Comparative Genomics

Lecture 7:
Phylogenetics II

Statistical Phylogenetics

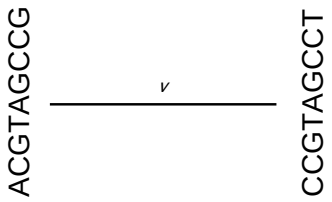
- Maximum Likelihood
- Bayesian Inference

Maximum Likelihood

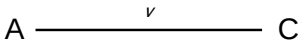
The best tree is the one with the highest probability of producing the observed data.

How can we calculate the probability that a tree generated some observed data?

How can we calculate the probability that one sequence evolves into another?



Assume that sites evolve independently.
Calculate each site separately and multiply all sites together.



Substitution Model: JC

$$Q = \begin{pmatrix} & [A] & [C] & [G] & [T] \\ [A] & -3\mu & \mu & \mu & \mu \\ [C] & \mu & -3\mu & \mu & \mu \\ [G] & \mu & \mu & -3\mu & \mu \\ [T] & \mu & \mu & \mu & -3\mu \end{pmatrix}$$

The Jukes-Cantor model
A simple discrete-state continuous time Markov model
Substitutions follow a Poisson process
Q is the instantaneous rate matrix

Substitution Probabilities: JC

$$P(v) = e^{-Qv} = \begin{pmatrix} \frac{1}{4}+\frac{3}{4}e^{-4v/3} & \frac{1}{4}-\frac{1}{4}e^{-4v/3} & \frac{1}{4}-\frac{1}{4}e^{-4v/3} & \frac{1}{4}-\frac{1}{4}e^{-4v/3} \\ \frac{1}{4}-\frac{1}{4}e^{-4v/3} & \frac{1}{4}+\frac{3}{4}e^{-4v/3} & \frac{1}{4}-\frac{1}{4}e^{-4v/3} & \frac{1}{4}-\frac{1}{4}e^{-4v/3} \\ \frac{1}{4}-\frac{1}{4}e^{-4v/3} & \frac{1}{4}-\frac{1}{4}e^{-4v/3} & \frac{1}{4}+\frac{3}{4}e^{-4v/3} & \frac{1}{4}-\frac{1}{4}e^{-4v/3} \\ \frac{1}{4}-\frac{1}{4}e^{-4v/3} & \frac{1}{4}-\frac{1}{4}e^{-4v/3} & \frac{1}{4}-\frac{1}{4}e^{-4v/3} & \frac{1}{4}+\frac{3}{4}e^{-4v/3} \end{pmatrix}$$

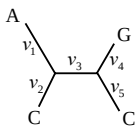
Substitution probabilities for the Jukes-Cantor model
Calculated by matrix exponentiation
v is branch length measured in expected substitutions per site
Probabilities drive system to stationarity

Assume that sites evolve independently.
Calculate each site separately and multiply all sites together.

A —— v —— C

The probability for this site is: $\frac{1}{4}-\frac{1}{4}e^{-4v/3}$

Now we can calculate the probability that a particular tree with a particular set of branch lengths generated some set of aligned sequences



Data

X The data

```
Taxon Characters
```


ACG TTA TTA AAT TGT CCT CTT TTC AGA

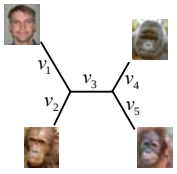

ACG TGT TTC GAT CGT CCT CTT TTC AGA


ACG TGT TTA GAC CGA CCT CGG TTA AGG


ACA GGA TTA GAT CGT CCG CTT TTC AGA

Model: tree and branch lengths

θ Parameters



topology (τ)

branch lengths (v_i)
(expected amount of change)

$\theta = (\tau, v)$

Model: molecular evolution

θ Parameters

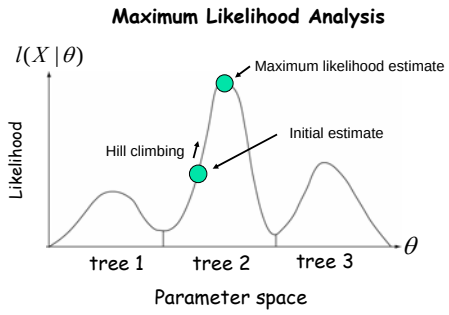
$$Q = \begin{pmatrix} & [A] & [C] & [G] & [T] \\ [A] & - & \mu & \mu & \mu \\ [C] & \mu & - & \mu & \mu \\ [G] & \mu & \mu & - & \mu \\ [T] & \mu & \mu & \mu & - \end{pmatrix}$$

instantaneous rate matrix
(Jukes-Cantor)

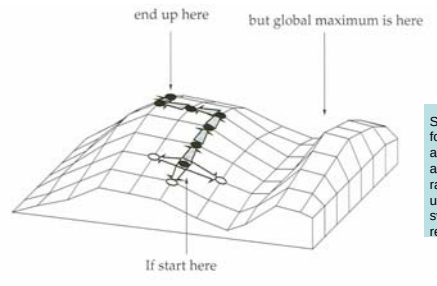
Maximum likelihood analysis

$l(X \mid \theta)$ Likelihood function

The probability (likelihood) of the data given the parameters (topology, branch lengths, substitution parameters,...)



Heuristic Search



Start at the tree found by stepwise addition, random addition, or at a random tree. Climb uphill using branch swapping (tree rearrangements).

Higher is better (better likelihood)

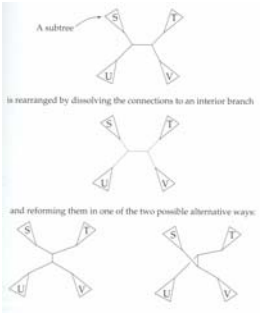
Stepwise Addition

At each level (A, B, C, etc), all trees are examined and only the best tree (or the *n* best trees) are kept as seed(s) in the next step of the algorithm.

Tree Rearrangements

- NNI: Nearest Neighbor Interchanges. Fastest but only modest tree changes
- SPR: Subtree Pruning and Regrafting: Slower but more substantial rearrangements
- TBR: Tree Bisection and Reconnection. Slowest but most comprehensive rearrangements

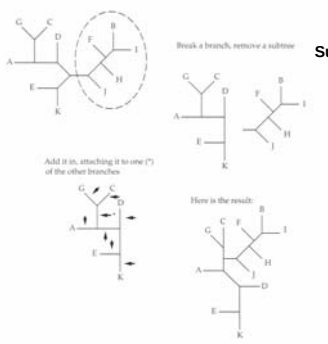
NNI



Nearest Neighbor Interchange

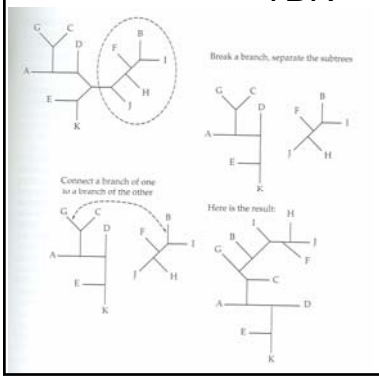
Delete one branch and try the two other alternative arrangements of the four subtrees surrounding that branch

SPR



Subtree Pruning and Regrafting

TBR

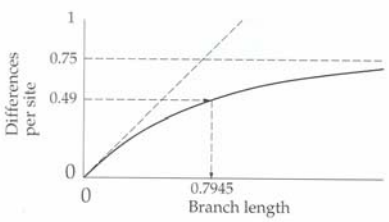


Tree Bisection Reconnection

Advantages of ML

- ML corrects for multiple hits. If this is not done, long branches can mislead a phylogenetic analysis ("long-branch attraction")
- ML can estimate evolutionary parameters of interest like substitution rates and stationary state frequencies

Saturation Plot



The longer the branch, the larger the discrepancy between the observed number of differences (parsimony length) and the true branch length (true number of changes)

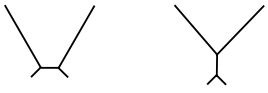
Maximum likelihood optimizes branch length by taking the overall difference between the sequences into account

ACGTAGCCG —*v*— CCGTAGCCT

ML is obtained when $p_{ii}(v) = \frac{m}{N}$

Where *m* is the number of the matching sites and *N* is total sequence length

Long-Branch Attraction



True tree Parsimony gives you this tree

Parsimony gives you the wrong tree if you do not weight branches according to their length.
Such weighting is difficult to do within the parsimony framework.

Evolutionary Models

- Different substitutions occur at different rates
- Different sites in a sequence evolve at different rates

Substitution Model: JC

$$Q = \begin{array}{c} \\ [A] \\ [C] \\ [G] \\ [T] \end{array} \begin{array}{c} \begin{array}{cccc} [A] & [C] & [G] & [T] \end{array} \\ \begin{pmatrix} -3\mu & \mu & \mu & \mu \\ \mu & -3\mu & \mu & \mu \\ \mu & \mu & -3\mu & \mu \\ \mu & \mu & \mu & -3\mu \end{pmatrix} \end{array}$$

The Jukes-Cantor model
All substitutions occur with the same rate

Substitution Model: K2P

$$Q = \begin{array}{c} \\ [A] \\ [C] \\ [G] \\ [T] \end{array} \begin{array}{c} \begin{array}{cccc} [A] & [C] & [G] & [T] \end{array} \\ \begin{pmatrix} - & \mu & \kappa\mu & \mu \\ \mu & - & \mu & \kappa\mu \\ \kappa\mu & \mu & - & \mu \\ \mu & \kappa\mu & \mu & - \end{pmatrix} \end{array}$$

The Kimura 2-parameter model
κ is the transition/transversion rate ratio

Substitution Model: HKY

$$Q = \begin{array}{c} \\ [A] \\ [C] \\ [G] \\ [T] \end{array} \begin{array}{c} \begin{array}{cccc} [A] & [C] & [G] & [T] \end{array} \\ \begin{pmatrix} - & \pi_C\mu & \pi_G\kappa\mu & \pi_T\mu \\ \pi_A\mu & - & \pi_G\mu & \pi_T\kappa\mu \\ \pi_A\kappa\mu & \pi_C\mu & - & \pi_T\mu \\ \pi_A\mu & \pi_C\kappa\mu & \pi_G\mu & - \end{pmatrix} \end{array}$$

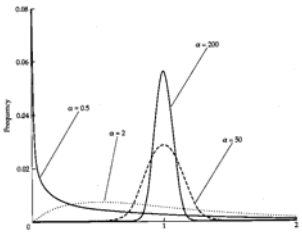
The Hasegawa-Kishino-Yano model
κ is the transition/transversion rate ratio
π_i is the stationary base frequency of nucleotide i

Substitution Model: GTR

$$Q = \begin{array}{c} \\ [A] \\ [C] \\ [G] \\ [T] \end{array} \begin{array}{c} \begin{array}{cccc} [A] & [C] & [G] & [T] \end{array} \\ \begin{pmatrix} - & \pi_C r_{AC}\mu & \pi_G r_{AG}\mu & \pi_T r_{AT}\mu \\ \pi_A r_{AC}\mu & - & \pi_G r_{CG}\mu & \pi_T r_{CT}\mu \\ \pi_A r_{AG}\mu & \pi_C r_{CG}\mu & - & \pi_T r_{GT}\mu \\ \pi_A r_{AT}\mu & \pi_C r_{CT}\mu & \pi_G r_{GT}\mu & - \end{pmatrix} \end{array}$$

The General Time Reversible model
r_{ij} is the rate of substitution between nucleotides i and j
π_i is the stationary base frequency of nucleotide i

Rate Variation Across Sites



Gamma distribution
The shape of the distribution is determined by a single parameter, the shape parameter α

Statistical Phylogenetics

- Maximum Likelihood
- Bayesian Inference

Max. Likelihood - Bayesian

- Both are parametric statistical approaches to phylogenetic inference
- Both are based on the same stochastic models of molecular evolution
- Both address the problem of long-branch attraction
- Both share model sensitivity

Max. Likelihood - Bayesian

Max. Likelihood

- Hill climbing is slow when there are many parameters
- Nuisance parameters cause problems
- Model testing is difficult
- Accepted philosophy

Bayesian MCMC

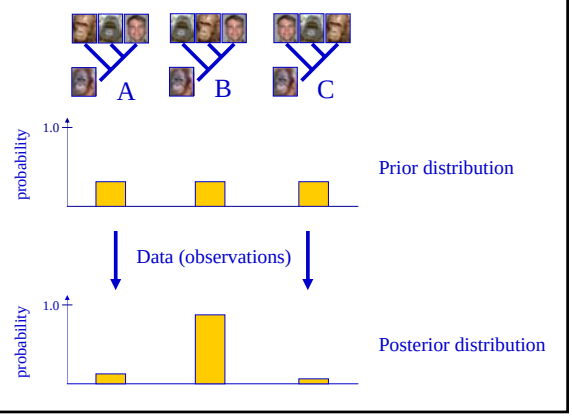
- Sampling of parameter values much faster
- No distinction between model and nuisance parameters
- Model testing easy
- Controversial philosophy

Bayesian Inference

You first specify some prior belief about the relative probability of the trees (and other parameters).

You then use some data and a stochastic model to update the prior to a posterior probability distribution on trees.

The posterior probability of a tree is the probability that it is correct given the prior and the model.



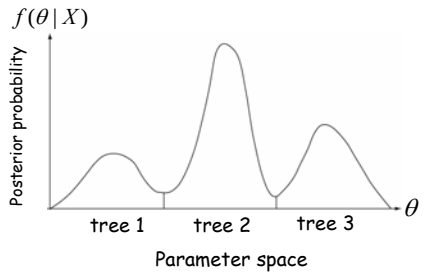
Bayes' theorem

Authored by Rev. Thomas Bayes; published posthumously in 1753

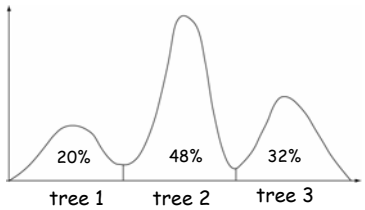
Posterior distribution | Prior distribution | Likelihood function

$$f(\theta \mid X) = \frac{p(\theta) l(X \mid \theta)}{\int p(\theta) l(X \mid \theta) d\theta}$$

Posterior probability distribution



We can focus on any parameter of interest (there are no nuisance parameters) by marginalizing the posterior over the other parameters (integrating out the uncertainty in the other parameters)



(Percentages denote marginal probability distribution on trees)

Why is it called marginalizing?

trees

joint probabilities

branch length vectors

	τ_1	τ_2	τ_3	
v_1	0.10	0.07	0.12	0.29
v_2	0.05	0.22	0.06	0.33
v_3	0.05	0.19	0.14	0.38
	0.20	0.48	0.32	

marginal probabilities

Estimating the Posterior

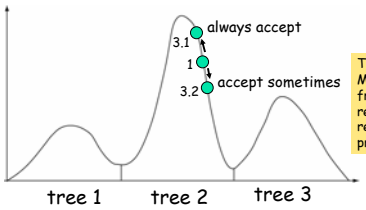
- Analytical calculation is impossible except for very simple examples
- Random sampling of parameter space is also impossible (huge space, most of it with very low probability)
- However, we can do dependent sampling using the Markov chain Monte Carlo (MCMC) technique

Markov chain Monte Carlo

- Set up a Markov process such that the stationary state is equivalent to the posterior probability distribution
- Regardless of starting state, if we run this simulation long enough, we will end up sampling from the posterior probability distribution
- Much of the difficulty of MCMC sampling is to find a process that converges quickly onto the stationary state

Markov chain Monte Carlo

1. Start at an arbitrary point
2. Make a small random move
3. Calculate height ratio (r) of new state to old state:
   1. $r > 1$ -> new state accepted
   2. $r < 1$ -> new state accepted with probability r. If new state not accepted, stay in the old state
4. Go to step 2



The proportion of time the MCMC procedure samples from a particular parameter region is an estimate of that region's posterior probability mass

Metropolis-Hastings Sampling

Assume that the current state has parameter values θ

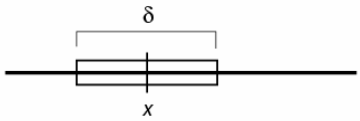
Consider a move to a state with parameter values θ^* according to proposal density q

Accept the move with probability

$$r = \min\left(1, \frac{p(\theta^*)}{p(\theta)} \frac{l(X \mid \theta^*)}{l(X \mid \theta)} \frac{q(\theta \mid \theta^*)}{q(\theta^* \mid \theta)}\right)$$

(prior ratio x likelihood ratio x proposal ratio)

Sliding Window Proposal



New values are picked uniformly from a sliding window of size δ centered on x.

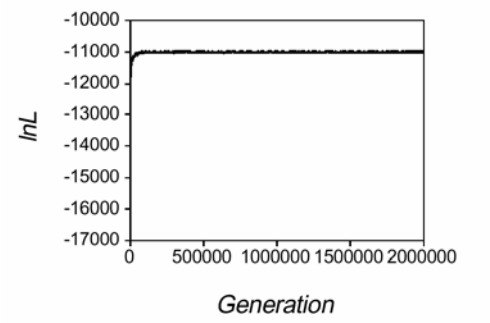
Tuning parameter: δ

Bolder proposals: increase δ

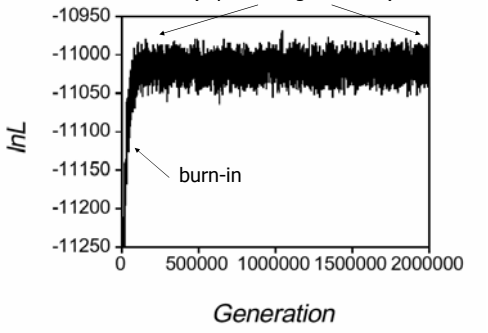
More modest proposals: decrease δ

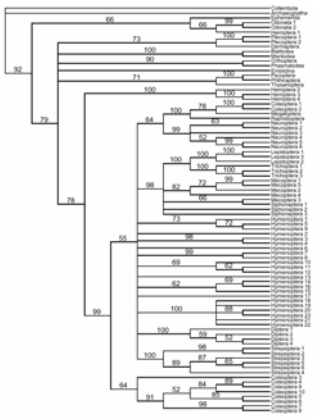
Works best when the effect on the probability of the data is similar throughout the parameter range

An Example of a Bayesian MCMC run



stationary phase sampled with thinning (rapid mixing essential)





Majority rule consensus tree from an MCMC run (insect 18S data, GTR + Γ)

Frequencies represent the posterior probability of the clades

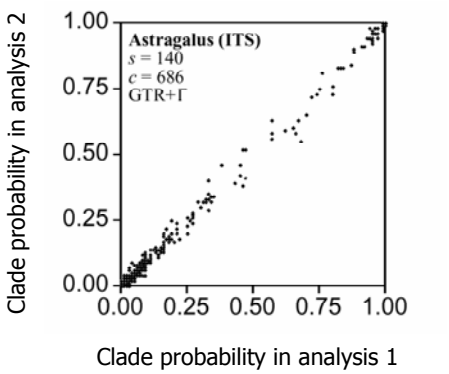
Probability of clade being true given data and model

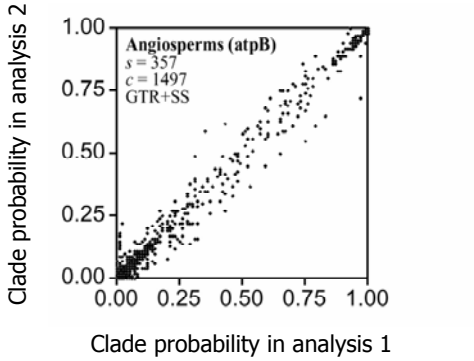
r_{AC}	1.35 (0.98, 1.82)
r_{AG}	3.24 (2.55, 4.06)
r_{AT}	1.64 (1.24, 2.11)
r_{CG}	1.18 (0.89, 1.56)
r_{CT}	5.93 (4.63, 7.54)
r_{GT}	1
α	0.32 (0.29, 0.35)
π_A	0.28 (0.26, 0.30)
π_C	0.20 (0.18, 0.22)
π_G	0.24 (0.22, 0.27)
π_T	0.28 (0.26, 0.30)

Mean and 95% credibility interval for model parameters

Assessing Convergence

- Look at the change in probability of the data given the parameters over MCMC generations
- Compare windows within the same run
- Compare independent runs starting from different randomly chosen topologies



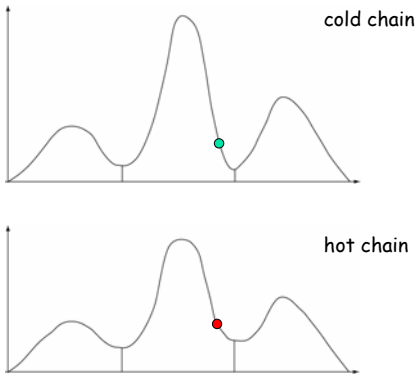


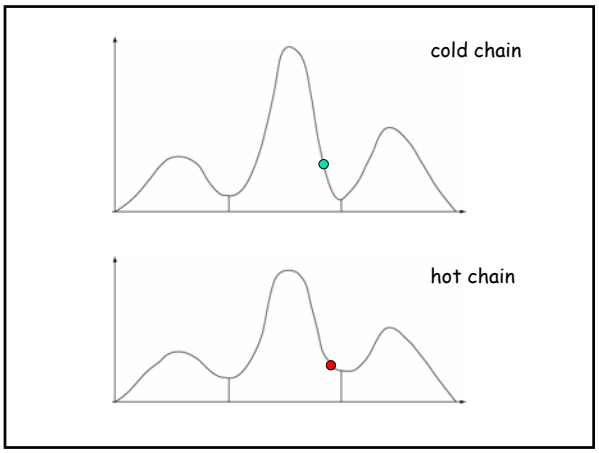
Improving Convergence

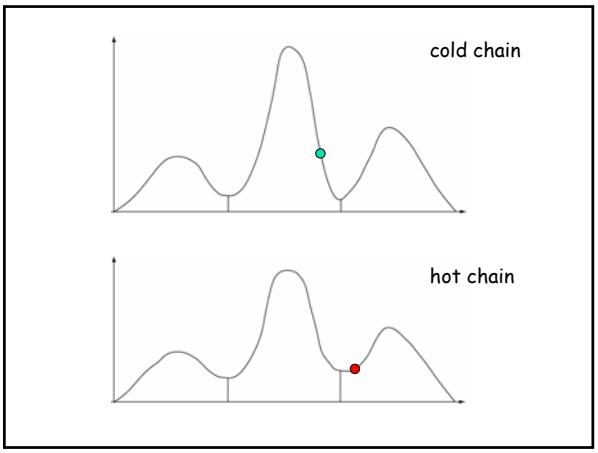
- Change tuning parameters of proposals to bring acceptance rate into the range 10 % to 70 %
- Use different proposal mechanisms
- Use Metropolis-coupled MCMC with heated chains

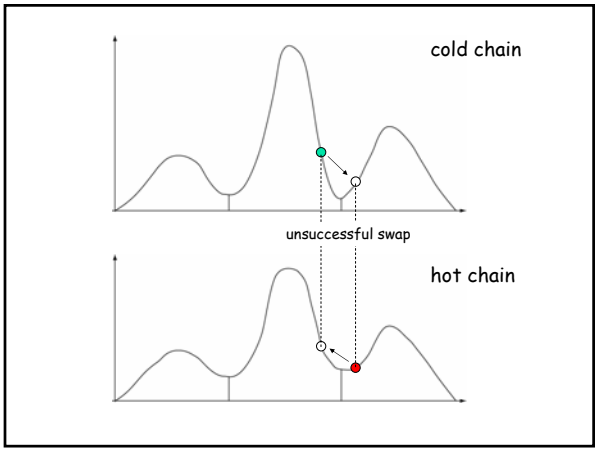
MCMCMC

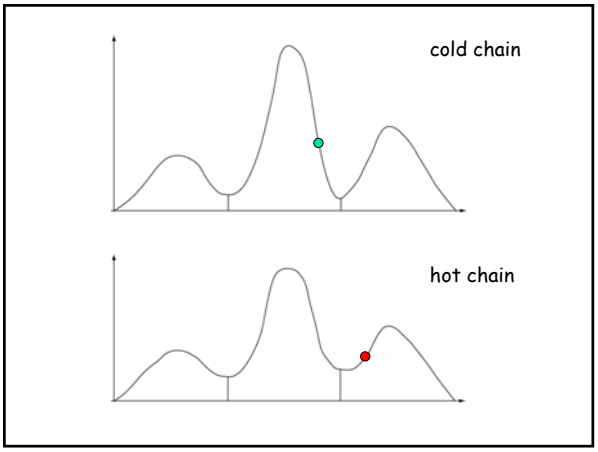
- In Metropolis-coupled Markov chain Monte Carlo, or MCMCMC, or $(MC)^3$ 'heated' chains are used to propose new states for the normal 'cold' chain
- The heated chains run in a landscape obtained by raising the posterior probability with a factor < 1
- The smaller the factor, the more heated the landscape becomes.
- With regular intervals, we try to swap states between the hot and the cold chains

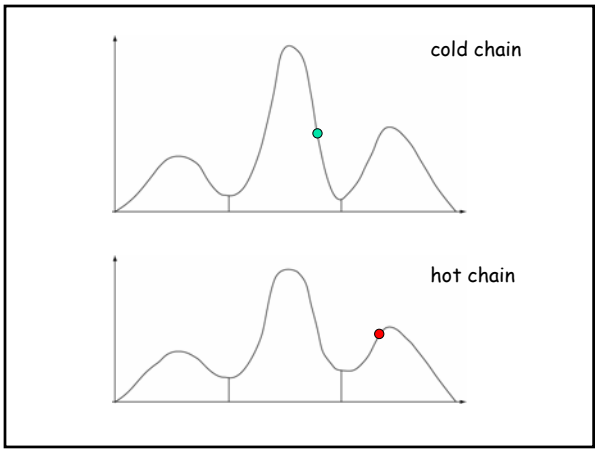


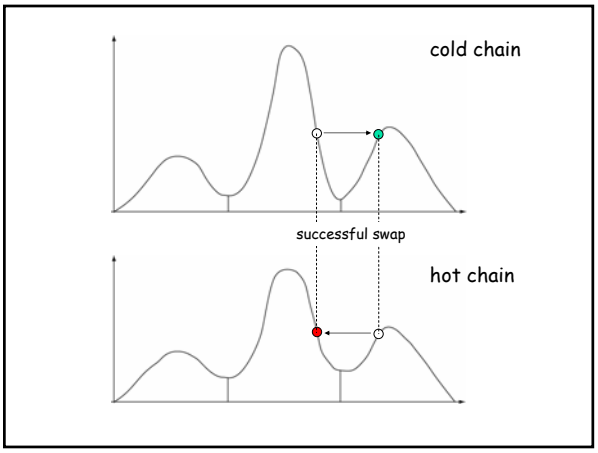

cold chain
hot chain

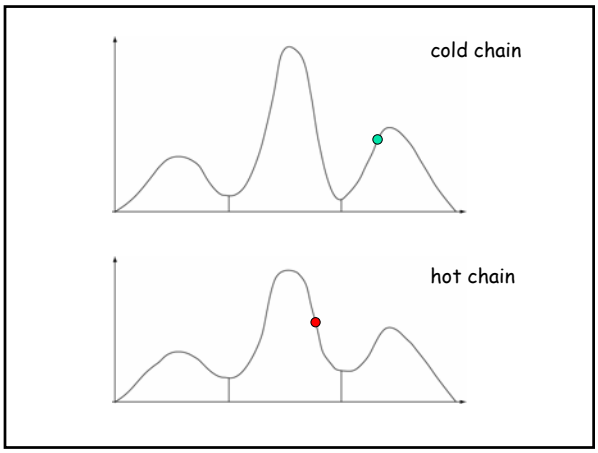

cold chain
hot chain

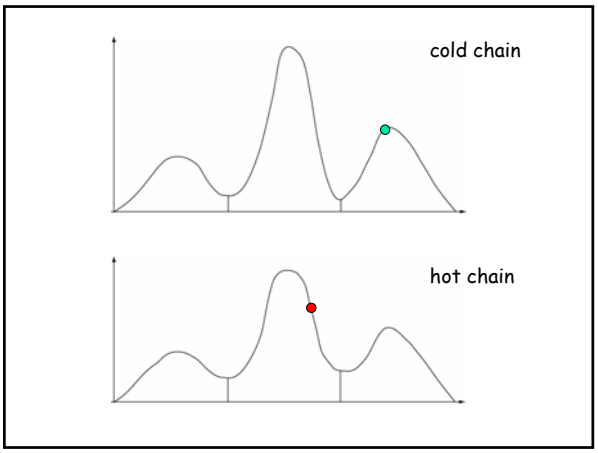

cold chain
unsuccessful swap
hot chain

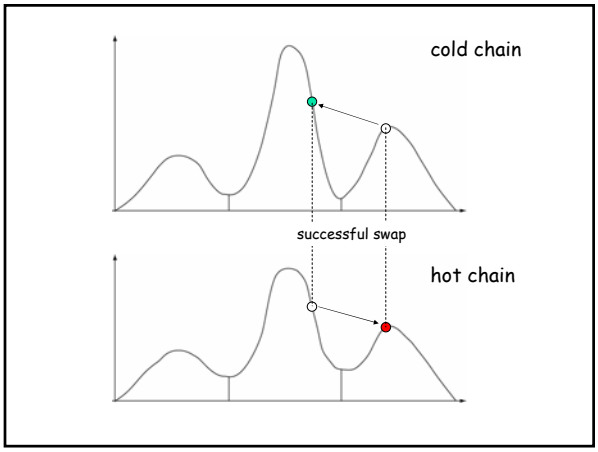

cold chain
hot chain

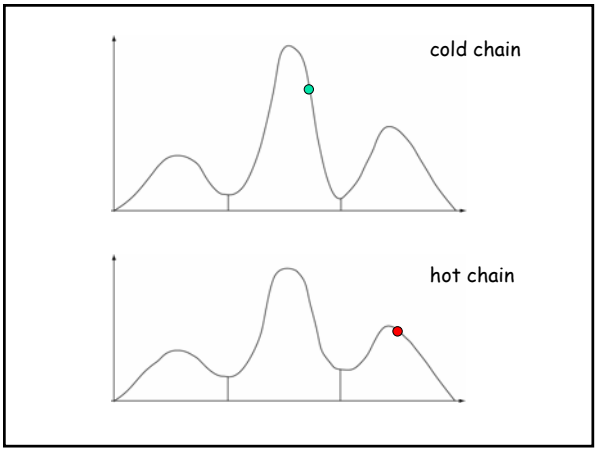

cold chain
hot chain


cold chain
successful swap
hot chain









Summary

- Statistical phylogenetics is based on evolutionary models (probability models) that account for variation in rates among types of substitutions and across sites
- Maximum likelihood inference, standard statistical approach but slow for the phylogeny problem
- Bayesian MCMC inference, fast but technically more complicated and more controversial approach