Bayesian Phylogenetics

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

October 6, 2011

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 Box B_1 contains four balls, three of which are black and one of which is white.
- Box B_2 has four balls, two of which are black and two of which are white.
- Box B_3 has four balls, one of which is black and three of which are white.

- If a ball is chosen *uniformly at random* from Box B_1, there is a 3/4 chance that it is black.
- But if a black ball is drawn, how likely is it that it came from Box B_1?

What is Bayes' Theorem?

- If a black ball is drawn, how likely is it that it came from Box B_1?
- To answer this question, we need to have prespecified probabilities of which box we pick to draw the ball from.
- 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.
- Do the problem with a probability tree.

Bayes' Theorem

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

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

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

Connection to Phylogeny

- In a Bayesian approach to phylogenetics, the boxes are like different tree topologies, only one of which is right.
- The colored balls are like site patterns, except that there are many more than two varieties and we are able to observe multiple independent draws from each box.
- Things are further complicated in that additional parameters such as branch lengths and likelihood model parameters affect the likelihood, but are also unknown.

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
Model	Likelihood	Likelihood and Prior Distribution

Bayesian Phylogenetic Methods

- Let's say we want to find the posterior probability of a clade.
- Then we are interested in computing

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

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

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

Bayesian Phylogenetic Methods

- So, we need to compute:

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

- However, P(data) is 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

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

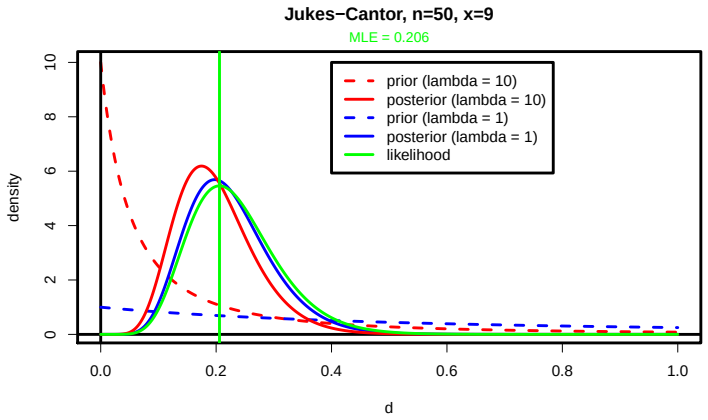
  where $\lambda > 0$ is a parameter.
- This density is what you get if you take the ratio of two independent exponential random variables, one with parameter λ and one with parameter 1.
- The median is $1/\lambda$, but the mean is $+\infty$.

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



What is Markov Chain Monte Carlo?

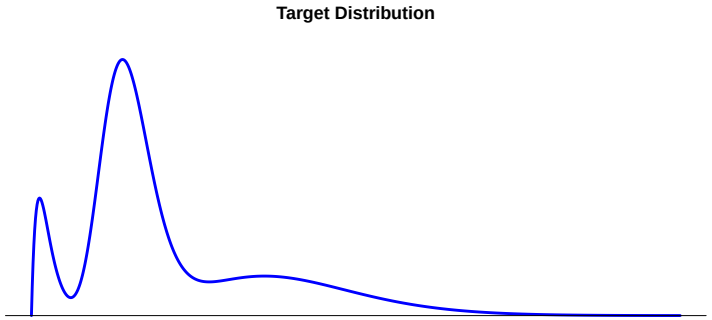
- Markov chain Monte Carlo (MCMC) is a method to take (dependent) samples from a distribution.
- The distribution need only be known up to a constant of proportionality.
- MCMC is especially useful for computation of Bayesian posterior probabilities.
- Simple summary statistics from the sample converge 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.

An MCMC Algorithm

1. Start at x_0; Set $i = 0$.
2. Propose x^* from the current x_i.
3. Calculate the acceptance probability.
4. Generate a random number.
5.
   1. If accepted, set $x_{i+1} = x^*$.
   2. If rejected, set $x_{i+1} = x_i$.
6. Increment i to $i + 1$.
7. Repeat steps 2 through 6 many times.

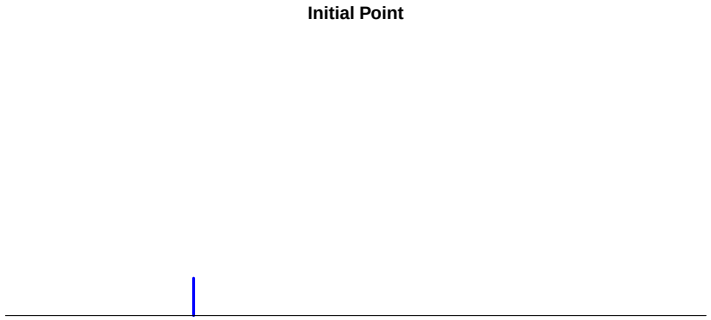
Example

- We have a function $h(\theta)$ from which we want to sample.
- We only need to know h up to a normalizing constant.



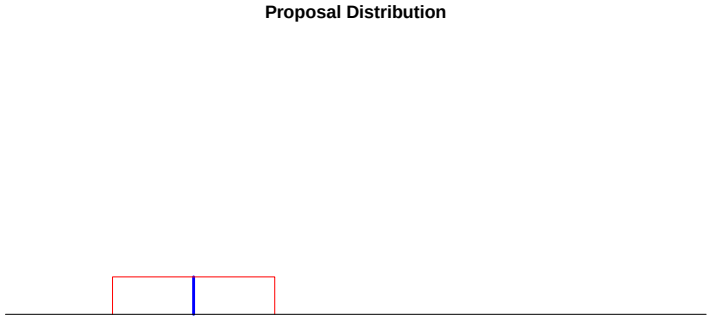
Initial Point

- We begin the Markov chain at a single point.
- We evaluate the value of h at this point.



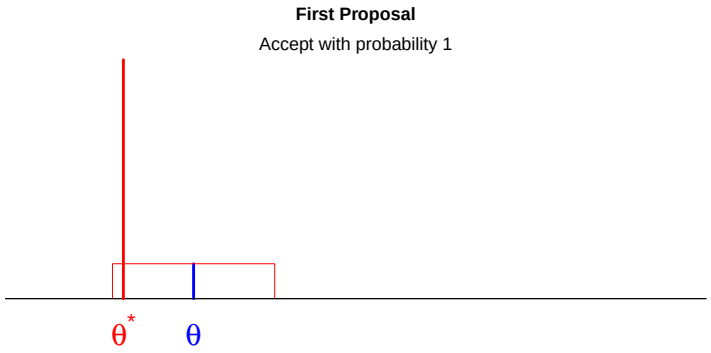
Proposal Distribution

- Given our current state, we have a proposal distribution for the next candidate state.



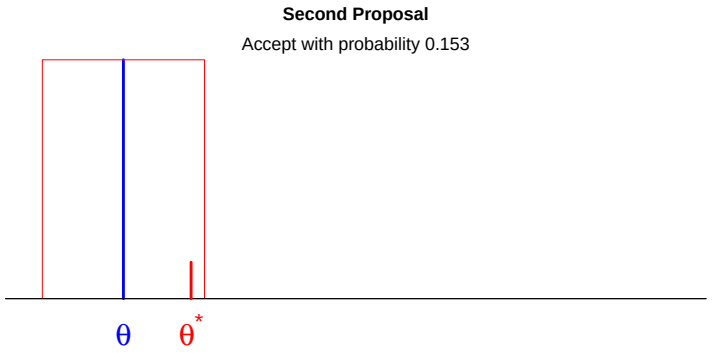
First Proposal

- We propose a *candidate* new point.
- Current state θ; Proposed state θ^*
- This proposal is accepted.



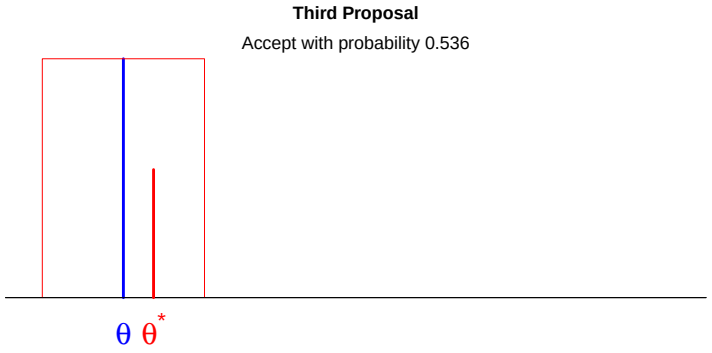
Second Proposal

- The proposal was accepted, so proposed state becomes current.
- Current state θ; Proposed state θ^*; Make another proposal.
- This proposal is rejected.



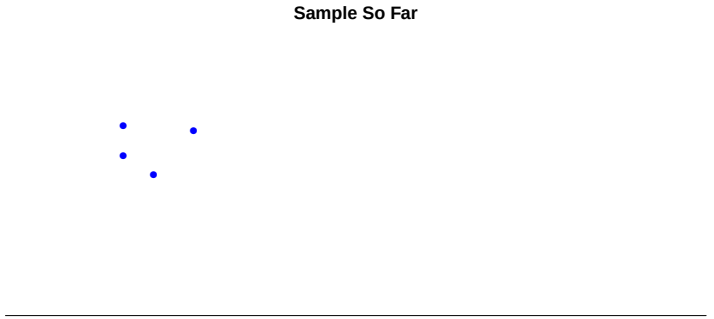
Third Proposal

- The proposal was rejected, so proposed state *is sampled again* and remains current.
- Current state θ; Proposed state θ^*; Make another proposal.
- This proposal is accepted.



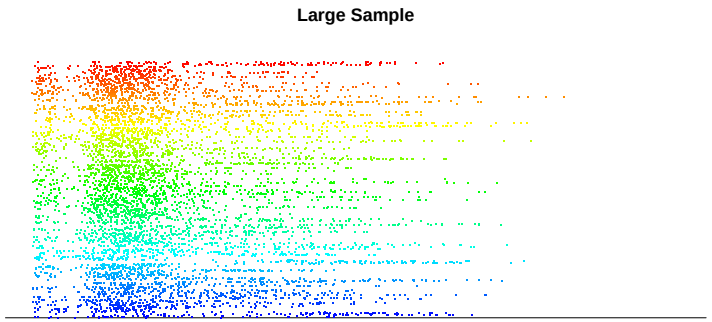
Beginning of Sample

- The first four sample points.
- Vertical position is random to separate points at the same point.

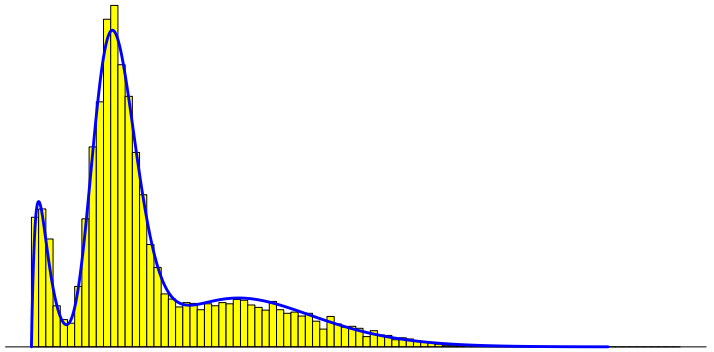


Larger Sample

- Repeat this for 10,000 proposals and show the sample.



Comparison to Target



Things to Note

- The resulting sample mimics the target sample very well.
- The shape of the proposal distribution *did not depend on the target distribution at all*: almost any type of proposal method would have worked.
- There is a lot of *autocorrelation*: MCMC produces *dependent samples*.
- The acceptance probabilities depend on the proposal distributions and *relative* values of the target.
- Summaries of the sample are *good estimates* of corresponding target quantities:
  - The sample mean converges to the mean of the target.
  - The sample median converges to the median of the target.
  - The sample tail area above 1.0 converges to the relative area above 1.0 in the target.

MCMC for Phylogenetics

- The model parameters for a Bayesian phylogenetics analysis typically includes:
  - a tree (topology and branch lengths);
  - substitution process parameters.
- There are most often multiple MCMC methods used in combination.
- For example, methods may:
  - Adust the stationary distribution, leaving other things fixed;
  - Adjust the rates, leaving the tree fixed;
  - Adjust some branch lengths, leaving the topology and Q fixed;
  - Adjust the tree in a small region, leaving the rest of the tree fixed;
  - and so on.

Cautions

- It is important to discard an initial portion of the sample as *burnin*.
- The MCMC sampler must be run for a long time after reaching stationarity.
- It is good practice to make several independent runs to assess agreement; chains can get stuck in local regions, leading to inaccurate inferences.
- Problems with many taxa or very long sequences are more likely to have computational problems.