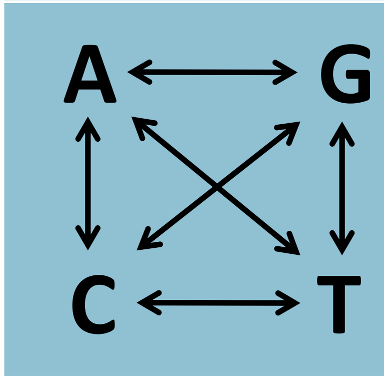
Lecture 2.1

Bayesian Phylogenetic Analysis

Phylogenetic methods

	Algorithm-based	Optimality criterion	Other
No explicit substitution model	Distance-based methods	Maximum parsimony	
	Distance-based methods	Maximum likelihood	Bayesian inference

The Bayesian framework

Bayesian phylogenetic analysis

- Bayesian phylogenetic analysis was developed in the mid 1990s
- Now one of the most widely used methods

MrBayes

BEAST 1

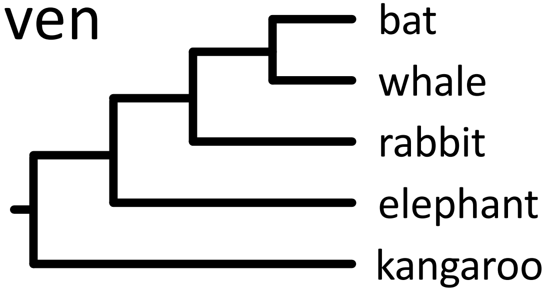
RevBayes

BEAST 2

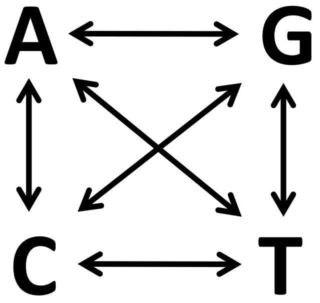
Bayesian phylogenetic analysis

Maximum likelihood

Given



+



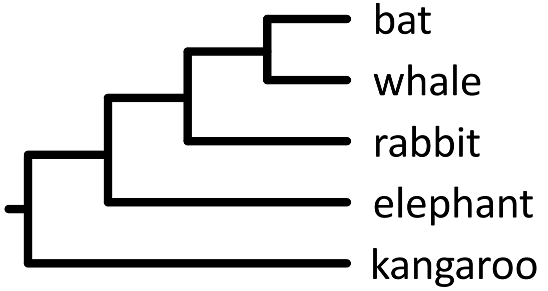
→

Probability of?

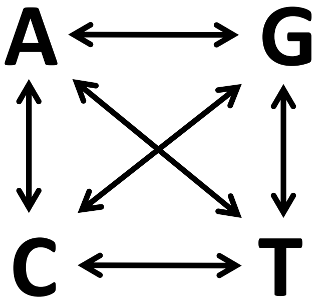
bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

Bayesian inference

Probability of?



+



←

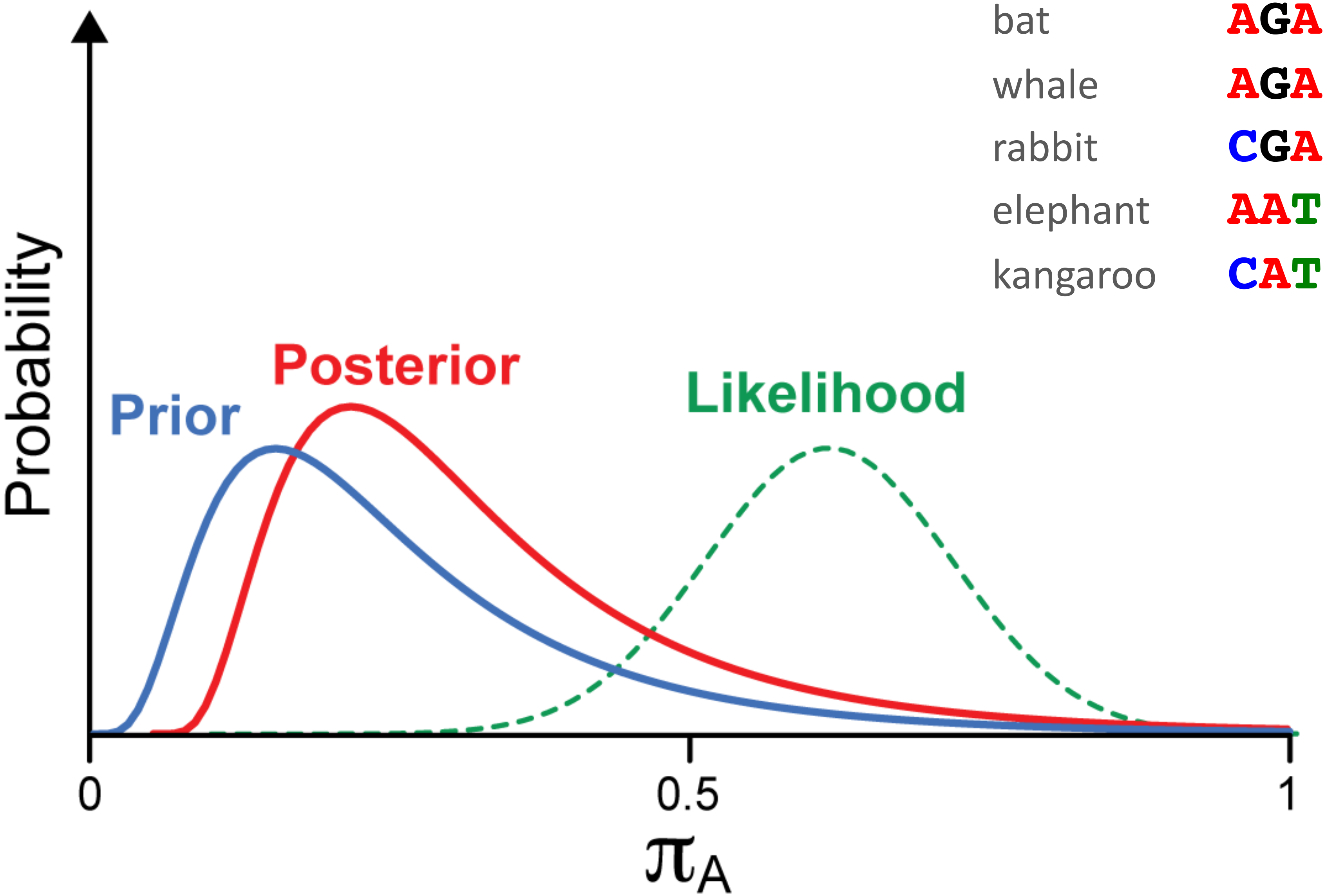
Given

bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

The Bayesian paradigm

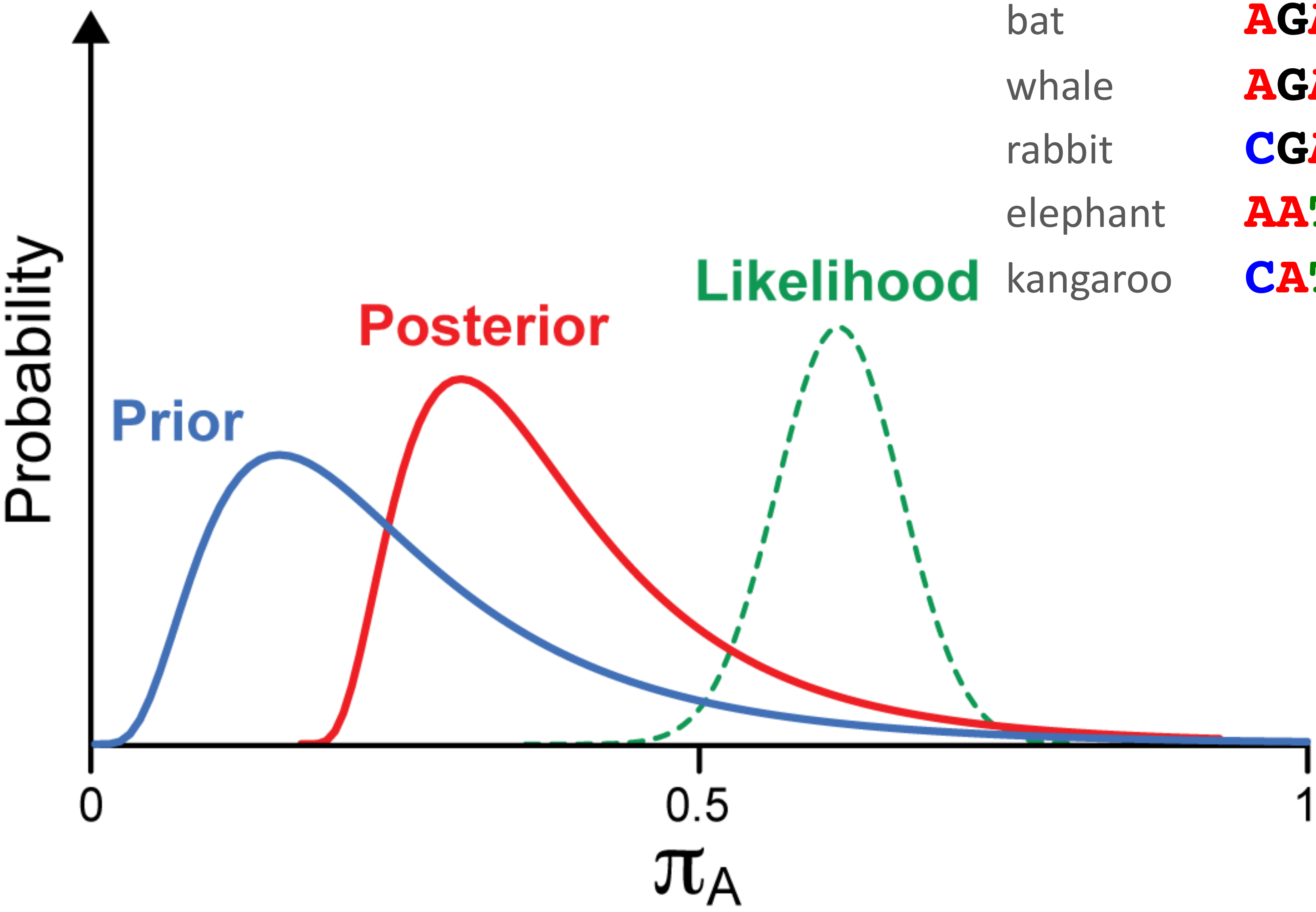
- Contrast with frequentist statistics (likelihood)
- Parameters have **distributions**
- Before the data are observed, each parameter has a **prior distribution**
- The **likelihood** of the data is computed
- The prior distribution is combined (updated) with the likelihood to yield the **posterior distribution**

Simple example



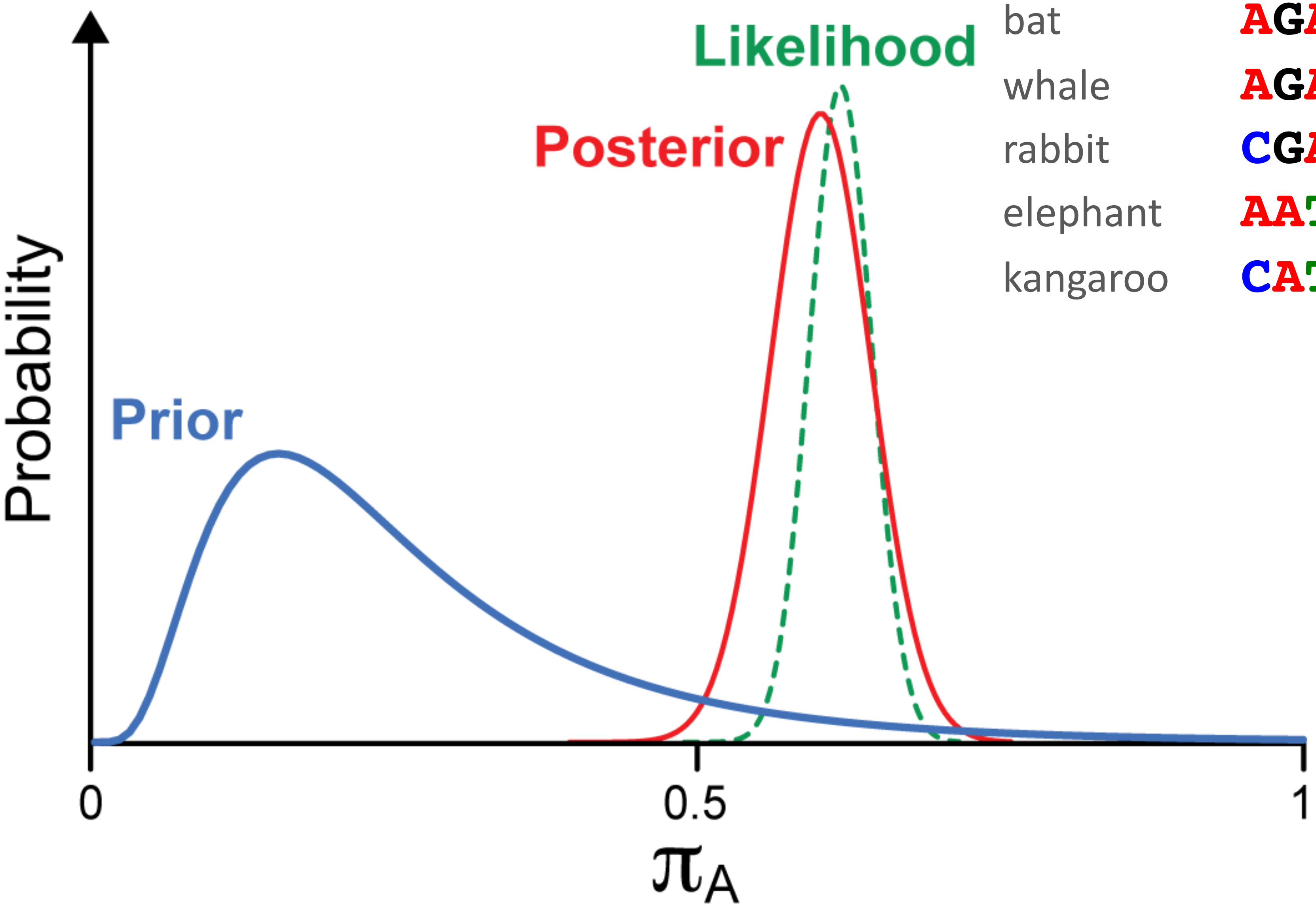
bat	AGA
whale	AGA
rabbit	CGA
elephant	AAT
kangaroo	CAT

Simple example



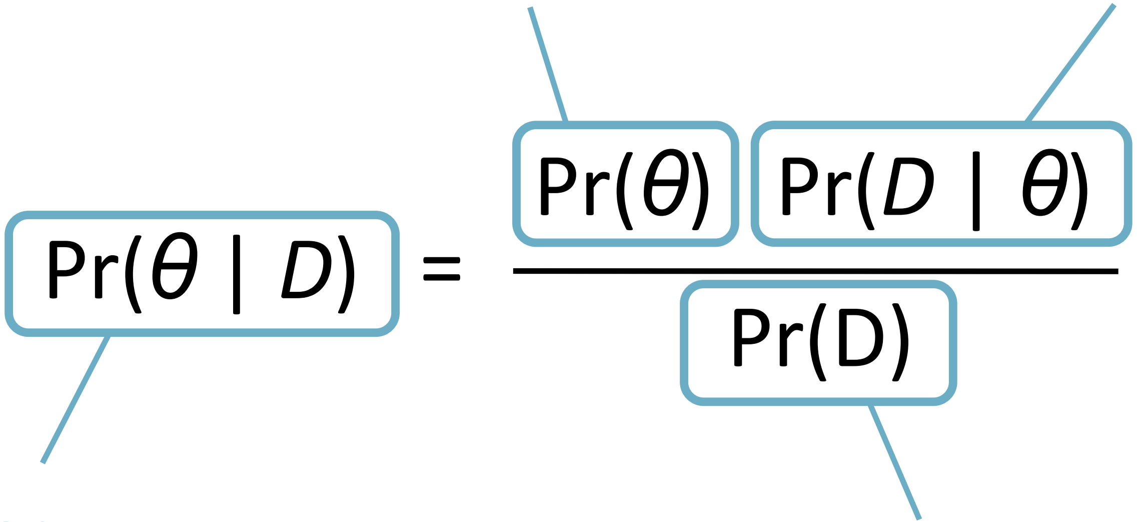
bat	AGAAA
whale	AGATA
rabbit	CGATA
elephant	AATGA
kangaroo	CATGA

Simple example



bat	AGAAACGATAAG
whale	AGATACTATAAG
rabbit	CGATACTAAACA
elephant	AATGAATAAACA
kangaroo	CATGACGAAGCA

Bayesian inference



Prior

Specified by user,
independent of data

Likelihood

Calculated from data

$$\boxed{\Pr(\theta \mid D)} = \frac{\boxed{\Pr(\theta)}\;\boxed{\Pr(D \mid \theta)}}{\boxed{\Pr(\mathrm{D})}}$$

Posterior

This is what we
want to estimate

normalising constant
marginal likelihood of the data
model likelihood

Bayesian inference

Prior prob of tree
Topology
Branch lengths

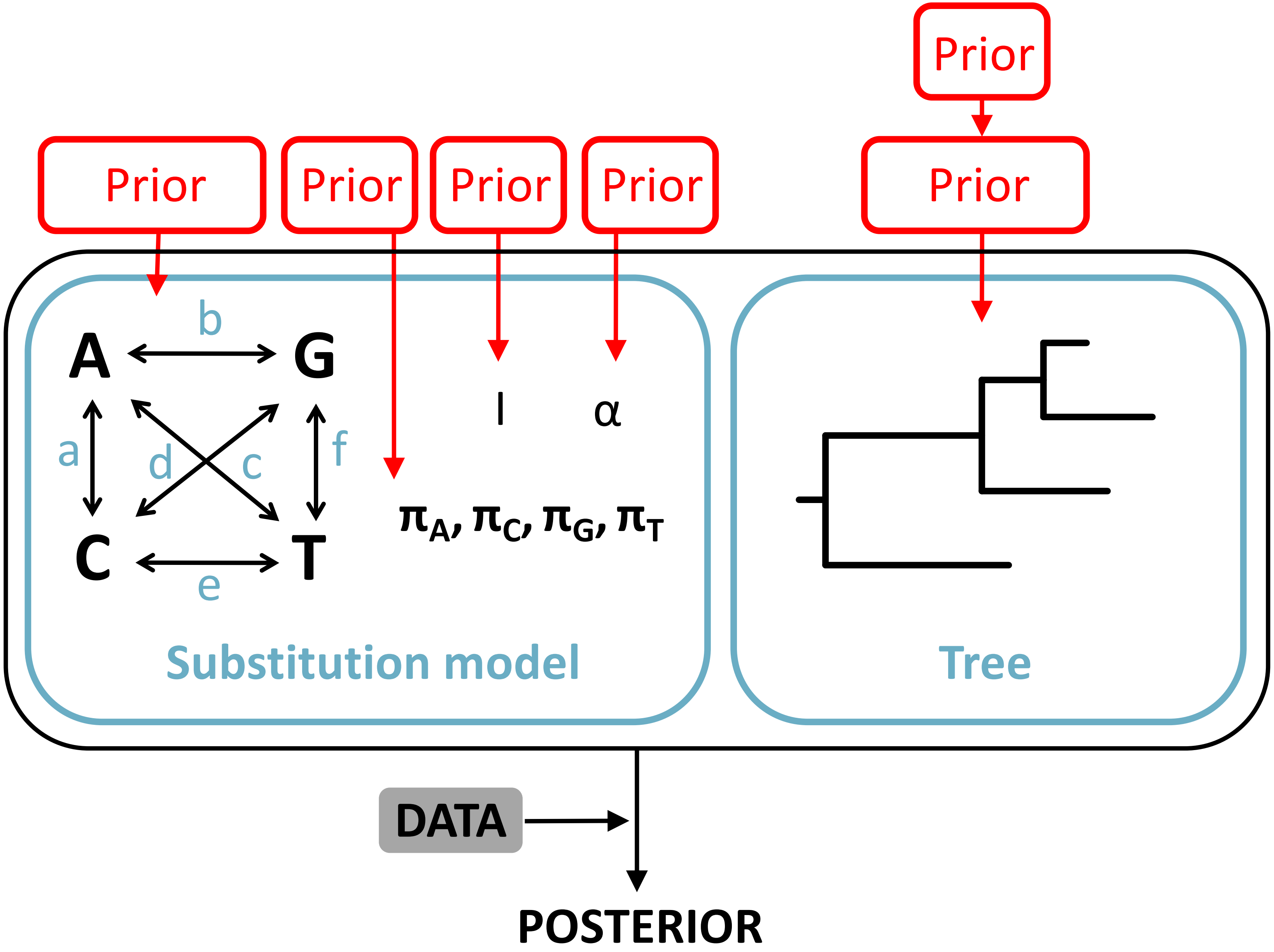
Prior prob of substitution model parameters
Rate parameters
Base frequencies

$$\boxed{\Pr(\tau,M \mid D)} = \frac{\boxed{\Pr(\tau)}\,\boxed{\Pr(M)}\,\boxed{\Pr(D \mid \tau,M)}}{\Pr(\mathrm{D})}$$

Posterior
This is what we want to estimate

Likelihood
Calculated from data

Bayesian hierarchical model



Priors

- Priors are chosen in the form of probability distributions
- Reflect our prior expectations (and uncertainty) about values of parameters (without knowledge of the data)
    - Past observations
    - Personal beliefs
    - Use of a biological model
- Uninformative priors

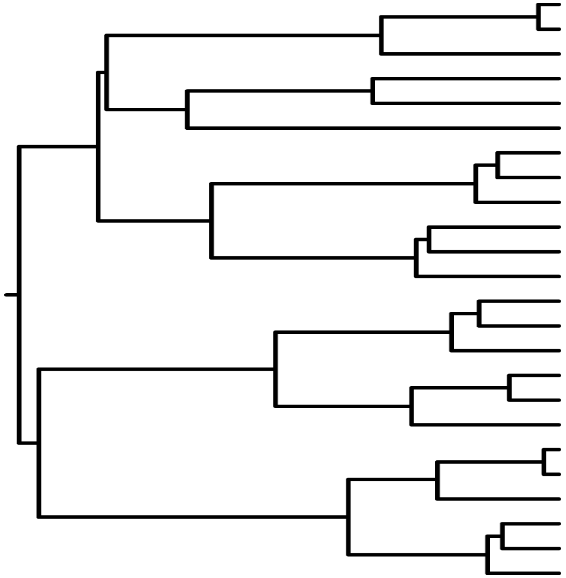
Priors

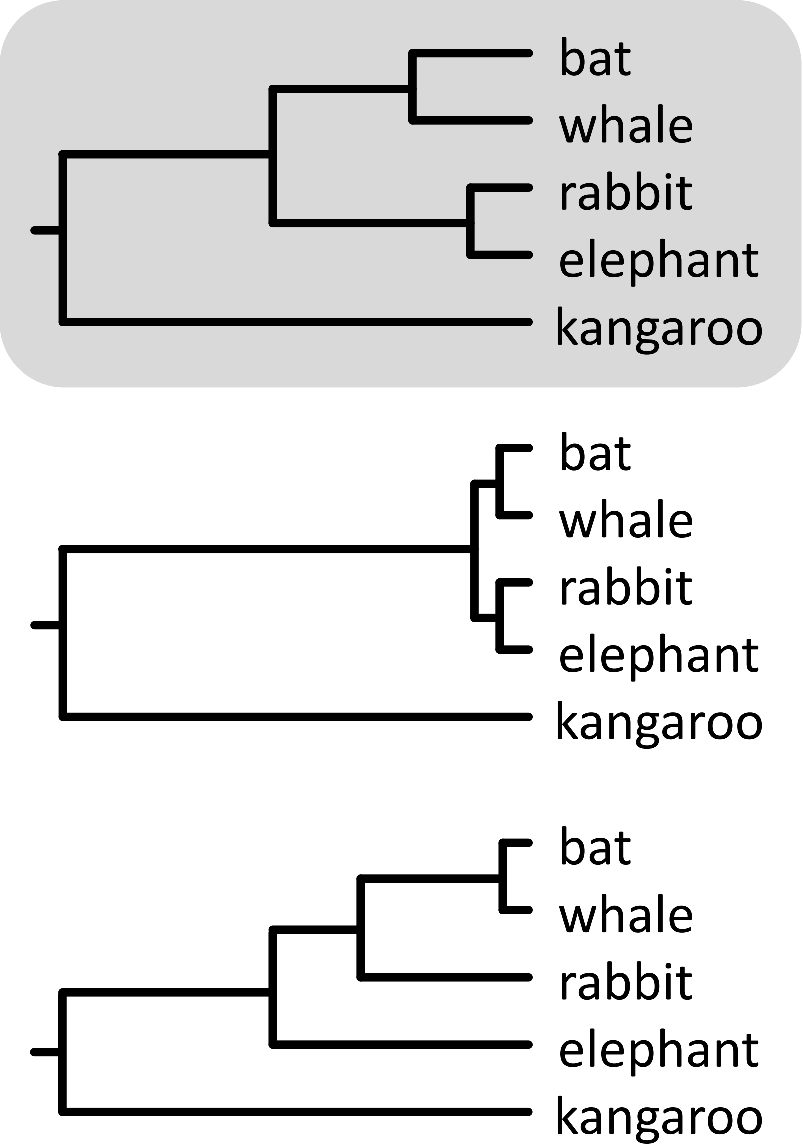
1. Use a **flat prior** for tree topology (*MrBayes*)
    - All trees have equal probability
    - Also need a prior for branch lengths or node times

2. Use a **biological model** to generate prior distribution (*BEAST* and *MrBayes*)
    - Among species: speciation model
    - Within species: coalescent model

Tree prior: Among species

- Tree shape described by a stochastic branching process
- Yule process
  - Lineages split at a constant rate
  - Simulates speciation process

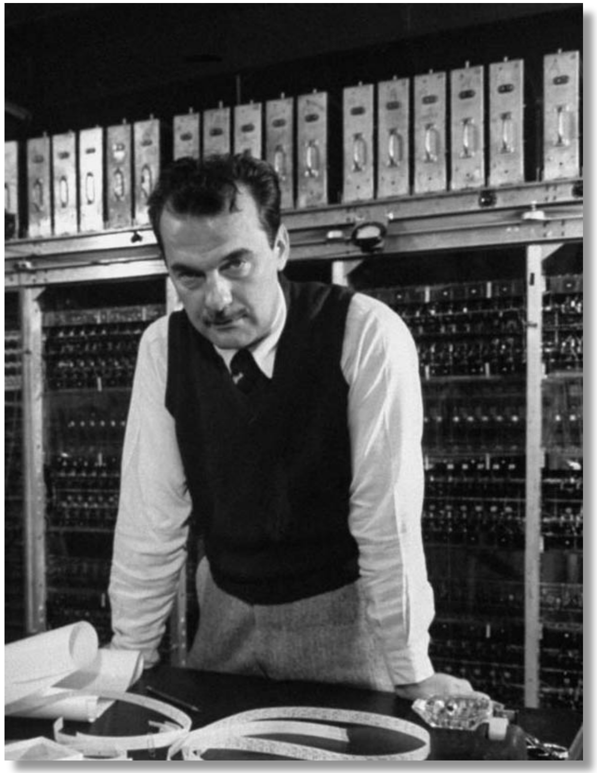




Markov Chain Monte Carlo Sampling

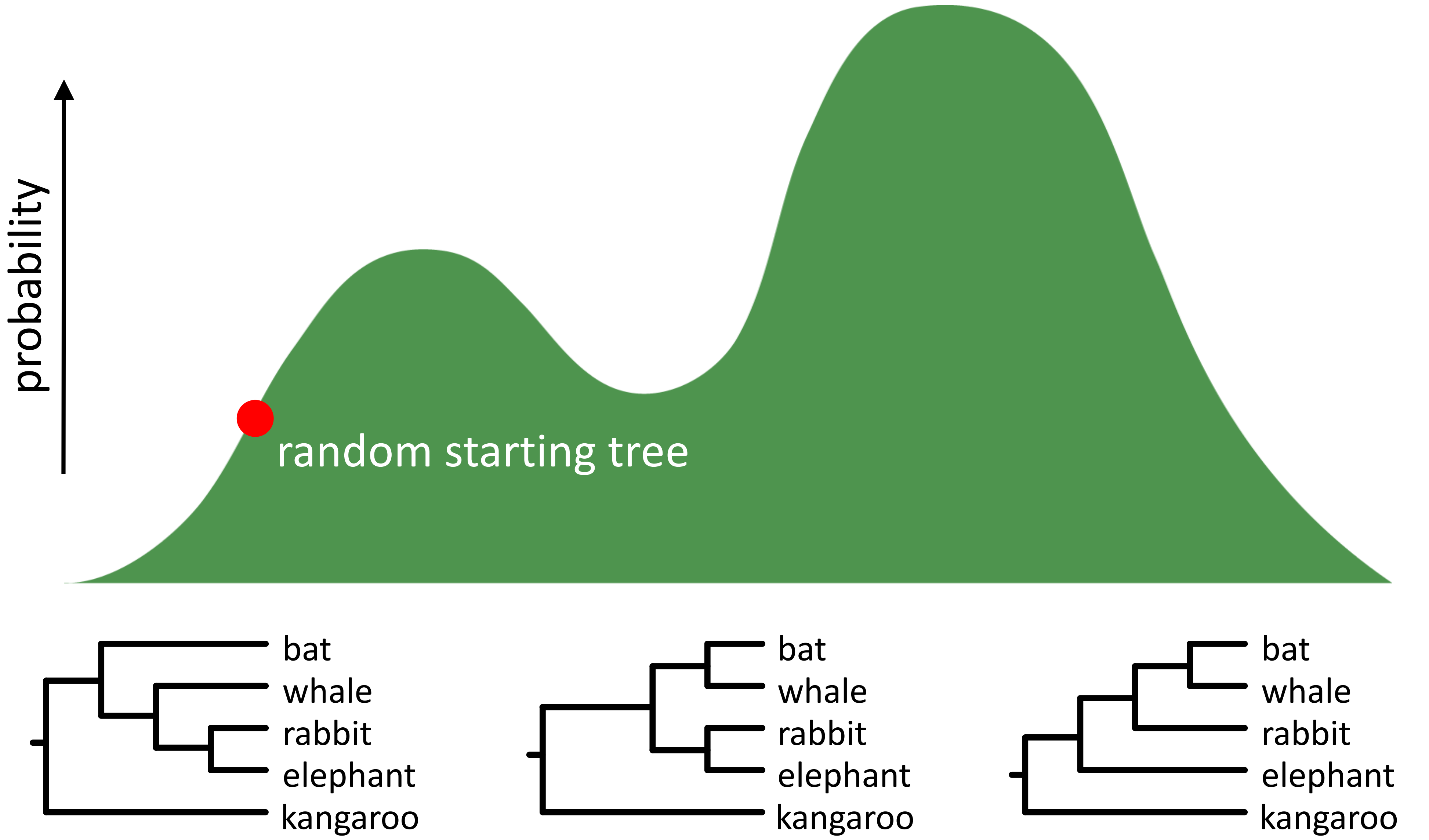
Estimating the posterior

- Impossible to obtain the posterior directly
- Instead, the posterior can be estimated using **Markov chain Monte Carlo simulation**
- This is usually done using the **Metropolis-Hastings algorithm**

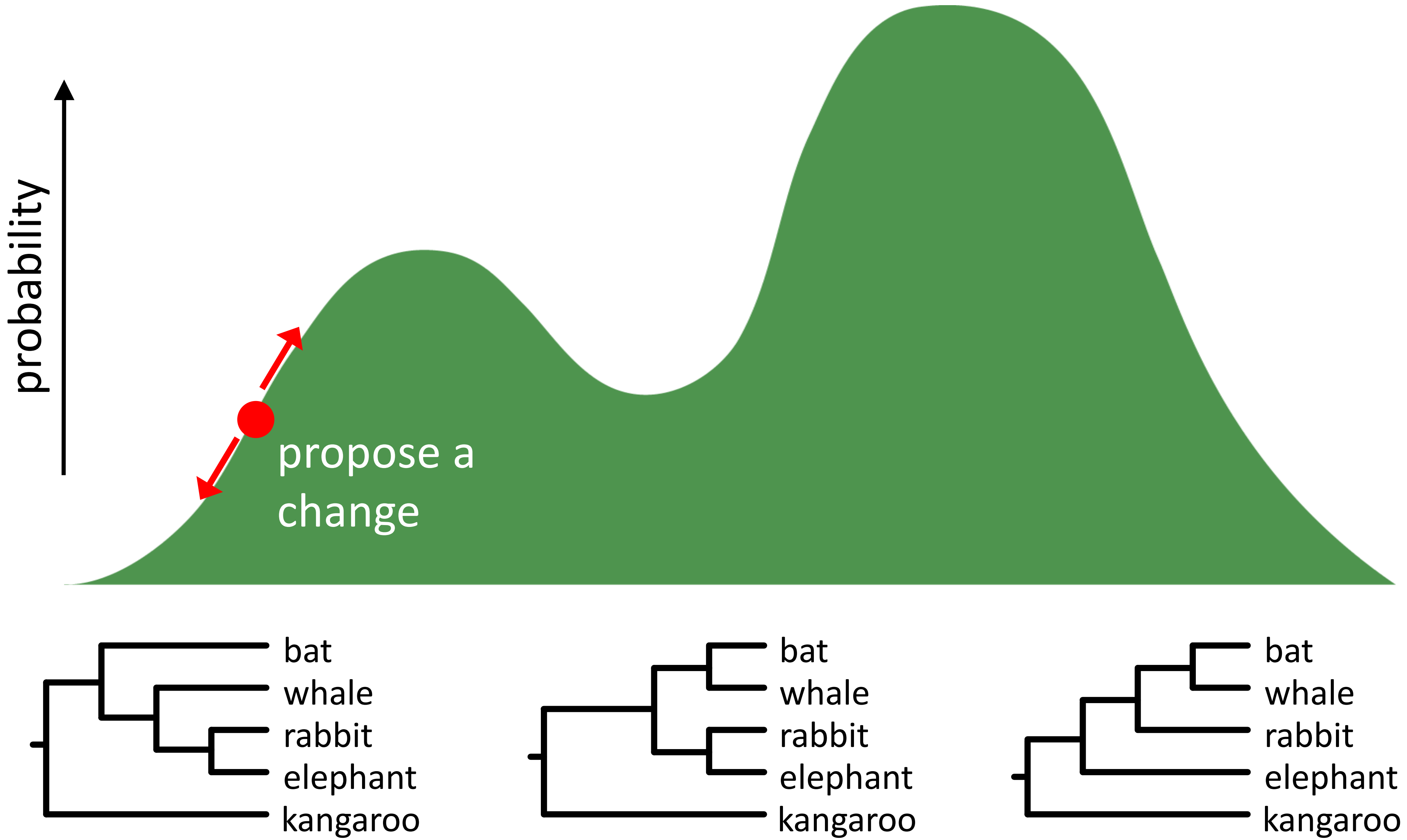


Nicholas Metropolis
Los Alamos, 1953

MCMC simulation

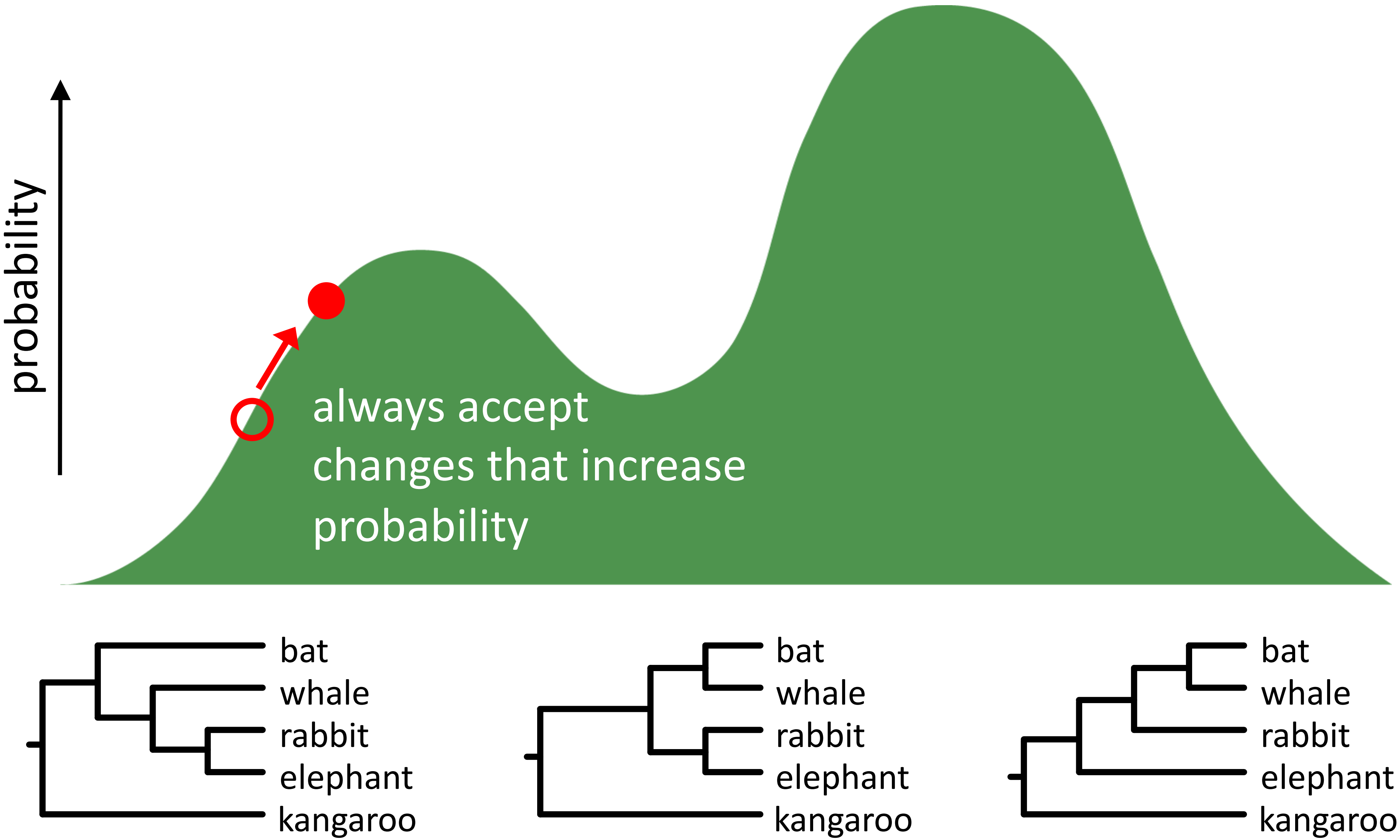


MCMC simulation



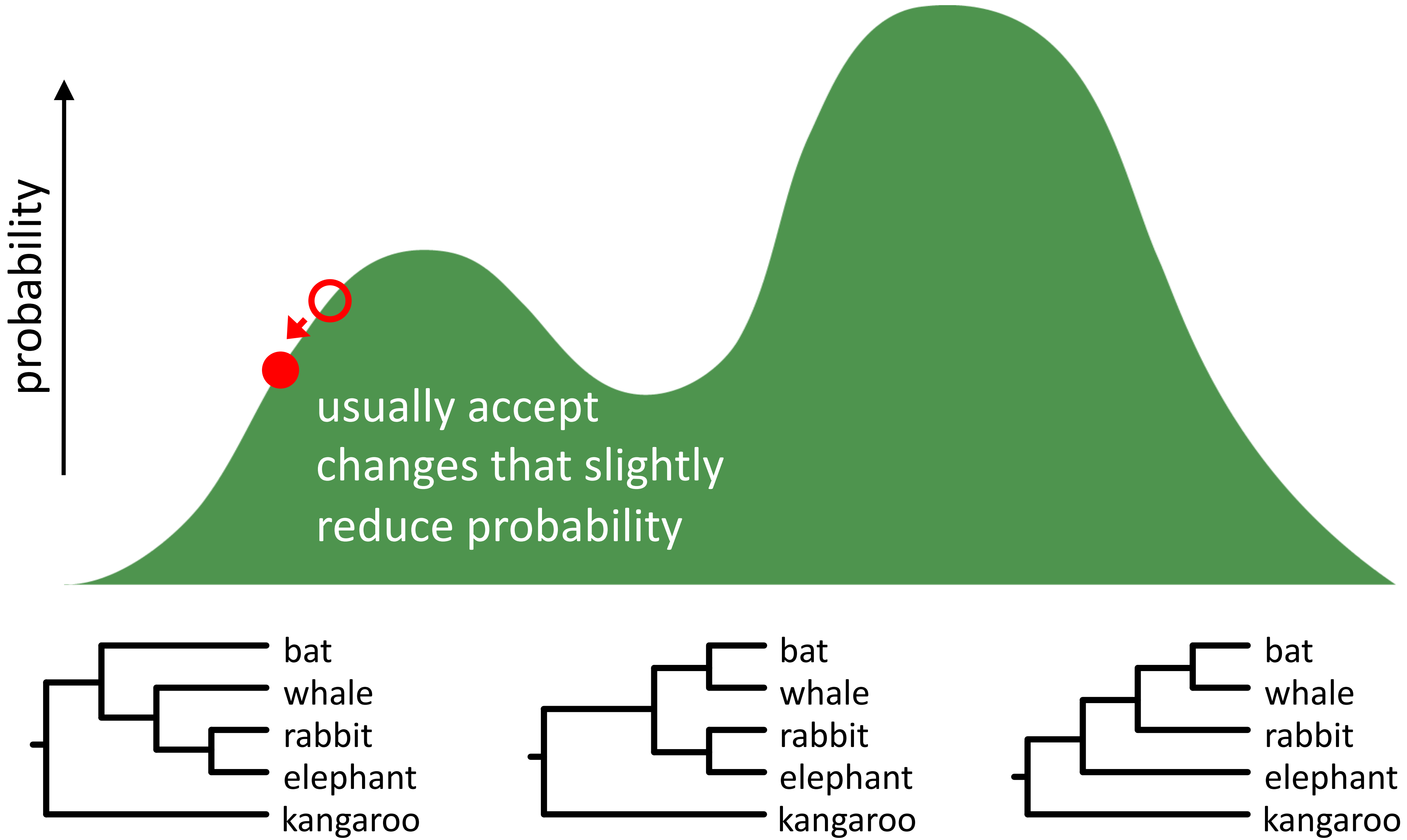
propose a change

MCMC simulation



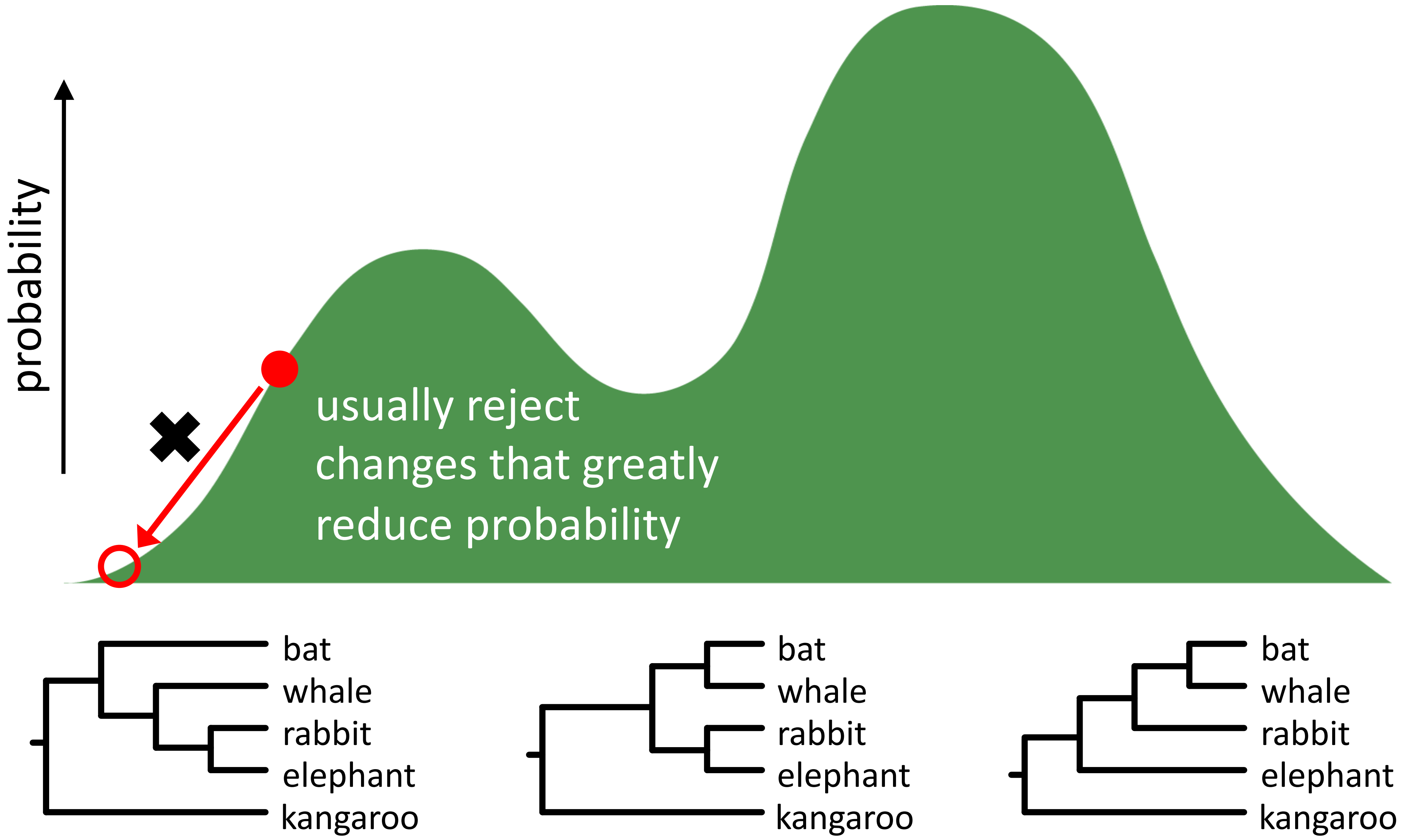
always accept changes that increase probability

MCMC simulation



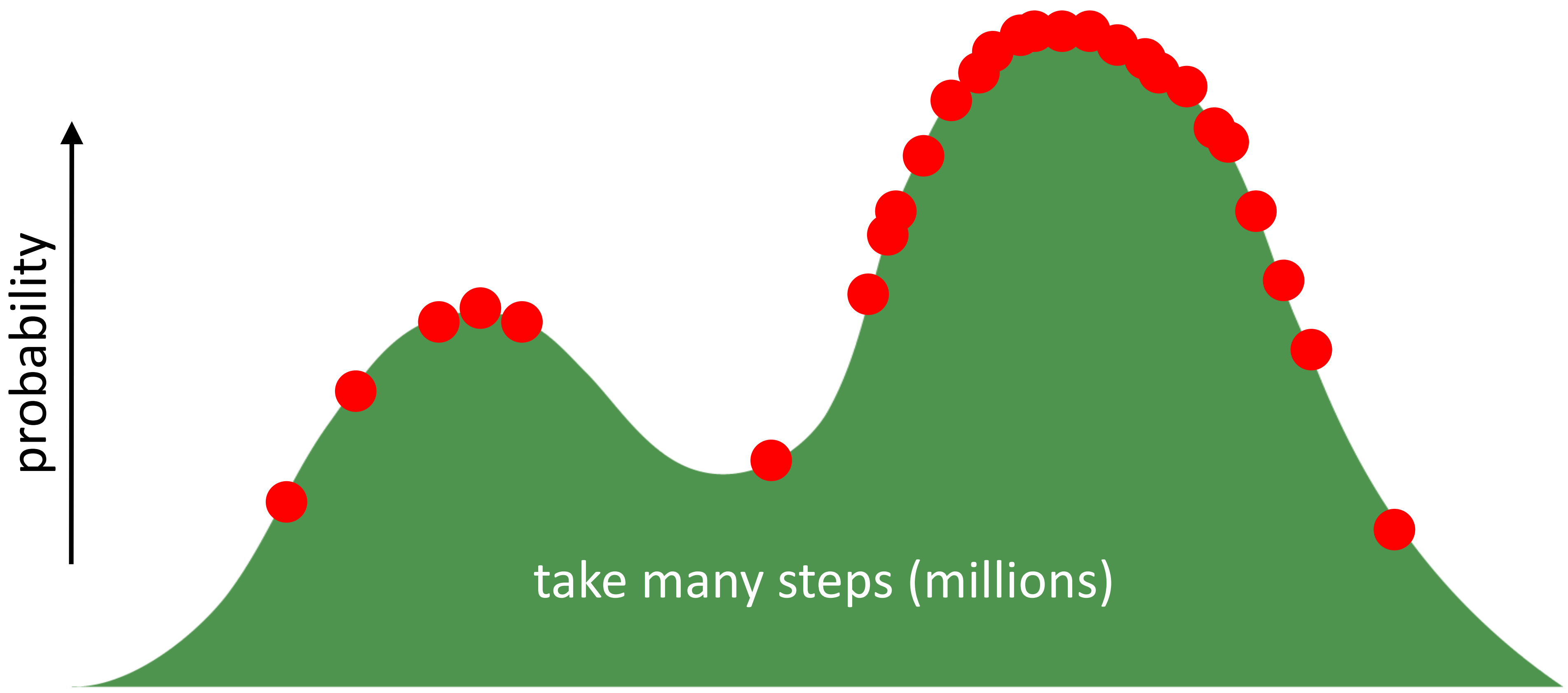
usually accept
changes that slightly
reduce probability

MCMC simulation

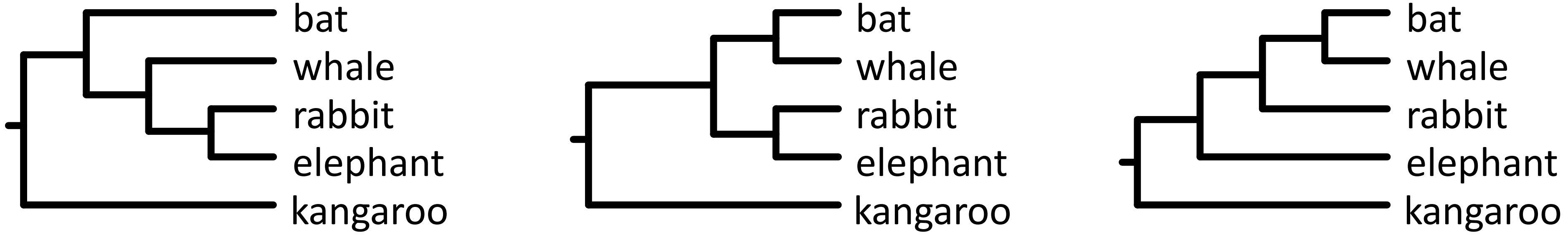


usually reject
changes that greatly
reduce probability

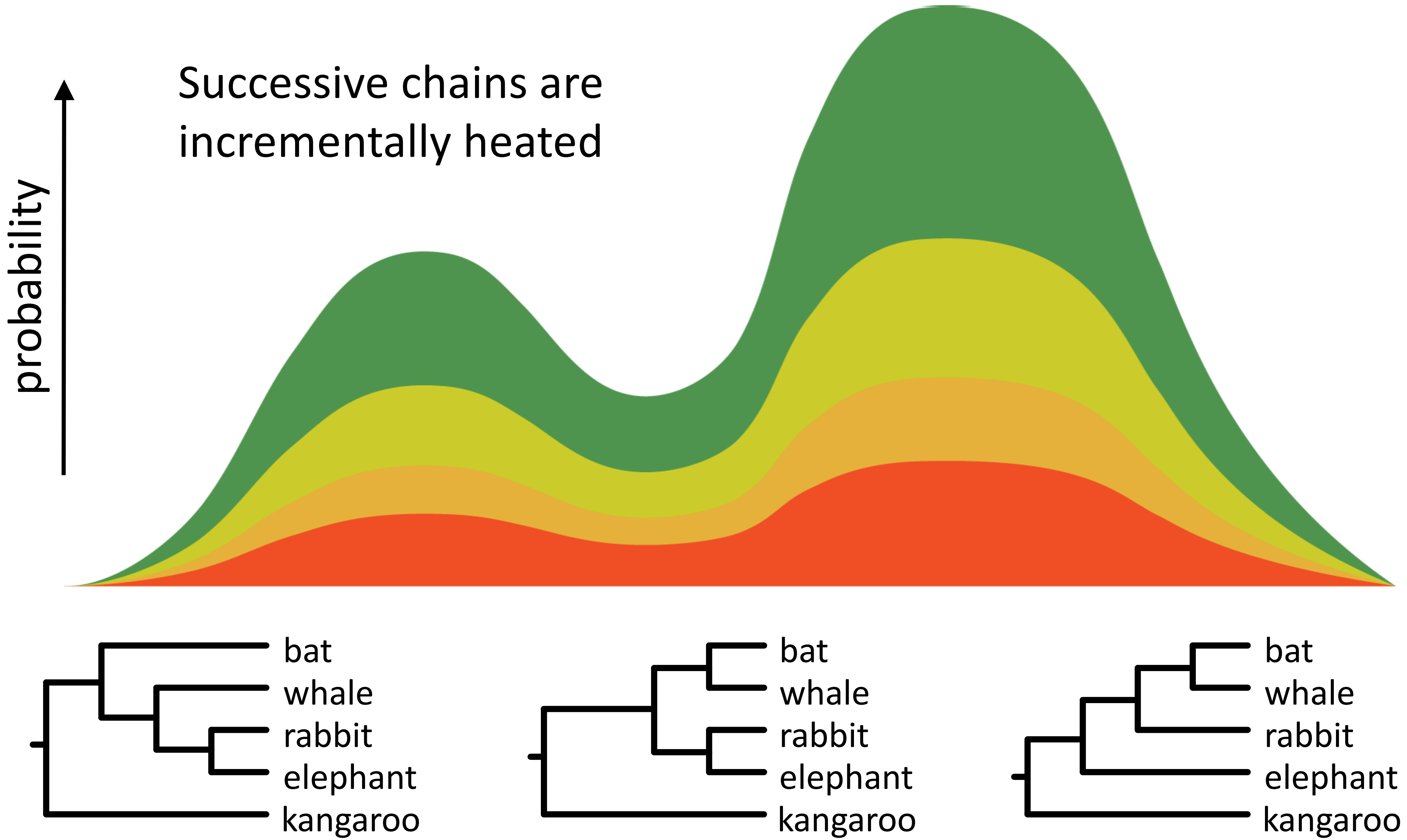
MCMC simulation



take many steps (millions)



Metropolis-coupled MCMC

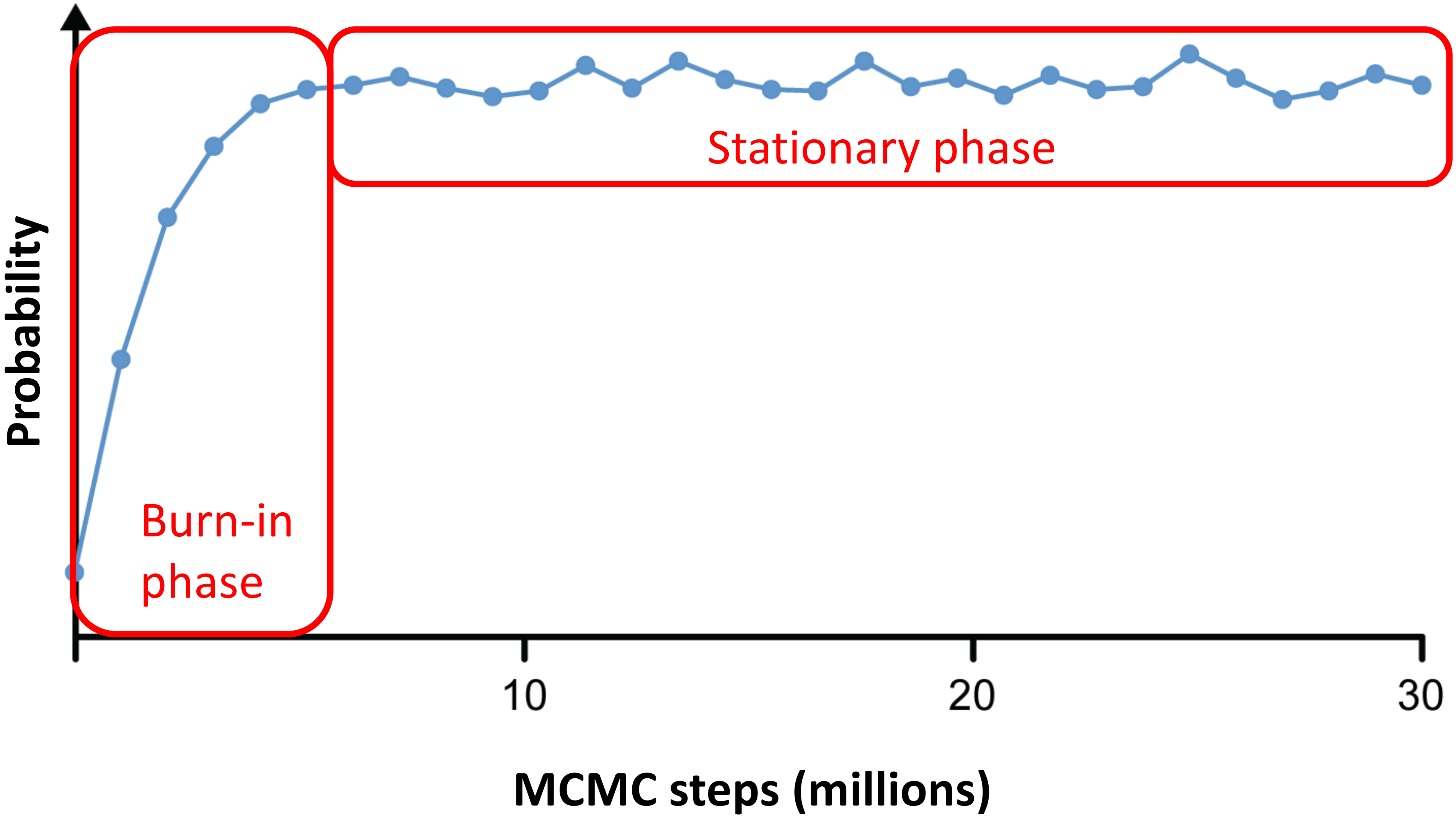


Successive chains are incrementally heated

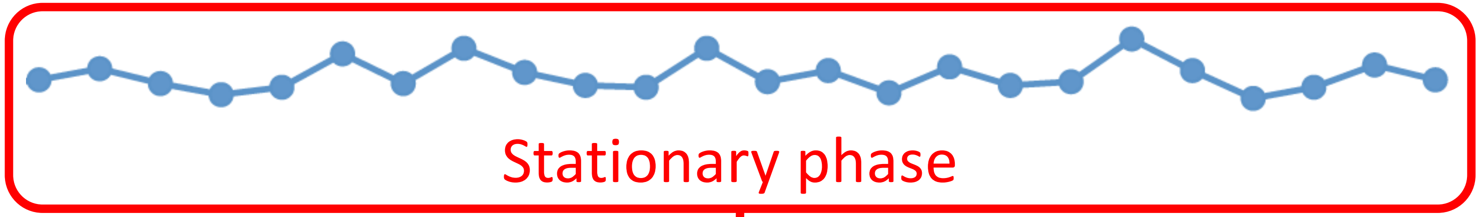
Samples from the MCMC

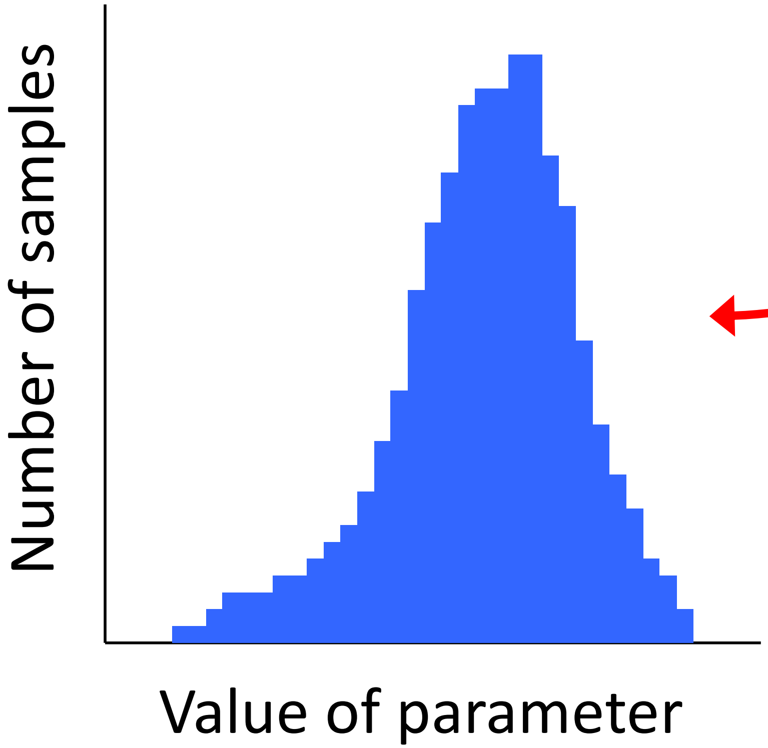
- Output from a Bayesian phylogenetic analysis:
  - A list of the **parameter values** visited by the Markov chain (.p file in *MrBayes*, .log file in *BEAST)*
  - A list of the **trees** visited by the Markov chain (.t file in *MrBayes*, .trees file in *BEAST*)

Samples from the MCMC



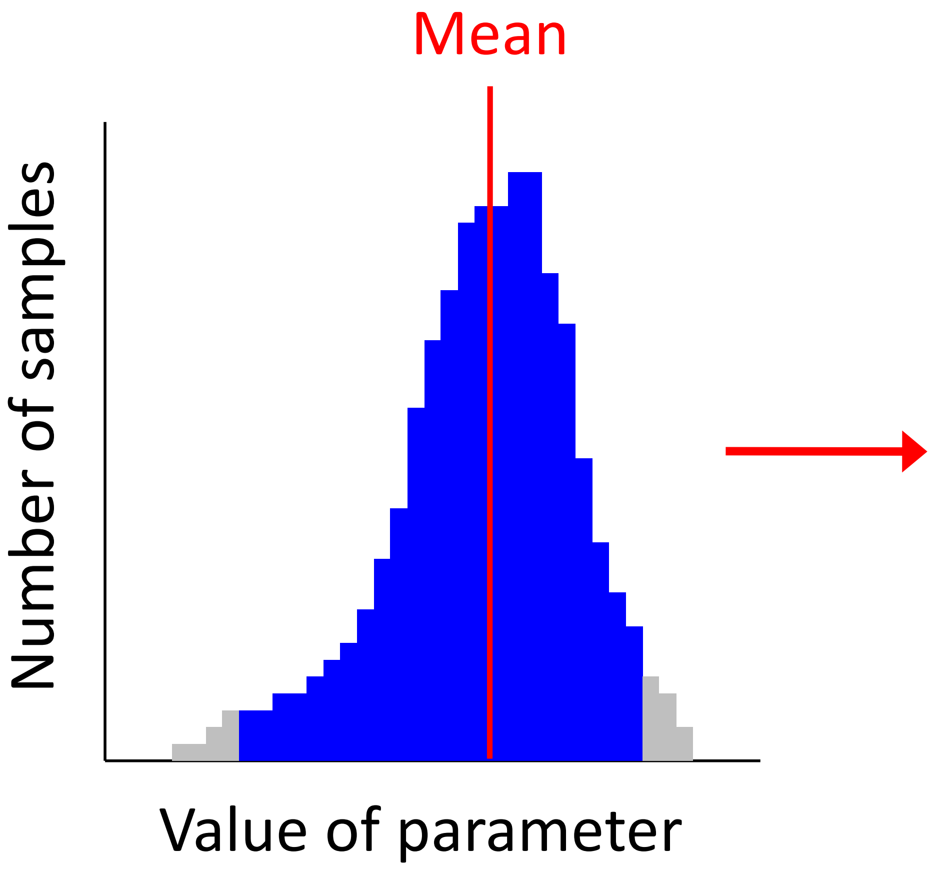
Samples from the MCMC





Using the Metropolis-Hastings algorithm, we sample parameter values and trees from the MCMC with a frequency proportional to their **posterior probability**

Samples from the MCMC



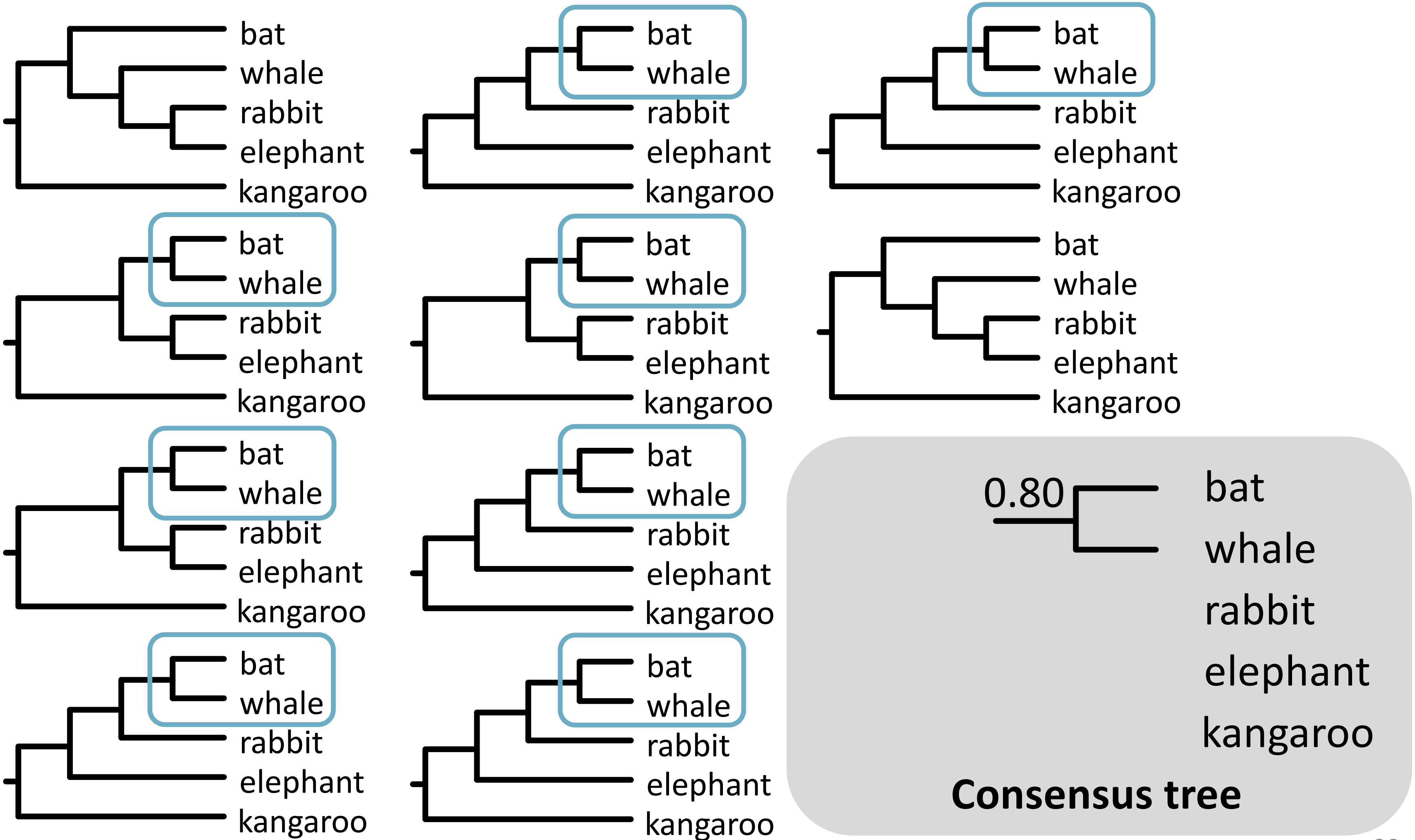
- Take the mean of the sampled values

  Mean posterior estimate

- Take the 'central' 95% of the sampled values

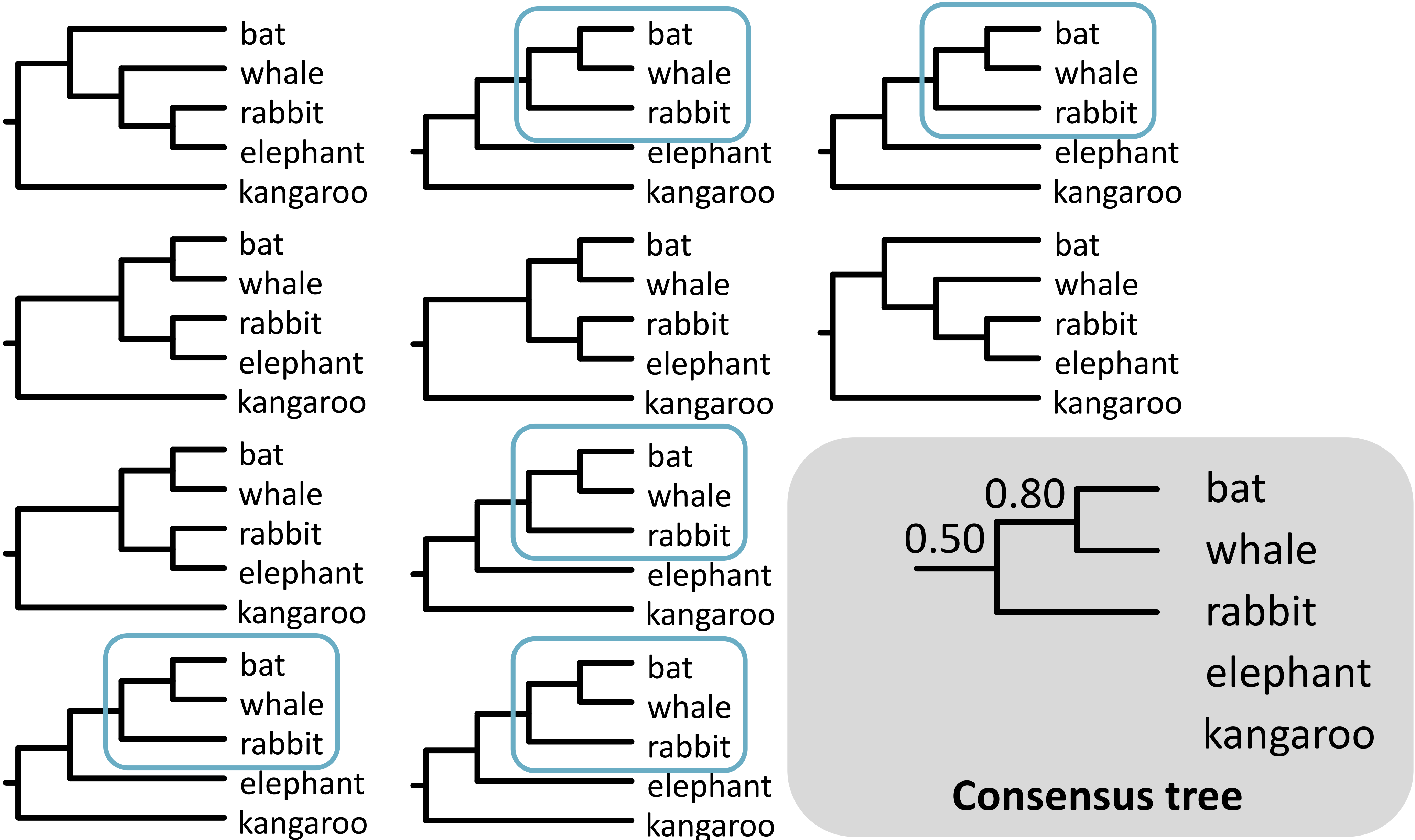
  95% credibility interval

Samples from the MCMC



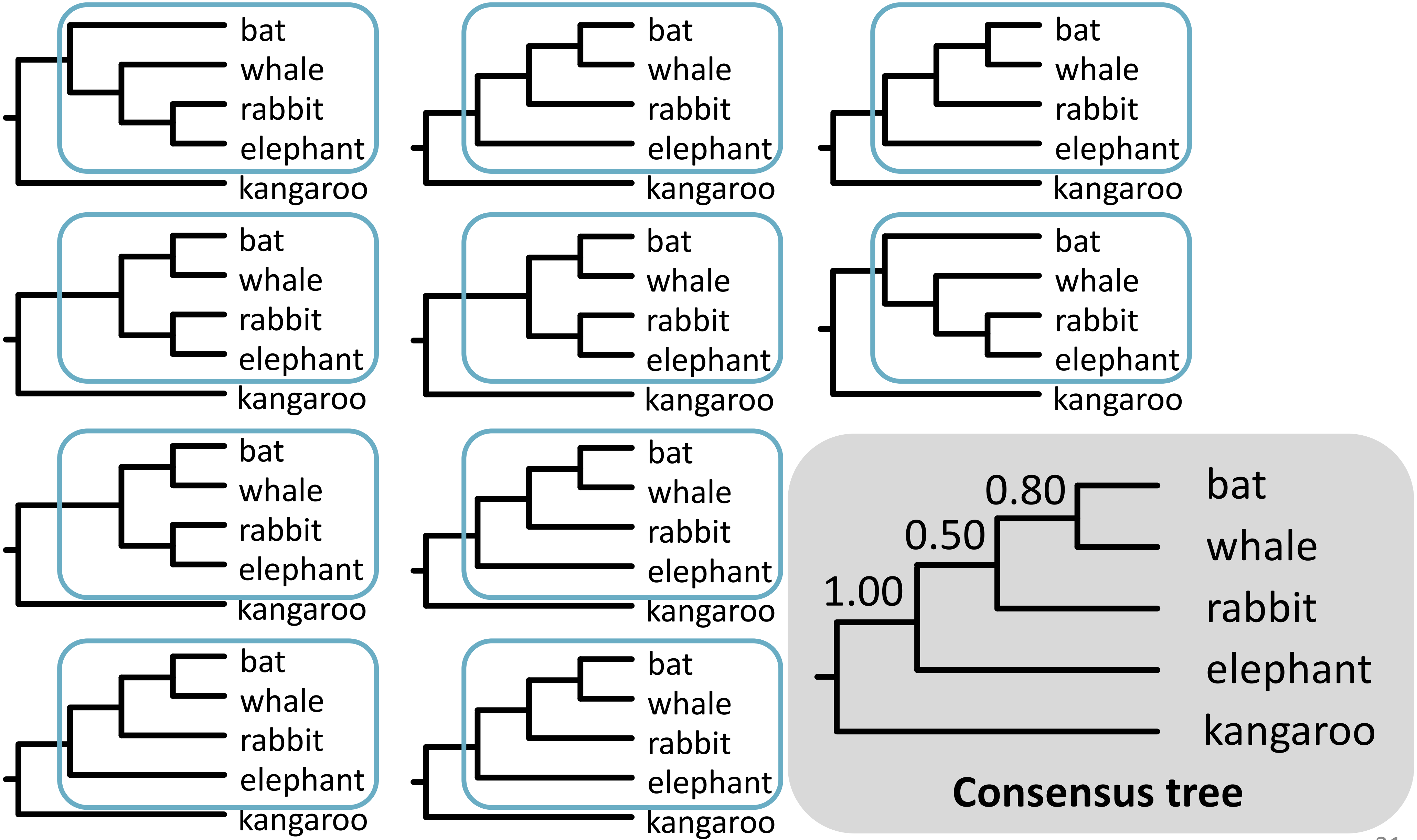
Consensus tree

Samples from the MCMC



Consensus tree

Bootstrapping



Consensus tree

Samples from the MCMC

- **Majority-rule consensus tree (*MrBayes*)**
  Shows all nodes with posterior probability >0.50

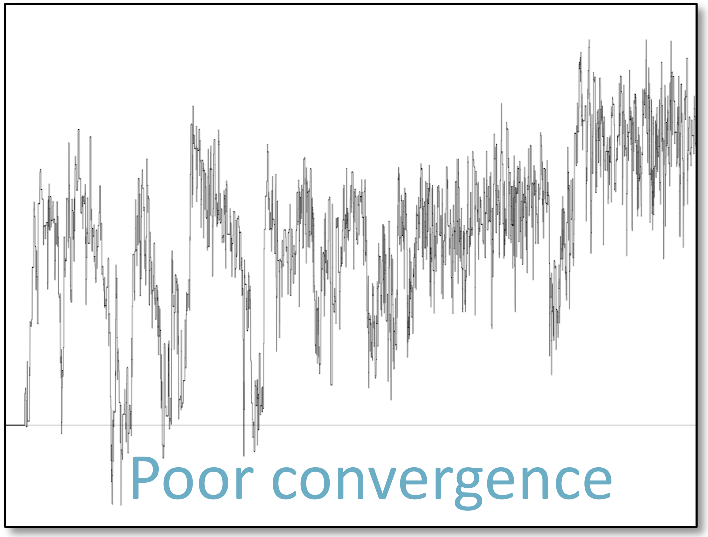
- **Maximum a posteriori (MAP) tree**
  Sampled tree with highest posterior probability

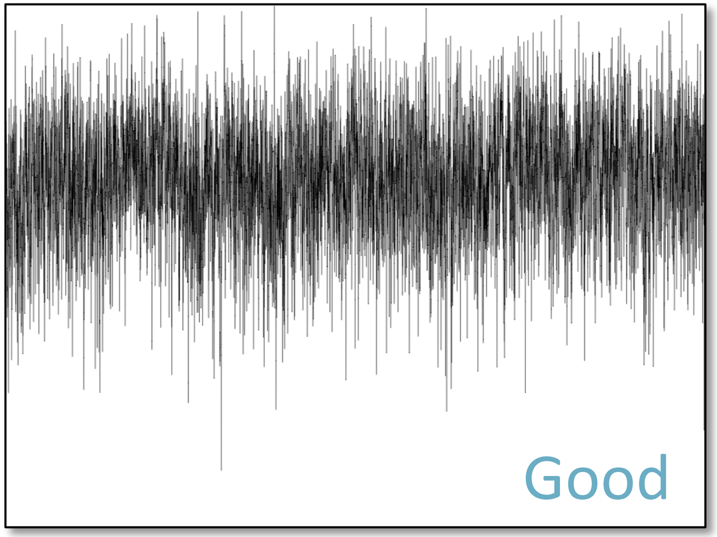
- **Maximum clade credibility (MCC) tree (*BEAST/TreeAnnotator*)**
  Sampled tree with highest sum or product of posterior node probabilities

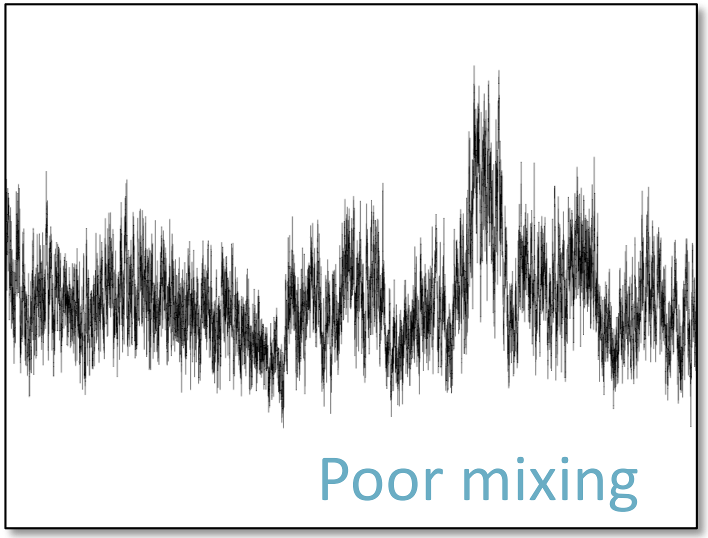
Diagnostics

1. **Convergence**
   Are we drawing samples from the stationary distribution?

2. **Sufficient sampling**
   Have we drawn enough samples to allow a reliable estimate of the posterior distribution?







Convergence

- Run at least 2 independent chains
- Likelihoods should be similar
- Estimates of model parameters should be similar

Sufficient sampling

- **Effective sample size (ESS)**
  Have we drawn enough independent samples to produce a reliable estimate of the posterior distribution?
- ESS is preferably **>200** for each parameter
- ESS can be increased by:
  - Increasing the length of the MCMC
    (and decreasing the frequency of sampling)
  - Modifying the MCMC proposals

Advantages and Problems

Advantages

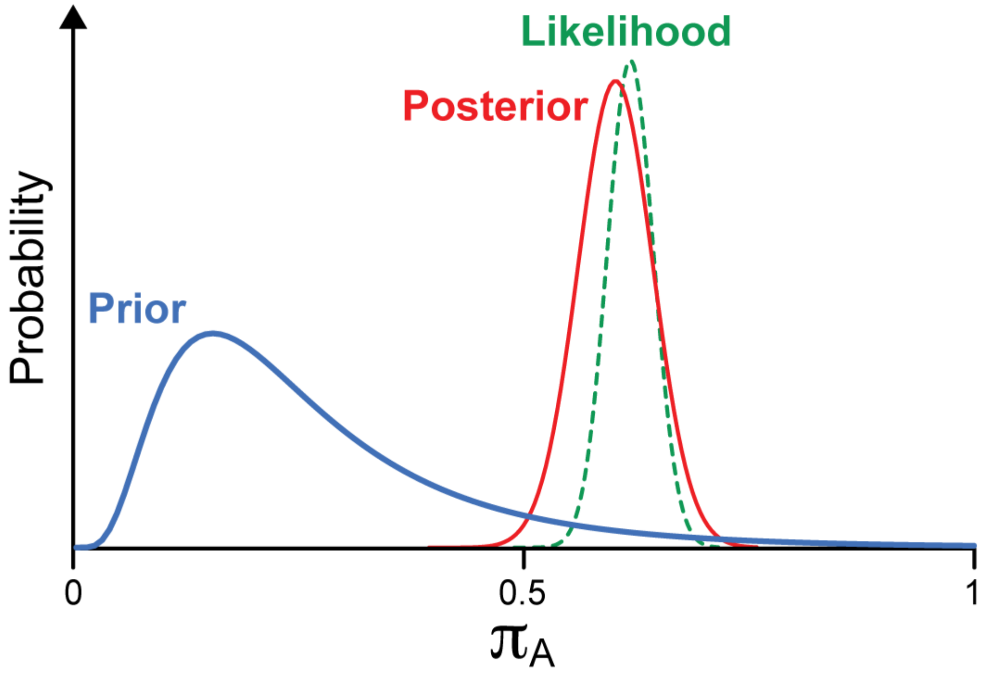
- Able to **implement highly parameterised models**
- **Estimating tree uncertainty** is straightforward
- **Posterior probabilities** have an intuitive interpretation
- Can incorporate **independent information** (in the prior)

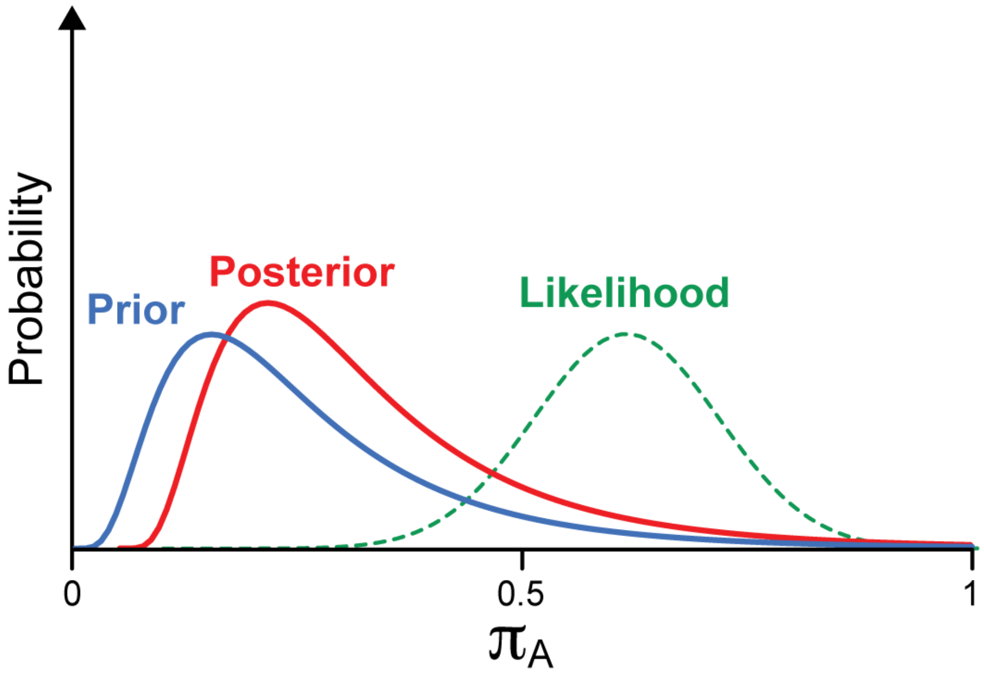
Nuisance parameters

- Integrate over ‘nuisance’ parameters
- Marginal distribution of a parameter of interest

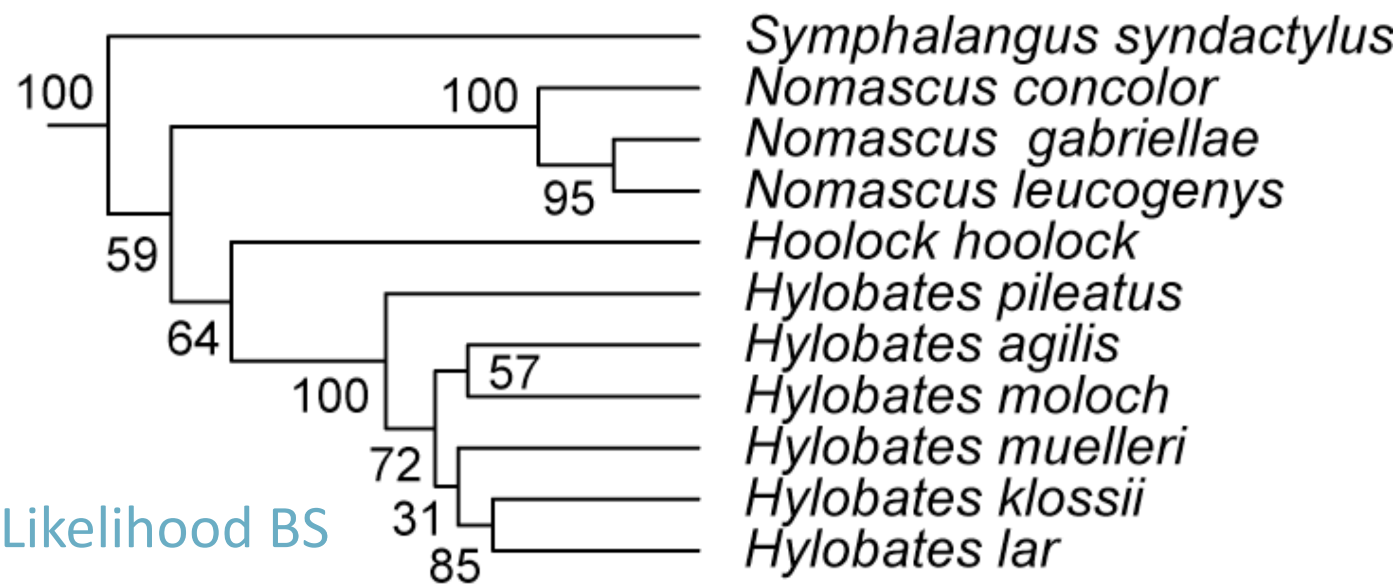
Influence of priors

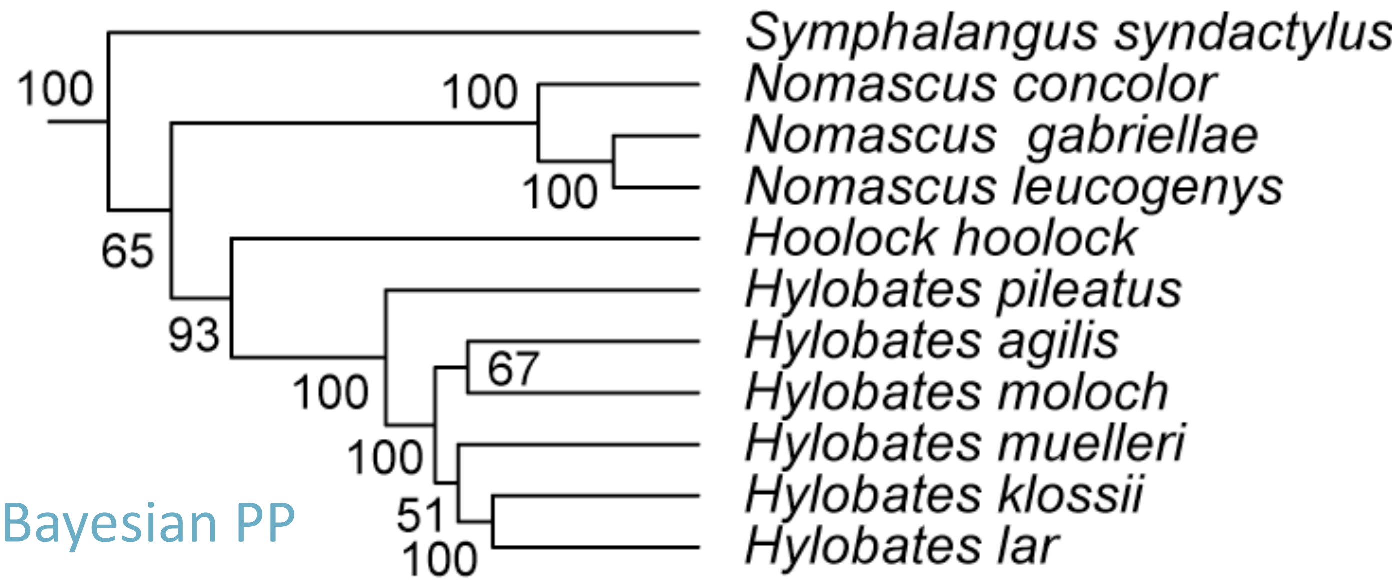
- Sensitivity of the posterior to the prior
- This problem can occur if the data are uninformative, the prior is strong, or both





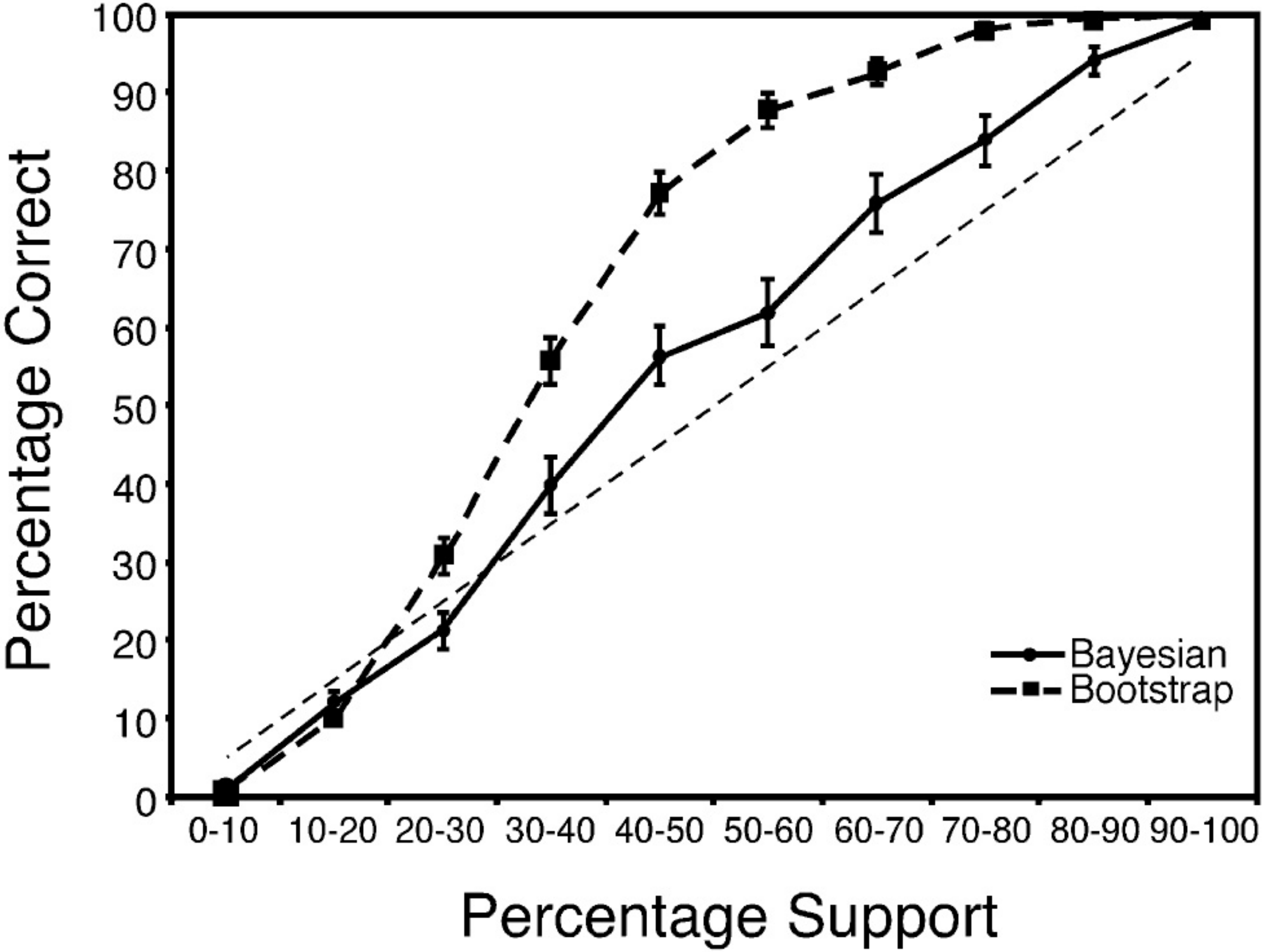
Node support





Chatterjee *et al*. (2009) *BMC Evol Biol*

Node support



Wilcox *et al*. (2002) *Mol Phylogenet Evol*

BEAST 1

- **B**ayesian **E**volutionary **A**nalysis by **S**ampling **T**rees
- Analyse population- or species-level data
- Simultaneous estimation of tree and node times
- Range of clock models
- Range of tree priors and demographic models

Beast2

Bayesian evolutionary analysis by sampling trees

- Re-write of *BEAST* to increase modularity
- Users can extend *BEAST* by adding packages
- Additional tree priors not available in *BEAST* 1
- Capacity to perform simulations

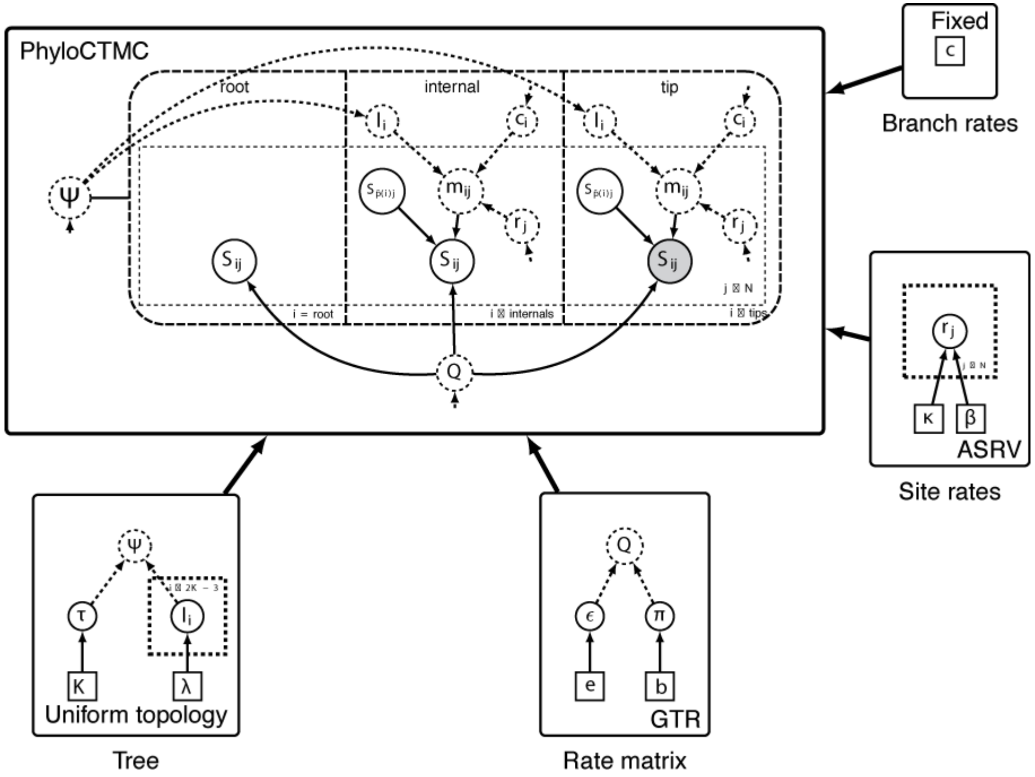
For a comparison of *BEAST* 1 and 2:
www.beast2.org/beast-features

MrBayes

- Primarily designed for species-level data
- Simultaneous estimation of tree and node times
- Range of clock models
- Range of tree priors
- Multiple chains and MCMC diagnostics

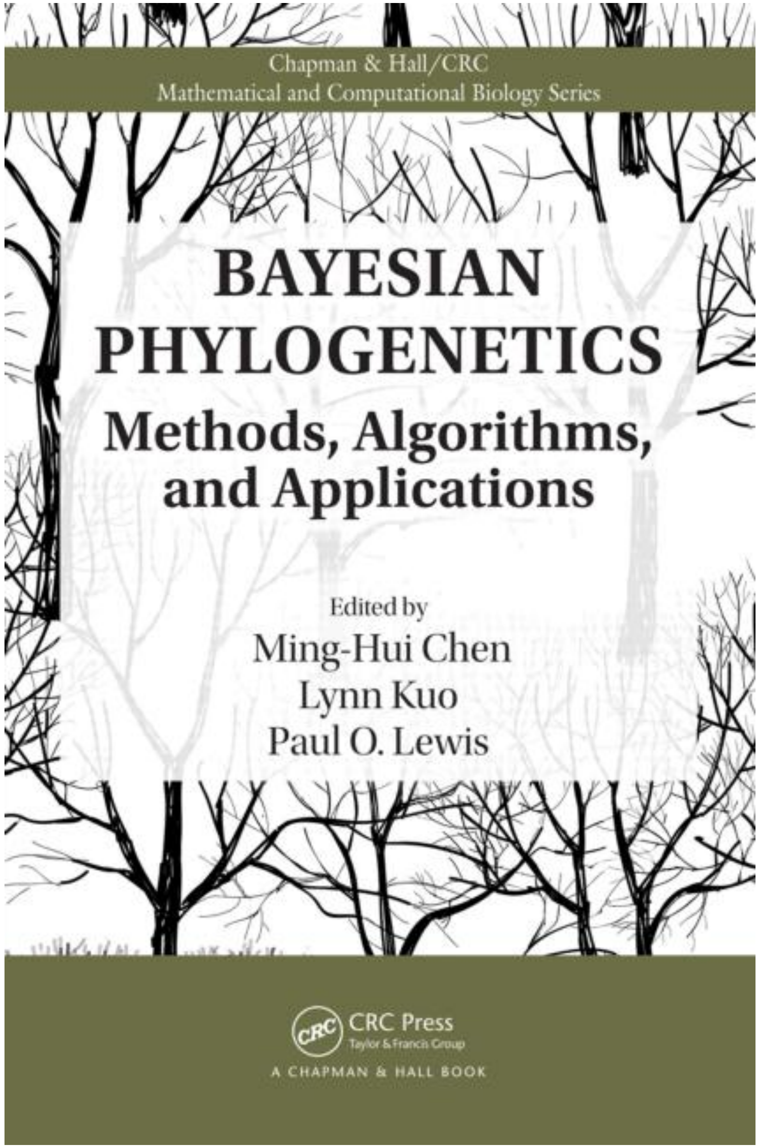
RevBayes

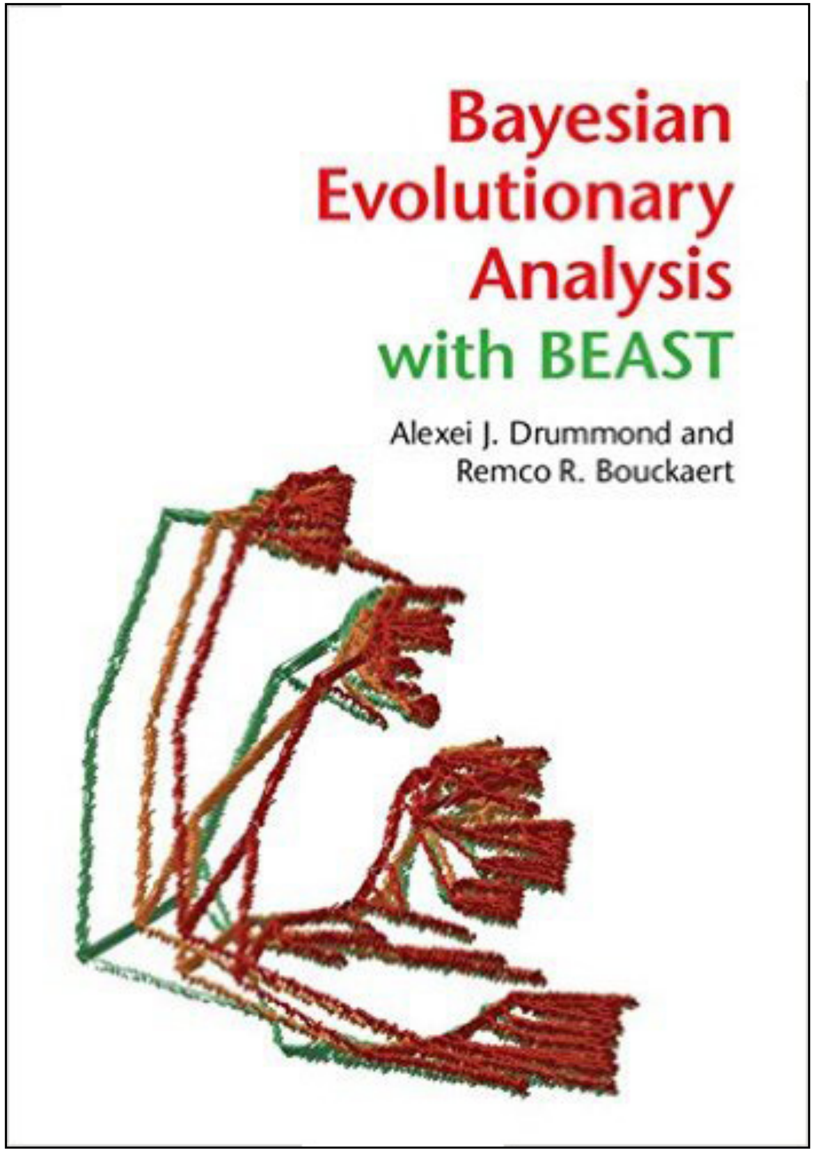
- Uses its own R-like language, Rev
- Interactive construction of graphical model
- Flexible and can be used for simulation and inference



- Analyses of large data sets on computing clusters
- Available priors similar to those in older versions of *MrBayes*
- Limited options, no molecular dating
- Likelihood component adapted from *RAxML*

Useful references


Chapman & Hall/CRC
Mathematical and Computational Biology Series
BAYESIAN PHYLOGENETICS
Methods, Algorithms, and Applications
Edited by
Ming-Hui Chen
Lynn Kuo
Paul O. Lewis
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK


Bayesian Evolutionary Analysis with BEAST
Alexei J. Drummond and
Remco R. Bouckaert