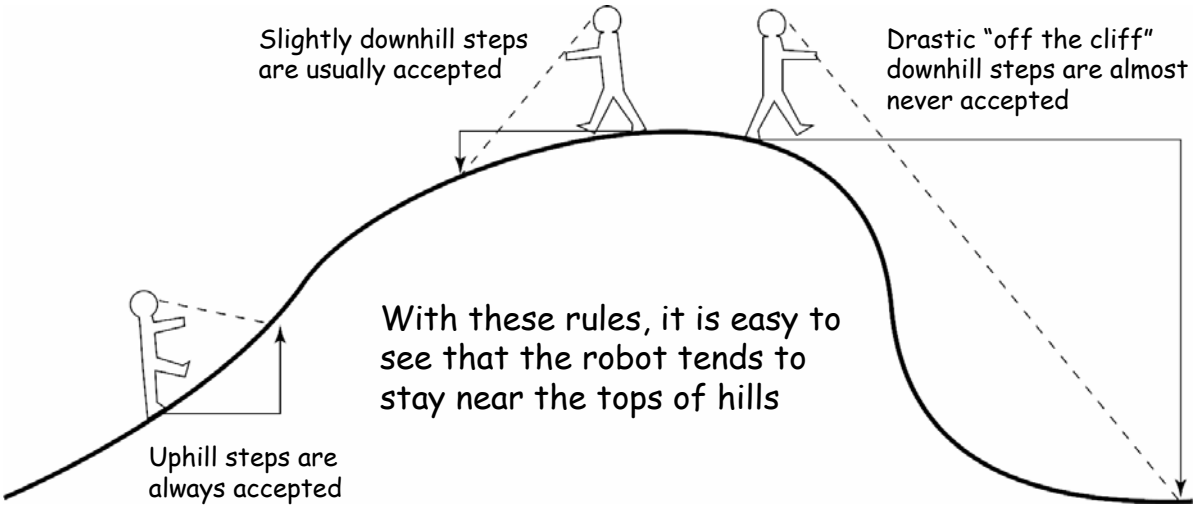
Excel spreadsheet version also available

Usually there are many parameters...

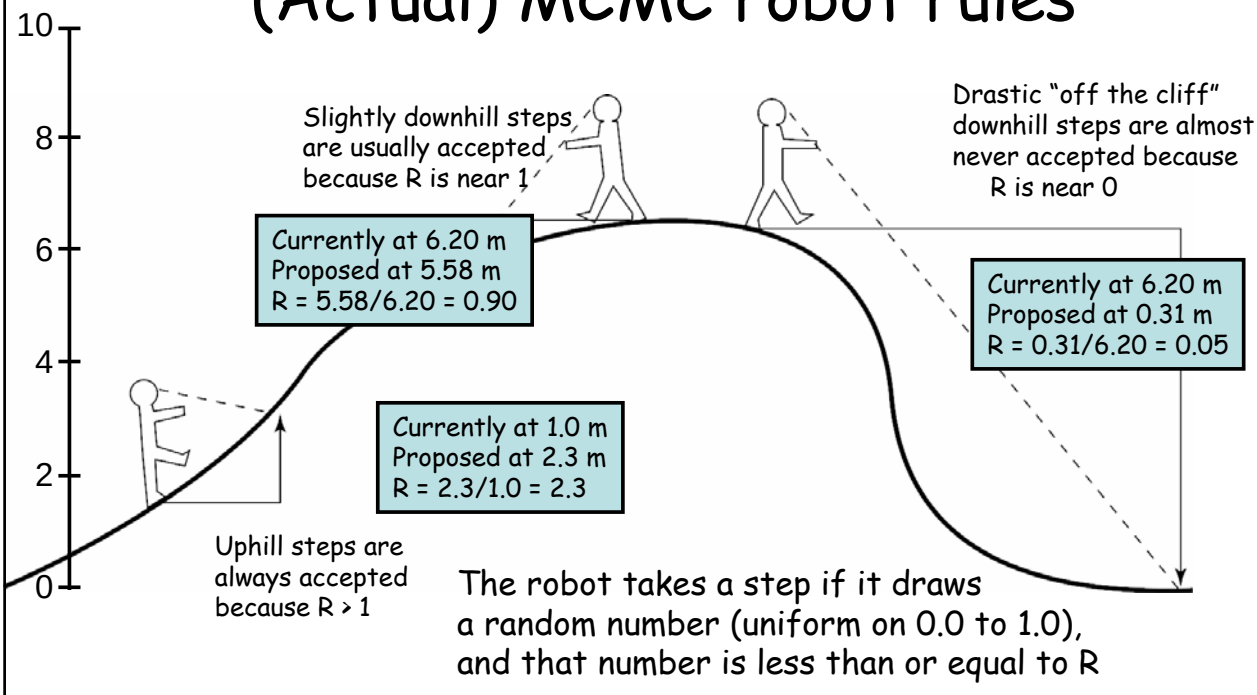


II. Markov chain Monte Carlo

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

A Thing Of Beauty

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

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

Target vs. proposal distributions

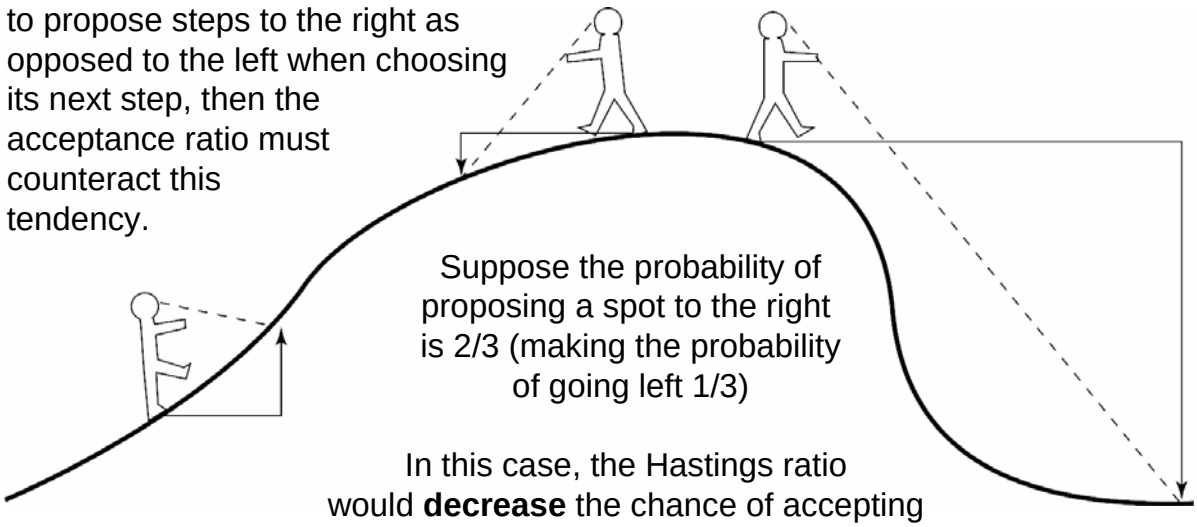
- The **target distribution** is the posterior distribution of interest
- The **proposal distribution** is used to decide where to go next; you have much flexibility here, and the choice affects only the *efficiency* of the MCMC algorithm

MCRobot

Windows program download from:
http://hydrodictyon.eeb.uconn.edu/people/plewis/

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 2/3 (making the probability of going left 1/3)

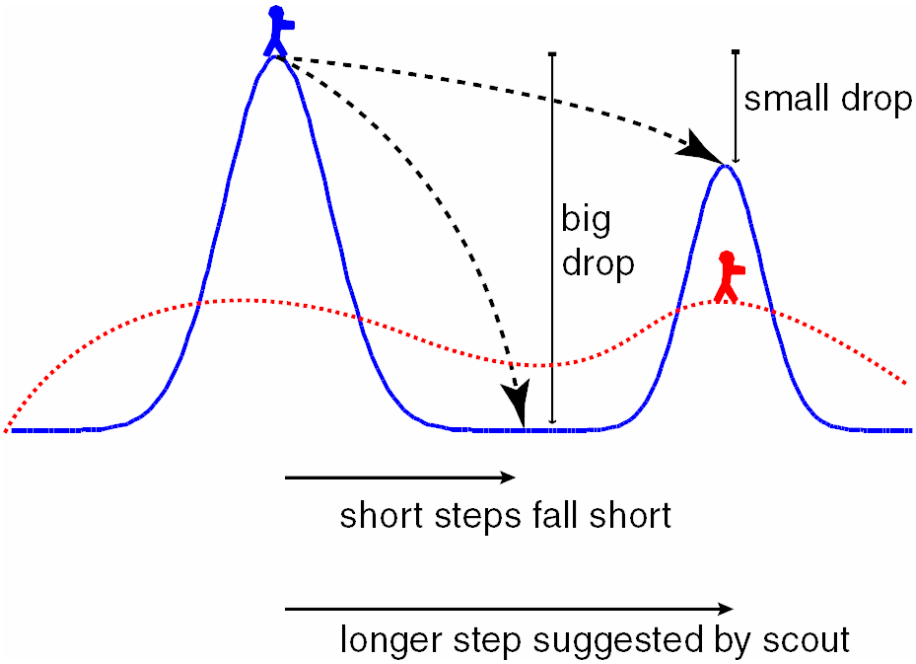
In this case, the Hastings ratio would **decrease** the chance of accepting moves to the **right** by **half**, and **double** the chance of accepting moves to the **left**

Metropolis-coupled Markov chain Monte Carlo (MCMCMC, or MC^3)

- MC^3 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)

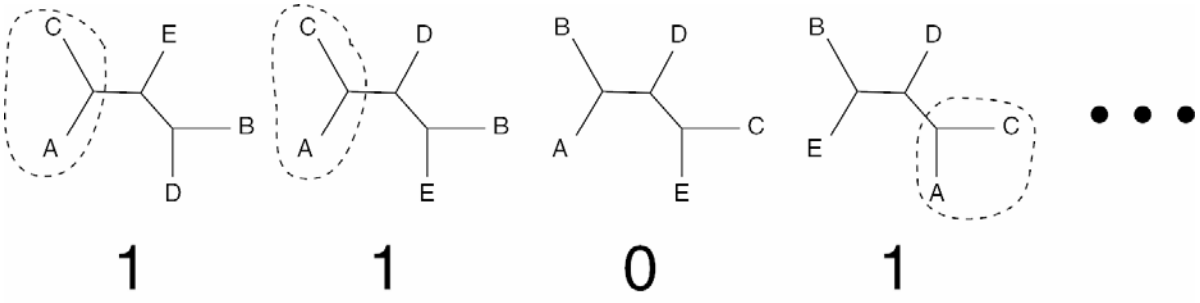
Geyer, C. J. 1991. Markov chain Monte Carlo maximum likelihood for dependent data. Pages 156-163 *in* Computing Science and Statistics (E. Keramidas, ed.).

Heated chains act as scouts for the cold chain



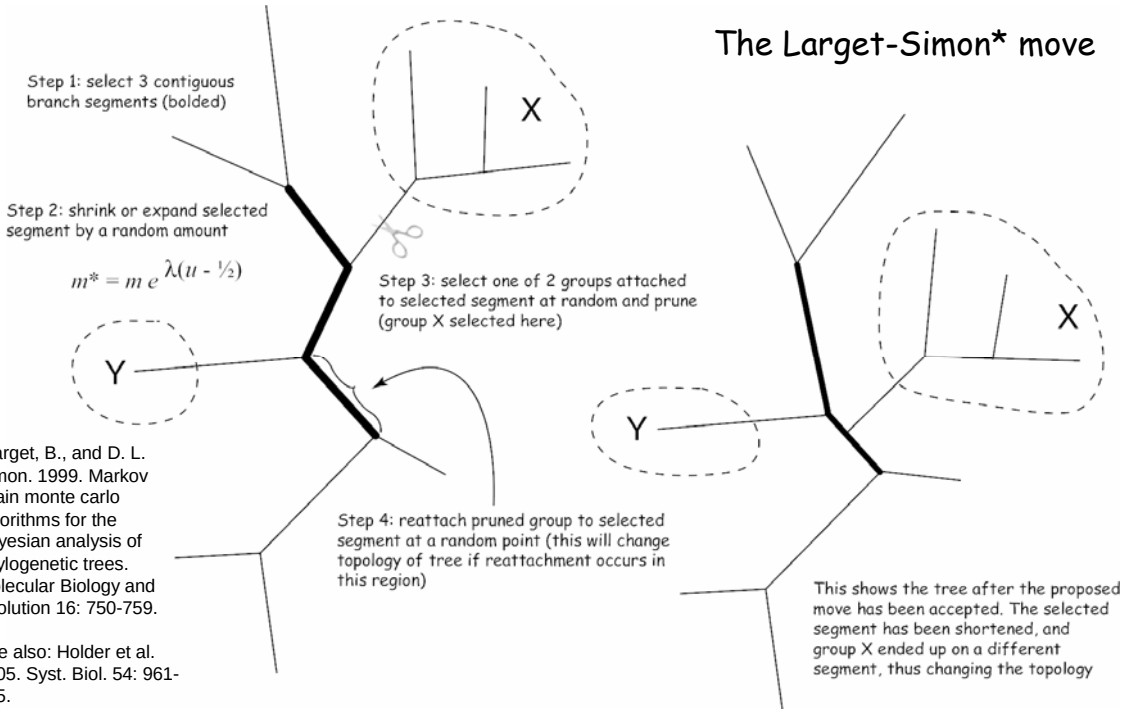
III. Bayesian phylogenetics

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



Imagine drawing tree topologies randomly from a bin in which the number of copies of any given topology is proportional to the (marginal) posterior probability of that topology. Approximating the posterior of any particular attribute of tree topologies (e.g. existence of group AC in this case) is simply a matter of counting.

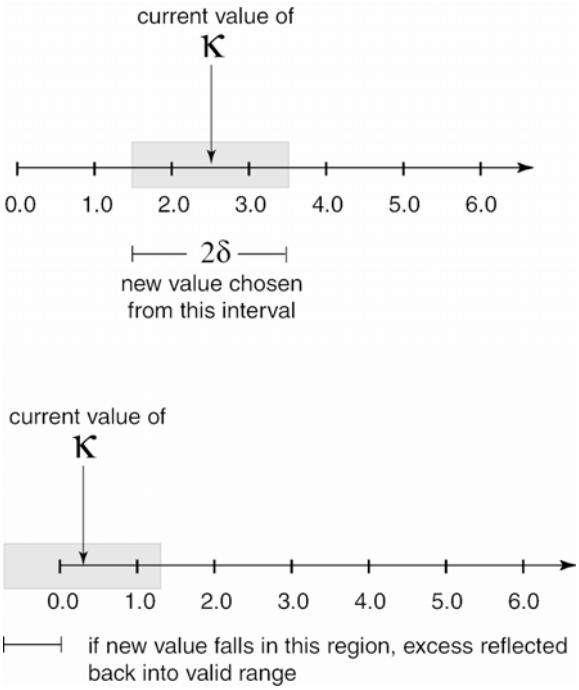
Moving through treespace



*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.

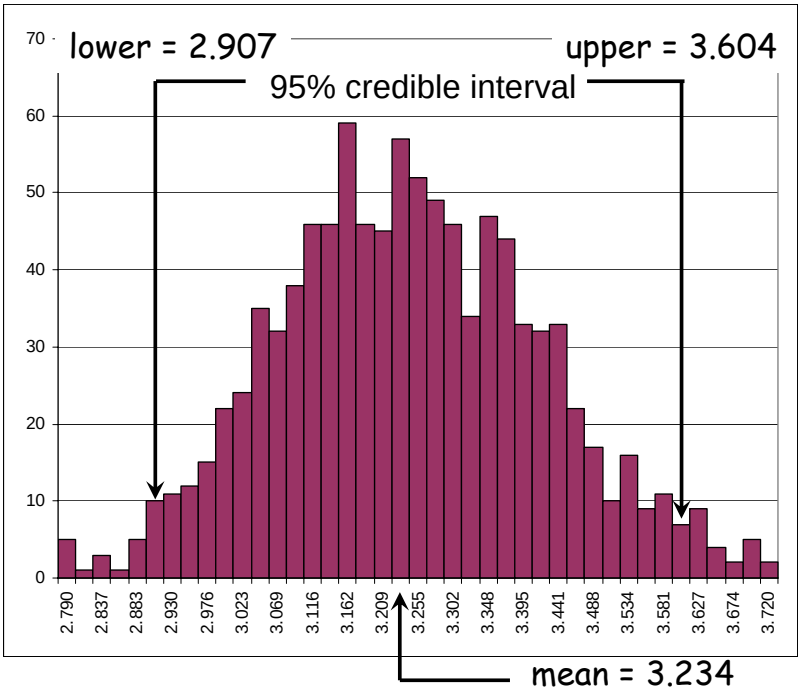
Proposal distribution is uniform from κ-δ to κ+δ

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.

Posteriors of model parameters



Histogram created from a sample of 1000 κ values.

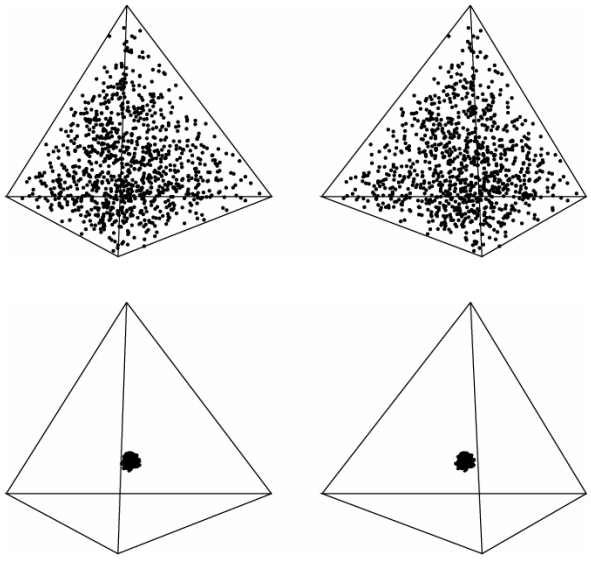
From: Lewis, L., and Flechtner, V. 2002. Taxon 51: 443-451.

IV. Prior distributions

Common Prior Distributions

- For **topologies**: discrete Uniform distribution
- For **proportions**: Beta(a,b) distribution
  - flat when a=b
  - peaked above 0.5 if a=b and both are greater than 1
- For **base frequencies**: Dirichlet(a,b,c,d) distribution
  - flat when a=b=c=d
  - all base frequencies close to 0.25 if a=b=c=d and large (e.g. 300)
- For **GTR model relative rates**: Dirichlet(a,b,c,d,e,f) distribution

4-parameter Dirichlet(a,b,c,d)



(stereo pairs)

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

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

(Thanks to Mark Holder for pointing out to me that a tetrahedron could be used for plotting a 4-dimensional Dirichlet)

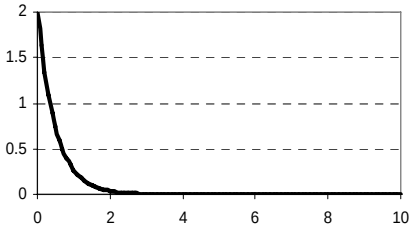
Common Priors (cont.)

- For other **model parameters** and **branch lengths**: Gamma(a,b) distribution
  - Exponential(λ) equals Gamma(1, λ^{-1}) distribution
  - Mean of Gamma(a,b) is $a \times b$
    - mean of an Exponential(10) distribution is 0.1
  - Variance of a Gamma(a,b) distribution is $a \times b^2$
    - variance of an Exponential(10) distribution is 0.01

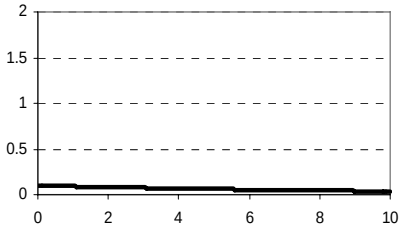
Note: be aware that in many papers the Gamma distribution is defined such that the second (scale) parameter is the *inverse* of the value b used in this slide! In this case, the mean and variance would be a/b and a/b^2, respectively.

Priors for model parameters with no upper bound

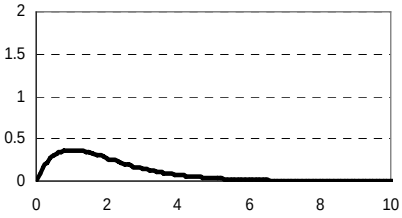
Exponential(2) = Gamma(1,½)



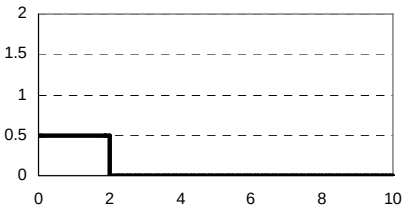
Exponential(0.1) = Gamma(1,10)



Gamma(2,1)



Uniform(0,2)



See chapter 18 in Felsenstein, J. (2004. Inferring Phylogenies.Sinauer) before using.

Learning about priors

Suppose you want to assume an Exponential distribution with mean 0.1 for the shape parameter of the discrete gamma distribution of among site rate heterogeneity. You use the command `help prset` in MrBayes (version 3.1.1) to find out how to do this, and this is what MrBayes says:

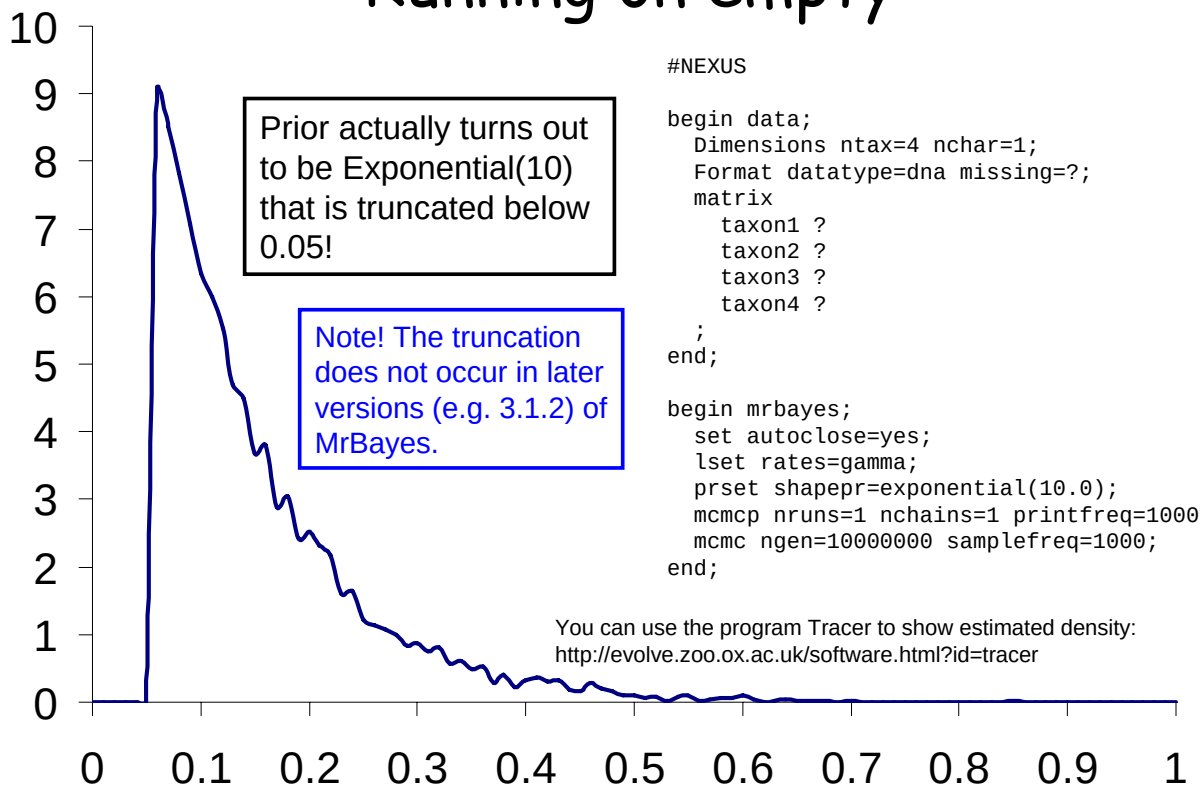
```
Shapepr -- This parameter specifies the prior for the gamma shape
           parameter for among-site rate variation. The options are:

           prset shapepr = uniform(<number>,<number>)
           prset shapepr = exponential(<number>)
           prset shapepr = fixed(<number>)
```

You type

prset shapepr=exponential(10.0);

but is mean of the prior going to be 10 or 0.1?
There is a way to find out...

Running on empty



Prior actually turns out to be Exponential(10) that is truncated below 0.05!

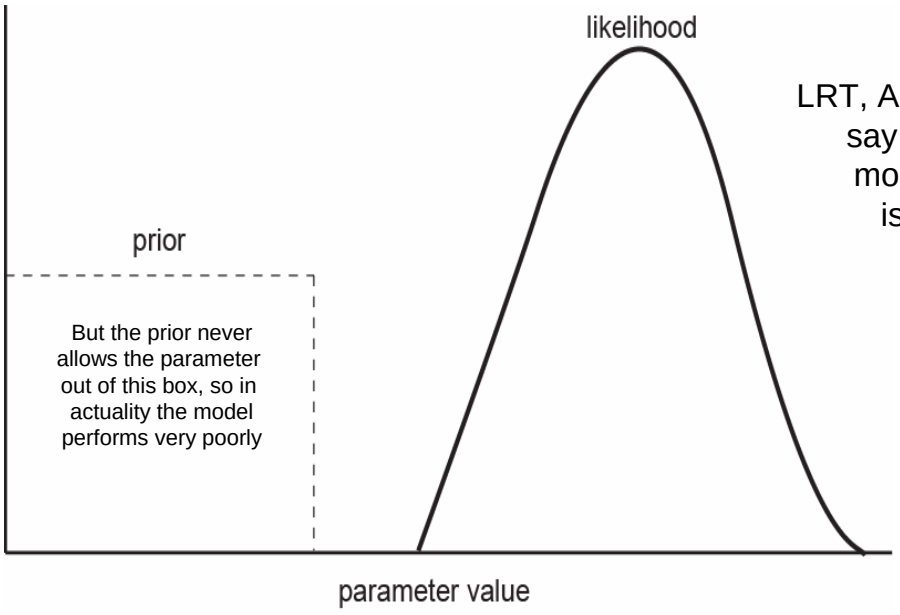
Note! The truncation does not occur in later versions (e.g. 3.1.2) of MrBayes.

```
#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;
```

You can use the program Tracer to show estimated density:
http://evolve.zoo.ox.ac.uk/software.html?id=tracer

V. Bayesian model selection

The choice of prior distributions can potentially turn a good model bad!



LRT, AIC and BIC all say this is a great model because it is able to attain such a high maximum likelihood score

Marginal probabilities of models

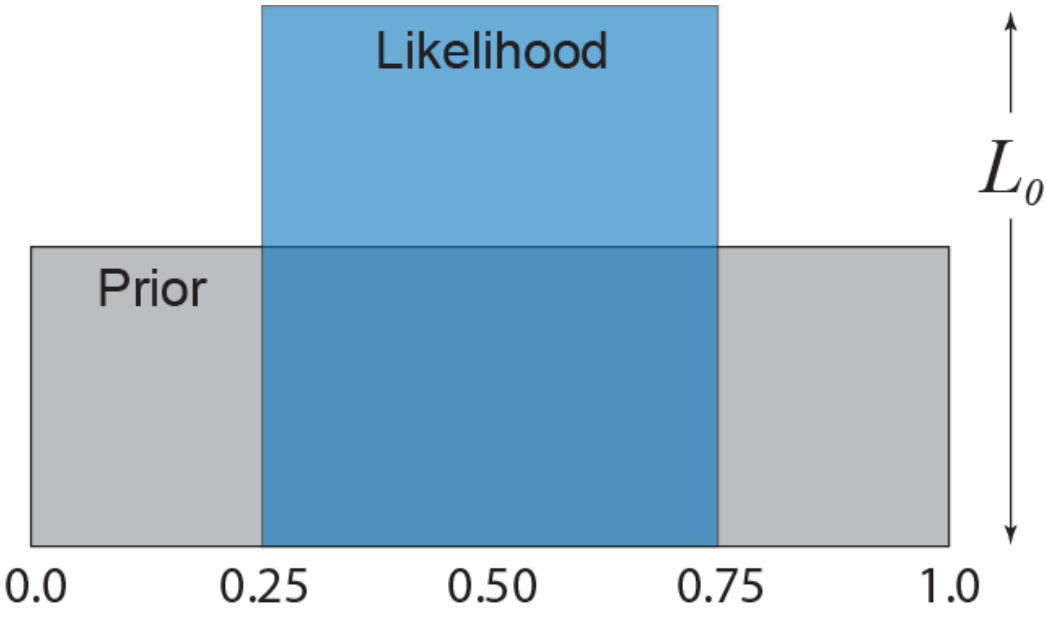
$$\Pr(D) = \int_{\theta} f(D|\theta)\ f(\theta)\ d\theta$$

Marginal probability of the data (denominator in Bayes' rule).
This is a weighted average of the likelihood, where the weights
are provided by the prior distribution.

$$\Pr(D|M) = \int_{\theta} f(D|\theta, M)\ f(\theta|M)\ d\theta$$

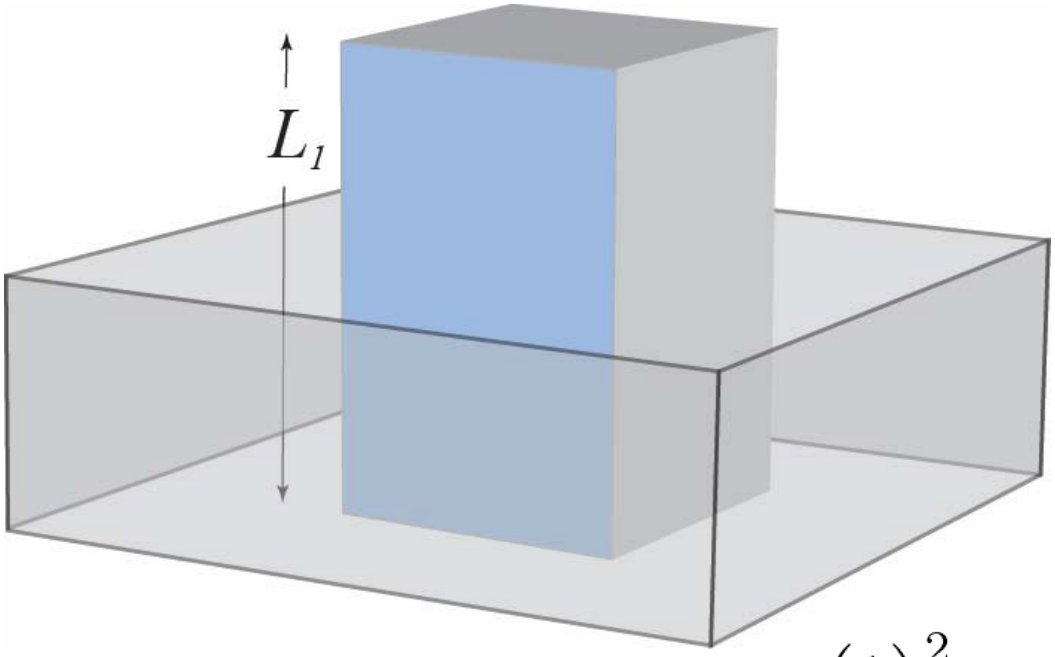
Often left out is the fact that we are also conditioning on M, the model used. $\Pr(D|M_1)$ is comparable to $\Pr(D|M_2)$ and thus the marginal probability of the data can be used to compare the average fit of different models as long as the data D is the same.

Bayes Factor: 1-param. model



Average likelihood = $\left(\frac{1}{2}\right) L_0$

Bayes Factor: 2-param. model



Average likelihood = $\left(\frac{1}{2}\right)^2 L_1$

Bayes Factor is ratio of marginal model likelihoods

1-parameter model M_0: (½) L_0

2-parameter model M_1: (¼) L_1

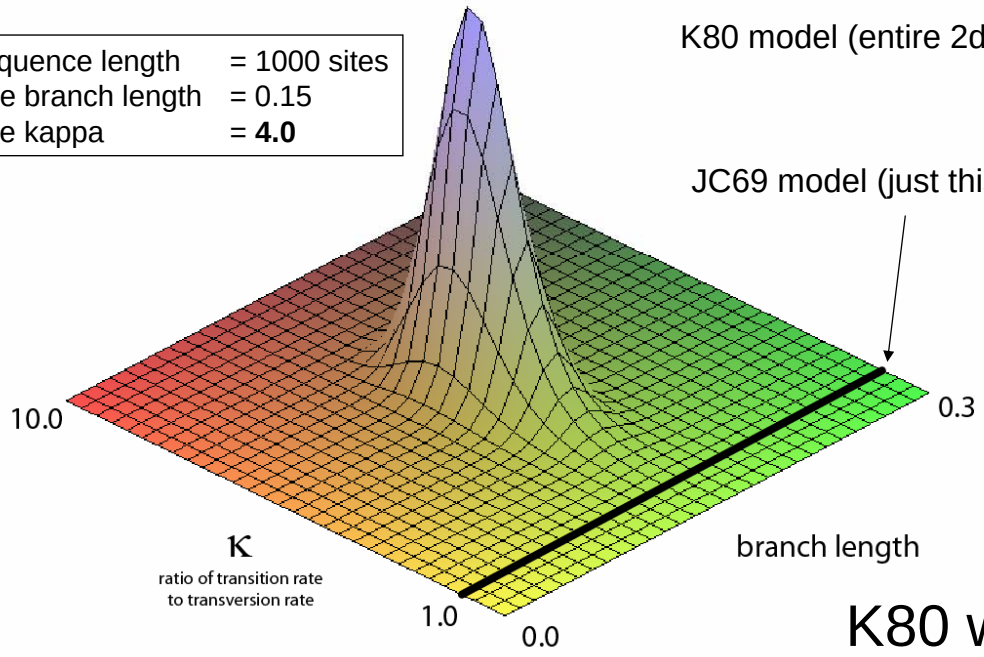
Bayes Factor favors M_0 unless L_1 is at least *twice* as large as L_0

All other things equal, more complex models are penalized by their extra dimensions

Recent work on Bayes factors with respect to phylogenetics:
Huelsenbeck, Larget & Alfaro. 2004. MBE 2004:1123-1133.
Lartillot & Phillippe. 2005. Syst. Biol. 55(2):195-207.

Marginal Likelihood of a Model

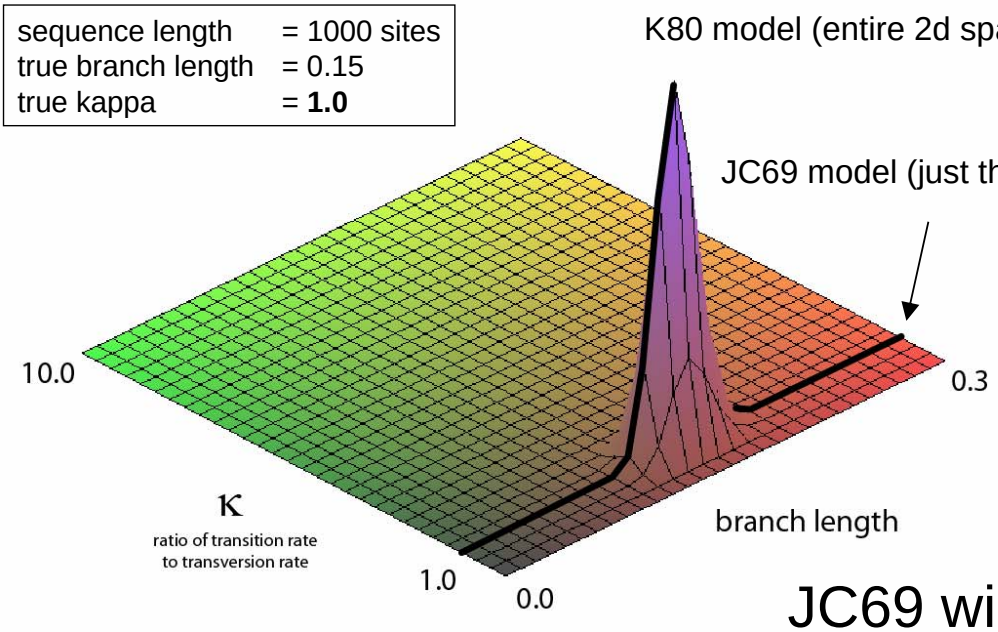
sequence length	= 1000 sites
true branch length	= 0.15
true kappa	= **4.0**



K80 wins

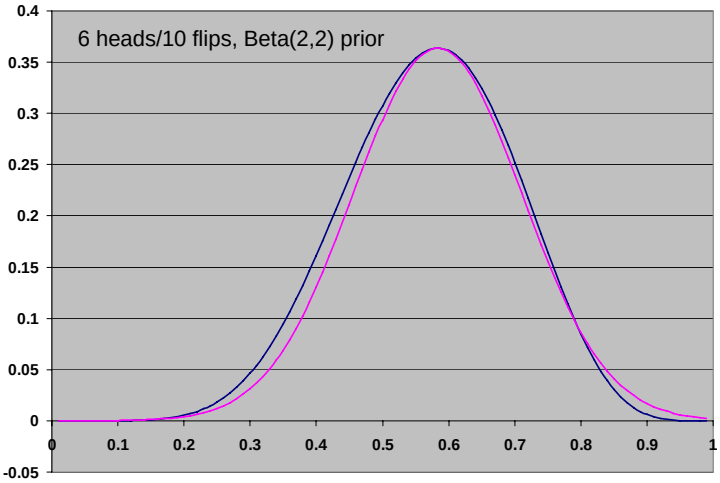
Marginal Likelihood of a Model

sequence length	= 1000 sites
true branch length	= 0.15
true kappa	= **1.0**



JC69 wins

Bayesian Information Criterion (BIC)



Area under pink curve is easy to calculate and is good approximation to the desired quantity (smaller by about 0.5%)

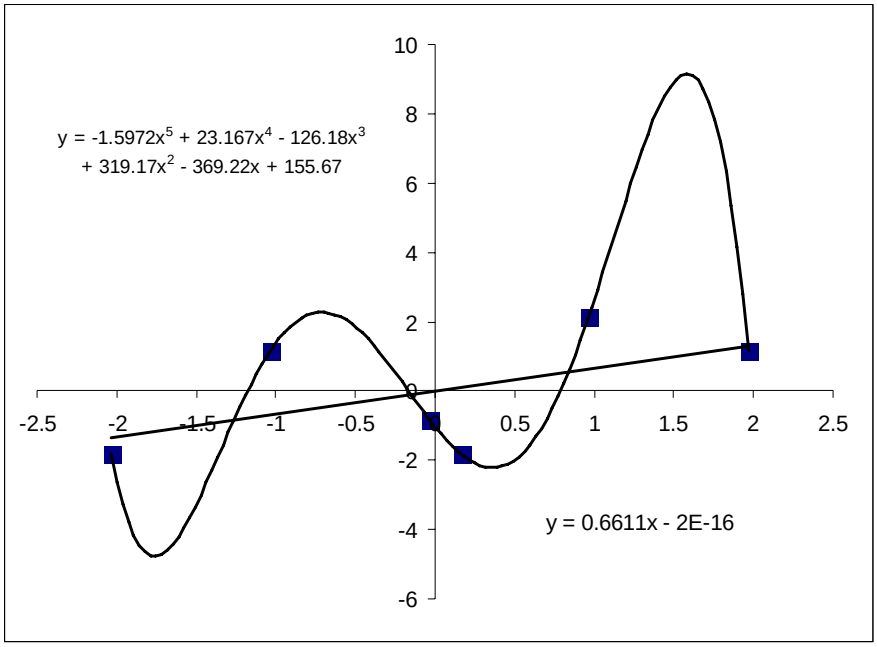
Blue curve is the **unnormalized posterior**

Area under this curve equals **marginal probability of the data** (the desired quantity)

Pink curve is a **normal distribution** scaled to match blue curve as closely as possible

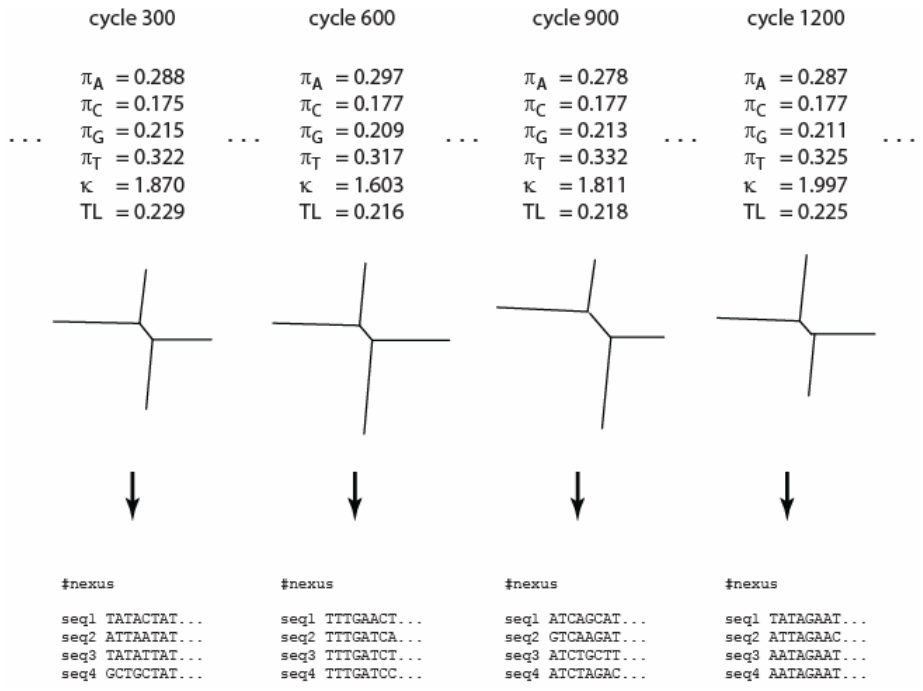
Assumes prior is a normal distribution *with variance equivalent to the amount of information in a single observation*

Goodness-of-fit ain't everything



(Thanks to Dave Swofford for introducing me to this excellent example)

Posterior Predictive Simulation



Parameter values sampled during MCMC

Tree topologies and branch lengths sampled during MCMC

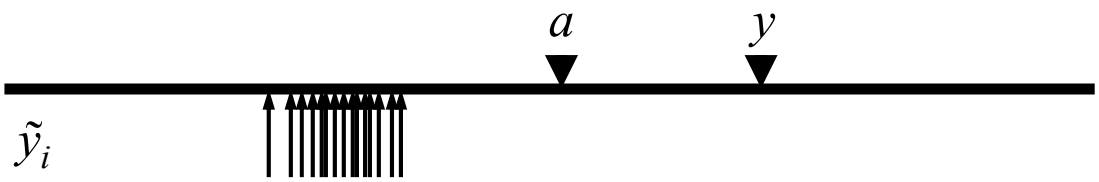
Posterior predictive data sets simulated

Minimum Posterior Predictive Loss Approach

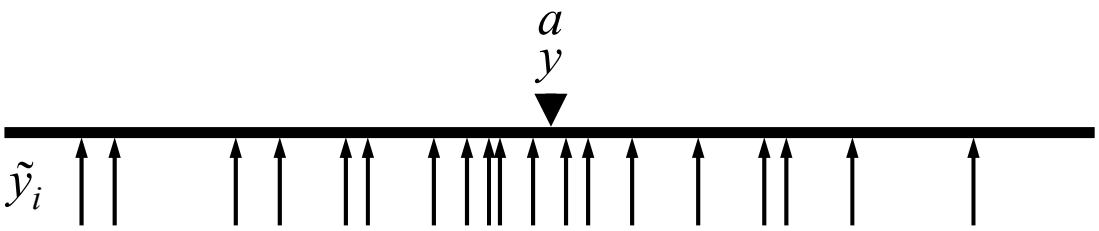
- Perform MCMC on original dataset ***y***
- During MCMC generate posterior predictive datasets ***ỹ***
- Find "average" dataset ***a*** that is as close as possible to both ***y*** and the ***ỹ***
- ***Gm*** measures distance between ***a*** and ***y***
- ***Pm*** measures expected distance between ***a*** and ***ỹ***
- Goal is to minimize the overall measure ***Dm*** = ***Gm*** + ***Pm***

Gelfand, A. E., and S. Ghosh. 1998. Model choice: a minimum posterior predictive loss approach. Biometrika 85:1-11.

More than one way to be bad

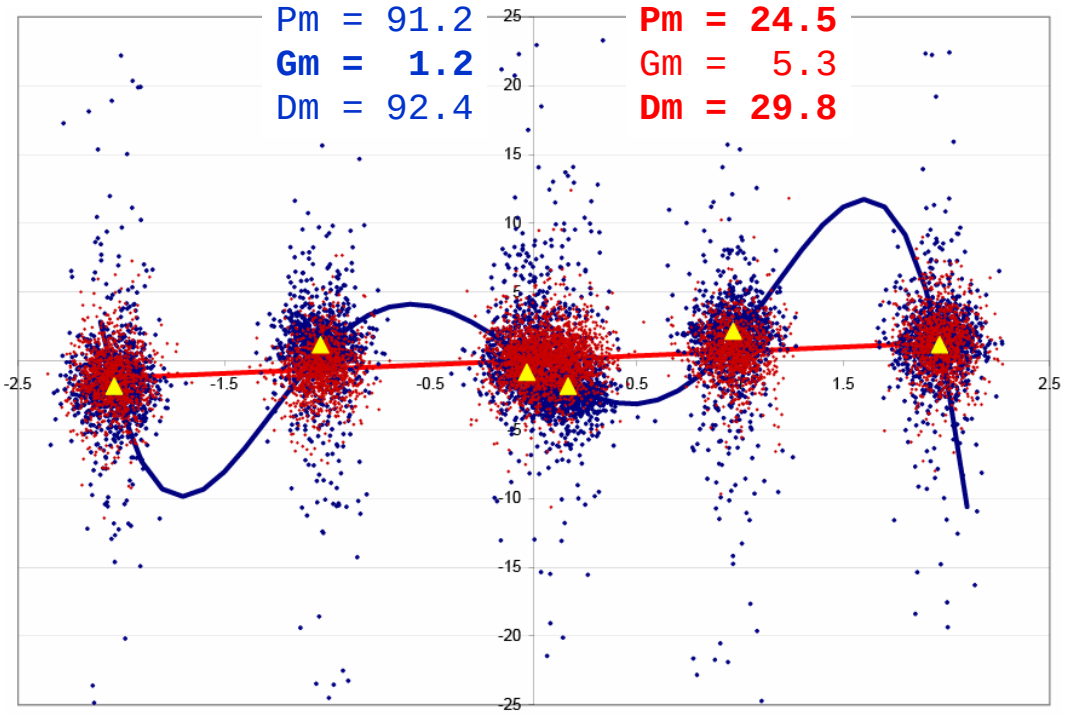


Poor *goodness-of-fit* (*Gm* large, *Pm* small)

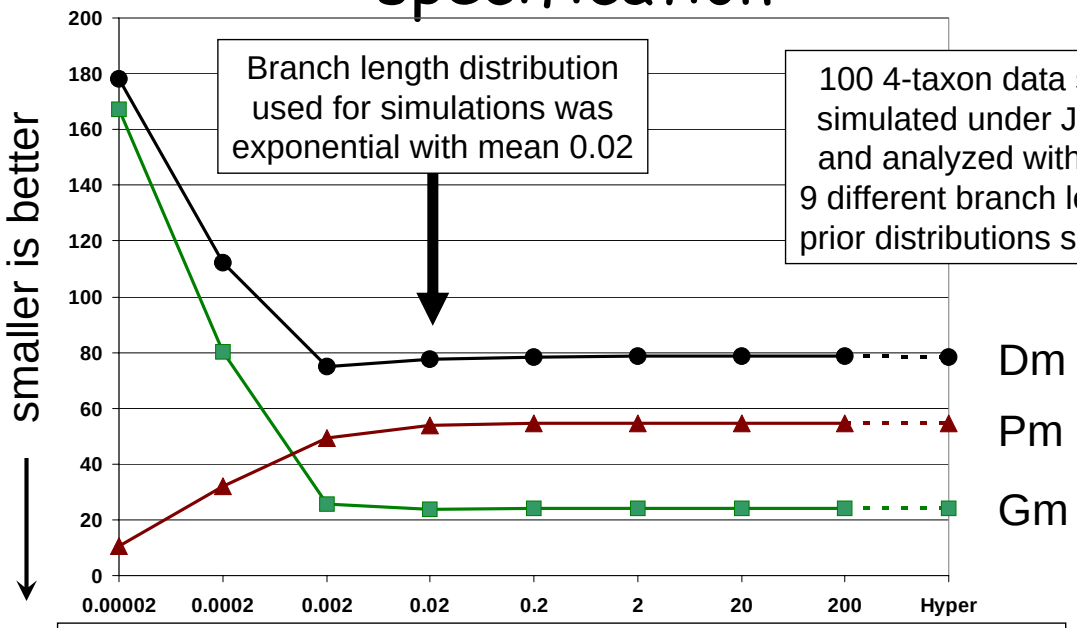


High *predictive variance* (*Gm* small, *Pm* large)

Regression Example Revisited



Models differ only in prior specification



Note: none of the standard model testing approaches (AIC, LRT, BIC) work here because these models differ only in their prior specification

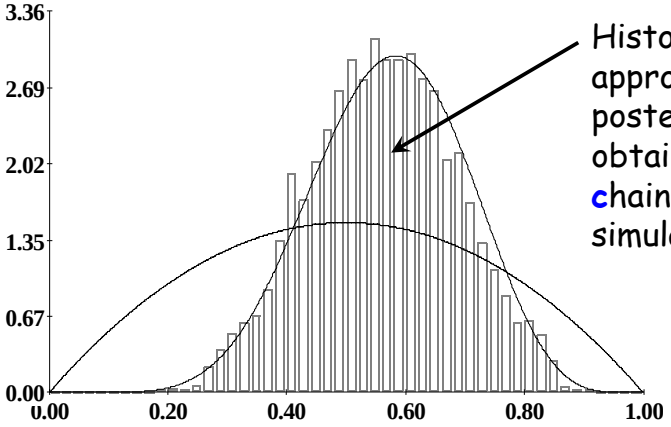
The End

Many thanks to NSF (CIPRES project) for my current funding, and
for UConn and the National Evolutionary Synthesis Center (NESCENT)
for funding my sabbatical this year.

NESCENT
Durham, NC

The following slides were not shown in the lecture, but they are relevant to the content and are included to provide a more complete record of the main points.

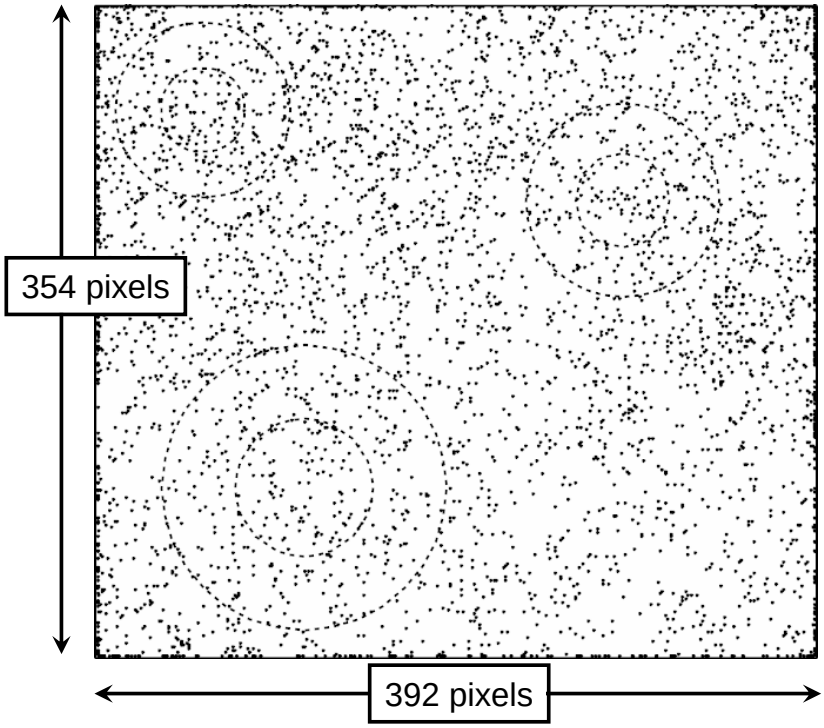
Markov chain Monte Carlo (MCMC)



Histogram is an approximation to the posterior distribution obtained using a **M**arkov **c**hain **M**onte **C**arlo simulation.

For coin flipping, we can compute posterior probabilities exactly (can even do the necessary integration in Excel!). For complex problems, we must settle for a **good approximation**.

Pure random walk



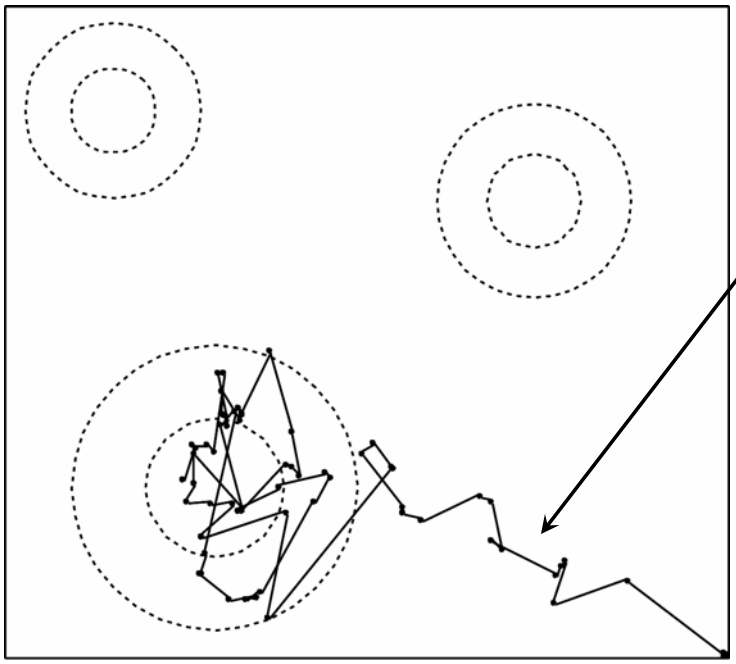
Proposal scheme:
- random direction
- gamma-distributed step length (mean 45 pixels, s.d. 40 pixels)
- reflection at edges

Target distribution:
- equal mixture of 3 bivariate normal "hills"
- inner contours: 50%
- outer contours: 95%

In this case, the robot is accepting every step

5000 steps shown

Burn-in



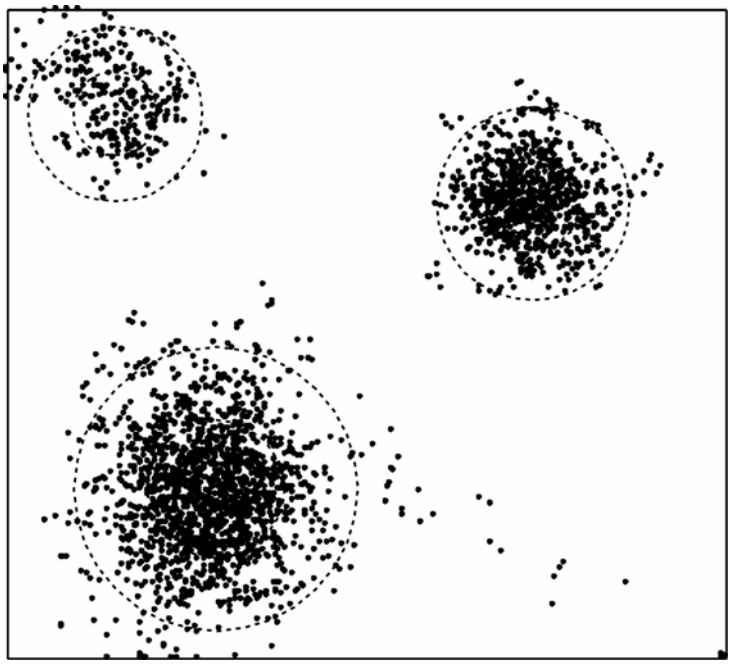
Robot is now following the rules and thus quickly finds one of the three hills.

Note that first few steps are not at all representative of the distribution.

100 steps taken

Starting point

Target distribution approximation



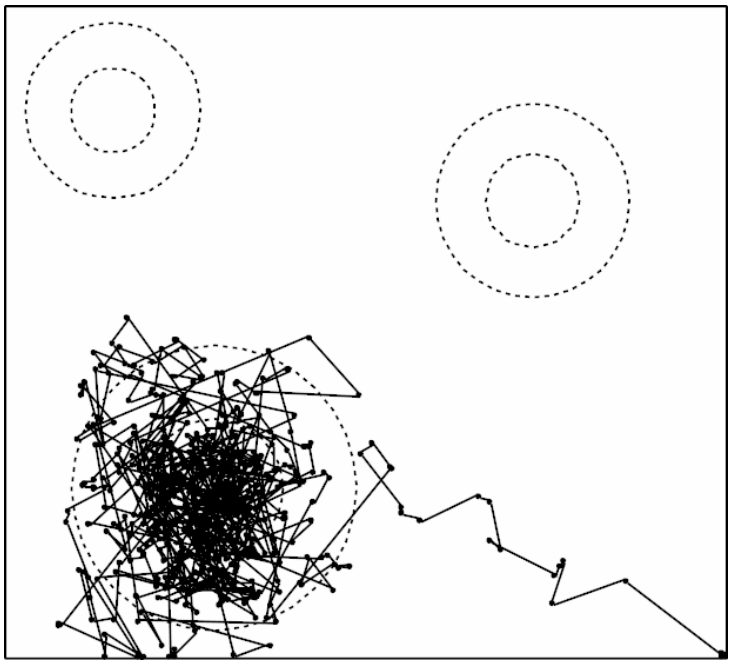
How good is the MCMC approximation?

- 51.2% of points are inside inner contours (cf. 50% actual)
- 93.6% of points are inside outer contours (cf. 95% actual)

Approximation gets better the longer the chain is allowed to run.

5000 steps taken

Just how long is a long run?



What would you conclude about the target distribution had you stopped the robot at this point?

1000 steps taken

The way to avoid this mistake is to perform several runs, each one beginning from a different randomly-chosen starting point.

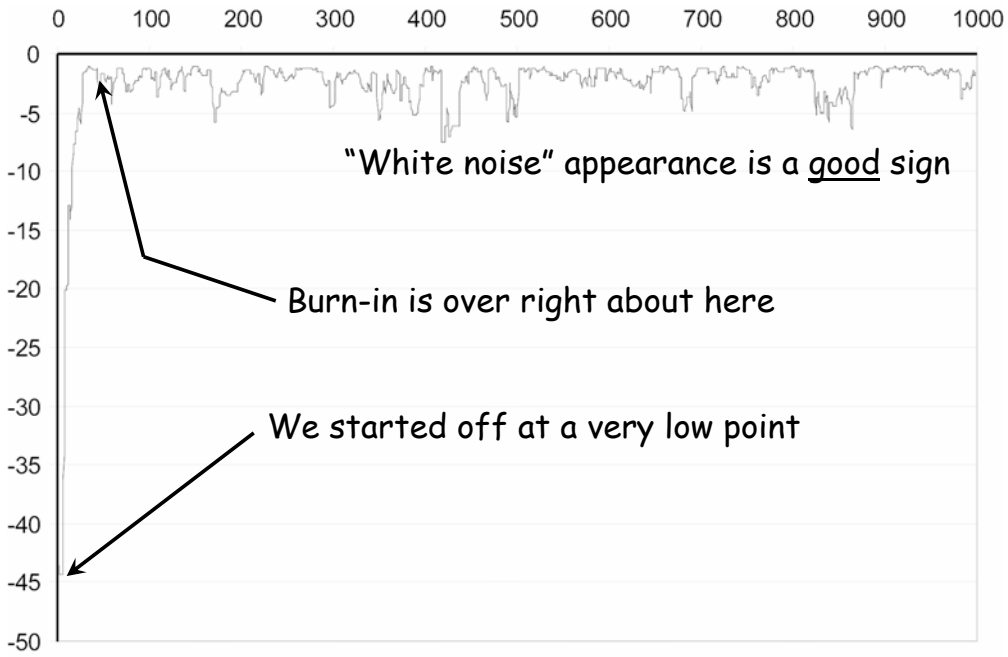
Results different among runs? Probably none of them were run long enough!

Cold vs. heated landscapes

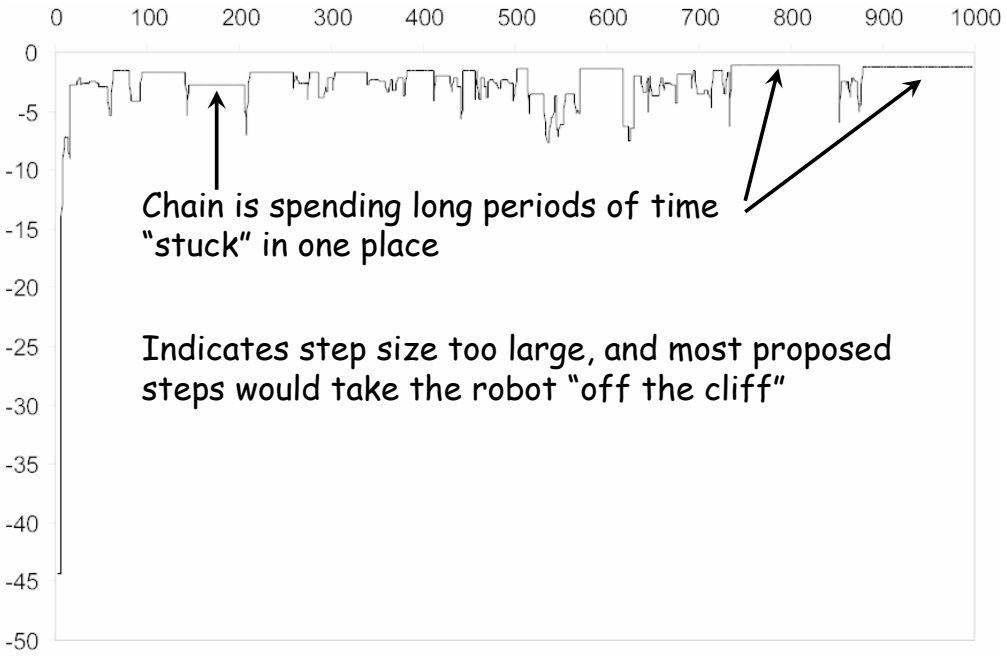
Cold landscape: note peaks separated by deep valleys

Heated landscape: note shallow (easy to cross) valleys

Trace plots


0
100
200
300
400
500
600
700
800
900
1000
0
-5
-10
-15
-20
-25
-30
-35
-40
-45
-50
"White noise" appearance is a good sign
Burn-in is over right about here
We started off at a very low point

Slow 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"