

# Phylogenetics:

## Bayesian Phylogenetic Analysis

---

COMP 571 - Spring 2015  
Luay Nakhleh, Rice University



# Bayes Rule

$$\mathbf{P}(X = x|Y = y) = \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)} = \frac{\mathbf{P}(X = x)\mathbf{P}(Y = y|X = x)}{\sum_{x'} \mathbf{P}(X = x')\mathbf{P}(Y = y|X = x')}$$



# Bayes Rule

- \* Example (from "Machine Learning: A Probabilistic Perspective")
- \* Consider a woman in her 40s who decides to have a mammogram.
- \* Question: If the test is positive, what is the probability that she has cancer?
- \* The answer depends on how reliable the test is!



# Bayes Rule

- \* Suppose the test has a sensitivity of 80%; that is, if a person has cancer, the test will be positive with probability 0.8.
- \* If we denote by  $x=1$  the event that the mammogram is positive, and by  $y=1$  the event that the person has breast cancer, then  $P(x=1|y=1)=0.8$ .



# Bayes Rule

- \* Does the probability that the woman in our example (who tested positive) has cancer equal 0.8?



# Bayes Rule

- \* No!
- \* That ignores the prior probability of having breast cancer, which, fortunately, is quite low:  $p(y=1)=0.004$



# Bayes Rule

- \* Further, we need to take into account the fact that the test may be a false positive.
- \* Mammograms have a false positive probability of  $p(x=1|y=0)=0.1$ .



# Bayes Rule

\* Combining all these facts using Bayes rule, we get (using  $p(y=0)=1-p(y=1)$ ):

$$\begin{aligned} p(y = 1 | x = 1) &= \frac{p(x=1|y=1)p(y=1)}{p(x=1|y=1)p(y=1) + p(x=1|y=0)p(y=0)} \\ &= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} \\ &= 0.031 \end{aligned}$$

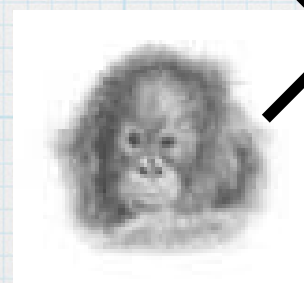
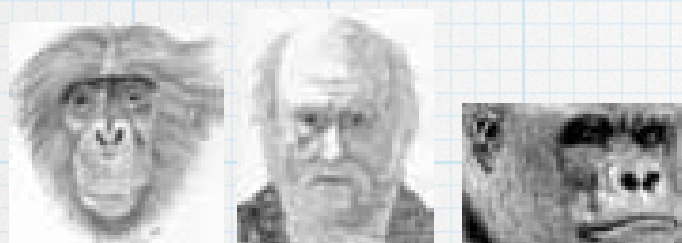


**\* How does Bayesian reasoning apply to phylogenetic inference?**

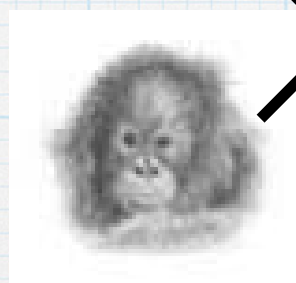
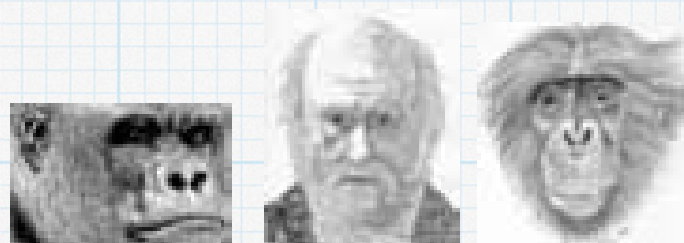


- \* Assume we are interested in the relationships between human, gorilla, and chimpanzee (with orangutan as an outgroup).
- \* There are clearly three possible relationships.

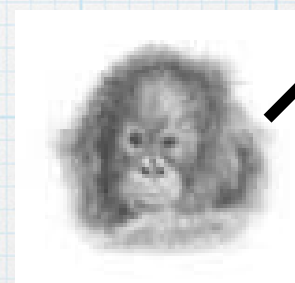
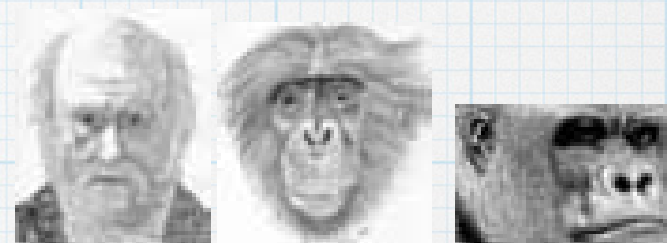




A



B

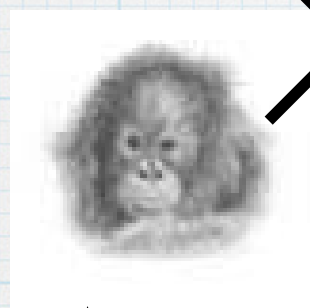
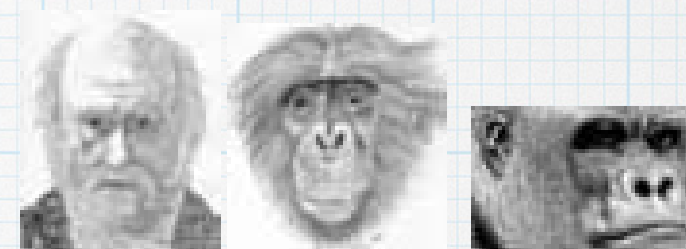
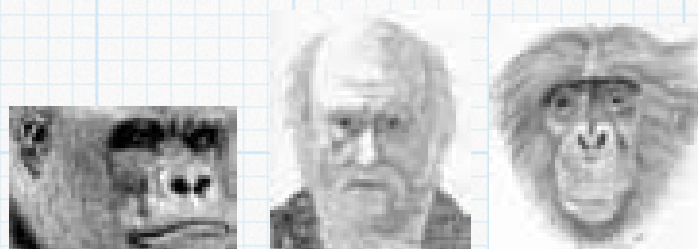
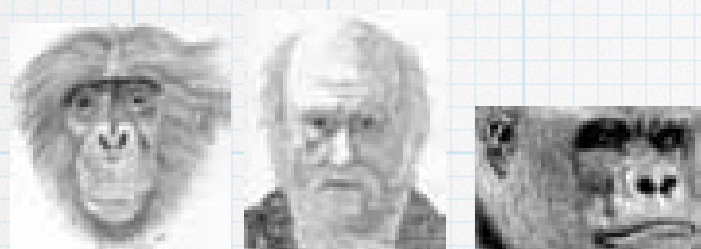


C

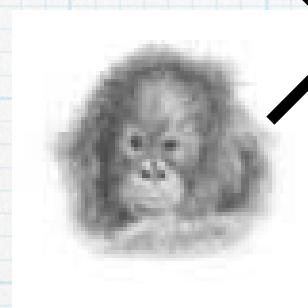


- \* Before the analysis, we need to specify our prior beliefs about the relationships.
- \* For example, in the absence of background data, a simple solution would be to assign equal probability to the possible trees.





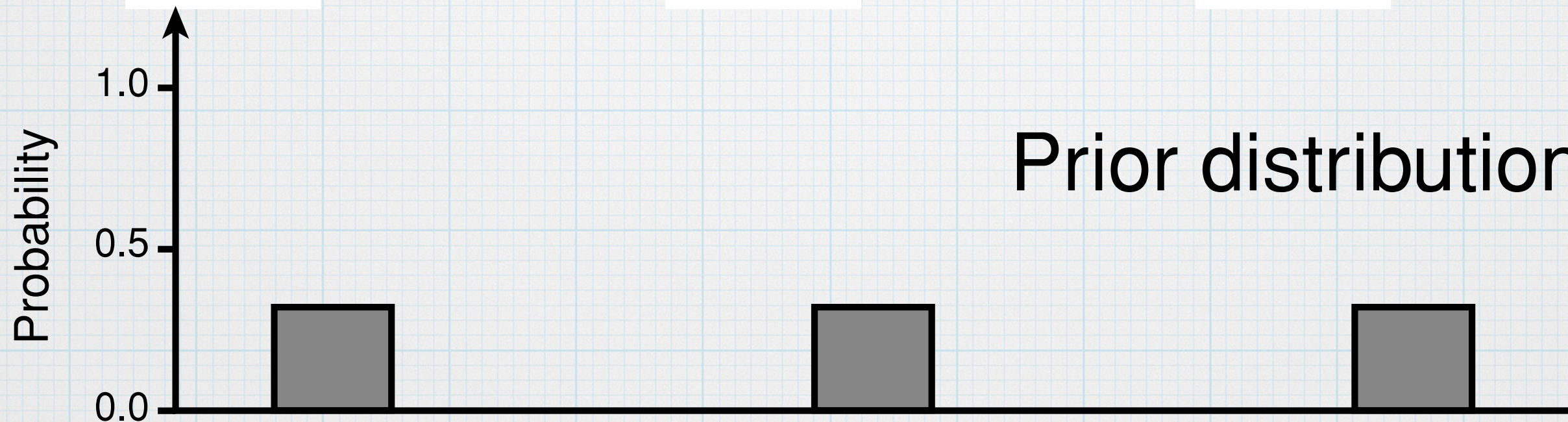
A



B



C



[This is an uninformative prior]



- \* To update the prior, we need some **data**, typically in the form of a **molecular sequence alignment**, and a **stochastic model of the process generating the data on the tree**.

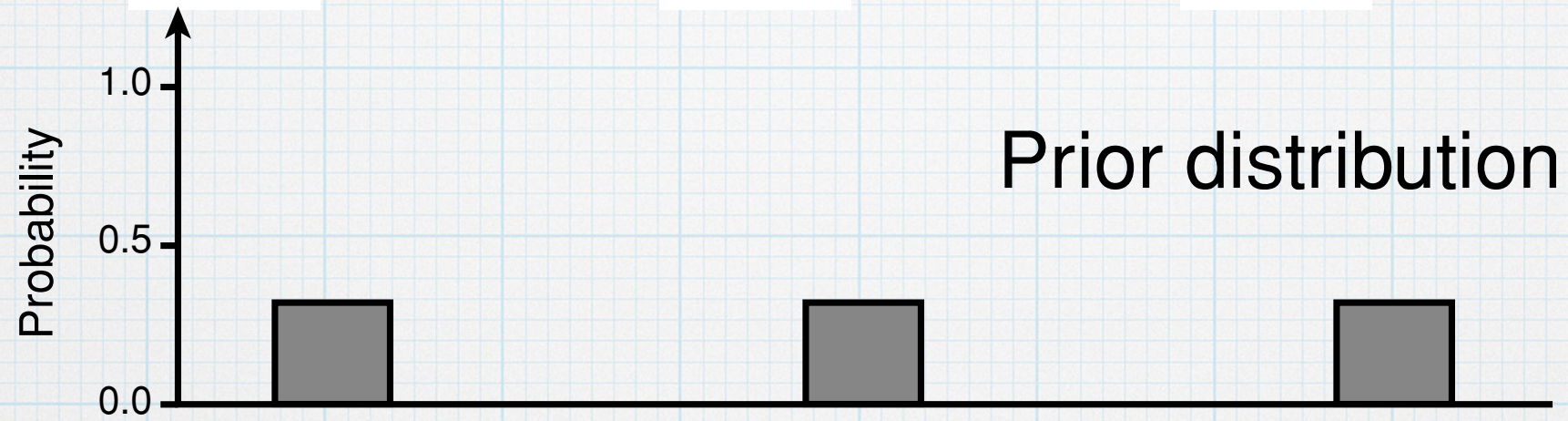
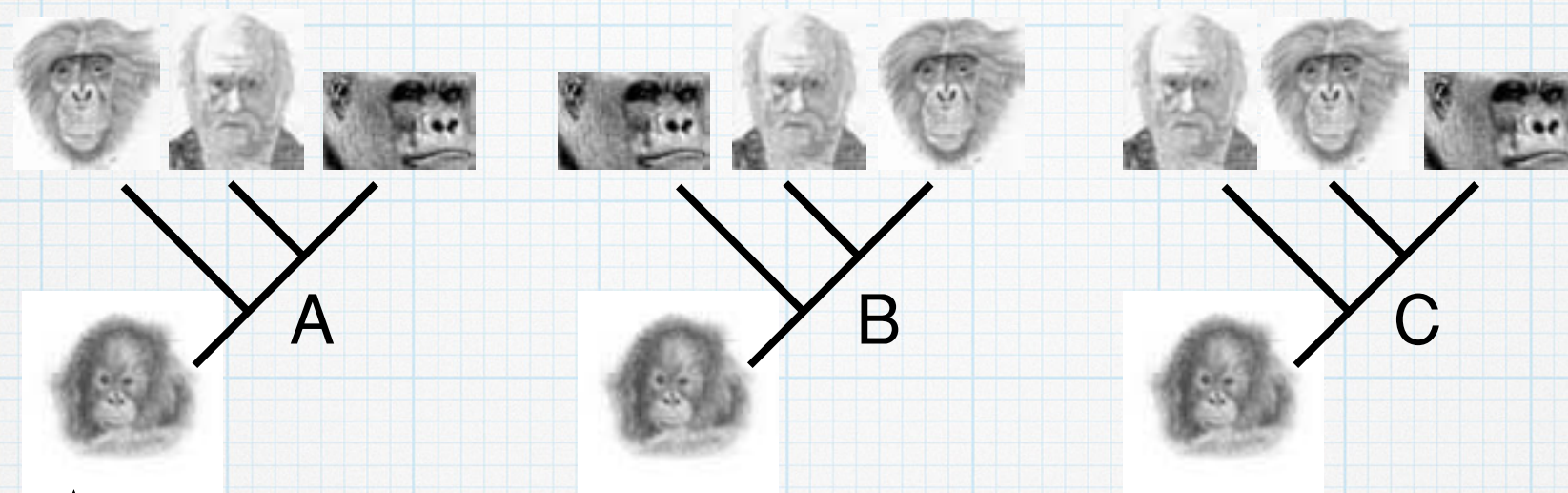


- \* In principle, Bayes rule is then used to obtain the posterior probability distribution, which is the result of the analysis.
- \* The posterior specifies the probability of each tree given the model, the prior, and the data.



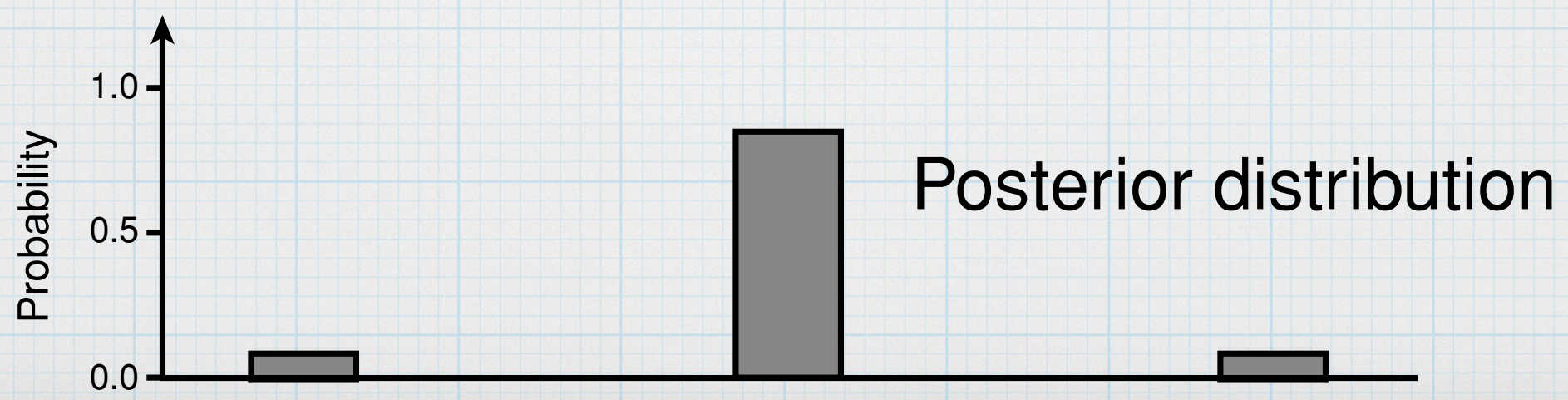
- \* When the data are informative, most of the posterior probability is typically concentrated on one tree (or, a small subset of trees in a large tree space).





Data (observations)

Two downward-pointing arrows indicate the flow from the prior distribution to the posterior distribution, with the text 'Data (observations)' centered between them.





- \* To describe the analysis mathematically, consider:
- \* the matrix of aligned sequences  $X$
- \* the tree topology parameter  $\tau$
- \* the branch lengths of the tree  $v$
- \* (typically, substitution model parameters are also included)

Let  $\theta = (\tau, v)$



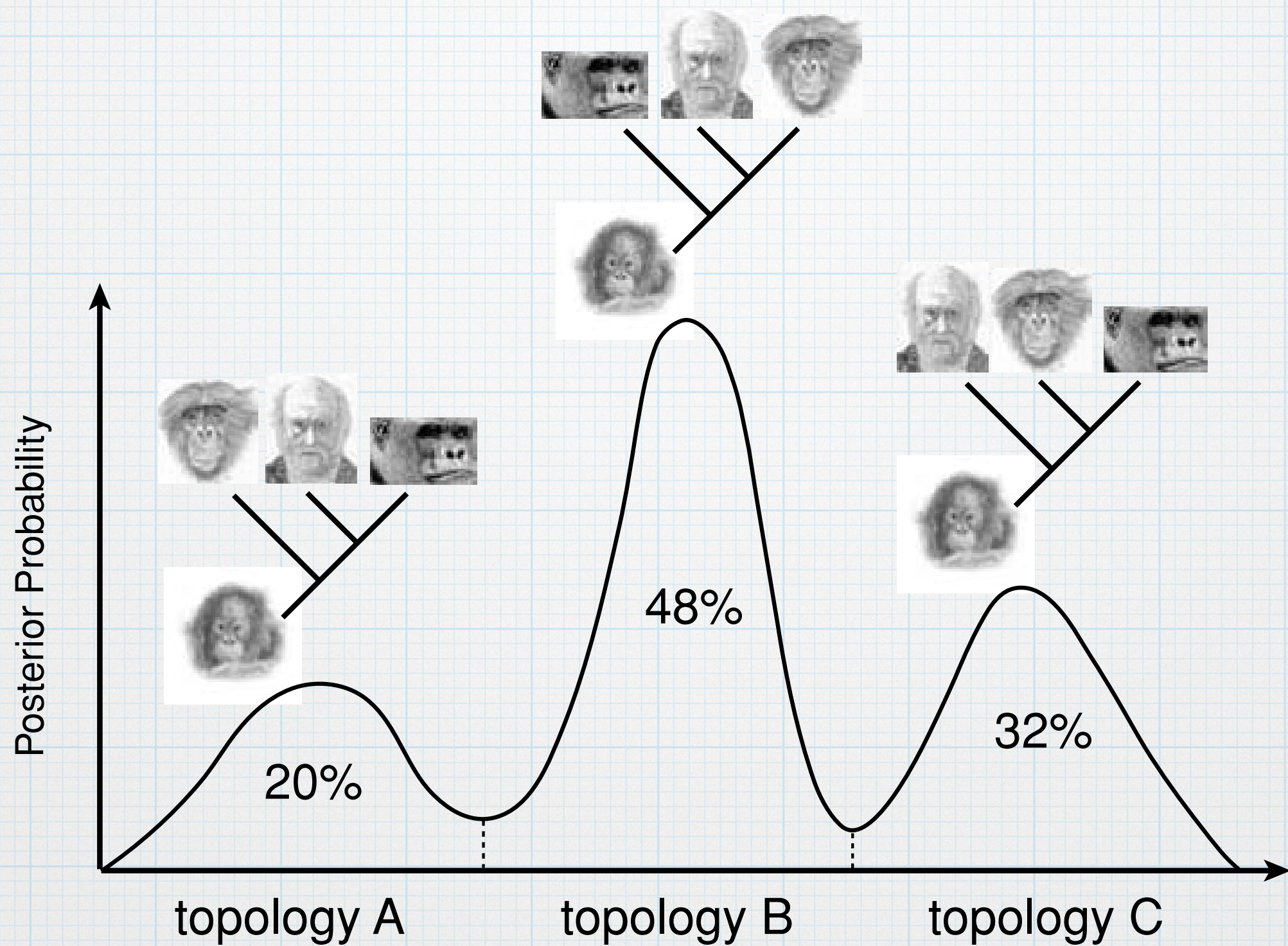
\* Bayes theorem allows us to derive the posterior distribution as

$$f(\theta|X) = \frac{f(\theta)f(X|\theta)}{f(X)}$$

where

$$\begin{aligned} f(X) &= \int f(\theta) f(X|\theta) d\theta \\ &= \sum_{\tau} \int_v f(v) f(X|\tau, v) dv \end{aligned}$$





**The marginal probability distribution on topologies**



# \* Why are they called marginal probabilities?

		Topologies			Joint probabilities
		$\tau_A$	$\tau_B$	$\tau_C$	
Branch length vectors	$\mathbf{v}^A$	0.10	0.07	0.12	0.29
	$\mathbf{v}^B$	0.05	0.22	0.06	0.33
	$\mathbf{v}^C$	0.05	0.19	0.14	0.38
		0.20	0.48	0.32	Marginal probabilities



# Markov chain Monte Carlo Sampling



- \* In most cases, it is impossible to derive the posterior probability distribution analytically.
- \* Even worse, we can't even estimate it by drawing random samples from it.
- \* The reason is that most of the posterior probability is likely to be concentrated in a small part of a vast parameter space.



- \* The solution is to estimate the posterior probability distribution using **Markov chain Monte Carlo sampling**, or **MCMC** for short.



- \* Markov chains have the property that they converge towards an equilibrium state regardless of starting point.
- \* We just need to set up a Markov chain that converges onto our posterior probability distribution!

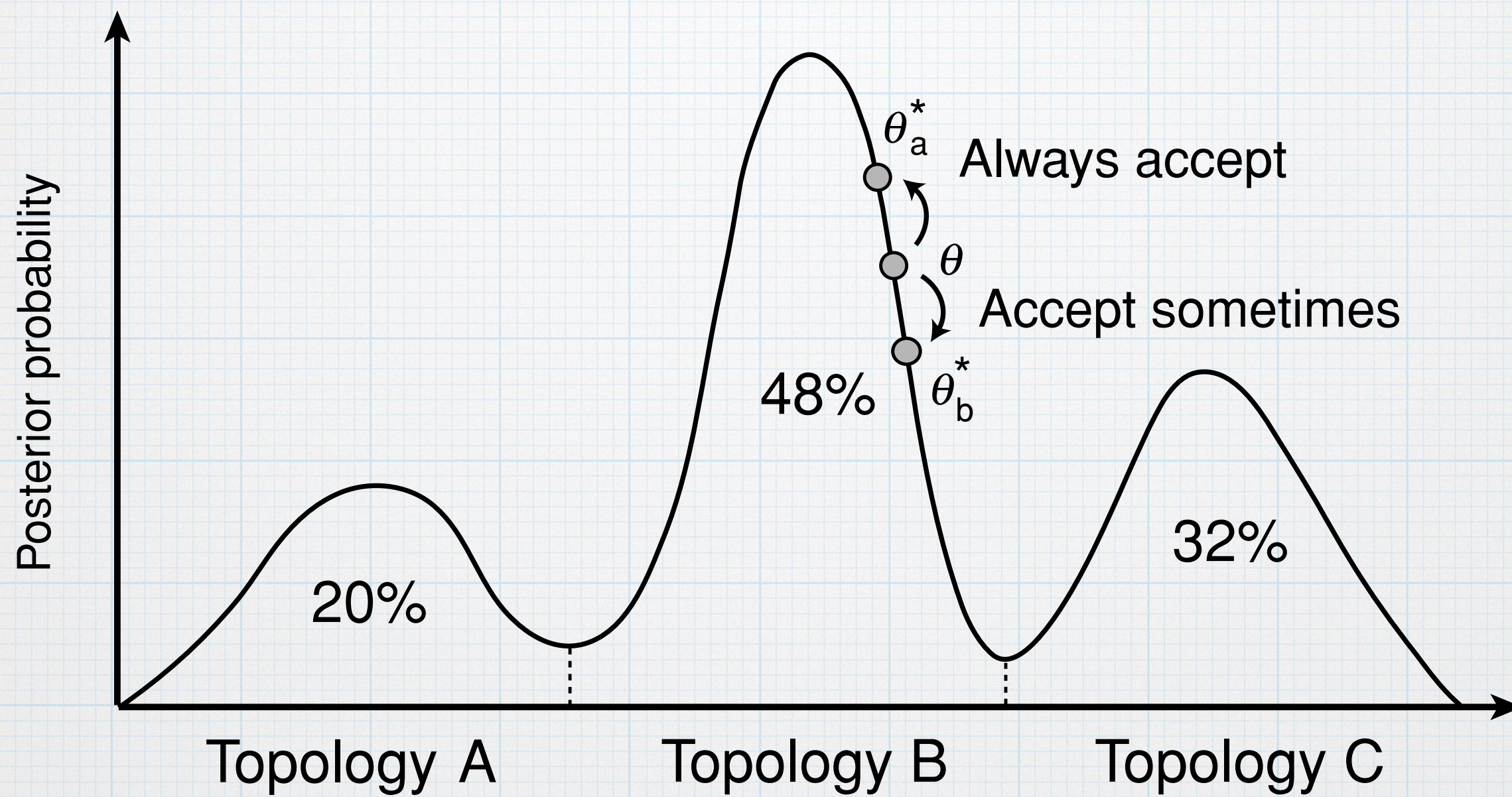


- \* This can be achieved using several different methods, the most flexible of which is known as the **Metropolis algorithm** and its extension, the **Metropolis-Hastings method**.



- \* The central idea is to make small random changes to some current parameter values, and then accept or reject the changes according to the appropriate probabilities







## Markov chain Monte Carlo steps

1. Start at an arbitrary point ( $\theta$ )
2. Make a small random move (to  $\theta^*$ )
3. Calculate height ratio ( $r$ ) of new state (to  $\theta^*$ ) to old state ( $\theta$ )
  - (a)  $r > 1$ : new state accepted
  - (b)  $r < 1$ : new state accepted with probability  $r$   
if new state rejected, stay in old state
4. Go to step 2



$$r = \min \left( 1, \frac{f(\theta^*|X)}{f(\theta|X)} \times \frac{f(\theta|\theta^*)}{f(\theta^*|\theta)} \right)$$

$$= \min \left( 1, \frac{f(\theta^*) f(X|\theta^*)/f(X)}{f(\theta) f(X|\theta)/f(X)} \times \frac{f(\theta|\theta^*)}{f(\theta^*|\theta)} \right)$$

$$= \min \left( 1, \frac{f(\theta^*)}{f(\theta)} \times \frac{f(X|\theta^*)}{f(X|\theta)} \times \frac{f(\theta|\theta^*)}{f(\theta^*|\theta)} \right)$$

**prior  
ratio**

**likelihood  
ratio**

**proposal  
ratio**

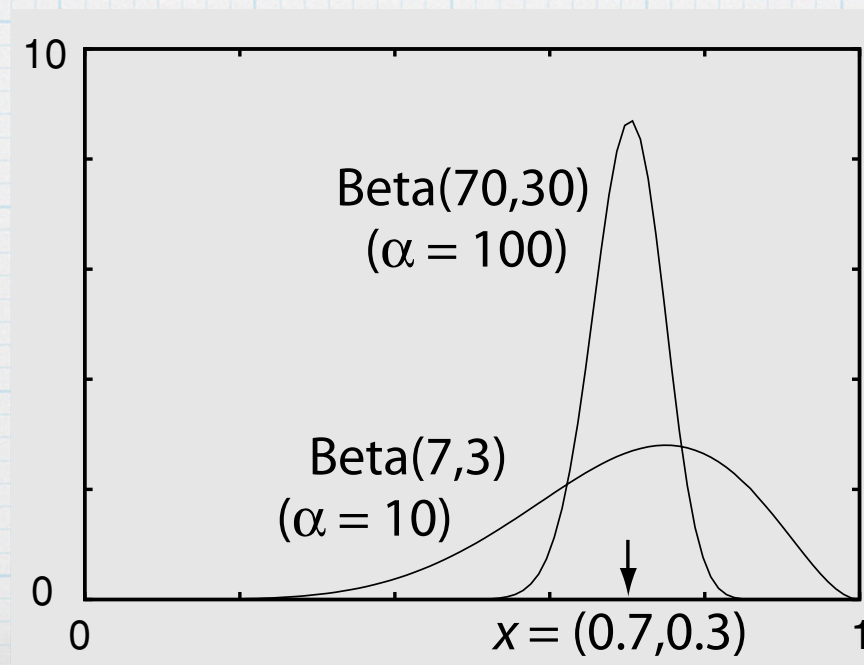


- \* An example of a proposal mechanism is the **beta proposal**:

- \* Assume the current values are  $(x_1, x_2)$ ;

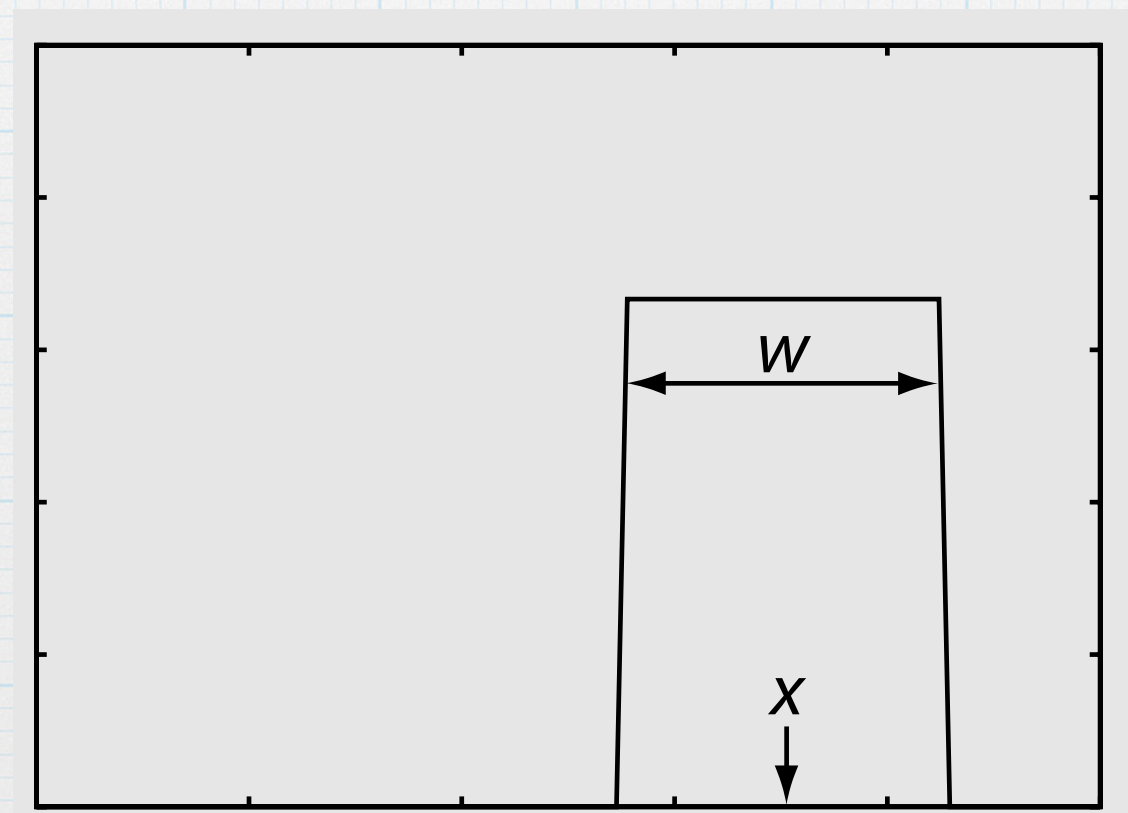
- \* Multiply them with a value  $\alpha$ ;

- \* Pick new values from  $\text{Beta}(\alpha x_1, \alpha x_2)$





- \* A simpler proposal mechanism is to define a continuous uniform distribution of width  $w$ , centered on the current value  $x$ , and the new value  $x^*$  is drawn from this distribution.





- \* More complex moves are needed to change tree topology.
- \* A common type uses operations such as SPR, TBR, and NNI.



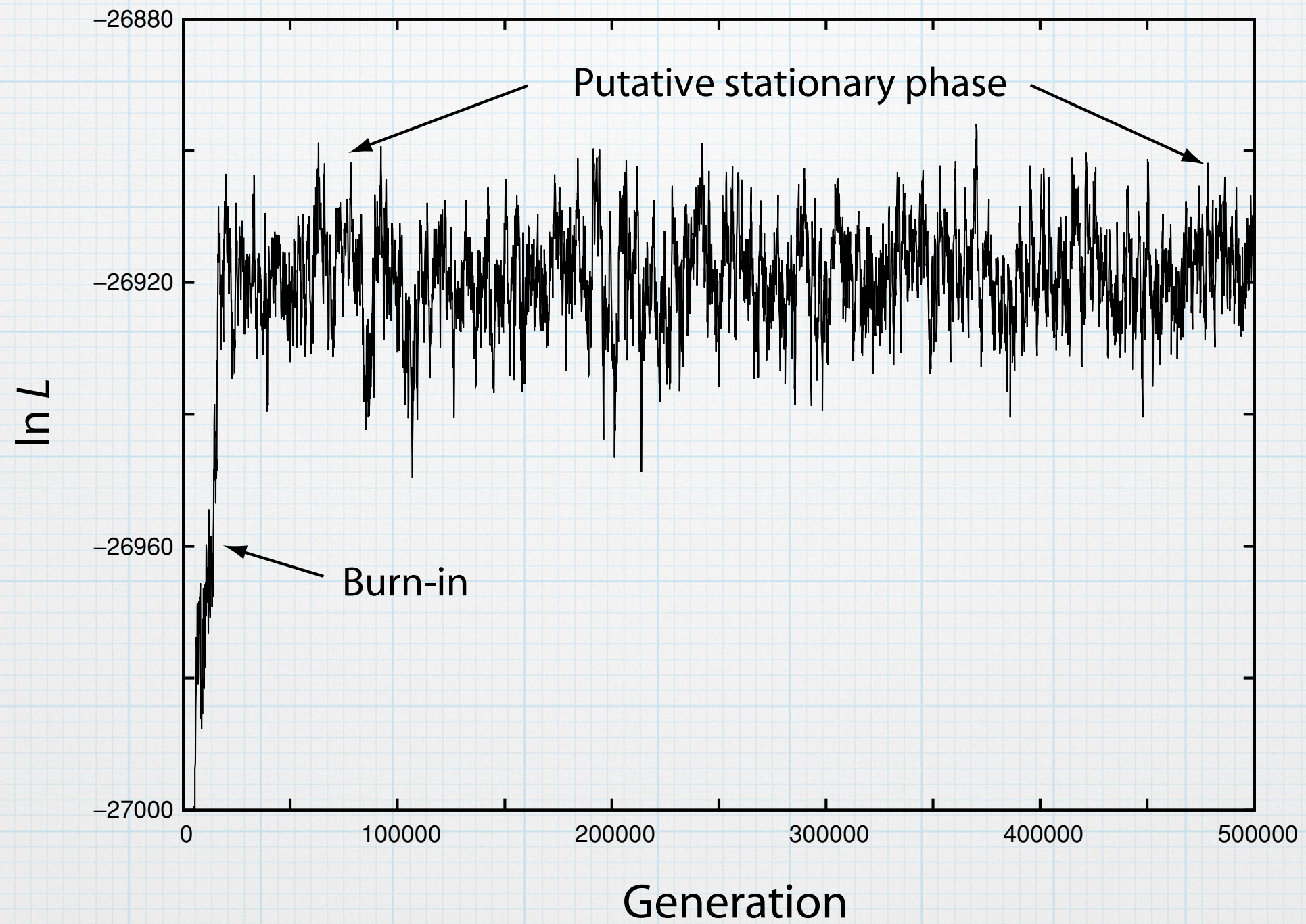
# Burn-in, mixing, and convergence



- \* If the chain is started from a random tree and arbitrarily chosen branch lengths, chances are that the initial likelihood is low.
- \* The early phase of the run in which the likelihood increases very rapidly towards regions in the posterior with high probability mass is known as the **burn-in**.

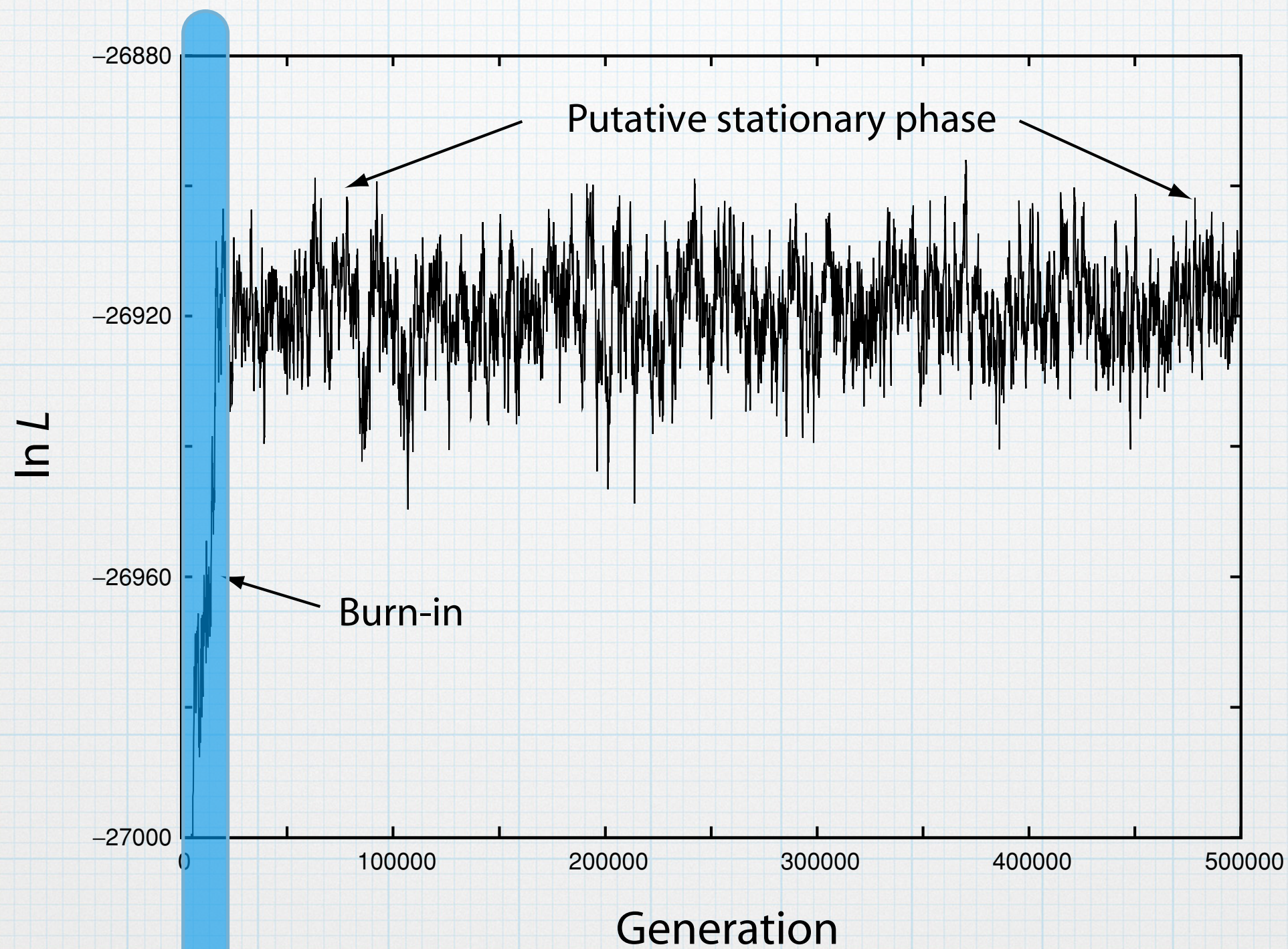


# Trace plot





# Trace plot



**samples in this region are discarded!**



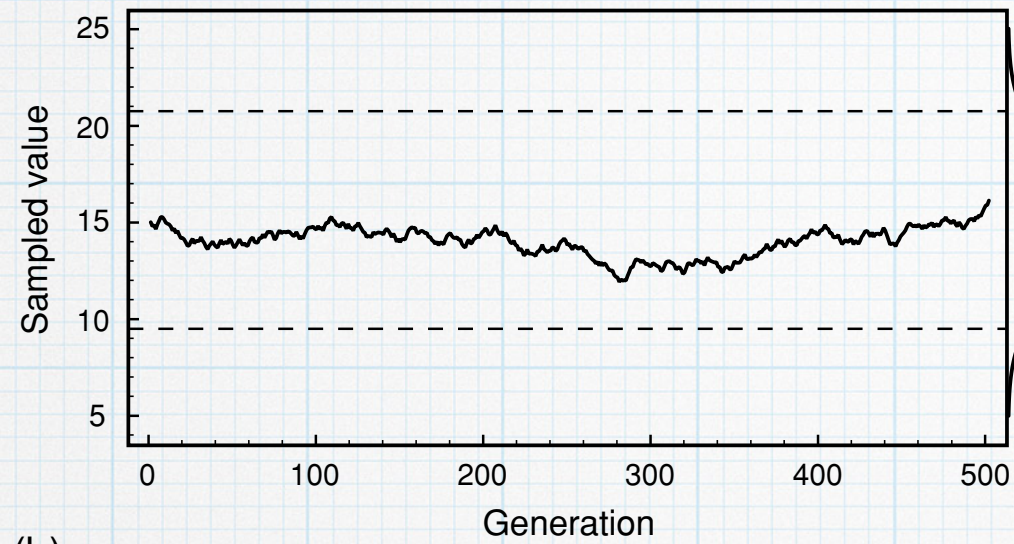
- \* As the chain approaches its stationary distribution, the likelihood values tend to reach a plateau.
- \* This is the first sign that the chain may have converged onto the target distribution.



- \* However, it is not sufficient for the chain to reach the region of high probability in the posterior; it must also cover this region adequately.
- \* The speed with which the chain covers the interesting regions of the posterior is known as its **mixing behavior**.
- \* The better the mixing, the faster the chain will generate an adequate sample of the posterior.

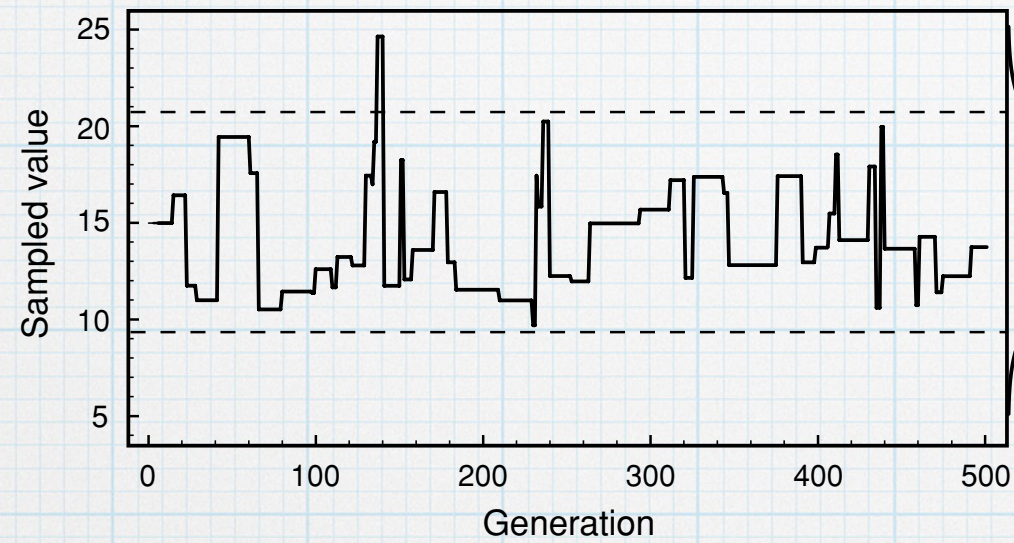


(a)



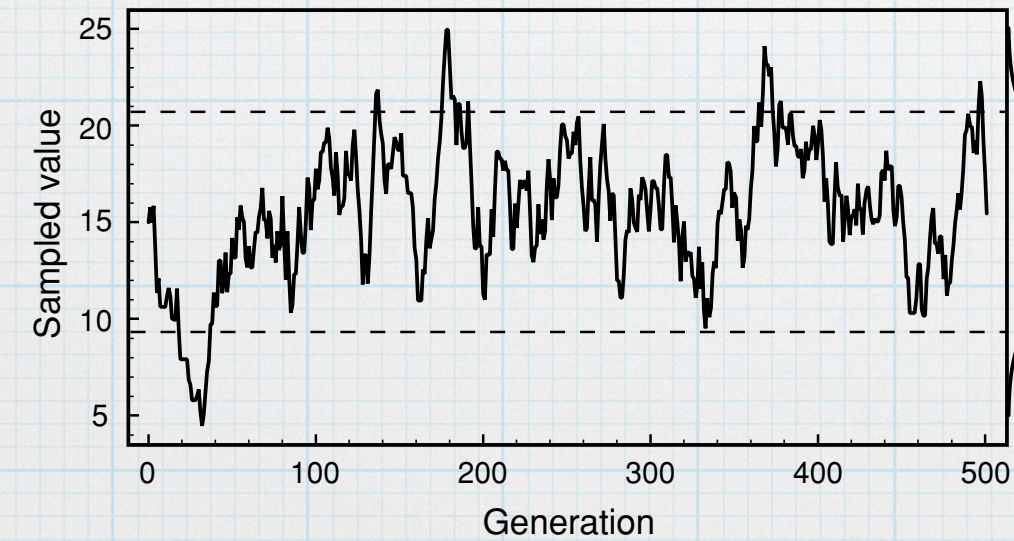
Too modest proposals  
Acceptance rate too high  
Poor mixing

(b)



Too bold proposals  
Acceptance rate too low  
Poor mixing

(c)



Moderately bold proposals  
Acceptance rate intermediate  
Good mixing



- \* In Bayesian MCMC sampling of phylogenetic problems, the tree topology is typically the most difficult parameter to sample from.
- \* Therefore, it makes sense to focus on this parameter when monitoring convergence.



**Summarizing the results**



- \* The stationary phase of the chain is typically sampled with some thinning, for instance every 50th or 100th generation.
- \* Once an adequate sample is obtained, it is usually trivial to compute an estimate of the marginal posterior distribution for the parameter(s) of interest.



- \* For example, this can take the form of a frequency histogram of the sampled values.
- \* When it is difficult to visualize this distribution or when space does not permit it, various summary statistics are used instead.



- \* The most common approach to summarizing topology posteriors is to give the frequencies of the most common splits, since there are much fewer splits than topologies.

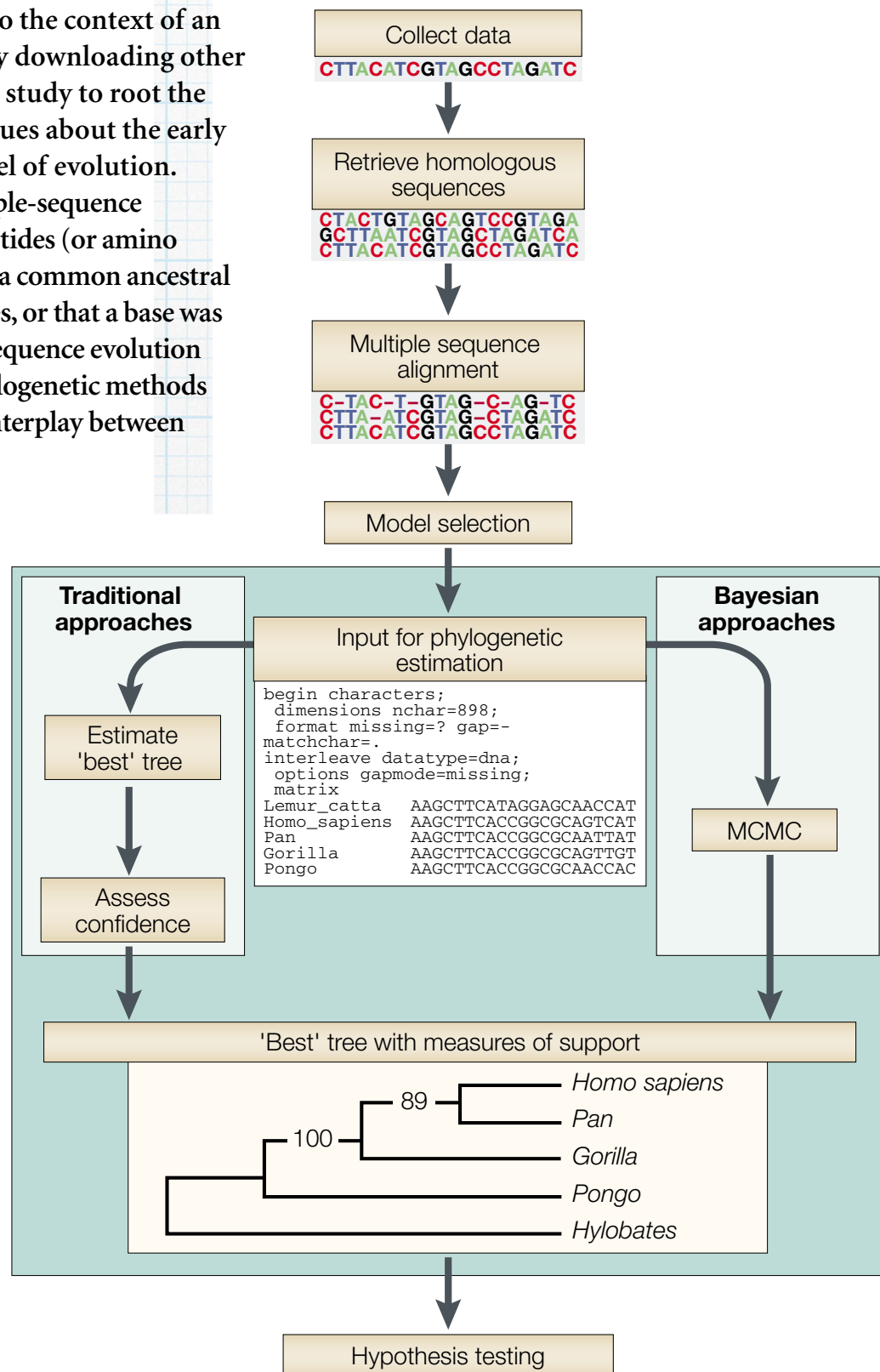


Source: Nat Rev Genet, 4:275, 2003

The flowchart puts phylogenetic estimation (shown in the green box) into the context of an entire study. After new sequence data are collected, the first step is usually downloading other relevant sequences. Typically, a few outgroup sequences are included in a study to root the tree (that is, to indicate which nodes in the tree are the oldest), provide clues about the early ancestral sequences and improve the estimates of parameters in the model of evolution.

In addition to the data, the scientist must choose a model of sequence evolution (even if this means just choosing a family of models and letting software infer the parameters of these models). Increasing model complexity improves the fit to the data but also increases variance in estimated parameters. Model selection<sup>60–63</sup> strategies attempt to find the appropriate level of complexity on the basis of the available data. Model complexity can often lead to computational intractability, so pragmatic concerns sometimes outweigh statistical ones (for example, NJ and parsimony are mainly justifiable by their speed).

As discussed in BOX 3, data and a model can be used to create a sample of trees through either Markov chain Monte Carlo (MCMC) or multiple tree searches on bootstrapped data (the ‘traditional’ approach). This collection of trees is often summarized using consensus-tree techniques, which show the parts of the tree that are found in most, or all, of the trees in a set. Although useful, CONSENSUS METHODS are just one way of summarizing the information in a group of trees. AGREEMENT SUBTREES are more resistant to ‘rogue sequences’ (one or a few sequences that are difficult to place on the tree); the presence of such sequences can make a consensus tree relatively unresolved, even when there is considerable agreement on the relationships between the other sequences. Sometimes, the bootstrap or MCMC sample might show substantial support for multiple trees that are not topologically similar. In such cases, presenting more than one tree (or more than one consensus of trees) might be the only way to appropriately summarize the data.





# Summary

Source: Nat Rev Genet, 4:275, 2003

Table 1   <b>Comparison of methods</b>			
<b>Method</b>	<b>Advantages</b>	<b>Disadvantages</b>	<b>Software</b>
Neighbour joining	Fast	Information is lost in compressing sequences into distances; reliable estimates of pairwise distances can be hard to obtain for divergent sequences	PAUP* MEGA PHYLIP
Parsimony	Fast enough for the analysis of hundreds of sequences; robust if branches are short (closely related sequences or dense sampling)	Can perform poorly if there is substantial variation in branch lengths	PAUP* NONA MEGA PHYLIP
Minimum evolution	Uses models to correct for unseen changes	Distance corrections can break down when distances are large	PAUP* MEGA PHYLIP
Maximum likelihood	The likelihood fully captures what the data tell us about the phylogeny under a given model	Can be prohibitively slow (depending on the thoroughness of the search and access to computational resources)	PAUP* PAML PHYLIP
Bayesian	Has a strong connection to the maximum likelihood method; might be a faster way to assess support for trees than maximum likelihood bootstrapping	The prior distributions for parameters must be specified; it can be difficult to determine whether the Markov chain Monte Carlo (MCMC) approximation has run for long enough	MrBayes BAMBE



# Acknowledgment

- \* Material in these slides are based on Chapter 7 in "The Phylogenetic Handbook", Lemey, Salemi, Vandamme (Eds.)