Molecular Phylogenetics Course

Statistical frameworks for modelling in phylogenetics

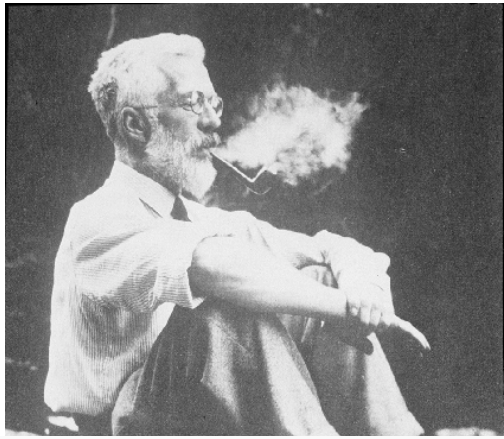
Jadranka Rota

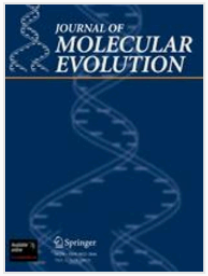
Some slides from Nicolas Chazot (Lund University, Sweden) and Paul Lewis (University of Connecticut, USA)

Maximum likelihood

Maximum Likelihood Estimation (MLE)

- **Statistical method for estimating parameters of a model (e.g. mean and variance of a normal distribution)**





Journal of Molecular Evolution
November 1981, Volume 17, Issue 6, pp 368–376 | Cite as

Evolutionary trees from DNA sequences: A maximum likelihood approach

Authors | Authors and affiliations

Joseph Felsenstein

Article

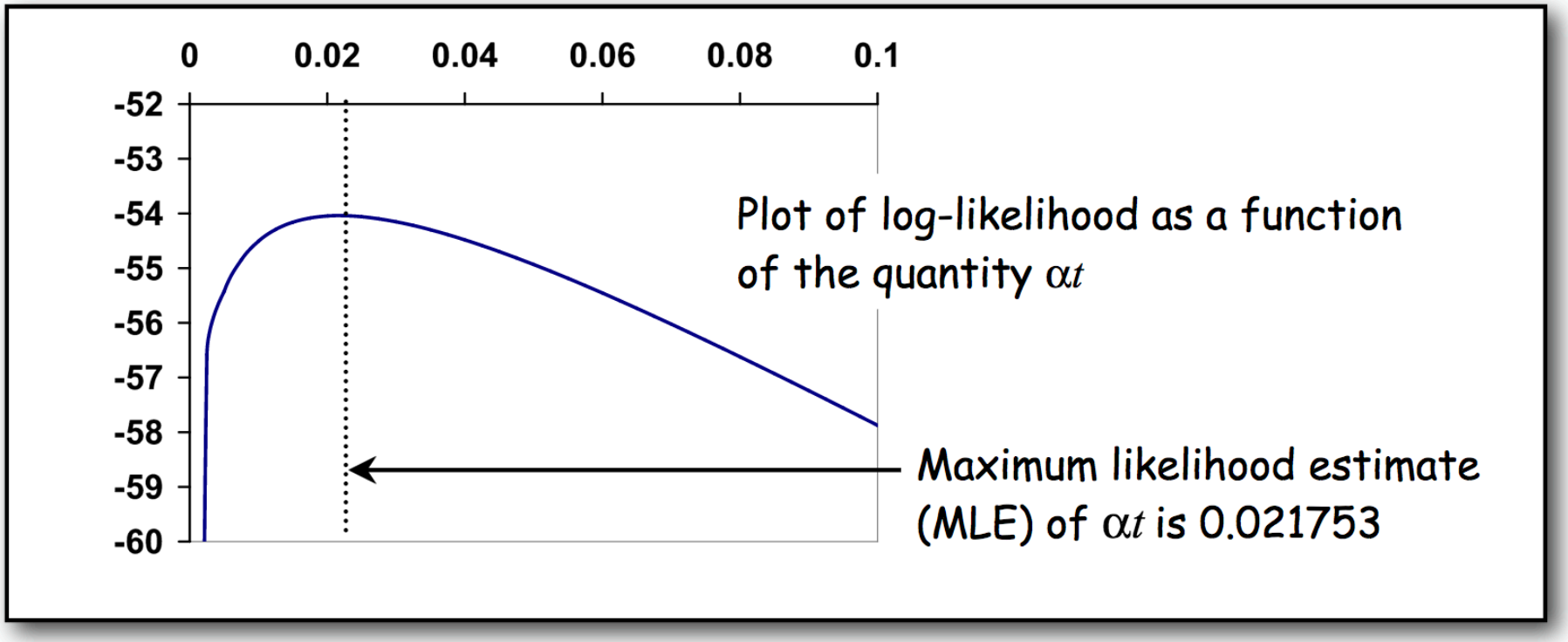
Likelihood of a hypothesis

- **Likelihood (L) is proportional to the probability (P) of observing the data (D) given a model (M) – *conditional probability***

  – $L(M) = Pr(D \mid M)$

- **We can examine this likelihood function to find where it is highest and identify the parameters of the model at this point -> Maximum Likelihood Estimates**

Likelihood function



Likelihood of a hypothesis

- **Likelihood (L) is proportional to the probability (P) of observing the data (D) given a model (M) – *conditional probability***

  – $L(M) = Pr(D \mid M)$

- **We can examine this likelihood function to find where it is highest and identify the parameters of the model at this point -> Maximum Likelihood Estimates**
- **In molecular phylogenetics, likelihood is the probability of observing the sequences given our model (e.g. GTR+G and our tree topology including the branch lengths)**

Maximum Likelihood

- **For reconstructing phylogenies**

Model

  - **which tree topology (τ), branch lengths, and parameters of DNA evolution models (θ) (e.g. transition/transversion ratio, base frequencies, …) are maximizing the probability of observing the sequences at hand?**

Data

$$L(\tau,\theta) = Pr(Data \mid \tau,\theta)$$

Likelihood of a single sequence

First 32 nucleotides of the $\psi\eta$-globin gene of gorilla:

GAAGTCCTTGAGAAATAAACTGCACACACTGG

$$L = \pi_G \pi_A \pi_A \pi_G \pi_T \pi_C \pi_C \pi_T \pi_T \pi_G \pi_A \pi_G \pi_A \pi_A \pi_A \pi_T \pi_A \pi_A \pi_A \pi_C \pi_T \pi_G \pi_C \pi_A \pi_C \pi_A \pi_C \pi_A \pi_C \pi_T \pi_G \pi_G$$
$$= \pi_A^{12} \pi_C^{7} \pi_G^{7} \pi_T^{6}$$

$$\ln L = 12 \ln(\pi_A) + 7 \ln(\pi_C) + 7 \ln(\pi_G) + 6 \ln(\pi_T)$$

We can already see by eye-balling this that the F81 model will fit better than the JC69 model because there are about twice as many As as there are Cs, Gs and Ts.

Likelihood ratio test

Find lnL under F81 model:

$$\begin{aligned}\ln L &= 12\ln(\pi_A) + 7\ln(\pi_C) + 7\ln(\pi_G) + 6\ln(\pi_T) \\ &= 12\ln(0.375) + 7\ln(0.21875) + 7\ln(0.21875) + 6\ln(0.1875) \\ &= -43.1\end{aligned}$$

Find lnL under JC69 model:

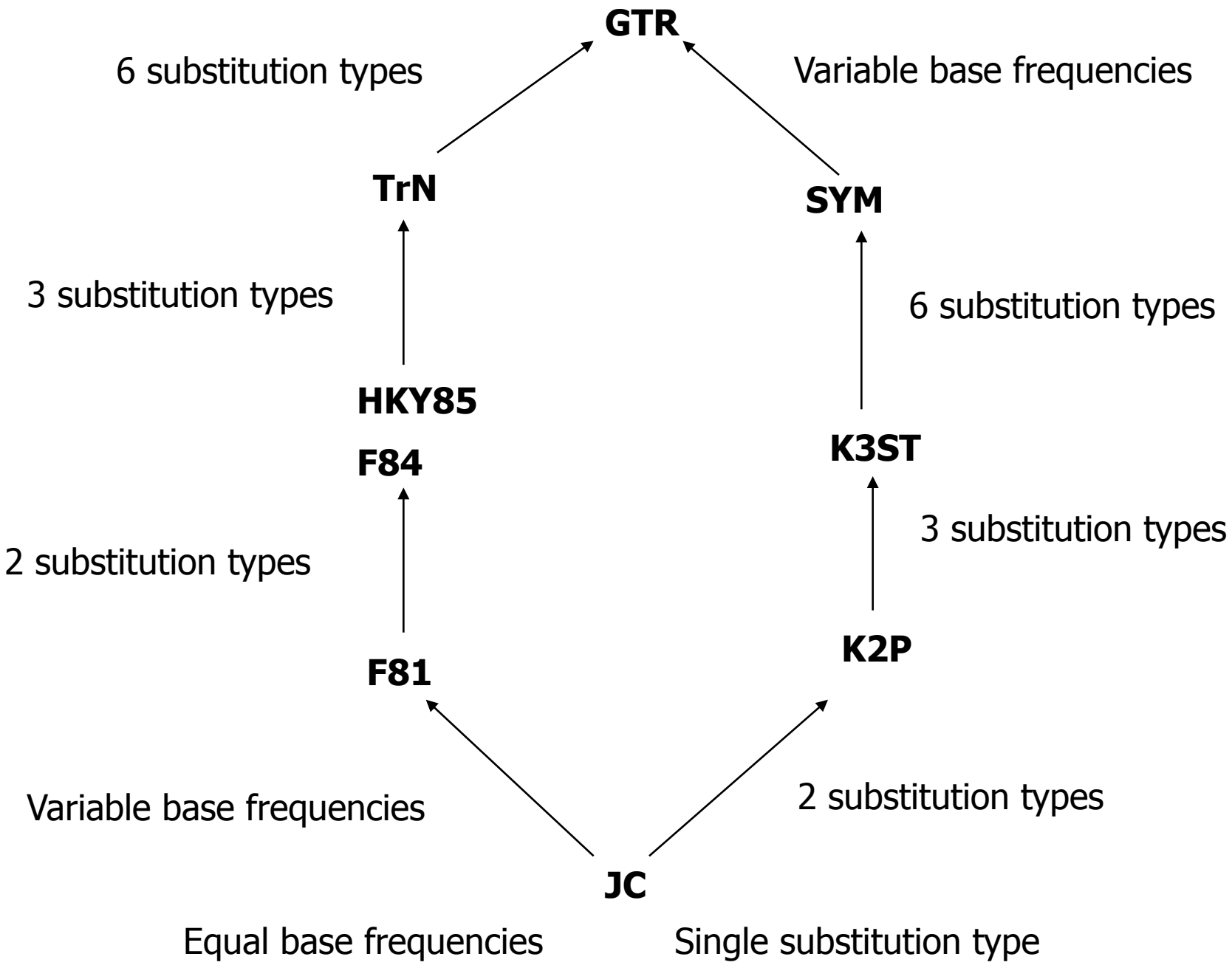
$$\begin{aligned}\ln L &= 12\ln(\pi_A) + 7\ln(\pi_C) + 7\ln(\pi_G) + 6\ln(\pi_T) \\ &= 12\ln(0.25) + 7\ln(0.25) + 7\ln(0.25) + 6\ln(0.25) \\ &= -44.4\end{aligned}$$

F81 does fit better (-43.1 > -44.4), but not significantly better (P = 0.457, chi-squared with 3 d.f.*)

Find likelihood ratio test statistic:

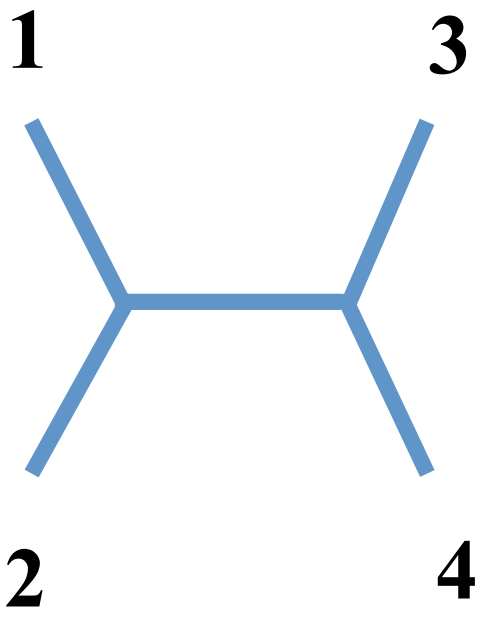
$$LR = -2(\ln L_{JC69} - \ln L_{F81}) = -2[-44.4 - (-43.1)] = 2.6$$

*The number of degrees of freedom equals the difference between the two models in the number of parameters. In this case, F81 has 3 parameters and JC69 has 0, so d.f. = 3 - 0 = 3


GTR
6 substitution types
Variable base frequencies
TrN
SYM
3 substitution types
6 substitution types
HKY85
K3ST
F84
3 substitution types
2 substitution types
K2P
F81
Variable base frequencies
2 substitution types
JC
Equal base frequencies
Single substitution type

Maximum Likelihood tree reconstruction

```
1  CGAGAC
2  AGCGAC
3  AGATTC
4  GGATAG
```



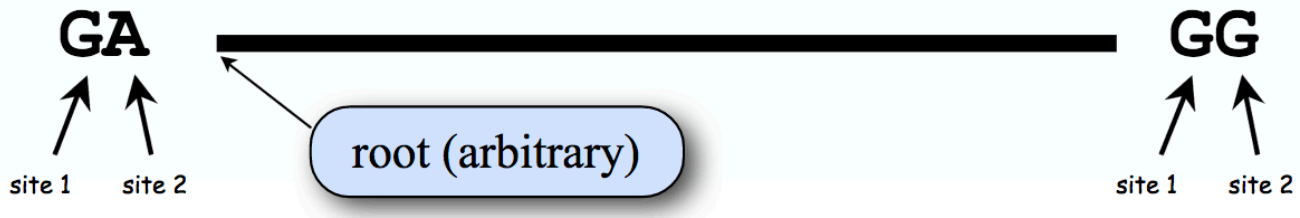
Tree A

What is the likelihood that Tree A (rather than another tree) could have generated the sequence alignment?

Likelihood of the simplest tree

sequence 1 ——— sequence 2

To keep things simple, assume that the sequences are only 2 nucleotides long:



$$
\begin{aligned}
L &= L_1 \, L_2 \\
&= \left[\left(\frac{1}{4}\right) \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)\right] \left[\left(\frac{1}{4}\right) \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)\right]
\end{aligned}
$$

Pr(G) Pr(G|G, αt) Pr(A) Pr(G|A, αt)

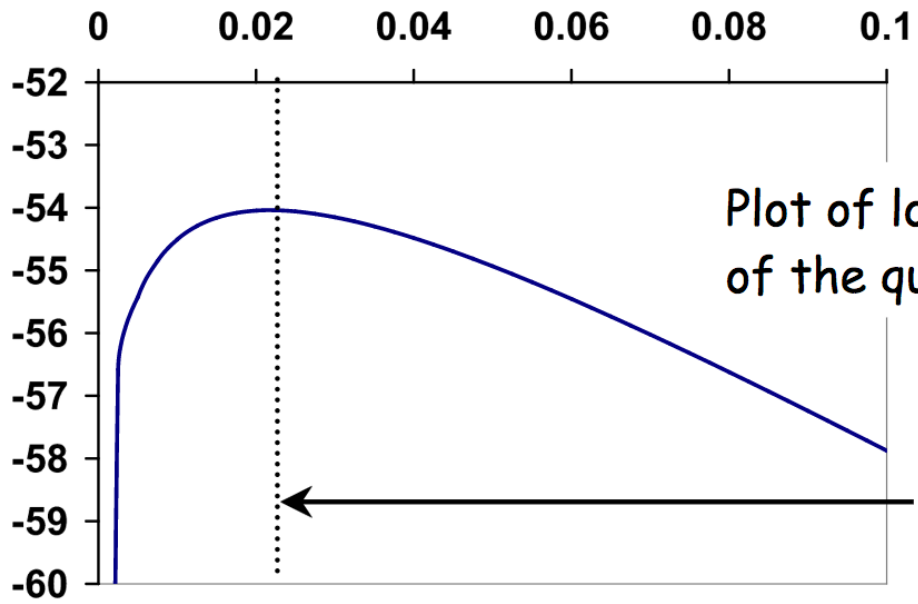
Note that we are NOT assuming independence here

Maximum likelihood estimation

First 32 nucleotides of the ψη-globin gene of gorilla and orangutan:

```
gorilla    GAAGTCCTTGAGAAATAAACTGCACACACTGG
orangutan  GGACTCCTTGAGAAATAAACTGCACACACTGG
```

$$L = \left[\left(\tfrac{1}{4}\right)\ \left(\tfrac{1}{4}+\tfrac{3}{4}e^{-4\alpha t}\right)\right]^{30}\left[\left(\tfrac{1}{4}\right)\ \left(\tfrac{1}{4}-\tfrac{1}{4}e^{-4\alpha t}\right)\right]^{2}$$

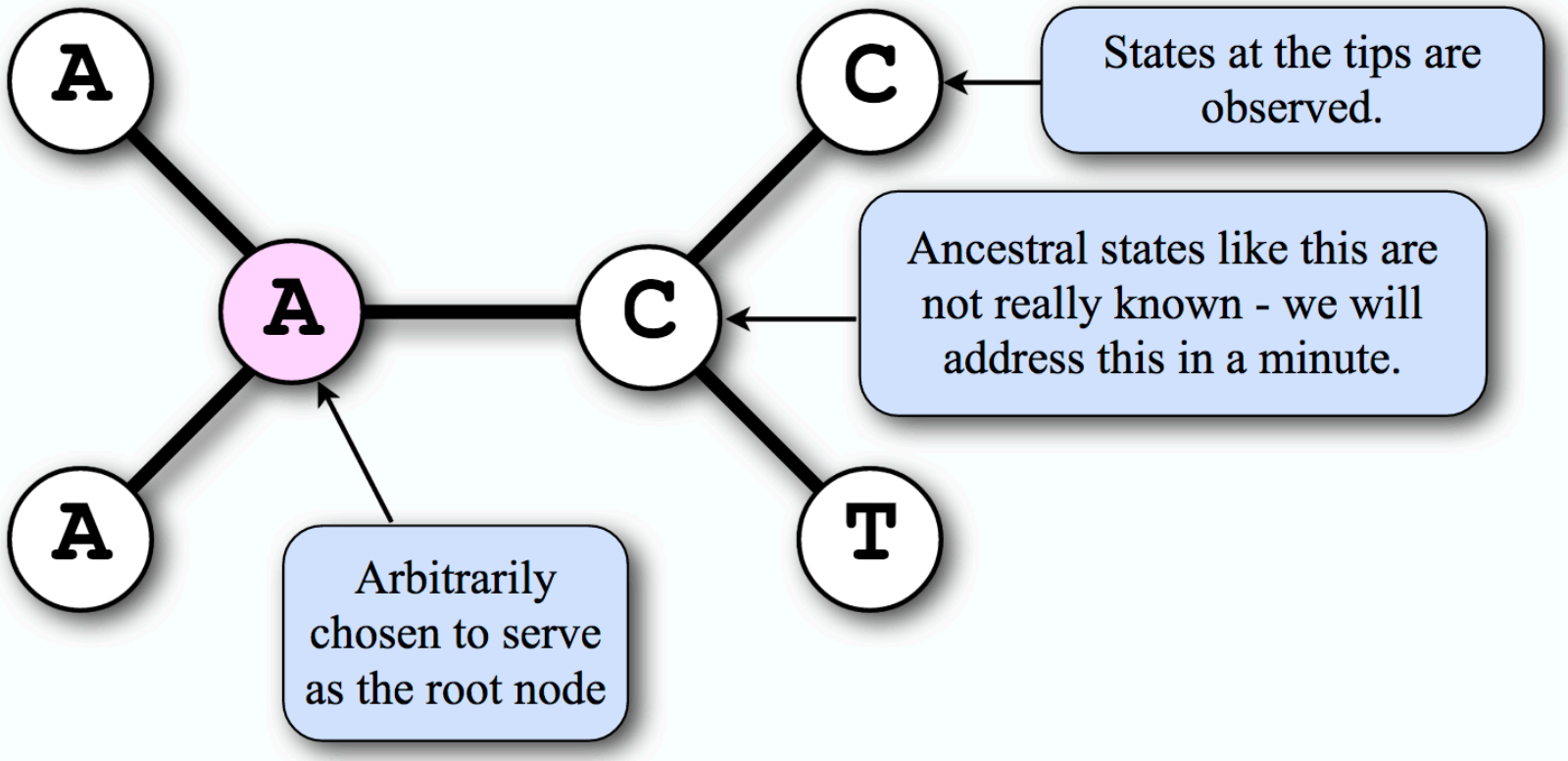


Plot of log-likelihood as a function of the quantity αt

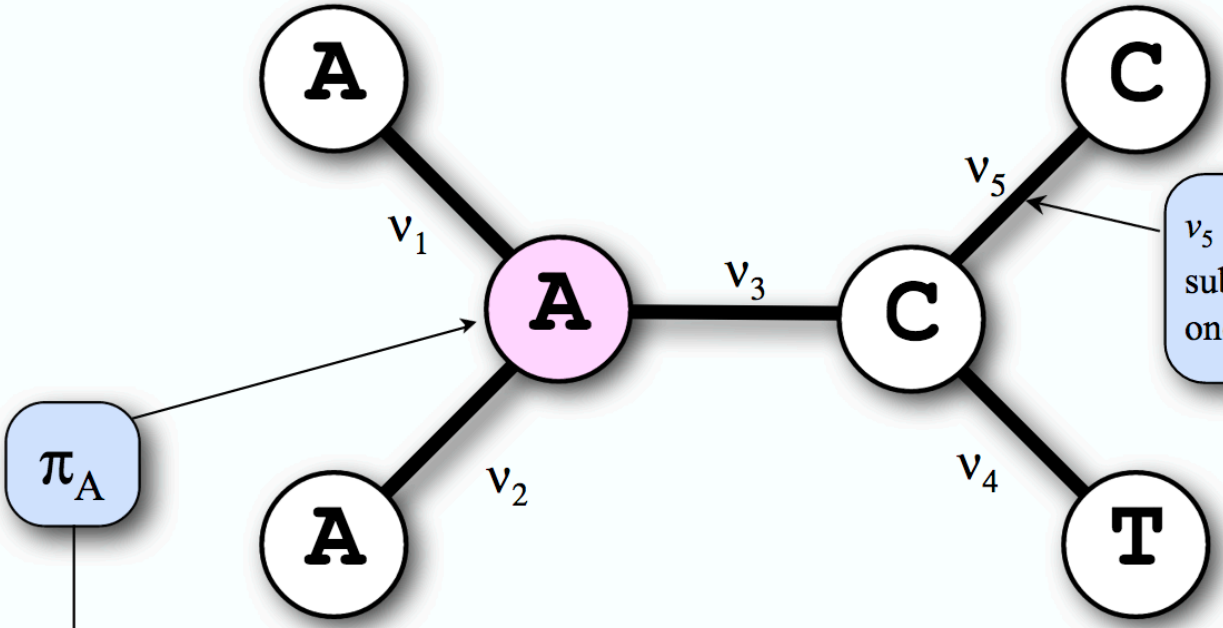
Maximum likelihood estimate (MLE) of αt is 0.021753

Likelihood of an unrooted tree

(data shown for only one site)



Likelihood for site k

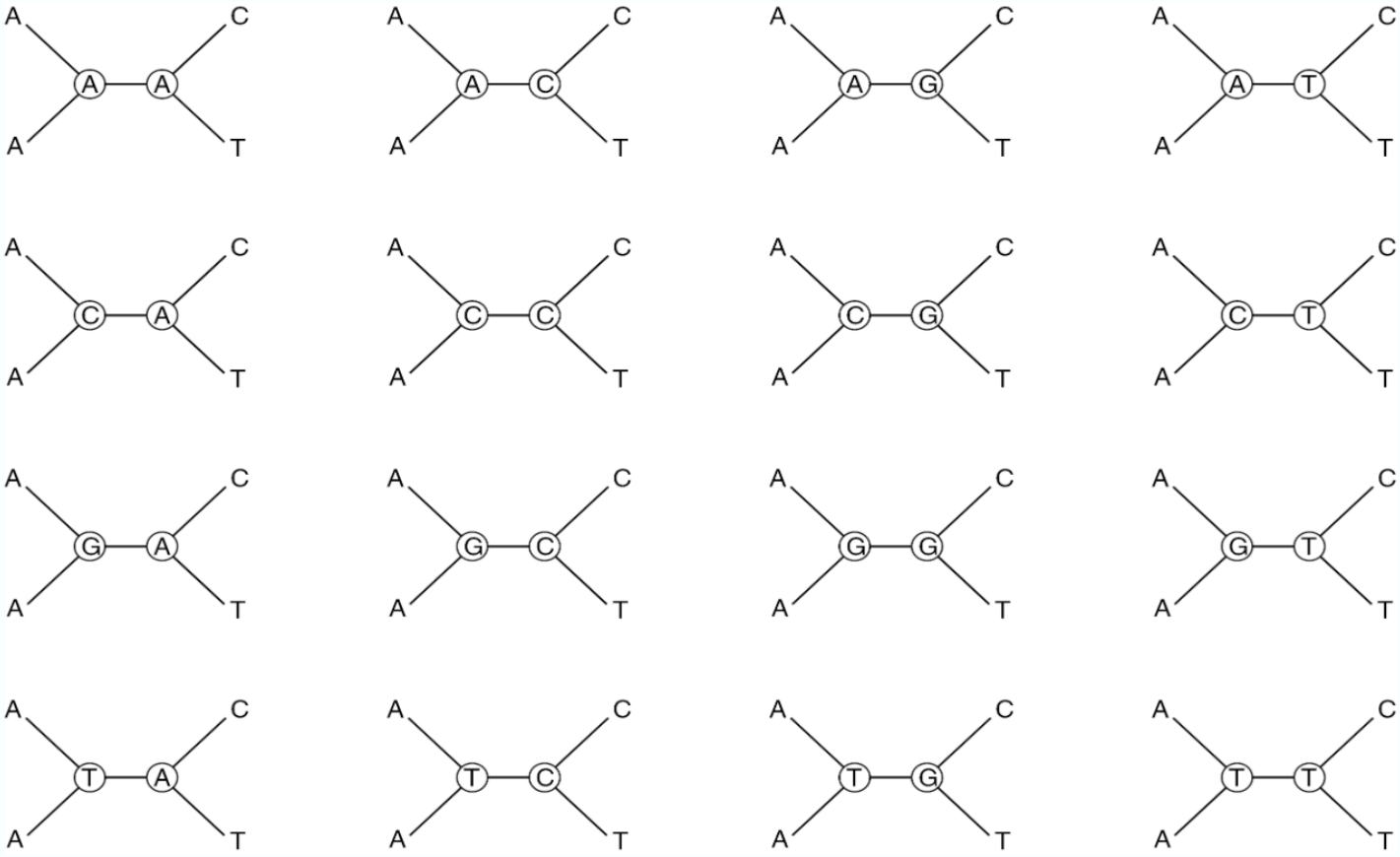


$$L_k = \frac{1}{4}\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_1/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_2/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4\nu_3/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4\nu_4/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_5/3}\right]$$

$P_{AA}(\nu_1)$ $P_{AA}(\nu_2)$ $P_{AC}(\nu_3)$ $P_{CT}(\nu_4)$ $P_{CC}(\nu_5)$

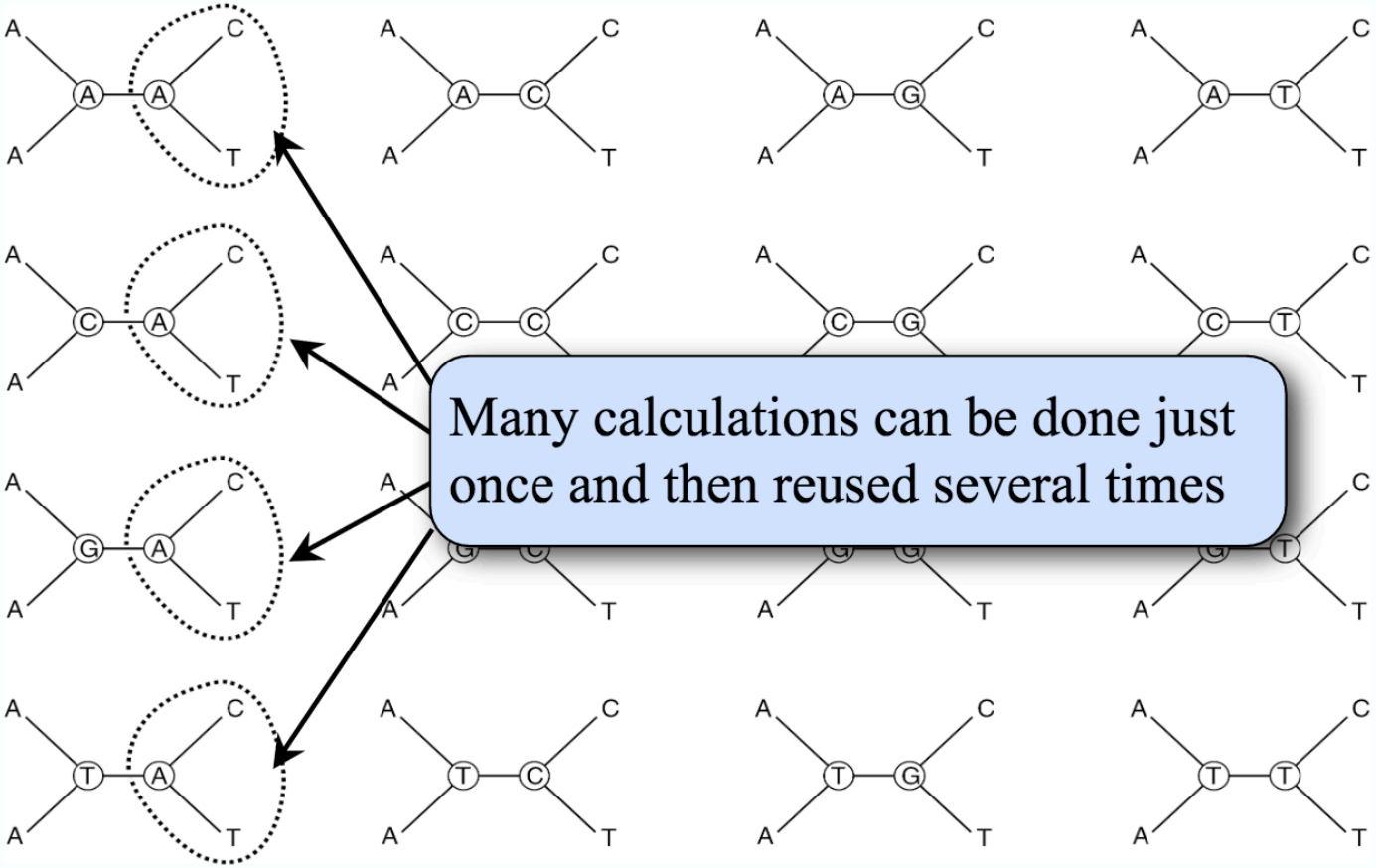
Note use of the AND probability rule

Brute force approach would be to calculate L_k for all 16 combinations of ancestral states and sum them



Note use of the OR probability rule

Pruning algorithm
(same result, less time)



Felsenstein, J. 1981. Evolutionary trees from DNA sequences:
a maximum likelihood approach. *Journal of Molecular Evolution* **17**:368-376

Finding the maximum likelihood of a tree

- **Problem: the number of possible trees** (e.g. for 10 taxa, 2 million unrooted trees possible)
  - for each tree topology we need to identify the maximum likelihood estimate for evolutionary parameters and branch lengths
  - then compare the likelihood among all the trees
    - This is simply computationally not feasible

- **Solution:**
  - There is currently no method that guarantees finding the best tree in the huge space of possible topology
  - Heuristic approaches are used:
    - e.g. NNI = Nearest Neighbour Interchange, SPR = subtree pruning and regrafting

Typical assumptions of ML substitution models

- The probability of any change is independent of the prior history of the site (a Markov Model)
- Substitution probabilities do not change with time or over the tree (a homogeneous Markov process)
- Change is time reversible e.g. the rate of change of A to T is the same as T to A

Advantages of ML

- **Appropriate for DNA sequences: can be reasonably modelled by stochastic processes**

  $\Rightarrow$statistical description of the stochastic processes

- **statistically well understood**
- **estimation method least affected by sampling error**
- **can evaluate different tree topologies (vs. NJ)**
- **use all the sequence information (vs. Distance)**

Disadvantages of ML (?)

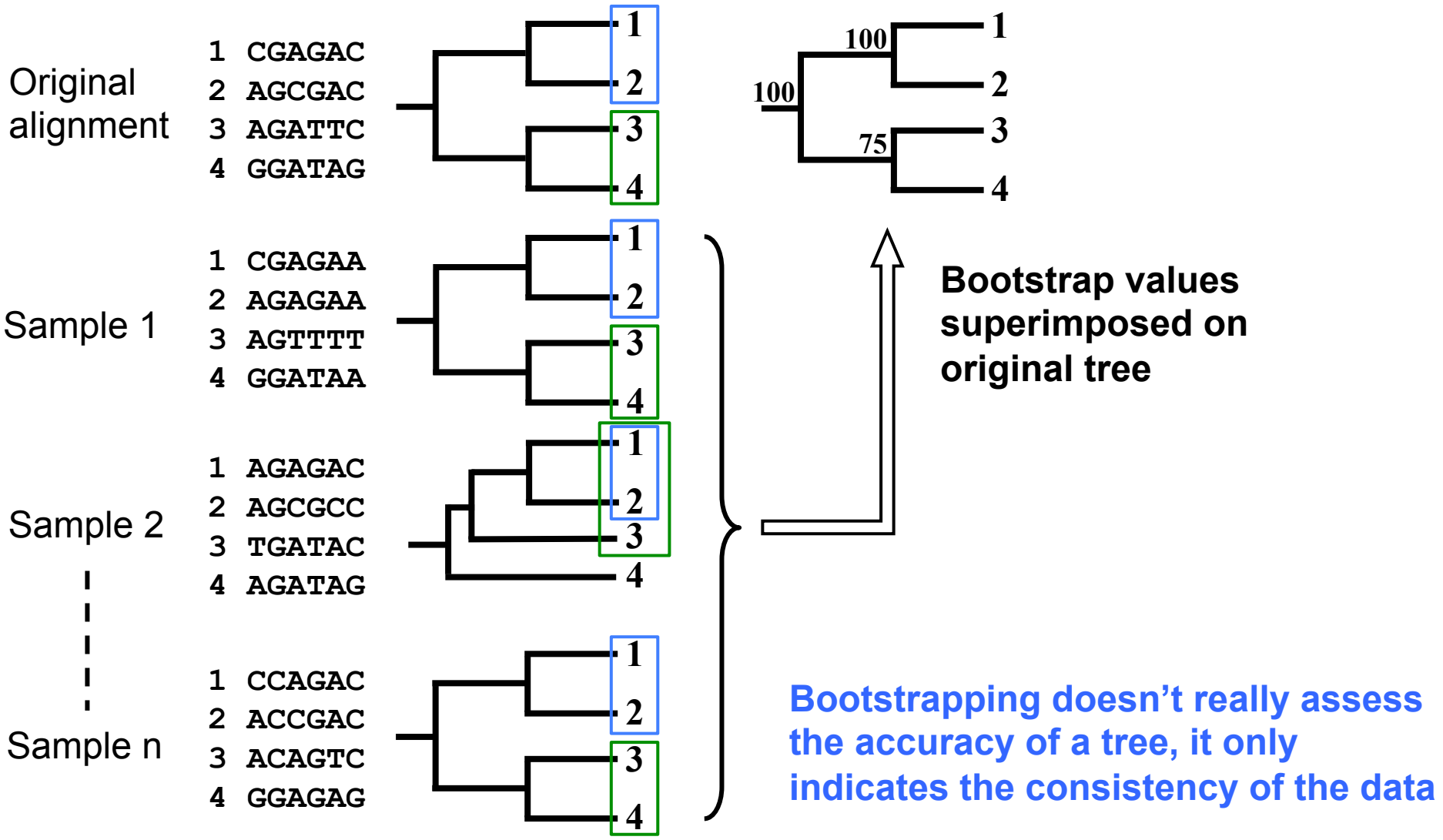
- **very computationally intensive (less of an issue nowadays)**
- **Potentially problematic when missing data not randomly distributed (*Simmons, M.P., 2011. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. Cladistics. 27:1-15*)**
- **questionably applicable to complex data like morphology**
- **philosophically less well established compared to parsimony**
- **Compared to Bayesian, there is only one tree per tree search and no direct test for robustness**

-> bootstrap!

Maximum Likelihood should be seen as a tree estimation procedure instead of a tree reconstruction

“we are making a *best estimate* of an evolutionary history based on incomplete information” Swofford, 1990

Bootstrap


Original alignment
1 CGAGAC
2 AGCGAC
3 AGATTC
4 GGATAG
Sample 1
1 CGAGAA
2 AGAGAA
3 AGTTTT
4 GGATAA
Sample 2
1 AGAGAC
2 AGCGCC
3 TGATAC
4 AGATAG
Sample n
1 CCAGAC
2 ACCGAC
3 ACAGTC
4 GGAGAG
100
100
75
Bootstrap values superimposed on original tree
Bootstrapping doesn't really assess the accuracy of a tree, it only indicates the consistency of the data

Some ML programs

- **RAxML**
- **PhyML**
- **IQ-Tree**

A Bayesian Approach to Phylogenetics

Bayesian inference in general

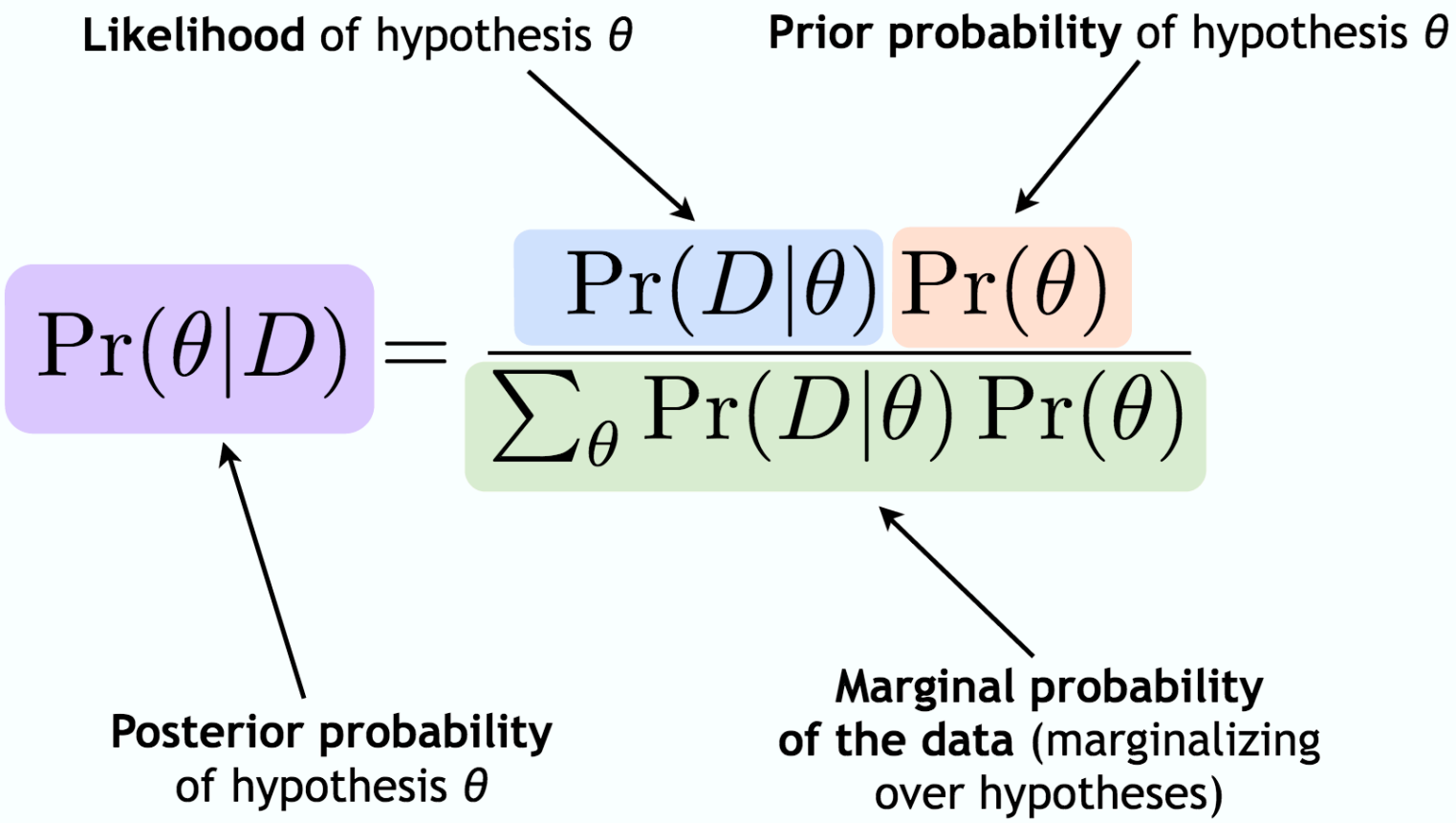
- **D stands for data**
- **Θ (Gr. theta) means any one of a number of things:**
  - **a discrete hypothesis**
  - **a distinct model (e.g. JC, HKY, GTR, etc.)**
  - **a tree topology**
  - **one of an infinite number of continuous model parameter values (e.g. ts:tv rate ratio)**

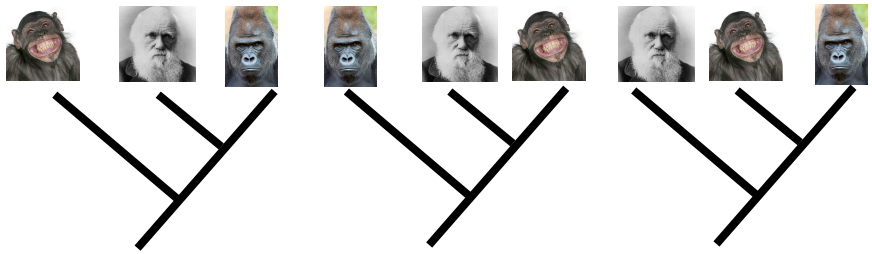
A Bayesian approach compared to ML

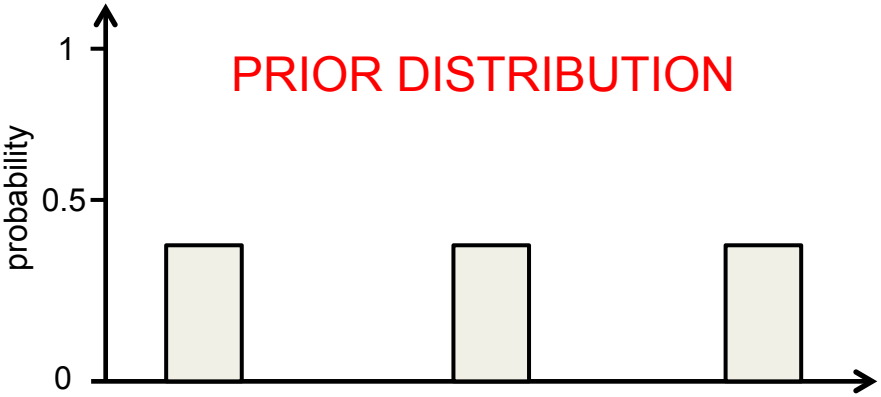
- **The likelihood is the probability of observing the data given a hypothesis**
  - **$L = Pr(D \mid \theta)$.**
- **In ML we search for the parameter values of the model that maximize the likelihood function**

- **In a Bayesian analysis, we get the probability of a hypothesis given the data (probability of the tree given the sequences)**
  - **We combine the likelihood of a given hypothesis with a prior expectation for this hypothesis to obtain a posterior probability of the hypothesis**

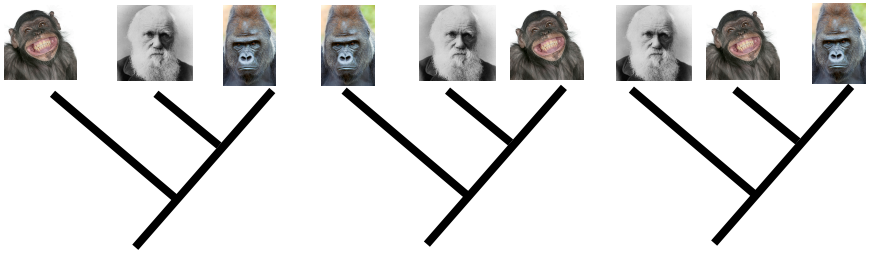
Bayes' rule in statistics



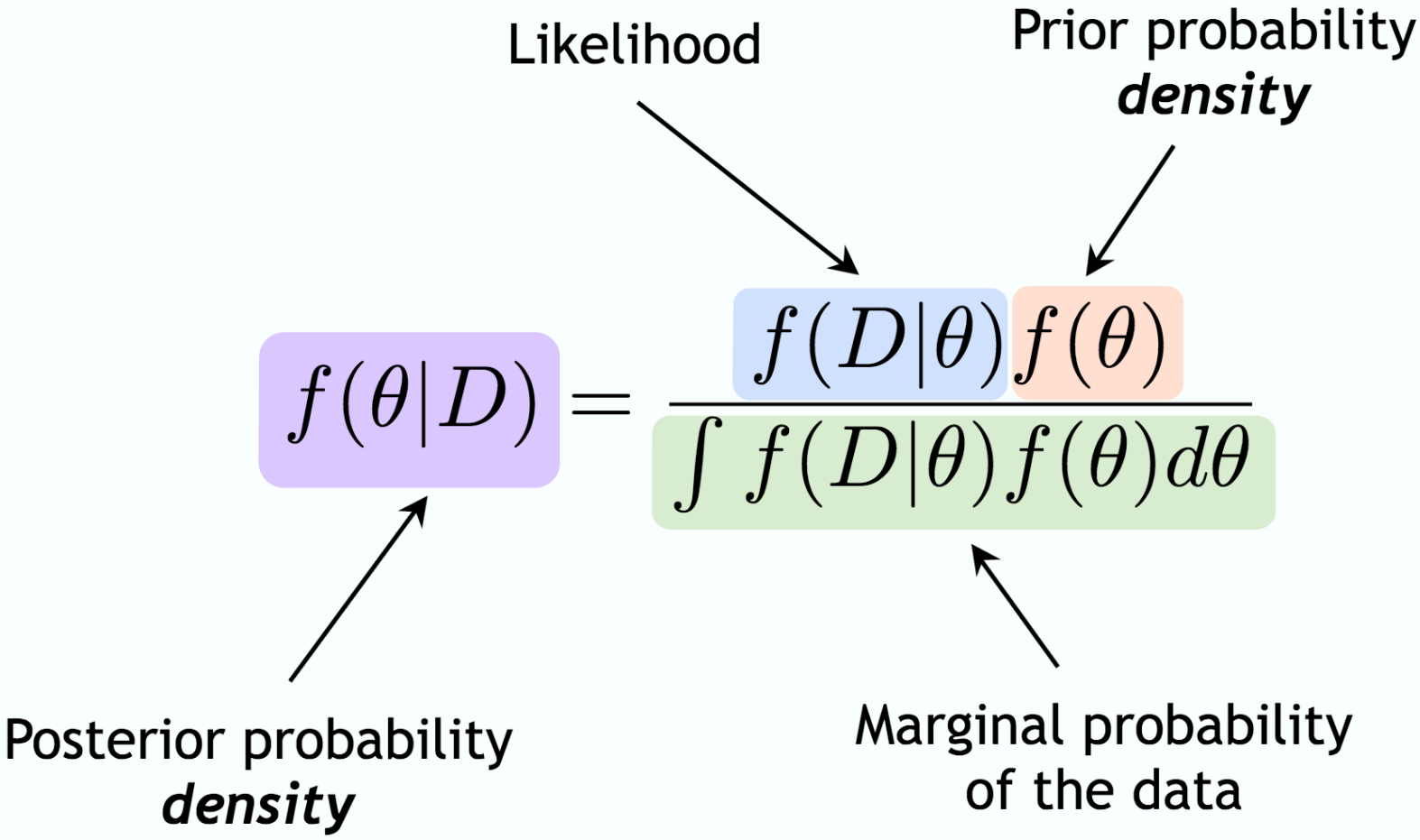


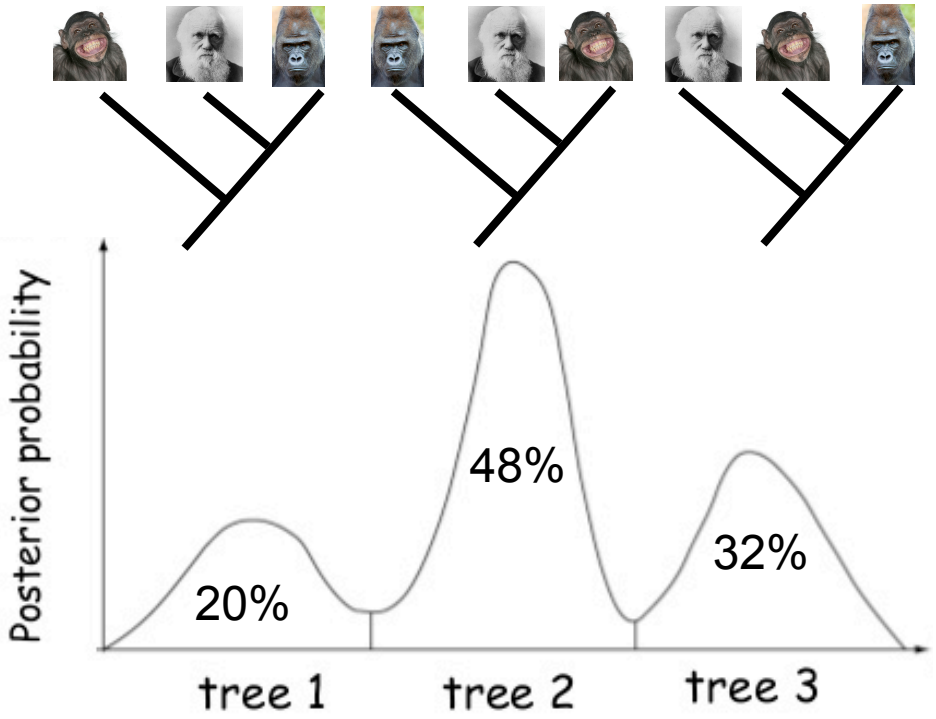


OBSERVATIONS



Bayes' rule: continuous case





Topologies

Branch length	t1	t2	t3	
vA	0.10	0.07	0.12	0.29
vB	0.05	0.22	0.06	0.33
vC	0.05	0.19	0.14	0.38
	0.20	0.48	0.32	

Joint probabilities

Marginal probabilities

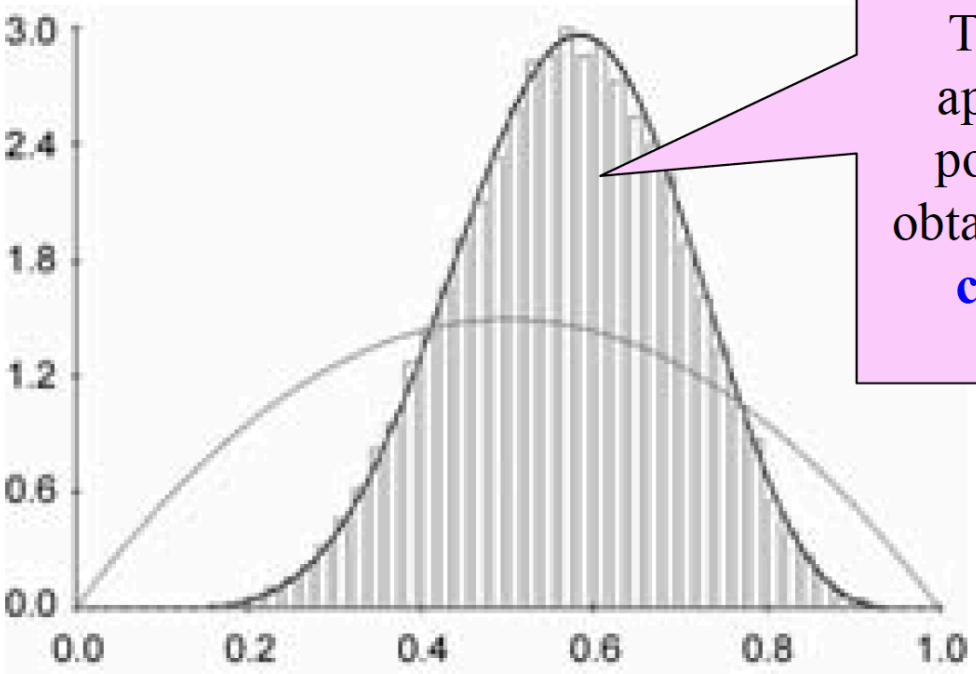
Problem: it is impossible, in most cases,
to derive the posterior probability analytically

or even estimate it by drawing random samples from it

We want something that will “walk” across this parameter space and actively search for the highest point in the parameter “landscape”

Markov chain Monte Carlo (MCMC)

Markov chain Monte Carlo (MCMC)



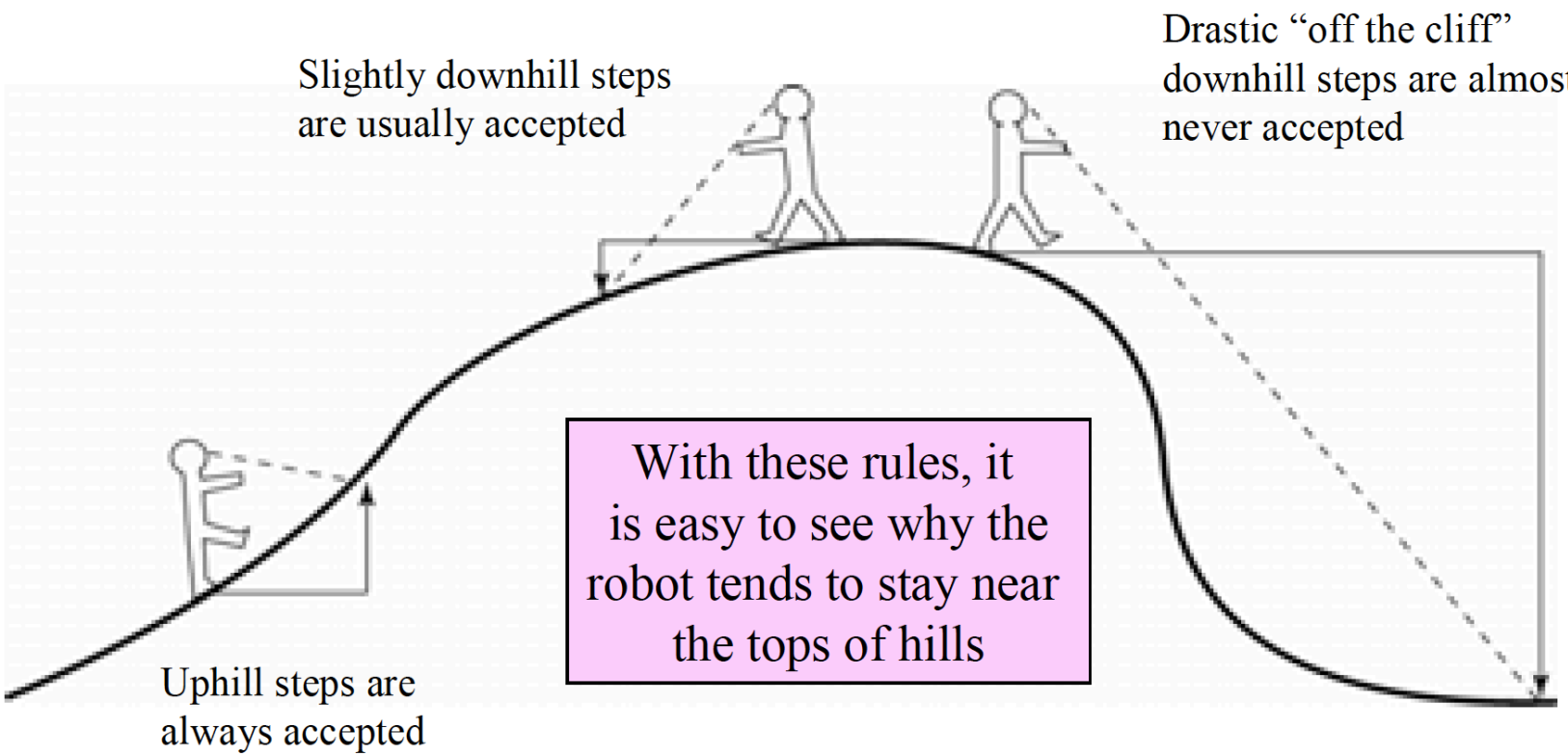
The histogram is an approximation to the posterior distribution obtained using a **M**arkov **c**hain **M**onte **C**arlo simulation.

For more complex problems, we might settle for a

good approximation

to the posterior distribution

MCMC robot's rules



How to decide?

θ = initial position in the parameter space
θ* = new position proposed randomly

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

- **Ratio of posterior probabilities R:**

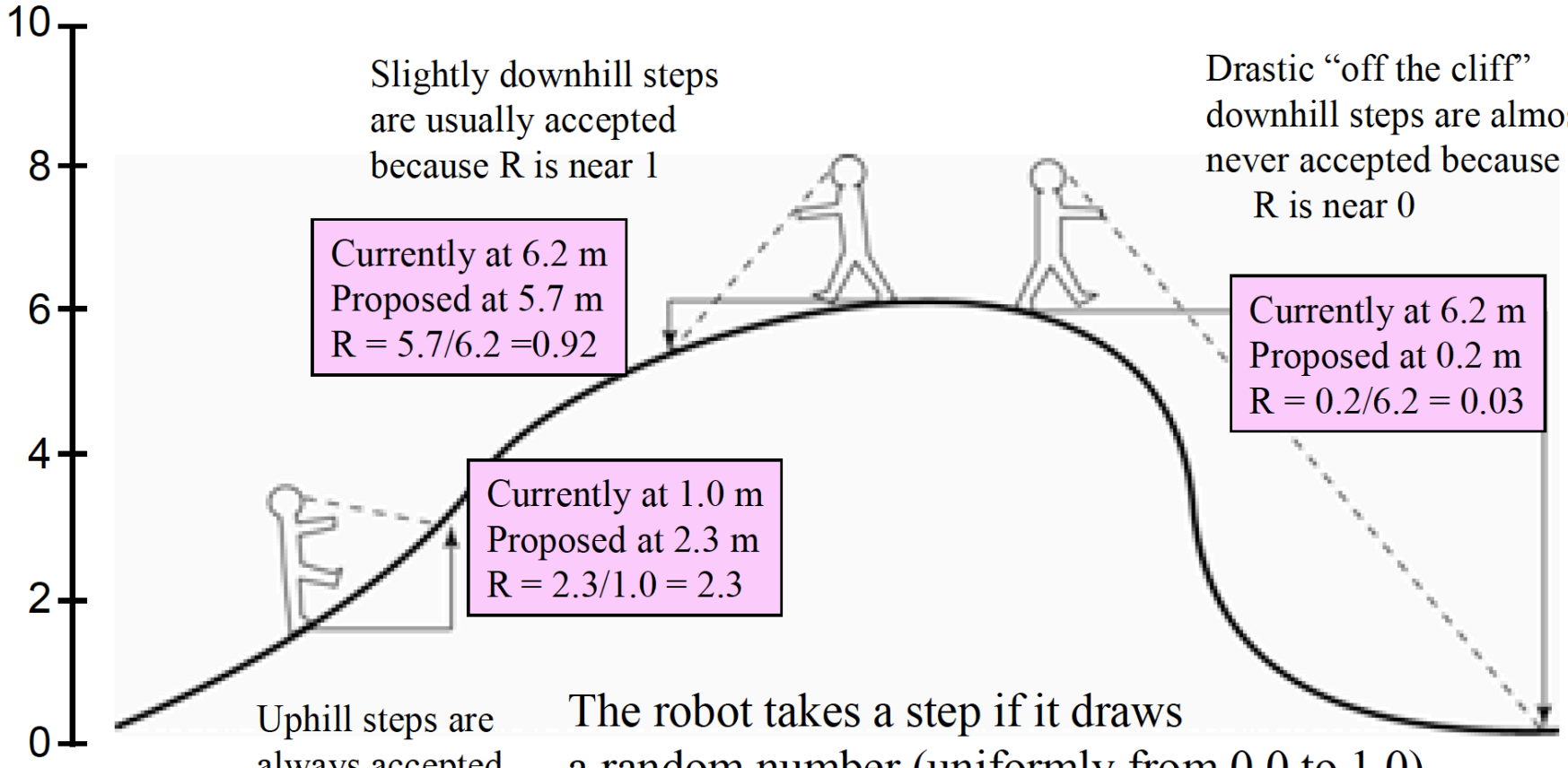
$$\frac{f(\theta^*|D)}{f(\theta|D)} = \frac{\frac{f(D|\theta^*)f(\theta^*)}{\cancel{f(D)}}}{\frac{f(D|\theta)f(\theta)}{\cancel{f(D)}}} = \frac{f(D|\theta^*)f(\theta^*)}{f(D|\theta)f(\theta)} = \mathrm{R}$$

- **Random number between [0,1]:** ***n***

if n ≤ R => new position accepted

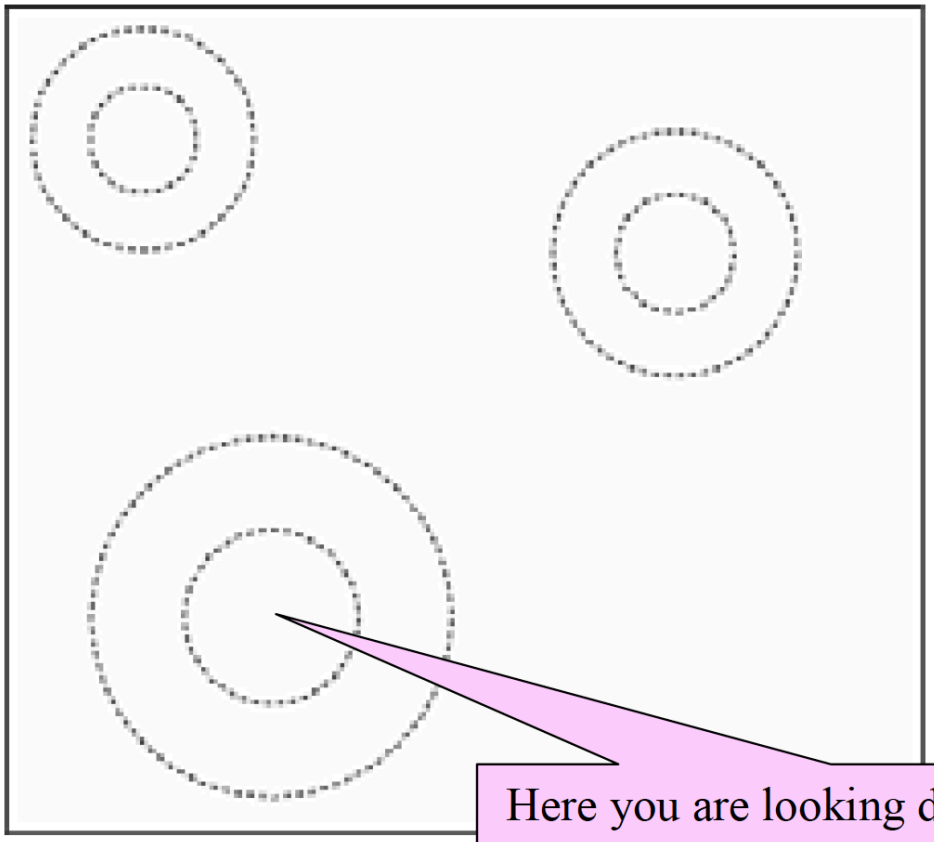
if n > R => new position rejected

(Actual) MCMC robot rules



The robot takes a step if it draws a random number (uniformly from 0.0 to 1.0), and that number is less than or equal to R

What MCRobot can teach us about Markov chain Monte Carlo



Here you are looking down from above at one of the three bivariate normal hills

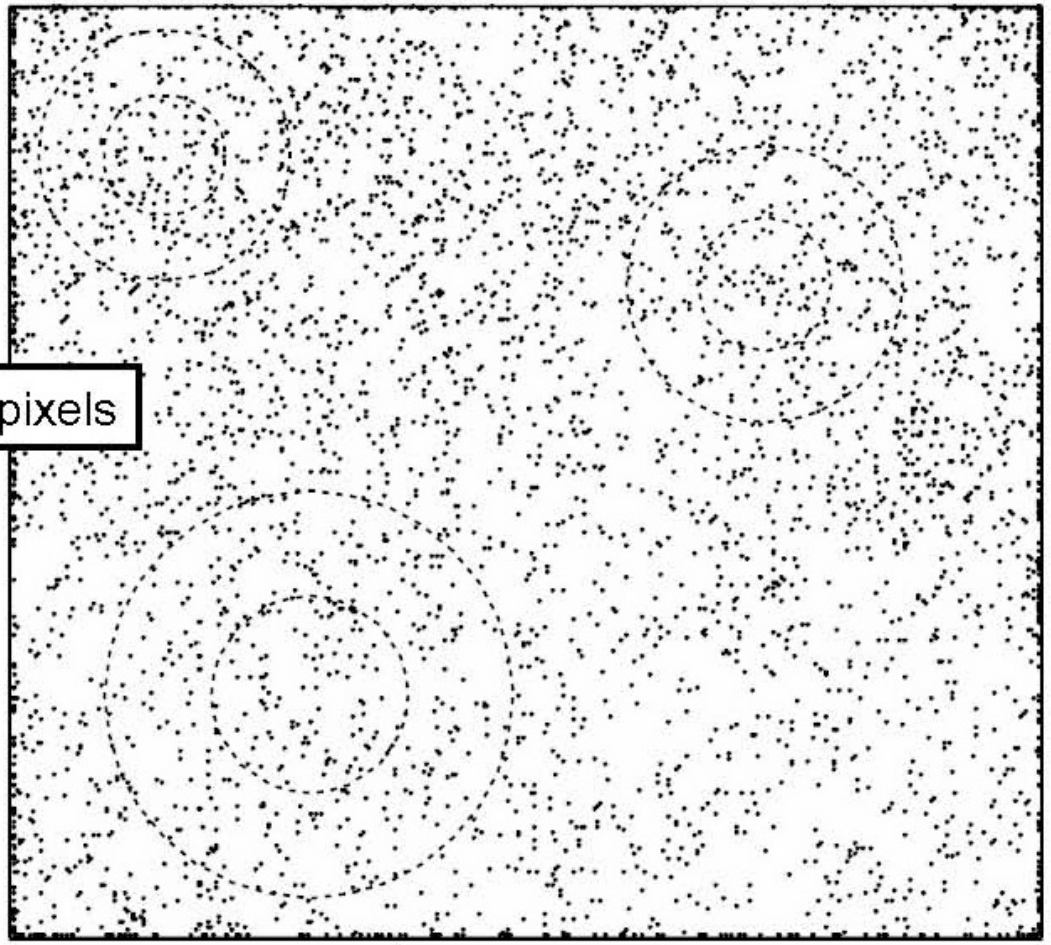
Posterior distribution:

- equal mixture of 3 bivariate normal "hills"
- inner contours: 50%
- outer contours: 95%

Proposal scheme:

- random direction
- gamma-distributed step length
- reflection at edges

Pure random walk



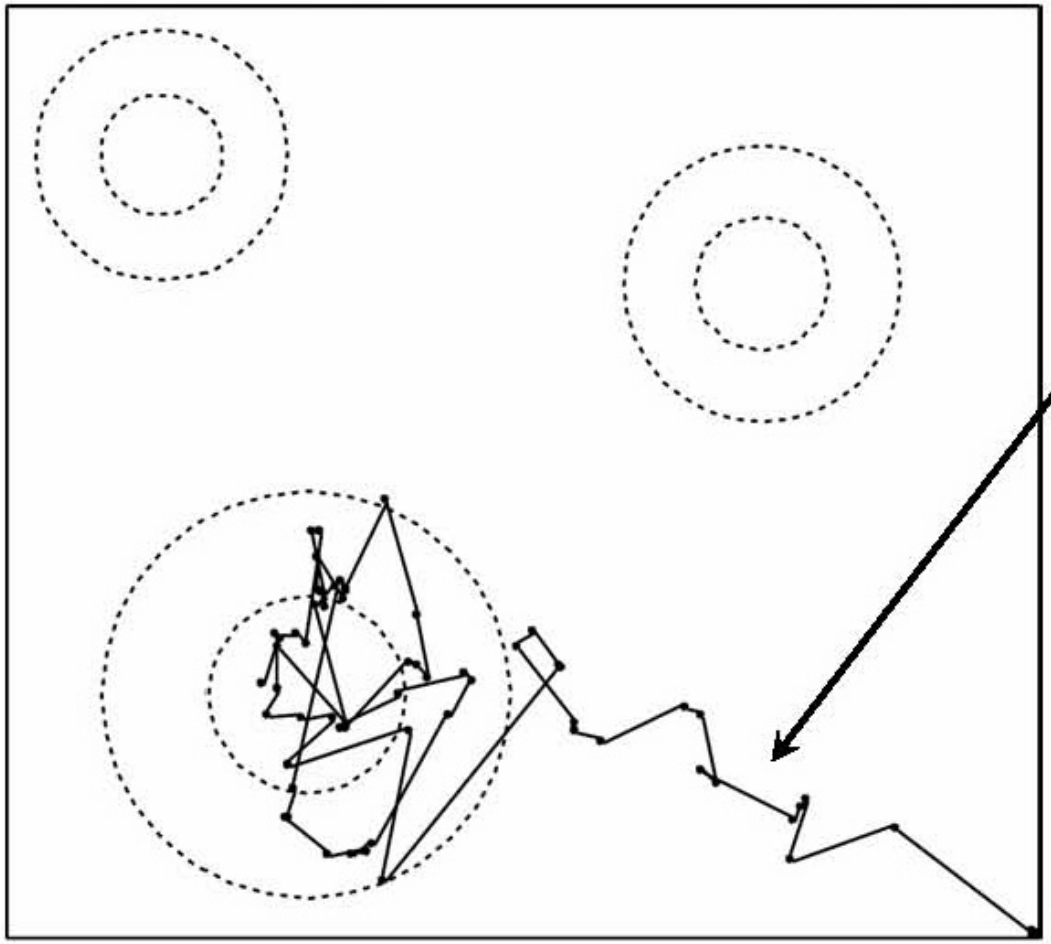
Proposal scheme:
- random direction
- gamma-distributed step length (mean 45 pixels, s.d. 40 pixels)
- reflection at edges

Target distribution:
- equal mixture of 3 bivariate normal "hills"
- inner contours: 50%
- outer contours: 95%

In this case, the robot is accepting every step

5000 steps shown

Burn-in



Robot is now following the rules and thus quickly finds one of the three hills.

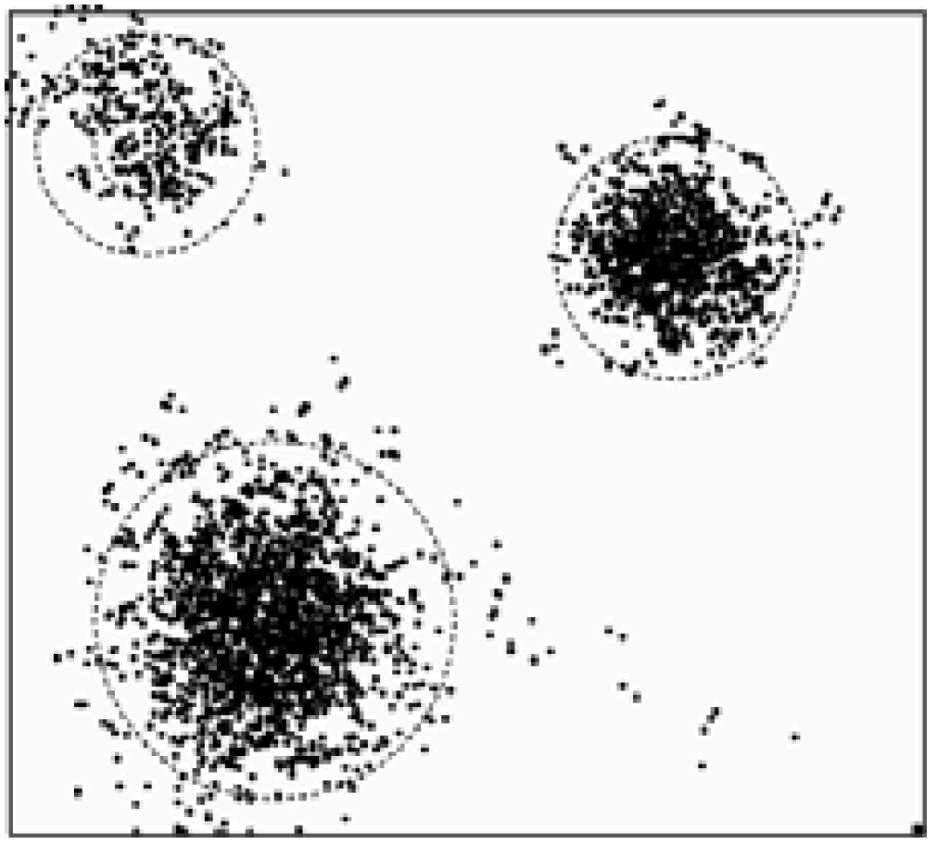
Note that first few steps are not at all representative of the distribution.

100 steps taken

Starting point

Target distribution approximation

5000 steps taken



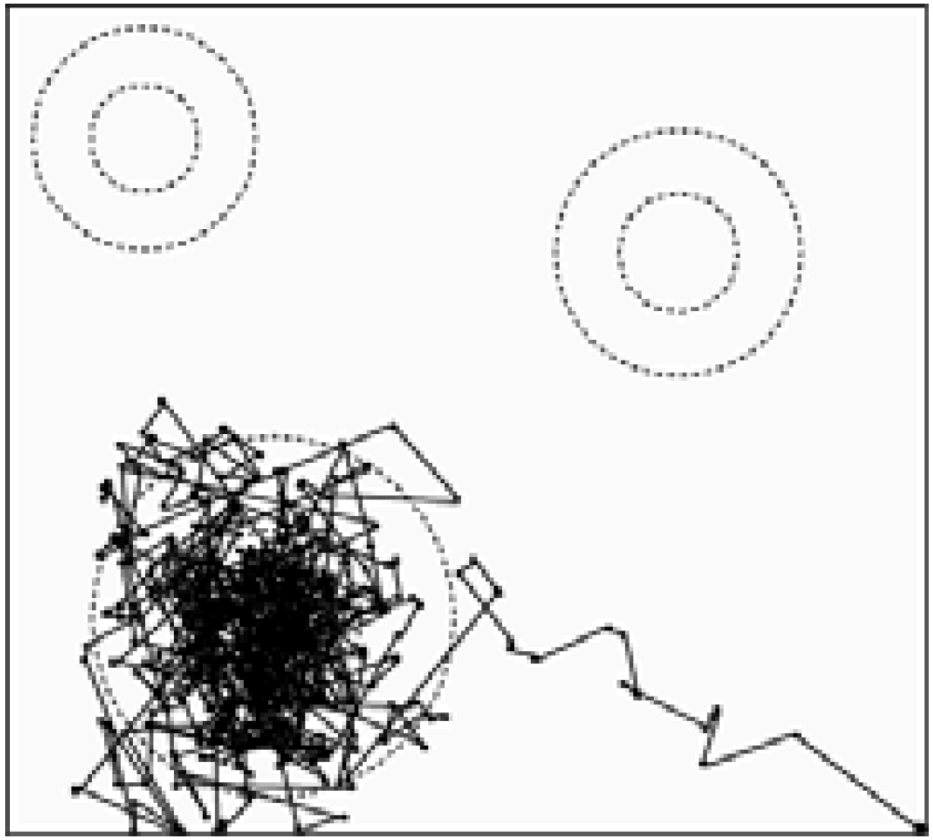
How good is the MCMC approximation?

- 51.2% of points are inside inner contours (cf. 50% actual)
- 93.6% of points are inside outer contours (cf. 95% actual)

Approximation gets better the longer the chain is allowed to run.

Just how long is a long run?

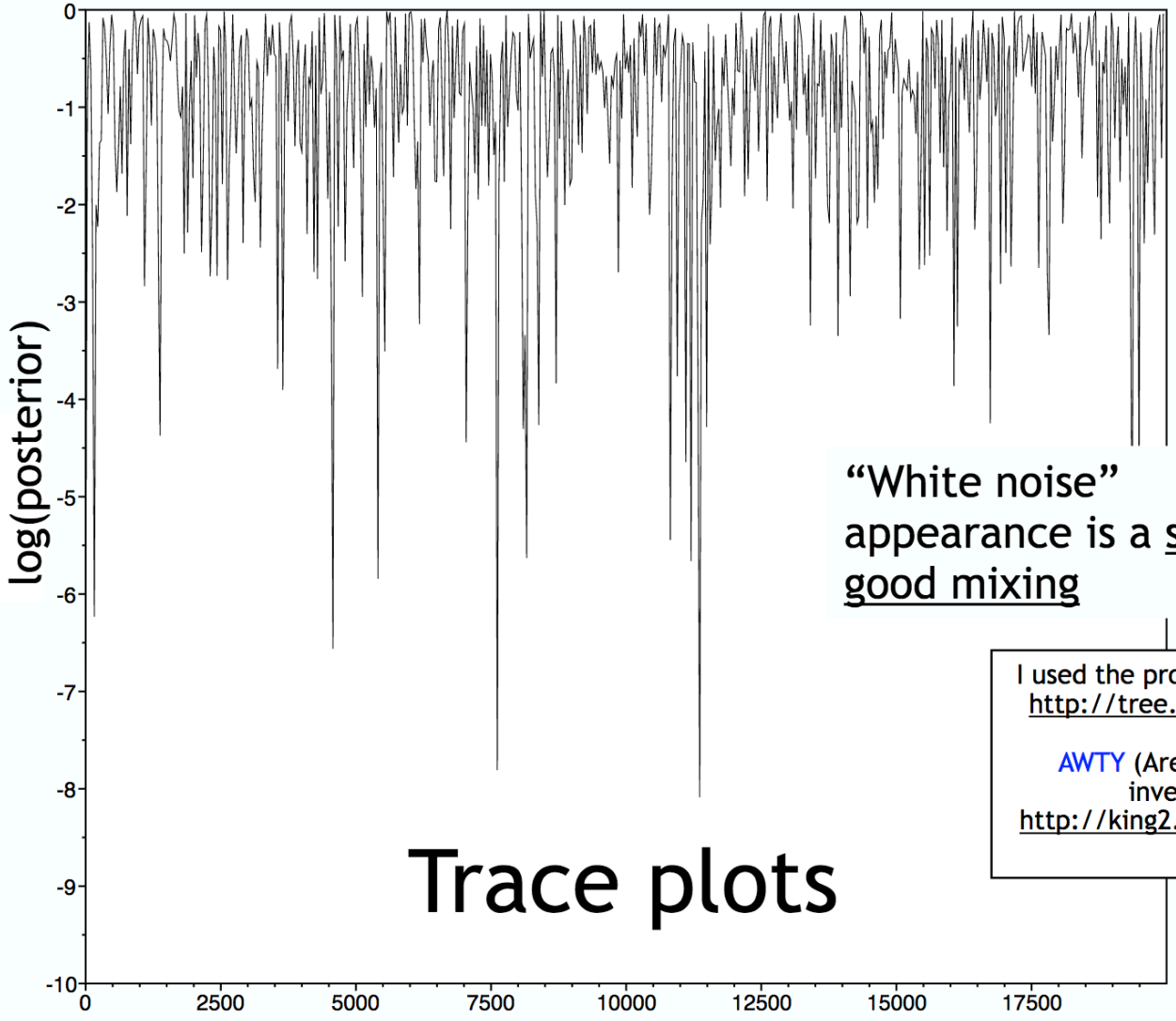
1000 steps taken



What would you conclude about the target distribution had you stopped the robot at this point?

The way to avoid this mistake is to perform several runs, each one beginning from a different randomly-chosen starting point.

Results different among runs? Probably none of them were long enough!

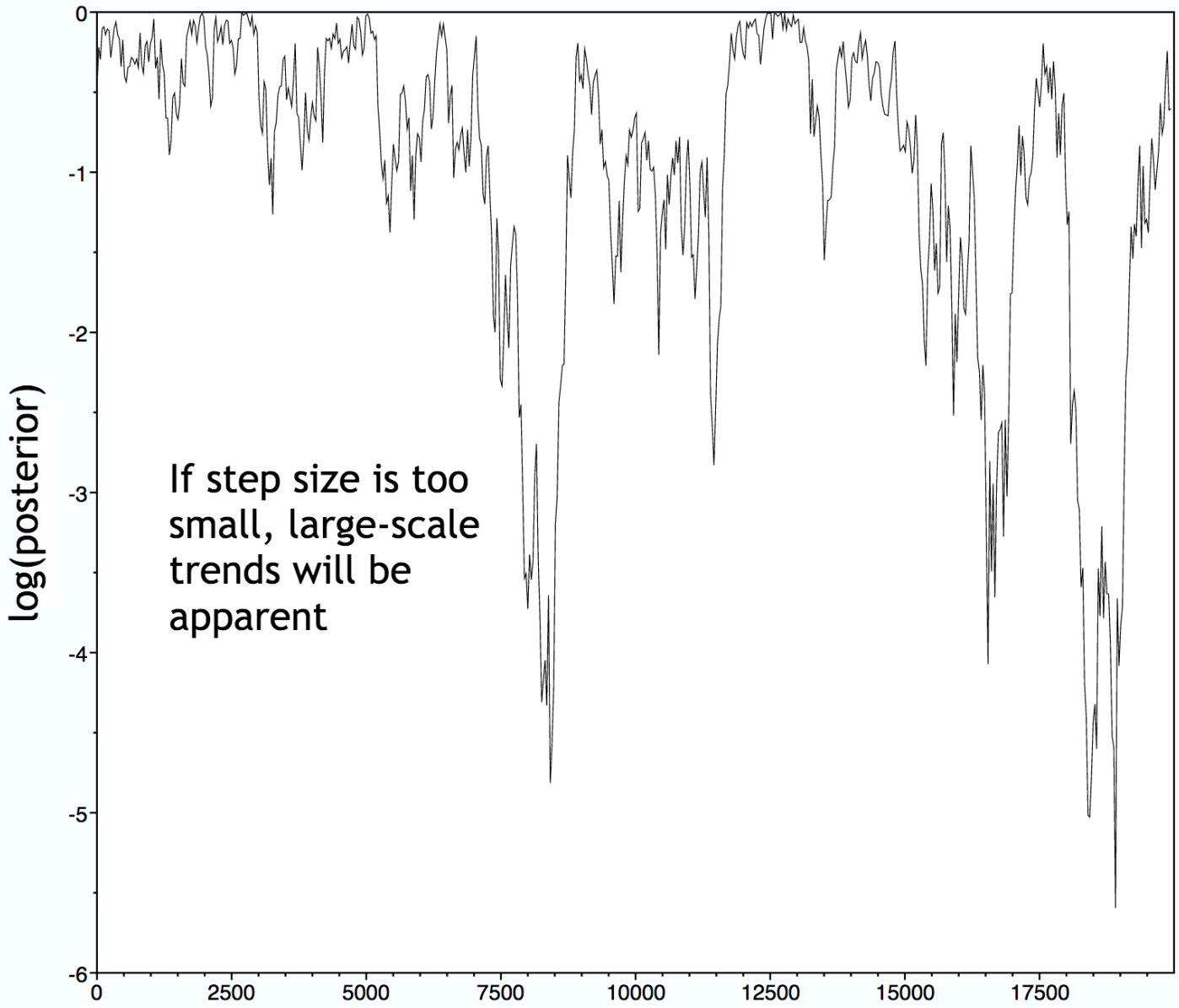


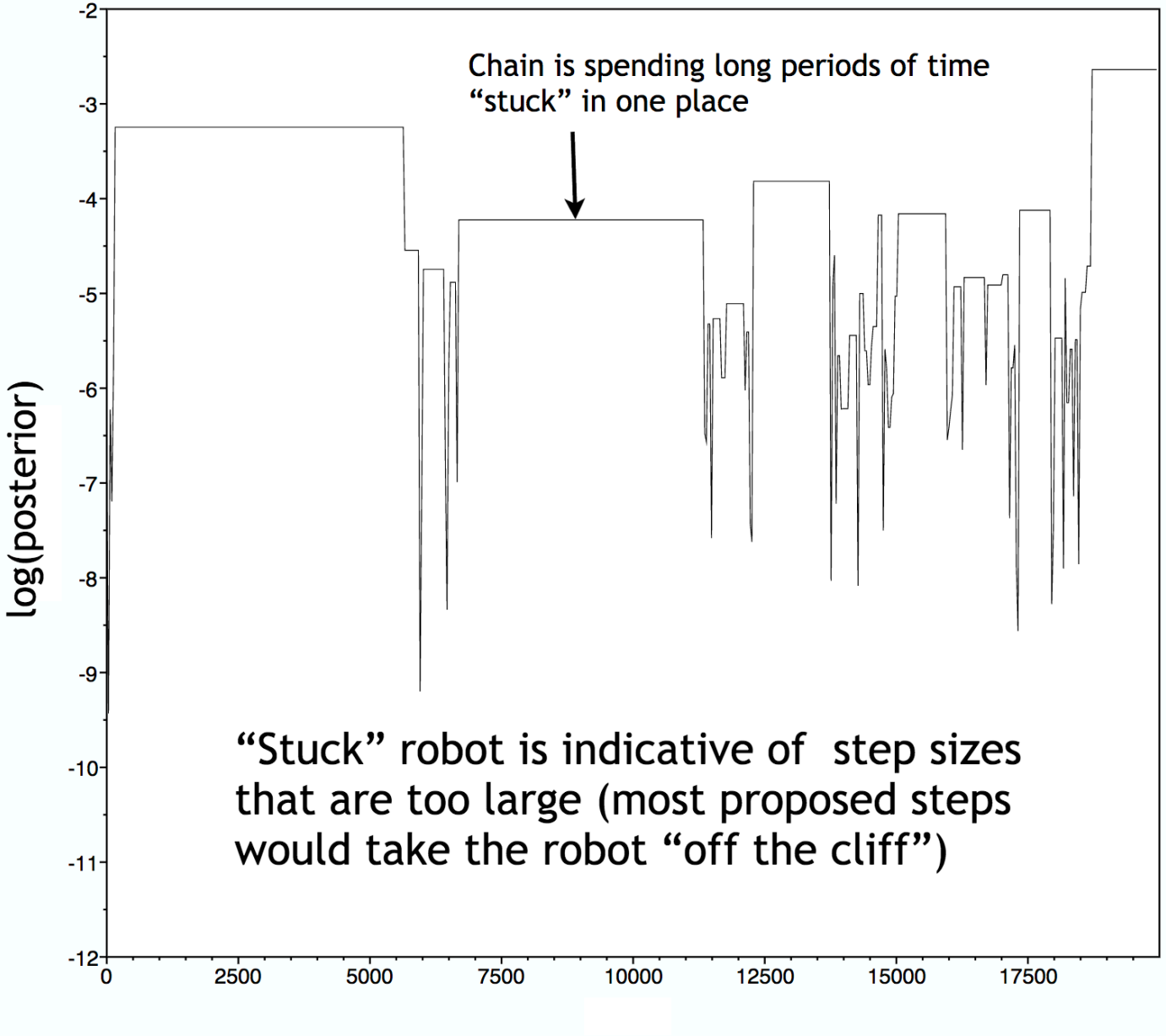
Trace plots

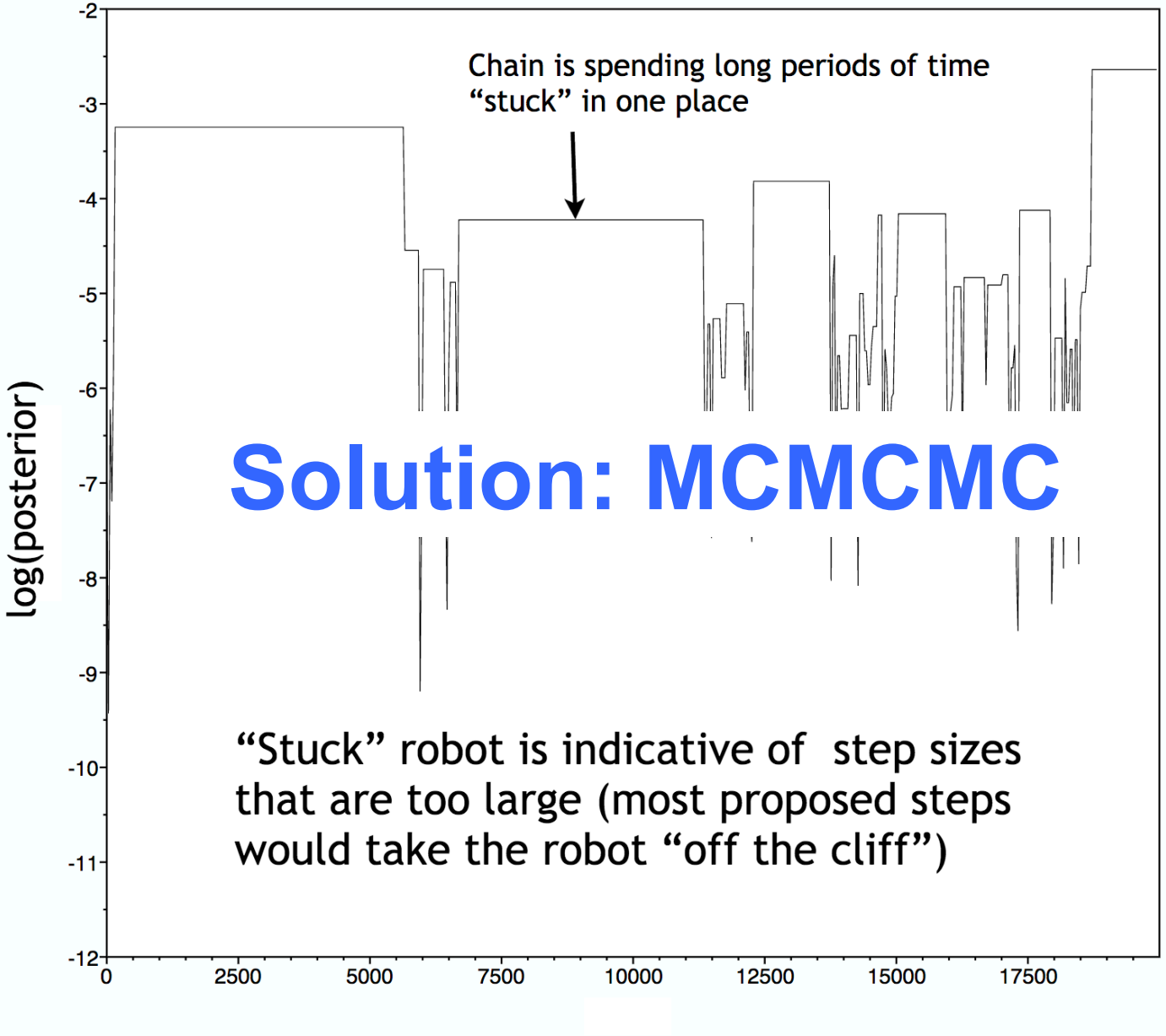
"White noise" appearance is a sign of good mixing

I used the program Tracer to create this plot:
http://tree.bio.ed.ac.uk/software/tracer/

AWTY (Are We There Yet?) is useful for investigating convergence:
http://king2.scs.fsu.edu/CEBProjects/awty/awty_start.php


log(posterior)
0
-1
-2
-3
-4
-5
-6
0
2500
5000
7500
10000
12500
15000
17500
If step size is too small, large-scale trends will be apparent

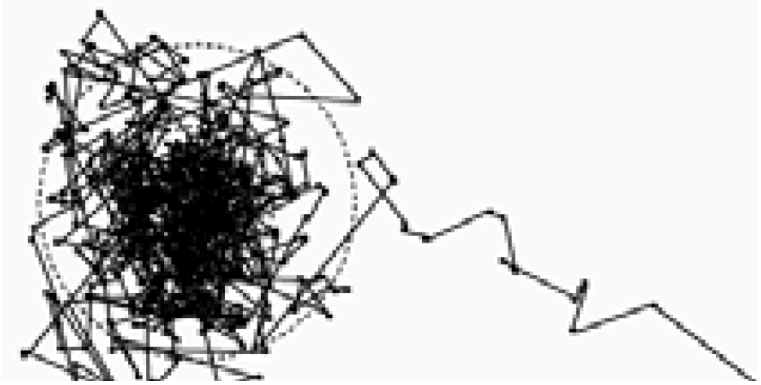

Chain is spending long periods of time
"stuck" in one place
log(posterior)
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
0
2500
5000
7500
10000
12500
15000
17500
"Stuck" robot is indicative of step sizes
that are too large (most proposed steps
would take the robot "off the cliff")



Solution: MCMCMC

"Stuck" robot is indicative of step sizes that are too large (most proposed steps would take the robot "off the cliff")

Metropolis-coupled Markov chain Monte Carlo (MCMCMC, or MC^3)



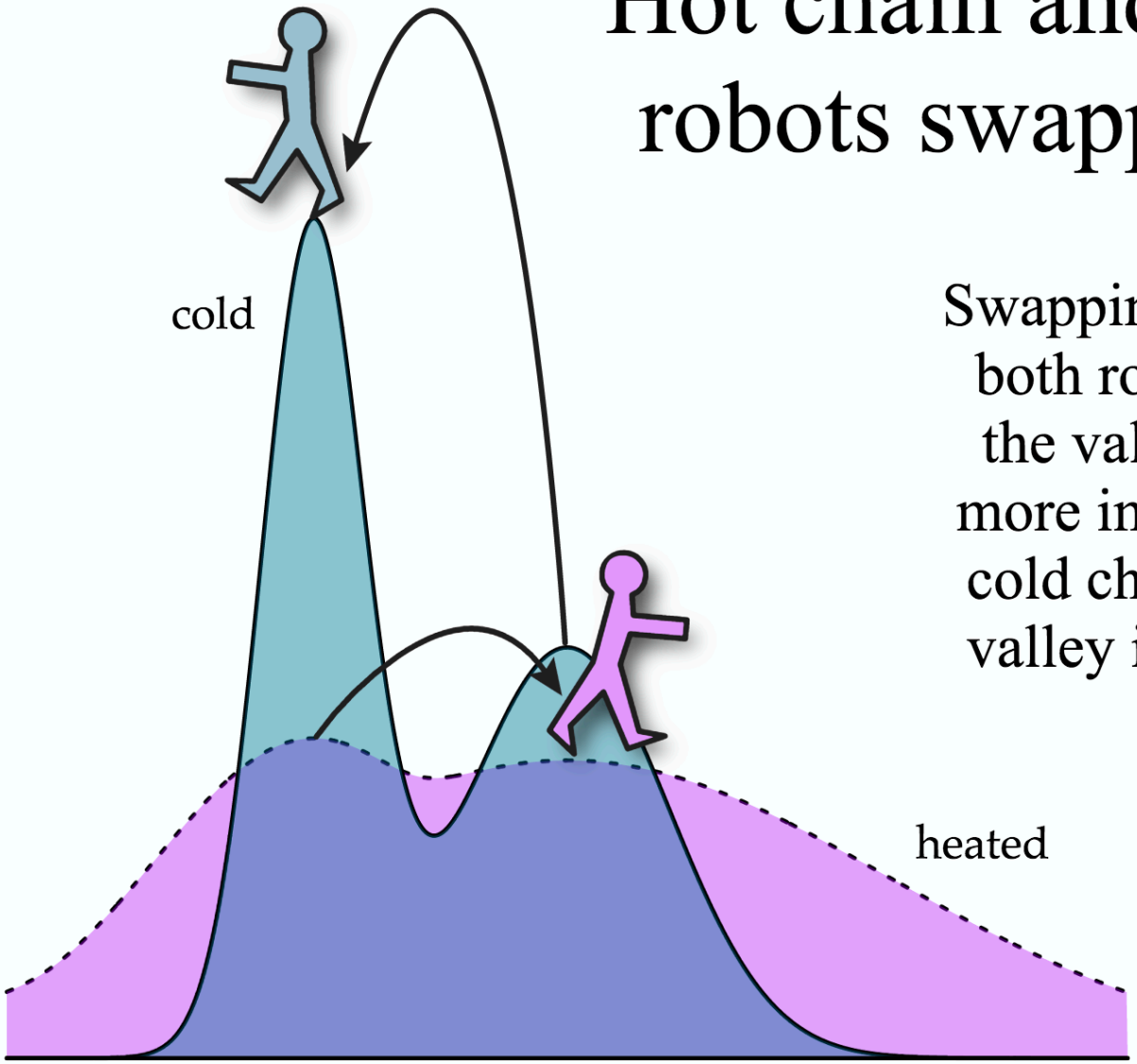
- **MC^3 involves running several chains simultaneously (one "cold" and several "heated")**
- **The cold chain is the one that counts, the heated chains are "scouts"**
- **Chain is heated by raising densities to a power less than 1.0 (values closer to 0.0 are warmer)**

Cold vs. heated landscapes

Cold landscape: note peaks separated by deep valleys

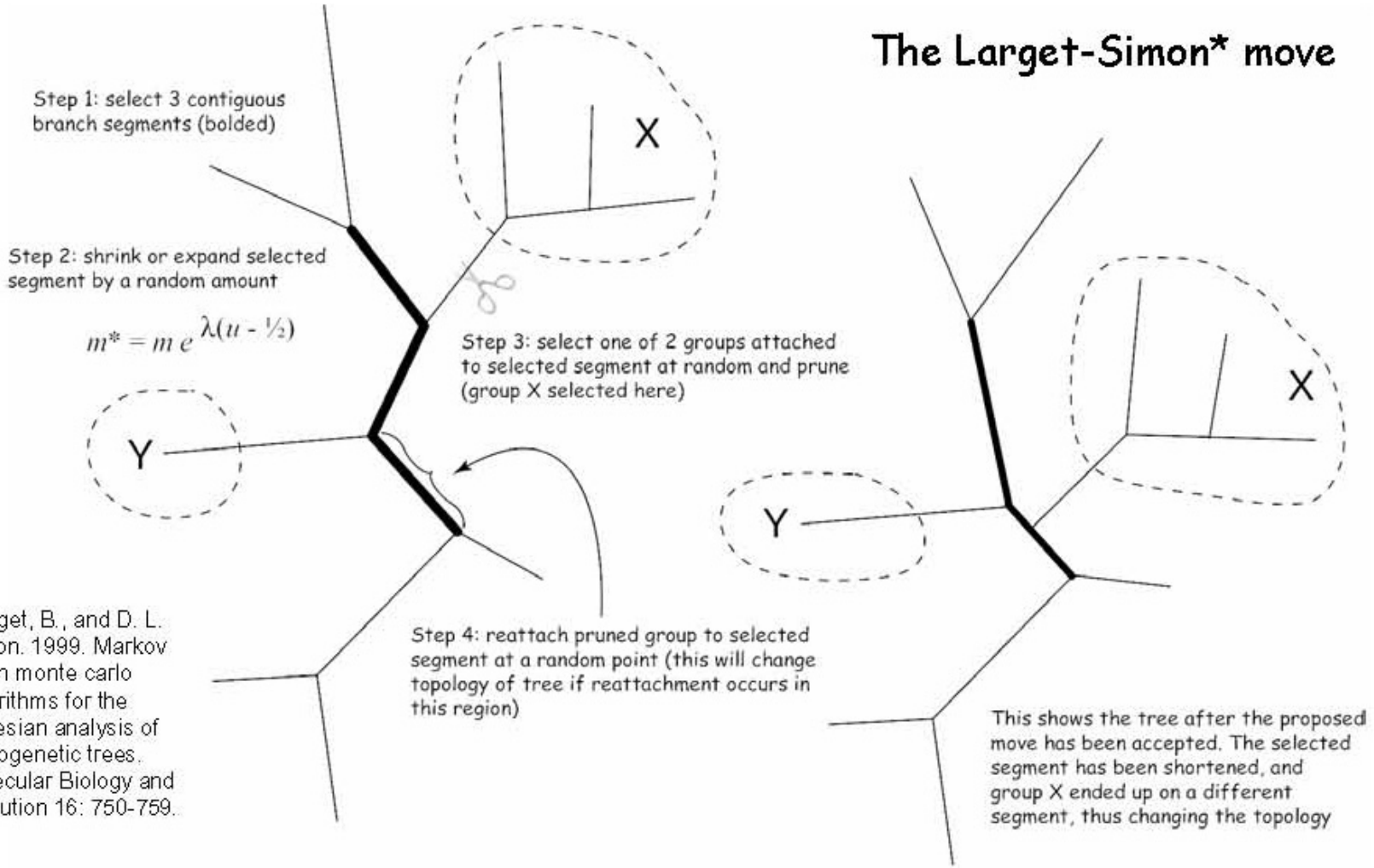
Heated landscape: note shallow (easy to cross) valleys

Hot chain and cold chain robots swapping places



Swapping places means both robots can cross the valley, but this is more important for the cold chain because its valley is much deeper

Moving through treespace



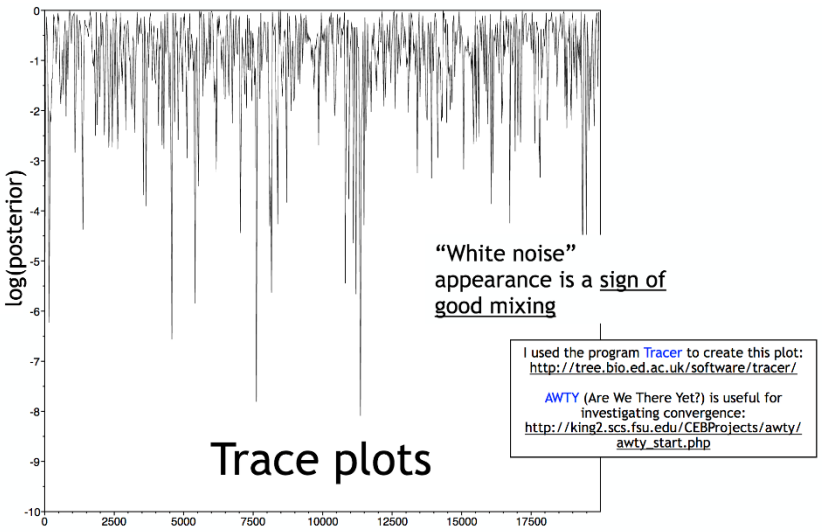
*Larget, B., and D. L. Simon. 1999. Markov chain monte carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution 16: 750-759.

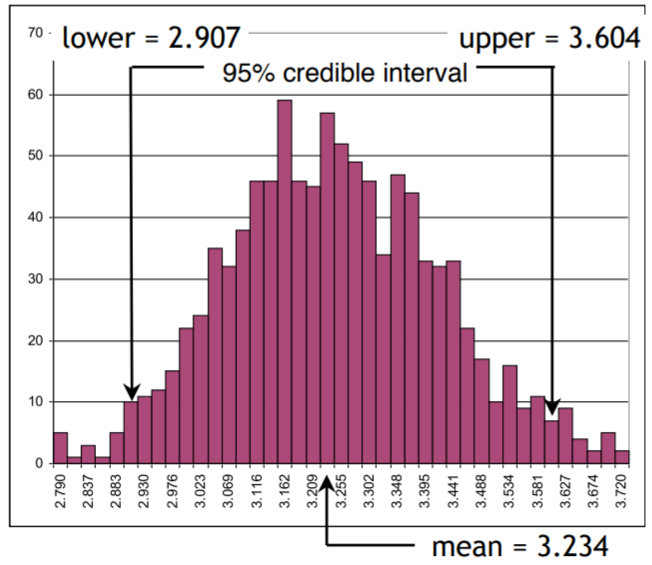
MCMC, in short:

- Start with random tree and arbitrary initial values for branch lengths and model parameters
- Each generation consists of one of these (chosen at random):
  - Propose a new tree (e.g. Larget-Simon move) and either accept or reject the move
  - Propose (and either accept or reject) a new model parameter value
- Every k generations, save tree topology, branch lengths and all model parameters (i.e. sample the chain)
- After n generations, summarize sample using histograms, means, credible intervals, etc.

Looking at the results

- **Graphically**





Histogram created from a sample of 1000 kappa values.

- **Statistically, by computing Effective Sample Size (ESS, minimum should be 200 or more)**
- **Checking convergence of each run, and all runs together**

Summarizing the results

- **Autocorrelation: between values that are sampled one after another, so need to sample values at a lower frequency – e.g. every 1000 steps**
- **MCMC easily runs over millions of states**

=> Synthetize/summarize the parameters we are interested in

By computing the marginal posterior distribution of these parameters

- mean, median or variance

- 95% credibility interval

By identifying one or more "best" topologies

e.g. the splits most frequently identified

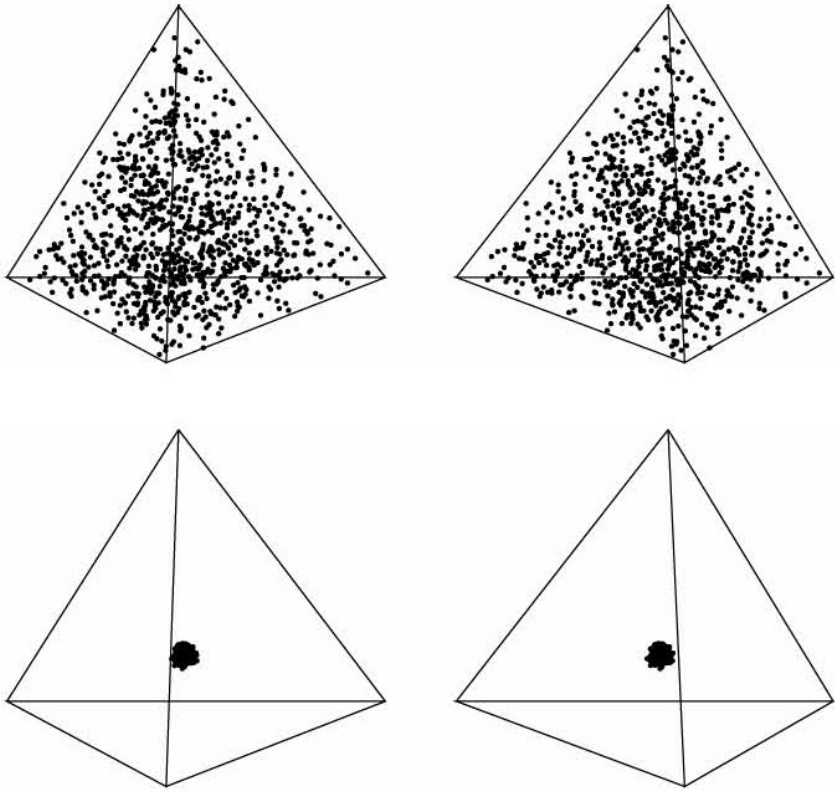
+ The number of times a clade in the tree is accepted during the MCMC defines the posterior probability of the clade, and therefore indicates the support for the node

Prior distributions

Prior Distributions

- For topologies: discrete Uniform distribution
- For proportions: Beta(a,b) distribution
  - flat when a=b
  - peaked above 0.5 if a=b and both are greater than 1
- For base frequencies: Dirichlet(a,b,c,d) distribution
  - flat when a=b=c=d
  - all base frequencies close to 0.25 if v=a=b=c=d and v large (e.g. 300)
- For GTR model relative rates: Dirichlet(a,b,c,d,e,f) distribution

4-parameter Dirichlet(a,b,c,d)



(stereo pairs)

Flat prior:
a = b = c = d = 1

Informative prior:
a = b = c = d = 300

Prior Distributions

- **For other model parameters and branch lengths: Gamma(a,b) distribution**
  - **Exponential(λ) equals Gamma(1, λ-1) distribution**
  - **Mean of Gamma(a,b) is ab (so mean of an Exponential(10) distribution is 0.1)**
  - **Variance of a Gamma(a,b) distribution is ab^2 (so variance of an Exponential(10) distribution is 0.01)**

10 important considerations

Top 10 List
(of important considerations)

1. Beware of arbitrarily truncated priors
2. Branch length priors particularly important
3. Beware of high posteriors for very short branch lengths
4. Partition with care (prefer fewer subsets) and run MCMC for longer
5. MCMC run length should depend on number of parameters
6. Pay attention to parameter estimates
7. Pay attention to the behavior of the MCMC for ALL parameters
8. Run without data to explore prior
9. Run long
10. Run several times and compare runs

3. Branch length priors

Syst. Biol. 59(1):108–117, 2010

DOI:10.1093/sysbio/syp080
Advance Access publication on November 17, 2009

Cryptic Failure of Partitioned Bayesian Phylogenetic Analyses: Lost in the Land of Long Trees

DAVID C. MARSHALL*

Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, U-3043, Storrs, CT 06269, USA;
**Correspondence to be sent to: Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, U-3043, Storrs, CT 06269, USA; E-mail: david.marshall@uconn.edu.*

Syst. Biol. 59(2):145–161, 2010

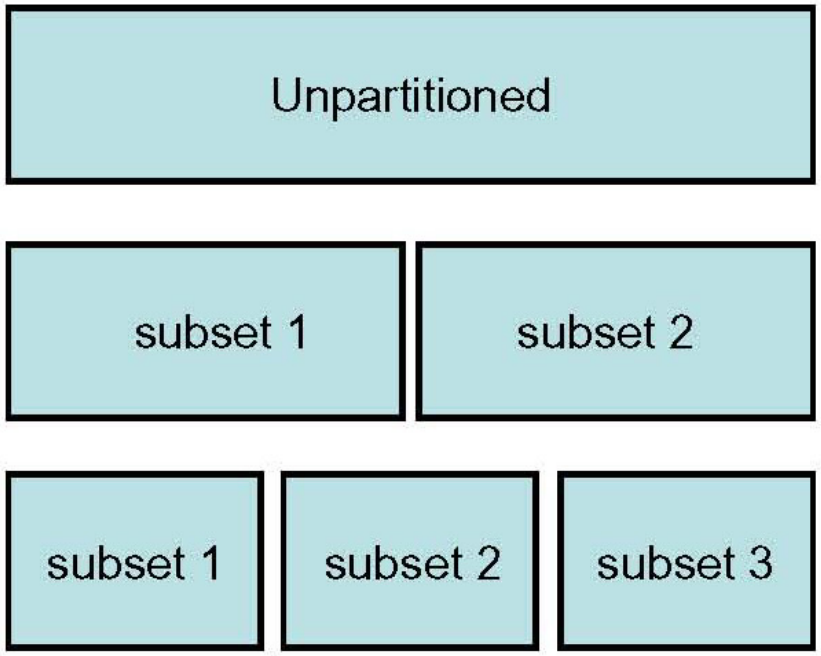
DOI:10.1093/sysbio/syp081
Advance Access publication on December 10, 2009

When Trees Grow Too Long: Investigating the Causes of Highly Inaccurate Bayesian Branch-Length Estimates

JEREMY M. BROWN[1,2,*], SHANNON M. HEDTKE[1], ALAN R. LEMMON[3,4], AND EMILY MORIARTY LEMMON[3,5]

4: Partition with care (prefer fewer subsets)
5: MCMC run length should depend on number of parameters

100 taxa, 2000 sites, F81 model



197 branch lengths, 3 base frequencies = 200 parameters
10 sites/parameter

394 branch lengths, 6 base frequencies = 400 parameters
5 sites/parameter

591 branch lengths, 9 base frequencies = 600 parameters
3.3 sites/parameter

- Partitioning reduces information for estimating some model parameters.
- Might want to run 3-subset case 3 times longer than the unpartitioned case.

Some Bayesian programs

- **MrBayes**
- **ExaBayes**
- **BEAST**

Maximum Likelihood and Bayesian methods: summary

- Both methods are very popular in molecular systematics
- Maximum likelihood is the most important method in phylogenomics
- Bayesian methods are able to take into account uncertainty in parameter estimates
- Bayesian methods can relax the assumption of a homogenous Markov model for rates of change in a tree