An Introduction to Phylogenetics

Bret Larget
larget@stat.wisc.edu

Departments of Botany and of Statistics
University of Wisconsin—Madison

January 27, 2010

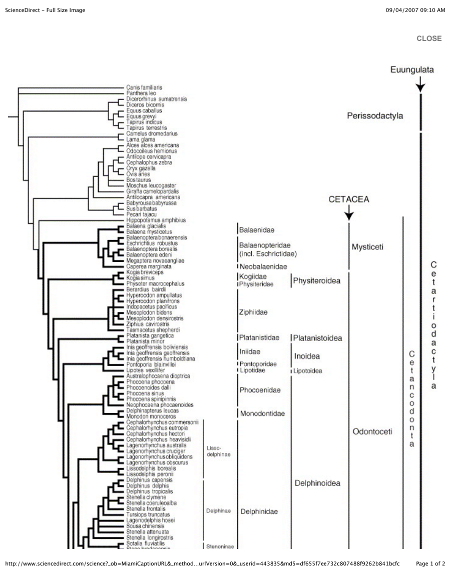
Modern Phylogenetics

- Phylogenies are usually estimated from aligned DNA sequence data.
- Phylogenetics is the primary tool for systematics.
- Phylogenetics is used for studying viruses such as HIV and Influenza..
- Phylogenetics has been used in court for forensic purposes.
- Phylogenetics is being used increasingly in comparative genomics and study of gene function.

Phylogenetics and Systematics

- Phylogenetic methods, particularly for molecular sequence data, have become the primary tool for systemicists to determine evolutionary relationships.
- These tools have been used to confirm expected relationships — for example, that chimpanzees are the closest living relative to humans —
- and have also been key in revealing several more surprising findings, including:
  - birds are descended from dinosaurs;
  - polar bears form a monophyletic group within brown bears;
  - the most closely related land mammal to whales is the hippopotamus.

Phylogenetic Tree of Whales



Phylogenetics and Forensics

- Phylogenetic trees have been used in several instances in the courts to provide evidence about the likely transmission of HIV.
- Examples include:
  - Confirming that a nurse contracted HIV from mishap with a broken glass blood collection tube from an infected patient and not from an alternative source;
  - Providing evidence of deliberate infection in a criminal case;
  - Indicated that an infected friend was likely not the direct source of infection in a case.

Forensic Phylogenetic Tree

Reviews

30 HIV Forensics

Figure 2

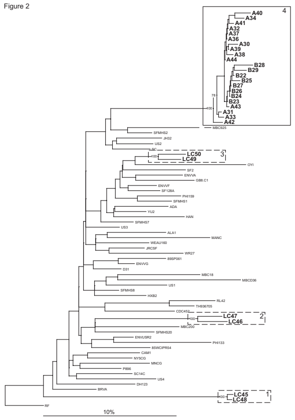


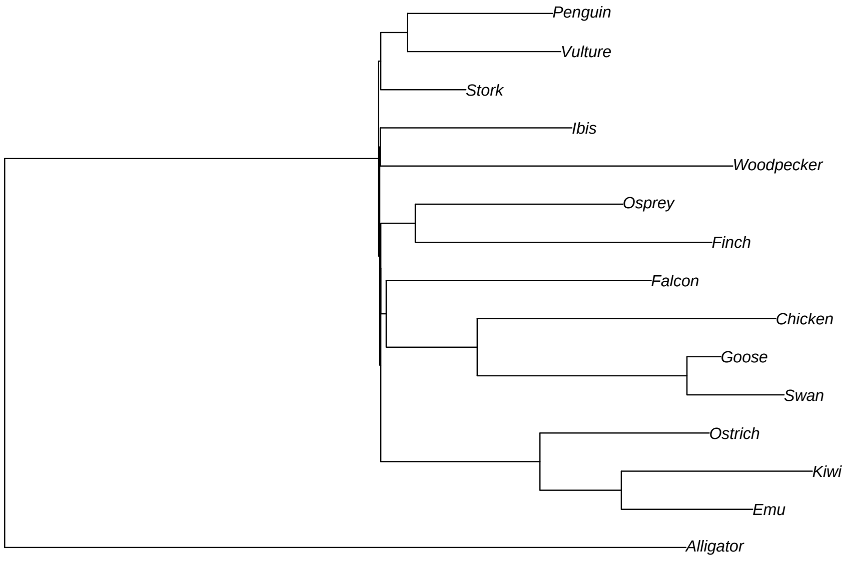
Figure 2. Neighbor-joining phylogram representing the reconstruction of the phylogenetic relationships between the env (C2-V5) sequences obtained from the index case (A31-44), the alleged recipient (B22-29), three local controls (LC45 and LC48; LC46 and LC47; and LC49 and LC50) and 48 sequences chosen from GenBank. Ten iterations of random sequence addition were used. Scale bar represents 10% genetic distance. Bootstrap values are shown at nodes with greater than 70% support.

DNA Data from a Sample of Birds

First 24 bases of 1558 from Cox I gene.

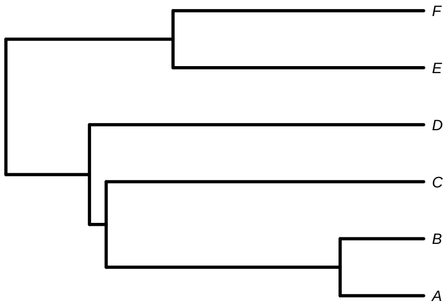
```
Alligator  GTG AAC TTC CAC --- CGT TGA CTC ...
Emu        GTG ACA TTC ATT ACT CGA TGA TTT ...
Kiwi       GTG ACC TTT ACT ACT CGA TGA CTC ...
Ostrich    GTG ACC TTC ATT ACT CGA TGA CTT ...
Swan       GTG ACC TTC ATC AAC CGA TGA CTA ...
Goose      GTG ACC TTC ATC AAC CGA TGA CTA ...
Chicken    GTG ACC TTC ATC AAC CGA TGA TTA ...
Woodpecker GTG ACC TTC ATC AAC CGA TGA TTA ...
Finch      ATG ACA TAC ATT AAC CGA TGA TTA ...
Ibis       GTG ACC TTC ATC AAC CGA TGA CTA ...
Stork      GTG ACC TTC ATT ACC CGA TGA CTA ...
Osprey     ATG ACA TTC ATC AAC CGA TGA CTA ...
Falcon     GTG ACC TTC ATC AAC CGA TGA CTA ...
Vulture    ATG ACA TTC ATC AAT CGA TGA CTA ...
Penguin    GTG ACC TTC ATT AAC CGA TGA CTA ...
```

An Estimated Phylogeny



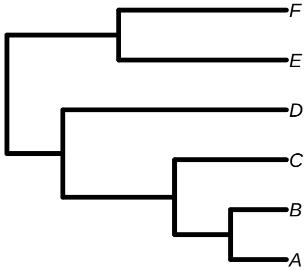
Activity 1: Example Tree

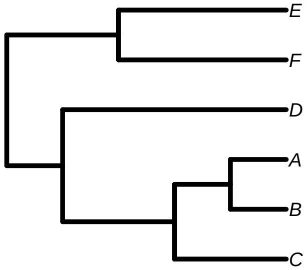
- How many descendent taxa does the common ancestor of taxa *A* and *C* have?
- Which taxon is sister to *A*?
- Which taxa are more closely related, *A* and *C* or *C* and *D*?
- Which taxa are more closely related, *A* and *E* or *D* and *E*?

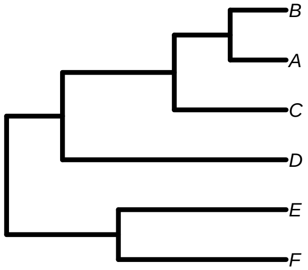


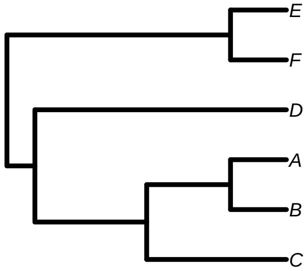
Activity 2: Compare Trees

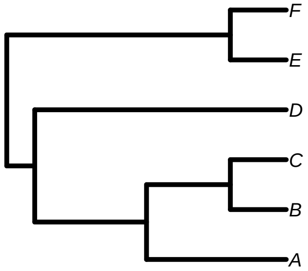
Which trees have the same tree topology?

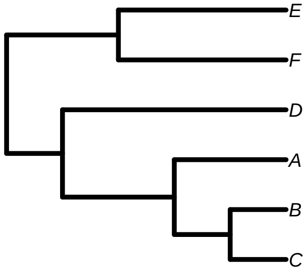






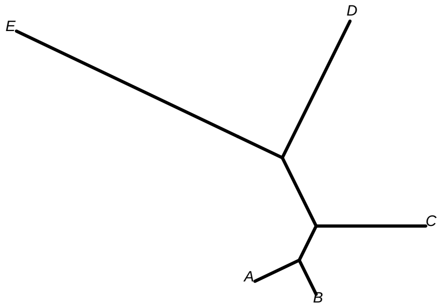






Activity 3: Unrooted Trees

- Some methods estimate unrooted trees.
- If *C* is the outgroup, what is the rooted tree topology?
- If taxon *C* is the outgroup, which node is sister to *B*?
- If taxon *A* is the outgroup, which node is sister to *B*?
- How many rooted tree topologies are consistent with this unrooted tree topology?



How Many Trees?

# of Taxa	# Unrooted Trees	# Rooted Trees
1	1	1
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425
11	34459425	654729075
12	654729075	13749310575
13	13749310575	316234143225

Formula for Counting Trees

- The number of rooted tree topologies with n taxa is $1 \times 3 \times \cdots \times (2n-3) \equiv (2n-3)!!$ for $n \geq 3$.
- There are more rooted trees with 51 species (2.7×10^{78}) than estimated # of hydrogen atoms in the universe (1.3×10^{77}).
- Biologists often estimate trees with more than 100 species.

Probabilistic Framework

Essentially, all models are wrong, but some are useful.

George Box

- Commonly used models of molecular evolution treat sites as independent.
- These common models just need to describe the substitutions among four bases — A, C, G, and T — at a single site over time.
- The substitution process is modeled as a continuous-time Markov chain.

Markov Property

- Use the notation $X(t)$ to represent the base at time t.
- $X(t) \in \{A, C, G, T\}$ for DNA.
- Formal statement:

$$\begin{aligned} & \mathrm{P}\left\{X(s+t)=j \mid X(s)=i, X(u)=x(u) \text{ for } u<s\right\} \\ = \quad & \mathrm{P}\left\{X(s+t)=j \mid X(s)=i\right\} \end{aligned}$$

- Informal understanding: given the present, the past is independent of the future
- If the expression does not depend on the time s, the Markov process is called homogeneous.

Rate Matrix

- A stationary, homogeneous, continuous-time, finite-state-space Markov chain is parameterized by a rate matrix where:
  - off-diagonal rates are nonnegative;
  - diagonal terms are negative row sums of off-diagonal elements;
  - consequently, row sums are zero.
- Example:

$$Q = \{q_{ij}\} = \begin{pmatrix} -1.1 & 0.3 & 0.6 & 0.2 \\ 0.2 & -1.1 & 0.3 & 0.6 \\ 0.4 & 0.3 & -0.9 & 0.2 \\ 0.2 & 0.9 & 0.3 & -1.4 \end{pmatrix}$$

Alarm Clock Interpretation

- How to simulate a continuous-time Markov chain beginning in state i.
  - time to the next transition $\sim \mathrm{Exponential}(q_i)$ where $q_i \equiv -q_{ii}$.
  - transition is to state j with probability

$$\frac{q_{ij}}{\sum_{k \neq i} q_{ik}}$$

Path Probability Density Calculation

- Example: Begin at A, change to G at time 0.3, change to C at time 0.8, and then no more changes before time $t = 1$.

$$\begin{aligned} \mathrm{P}\{\text{path}\} &= \mathrm{P}\{\text{begin at A}\} \\ &\quad \times \left(1.1\mathrm{e}^{-(1.1)(0.3)} \cdot \frac{0.6}{1.1}\right) \\ &\quad \times \left(0.9\mathrm{e}^{-(0.9)(0.5)} \cdot \frac{0.3}{0.9}\right) \\ &\quad \times \left(\mathrm{e}^{-(1.1)(0.2)}\right) \end{aligned}$$

Probability Transition Matrices

- The transition matrix is $P(t) = \mathrm{e}^{Qt}$ where

$$\mathrm{e}^{A} = \sum_{k=0}^{\infty} \frac{A^k}{k!} = I + A + \frac{A^2}{2} + \frac{A^3}{6} + \cdots$$

- A probability transition matrix has non-negative values and each row sums to one.
- Each row contains the probabilities from a probability distribution on the possible states of the Markov process.

Examples

$$P(0.1) = \begin{pmatrix} 0.897 & 0.029 & 0.055 & 0.019 \\ 0.019 & 0.899 & 0.029 & 0.053 \\ 0.037 & 0.029 & 0.916 & 0.019 \\ 0.019 & 0.080 & 0.029 & 0.872 \end{pmatrix} \qquad P(0.5) = \begin{pmatrix} 0.605 & 0.118 & 0.199 & 0.079 \\ 0.079 & 0.629 & 0.118 & 0.174 \\ 0.132 & 0.118 & 0.671 & 0.079 \\ 0.079 & 0.261 & 0.118 & 0.542 \end{pmatrix}$$

$$P(1) = \begin{pmatrix} 0.407 & 0.190 & 0.276 & 0.126 \\ 0.126 & 0.464 & 0.190 & 0.219 \\ 0.184 & 0.190 & 0.500 & 0.126 \\ 0.126 & 0.329 & 0.190 & 0.355 \end{pmatrix} \qquad P(10) = \begin{pmatrix} 0.200 & 0.300 & 0.300 & 0.200 \\ 0.200 & 0.300 & 0.300 & 0.200 \\ 0.200 & 0.300 & 0.300 & 0.200 \\ 0.200 & 0.300 & 0.300 & 0.200 \end{pmatrix}$$

Stationary Distribution

- Well-behaved continuous-time Markov chains have a stationary distribution π. (For finite-state-space chains, irreducibility is sufficient.)
- When the time t is large enough, the probability $P_{ij}(t)$ will be close to π_j for each i. (See $P(10)$ from earlier.)
- The stationary distribution can be thought of as a long-run average — the proportion of time the state spends in state i converges to π_i.
- The stationary distribution satisfies $\pi^\top Q = 0^\top$.
- Also, $\pi^\top P(t) = \pi^\top$ for any time t.

Numerical Example

$$\pi^\top Q \quad = \quad 0^\top$$

$$\begin{pmatrix} 0.2 & 0.3 & 0.3 & 0.2 \end{pmatrix} \begin{pmatrix} -1.1 & 0.3 & 0.6 & 0.2 \\ 0.2 & -1.1 & 0.3 & 0.6 \\ 0.4 & 0.3 & -0.9 & 0.2 \\ 0.2 & 0.9 & 0.3 & -1.4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$$

Usual Parameterization

- The matrix $Q = \{q_{ij}\}$ is typically scaled and parameterized

  $$q_{ij} = r_{ij}\pi_j/\mu$$

  for $i \neq j$ where

  $$\mu = \sum_i \pi_i \sum_{j \neq i} r_{ij}\pi_j$$

  which guarantees that π will be the stationary distribution when $r_{ij} = r_{ji}$.
- With this scaling, there is one expected transition per unit time.

Time-reversibility

- A continuous-time Markov chain is time-reversible if the probability of a sequence of events is the same going forward as it is going backwards.
- The matrix Q is the matrix for a time-reversible Markov chain when $\pi_i q_{ij} = \pi_j q_{ji}$ for all i and j.
- That is, the overall rate of substitutions from i to j equals the overall rate of substitutions from j to i for every pair of states i and j.
- The matrix equivalent is $\Pi Q = Q^\top \Pi$ where $\Pi = \mathrm{diag}(\pi)$.

General Time-Reversible Model

- The GTR model is the most general basic time-reversible continuous-time Markov model for nucleotide substitution.
- The model is typically parameterized with 8 free parameters where

$$q_{ij} = \begin{cases} r_{ij}\pi_j/\mu & \text{for } i \neq j \\ -\sum_{j \neq i} q_{ij} & \text{for } i = j \end{cases}$$

  with $\mu = \sum_i \pi_i \sum_{j \neq i} r_{ij}\pi_j$.
  - The stationary distribution π has three free parameters as π sums to one;
  - The vector $r = (r_{AC}, r_{AG}, \ldots, r_{GT})$ is usually constrained to five degrees of freedom (either by setting $r_{GT} = 1$ or constraining the sum).
- Many other popular models are special cases.
- These models are often named by the initials of the authors and the year in which they were published.

Rate Variation Among Sites

- A common extension to the standard CTMC models is to assume that there is rate variation among sites.
- At these sites, the Q matrix is multiplied by a site-specific rate.
- The two most popular extensions are:
  - Invariant sites: some sites have rate 0
  - Gamma-distributed rates: rates are drawn from a mean 1 gamma distribution
- For computational tractability, the Gamma distribution is typically replaced by a mean 1 discrete distribution with four distinct rates based on quantiles of a Gamma distribution.

Other Extensions

- There are many other model extensions in common use and under development.
- It is common to partition sites (by gene, by codon position, by genomic location) and to use different models for each part.
- The covarion model allows different lineages to have different rates at the same site.
- This is typically modeled with a hidden Markov model where the site can turn “off”.
- There are models for amino acid substitution, models for codons, models for RNA pairs, models that incorporate protein structure information, and so on.
- Current models still do not capture much of the important biological processes that affect evolution of molecular sequences.

Distance Between Pairs of Taxa

- In a two-taxon tree, the distance between two taxa can be estimated under any model by maximum likelihood.
- If the distance is t and at site i one species has base A and the other has base C, the contribution to the likelihood at this site j is

$$L_j(t) = \pi_A P_{AC}(t) = \pi_C P_{CA}(t)$$

  for a time-reversible model.
- The overall likelihood is

$$L(t) = \prod_j L_j(t)$$

- and the log-likelihood is

$$\ell(t) = \sum_j \log L_j(t) = \sum_j \left(\log \pi_{x[j]} + \log P_{x[j]y[j]}(t)\right)$$

Distance Between Pairs of Taxa

- For models with free π, it is common to estimate π with observed base frequencies.
- Other parameters are usually estimated by maximum likelihood.
- The simplest models have closed form solutions, others require numerical optimization.

Notation for the Alignment

- An alignment of m taxa and n sites will have mn nucleotide bases.
- Let the observed base for the ith taxon and the jth site be x_{ij}.

Notation for the Tree

- With a time-reversible model, the location of a root (where the CTMC begins at stationarity) does not affect the likelihood calculation.
- We can assume an unrooted tree without loss of generality.
- An unrooted tree with m taxa will have $m - 2$ internal nodes.
- Number these nodes $i = 1, \ldots, 2m - 2$ with the first m for leaf nodes and the last $m - 2$ for internal nodes.
- For calculation purposes, we will denote node ρ (which could be any node) as the root.
- There are $2m - 3$ edges in the tree, numbered $e = 1, \ldots, 2m - 3$.
- Relative to root node ρ, edge e connects parent node $p(e)$ and child node $c(e)$ where $p(e)$ is closer to ρ than $c(e)$.
- Edge e has length t_e.

Notation for Unobserved Data

- The likelihood for a tree is computed by summing over all possible bases at the internal nodes for each of the n sites.
- For each site, there are 4^{m-2} possible allocations of bases at internal nodes we will index by k.
- Internal node i is set to nucleotide b_{ik} at the kth allocation, $i = m + 1, \ldots, 2m - 2$.
- Let $z(i, j, k)$ be the nucleotide at node i, site j, and allocation k.

$$z(i,j,k) = \begin{cases} x_{ij} & \text{if } i \leq m \text{ (i is a leaf node)} \\ b_{ik} & \text{if } i > m \text{ (i is an internal node)} \end{cases}$$

Likelihood of a Tree

- Let $P(t)$ be the 4×4 probability transition matrix over an edge of length t.
- The likelihood of the tree is

$$\prod_j \sum_k \left(\pi_{z(\rho,j,k)} \prod_e P_{z(p(e),j,k)z(c(e),j,k)}(t_e) \right)$$

- Notice that the sum is over the 4^{m-2} possible allocations.
- A naive calculation would not be tractible for large trees.

Felsenstein's Pruning Algorithm

- Felsenstein's pruning algorithm is an example of dynamic programming.
- By saving partial calculations, the time complexity of the likelihood evaluation grows linearly with the number of sites, not exponentially.
- For each site and node, the algorithm depends on calculating the probability in the subtree rooted at that node for each possible base.
- The algorithm begins at the leaves of the tree and recurses to the root.
- The likelihood of the site is a weighted average of the conditional subtree probabilities at the root weighted by the stationary distribution.

Maximum Likelihood Estimation for one Tree

- For a single tree topology, the ML estimation requires optimization of branch lengths and of any parameters in the substitution model.
- Numerical optimization methods are required even for simple models and small trees.

Tree Search

- The search for the maximum likelihood tree conceptually requires obtaining the maximum likelihood for each possible tree topology and then picking the best of these.
- For more than a dozen or so taxa, exhaustive search is non feasible.
- Heuristic search algorithms typically define a neighborhood structure for possible topologies.
- The search goes through neighbors and jumps to the first neighbor with a higher likelihood.
- When all neighbors are inferior to the current tree, the search stops.
- Much improvement has been made in recent years (RAxML and GARLI are two modern ML programs).

Bayesian Inference

- In Bayesian inference, the posterior distribution is proportional to the product of the likelihood and the prior distribution.
- For parameters θ and data D,

$$\mathrm{P}\{\theta \mid D\} = \frac{\mathrm{P}\{D \mid \theta\}\,\mathrm{P}\{\theta\}}{\mathrm{P}\{D\}}\ .$$

- The denominator is the marginal likelihood of the data, which is the integral of the likelihood against the prior distribution.

Bayesian Phylogenetics

- For a phylogenetic problem, the parameter θ typically includes the tree topology, the edge lengths, and parameters for the substitution model.

$$\theta = (\tau, \nu, \phi)$$

- Often we assume independence of these components:

$$\mathrm{P}\{\theta\} = \mathrm{P}\{\tau\}\,\mathrm{P}\{\nu\}\,\mathrm{P}\{\phi\}\ .$$

- In a typical phylogenetic problem, the marginal likelihood cannot be computed as

$$\mathrm{P}\{D\} = \int_{\Theta} \mathrm{P}\{D \mid \theta\}\,\mathrm{P}\{\theta\}\,\mathrm{d}\theta$$

  is a sum of very many terms (one for each topology) where each term is a high-dimensional integral of a complicated function.

Phylogenetic Inference

- We may be interested in the posterior distribution of the tree topology, $\mathrm{P}\{\tau \mid D\}$.
- When this posterior distribution is diffuse, we can summarize it by computing posterior distributions of clades.
- The posterior probability of a clade C is the sum of the posterior probabilities of all tree topologies that contain it.

$$\mathrm{P}\{C \mid D\} = \sum_{\tau : C \in \tau} \mathrm{P}\{\tau \mid D\}$$

- A consensus tree which includes as many clades with high posterior probability as possible is often used as a single tree summary of a distribution of the tree topology.

Sample-based Inference

- Any aspect of a posterior distribution can be estimated from a sample drawn from the distribution.
- For example, the sample proportion of trees with topology τ_0 is an estimate of $\mathrm{P}\{\tau_0 \mid D\}$.
- Also, the sample mean of a transition/transversion parameter κ is an estimate of the posterior mean $\mathrm{E}[\kappa \mid D]$.
- But how do we sample from a complicated posterior distribution?

Markov Chain Monte Carlo

- Markov chain Monte Carlo (MCMC) is a mathematical method for obtaining dependent samples from a target distribution (such as a posterior distribution).
- The idea is to construct a Markov chain whose state space is the parameter space Θ where the stationary distribution of the Markov chain matches the target distribution, say $\mathrm{P}\{\theta \mid D\}$.
- Simulating the Markov chain produces a sample

$$\theta_0, \theta_1, \ldots$$

which, after discarding an initial burn-in portion, may be treated as a dependent sample from the target distribution.

Metropolis-Hastings

- For notational convenience, let the target distribution be $\pi(\theta) = \mathrm{P}\{\theta \mid D\}$.
- The most common form of MCMC uses the Metropolis-Hastings algorithm in which a proposal distribution q which can depend on the most recently sampled θ_i generates a proposal θ^* which is accepted with some probability.
- When accepted, $\theta_{i+1} = \theta^*$.
- When rejected, $\theta_{i+1} = \theta_i$.
- The proposal distribution q is essentially arbitrary provided it can move around the entire space Θ.

Metroplis-Hastings Algorithm

- The acceptance probability is

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

where $|J|$ is a Jacobian.
- Notice the target density appears only as a ratio — this means that it only need be known up to scalar, and we can simply evaluate $h(\theta) = \mathrm{P}\{D \mid \theta\}\,\mathrm{P}\{\theta\}$ since

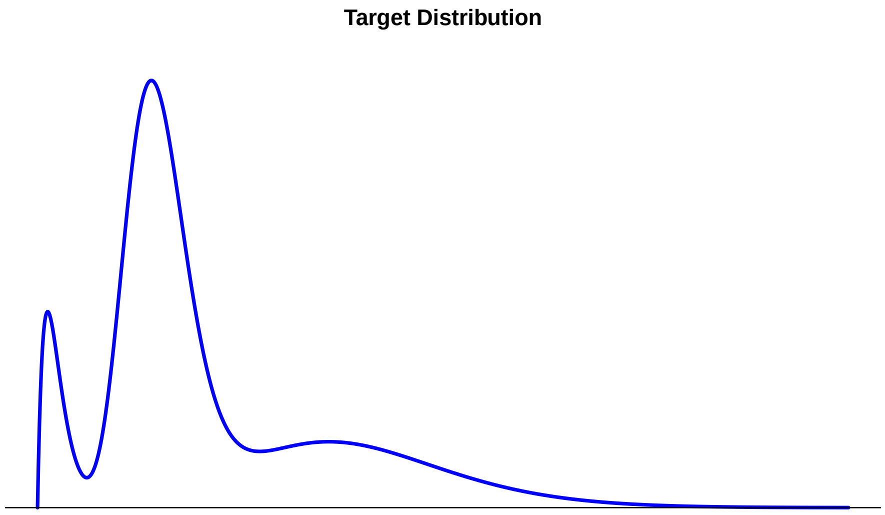
$$\frac{\pi(\theta^*)}{\pi(\theta)} = \frac{\mathrm{P}\{D \mid \theta^*\}\,\mathrm{P}\{\theta^*\}/\mathrm{P}\{D\}}{\mathrm{P}\{D \mid \theta\}\,\mathrm{P}\{\theta\}/\mathrm{P}\{D\}} = \frac{h(\theta^*)}{h(\theta)}$$

- Note that the proposal ratio

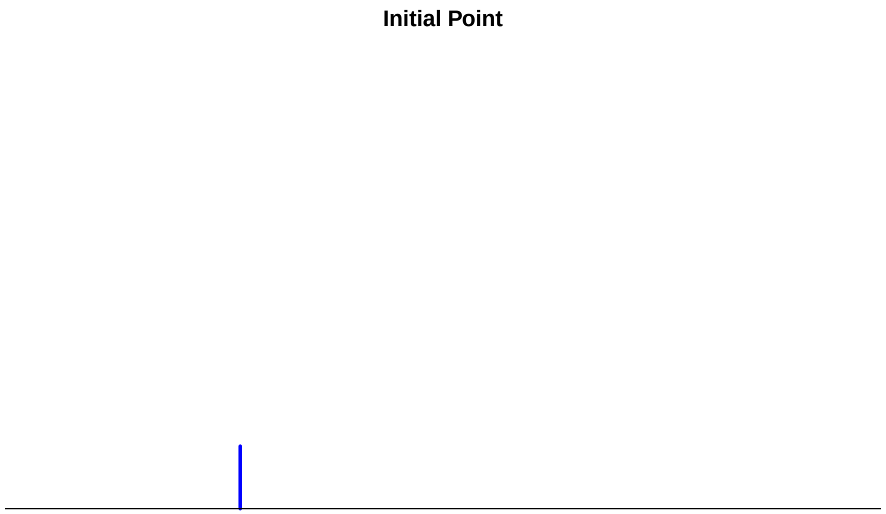
$$\frac{q(\theta \mid \theta^*)}{q(\theta^* \mid \theta)}$$

can be tricky to compute.

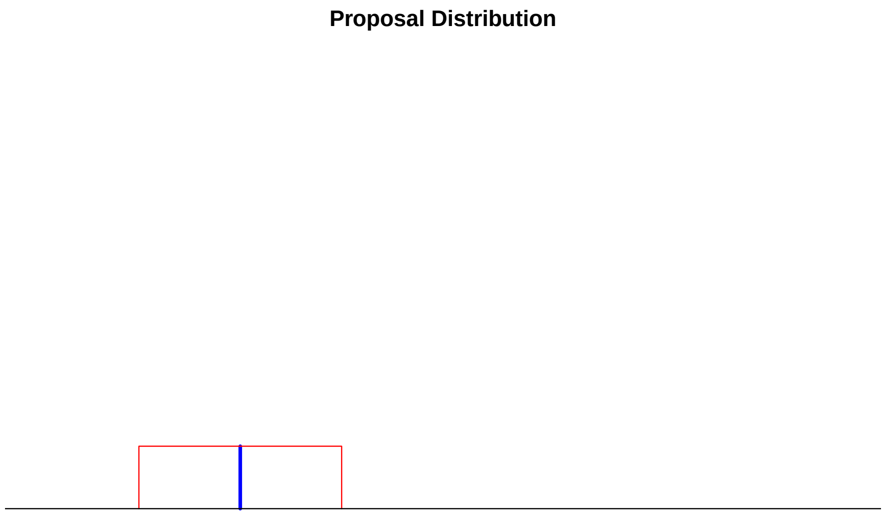
MCMC Example



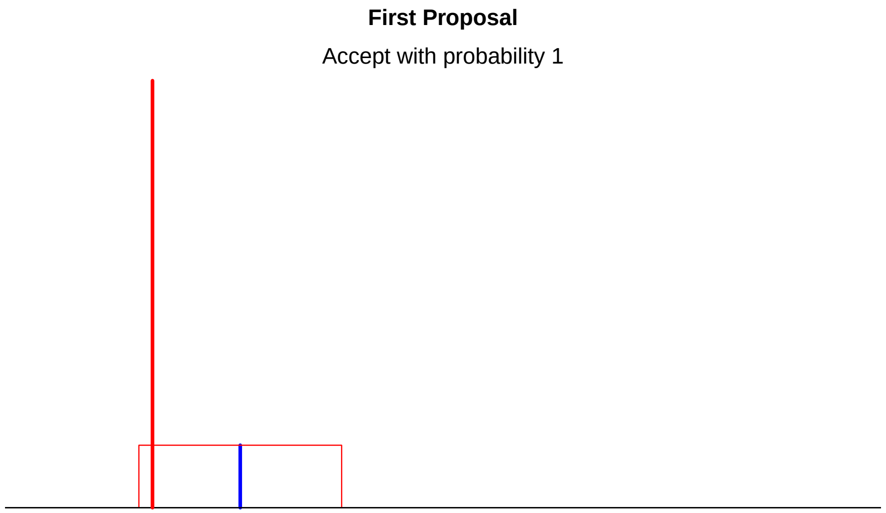
First Point



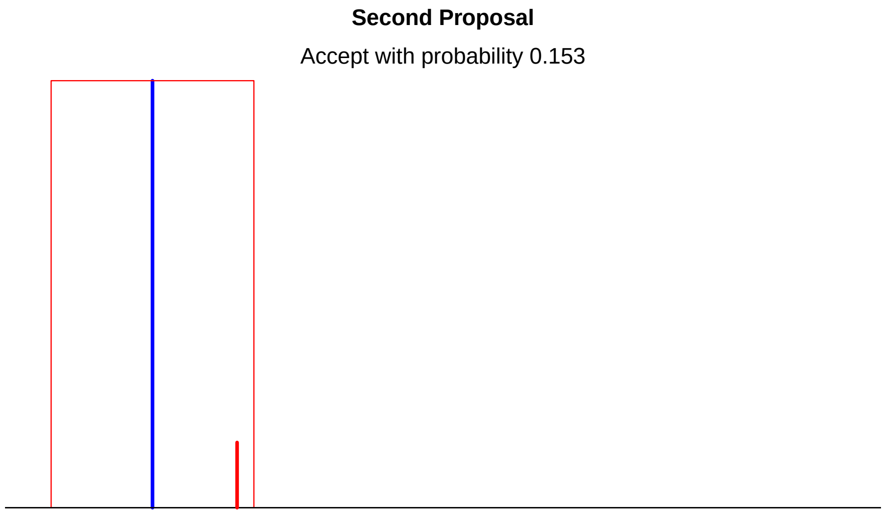
Proposal Distribution



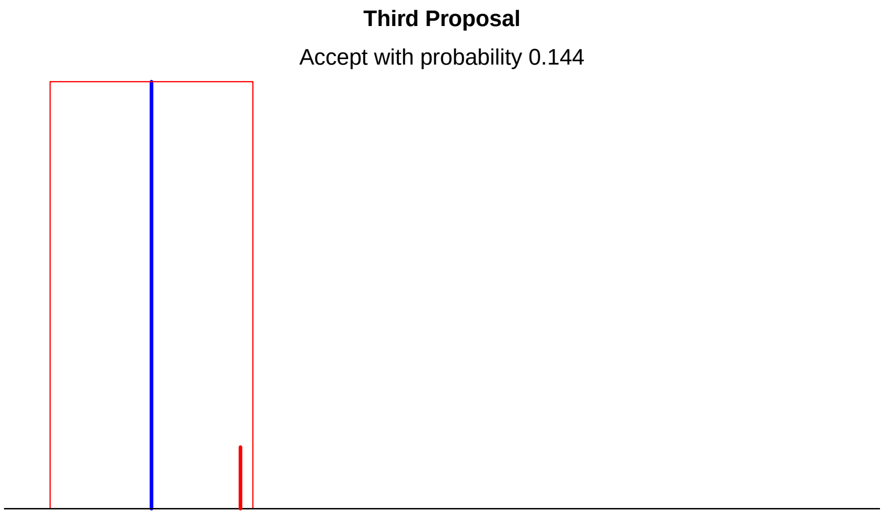
First Proposal



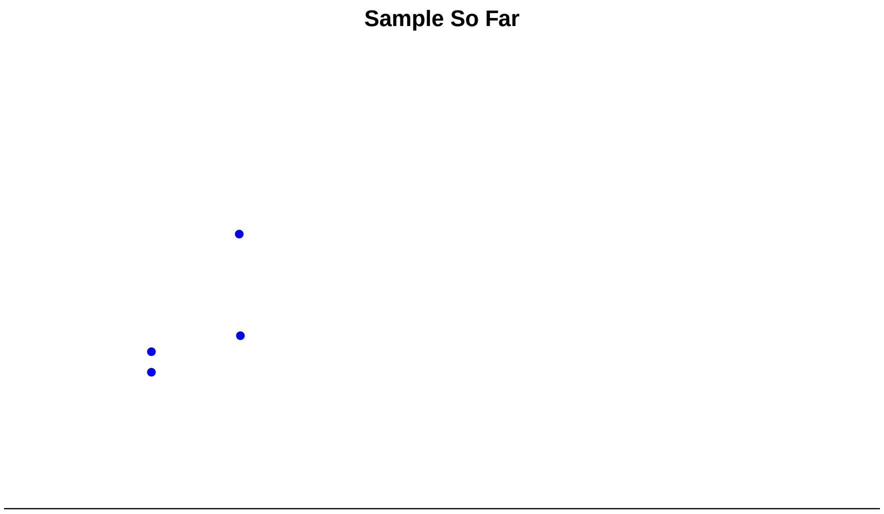
Second Proposal



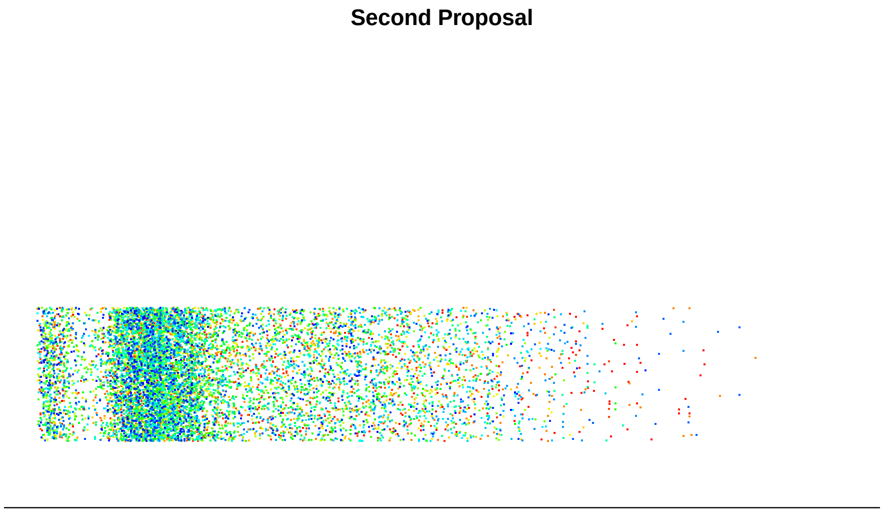
Third Proposal



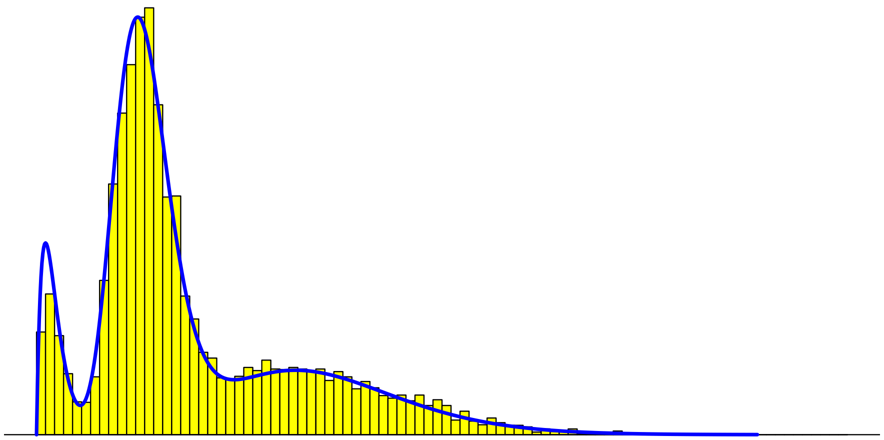
Beginning of Sample



Larger Sample



Comparison to Target



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.