



MCMC Bayesian Methods to Estimate the Distribution of Gene Trees

Dennis Pearl April 27, 2010

Reference:

Ronquist, van der Mark & Huelsenbeck, *chapter 7*
of The Phylogenetic Handbook 2nd edition



**"My average student is doing great.
Half my class thinks $2+2=3$ and the
other half thinks $2+2=5$."**



Outline

- Phylogenetics Background
 - PGK example
- Bayesian Phylogenetics
- Diagnostics

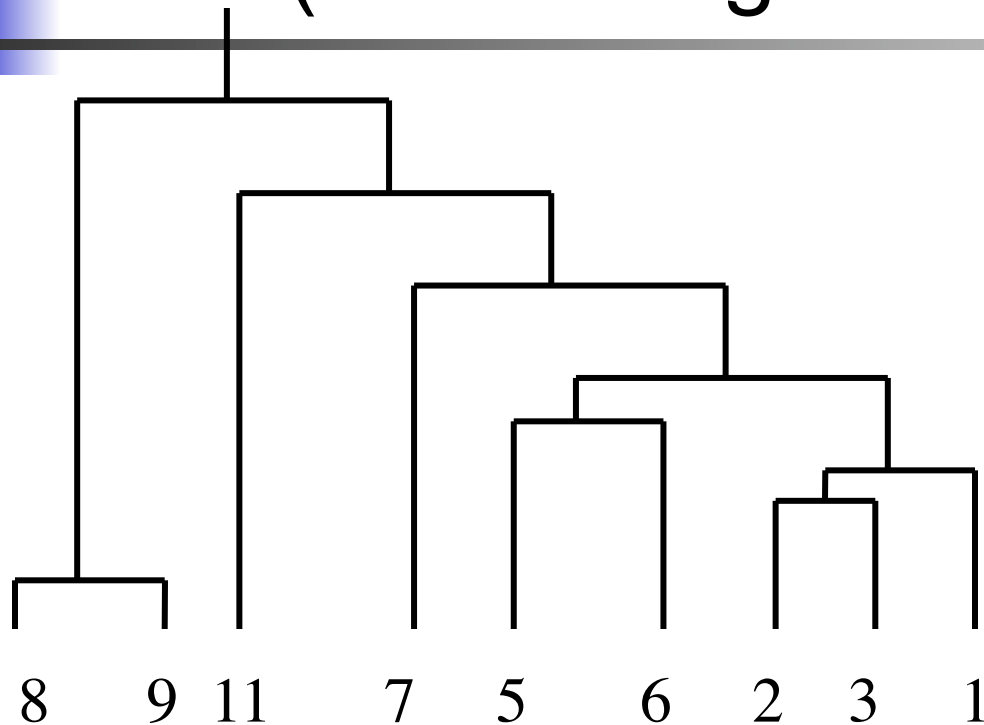
Gene Evolution Example I

Recall: Leitner's Swedish Social Network

The Data: HIV *env* sequences from 9 linked patients.

- Index case patient 1 transmitted the virus to female patients 8, 11, 7, 5, and 2.
- Patient 5 transmitted the virus to male patient 6.
- Patient 2 transmitted the virus to child patient 3
- Patient 8 transmitted the virus to child patient 9
- Times for each transmission are known within a few months.

The Swedish Social Network - (Tree of highest likelihood)

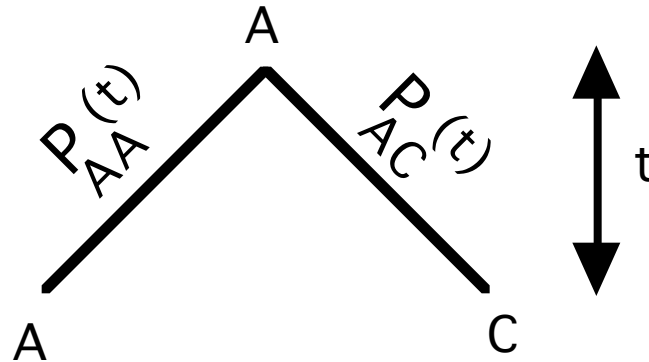


$P(\text{Tree}) =$

$P(\text{Topology})P(\text{Split Times}|\text{Topology})P(\text{Sequences}|\text{Topology, Times})$

Markov Probability Model for calculating $P(\text{Sequences}|\text{Topology, Times})$

- Define $X(t)$ = letter at time t ,
- $P_{ij}(t) = P(X(t+s) = j | X(s) = i)$ (stationarity)



- Matrix notation $\mathbf{P}(t)$ (note: rows add to 1)



A Gene Evolution Example: PGK

The Data: complete Phosphoglycerate Kinase (PGK) sequences in public databases with duplicate species removed.

PGK was chosen for its role in DNA metabolism

131 Amino acid sequences of length 411 ± 23

33 Eukaryotes

98 Prokaryotes (15 Archaea; 83 Bacteria)

(note: there are $\approx 10^{257}$ possible tree topologies relating 131 sequences)

Reference: Pollack, Li, & Pearl, *Molecular Phylogenetics and Evolution*, 2005.

The age of the universe:

$$10^{10} \text{ years}$$

The radius of the universe:

$$6 \times 10^{12} \text{ miles/year} (10^{10} \text{ years}) = 6 \times 10^{22} \text{ miles}$$

The volume of the universe:

$$5.4 \times 10^{69} \text{ miles}^3 \text{ or } 3.9 \times 10^{79} \text{ meters}^3$$

of atoms that could fit in a cubic meter:

$$(5 \times 10^9)^3 = 1.25 \times 10^{29}$$

of atoms that could fit in the universe:

$$5 \times 10^{108} \text{ or about } 2^{261}$$

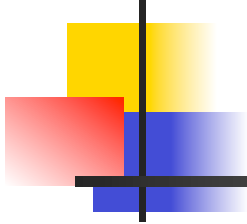
Compare: there are at least $10^{257} \approx 2^{853}$ possible topologies for the PGK data



PGK Gene Evolution Example

HUMAN	...	DFNVPMKNN-QITNNQRIKAAVPSIKFCLDNGAKSVVLMShLGR	...
RAT	...	DFNVPMKNN-QITNNQRIKAAVPSIKFCLDNGANSVVLMShLGR	...
TOBACCO	...	DLNVPLDDNQNITDDTRIRAAVPTIKHLMANGAK-VILSSHLGR	...
WHEAT	...	DLNVPLDDNQNITDDTRIRAAIPTIKYLLSNGAK-VILTShLGR	...
CHICKEN	...	DFNVPMKDH-KITNNQRIKAAVPTIKHCLDHGAKSVVLMShLGR	...

•	•	•
•	•	•
•	•	•



PGK Gene Evolution Example

HUMAN	...	DFNVPMKNN-QITNNQRIKAAVPSIKFCLDNGAKSVVLMShLGR	...
RAT	...	DFNVPMKNN-QITNNQRIKAAVPSIKFCLDNGANSVVLMShLGR	...
TOBACCO	...	DLNVPLDDN Q NITDDTRIRAAVPTIKHLMANGAK-V I LSSHLGR	...
WHEAT	...	DLNVPLDDN Q NITDDTRIRAAIPTIKYLLSNGAK-V I LTShLGR	...
CHICKEN	...	DFNVPMKDH-KITNNQRIKAAVPTIKHCLDHGAKSVVLMShLGR	...
		•	•
		•	•
		•	•

rooted display of 90% central tree based on R-F distance

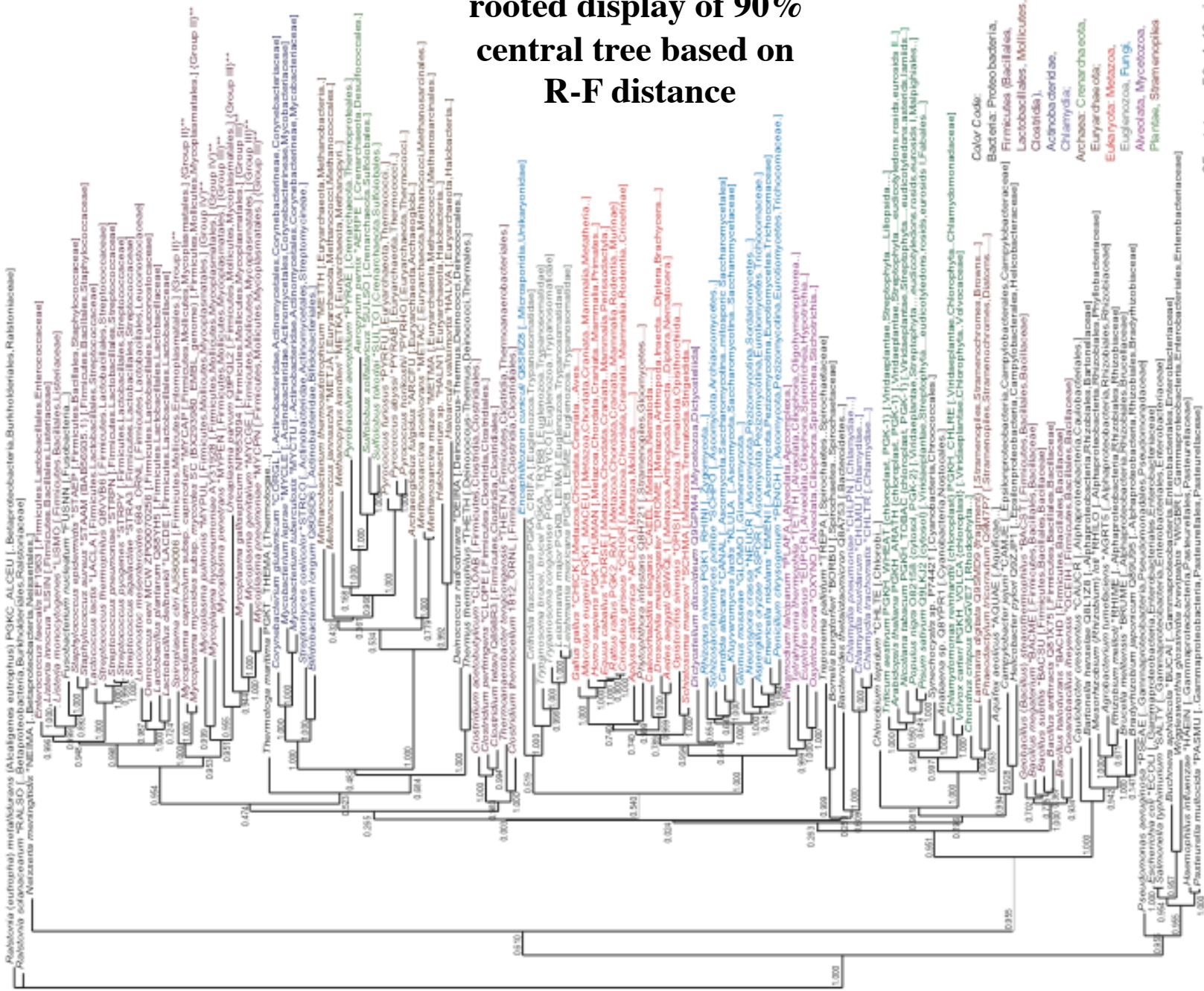


FIGURE 1. * = PGK. ** Grouping of Mollicutes according to Johansson et al. (1998) (see, Discussion)

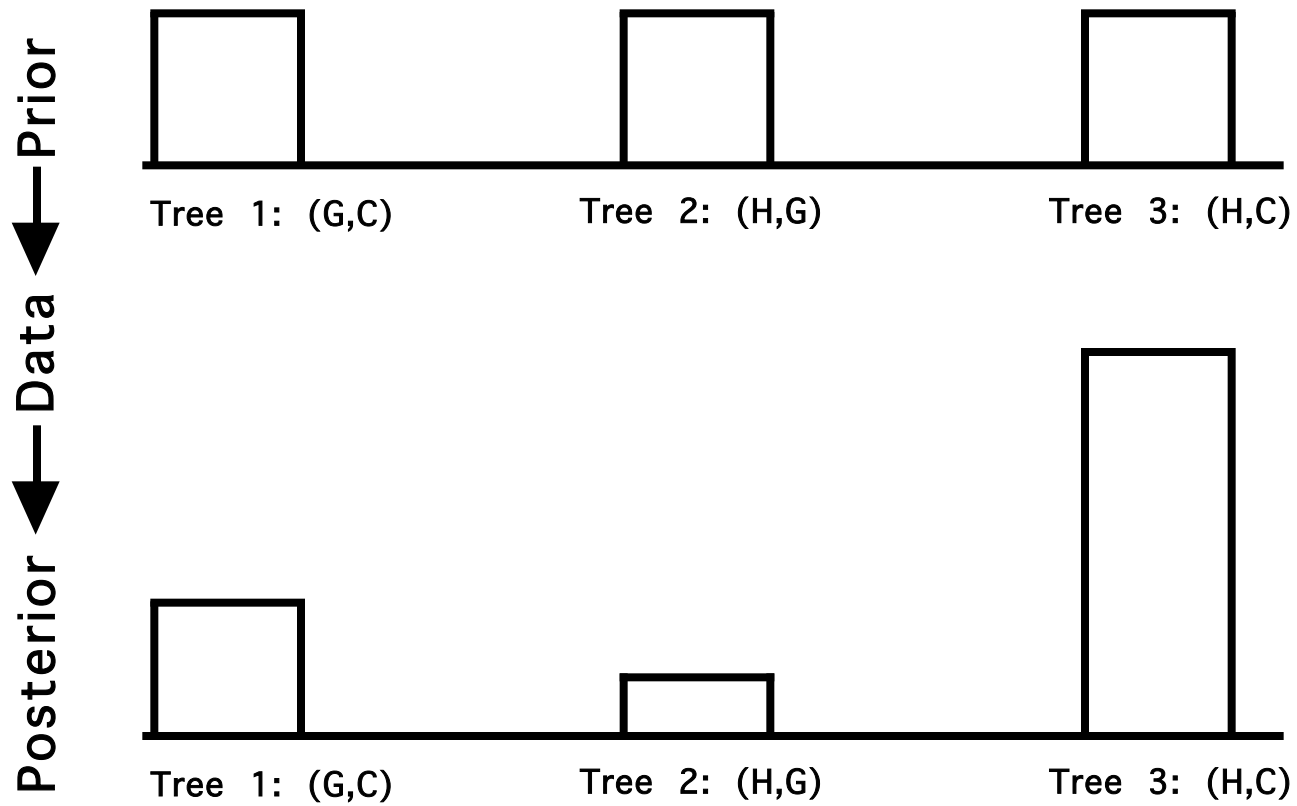
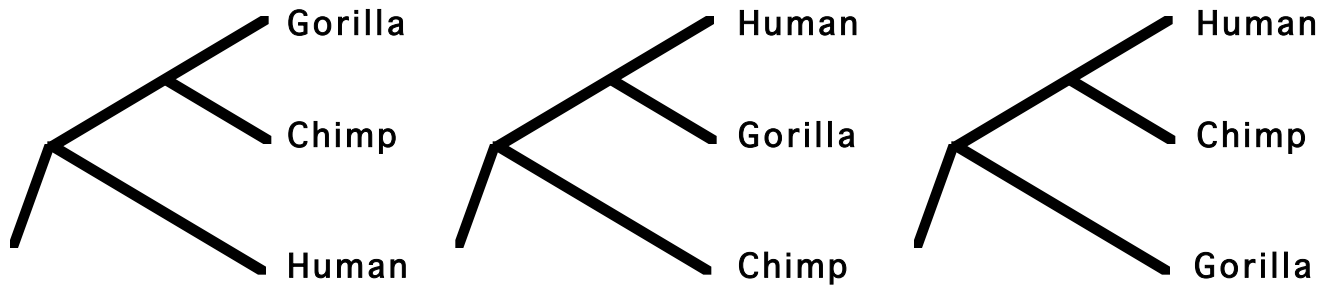
Bayesian Phylogenetics & MCMC

Idea: Find the distribution of trees that would produce the observed data.

$$P(\text{Hypothesis} \mid \text{Data}) = \frac{P(H)P(D \mid H)}{P(D)} = \frac{P(H)P(D \mid H)}{\int p(h)P(D \mid H = h)dh}.$$

In words:

Posterior probabilities are proportional to prior probabilities times the likelihood of the data.



Bayes Theorem tells how prior probabilities can be updated once data is known.



Probability framework for trees

T = topology, \mathbf{t} = branch lengths, \mathbf{v} = internal node sequences, θ = parameters of model of evolution, \mathbf{a} = alignment, and \mathbf{d} = sequence data.

Usually, the sequences at the internal nodes are ignored, so the probability structure is

$$P(T, \mathbf{t}, \theta, \mathbf{a}, \mathbf{d}) = P(\mathbf{d} | T, \mathbf{t}, \mathbf{a}, \theta) P(T, \mathbf{t}, \mathbf{a}, \theta).$$

The first term on the right is the likelihood in the ML approach to phylogeny estimation (which for sequence data is found using the peeling algorithm to sum over all possible values of \mathbf{v}).



Probability framework for trees

The alignment is usually considered as given and the data is fixed, so the purpose of a Bayesian phylogenetic analysis is then to learn about:

$$P(\mathbf{G}|\mathbf{d},\mathbf{a}) = \underset{\substack{\text{posterior} \\ \swarrow}}{P(\mathbf{T},\mathbf{t},\boldsymbol{\theta}|\mathbf{d},\mathbf{a})} = \frac{\underset{\substack{\text{prior} \\ \downarrow}}{P(\mathbf{T},\mathbf{t},\boldsymbol{\theta}|\mathbf{a})} \underset{\substack{\text{likelihood} \\ \downarrow}}{P(\mathbf{d}|\mathbf{T},\mathbf{t},\boldsymbol{\theta},\mathbf{a})}}{\sum_T \int_{\mathbf{t}} P(\mathbf{T},\mathbf{t},\boldsymbol{\theta}|\mathbf{a}) P(\mathbf{d}|\mathbf{T},\mathbf{t},\boldsymbol{\theta},\mathbf{a}) dt}$$

Denominator is a constant (impossible to calculate)



Markov Chain Monte Carlo

Idea: Even though a distribution depends on an impossible to calculate normalizing constant, ratios of values are still accessible. This fact can be exploited to create a Markov chain whose equilibrium distribution is the posterior distribution of interest.



The Metropolis-Hastings Algorithm

- Start with an initial tree x
- Define a density $q(x,y)$ that specifies probabilities of moves from x to a proposed tree y

- Accept y with probability $\min \left\{ \frac{\text{posterior}(y)q(y,x)}{\text{posterior}(x)q(x,y)}, 1 \right\}$

- **Theory:** Repeating these steps many times will create a Markov Chain whose stationary distribution is the desired posterior.



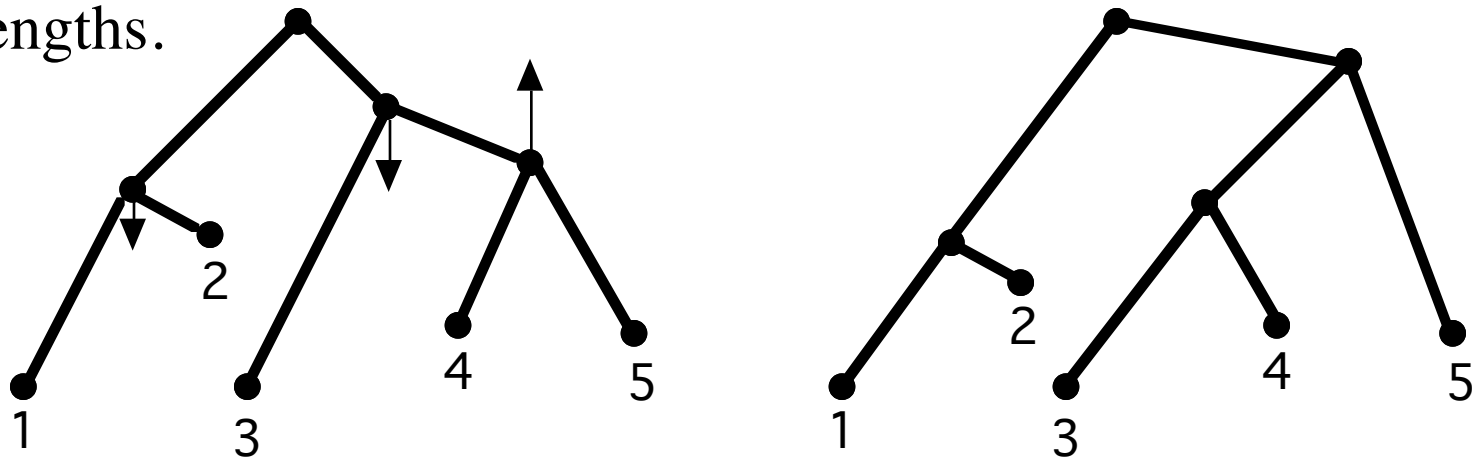
The Metropolis-Hastings Algorithm

To complete the algorithm we need to specify:

- a prior distribution,
- a model for calculating the likelihood,
- a method for proposing new trees for the chain, and
- how long to run the chain to gain a practical approximation to the equilibrium distribution.
 - Need a “Burn-In” for the chain to reach equilibrium
 - Need a long chain at the equilibrium state to provide sufficient samples from the posterior to approximate the full distribution.

MCMC Algorithm (Mau, 1996)

- Prior: $T \sim$ Uniform on “histories,” $t|T$ & θ also “uninformative”
- Move in tree space based on simultaneous perturbations in the branch lengths.



- Calculations require use of Felsenstein’s peeling technique at each step but moves rapidly through topology space.

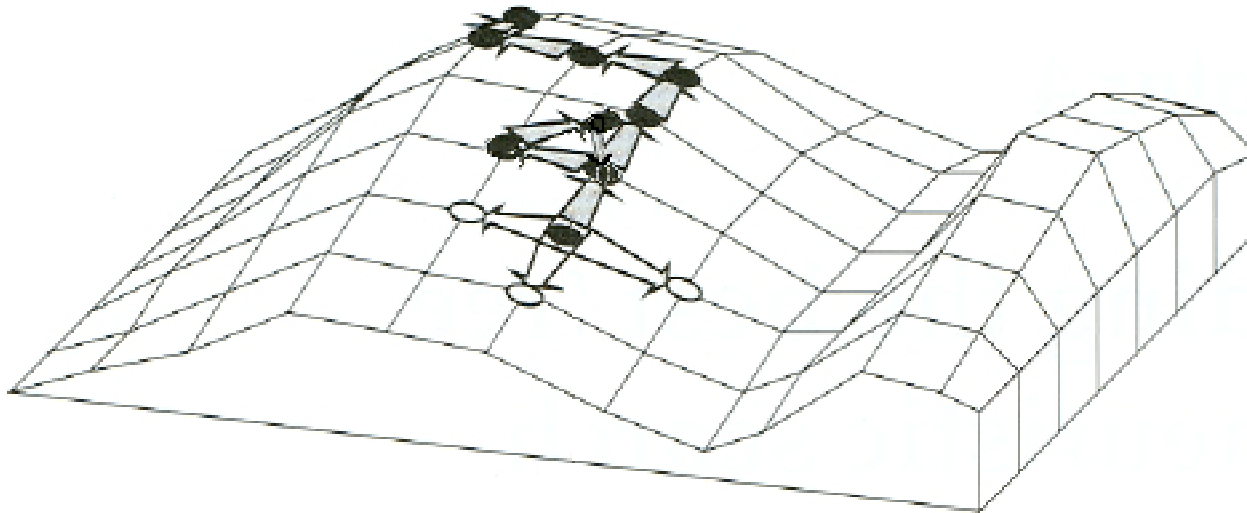


MCMC Algorithm (Li, 1996)

- Prior: $T \sim \text{Uniform on Topologies}$, $t|T \sim \text{Uniform}$, θ and \mathbf{v} estimated using data (empirical Bayes)
- Move in tree space based on random NNI movements for topologies while branch lengths and internal node sequences are chosen to target the posterior.
- Calculations at each step are rapid since they are only local to a single node but the chain moves slowly through tree space.

MCMC: Choosing a proposal mechanism

Idea: Algorithm must balance desire to accept a high percentage of proposed trees with computational efficiency and with need to visit every island of high probability.





Implementation: MrBayes (open source)

- Wide variety of nucleotide, amino acid, and codon models
- Variety of proposal distribution options
- Parallel “hot” and “cold” chains to balance efficiency while covering large tree spaces.
 - ”Cold” chain runs as usual for the desired posterior.
 - ”Hot” chains use the posterior raised to a power < 1 .
 - Periodically an attempt is made to accept the current trees from the hot chains into the cold chain.

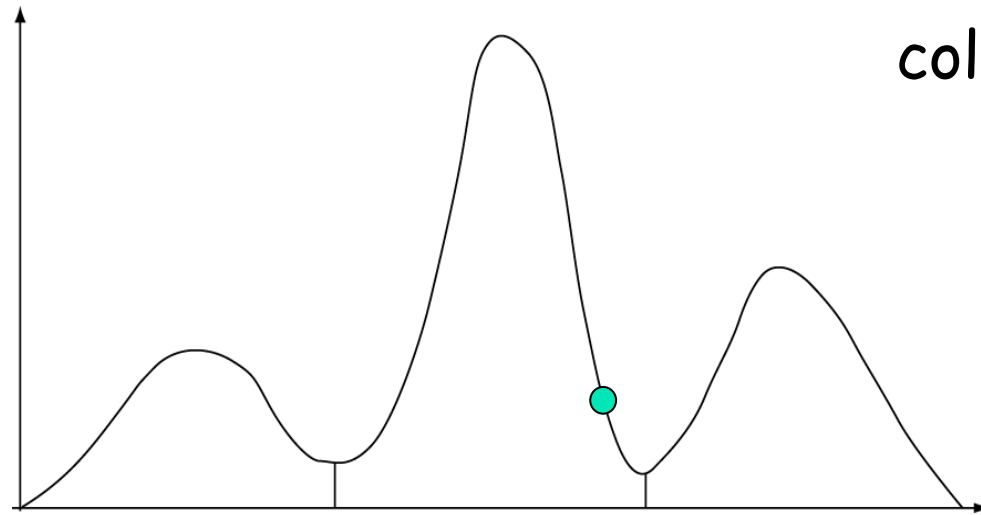
Metropolis-
coupled
Markov chain
Monte Carlo

a. k. a.

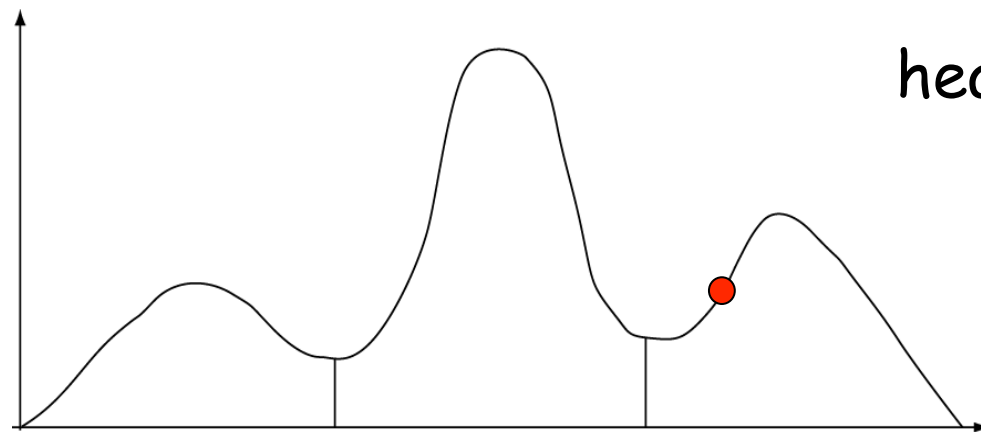
MCMCMC

a. k. a.

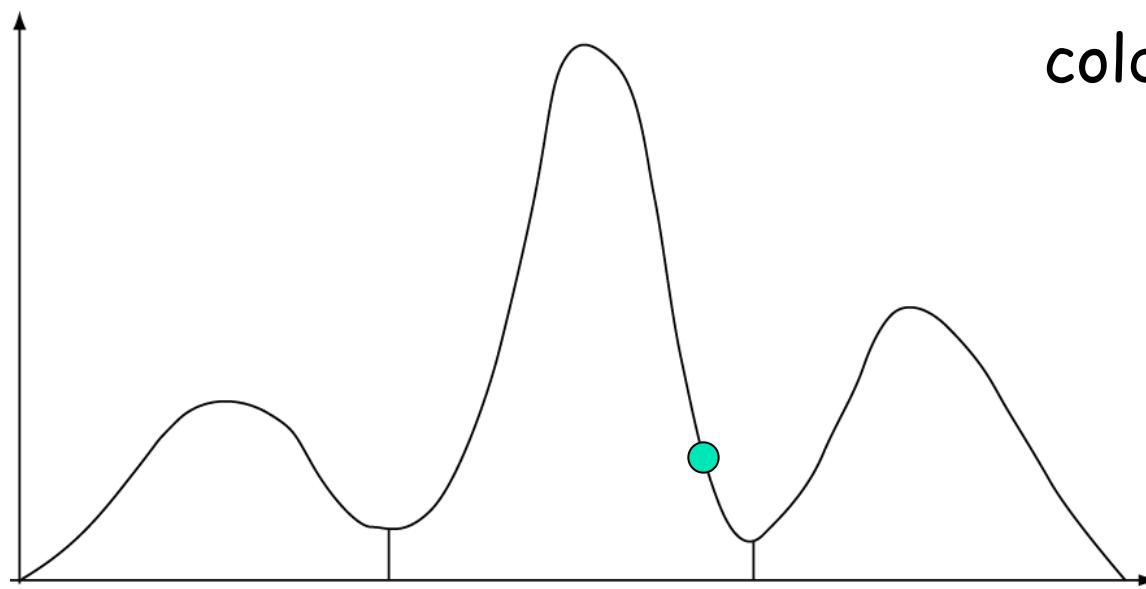
$(MC)^3$



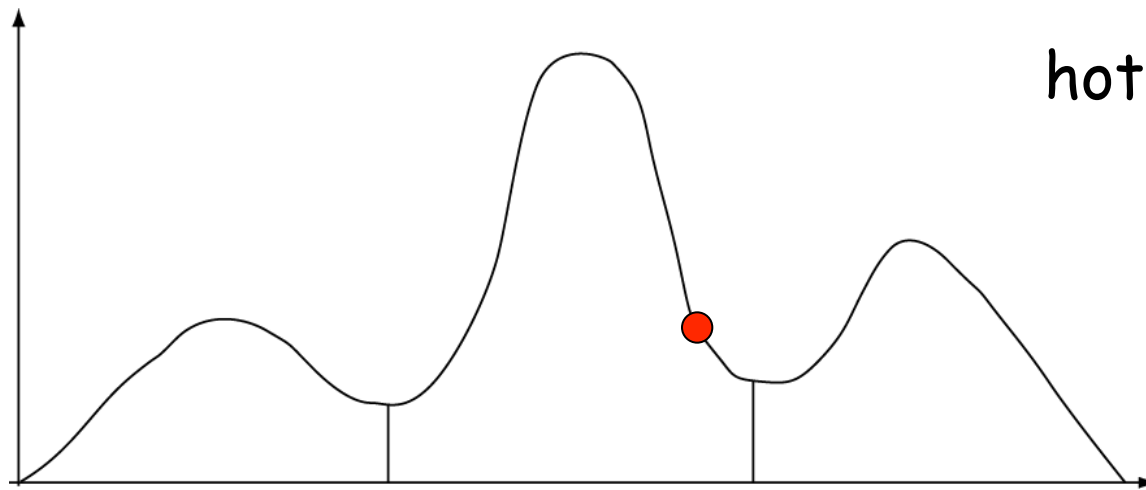
cold chain



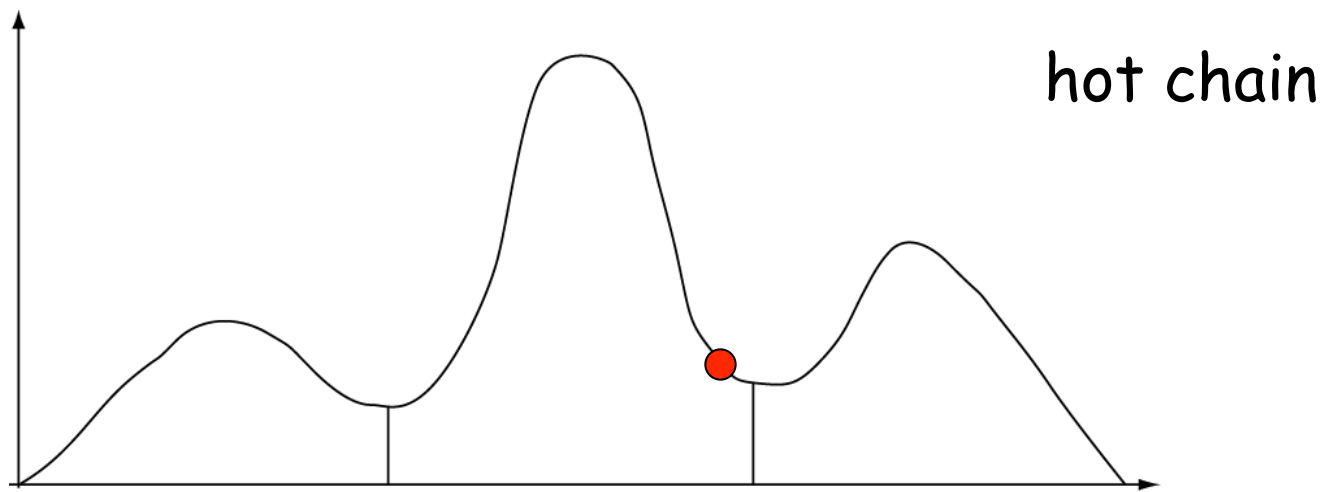
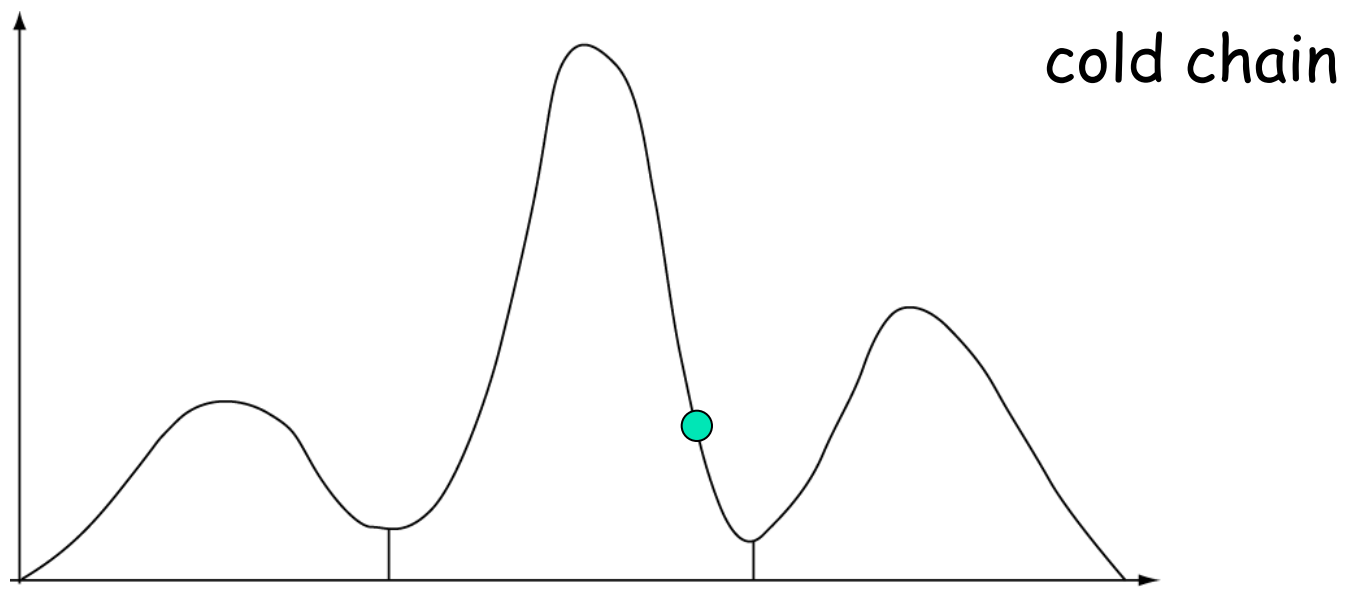
heated chain

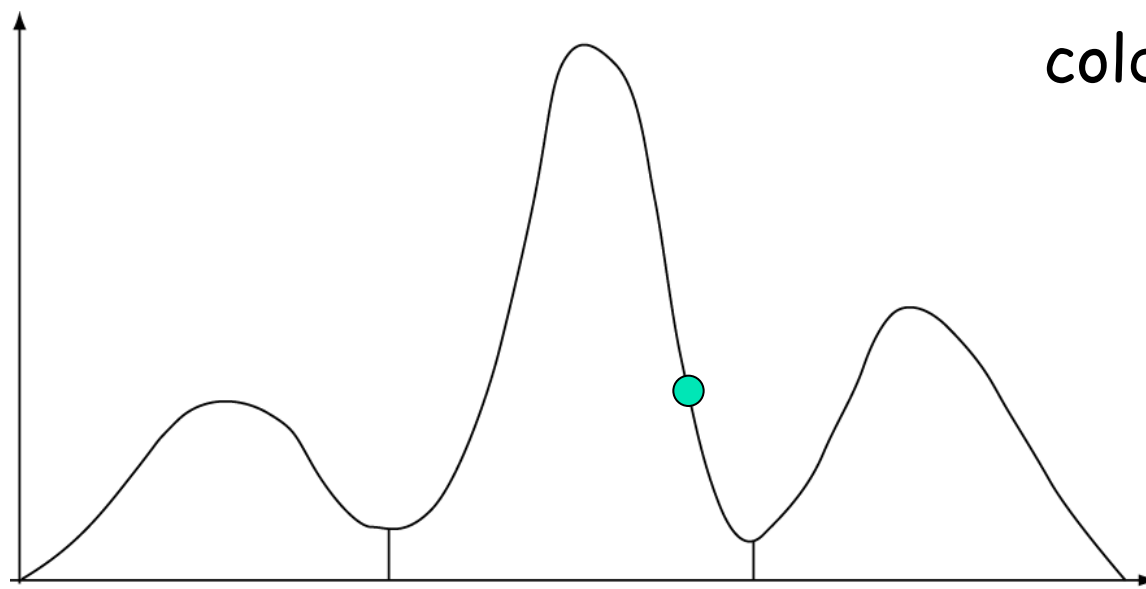


cold chain

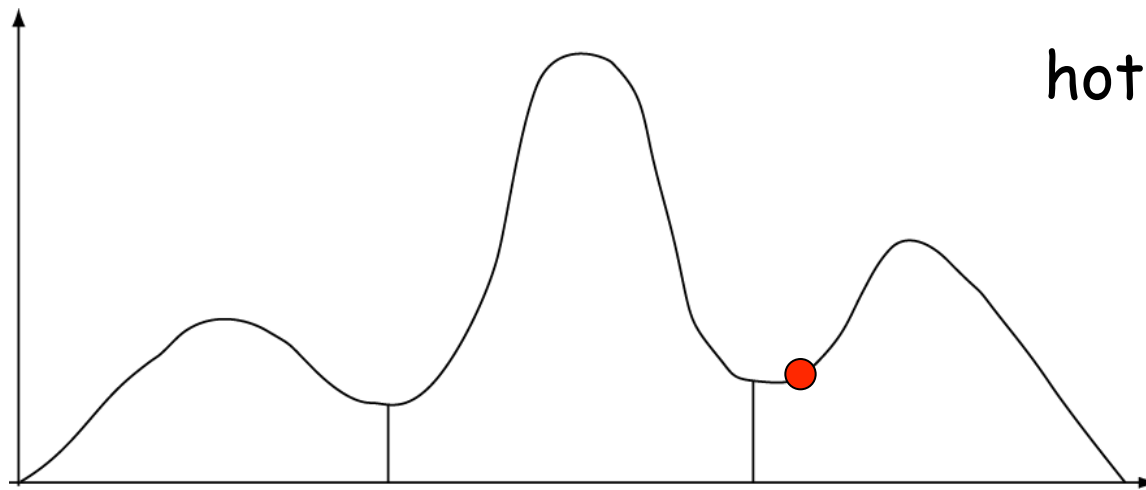


hot chain

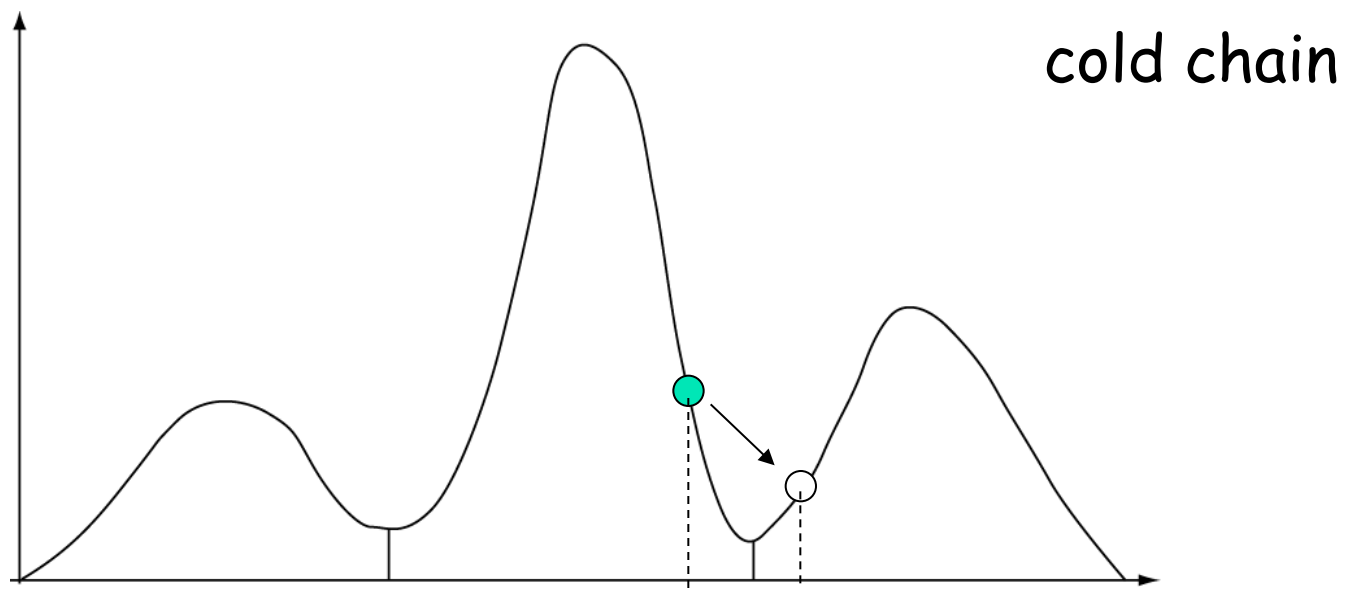




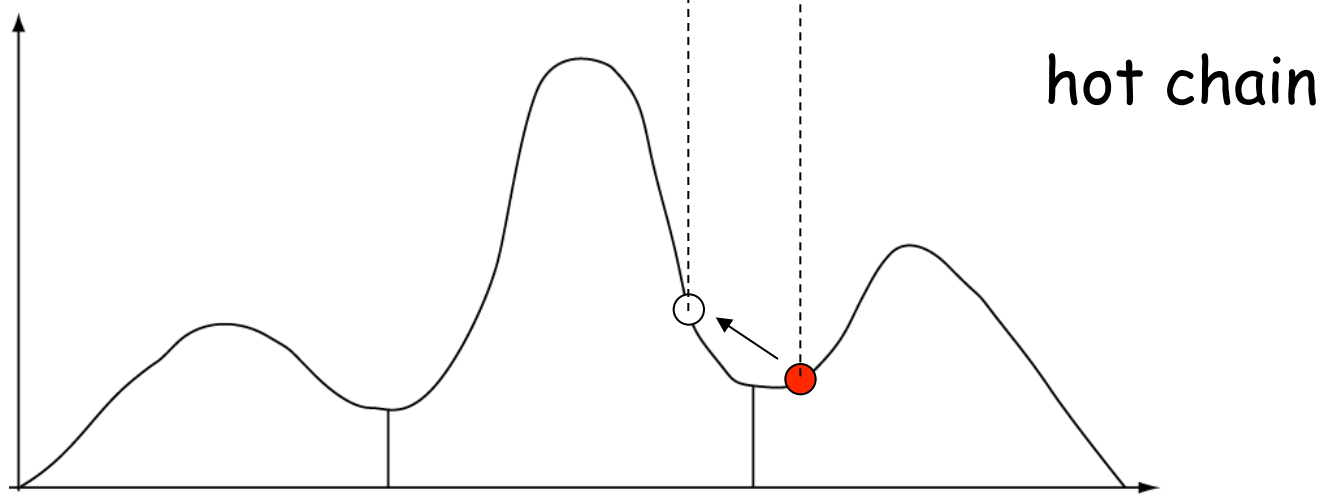
cold chain

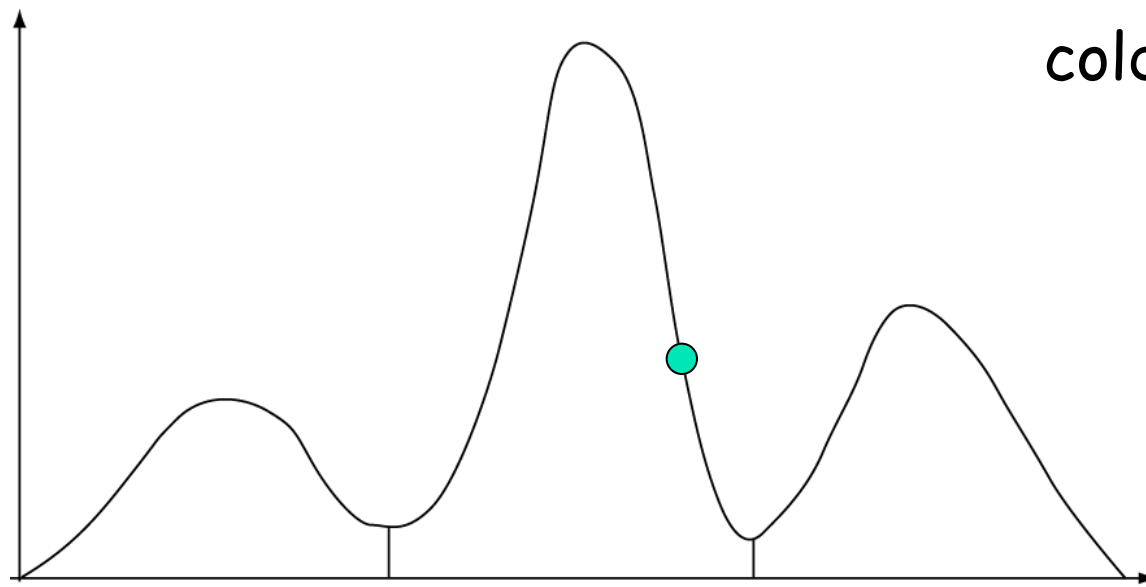


hot chain

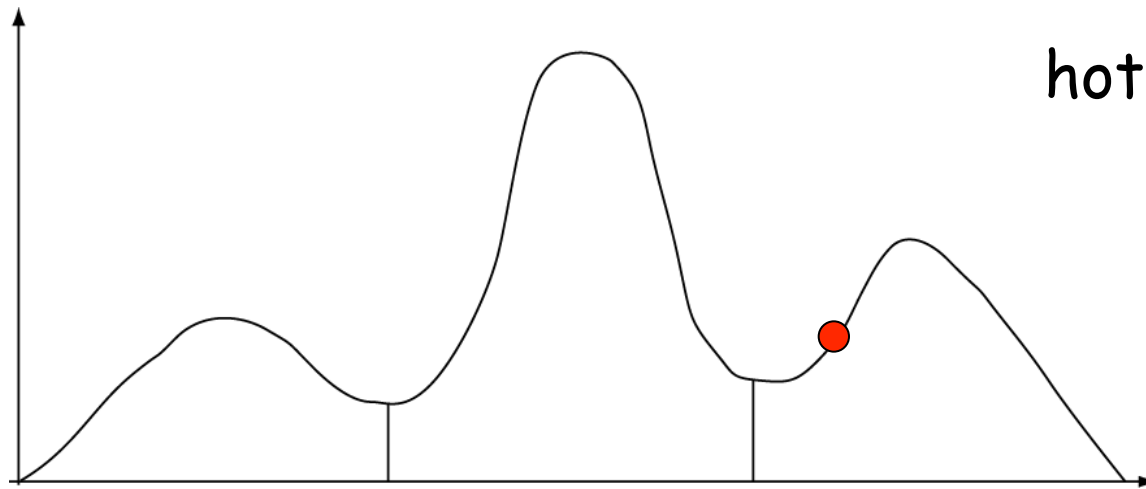


unsuccessful swap

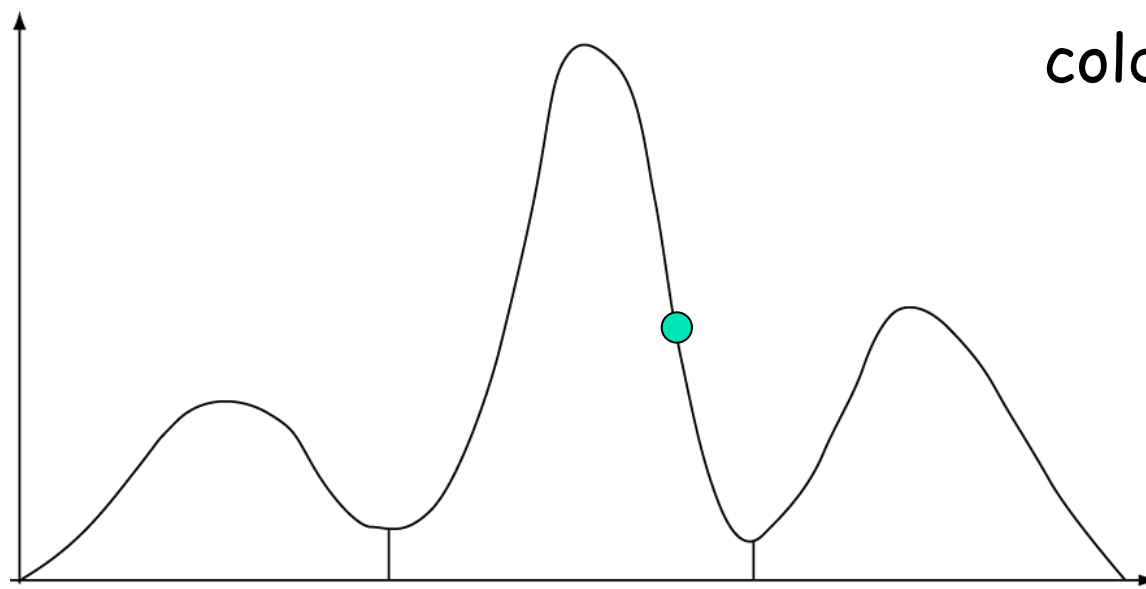




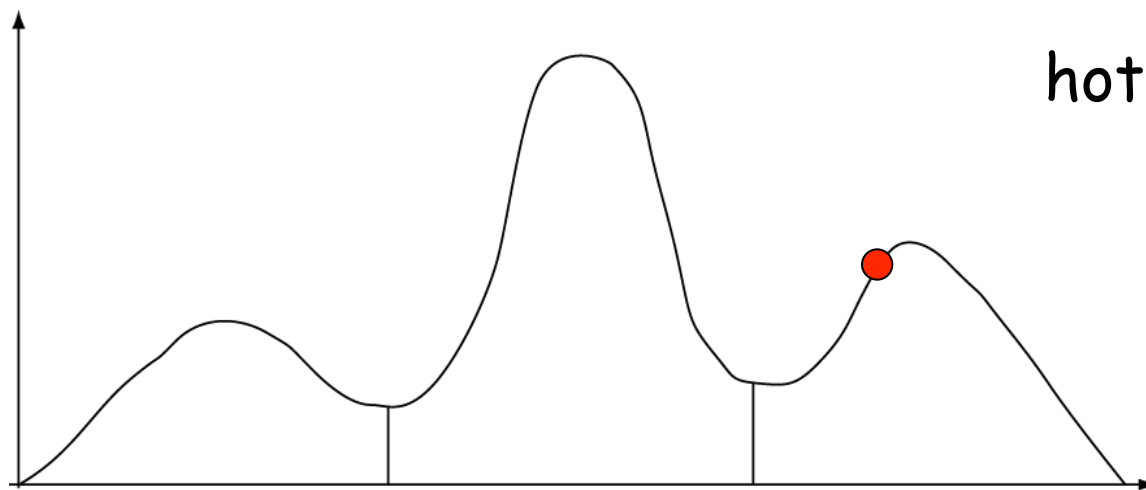
cold chain



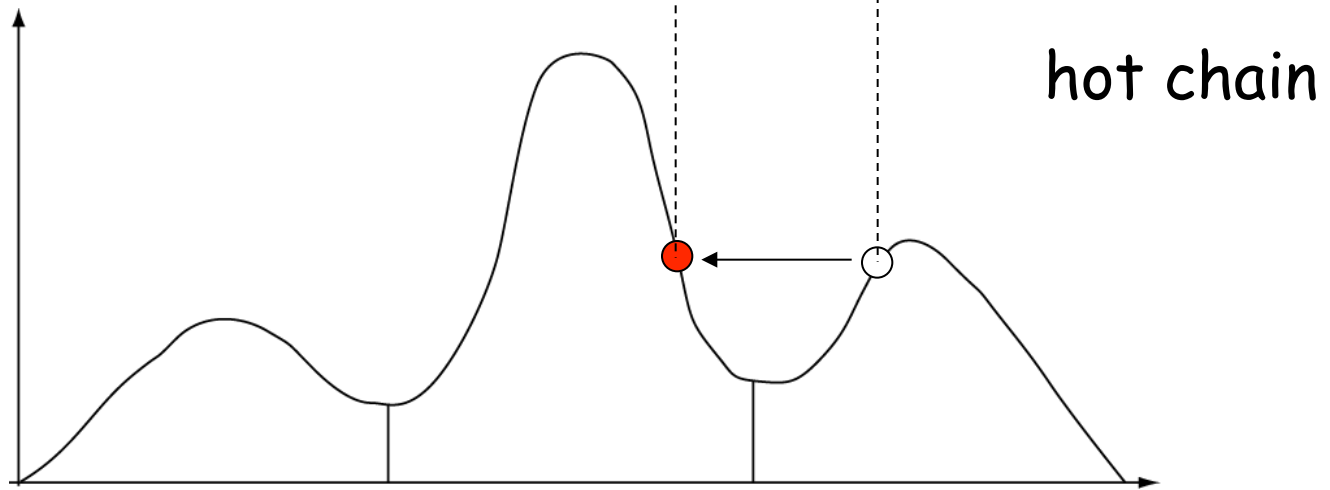
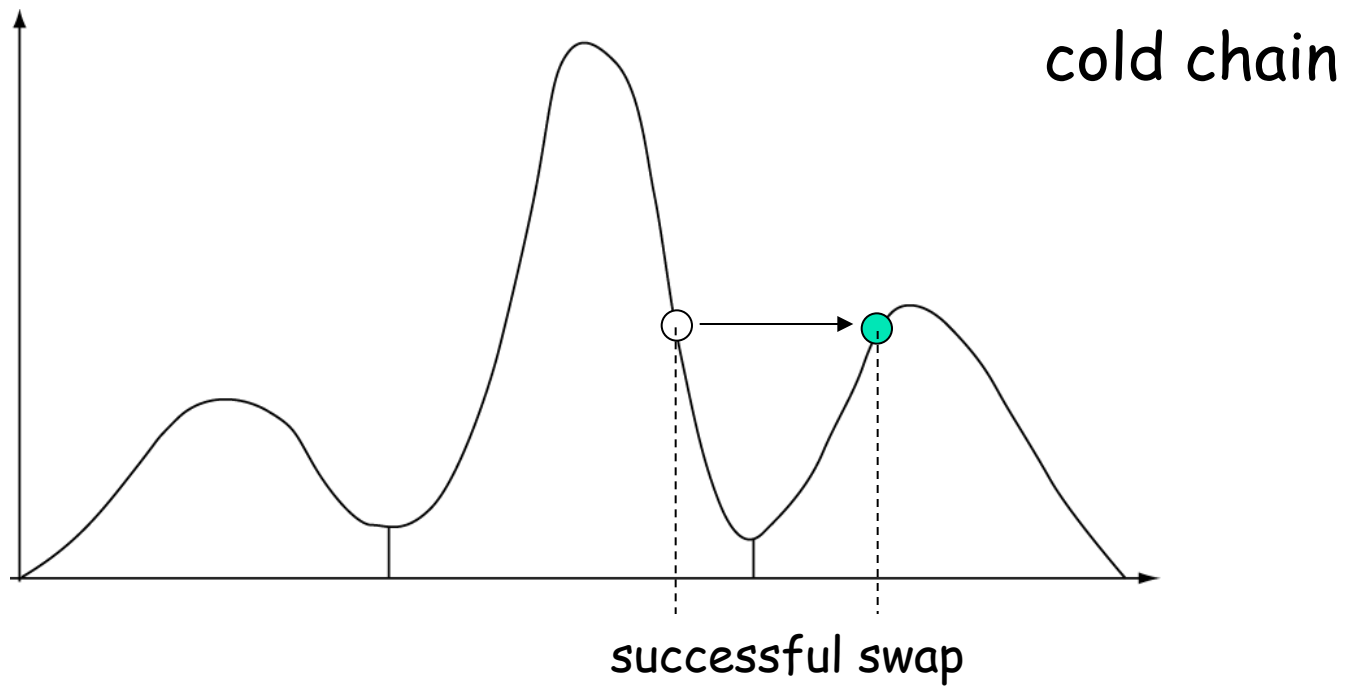
hot chain

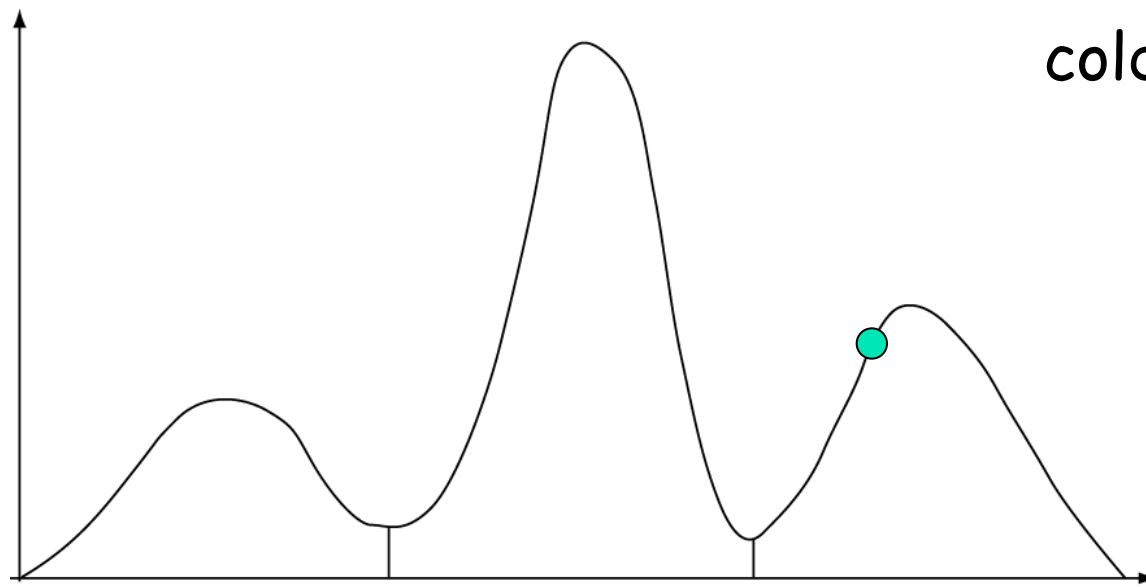


cold chain

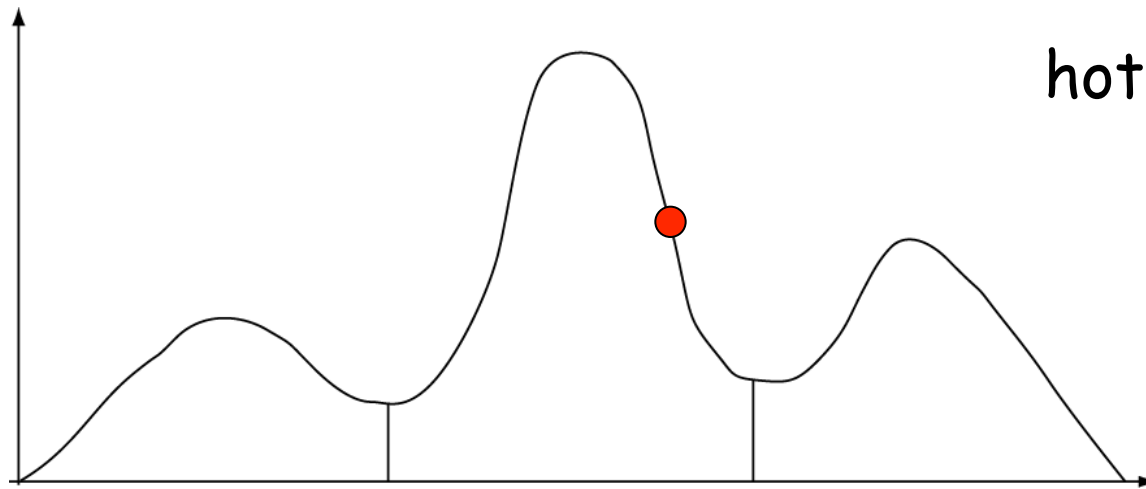


hot chain

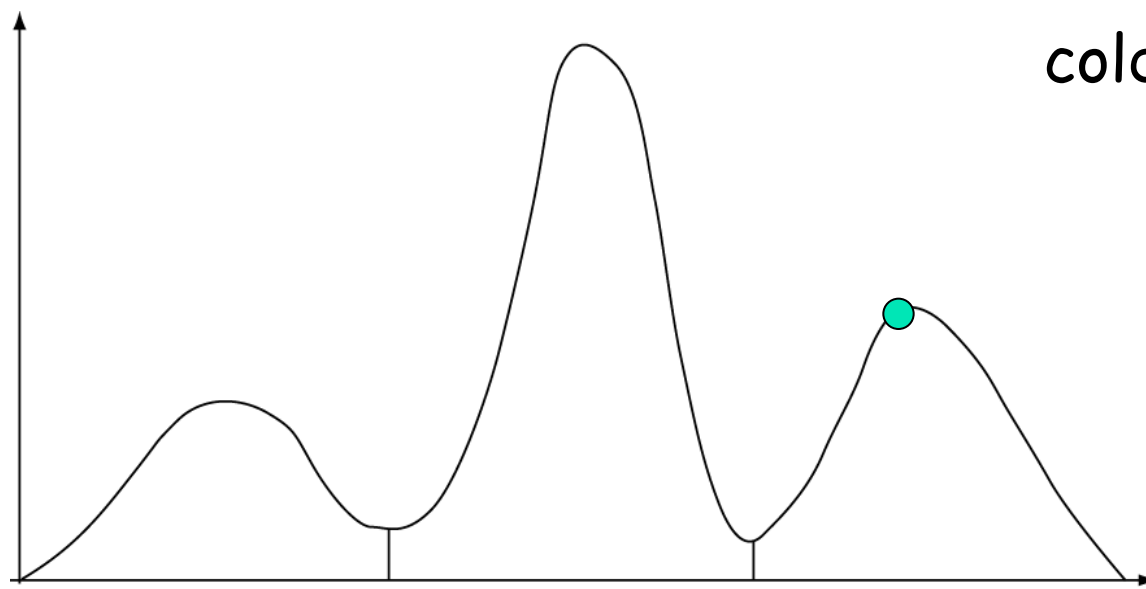




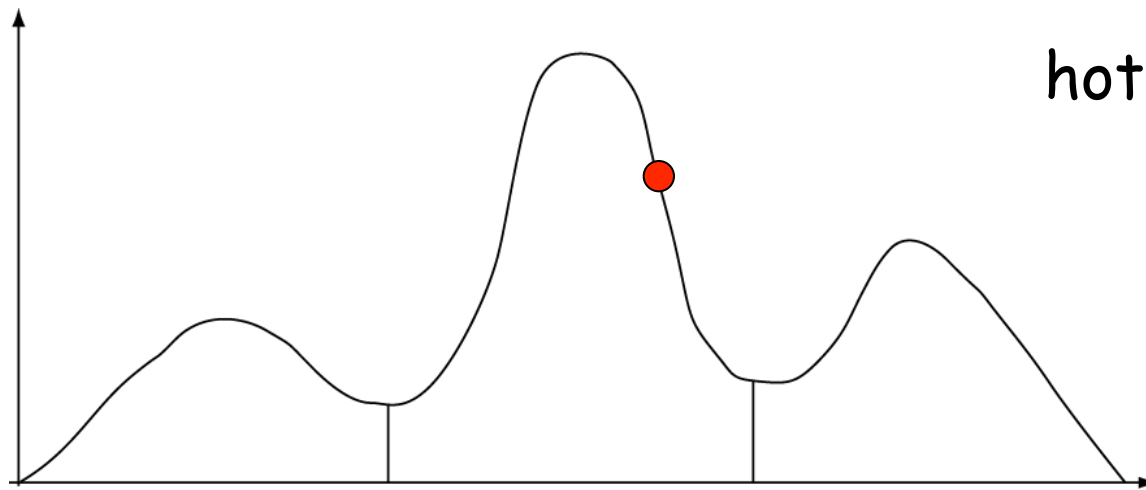
cold chain



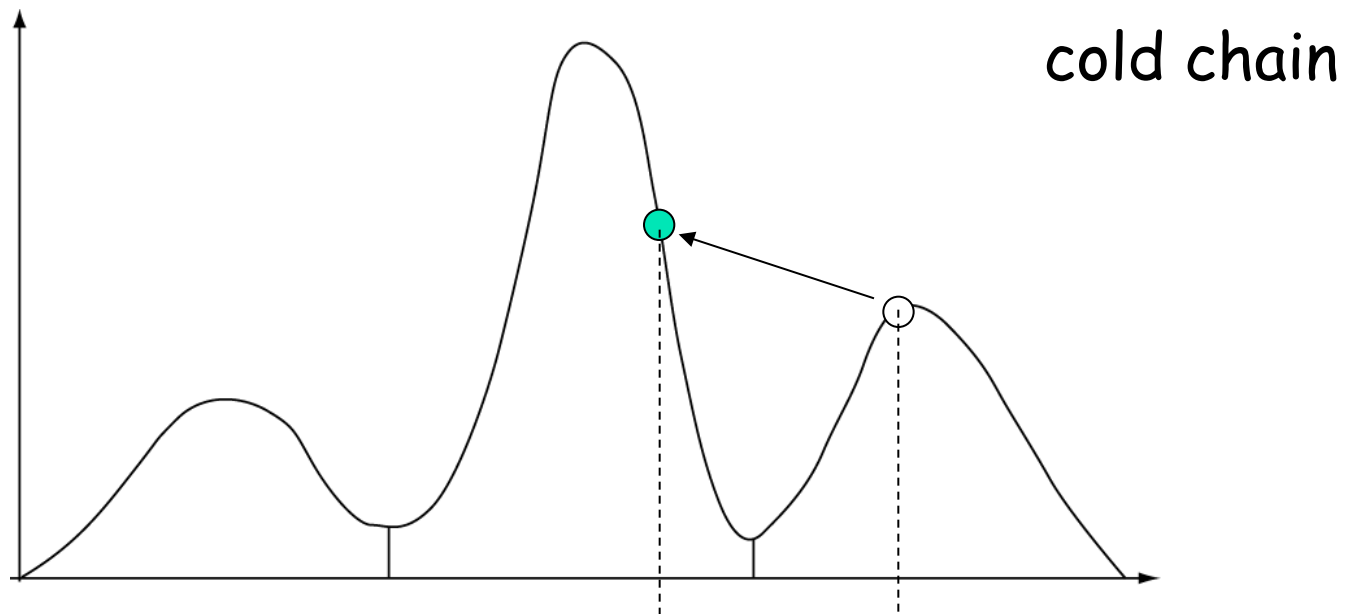
hot chain



cold chain

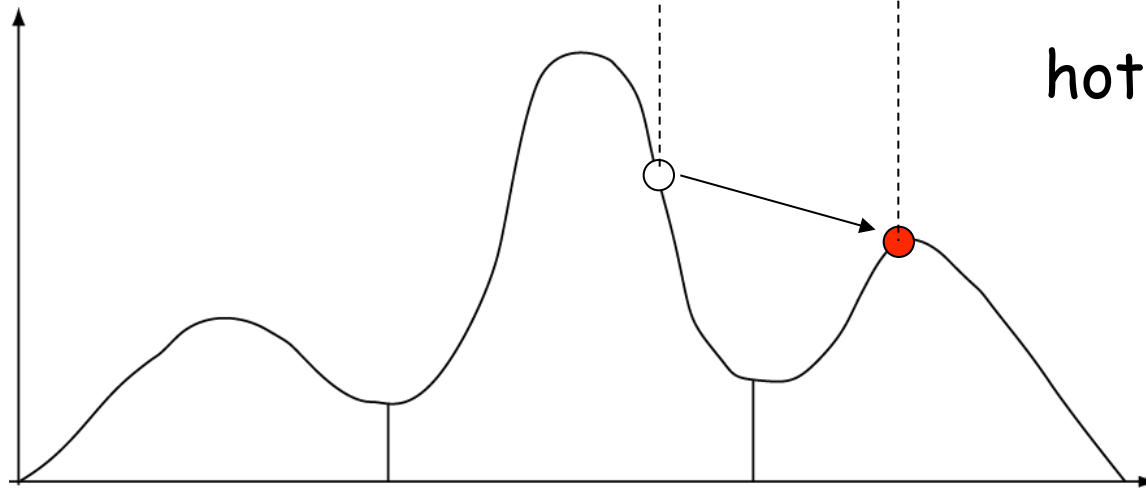


hot chain

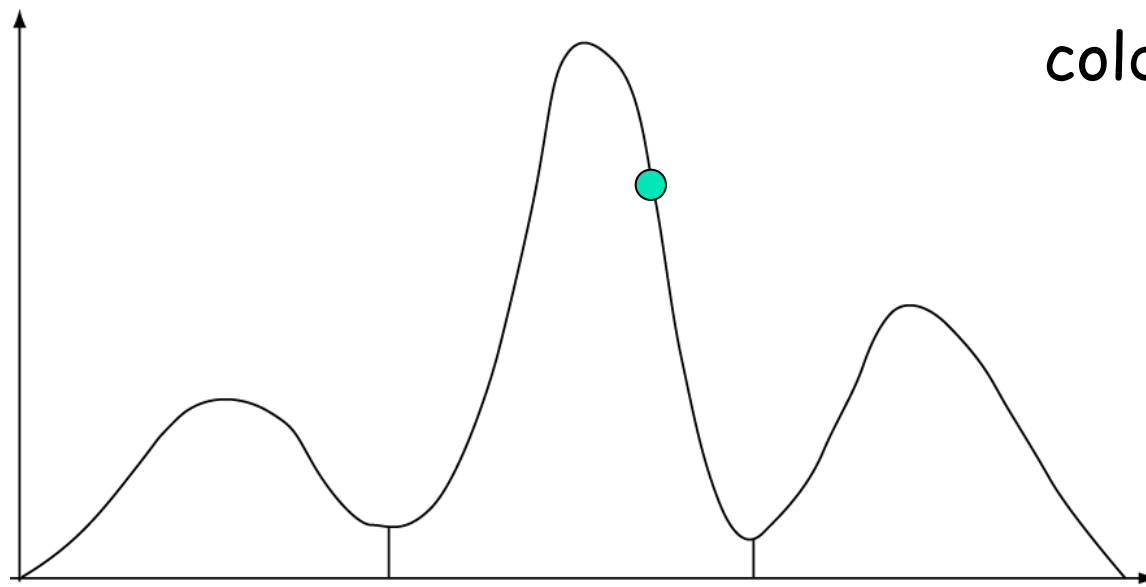


cold chain

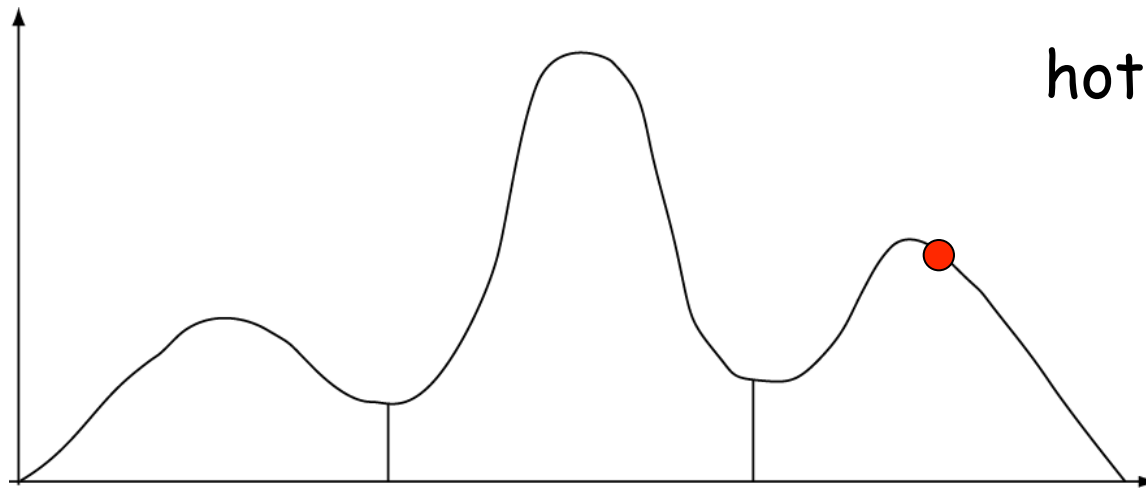
successful swap



hot chain



cold chain



hot chain



Summary of Steps in Bayesian Analyses

- Read the data
- Set the model (data|gene tree)
- Set the Prior
- Set the MCMC rules
- Run the MCMC
- Check convergence
- Summarize results



The Metropolis-Hastings Algorithm

- Start with an initial tree x
- Define a density $q(x,y)$ that specifies probabilities of moves from x to a proposed tree y

- Accept y with probability $\min \left\{ \frac{\textit{posterior}(y)q(y,x)}{\textit{posterior}(x)q(x,y)}, 1 \right\}$

- **Theory:** Repeating these steps **many times** will create a Markov Chain whose stationary distribution is the desired posterior.



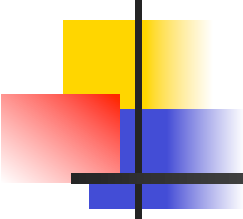
The Metropolis-Hastings Algorithm

- Start with an initial tree x
- Define a density $q(x,y)$ that specifies probabilities of moves from x to a proposed tree y

- Accept y with probability $\min \left\{ \frac{\text{posterior}(y)q(y,x)}{\text{posterior}(x)q(x,y)}, 1 \right\}$

- **Theory:** Repeating these steps **many times** will create a Markov Chain whose stationary distribution is the desired posterior.

How many?



Diagnostics: Ways to examine whether an MCMC phylogenetic analysis has properly described the desired posterior distribution.

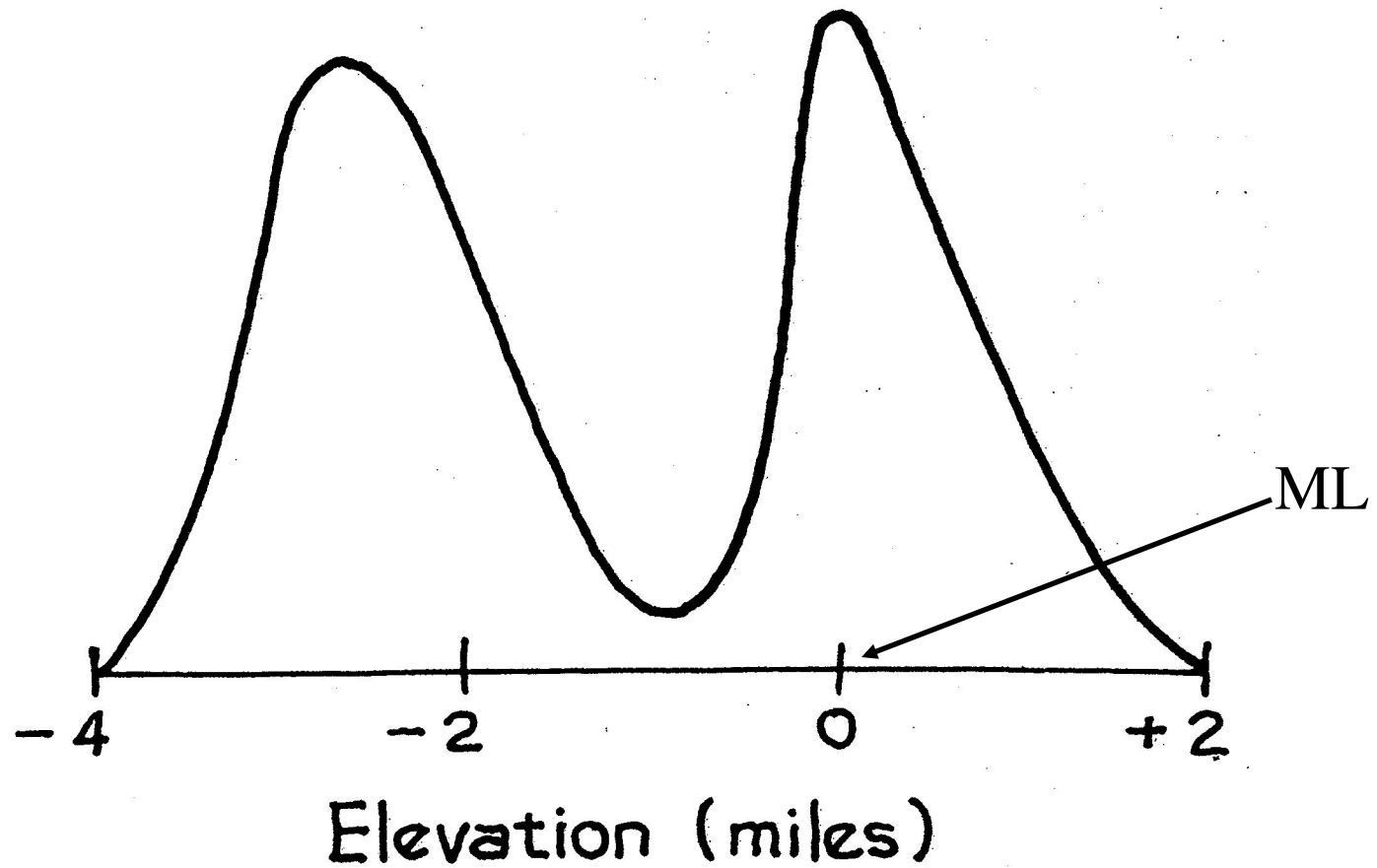
- The issues addressed
- Some Methods



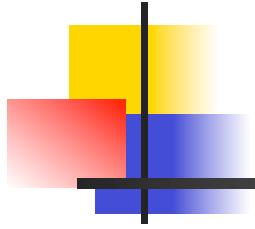
The Issues

Idea: We need to run the burn-in long enough to reach the stationary distribution and we need to run the chain long enough to get a precise measure of our endpoint.

- Is the dependence structure so strong that the effective sample size is too small?
- Is the burn-in long enough to eliminate the effect of the initial starting point?
- Does the chain make regular visits to all parts of the parameter space?



Distribution of elevation of a randomly selected spot on Earth
(from Freedman, Pisani, & Purves, *Statistics*, 3rd ed. 1998)

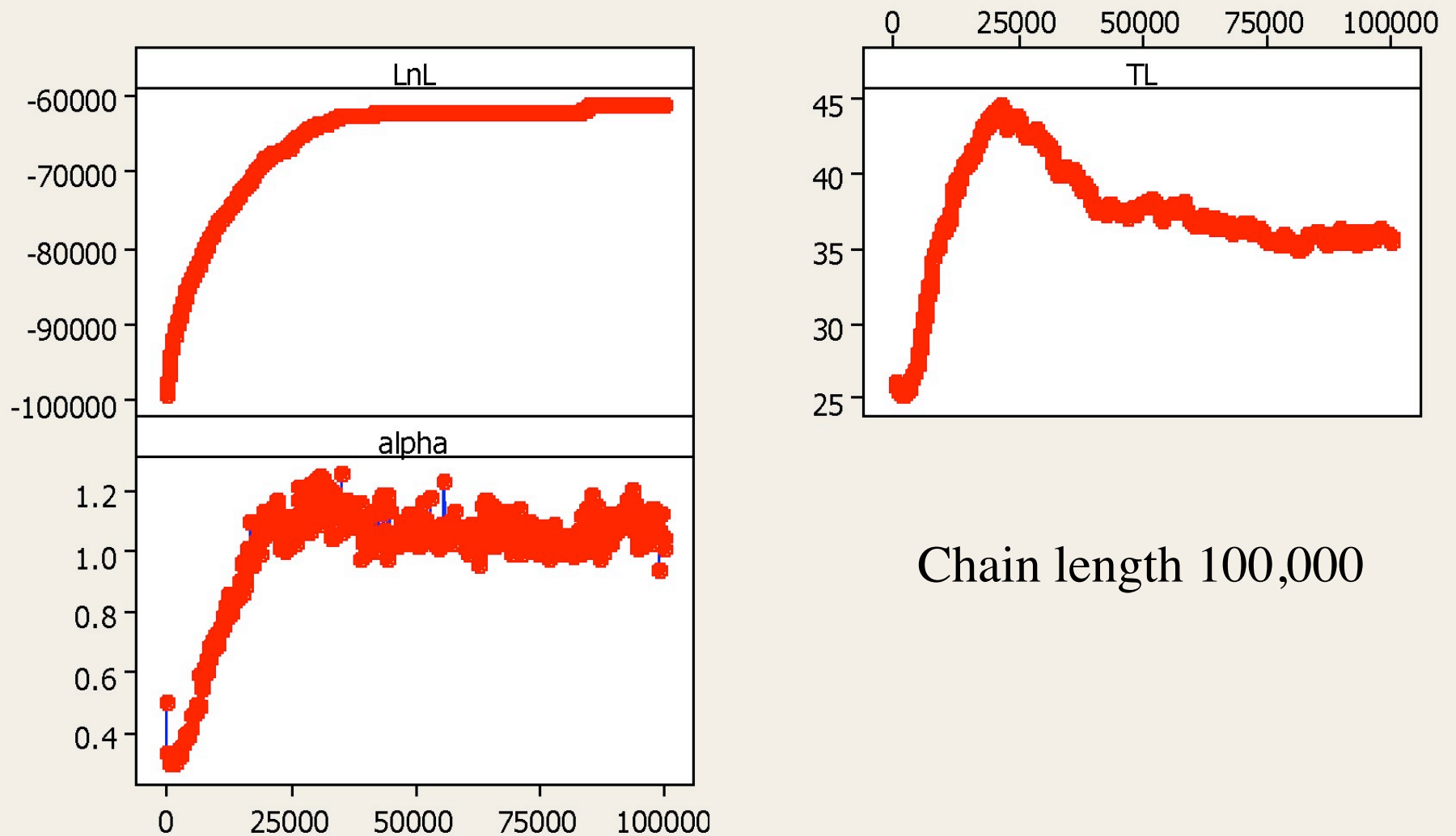


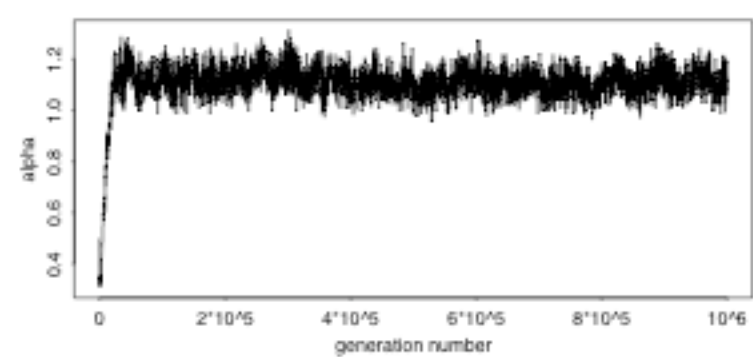
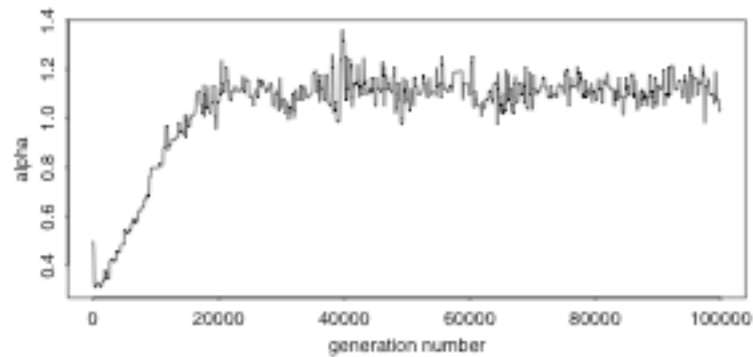
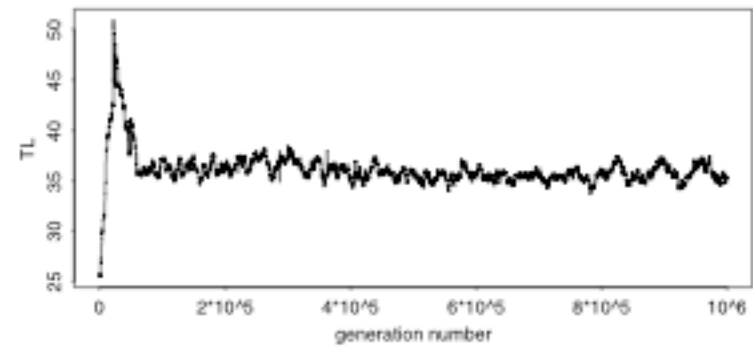
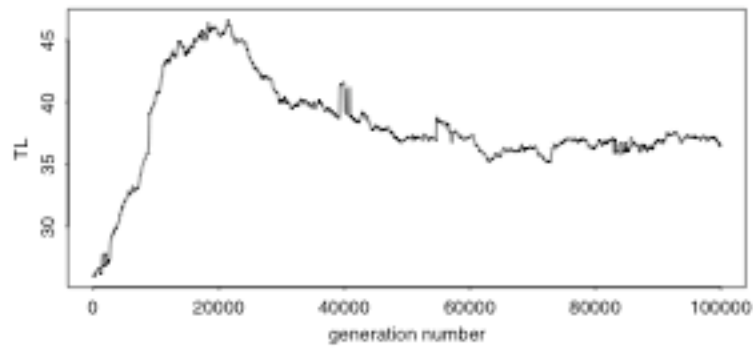
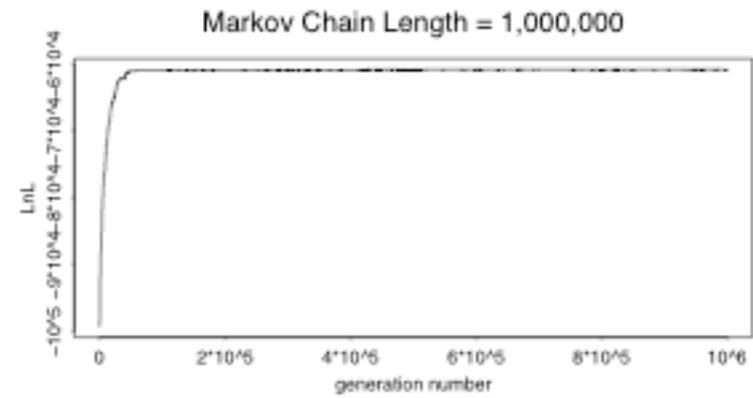
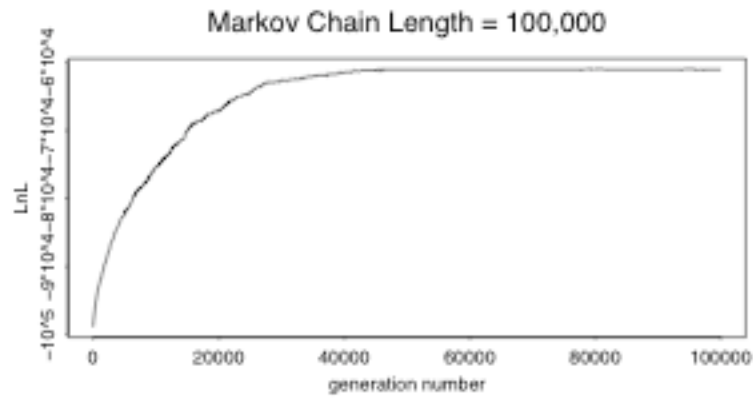
Applicable standard techniques

- Autocorrelation plots
- Multiple runs with different starting points
- Log-posterior & other time series plots

Idea: Time series diagnostics can be applied to any characteristic of the model or tree since all must reach their equilibrium distributions.

Scatterplot of LnL (log-likelihood), TL (tree length), alpha (gamma shape parameter) vs Generation





Time series diagnostics comparison of 10^5 versus 10^6 chain lengths

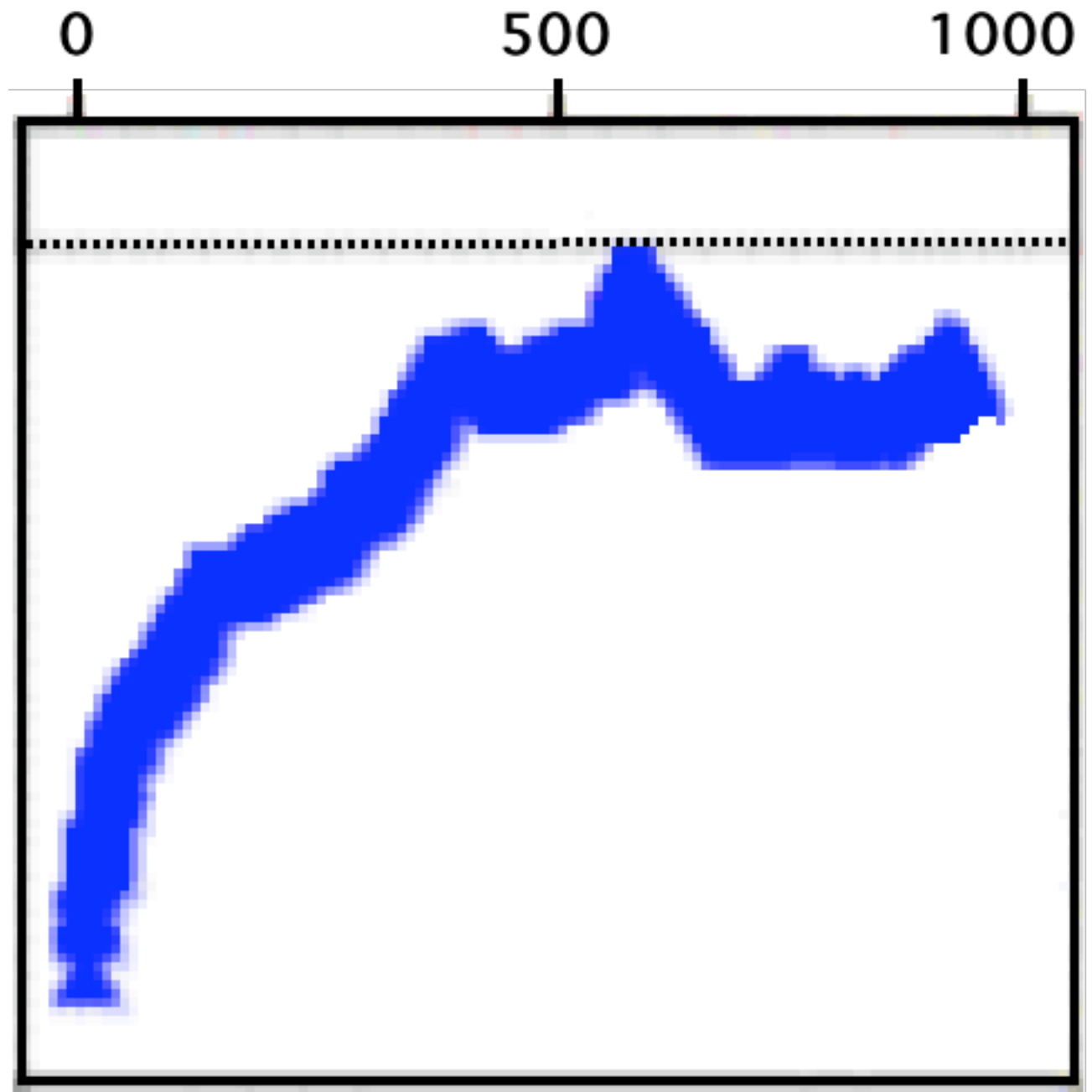


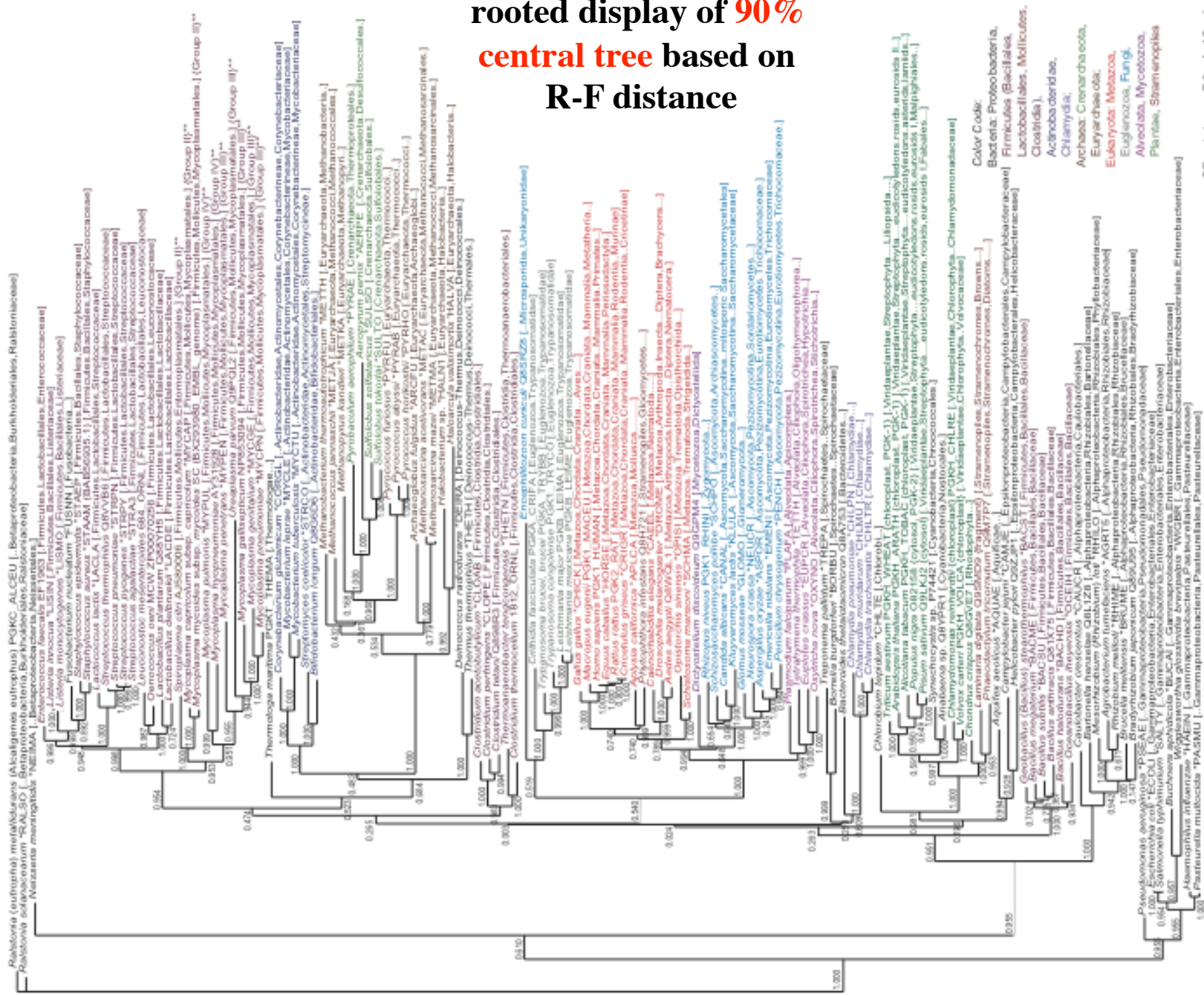
Need for multiple diagnostics

Trees with similar posterior probabilities or similar lengths are not necessarily close in tree space.

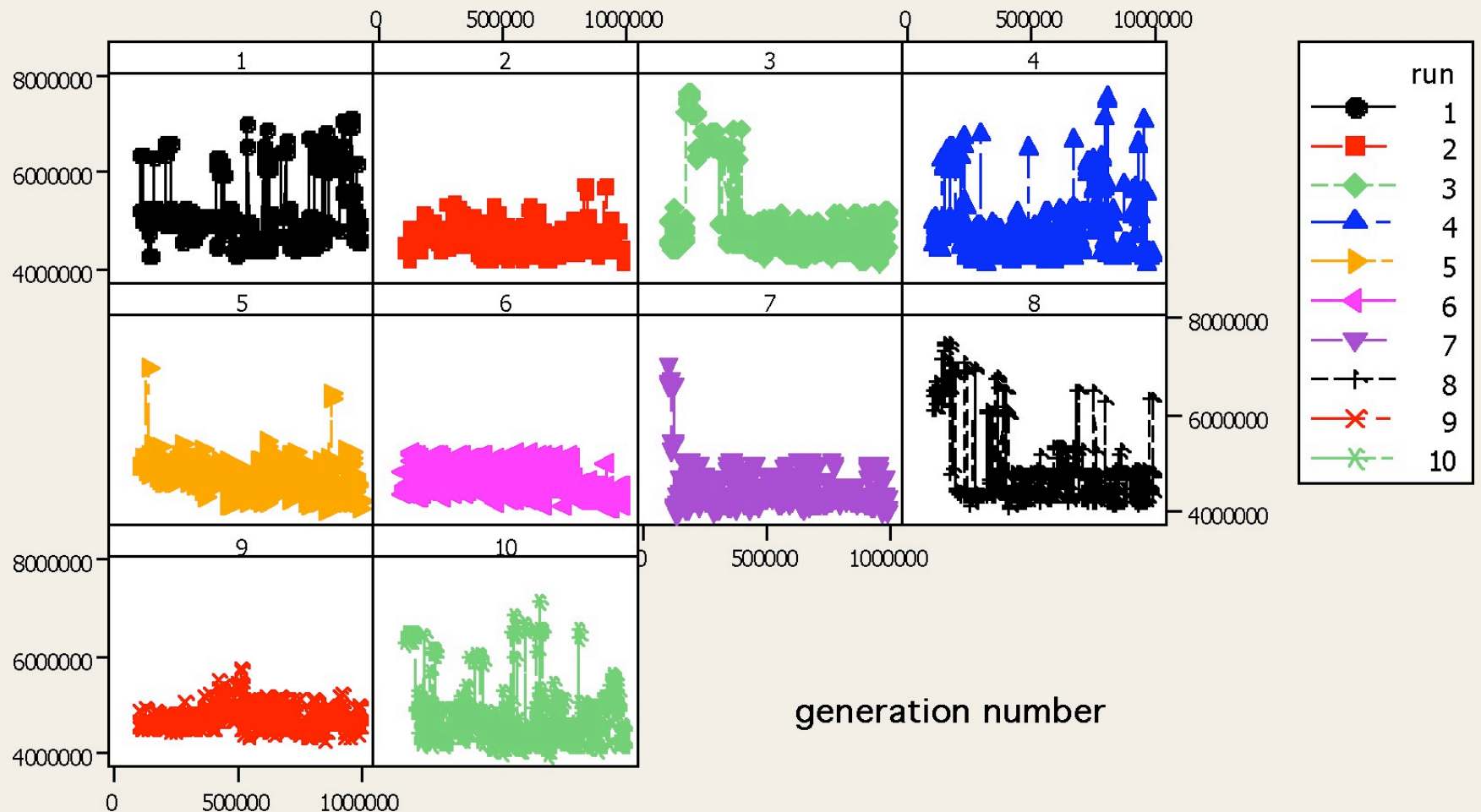
How far apart in the chain do trees have to be for them to be essentially independent with respect to distance in tree space?

**Lag Quartet
Distance plot
for chain length
= 1 million**





Scatterplot of 90% radius vs generation



Even chains of length 1 million did not always visit both islands of high posterior probability.



Scaled Regeneration Quantile (SRQ) plots

Mykland, Tierney, & Yu (1995, *JASA*)

T_i = the time in the chain when it returns to a specific topology,
 $i = 1, \dots, N$

A Scaled Regeneration Quantile plot is a plot of T_i/T_N versus i/N .

Interpretation: Large deviations from the 45° line indicates a problem with mixing.

The slope of the line connecting the SRQ plot at i/N and the SRQ plot at j/N is the ratio of the estimated probability of the specific topology based on the entire chain to the estimate based on the chain between tours i and j .

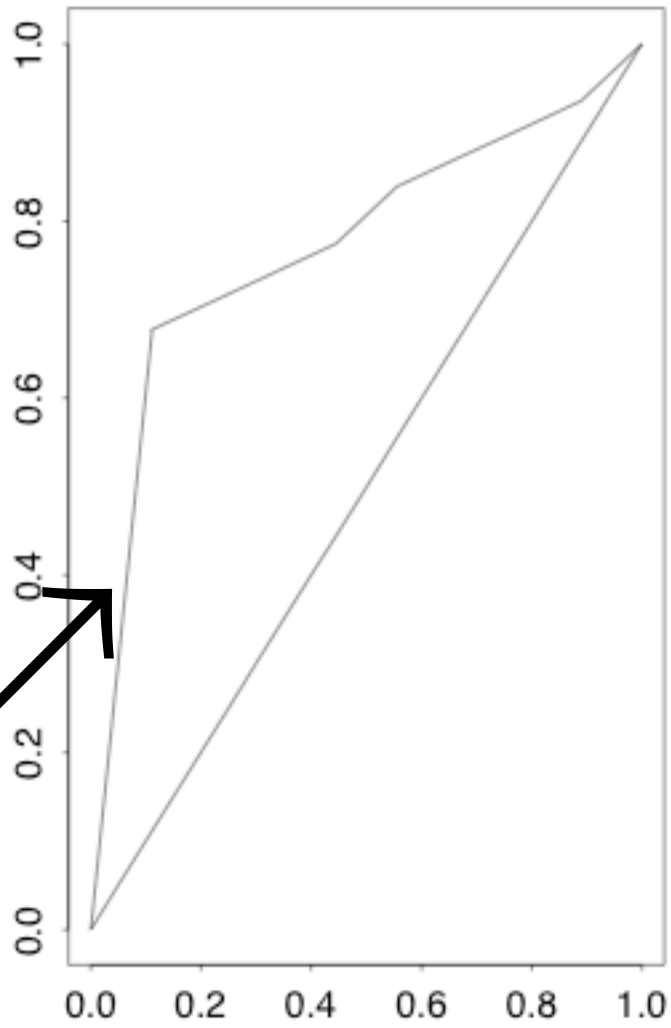


SQR Plots: adaptation to large tree spaces

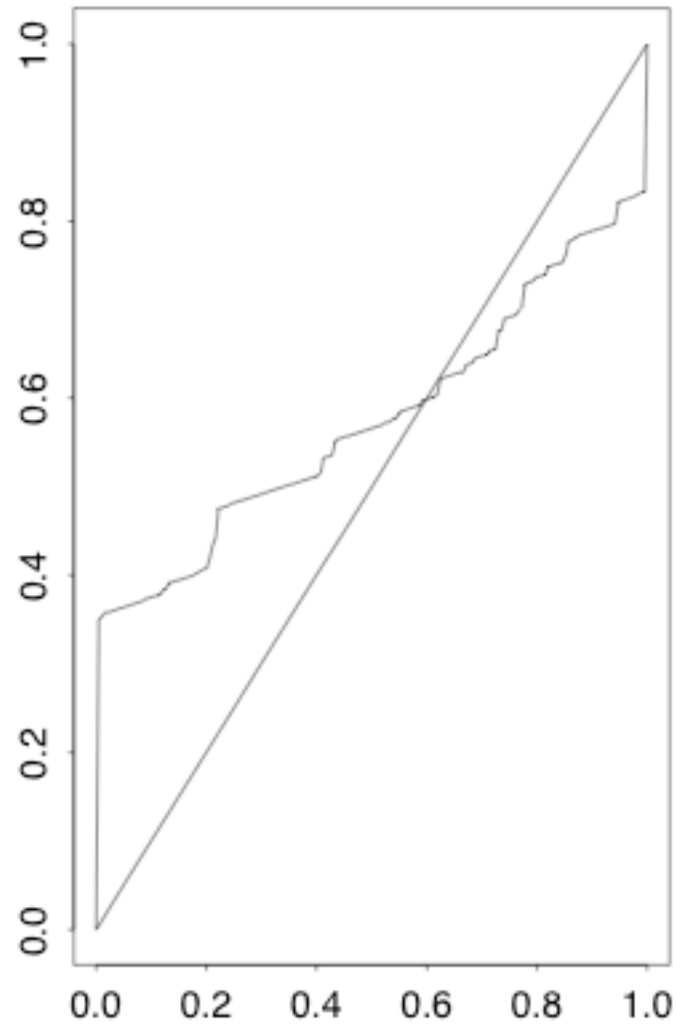
Application when any individual phylogeny is rare:

Form a ball around a specific tree that captures 5% of the distribution and mark returns to this ball.

Chain length: 100,000



Chain length: 1,000,000



Steep slope indicates topology was rarely visited during this period



Distance Density Plots

Theory (Maa, Pearl, and Bartoszynski, 1996) :

For independent replicate X_1, X_2, \dots from distribution F and Y_1, Y_2, \dots from G , then $F = G$ **if and only if** the one-dimensional distances $d(X_1, X_2)$, $d(Y_1, Y_2)$, and $d(X_3, Y_3)$ all have the same distributions.

Note: “d” can be any arbitrary distance, and F & G can be continuous, discrete, vector-valued, or even tree-valued.



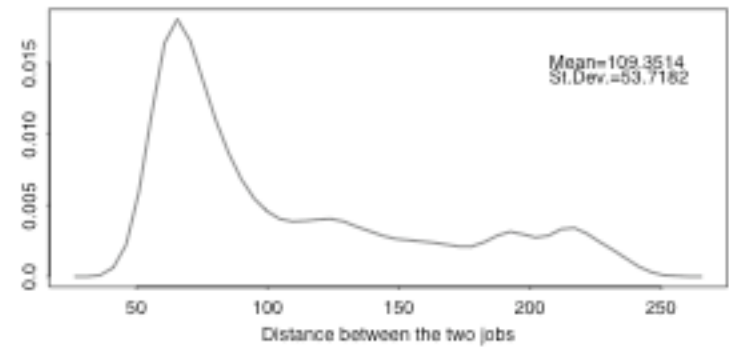
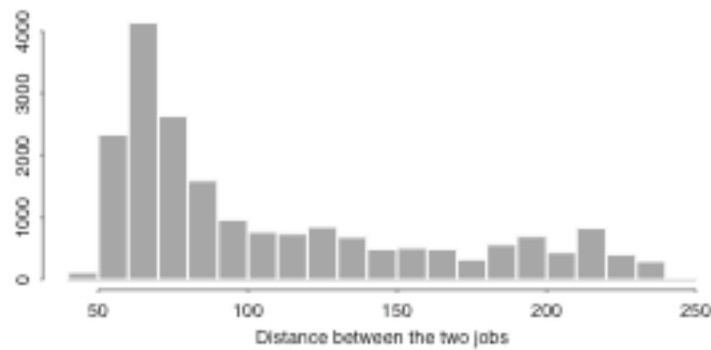
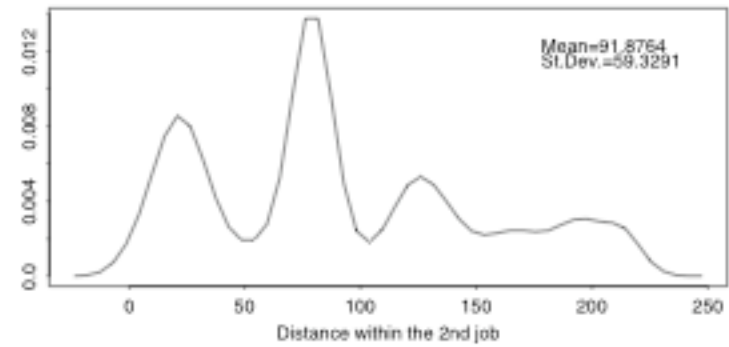
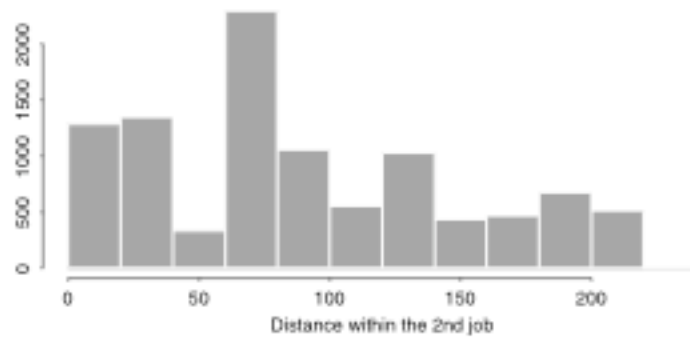
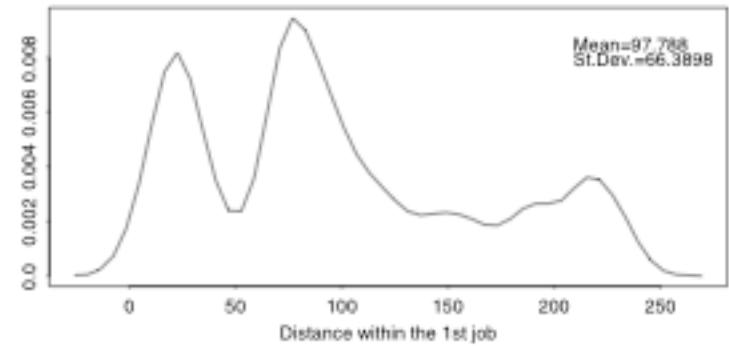
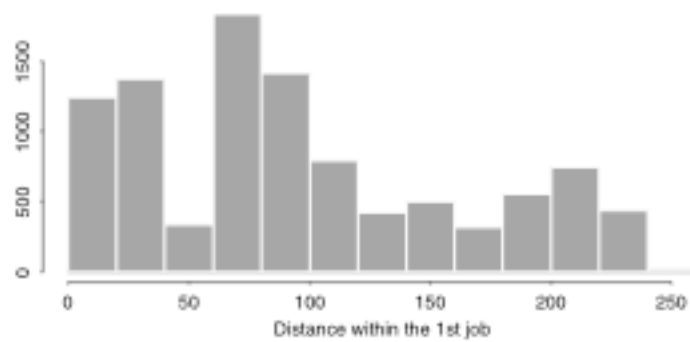
Distance Density Plots

Application to MCMC phylogenetics (Li, Pearl, and Doss, 2000):

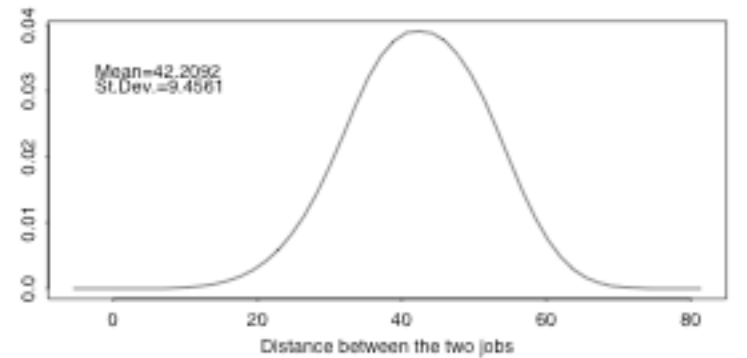
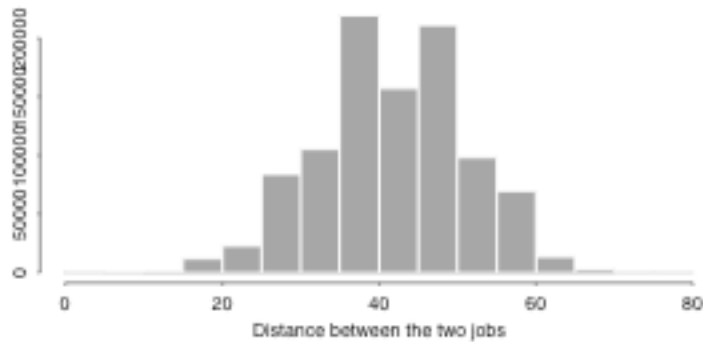
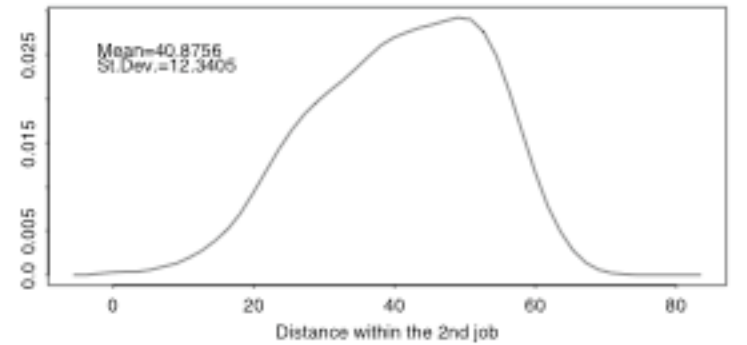
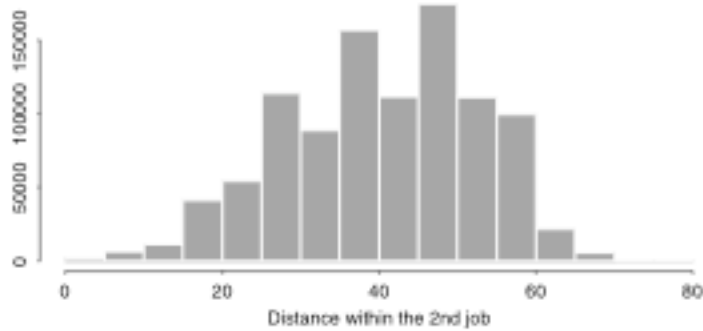
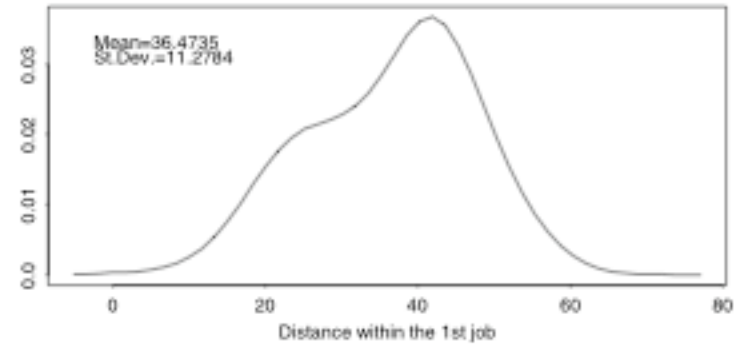
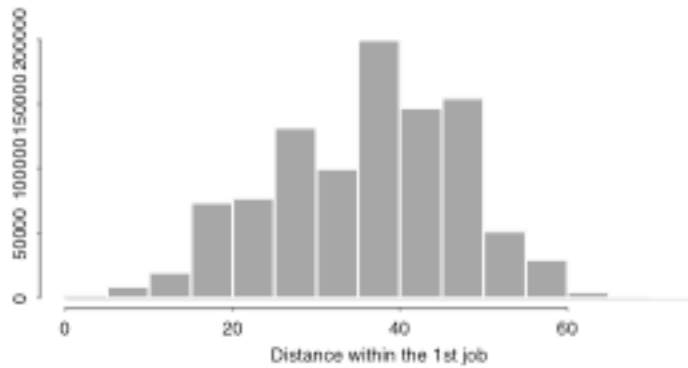
Run the chain twice independently, then compare the distributions of distances for random pairs of trees drawn

- i) both from within the first chain,
- ii) both from within the second chain, and
- iii) one from each chain.

If these three distributions don't look the same then the two chains did not converge to the same distribution.



Distance histograms chain lengths 100,000 (quartet distance)

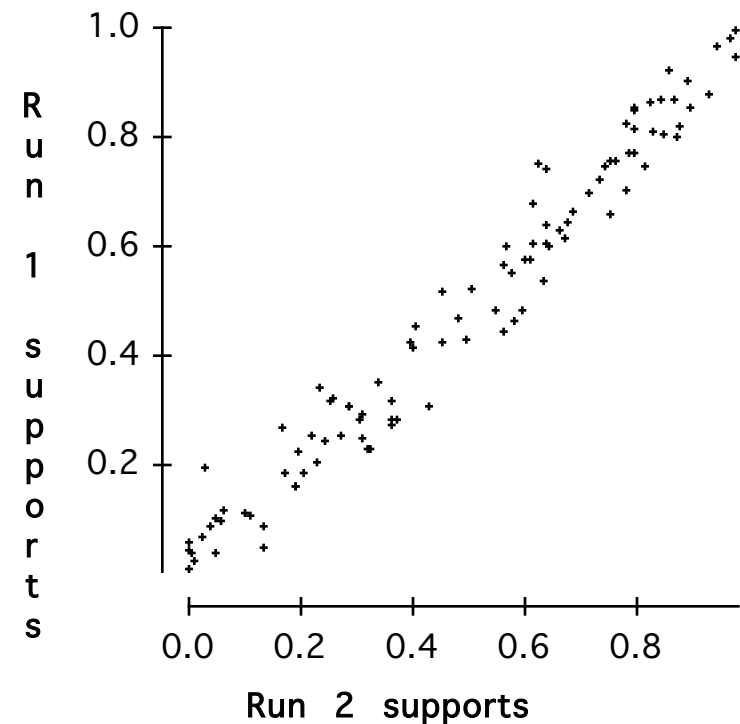


Distance histograms chain lengths 1,000,000 (quartet distance)

Split Support Correlation Plots

Idea: Partition support values should be nearly the same in two runs of an MCMC phylogenetic analysis.

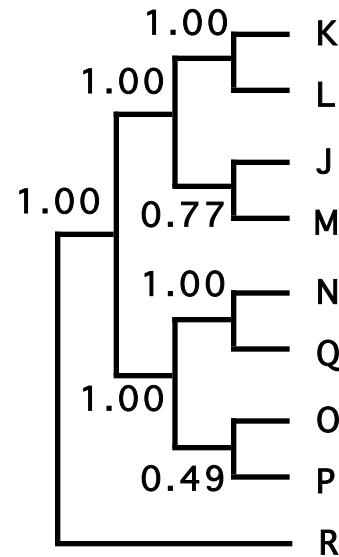
MrBayes look at the average standard deviation of split probabilities across samples.



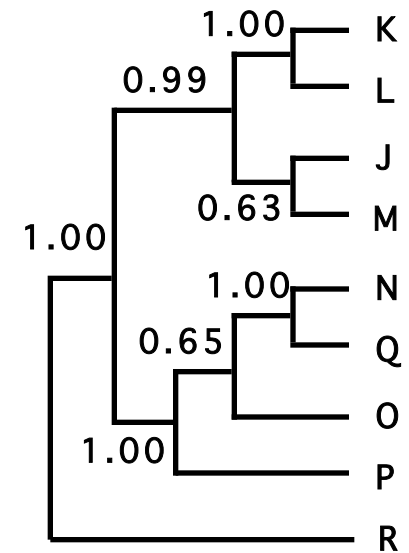


For ignoring the MCMC
diagnostics, he was
sentenced to the
Markov chain gang.

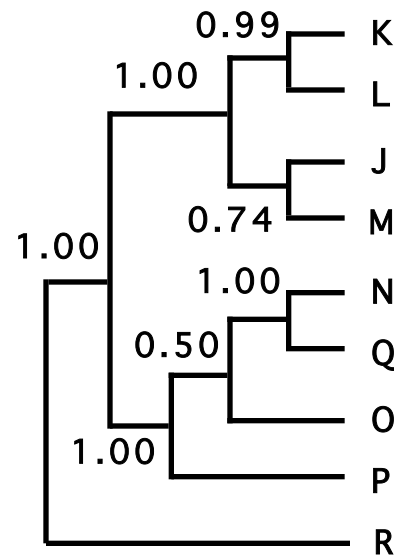
Hillis et al, 1992 produced a laboratory generated set of nine Bacteriophage T7 DNA sequences with a known phylogeny (1091 sites, 63 informative for parsimony). The true tree had an estimated 47% posterior probability and the top four, out of 135135 topologies, made up 96% of the distribution.



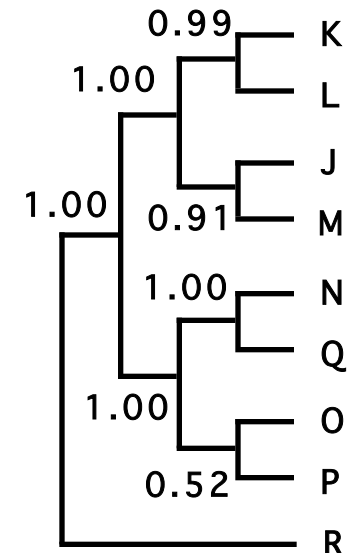
Parsimony



Neighbor-Joining



Maximum Likelihood

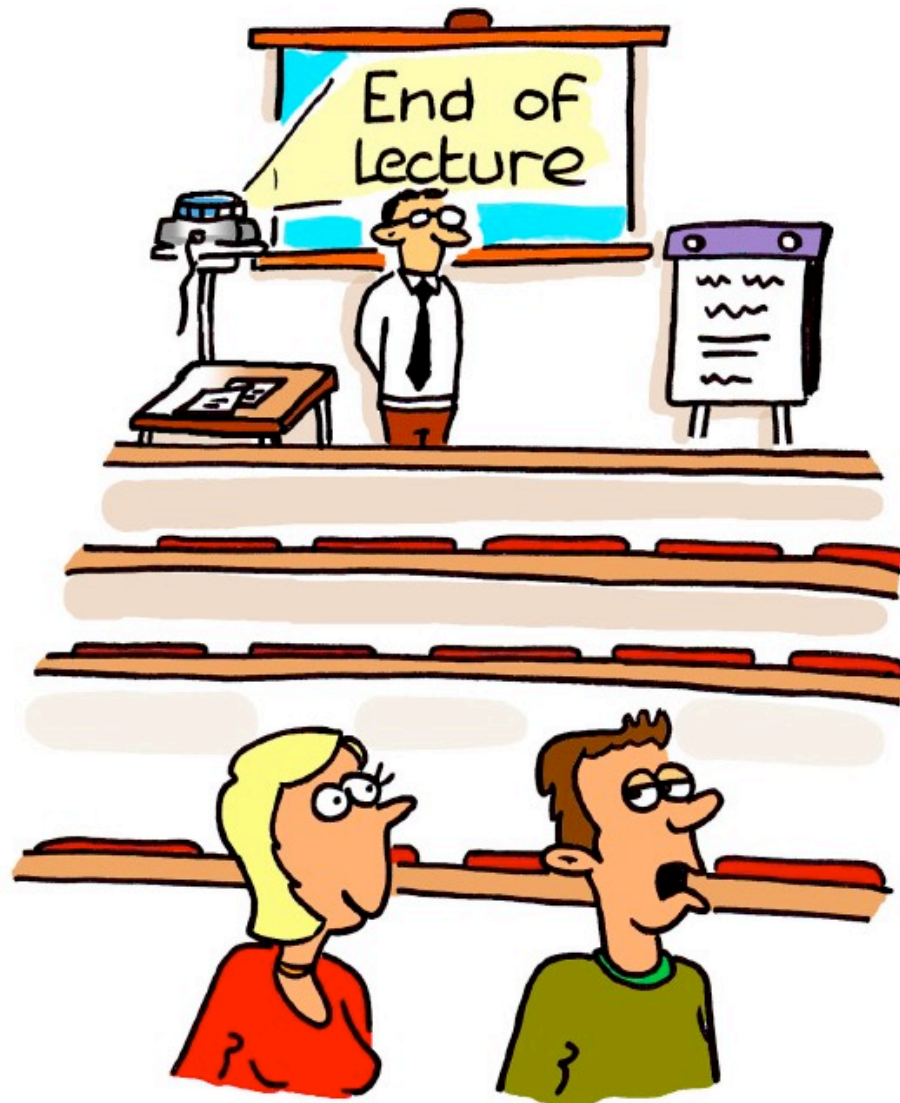


MCMC (Li Algorithm)
chain length = 100000



Summary

- An estimate of a phylogeny must be made together with an analysis of its variability.
- Uncertainty in the evolutionary history can be summarized as a distribution of trees given the data (Bayesian approach)
- The MCMC Algorithm provides a means to find this distribution
- Checking for convergence of the chain & summarizing the estimated posterior distribution are still difficult.



"He always allows time for questions. Someday he'll allow time for answers."