

Bayesian Phylogenetic Inference: Introduction

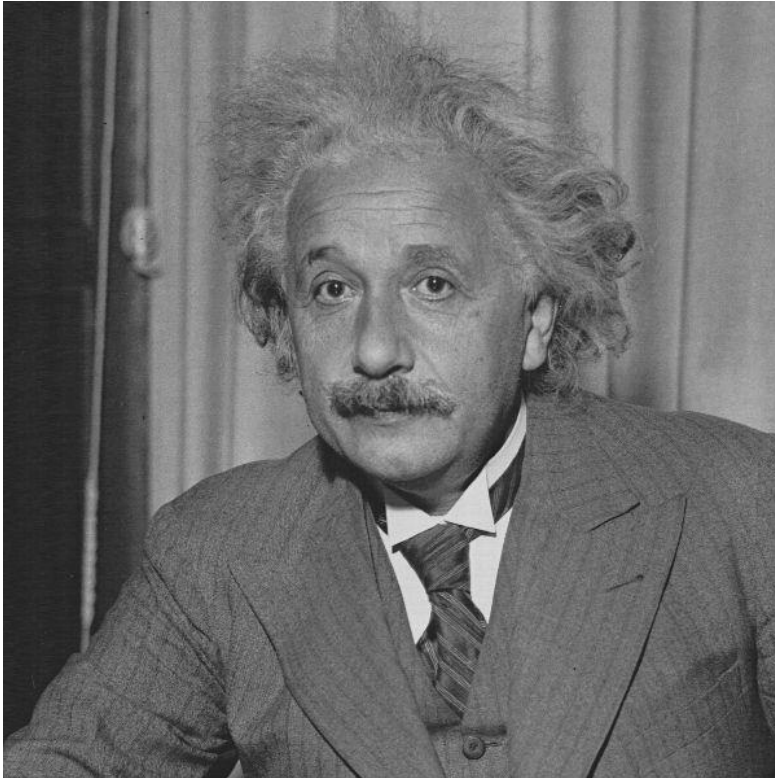
Fred(rik) Ronquist
Swedish Museum of Natural History,
Stockholm, Sweden

BIG 4 Workshop, October 9-19, Tovetorp, Sweden

Topics

- Probability 101
- Bayesian Phylogenetic Inference
- Markov chain Monte Carlo
- Bayesian Model Choice and Model Averaging

1. Probability 101



Do not worry about your
difficulties in Mathematics. I
can assure you that mine
are still greater.

Albert Einstein

the **ABCs** *of* College Drinking

25 TIPS FOR NAVIGATING THE COLLEGIATE PARTY SCENE



Jim Matthews, M.Ed.

Author of "Beer, Booze and Books...a sober look at higher education"

Continuous probability distributions

α Uniform distribution

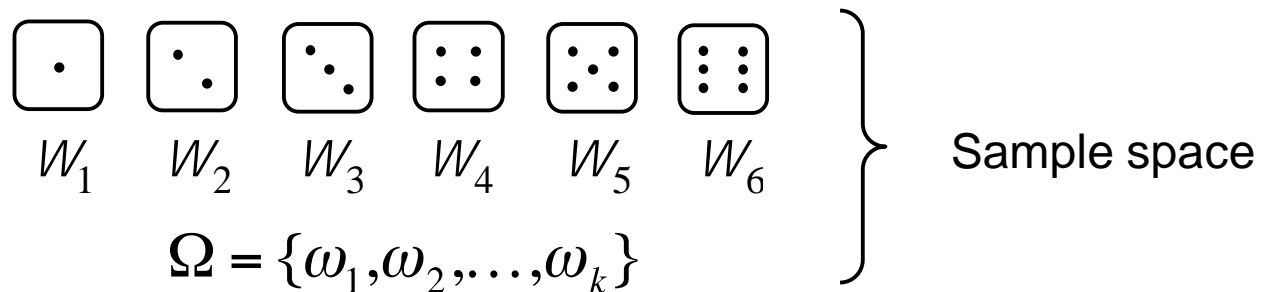
β Beta distribution

γ Gamma distribution

δ Dirichlet distribution

ε Exponential distribution

Random variable X



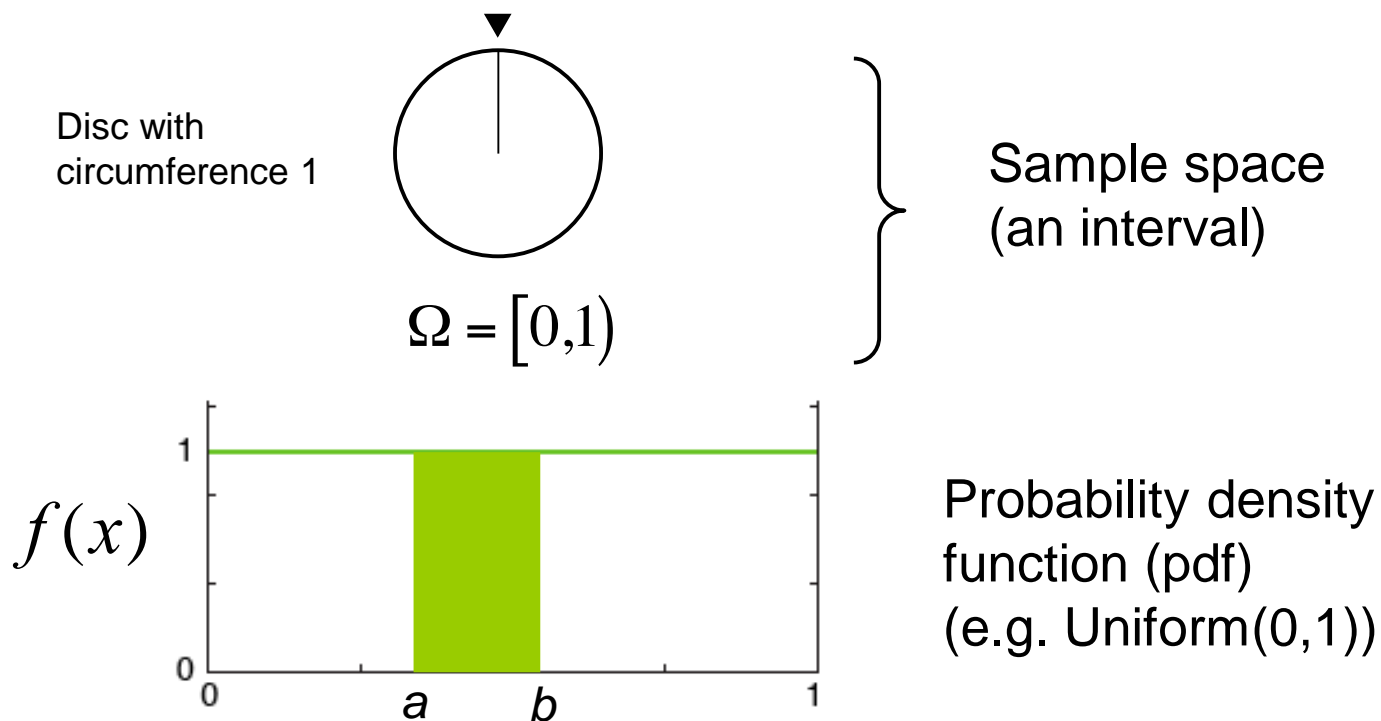
$$E = \{W_1, W_3, W_5\}$$

Event (subset of outcomes;
e.g., face with odd number)

$$\Pr(E) = \sum_{W \in E} m(W)$$

Probability

Random variable X



$$E = [a, b)$$

Event (a subspace of the sample space)

$$\Pr(E) = \int_{x \in E} f(x) dx$$

Probability

Continuous Distributions

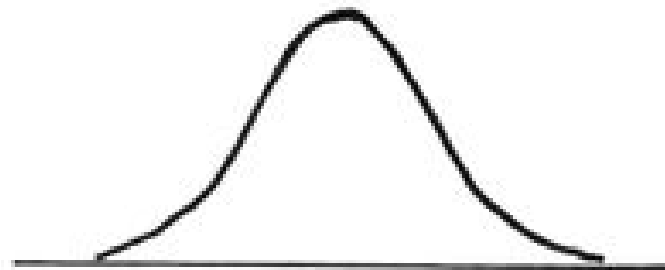
- Uniform distribution
- Beta distribution
- Gamma distribution
- Dirichlet distribution
- Exponential distribution
- Normal distribution
- Lognormal distribution
- Multivariate normal distribution

Discrete Distributions

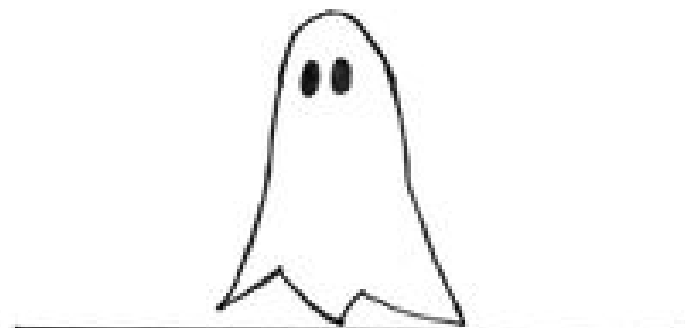
- Bernoulli distribution
- Categorical distribution
- Binomial distribution
- Multinomial distribution
- Poisson distribution

Stochastic Processes

- Markov chain
- Poisson process
- Birth-death process
- Coalescence
- Dirichlet Process Mixture



NORMAL DISTRIBUTION



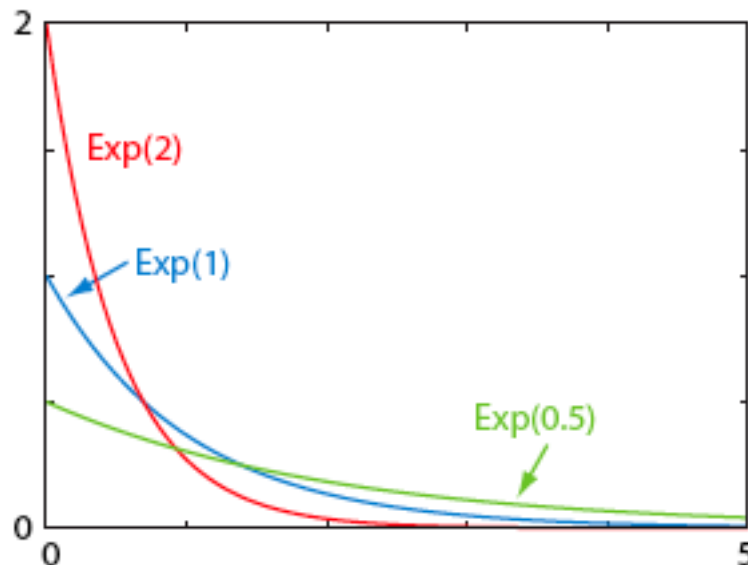
PARANOORMAL DISTRIBUTION

Exponential distribution $X \sim \text{Exp}(\lambda)$

Parameters: λ = rate (of decay)

Probability density function: $f(x) = \lambda e^{-\lambda x}$

Mean: $1/\lambda$



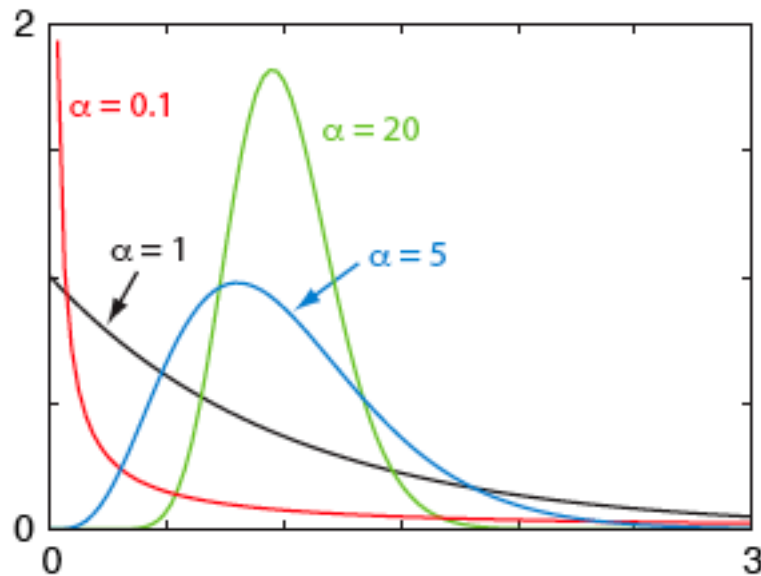
Gamma distribution $X \sim \text{Gamma}(a, b)$

Parameters: a = shape b = inverse scale

Probability density function: $f(x) \propto x^{a-1} e^{-bx}$

Mean: a/b

Scaled gamma: $a = b$



Scaled Gamma

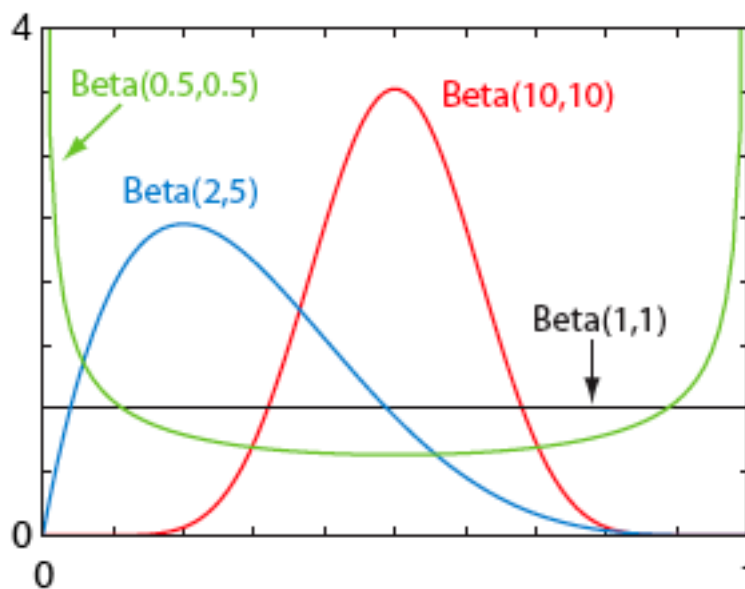
Beta distribution

$$X \sim \text{Beta}(a_1, a_2)$$

Parameters: a_1, a_2 = shape parameters

Probability density function: $f(x) \propto x^{a_1-1}(1-x)^{a_2-1}$

Mode: $\frac{a_1 - 1}{a_1 + a_2 - 2}$ Defined on two proportions of a whole
(a simplex)

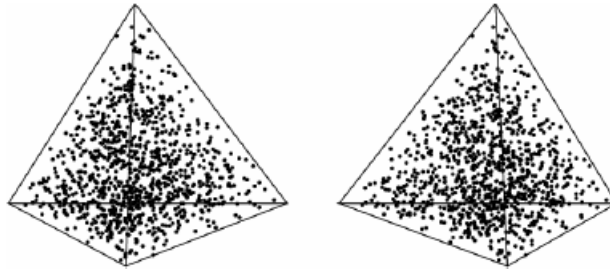


Dirichlet distribution $X \sim \text{Dir}(a) : a = \{a_1, a_2, \dots, a_k\}$

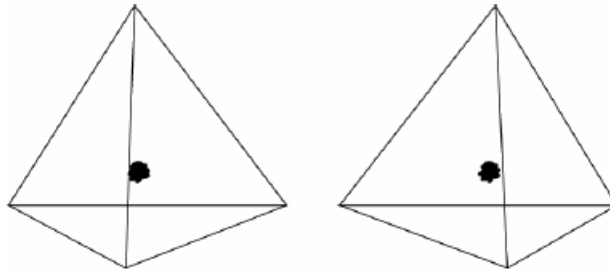
Parameters: a = vector of k shape parameters

Probability density function: $f(x) \propto \prod_i x_i^{a_i - 1}$

Defined on k proportions of a whole

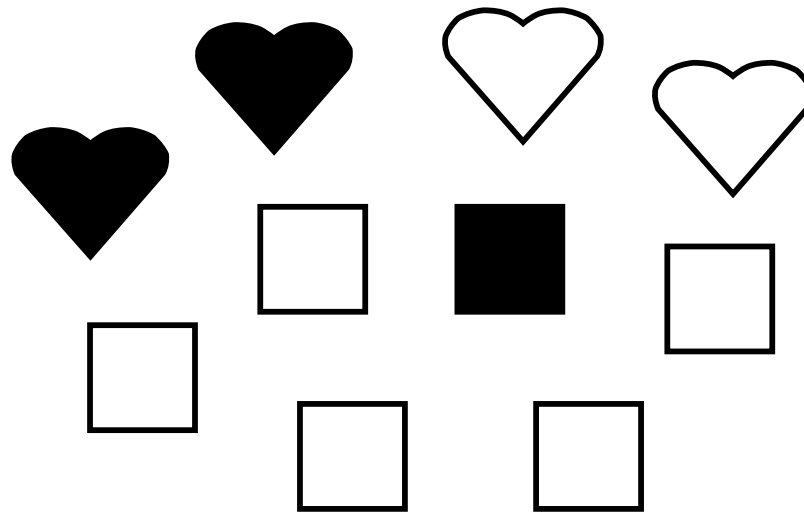


$\text{Dir}(1,1,1,1)$



$\text{Dir}(300,300,300,300)$

Conditional Probability



$$\Pr(H) = \frac{4}{10} = 0.4$$

$$\Pr(D) = \frac{3}{10} = 0.3$$

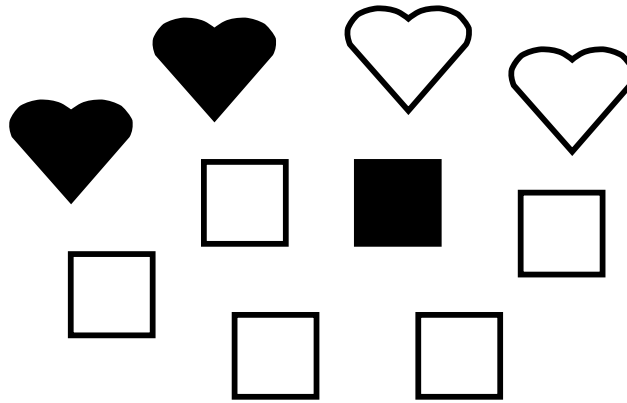
$$\text{Joint probability: } \Pr(D, H) = \frac{2}{10} = 0.2$$

$$\text{Conditional probability: } \Pr(D | H) = \frac{2}{4} = 0.5$$

Reverend Thomas Bayes
(1701-1760)



$$\Pr(A | B) \supset \Pr(B | A) ?$$



$$\Pr(D, H) = \Pr(D) \Pr(H | D) = \frac{3}{10} \cdot \frac{2}{3} = \frac{2}{10} = 0.2$$

$$= \Pr(H) \Pr(D | H) = \frac{4}{10} \cdot \frac{2}{4} = \frac{2}{10} = 0.2$$

$$\Pr(D) \Pr(H | D) = \Pr(H) \Pr(D | H)$$

$$\Pr(H | D) = \frac{\Pr(H) \Pr(D | H)}{\Pr(D)}$$

Bayes' rule

Maximum Likelihood Inference

Data D ; Model M with parameters θ

We can calculate $\Pr(D | q)$ or $f(D | q)$

Define the likelihood function $L(q) \propto f(D | q)$

Maximum likelihood: find the value of θ that maximizes $L(\theta)$

Confidence: asymptotic behavior, more samples, bootstrapping

Bayesian Inference

Data D ; Model M with parameters θ

We can calculate $\Pr(D | q)$ or $f(D | q)$

We are actually interested in $\Pr(q | D)$ or $f(q | D)$

Bayes' rule:

$$f(q | D) = \frac{f(q)f(D | q)}{f(D)}$$

Posterior density \swarrow
 Prior density \downarrow
 "Likelihood" \swarrow
 Normalizing constant \swarrow
 Marginal likelihood of the data \swarrow
 Model likelihood

$$f(D) = \int f(q)f(D | q) dq$$

Coin Tossing Example

DID THE SUN JUST EXPLODE?

(IT'S NIGHT, SO WE'RE NOT SURE.)



FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.
SINCE $p < 0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.

A stick figure on the left stands next to the detector, which is on a stand.

BAYESIAN STATISTICIAN:






































BET YOU \$50 IT HASN'T.




































A stick figure on the right stands next to the detector, which is on a stand.



What is the probability of your favorite team winning the next ice hockey World Championships?

World Championship Medalists

	Gold	Silver	Bronze	Medals	
2007					3
2008					5
2009					5
2010					1
2011					8
2012					5
2013					2
2014				other	1
2015					
2016					

Prior		Data 1		Posterior 1		Data 2		Posterior 2	
	3		in		3		out		0
	5		in		5		won		5
	5		out		0		out		0
	1		out		0		out		0
	8		in		8		out		0
	5		out		0		out		0
	2		in		2		out		0
other	1	other	out	other	0	other	out	other	0

$f(\theta)$



$f(\theta | D_1)$



$f(\theta | D_2)$

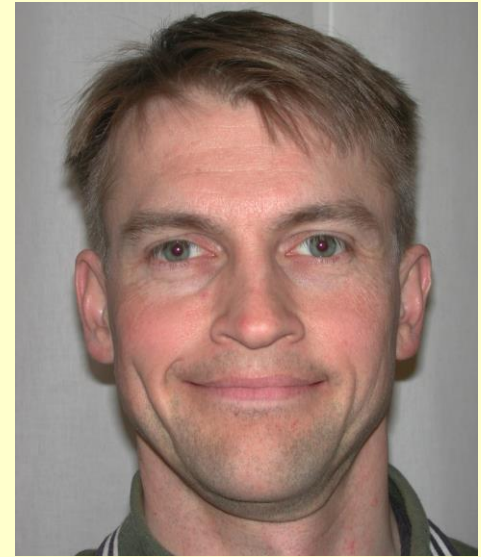
$f(\theta) \longrightarrow f(\theta | D_1 + D_2)$

Learn more:

- Wikipedia (good texts on most statistical distributions, sometimes a little difficult)
- Grinstead & Snell: Introduction to Probability. American Mathematical Society. Free pdf available from:
http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/amsbook.mac.pdf
- Team up with a statistician or a computational / theoretical evolutionary biologist!

2. Bayesian Phylogenetic Inference

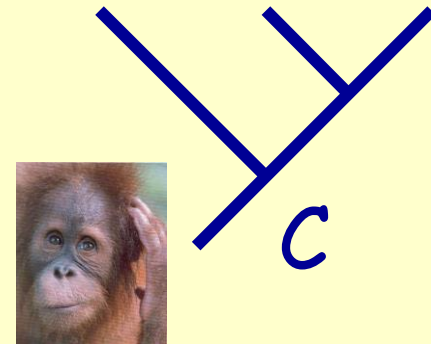
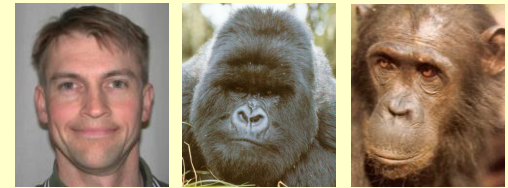
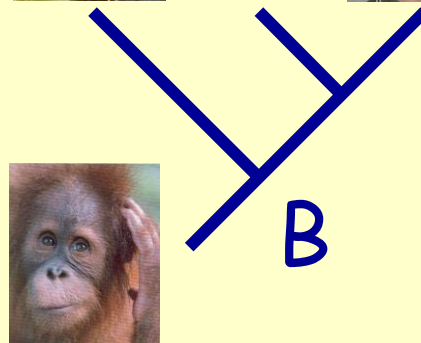
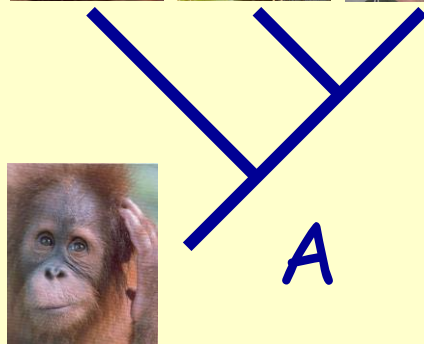
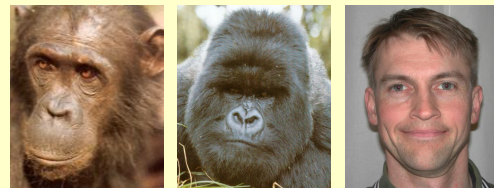
Infer relationships among three species:

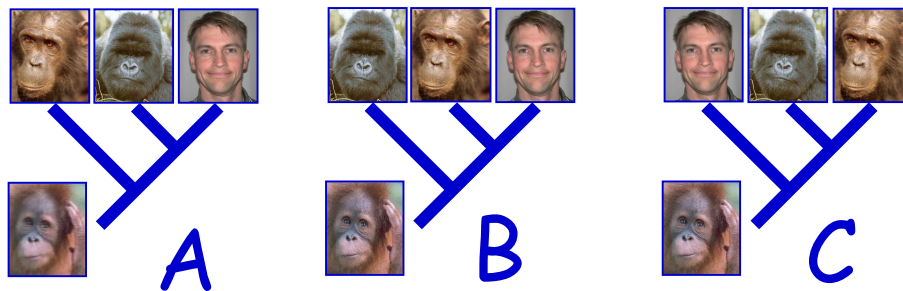


Outgroup:

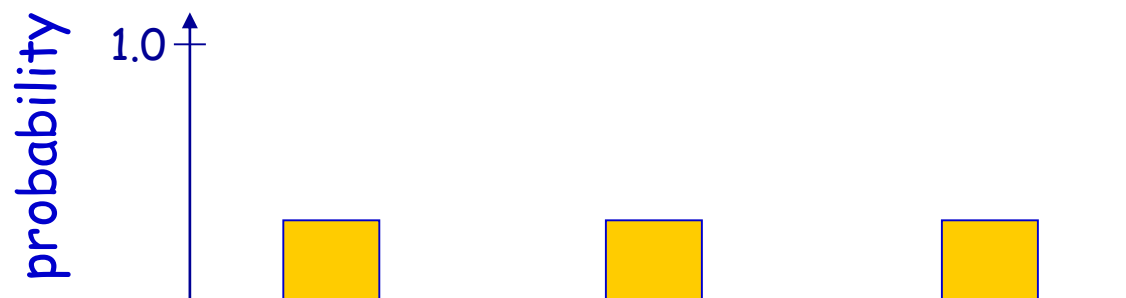


Three possible trees (topologies):



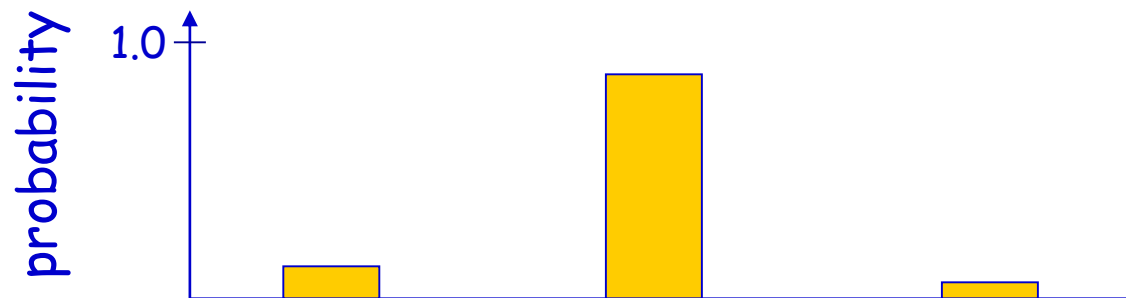


Model



Prior distribution

Data (observations)



Posterior distribution

D The data

Taxon Characters



ACG TTA TTA AAT TGT CCT CTT TTC AGA



ACG TGT TTC GAT CGT CCT CTT TTC AGA



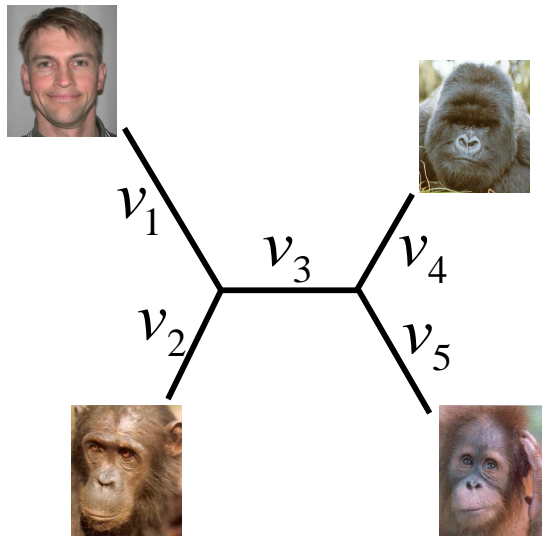
ACG TGT TTA GAC CGA CCT CGG TTA AGG



ACA GGA TTA GAT CGT CCG CTT TTC AGA

Model: topology AND branch lengths

q Parameters



topology (t)

branch lengths (v_i)
(expected amount of change)

$$q = (t, v)$$

Model: molecular evolution

q Parameters

$$Q = \begin{matrix} & \begin{matrix} [A] & [C] & [G] & [T] \end{matrix} \\ \begin{matrix} [A] \\ [C] \\ [G] \\ [T] \end{matrix} & \begin{bmatrix} - & m & m & m \\ m & - & m & m \\ m & m & - & m \\ m & m & m & - \end{bmatrix} \end{matrix}$$

instantaneous rate matrix
(Jukes-Cantor)

Model: molecular evolution

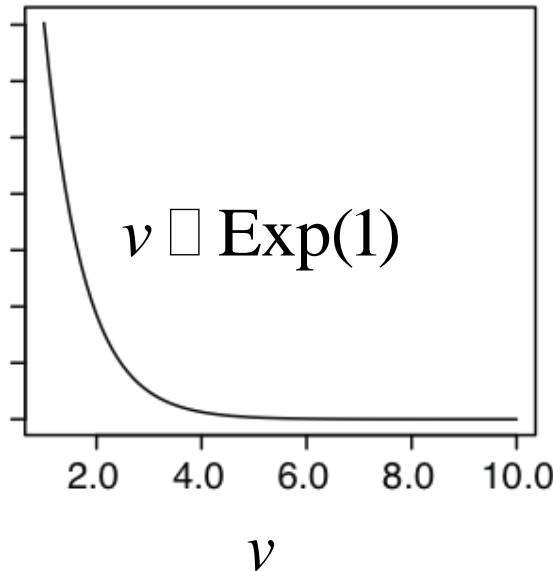
Probabilities are calculated using the transition probability matrix P

$$P(v) = e^{Qv} = \begin{cases} \frac{1}{4} - \frac{1}{4}e^{-4v/3} & \text{(change)} \\ \frac{1}{4} + \frac{3}{4}e^{-4v/3} & \text{(no change)} \end{cases}$$

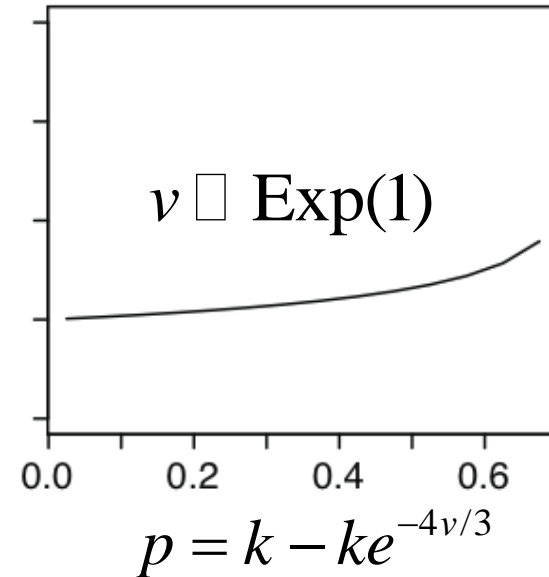
Priors on parameters

- Topology
 - all unique topologies have equal probability
- Branch lengths
 - exponential prior (puts more weight on small branch lengths)

Scale matters in priors



Branch length



Prob. of substitution

The effect on data likelihood is most important

Jeffrey's uninformative priors formalize this

Bayes' theorem

D = Data

θ = Model parameters

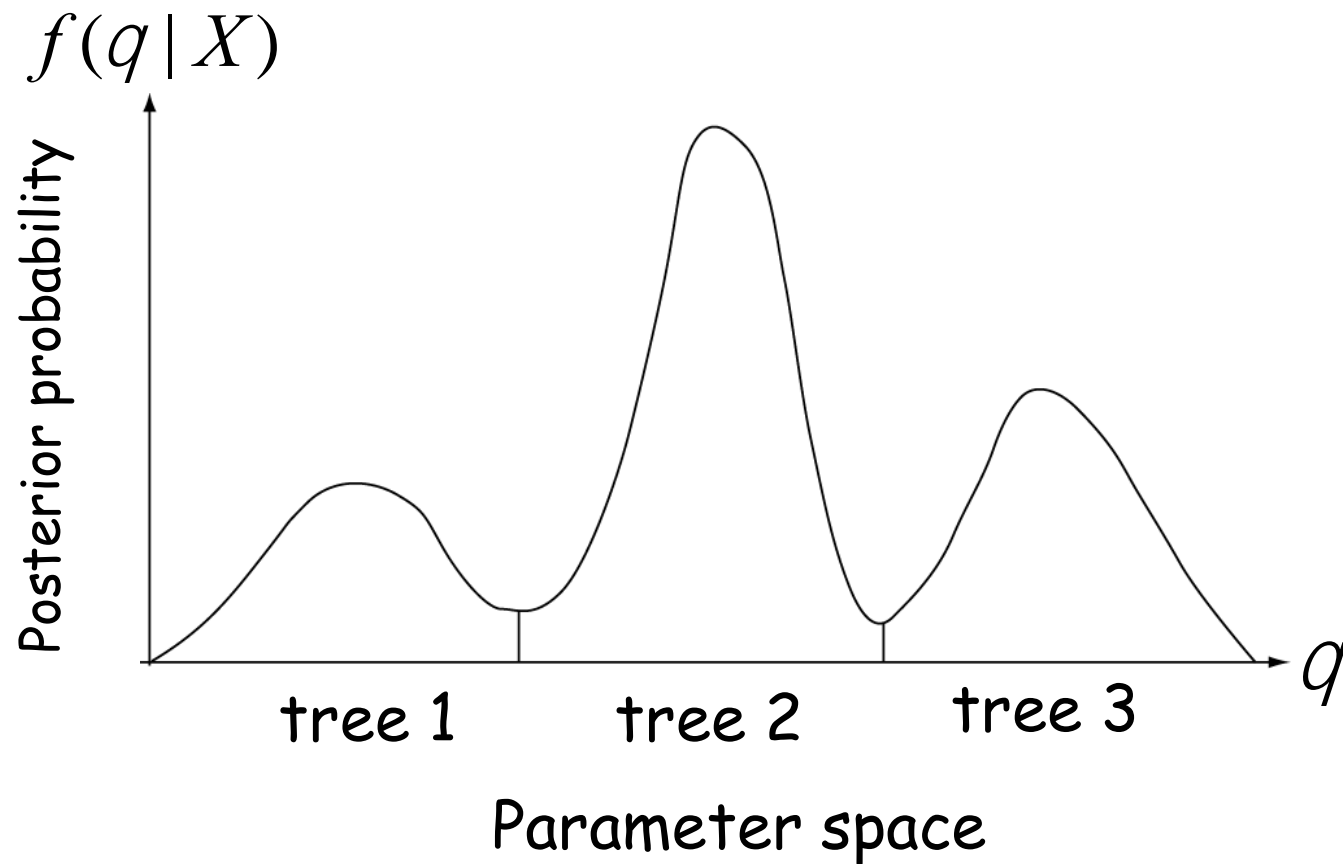


Posterior distribution Prior distribution "Likelihood"

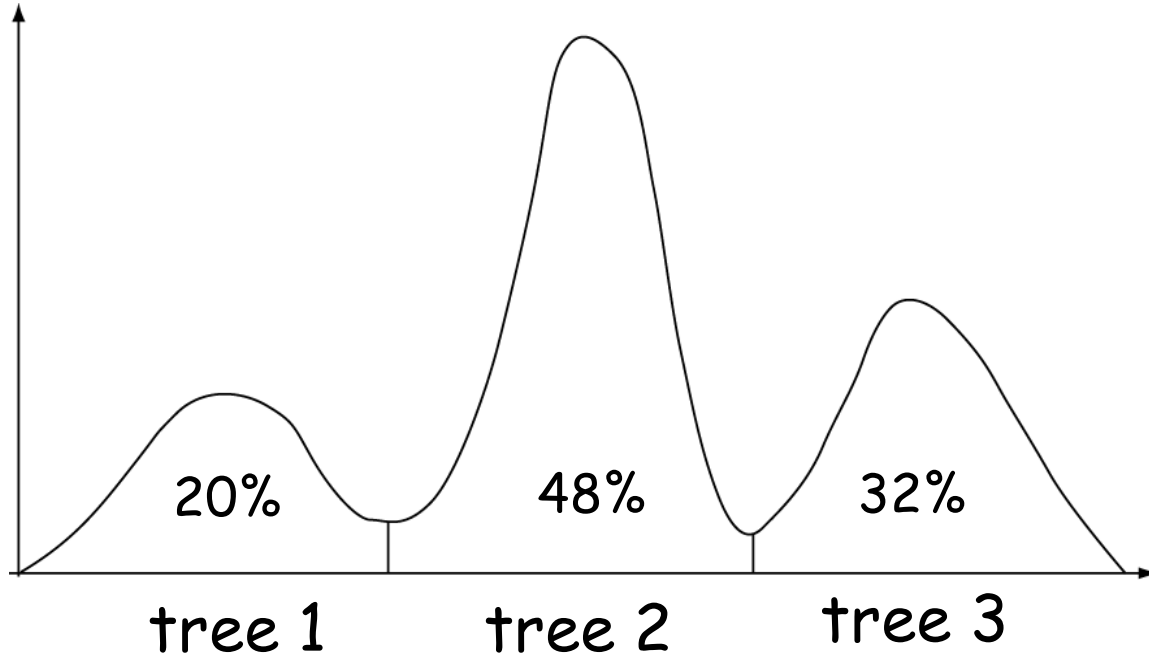
$$f(\theta | D) = \frac{f(\theta) f(D | \theta)}{\int f(\theta) f(D | \theta) d\theta}$$

Normalizing constant

Posterior probability distribution

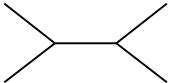
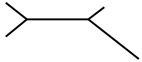
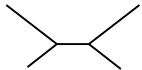


We can focus on any parameter of interest (there are no nuisance parameters) by marginalizing the posterior over the other parameters (integrating out the uncertainty in the other parameters)



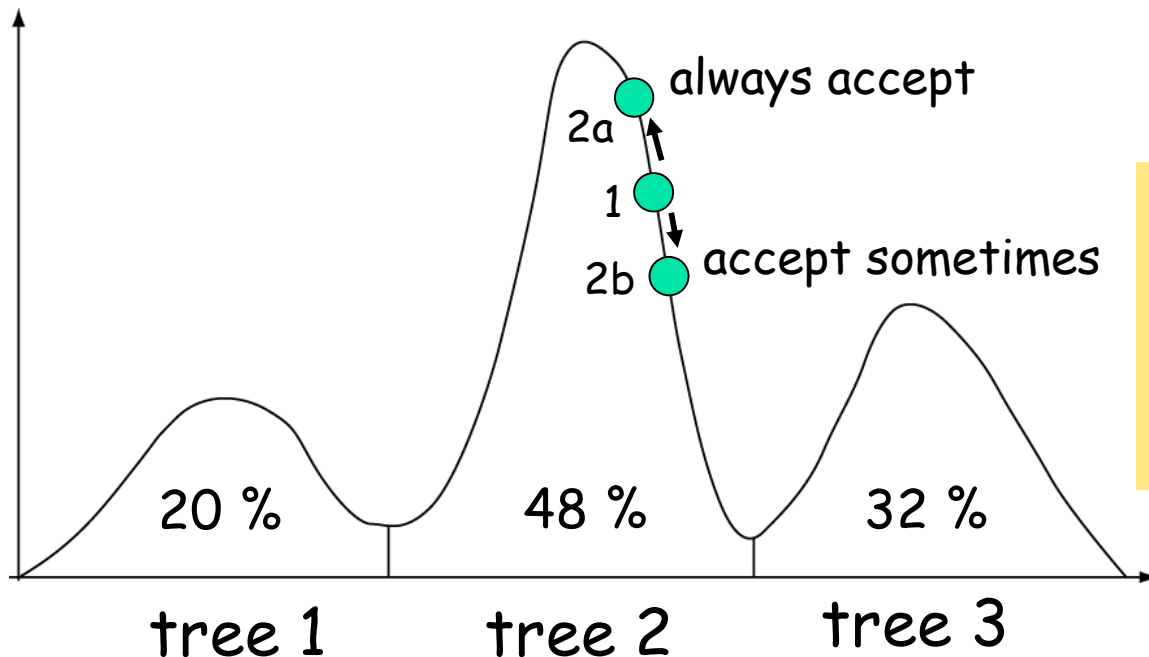
(Percentages denote marginal probability distribution on trees)

Why is it called marginalizing?

		trees			joint probabilities	
		t_1	t_2	t_3		
branch length vectors	n^1	0.10	0.07	0.12	0.29	
	n^2	0.05	0.22	0.06	0.33	
	n^3	0.05	0.19	0.14	0.38	
		0.20	0.48	0.32		
		marginal probabilities				

Markov chain Monte Carlo

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

Metropolis algorithm

Assume that the current state has
parameter values θ

Consider a move to a state with parameter
values θ^*

The height ratio r is

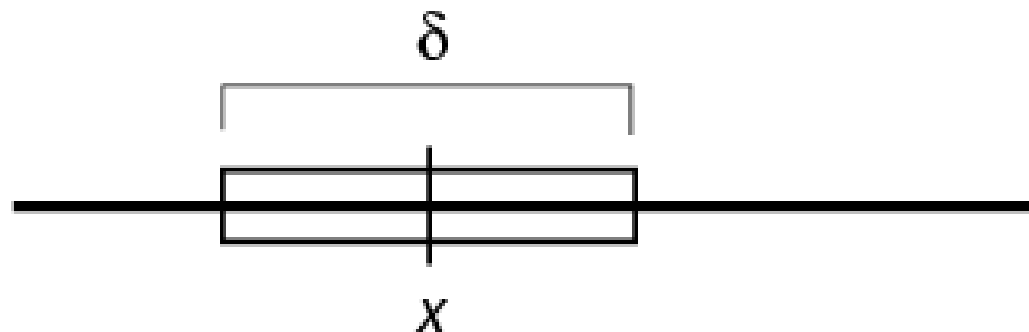
$$r = \frac{f(q^* | D)}{f(q | D)} = \frac{f(q^*)f(D | q^*) / f(D)}{f(q)f(D | q) / f(D)} = \frac{f(q^*)}{f(q)} \cdot \frac{f(D | q^*)}{f(D | q)}$$

(prior ratio x likelihood ratio)

MCMC Sampling Strategies

- Great freedom of strategies:
 - Typically one or a few related parameters changed at a time
 - You can cycle through parameters systematically or choose randomly
 - One "generation" or "iteration" or "cycle" can include a single randomly chosen proposal (or move, operator, kernel), one proposal for each parameter, a block of randomly chosen proposals

Sliding Window Proposal



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

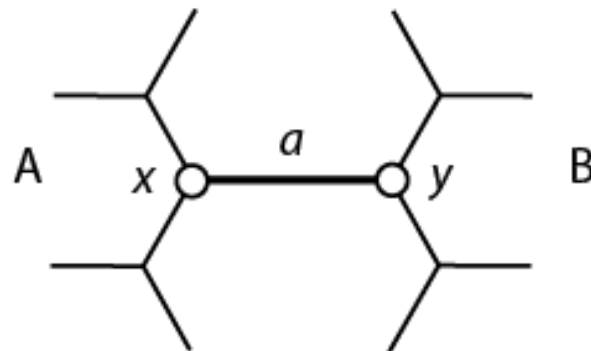
Tuning parameter: δ

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

Extending TBR



An internal branch a is chosen at random

The length of a is changed using a multiplier with tuning parameter λ

The node x is moved, with one of the adjacent branches, in subtree A, one node at a time, each time the probability of moving one more branch is p (the extension probability).

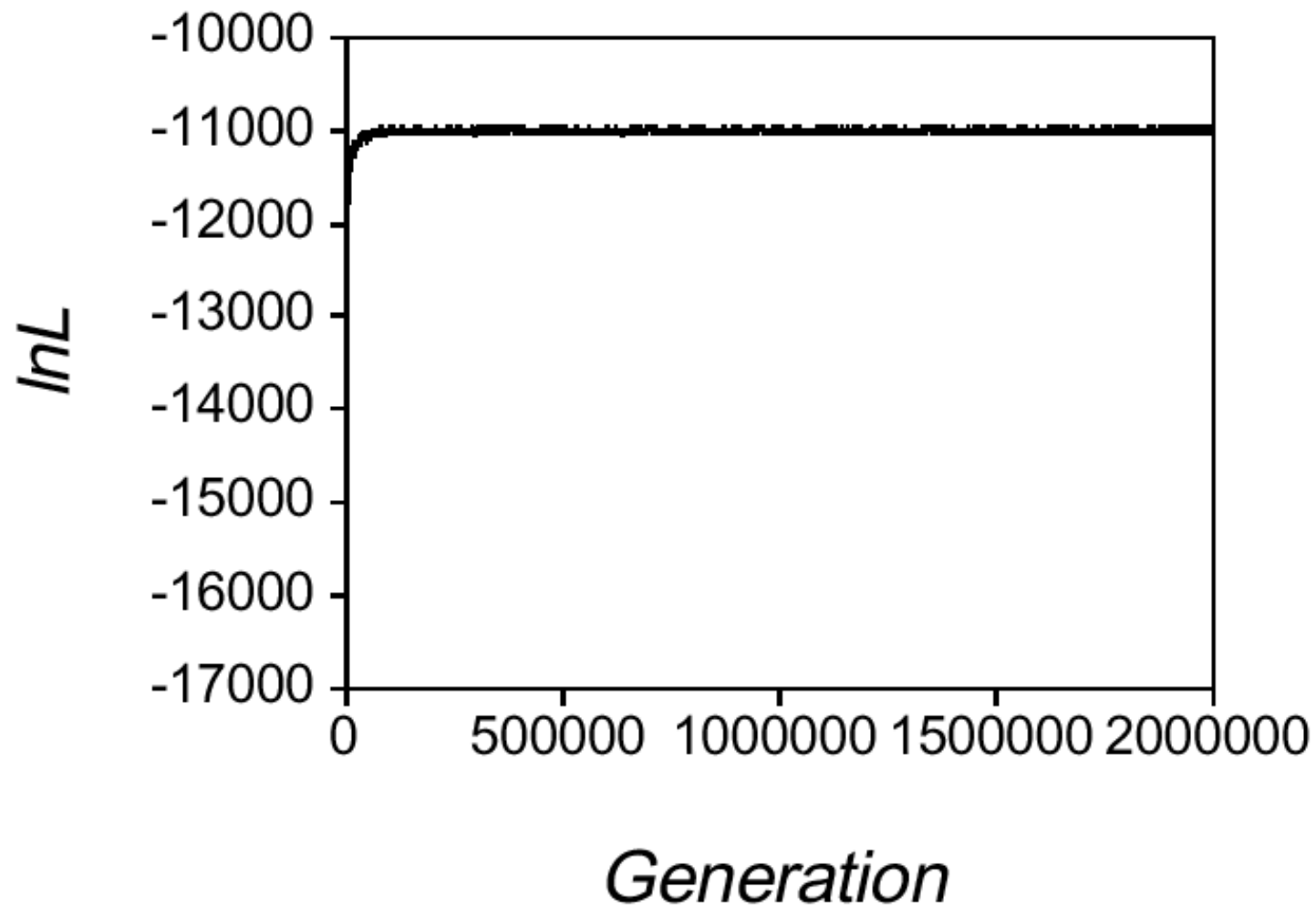
The node y is moved similarly in subtree B.

Bolder proposals: increase p

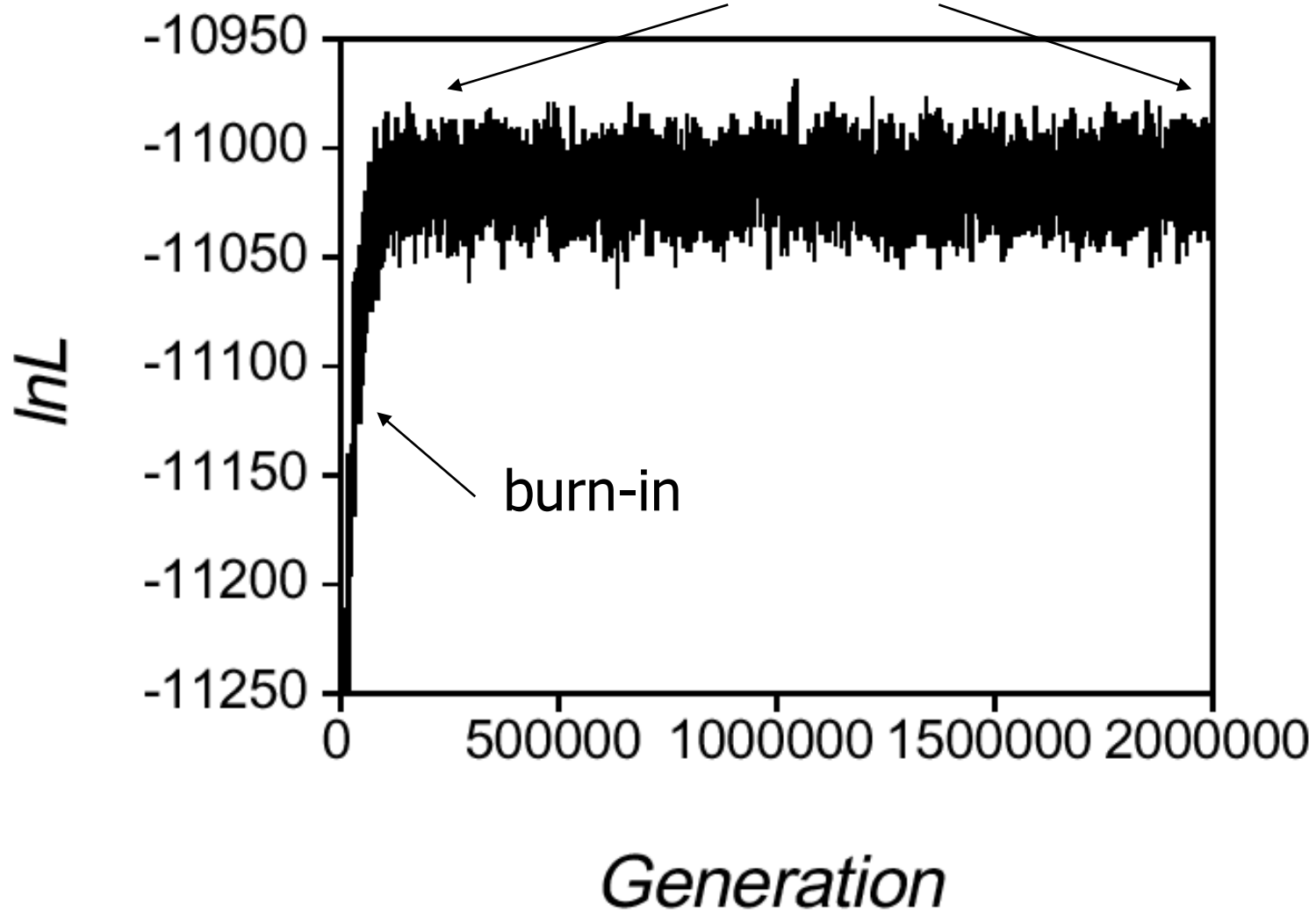
More modest proposals: decrease p

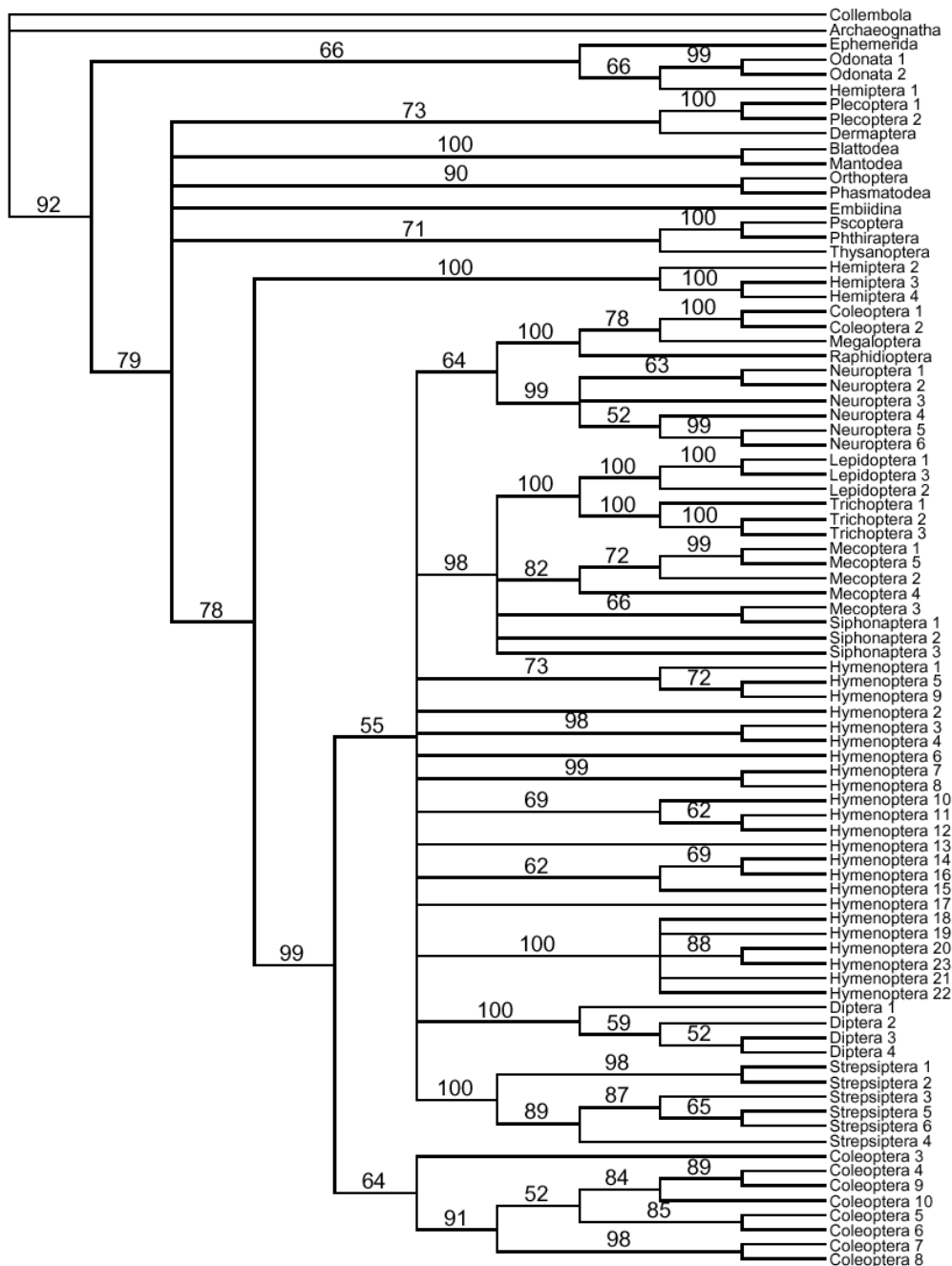
Changing λ has little effect on the boldness of the proposal.

Trace Plot



stationary phase sampled with thinning
(rapid mixing essential)





Majority rule
 consensus tree

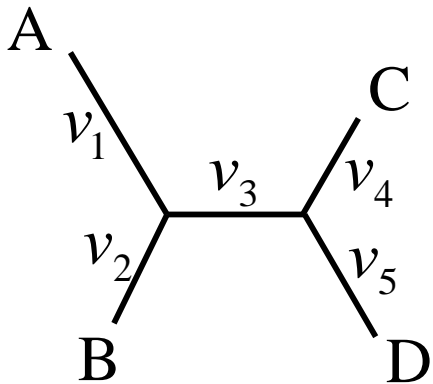
Frequencies
 represent the
 posterior
 probability of
 the clades

Probability of
 clade being true
 given data and
 model

Summarizing Trees

- Maximum posterior probability tree (MAP tree)
 - can be difficult to estimate precisely
 - can have low probability
- Majority rule consensus tree
 - easier to estimate clade probabilities exactly
 - branch length distributions can be summarized across all trees with the branch
 - can hide complex topological dependence
 - branch length distributions can be multimodal
- Credible sets of trees
 - Include trees in order of decreasing probability to obtain, e.g., 95 % credible set
- "Median" or "central" tree

Adding Model Complexity

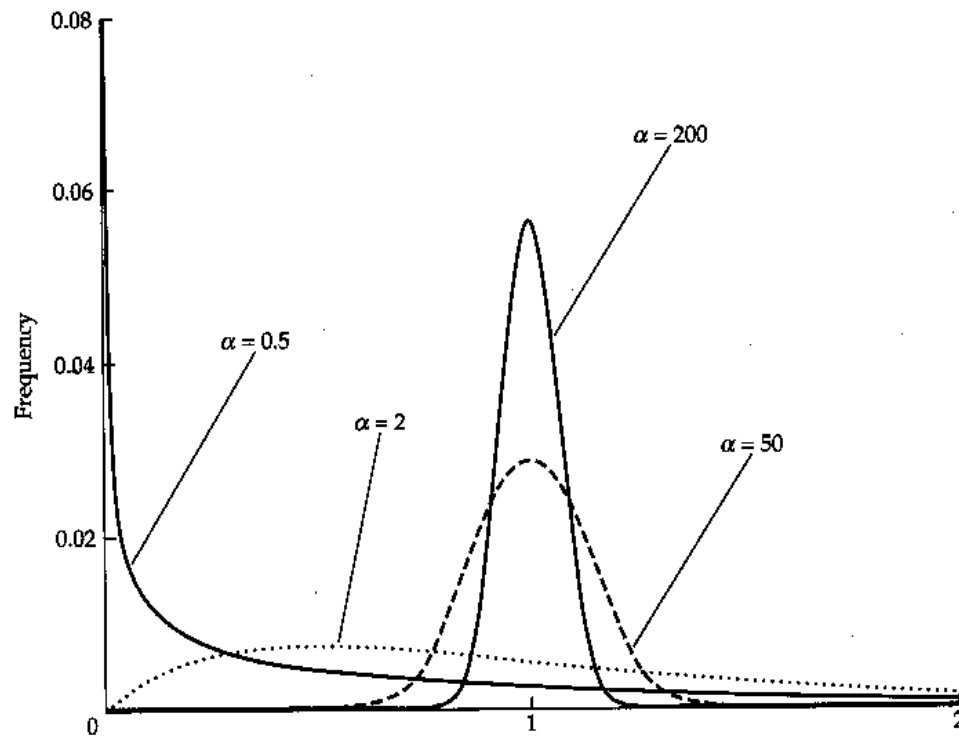


topology (t)
branch lengths (v_i)

$$Q = \begin{matrix} & \begin{matrix} - & \rho_C r_{AC} & \rho_G r_{AG} & \rho_T r_{AT} \end{matrix} \\ \begin{matrix} \rho_A r_{AC} \\ \rho_A r_{AG} \\ \rho_A r_{AT} \end{matrix} & \begin{matrix} - & \rho_G r_{CG} & \rho_T r_{CT} \\ \rho_C r_{CG} & - & \rho_T r_{GT} \\ \rho_C r_{CT} & \rho_G r_{GT} & - \end{matrix} \end{matrix}$$

General Time Reversible
substitution model

Adding Model Complexity



Gamma-shaped
rate variation
across sites

Priors on Parameters

- Stationary state frequencies
 - Flat Dirichlet, $\text{Dir}(1,1,1,1)$
- Exchangeability parameters
 - Flat Dirichlet, $\text{Dir}(1,1,1,1,1,1)$
- Shape parameter of scaled gamma distribution of rate variation across sites
 - Uniform $\text{Uni}(0,50)$

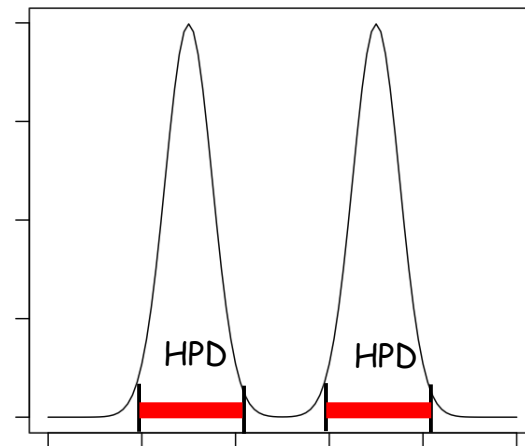
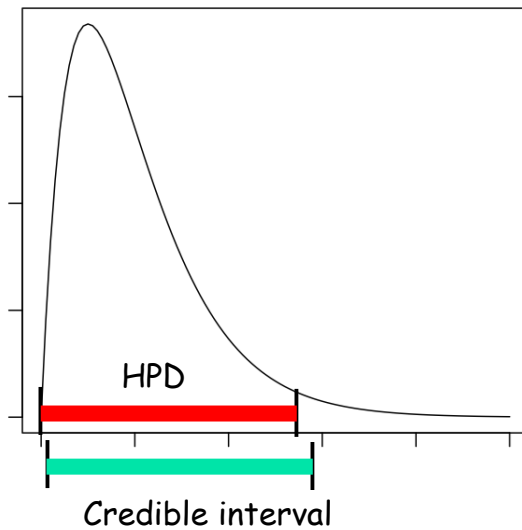
r_{AC}	1.35 (0.98, 1.82)
r_{AG}	3.24 (2.55, 4.06)
r_{AT}	1.64 (1.24, 2.11)
r_{CG}	1.18 (0.89, 1.56)
r_{CT}	5.93 (4.63, 7.54)
r_{GT}	1
α	0.32 (0.29, 0.35)
π_A	0.28 (0.26, 0.30)
π_C	0.20 (0.18, 0.22)
π_G	0.24 (0.22, 0.27)
π_T	0.28 (0.26, 0.30)

Mean and 95%
credibility
interval for model
parameters

Summarizing Variables

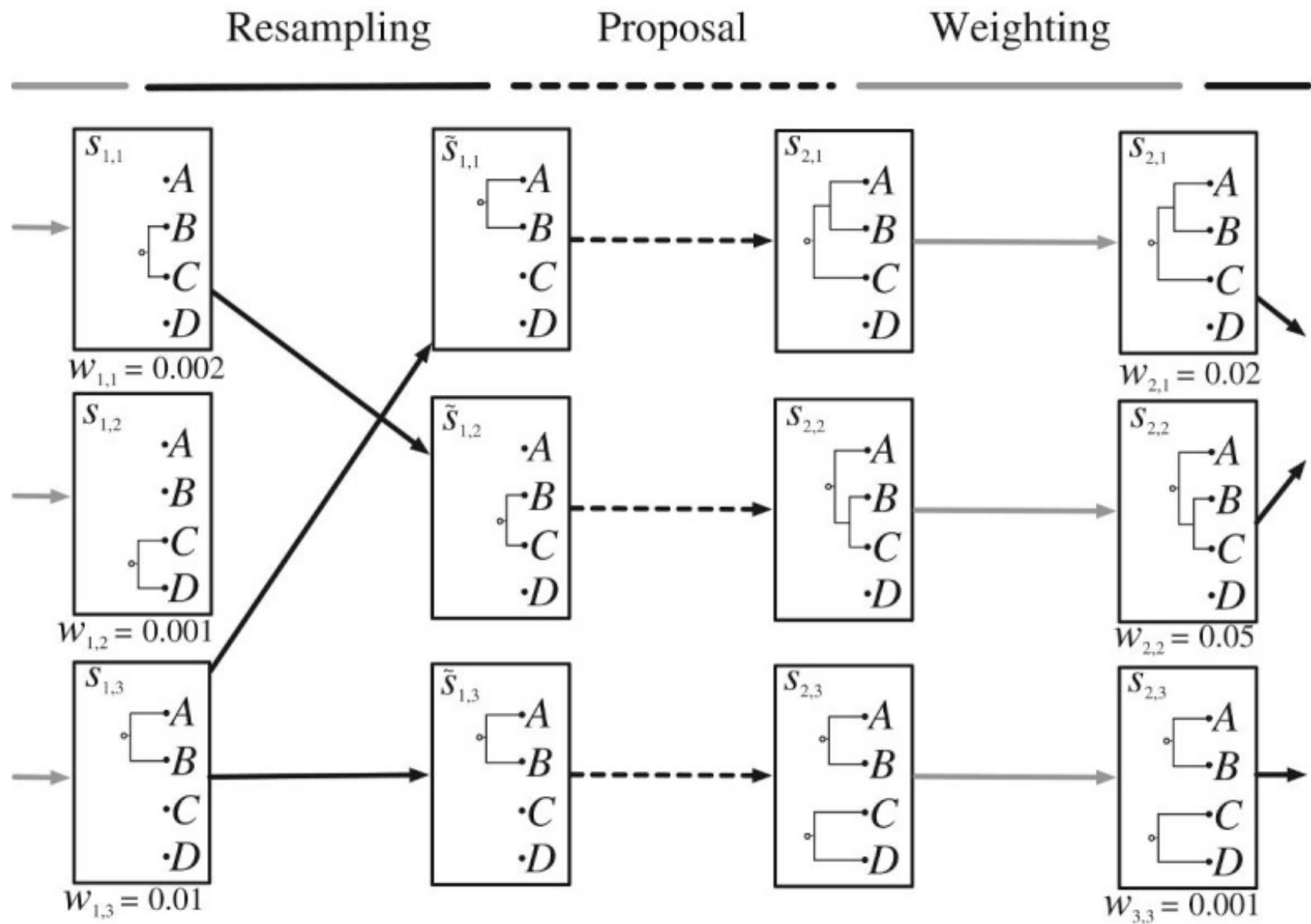
- Mean, median, variance common summaries
- 95 % credible interval: discard the lowest 2.5 % and highest 2.5 % of sampled values
- 95 % region of highest posterior density (HPD): find smallest region containing 95 % of probability

Credible intervals and HPDs



Other Sampling Methods

- **Gibbs sampling:** sample from the conditional posterior (a variant of the Metropolis algorithm)
- **Metropolized Gibbs sampling:** more efficient variant of Gibbs sampling of discrete characters
- **Slice sampling:** less prone to get stuck in local optima than the Metropolis algorithm
- **Hamiltonian sampling.** A technique for decreasing the problem with sampling correlated parameters.
- **Simulated annealing:** increase "greediness" during the burn-in phase of MCMC sampling
- **Data augmentation techniques:** add parameters to facilitate probability calculations
- **Sequential Monte Carlo techniques:** generate a sample of complete state by building sets of particles from incomplete states



Sequential Monte Carlo Algorithm for Phylogenetics

3. Markov chain Monte Carlo

Convergence and Mixing

- Convergence is the degree to which the chain has converged onto the target distribution
- Mixing is the speed with which the chain covers the region of interest in the target distribution

Assessing Convergence

- Plateau in the trace plot
- Look at sampling behavior within the run (autocorrelation time, effective sample size)
- Compare independent runs with different, randomly chosen starting points

Convergence within Run

Autocorrelation time $t = 1 + 2 \sum_{k=1}^{\infty} \rho_k(q)$

where $\rho_k(q)$ is the autocorrelation in the MCMC samples for a lag of k generations

Effective sample size (ESS) $e = \frac{n}{t}$

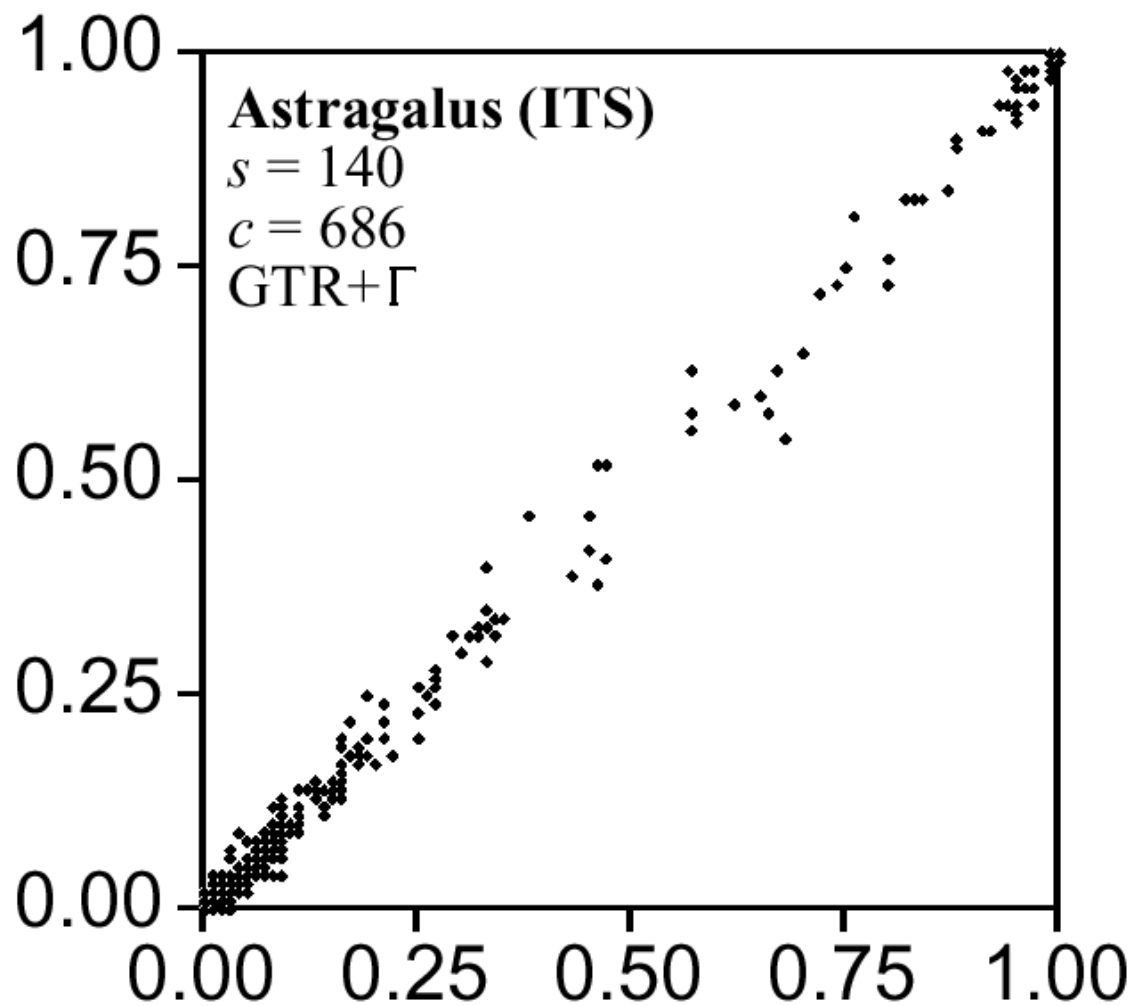
where n is the total sample size (number of generations)

Good mixing when t is small and e large

Convergence among Runs

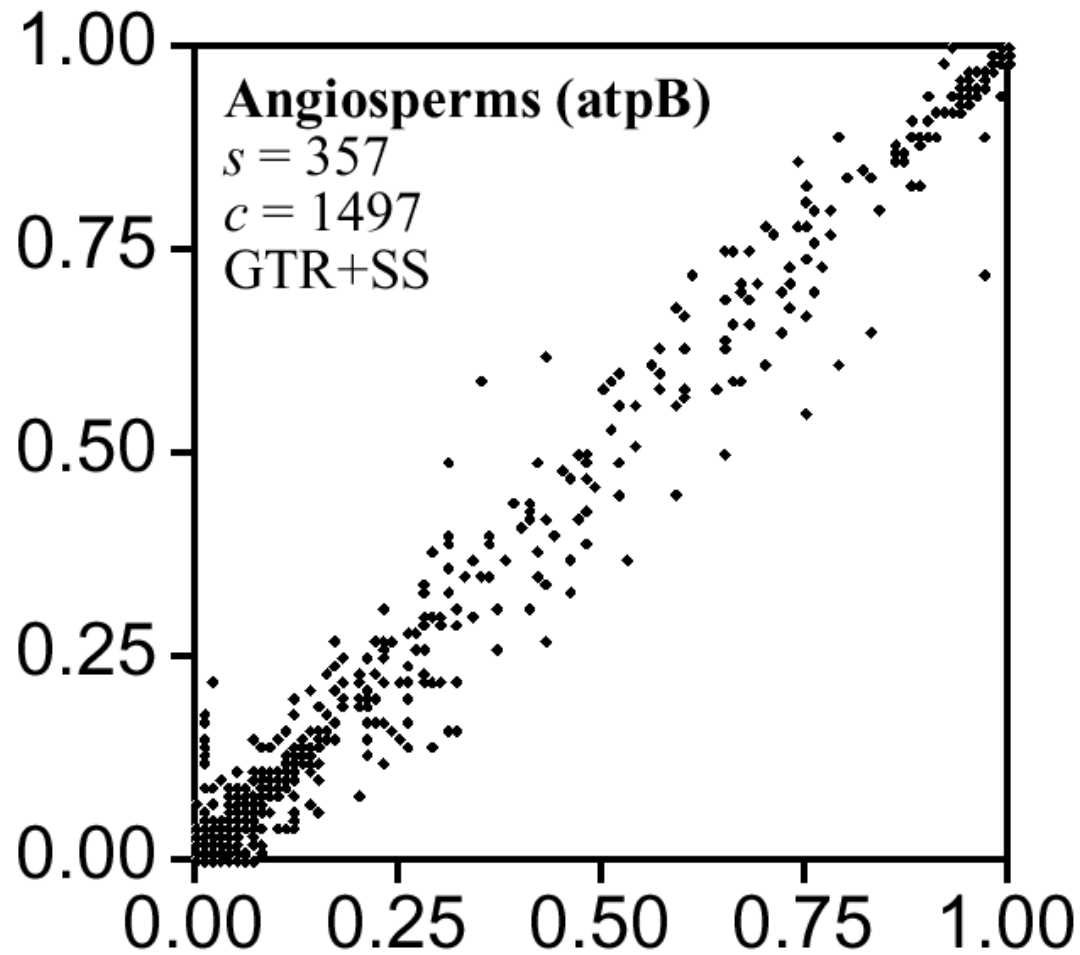
- Tree topology:
 - Compare clade probabilities (split frequencies)
 - Average standard deviation of split frequencies above some cut-off (min. 10 % in at least one run). Should go to 0 as runs converge.
- Continuous variables
 - Potential scale reduction factor (PSRF).
Compares variance within and between runs.
Should approach 1 as runs converge.
- Assumes overdispersed starting points

Clade probability in analysis 2



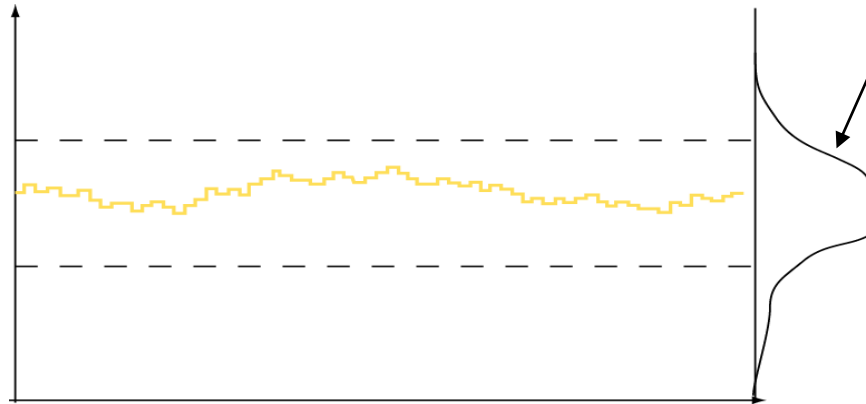
Clade probability in analysis 1

Clade probability in analysis 2



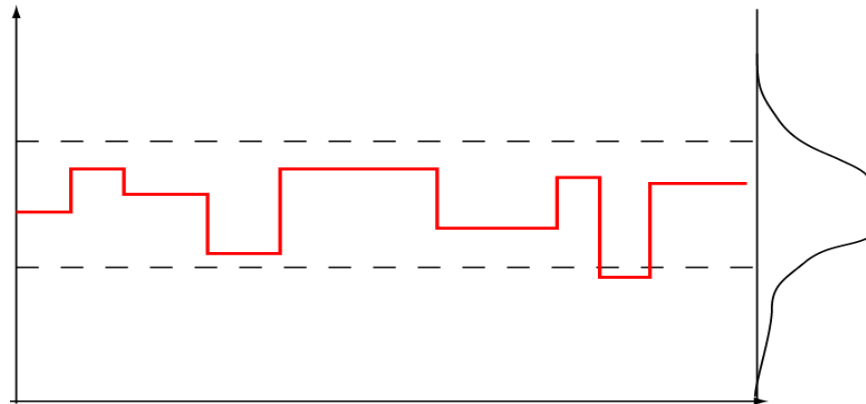
Clade probability in analysis 1

Sampled value

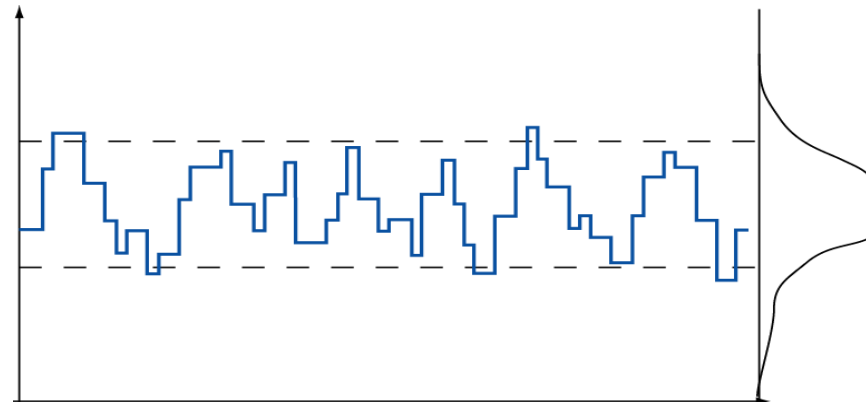


Target distribution

Too modest proposals
Acceptance rate too high
Poor mixing



Too bold proposals
Acceptance rate too low
Poor mixing



Moderately bold proposals
Acceptance rate intermediate
Good mixing

Tuning Proposals

- Manually by changing tuning parameters
 - Increase the boldness of a proposal if acceptance rate is too high
 - Decrease the boldness of a proposal if acceptance rate is too low
- Auto-tuning
 - Tuning parameters are adjusted automatically by the MCMC procedure to reach a target proposal rate

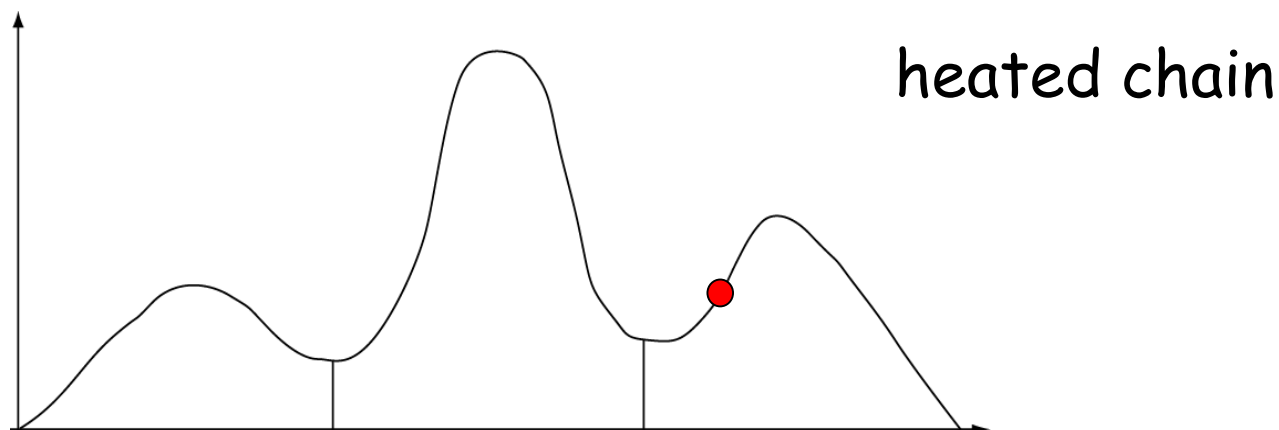
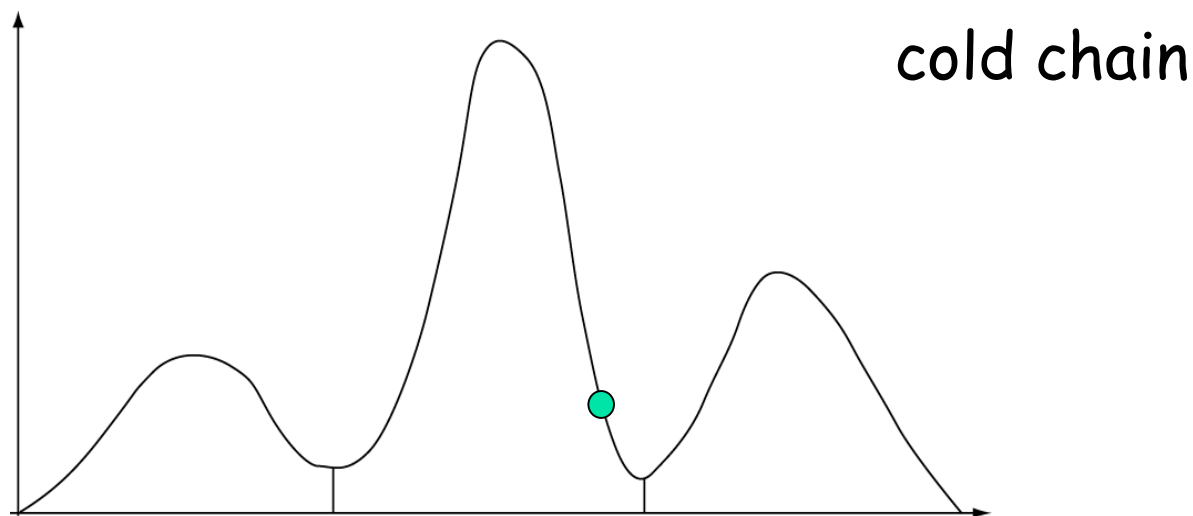
Metropolis-
coupled
Markov chain
Monte Carlo

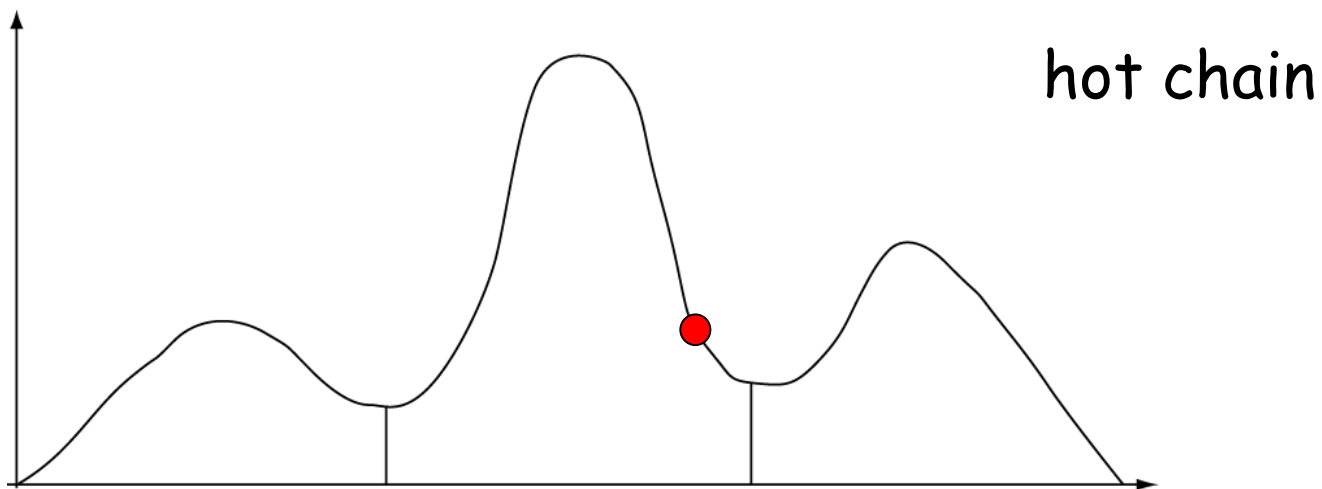
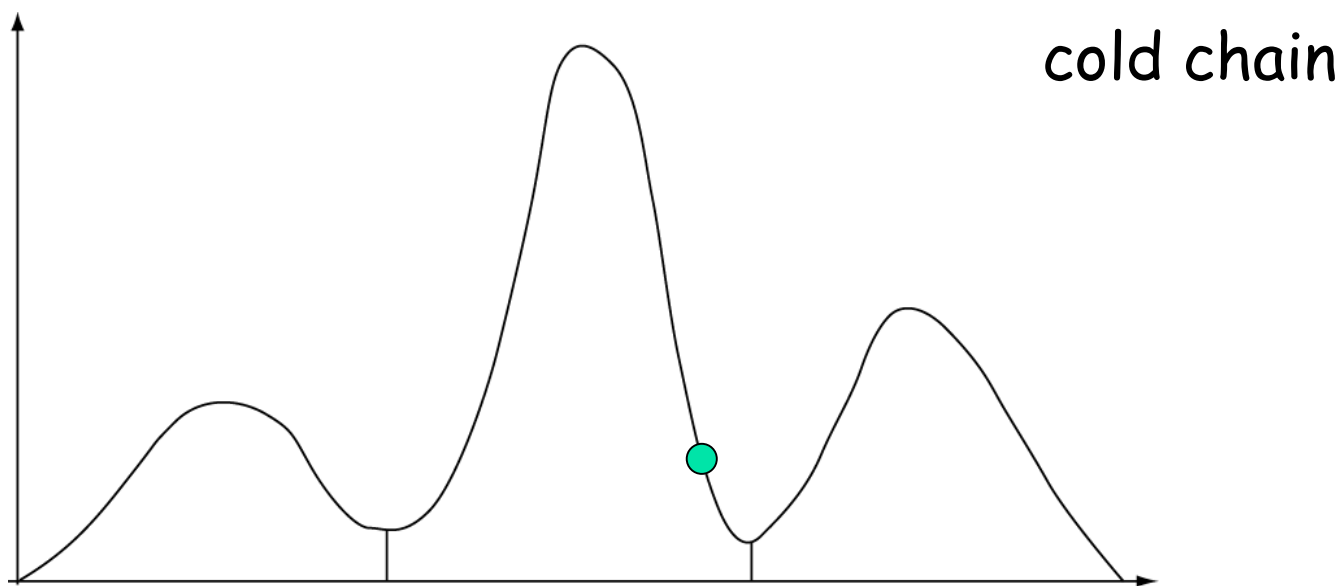
a. k. a.

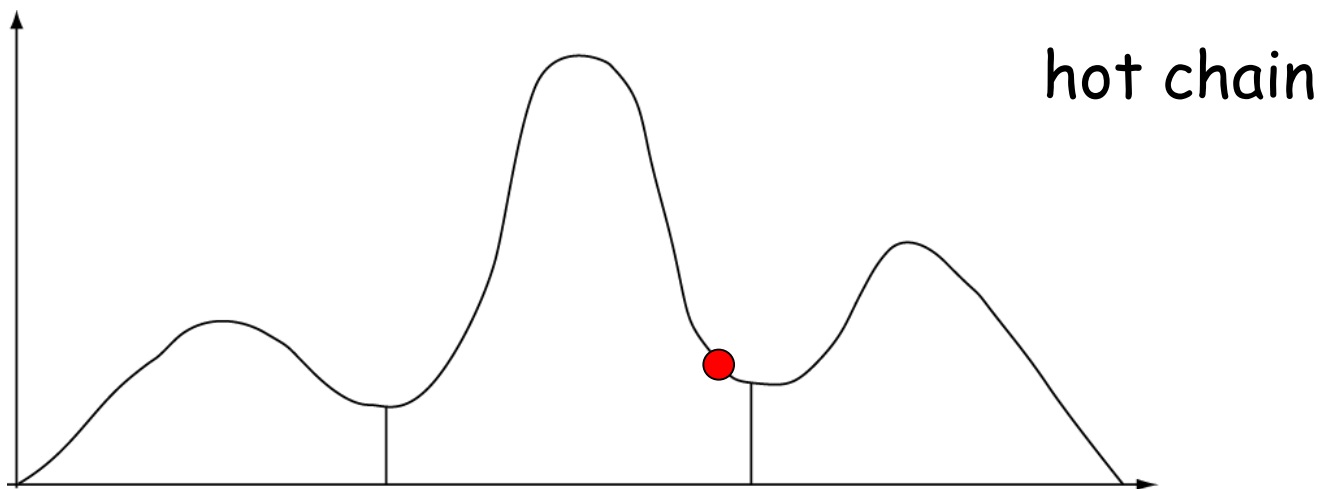
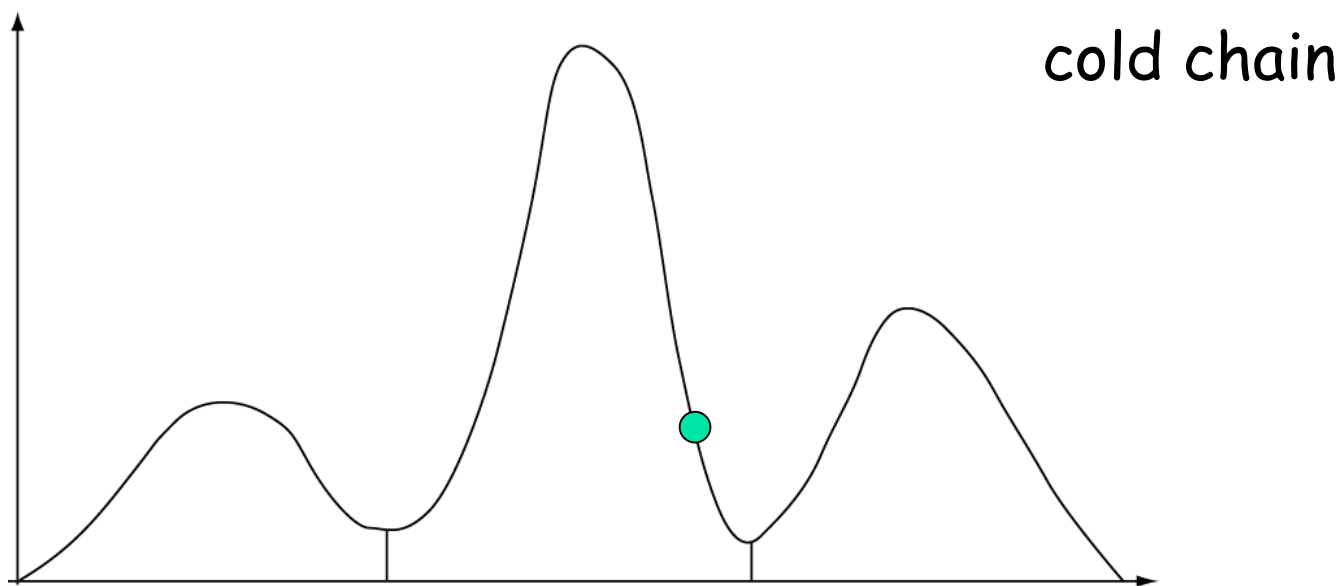
MCMCMC

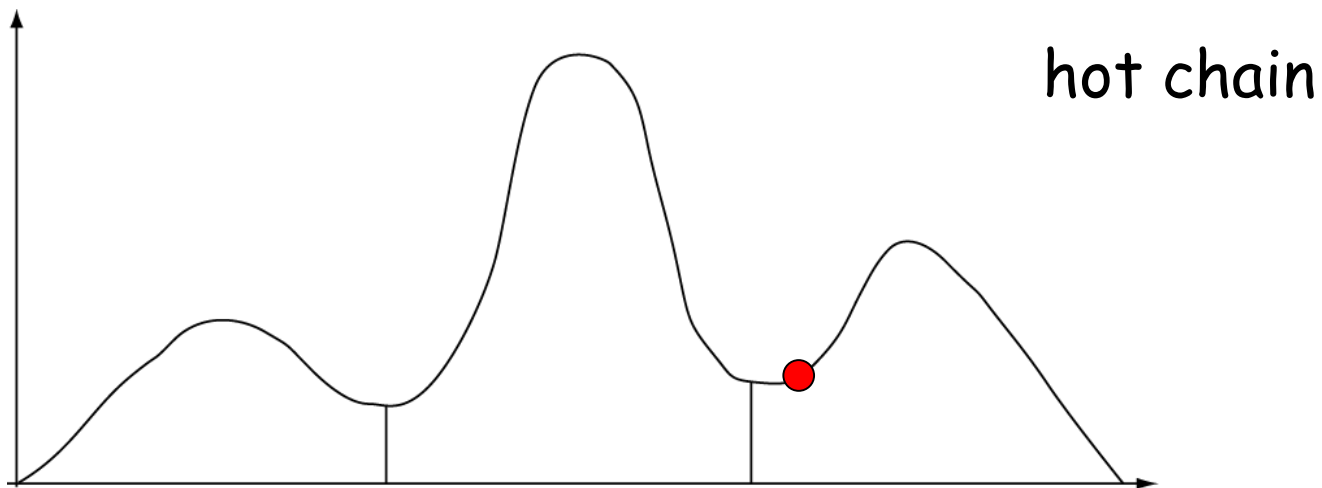
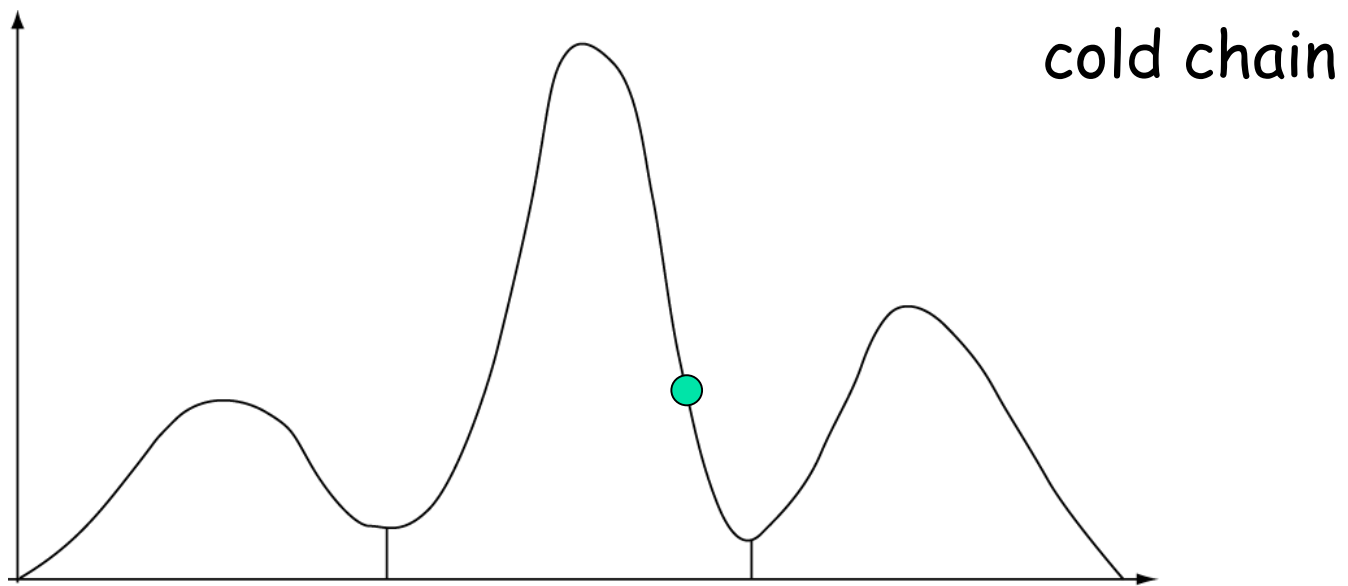
a. k. a.

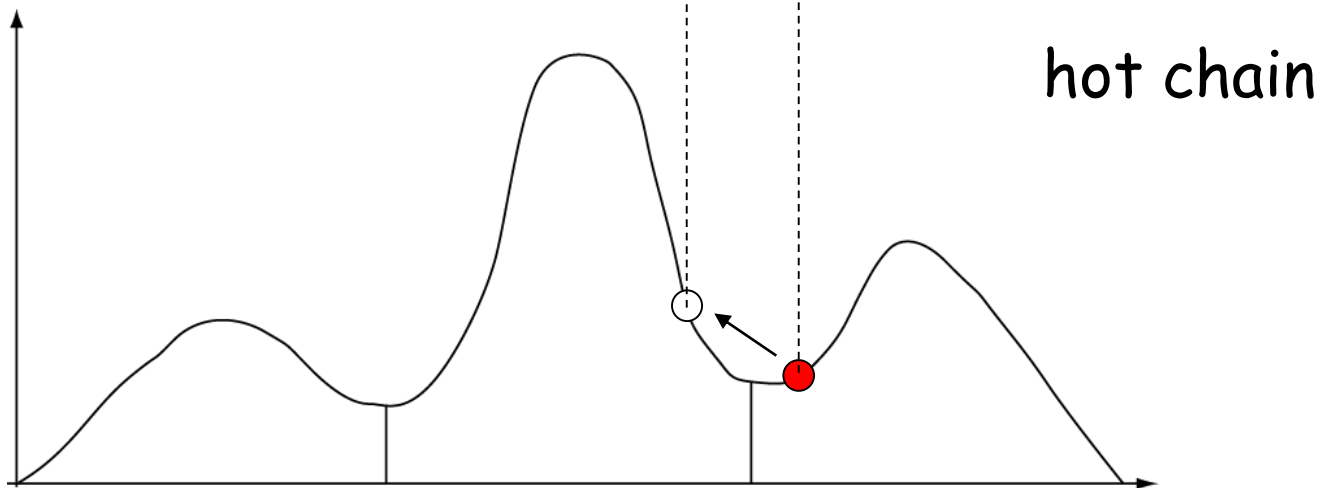
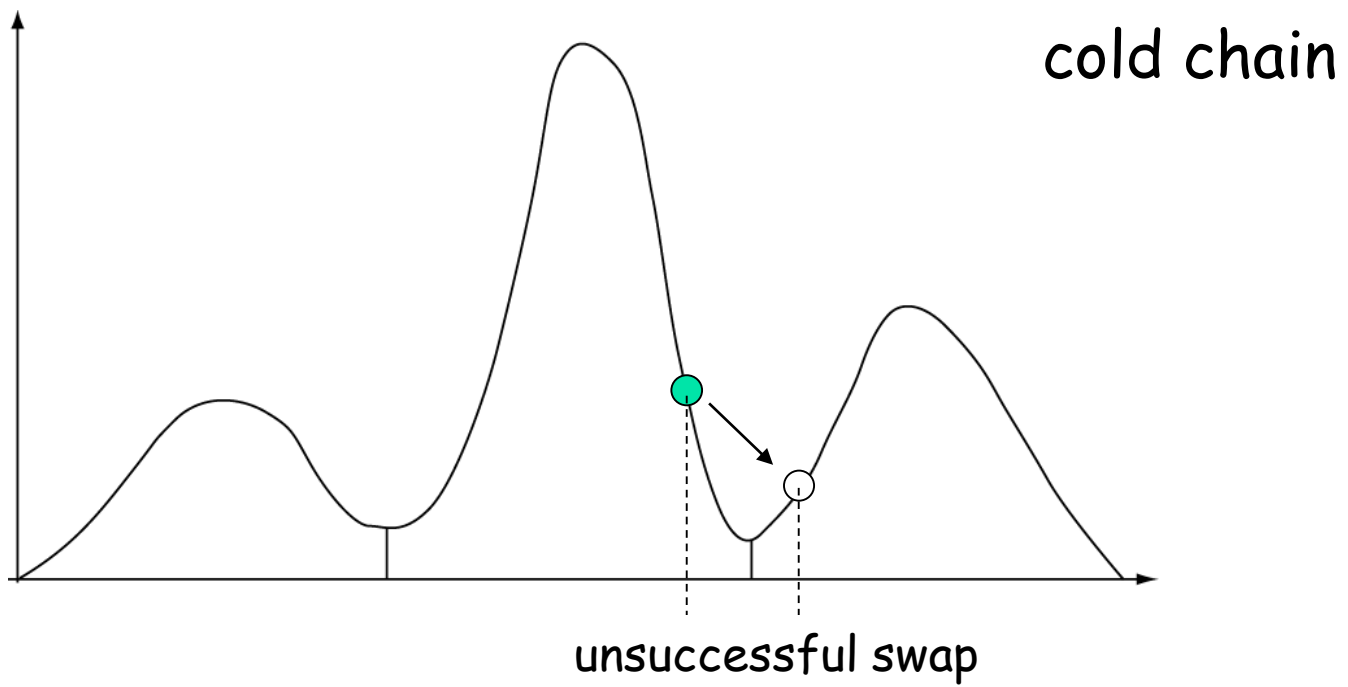
(MC)³

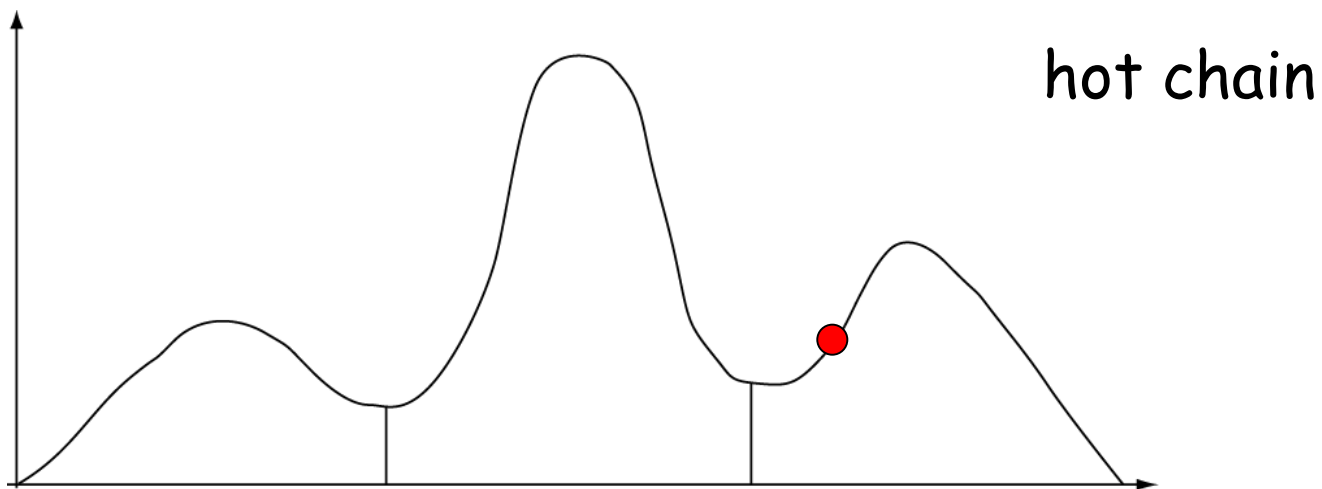
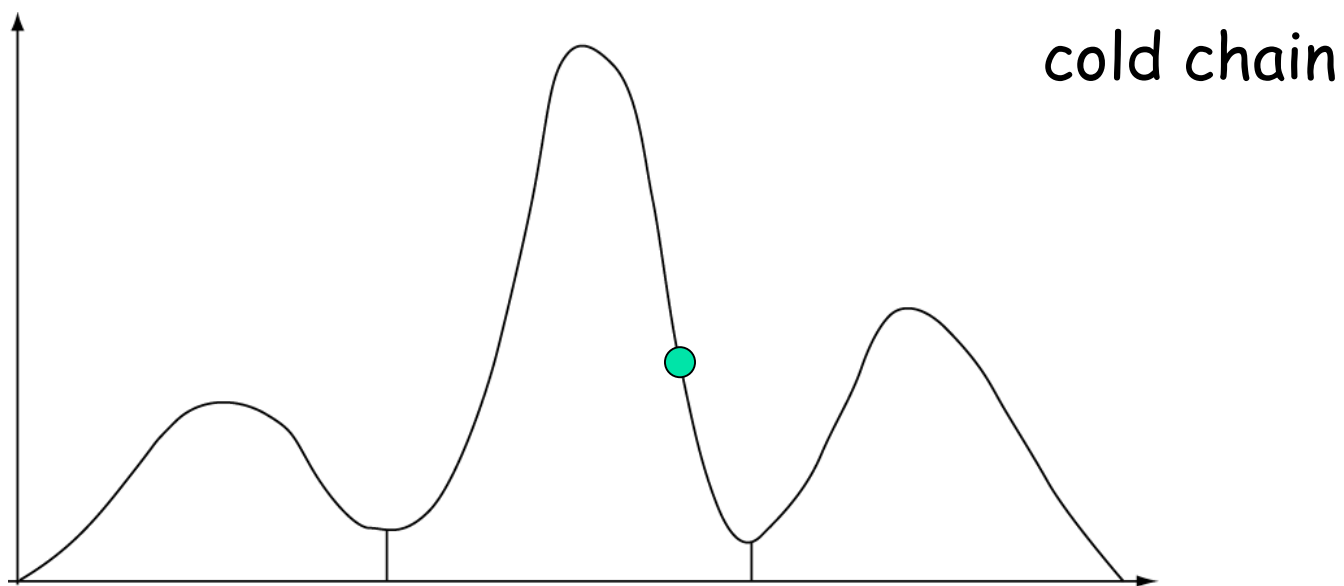


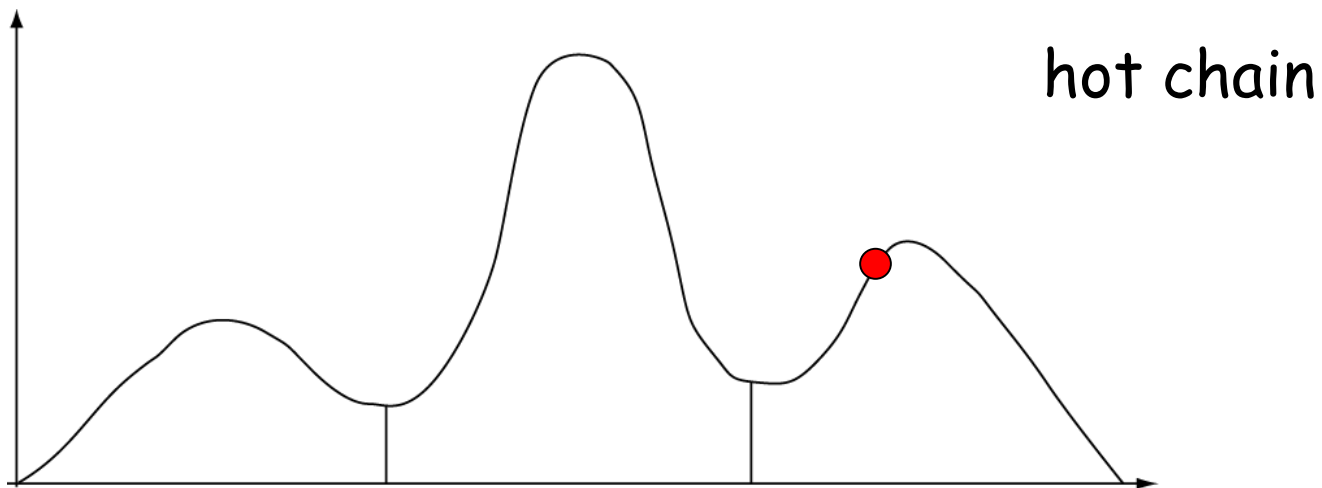
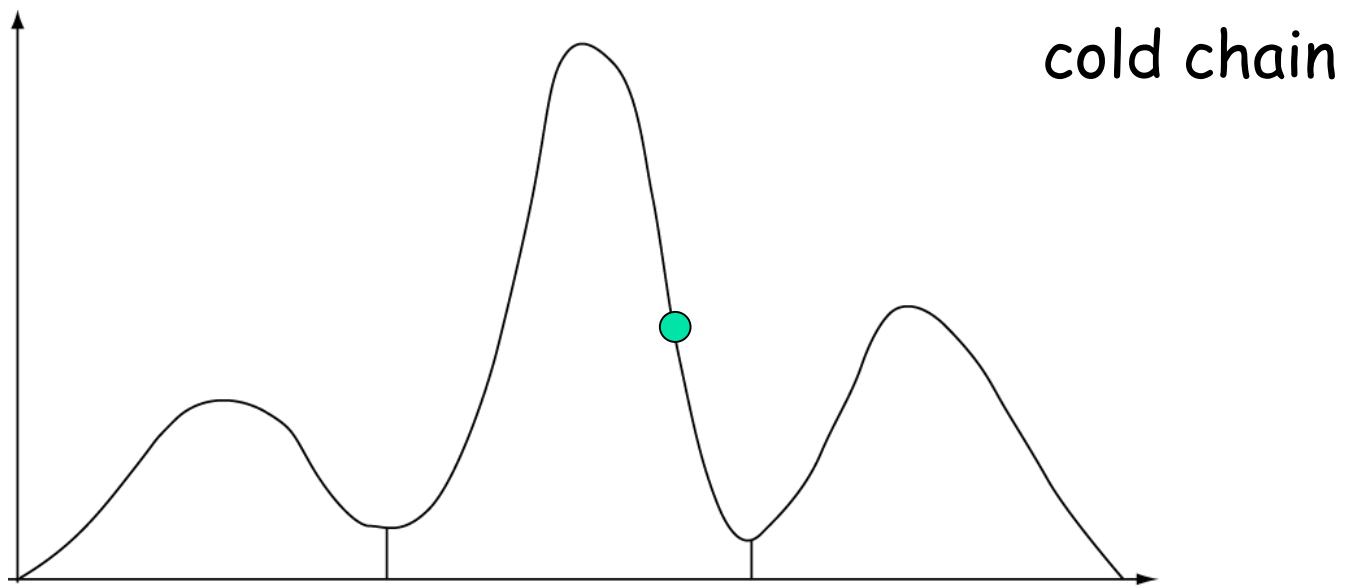


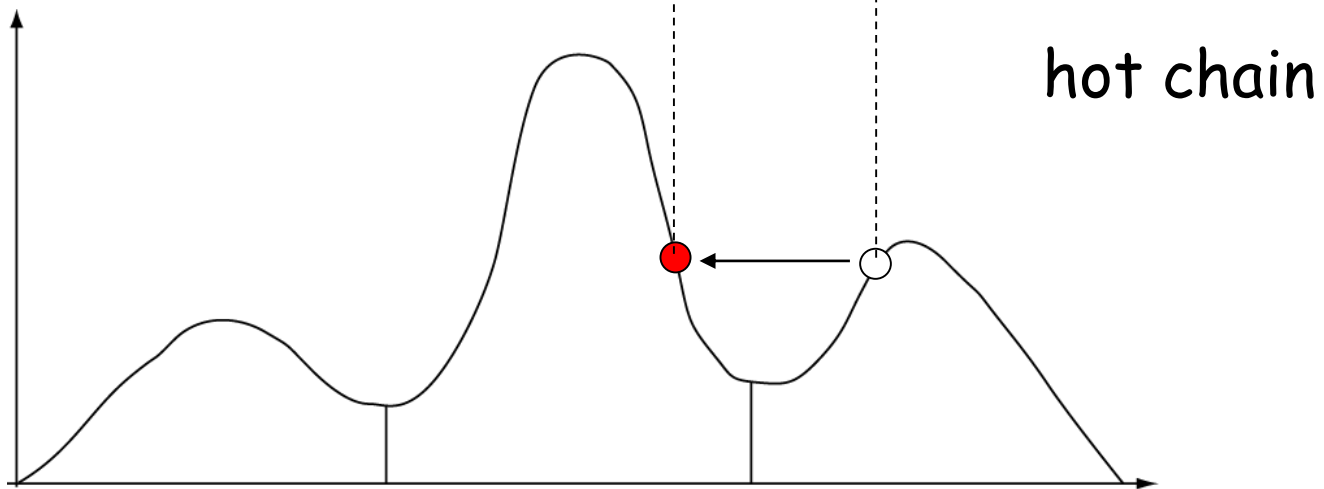
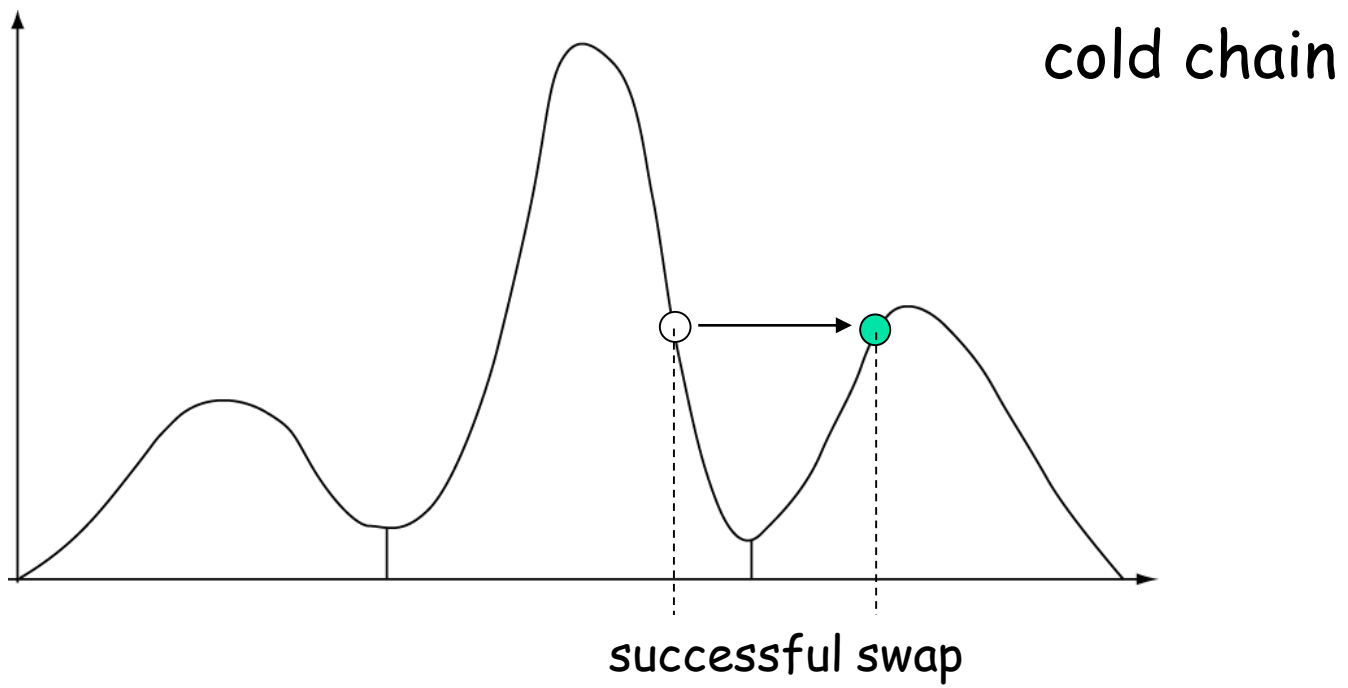


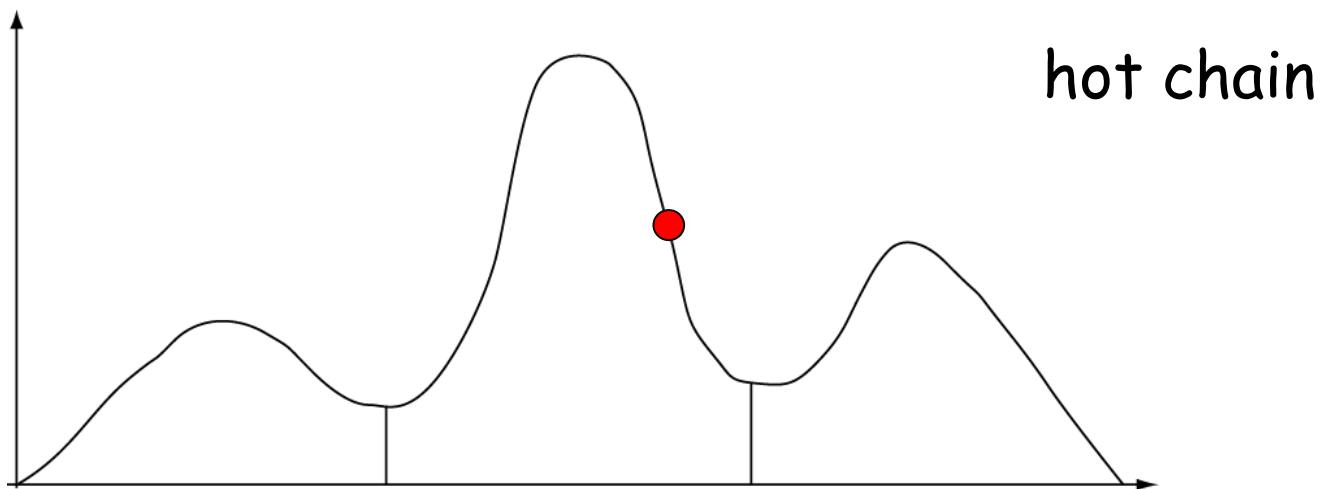
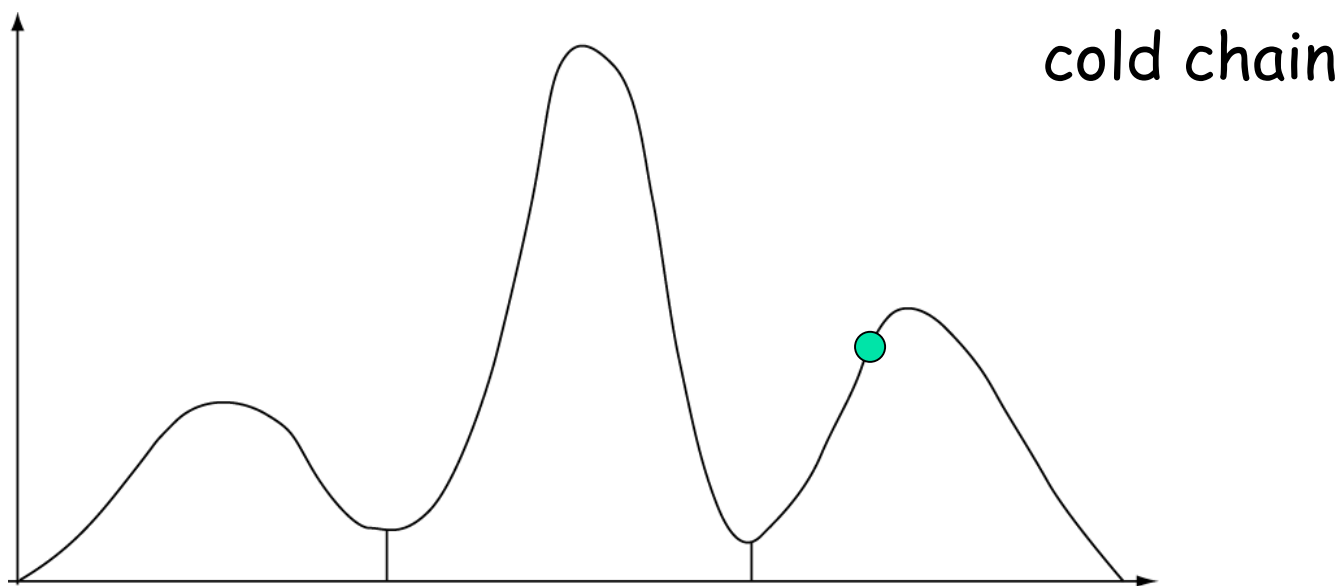


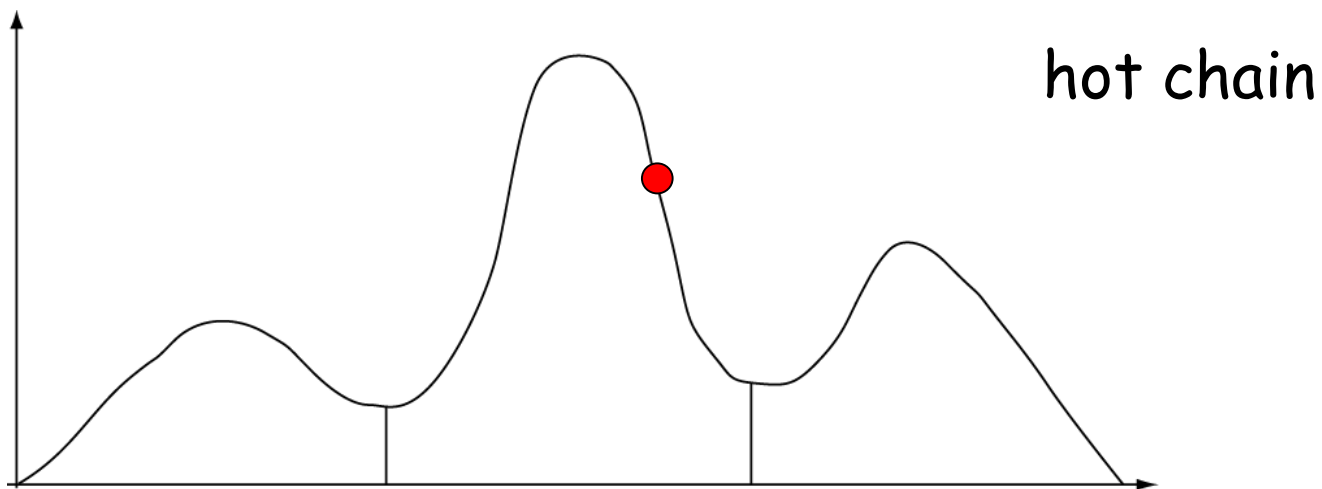
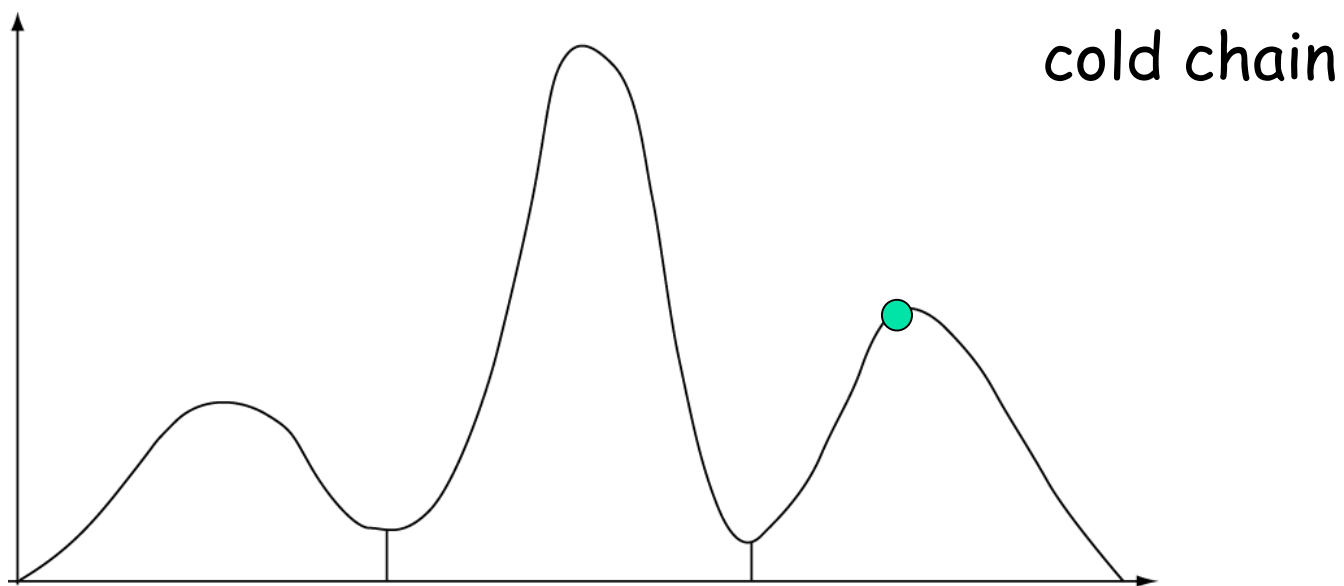


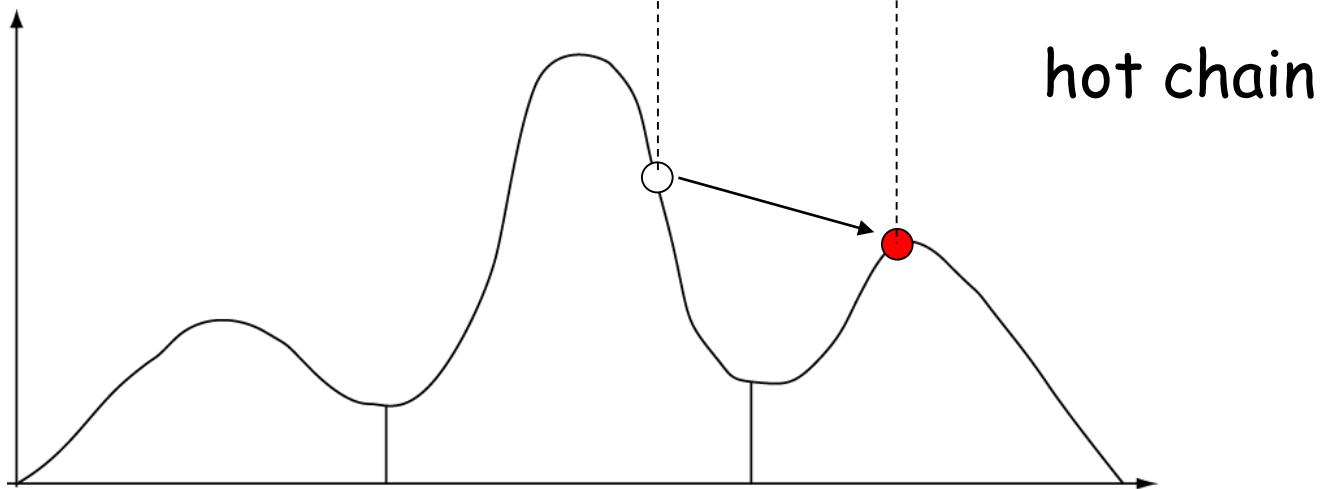
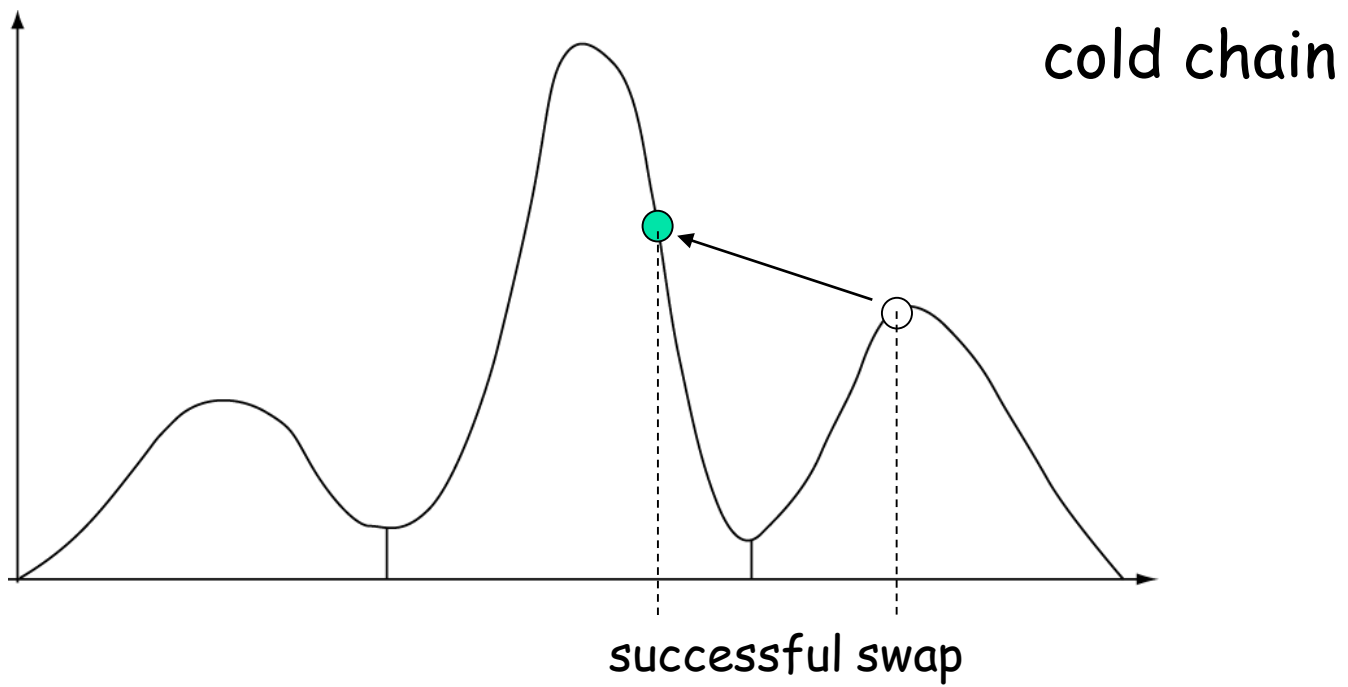


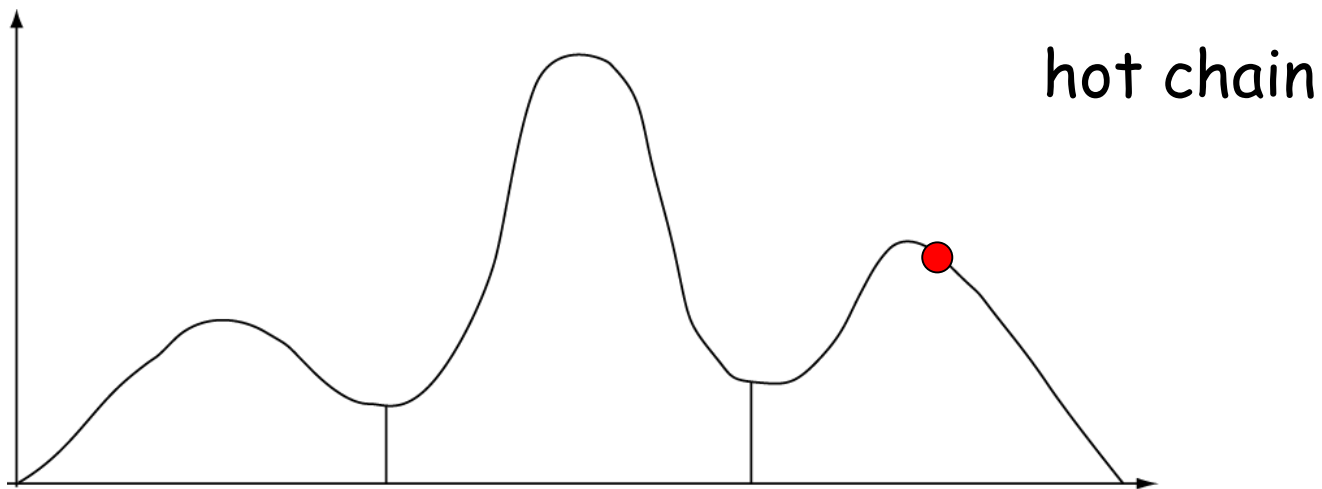
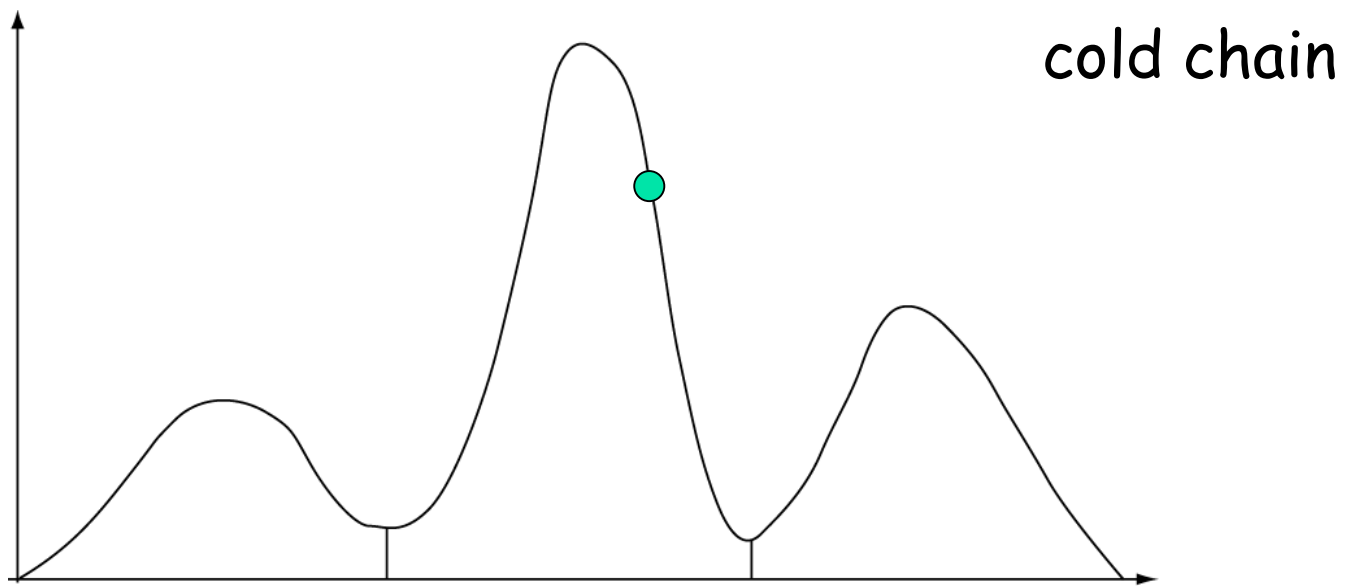












Incremental Heating

T is temperature, λ is heating coefficient

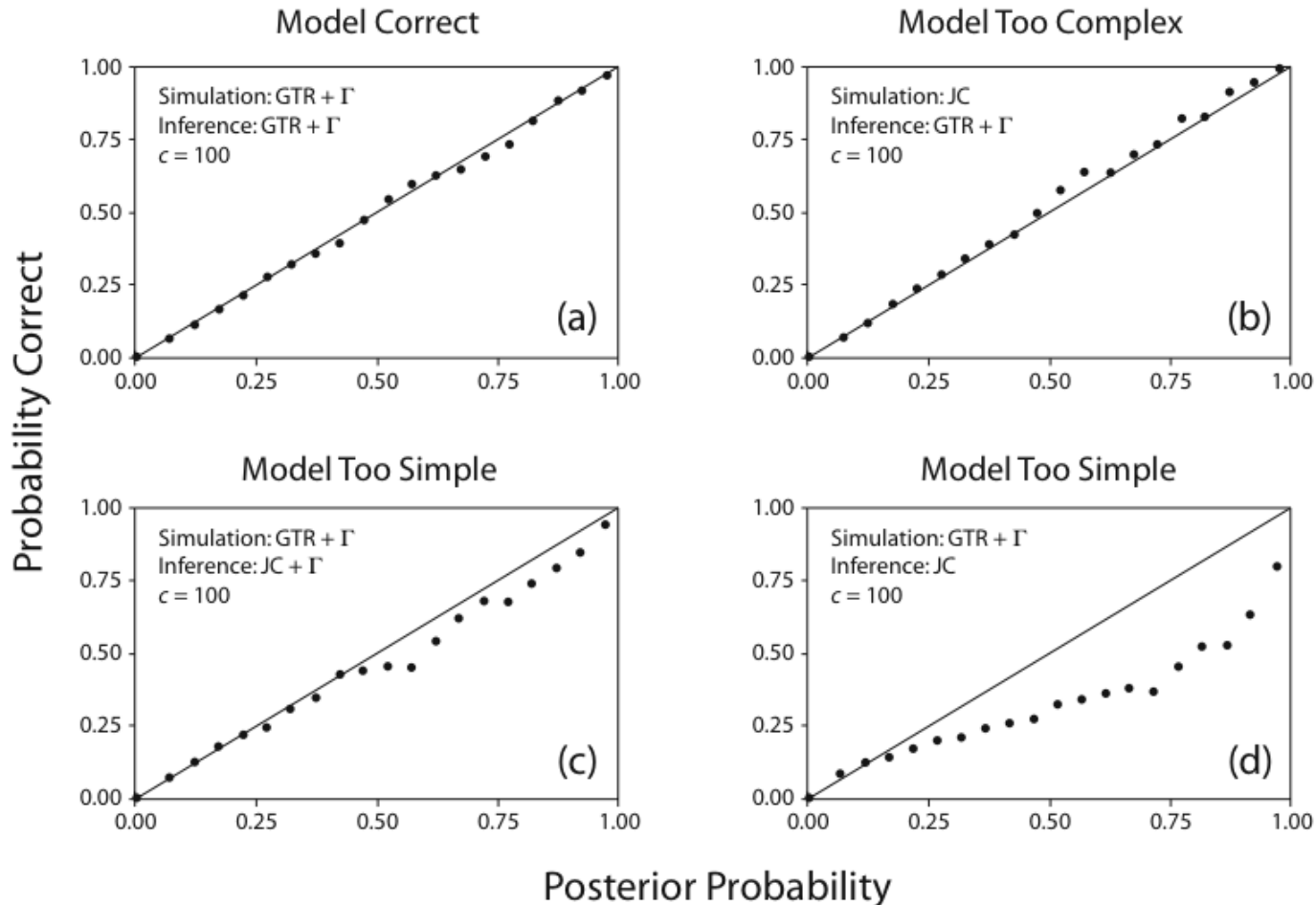
$$T = 1/(1 + \lambda i) \quad i = \{0, 1, \dots, n - 1\}$$

Example for $\lambda = 0.2$:

i	T	Distr.	
0	1.00	$f(q X)^{1.00}$	← cold chain
1	0.83	$f(q X)^{0.83}$	↙ heated chains ← ↘
2	0.71	$f(q X)^{0.71}$	
3	0.62	$f(q X)^{0.62}$	

4. Bayesian Model Choice

Bayesian Model Sensitivity



Models, models, models

- Alignment-free models
- Heterogeneity in substitution rates and stationary frequencies across sites and lineages
- Relaxed clock models
- Models for morphology and biogeography
- Sampling across model space, e.g. GTR space and partition space
- Models of dependence across sites according to 3D structure of proteins
- Positive selection models
- Aminoacid models
- Models for population genetics and phylogeography

Bayes' Rule

$$f(q|D) = \frac{f(q)f(D|q)}{\int f(q)f(D|q) \, dq} = \frac{f(q)f(D|q)}{f(D)}$$

Marginal likelihood (of the data)



We have implicitly conditioned on a model:

$$f(q|D, M) = \frac{f(q|M)f(D|q, M)}{f(D|M)}$$

Bayesian Model Choice

Posterior model odds:

$$\frac{f(M_1)f(D|M_1)}{f(M_0)f(D|M_0)}$$

Bayes factor:

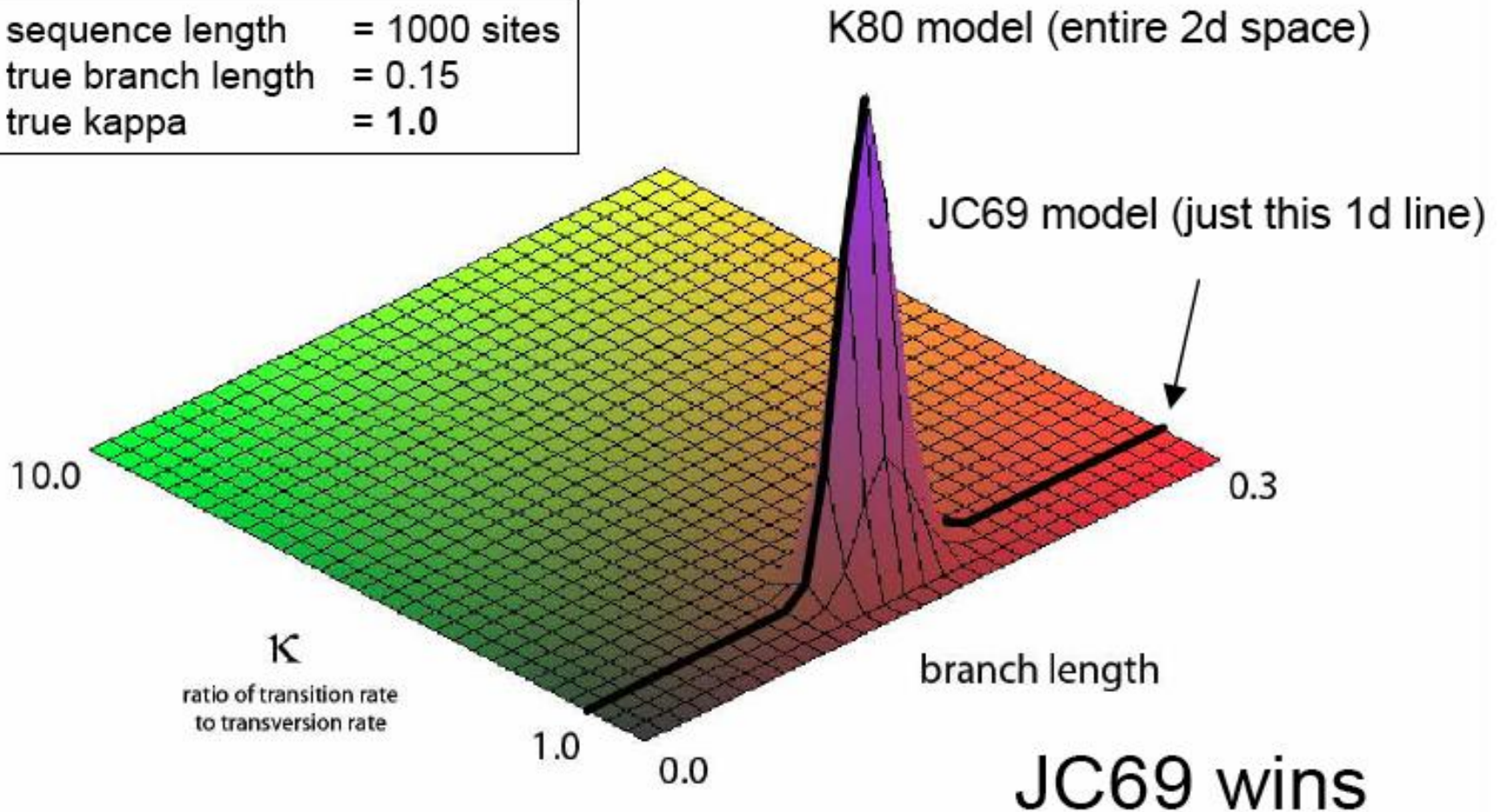
$$B_{10} = \frac{f(D|M_1)}{f(D|M_0)}$$

Bayesian Model Choice

- The normalizing constant in Bayes' theorem, the marginal likelihood of the data, $f(D)$ or $f(D|M)$, can be used for model choice
- $f(D|M)$ can be estimated by taking the harmonic mean of the likelihood values from the MCMC run. Thermodynamic integration and stepping-stone sampling are computationally more complex but more accurate methods
- Any models can be compared: nested, non-nested, data-derived; it is just a probability comparison
- No correction for number of parameters
- Can prefer a simpler model over a more complex model
- Critical values in Kass and Raftery (1997)

Simple Model Wins (from Lewis, 2008)

sequence length	= 1000 sites
true branch length	= 0.15
true kappa	= 1.0



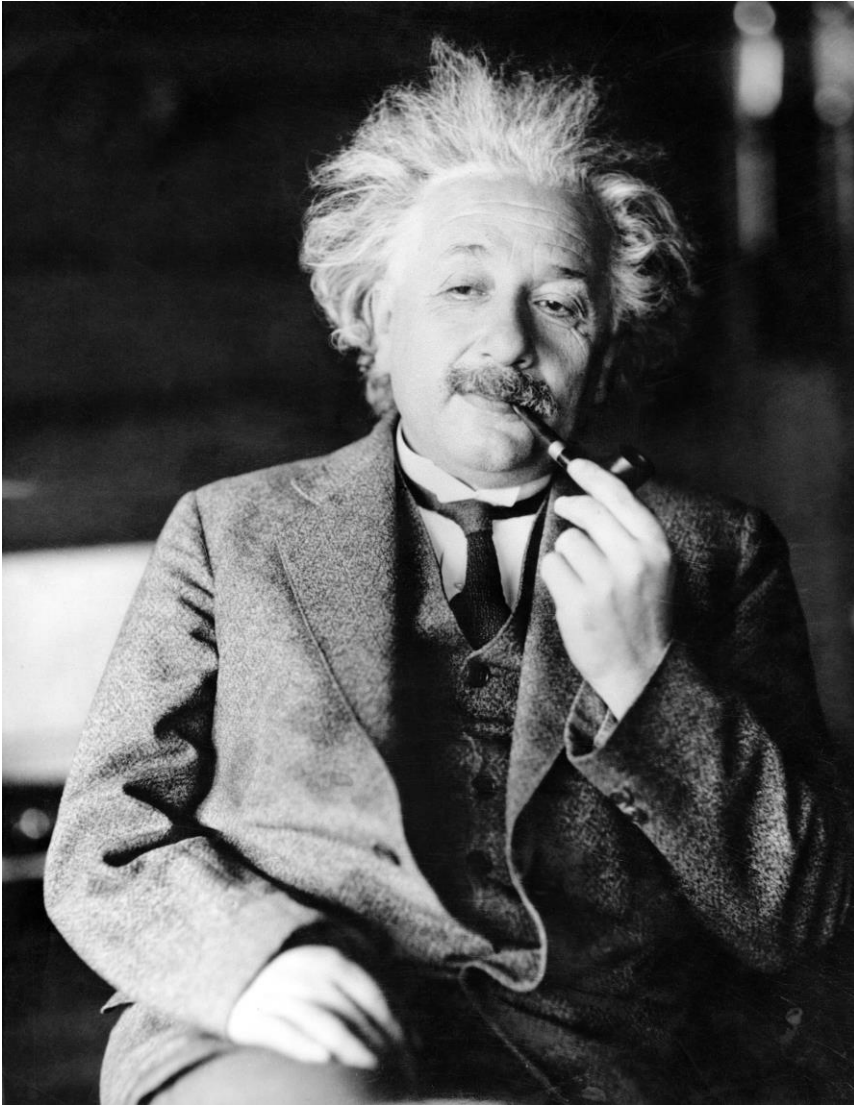
Bayes Factor Comparisons

Interpretation of the Bayes factor

$2\ln(B_{10})$	B_{10}	Evidence against M_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

Bayesian Software

- Model testing
 - ModelTest
 - MrModelTest
 - MrAIC
- Convergence diagnostics
 - AWTY
 - Tracer
- Phylogenetic inference
 - MrBayes
 - BEAST
 - BayesPhylogenies
 - PhyloBayes
 - Phycas
 - BAMBE
 - RevBayes
- Specialized inference
 - PHASE
 - BALiPhy
 - BayesTraits
 - Badger
 - BEST
 - *BEAST
 - CoEvol
- Tree drawing
 - TreeView
 - FigTree



Listening to lectures,
after a certain age,
diverts the mind too much
from its creative pursuits.
Any scientist who attends
too many lectures and
uses her own brain too
little falls into lazy habits
of thinking.

after Albert Einstein