Bayesian Phylogenetics

Bret Larget
larget@stat.wisc.edu

Departments of Botany and of Statistics
University of Wisconsin—Madison

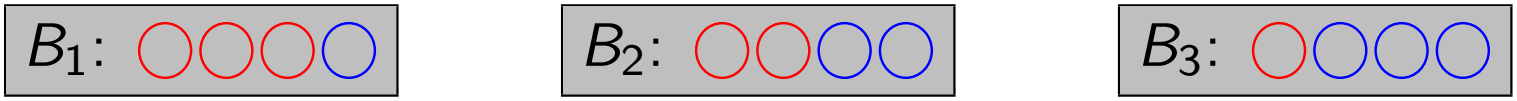
February 11, 2008

Who was Bayes?

- The Reverand Thomas Bayes was born in London in 1702.
- He was the son of one of the first Noncomformist ministers to be ordained in England.
- He became a Presbyterian minister in the late 1720s, but was well known for his studies of mathematics.
- He was elected a Fellow of the Royal Society of London in 1742.
- He died in 1761 before his works were published.

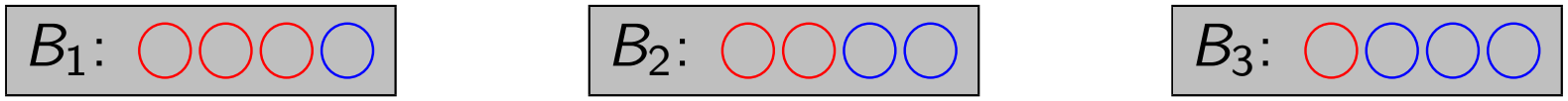
What is Bayes' Theorem?

- Bayes' Theorem explains how to calculate inverse probabilities.
- For example, suppose that boxes contains colored balls as shown below.



- Given a box, a ball is chosen *uniformly at random*.
- For example, if a ball is chosen from Box B_1, there is a $3/4$ chance that it is red.
- The inverse problem states if a red ball is drawn, how likely is it that it came from Box B_1?

What is Bayes' Theorem?



- If a red ball is drawn, how likely is it that it came from Box B_1?
- To answer this question, we need *a prior distribution* for the selection of the box.
- The answer will be different if we believe *a priori* that Box B_1 is 10% likely to be the chosen box than if we believe that all three boxes are equally likely.

Bayes' Theorem

- Bayes' Theorem states that if a complete list of mutually exclusive events $B_1, B_2, \ldots$ have prior probabilities $\Pr(B_1), \Pr(B_2), \ldots$, and if the *likelihood* of the event A given event B_i is $\Pr(A \mid B_i)$ for each i, then

$$\Pr(B_i \mid A) = \frac{\Pr(A \mid B_i)\Pr(B_i)}{\sum_j \Pr(A \mid B_j)\Pr(B_j)}$$

- The *posterior probability* of B_i given A, written $\Pr(B_i \mid A)$, is proportional to the product of the *likelihood* $\Pr(A \mid B_i)$ and the *prior probability* $\Pr(B_i)$ where the normalizing constant $\Pr(A) = \sum_j \Pr(A \mid B_j)\Pr(B_j)$ is the prior probability of A.

Connection to Phylogeny

- In a Bayesian approach to phylogenetics, the *boxes are like different tree topologies*.
- The *colored balls are like site patterns*, except:
  - there are many more than two colors; and
  - we observe multiple draws from each box.
- Additional parameters such as branch lengths and substitution model parameters affect the likelihood, are unknown, and add to the complexity.

Prior and Posterior Distributions

- A *prior distribution* is a probability distribution on parameters *before* any data is observed.
- A *posterior distribution* is a probability distribution on parameters *after* data is observed.

Bayesian Methods vs. Maximum Likelihood

	Maximum Likelihood	Bayesian
Probability	Only defined in the context of long-run relative frequencies	Describes everything that is uncertain
Parameters	Fixed and Unknown	Random
Nuisance Parameters	Optimize them	Average over them
Testing	p-values	Bayes' factors
Nature of Method	Objective	Subjective

Bayesian Phylogenetic Methods

- Let's say we want to find the posterior probability of a clade.
- We would need to sum the posterior probabilities of all trees with the clade.

$$\begin{aligned}\Pr(\text{clade} \mid \text{data}) &= \sum_{\text{tree with clade}} \Pr(\text{tree} \mid \text{data}) \\ &= \sum_{\text{tree with clade}} \frac{\Pr(\text{data} \mid \text{tree}) \Pr(\text{tree})}{\Pr(\text{data})}\end{aligned}$$

- But we need to know the parameters including branch lengths (params) to compute the likelihood.

$$\begin{aligned}&\sum_{\text{tree with clade}} \Pr(\text{data} \mid \text{tree}) \Pr(\text{tree}) \\ &= \sum_{\text{tree with clade}} \int \Pr(\text{data}, \text{params} \mid \text{tree}) \Pr(\text{tree}) \mathrm{d}\text{params} \\ &= \sum_{\text{tree with clade}} \Pr(\text{tree}) \int \Pr(\text{data} \mid \text{params}, \text{tree}) \Pr(\text{params} \mid \text{tree}) \mathrm{d}\text{params}\end{aligned}$$

Bayesian Phylogenetic Methods

- So, we need to compute:

$$\frac{\sum_{\text{tree with clade}} \Pr(\text{tree}) \int \Pr(\text{data} \mid \text{params}, \text{tree}) \Pr(\text{params} \mid \text{tree}) \mathrm{d}\text{params}}{\Pr(\text{data})}$$

- However, the denominator Pr(data) and the integral in the numerator are generally not computable.
- Solution? Markov chain Monte Carlo.

Metropolis-Hastings Example

- Assume a Jukes-Cantor likelihood model for two species where we observe 50 sites, 9 of which differ.
- The likelihood for the distance d is

$$L(d) = \left(\frac{1}{4}\right)^{50} \times \left(\frac{1}{4} - \frac{1}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{9} \times \left(\frac{1}{4} + \frac{3}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{41}$$

- Assume a prior for d with the form

$$p(d) = \frac{\lambda}{(1+\lambda d)^2}, \quad d > 0$$

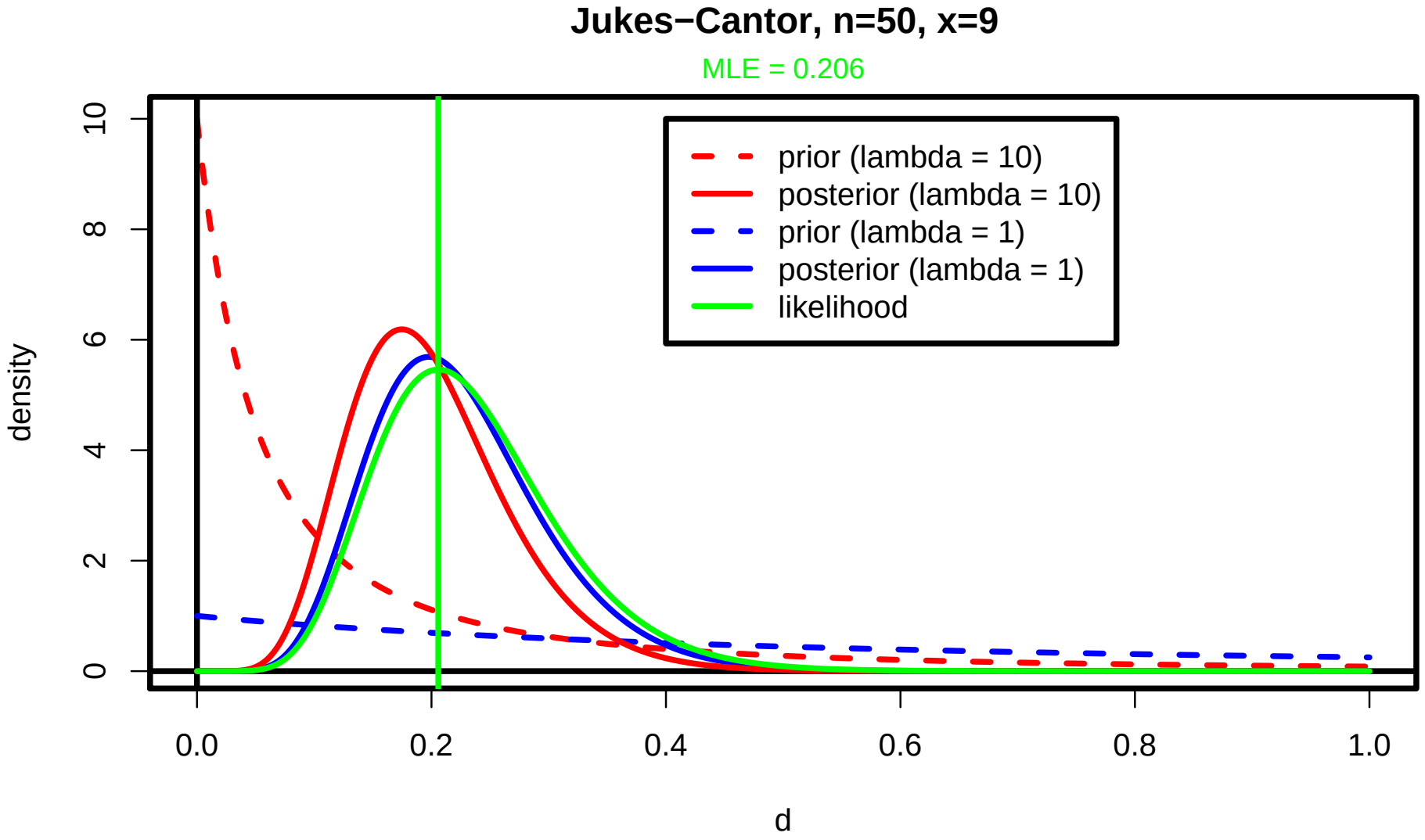
  where $\lambda > 0$ is a parameter.
- This density is what you get if you take the ratio of two independent exponential random variables, one with parameter λ and one with parameter 1.
- The median is $1/\lambda$, but the mean is $+\infty$.

Example

- An exact expression for the posterior density of d is

$$p(d \mid x) = \frac{\left(\frac{\lambda}{(1+\lambda d)^2}\right)\left(\left(\frac{1}{4}\right)^{50}\left(\frac{1}{4}-\frac{1}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{9}\left(\frac{1}{4}+\frac{3}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{41}\right)}{\int_0^\infty \left(\frac{\lambda}{(1+\lambda d)^2}\right)\left(\left(\frac{1}{4}\right)^{50}\left(\frac{1}{4}-\frac{1}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{9}\left(\frac{1}{4}+\frac{3}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{41}\right)\mathrm{d}d}$$

Graph


Jukes−Cantor, n=50, x=9
MLE = 0.206
prior (lambda = 10)
posterior (lambda = 10)
prior (lambda = 1)
posterior (lambda = 1)
likelihood
density
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0
d

What is Markov Chain Monte Carlo?

- Markov chain Monte Carlo (MCMC) is a method to take (dependent) samples from a distribution.
- The distribution need only be known up to a constant of proportionality.
- MCMC is especially useful for computation of Bayesian posterior probabilities.
- Simple summary statistics from the sample converge to posterior probabilities.
- Metropolis-Hastings is a form of MCMC that works using any Markov chain to propose the next item to sample, but rejecting proposals with specified probability.

Typical Problem

- We want to make inferences on the basis of a posterior distribution $p(\theta \mid x)$.
- We cannot calculate desired quantities analytically, so instead we wish to sample from $p(\theta \mid x)$ and use sample statistics as estimates for the true posterior values— for example, a sample mean is an estimate of an expected value.
- But, we also may not be able to take a simple random sample of θ values from the posterior distribution.
- A computational method called *Markov chain Monte Carlo* has proven to be remarkably successful for obtaining *dependent* samples from probability distributions.
- The idea is that each sampled point depends on the most recently sampled point.
- If this is done carefully, sample statistics will converge to the desired posterior values.

Metroplis-Hastings MCMC

- Markov chain Monte Carlo (MCMC) takes (dependent) samples from a distribution.
- The distribution *need only be known up to a constant of proportionality* as the algorithm depends only on *ratios*
- A *proposal method* is needed that describes a probability distribution for proposing new parameter values given current ones.
- In theory, just about any proposal distribution is correct (given an infinite sample size)— the *art* is in designing (and correctly implementing) a method so that feasible sample sizes are adequate.
- If $q(\theta^* \mid \theta)$ is the probability of proposing θ^* given the current state θ, and if $h(\theta) \propto p(\theta \mid x)$ is proportional to the posterior distribution, then the probability of accepting a proposed θ^* is

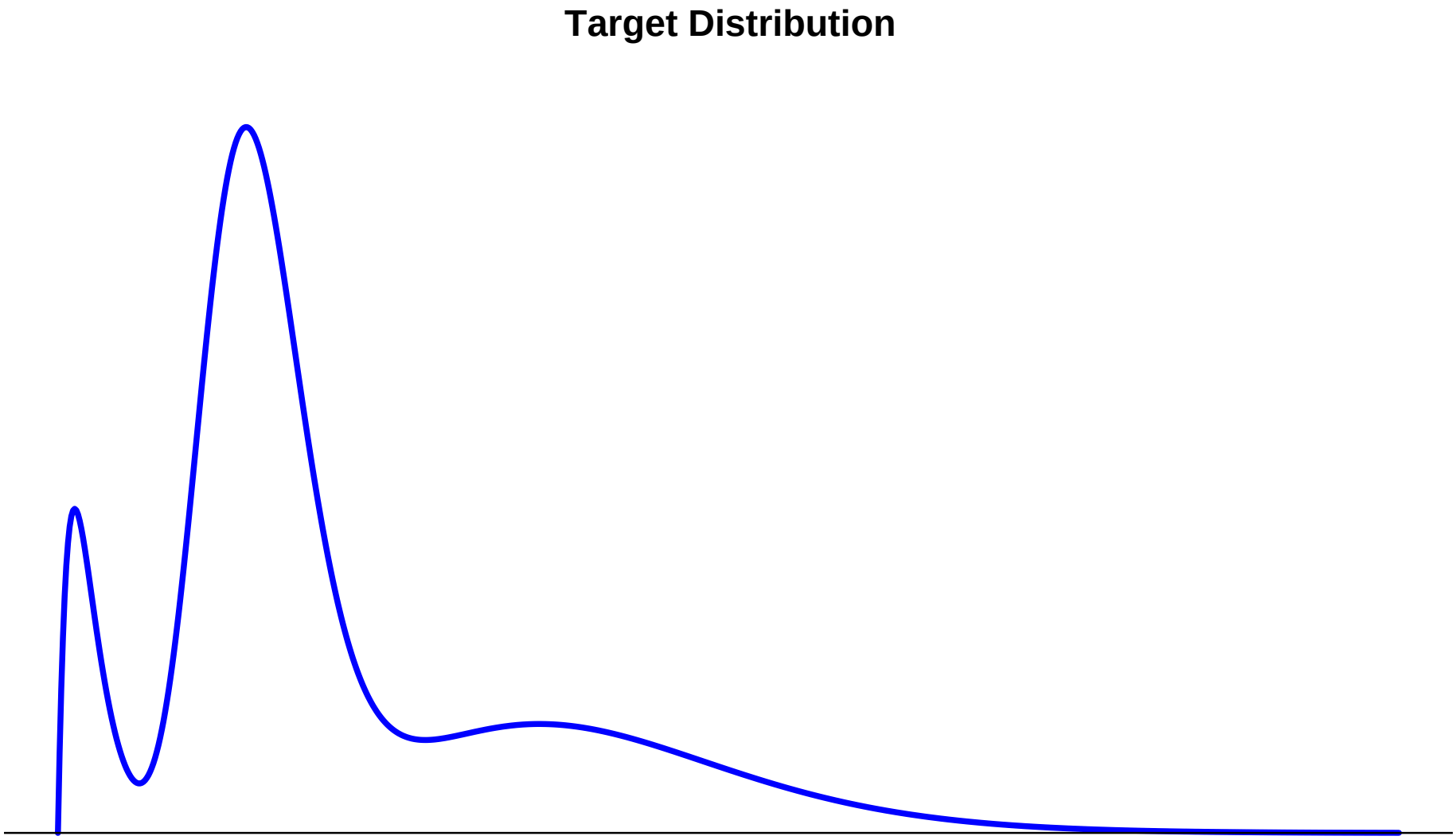
$$\min \left\{ 1, \frac{h(\theta^*)}{h(\theta)} \times \frac{q(\theta \mid \theta^*)}{q(\theta^* \mid \theta)} \right\}$$

- If a proposal is not accepted, the current value θ is *sampled again*.

An MCMC Algorithm

1. Start at θ_0; Set $i = 0$.
2. Propose θ^* from the current θ_i.
3. Calculate the acceptance probability.
4. Generate a random number.
5.
    1. If accepted, set $\theta_{i+1} = \theta^*$.
    2. If rejected, set $\theta_{i+1} = \theta_i$.
6. Increment i to $i + 1$.
7. Repeat steps 2 through 6 many times.

MCMC Example



First Point

Initial Point

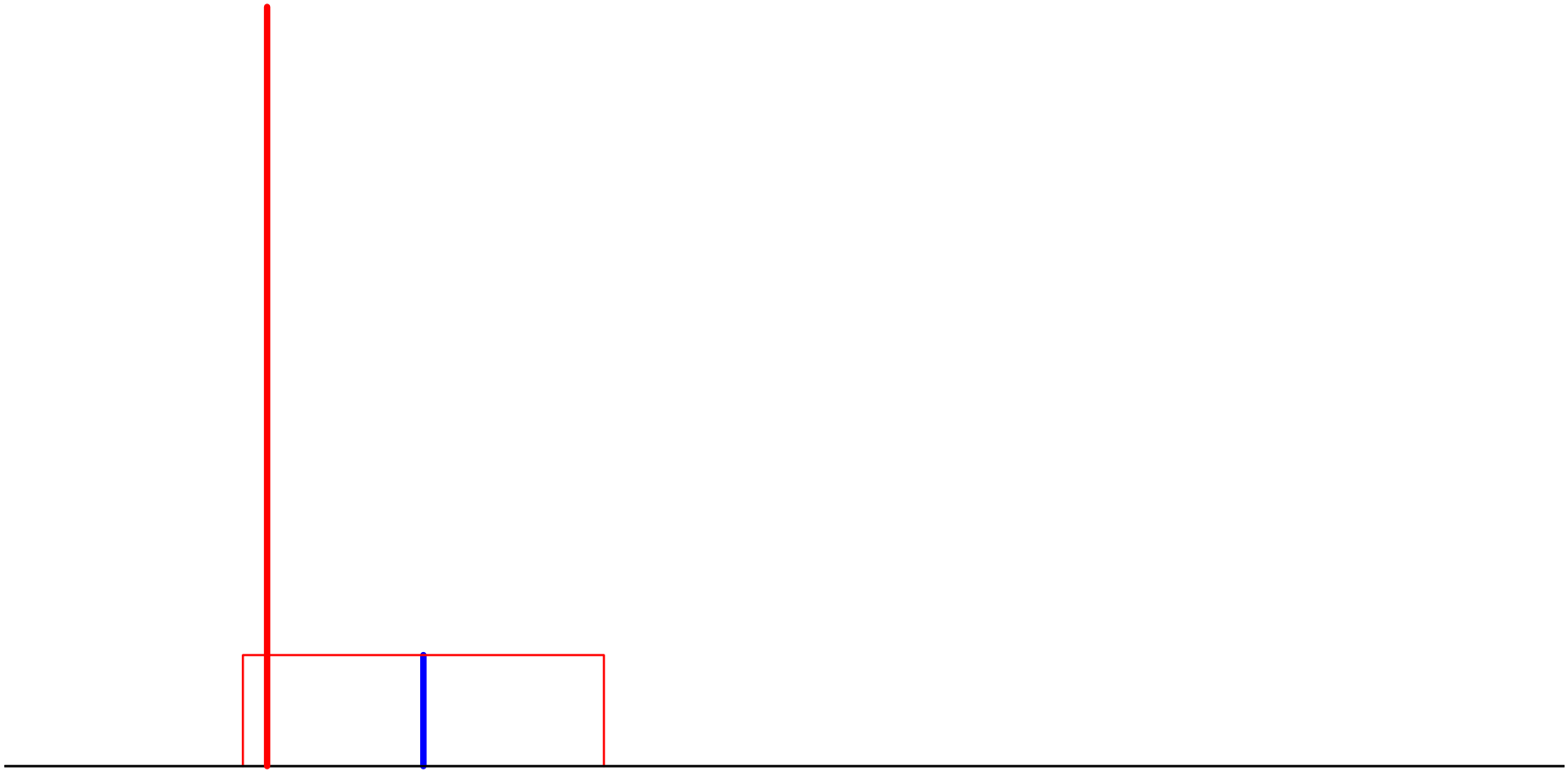
Proposal Distribution

Proposal Distribution

First Proposal

First Proposal

Accept with probability 1



Second Proposal

Second Proposal

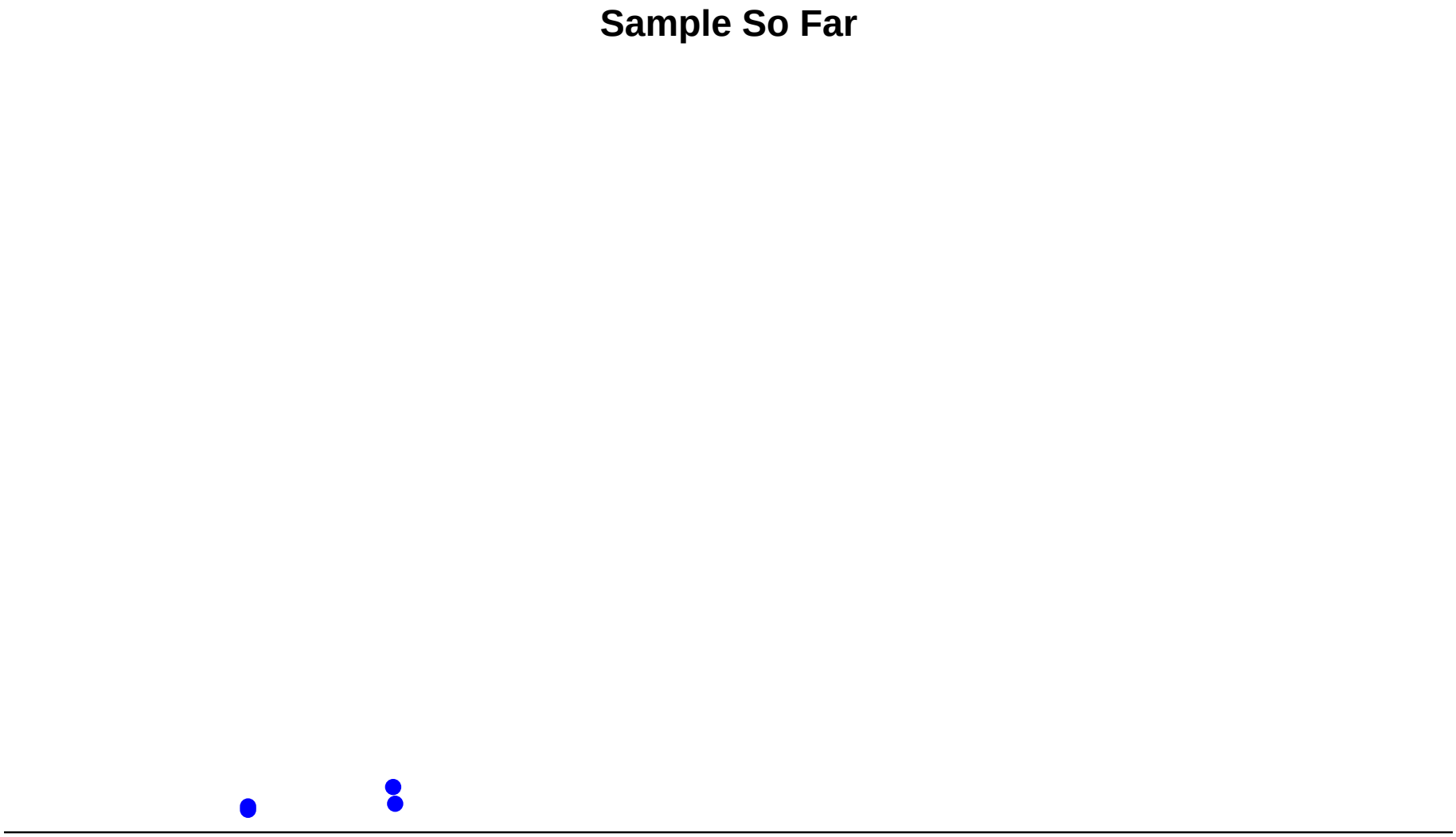
Accept with probability 0.153

Third Proposal

Third Proposal

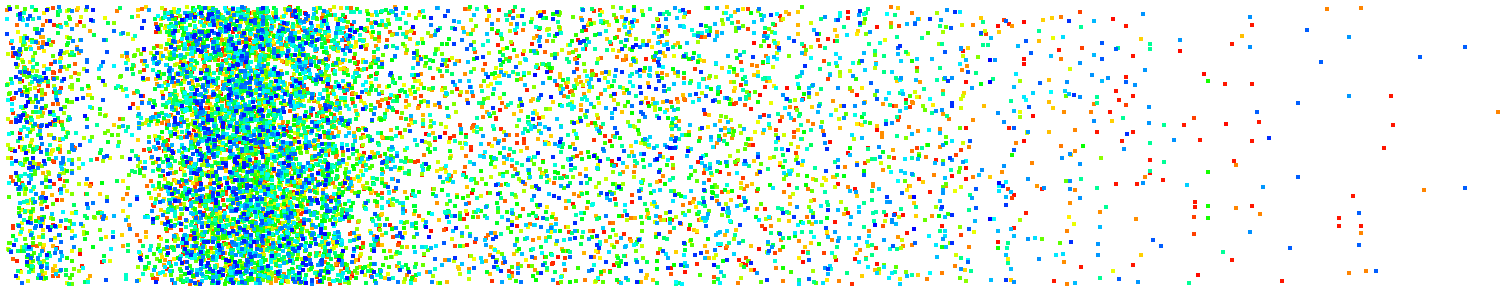
Accept with probability 0.144

Beginning of Sample

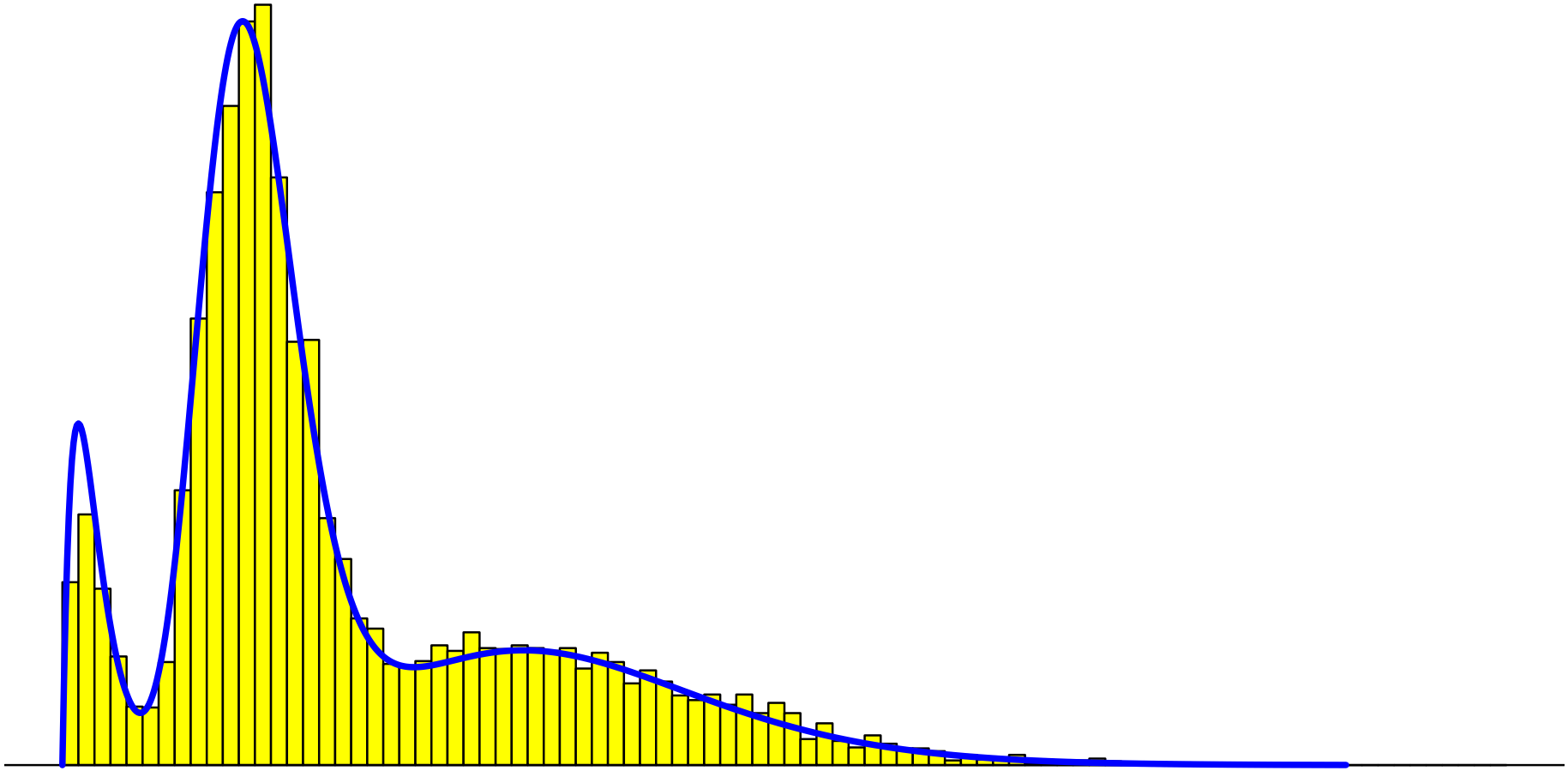


Larger Sample

Second Proposal



Comparison to Target



MCMC for Phylogenies

- The parameter space includes:
  - The tree topology;
  - The branch lengths;
  - Substitution model parameters;
- In practice, we use several MCMC proposals that leave some parameters fixed while changing others.

Bayesian Inference

- The result of an MCMC analysis is a *sample from the posterior distribution*.
- Sample statistics are estimates of corresponding posterior estimates.
  - The sample proportion of a give tree topology converges to the posterior probability of that tree topology;
  - The proportion of trees with a given clade converge to the posterior probability of that clade;
  - The ends of the middle 95% of the sample for the transition/transversion bias κ is an interval estimate for κ.

Summarizing a Posterior Distribution

- A *consensus tree* from an MCMC sample is simply a summary of the posterior distribution of the topology.
- Other summaries are possible.
- This consensus tree is not an *optimal tree* according to some criterion such as maximum likelihood or parsimony.

Cautions

- MCMC does not always converge;
- Should always run several chains with different random numbers and compare answers;
- If the true tree has some very short internal edges, Bayesian inference can mislead;
- Different likelihood models can lead to different results.

Bayesian Inference

- Development of Bayesian methods has led to continual improvement in our ability to model and learn about molecular evolution.
- Bayesian Inference uses likelihood, but requires a *prior distribution*.
- Bayesian inference is computationally intensive, but can be less so than ML plus bootstrapping.
- Bayesian inference directly measures items of interest on an easily interpretable probability scale.
- Some folks dislike the requirement of specifying a prior distribution.